JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

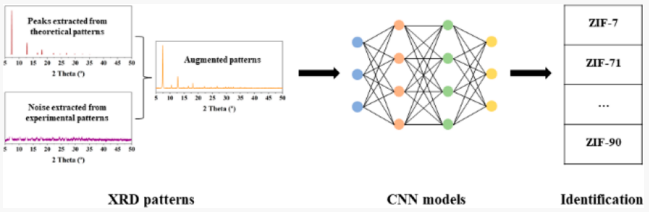
Rapid Identification of X-ray Diffraction Patterns Based on Very Limited Data by Interpretable Convolutional Neural Networks

Hong Wang, Yunchao Xie, Dawei Li, Heng Deng, Yunxin Zhao, Ming Xin, and Jian Lin*

Cite This: J. Chem. Inf. Model. 2020, 60, 2004−2011 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Large volumes of data from material characterizations call for rapid and automatic data analysis to accelerate materials discovery. Herein, we report a convolutional neural network (CNN) that was trained based on theoretical data and very limited experimental data for fast identification of experimental X-ray diffraction (XRD) patterns of metal−organic frameworks (MOFs). To augment the data for training the model, noise was extracted from experimental data and shuffled; then it was merged with the main peaks that were extracted from theoretical spectra to synthesize new spectra. For the first time, one-to-one material identification was achieved. Theoretical MOFs patterns (1012) were augmented to a whole data set of 72 864 samples. It was then randomly shuffled and split into training (58 292 samples) and validation (14 572 samples) data sets at a ratio of 4:1. For the task of discriminating, the optimized model showed the highest identification accuracy of 96.7% for the top 5 ranking on a test data set of 30 hold-out samples. Neighborhood component analysis (NCA) on the experimental XRD samples shows that the samples from the same material are clustered in groups in the NCA map. Analysis on the class activation maps of the last CNN layer further discloses the mechanism by which the CNN model successfully identifies individual MOFs from the XRD patterns. This CNN model trained by the data augmentation technique would not only open numerous potential applications for identifying XRD patterns for different materials, but also pave avenues to autonomously analyze data by other characterization tools such as FTIR, Raman, and NMR spectroscopies.

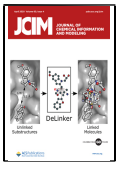


INTRODUCTION

High-throughput-synthesis techniques have shown great potential in accelerating material innovation.[1] Large volumes of characterization data including X-ray diffraction (XRD), Raman, nuclear magnetic resonance (NMR), and Fourier transform infrared (FTIR) patterns are collected during or after the synthesis. Among them, XRD is a powerful technique to characterize crystallographic structures, grain size, and molecular structures.[2] Typically, experimental XRD samples are analyzed via comparing descriptors such as peak positions, intensities, and full widths at half-maxima (FWHM) against a known database such as the Crystallography Open Database and Inorganic Crystal Structure Database, allowing scientists to identify the compounds of interest and to map phase diagrams of combinatorial materials. However, the tedious and time-consuming procedure due to the manual analysis at a relatively low speed severely hinders fast decision-making.[2,3] To fully exploit the characterization tools, it is becoming urgent to develop new data assessment tools with automation and recommendation functions, especially with the emergence of self-driven laboratories enabled by robots.[4−6] Despite recent progress, it has been and continues to be a grand challenge.

Recently, machine learning (ML) models have shown great potential in managing the large volumes of characterization data for rapidly and automatically identifying composition−phase maps as well as constructing composition−structure−property relationships, thereby speeding up the materials discovery.[1,7−15] For instance, Iwasaki et al., Kusne et al., Stanev et al., Aguiar et al., and Yoon et al. implemented machine learning techniques such as cluster analysis and mean shift for phase or crystallographic classification.[16−20] Ziatdinov et al. applied deep learning to resolve scanning transmission electron microscopy images.[21] Lee et al. demonstrated a deep learning technique for application in phase identification of inorganic compunds.[22] Park et al. demonstrated well-trained convolutional neural networks (CNNs) which exhibited satisfactory accuracy in classifying XRD patterns based on a theoretical database.[23] Oviedo and colleagues proposed a machine learning approach to predict crystallographic dimensionality and space groups from a limited number of thin-film XRD patterns.[2] Ziletti's research group developed a robust CNN model to classify crystal structures and also unfolded the internal behavior of the classification model through visualization.[14] Miller's research group implemented a CNN to

Received: January 8, 2020
Published: March 25, 2020



https://dx.doi.org/10.1021/acs.jcim.0c00020
J. Chem. Inf. Model. 2020, 60, 2004−2011

