Space-Group Identification from Multi-Phase Powder XRD via Latent Phase Modeling

Tomoya Murata [1]

[1] *Toyota Motor Corporation, Toyota City, Aichi, Japan*. Correspondence to: tomoya_murata_aa@mail.toyota.co.jp.

1. Introduction

Powder X-ray diffraction (XRD) remains one of the most widely used experimental techniques for probing the crystal structure and symmetry of materials. Diffraction patterns encode lattice periodicity and crystallographic symmetry through the positions and intensities of Bragg peaks, thereby enabling the identification of crystalline phases and their associated symmetries. Conventional phase identification and structure analysis rely on expert-driven procedures such as peak indexing, extinction rule analysis, and Rietveld refinement [1, 2, 3, 4]. These methods have long served as the foundation of crystallographic analysis and are highly effective for well-resolved single-phase samples.

Motivated by the growing availability of large crystallographic databases and rapid advances in machine learning, substantial efforts have been devoted to automating powder XRD analysis using data-driven models [5, 6, 7, 8, 9, 10]. In particular, space-group prediction has attracted considerable attention as a critical early step in crystal structure determination, as accurate symmetry assignment provides strong constraints that substantially reduce the complexity of subsequent structure solution and refinement.

Despite substantial methodological differences, both conventional crystallographic methods and existing learning-based approaches to space-group prediction implicitly assume single-phase diffraction patterns with known phase identities, whether through explicit phase models [1, 2, 3, 4] or training on predefined crystallographic libraries [5, 6, 7, 8].

This limitation reflects a mismatch in problem formulation. In realistic experimental settings, powder XRD measurements are often multi-phase and may include crystalline phases that are not present in reference databases. Consequently, the observed diffraction profile arises from a superposition of signals, leading to composite peaks and ambiguous labels. Classical crystallographic approaches, including Rietveld refinement and related quantitative phase analysis, typically assume that all constituent phases are known a priori and focus on refining structural parameters, with phase fraction estimation as a secondary objective [2, 11, 12]. Recently, machine learning methods have been proposed to identify combinations of known crystalline phases from multi-phase diffraction spectra within fixed chemical systems [13, 14, 15, 16]. These approaches focus on phase identification within reference libraries and are not designed to infer crystallographic space groups from multi-phase diffraction profiles, particularly in the presence of unknown phases.

As a result, space-group prediction from multi-phase powder XRD profiles, in a chemistry-agnostic setting and in the presence of unknown crystalline phases, has not yet been systematically formulated or investigated in previous studies.

From a physical standpoint, a multi-phase powder XRD pattern can be viewed as the superposition of diffraction signals generated by multiple crystalline phases, each governed by its own space-group symmetry. This observation suggests that models which explicitly represent and disentangle latent phases in a representation space are naturally aligned with the generative process of diffraction and are therefore well suited to this problem.

In this work, we propose a latent phase decomposition framework for multi-label space-group prediction from multi-phase XRD profiles (Fig. 1). Instead of directly predicting space-group labels independently from the mixed diffraction signal, our method introduces a small number of latent space phase components that are jointly learned under supervision from space-group labels to capture phase-wise contributions relevant symmetry inference. These latent components serve as task-oriented intermediate representations, while avoiding explicit reconstruction of individual diffraction patterns—an intermediate objective that is inherently ambiguous and difficult in multi-phase XRD and not required for accurate symmetry inference. We evaluate the proposed method on a large-scale simulated multi-phase XRD dataset constructed from crystallographic databases, covering diverse phase compositions and the presence of out-of-library materials.

2. Results

2.1 Problem Setup and Models

We focus on the problem of identifying the set of space groups present in a multi-phase XRD profile. Given a diffraction intensity vector sampled on a 2θ grid, the task is formulated as multi-label classification, where the model outputs a multi-hot vector indicating the presence of one or more space groups.

All models share a common but learnable 1D convolutional front-end that extracts local peak features from the input profile, providing a unified low-level representation.

To study the impact of architectural inductive biases and explicit phase modeling, we evaluate three model variants: (i) a *multi-label presence-based CNN* model, which encodes the local peak features using a convolutional encoder and directly predicts space-group probabilities from a single global representation, serving as a naive multi-label baseline without explicit phase decomposition; (ii) a *multi-label presence-based Transformer* model, which replaces the

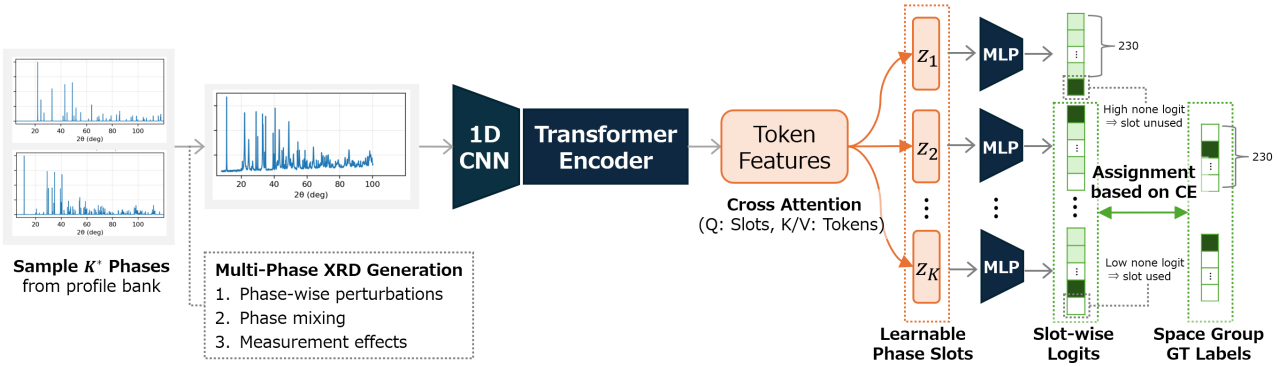

Fig. 1: Illustration of the proposed phase-clustering approach for multi-label space-group classification. Token features are aggregated into learnable phase slots via cross-attention, and slot-wise predictions are matched to ground-truth phases using a permutation-invariant cross-entropy assignment with a NONE class.

