PHYSICS-ROUTED MOE WITH CRYSTALLOGRAPHIC PRIORS FOR LONG-TAILED SPACE-GROUP RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Determining a crystal's space group from a powder X-ray diffraction (PXRD) pattern is a 230-way fine-grained classification problem characterized by severe class imbalance and substantial similarity among diffraction patterns. Many rare space groups have fewer than 20 training examples and are easily confused with frequent groups in the same crystal system. We present Phys-MoE, a Mixture-of-Experts classifier that incorporates crystallographic priors into routing and representation learning. Phys-MoE consists of three components: a hierarchical router that masks space groups inconsistent with the predicted crystal system, a rule-biased gating mechanism that encourages expert specialization across symmetry blocks, and a rule-semantic contrastive objective based on the Jaccard similarity between symmetry-operation sets. Experiments on CCDC, CoREMOF, and an inorganic distribution-shift benchmark demonstrate the effectiveness and generalizability of Phys-MoE. On CCDC, Phys-MoE achieves 81.70% overall Top-1 accuracy and 26.19% accuracy for space groups with fewer than 20 training examples, establishing a new state of the art for long-tailed space-group recognition from PXRD patterns. Ablation studies demonstrate that the three crystallographic components work jointly to improve recognition, while error and oracle analyses identify crystal-system prediction as a remaining bottleneck for rare classes. These results show that incorporating crystallographic structure into expert routing and representation learning provides an effective approach to accurate and reliable space-group recognition under long-tailed data distributions.

1 INTRODUCTION

Space-group identification from powder X-ray diffraction (PXRD) is a step in crystal structure analysis (Khan et al., 2020; Giacovazzo et al., 2002). It requires distinguishing 230 space groups (Nespolo, 2017) whose diffraction profiles can share similar features. Training examples are unevenly distributed across these groups, leaving rare classes with little supervision.

Learning-based PXRD analyzers predict crystal symmetry directly from diffraction profiles (Vecsei et al., 2019). Salgado et al. (2023) used convolutional networks for crystal-system and space-group classification. Yu et al. (2024) proposed XRDMamba, combining selective state-space modeling with a crystal-plane-informed representation for space-group identification. Yu et al. (2026) introduced XRDecoupler to address confusion among groups that share crystallographic properties. It uses crystal systems, Bravais lattice types, and point groups as superclass guidance, together with hierarchical pattern learning and multi-objective optimization. Generic long-tailed methods use class-frequency information and multiple experts to mitigate imbalance (Ren et al., 2020; Menon et al., 2020; Cao et al., 2019; Chen et al., 2025). Rare-space-group recognition remains difficult for these methods under the extreme imbalance in our PXRD benchmark.

In the CCDC-derived benchmark (Groom et al., 2016), the head-to-tail class-count ratio exceeds 10^4:1, and some space groups have fewer than 20 training examples. These groups are often confused with frequent groups that share crystallographic properties. Class rebalancing adjusts the contribution of each class to training; crystal-system constraints narrow the candidate labels. Distinguishing the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

remaining groups requires learning from their shared properties while preserving the features that separate them.

We introduce Phys-MoE, a Mixture-of-Experts classifier that uses crystallographic relations in label selection, expert weighting, and representation learning. It has a shared PXRD encoder, a crystal-system predictor, and multiple expert classification heads. Hierarchical Routing (HR) restricts candidate space groups using the ground-truth crystal system during training and the predicted system at inference. Rule-Guided Experts (RGE) uses the predicted system distribution to bias expert weights. Rule-Semantic Contrastive Learning (RSCL) uses a fixed symmetry-operation similarity to select same-class and rule-neighbor positives and reduce contrastive repulsion between related classes.

Phys-MoE achieves the highest overall Top-1 accuracy among the evaluated PXRD-only methods on CCDC, CoREMOF, and Inorganic. On CCDC, it reaches 81.70% overall Top-1 accuracy and 26.19% class-balanced Top-1 accuracy for space groups with fewer than 20 training examples. Incremental and counterfactual ablations measure the gains from the crystallographic components after the shared-backbone and loss adaptation, and show how the effect of RSCL changes with the other components. We also examine routing errors, operation-set similarity of incorrect predictions, and sensitivity to controlled PXRD perturbations.

Our main contributions are:

❶ We introduce Phys-MoE, which integrates crystallographic label constraints, system-conditioned expert weighting, and rule-semantic representation learning for long-tailed space-group recognition.

❷ We demonstrate improved overall recognition on three PXRD benchmarks and improved extreme-tail recognition on CCDC, with paired-bootstrap confidence intervals for model comparisons.

❸ We assess the crystallographic components on an architecture-adapted baseline and analyze component interactions, routing errors, and sensitivity to controlled PXRD perturbations.

2 Related Work

► **Learned PXRD Analyzers.** Alongside powder indexing methods (Werner et al., 1985), learned PXRD analyzers include CNN classifiers (Park et al., 2017; Zaloga et al., 2020), sequence models (Yu et al., 2024), and models for inorganic-pattern classification (Salgado et al., 2023) and structure-type assignment (Chen et al., 2024). XRDecoupler (Yu et al., 2026) uses crystallographic superclass supervision to address sub-property confusion (SPC). XQueryer (Cao et al., 2025b) combines PXRD with elemental information for crystal-structure identification, while SimXRD-4M (Cao et al., 2025a) provides a large simulated-data benchmark for crystal-symmetry classification. Appendix D describes their inputs, prediction targets, and evaluation settings in relation to our comparisons.

