THE JOURNAL OF PHYSICAL CHEMISTRY C
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

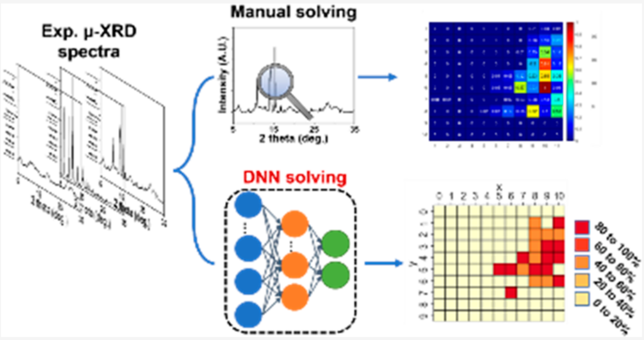
Machine Learning Automated Analysis of Enormous Synchrotron X-ray Diffraction Datasets

Xiaodong Zhao,[#] YiXuan Luo,[#] Juejing Liu,[#] Wenjun Liu, Kevin M. Rosso, Xiaofeng Guo,* Tong Geng,* Ang Li,* and Xin Zhang*

Cite This: J. Phys. Chem. C 2023, 127, 14830−14838 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: X-ray diffraction (XRD) data analysis can be a time-consuming and laborious task. Deep neural network (DNN) based models trained with synthetic XRD patterns have been proven to be a highly efficient, accurate, and automated method for analyzing common XRD data collected from solid samples in ambient environments. However, it remains unclear whether synthetic XRD-based models can be effective in solving micro(μ)-XRD mapping data for in situ experiments involving liquid phases, which always have lower quality and significant artifacts. In this study, we collected μ-XRD mapping data from a $LaCl_3$-calcite hydrothermal fluid system and trained two categories of models to analyze the experimental XRD patterns. The models trained solely with synthetic XRD patterns showed low accuracy (as low as 64%) when solving experimental μ-XRD mapping data. However, the accuracy of the DNN models significantly improved (90% or above) when we trained them with a data set containing both synthetic and a small number of labeled experimental μ-XRD patterns. This study highlights the importance of labeled experimental patterns in training DNN models to solve μ-XRD mapping data from in situ experiments involving liquid phases.

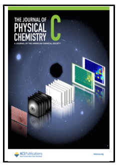


INTRODUCTION

X-ray diffraction (XRD) is an important workhorse technique that is widely used for identifying and characterizing crystalline materials, including their structures, phase compositions, crystallinities, unit cell parameters, and atomic displacement parameters.[1,2] Recently, X-ray microbeam techniques have expanded the information richness of XRD analyses to include additional insight into a reaction system based on examining spatial relationships between reactant and product solid phases. For example, the micro-XRD (μ-XRD) technique, which uses a focused X-ray beam with a typical size of submicron to microns, has brought the power of XRD mapping to laboratory-scale equipment.[3] The μ-XRD technique can be used to study changes in the structural characteristics of reacting mineral solids or precipitates in situ over the course of reaction, in areas as small as tens of micrometers.[4−8] This enables XRD to be useful for distinguishing homogeneous versus heterogeneous mineral nucleation and growth pathways. Similar microbeam XRD techniques are now also widely available at synchrotron X-ray user facilities, where one can take advantage of even higher spatial resolution and brightness for higher fidelity studies and XRD mapping studies. Combining XRD mapping techniques with compatible reaction cells, such as a hydrothermal diamond anvil cell (HDAC) has broadened the range of possible in situ studies. Such techniques have been used in numerous studies to provide insights into the structure and behavior of minerals and materials under extreme conditions such as high temperatures (up to 600 °C) and high pressures (up to more than 1 GPa).[9−13]

Because modern XRD mapping capabilities can quickly generate enormous quantities of data (e.g., thousands of individual XRD patterns comprising one map), one of the emerging challenges is efficient data analysis with the same level of accuracy as manual comparisons of material diffractograms with those available in standard XRD databases. XRD mapping data interpretation requires a high level of expertise, including coordination between data collection and analysis, skill to collect high-quality data, recognition of systematic errors, and preconception of potential phases. Although commercial software, such as Jade,[14] is useful for unknown phase identification based on automated search and matching, this approach is intrinsically imprecise and capable of missing

Received: May 26, 2023
Revised: July 3, 2023
Published: July 24, 2023



or mismatching phases. In total, manual analysis of XRD data is a labor-intensive process that requires a profound understanding of the materials for reliable results. However, in mapping mode, to study spatial relationships, the amount of XRD data increases exponentially, making manual analysis of "big XRD data" impractical.[15−17]

A promising solution to the challenges posed by massive XRD data sets is the use of deep neural network (DNN) based models, which can automatically extract information from XRD patterns.[18−24] These models leverage the advantages of combining 1D convolution layers and dense layers to extract and learn features from labeled XRD patterns in the training data set. Several studies have reported using DNN models to extract various types of information from experimental XRD patterns, such as phase identity and ratio, unit cell parameters, and even the ability to distinguish potential perovskite materials.[24−27] However, since DNN model training usually requires a large amount of diverse data, a common practice is to generate theoretical XRD patterns from crystallographic structures.[25−28] Previous studies have shown that models trained on theoretical XRD patterns can analyze XRD patterns collected from solid samples in ambient environments.[24−27] However, it remains unclear whether these models can also accurately and efficiently analyze data collected from the μ-XRD mapping of in situ experiments involving complex mixtures of solids and liquid.

Combining μ-XRD mapping with solid−liquid mixed samples in high-temperature and high-pressure environments presents the most challenging scenario for obtaining high-quality XRD data interpretation. μ-XRD is typically used to analyze micrograins in samples, but when dealing with mixed solid−liquid samples, common issues such as overexposure, imperfect diffraction, and preferred orientation can lead to artifacts in the XRD pattern.[29] If the samples are entirely solid, these issues can still be manageable, and data of reasonable quality can be retrieved.[4−8] However, the presence of the liquid phase and extreme environments significantly amplifies these adverse effects, leading to distorted XRD data. This poses a major challenge that could potentially prevent DNN models trained solely on synthetic XRD patterns from recognizing experimental μ-XRD mapping data.