Table 1: Quantitative results for in-library (train) and out-of-library (test) evaluation. Micro-F1 is micro-averaged over all space-group classes. For top-K metrics, K is set to the number of ground-truth phases (K^*) or $3K^*$; cover@K and recall@K denote set-level coverage and recall within the top-K predictions, respectively.

	Train (in-library)					Test (out-of-library)				
Method	micro-F1	cover@K^*	recall@K^*	cover@$3K^*$	recall@$3K^*$	micro-F1	cover@K^*	recall@K^*	cover@$3K^*$	recall@$3K^*$
Multi-label (CNN)	0.778	0.561	0.806	0.790	0.923	0.665	0.424	0.690	0.680	0.859
Multi-label (Transformer)	**0.883**	**0.737**	**0.900**	**0.893**	**0.965**	0.676	0.436	0.692	0.645	0.829
Phase-cluster (Transformer)	0.858	0.689	0.877	0.822	0.940	**0.688**	**0.456**	**0.712**	**0.695**	**0.867**

CNN encoder with a Transformer encoder to better capture long-range dependencies in the diffraction pattern, while still performing global presence prediction; (iii) a *latent phase-cluster-based Transformer* model, which introduces a fixed number of latent phase slots and aggregates token-level features into these slots via attention-based clustering, producing slot-wise space-group predictions and modeling latent phase structure in the multi-phase XRD signal.

The first two variants are designed to isolate the effect of encoder choice, testing the architectural superiority of convolutional versus Transformer-based representations under an identical naive multi-label prediction paradigm. In contrast, the latter variant tests the hypothesis that explicitly modeling latent phases independent label combinations— provides a more physically meaningful inductive bias for space-group identification.

2.2 Dataset Construction

In the absence of large-scale experimental multi-phase XRD datasets with reliable space-group annotations, we constructed a synthetic dataset based on crystal structures from the Open Quantum Materials Database (OQMD) [17], enabling systematic evaluation under controlled conditions.

Single-phase diffraction patterns were precomputed for a large collection of crystalline materials. Multi-phase samples were generated by randomly selecting K^* materials ($K^* = 1, \ldots, 5$) according to the probability distribution $[0.20, 0.40, 0.25, 0.10, 0.05]$ and mixing their intensity profiles using Dirichlet-distributed weights to model variable phase fractions.

To mimic experimental variability, we applied a suite of physically motivated augmentations to the clean XRD profiles generated from CIF files. These augmentations included a global zero-shift in diffraction angle, peak broadening arising from finite crystallite size and microstrain, smooth background variations, additive measurement noise, and random masking of the measurement window. The resulting dataset was split into training, validation, and test sets with ratios of 0.95, 0.025, and 0.025, respectively.

2.3 Quantitative Evaluation

We evaluated model performance using the micro-averaged F1-score for multi-label space-group prediction, which is well suited to the highly imbalanced label distribution and variable phase cardinality inherent to multi-phase XRD data. In addition, interpretable top-K set-based metrics are reported, where the K highest-scoring predicted space groups were compared against the ground-truth set to assess coverage and recall.

Table 1 reveals two main observations. First, models employing a Transformer encoder consistently outperformed their CNN-based counterparts across almost all evaluation metrics, underscoring the importance of capturing long-range dependencies in diffraction patterns. Second, when comparing presence-based and phase-clustered formulations, the latent phase slots–based Transformer demonstrated small but consistent improvement of robustness on out-of-library data over the presence-based model, supporting the hypothesis that explicitly separating latent phases provides a more transferable and physically meaningful inductive bias for space-group identification in multi-phase diffraction patterns.

References

[1] Hugo M Rietveld. A profile refinement method for nuclear and magnetic structures. *Applied Crystallography*, 2(2):65–71, 1969.

[2] Robert Alan Young. *The rietveld method*, volume 5. International union of crystallography, 1993.

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

[4] A. Le Bail, H. Duroy, and J.L. Fourquet. Ab-initio structure determination of lisbwo6 by x-ray powder diffraction. *Materials Research Bulletin*, 23(3):447–452, 1988.

[5] 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*, 99(24):245120, 2019.

[6] Felipe Oviedo, Zekun Ren, Shijing Sun, Charles Settens, Zhe Liu, Noor Titan Putri Hartono, Savitha Ramasamy, Brian L DeCost, Siyu IP Tian, Giuseppe Romano, et al. Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks. *npj Computational Materials*, 5(1):60, 2019.

[7] 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*, 10(1):21790, 2020.

[8] Alexander N Zaloga, Vladimir V Stanovov, Oksana E Bezrukova, Petr S Dubinin, and Igor S Yakimov. Crystal symmetry classification from powder x-ray diffraction patterns using a convolutional neural network. *Materials Today Communications*, 25:101662, 2020.

[9] Eric A Riesel, Tsach Mackey, Hamed Nilforoshan, Minkai Xu, Catherine K Badding, Alison B Altman, Jure Leskovec, and Danna E Freedman. Crystal structure determination from powder diffraction patterns with generative machine learning. *Journal of the American Chemical Society*, 146(44):30340–30348, 2024.

[10] Kamal Choudhary. Diffractgpt: Atomic structure determination from x-ray diffraction patterns using a generative pretrained transformer. *The Journal of Physical Chemistry Letters*, 16(8):2110–2119, 2025.

[11] RJ Hill and CJ Howard. Quantitative phase analysis from neutron powder diffraction data using the rietveld method. *Applied Crystallography*, 20(6):467–474, 1987.

[12] William IF David, Kenneth Shankland, Lynne B McCusker, and Ch Baerlocher. *Structure determination from powder diffraction data*, volume 13. OUP Oxford, 2006.

[13] Jin-Woong Lee, Woon Bae Park, Jin Hee Lee, Satendra Pal Singh, and Kee-Sun Sohn. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic xrd powder patterns. *Nature communications*, 11(1):86, 2020.

[14] 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.

[15] Jaimie Greasley and Patrick Hosein. Exploring supervised machine learning for multi-phase identification and quantification from powder x-ray diffraction spectra. *Journal of Materials Science*, 58(12):5334–5348, 2023.

[16] Titouan Simonnet, Sylvain Grangeon, Francis Claret, Nicolas Maubec, Mame Diarra Fall, Rachid Harba, and Bruno Galerne. Phase quantification using deep neural network processing of xrd patterns. *IUCrJ*, 11(5), 2024.

