MACHINE LEARNING BASED PREDICTION OF NONCENTROSYMMETRIC CRYSTAL MATERIALS

Yuqi Song, Joseph Lindsay, Yong Zhao,Alireza Nasiri,Steph-Yves Louis
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC, 29201

Jie Ling
Department of Chemistry and Biochemistry
Claflin University
Orangeburg, SC, 29115

Ming Hu
Department of Mechanical Engineering
University of South Carolina
Columbia, SC, 29201

Jianjun Hu *
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC, 29201
* Correspondence author: `jianjunh@cse.sc.edu`

April 14, 2020

ABSTRACT

Noncentrosymmetric materials play a critical role in many important applications such as laser technology, communication systems,quantum computing, cybersecurity, and etc. However, the experimental discovery of new noncentrosymmetric materials is extremely difficult. Here we present a machine learning model that could predict whether the composition of a potential crystalline structure would be centrosymmetric or not. By evaluating a diverse set of composition features calculated using matminer featurizer package coupled with different machine learning algorithms, we find that Random Forest Classifiers give the best performance for noncentrosymmetric material prediction, reaching an accuracy of 84.8% when evaluated with 10 fold cross-validation on the dataset with 82,506 samples extracted from Materials Project. A random forest model trained with materials with only 3 elements gives even higher accuracy of 86.9%. We apply our ML model to screen potential noncentrosymmetric materials from 2,000,000 hypothetical materials generated by our inverse design engine and report the top 20 candidate noncentrosymmetric materials with 2 to 4 elements and top 20 borate candidates.

1 Introduction

Nonlinear optical materials (NLO), in which light waves interact with each other, are one of the key enablers for next generation of new lasers, fast telecommunication, quantum computing, quantum encryption, dynamic or optical storage data, and many other applications [1, 2, 3, 4]. NLO materials are most broadly defined as those compounds capable of altering the frequency of light. Depending on the chemical and physical construct of the materials they can combine multiple photons to generate shorter wavelength photons or split one photon into several new photons of longer wavelengths. These new photons can be employed to perform all of the above applications as well as many others. The classes of NLO materials range broadly from inorganic oxides such as $KTiOPO_4$ and $LiNbO_3$ to semiconductors like to periodically poled GaAs, to organic polymers to metal organic framework (MOFs), and to simple small organic molecules like stilbene. This broad range of materials has many different properties and characteristics but all are united

by one common factor, i.e. their lattice structure must not contain a center of symmetry and must be acentric [1, 2]. This is a rigorous requirement that can only be met in well-ordered lattice structures, meaning ordered crystals. It is generally difficult to design and grow acentric single crystals and less than 15% of all known structures are acentric. This demands exceptional determination on the part of the synthetic and crystal growth experimentalists. The process is made even more difficult by the fact that the NLO processes that enable frequency modification are inherently inefficient. Moreover, the ability to prepare new NLO materials and study their properties is not trivial and requires patient and detailed investigations. The payoff is enormous however, as the materials enable the development of devices used in next generation laser surgery, imaging, optical communication, advanced spectroscopy, optical data storage and a vast array of applications dependent on the interaction of light with matter. In Figure 1, We show the crystal structures of a centrosymmetric material and a noncentrosymmetric material, namely ScBO3 and SrB12O7.
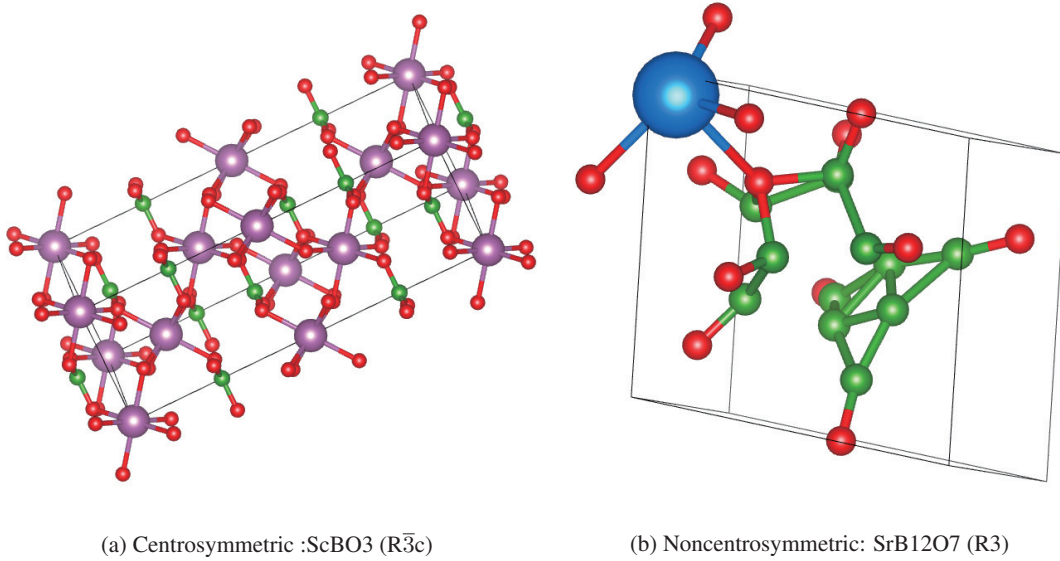

(a) Centrosymmetric :ScBO3 (R$\bar{3}$c)   (b) Noncentrosymmetric: SrB12O7 (R3)

Figure 1: Crystal Structures of centrosymmetric and noncentrosymmetric materials. (a) The crystal structures of ScBO3 of space group R$\bar{3}$c, where the purple nodes represent Sc atoms, the green nodes represent B atoms and red nodes are O atoms. (b) The crystal structure of SrB12O7 of space group R3, where the blue node represents Sr atom, the green nodes represent B atoms, and the red nodes are O atoms.