Here we show the limitation of DNN models, trained solely by synthetic XRD patterns, to analyze μ-XRD data collected from a hydrothermal fluid environment. To achieve this, we trained multiple models to solve μ-XRD mapping data from a $LaCl_3$-calcite system reacted at 200 °C. The accuracy of this approach was evaluated by comparing model-driven results with findings determined manually. Two training data sets were generated, one including only theoretical XRD patterns and the other including a small number of labeled experimental patterns in addition to the theoretical data. We trained two types of models, three binary classification models to identify the existence of specified phases and two multiclass multilabel models to extract all types and ratios of potential phases in the $LaCl_3$-calcite system. Our results showed that all DNN models trained solely by synthetic XRD data performed poorly for resolving the key information present in the μ-XRD mapping data. Accurate and robust models were achieved only when a small number of experimental XRD patterns were included in the training data set. This finding underscores the importance of labeled experimental data for DNN model training to solve μ-XRD mapping data collected from hydrothermal fluid systems.

METHODOLOGY AND EXPERIMENT

Computation Platform. A conventional server cluster computer (processor: Intel Xeon Gold 6330 CPU, graphical processor unit, Nvidia A100-PCIE-40GB; memory, 40GB) was used to train all the models.

Software Libraries. Software libraries used in this study were PyTorch, NumPy, Pandas, and SciPy. The python packages NumPy and Pandas were used to write code to load and preprocess the raw theoretical patterns (described below).[30,31] PyTorch and TensorFlow2 were used to construct the NN and produce the models.[32−34] The code was written using Python 3.9.12.

Data Augmentation and Preprocessing. Two categories of data were used to build training and evaluation data sets: theoretical XRD patterns and a small number of labeled experimental data. The theoretical XRD patterns were generated by mixing two and three end member patterns (bastnaesite, calcite, and rhenium (Re) metal). This mixing was conducted multiple times for every combination of phases with different ratios.[26−28] Small number of labeled experimental XRD patterns were also produced for generating data sets (see additional information in the Supporting Information for detailed composition of training data sets).

To unify the scale of theoretical data and experimental data, we extracted their spectrum intersection ($2\theta = 5°$ to $38°$) and obtain the value over each 0.01° interval as the features, which means the number of features should be 3501. For the data that have a random scale interval, we used 1D linear interpolation to fill in gaps. The diffraction amplitude is different between the theoretical XRD pattern and experimental data. To ensure that the gradient moves smoothly toward the minima while maintaining the same rate for all the features, we scaled the features' value to [0,1] by data normalization as follows:

$$X = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where X is the data sample.

The content above illustrates how we preprocessed the raw data for NN models. However, as a data-driven domain, machine learning requires numerous data samples, among which experimental data are essential but hard to obtain by artificial labeling. Therefore, to enrich the existing unbalanced and limited data set, we propose a data synthesis algorithm to generate artificial data points. In this data synthesis algorithm, E and T are the normalized experimental and theoretical data sets, respectively. By randomly selecting one positive sample e_i from E and one negative sample t_i from T, the artificially created positive data sample can be represented by

$$X_{ac} = e_i + \varepsilon \cdot t_i$$

where ε is a random number and $\varepsilon \in (0,1)$. Additionally, to unify the scale of the original data and artificially create data, X_{ac} also needs to be normalized using the approach mentioned above. Through the novel data synthesis approach, we can freely create a balanced and abundant data set for training and validating, which greatly benefits the training process.

NN Model Architecture. All DNN models trained in this study used multiple layers of convolution and maxpooling to extract the features from the 1D XRD pattern. Based on the previously studies, the combination of these two layers improves the accuracy of models.[25−28] As shown in Figure

S1, the binary classification DNN models determine the existence of phases from input XRD pattern mainly consists of 3 different kinds of layers including a 1d convolution layer, a maxpooling layer, and a fully connected layer. At the start of the model, normalized data points are fed into a 1d convolution layer (input channel of 1, output channel of 4, kernel size of 5, stride size of 1, and padding size of 1) which effectively extracts the features from the input. Next, a maxpooling layer (kernel size of 16, stride size of 1, and padding size of 1) will process the output from the previous convolution layer in which the large kernel size can enlarge the field of vision for peak detection. Then a combination of the convolution layer and the maxpooling layer with the same configuration will be added to the model. To make a prediction based on the features obtained from previous layers, we applied 6 fully connected layers to the model, among which the output nodes were set to be 1024, 512, 256, 128, 64, and 2. After forward propagation, the NN model will output 2 values denoting the class probabilities of the element, between which the class with a higher value will be considered as the classification result.

The architectures of the two multiclass and multilabel models capable of retrieving all potential phases and phase ratios from the input XRD pattern in the $LaCl_3$-calcite hydrothermal fluid system were obtained from a previous study (see Figure S2).[26] The data sets for training and evaluating the two models are similar to ones described above, except the range of data is from 5.00° to 35.00° in terms of 2θ (2501 points). The only difference of the two models is that one was trained with data set containing a small amount of labeled experimental data, and another by the data set without any labeled experimental data.

Training Notes. During the training phase for the binary classification models, we used CrossEntropyLoss (see eq 1) as the loss function, which computes the cross-entropy loss between input logits and target.

$$\text{Loss} = -y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i)) \tag{1}$$

To minimize the training loss, we used Adam, an algorithm for first-order gradient-based optimization of stochastic objective functions, to optimize the parameters in the models. With continuous iteration, Adam can adjust the parameters in our models to help them better fit observed data and determine the relation between input features and ground truth.