[17] Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Computational Materials*, 1(1):1–15, 2015.

[18] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[19] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[20] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

[21] Pradyumna Reddy, Scott Wisdom, Klaus Greff, John R Hershey, and Thomas Kipf. Audioslots: A slot-centric generative model for audio separation. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE, 2023.

Appendix A. Methods

1.1 Synthetic Multi-Phase XRD Generation

To train and evaluate multi-label space-group prediction under controlled phase combinations, we synthesized multi-phase XRD profiles from a bank of precomputed clean single-phase diffraction patterns. The bank was generated from CIF files in the OQMD database, restricted to structures with energy above hull ≤ 0.1 eV/atom, using pymatgen's `XRDCalculator`. Each clean single-phase profile was stored on a shared fixed 2θ grid with global bounds $[5^\circ, 120^\circ]$, discretized into 2,048 uniform bins.

Phase set sampling. For each synthetic sample, we first drew the number of phases $K^* \in \{1, \ldots, 5\}$ from a categorical distribution with probabilities $[0.20, 0.40, 0.25, 0.10, 0.05]$. We then sampled a set of K^* material IDs $\{m_k\}_{k=1}^{K^*}$ using the mixture sampler, which randomly selected candidate phases from the split-specific ID pool (train/val/test). Given the selected phases, we sampled mixture proportions $\mathbf{w} = (w_1, \ldots, w_{K^*})$ from a Dirichlet distribution with concentration parameter $\alpha = 1.0$, followed by the enforcement of a small minimum phase weight of 0.03 to avoid degenerate mixtures. This procedure preserved variability in phase proportions while preventing numerically negligible components.

Phase-wise structural perturbations. Let $x_k(2\theta)$ denote the clean single-phase profile of phase k retrieved from the bank. Before mixing, we applied two independent phase-wise perturbations to each component to mimic non-structural variability arising from specimen preparation and microstructure.

First, we applied a smooth low-frequency multiplicative envelope intended to emulate preferred orientation effects. This perturbation was applied with probability 0.6 and sampled with a bounded strength parameter from $[-0.18, 0.18]$. The envelope was constructed from a low-order hybrid basis, combining Chebyshev polynomials up to order 3 with a small number of Fourier terms, normalized to unit mean and clipped to the range $[0.85, 1.2]$ to preserve the overall intensity scale.

Second, we applied phase-specific peak broadening to emulate microstructure-driven effects. Broadening is modeled as a position-dependent pseudo-Voigt profile using the Thompson–Cox–Hastings (TCH) approximation, where the Gaussian component is induced by microstrain and the Lorentzian component by finite crystallite size. Specifically, we sampled a phase-wise microstrain $\epsilon \in [0, 0.0025]$ and used a Gaussian FWHM proportional to $\epsilon \tan\theta$, and we sampled a crystallite size L log-uniformly from $10^{[2.0,2.4]}$ nm and used a Scherrer-type Lorentzian contribution with wavelength $\lambda = 1.54184$ Å. The resulting angle-dependent broadening was applied by variable-width convolution on the fixed grid using a LUT-based approximation (log-spaced width levels with per-bin interpolation) for computational efficiency.

Mixture formation and global measurement effects. We then formed the mixture profile by a convex combination,

$$x_{\mathrm{mix}}(2\theta) = \sum_{k=1}^{K^*} w_k \, x_k(2\theta). \tag{A1}$$

Next, we applied a global 2θ zero-shift Δ sampled uniformly from $[-0.06^\circ, 0.06^\circ]$, modeling systematic instrument miscalibration. We further applied an instrument-resolution broadening to the mixture only, modeled as a Gaussian Caglioti-like form with randomized increments $\Delta U \in [0, 0.004]$, $\Delta V = 0$, and $\Delta W \in [0.002, 0.0018]$.

Measurement windowing, background, and noise. To reflect incomplete scan ranges, we sampled a measurement window $[2\theta_{\mathrm{lo}}, 2\theta_{\mathrm{hi}}]$ with $2\theta_{\mathrm{lo}} \sim \mathcal{U}(5^\circ, 20^\circ)$ and $2\theta_{\mathrm{hi}} \sim \mathcal{U}(80^\circ, 120^\circ)$, yielding a binary mask $m(2\theta) \in \{0, 1\}$. We then added measurement-level background and noise to the mixture.

The background consisted of (i) a smooth multiplicative baseline constructed from a softplus-transformed polynomial of order 4 with coefficients sampled from $[-0.12, 0.12]$ and scaled by a random amplitude in $[0.06, 0.18]$, and (ii) an additive floor given by a constant offset in $[0.003, 0.03]$ plus a small slope in $[-0.01, 0.01]$ over the normalized angle coordinate. Finally, we added i.i.d. Gaussian noise with per-sample standard deviation sampled from $[0.0008, 0.0025]$ and clipped intensities to be non-negative.

Normalization and returned targets. After masking and background and noise injection, we normalized the mixture profile by its 99th percentile intensity, yielding the final input vector.

1.2 1D CNN Front-End for Peak Feature Extraction

Given an input XRD intensity profile sampled on a fixed 2θ grid, we employed a lightweight one-dimensional convolutional front-end to extract local peak features and to compress the long input sequence into a shorter latent sequence for subsequent global reasoning modules. Let $\mathbf{x} \in \mathbb{R}^N$ denote the normalized intensity vector (after masking and normalization). We reshaped it as a single-channel signal $\mathbf{x} \in \mathbb{R}^{1\times N}$ and applied a stack of strided convolutional blocks, producing a feature map $\mathbf{H} \in \mathbb{R}^{C_{\mathrm{last}}\times L}$.

