# MACHINE LEARNING FOR XRD SPECTRA INTERPRETATION IN HIGH-THROUGHPUT MATERIAL SCIENCE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Experimentally synthesizing predicted materials in a reproducible manner remains a key bottleneck in materials science progress. Autonomous synthesis and closed loop integration of prediction and characterization can address these issues, however, this requires autonomous characterization methods for all analysis including crystallographic phase identification which currently remains a rate-limiting step. Here we benchmark several machine learning techniques for X-ray Diffraction spectra interpretation (spectral clustering, convolutional neural networks, and invertible neural networks) and compare the relative strengths and weaknesses of each approach. Future work will involve deploying these techniques across the entire high-throughput experimental materials database.

## 1 INTRODUCTION

Discovery of new materials can transform a broad range of applications from new semiconductors for solar cells and power electronics, to solid state ion conductors for fast-charging Li-ion batteries and sustainable hydrogen fuel production. However, while rapid progress has been made on the computational prediction of such materials, experimentally synthesizing such materials and reproducing lab-scale synthesis in industrial settings remain critical bottlenecks in materials science progress (De Yoreo et al.).

High-throughput methods and autonomous synthesis capabilities can play a key role in addressing these difficulties. Combinatorial synthesis experiments can create a multitude of samples within a single experiment using gradients in chemical composition, substrate temperature, film thickness, and other synthesis parameters across a substrate (Talley et al., 2021). Meanwhile, autonomous experimentation has shown promising results in organic and polymer synthesis, increasing reproducibility and the pace of material optimization (Granda et al., 2018; Langner et al., 2020). Together, these capabilities allow rapid experimental testing of huge volumes of materials, but closing the experimental feedback loop and informing future synthesis prediction will require automated data analysis methods on the same scale.

While some characterization methods directly measure quantities of interest and require little post-processing analysis (e.g. X-Ray Fluorescence (XRF) measurements), others like X-Ray Diffraction (XRD) are less directly interpretable and necessitate further analysis to perform phase identification, often leading to a bottleneck in autonomous synthesis and characterization loops. In this context, a plethora of methods for automating the analysis of XRD data spectra have been recently proposed, such as deep-learning-based (Lee et al., 2021; Szymanski et al., 2021; Wang et al., 2020) and analytical (Baptista de Castro et al., 2022; Kikkawa et al., 2020) approaches.

Here we discuss several machine learning based approaches for XRD phase identification and present initial results on comparing and benchmarking the diverse set of approaches on common synthetic and experimental datasets. In Section 2 we discuss synthetic XRD data generation, the High Throughput Experimental Materials Database (HTEM DB), and the spectral clustering, convolutional neural network (CNN), and invertible neural network (INN) methodologies. In Section 3 we present our initial benchmarking results and examine key difficulties of implementing machine learning methods in this space. Finally, we conclude in Section 4 with a brief discussion of future work.

# 2 MATERIALS METHODS

## 2.1 SYNTHETIC DATA GENERATION

Synthetic XRD training and test data was generated by defining a chemical system of interest, searching for theoretical reference phases from the Materials Project database (MP), and augmenting the reference phases to create synthetic spectra. This data augmentation procedure leverages the work of Szymanski et al. (2021), and includes peak shifting, broadening and intensity variation.

For this study we concentrated on the Zn-Ni-Co-O system and to simplify the problem, only oxide phases were considered, not elemental and intermetallic phases, and only entries that have a corresponding match in the ICSD database were included, constraining our space to only compounds that have been experimentally observed. This procedure resulted in 22 reference phases. From this set of reference phases, 250 augmented spectra were generated per reference phase. We then separate out $10\%$ of the data to be used as a test data set while the remainder is used for training the ML models considered in this work.

## 2.2 EXPERIMENTAL DATA

Corresponding experimental data was gathered from the HTEM DB, which contains synthesis and characterization data on inorganic materials synthesized in thin film form using combinatorial physical vapor deposition (PVD) methods. The database contains over 300,000 samples across 7,327 sample libraries with more than 33 elements quantified in composition measurements (Talley et al., 2021). For this study we use XRD and XRF data from the same Zn-Ni-Co-O system; spanning 15 combinatorial libraries with 44 samples each, for a total of 660 X-ray diffraction spectra. Previous manual analysis of the data found the material phases observed for this system are wurtzite ZnO, rocksalt NiO, and $Co_3O_4$ spinel (Zakutayev et al., 2011).