In addition to the choice of loss function and optimizer, we also need to set up other hyperparameters to initialize the training of models. To balance the trade-off between the rate of convergence and overshooting, we set the learning rate to be 1×10^{-5}. Furthermore, to prevent overfitting the model, the weight decay of the optimizer is set to be 1×10^{-8}. Besides, for the training and validation sets, the batch size was set to 60. Both training and evaluation data sets were shuffled at each training epoch. To ensure fairness during training, we set the training epoch to 200 for each element.

For the multiclass and multilabel models, we modified the cross entropy loss function and accuracy metric function based on the previous study.[26] The Adam optimizer was used as the gradient decent function with the initial learning rate as 1×10^{-2} and a gradually decrease to 1×10^{-8}. Similar to the binary classification models above, the training and evaluation data set were shuffled during each epoch. These models were trained by 256 epochs.

Training and Evaluation of Different NN Models. Even though we artificially created data for training and validation, the samples in the test set were all from experimental data collected from the $LaCl_3$-calcite hydrothermal fluid system, which means that the test set is still unbalanced. While evaluating a data set owning many more negative samples than positive samples, accuracy is not enough to fairly demonstrated model performance.

$$\text{Acc} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Hence, in addition to the accuracy, we also applied the area under the receiver operating characteristic (AUROC) and the area under the precision-recall curve (AURPC) to prove the superiority of our model (see Figure S3 and additional information in the Supporting Information for a detailed algorithm).

The evaluation of the two comprehensive models was performed by a direct comparison method. The model driven spatial distribution of different phases was compared to the manual solving results.

Theoretical XRD Pattern Generation. The three main features of XRD patterns we considered in this work are *i*) peak position, *ii*) peak intensity and *iii*) peak profile shape. A Python script was used to manipulate General Structure Analysis System-II (GSAS-II) to generate theoretical XRD patterns without the use of the graphic user interface. GSAS-II is a software package developed by Toby and Von Dreele for X-ray and neutron diffraction analysis.[35] For the same compound, the three features of XRD can be impacted by the measuring condition, the experiment where it is performed, the source of X-ray, and sample nature.

The principle of the XRD method is based on the diffraction of X-rays by crystalline structure, with the diffraction rule described by Bragg's law,

$$n\lambda = 2d_{hkl}\sin(\theta)$$

The d spacing is defined by the unit cell parameters (a, b, c, α, β, γ) and Miller indices (h, k, l).

$$\begin{aligned}\frac{1}{{d_k}^2} = \frac{1}{\nu^2}[&H^2b^2c^2\sin^2\alpha + K^2a^2c^2\sin^2\beta + L^2a^2b^2\sin^2\gamma \\ &+ 2HKabc^2(\cos\alpha\cos\beta - \cos\gamma) \\ &+ 2KLa^2bc(\cos\beta\cos\gamma - \cos\alpha) \\ &+ 2HLab^2c(\cos\gamma\cos\alpha - \cos\beta)]\end{aligned}$$

where

$$\nu = abc(1 + 2\cos\alpha\cos\beta\cos\gamma - \cos^2\alpha - \cos^2\beta - \cos^2\gamma)^{1/2}$$

In order to change the d spacing, the unit cell parameters a, b, and c of different crystalline phases were modified randomly from 1% to 10%, so the peak position of the theoretical XRD patterns were modulated accordingly.

The diffracted intensities $I_{(hkl)}$ are proportional to the square of the structural factor $F_{(hkl)}$,

$$F_{(hkl)} = \sum_{j=1}^{N} f_j \times \exp(-i2\pi(hx_j + ky_j + lz_j))$$

where f_j is the atomic form factor for element j, hkl is the Miller indices, and (x, y, z) are the coordinates of atoms of element j in the unit cell.

The total intensity $I_{(hkl)}$ is formulated as low:

$$I_{(hkl)} = K \times |F_{(hkl)}|^2 \times f_a \exp \frac{-B \sin^2(\theta)}{\lambda^2} \times A \times L(\theta) \times P(\theta) \times m$$

where K is a constant, $f_a \exp \frac{-B \sin^2(\theta)}{\lambda^2}$ stands for the thermal displacement off the equilibrium position due to temperature effect, and A is the absorption factor,

The peak shape function G_k is a key component to simulate an XRD pattern. The XRD pattern profile is generally controlled by instrument factors and sample factors. The instrument effect can be approximated by Gaussian function due to the similarity in the peak shape, and the sample effect can be reproduced by Lorentzian function. By modulating these factors through the pseudo-Voigt and Pearson VII function, an XRD pattern can be easily synthesized. In this work, the pseudo-Voigt regime was used to simulate the profile function of the XRD peaks by tuning the Gaussian (G) and Lorentzian (L) components. The Gaussian shape is elucidated by the Cagliotti function

$$G \approx {H_k}^2 \approx U \times \tan^2 \theta + V \times \tan \theta + W + \frac{P}{\cos^2 \theta}$$

where U, V, W, and P parameters were used to control he peak profile.

Then other important factors, including the effect from crystallite size broadening, strain broadening, are controlled by the Lorentzian terms X and Y. Overall, the pseudo-Voigt function is a linear combination of the Gaussian function and the Lorentzian function by the ratio of

$$pV(x) = \eta G(x) + (1 - \eta)L(x)$$

Above all, the peak center (position of maximum), height (height of the peak at the maximum), and fwhm (full width at half-maximum of the peak) of theoretical XRD patterns were achieved by modifying the unit cell parameters (a, b, c), Gaussian profile parameters (U, V, W), and Lorentzian parameters (X, Y).

