XRDMamba: Large-scale Crystal Material Space Group Identification with Selective State Space Model

Liheng Yu
yuliheng@mail.ustc.edu.cn
University of Science and Technology of China
Suzhou, Jiangsu, China

Pengkun Wang*
pengkun@ustc.edu.cn
University of Science and Technology of China
Suzhou, Jiangsu, China

Zhe Zhao
zz4543@mail.ustc.edu.cn
University of Science and Technology of China
Suzhou, Jiangsu, China

Zhongchao Yi
zhongchaoyi@mail.ustc.edu.cn
University of Science and Technology of China
Suzhou, Jiangsu, China

Sun Nan
sunnan@hust.edu.cn
Huazhong University of Science and Technology
Wuhan, Jiangsu, China

Di Wu
wdcxy@mail.ustc.edu.cn
University of Science and Technology of China
Suzhou, Jiangsu, China

Yang Wang*
angyan@ustc.edu.cn
University of Science and Technology of China
Suzhou, Jiangsu, China

Abstract

In material science, the properties of crystalline materials largely depend on their structures, and space group is a key descriptor of crystal structure. With the rapid advancement of deep learning, the traditional artificial structure analysis method based on X-ray diffraction (XRD) has become cumbersome and is being gradually supplanted by neural networks. However, existing models are too simplistic and lack a comprehensive understanding of material structure. Our approach XRDMamba integrates chemical knowledge and presents a fresh crystal planes perspective on XRD data. We also introduce a knowledge-driven model for space group identification tasks. We have thoroughly analyzed our approach through numerous experiments, observing its SOTA performance and excellent generalization capabilities. The code is available in https://github.com/baigeiguai/XRDMamba.

CCS Concepts

• **Computing methodologies** → **Knowledge representation and reasoning**.

Keywords

Crystal material, Space group, X-ray diffraction, State space model

*Corresponding Author

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3680006

ACM Reference Format:
Liheng Yu, Pengkun Wang, Zhe Zhao, Zhongchao Yi, Sun Nan, Di Wu, and Yang Wang. 2024. XRDMamba: Large-scale Crystal Material Space Group Identification with Selective State Space Model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3680006

1 Introduction

In the field of materials science, the properties of materials are largely determined by their atomic arrangement and crystal structure. Determining the crystal structure of a material is crucial for understanding its mechanical, electromagnetic, and thermodynamic properties [13, 19, 21]. In crystallography, the space group is a way to describe the structural characteristics of a crystal [11, 14]. There are theoretically 230 space groups, which represent different combinations of atomic symmetries and unit cell arrangements in a crystal. In material analysis, powder X-ray diffraction (XRD) [1] is a critical technique for characterizing materials. The diffraction pattern generated by XRD encodes information about the crystal symmetry, lattice parameters, crystal type, and atomic stacking of nano-scale domains [4, 16]. Traditional XRD-based crystal structure determination methods, such as indexing techniques, require extensive manual operations and prior knowledge obtained from the material [2, 8–10, 12, 16]. Unfortunately, manual crystal structure determination becomes more challenging, time-consuming, and less accurate when dealing with materials containing small amounts of impurity phases or characterizing materials for which no background knowledge is available.

Recently, due to the need for automated XRD analysis, deep learning-based space group identification(as shown in Figure 1) has attracted the attention of researchers [15, 20]. Many studies primarily employ end-to-end black-box models, which treat the

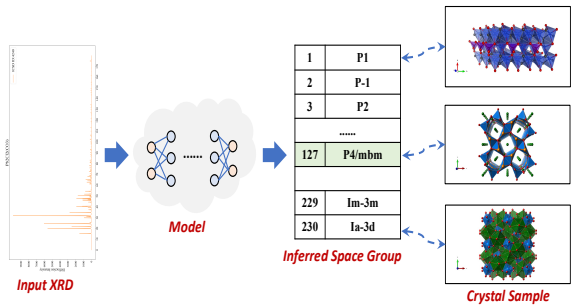

Figure 1: Motivation of XRDMamba. We employ deep learning models to replace experienced scientists in swiftly determining the corresponding space group type of a given crystal.

