ACS APPLIED MATERIALS & INTERFACES

www.acsami.org

Research Article

DeepXRD, a Deep Learning Model for Predicting XRD spectrum from Material Composition

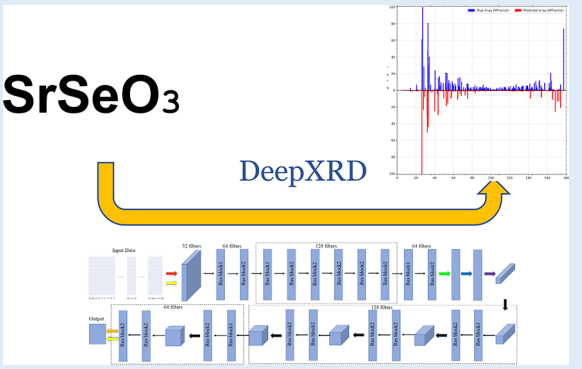
Rongzhi Dong, Yong Zhao, Yuqi Song, Nihang Fu, Sadman Sadeed Omee, Sourin Dey, Qinyang Li, Lai Wei, and Jianjun Hu*

Cite This: *ACS Appl. Mater. Interfaces* 2022, 14, 40102–40115

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: One of the long-standing problems in materials science is how to predict a material's structure and then its properties given only its composition. Experimental characterization of crystal structures has been widely used for structure determination, which is, however, too expensive for high-throughput screening. At the same time, directly predicting crystal structures from compositions remains a challenging unsolved problem. Herein we propose a deep learning algorithm for predicting the XRD spectrum given only the composition of a material, which can then be used to infer key structural features for downstream structural analysis such as crystal system or space group classification or crystal lattice parameter determination or materials property prediction. Benchmark studies on two data sets show that our DeepXRD algorithm can achieve good performance for XRD prediction as evaluated over our test sets. It can thus be used in high-throughput screening in the huge materials composition space for materials discovery.

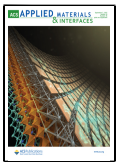


KEYWORDS: *inorganic materials, XRD spectrum, crystal structure prediction, deep learning, residual connection, materials screening*

1. INTRODUCTION

One of the major goals of materials science is to elucidate the composition−processing−structure−property−performance relationships so that materials with desired functions can be designed and synthesized.[1] Traditionally, the problem is studied as a forward problem in which the cause-and-effect relationships are uncovered from composition to processing and structure and then to performance.[2] One starts with a tentative composition/recipe and then utilizes some known processing processes with adjustments to synthesize the material sample, whose structure is then derived using the structural characterization data, which are typically generated via scanning X-ray diffraction (XRD) or Raman spectroscopy experiments. By analyzing the structural characteristics, one can estimate its potential properties and performance. On the other hand, the materials discovery can be formulated as an inverse design problem, in which one starts from a performance target and tries to find/search for the best composition and processing to achieve the desired performance. In both processes, one of the major bottlenecks is how to get the structure for a given composition. As shown in Figure 1, currently the experimental approaches are infeasible for large-scale screening of the vast chemical design space in which millions of possible compositions may be generated by modern generative models.[3] On the other hand, computational crystal structure prediction algorithms such as USPEX and CALYPSO[4,5] can only be applied to relatively small systems. The template-based crystal structure prediction methods such as those in refs 6 and 7 are limited to predicting structures with known structure prototypes. In this paper, we aim to explore whether we can develop a deep learning algorithm to predict the XRD spectrum from the composition alone, which can then be used for fast large-scale structure-oriented screening in modern computational generative materials design. Since our predicted XRDs can be fed to downstream algorithms to predict their structural dimensionality,[8] crystal systems,[9] and space groups,[10] our XRD prediction algorithm can be very useful for screening potential new materials with only their composition information, which can significantly narrow down the crystal structure prediction and Density Functional Theory (DFT) calculations effort. Our XRD prediction algorithm can also be potentially used for the unsupervised discovery of new materials with similar properties to known materials.[11] The third major application is to use our predicted XRD as reference XRDs for traditional XRD-based phase mapping

Received: April 1, 2022
Accepted: August 19, 2022
Published: August 26, 2022



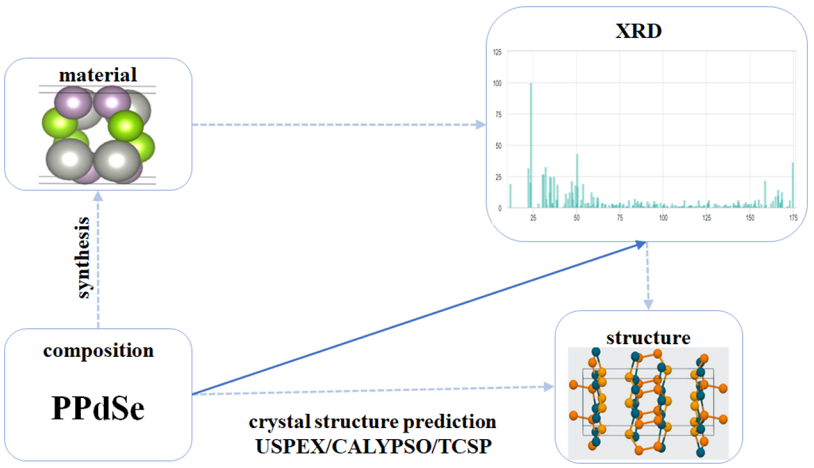


Figure 1. Composition−processing−structure−property relationships in materials science research. Experimental methods use synthesis and XRD characterization to obtain the crystal structures, while crystal structure prediction algorithms aim to directly predict the structures from compositions. This work tries the third approach: predicting the XRD spectrum from compositions.

analysis[12] with uncharacterized component phases. Our predicted XRDs can greatly expand the range of mixed materials phase mapping techniques. A high-quality predicted XRD spectrum may also be used to reconstruct crystal structures using well-established methods in the crystallography community.[13]

First-principles calculations such as DFT have been widely used for crystal structure prediction to study the crystal structures of inorganic materials.[14,15] Although the first-principles calculations are powerful, they are susceptible to the constraints of their excessive calculation cost, which limits the size of the material design space or the number of materials they can screen. To address this problem, machine learning (ML) has been increasingly applied to materials science fields, leading to the emergence of "materials informatics",[16] in which materials learning methods are developed to obtain prior knowledge and predictive models from known material data sets, and then predict complex material properties based on these models. In the past few years, ML has succeeded in predicting new features,[17] guiding chemical synthesis, and discovering suitable compounds with target properties.[18−21]

Several composition-based machine learning models have been proposed to predict structural properties such as crystal systems,[22] space groups,[23] lattice constants,[24] or Vickers hardness,[25,26] with varying performances. Composition-based ML models have also been extensively used for material property prediction. Well-known composition descriptors such as Magpie,[27] Matminer,[28] and composition-graph-based embeddings[29] have all been proposed for structure or property prediction. While these composition-based ML models for such tasks have been criticized for lack of high performance compared to structural descriptors-based materials property prediction models, they have a unique advantage for de novo discovery of new materials of which the crystal structures are usually not available and then only composition-based ML models can be used.[3] In addition, such models can be used as the first level coarse screening of millions of generated hypothetical materials from generative machine learning models.[3]

Several recent works have applied machine learning to XRD data. Suzuki et al.[30] used Random forest models to predict crystal systems and 230 space groups from XRD. A deep learning approach has also been proposed for space group classification from XRD.[31] Oviedo et al.[8] proposed a physics-informed data augmentation method that extends small, targeted experimental and simulated data sets and developed a convolutional neural network for classifying seven space groups. Convolutional neural networks have also been used to map the XRD patterns to materials with one-to-one mapping.[32] More recently, a deep neural network model[33] has been reported in *Science* to autonomously identify the crystal symmetry (systems) from electron backscatter diffraction and achieved high accuracy. Lee et al. proposed a deep learning algorithm to identify phases for multiphase inorganic compounds using simulated XRD data sets.[34] A similar effort but using nonnegative matrix factorization has also been developed for the phase identification problem.[12] Considering the limited data availability, Wang et al. proposed a data augmentation technique and used a convolution neural network for one-to-one phase identification.[32] Another related work is the XRD-based phase attribution or phase diagram reconstruction.[35,36] In a typical XRD spectrum, the most critical information on the structure is encoded in the peak positions and corresponding intensities. In many cases, small peak shifts may also happen. In addition, the XRD spectrum has been shown to be used to predict the space group of the sample with high accuracy. The XRD spectrum has also been used as a feature for unsupervised clustering to find new lithium superionic materials.[11]

Instead of trying to reconstruct the 3D coordinates of the crystal structure using DFT-based evolutionary search algorithms that are feasible only for small systems, in this work, we aim to emulate the traditional XRD-based structure characterizing approach: building a deep learning-based prediction model to predict the XRD given its composition. Experimentalists have been using XRD to analyze materials' properties for a long time. XRD predicted by our models can then be used by them for quick downstream analysis such as structure determination or property prediction. Similar to the experimental crystallography community practice, given an XRD from our model, there is a large set of algorithms such as the Rietveld refinement or the Rietveld method[37] that can be used to determine crystal structures. XRD diffraction patterns have been used to achieve over 90% accuracy for crystal system classification, except for triclinic cases, and with 88% accuracy for space group classification with five candidates.[10] In ref 38,

https://doi.org/10.1021/acsami.2c05812
ACS Appl. Mater. Interfaces 2022, 14, 40102−40115