Experimental XRD Acquisition and Preprocessing. Synchrotron XRD was performed at the Advanced Photon Source (APS) at the Argonne National Laboratory. A microdiffraction μ-XRD technique was employed to investigate the spatial correlation in the chamber at the beamline 34-ID-E. The X-ray beam size was 300 nm, and synchrotron X-ray energy was at 22 keV. To conduct the mapping, X-ray sampling was employed on a mesh pattern measuring 11 × 10 within the designated region. The small size of the X-ray beam (300 nm) facilitated the scanning process by ensuring that each mesh point received sufficient coverage. Additionally, the high energy of the X-ray beam (22 keV) and high flux of highly coherent synchrotron X-ray contributed to achieving a high resolution for each phase and minimizing the risk of overlooking any phases during the scanning process. Hence, a comprehensive sampling map with a spatial resolution can be generated (see Figure S4). By this approach, the correlation between the primary calcite and 110 μ-XRD patterns was collected at a temperature of 200 °C. Fewer data were collected under higher temperatures due to the shorter data acquisition time used to prevent liquid leakage during heating. Collected 2D diffraction images were calibrated, masked, and integrated by Dioptas software.[35,36] The background of pristine 1D patterns was automatically subtracted through Dioptas software, the polynomial order was up to 50th order with a smoothing width at 0.1 Å, and the iteration was 150 times. The concentration of $LaCl_3$ for collecting μ-XRD from the $LaCl_3$-calcite hydrothermal fluid system was 0.1 M with calcite. The temperature was set to 200 °C. To produce a high pressure, the diamond culet was 1 mm, and the gasket used in the experiment had a diameter of 500 μm. During the experimental setup, the calcite and $LaCl_3$ solution were manually introduced to the system under a microscope. However, due to the limited space, it was impossible to control the exact concentration of calcite within this confined area. Therefore, while the presence of both calcite and the $LaCl_3$ liquid solution was ensured, the specific concentration of calcite could not be manipulated or precisely determined.

RESULTS AND DISCUSSION

The XRD mapping generates a substantial amount of data, as shown in Figure S4. Therefore, a high-throughput automatic analysis method is necessary to process the μ-XRD data and obtain phase type and ratio information. To observe the relationships among different mineral phases, we collected μ-XRD mapping data from the $LaCl_3$-calcite hydrothermal fluid system with a specified spatial configuration, resulting in a matrix with 11 columns and 10 rows for a total of 110 XRD patterns (Figure 1). We then used two methods to analyze the type and ratio of phases in each XRD pattern: manual mapping and an automatic analysis based on DNN models. When we trained the model using a data set containing both theoretical XRD patterns and a small number of labeled XRD patterns, the phase information derived from the DNN model matched well with the manual analysis method.

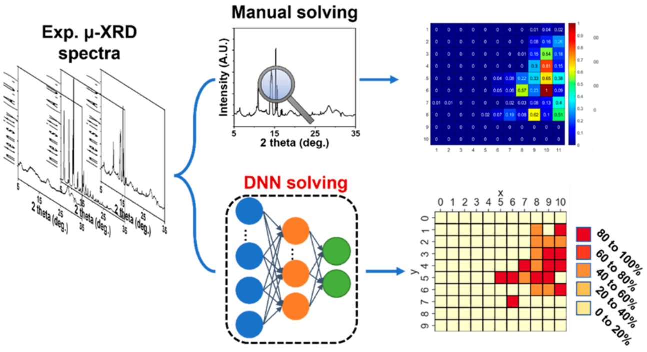


Figure 1. Illustration of solving μ-XRD mapping data by manual resolving and the deep neural network based automatic method.

The manually analyzed μ-XRD mapping result collected from the $LaCl_3$-calcite hydrothermal fluid system at 200 °C shows an overall spatial overlap between calcite and La-bastnaesite (see Figure 2), which indicates a heterogeneous nucleation mechanism. To produce an accurate phase distribution, the pristine XRD data could not be used due to the nonstatistical issue of micro-XRD. During the micro-XRD measurements, the X-ray is only irradiating a small volume, which may only contain a limited number of crystalline powders and result in a nonstatistical problem in the XRD pattern. To tackle this systematic error, first, an attentive masking was done on each 2D XRD figure to eliminate the overexposure (like strong texture). Second, the phase quantity was determined by the sum of at least three characteristic peak

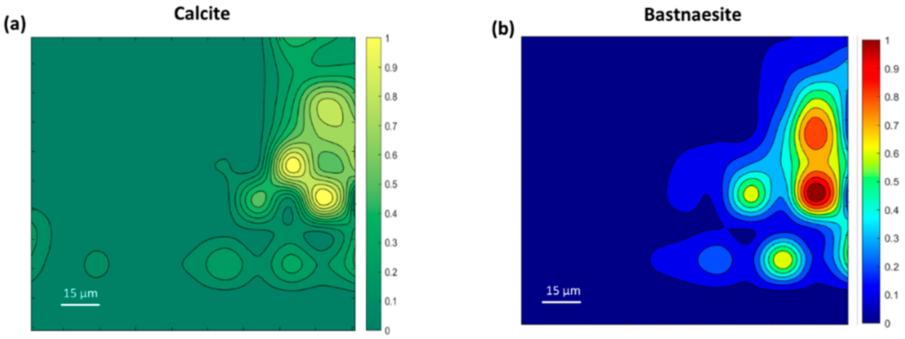


Figure 2. Heatmaps of crystalline phase occupancy in the $LaCl_3$-calcite hydrothermal fluid system at 200 °C. (a) Calcite occupancy. The bright yellow color represents a high enrichment of calcite in the area. (b) La-bastnaesite occupancy. The bright red color represents a high enrichment of bastnaesite in the area.

intensities, instead of one most characteristic peak, so that an unsolicited increase in one peak can be partly minimized during the multiple peak summation. For the calcite, the peak intensities of (102), (104), and (202) from ranges 8.30°−8.38°, 10.56°−10.64°, and 15.40°−15.44° were summed up, and for bastnaesite, the peaks of (002), (300), (2$\bar{1}$3), (6$\bar{3}$0), and (304) at around 6.42°−6.46°, 8.80°−8.84°, 10.91°−10.95°, 15.28°−15.32°, and 15.64°−15.68° were accumulated. All peak intensities were obtained by a Matlab code and manually cross-checked to ensure the accuracy. The processing of the pristine data is expensive in time, which highly motivates an automotive methodology without or with little manual perturbation.