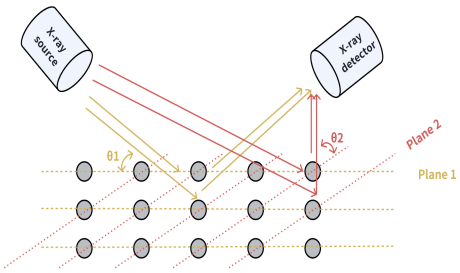

Figure 2: Principle of XRD spectrum obtained by diffraction of crystal planes.

XRD spectrum as a sequence and use common sequence representation models for encoding, resulting in a multi-classification result. For example, Salgado [17] attempted to characterize XRD using simple MLP and convolutional neural network (CNN) structures. However, such modeling approach is overly simplistic and lacks certain chemical knowledge. The performance of these models significantly decreases when dealing with datasets containing a large number of crystal classes or crystals with complex structures. To address this issue, RCNet [3] reduces the identification difficulty of the task by customizing crystal structure classes and introduces residual convolutional networks to improve identification accuracy. Unfortunately, this classification method fails to meet the requirements of all scenarios. Essentially, this approach can be seen as a secondary choice or a compromise.

To address the aforementioned issues, we combine knowledge from the field of materials science and propose a novel perspective for processing XRD spectral data, starting from the perspective of crystal planes. This perspective provides a better modeling of XRD spectra and offers improved inputs for subsequent model training. It is worth noting that the recent popular deep learning architecture, Mamba [7], has introduced a selective mechanism into the state space model, enabling it to discern the importance of information similar to an attention mechanism. For XRD spectra of crystal materials, the length of spectral sequences is often in the thousands or tens of thousands. Therefore, when directly applying self-attention models like Transformer, their high-dimensional complexity often leads to extremely high computational requirements. In contrast, the selective state space model [6] (SSM) is beneficial for improving this problem. Thus, based on the aforementioned novel analytical perspective, we propose a complex selective SSM-based architecture, called XRDMamba, which can better encode the structural information of crystals. Furthermore, we validate our approach on a large-scale CCDC crystal database. Comprehensive experimental results demonstrate that XRDMamba outperforms other baselines significantly and exhibits impressive performance even on out-of-domain data.

Our contributions in this paper are summarized as follows:

- *New insight*: for the first time, we incorporate chemical knowledge into DL-based space group identification and propose a novel perspective for analyzing XRD spectra from the viewpoint of crystal planes.
- *New model*: we introduce a knowledge-driven model, called XRDMamba, which accurately encodes XRD spectral data.
- *Compelling empirical results*: we conduct extensive experiments on a large-scale CCDC dataset, demonstrating the effectiveness and generalization of our approach.

2 Preliminaries

2.1 Problem Statement

As shown in Figure 2, the input for the analysis task is an XRD spectrum, which is a curve of length n obtained from diffraction experiments on a crystalline powder. The curve is represented as $S = [S_1, S_2, ..., S_\theta, ..., S_n]$, $(\theta \in [0^\circ, \delta, 2\delta, 3\delta, ..., 180^\circ])$, where θ represents the diffraction angle of the X-ray and S_θ represents the corresponding diffraction intensity when the incident angle is θ. The interval δ is the interval used for recording the diffraction intensities. In the analysis of crystal structures, the diffraction intensities at high angles are often very small and have minimal impact on the analysis. Therefore, researchers typically extract key segments for processing. The range of θ is usually defined as an interval $R(\theta_l, \theta_r, \delta)$, where $0^\circ \le \theta_l \le \theta_r \le 180^\circ$ (typically $\theta_l \le 5^\circ$ and $50^\circ \le \theta_r \le 130^\circ$). Here, $R(\theta_l, \theta_r, \delta)$ represents $[\theta_l, \theta_l + \delta, \theta_l + 2\delta, ..., \theta_r]$. The output of the analysis task is the space group class $Y \in [0, 229]$, which describes the 230 different theoretical crystal structures. Previous methods [3, 17] directly used S to represent the XRD curve and trained a mapping $f : X \to Y$ using simple MLP or one-dimensional convolution models.

