Vision Transformers for X-ray Diffraction Patterns Analysis

Titouan Simonnet, Mame Diarra Fall, Sylvain Grangeon, Bruno Galerne

ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10887635

Vision Transformers for X-ray Diffraction Patterns Analysis

Titouan Simonnet*†, Mame Diarra Fall*, Sylvain Grangeon†, Bruno Galerne*‡
*Institut Denis Poisson, Université d'Orléans, Université de Tours, CNRS, France
†BRGM, 3, Avenue Claude Guillemin, 45060 Orléans cedex 2, France
‡Institut universitaire de France (IUF)

***Abstract*—Understanding materials properties depends largely on the ability to determine its components, and in particular its mineral phases. Powder X-ray diffraction (XRD) is a powerful tool for such purposes. This paper presents a Transformer-based vision model (ViT) for mineral phase identification, and proportion inference to quantify the mineral phases present in a material. Our analysis shows that the tokenization strategy is a critical step for XRD pattern analysis. The results obtained for both tasks are excellent and more robust than those obtained with a CNN. The proposed approach also makes it possible to introduce visualization tools for signal analysis, to better understand how information flows through the model and how data is classified or quantified.**

***Index Terms*—Vision Transformers, Deep Learning, XRD patterns**

I. Introduction

Powder X-Ray Diffraction (XRD) analysis [1] is a key technique for identifying and quantifying mineral phases in natural (e.g. soils, sediments) and synthetic (e.g. cement materials, batteries) materials, and thus for understanding their chemical and mechanical characteristics. The signal resulting from the X-ray scattering is known as an XRD pattern. Two types of signals are considered in the present study. Firstly, single-phase XRD patterns that consists of a single mineral phase component and yield to a classification problem referred to as phase identification. Secondly, multi-compound XRD patterns the analysis of which aims at retrieving the proportions of each mineral phase in the material.

The so-called Rietveld refinement of an experimental XRD pattern is probably the most widely used method for refining structural parameters and quantifying mineral phases [2]. However, the application of this method requires preliminary phase identification by an expert user, which is time-consuming or virtually impossible in the case of XRD computed tomography (XRD-CT) [3]–[5].

Deep learning [6] is becoming a state-of-the-art technique for large-scale data analysis, for example in natural language processing, computer vision, and image and signal processing. XRD pattern analysis is no exception, and numerous solutions using Neural Networks (NN) have been proposed in recent years [7]. Artificial Neural Network was first used to analyze XRD patterns of clay [8]. Next, classification methods were used to identify mineral phases [9], or to classify mineral phases according to various factors such as space group or crystal size [10]–[13]. However, to date, there are only a few methods for quantifying mineral phases in mixtures. The authors of [14] have simplified the problem by dividing the proportion space into different classes, thus turning the problem into a classification one. Conventional neural networks (CNNs) show promising results for identifying mineral phases from their XRD patterns, but remain less accurate for quantifying phases within a mixture [15]. Several elements of the signal structure contribute to the limitations of these models. The main difficulty lies in the fact that the recognition of mineral phases from the signal is not solely based on one or two main peaks, but rather on the simultaneous presence of several characteristic peaks. Besides, the particular structure of XRD pattern is not invariant by translation and necessitate analysis relative to peak absolute positions as well as peak co-occurence.

These observations lead us to look for more robust architectures and Transformers seem a strong alternative model. They are built on the principle of self-attention [16] to treat data as a sequence where the data position is explicitly encoded. They are a key component, for example in language modeling [16] and computer vision with visual Transformers (ViT) [17]. They have recently emerged in signal processing, for example in hyperspectral unmixing [18] or electrocardiogram analysis [19]. As far as XRD is concerned, research is still in its infancy, for example with the use of ViT to identify phases [20]. In addition to providing powerful modeling capabilities, ViTs offer an Attention Rollout [21] to observe how information propagates through the model.

