ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11464651

CHEMICAL SIGHT NET: INCORPORATING CRYSTALLOGRAPHIC PRIORS FOR ACCURATE SPACE GROUP DETERMINATION FROM PXRD

Di Wu[1], Chen Zhang[1], Yudong Zhang[12], Pengkun Wang[12†], Yang Wang[12]

[1] University of Science and Technology of China
[2] Suzhou Institute for Advanced Research, USTC

ABSTRACT

Accurate space group determination from PXRD patterns is critical for materials science but remains challenging due to intensity variations and long-tailed space group distributions. To address this, we propose Chemical Sight Net (CSN), a deep learning framework that incorporates crystallographic domain knowledge. CSN employs a chemically informed embedding mechanism to preserve subtle peak information and integrates symmetry constraints into the loss function. This approach enhances feature representation and improves generalization, especially for underrepresented space groups. Experiments demonstrate state-of-the-art performance across benchmarks, achieving a Top-1 accuracy of 73.4%. This validates our principled approach of embedding scientific priors into deep learning, enabling interpretable and data-efficient PXRD analysis.

***Index Terms*—** PXRD, space group, long-tailed distribution, prior chemical knowledge

1. INTRODUCTION

Powder X-ray diffraction (PXRD) technology, which carries optical signatures of atomic arrangements and symmetry within materials, plays a decisive role in the precise determination of crystal structures for materials discovery and synthesis research [1, 2, 3]. Within this process, space group identification—a fundamental step in crystallographic symmetry analysis—has long attracted significant scholarly attention. Conventional PXRD-based space group determination methods rely heavily on manual expertise, requiring iterative trial-and-error and complex computations [4, 5]. This approach is not only labor-intensive and time-consuming but also constitutes a critical bottleneck limiting the efficiency of new material development [6].

Deep learning is increasingly emerging as a new paradigm for space group classification from PXRD data, enabling end-to-end mapping from diffraction patterns to space groups while demonstrating promising efficiency and accuracy. Pioneering studies demonstrated the feasibility of end-to-end mapping from PXRD patterns to space groups using deep neural networks (DNNs) and convolutional neural networks (CNNs) [7, 8, 9, 10]. *However, these methods exhibit limited generalizability: their predictive capabilities are often confined to specific subsets of space groups or narrow material systems, hindering broad applicability across diverse crystalline structures*.

Recently, a new architecture named XRDMamba has been developed, achieving higher accuracy in space group classification on expanded datasets and improving model adaptability to diverse crystal structures[11]. However, existing methods still face two critical **challenges**:

❶ The critical information in PXRD data resides in the correspondence between diffraction angles and intensity values[12, 13]. Conventional analytical approaches predominantly focus on high-intensity diffraction peaks, often overlooking structural information embedded in low-intensity peaks[14]. AI methods also tend to overlook low-intensity peak data due to this data format.[15, 16, 17]. *However, simply amplifying the weights of weak peaks would distort the original intensity distribution, thereby losing crucial comparative information between peaks*. Consequently, ***developing specialized frameworks that preserve the physical characteristics of PXRD while enabling intelligent feature extraction has become a critical technical challenge.***.

❷ Crystallographic data inherently exhibit a long-tail distribution. This phenomenon arises because *low-symmetry space groups tend to form more stable structures that are easier to synthesize, whereas high-symmetry space groups often require more complex structural arrangements that present greater synthetic challenges*. Consequently, the data distribution becomes severely imbalanced: while head-class samples may exceed 300,000 instances, tail-

†:Corresponding Author. The authors gratefully acknowledge the support from the National Natural Science Foundation of China (NSFC) under Grant Nos. 62402472, and 12227901. This work was also supported by the Natural Science Foundation of Jiangsu Province (No. BK20240461), the Key Basic Research Foundation of Shenzhen (No. JCYJ20220818100005011), the Research Grants Council of the Hong Kong Special Administrative Region (GRF Project No. CityU 11215723), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No. YSBR-005), and the Academic Leaders Cultivation Program at USTC.

class samples often dwindle to single-digit quantities[18]. ***This extreme scarcity of tail-class samples constitutes a classic long-tail problem in crystallographic symmetry identification***.