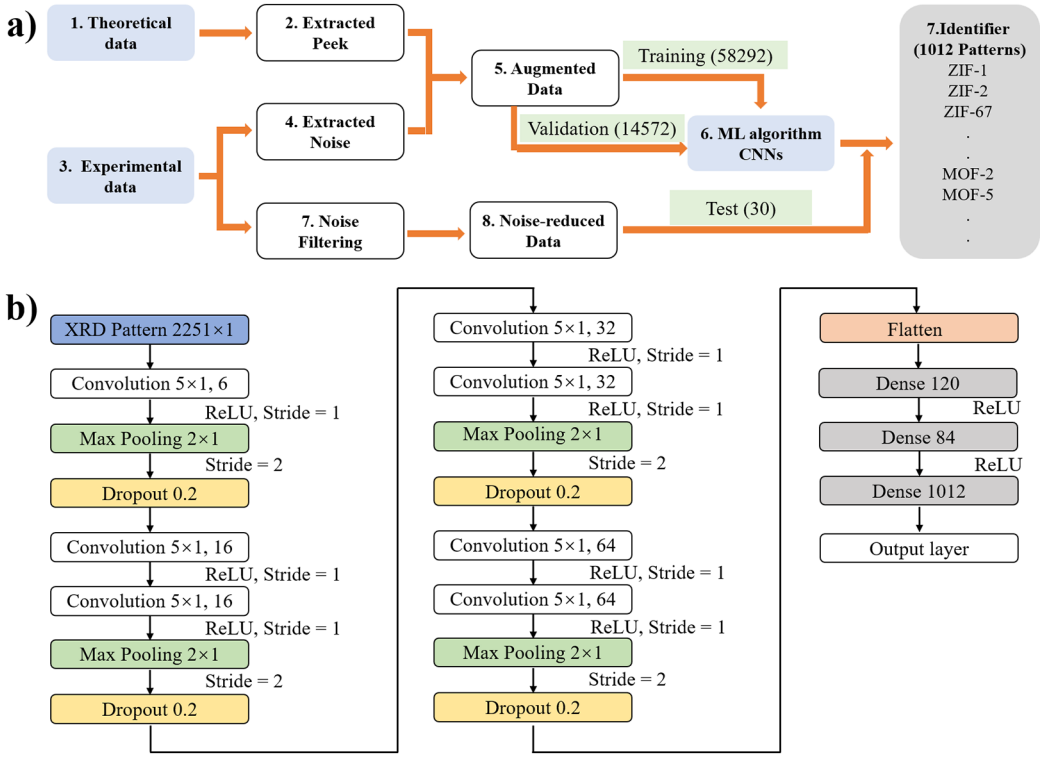


Figure 1. (a) Flowchart showing the process of XRD pattern identification. (b) Architecture of proposed convolutional neural network.

determine crystallography trained on imaging and diffraction data.[15] However, these approaches were applied to identify several classes or crystal systems into which target materials are grouped. One-by-one identification of individual spectrum from millions of spectrum databases is still challenging. Another big challenge for developing machine learning enabled methodology is the lack of experimental data for training the models. Although a technique of Gaussian mixture was employed to augment the theoretical data,[24] it may not fully reflect the real experiments when distinguished features arise from the experiments. It is envisioned that directly incorporating experimental data into theoretical data is a better approach. Finally, the deep learning models like CNN are usually treated as a "black box". Interpreting the underlying mechanism of such a black box for decision-making or obtaining the final desired results is still an open problem. Therefore, developing a procedure that can better interpret the deep learning models when they are applied to material research has recently seen a resurgence.

In this paper, we propose a CNN model that was trained for rapid one-to-one identification of experimental XRD samples of metal−organic frameworks (MOFs). To increase robustness of the CNN model, noise was extracted from the experimental spectra to augment the theoretical spectra for training. In the cases of very noisy experimental spectra, the fast Fourier transform (FFT) was applied to reduce the noise before they were input into the CNN for improving the prediction accuracy. Theoretical MOF patterns (1012) were augmented to a whole data set of 72 864 samples. It was then randomly shuffled and split into training (58 292 samples) and validation (14 572 samples) data sets at a ratio of 4:1. For the task of discriminating, the optimized model showed the highest identification accuracy of 96.7% for the top 5 ranking on a test data set of 30 hold-out samples. The training, validation, and testing data allocation is illustrated in Figure S1. Data dimension reduction analysis on the experimental XRD samples by neighborhood component analysis (NCA) shows that the samples from the same MOF are clustered in individual groups in the NCA map, while the XRD samples from different MOFs but with very similar characteristics may have overlapping. Further analysis on the class activation maps (CAMs) of the last layer before the flatten layer of the CNN model shows that the grouped samples that are highly distinguishable in the NCA map exhibit very different activation characteristics. This observation can well explain why the CNN can identify individual spectra from the library.

Our work can be summarized as follows. First, to the best of our knowledge, this is the first demonstration that a CNN enables one-by-one identification of XRD patterns for individual materials. The previously reported machine learning algorithms only classify several classes or crystal systems into which target materials are grouped. Second, the model was trained by theoretical data combined with very limited experimental data. Third, the noise-based data augmentation technique is very easy and straightforward to implement, but it results in very effective outcomes. Fourth, the trained CNN model can successfully and robustly perform one-by-one classification with the help of a noise filtering procedure even though the experimental XRD samples exhibit peak shift, scaling in peak intensities, or FWHM broadening compared to the theoretical spectra. Last but not least, the study on the CAMs of the convolutional layers discloses the mechanism of how the CNN model makes the decision, thus shedding new light on the interpretable deep learning for materials characterization data analysis. Consequently, the proposed solution is of great interest and appears to be very promising, not only because of the applications of XRD to characterize different types of materials, but also because of the possible

https://dx.doi.org/10.1021/acs.jcim.0c00020
J. Chem. Inf. Model. 2020, 60, 2004−2011