Our proposal in this paper is twofold. Firstly, we aim to assess the capabilities of a ViT in phase identification tasks using single-phase XRD patterns. Secondly, we tackle the more complex problem of proportion estimation for multi-component XRD patterns using a ViT. To our knowledge, this paper is the first to propose such an approach. Another important innovation is the integration of visualization tools to understand the behavior of Transformers on 1D signals.

II. Vision Transformers for XRD patterns

A. Adapting ViT architecture

We use the traditional architecture of a ViT [17], adapted to our 1D signal problem. Figure 1 outlines the model, which is almost identical to that of Chen *et al.* [20], and consists of multiple layers, starting with normalization, followed by a multi-head attention (MHA) block. At MHA output, there

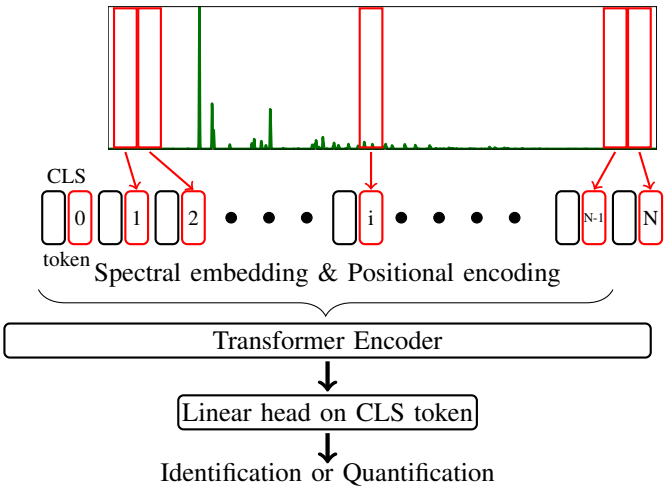


Fig. 1: Schematic representation of the Visual Transformer architecture [17]. The input embedding for XRD patterns simply consists in taking the signal segment of length $C = 80$ represented by the red rectangle areas.

is again normalization followed by passage through a Multi-Layer Perceptron (MLP). Residual connections are added to ensure the model stability. Finally, an encoded output is obtained and its class token (CLS) component is used to perform the final task (classification or proportion estimation) by passing it through a linear head.

Before the original XRD pattern enters in the encoder block, it must first undergo spectral embedding (or tokenization) which is a critical step for signal processing. According to [20], the most efficient strategy for XRD patterns is to partition the signal into N segments of length C (C being an hyperparameter). In most cases, the final part of the signal is not included in the tokenization, unless the signal length is a multiple of C. Anticipating on our results presented in Section III, our experiments show that choosing a good value for the length C is critical: For our signal of length 2905, the ViT architecture is most efficient for $C = 80$ (Table I), resulting in a partition of 36 segments to which the CLS token is added. It is worth mentioning that other approaches can be considered for tokenizing the input signal, e.g. using a CNN [18], [19]. However, as Table I shows, this method does not appear to be effective for our application.

To complete the tokenization, positional encoding vectors are concatenated to each token. In general, in image and signal processing, each positional encoding vector is a learnable parameter, and the initialization values are zeros [17], [20].

B. Explainability and visualization for ViT

Positional encoding. The advantage of ViTs in the application under consideration lies in their ability to provide insight into signal understanding through visualization. Firstly, by leveraging the positional encoding learnable parameters, these vectors encode the positions of each segment after spectral embedding. Each vector corresponds to a diffraction angular range, allowing us to observe the proximity (via cosine similarity [17]) between these vectors and uncover dependencies between different angular ranges. We can for example determine whether the presence of a peak in one angular range influences the presence of a peak in another range.

Attention Rollout. Another visualization tool is Attention Rollout [21], which uses the attention matrix to observe how information propagates through the layers of the self-attention part of the model. Denoting $FA_\ell = (A_\ell + I)/\max(A_\ell + I)$, where A_ℓ is the fused attention matrix on the all attention head (using the minimum) and I the identity matrix, the Attention Rollout matrix AR_L is recursively defined as follows:

$$AR_1 = FA_\ell; AR_\ell = (FA_\ell + I) \times AR_{\ell-1},\ \ell \in [\![2\ ;\ L]\!]. \quad (1)$$

The identity matrix is added to account for residual connections between layers [1]. Initialization consists in retrieving the attention matrix A of the first layer. The matrix of interest is that of the output corresponding to the last layer $L = 12$. From this matrix, we retain only the first row associated with the CLS token, as this will be used exclusively for the final task (e.g., classification or proportion inference). This row allows us to identify the segments that have contributed most to the value of the CLS token and, consequently, the most relevant angular ranges.