► **Long-tailed Recognition with Contrastive Objectives.** Generic long-tailed approaches include class-balanced losses (Cui et al., 2019) and decoupled representation and classifier training (Kang et al., 2019). Balanced Softmax (Ren et al., 2020) and Logit Adjustment (Menon et al., 2020) incorporate class-frequency information into classification objectives or prediction scores, while LDAM-DRW (Cao et al., 2019) combines class-dependent margins with deferred reweighting. Beyond classifier correction, balanced contrastive learning (Zhu et al., 2022) addresses imbalance in representation learning by balancing the contributions of different classes.

► **Mixture of Experts under Data Scarcity.** Mixture-of-Experts models distribute computation or prediction across specialized branches (Shazeer et al., 2017; Fedus et al., 2022). For long-tailed recognition, ICL (Chen et al., 2025) combines independent expert optimization with competition and collaborative learning. Physics-informed MoE has also been used to enforce PDE constraints on decomposed spatial domains (Chalapathi et al., 2024). Phys-MoE studies system-conditioned expert weighting for discrete space-group labels with highly unequal class frequencies.

3 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a training dataset, where $x_i \in \mathbb{R}^L$ is a PXRD pattern sampled at L positions and $y_i \in \{1, \ldots, 230\}$ is its space-group label. The class distribution is long-tailed, with

limited training examples for many space groups. We use i, j, r to index samples, g, g' to index space groups, s to index crystal systems, and k to index experts.

► **Crystal-system Hierarchy.** Each space group belongs to exactly one of seven crystal systems, defining a fixed mapping

$$T : \{1, \ldots, 230\} \to \{1, \ldots, 7\}.$$

The crystal-system label of sample i is therefore $s_i = T(y_i)$. This mapping provides auxiliary supervision for crystal-system prediction and defines the candidate-label constraint used by hierarchical routing.

► **Symmetry-operation Similarity.** Let $\mathcal{O}_g$ denote the finite symmetry-operation set used to represent space group g. We define a rule-similarity matrix $S \in [0, 1]^{230\times230}$ using Jaccard similarity:

$$S(g, g') = \frac{|\mathcal{O}_g \cap \mathcal{O}_{g'}|}{|\mathcal{O}_g \cup \mathcal{O}_{g'}|}. \tag{1}$$

The operation sets are constructed from the *International Tables for Crystallography* (Brock et al., 2016), with their representation detailed in Appendix C. The matrix S is computed before training and held fixed. For samples i and j, their class-level rule similarity is $S(y_i, y_j)$.

► **Goal.** Given a PXRD pattern x, the task is to estimate a space-group distribution $p_\theta(g \mid x)$ and predict

$$\hat{g}(x) = \arg \max_{g\in\{1,\ldots,230\}} p_\theta(g \mid x),$$

where θ denotes the trainable model parameters. Phys-MoE uses the fixed hierarchy T and rule-similarity matrix S to guide classification and representation learning under the long-tailed training distribution.

4 METHOD: PHYS-MOE

4.1 FRAMEWORK

Phys-MoE consists of a shared PXRD encoder and projection, a seven-way crystal-system head, $K = 8$ expert classification heads, and a rule-semantic contrastive objective (Figure 1). The encoder and projection jointly produce $z_i = \phi_\theta(x_i) \in \mathbb{R}^d$, where d is the representation dimension. The system head outputs

$$q_i = \mathrm{Softmax}(W_{\mathrm{sys}}(z_i)), \tag{2}$$

where $W_{\mathrm{sys}} : \mathbb{R}^d \to \mathbb{R}^7$ is a learned affine map and $q_{i,s}$ is the predicted probability of crystal system s.

Rule-Guided Experts (RGE) produces aggregated space-group logits $\ell_i \in \mathbb{R}^{230}$, which are subsequently constrained by Hierarchical Routing (HR). Rule-Semantic Contrastive Learning (RSCL) uses S to guide representation learning during training only. We write feature, probability, and logit vectors as row vectors. Sample subscripts are omitted in Figure 1 for clarity. Training losses are evaluated on a mini-batch with sample-index set $\mathcal{I}$ and size $m = |\mathcal{I}|$.

4.2 HIERARCHICAL ROUTING

HR uses the fixed mapping T to restrict candidate space groups to a selected crystal system. For system s, define the additive logit mask $M_s \in \{0, -\infty\}^{230}$:

$$M_s(g) = \begin{cases} 0, & T(g) = s, \\ -\infty, & T(g) \neq s. \end{cases} \tag{3}$$

During training, the mask is constructed from the ground-truth system $s_i = T(y_i)$. The crystal-system and masked space-group losses are

$$\begin{aligned} \mathcal{L}_{\mathrm{sys}} &= -\frac{1}{m} \sum_{i\in\mathcal{I}} \log q_{i,s_i}, \\ \mathcal{L}_{\mathrm{sg}} &= \frac{1}{m} \sum_{i\in\mathcal{I}} \mathrm{CE}(\ell_i + M_{s_i}, y_i), \end{aligned} \tag{4}$$