Although the structure-property relation between NLO effects and microstructure can be used as a guide, new NLO crystals are still mainly explored using "trial and error" Edisonian approaches. A reliable determination of lattice symmetry is a crucial first step for materials characterization and analytics. Recently, a deep learning-based approach to automatically classify structures given a crystal structure (even with defects) has been recently proposed [5]. Similarly, Kaufmann et. al. [6] proposed a crystal symmetry determination method from electron diffraction using machine learning. However, these methods cannot be applied for large-scale composition based screening as they both require experimental data. On the other hand, direct numerical calculation of the optical properties of a single crystalline material from its atomic structure by accurate first-principles without any other inputs has just been made available for a few years. Studies have focused on properties such as second harmonic generation (SHG) coefficients [7] and other important optical properties such as energy band gap, refractive indices [8], and birefringence. While first-principles calculations make it possible to predict some optical properties without any experimental data, such computation is usually tedious and very computationally demanding even for not too complicated primitive cells. Consider this: four-element compounds with different ratios can lead to a search space of 32.4 billion combinations. Currently, Density Functional Theory (DFT) based first-principles methods for optical properties calculation is out of the question for high-throughput screening of NLO materials. Especially, these methods cannot be used for discovery of new NLO materials as they all require the knowledge of the crystal structure information which is usually not available and computational prediction of crystal structures from composition is feasible only for a small subset of materials with simple compositions [9]. In-depth understandings of the mechanism on how compositions form specific structures which further determines the NLO behavior would provide the guide for experimental explorations, and save enormous human and materials resources. On the other hand, data driven computational prediction models for noncentrosymmetric materials discovery can be used as the first step for nonlinear optical materials discovery.

In the past five years, machine learning (ML) has been increasingly applied to materials informatics problems from property prediction [10, 11], to materials structure prediction, to computational screening [12], and inverse materials design [13, 14]. Among these ML algorithms and models, Random Forest (RF) models have shown great success for predicting a variety of materials properties such as the critical temperatures of superconducting materials [15, 16] and for predicting the ability of a given composition to form an amorphous ribbon of metallic glass via melt spinning [17, 18]. In [19], Furmanchuk et al. utilized a RF regression model to predict the bulk modulus. RF models have also been widely used in other research areas. For example, a RF based approach showed its superiority in automatically selecting molecular descriptors for ligands of kinases and nuclear hormone receptors [20]. On the other hand, recent years have observed tremendous success of deep learning [21] based neural network models in applications such as image recognition, automatic machine translation, robotics [22], and autonomous driving [23]. More importantly, their success in materials discovery problems such as the prediction of crystal stability [24] and superconductor critical temperatures [25], makes it promising for other applications in materials discovery. In our previous work, we have applied machine learning and deep learning for crystal space group and crystal system prediction from composition [26] and for formation energy prediction [10].

Herein, we propose and evaluate two machine learning models including RF and multi-layer perceptron (MLP) neural network models for noncentrosymmetric classification given only material composition. The Magpie composition descriptors are used in our study. Cross-validation and hold-out experiments show that RF with Magpie features achieved the best results. A further application of our RF noncentrosymmetric prediction model to screening two million hypothetical materials generated by our generative ML model [14] allows us to identify and predict dozens of potential novel noncentrosymmetric materials with high confidence scores.

Our contributions can be summarized as follows:

(1) We propose two machine learning algorithms (RF and MLP) for predicting noncentrosymmetric materials given only their composition.

(2) We evaluate and compare the performances of different machine learning algorithms for noncentrosymmetric materials classification.

(3) We apply our prediction models to screen the 2 million hypothetical materials generated by a generative adversarial network (GAN) based predictors and identify a list of top candidate materials with highly probable noncentrosymmetry structures.

2 Materials and Methods

2.1 Feature Calculation

To accomplish the goal of noncentrosymmetry classification, one of the key steps is to identify the most relevant features of a chemical composition that correlates with symmetry tendency of its formed structure. To do this, we have tried the myriad of featurizers provided by the matminer library [27], which is a Python-based software platform to facilitate data-driven methods of analyzing and predicting materials properties, such as composition, crystal structure, band structure, and more. The matminer featurizers package has a total of 5 different classes of featurizers present in the library's current deployment, ranging from composition descriptors to structural ones.

We use the composition featurizer's Element Property module to calculate the Magpie elemental descriptors for training our ML models. The Magpie feature set has 132 elemental descriptors [18], composed of 6 statistics of a set of elemental properties such as atomic number in the material, space group of the material, the Magnetic Moment calculated by Density Functional Theory (DFT). Magpie feature set was selected based on our evaluations of a couple of descriptors.

2.2 Machine learning models

We evaluate two machine learning models for noncentrosymmetry prediction, namely, a Random Forest (RF) classifier, and a Deep Neural Network (DNN) classifier.

Random Forest [28, 29] is a supervised learning method that can be applied to solve classification or regression problems. It is an ensemble algorithm that constructs a multitude of many decision trees at training time and outputs the class that is the mode of the classification of the individual trees. RF classifiers have shown strong prediction performance when combined with composition features in our previous studies [30]. In our RF classifier model, we set the number of trees to be 200. This algorithm was implemented using the Scikit-Learn library in Python 3.6.