III. Phase identification for single phase XRD patterns

A. Dataset

All our XRD patterns are signals $\mathbf{x} \in \mathbb{R}^d$ with signal length $d = 2905$. We first consider a classification problem: given a single-phase XRD pattern, the objective is to determine the associated mineral phase among a set of K classes. Synthetic XRD patterns are considered here. The problem involves $K = 6$ different classes: Calcite ($CaCO_3$, space group R -3 c), Halite (NaCl, F m 3 m), Hematite (Fe_2O_3, R -3 c), Dolomite ($CaMgC_2O_6$, R -3), Gibbsite (AlO_3H_3, P 1 21/n 1), and Quartz (SiO_2, P 32 2 1). Each signal consists of 2905 points over an angular range from 4.0001° to 90.020055° (2θ CuKα) and the X-Ray wavelength is 1.5418. We consider 1800 signals for network training and 600 for model testing.

B. Results

The ViTs are trained using Cross-Entropy loss for 100 epochs using Adam as optimizer, with a constant learning rate of 0.001 and a batch size of 64. Results for this classification task show very high performance (Table I). On the 600 synthetic data points in the test set, the CNN achieves a classification accuracy of 96%. All the ViT variants perform slightly worse than the CNN to noticeable exception of ViT-80 that further improves on these already excellent results, achieving an accuracy of 99.8%. Indeed only one XRD pattern was misclassified by this model, showcasing the its potential for this type of data. We emphasize that the CNN was trained with the same loss function, number of epochs, batch size and optimizer as the ViTs.

[1] https://jacobgil.github.io/deeplearning/vision-transformer-explainability

TABLE I: CNN and ViT performance for phase identification.

Method	Accuracy↑	F1 score↑	Recall↑
CNN	0.960	0.960	0.960
ViT-20	0.947	0.947	0.947
ViT-40	0.952	0.952	0.952
ViT-80	**0.998**	**0.998**	**0.998**
ViT-100	0.952	0.952	0.952
ViT-CNN	0.168	0.168	0.168

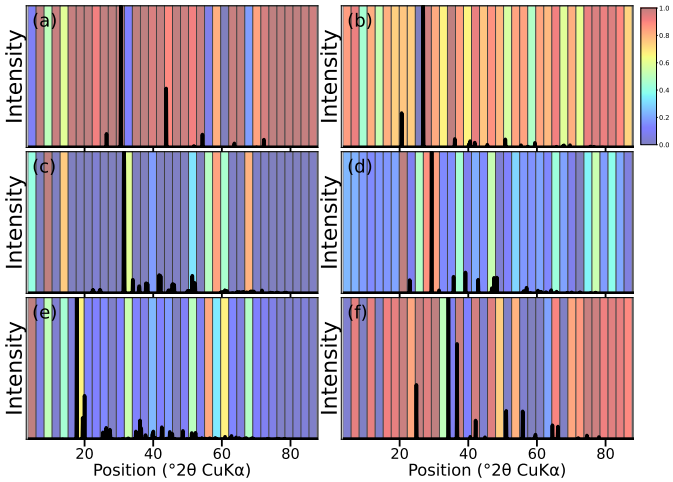


Fig. 2: Attention Rollout visualization for six XRD patterns from the classification test set: (a) Halite, (b) Quartz, (c) Dolomite, (d) Calcite, (e) Gibbsite and (f) Hematite.

In addition to this impressive performance, ViT also allows for visualization of the propagation of information. Starting from now we only consider the ViT-80 variant which is simply referred as ViT.