2.2 Mamba Architecture

The Mamba architecture is a sequence transformation model that effectively captures dependency information in long sequences by incorporating a data-dependent selection mechanism and hardware-aware parallel algorithms into the state space model S4. It maintains nearly linear computational efficiency. The Mamba architecture consists of multiple stacked Mamba blocks. Each Mamba block takes a sequence $X \in \mathbb{R}^{B \times L \times D}$ as input, where B represents the batch size, L represents the sequence length, and D represents the dimension of each item in the sequence. Within each Mamba block, the data undergoes operations such as projection, state space modeling, and residual connections, resulting in an output $Y \in \mathbb{R}^{B \times L \times D}$. Therefore, the transformation of Mamba can be formulated simply as sequence to sequence: $Y = Mamba(X)$.

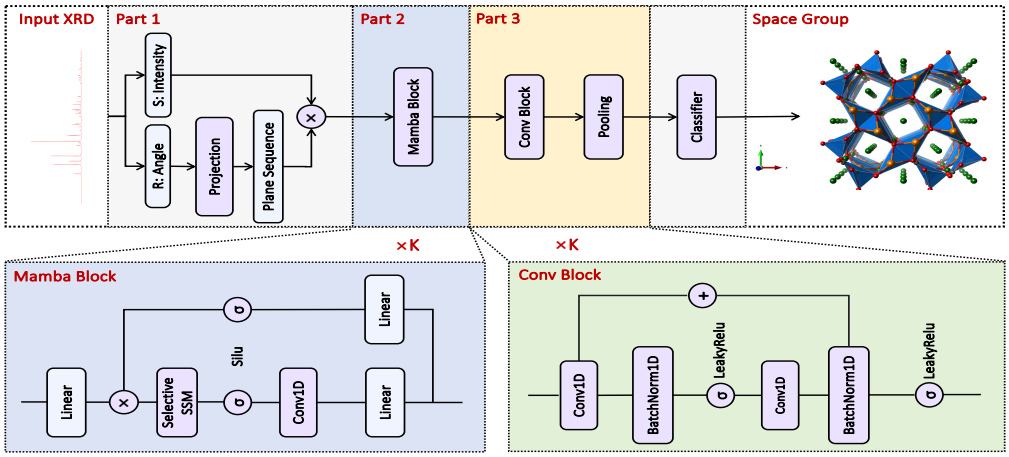

Figure 3: Overview of XRDMamba. From the perspective of chemical knowledge driven, it combines selective SSM to realize efficient space group identification.

3 Methodology

3.1 Crystal Plane Perspective

In theory, during the diffraction process, each diffraction angle corresponds to a crystal plane in the crystal. The diffraction intensity at a specific angle represents the presence of that type of crystal plane, where a higher intensity indicates a greater quantity of such planes. Therefore, the positions of peaks in XRD indicate the presence of important crystal planes in the crystal. We believe that these crystal planes largely reflect the structural information, while the peak intensity merely reflects the significance of each crystal plane in the crystal (an intensity of 0 indicates its non-existence). Hence, we propose using each angle θ in $R(\theta_l, \theta_r, \delta)$ to represent a specific crystal plane and S_θ to denote the importance of that crystal plane in the crystal. We differentiate the XRD spectrum data of a crystal into two sequences of length n: S and $R(\theta_l, \theta_r, \delta) \in \mathbb{R}^n$, which are processed separately. From this perspective, we aim to train a model that maps $(S, R(\theta_l, \theta_r, \delta))$ to Y, denoted as

$$f : (S, R(\theta_l, \theta_r, \delta)) \rightarrow Y \tag{1}$$

3.2 XRDMamba

From our perspective, as shown in Figure 3, we propose a knowledge-driven model called XRDMamba. The entire model consists of three components: the crystal plane representation module, the representation transformation module, and the classifier.

Part 1 - Crystal Plane Representation Module. In this module, our main objective is to tokenize the crystal planes and assign higher attention to the crystal planes corresponding to the peaks in XRD. Specifically, we start by using a linear projection layer to embed each diffraction angle θ into a k-dimensional vector, resulting in an $n \times k$ matrix that represents the n different crystal planes. Next, we consider the peak intensity S_θ at each θ angle as a measure of the importance of each crystal plane, which is used to weight the representation of the crystal planes. Therefore, we multiply the peak intensity vector with the crystal plane matrix to obtain an $n \times k$ feature matrix S'. More specifically,

$$S' = Projection_{1 \rightarrow k}(R(\theta_l, \theta_r, \delta)) \times S_\theta \tag{2}$$

where $S_\theta, R, S' \in \mathbb{R}^n$.