Deep learning excels at identifying patterns in unstructured data by building multiple layers to progressively extract higher-level features from the raw input to do the predictive task [31]. For instance, Xie et al. [32] proposed a graph

convolutional neural network model for property predictions of materials and provided a universal and interpretable representation of crystalline materials. In this paper, we aimed to explore whether DNNs can achieve better predictive performance than RF models in noncentrosymmetry prediction. Therefore, we designed a MLP neural network classifier made of five fully connected layers, with four layers using LeakyReLU as their activation function and Sigmoid in the final layer for classifying. A dropout layer with a 0.05 drop rate was added to avoid overfitting. An Adam optimizer and binary cross entropy function are selected for training the DNN. In addition, the epoch, batch and learning rate are set to 50, 500, 0.001, respectively.

2.3 Hyper-parameter tuning

Due to various hyperparameters and the impact of their combinations on the training process and the final performance of machine learning models, manual parameter tuning is time-consuming. Hence, automatic hyperparameter tuning method is needed for finding suitable parameters. To ensure fair comparison of the ML models, we use the Bayesian optimization [33] algorithm to find optimal hyper-parameters for RF models, which has been proven to be an effective tool. This method requires that the objective be a scalar value depending on the hyperparamter configuration x, where the maximum is sought for an expensive function $f : \mathcal{X} \to \mathbb{R}$.

$$\mathbf{x}_{opt} = \underset{\mathbf{x} \in \mathcal{X}}{\arg \max} f(\mathbf{x}) \tag{1}$$

We use the hyperopt package library [34] to optimize n_estimators, max_depth and max_features in RF models by supplying an optimization function which maximizes its precision.

3 Results and Discussion

Herein, we describe the datasets, the evaluation criteria, and the experimental results. We analyze and compare the prediction performance of RF and DNN models. Besides, we discuss the application of our model to screening new hypothetical noncentrosymmetric materials. Our experiments on classifying noncentrosymmetry from composition include three parts: cross-validation experiments, holdout experiments on Borates, and screening a two million hypothetical materials.

3.1 Datasets

Crystal structures with different space groups have different centrosymmetric tendencies. It is known that there are 138 noncentrosymmetric space groups and 92 centrosymmetric space groups, the detailed space group IDs and names and their centrosymmetric property are summarized in Table 1.

4

Table 1: Space groups with noncentrosymmetric and centrosymmetric structures

	group IDs	group names
centrosymmetric	2, 10-15, 47-74, 83-88, 123-142, 147-148, 162-167, 175-176, 191-194, 200-206, 221-230	$P\bar{1}$, P2/m, $P2_1$/m, C2/m, P2/c, $P2_1$/c, C2/c, Pmmm, Pnnn, Pccm, Pban, Pmma, Pnna, Pmna, Pcca, Pbam, Pccn, Pbcm, Pnnm, Pmmn, Pbcn, Pbca, Pnma, Cmcm, Cmca, Cmmm, Cccm, Cmma, Ccca, Fmmm, Fddd, Immm, Ibam, Ibca, Imma, P4/m, $P4_2$/m, P4/n, $P4_2$/n, I4/m, $I4_1$/a, P4/mmm, P4/mcc, P4/nbm, P4/nnc, P4/mbm, P4/mnc, P4/nmm, P4/ncc, $P4_2$/mmc, $P4_2$/mcm, $P4_2$/nbc, $P4_2$/nnm, $P4_2$/mbc, $P4_2$/mnm, $P4_2$/nmc, $P4_2$/ncm, I4/mmm, I4/mcm, $I4_1$/amd, $I4_1$/acd, $P\bar{3}$, $R\bar{3}$, $P\bar{3}$1m, $P\bar{3}$1c, $P\bar{3}$m1, $P\bar{3}$c1, $R\bar{3}$m, $R\bar{3}$c, P6/m, $P6_3$/m, P6/mmm, P6/mcc, $P6_3$/mcm, $P6_3$/mmc, $Pm\bar{3}$, $Pn\bar{3}$, $Fm\bar{3}$, $Fd\bar{3}$, $Im\bar{3}$, $Pa\bar{3}$, $Ia\bar{3}$, $Pm\bar{3}$m, $Pn\bar{3}$n, $Pm\bar{3}$n, $Pn\bar{3}$m, $Fm\bar{3}$m, $Fm\bar{3}$c, $Fd\bar{3}$m, $Fd\bar{3}$c, $Im\bar{3}$m, $Ia\bar{3}$d
noncentrosymmetric	1, 3-9, 16-46, 75-82, 89-122, 143-146, 149-161, 168-174, 177-190, 195-199, 207-220	P1, P2, $P2_1$, C2, Pm, Pc, Cm, Cc, P222, $P222_1$, $P2_12_12$, $P2_12_12_1$, $C222_1$, C222, F222, I222, $I2_12_12_1$, Pmm2, $Pmc2_1$, Pcc2, Pma2, $Pca2_1$, Pnc2, $Pmn2_1$, Pba2, $Pna2_1$, Pnn2, Cmm2, $Cmc2_1$, Ccc2, Amm2, Aem2, Ama2, Aea2, Fmm2, Fdd2, Imm2, Iba2, Ima2, P4, $P4_1$, $P4_2$, $P4_3$, I4, $I4_1$, $P\bar{4}$, $I\bar{4}$, P422, $P42_12$, $P4_122$, $P4_12_12$, $P4_222$, $P4_22_12$, $P4_322$, $P4_32_12$, I422, $I4_122$, P4mm, P4bm, $P4_2$cm, $P4_2$nm, P4cc, P4nc, $P4_2$mc, $P4_2$bc, I4mm, I4cm, $I4_1$md, $I4_1$cd, $P\bar{4}$2m, $P\bar{4}$2c, $P\bar{4}2_1$m, $P\bar{4}2_1$c, $P\bar{4}$m2, $P\bar{4}$c2, $P\bar{4}$b2, $P\bar{4}$n2, $I\bar{4}$m2, $I\bar{4}$c2, $I\bar{4}$2m, $I\bar{4}$2d, P3, $P3_1$, $P3_2$, R3, P312, P321, $P3_1$12, $P3_1$21, $P3_2$12, $P3_2$21, R32, P3m1, P31m, P3c1, P31c, R3m, R3c, P6, $P6_1$, $P6_5$, $P6_2$, $P6_4$, $P6_3$, $P\bar{6}$, P622, $P6_122$, $P6_522$, $P6_222$, $P6_422$, $P6_322$, P6mm, P6cc, $P6_3$cm, $P6_3$mc, $P\bar{6}$m2, $P\bar{6}$c2, $P\bar{6}$2m, $P\bar{6}$2c, P23, F23, I23, $P2_13$, $I2_13$, P432, $P4_232$, F432, $F4_132$, I432, $P4_332$, $P4_132$, $I4_132$, $P\bar{4}$3m, $F\bar{4}$3m, $I\bar{4}$3m, $P\bar{4}$3n, $F\bar{4}$3c, $I\bar{4}$3d