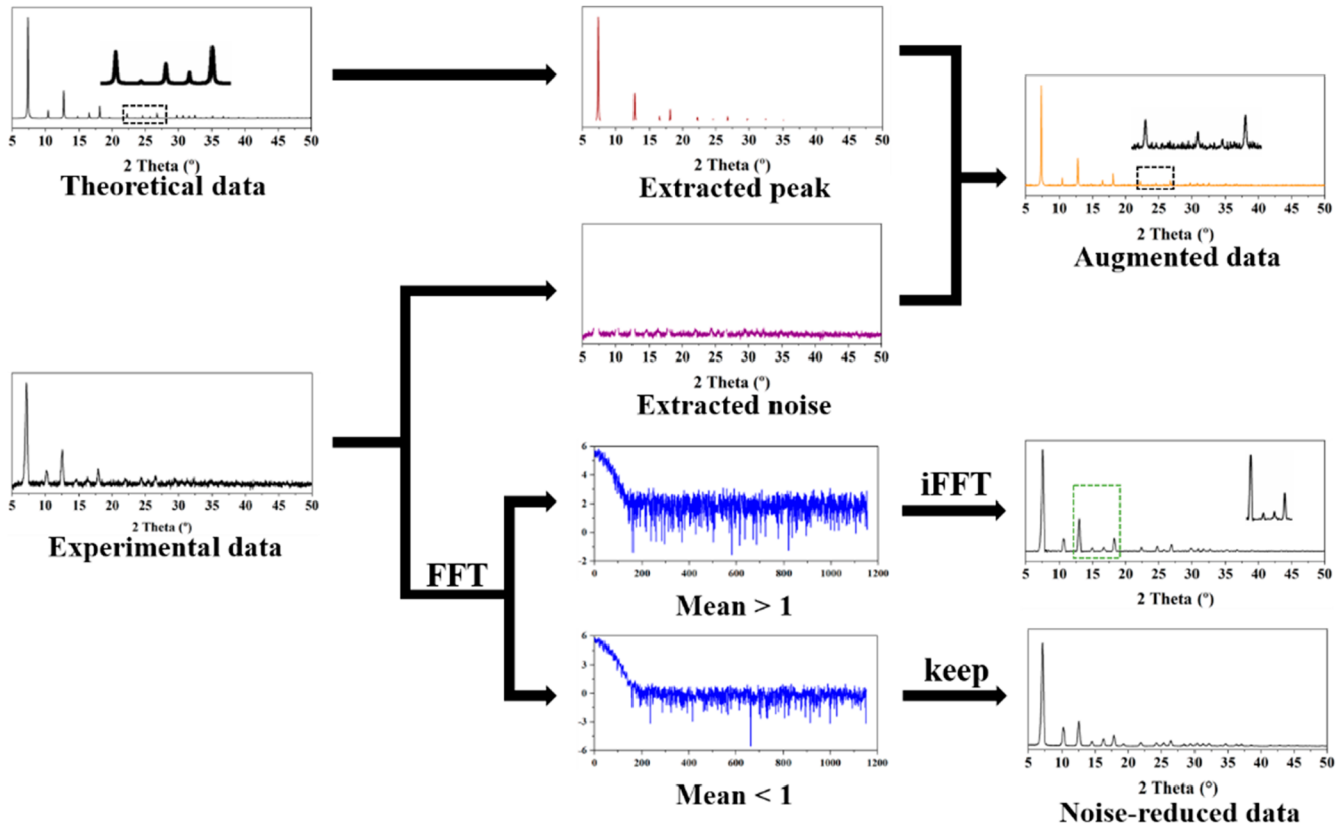


Figure 2. Flowchart showing process of augmenting theoretical data and filtering noise of experimental data.

extension to samples collected by other characterization techniques including Raman, NMR, and FTIR spectroscopies.

RESULTS AND DISCUSSION

The flowchart showing the procedure for rapid identification of XRD patterns enabled by the CNN is illustrated in Figure 1a. First, theoretical CIF files of MOFs downloaded from Cambridge Crystallographic Data Centre (CCDC)[25] were converted into theoretical XRD patterns. Experimental XRD samples were collected from as-synthesized MOFs powders in a Bruker D8 Advance. Detailed synthesis and characterization can be found in the Experimental Section in the Supporting Information. Then, the noise was extracted from experimental data and shuffled, and then merged with the main peaks that were extracted from theoretical data to obtain new synthesized data. By this data augmentation method, sufficient training data were realized. The experimental XRD samples that were used as the testing data sets were filtered to reduce the noise level if needed. The detailed procedure of data augmentation and noise reduction is described in the following paragraph. A CNN was built from scratch based on famous networks whose fully connected layers are from Lenet5 and convolutional blocks are from VGG16.[26,27] Its architecture is shown in Figure 1b and Table S1. Basically, the CNN consists of one input layer, four convolutional layers, three fully connected layers, and one output layer.[28,29] The first layer inputs are synthesized XRD samples with 2θ ranging from 5 to 50°. Then the data are fed subsequently into four convolution layers with kernel filters followed by a "max pooling" layer. The kernels for these convolutional layers are 6, 16, 32, and 64 filters with a size of 5 × 1 and a stride of 1. The max pooling layer has two filters with a stride of 2 and a dropout with a dropout rate of 0.2.[29−31] All convolutional layers were activated by the function of the rectified linear unit (ReLU). After one flatten layer, the data were fed to three dense layers with sizes of 120, 84, and 1012, respectively. The Adam optimizer was applied to minimize the categorical cross-entropy loss function.[32,33] The hyperparameters of the CNN are shown in Table S2.[32]