## 2.3 XRD INTERPRETATION METHODS

### 2.3.1 SPECTRAL CLUSTERING

We first discuss a purely data-driven method for identifying groups of similar samples from XRD and XRF data, using the Spectral Clustering algorithm. Here an affinity metric is calculated between each pair of samples in a compositional space as

$$A_{ij} = \exp\left(-\epsilon(1 - CC_{ij}^{XRD} \cdot (1 - dd_{ij}^{COMP}))^2\right) \tag{1}$$

where $CC_{ij}^{XRD}$ is the normalized cross-correlation of the XRD patterns and $dd_{ij}^{COMP}$ is the normalized compositional distance (Hattrick-Simpers et al., 2019). The spectral clustering algorithm then calculates the best set of groups given the affinity matrix and a given number of clusters and the optimal number of clusters is chosen using a silhouette analysis.

## 2.4 CNN

A recent approach to perform phase identification from experimental XRD spectra based on a convolutional neural network (CNN) was proposed by Szymanski et al. (2021). It leveraged reference data obtained from the ICSD database, a data augmentation procedure to disturb reference XRD patterns based on experimental artifacts that occur in diffractometry experiments, and a branching algorithm to pick up phases with highest probability. Results from simulated and experimental data showed high accuracy for phase identification in the Li-Mn-Ti-O-F chemical system, making it a promising approach to streamline an autonomous synthesis/characterization loop. Here we apply this methodology to the Zn-Ni-Co-O System. Training proceeded during 100 epochs, and one-hot encoding was employed for labeling XRD data and their corresponding reference phases. Details of this CNN architecture can be found in Szymanski et al. (2021), and were kept the same for this work.

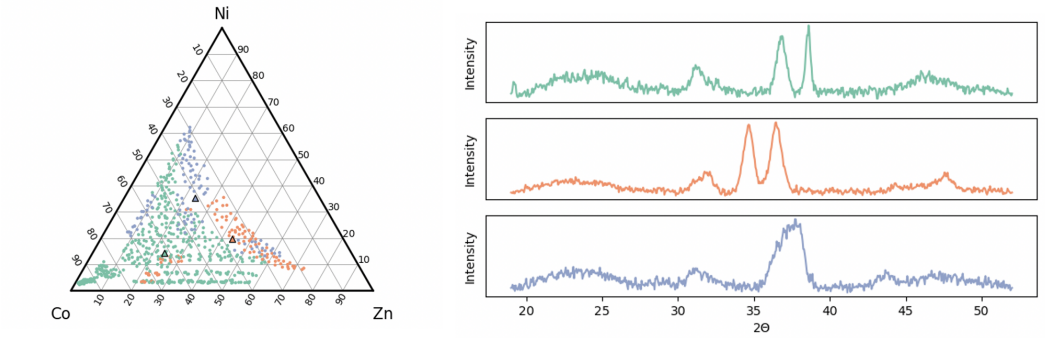


Figure 1: Spectral clustering applied to ZnO-NiO-CoO HTEM data. *Left:* Compositional space diagram showing clusters in different colors. *Right:* Representative XRD spectra chosen from each clusters center.

### 2.5 INN

Invertible neural networks (INNs) are a type of neural architecture that can be run and trained either forward or in reverse and have recently become notable for their particular use in solving inverse problems where an easily computatable forward mapping $y = f(x)$ is ill-posed and potentially degenerate in reverse (Ardizzone et al., 2018). Here we show preliminary results of applying this framework to the XRD interpretation problem considering the observed quantity $y$ the spectral measurements and $x$ as the original material phase. To construct a smaller dimensional form of the XRD spectra without involving a separate convolutional structure we pre-process the XRD spectra using a continuous wavelet transform algorithm to identify the wavelength of the top 5 spectral peaks and their magnitudes to form the $y$ vector of $N_y = 10$, add a $N_z = 6$ dimensional latent space, and zero pad for a total of $N_{zy} = 24$. The $x$ space is formed by one-hot-encoding the material phase labels and zero-padding for a total of $N_x = 24$. The network architecture is formed by stacking 3 invertible blocks consisting of 2 fully connected layers of size 128.

## 3 Results

### 3.1 Spectral Clustering

Results for the application of the spectral clustering algorithm to the HTEM experimental sample libraries in the ZnO-NiO-CoO compositional space are shown in Figure 1. Despite the silhouette analysis, we do find that the shape and distribution of clusters is quite sensitive to the number of clusters.