To address these challenges, we propose Chemical Sight Net (**CSN**), a novel hybrid deep learning framework designed for robust and highly accurate space group classification using PXRD patterns. The overall architecture of CSN is shown in Figure 1. It mainly includes the following two **key contributions**:

❶ PXRD embedding with chemical awareness: we introduce a chemically-weighted embedding module that assigns learnable chemical weights to diffraction peaks, which can enable the extraction of both dominant crystallographic plane signatures and subtle correlations between prominent and minor peaks.

❷ Handling Long-Tailed Distribution from a Chemical Perspective: CSN transforms the space group classification problem into loss computation within a latent space of symmetry rules. This knowledge-guided learning process significantly enhances the model's generalization capability, particularly for tail classes with scarce samples.

In summary, CSN integrates crystallographic priors into deep learning, enhancing the physical interpretability and generalization of PXRD analysis under long-tail distributions while retaining end-to-end efficiency. This study establishes a new paradigm for fusing diffraction physics with AI. Experimental results demonstrate state-of-the-art performance, particularly in accurately recognizing space groups with scarce samples.

2. PROPOSED METHOD

► **Problem Definition.** The input for the PXRD-based space group prediction task is an XRD spectrum, which is a curve of length $N \times 2$ obtained from diffraction experiments on a crystalline powder. The curve is represented by the vertical coordinate (***intensity***): $S = [S_1, S_2, ..., S_\theta, ..., S_n]$, and the horizontal coordinate (***angle***): $\theta \in [0°, \delta, 2\delta, 3\delta, ..., 90°]$. Here, θ represents the diffraction angle of the X-ray, and S_θ denotes the corresponding diffraction intensity when the incident angle is θ. The interval $\delta = 0.1°$ is the angular sampling interval used for recording diffraction intensities. The output of the task is the space group class $Y \in [0, 229]$, corresponding to 230 theoretical crystal structures. Additionally, we define symmetry rules as $R \in [r_1, r_2, ..., r_{46}]$, representing 32 spatial symmetry operations and 14 Bravais lattices, where $r_i \in \{0, -1, 1\}$. Here, 0 indicates the absence of the rule, 1 indicates its presence, and -1 represents inverse operations of certain rotations and translations. The objective of our model is to learn a mapping function:

$$f : (S, \theta) \rightarrow Y. \tag{1}$$

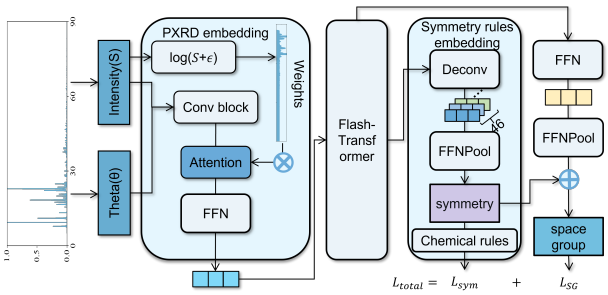

Fig. 1: Overview of the proposed CSN framework. CSN integrates crystallographic prior knowledge through three core modules: ❶ an attention-based PXRD embedding layer for diffraction data representation, ❷ Flash-Transformer blocks for linking local feature extraction with global dependencies, and ❸ a knowledge embedding output layer that incorporates chemical relationships between symmetry operations and space groups.

► **Overview of CSN.** CSN primarily consists of three modules: ❶ Processes raw diffraction patterns through spectral attention mechanisms to capture intensity-sensitive features while preserving angular relationships, ❷ Flash-Transformer[19] blocks that establish connections between local information extraction and global relationships, and ❸ an output layer that embeds the chemical relationships between symmetry operations and space groups into the network. We will subsequently detail the PXRD embedding methodology and the knowledge embedding approach for symmetry rules.