Figure 2 exhibits the detailed procedure of preprocessing and augmenting theoretical data, and reducing the noise level of experimental data. The data used for training were synthesized by merging extracted peaks from the theoretical data and baseline noise from the experimental data. The main peaks were extracted from a theoretical spectrum containing the largest 400 points. The noise was collected from the baseline of raw experimental data after the main peaks were removed. As shown in Figure S2, it is found that the noise extracted from experimental XRD patterns does not follow Gaussian or Poisson distributions. To well preserve the noise characteristics, we randomly shuffled the noise samples, which were combined with the theoretical XRD pattern for data augmentation. Then these two components were superimposed to form a new spectrum. Since the noise can be randomly sampled and shuffled, the data can be largely augmented for training the CNN model. These training data were synthesized from a library of a total of 1012 theoretical patterns. Next, FFT and inverse FFT (iFFT) were applied to smooth the experimental data for the purpose of reducing their noise level. A total of 30 experimental samples collected from 10 types of MOFs after noise filtering were used as the testing data sets. FFT converts the signal in an original domain to a representation in a frequency domain.[34] After the raw XRD

https://dx.doi.org/10.1021/acs.jcim.0c00020
J. Chem. Inf. Model. 2020, 60, 2004−2011

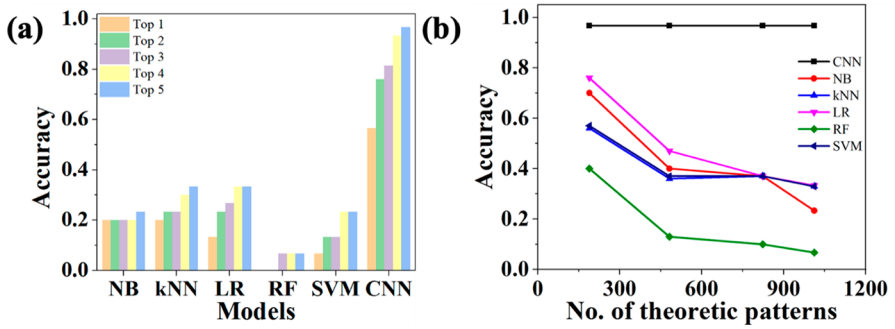


Figure 3. (a) Comparison of various ML models for identifying XRD patterns among the library with 1012 MOFs. NB, naïve Bayes; kNN, *k*-nearest neighbors; LR, logistic regression; RF, random forest; SVM, support vector machine; CNN, convolutional neural network. (b) Top 5 accuracy of various ML models trained with different numbers of theoretical patterns.

data were converted to FFT data in the frequency range of 0−200 discrete frequency bins, the data beyond the frequency of 200 is the white noise. Here, we use the mean peak value of the white noise as a criterion to determine whether iFFT should be applied to reduce the noise of experimental XRD samples. If the mean value is larger than 1, it means that the noise has high intensity and the iFFT is applied to the first original 200 discrete frequency bins with the rest of bins set to 0. Otherwise, the original data can be directly used as testing data. Compared to previous reports on the procedure of preprocessing XRD samples, only the threshold for noise filtering needs to be set by humans and intervention steps such as background removal, smoothing and interpolation, region exclusion, and peak shifting were not involved,[3,18] thereby significantly improving autonomy of the model for data analysis.

After the data preprocessing, the synthesized data and noise-reduced experimental data were used as the training and test data sets, respectively, to train and test various supervised ML models for evaluating one-by-one identification performance. We first tested five classical ML algorithms such as naïve Bayes (NB), *k*-nearest neighbors (kNN), logistic regression (LR), random forest (RF), and support vector machine (SVM). The hyperparameters are shown in Table S3, and their classification accuracies are shown in Figure 3a and summarized in Table S4. Here, we define top 1 to top 5 as the ranking positions of identification results of the testing data among the library consisting of a total 1012 MOFs (Table S5). For example, top 1 means that an ML model algorithm can successfully rank an MOF sample at the first position. Different from previous studies which map the samples into seven crystal systems or 230 space groups among thousands of materials,[2,23] it is much more challenging to reach the goal of one-by-one material classification. One-by-one classification mission is similar to the large-scale image-classification challenge attempted by the ImageNet, which classifies high-resolution images into 1000 different categories. It can be seen that all five classical ML models exhibit <50% identification accuracy for top 1-to-top 5 rankings. In comparison, the best result of CNN performed much better with 56.7, 76.7, 90, and 93.3% accuracies for top 1, top 2, top 3, and top 4 rankings, respectively (Table S6), and reached 96.7% accuracy for the top 5 ranking. It demonstrates that high-level hidden and meaningful features learned by the CNN help identify XRD patterns with a much higher accuracy than the classical ML algorithms.[35] In addition, the classification accuracy of top 5 over the classical ML algorithms and CNN model under different sized theoretical data is also investigated and shown in Figure 3b. It is obvious that the classification accuracies of the classical ML algorithms decreases sharply when the data size increases, whereas our CNN model is robust enough to deliver a predictive accuracy of 96.7% even when the number of theoretical patterns in the library increases from 189 to 1012 (Figure 3b). It is likely that, as the library size further increases, the prediction accuracy would be well maintained.