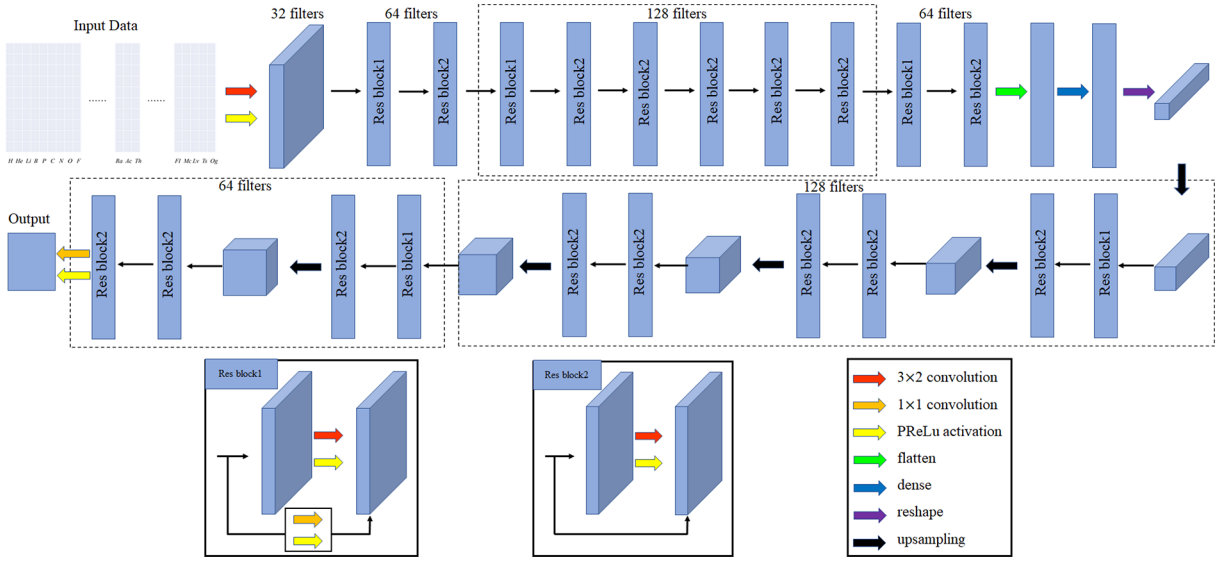


Figure 2. XRD spectrum prediction framework. The input is the one-hot encoding of a given formula through the first row's encoder to extract the key information and then follow the steps shown in the second row to decode and reconstruct the given material's probable XRD spectrum, which is our output. To apply skip connections on layers with filters of different sizes and the same sizes, we use two different kinds of residual blocks with their structures shown in the left and middle bottom. To connect layers with different filters, we add convolution operations when using skip connections to make sure the two layers keep the same shape so that they can be added together, as shown in Res block1. For layers with the same filters, we can just skip connections and add them together as shown in Res block2. We use colorful arrows to represent different operations in our prediction model as shown in the bottom right.

neural networks were used for space group classification with an accuracy of around 54% on experimental data, which can be improved to 82% at the cost of having half of the experimental data unclassified.

In this paper, we aim to develop deep learning-based models to simulate the relationship between materials composition and XRD spectrum with the understanding that many chemically and structurally similar materials share close XRD spectra.

Our contributions can be summarized as follows:

- We develop two benchmark data sets for the composition-based XRD prediction problems: ABC_3-XRD with 4270 samples and the Ternary-XRD data set with 43,223 samples.
- We propose a deep learning-based neural network model for predicting XRD spectra from material compositions.
- We evaluate four different loss functions based on different distance measures for calculating XRD similarities and find that the Pearson loss function achieves the best result. We find mean square error (MSE) is not a good choice for training deep learning models for XRD prediction due to their sensitivity to the peak intensities.
- We conduct extensive experiments over the two data sets and show that our proposed framework is capable to achieve good performance for test sets.

The remainder of this paper is organized as follows. Section 2 focuses on the research framework, materials representation, and evaluation indicators. Section 3 describes our experiments and highlights our prediction performance. The last section concludes the paper.

2. MATERIALS AND METHODS

2.1. XRD Spectrum Prediction Problem. In our composition-to-XRD mapping problem, the goal is to design a model that could learn from inorganic materials' compositions and then predict their probable XRD spectra. We prepare two data sets for training our models. The smaller data set has 4270 different inorganic materials with the prototype of ABC_3, where A, B, and C are three different elements. The larger data set has 43 223 samples of ternary materials. Each independent material has a corresponding XRD spectrum. In the case of polymorphism, where one composition corresponds to multiple phases, we pick the structure with the lowest formation energy. According to the composition of a material, we need to predict what the XRD spectrum is. To evaluate our model performance, for each data set, we randomly select 20% as the test set from all samples 70% as the training set, and 10% as the validation set. The training set is used to train our prediction model and use the validation set to tune the hyper-parameters. Finally, for a given target formula, we use our model to predict its XRD and compare it with the true XRD.

The main components of our deep learning framework are shown in Figure 2. We use a deep residual network (ResNet)[39] model trained with one-hot composition encoding features to learn the relationship between material composition and the XRD spectrum. For a given material formula, we use its one-hot matrix as the DeepXRD model's input. Due to the one-hot matrix being sparse, as shown in the first row of Figure 2, after the first convolution layer, we use several ResNet blocks with different filters, a flattened layer, and a dense layer as the encoding part to abstract the key information on input matrix. To reconstruct the XRD spectrum, we use several upsampling layers to magnify the key information to a 32×32 matrix as our output, and this progress is shown in the second row of Figure 2. The last row shows two different ResNet block structures. For skip connection, the two inputs must have the same shape, if the two layers have different filter numbers, we must add a convolution step to make the two inputs keep with the same shape, as Res block1 shown in the left bottom of Figure 2, and if the two layers have the same filter number, we can just use skip connection as Res block2 at the middle bottom. We introduced all operations used in our model in the right bottom; we have a 3×2 convolution layer, a 1×1 convolution layer, a flattened layer, a dense layer, a reshaping layer, and an upsampling layer. The activation function we used in our model is PReLu. The operations used in each layer are indicated by arrows of different colors.

https://doi.org/10.1021/acsami.2c05812
ACS Appl. Mater. Interfaces 2022, 14, 40102−40115

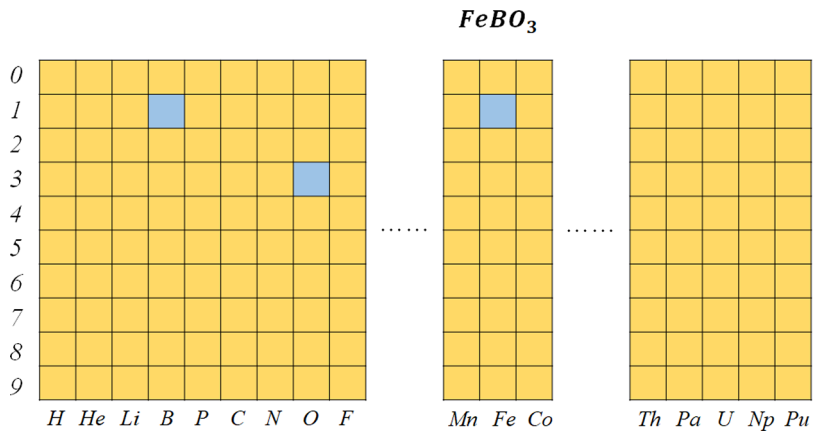


Figure 3. One-hot representation of formula $FeBO_3$; blue cells indicate 1 and yellow cells indicate 0.

2.2. Materials Representation. We use the one-hot representation of the formula for XRD prediction, which has been used in ref 3. The advantage of one-hot encoding is that it can encode a discrete material's elemental composition with a discrete 2D matrix of binary values 0 and 1, which are extremely suitable for the convolution neural network layers to extract hierarchical patterns from it. This coding method is also suitable for the characterization of the elements in the molecular formula of the material. As shown in Figure 3, each formula can be encoded as a 2D matrix of dimension $N \times M$, where N is the maximum number of atoms for an element in the formula and M is the number of elements considered in our data sets. All elements not included in the material formula are set to 0, and the column corresponding to each element in the formula has a nonzero value of 1, which is assigned to the column cell in the row j+1 where j is the number of atoms with this element in the material formula. For example, for $AcBO_3$, the one-hot code corresponding to this formula is a two-dimensional matrix with 10 rows and 84 columns (the samples in our data sets are composed of 84 different elements), in which the column Fe, B, and O have values 1 on row 4, 2, and 2, respectively, and the remaining values of the matrix are all set to 0.

2.3. ResNet Model for XRD Prediction from Composition. Figure 2 shows the whole neural network architecture of our DeepXRD model, which contains an input part, an encoding module, and a decoding module. The input data of our model is the one-hot representation of a given material formula with the dimension of 10 × 84, and the output is a matrix that represents the corresponding XRD spectrum with a dimension of 32 × 32. Since the one-hot encoding matrix is very sparse, we use the encoding module to extract the key information on materials and then use the decoding part to reconstruct the XRD spectrum. Our network is mainly composed of two types of residual blocks. Our Res blocks are shown at the bottom of Figure 2. Figure 4 shows the basic residual block with the shortcut connections. In convolutional neural networks, the output from the layer and the identity input may have different dimensions, so we add convolution operations in the shortcut connection such that the input is converted to the same dimensions. The Res block1 is used to connect layers with different numbers of filters so we add convolution operations when making skip connections to make sure the two parts we add together have the same dimension. Res block2 is used for layers with the same number of filters, so we can just add them together. The use of residual blocks aims to address the vanishing gradient problem in training deep convolutional neural networks.[39]

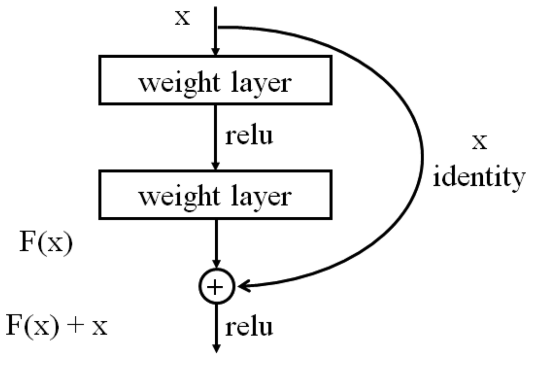


Figure 4. Building block of ResNet, introduced by ref 39.