(a) Centrosymmetric space groups   (b) Non-centrosymmetric space groups
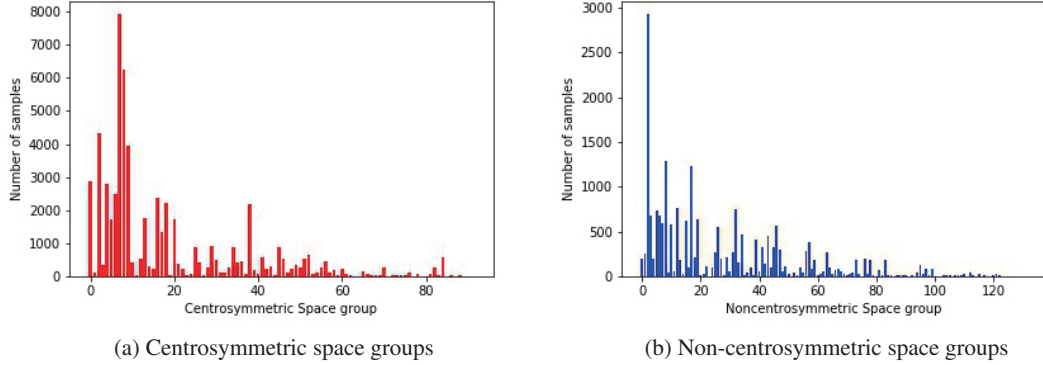
Figure 2: Sample distribution of noncentrosymmetric and centrosymmetric space groups in MPF dataset

We first downloaded the composition formulas of 97,217 crystal materials from the Materials Project database. We then remove those compositions belonging to multiple space groups with conflicting centrosymmetric tendencies. In total, we collecte 82,506 material compositions and assign the noncentrosymmetric property labels according to their corresponding space group. The dataset is called **MPF**, which have 60,587 positive (noncentrosymmetric) samples and 21,919 negative (centrosymmetric) samples, as shown in Table 2. The distribution of noncentrosymmetric and centrosymmetric space groups in MPF dataset are shown in Figure 2. We find that the distribution of samples over different space groups are not well balanced.

In order to evaluate the extrapolation prediction performance of our machine learning prediction model of noncentrosymmetry, we select all the 315 borate compounds from MPF dataset and assign them as the hold-out test dataset

Borates315. Borates contain boron (B) element and oxygen (O) element, which are a ubiquitous family of flame retardants found as boric acid and as a variety of salts. Previous research found that compared to other material family, borates tend to have higher percentage of nonlinear proprieties, which makes it a good hold-out test set. [35]. We further find that most borate materials include 3 elements. It is interesting to see if ML models trained with 3-element training samples can achieve better prediction performance. We select all 3-element materials from the MPF dataset and assigned them to the **MP3** dataset, which includes 30,762 centrosymmetric materials and 8,964 noncentrosymmetric materials as shown in Table 2. The motivation is to check if our classification models trained with MP3 dataset can achieve better performance when testing on the hold-out borates dataset.

Table 2: Dataset

	#symmetry	#non symmetry	#total
MPF	63,376	19,130	82,506
MP3	30,762	8,964	39,726
Borates315	250	65	315

3.2 Evaluation criteria

To evaluate the prediction performance of our model, precision, recall, accuracy, F1 score, and receiver operating characteristic area under the curve (ROC AUC) are used as performance metrics in this study.

The formula for these performance metrics are given as follows, where TP is number of true positives, FP is number of false positives, TN is number of true negatives, FN is number of false negatives, TPR is the true positive rate (also referred to as recall) of TP, and FPR refers to false positive rate of FP.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{5}$$

$$AUC = \int_{x=0}^{1} TPR(FPR^{-1}(x))dx \tag{6}$$

3.3 Prediction performance

To evaluate how our machine learning models can predict whether a crystal material's structure is noncentrosymmetry or not, we used two evaluation approaches: one is cross-validation over the MPF dataset and the other is the hold-out evaluation trained with non-borates datasets MPF and MP3 and tested on the Borates315 dataset. This hold-out test is especially important as the cross-validation performance can usually be over-estimated due to the redundancy of the training samples in most of the large-scale datasets such as the Materials Projects and the OQMD [36].