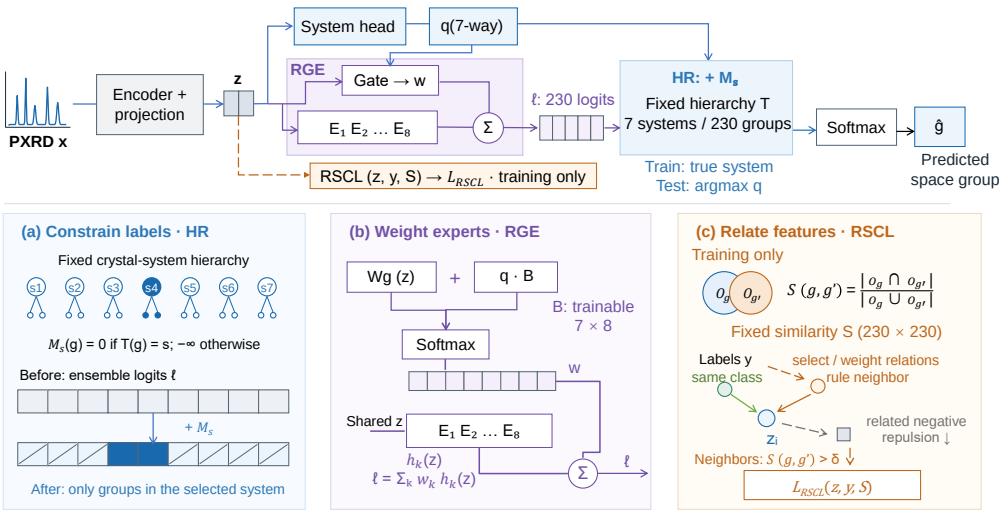

Figure 1: **Overview of Phys-MoE.** A shared encoder and projection produce a representation z for crystal-system prediction and eight expert heads. **(a) Hierarchical Routing (HR)** applies the additive mask M_s to the aggregated logits ℓ, retaining only space groups in the selected crystal system. Training uses the true system; inference uses $\arg\max_s q_s$. **(b) Rule-Guided Experts (RGE)** combines feature-based gating with the predicted system distribution q through a trainable bias matrix B. The resulting weights w aggregate the expert logits before hierarchical masking. **(c) Rule-Semantic Contrastive Learning (RSCL)** uses the fixed symmetry-operation Jaccard matrix S to select same-class and rule-neighbor positives and to reweight non-self pairs in the contrastive denominator. RSCL is used only during training.

where $\mathrm{CE}(v, y) = -\log[\mathrm{Softmax}(v)]_y$ denotes cross-entropy applied to a logit vector v and label y.

At inference, the selected system is $\hat{s}_i = \arg\max_s q_{i,s}$, and the space-group distribution is

$$p_\theta(\cdot \mid x_i) = \mathrm{Softmax}(\ell_i + M_{\hat{s}_i}). \tag{5}$$

The final prediction is $\hat{g}_i = \arg\max_g p_\theta(g \mid x_i)$. HR retains the space groups within the selected system for classification. An incorrect system prediction excludes the true space group; Section 5.6 measures the resulting errors.

4.3 Rule-Guided Experts

RGE introduces system-conditioned weighting at the expert decision heads. All K experts share the encoder and projection. Expert E_k produces the logit vector

$$h_k(z_i) = z_i W_k + b_k, \tag{6}$$

where $W_k \in \mathbb{R}^{d\times 230}$ and $b_k \in \mathbb{R}^{230}$ are its trainable parameters.

The gating logits combine a feature-based term with a crystal-system-conditioned bias:

$$a_i = W_g(z_i) + q_i B, \qquad w_i = \mathrm{Softmax}(a_i), \tag{7}$$

where $W_g : \mathbb{R}^d \to \mathbb{R}^K$ is a learned affine gating map, $B \in \mathbb{R}^{7\times K}$ is the trainable rule-bias matrix, and $w_{i,k}$ is the weight assigned to expert k. The matrix B is initialized with the ± 0.5 bias pattern associated with the CCDC-calibrated expert allocation and optimized jointly with the model (Appendix C). Unlike B, the rule-similarity matrix S remains fixed.

The expert logits are aggregated before applying HR:

$$\ell_i = \sum_{k=1}^{K} w_{i,k} h_k(z_i). \tag{8}$$

The gate uses the full predicted distribution q_i during both training and inference, whereas HR uses a single selected system. Expert allocation and bias initialization are chosen using the CCDC class distribution.

4.4 Rule-Semantic Contrastive Learning

While HR and RGE guide classification decisions, RSCL uses relationships between space groups to guide the shared representation. For anchor sample i, the positive set includes other samples from the same class and samples whose labels are rule neighbors:

$$\mathcal{P}(i) = \{j \in \mathcal{I} \setminus \{i\} \mid y_j = y_i \text{ or } S(y_i, y_j) > \delta\}, \tag{9}$$

where $\delta \in [0, 1]$ is the rule-neighbor threshold.

For each non-self pair, define the rule-dependent denominator weight

$$\omega_{ij} = 1 - S(y_i, y_j), \qquad j \neq i. \tag{10}$$

Let $\mathcal{I}_+ = \{i \in \mathcal{I} : |\mathcal{P}(i)| > 0\}$ denote anchors with at least one positive. The rule-weighted contrastive objective is

$$\mathcal{L}_{\mathrm{RSCL}} = -\sum_{i \in \mathcal{I}_+} \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{\exp(\langle z_i, z_j \rangle / \tau)}{\epsilon + \sum_{r \in \mathcal{I} \setminus \{i\}} \omega_{ir} \exp(\langle z_i, z_r \rangle / \tau)}, \tag{11}$$

where $\langle \cdot, \cdot \rangle$ is the feature inner product, $\tau > 0$ is the contrastive temperature, and $\epsilon > 0$ is a numerical stabilizer. Anchors without positives do not contribute; if $\mathcal{I}_+$ is empty, the contrastive loss is zero.

The weights ω_{ir} apply to all non-self pairs in the denominator, including pairs selected as positives. RSCL therefore combines rule-informed positive selection with rule-dependent denominator weighting. It is used only during training and introduces no additional inference branch. Figure 1 abbreviates this batch-level objective as $\mathcal{L}_{\mathrm{RSCL}}(z, y, S)$.

4.5 Composite Objective

The expert heads are trained with an asymmetric Diverse Optimization Learning (DoL) objective. Let $\mathcal{K}_{\mathrm{head}}$, $\mathcal{K}_{\mathrm{tail}}$, and $\mathcal{K}_{\mathrm{trans}}$ denote the mutually disjoint index sets of head, tail, and transition experts, respectively; their union is $\{1, \ldots, K\}$. Their allocation is specified in Appendix C.