Backbone architecture. The CNN backbone was defined by a sequence of channel widths $\{C_\ell\}_{\ell=1}^{L_b}$ and strides $\{s_\ell\}_{\ell=1}^{L_b}$, where each block maps

$$\mathbf{H}_{\ell-1} \in \mathbb{R}^{C_{\ell-1}\times L_{\ell-1}} \;\rightarrow\; \mathbf{H}_\ell \in \mathbb{R}^{C_\ell \times L_\ell}, \qquad L_\ell \approx \left\lceil \frac{L_{\ell-1}}{s_\ell} \right\rceil.$$

In our default setting, we used four stages with `cnn_channels` $= [64, 128, 256, 256]$ and `cnn_strides` $= [2, 2, 2, 2]$, resulting in an overall downsampling factor of $2^4 = 16$ and an output tensor $\mathbf{H} \in \mathbb{R}^{256\times(N/16)}$ (up to boundary effects from padding). This strided design reduced the computational cost of downstream

modules while retaining fine-grained local information around Bragg peaks.

Multi-scale dilated residual block. To robustly capture peak patterns of varying widths and overlaps, we adopted a multi-scale dilated residual block as the default building block [18]. Given an input feature map $\mathbf{H} \in \mathbb{R}^{C_{\text{in}} \times L}$, the block applies multiple one-dimensional convolutions in parallel, each using a different kernel size k_j and dilation d_j:

$$\mathbf{U}_j = \text{Conv1D}_{k_j,d_j,s}(\mathbf{H}), \qquad j = 1, \ldots, J.$$

The branch outputs are concatenated along the channel dimension, batch-normalized, and passed through a GELU nonlinearity. A 1×1 convolution then mixes the concatenated channels and projects back to C_{out} channels:

$$\mathbf{U} = \phi\big(\text{BN}([\mathbf{U}_1; \ldots; \mathbf{U}_J])\big), \qquad \tilde{\mathbf{H}} = \text{Conv1D}_{1\times1}(\mathbf{U}),$$

followed by dropout. A residual connection is added to preserve information and stabilize optimization; when the stride or channel size changes, a 1×1 projection is used on the shortcut path:

$$\mathbf{H}_{\text{out}} = \tilde{\mathbf{H}} + \text{Shortcut}(\mathbf{H}).$$

This design provides a compact approximation of multi-resolution receptive fields: smaller kernels emphasize sharp local peak structures, while larger kernels and larger dilations integrate broader contexts useful for modeling peak shoulders, overlaps, and background variations.

Configuration. In the experiments reported in this paper, the multi-scale branches used kernel sizes $[3, 7, 15]$ and dilations $[1, 2, 4]$, with a dropout rate of 0.1. Padding was set per branch to preserve the "same-ish" output length under dilation, ensuring that the temporal resolution was controlled primarily by the block stride. The CNN output was then forwarded to the subsequent modules (e.g., attention and phase-aware heads), which operated on the reduced-length sequence while benefiting from peak-centric local descriptors extracted by the convolutional front-end.

1.3 Encoder Architectures

After the 1D CNN front-end, we obtained a length-reduced token sequence $\mathbf{T} \in \mathbb{R}^{B\times L\times d}$, where $L \approx N/16$. Since the simulated profiles could be windowed, we propagated a downsampled measurement mask to token resolution, $\mathbf{m} \in [0, 1]^{B\times1\times L}$, computed by staged max-pooling with the same strides as the CNN. We used a boolean valid-token indicator to (i) construct a padding mask for attention-based encoders and (ii) prevent feature leakage into invalid regions for convolutional encoders.

In addition, to preserve the geometric meaning of the token positions after downsampling, we added a lightweight positional conditioning derived from the 2θ bin centers. We computed a 2θ map $\boldsymbol{\theta} \in \mathbb{R}^{B\times1\times N}$, downsampled it by staged average pooling to $\boldsymbol{\theta}_{\text{tok}} \in \mathbb{R}^{B\times1\times L}$, and added a learned projection of two per-token scalars—the normalized angle and the local mask mean—to the token embeddings:

$$\mathbf{T} \leftarrow \mathbf{T} + \text{MLP}\Big(\big[\tilde{\theta}_{\text{tok}},\, \mathbf{m}\big]\Big),$$

where $\tilde{\theta}$ denotes 2θ scaled to $[-1, 1]$ using the global scan bounds.

Transformer encoder (default). Our primary encoder was a stack of 18 Transformer encoder layers, each comprising multi-head self-attention with 12 heads and a position-wise feed-forward network. We adopted RoPE-based positional encoding to inject relative position information along the 2θ axis. Given a padding mask, self-attention was restricted to valid tokens, yielding contextualized token representations. We set the model dimension to $d = 768$ and the feed-forward hidden width to 4,096, and used a dropout rate of 0.1 for both attention and feed-forward sublayers.

Naive token CNN encoder (alternative). To provide a convolutional alternative to global self-attention, we also considered a simple one-dimensional token CNN encoder operating on $\mathbf{T}$ along the token dimension. The encoder consisted of 18 residual blocks; each block applied two same-length 1D convolutions with GELU activation and dropout with rate 0.1. To mitigate mask-induced leakage, after each residual block we enforced the invalid positions to remain zero by multiplying with the valid mask. A final LayerNorm in (B, L, d) space was applied to obtain the encoded sequence $\mathbf{Z} \in \mathbb{R}^{B\times L\times d}$. We used kernel size 7 and a cyclic dilation schedule $1, 2, 4, 8, \ldots$ (period 4), matching the depth of the Transformer preset in a parameter-efficient manner while retaining a purely convolutional inductive bias.

Pooling for mixture-level prediction. Both encoder variants produce $\mathbf{Z} \in \mathbb{R}^{B\times L\times d}$, which is aggregated into a single mixture representation via masked mean pooling:

$$\mathbf{g} = \frac{\sum_{\ell=1}^{L} \mathbf{m}_\ell\, \mathbf{Z}_\ell}{\sum_{\ell=1}^{L} \mathbf{m}_\ell + \varepsilon} \in \mathbb{R}^{B\times d}.$$

The pooled vector $\mathbf{g}$ is then fed to the multi-label heads for space-group presence prediction.

1.4 Phase Clustering

To explicitly represent latent phases in multi-phase XRD profiles, we introduce a *phase clustering* module that aggregates token-level representations into a fixed set of latent phase slots. Given the encoded token sequence $\mathbf{Z} \in \mathbb{R}^{B\times L\times d}$ produced by the encoder, the module outputs slot-wise representations $\mathbf{S} \in \mathbb{R}^{B\times K\times d}$, where K denotes a predefined upper bound on the number of phases present in a mixture.