It is usually difficult for the deep neural networks to afford insights toward interpreting mechanism since they introduce the complexity of interactions and nonlinearities.[36] Thus, they were also known as "black boxes" for a long time. To reduce the dimensionality of the data for better understanding of how the CNN makes the decision, neighborhood component analysis (NCA) was employed to analyze the experimental XRD samples.[37] As shown in the NCA map (Figure 4), these

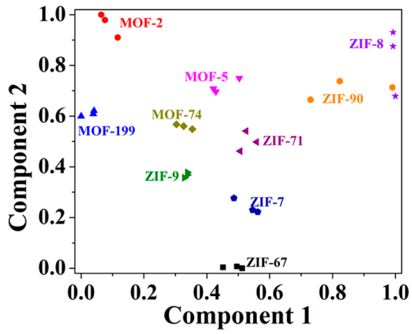


Figure 4. Neighborhood component analysis (NCA) map for clustering of XRD patterns of all 30 MOFs.

XRD samples from the same MOFs are clustered into 10 separate groups, while the XRD samples from different MOFs but with very similar pattern characteristics may result in overlapped or very close groups in the component map. This indicates that characteristics of main peaks such as position, intensity, and full width at half-maximum—usually the main criteria to distinguish the XRD patterns—are reduced to show distinguished features as shown in the NCA map.

In order to understand the mechanism of how the CNN model distinguishes individual experimental samples, their CAMs in the 14th layers and corresponding XRD patterns were shown in Figure 5 and Figure S3. Even though CAMs had some limitations reported in a recent paper,[38] in an image classification task, CAMs can still be used to reflect the main discriminative features of images, which helps to interpret and

https://dx.doi.org/10.1021/acs.jcim.0c00020
J. Chem. Inf. Model. 2020, 60, 2004−2011

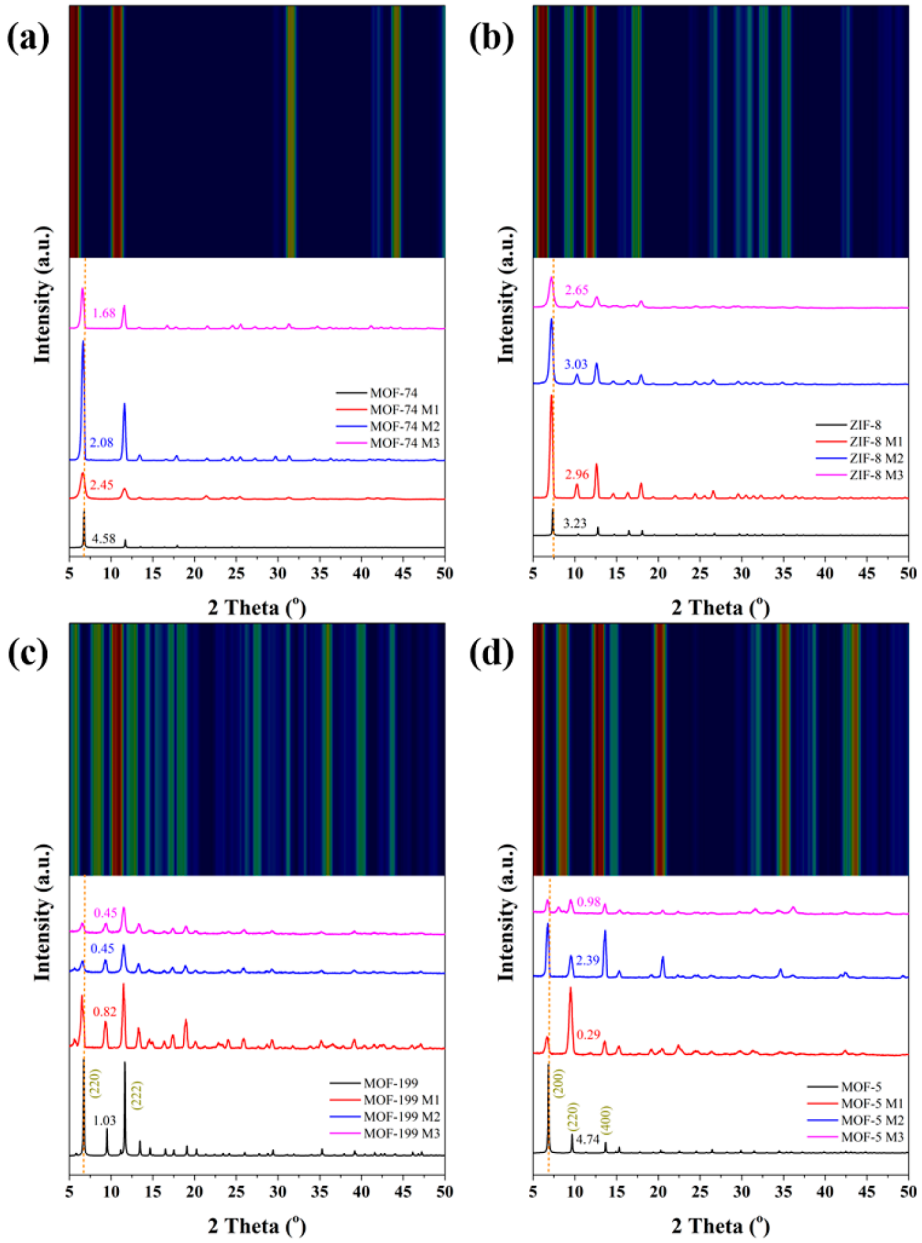


Figure 5. CAMs of the 14th layer output from the CNN model and corresponding XRD patterns: (a) MOF-74; (b) ZIF-8; (c) MOF-199; (d) MOF-5.