For sample i and expert k, define the probability assigned to the true space group before hierarchical masking as

$$\rho_{ik} = [\mathrm{Softmax}(h_k(z_i))]_{y_i}.$$

Head experts use cross-entropy, while tail and transition experts use focal loss:

$$\mathcal{L}_{\mathrm{DoL}} = \frac{1}{mK} \sum_{i \in \mathcal{I}} \sum_{k=1}^{K} \begin{cases} -\log \rho_{ik}, & k \in \mathcal{K}_{\mathrm{head}}, \\ -(1 - \rho_{ik})^{\gamma} \log \rho_{ik}, & k \in \mathcal{K}_{\mathrm{tail}} \cup \mathcal{K}_{\mathrm{trans}}, \end{cases} \tag{12}$$

where $\gamma = 2$ is the focal-loss focusing parameter.

The complete training objective is

$$\mathcal{L} = \lambda_{\mathrm{DoL}} \mathcal{L}_{\mathrm{DoL}} + \lambda_{\mathrm{hier}} (\mathcal{L}_{\mathrm{sys}} + \mathcal{L}_{\mathrm{sg}}) + \lambda_{\mathrm{RSCL}} \mathcal{L}_{\mathrm{RSCL}}, \tag{13}$$

where the nonnegative coefficients λ_{DoL}, λ_{hier}, and λ_{RSCL} control the contributions of expert optimization, hierarchical supervision, and rule-semantic contrastive learning, respectively.

5 Experiments

5.1 Experimental Setup

► **Datasets.** We evaluate on CCDC, derived from the Cambridge Structural Database (Groom et al., 2016); CoREMOF, a computation-ready metal–organic framework database (Chung et al., 2019); and Inorganic (Belsky et al., 2002; Salgado et al., 2023). Their PXRD patterns are simulated from

Table 1: Overall performance (%) on CCDC, CoREMOF, and Inorganic. Models are trained and evaluated separately on each dataset. F1 and Recall are macro-averaged. Bold indicates the best result for each dataset and metric.

Method	Dataset	Top-1	Top-2	Top-5	F1	Recall
CNN	CCDC	30.20±0.99	40.50±1.32	60.00±2.09	5.05±0.31	5.79±0.33
	CoREMOF	27.31±1.01	41.82±1.22	58.07±1.97	4.46±0.28	4.77±0.24
	Inorganic	39.29±1.04	55.27±1.41	62.50±2.33	15.85±0.56	15.32±0.60
XRDMamba	CCDC	62.71±1.01	77.34±1.33	89.61±2.14	46.77±1.03	45.35±0.98
	CoREMOF	31.74±0.98	42.19±1.21	54.57±1.87	14.17±0.54	14.69±0.55
	Inorganic	37.52±1.00	51.54±1.32	**66.15±2.29**	20.24±1.05	**21.57±0.87**
XRDecoupler	CCDC	75.09±1.11	88.11±1.76	92.26±2.31	46.18±1.12	48.43±1.19
	CoREMOF	32.71±0.99	40.89±1.20	51.92±1.67	15.11±0.60	15.67±0.66
	Inorganic	35.54±1.09	45.25±1.26	58.83±1.87	19.07±0.84	20.61±0.88
ICL-MoE	CCDC	33.13±1.01	43.05±1.28	63.36±2.12	1.98±0.11	3.11±0.18
	CoREMOF	17.75±0.58	31.27±0.97	48.87±1.34	0.68±0.13	0.90±0.15
	Inorganic	28.20±0.87	47.25±1.39	63.51±1.97	3.13±0.22	1.95±0.18
Phys-MoE	CCDC	**81.70±1.21**	**89.66±1.59**	**93.93±2.01**	**49.36±1.10**	**50.11±0.94**
Phys-MoE	CoREMOF	**44.78±1.17**	**53.00±1.43**	**62.36±1.89**	**16.35±0.68**	**17.53±0.59**
Phys-MoE	Inorganic	**51.88±1.21**	**57.61±1.39**	63.55±1.44	**21.79±0.58**	21.40±0.55

crystal structures. CCDC is the primary long-tailed benchmark; CoREMOF and Inorganic provide different material compositions and class distributions.

Let n_g denote the number of training examples for space group g. We divide the classes into four disjoint frequency buckets: Many-shot ($n_g > 1000$), Medium-shot ($100 \leq n_g \leq 1000$), Tail ($20 \leq n_g < 100$), and Extreme-Tail ($n_g < 20$). For cross-distribution evaluation of CCDC-trained models, bucket membership is determined by the CCDC training counts.

► **Metrics.** We report sample-weighted Top-1, Top-2, and Top-5 accuracy, together with macro-averaged F1 and recall. Bucket-wise accuracy is computed by first calculating Top-1 accuracy for each evaluated space group and then averaging the class accuracies equally within each bucket.

► **Baselines.** Baselines are a 1D-CNN (Salgado et al., 2023), XRDMamba (Yu et al., 2024), XRDecoupler (Yu et al., 2026), and the ICL-MoE adaptation (Chen et al., 2025). CCDC comparisons also include Balanced Softmax (Ren et al., 2020), Logit Adjustment (Menon et al., 2020), and LDAM-DRW (Cao et al., 2019). Section 5.4 traces the adaptation from ICL-MoE.

► **Evaluation Protocols.** Main benchmark models are trained and evaluated separately on each dataset. Controlled routing studies train CCDC models with eight experts across repeated runs, then freeze each model for evaluation on all three datasets. Reported accuracies are means across runs. Training settings and controlled-study splits are in Appendices A and B.