Part 2 - Representation Transformation Module. We aim to further characterize S' to learn the structural information of XRD and obtain an embedding $E \in \mathbb{R}^D$ with a dimension of D. Specifically, we apply several Mamba blocks for seq2seq transformations [18]. The selective state space model (SSM) adaptively selects the important crystal planes from the data and retains long-range contextual sequence information, which helps the model learn the correlations between crystal planes. Next, we use a stack of one-dimensional residual convolutional blocks and pooling layers for further characterization, as illustrated in Figure 3. The residual convolutional block consists of residual connections, one-dimensional convolutions, and normalization modules. The kernels in the one-dimensional convolutional layers and pooling layers are small-sized, and the number of channels in the convolutional blocks gradually increases. This approach of stacking multiple layers with small kernels and reducing dimensionality layer by layer enables the aggregation of information from all crystal planes, resulting in a final representation E that captures the characteristics of the crystal. The specific formula is as follows:

$$E = ResNet(Mamba(S')) \tag{3}$$

Part 3 - Classifier. We employ a simple MLP (Multi-Layer Perceptron) model as the classifier to classify E into 230 classes, yielding the final classification result as $Y = \text{MLP}(E)$.

4 Experimental

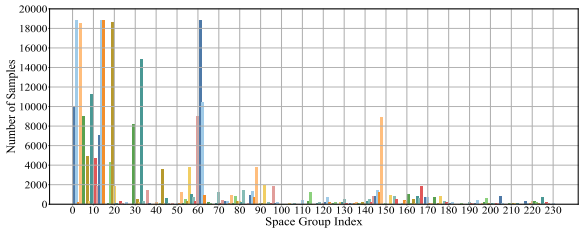

Figure 4: Distribution of metal-organic frameworks data acquired from the Cambridge Structural Database.

Dataset & Baseline. We obtained a dataset of over 280,000 metal-organic frameworks (MOFs) from the Cambridge Structural Database [5] for our experiments. The dataset covers 225 out of 230 classes, as shown in Figure 4. We used 50% of the data for model training, and the remaining 50% for evaluating the model's performance, which we refer to as MOF. However, due to the imbalanced nature of the test data, we further extracted a balanced dataset from the test set. Specifically, we sampled 10 samples from each class with more than 10 test samples, resulting in a balanced test dataset of over 200 samples, referred to as MOF-Balanced in the subsequent analysis. On the above two datasets, we selected four state-of-the-art methods in the field as our baselines, including MLP [17], CNN [17], NoPoolCNN [17] (CNN without pooling layers), and RCNet [3] (introduces residual connections based on NoPoolCNN).
Implementation. In the model, we set the number of stacked Mamba blocks to be 8. The dimensionality k representing the crystal planes is set to be 8. The hidden state dimensionality is set to be 16. We use the Adam optimizer with weight decay set to $1e^{-5}$ and a learning rate of $5e^{-5}$ to optimize the model. The model is trained for 100 iterations using cosine annealing learning rate scheduling, with a warm-up period of 20 iterations.

4.1 Benchmark Results

We compared the performance of XRDMamba and the SOTA methods on two test sets, MOF and MOF-Balanced. Table 1 presents the Top-1 accuracy and Top-2 accuracy of each model on the two test sets. Specifically, for simple models like MLP, CNN, and NoPoolCNN, their accuracy is generally below 50%, indicating unsatisfactory performance. After introducing residual connections, RCNet shows an improvement of more than 10% across all metrics. Our method incorporates new crystal plane perspectives and the Mamba module, resulting in an improved residual convolutional model that achieves the best performance on the MOF and MOF-Balanced datasets. It is evident that incorporating chemical knowledge and modeling is crucial for this task. The outstanding performance of XRDMamba can assist scientists in making rapid judgments on crystal structures to a certain extent.

Table 1: Accuracy (%) on CCDC dataset wtih state-of-the-art methods. Blod indicates the best performance. (+) indicates the the relative gain.