improve classification accuracy.[2,39] In our case, CAMs may afford a clear and direct impression on the characteristics of the XRD patterns that were the most relevant to the specific class. It is found that the red regions in the CAMs correspond to the main peaks of the XRD patterns. Hence, it is deduced that the CNN model can well distinguish XRD patterns according to the main peaks, i.e., the most important peaks. Further observation shows that MOF-74 (Figure 5a) only exhibits six colorful regions including two dominated red regions, and ZIF-8 (Figure 5b) exhibited much more colorful regions. This is similar to the way that a professional material scientist analyzes the XRD data. A traditional way to identify the compounds of interest and to map phase diagrams of combinatorial materials is to match descriptors of their XRD patterns such as peak positions, intensities, and FWHM with a known database. Thus, it is straightforward to train machine learning models via data augmentation by peak scaling, peak elimination, and peak shift.[2,23] The models successfully identify several crystal systems into which target materials are grouped. However, it is much more challenging to reach the goal of one-by-one classification, i.e., to assign a correct label to individual XRD pattern instead of seven crystal systems or 230 space groups among thousands of data sets. In our case, the trained CNN model can successfully and robustly perform one-by-one classification even though the experimental XRD samples exhibit peak shift, scaling in peak intensities, or FWHM broadening compared to the theoretical patterns. As is evident in Figure 5 and Figure S3, all XRD data of the MOFs exhibit peak shift, the existence of noise, and peak intensity scaling compared to their theoretical patterns. Our model can tackle these abnormal phenomena and reach 96.7% one-by-one classification accuracy. In addition, amazingly, the CNN can

https://dx.doi.org/10.1021/acs.jcim.0c00020
J. Chem. Inf. Model. 2020, 60, 2004−2011

still identify the patterns even when intensities of main peaks are largely changed. Take MOF-199 (Figure 5c) and MOF-5 (Figure 5d) as examples. The ratios of (220) to (222) peaks for experimental data of three MOF-199 samples are varied from 0.82 to 0.45. The trained model can identify them in the top 3 (Table S6). The CAM on MOF-199 agrees well with this result, which shows that the dominated red region shifts to (222) from (220). Similarly, the ratios of (200) to (220) for MOF-5 are 0.29, 2.40, and 0.99 for M1, M2, and M3 of MOF-5, respectively. The model also can rank MOF-5 in the top 3 (Table S6). For MOF-5 M2 sample, the peak corresponding to the (400) plane increases to be the second highest peak. This result agrees well with the CAM analysis result, and it shows that the two dominated regions correspond to the (200) and (400) peaks. In addition, we also observed that the lower crystallinity and existence of noise greatly affect the classification accuracy. Here, ZIF-9 was chosen as an example (Figure S3f). The CNN model can classify the ZIF-9 M2 and ZIF-9 M3 in the top 2 and top 4 rankings, but it cannot distinguish ZIF-9 M1. The CAM analysis shows only one red dominated region, corresponding to the first highest peak. Due to the lower intensity and the existence of noise, the CNN could not learn the main features which can effectively assign all three samples to ZIF-9.

CONCLUSION

In summary, we demonstrated a CNN model trained by the theoretical XRD patterns augmented by noise from limited experimental data for rapid one-to-one identification of individual MOFs. The CNN model is employed to identify XRD patterns instead of categorizing them into groups or crystallinity systems. The optimized CNN model showed the highest identification accuracy of 96.7% for the top 5 rankings among a data set of 1012 XRD patterns. The advantages of the proposed CNN model can be summarized as follows: (1) It is a one-by-one identification instead of predicting several crystal groups. (2) The model was trained based on very limited theoretical data. (3) Simple and straightforward noise-based data augmentation—not like the past technique that employed multistep operations (peak scaling, elimination, and shifting)—was deployed; thus it is easy to operate and requires less hyperparameter tuning. (4) The procedure of noise filtering can greatly increase the classification accuracy of the CNN model. Finally, the proposed CNN model has potential not only in numerous applications of XRD in materials science, but also in the possible expansion of the solution to several other characterization techniques such as Raman, NMR, and FTIR spectroscopies.

METHODS

Chemicals. $Zn(NO_3)_2{\cdot}6H_2O$ (Sigma-Aldrich), $ZnCl_2{\cdot}6H_2O$ (Fluka), $Zn(CH_3COO)_2{\cdot}2H_2O$ (Fisher), $Co(NO_3)_2{\cdot}6H_2O$ (Sigma-Aldrich), $CoCl_2{\cdot}6H_2O$ (Sigma-Aldrich), $Co(CH_3COO)_2{\cdot}4H_2O$ (Fisher), $Cu(NO_3)_2{\cdot}3H_2O$ (Fisher), 2-methylimidazole (Fisher), benzimidazole (Fisher), 4,5-dichloroimidazole (Fisher), 2-imidazolecarboxaldehyde (Fisher), terephthalic acid (BDC; Fisher), 1,3,5-benzenetricarboxylic acid (BTC; Fisher), cetyltrimethylammonium bromide (CTAB; Sigma-Aldrich), diethylamine (DEA; Fisher), triethylamine (TEA; Fisher), polyvinylpyrrolidone (PVP, M_w = 360 000 or 40 000; Sigma-Aldrich), *N*,*N*-dimethylformamide (DMF; Fisher), methanol (Fisher), and ethanol (Fisher) were used without any further purification.