► **Statistical Reporting.** Table 1 reports means and standard deviations over five randomly selected training seeds. Bucket-wise and controlled comparisons use 2,000 paired-bootstrap resamples (seed 42) and percentile 95% intervals. Overall comparisons resample examples; bucket comparisons resample within classes and average classes equally.

5.2 Main Results

► **Overall Performance.** Table 1 summarizes the five-seed results on the three PXRD benchmarks. Phys-MoE achieves mean Top-1 accuracies of $81.70 \pm 1.21\%$ on CCDC, $44.78 \pm 1.17\%$ on CoREMOF, and $51.88 \pm 1.21\%$ on Inorganic, improving over the best baseline mean on each dataset by 6.61, 12.07, and 12.59 percentage points, respectively. It also achieves the highest mean macro-F1 on all three datasets and the highest mean values across all five metrics on CCDC and CoREMOF.

► **Long-Tailed Performance.** The class-balanced CCDC results show that Phys-MoE has the highest accuracy in each bucket, with the largest gain in Extreme-Tail: 26.19%, a 16.19-point improvement over the best baseline in the table (95% CI: [4.20, 28.18]). Medium-shot accuracy reaches 66.54%, with a 5.53-point gain (95% CI: [1.04, 10.02]); the gains on Many-shot and Tail are 0.75 and approximately 7.4 points.

Table 2: CCDC class-balanced Top-1 (%). Gains and paired-bootstrap 95% CIs are relative to the best baseline in each bucket.

Method	Many	Medium	Tail	Extreme
CNN	38.89	7.13	3.80	0.00
XRDMamba	75.01	57.43	34.42	2.00
XRDecoupler	80.40	61.01	39.13	10.00
ICL-MoE	43.69	5.23	0.00	0.00
Phys-MoE	**81.15**	**66.54**	**46.5**	**26.19**
Gain	+0.75	+5.53	+7.4	+16.19
95% CI	[−1.62, 3.10]	[1.04, 10.02]	–	[4.20, 28.18]

5.3 Ablation: Leave-One-Out

We disable one component at a time on top of full Phys-MoE: ❶ no HR removes the hierarchical mask and $\mathcal{L}_{\text{sys}}$; ❷ no RGE replaces the rule-guided experts with standard independent experts; ❸ no RSCL removes $\mathcal{L}_{\text{RSCL}}$. Removing any component lowers both Tail and Extreme-Tail accuracy (Table 3). Removing HR produces the largest decrease: 13.1 points on Tail and 10.07 points on Extreme-Tail. Removing RGE lowers the two accuracies by approximately 9.7 and 5.87 points, while removing RSCL lowers them by 9.2 and 6.76 points. The no-RSCL configuration corresponds to Step C in Section 5.4.

Table 3: CCDC leave-one-out ablation (Top-1, %).

HR	RGE	RSCL	Tail	Extreme-Tail
✔	✔	✔	**46.5**	**26.19**
✘	✔	✔	33.40	16.12
✔	✘	✔	36.78	20.32
✔	✔	✘	37.30	19.43

5.4 From ICL-MoE to Phys-MoE, One Step at a Time

We trace the conversion from the ICL-MoE adaptation (Chen et al., 2025) to Phys-MoE (Figure 2a). Step A replaces independent feature towers with a shared backbone and lightweight expert heads, and adopts asymmetric DoL (CE on head experts and focal loss on the other experts). Steps B, C, and D add HR, RGE, and RSCL, respectively.

The backbone/loss adaptation (0→A) adds 39.67 Overall and 14.29 Extreme-Tail points; the crystallographic components (A→D) add a further 8.90 and 11.90 points. Step A′ adds RSCL directly to A without HR or RGE. Figure 2b and Appendix Table 9 compare this configuration with the complete model.

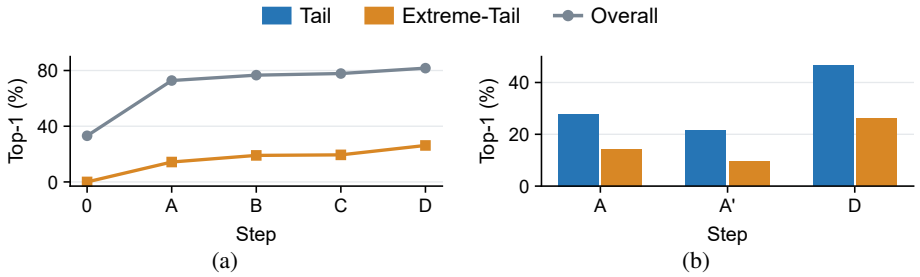

Figure 2: Incremental construction on CCDC. (a) Steps 0–D; (b) A, A′, and D. Full values are in Appendix Table 9.

5.5 Generic Long-Tailed Baselines and Error Analysis

► **Generic Long-Tailed Comparisons.** We evaluate Balanced Softmax, Logit Adjustment, and LDAM-DRW in the reported CCDC setting (Appendix Table 13). LDAM-DRW reaches 72.01% Overall and 16.67% Extreme-Tail accuracy; Phys-MoE improves these results by 9.69 and 9.52 points. Adding HR, RGE, and RSCL to Balanced Softmax raises Extreme-Tail accuracy from 4.00% to 14.29%.

► **Error Location and Severity.** Among class-averaged outcomes, intra-system wrong-group predictions account for 10.3% in Many-shot and 47.1% in Extreme-Tail (Figure 3a). We further characterize incorrect predictions using the operation-set similarity $S(y, \hat{g})$. Figure 3b shows that low-similarity predictions ($S < 0.2$) make up 5.69% of Phys-MoE errors, compared with 9.01% for LDAM-DRW. Cross-system predictions account for 8.11% and 22.20% of all test examples, respectively (Figure 3c).