Method	Top-1 Accuracy		Top-2 Accuracy	
	MOF-Balanced	MOF	MOF-Balanced	MOF
MLP [17]	4.1 (+0.0)	9.1 (+0.0)	5.4 (+0.0)	15.1 (+0.0)
CNN [17]	22.9 (+18.8)	39.0 (+29.9)	32.4 (+27.0)	56.4 (+41.3)
NoPoolCNN [17]	33.8 (+29.7)	38.2 (+29.1)	40.7 (+35.3)	51.8 (+36.7)
RCNet [3]	44.5 (+40.4)	59.0 (+49.9)	55.5 (+50.1)	73.7 (+58.6)
XRDMamba	**48.7 (+44.6)**	**72.2 (+63.1)**	**61.7 (+56.3)**	**85.2 (+70.1)**

4.2 Ablation Study

We conducted further ablation experiments, and the results are shown in Table 2. Firstly, we used the traditional perspective by directly using the one-dimensional diffraction intensity as the model input and employed a residual convolutional network as the model backbone, which yielded relatively average results. Next, we introduced the crystal plane perspective while still using a residual convolutional network as the model backbone, and we observed a significant improvement in the F1 score in the test results. Finally, when the Mamba module was introduced, the model achieved an additional performance improvement of approximately 6%. This reflects that XRDMamba, incorporating chemical knowledge, can effectively learn the underlying patterns in XRD data.

Table 2: Ablation study (%) on CCDC dataset.

ResNet	Crystal Plane	SSM	Top-1 Accuracy	F1-Score	Top-2 Accuracy
✔			66.1	44.5	81.3
✔	✔		66.6	**47.7**	80.5
✔	✔	✔	**72.2**	47.6	**85.2**

4.3 Generalization Analysis

According to [17], we obtained data for over 8,000 inorganic crystals, which essentially belong to out-of-domain data. We tested and compared the model trained on the MOF training set using this out-of-domain data. The results, as shown in Table 3, indicate that XRDMamba demonstrates stable and excellent performance on out-of-domain data compared to all other SOTA models. This highlights its superior generalization ability.

Table 3: Generalization analysis (%) wtih SOTA methods.

Method	Top-1 Accuracy	F1-Score	Top-2 Accuracy
MLP [17]	15.5	8.5	21.4
CNN [17]	29.6	7.7	44.6
NoPoolCNN [17]	30.4	15.9	41.6
RCNet [3]	41.7	19.4	52.4
XRDResNet	50.5	22.6	62.6
XRDMamba	**54.5**	**24.1**	**64.7**

4.4 Crystal Case Study

We conducted visual analysis on some representative samples. As shown in Figure 5, for the crystal structure on the left, XRDMamba incorrectly predicted its space group type. However, we found that the Top-2 result was consistent with the correct space group. Further analysis revealed that $F4_132$ and $I4_132$ belong to the same crystal system and the same point group, making it challenging to effectively differentiate them solely based on the low-dimensional information from XRD. On the other hand, the crystal structure on the right belongs to a rare space group type (only two structures are included in the dataset). XRDMamba successfully predicted its space group type, demonstrating its robustness in dealing with challenging crystal materials.

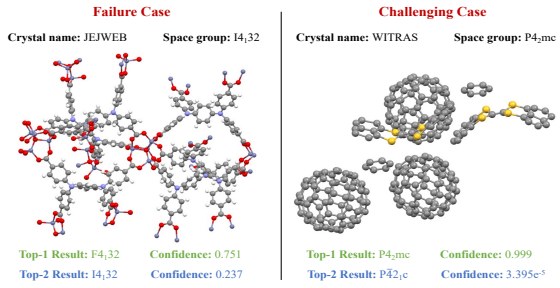

Figure 5: Typical case of crystal structure identification.

5 Conclusion

In this paper, we propose a chemically informed deep learning model, XRDMamba, that achieves accurate large-scale crystal material space group identification. We conducted comprehensive experiments on the renowned Cambridge Structural Database, demonstrating the superiority of XRDMamba.

References

[1] Asif Ali, Yi Wai Chiang, and Rafael M Santos. 2022. X-ray diffraction techniques for mineral characterization: A review for engineers of the fundamentals, applications, and research directions. *Minerals* 12, 2 (2022), 205.