In our DeepXRD model, we choose the Parametric Rectified Linear Unit (PReLu)[40] instead of the Rectified Linear Unit(ReLu) as the activation function. The basic ReLU has an output of 0 if the input is less than 0, which could cause the dying ReLU problem[41] where some ReLU neurons essentially remain inactive for all inputs. Due to the slope of ReLU in the negative part is also 0, once a neuron gets negative, it is unlikely to recover anymore. Therefore, no gradient flows and a large part of neural networks may do nothing. Parametric ReLU (PReLU) has a small slope for negative input values, which fixes the dying ReLU problem and can also speed up training. PReLU improves model fitting with nearly zero extra computational cost and little overfitting risk.[40] As shown in Figure 2, after the first convolution layer, there are 10 residual blocks with different numbers of filters, which reduce the feature map matrix size to 11 × 2. Flatten, dense, and reshape layers are then used to convert the feature map matrix to 1 × 1 × 128. The decoding module uses upsampling layers and Res blocks layers to increase the feature map matrix size to get the final XRD value matrix. We compare four different loss functions and finally choose the Pearson product-moment correlation as the loss function for our DeepXRD model. The parameters of each layer are shown in Table 1.

2.4. Materials Data Sets. To evaluate the performance of our DeepXRD algorithm, we prepared two data sets. The first data set, ABC_3-XRD, contains 4270 material compositions of the prototype ABC_3 along with the XRD spectra calculated for their crystal structures with the lowest formation energy as downloaded from the Materials Project database.[42] Since all the materials share the same prototype except the elements, we expect that the algorithm will achieve better performance over this data set. The second data set Ternary-XRD contains 43 223 compositions of diverse prototypes along with their computed XRD as downloaded from the Materials Project database.[42] Each XRD data set contains the corresponding XRD intensity at $2\theta(180)$ degrees. We used an average mapping method to sample the raw XRD spectrum so that all XRD sample contains 1024 points ranging from 0 to 180°.

In the crystallography community, it is well-known that compositions in a given materials families tend to have similar XRD spectra. It is interesting to check whether XRD data also form clusters. We thus visualize the distribution of the ABC_3-XRD data set using the t-distributed Stochastic Neighbor Embedding (t-SNE)[43] technique to map high-dimensional XRD spectra into a two-dimensional map, with each point in Figure 5a corresponding to one XRD spectrum. From the data distribution figure, we find that there are several loose cluster

https://doi.org/10.1021/acsami.2c05812
ACS Appl. Mater. Interfaces 2022, 14, 40102−40115

Table 1. Layers and Parameters of the DeepXRD Model

layer	input shape	filter	layer	input shape	filter
conv1	[batch, 84,10,1]	32	Res block1	[batch, 2,2,128]	128
Res block1	[batch, 84,10,32]	64	Res block2	[batch, 2,2,128]	128
Res block2	[batch, 42,5,64]	64	upsampling	[batch, 2,2,128]	128
Res block1	[batch, 42,5,64]	128	Res block2	[batch, 4,4,128]	128
Res block2	[batch, 21,3,128]	128	Res block2	[batch, 4,4,128]	128
Res block2	[batch, 21,3,128]	128	upsampling	[batch, 4,4,128]	128
Res block2	[batch, 21,3,128]	128	Res block2	[batch, 8,8,128]	128
Res block2	[batch, 21,3,128]	128	Res block2	[batch, 8,8,128]	128
Res block2	[batch, 21,3,128]	128	upsampling	[batch, 8,8,128]	128
Res block2	[batch, 21,3,128]	64	Res block1	[batch, 16,16,64]	64
Res block2	[batch, 11,2,64]	64	Res block2	[batch, 16,16,64]	64
flatten	[batch, 11,2,64]		upsampling	[batch, 16,16,64]	64
dense	[batch, 1408]		Res block1	[batch, 32,32,64]	64
reshape	[batch, 128]		Res block2	[batch, 32,32,64]	64
upsampling	[batch, 1,1,128]	128	conv	[batch, 32,32,64]	1

sets with the same crystal system, which means these materials may also have similar structures or chemical properties. Furthermore, for cluster sets containing several systems, the corresponding XRD spectra with different crystal systems may have a similar shape. Figure 5b shows the local zoomed region of Figure 5a, and this cluster is mainly composed of cubic and orthorhombic. We find that these materials from both cubic and orthorhombic have similar XRD shapes. The biggest peak positions are almost the same as each other, and the other peak's variation tendencies are also similar.

2.5. Model Evaluation. *2.5.1. Postprocessing.* There is always some noise in XRD caused by instrumental errors. To ignore the influence of these noises, we set a threshold of XRD magnitude values to distinguish between the real peaks and the noises. When using XRD spectra to predict structural information, many scientists only focus on the main peaks. Inoue focused on five major diffraction peaks when clarifying the relationship between crystal morphology and XRD peak intensity of $CaSO_4 \cdot 2H_2O$,[44] and Suzuki's crystal structure prediction model[10] focused on ten main peaks. Oviedo also used peak elimination and pattern shifting in their paper when predicting crystallographic dimensionality and space group.[8] As the largest XRD magnitude value is 100 counts per second (cps), and peaks with intensities less than 10 cps will never be the main peaks in our data set, we set 10 cps as the threshold value: if the XRD magnitude values are less than 10 cps, they are more likely caused by noise rather than main components. Therefore, we can just ignore them and consider values greater than 10 cps as true peaks. We also introduce a peak-alignment operation to consider the allowable peak shifts in experimental XRD-based structure characterization. In the performance evaluation stage, we have corresponding ground truth XRD peaks so we can shift predicted XRD peaks within a threshold distance to do peak alignment: for a given ground truth peak, we first find all peaks within 2° of peak position and then consider the largest peak as the corresponding peak of the true peak for prediction error calculation. In the testing stage, we do not have ground truth XRD value, so we can just do peak merging: as all peak positions range from 0 to 2θ (in our data set, 2θ is 180°), we can safely consider that there is just one peak if the distance between several peaks is less than 2°. In this case, we assume that the main peak's position is in the middle of these peaks. By disregarding noises and applying peak alignment, the predicted XRD values can be more realistic.

To evaluate the performance of our DeepXRD, we introduce a series of XRD dissimilarity/distance metrics including Cosine matrix, Pearson product-moment correlation, Jensen−Shannon divergence (JSD), and dynamic time warping (DTW), which have been evaluated in the literature.[45] It is found that the Cosine and Pearson similarity measures can get the best XRD clustering performance when peak height changes and peak shifting are present in the data (due to lattice constant changes) and the magnitude of peak shifting is unknown. In another study,[46] 49 metrics are evaluated to check their sensitivity when used for XRD clustering. It is shown that when the prior knowledge of the maximum peak shifting is available, dynamic time warping in a normalized constrained mode provides the best

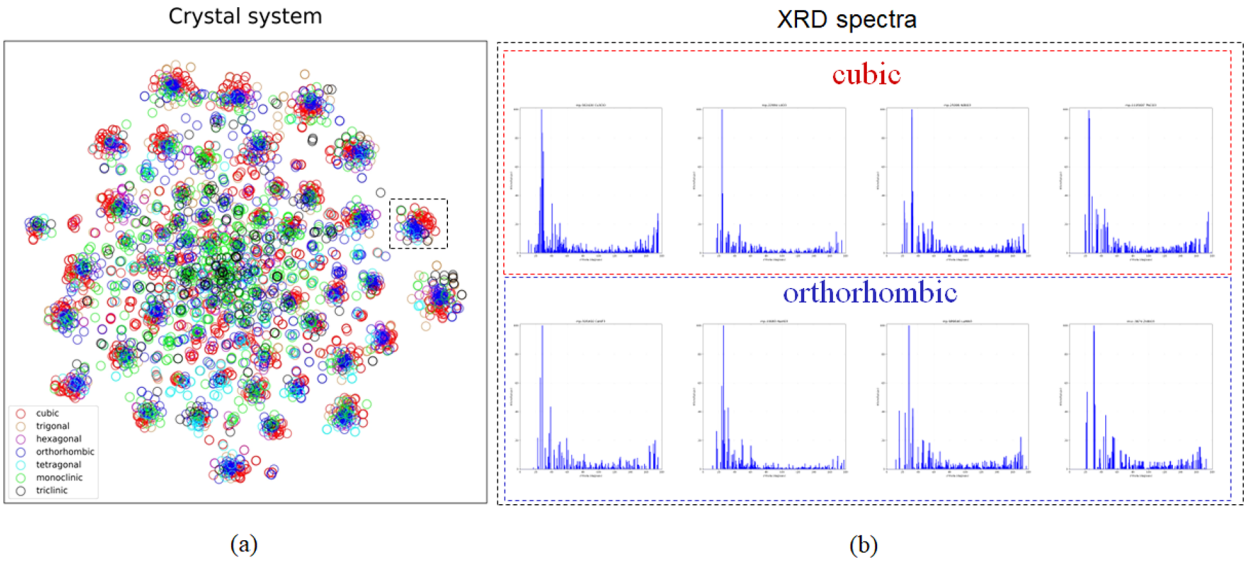


Figure 5. XRD distribution of the ABC_3-XRD data set using t-SNE visualization of XRD patterns. Each point corresponds to one XRD pattern. XRD patterns form loose clusters for each crystal system. (a) Distribution of ABC_3 samples, (b) XRD spectra of the zoomed region as marked in panel a.

https://doi.org/10.1021/acsami.2c05812
ACS Appl. Mater. Interfaces 2022, 14, 40102−40115

clustering performance. For two diffraction patterns s and t, the dissimilarity measure is defined as $D(s, t)$. For $D(s, t) = 0$, the two diffraction patterns are assumed to be identical, and the corresponding samples are assumed to share the same structure. Larger values of the dissimilarity measure imply greater dissimilarity between the samples' structures.