C. Visualization

Figure 2 shows which angular ranges were decisive for the classification of each class. In most cases, this result is straightforwardly interpretable, such as for Calcite, Gibbsite, or Dolomite, with values close to 1 on the signal peaks and low values elsewhere. This indicates that ViT mainly uses angular ranges with peaks to assign a data point to the right class. However, in other cases, these graphs are more difficult to interpret, with a majority of angular ranges used for classification. The absence of peaks in a number of cases may be one of the reasons why a class distinguishes itself in this way.

Given these results and the significant improvement provided by ViT, it was justified to test it on a more complex problem of mineral phase quantification.

IV. Phase quantification in multi-compounds XRD patterns

In this section, we consider a more challenging problem: the signal $\mathbf{x}$ no longer corresponds to a single mineral phase, but rather to a linear combination of several phases, i.e $\mathbf{x} = \sum_{i=1}^{K} p_i \mathbf{c_i}$, where p_i and $\mathbf{c_i}$ are both related to the i^{th} class ($i = 1, \ldots, K$), and represent respectively the proportion of mineral phase in the mixture and the associated XRD pattern. K denotes the total number of mineral phases considered for the mixture. Hence, $\mathbf{p} = (p_1, \ldots, p_k)$ forms a vector of proportions, with each $p_i \geq 0$ and $\sum_{i=1}^{K} p_i = 1$. Each signal $\mathbf{c_i}$ lives in $\mathbb{R}^d$ and, due to intraclass variation, the matrix $\mathbf{c} \in \mathbb{R}^{K \times d}$ is unknown. This intra-class variability generates significant fluctuations between the different XRD patterns associated with the same mineral phases. This is mainly due to variations in unit-cell parameters and crystallite size. Finally, considering a XRD pattern $\mathbf{x}$, the challenging task presented here is to recover the proportion vector $\mathbf{p} \in \mathbb{R}^K$.

A. Datasets

The data are multi-compound XRD patterns composed of $K = 4$ mineral phases (Calcite, Dolomite, Gibbsite, and Hematite) provided as supplementary material from [22]. The angular range and number of acquisition points are identical to those of the previous dataset. Independent synthetic data were used to create the training (10,000), validation (2,500), and testing (2,500) sets. This synthetic dataset is complemented by a set of 32 laboratory XRD patterns we additionally use to test the network in a realistic setting.

B. Results

The ViT was trained over 1000 epochs with the same parameters as the classification task. Only the loss function and the batch size (128) were different. A novelty is introduced: the ViT was trained with a specific loss for proportion estimation proposed in [15]. It combines Dirichlet modeling and Mean Square Error and appears to be the most efficient and stable loss for proportion inference.

In order to compare with the CNN method used in [15], we use the same evaluation metrics. The aim is to compare the true proportion vector with that predicted by the model. The Root Mean Square Error (RMSE) calculates the average error over the entire proportion vector. The Mean Maximal Absolute Error (MMAE) allows comparison by considering the maximum error for each data. Finally, we evaluate the network ability to correctly identify the mineral phases of a mixture through the Rate of Recovered Support (RRS).

Table II reports these measures, comparing the performances of the CNN and the ViT. Regarding computation time (performed on a NVIDIA Tesla P4), the training time for the CNN is faster since it only requires 100 epochs while the ViT training necessitates 1,000 epochs. Regarding performances, whatever the measure, the CNN is the best model on the synthetic testing dataset, with errors of less than 1% for phase quantification. In contrast, the ViT is less efficient on synthetic data, with errors three times higher (MMAE and RMSE). However, this relative counterperformance of ViT is counterbalanced when applied on real data. Indeed, ViT obtains better results on real laboratory data, suggesting that it does not adapt too much to the synthetic data. Compared with the CNN, there is a 1% improvement in MMAE and RMSE, and a significant increase of over 25% in RRS. These are excellent results, but still far from the Rietveld method, which is very effective on this type of laboratory data. Rietveld