The heatmap shows that the calcite phase is centered in the right bottom corner of the scanned zone, which is consistent with the X-ray absorption imaging shown in Figure S4. This is because the calcite solid was initially introduced to that position and La^{3+} existed as an ion form in the solution. With the dissolution of calcite at 200 °C, the released carbonate ion facilitates the precipitation of La^{3+} in the form of the bastnaesite ($LaCO_3OH$) solid form. Notably, the center of the bastnaesite precipitation zone overlapped with the calcite-enriched area, indicating a spatial dependence of La-bastnaesite mineralization in conjunction with the presence of calcite. Additionally, Re metal from the HDAC gasket was found at the bottom site. It should be noted that this current pilot test only covers data from 200 °C, but a series of in situ experiments with varying temperature steps would produce a substantially larger dataset than the one presented in this study more than current data amount. Therefore, it is crucial to develop an automatic method to analyze future μ-XRD mapping data from hydrothermal fluid systems containing other rare earth elements (REEs) besides lanthanum.

As a proof-of-concept, we trained three DNN models for binary classification to exclusively identify bastnaesite, calcite, and Re metal in the input XRD pattern, outputting true or false for each phase's presence or absence. Two kinds of data sets were used to train these models. The first data set was generated using a previously reported method without labeled experimental data.[26] The second data set was created by adding a small number of labeled experimental diffraction patterns to the first data set. To evaluate the impact of different training data sets on model performance, we used these models to analyze experimental data that was manually solved (see Figure 2) but not used in training or evaluation of the models. We used three metrics to evaluate performance: area under the receiver operating characteristic (AUROC), area under the precision-recall curve (AUPRC), and accuracy (see the method section for more detailed information).

Comparing the performance of DNN models trained with and without experimental data inputs (Table 1) reveals that using solely synthetic diffraction patterns is insufficient to train models capable of identifying phases from our experimental data, despite being reported in multiple studies.[25−28,37,38] As shown in Table 1, most models trained without labeled experimental data exhibit a lower phase identification performance, including lower AUROC, AUPRC, and accuracy values. The only exception is bastnaesite-focused models, with two models having the same AUROC of 0.96. As the AUROCs are close to 1.00, the models confidently judge whether the experimental data contains bastnaesite. For AUPRC and accuracy, the model trained without experimental data slightly underperforms compared to the model trained with experimental data. Overall, the two models behave similarly in terms of deciding the existence of the bastnaesite phase from experimental data, plausibly due to the significant difference of the XRD patterns among the bastnaesite, calcite, and Re metal phases (see Figure S5).

Table 1. Performance of DNN Models in Terms of Phase Identification Trained by Datasets with and without Experimental Data (w/exp. and w/o exp.)[a]

	AUROC		AUPRC		Accuracy (%)	
Phase	w/o exp.	w/exp.	w/o exp.	w/exp.	w/o exp.	w/exp.
Bastnaesite	0.96	0.96	0.88	0.92	89	92
Calcite	0.66	0.95	0.62	0.88	64	90
Re	0.50	0.96	0.26	0.99	74	95

[a] See additional information in the Supporting Information for detailed composition of training datasets.

The performance gap between the models trained with and without labeled experimental data is substantial for the remaining two models. The calcite focused model trained without labeled experimental data exhibits an AUROC of 0.66 and an AUPRC of 0.62, which are significantly lower than the corresponding values of 0.95 and 0.88 obtained by the model trained with labeled experimental data. Additionally, the accuracy of the model trained without labeled data is only 64%, which is close to random guessing (50%). In contrast, the model trained with labeled data achieves an accuracy of 90%. Similar outcomes are observed in the Re metal focused model.

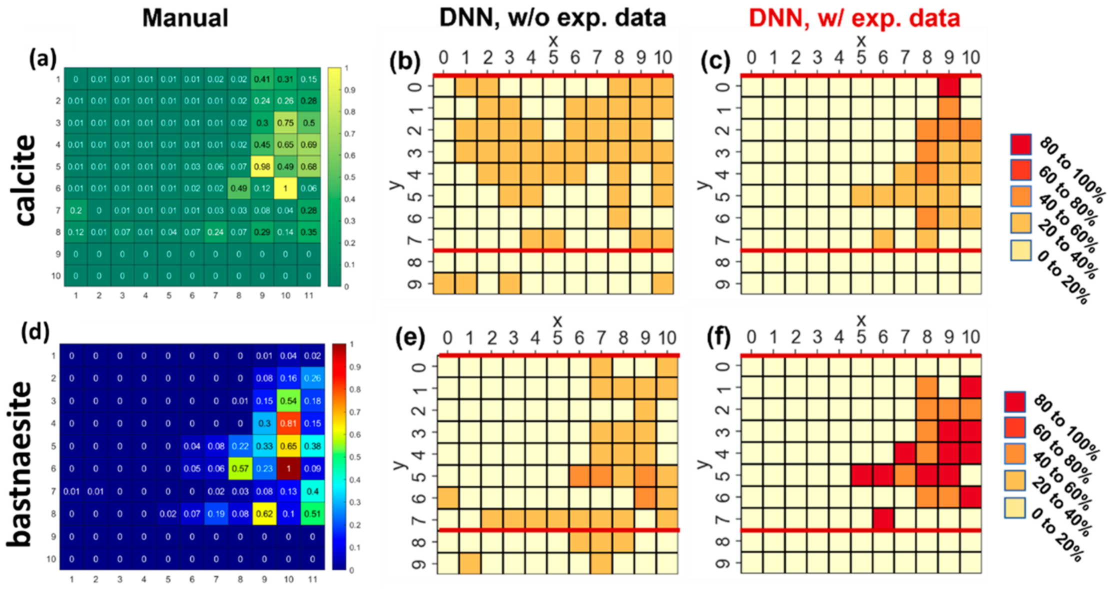


Figure 3. Utilization of DNN based models to retrieve phase types and ratios from μ-XRD mapping data obtained from the $LaCl_3$-calcite hydrothermal fluid system at 200 °C. (a) Distribution of calcite obtained by manual mapping, (b) DNN model trained without labeled experimental data, and (c) DNN model trained with labeled data. (d) Distribution of bastnaesite obtained by manual mapping, (e) DNN model trained without labeled data, and (f) DNN model trained with labeled data. The red lines in the heatmaps retrieved by DNN based models mark the region shown in the manual analyses.