The evaluation measures used in our work are shown as follows: we use MSE, mean squared logarithmic error (MSLE), Cosine metric, Pearson product-moment correlation, JSD, and DTW[47] to compute the XRD dissimilarity.

$$D_{\mathrm{MSE}}(\boldsymbol{s}, t) = \frac{1}{n}\sum_{i=1}^{n}(s_i - t_i)^2 \tag{1}$$

$$D_{\mathrm{MSLE}}(\boldsymbol{s}, t) = \frac{1}{n}\sum_{i=1}^{n}(\log(s_i + 1) - \log(t_i + 1))^2 \tag{2}$$

$$D_{\mathrm{Cosine}}(\boldsymbol{s}, t) = 1 - \frac{\sum_{i=1}^{n}(s_i t_i)}{(\sum_{i=1}^{n} s_i^2)^{1/2}(\sum_{i=1}^{n} t_i^2)^{1/2}} \tag{3}$$

$$D_{\mathrm{Pearson}}(\boldsymbol{s}, t) = 1 - \frac{\sum_{i=1}^{n}(s_i - \bar{s})(t_i - \bar{t})}{(\sum_{i=1}^{n}(s_i - \bar{s})^2)^{1/2}(\sum_{i=1}^{n}(t_i - \bar{t})^2)^{1/2}} \tag{4}$$

$$D_{\mathrm{JSD}}(\boldsymbol{s}, t) = \frac{1}{2}\sum_{n}^{i=1} s_i \log\left(\frac{2s_i}{s_i + t_i}\right) + \frac{1}{2}\sum_{n}^{i=1} t_i \log\left(\frac{2t_i}{s_i + t_i}\right) \tag{5}$$

3. RESULTS AND DISCUSSION

3.1. Prediction Performance of the Composition Descriptor-Based XRD Predictor. In XRD-based crystal structure characterization, the peak positions rather than their magnitude values mainly reflect the structural or chemical properties of the materials. Small XRD magnitude values are usually caused by noise, and therefore independent of the material properties. Thus, we can mainly focus on the peak positions of our predicted results. For each formula, after generating XRD values, we first select its peak position (where magnitude values are larger than 10 cps) and then use peak merging to combine peaks if their position's distance is less than 2°. To determine the training set size used to train our DeepXRD model, we compared different train test split ratios. Figure 6 shows the training error changes versus training set ratios on both ABC_3-XRD and Ternary-XRD data sets. When we increase the training set ratio, training error reduces, and after 0.7, both data sets' reduction trend gradually becomes less apparent. Thus, we use 70% samples from two data sets as the training set. Table 2 uses Cosine, Pearson, JSD, and DTW algorithms to evaluate mean distance errors of testing samples to show the predicted performance of the DeepXRD model. We use Pearson as the module's loss function. The first row is the distance between all predicted XRD positions and target positions. In the second row, we only focus on the peak positions between the predicted and target XRD spectra. In the last row, we use peak merging to apply shifting on the predicted peak position to make it clear. By focusing on peak position, the errors calculated by Cosine, Pearson, and JSD functions have increased from 0.884, 0.885, and 0.773 to 0.943, 0.950, and 0.811, respectively. The increase may be caused by the accurate nonpeak positions we ignored in this step. However, after merging peaks, the errors of these three functions have reduced to 0.633, 0.633, and 0.550 respectively. The decrease shows that although predicted peak positions are not totally exact, most of them are very close to the ground truth positions. The errors calculated by the DTW function are kept reducing from 4.54 to 3.92 and then to 3.57. DTW can warp the sequence so it can calculate the distance of peaks of a similar wave shape. Table 2 shows that our DeepXRD model can find the key peak position of materials only through its composition. The distance error distribution of all testing samples is shown in Figure 7. The first row shows the predicted peak position distance distribution, and the second row shows the distance errors after ignoring the noise and applying peak merging. The distance distribution figures show that by using peak merging, predicted XRD peak positions can be more similar to true peak positions thus the distribution of errors is more close to 0.

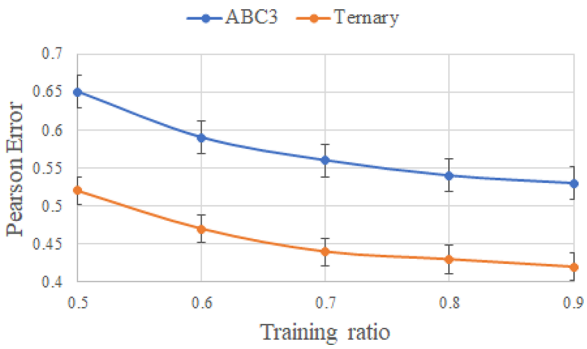


Figure 6. Training error versus training set ratio on ABC_3-XRD and Ternary-XRD data sets, respectively.

Table 2. Testing Errors (cps) Evaluated by Four Different Performance Measures for Composition-Based XRD Prediction

	Cosin	Pearson	JSD	DTW
prediction distance	0.884	0.885	0.773	4.54
peak distance	0.943	0.950	0.811	3.92
peak alignment distance	0.633	0.633	0.550	3.57

To choose the best loss function for the DeepXRD model, we use the peak match percentage as the criterion to evaluate four different loss functions' performance on the ABC_3-XRD and Ternary-XRD data sets, respectively, while keeping other hyper-parameters such as the batch size, learning rate, and training epoch unchanged. We compare two traditional loss functions, MSE and MSLE with Cosine and Pearson, which have been shown to perform better in XRD similarity studies.[45,46] Table 3 shows that all loss functions achieve better peak match percentages on the smaller ABC_3-XRD data set. The peak match accuracy improves by about 1% compared with the larger Ternary-XRD data set. We also find that the models trained with the Pearson loss function achieve the best match percentages on both data sets, which are 0.681 and 0.678, respectively. Compared with MSE's 0.626 and 0.612, MSLE's 0.644 and 0.631, Cosine's 0.673 and 0.667, the match percentages have improved by 6, 4, and 1%, respectively. The results of Table 3 prove that for the XRD prediction problem, Cosine and Pearson loss functions are better than traditional MSE and MSLE loss functions because they focus more on the shape rather than the exact values.

Figure 8 shows an example of the prediction results of $BaTbO_3$. Figure 8a shows the predicted XRD values and the target values, we find that our DeepXRD model can find the

https://doi.org/10.1021/acsami.2c05812
ACS Appl. Mater. Interfaces 2022, 14, 40102−40115

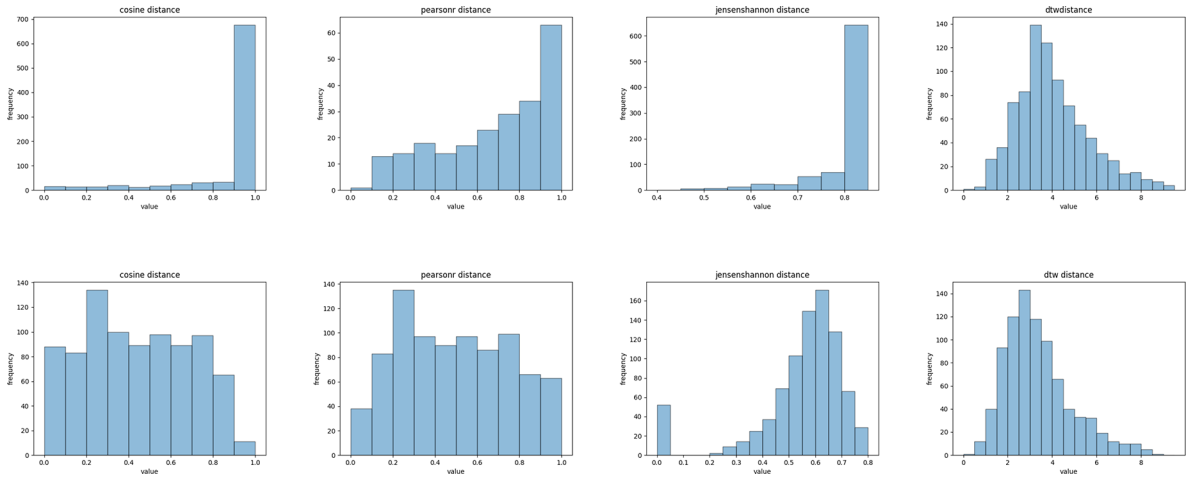


Figure 7. Histogram of peak distance errors over the testing samples.

Table 3. Prediction Performance (Peak Position Match Percentage) of Different Loss Functions

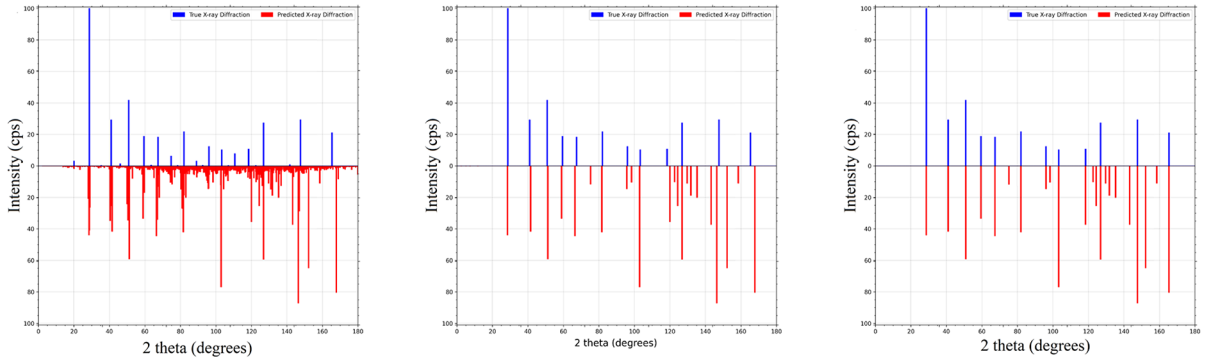
	performance measure			
data set	MSE	MSLE	Cosine	Pearson
ABC_3-XRD	0.626	0.644	0.673	**0.681**
Ternary-XRD	0.612	0.631	0.667	**0.678**

positions for most peaks. Figure 8b focuses only on all peak positions in the predicted and target XRD, and we use peak alignment to fine-tune peak positions as Figure 8c shows. Our algorithm has found almost all XRD peak positions, although the peak magnitudes maybe not be very accurate.