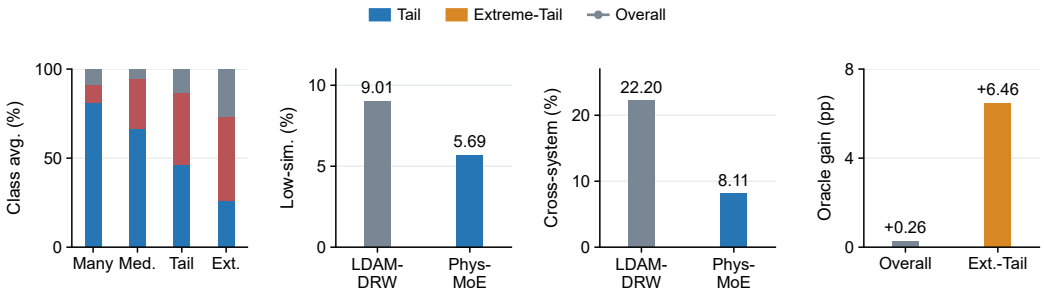

Figure 3: CCDC errors and routing. (a) Correct, intra-system, and cross-system outcomes, averaged equally across space groups within each bucket. (b) Low-similarity predictions, $S(y, \hat{g}) < 0.2$ (% of errors). (c) Cross-system predictions (% of all examples). (d) Top-1 gain with the true system at inference. Values and Top-2 routing intervals: Appendix Tables 14, 10, and 11.

5.6 Crystal-system Routing: Accuracy and Propagation

► **System Prediction and Oracle Analysis.** The crystal-system head reaches 91.89% Top-1 on CCDC. Replacing its prediction with the true system at inference raises Overall space-group accuracy from 81.70% to 81.96% and Extreme-Tail accuracy from 26.19% to 32.65% (Figure 3d). In the class-averaged decomposition, cross-system predictions account for 26.7% of Extreme-Tail outcomes, compared with 8.5% for Many-shot (Figure 3a). A wrong system excludes the true group under HR, which is consistent with the larger oracle gain for Extreme-Tail. Overall Top-1 counts test examples, whereas the Extreme-Tail metric averages over rare groups; the two metrics therefore expose different effects of the same routing decision. Within-system mistakes remain the largest Extreme-Tail error category, so correct system selection still leaves fine-grained groups to distinguish. We also evaluate Top-2 crystal-system routing with the trained model fixed; its results are in Appendix Table 11. Table 4 reports system-head accuracy across the seven crystal systems; class-averaged prediction outcomes are in Appendix Table 14. System-head accuracy varies across the hierarchy: Hexagonal and Monoclinic reach 83.25% and 84.38%, whereas Tetragonal and Cubic reach 97.30% and 97.73% (Table 4). These differences matter for HR because the predicted system determines which of the 230 space-group logits remain eligible at inference.

Table 4: Crystal-system Top-1 on CCDC (%).

Crystal system	Test share	Top-1
Triclinic	25.9	91.81
Monoclinic	22.2	84.38
Orthorhombic	33.1	95.55
Tetragonal	9.5	97.30
Trigonal	5.6	93.41
Hexagonal	2.3	83.25
Cubic	1.5	97.73
Overall	100	**91.89**

► **Probability Routing and Multi-mask Training.** We compare single- and multi-mask training under hard and probability routing using frozen CCDC-trained models. Probability routing combines the within-system space-group distributions with the predicted crystal-system probabilities. These comparisons separate the training constraint from the inference rule: multi-mask training includes true-system, two-system, and unmasked terms in the space-group loss, while hard and probability routing change how system predictions determine the final group scores. The CCDC-trained checkpoints are evaluated without refitting on CoREMOF and Inorganic. The training objective and additional routing variants are detailed in Appendix B; the comparison appears in Appendix Table 12.

5.7 INSIDE THE MODEL

► **Expert Weighting across Crystal Systems.** Table 5 reports the mean gate weights for each crystal system; Appendix Figure 4 shows baseline and Phys-MoE feature projections for head-class $Pnma$ and Extreme-Tail $Pn\bar{3}n$. The mean gate weights emphasize Experts 1–3 for Cubic and Hexagonal, Experts 4–6 for Trigonal, Tetragonal, and Orthorhombic, and Experts 7–8 for Monoclinic and Triclinic. Each system distributes weight across several experts rather than selecting a single fixed expert.

Table 5: Mean expert activation probability by crystal system on the evaluated set. Values are rounded to two decimals.

System	E1	E2	E3	E4	E5	E6	E7	E8
Cubic	0.20	0.29	0.25	0.09	0.02	0.02	0.01	0.12
Hexagonal	0.24	0.26	0.16	0.14	0.12	0.03	0.03	0.03
Trigonal	0.05	0.09	0.07	0.24	0.26	0.18	0.05	0.06
Tetragonal	0.07	0.12	0.03	0.24	0.23	0.18	0.10	0.03
Orthorhombic	0.01	0.13	0.13	0.22	0.20	0.16	0.09	0.06
Monoclinic	0.02	0.09	0.01	0.16	0.05	0.12	0.27	0.30
Triclinic	0.06	0.02	0.11	0.09	0.11	0.10	0.23	0.27

The largest gate weight is E2 for Cubic and Hexagonal, E5 for Trigonal, E4 for Tetragonal and Orthorhombic, and E8 for Monoclinic and Triclinic. These per-system maxima range from 0.22 to 0.30 (Table 5). Monoclinic and Triclinic place 0.57 and 0.50 total weight on E7–E8, respectively, with the remaining weight distributed across the other experts. ► **Controlled PXRD Perturbations.** On the original CCDC evaluation split, we test Gaussian intensity noise, peak shifts, missing peaks, preferred-orientation reweighting, and impurity mixtures, using five perturbation seeds per condition. At the reported medium severity, Overall Top-1 ranges from 78.50% to 81.10% and Extreme-Tail Top-1 from 22.00% to 25.00% across these conditions, compared with 81.70% and 26.19% on the clean set (Table 6).