2.1. Attention-based PXRD Embedding

To prevent ***low-intensity peaks from being overlooked***, we propose an attention-based PXRD embedding module. The input pipeline uses 1D-convolutions for spectral feature extraction, capturing local patterns like doublet features while optimizing sequence length. Instead of conventional pooling, which discards weak signals, we employ a self-attention variant that ***assigns adaptive weights based on signal intensity***. By applying logarithmic preprocessing to peak intensities before computing the attention matrix, we ensure that intensity information guides the attention distribution. This mechanism prevents critical low-intensity diffraction peaks from being overshadowed by dominant signals. The formulation is as follows:

$$Q = W^Q \cdot I, \quad K = W^K \cdot I, \quad V = W^V \cdot I, \tag{2}$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}} \odot (W_{log} \cdot \mathbf{1}^\top)\right) V \tag{3}$$

$$\text{where} I = \text{Conv1D}(S|\theta) \in \mathbb{R}^{N \times L}. \tag{4}$$

where $\mathbf{1}^\top \in \mathbb{R}^{1\times L}$ is an all-ones row vector used for broadcasting W_{log} to the dimensions of the attention matrix.

2.2. Symmetry Rules Embedding

The symmetry rule embedding process is constructed by using neural networks to simulate the correspondence between crystal symmetry invariance and space groups. We diverge the Transformer output into two parallel processing streams: the ***first*** stream utilizes a standard feed-forward network with pooling operations for direct space group prediction.

$$Y_{\text{sg}} = \text{Pooling}\left(\text{FFN}(X_{\text{t}})\right), \tag{5}$$

where X_{t} is the output of the Flash-Transformer [19].

The ***second*** stream employs deconvolution operations for dimensional expansion, transforming the $N \times C$ feature matrix into an $N \times 46 \times L$ third-order tensor, thereby decomposing the single feature sequence into 46 independent subsequences, ***each corresponding to a specific symmetry rule***.

$$X_{\text{dec}} = \text{Deconv}(X_{\text{t}}) \in \mathbb{R}^{N\times 46\times L}. \tag{6}$$

To maintain mathematical rigor for Binary Cross-Entropy (BCE) optimization, we decompose each symmetry rule $R \in \{-1, 0, 1\}$ into two binary indicators representing "Presence" and "Directionality"[cite: 67, 68, 115]. The predicted rule vector and the corresponding loss are:

$$R_{\text{pred}} = \left[\text{FFN}(X_{\text{dec}}^{(1)}), \ldots, \text{FFN}(X_{\text{dec}}^{(46)})\right] \in \mathbb{R}^{46\times 2} \tag{7}$$

$$\mathcal{L}_{\text{sym}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{46}\sum_{k=1}^{2}\big[R_{i,j,k}\log(\hat{R}_{i,j,k}) + (1-R_{i,j,k})\log(1-\hat{R}_{i,j,k})\big] \tag{8}$$

where $k \in \{1, 2\}$ denotes the decomposed binary properties[cite: 106, 107, 120]. This maps targets into the $[0, 1]$ domain, ensuring the loss remains well-defined while capturing the orientation of symmetry operations[cite: 108, 116].

The rationale for decomposing space group classification into distinct symmetry rules is evidenced by the hierarchical relationships in Figure 2. While one-hot encodings treat groups like `P6`, `P6-m`, and `P6-mmm` as independent categories, crystallography reveals an evolutionary symmetry path. For instance, `P6-m` adds a horizontal mirror and inversion center to `P6`, whereas `P6-mmm` incorporates further vertical mirrors and two-fold axes. Our rule-based approach captures these underlying connections, enabling the model to learn structural hierarchies rather than isolated labels.

Using one-hot encodings during training forces the model to treat these structurally related space groups as completely independent categories due to the gradient signals from the loss function, which contradicts crystallographic principles. In contrast, ***the symmetry rules embedding approach guides the model to learn fine-grained symmetry relationships through gradient updates***, thereby capturing more accurate chemical knowledge. Moreover, ***this method mitigates the long-tailed distribution issue in space group data***. For instance, in the CCDC database, space group `Ccc2` has only 133 samples, while `Cc` has over 10,000. By sharing symmetry rules, tail classes (e.g., `Ccc2`) implicitly receive supplementary training signals from head classes (e.g., `Cc`) with overlapping symmetry characteristics.