3.2. Hyper-Parameter Tuning of DeepXRD Models. To obtain the best hyper-parameters for the DeepXRD model, we compare peak position match percentages under different hyper-parameter combinations on two data sets. For a given formula, we first predict its XRD spectrum and then determine all peak positions with intensity greater than 10 cps. The performance results are shown in Table 4. For models with 10, 15, 20, and 25 ResNet layers, we calculate and compare the peak position match percentages with different learning rates and batch sizes. From Table 4, we find that for the ABC_3-XRD data set, the model with 20 ResNet layers, learning rate 0.001, and batch size 64 achieves the best performance. After peak shift operations, the final average peak match percentage is 68%. For the Ternary-XRD data set (Table 5), the model with 20 ResNet layers, learning rate 0.001, and batch size 128 achieves the best performance with peak match percentage 63%.

3.3. Case Studies of DeepXRD for XRD Spectrum Prediction. To evaluate the performance of our DeepXRD model, we randomly select three target compositions and their XRD spectra as design targets from the *ABC*$_3$-XRD test set. The XRD of the formulas predicted by the DeepXRD model (red color) are shown in Figure 9 together with the target XRD (blue color). The first row shows the structure of the given formulas. The second row shows XRD values predicted by the DeepXRD model. The third row shows the results that ignore the noise and only focus on peak positions and peak values. The last row shows predicted and ground truth XRD peak positions and peak values after peak alignment.

As most ABC_3 structures are ABO_3, we fix the C in ABC_3 as oxygen and choose different A and B values to evaluate how DeepXRD models perform on different materials. In the 3D space of crystal materials, there are seven crystal systems: triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal, and cubic. We choose three samples from the most common crystal systems of ABO3: cubic, orthorhombic, and monoclinic, respectively, to evaluate the predictive performance of our model. The crystal system of $SrSeO_3$ is monoclinic, and the structure of $SrSeO_3$ is shown in Figure 9a. It has two high peaks: one is between 25 and 35°; another is around 175 to 180°, and also some small peaks and noises, as



(a) predicted XRD VS ground truth (b) XRD peak VS ground truth (c) shifted peak XRD VS ground truth

Figure 8. XRD prediction performance of $BaTbO_3$.

https://doi.org/10.1021/acsami.2c05812
ACS Appl. Mater. Interfaces 2022, 14, 40102−40115

Table 4. Prediction Performance (Peak Position Match Percentage) Of Different Parameter Settings on the ABC_3-XRD Data Set

	learning rate											
	0.001			0.002			0.003			0.004		
ResNet layers	batch size 32	batch size 64	batch size 128	batch size 32	batch size 64	batch size 128	batch size 32	batch size 64	batch size 128	batch size 32	batch size 64	batch size 128
10	0.58	0.63	0.57	0.6	0.64	0.57	0.58	0.64	0.60	0.59	0.61	0.55
15	0.58	0.63	0.63	0.57	0.62	0.59	0.58	0.63	0.65	0.56	0.59	0.53
20	0.57	**0.68**	0.59	0.58	0.65	0.58	0.56	0.61	0.57	0.54	0.59	0.53
25	0.54	0.61	0.59	0.57	0.59	0.59	0.59	0.59	0.56	0.59	0.56	0.53

Table 5. Prediction Performance (Peak Position Match Percentage) Of Different Parameter Settings on the Ternary-XRD Data Set

	learning rate											
	0.001			0.002			0.003			0.004		
ResNet layers	batch size 64	batch size 128	batch size 256	batch size 64	batch size 128	batch size 256	batch size 64	batch size 128	batch size 256	batch size 64	batch size 128	batch size 256
10	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.59	0.57	0.55	0.56	0.57
15	0.57	0.59	0.57	0.57	0.57	0.57	0.57	0.58	0.58	0.56	0.56	0.54
20	0.61	**0.63**	0.6	0.57	0.58	0.58	0.57	0.59	0.58	0.53	0.58	0.57
25	0.6	0.61	0.59	0.53	0.59	0.56	0.52	0.58	0.51	0.56	0.59	0.57

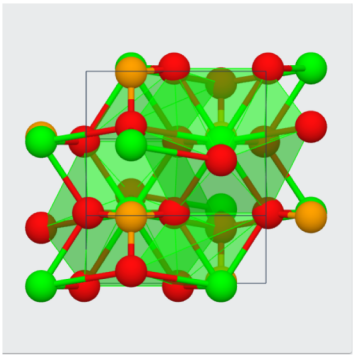
shown in Figure 9d. The predicted peaks (red color) match well with the approximate positions of the first two high peaks along with some small peaks, and the magnitude of the highest peak is accurately predicted. The last high peak located between 175 and 180° is missed, which is probably due to we do not have many training samples containing elements Sr and Se that generate peaks in this area. That is why our model does not predict a peak at around 180°. After filtering the noise signals in the XRD values (values smaller than 10 cps) we obtain Figure 9g, which shows the peak matches much better than in panel d by focusing only on significant peaks. From this figure, we find that there are small gaps in terms of the peak positions, which is related to the common X-ray diffraction peak shifting phenomena in crystallography.[48] To consider this factor, we conduct a peak shifting operation to match and adjust the predicted peaks within a distance smaller than 2° to the ground truth peaks, which makes our predictions much closer to the true values, as shown in Figure 9j.

The crystal system of $YAlO_3$ is orthorhombic, and its structure is shown in Figure 9b. It has only one high peak, around 30 to 35°, along with several small peaks. As shown in Figure 9e, the predicted peaks (red color) successfully match the approximate positions of almost all peaks with a value greater than 20. After ignoring the noises in the XRD spectrum (values smaller than 10), Figure 9h shows our predicted peaks and their matches with the ground truth more clearly. Figure 9k shows the matches after peak alignment: our algorithm predicts the same number of peaks as the true ones and only the smallest three of the 11 peaks are not aligned. Figure 9c shows the structure of $BaZrO_3$, which is a cubic material with scattered peaks and two high peaks around 30 and 165°, respectively. It also has several median peaks and almost no noise. As shown in Figure 9f, the predicted result (red color) successfully matches the approximate positions and magnitudes of the highest peak and almost all peak positions within the first 90°. Ignoring the noises in the XRD values (values smaller than 10 cps) does not improve the prediction results very much as shown in Figure 9i. Peak alignment can help to adjust the predicted peak positions in the second half as shown in Figure 9l. Our algorithm misses only the smallest three peaks.

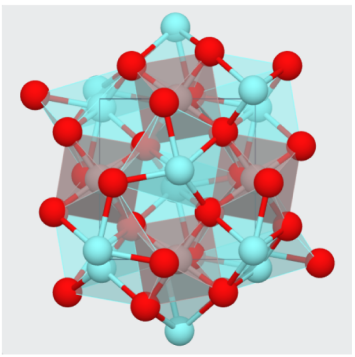
From Figure 9, we can find that the XRD distributions predicted by our DeepXRD model are very similar to the true XRD spectrum. When we focus only on the peak positions, the prediction errors can be further reduced. The peak position matches the ground truths and can be fine-tuned with the peak alignment operations. We also find that if the material crystal system or composition elements of a test material composition are infrequent in the training set, the predicted accuracy may become lower.

To further evaluate the DeepXRD model's performance on the ABC_3-XRD data set, we choose test samples with different A, B, and C elements. Figure 10 shows their predicted XRD (red color) and ground truth XRD spectra (blue color). The first test sample Ca_3SiO is orthorhombic (Figure 10a), which has two high-intensity peaks within the interval of 30 to 40° and a median peak around 50° together with several small peaks. Our predicted XRD spectrum matches almost all peaks of Ca_3SiO as shown in Figure 10d. The second test case is $CsInBr_3$, which has a cubic structure as shown in Figure 10b. $CsInBr_3$ has 5 high peaks: the first four peak positions are around 20 to 50° and the last peak is located at 160 to 165°. Figure 10e shows that for peaks with magnitude values greater than 20 cps, our predicted peak intensity and positions are similar to the true ones except for the last high peak. The third test sample $CsCaCl_3$, as shown in Figure 10c, is tetragonal. Figure 10f shows that $CsCaCl_3$'s main peak is around 20 to 60°. Our model accurately predicted the exact positions of these peaks. The fourth test sample $NdLuS_3$ is orthorhombic, and its structure is shown in Figure 10g. Its highest peak is the first peak located around 20° and the remaining peaks are around 20 to 60° as shown in Figure 10j. Our predicted XRD spectrum matches almost all peak positions with only several peak differences in intensities. Figure 10h shows the fifth test case Ca_3BiSb, which is a cubic with two high peaks: the first and the last one of all peaks with intensity greater than 80 cps. Our predicted XRD matches the first and highest peak very well and matches the second highest peak with only a small

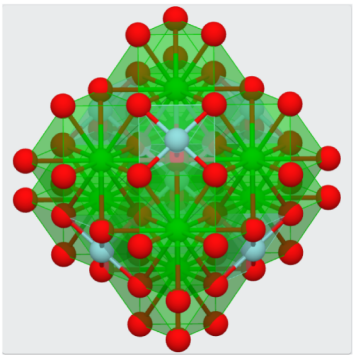
https://doi.org/10.1021/acsami.2c05812
ACS Appl. Mater. Interfaces 2022, 14, 40102−40115