Table 6: Controlled CCDC PXRD perturbations at medium severity (Top-1, %). Drops are percentage points relative to the clean benchmark; five perturbation seeds are used per condition.

Setting	Overall		Extreme-Tail	
	Top-1	Drop	Top-1	Drop
Clean benchmark	81.70	–	26.19	–
Gaussian intensity noise	81.10	0.60	25.00	1.19
Peak shift	80.80	0.90	24.50	1.69
Missing peaks	80.20	1.50	24.00	2.19
Preferred orientation	79.80	1.90	23.50	2.69
Impurity/multi-phase	78.50	3.20	22.00	4.19

The two metrics rank the perturbations identically, from intensity noise to impurity/multi-phase mixtures. Across all five conditions, the difference between the Extreme-Tail and Overall drops remains below one percentage point (Table 6).

6 CONCLUSION

Phys-MoE uses crystallographic hierarchy and operation-set relations for PXRD space-group recognition, reaching 81.70% overall and 26.19% Extreme-Tail Top-1 on CCDC. Ablations show gains from HR, RGE, and RSCL after the shared-backbone adaptation; error analyses identify intra-system confusion and system-prediction failures. The controlled perturbation results are in Table 6, and additional routing studies are in the Appendix. Future work should address rare-group routing across material distributions and evaluate measured spectra and mixed phases.

AI USE STATEMENT

Generative AI was used to assist with manuscript translation and editing, literature and citation checks, feedback on experimental design, interpretation and presentation of reported results, and preparation of LaTeX and figure-generation code. The authors are responsible for the scientific content and its verification.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 define the label hierarchy, rule similarity, routing, expert weighting, and training objectives. Section 5 describes the datasets, baselines, and metrics. Appendix A gives the training configuration, and Appendix B specifies the controlled splits, perturbations, uncertainty estimates, and routing variants, together with numerical tables for the ablation and error analyses.

REFERENCES

Alec Belsky, Mariette Hellenbrandt, Vicky Lynn Karen, and Peter Luksch. New developments in the inorganic crystal structure database (icsd): accessibility in support of materials research and design. *Structural Science*, 58(3):364–369, 2002.

Carolyn Pratt Brock, T Hahn, H Wondratschek, U Müller, U Shmueli, E Prince, A Authier, V Kopsky, D Litvin, E Arnold, et al. International tables for crystallography volume a: Space-group symmetry, 2016.

Bin Cao, Yang Liu, Zinan Zheng, Ruifeng Tan, Jia Li, and Tong-yi Zhang. Simxrd-4m: big simulated x-ray diffraction data and crystal symmetry classification benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 70721–70745, 2025a.

Bin Cao, Zinan Zheng, Yang Liu, Longhan Zhang, Lawrence WY Wong, Lu-Tao Weng, Jia Li, Haoxiang Li, and Tong-Yi Zhang. Xqueryer: an intelligent crystal structure identifier for powder x-ray diffraction. *National Science Review*, 12(12):nwaf421, 2025b.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Nithin Chalapathi, Yiheng Du, and Aditi Krishnapriyan. Scaling physics-informed hard constraints with mixture-of-experts. In *International Conference on Learning Representations*, volume 2024, pp. 36015–36033, 2024.

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*, 146 (12):8098–8109, 2024.

Yanhao Chen, Zhongquan Jian, Nianxin Ke, Shuhao Hu, Junjie Jiao, Qingqi Hong, and Qingqiang Wu. Enhancing mixture of experts with independent and collaborative learning for long-tail visual recognition. In *IJCAI*, pp. 828–836, 2025.

Yongchul G Chung, Emmanuel Haldoupis, Benjamin J Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S Camp, et al. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: Core mof 2019. *Journal of Chemical & Engineering Data*, 64(12):5985–5998, 2019.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9260–9269. IEEE, 2019.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Carmelo Giacovazzo, Hugo L Monaco, Davide Viterbo, F Scordari, Gastone Gilli, G Zanotti, and M Catti. *Fundamentals of crystallography*, volume 7. Oxford university press Oxford, 2002.

Colin R Groom, Ian J Bruno, Matthew P Lightfoot, and Suzanna C Ward. The cambridge structural database. *Structural Science*, 72(2):171–179, 2016.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Hayat Khan, Aditya S Yerramilli, Adrien D'Oliveira, Terry L Alford, Daria C Boffito, and Gregory S Patience. Experimental methods in chemical engineering: X-ray diffraction spectroscopy—xrd. *The Canadian journal of chemical engineering*, 98(6):1255–1266, 2020.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Massimo Nespolo. International tables for crystallography, volume a, space-group symmetry. edited by mois i. aroyo. wiley, 2016. pp. xxi+ 873. price gbp 295.00, eur 354.00 (hardcover). isbn 978-0-470-97423-0. *Foundations of Crystallography*, 73(3):274–276, 2017.

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*, 4(4):486–494, 2017.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020.

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*, 9(1):214, 2023.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

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.

P-E Werner, L Eriksson, and M Westdahl. Treor, a semi-exhaustive trial-and-error powder indexing program for all symmetries. *Applied Crystallography*, 18(5):367–370, 1985.

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*, pp. 4233–4237, 2024.

Liheng Yu, Zhe Zhao, Xucong Wang, Di Wu, and Pengkun Wang. Rethinking crystal symmetry prediction: A decoupled perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16208–16216, 2026.