As a purely data-driven method, the clustering analysis does not directly predict the reference phase associated with each sample; however, this weakness has the corresponding advantages of not needing pre-labeled data for training or suffering from domain-adaptation issues. Without directly predicting phase labels, the results of the clustering algorithm can still be used to build confidence/statistical uncertainty in other methods; one would expect that samples from the same cluster should be identified as the same phase, and the distribution of a directly predicted phase over a given cluster may be informative. Furthermore, individual spectra associated with the cluster center can be used as a set of representative spectra across the material phase space for quick algorithm testing.

### 3.2 Deep-Learning Methods with Synthetic Data

Table 1 shows the results of applying the CNN and INN approaches on the synthetic dataset discussed in Section 2, after 10 independent runs. The CNN method achieved accuracy, compared to those in Szymanski et al. (2021), and the model that yielded the best accuracy for the test set was used to perform phase identification in experimental XRD data from the HTEM database. While the INN method does not perform as well as the CNN technique here, there is still a definitive sign

Table 1: Accuracy comparison, expressed in %, between the CNN and INN approaches on synthetic test and train sets, after 10 independent training runs.

| | Train set | | | Test set | | |
|---|---|---|---|---|---|---|
| Method | Average | Best | Std. dev. | Average | Best | Std. dev. |
| CNN | 94.7 | 95.2 | 0.4 | 92.7 | 94.5 | 1.2 |
| INN | 49.5 | 51.3 | 1.4 | 49.3 | 51.4 | 1.7 |

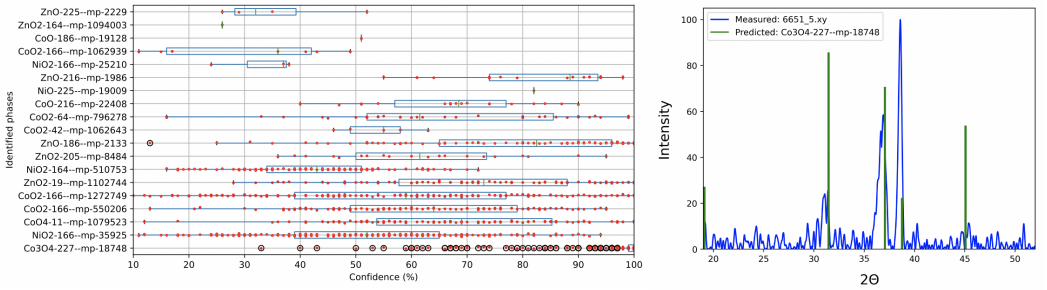


Figure 3: *Left:* Confidence levels for all phases predicted by the CNN approach in the Zn-Ni-Co-O system from the HTEM DB *Right:* CNN prediction for sample 5 - library 6651 - from the HTEM DB

of learning in the confusion matrix (see Figure 2) with most errors arising from poor ZnO identification. Furthermore, there is still substantial room for optimization of the architecture and future work will include applying convolutional kernels to the INN framework rather than pre-processing the data with peak-fitting.

### 3.3 Applying Deep-Learning Methods to HTEM

Figure 3 show the confidence levels for the predicted phases in experimental HTEM data. The $Co_3O_4$ spinel phase was correctly identified for most XRD spectra but the CNN model had difficulties in indentifying the ZnO wurtzite and NiO rocksalt (ZnO-186 and NiO-225, respectively) phases, most likely because these phases are impurity phases in this dataset. Figure 3 also shows an experimental sample from the HTEM DB with corresponding CNN phase prediction. In the INN case, the algorithm over-identified $ZnO_2$ in the HTEM sample data, likely suffering from domain adaptation issues.

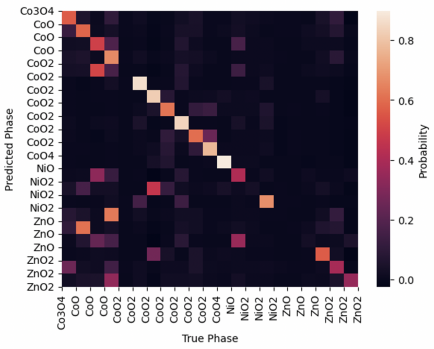


Figure 2: Confusion matrix showing probability of the predicted phase vs true phase for the INN approach.

## 4 Conclusions and Future Work

Here we have demonstrated several machine learning based approaches to interpreting XRD spectral data. Future work in this space will include further optimization of the ML approaches, particularly for the INN technique, and application of these methods to the HTEM database beyond the ZnO-CoO-NiO space. In a broader context these techniques will form a foundation for a key high-throughput analysis technique for enabling closed loop autonomous synthesis and materials discovery feedback.