Latent phase slots. We maintain K learnable slot embeddings $\{\mathbf{s}_k\}_{k=1}^{K} \subset \mathbb{R}^d$, initialized as trainable parameters. Each slot is intended to capture a coherent representation of a single latent crystalline

phase. Given token features $\mathbf{Z}$ and a valid-token mask, the module aggregates token-level information into slot-wise features $\mathbf{S} = \{\mathbf{S}_k\}_{k=1}^{K}$ using masked cross-attention, as described below.

Token-to-slot aggregation. Aggregation from tokens to slots is implemented via multi-head cross-attention, with latent slots acting as queries and token features serving as keys and values. Invalid tokens are excluded through masking. Intuitively, each slot learns to selectively attend to a subset of tokens corresponding to a consistent diffraction pattern, thereby isolating a single latent phase.

The aggregation can be iterated over multiple update layers. Denoting the slot representations at layer t by $\mathbf{S}^{(t)}$, each update consists of a cross-attention block followed by a slot-wise feed-forward network with residual connections and layer normalization:

$$\mathbf{S}^{(t+1)} = \mathrm{FFN}\Big(\mathbf{S}^{(t)} + \mathrm{Attn}\big(\mathbf{S}^{(t)}, \mathbf{Z}, \mathbf{Z}\big)\Big).$$

This formulation is permutation-invariant with respect to the token order and enables each slot to form a phase-specific summary by attending to relevant regions of the 2θ axis.

Slot-wise classification and gating. Each slot representation $\mathbf{S}_k$ is mapped to space-group logits using a linear classification head, producing $\boldsymbol{\pi}_k \in \mathbb{R}^{C+1}$ over C candidate space groups plus an additional *NONE* class. Applying a softmax yields probabilities $\pi_{k,c}$ for each space group c and $\pi_{k,\mathrm{none}}$ for the NONE class. The NONE probability is interpreted as the likelihood that the slot does not correspond to a physical phase, and we define a continuous slot-activation gate

$$g_k = 1 - \pi_{k,\mathrm{none}} \in [0, 1].$$

This gating mechanism allows the model to flexibly adapt to mixtures containing fewer than K phases without requiring hard slot assignments.

Permutation-invariant assignment loss. Supervision is provided at the level of ground-truth phases, each annotated with a space-group label and an optional weight. Because the ordering of latent slots is arbitrary, we employ a permutation-invariant assignment loss. For a sample with M ground-truth phases, we enumerate all injective mappings from the M phases to the K slots and compute the total cross-entropy cost for each mapping, including a NONE-class penalty for unused slots. The minimum-cost mapping is selected, and its normalized cost is used as the training loss.

This exhaustive assignment is computationally feasible because K is small in practice (typically $K \leq 5$), and it ensures that training is invariant to permutations of the latent phase slots while providing explicit phase-level supervision.

1.5 Training Objective

For models without phase clustering, we train a multi-label classifier over the $C = 230$ space-group classes using a presence-only objective. Let $\mathbf{y} \in \{0, 1\}^C$ denote the ground-truth indicator vector of space groups present in a synthesized multi-phase XRD profile, and let $\boldsymbol{\ell} \in \mathbb{R}^C$ be the output logits of the global presence head. The per-sample loss is defined as the mean binary cross-entropy over all classes:

$$\mathcal{L}_{\mathrm{pres}} = -\frac{1}{C}\sum_{c=1}^{C}\Big[y_c \log \sigma(\ell_c) + (1 - y_c)\log\big(1 - \sigma(\ell_c)\big)\Big],$$

where $\sigma(\cdot)$ denotes the sigmoid function.

For models equipped with phase clustering, we instead optimize a permutation-invariant assignment objective that encourages each latent slot to represent either a single physical phase or a dedicated NONE category. Let $\boldsymbol{\pi}_k \in \Delta^{C+1}$ denote the predicted categorical distribution over the C space groups plus NONE for slot k, and let $\mathcal{Y} = \{(c_m, w_m)\}_{m=1}^{M}$ denote the set of M ground-truth phases with their associated space-group labels and mixture weights. Because the correspondence between latent slots and ground-truth phases is unknown, we compute the minimum total cross-entropy cost over all injective assignments from phases to slots, assigning all remaining (unused) slots to the NONE class.

Formally, let $\mathcal{I}_{K,M}$ denote the set of injections $g : \{1, \ldots, M\} \to \{1, \ldots, K\}$. The assignment loss is defined as

$$\mathcal{L}_{\mathrm{assign}} = \min_{g \in \mathcal{I}_{K,M}}\left(\sum_{m=1}^{M}\mathrm{CE}\big(\boldsymbol{\pi}_{g(m)}, c_m\big) + \sum_{k \in \bar{g}}\mathrm{CE}(\boldsymbol{\pi}_k, \mathrm{NONE})\right),$$

where $\mathrm{CE}(\boldsymbol{\pi}, c) = -\log \boldsymbol{\pi}[c]$, and $\bar{g}$ denotes the complement of the assigned slots $g(\{1, \ldots, M\})$. In practice, the minimum is evaluated exactly by enumerating all admissible assignments, which is computationally feasible due to the small number of slots ($K = 5$). For the phase-clustering configuration considered in this work, we optimize $\mathcal{L}_{\mathrm{assign}}$ exclusively, without an additional global presence loss.

The overall training objective is therefore given by

$$\mathcal{L} = \lambda_{\mathrm{pres}}\mathcal{L}_{\mathrm{pres}} + \lambda_{\mathrm{assign}}\mathcal{L}_{\mathrm{assign}},$$

with $(\lambda_{\mathrm{pres}}, \lambda_{\mathrm{assign}}) = (1, 0)$ for the presence-only CNN and Transformer baselines, and $(0, 1)$ for the phase-clustering model.

1.6 Training Protocol

All models are trained on synthesized multi-phase XRD profiles using the AdamW optimizer with learning rate 1×10^{-4} and weight decay 10^{-2}, and a batch size of 256. We employ a cosine learning-rate schedule with linear warmup over the first 4,000 optimization steps, followed by cosine decay for a total of 1,000,000 steps. Gradients are clipped to a global ℓ_2 norm of 1.0. Unless otherwise specified, training is performed without automatic mixed precision.