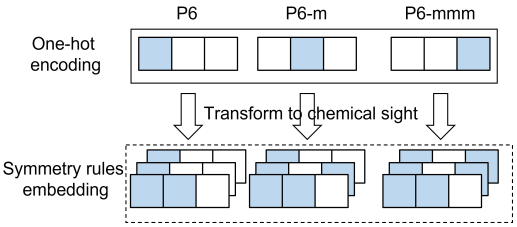

Fig. 2: Space Group Classification Encoding from Multiple Perspectives.

	Top-1 Accuracy (%)		Top-2 Accuracy (%)	
Method	MOF-Balanced	MOF	MOF-Balanced	MOF
MLP	4.1 (+0.0)	9.1 (+0.0)	5.4 (+0.0)	15.1 (+0.0)
CNN	22.9 (+18.8)	39.0 (+29.9)	32.4 (+27.0)	56.4 (+41.3)
NoPoolCNN	33.8 (+29.7)	38.2 (+29.1)	40.7 (+35.3)	51.8 (+36.7)
RCNet	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)
CSN	**50.5 (+46.4)**	**73.3 (+64.2)**	**66.1 (+60.7)**	**86.3 (+71.2)**

Table 1: Accuracy (%) on CCDC dataset with SOTAs. Bold indicates the best performance. (+) indicates the relative gain.

Notably, we do not entirely abandon one-hot encoding. The final loss function combines the conventional space group classification loss (based on one-hot labels) and the symmetry rules embedding loss, ensuring that the model both recognizes symmetry relationships and maintains discriminative capability among symmetry-similar space groups.

3. REUSLT AND ANALYSIS

► **Dataset and Baseline.** For our experimental study, we curated a dataset comprising over 200,000 metal-organic frameworks (MOFs) from the Cambridge Structural Database [18]. This dataset spans 225 out of 230 space group classes, the remaining five categories currently lack registered synthetic samples. We allocated 50% of the data for model training and reserved the remaining 50% for performance evaluation, referring to this split as the **MOF** dataset. To address class imbalance in the test set, we constructed a balanced subset following the methodology of XRD-Mamba: from each class containing more than 10 test samples, we randomly selected 10 instances, resulting in a balanced evaluation set of over 200 samples, denoted as MOF-Balanced. We compared our ap-

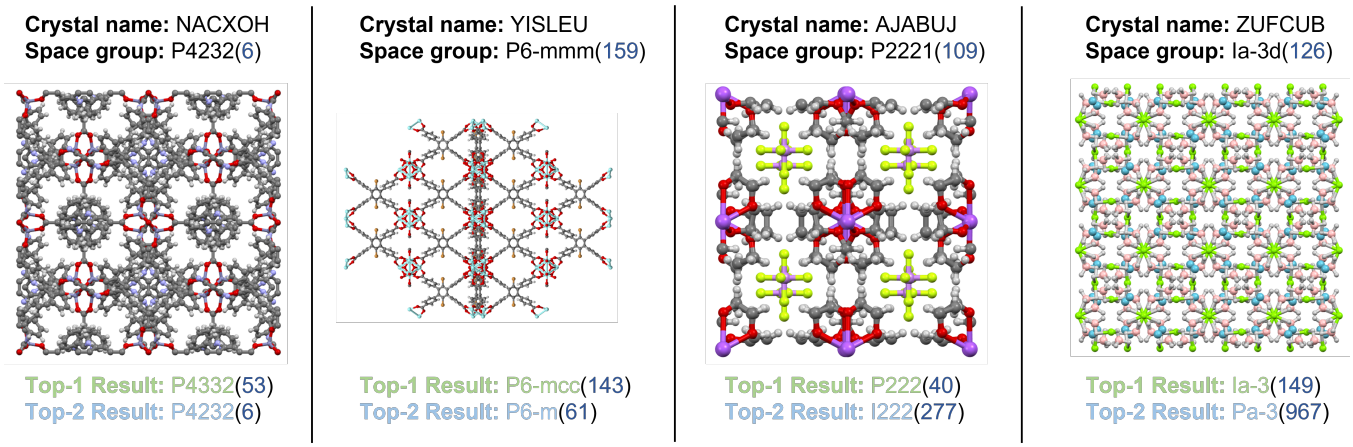

Fig. 3: Typical predicted case of CSN.