[2] Angela Altomare, Rocco Caliandro, Mercedes Camalli, Corrado Cuocci, I da Silva, Carmelo Giacovazzo, Anna GraziaGiuseppina Moliterni, and Riccardo Spagna. 2004. Space-group determination from powder diffraction data: a probabilistic approach. *Journal of applied crystallography* 37, 6 (2004), 957–966.

[3] Litao Chen, Bingxu Wang, Wentao Zhang, Shisheng Zheng, Zhefeng Chen, Mingzheng Zhang, Cheng Dong, Feng Pan, and Shunning Li. 2024. Crystal Structure Assignment for Unknown Compounds from X-ray Diffraction Patterns with Deep Learning. *Journal of the American Chemical Society* 146, 12 (2024), 8098–8109.

[4] Marc De Graef and Michael E McHenry. 2012. *Structure of materials: an introduction to crystallography, diffraction and symmetry*. Cambridge University Press.

[5] Colin R Groom, Ian J Bruno, Matthew P Lightfoot, and Suzanna C Ward. 2016. The Cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* 72, 2 (2016), 171–179.

[6] Albert Gu. 2023. *Modeling Sequences with Structured State Spaces*. Stanford University.

[7] Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).

[8] Scott Habershon, Eugene Y Cheung, Kenneth DM Harris, and Roy L Johnston. 2004. Powder diffraction indexing as a pattern recognition problem: a new approach for unit cell determination based on an artificial neural network. *The Journal of Physical Chemistry A* 108, 5 (2004), 711–716.

[9] Armel Le Bail. 2004. Monte carlo indexing with mcmaille. *Powder Diffraction* 19, 3 (2004), 249–254.

[10] Armel Le Bail, Huguette Duroy, and Jean Louis Fourquet. 1988. Ab-initio structure determination of LiSbWO6 by X-ray powder diffraction. *Materials Research Bulletin* 23, 3 (1988), 447–452.

[11] Clare F Macrae, Paul R Edgington, Patrick McCabe, Elna Pidcock, Greg P Shields, Robin Taylor, Matthew Towler, and JVD Streek. 2006. Mercury: visualization and analysis of crystal structures. *Journal of applied crystallography* 39, 3 (2006), 453–457.

[12] Marcus A Neumann. 2003. X-Cell: a novel indexing algorithm for routine tasks and difficult cases. *Journal of applied crystallography* 36, 2 (2003), 356–365.

[13] John Frederick Nye. 1985. *Physical properties of crystals: their representation by tensors and matrices*. Oxford university press.

[14] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. 2013. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* 68 (2013), 314–319.

[15] Woon Bae Park, Jiyong Chung, Jaeyoung Jung, Keemin Sohn, Satendra Pal Singh, Myoungho Pyo, Namsoo Shin, and K-S Sohn. 2017. Classification of crystal structure using a convolutional neural network. *IUCrJ* 4, 4 (2017), 486–494.

[16] GS Pawley. 1981. Unit-cell refinement from powder diffraction scans. *Journal of Applied Crystallography* 14, 6 (1981), 357–361.

[17] Jerardo E Salgado, Samuel Lerman, Zhaotong Du, Chenliang Xu, and Niaz Abdolrahim. 2023. Automated classification of big X-ray diffraction data using deep learning models. *npj Computational Materials* 9, 1 (2023), 214.

[18] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems* 27 (2014).

[19] Jin Chong Tan and Anthony K Cheetham. 2011. Mechanical properties of hybrid inorganic–organic framework materials: establishing fundamental structure–property relationships. *Chemical Society Reviews* 40, 2 (2011), 1059–1080.

[20] Pascal Marc Vecsei, Kenny Choo, Johan Chang, and Titus Neupert. 2019. Neural network based classification of crystal symmetries from x-ray diffraction patterns. *Physical Review B* 99, 24 (2019), 245120.

[21] Angelo Ziletti, Devinder Kumar, Matthias Scheffler, and Luca M Ghiringhelli. 2018. Insightful classification of crystal structures using deep learning. *Nature communications* 9, 1 (2018), 2775.