Although the model trained without experimental data has a higher accuracy (74% for Re vs 64% for calcite), the extremely low AUROC and AUPRC values indicate that this model lacks robustness. Consequently, the models trained without experimental data could not identify phases from the experimental data.

The poor performance of the models trained without labeled data can be attributed to the significant differences between the synthetic XRD patterns used in training and the μ-XRD data obtained from experiments. In previous studies, experimental data used for evaluating DNN models were typically collected from solid samples in ambient environments.[26−28] In such cases, high-quality X-ray diffraction peaks from different planes were easily obtainable, which could be fitted to the synthetic data generated from the crystallographic structure. However, μ-XRD data obtained from hydrothermal fluid systems only exhibits a few primary diffraction peaks with low-index (*hkl*) lattice planes, even under the best circumstances, due to several adverse factors, including the small diffraction volume with microbeam, low exposure time, poorly crystalline samples, preferred orientation or large crystal compared with beam size, and overexposure in 2D images (extra bright spot, see Figure S6). The differences between the theoretical and experimental XRD patterns, including intensity distortions of primary peaks and the absence of many weak peaks with higher indices, exceed what the model trained solely with theoretical data can handle, even if they correspond to the same crystal structure. Therefore, incorporating labeled experimental data into the training data set is essential to enhance the DNN-based XRD analysis model's robustness to data obtained from hydrothermal fluid systems.

Two multiclass and multilabel deep neural network (DNN) models were employed (Figure S2 and the methodology section for model architecture) to ascertain the proportions of the phases in experimental X-ray diffraction (XRD) patterns. Our approach builds upon a previously reported method, which has been improved in this study.[26,27] In the original method, the training data set is generated by linearly combining randomly selected theoretical XRD patterns, ranging from one to four, from a comprehensive list of all possible phases present in the sample. The selection criterion for potential phases is that they must solely comprise elements found in the sample (specifically, La, C, O, and H in our case). Prior to mixing, the chosen theoretical XRD patterns were subjected to a Voigt filter (a convolution of Gaussian and Lorentzian filters) to replicate instrument-induced effects.[39] This methodology enables the production of numerous XRD patterns, amounting to hundreds of thousands, for training DNN models. Nonetheless, despite the substantial size of the XRD training data set, training a model to accurately determine the exact phase composition, i.e., a regression model, proves challenging due to the diverse effects stemming from instrumentation and artifacts. We transformed this regression problem to a classification problem. The model is trained to predict the range of phase ratios (labels) represented in the XRD patterns, such as 0% to 20%, 20% to 40%, and so on until 80% to 100%, using input 1D XRD patterns as features.

The accuracy of the aforementioned method was improved by adding a small number of phase ratio labeled experimental XRD patterns into the training and validation data set. The model trained with a data set containing a small number of labeled XRD patterns showed significantly better accuracy than that trained without labeled experimental data. As shown in Figure 3a,b, there is a significant difference in the spatial distribution of calcite between manual solving and DNN models trained without labeled data. The manual solving results indicate that calcite is located on the right side of the area, while the DNN models trained without experimental data suggest that calcite presents everywhere. However, after incorporating a small amount of labeled experimental data, the model successfully retrieved the rough distribution of calcite in the area (see Figure 3c).

A similar trend is observed in the retrieval of bastnaesite from experimental data (see Figure 3d−f). Although the model trained without labeled data roughly retrieved the distribution of bastnaesite in the area, it falsely predicted bastnaesite at the bottom of the region of interest (right above the red line). Adding labeled data significantly improved the model's performance, resulting in the retrieval of bastnaesite phase distribution similar to the manual solution. Moreover, the model trained with labeled data is less likely to falsely predict phases not seen in the manual solution results in the $LaCl_3$-calcite hydrothermal fluid system, such as $La_2O_2CO_3$ with *Ama2* or *C12c1* space group and $LaOHCO_3$ (see Figures S7 and S8). Overall, these two models further underscore the importance of using labeled experimental data in the training data set when using DNN to automatically solve μ-XRD data, particularly from hydrothermal fluid systems.

■ CONCLUSION

In this study, we have demonstrated the importance of including labeled experimental data in the training data set when training DNN models to obtain phase information from μ-XRD mapping, in this case illustrated using a $LaCl_3$-calcite hydrothermal reaction system. Our simple binary phase experiment showed that models trained with a small amount of experimental data outperformed models trained without experimental data when judging the existence of phases from experimental μ-XRD data. All three statistical parameters (AUROC, AUPRC, and accuracy) of models trained with labeled experimental data were higher/better than those trained without labeled data. This trend was maintained when using the same data sets to train two more comprehensive models to retrieve not only the type but also the ratio of all possible phases. The model trained with labeled data correctly retrieved the spatial distribution of calcite and bastnaesite from μ-XRD maps. This study emphasizes that training DNN models with synthetic XRD patterns is not a universal solution to analysis of XRD mapping data. It also highlights the necessity to build a robust experimental mineral diffractogram data set in the environment of interest, which would be of great importance for future DNN studies attempting to identify, for example, REE minerals and their formation pathways in hydrothermal fluid environments.

■ ASSOCIATED CONTENT

Data Availability Statement

The code that can be used to replicate the results presented in this work is available at https://github.com/YixuanLuoBanksy/ML-Chemistry-2023 and https://github.com/pnnl.

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcc.3c03572.