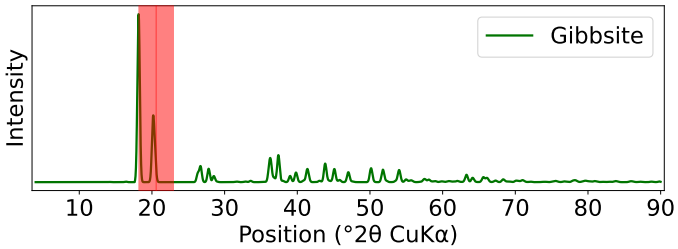
TABLE II: Performance comparison between CNN and ViT for proportion inference on XRD patterns (measures in percentage)

			Simulated data			Real data		
Method	Parameters	Training time (GPU)	RMSE ↓	MMAE ↓	RRS↑	RMSE↓	MMAE↓	RRS↑
CNN	832,868	13 minutes	**0.49%**	**0.55%**	**97.4%**	5.16%	6.96%	71.87%
ViT	934,564	170 minutes	1.23%	1.56%	87.2%	4.56%	5.81%	96.9%
Rietveld			Depends on the sample			**1.3%**	**2.07%**	**100%**

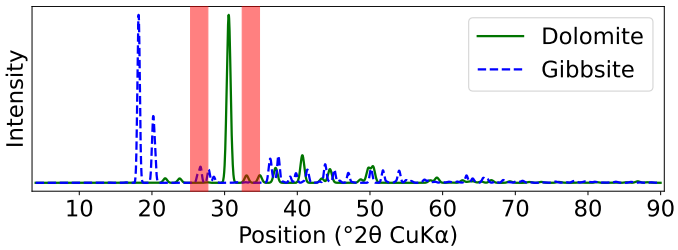
refinement was not tested on synthetic data, as the data generation method is similar to that used by Rietveld to refine quantification parameters.

C. Visualization

This study can be supplemented by visualization; here we focus on positional encoding. As mentioned above, this allows us to observe the links between the different angular ranges. Figure 3 shows the two most important links between tokens 6-7 (a) and 9-12 (b) respectively, with cosine similarity values close to 0.4 in absolute value. The most positively correlated corresponds to the co-occurence of two characteristic peaks for the Gibbsite class (Figure 3 (a)) while the most negatively correlated positions corresponds to an alternation of secondary peaks and flat areas for the Dolomite and Gibbsite classes (Figure 3 (b)).



(a) token 6-7, cossim = 0.42



(b) token 9-12, cossim = - 0.41

Fig. 3: Relationship between the most positively correlated tokens (a), the most negatively correlated tokens (b).

V. Conclusion

In this paper, we have proposed and demonstrated the suitability of Visual Transformers for processing X-ray diffraction signals. The choice of Transformers was motivated on one hand by the particular structure of these signals that is not invariant by translation and necessitate analysis relative to peak absolute positions, and on the other by the ability of ViT to outperform other types of architectures in various tasks. The promising results obtained for the two tasks considered here demonstrate the potential of Transformers, which outperform a CNN for the classification task and also appear to better generalize for the proportion inference task. This is illustrated by the results on real data, which are better than those obtained with a CNN despite lower performance on simulations. In addition to performance, the visualization tools show us how the network analyse the signals of interest.

Our work adds yet another example of a practical application for which ViT is highly effective, illustrating the versatility of this architecture. Regarding XRD pattern analysis, this work opens the way for further investigations on scaling these promising results to more complex databases and possibly training an XRD fundation model using all available phase descriptions.

References

[1] M. T. Fernandez-Diaz and M. H. Lemée-Cailleau, "Max von laue–hundred years of crystal diffraction," *Neutron News*, vol. 24, no. 2, pp. 11–12, 2013.

[2] H. M. Rietveld, "A profile refinement method for nuclear and magnetic structures," *Journal of applied Crystallography*, vol. 2, no. 2, pp. 65–71, 1969.

[3] S. DM. Jacques, M. Di Michiel, S. AJ. Kimber, X. Yang, R. J. Cernik, A. M. Beale, and S. JL. Billinge, "Pair distribution function computed tomography," *Nature Communications*, vol. 4, no. 1, pp. 2536, 2013.