3.3.1 10-fold cross-validation performance

We set the maximum tree depth to be 20 and the number of decision trees as 200. This was later expanded to include the minimum number of samples per leaf node, the minimum number of samples required to split a node, and the maximum number of leaf nodes. With these 5 settings tuned per featurizer iteration, we then train the final prediciton RF models and make prediction, and caculate the performance scores. To further verify the performance of our RF-based models, we compare it with those of the DNN-based models. Table 3 shows the performances we achieved on two datasets using four evaluation criteria.

Table 3: Ten-fold cross-validation performance of ML models for noncentrosymmetry prediction

Model	Dataset	Precision	Recall	Accuracy	F1 score
RF-based	MPF	0.834	0.754	0.848	0.781
RF-based	MP3	**0.845**	0.755	**0.869**	**0.786**
DNN-based	MPF	0.773	0.769	0.785	0.771
DNN-based	MP3	0.784	**0.780**	0.792	0.782

Firstly, we found that the precision and accuracy of the RF model are significantly better in comparison with DNN models: the 10 fold cross-validation accuracy of RF model on the MPF dataset is 0.848 compared to 0.785, which indicates 7.89% improvement. The F1 score of RF model is 0.781 compared to 0.771 of DNN. Although DNN achieves better Recall score, the F1-score of RF is higher than DNN's. This validates the effectiveness of our RF-based model for predicting the noncentrosymmetric property for a given material. This is consistent with a recent evaluation of different ML methods for materials property prediction [37].

Secondly, comparing the results of the same RF and DNN model on the MPF dataset and the MP3 dataset, we found that each model achieved better prediction performance for the MP3 dataset. Particularly, the precision, accuracy and F1 score of the RF classifier increase to 0.845, 0.869 and 0.786, respectively.

3.3.2 Hold out experiment results

To explore the effectiveness of our model for extrapolative prediction of noncentrosymmetry where the test samples may not have the same distribution with the training set, we conducted a hold-out test over the Borates315 dataset.The training dataset is generated by filtering out all the samples of the Borates315 dataset from the MPF dataset and keeping the remaining ones, which includes 82,191 samples. Similarly, we also conduct a hold-out test for the MP3 dataset for which the training set is generated by removing all borates in the MP3 dataset. The number of samples of the no-borates 3-element training set is 39411. Their ROC curves and AUC scores are shown in Figure 3.

7

(a) Cross Validation performance over MPF dataset

(b) Holdout performance over MPF dataset

(c) Cross Validation performance over MP3 dataset
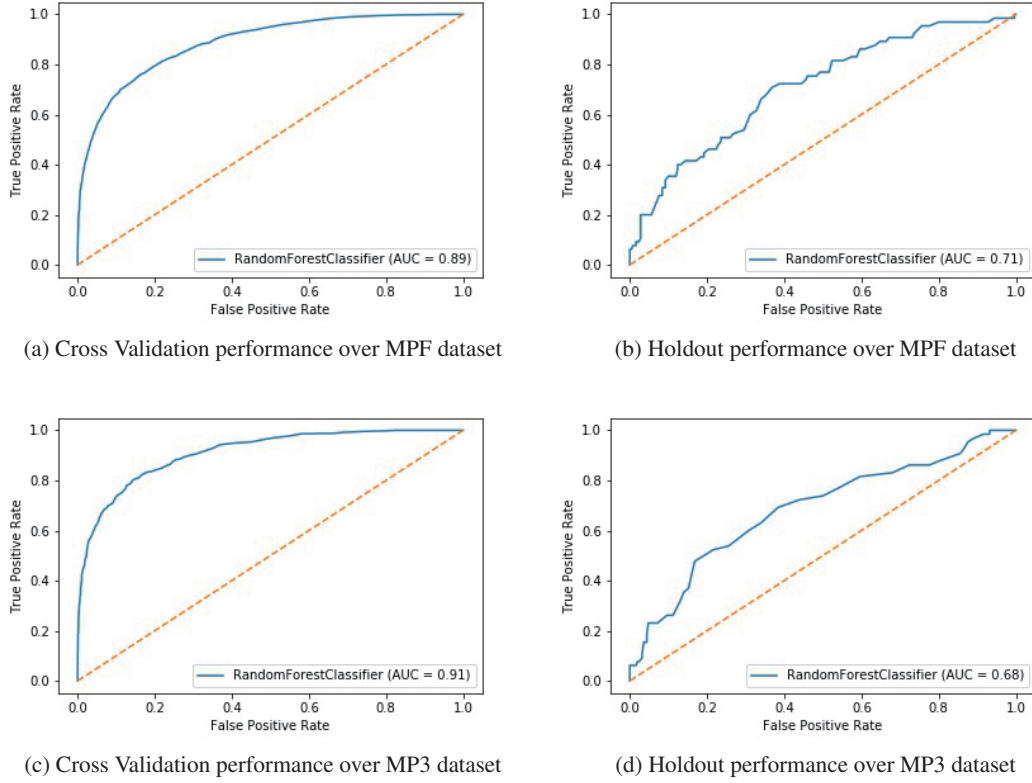

(d) Holdout performance over MP3 dataset

Figure 3: ROC curves for cross-validation and hold-out experiments for the RF prediction models trained with the whole dataset and the 3-element dataset.