## REFERENCES

Lynton Ardizzone, Jakob Kruse, Sebastian Wirkert, Daniel Rahner, Eric W Pellegrini, Ralf S Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Köthe. Analyzing inverse problems with invertible neural networks. *arXiv preprint arXiv:1808.04730*, 2018.

Pedro Baptista de Castro, Kensei Terashima, Miren Garbine Esparza Echevarria, Hiroyuki Takeya, and Yoshihiko Takano. Xerus: An open-source tool for quick xrd phase identification and refinement automation. *Advanced Theory and Simulations*, 5(5):2100588, 2022. doi: https://doi.org/10.1002/adts.202100588.

Jim De Yoreo, David Mandrus, Lynda Soderholm, Tori Forbes, Mercouri Kanatzidis, Jonah Erlebacher, Julia Laskin, Uli Wiesner, Ting Xu, Simon Billinge, Sarah Tolbert, Michael Zaworotko, Giulia Galli, Julia Chan, John Mitchell, Linda Horton, Arvind Kini, Bonnie Gersten, George Maracas, Raul Miranda, Mick Pechan, and Katie Runkles. Basic research needs workshop on synthesis science for energy relevant technology. doi: 10.2172/1616513.

Jarosław M Granda, Liva Donina, Vincenza Dragone, De-Liang Long, and Leroy Cronin. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature*, 559 (7714):377–381, 2018.

Jason R Hattrick-Simpers, Zachary T Trautt, Kamal Choudhary, Aaron G Kusne, Feng Yi, Martin L Green, Sara Barron, Andriy Zakutayev, Nam Nguyen, Caleb Phillips, et al. An inter-laboratory comparative high throughput experimental materials study of zn-sn-ti-o thin films. 2019.

Nobuaki Kikkawa, Akitoshi Suzumura, Kazutaka Nishikawa, Shin Tajima, and Seiji Kajita. Extraction of component bases from mixed spectra using non-negative matrix factorization with dissimilarity regularization. *Chemometrics and Intelligent Laboratory Systems*, 206:104096, 2020. ISSN 0169-7439. doi: https://doi.org/10.1016/j.chemolab.2020.104096.

Stefan Langner, Florian Häse, José Darío Perea, Tobias Stubhan, Jens Hauch, Loïc M Roch, Thomas Heumueller, Alán Aspuru-Guzik, and Christoph J Brabec. Beyond ternary opv: high-throughput experimentation and self-driving laboratories optimize multicomponent systems. *Advanced Materials*, 32(14):1907801, 2020.

Jin-Woong Lee, Woon Bae Park, Minseuk Kim, Satendra Pal Singh, Myoungho Pyo, and Kee-Sun Sohn. A data-driven xrd analysis protocol for phase identification and phase-fraction prediction of multiphase inorganic compounds. *Inorg. Chem. Front.*, 8:2492–2504, 2021. doi: 10.1039/D0QI01513J.

Nathan J. Szymanski, Christopher J. Bartel, Yan Zeng, Qingsong Tu, and Gerbrand Ceder. Probabilistic deep learning approach to automate the interpretation of multi-phase diffraction spectra. *Chemistry of Materials*, 33(11):4204–4215, 2021. doi: 10.1021/acs.chemmater.1c01071.

Kevin R. Talley, Robert White, Nick Wunder, Matthew Eash, Marcus Schwarting, Dave Evenson, John D. Perkins, William Tumas, Kristin Munch, Caleb Phillips, and Andriy Zakutayev. Research data infrastructure for high-throughput experimental materials science. *Patterns*, 2(12):100373, 2021. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2021.100373.

Hong Wang, Yunchao Xie, Dawei Li, Heng Deng, Yunxin Zhao, Ming Xin, and Jian Lin. Rapid identification of x-ray diffraction patterns based on very limited data by interpretable convolutional neural networks. *Journal of Chemical Information and Modeling*, 60(4):2004–2011, 2020. doi: 10.1021/acs.jcim.0c00020.

A. Zakutayev, J.D. Perkins, P.A. Parilla, N.E. Widjonarko, A.K. Sigdel, J.J. Berry, and D.S. Ginley. Zn–ni–co–o wide-band-gap p-type conductive oxides with high work functions. *MRS Communications*, 1(1):23–26, 2011. doi: 10.1557/mrc.2011.9.