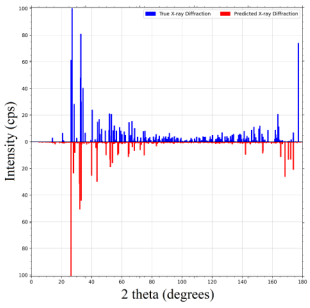
(a) $SrSeO_3$



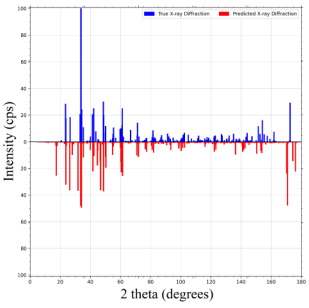
(b) $YAlO_3$



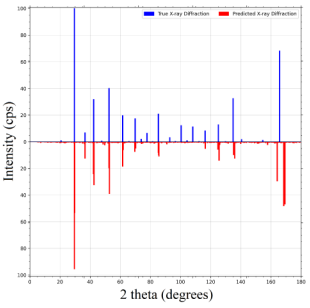
(c) $BaZrO_3$



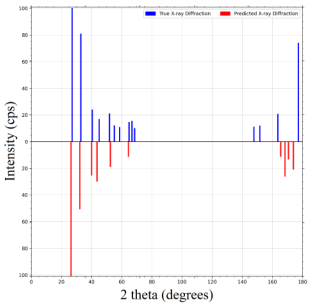
(d) $SrSeO_3$ predicted XRD



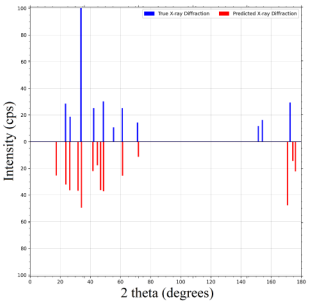
(e) $YAlO_3$ predicted XRD



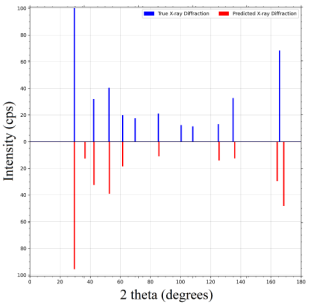
(f) $BaZrO_3$ predicted XRD



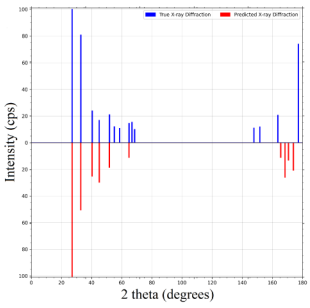
(g) $SrSeO_3$ predicted XRD peak



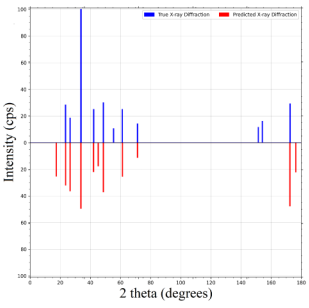
(h) $YAlO_3$ predicted XRD peak



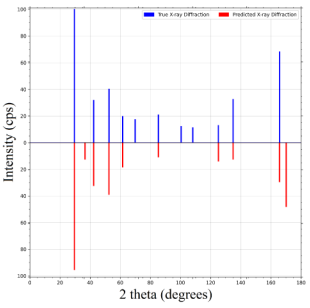
(i) $BaZrO_3$ predicted XRD peak



(j) $BaSnO_3$ alignment XRD peak



(k) $YAlO_3$ alignment XRD peak



(l) $BaZrO_3$ alignment XRD peak

Figure 9. Prediction performance of peak positions by DeepXRD. Structure and predicted peak positions of (a, d, g, j) $BaSnO_3$, (b, e, h, k) $YAlO_3$, and (c, f, i, l) $BaZrO_3$.

distance that can be aligned by the shifting operation of the peak alignment process. As shown in Figure 10k, the trends of the other peaks we predicted are the same as the ground truth peaks. The last cubic test case $MgAgF_3$ is shown in Figure 10i. We find that due to fewer training samples containing elements Ag and Mg, the positions of the two highest peaks of $MgAgF_3$ (Figure 10l) are predicted with larger offsets than in previous examples. However, the positions of other peaks and intensities are still very close to the ground truth.

On the larger Ternary-XRD data set, we also choose several test samples with different elements and crystal systems to evaluate our DeepXRD model's performance. Their predicted

https://doi.org/10.1021/acsami.2c05812
ACS Appl. Mater. Interfaces 2022, 14, 40102−40115

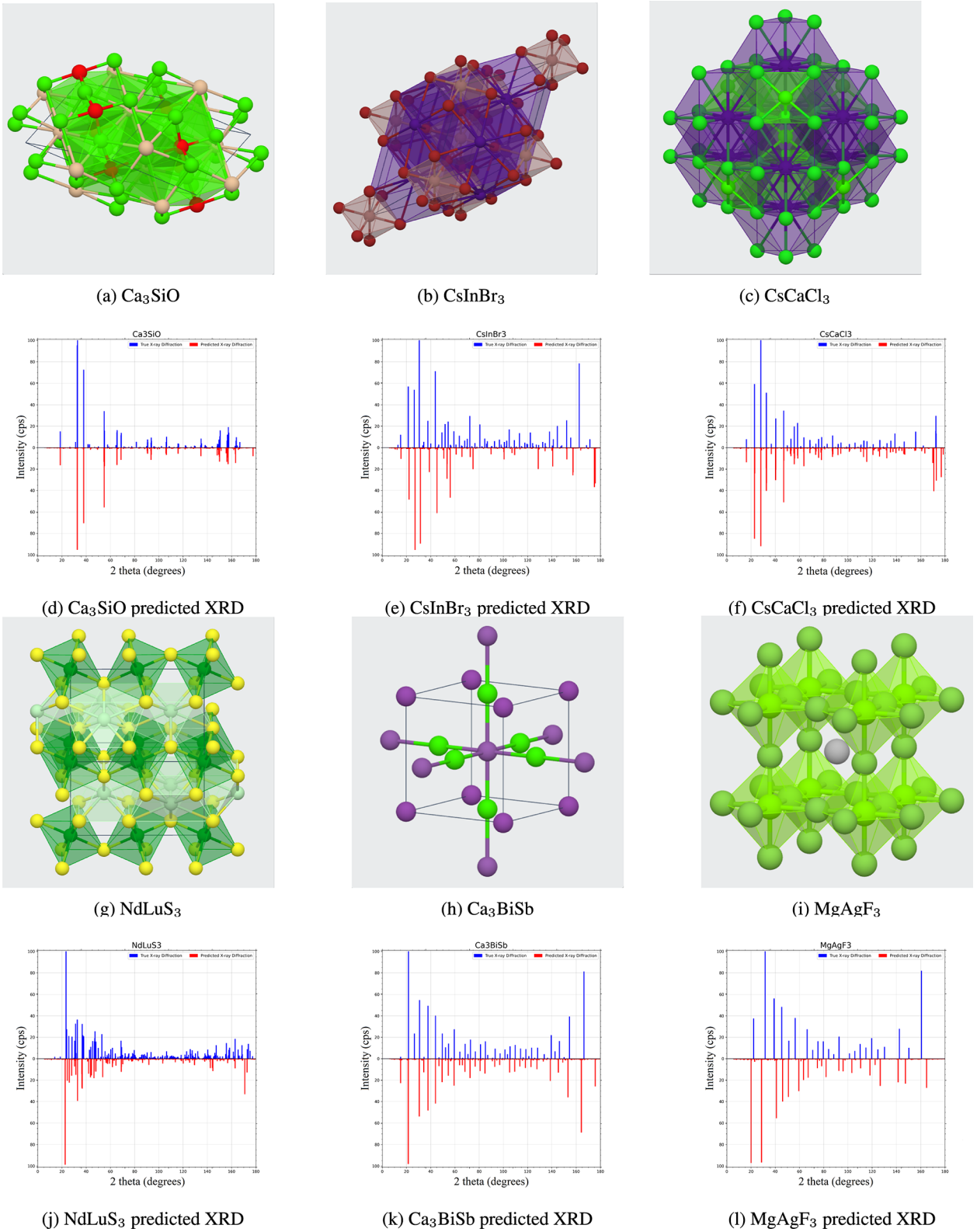


Figure 10. Prediction performance of DeepXRD. Structure and predicted XRD of (a, d) Ca_3SiO, (b, e) $CsInBr_3$, (c, f) $CsCaCl_3$, (g, j) $NdLuS_3$, (h, k) Ca_3BiSb, and (i, l) $MgAgF_3$.

XRD (red color) and truth XRD spectra (blue color) are shown in Figure 11. Comparing Figure 11 with Figure 10, we can find that the peaks of these ternary materials are more complex than ABC_3. It is also found that our predicted XRD spectra are denser than those of the ABC_3-XRD data set.

As shown in Figure 11a, $Fe_3(OF_2)_2$ is monoclinic. Its highest peak is located around 25°, and most of the peaks are within the first 90°. Our predicted XRD matches almost all peaks of the first half and misses those peaks within the interval of 160 to 180° again as shown in Figure 11d. The next test case is

https://doi.org/10.1021/acsami.2c05812
ACS Appl. Mater. Interfaces 2022, 14, 40102−40115