In Figure 3, each dotted yellow line corresponds to the ROC curve of a random predictor with AUC value of 0.5. Each blue curve represents the ROC curve of the classifier. As is well known the higher value of AUC, the better performance of the classifier. Among the four sub-figures, figure (c) shows the best result, with AUC reaching 0.91. Furthermore, comparing (a) (c) with (b) (d), we can find AUC scores of cross-validation experiments are higher than those of hold-out experiments over the same two datasets, which suggests the over-estimation of model performance due to dataset sample redundancy. Meanwhile, although the performance of hold out experiments is not as good as cross validation experiments, it only uses the non-borate materials as the training data for predicting the 315 borate materials, which interprets the 0.71 and 0.68 AUC are acceptable since this is extrapolation prediction performance. Based on this analysis, we use the RF model to predict and screen hypothetical materials from a large generated materials as discussed in detail in Section 3.4.

3.3.3 The stability of our model

To evaluate the stability of our RF model performance, we made the following Box plot, which shows that the fluctuations of precision and F1 scores for the 10-fold cross-validation experiments are less than 0.01. However, we found that the precision scores of the hold out experiments over the MPF dataset range from 0.61 to 0.67, and the F1 scores are between 0.58 to 0.64. This shows that the prediction performance of our RF models with 10-fold cross-validation experiments are more stable than those of the hold out tests.
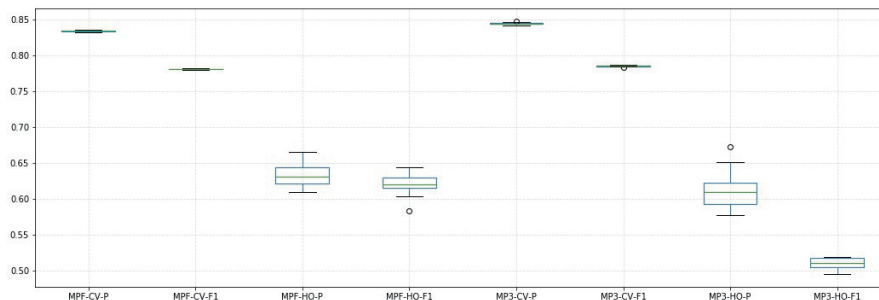
Figure 4: The stability of RF models. (MPF and MP3 are datasets; CV and H are abbreviations of cross validation and hold out; P and F1 represent Precision and F1 score respectively.)

3.3.4 Feature importance ranking in noncentrosymmetry prediction

There are 132 descriptors in the Magpie feature set. To gain further understanding of how different descriptors affect the ML model performance, we calculated the importance scores for all descriptors in the prediction of the RF model and sorted them by their scores. The top 15 descriptors are shown in Figure 5 and the corresponding description of them are presented in Table 4.

As can be seen from Figure 5, the importance scores of top 15 features are above 0.014. The top six features have significantly higher scores than the remaining nine features, which shows they make more contributions to predicting the non-centrosymmetry. Combined with Table 4, we find that range of atomic number, maximum melting temperature, mean number of valence, range of number of valence, mean number of Ns valence and minimum number of Nd valence are the six most important factors. We also find that the importance of valence number to noncentrosymmetry prediction is consistent with the physical knowledge: first the distribution of valence electrons have strong effect on chemical bond formation (strong covalent bonds or weaker ionic bonds), and thus the final crystal structure formation. Second, previous study [38] shows that the valence electrons of the atoms is involved in its nonlinear optical behavior: they construct the free electron gas, which can be polarized by the oscillating electric field and determine the harmonic excitation frequency by counting linear and nonlinear reflected waves.
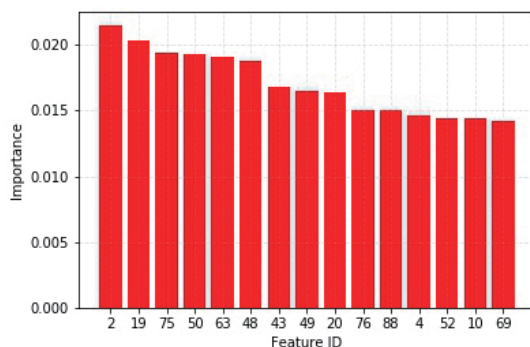

Figure 5: Ranking of top 15 features in terms of their importance scores

Table 4: Top 15 features in noncentrosymmetry prediction

Feature ID	Feature Name	Feature Description
2	Range Number	Range of Atomic Number
19	Maximum MeltingT	Maximum Melting Temperature
75	Mean NValence	Mean # Valence
50	Range NsValence	Range of # Valence s-orbitals
63	Mean NdValence	Mean # Valence d-orbitals
48	Minimum NsValence	Minimum # Valence s-orbitals
43	Maximum Electronegativity	Maximum Electro-negativity
49	Maximum NsValence	Maximum # Valence s-orbitals
20	Range MeltingT	Range of Melting Temperature
76	Avg_dev NValence	Mean absolute deviation of # Valence
88	Avg_dev NpUnfilled	Mean absolute deviation of # Unfilled s Orbitals
4	Avg_dev Number	Mean absolute deviation of Atomic Number
52	Avg_dev NsValence	Mean absolute deviation of # Valence s-orbitals
10	Avg_dev MendeleevNumber	Mean absolute deviation of Mendeleev Number
69	Mean NfValence	Mean # Valence f-orbitals

3.4 Predicting new noncentrosymmetric materials

To identify interesting hypothetical new NLO noncentrosymmetric materials, we applied our RF-based noncentrosymmetric materials prediction model to screen the two million hypothetical materials generated by our Generative Adversarial Network (GAN) based new materials composition generator [14]. After predicting the probability of each candidate belonging to noncentrosymmetric materials, we sort them by the probability scores and report top 20 hypothetical noncentrosymmetric materials with 2, 3 and 4 elements here in Table 5. Furthermore, as we mentioned above that most borate materials are NLO materials. So we also reported top 20 borate materials with highest proability here. Please note that materials containing lanthanide and actinide elements have been filtered in these results because they are very rare.