MOFs Synthesis. Here, all MOFs were synthesized by three different methods according to the reported literature. The detailed synthesis methods are described in the Supporting Information.

Characterization. Powder X-ray diffraction (XRD) patterns were obtained on a Bruker D8 Discover diffractometer (Cu Kα, λ = 0.15406 nm).

MACHINE LEARNING MODELS

Five classical machine learning algorithms, i.e., naïve Bayes (NB), *k*-nearest neighbors (kNN), logistic regression (LR), random forest (RF), and support vector machine (SVM), were well-trained and employed for identifying XRD patterns. Feed-forward convolutional neural networks were constructed using Keras software library with the TensorFlow back end.[40] The machine learning models were performed using Python with Scikit-Learn on a high-performance computer with Intel i7-9700k CPU, 16 GB DDR4 memory at 2.4 GHz, and Nvidia EVGA GeForce RTX 2070 GPU.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00020.

Experimental section; summary of the architecture of convolutional neural networks; hyperparameters of CNN and classical ML models; comparison of classical ML and CNN models; definition of top 1 to top 5 for presenting identification results using ZIF-90 as an example; one-by-one identification of XRD spectra by a CNN model; training, validation, and testing data allocations; noise histogram extracted from experimental data; CAMs of the 14th layer output from the CNN model and corresponding XRD spectra (PDF)

AUTHOR INFORMATION

Corresponding Author

Jian Lin − *Department of Mechanical and Aerospace Engineering, Department of Electrical Engineering and Computer Science, and Department of Physics and Astronomy, University of Missouri, Columbia, Missouri 65211, United States;* orcid.org/0000-0002-4675-2529; Email: LinJian@missouri.edu

Authors

Hong Wang − *Department of Mechanical and Aerospace Engineering, University of Missouri, Columbia, Missouri 65211, United States*

Yunchao Xie − *Department of Mechanical and Aerospace Engineering, University of Missouri, Columbia, Missouri 65211, United States;* orcid.org/0000-0001-6216-1211

Dawei Li − *Department of Mechanical and Aerospace Engineering, University of Missouri, Columbia, Missouri 65211, United States*

Heng Deng − *Department of Mechanical and Aerospace Engineering, University of Missouri, Columbia, Missouri 65211, United States*

Yunxin Zhao − *Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri 65211, United States*

https://dx.doi.org/10.1021/acs.jcim.0c00020
J. Chem. Inf. Model. 2020, 60, 2004−2011

Ming Xin − *Department of Mechanical and Aerospace Engineering, University of Missouri, Columbia, Missouri 65211, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00020

Author Contributions

H.W. and Y.X. contributed equally to this work. H.W. proposed the data augmentation method. He designed and trained CNN as well as implementing NCA and activation mapping analysis. Y.X. synthesized and characterized MOFs. He also downloaded and processed the theoretical data. D.L. initially explored the idea by testing different types of machine learning models. Y.Z. offered valuable suggestions in developing CNNs. M.X. provided valuable discussions and inspired solutions to the problems. J.L. conceived the idea, organized the research scopes, and oversaw all phases of the project. Y.X. wrote the manuscript which was edited by J.L. All authors commented on the manuscript.

Notes

The authors declare no competing financial interest.

The Python scripts for preprocessing, augmentation, and classification are available at https://github.com/linresearchgroup/MOFs

■ ACKNOWLEDGMENTS

This work was supported by grants from the U.S. Department of Energy (Award No. DE-FE0031645) and the National Science Foundation (Award Nos. 1825352 and 1933861).

■ REFERENCES

(1) Correa-Baena, J.-P.; Hippalgaonkar, K.; van Duren, J.; Jaffer, S.; Chandrasekhar, V. R.; Stevanovic, V.; Wadia, C.; Guha, S.; Buonassisi, T. Accelerating Materials Development *via* Automation, Machine Learning, and High-Performance Computing. *Joule* **2018**, *2*, 1410−1420.

(2) 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. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Comput. Mater.* **2019**, *5*, 60.

(3) Vecsei, P. M.; Choo, K.; Chang, J.; Neupert, T. Neural network based classification of crystal symmetries from x-ray diffraction patterns. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2019**, *99*, 245120.

(4) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019**, *365*, No. eaax1566.

(5) Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J. M.; Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone, D.; Cronin, L. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **2019**, *363*, No. eaav2211.

(6) Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **2018**, *559*, 377−381.

(7) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **2017**, *8*, 15679.

(8) de Pablo, J. J.; Jackson, N. E.; Webb, M. A.; Chen, L.-Q.; Moore, J. E.; Morgan, D.; Jacobs, R.; Pollock, T.; Schlom, D. G.; Toberer, E. S.; Analytis, J.; Dabo, I.; DeLongchamp, D. M.; Fiete, G. A.; Grason, G. M.; Hautier, G.; Mo, Y.; Rajan, K.; Reed, E. J.; Rodriguez, E.; Stevanovic, V.; Suntivich, J.; Thornton, K.; Zhao, J.-C. New frontiers for the materials genome initiative. *npj Comput. Mater.* **2019**, *5*, 41.