During training, models are evaluated every 2,000 steps, and the checkpoint with the highest validation micro-averaged F1-score is selected for reporting. Additional checkpoints are saved at fixed intervals of 50,000 steps to ensure reproducibility.

1.7 Evaluation Protocol

We evaluate all models on two data splits: an *in-library* split drawn from the training library and a held-out *out-of-library* split, following the dataset construction described earlier. To ensure fair comparisons across training runs, evaluation on the in-library split is performed using a fixed-size subset whose cardinality matches that of the out-of-library split. This subset is sampled once using a fixed random seed and reused across all experiments.

Presence prediction. For presence-only models (CNN and Transformer), per-class presence probabilities are obtained as $p_c = \sigma(\ell_c)$. At evaluation time, we sweep a global threshold $\tau \in [0, 1]$ on a uniform grid of 101 values and convert scores to a binary prediction set $\hat{\mathbf{y}}(\tau) = \mathbb{1}[p_c \geq \tau]$. Reported metrics correspond to the threshold that maximizes the micro-averaged F1-score on the evaluated split. The same predictions are used to compute the set-based top-K metrics described below.

Slot-union prediction for phase clustering. For phase-clustering models, inference yields slot-wise categorical distributions over C+1 classes. We convert these predictions into a global multi-label output using a *slot-union* rule (`pred_mode=slot_union`). Specifically, for each slot k, we identify the most probable space-group class and include it in the predicted set if the slot activation gate exceeds a specified threshold. We sweep the gate threshold over the range $[0.10, 0.90]$ with step size 0.02 and select the value that maximizes the micro-averaged F1-score. Unless otherwise noted, the space-group confidence threshold is fixed to zero, such that slot selection is controlled solely by the gate. We report both the performance at a fixed gate threshold of 0.5 and the best-threshold performance.

1.8 Quantitative Evaluation

All models are evaluated under a unified multi-label prediction setting, where each XRD sample is associated with a variable-size set of ground-truth space groups. Let $\mathbf{y} \in \{0, 1\}^C$ denote the binary indicator vector over $C = 230$ space groups, and let $\hat{\mathbf{s}} \in \mathbb{R}^C$ denote the model output scores, obtained either from a global presence head (presence-only models) or via slot-based aggregation (phase-clustering models).

Micro-averaged metrics. To summarize performance across all samples and space-group classes, we report micro-averaged precision, recall, and F1-score. Given a binarization threshold τ, predictions are defined as $\hat{y}_c(\tau) = \mathbb{1}[\hat{s}_c \geq \tau]$. Micro-averaged statistics are computed by aggregating true positives, false positives, and false negatives across all samples and classes prior to forming the corresponding ratios. This metric appropriately balances frequent and rare space groups and is therefore well suited to the highly imbalanced label distribution and variable label cardinality characteristic of multi-phase XRD data.

For presence-only models, the threshold τ is selected independently for each evaluation split by sweeping $\tau \in [0, 1]$ on a grid of 101 values and choosing the value that maximizes the micro-averaged F1-score, thereby avoiding bias from threshold transfer.

Set-based top-K metrics. In addition to thresholded micro-F1, we report set-based top-K metrics that provide a more interpretable assessment of partial recovery of the ground-truth phase set. For each sample, let K^* denote the number of true space groups. We rank the predicted scores $\hat{\mathbf{s}}$ in descending order and select the top $K = K^*$ or $K = 5K^*$ predicted classes. We then compute: (i) *coverage*, defined as the fraction of samples for which all ground-truth space groups appear within the top-K set; (ii) *recall@K*, defined as the fraction of ground-truth labels recovered within the top-K set; and (iii) *intersection-over-union* (IoU) between the predicted and ground-truth sets. Metrics are averaged over samples and reported as `cover@`K^*, `recall@`K^*, and their corresponding $5K^*$ variants.

Evaluation splits. We report results under two evaluation regimes designed to probe generalization under distribution shift. The *in-library* setting evaluates performance on a held-out subset of the training library, where all constituent phases have been observed during training. The *out-of-library* setting evaluates performance on mixtures composed entirely of phases absent from the training library.

Together, these metrics and evaluation protocols provide a comprehensive assessment of both in-distribution accuracy and robustness to systematic extrapolation toward previously unseen phases.

Appendix B. Discussion

This study systematically investigated multi-label space-group prediction from multi-phase powder XRD profiles under a chemistry-agnostic setting, in which constituent phases may be absent from the training library. To the best of our knowledge, the problem of directly inferring multiple crystallographic space groups from a single mixed diffraction profile has not been systematically formulated or studied in prior work, which has predominantly focused on single-phase settings or phase identification within predefined libraries. The results demonstrate that both the choice of encoder architecture and the explicit modeling of latent phases play critical and complementary roles in achieving robust generalization in this previously unexplored regime.

Impact of encoder architecture. Across all evaluation regimes, Transformer-based models consistently outperformed their CNN-based counterparts, confirming the importance of modeling long-range dependencies in diffraction patterns. Unlike local convolutional filters, self-attention enables the encoder to integrate information across widely separated 2θ regions, which is essential for capturing corre-

lated peak patterns, extinction rules, and symmetry-dependent intensity distributions. These observations are consistent with prior findings in single-phase XRD analysis and extend them to the substantially more challenging multi-phase setting.

Benefits of explicit latent phase modeling. While presence-based models achieved strong in-library performance, their generalization degraded markedly on out-of-library mixtures. This behavior is consistent with the implicit assumption underlying naive multi-label classification, namely that a mixed diffraction signal can be directly mapped to a set of labels without explicitly accounting for its compositional origin. In contrast, the proposed phase-clustering model introduces a structured intermediate representation that more closely reflects the physical generative process of diffraction as a superposition of phase-specific signals. By aggregating token features into latent phase slots and supervising them via permutation-invariant assignment, the model is encouraged to isolate phase-consistent representations that remain stable under changes in phase composition. This inductive bias leads to improved robustness on out-of-library data, particularly in set-based coverage and recall metrics that directly measure phase recovery.