> Architecture of convolution neural network used in binary classification and architecture of convolution neural network used in retrieving phase type and ratio; an illustration of a typical confusion matrix; the schematic representation of the synchrotron XRD data collection and processing progress; representations of the theoretical XRD pattern and comparison between the experimental and theoretical XRD patterns; results of DNN model predicted material phases with training sets with and without experimental data (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Xiaofeng Guo − *Department of Chemistry, Washington State University, Pullman 99164, United States; Materials Science and Engineering Program, Washington State University, Pullman 99164, United States;* orcid.org/0000-0003-3129-493X; Email: x.guo@wsu.edu

Tong Geng − *Department of Electrical and Computer Engineering, University of Rochester, New York 14627, United States*; Email: tgeng@UR.Rochester.edu

Ang Li − *Pacific Northwest National Laboratory, Richland, Washington 99354, United States*; Email: ang.i@pnnl.gov

Xin Zhang − *Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0000-0003-2000-858X; Email: xin.zhang@pnnl.gov

Authors

Xiaodong Zhao − *Pacific Northwest National Laboratory, Richland, Washington 99354, United States; Department of Chemistry, Washington State University, Pullman 99164, United States*

YiXuan Luo − *Department of Electrical and Computer Engineering, University of Rochester, New York 14627, United States*

Juejing Liu − *Pacific Northwest National Laboratory, Richland, Washington 99354, United States; Materials Science and Engineering Program, Washington State University, Pullman 99164, United States*

Wenjun Liu − *Advanced Photon Source, Argonne National Laboratory, Lemont 60439, United States*

Kevin M. Rosso − *Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0000-0002-8474-7720

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcc.3c03572

Author Contributions

[#]X.D., Y.L., and J.L. contributed equally

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The research described in this paper was conducted under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory (PNNL), a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy. K.M.R. acknowledges support from the U.S. Department of Energy (DOE) Office of Science, Office of Basic Energy Sciences, Chemical Sciences, Geosciences, and Biosciences Division through its Geosciences program at PNNL (FWP #56674). A portion of the work was performed with the user proposal 51382 using the Environmental and Molecular Sciences Laboratory (EMSL), a national scientific user facility at PNNL sponsored by the DOE's Office of Biological and Environmental Research. PNNL is a multiprogram national laboratory operated by Battelle Memorial Institute under contract no. DE-AC05-76RL01830 for the DOE. This research used resources of the Advanced Photon Source, a DOE Office of Science user facility operated for the

DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. X.G. also acknowledge the support of this work by the National Science Foundation (NSF), Division of Earth Sciences, under award No. 2149848.

REFERENCES

(1) Ergun, S. Analysis of coherence, strain, thermal vibration and preferred orientation in carbons by X-Ray diffraction. *Carbon* **1976**, *14* (3), 139−150.

(2) Kaduk, J. A.; Reid, J. Typical values of Rietveld instrument profile coefficients. *Powder Diffraction* **2011**, *26* (1), 88−93.

(3) Flemming, R. L. Micro X-ray diffraction (μXRD): a versatile technique for characterization of Earth and planetary materials. *Canadian Journal of Earth Sciences* **2007**, *44* (9), 1333−1346.

(4) O'Brien, M. G.; Jacques, S. D. M.; Di Michiel, M.; Barnes, P.; Weckhuysen, B. M.; Beale, A. M. Active phase evolution in single Ni/Al2O3methanation catalyst bodies studied in real time using combined μ-XRD-CT and μ-absorption-CT. *Chemical Science* **2012**, *3* (2), 509−523.

(5) Shui, J. L.; Okasinski, J. S.; Kenesei, P.; Dobbs, H. A.; Zhao, D.; Almer, J. D.; Liu, D. J. Reversibility of anodic lithium in rechargeable lithium-oxygen batteries. *Nat. Commun.* **2013**, *4*, 2255.

(6) Dejoie, C.; Sciau, P.; Li, W.; Noe, L.; Mehta, A.; Chen, K.; Luo, H.; Kunz, M.; Tamura, N.; Liu, Z. Learning from the past: rare epsilon-Fe2O3 in the ancient black-glazed Jian (Tenmoku) wares. *Sci. Rep.* **2014**, *4*, 4941.

(7) Corkhill, C. L.; Crean, D. E.; Bailey, D. J.; Makepeace, C.; Stennett, M. C.; Tappero, R.; Grolimund, D.; Hyatt, N. C. Multi-scale investigation of uranium attenuation by arsenic at an abandoned uranium mine, South Terras. *npj Materials Degradation* **2017**, *1*, 19.

(8) Zhou, G.; Zhu, W.; Shen, H.; Li, Y.; Zhang, A.; Tamura, N.; Chen, K. Real-time microstructure imaging by Laue microdiffraction: A sample application in laser 3D printed Ni-based superalloys. *Sci. Rep.* **2016**, *6*, 28144.

(9) Li, J.; Li, S. Application of Hydrothermal Diamond Anvil Cell to Homogenization Experiments of Silicate Melt Inclusions. *Acta Geologica Sinica - English Edition* **2014**, *88* (3), 854−864.

(10) Louvel, M.; Drewitt, J. W. E.; Ross, A.; Thwaites, R.; Heinen, B. J.; Keeble, D. S.; Beavers, C. M.; Walter, M. J.; Anzellini, S. The HXD95: a modified Bassett-type hydrothermal diamond-anvil cell for in situ XRD experiments up to 5 GPa and 1300 K. *Journal of Synchrotron Radiation* **2020**, *27* (2), 529−537.

(11) Li, J.; Bassett, W. A.; Chou, I.-M.; Ding, X.; Li, S.; Wang, X. An improved hydrothermal diamond anvil cell. *Rev. Sci. Instrum.* **2016**, *87* (5), 053108.

(12) Maneta, V.; Anderson, A. J. Monitoring the crystallization of water-saturated granitic melts in real time using the hydrothermal diamond anvil cell. *Contributions to Mineralogy and Petrology* **2018**, *173* (10), 83.