Table 5: Predicted hypothetical noncentrosymmetric materials with 2, 3, and 4 elements and predicted noncentrosymmetric borates (only top 20 are listed here)

2 element	score	3 element	score	4 element	score	Borate	Score
Li4Ge	0.935	AlCuSe3	0.960	LaCeNdS4	0.975	CB2O6	0.840
Cu2S3	0.875	Cu2AsS3	0.955	LaCeNdSe4	0.965	N2B4O7	0.715
NO5	0.835	Cu3As2S4	0.945	CeNdEuS4	0.960	CB4O6	0.700
Li4Pb	0.830	Y2CeO5	0.945	CuZnInS3	0.955	S3B2O8	0.670
Li4Sn	0.800	CeTb2S4	0.935	AlCuZnTe4	0.925	CB2O4	0.665
Cl3S	0.745	DyErC3	0.930	MnNiAgSn	0.925	NCB4O6	0.665
SbC	0.740	MnDy2S4	0.925	AlCuInSe2	0.915	CoIB4O6	0.660
Pd2S	0.735	LaSm2S4	0.920	MnCoRuSn	0.915	EuB4O6	0.655
AsC	0.720	ZnGaSe2	0.920	LaNdUTe4	0.915	ZnSnO6B4	0.650
SeO6	0.715	AlCu2Te3	0.915	Cu2ZnInS6	0.900	As2B2O7	0.635
Ni3Ge2	0.715	AlCu2S4	0.910	NiCuSnSe3	0.895	PB2O6	0.630
Cl5S	0.710	CoCd2S3	0.905	MnCoAgSn	0.895	ZnB2O4	0.625
Zr2S3	0.695	NbSnIr	0.900	TiCoRhSn	0.880	SB2O6	0.620
S2O5	0.690	NbWTe4	0.900	MnFeSbO6	0.875	MnZnLaEuO6B2	0.610
LiOs	0.690	VSnAu	0.900	MnCu2AgS4	0.875	Zn3SB2O6	0.600
NH2	0.690	CrCu2S3	0.895	FeLaPbO6	0.875	Sr2TaB2O6	0.600
CrI	0.685	SnTaOs	0.890	V2Ni2RuSn2	0.875	PbB4O6	0.595
F3N	0.680	NdDySi3	0.885	MnFeBi2O6	0.875	AlB2O4	0.585
Cl6S	0.675	Dy2GeS4	0.885	TiCoBi2O6	0.870	NbRuCl2B4O6	0.585
S2O3	0.660	Mg6MnSn	0.885	SrLaNdS4	0.865	C3B4O6	0.580

As shown in Table 5, the probability score range of top 20 2-element materials, 3-element materials, 4-element materials and borate materials are 0.935 to 0.660, 0.960 to 0.885, 0.975 to 0.865 and 0.885 to 0.670, respectively. It is clear that the predicted noncentrosymmetic probabilities of 3 element materials are higher than those of 2-element materials and 4-element materials. As those material are generated and hypothetical, we can only give the predicted noncentrosymmetry scores, which may guide experimental work to verify them in future research, which may further validate the effectiveness and the predictive capability of our models. More prediction results can be provided by the corresponding author upon reasonable request.

4   Conclusions

Computational prediction of noncentrosymmetry of a given composition can be used for fast screening new nonlinear optical materials. Here we developed and evaluated two machine learning models including a Random Forest Classifier and a neural network model for computational prediction of materials noncentrosymmetry given only their composition information. By using the Magpie composition features, our best prediction model based on Random forest can achieve an accuracy of 84.8% when evaluated using 10-fold cross-validation over the Material Projects database. Further experiments showed that when the prediction model is trained only on 3-element samples, it can achieve even higher performance for the test set, which is made of mostly 3-element materials. A feature importance calculation shows the top six contribution factors for predicting noncentrosymmetry, many of which are related to the distribution of valence electrons. which is consistent with current physichochemical principles. Our developed model can be applied to discovering novel nonlinear materials as we conduct large-scale screening over two million hypothetical materials.

5   Author contributions

conceptualization, J.H. and J.L.; methodology, J.H., S.Y., Y.Z., J.L.; software, Y.S., Y.Z., J.L.; validation, Y.S and J.L.; investigation, J.H., Y.S., J.L.; data curation, A.N. and J.L.; writing–original draft preparation, J.H., Y.S., and J.L.; writing–review and editing, J.H.,Y.S., J.L., M.H.; visualization, Y.S.; supervision, J.H. ; project administration, J.H.; funding acquisition, J.H., M.H. and J.L.

6   Acknowledgements

The authors declare no conflict of interest.

Data Availability:

The data required to reproduce these findings are downloaded from Materials Project database at https://materialsproject.org/.

References

[1] Kang Min Ok, Eun Ok Chi, and P Shiv Halasyamani. Bulk characterization methods for non-centrosymmetric materials: second-harmonic generation, piezoelectricity, pyroelectricity, and ferroelectricity. *Chemical Society Reviews*, 35(8):710–717, 2006.

[2] P Shiv Halasyamani and Kenneth R Poeppelmeier. Noncentrosymmetric oxides. *Chemistry of Materials*, 10(10):2753–2769, 1998.