Relation to classical crystallographic workflows. Although the proposed method does not perform explicit pattern decomposition or phase-fraction refinement, it introduces a task-oriented abstraction that is conceptually aligned with classical crystallographic reasoning. The latent phase slots act as symmetry-aware carriers of phase-specific information, while avoiding the ill-posed objective of reconstructing individual diffraction patterns from a mixed signal. In this sense, the model bridges traditional symmetry analysis and modern data-driven learning, offering a complementary approach rather than a replacement for established refinement techniques.

Limitations and future directions. Several limitations warrant discussion. First, the evaluation relied on synthetically generated multi-phase XRD data, which, while physically informed, cannot fully capture all experimental artifacts, such as preferred orientation under extreme texture, anisotropic strain, or instrument-specific aberrations associated with individual diffractometers. Second, the number of latent phase slots was fixed *a priori*, which may limit scalability to mixtures containing a larger or unknown number of phases.

In addition, although the proposed framework improves robustness relative to naive multi-label baselines, the overall prediction accuracy remains below the level required for fully reliable automated space-group assignment in practical experimental workflows. This observation indicates that further advances in model architecture, training objectives, and data efficiency are necessary. Promising directions include more expressive encoders, improved phase–slot assignment strategies, curriculum or self-supervised pretraining schemes, and the incorporation of uncertainty-aware or physically constrained learning objectives.

Future work may also explore adaptive slot allocation, hierarchical phase representations, or hybrid models that combine latent phase modeling with partial pattern reconstruction. Finally, extending the framework to incorporate compositional constraints or joint prediction of space group and lattice constants represents a promising direction toward fully automated multi-phase crystallographic analysis.

Overall, this work establishes latent phase clustering as an effective and physically motivated inductive bias for space-group prediction from multi-phase XRD profiles, particularly in regimes characterized by compositional complexity and incomplete reference libraries.

Appendix C. Relation to Multi-Instance and Slot-Based Representation Learning

Related ideas of decomposing a complex input into a small set of latent components have been explored in the machine learning literature, most notably in multi-instance learning and slot-based representation learning frameworks [19, 20, 21].

Multi-instance learning models typically assume that each input consists of a collection of instances, where the global label is determined by an unknown aggregation of instance-level contributions [19]. Slot-based models, on the other hand, aim to decompose perceptual signals such as images or audio into a fixed number of latent slots, each corresponding to an object or source, often trained using reconstruction-based or generative objectives [20, 21].

Despite this conceptual similarity, these approaches are not directly applicable to powder X-ray diffraction data. First, powder XRD profiles do not admit a natural instance decomposition analogous to pixels, patches, or time–frequency bins, as diffraction peaks arise from a global, physically coupled superposition of scattering contributions. Second, existing slot-based and multi-instance frameworks are primarily designed for perceptual domains and do not incorporate crystallographic constraints such as space-group symmetry, systematic absences, or the discrete nature of symmetry labels.

In contrast, the proposed framework introduces latent phase components that are explicitly supervised by crystallographic space-group labels and are learned as task-oriented intermediate representations for symmetry inference. This design aligns the latent decomposition with the physical generative process of multi-phase diffraction, while avoiding explicit reconstruction of individual phase diffraction patterns, which is inherently ambiguous in powder XRD.

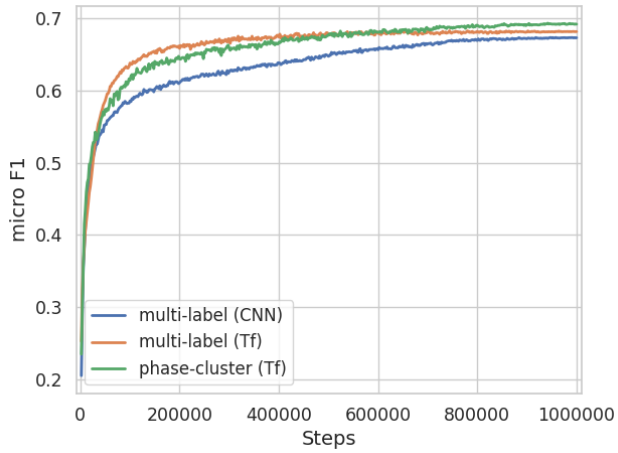

Fig. A1: Evolution of validation micro-F1 during training for the three model variants.

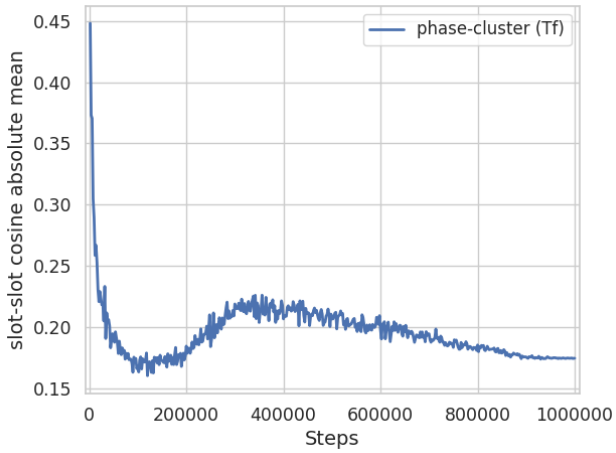

Fig. A2: Training dynamics of slot diversity in the phase-clustered Transformer, measured by the mean absolute cosine similarity between different latent phase slots.

Appendix D. Supplemental Results

4.1 Validation Micro-F1 Trajectories During Training

Fig. A1 shows the evolution of micro-averaged F1-score on the validation set as a function of training steps for the three model variants. All models exhibit a rapid increase in performance during the early training phase, followed by a more gradual improvement as training progresses.

Comparing encoder architectures, Transformer-based models consistently achieve higher validation micro-F1 than the CNN-based baseline throughout training, indicating faster convergence and a higher asymptotic performance. The phase-clustered Transformer initially lags slightly behind the presence-based Transformer in the early stages, but gradually closes the gap and converges to a better validation performance at later training steps.

This behavior suggests that explicit latent phase modeling introduces additional optimization complexity, which may slow early learning, but does not hinder final performance.

4.2 Slot Diversity in the Phase-Clustered Model

Fig. A2 reports the mean absolute cosine similarity between different latent phase slots in the phase-clustered Transformer during training. This quantity serves as a proxy for slot diversity, where lower values indicate more distinct slot representations.