(13) Kalintsev, A.; Migdisov, A.; Alcorn, C.; Baker, J.; Brugger, J.; Mayanovic, R. A.; Akram, N.; Guo, X.; Xu, H.; Boukhalfa, H.; et al. Uranium carbonate complexes demonstrate drastic decrease in stability at elevated temperatures. *Communications Chemistry* **2021**, *4* (1), 120.

(14) *JADE 9.5*; Materials Data: Livermore, CA, 2019.

(15) Wang, C.; Steiner, U.; Sepe, A. Synchrotron big data science. *Small* **2018**, *14* (46), 1802291.

(16) Wang, C.; Yu, F.; Liu, Y.; Li, X.; Chen, J.; Thiyagalingam, J.; Sepe, A. Deploying the big data science center at the shanghai synchrotron radiation facility: the first superfacility platform in China. *Machine Learning: Science and Technology* **2021**, *2* (3), 035003.

(17) Dixit, M. B.; Verma, A.; Zaman, W.; Zhong, X.; Kenesei, P.; Park, J. S.; Almer, J.; Mukherjee, P. P.; Hatzell, K. B. Synchrotron imaging of pore formation in Li metal solid-state batteries aided by machine learning. *ACS Applied Energy Materials* **2020**, *3* (10), 9534−9542.

(18) 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), 1−11.

(19) Venderley, J.; Mallayya, K.; Matty, M.; Krogstad, M.; Ruff, J.; Pleiss, G.; Kishore, V.; Mandrus, D.; Phelan, D.; Poudel, L. Harnessing interpretable and unsupervised machine learning to address big data from modern X-ray diffraction. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119* (24), No. e2109665119.

(20) Maruyama, S.; Ouchi, K.; Koganezawa, T.; Matsumoto, Y. High-Throughput and Autonomous Grazing Incidence X-ray Diffraction Mapping of Organic Combinatorial Thin-Film Library Driven by Machine Learning. *ACS Comb. Sci.* **2020**, *22* (7), 348−355.

(21) Li, J.; Huang, X.; Pianetta, P.; Liu, Y. Machine-and-data intelligence for synchrotron science. *Nature Reviews Physics* **2021**, *3* (12), 766−768.

(22) Li, Z.; Qin, L.; Guo, B.; Yuan, J.; Zhang, Z.; Li, W.; Mi, J. Characterization of the convoluted 3d intermetallic phases in a recycled al alloy by synchrotron x-ray tomography and machine learning. *Acta Metallurgica Sinica (English Letters)* **2022**, *35*, 115−123.

(23) Wang, B.; Guan, Z.; Yao, S.; Qin, H.; Nguyen, M. H.; Yager, K.; Yu, D. Deep learning for analysing synchrotron data streams. *New York Scientific Data Summit (NYSDS)* **2016**, DOI: 10.1109/NYSDS.2016.7747813.

(24) Massuyeau, F.; Broux, T.; Coulet, F.; Demessence, A.; Mesbah, A.; Gautier, R. Perovskite or Not Perovskite? A Deep-Learning Approach to Automatically Identify New Hybrid Perovskites from X-ray Diffraction Patterns. *Adv. Mater.* **2022**, *34* (41), 2203879.

(25) Schuetzke, J.; Benedix, A.; Mikut, R.; Reischl, M. Enhancing deep-learning training for phase identification in powder X-ray diffractograms. *IUCrJ.* **2021**, *8* (3), 408−420.

(26) 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* (1), 86.

(27) Lee, J.-W.; Park, W. B.; Kim, M.; Pal Singh, S.; Pyo, M.; Sohn, K.-S. A data-driven XRD analysis protocol for phase identification and phase-fraction prediction of multiphase inorganic compounds. *Inorganic Chemistry Frontiers* **2021**, *8* (10), 2492−2504.

(28) 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* (7), 2200042.

(29) Yanxon, H.; Weng, J.; Parraga, H.; Xu, W.; Ruett, U.; Schwarz, N. Artifact identification in X-ray diffraction data using machine learning methods. *Journal of Synchrotron Radiation* **2023**, *30* (1), 137−146.

(30) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; et al. Array programming with NumPy. *Nature* **2020**, *585* (7825), 357−362.

(31) *pandas-dev/pandas: Pandas 1.2.2*; Zenodo, 2021 (accessed 2023-02-05).

(32) *TensorFlow*; Zenodo, 2022. DOI: DOI: 10.5281/zenodo.5949125 (accessed 2023-02-05).

(33) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. TensorFlow: A System for Large-Scale Machine Learning. *arXiv:1605.08695* **2016**, DOI: 10.5555/3026877.3026899.

(34) Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv:1912.01703* **2019**, DOI: 10.48550/arXiv.1912.01703.

(35) Toby, B. H.; Von Dreele, R. B. GSAS-II: the genesis of a modern open-source all purpose crystallography software package. *J. Appl. Crystallogr.* **2013**, *46* (2), 544−549.

(36) Prescher, C.; Prakapenka, V. B. DIOPTAS: a program for reduction of two-dimensional X-ray diffraction data and data exploration. *High Pressure Research* **2015**, *35* (3), 223−230.

(37) Szymanski, N. J.; Bartel, C. J.; Zeng, Y.; Tu, Q.; Ceder, G. Probabilistic Deep Learning Approach to Automate the Interpretation of Multi-phase Diffraction Spectra. *Chem. Mater.* **2021**, *33* (11), 4204−4215.

(38) Chen, D.; Bai, Y.; Ament, S.; Zhao, W.; Guevarra, D.; Zhou, L.; Selman, B.; van Dover, R. B.; Gregoire, J. M.; Gomes, C. P. Automating crystal-structure phase mapping by combining deep learning with constraint reasoning. *Nature Machine Intelligence* **2021**, *3* (9), 812−822.

(39) *XRD*; 2022. https://github.com/PicricAcid/XRD (accessed 2023-02-05).