[4] K. MØ. Jensen, X. Yang, J. V. Laveda, W. G. Zeier, K. A. See, M. Di Michiel, B. C. Melot, S. A. Corr, and S. JL. Billinge, "X-ray diffraction computed tomography for structural analysis of electrode materials in batteries," *Journal of The Electrochemical Society*, vol. 162, no. 7, pp. A1310, 2015.

[5] F. Claret, S. Grangeon, A. Loschetter, C. Tournassat, W. De Nolf, N. Harker, F. Boulahya, S. Gaboreau, Y. Linard, X. Bourbon, et al., "Deciphering mineralogical changes and carbonation development during hydration and ageing of a consolidated ternary blended cement paste," *IUCrJ*, vol. 5, no. 2, pp. 150–157, 2018.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT press, 2016.

[7] V. A. Surdu and R. Győrgy, "X-ray diffraction data analysis by machine learning methods—a review," *Applied Sciences*, vol. 13, no. 17, pp. 9992, 2023.

[8] D. T Griffen, "Quantitative phase analysis of clay minerals by x-ray powder diffraction using artificial neural networks. i. feasibility study with calculated powder patterns," *Clay minerals*, vol. 34, no. 1, pp. 117–126, 1999.

[9] J. W. Lee, W. B. Park, M. Kim, S. P. Singh, M. Pyo, and K. S. Sohn, "A data-driven xrd analysis protocol for phase identification and phase-fraction prediction of multiphase inorganic compounds," *Inorganic Chemistry Frontiers*, vol. 8, no. 10, pp. 2492–2504, 2021.

[10] F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. TP. Hartono, S. Ramasamy, B. L. DeCost, S. IP. Tian, G. Romano, et al., "Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks," *npj Computational Materials*, vol. 5, no. 1, pp. 1–9, 2019.

[11] P. M. Vecsei, K. Choo, J. Chang, and T. Neupert, "Neural network based classification of crystal symmetries from x-ray diffraction patterns," *Physical Review B*, vol. 99, no. 24, pp. 245120, 2019.

[12] A. Zaloga, V. Stanovov, O. Bezrukova, P. Dubinin, and I. Yakimov, "Crystal symmetry classification from powder x-ray diffraction patterns using a convolutional neural network," *Materials Today Communications*, vol. 25, pp. 101662, 2020.
[13] W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin, and K. S. Sohn, "Classification of crystal structure using a convolutional neural network," *IUCrJ*, vol. 4, no. 4, pp. 486–494, 2017.
[14] J. W. Lee, W. B. Park, J. H. Lee, S. P. Singh, and K. S. Sohn, "A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic xrd powder patterns," *Nature communications*, vol. 11, no. 1, pp. 1–11, 2020.
[15] T. Simonnet, M. D. Fall, B. Galerne, F. Claret, and S. Grangeon, "Proportion inference using deep neural networks. applications to x-ray diffraction and hyperspectral imaging," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 1310–1314.
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[17] A. Dosovitskiy, L. Beyer, A.r Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.
[18] P. Ghosh, S. K. Roy, B. Koirala, B. Rasti, and P. Scheunders, "Hyperspectral unmixing using transformer network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.
[19] C. Che, P. Zhang, M. Zhu, Y. Qu, and B. Jin, "Constrained transformer network for ecg signal processing and arrhythmia classification," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 184, 2021.
[20] Z. Chen, Y. Xie, Y. Wu, Y. Lin, S. Tomiya, and J. Lin, "An interpretable and transferrable vision transformer model for rapid materials spectra classification," *Digital Discovery*, vol. 3, no. 2, pp. 369–380, 2024.
[21] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online, July 2020, pp. 4190–4197, Association for Computational Linguistics.
[22] T. Simonnet, S. Grangeon, F. Claret, N. Maubec, M. D. Fall, R. Harba, and B. Galerne, "Phase quantification using deep neural network processing of xrd patterns," *IUCrJ*, vol. 11, no. 5, 2024.