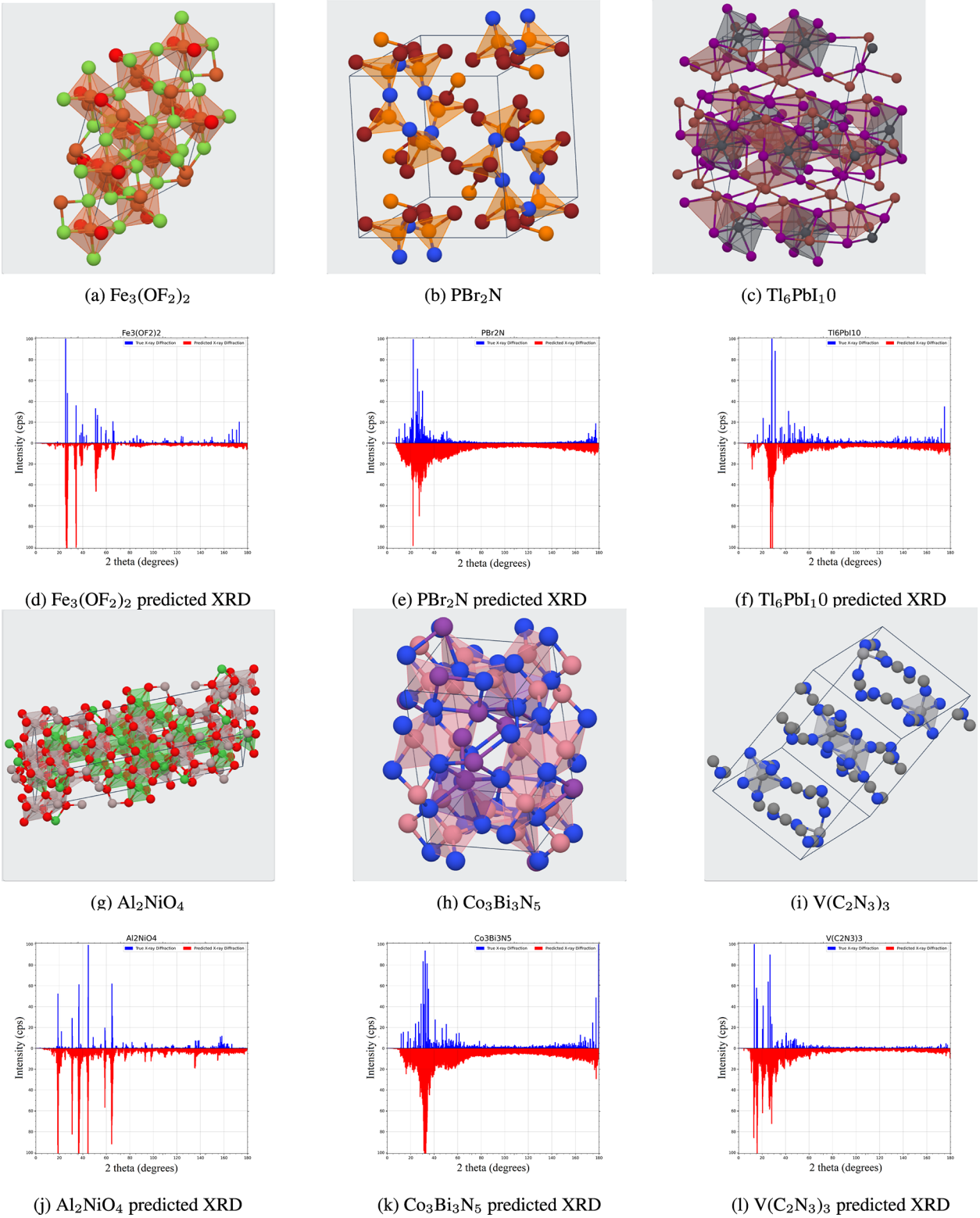


Figure 11. Prediction performance of DeepXRD. Structure and predicted XRD of (a, d) $Fe_3(OF_2)_2$, (b, e) PBr_2N, (c, f) Tl_6PbI_{10}, (g, j) Al_2NiO_4, (h, k) $Co_3Bi_3N_5$, and (i, l) $V(C_2N_3)_3$.

PBr_2N with a triclinic structure as shown in Figure 11b. PBr_2N's highest peak position is around 20° followed by a series of peaks with intensity larger than 40 cps. Figure 11e shows that almost all true peaks are located within the interval of 20 to 40°, the same as the predicted ones. Figure 11c shows the structure of Tl_6PbI_{10}, which has a trigonal crystal system. This crystal has three very close main peaks around 30° (Figure 11f). Our model accurately predicts the positions and intensities of all these peaks. Another monoclinic test case Al_2NiO_4 is shown in Figure 11g. Its peaks are discrete within the first 90° as shown in Figure 11 j. Our predicted XRD could match almost all the peak positions with intensity errors for

https://doi.org/10.1021/acsami.2c05812
ACS Appl. Mater. Interfaces 2022, 14, 40102−40115

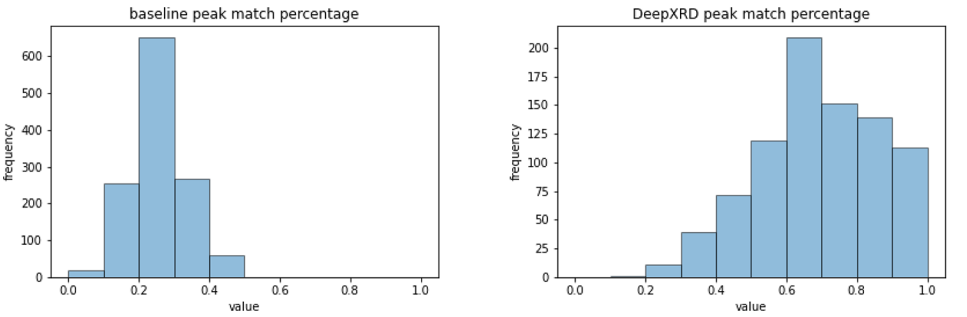
Table 6. Prediction Performance (Peak Position Match Percentage) of DeepXRD and Baseline Model on the ABC_3-XRD Data Set

methods	average match percentage	max match percentage
DeepXRD	0.68	N/A
baseline model	0.26	0.73

only a few peaks. Figure 11h shows a triclinic test case $Co_3Bi_3N_5$, which has a series of close high peaks and two single peaks around 180°. Our predicted XRD matches the first series of peaks very well as shown in Figure 11k. The trends of the other predicted peaks are also similar to the true ones. The last case is the orthorhombic $V(C_2N_3)_3$ as shown in Figure 11I with the peaks gathered between 0 and 30° and our predicted XRD spectrum has the same distribution (Figure 11l).

3.4. Discussion. From all three case studies discussed before, we show that for a given material formula, our DeepXRD model can predict its probable XRD spectrum only based on its composition. Even though the predicted peaks may not be at the exact positions compared to the ground truths, they are within the minor shifting range. For test cases with good performance, our model can find most peak positions and corresponding values but a few peaks may still not match, which may be caused by infrequent elemental combinations that only appear in limited times during model training. In machine learning studies, models trained with a larger data set usually achieve better prediction performance. However, in our XRD study, the predicted XRD peaks of the test samples in the small ABC_3-XRD data set to match their target XRD spectra better. This is due to the fact that the smaller ABC_3-XRD data set contains more similar compositions and structures compared to the samples of the Ternary-XRD data set, which are much more diverse. During model training, we find that our model can easily overfit, which is probably due to the XRD spectrum data containing noise and being sensitive to its composition change, which means that even if there is one elemental change on the input formula, the corresponding XRD spectrum may change dramatically. If the training process focuses too much on the training set, the model tends to adapt to the details or even noises within the XRD spectra of the training set, which makes the model to be not generalizable to the test samples. To deal with this overfitting problem, we add dropout layers in our model to control model overfitting and use early stopping to avoid our model's overfitting. We also find that the performance of our XRD prediction models may be significantly improved by designing a smoother loss function: instead of directly comparing the magnitudes at the sampling points, a loss function that allows a certain degree of peak shifting should lead to a smoother loss landscape so that similar compositions can have similar distribution despite some those peak shifts.

To further estimate our DeepXRD's performance compared with other methods, we established a baseline using Pymatgen,[49] and showed the prediction performance of both DeepXRD and the baseline model on the ABC_3-XRD data set. The input of the baseline model is the corresponding structures of formulas in the ABC_3-XRD data set and we download it from the Materials Project database.[42] Then we can get the particular A, B, and C of each formula, elements that appear once in the formula are considered to be A or B, and the element that appears 3 times are treated as C. Take $SrSiO_3$ as an example, the A and B are Sr and Si (the order does not matter), and C is O. After we have this A\B\C element information, we then use the element substitution method to replace elements from known structures with other elements to get new structures. If we substitute Si in $SrSiO_3$ to Zr, we can get a new unstable structure of $SrZrO_3$. Due to the substituted atom may have different volumes from the original one, this new structure needs to be adjusted based on atom information. After structure adjustment, we can use the XRD calculator function in Pymatgen to get the probable XRD pattern of this substituted new structure. And this calculated XRD pattern can then be applied with our peak selection and alignment operation to get the final peak positions and intensities. The prediction performance results are shown in Table 6. The average match percentage shows the peak match percentage of the target XRD pattern and the calculated XRD pattern based on a randomly selected structure with the prototype ABC_3. The max match percentage shows the best possible peak match percentage when we select the most similar structure to the target one. Table 6 shows that when randomly selecting a structure as the template to apply element substitution, the final XRD peak position match percentage is only 0.26, which is much smaller than DeepXRD's 0.68. And the max match percentage is 0.73, which means even if we spend a lot of computational cost to find a similar structure to the target formula, we can only make a 0.05 improvement. The average peak match percentage distributions of the two methods are shown in Figure 12. Figure 12a shows the average peak match percentage distribution of the baseline model, most percentages are located in the 0.1 to 0.4 interval, and almost all of them are smaller than 0.5. Figure 12a shows the average peak match percentage distribution of the



(a) baseline average peak match percentage (b) DeepXRD average peak match percentage

Figure 12. Average peak match percentage distribution of baseline and DeepXRD.

https://doi.org/10.1021/acsami.2c05812
ACS Appl. Mater. Interfaces 2022, 14, 40102−40115

DeepXRD model, most percentages are located in the 0.6 to 1.0 interval, and only a few of them are smaller than 0.4. This distribution figure also supports that XRD patterns predicted by the DeepXRD method are more similar to the ground truth.

4. CONCLUSION

We propose a deep neural network-based model for predicting materials' XRD spectra given their composition only. These models can be used to conduct high-throughput screening of the almost infinite composition design space for structures with specific structural features or symmetry. When evaluated on two data sets with a more homogeneous ABC_3-XRD data set and a larger Ternary-XRD data set with more diverse structures, we show that our DeepXRD algorithm can make an accurate prediction of XRD spectra for a large category of material formulas. When we want to find materials with target XRD spectra, we can use the DeepXRD model as preliminary screening to narrow down the candidates. Based on the predicted XRD spectra, we may further estimate the material's structure. Based on our successful case studies, we believe that our DeepXRD model and its future variants are of great significance to be used for guiding the discovery of new materials.

AUTHOR INFORMATION

Corresponding Author

Jianjun Hu − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States;* orcid.org/0000-0002-8725-6660; Email: jianjunh@cse.sc.edu

Authors

Rongzhi Dong − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Yong Zhao − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Yuqi Song − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Nihang Fu − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Sadman Sadeed Omee − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Sourin Dey − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Qinyang Li − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Lai Wei − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States;* orcid.org/0000-0003-0344-8540

Complete contact information is available at: https://pubs.acs.org/10.1021/acsami.2c05812

Author Contributions

Conceptualization, J.H.; methodology, J.H. and R.D.; software, R.D., Y.Z., and J.H.; validation, R.D. and J.H.; investigation, R.D., J.H., N.F., S.O., S.D., Q.L., L.W.; resources, J.H.; writing−original draft preparation, R.D. and J.H.; writing−review and editing, J.H; visualization, R.D.; supervision, J.H.; funding acquisition, J.H.