At the beginning of training, the average cosine similarity is high, indicating that slot embeddings are initially poorly differentiated. As training proceeds, the similarity rapidly decreases, suggesting that the slots quickly specialize and capture distinct aspects of the input representation. After this initial separation phase, the similarity stabilizes and exhibits a slow downward trend, with mild non-monotonic fluctuations at intermediate training stages.

The absence of a collapse toward high similarity indicates that the slots do not degenerate into redundant representations. Instead, the observed dynamics are consistent with a stable equilibrium in which slots remain distinct while still sharing common structure induced by the downstream supervision. These results support the interpretation that the phase-clustering mechanism learns a non-trivial decomposition of the token representation space rather than acting as a redundant partition.

4.3 Stratified Analysis by Phase Count and Mixing Ratio

We further analyze performance as a function of the number of phases K and the minimum phase mixing ratio. Fig. A3 reports stratified performance as a function of the number of phases K and whether the mixture contains a low-fraction phase (minimum phase fraction $\leq 10\%$). Three observations emerge. (i) As K increases, performance consistently degrades across all methods, reflecting the increased ambiguity due to peak overlap and label cardinality. Notably, in the single-phase regime ($K = 1$), the CNN-based presence model attains the highest micro-F1 on the test split, whereas in multi-phase regimes ($K \geq 2$) Transformer-based encoders tend to achieve higher micro-F1 than CNN-based models, suggesting better generalization under compositional complexity. (ii) Mixtures containing very low-fraction phases are substantially more challenging: all methods show lower micro-F1 and reduced top-K coverage/recall when the minimum phase fraction is $\leq 10\%$, indicating that recovering space groups of trace phases remains difficult. (iii) The relative ordering between the presence-based and phase-clustered variants is less stable in high-K strata, partly because the number of samples decreases rapidly for large K, increasing variance. Nevertheless, across a wide range of multi-phase conditions, phase clustering is often competitive and can provide modest gains, supporting its potential as an inductive bias for multi-phase generalization.

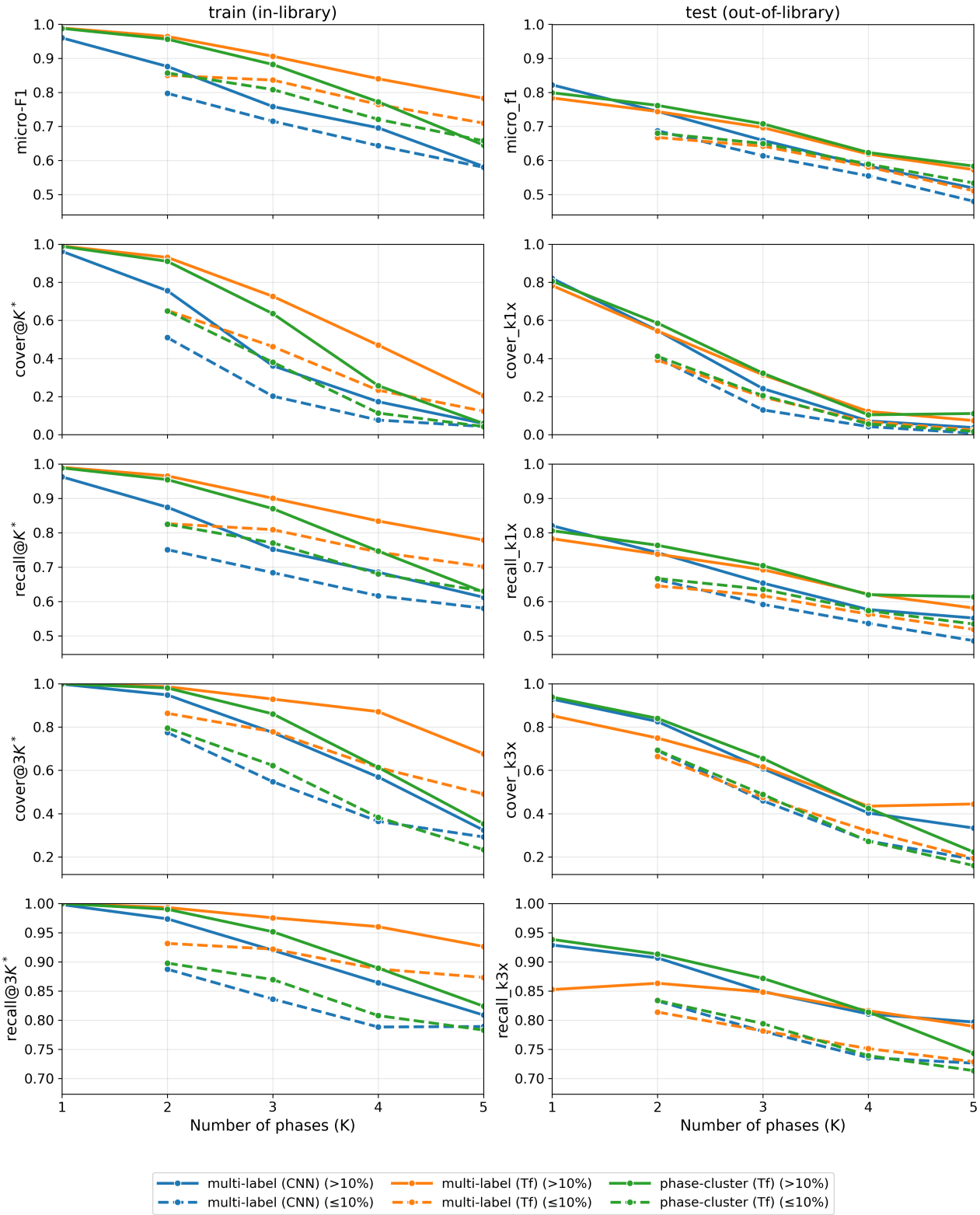

Fig. A3: Stratified performance as a function of the number of phases K and the minimum phase fraction. Results are shown for train (in-library, left) and test (out-of-library, right) splits, further stratified by whether the minimum phase fraction exceeds 10% (solid lines) or not (dashed lines). Metrics include micro-F1, cover@K^*, recall@K^*, cover@$3K^*$, and recall@$3K^*$.