Components		Top-1 Accuracy (%)	
PxrdE	Sym-rule	MOF	MOF-Balanced
✘	✘	64.5	41.5
✔	✘	70.1	44.3
✘	✔	68.1	49.3
✔	✔	73.4	50.6

Table 2: Ablation study on CCDC dataset.

proach against five state-of-the-art methods: MLP [7], CNN [7], NoPoolCNN [7], RCNet [10], and XRDMamba[11].

► **Benchmark Results.** We conducted a systematic evaluation of CSN against current state-of-the-art (SOTA) methods on the aforementioned MOF and MOF-Balanced test sets. Table 1 presents the Top-1 and Top-2 accuracy of each model on both test sets. Notably, conventional models, including MLP, CNN, and NoPoolCNN, demonstrated suboptimal performance across all metrics. Based on the results, our CSN framework achieves superior performance by innovatively integrating PXRD peak intensity attention with crystalline symmetry rule mining: obtaining breakthrough results on both MOF and MOF-Balanced datasets- all accuracy metrics are above 50%. Experiments demonstrate that feature extraction based on PXRD data characteristics can more effectively capture structural information, while embedding the intrinsic regularities of symmetry and space groups enables the model to exhibit exceptional generalization capability on balanced datasets.

► **Ablation Study.** Systematic ablation results (Table 2) validate the contribution of each CSN module. Replacing the baseline 1D-CNN with our PXRDEmbedding module significantly improved feature extraction across both datasets. Furthermore, substituting the standard FFN-pooling with the Sym-rule block yielded a 3.6% and 7.8% accuracy gain on MOF and MOF-Balanced sets, respectively. The more pronounced improvement on the balanced set (+7.8% vs. +3.6%) highlights the module's effectiveness in mitigating long-tail distribution issues. Overall, the full CSN architecture achieved a 9% total gain over the baseline, confirming that integrating PXRD feature engineering with symmetry constraints enables effective learning of the underlying physical patterns in XRD data.

► **Case Study.** We selected four representative prediction cases for visual analysis. As shown in Figure 3, each case displays: crystal name, correct space group (with sample count), structural diagram, and Top-1 to Top-2 predictions. The model consistently predicts space groups sharing symmetry rules with ground truth—exactly the desired behavior. ***Even when incorrect, predictions remain scientifically meaningful. This indicates our model captures underlying symmetry information rather than treating space groups as isolated one-hot entities*** . Case analysis shows: For NACXOH (Case 1), the correct group appears in Top-2. For other cases (2-4), the model avoids head-class bias and predicts groups with high symmetry similarity, effectively mitigating the long-tail problem.

4. CONCLUSIONS

This paper introduces CSN, a physics-informed model that predicts crystal space groups from PXRD data with high accuracy, even for scarce samples. By embedding crystallographic priors, CSN achieves robust performance and remains chemically meaningful even in misclassifications, offering chemists a valuable diagnostic tool. Ultimately, our framework establishes a new paradigm for integrating scientific knowledge into AI to interpret complex diffraction patterns.

5. REFERENCES

[1] Vitalij K Pecharsky and Peter Y Zavalij, Fundamentals of powder diffraction and structural characterization of materials, Springer, 2003.

[2] Zhengyang Zhou, Lukáš Palatinus, and Junliang Sun, "Structure determination of modulated structures by powder x-ray diffraction and electron diffraction," Inorganic Chemistry Frontiers, vol. 3, no. 11, pp. 1351–1362, 2016.

[3] Anders S Larsen, Toms Rekis, and Anders Ø Madsen, "Phai: A deep-learning approach to solve the crystallographic phase problem," Science, vol. 385, no. 6708, pp. 522–528, 2024.