(9) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.

(10) Oliynyk, A. O.; Mar, A. Discovery of Intermetallic Compounds from Traditional to Machine-Learning Approaches. *Acc. Chem. Res.* **2018**, *51*, 59−68.

(11) Li, F.; Han, J.; Cao, T.; Lam, W.; Fan, B.; Tang, W.; Chen, S.; Fok, K. L.; Li, L. Design of self-assembly dipeptide hydrogels and machine learning via their chemical features. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 11259−11264.

(12) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2016**, *2*, 16028.

(13) Wu, Z.; Kan, S. B. J.; Lewis, R. D.; Wittmann, B. J.; Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 8852−8858.

(14) Ziletti, A.; Kumar, D.; Scheffler, M.; Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat. Commun.* **2018**, *9*, 2775.

(15) Aguiar, J. A.; Gong, M. L.; Unocic, R. R.; Tasdizen, T.; Miller, B. D. Decoding crystallography from high-resolution electron imaging and diffraction datasets with deep learning. *Sci. Adv.* **2019**, *5*, No. eaaw1949.

(16) Iwasaki, Y.; Kusne, A. G.; Takeuchi, I. Comparison of dissimilarity measures for cluster analysis of X-ray diffraction data from combinatorial libraries. *npj Comput. Mater.* **2017**, *3*, 4.

(17) Kusne, A. G.; Keller, D.; Anderson, A.; Zaban, A.; Takeuchi, I. High-throughput determination of structural phase diagram and constituent phases using GRENDEL. *Nanotechnology* **2015**, *26*, 444002.

(18) 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 Comput. Mater.* **2018**, *4*, 43.

(19) Aguiar, J. A.; Gong, M. L.; Tasdizen, T. Crystallographic prediction from diffraction and chemistry data for higher throughput classification using machine learning. *Comput. Mater. Sci.* **2020**, *173*, 109409.

(20) Yoon, C. H.; Schwander, P.; Abergel, C.; Andersson, I.; Andreasson, J.; Aquila, A.; Bajt, S.; Barthelmess, M.; Barty, A.; Bogan, M. J.; et al. Unsupervised classification of single-particle X-ray diffraction snapshots by spectral clustering. *Opt. Express* **2011**, *19*, 16542−16549.

(21) Ziatdinov, M.; Dyck, O.; Maksov, A.; Li, X.; Sang, X.; Xiao, K.; Unocic, R. R.; Vasudevan, R.; Jesse, S.; Kalinin, S. V. Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations. *ACS Nano* **2017**, *11*, 12742−12752.

(22) 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.

(23) 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.

(24) Suzuki, Y.; Hino, H.; Kotsugi, M.; Ono, K. Automated estimation of materials parameter from X-ray absorption and electron energy-loss spectra with similarity measures. *npj Comput. Mater.* **2019**, *5*, 39.

(25) Moghadam, P. Z.; Li, A.; Wiggin, S. B.; Tao, A.; Maloney, A. G. P.; Wood, P. A.; Ward, S. C.; Fairen-Jimenez, D. Development of a Cambridge Structural Database Subset: A Collection of Metal−Organic Frameworks for Past, Present, and Future. *Chem. Mater.* **2017**, *29*, 2618−2625.

(26) Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*, 2014. https://arxiv.org/abs/1409.1556.
(27) Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278−2324.
(28) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436−444.
(29) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929−1958.
(30) Kubo, Y.; Tucker, G.; Wiesler, S. Compacting Neural Network Classifiers *via* Dropout Training. *arXiv*, 2016. https://arxiv.org/abs/1611.06148.
(31) Gal, Y.; Ghahramani, Z. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. *arXiv*, 2015. https://arxiv.org/abs/1512.05287.
(32) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv*, 2014. https://arxiv.org/abs/1412.6980.
(33) Sutskever, I.; Martens, J.; Dahl, G.; Hinton, G. On the importance of initialization and momentum in deep learning. *Proc. Mach. Learn. Res.* **2013**, *28*, 1139−1147.
(34) Kent, R. D.; Read, C.; Kent, R. D. *The Acoustic Analysis of Speech*. Singular Publishing Group: San Diego, 1992; Vol. *58*.
(35) Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L. Deep learning for sensor-based activity recognition: A survey. *Pattern Recog. Lett.* **2019**, *119*, 3−11.
(36) Yosinski, J.; Clune, J.; Nguyen, A.; Fuchs, T.; Lipson, H. Understanding Neural Networks Through Deep Visualization. *arXiv*, 2015. https://arxiv.org/abs/1506.06579.
(37) Goldberger, J.; Hinton, G. E.; Roweis, S. T.; Salakhutdinov, R. R.Neighbourhood components analysis. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: 2005; pp 513−520.
(38) Fu, K.; Dai, W.; Zhang, Y.; Wang, Z.; Yan, M.; Sun, X. MultiCAM: Multiple Class Activation Mapping for Aircraft Recognition in Remote Sensing Images. *Remote Sensing* **2019**, *11*, 544.
(39) Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27−30, 2016*; IEEE: 2016; pp 2921−2929.
(40) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mane, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viegas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv*, 2016. https://arxiv.org/abs/1603.04467.

https://dx.doi.org/10.1021/acs.jcim.0c00020
J. Chem. Inf. Model. 2020, 60, 2004−2011