[3] Walter Kohn. Nobel lecture: Electronic structure of matter—wave functions and density functionals. *Reviews of Modern Physics*, 71(5):1253, 1999.

[4] Hossin A. Abdeldayem and Donald O. Frazier, editors. *Nonlinear Optics and Applications*. Research Signpost, 2007.

[5] 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.

[6] 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.

[7] A Diatta, J Rouquette, P Armand, and P Hermet. Density functional theory prediction of the second harmonic generation and linear pockels effect in trigonal bazno2. *The Journal of Physical Chemistry C*, 122(37):21277–21283, 2018.

[8] Bartłomiej Dec and Robert Bogdanowicz. Dft studies of refractive index of boron-doped diamond. *Photonics Letters of Poland*, 10(2):39–41, 2018.

[9] Artem R Oganov, Chris J Pickard, Qiang Zhu, and Richard J Needs. Structure prediction drives materials discovery. *Nature Reviews Materials*, 4(5):331–348, 2019.

[10] Zhuo Cao, Yabo Dan, Zheng Xiong, Chengcheng Niu, Xiang Li, Songrong Qian, and Jianjun Hu. Convolutional neural networks for crystal material property prediction using hybrid orbital-field matrix and magpie descriptors. *Crystals*, 9(4):191, 2019.

[11] Kam Hamidieh. A data-driven statistical model for predicting the criticaltemperature of a superconductor. Technical report, Statistics Department, University of Pennsylvania, Wharton, PA, 2018.

[12] Kamal Choudhary, Marnik Bercx, Jie Jiang, Ruth Pachter, Dirk Lamoen, and Francesca Tavazza. Accelerated discovery of efficient solar cell materials using quantum and machine-learning methods. *Chemistry of Materials*, 31(15):5900–5908, 2019.

[13] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[14] Yabo Dan, Yong Zhao, Xiang Li, Shaobo Li, Ming Hu, and Jianjun Hu. Generative adversarial networks (gan) based efficient sampling of chemical space for inverse design of inorganic materials. *arXiv preprint arXiv:1911.05020*, 2019.

[15] Valentin Stanev, Corey Oses, A. Kusne, Efrain Rodriguez, Johnpierre Paglione, Stefano Curtarolo, and I. Takeuchi. Machine learning modeling of superconducting critical temperature. *npj Computational Materials*, 4, 09 2017.

[16] Kaname Matsumoto and Tomoya Horide. Acceleration search method of higher tc superconductors by machine learning algorithm. *Applied Physics Express*, 12, 06 2019.

[17] Logan Ward. A general-purpose machine learning framework for predicting properties of inorganic materials. In Pat Langley, editor, *Nature News*. Nature Publishing Group, 2016.

[18] Logan Ward, Stephanie C O'Keeffe, Joseph Stevick, Glenton R Jelbert, Muratahan Aykol, and Chris Wolverton. A machine learning approach for engineering bulk metallic glass alloys. *Acta Materialia*, 159:102–111, 2018.

[19] Al'ona Furmanchuk, Ankit Agrawal, and Alok Choudhary. Predictive analytics for crystalline materials: bulk modulus. *RSC advances*, 6(97):95246–95251, 2016.

[20] Gaspar Cano, Jose Garcia-Rodriguez, Alberto Garcia-Garcia, Horacio Perez-Sanchez, Jón Atli Benediktsson, Anil Thapa, and Alastair Barr. Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert Systems with Applications*, 72:151–159, 2017.

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[22] Zhidong Su, Yang Li, and Guanci Yang. Dietary composition perception algorithm using social robot audition for mandarin chinese. *IEEE Access*, 8:8768–8782, 2020.

[23] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

[24] Weike Ye, Chi Chen, Zhenbin Wang, Iek-Heng Chu, and Shyue Ong. Deep neural networks for accurate predictions of crystal stability. *Nature Communications*, 9, 12 2017.

[25] Shaobo Li, Yabo Dan, Xiang Li, Tiantian Hu, Rongzhi Dong, Zhuo Cao, and Jianjun Hu. Critical temperature prediction of superconductors based on atomic vectors and deep learning. *Symmetry*, 12(2):262, 2020.

[26] 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*, 2020.

[27] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils ER Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018.

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

[29] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[30] Zhuo Cao, Yabo Dan, Zheng Xiong, Chengcheng Niu, Xiang Li, Songrong Qian, and Jianjun Hu. Convolutional neural networks for crystal material property prediction using hybrid orbital-field matrix and magpie descriptors. *Crystals*, 9:191, 04 2019.

[31] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12):2295–2329, 2017.

[32] 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.

[33] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[34] James Bergstra, Dan Yamins, and David D Cox. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in science conference*, pages 13–20. Citeseer, 2013.

[35] Rimma Bubnova, Sergey Volkov, Barbara Albert, and Stanislav Filatov. Borates—crystal structures of prospective nonlinear optical materials: High anisotropy of the thermal expansion caused by anharmonic atomic vibrations. *Crystals*, 7(3):93, 2017.

[36] Zheng Xiong, Yuxin Cui, Zhonghao Liu, Yong Zhao, Ming Hu, and Jianjun Hu. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Computational Materials Science*, 171:109203, 2020.

[37] Matthew C Robinson, Robert C Glen, and Alpha A Lee. Validating the validation: Reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. *arXiv preprint arXiv:1905.11681*, 2019.

[38] Jeffery J Maki, Martti Kauranen, and André Persoons. Surface second-harmonic generation from chiral materials. *Physical review B*, 51(3):1425, 1995.