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

[5] Jin Chong Tan and Anthony K Cheetham, "Mechanical properties of hybrid inorganic–organic framework materials: establishing fundamental structure–property relationships," Chemical Society Reviews, vol. 40, no. 2, pp. 1059–1080, 2011.

[6] Jing Wei, Xuan Chu, Xiang-Yu Sun, Kun Xu, Hui-Xiong Deng, Jigen Chen, Zhongming Wei, and Ming Lei, "Machine learning in materials science," InfoMat, vol. 1, no. 3, pp. 338–358, 2019.

[7] Woon Bae Park, Jiyong Chung, Jaeyoung Jung, Keemin Sohn, Satendra Pal Singh, Myoungho Pyo, Namsoo Shin, and K-S Sohn, "Classification of crystal structure using a convolutional neural network," IUCrJ, vol. 4, no. 4, pp. 486–494, 2017.

[8] Angelo Ziletti, Devinder Kumar, Matthias Scheffler, and Luca M Ghiringhelli, "Insightful classification of crystal structures using deep learning," Nature communications, vol. 9, no. 1, pp. 2775, 2018.

[9] Jerardo E Salgado, Samuel Lerman, Zhaotong Du, Chenliang Xu, and Niaz Abdolrahim, "Automated classification of big x-ray diffraction data using deep learning models," npj Computational Materials, vol. 9, no. 1, pp. 214, 2023.

[10] Litao Chen, Bingxu Wang, Wentao Zhang, Shisheng Zheng, Zhefeng Chen, Mingzheng Zhang, Cheng Dong, Feng Pan, and Shunning Li, "Crystal structure assignment for unknown compounds from x-ray diffraction patterns with deep learning," Journal of the American Chemical Society, vol. 146, no. 12, pp. 8098–8109, 2024.

[11] Liheng Yu, Pengkun Wang, Zhe Zhao, Zhongchao Yi, Sun Nan, Di Wu, and Yang Wang, "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, 2024, pp. 4233–4237.

[12] Christopher G Pope, "X-ray diffraction and the bragg equation," Journal of chemical education, vol. 74, no. 1, pp. 129, 1997.

[13] T Dan Vu, Firas Krichen, Maud Barre, Sandrine Coste, Alain Jouanneaux, Emmanuelle Suard, Andrew Fitch, and François Goutenoire, "Ab initio structure determination of la34mo8o75 using powder x-ray and neutron diffraction data," Crystal Growth & Design, vol. 19, no. 11, pp. 6074–6081, 2019.

[14] Jürgen Brüning and Martin U Schmidt, "The determination of crystal structures of active pharmaceutical ingredients from x-ray powder diffraction data: a brief, practical introduction, with fexofenadine hydrochloride as example," Journal of Pharmacy and Pharmacology, vol. 67, no. 6, pp. 773–781, 2015.

[15] Yuta Suzuki, Hideitsu Hino, Takafumi Hawai, Kotaro Saito, Masato Kotsugi, and Kanta Ono, "Symmetry prediction and knowledge discovery from x-ray diffraction patterns using an interpretable machine learning approach," Scientific reports, vol. 10, no. 1, pp. 21790, 2020.

[16] Byung Do Lee, Jin-Woong Lee, Woon Bae Park, Joonseo Park, Min-Young Cho, Satendra Pal Singh, Myoungho Pyo, and Kee-Sun Sohn, "Powder x-ray diffraction pattern is all you need for machine-learning-based symmetry identification and property prediction," Advanced Intelligent Systems, vol. 4, no. 7, pp. 2200042, 2022.

[17] Pascal Marc Vecsei, Kenny Choo, Johan Chang, and Titus Neupert, "Neural network based classification of crystal symmetries from x-ray diffraction patterns," Physical Review B, vol. 99, no. 24, pp. 245120, 2019.

[18] Colin R Groom, Ian J Bruno, Matthew P Lightfoot, and Suzanna C Ward, "The cambridge structural database," Structural Science, vol. 72, no. 2, pp. 171–179, 2016.

[19] Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc Le, "Transformer quality in linear time," in International conference on machine learning. PMLR, 2022, pp. 9099–9117.