Notes

The authors declare no competing financial interest.

The two benchmark data sets that support the findings of this study can be downloaded from figshare.com at https://figshare.com/account/projects/141044/articles/20021369 and https://figshare.com/account/projects/141044/articles/20022359. Source code is available from the corresponding author upon reasonable request.

ACKNOWLEDGMENTS

Research reported in this work was supported in part by NSF under grant and 1940099, 2110033, and 1905775 and by NSF SC EPSCoR Program under NSF Award OIA-1655740. The views, perspective, and content do not necessarily represent the official views of the SC EPSCoR Program nor those of the NSF.

REFERENCES

(1) Hattrick-Simpers, J. R.; Gregoire, J. M.; Kusne, A. G. Perspective: composition-structure-property mapping in high-throughput experiments: turning data into knowledge. *APL Materials* **2016**, *4*, 053211.

(2) Arroyave, R.; McDowell, D. L. Systems approaches to materials design: past, present, and future. *Annu. Rev. Mater. Res.* **2019**, *49*, 103−126.

(3) Dan, Y.; Zhao, Y.; Li, X.; Li, S.; Hu, M.; Hu, J. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Computational Materials* **2020**, *6*, 1−7.

(4) Glass, C. W.; Oganov, A. R.; Hansen, N. USPEX-Evolutionary crystal structure prediction. *Computer physics communications* **2006**, *175*, 713−720.

(5) Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. CALYPSO: A method for crystal structure prediction. *Comput. Phys. Commun.* **2012**, *183*, 2063−2070.

(6) Wei, L.; Fu, N.; Siriwardane, E.; Yang, W.; Omee, S. S.; Dong, R.; Xin, R.; Hu, J. TCSP: a Template based crystal structure prediction algorithm and web server for materials discovery. *arXiv*, 2021, 2111.14049.

(7) Kusaba, M.; Liu, C.; Yoshida, R. Crystal structure prediction with machine learning-based element substitution. *arXiv*, 2022, 2201.11188.

(8) Oviedo, F.; Ren, Z.; Sun, S.; Settens, C.; Liu, Z.; Hartono, N. T. P.; Ramasamy, S.; DeCost, B. L.; Tian, S. I. P.; Romano, G.; Gilad Kusne, A.; Buonassisi, T.; et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Computational Materials* **2019**, *5*, 1−9.

(9) Lee, B. D.; Lee, J.-W.; Park, W. B.; Park, J.; Cho, M.-Y.; Pal Singh, S.; Pyo, M.; Sohn, K.-S. Powder X-Ray Diffraction Pattern Is All You Need for Machine-Learning-Based Symmetry Identification and Property Prediction. *Advanced Intelligent Systems* **2022**, *4*, 2200042.

(10) Suzuki, Y.; Hino, H.; Hawai, T.; Saito, K.; Kotsugi, M.; Ono, K. Symmetry prediction and knowledge discovery from X-ray diffraction patterns using an interpretable machine learning approach. *Sci. Rep.* **2020**, *10*, 1−11.

(11) 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*, 5260.

(12) Stanev, V.; Vesselinov, V. V.; Kusne, A. G.; Antoszewski, G.; Takeuchi, I.; Alexandrov, B. S. Unsupervised phase mapping of X-ray

diffraction data by nonnegative matrix factorization integrated with custom clustering. *npj Computational Materials* **2018**, *4*, 1−10.
(13) Harris, K. D. M.; Tremayne, M.; Kariuki, B. M. Contemporary advances in the use of powder X-ray diffraction for structure determination. *Angew. Chem., Int. Ed.* **2001**, *40*, 1626−1651.
(14) Sikam, P.; Moontragoon, P.; Ikonic, Z.; Kaewmaraya, T.; Thongbai, P. The study of structural, morphological and optical properties of (Al, Ga)-doped ZnO: DFT and experimental approaches. *Appl. Surf. Sci.* **2019**, *480*, 621−635.
(15) Khalid, M.; Ullah, M. A.; Adeel, M.; Usman Khan, M.; Tahir, M. N.; Braga, A. A. C. Synthesis, crystal structure analysis, spectral IR, UV-Vis, NMR assessments, electronic and nonlinear optical properties of potent quinoline based derivatives: interplay of experimental and DFT study. *Journal of Saudi Chemical Society* **2019**, *23*, 546−560.
(16) Rajan, K. Materials informatics. *Mater. Today* **2005**, *8*, 38−45.
(17) Ward, L.; Wolverton, C. Atomistic calculations and materials informatics: A review. *Curr. Opin. Solid State Mater. Sci.* **2017**, *21*, 167−176.
(18) Gomez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernandez-Lobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* **2018**, *4*, 268−276.
(19) Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Science advances* **2018**, *4* (3), eaap7885.
(20) Lu, S.; Zhou, Q.; Ouyang, Y.; Guo, Y.; Li, Q.; Wang, J. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat. Commun.* **2018**, *9* (1), 1−8.
(21) Collins, Sean P; Daff, Thomas D; Piotrkowski, Sarah S; Woo, Tom K Materials design by evolutionary optimization of functional groups in metal-organic frameworks. *Science advances* **2016**, *2*, e1600954.
(22) Liang, H.; Stanev, V.; Kusne, A. G.; Takeuchi, I. CRYSPNet: Crystal Structure Predictions via Neural Network. *arXiv*, 2020, 2003.14328.
(23) Zhao, Y.; Cui, Y.; Xiong, Z.; Jin, J.; Liu, Z.; Dong, R.; Hu, J. Machine Learning-Based Prediction of Crystal Systems and Space Groups from Inorganic Materials Compositions. *ACS omega* **2020**, *5*, 3596−3606.
(24) Takahashi, K.; Takahashi, L.; Baran, J. D.; Tanaka, Y. Descriptors for predicting the lattice constant of body centered cubic crystal. *J. Chem. Phys.* **2017**, *146*, 204104.
(25) Swetlana, S.; Khatavkar, N.; Singh, A. K. Development of Vickers hardness prediction models via microstructural analysis and machine learning. *J. Mater. Sci.* **2020**, *55*, 15845−15856.
(26) Khatavkar, N.; Swetlana, S.; Singh, A. K. Accelerated prediction of Vickers hardness of Co-and Ni-based superalloys from microstructure and composition using advanced image processing techniques and machine learning. *Acta Mater.* **2020**, *196*, 295−303.
(27) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2016**, *2*, 1−7.
(28) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E.R.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **2018**, *152*, 60−69.
(29) Goodall, R. E. A.; Lee, A. A. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *arXiv*, 2019, 1910.00617.
(30) Suzuki, Y.; Hino, H.; Takeichi, Y.; Hawai, T.; Kotsugi, M.; Ono, K. Machine Learning-based Crystal Structure Prediction for X-Ray Microdiffraction. *Microscopy and Microanalysis* **2018**, *24*, 144−145.
(31) 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.
(32) Wang, H.; Xie, Y.; Li, D.; Deng, H.; Zhao, Y.; Xin, M.; Lin, J. Rapid identification of X-ray diffraction patterns based on very limited data by interpretable convolutional neural networks. *J. Chem. Inf. Model.* **2020**, *60*, 2004−2011.
(33) Kaufmann, K.; Zhu, C.; Rosengarten, A. S.; Maryanovsky, D.; Harrington, T. J.; Marin, E.; Vecchio, K. S. Crystal symmetry determination in electron diffraction using machine learning. *Science* **2020**, *367*, 564−568.
(34) Lee, J.-W.; Park, W. B.; Lee, J. H.; Singh, S. P.; Sohn, K.-S. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns. *Nat. Commun.* **2020**, *11*, 86.
(35) Xiong, Z.; He, Y.; Hattrick-Simpers, J. R.; Hu, J. Automated phase segmentation for large-scale X-ray diffraction data using a graph-based phase segmentation (GPhase) algorithm. *ACS Comb. Sci.* **2017**, *19*, 137−144.
(36) Li, S.; Xiong, Z.; Hu, J. Inferring phase diagrams from X-ray data with background signals using graph segmentation. *Mater. Sci. Technol.* **2018**, *34*, 315−326.
(37) Ozaki, Y.; Suzuki, Y.; Hawai, T.; Saito, K.; Onishi, M.; Ono, K. Automated crystal structure analysis based on blackbox optimization. *npj Computational Materials* **2020**, *6*, 1−7.
(38) Vecsei, P. M.; Choo, K.; Chang, J.; Neupert, T. Neural network based classification of crystal symmetries from x-ray diffraction patterns. *Phys. Rev. B* **2019**, *99*, 245120.
(39) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, 2016; pp 770−778.
(40) He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: Piscataway, NJ, 2015; pp 1026−1034.
(41) Lu, L., Shin, Y., Su, Y., Karniadakis, G. E. Dying relu and initialization: Theory and numerical examples. *arXiv*, 2019, 1903.06733.
(42) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **2013**, *1* (1), 011002.
(43) van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **2008**, *9*, 2579−2605.
(44) Inoue, M.; Hirasawa, I. The relationship between crystal morphology and XRD peak intensity on casO4.2h2o. *Journal of crystal growth* **2013**, *380*, 169−175.
(45) Iwasaki, Y.; Kusne, A. G.; Takeuchi, I. Comparison of dissimilarity measures for cluster analysis of X-ray diffraction data from combinatorial libraries. *npj Computational Materials* **2017**, *3*, 1−9.
(46) Hernandez-Rivera, E.; Coleman, S. P.; Tschopp, M. A. Using similarity metrics to quantify differences in high-throughput data sets: application to X-ray diffraction patterns. *ACS combinatorial science* **2017**, *19*, 25−36.
(47) Salvador, S.; Chan, P. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis* **2007**, *11*, 561−580.
(48) Shiah, R. T. S.; Vook, R. W. Kinematic theory of the effect of surface relaxation on X-ray diffraction peak shifts. *Surf. Sci.* **1973**, *38*, 357−372.
(49) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314−319.