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.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10795–10816, 2023.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6898–6907. IEEE, 2022.

Appendix

Physics-Routed MoE with Crystallographic Priors for Long-Tailed Space-Group Recognition

The Appendix contains the following material:

- **Section A: Experimental Settings.** Training configuration for the original benchmarks.
- **Section B: Additional experiments and diagnostics.** Uncertainty estimates, controlled perturbation settings, generic long-tailed comparisons, error diagnostics, frozen transfer, routing alternatives and multi-mask experiments.
- **Section C: Architectural Details.** Implementation details for the rule matrix, expert partitioning, and forward pass.
  - Section C.1: Physics Rule Matrix.
  - Section C.2: Expert Partitioning and Rule-bias Initialization.
  - Section C.3: Forward Pass.
- **Section D: Comparison scope.** PXRD-only comparison settings and their relation to recent methods.
- **Section E: Feature visualization.** t-SNE feature projections.
- **Section F: Model size.** Parameter counts and their counting conventions.
- **Section G: Limitations.** Four limitations of Phys-MoE and directions for further evaluation.

A Experimental Settings

Table 7 summarizes the original benchmark training configuration. We optimize cross-entropy, focal, hierarchical, and RSCL objectives using SGD with cosine annealing. The controlled CCDC-trained experiments and their data splits are specified in Appendix B.

Table 7: Phys-MoE training hyperparameters.

Configuration	Value	Configuration	Value
Optimizer	SGD	Batch size m	256
Init Learning Rate	5×10^{-4}	Total Epochs	100
Weight Decay	1×10^{-4}	Warmup Epochs	5
LR Scheduler	Cosine Annealing	Grad Clip Norm	1.0
Input shape $L \times 1$	8500×1	Experts K	8
Backbone feature dim	1024	Projected dim d	256
Crystal Systems	7	Space Groups	230
Frontend	ResTCN (12-layer)	Activation	LeakyReLU/GELU
λ_{DoL}	0.5	λ_{hier}	1.5
λ_{RSCL}	0.2	Contrastive τ	0.07

B Additional experiments and diagnostics

Statistical reporting. Table 1 reports the mean and standard deviation over five independently trained random seeds. Table 8 reports the same five-seed mean together with a paired-bootstrap 95% confidence interval. In each of 2,000 bootstrap replicates (seed 42), test examples are sampled with replacement within each dataset using the same sampled indices for all methods. Each metric is computed separately for each of the five trained models and then averaged over seeds; the 2.5th and 97.5th percentiles of these replicate means form the interval. For Table 2, examples are resampled within each true space group and class accuracies are averaged equally within each frequency bucket.

Table 8: Five-seed mean evaluation: point estimates and bootstrap 95% intervals. Each entry gives the estimate followed by [lower, upper] (%).

Method	Dataset	Top-1	Top-2	Top-5	F1	Recall
CNN	CCDC	30.20 [28.61,31.79]	40.50 [38.80,42.20]	60.00 [58.31,61.69]	5.05 [4.03,6.07]	5.79 [4.70,6.88]
CNN	CoREMOF	27.31 [22.55,32.07]	41.82 [36.55,47.09]	58.07 [52.80,63.34]	4.46 [1.48,7.44]	4.77 [1.70,7.84]
CNN	Inorganic	39.29 [33.85,44.73]	55.27 [49.73,60.81]	62.50 [57.11,67.89]	15.85 [10.36,21.34]	15.32 [9.91,20.73]
XRDMamba	CCDC	62.71 [61.04,64.38]	77.34 [75.89,78.79]	89.61 [88.55,90.67]	46.77 [44.44,49.10]	45.35 [43.03,47.67]
XRDMamba	CoREMOF	31.74 [26.77,36.71]	42.19 [36.92,47.46]	54.57 [49.25,59.89]	14.17 [9.14,19.20]	14.69 [9.59,19.79]
XRDMamba	Inorganic	37.52 [32.13,42.91]	51.54 [45.98,57.10]	66.15 [60.88,71.42]	20.24 [14.20,26.28]	21.57 [15.39,27.75]
XRDecoupler	CCDC	75.09 [73.59,76.59]	88.11 [86.99,89.23]	92.26 [91.34,93.18]	46.18 [43.85,48.51]	48.43 [46.10,50.76]
XRDecoupler	CoREMOF	32.71 [27.70,37.72]	40.89 [35.64,46.14]	51.92 [46.59,57.25]	15.11 [9.95,20.27]	15.67 [10.43,20.91]
XRDecoupler	Inorganic	35.54 [30.21,40.87]	45.25 [39.71,50.79]	58.83 [53.35,64.31]	19.07 [13.17,24.97]	20.61 [14.53,26.69]
ICL-MoE	CCDC	33.13 [31.50,34.76]	43.05 [41.34,44.76]	63.36 [61.69,65.03]	1.98 [1.33,2.63]	3.11 [2.30,3.92]
ICL-MoE	CoREMOF	17.75 [13.67,21.83]	31.27 [26.32,36.22]	48.87 [43.53,54.21]	0.68 [0.00,1.86]	0.90 [0.00,2.26]
ICL-MoE	Inorganic	28.20 [23.19,33.21]	47.25 [41.69,52.81]	63.51 [58.15,68.87]	3.13 [0.51,5.75]	1.95 [0.00,4.03]
Phys-MoE	CCDC	81.70 [80.36,83.04]	89.66 [88.61,90.71]	93.93 [93.10,94.76]	49.36 [47.03,51.69]	50.11 [47.78,52.44]
Phys-MoE	CoREMOF	44.78 [39.47,50.09]	53.00 [47.67,58.33]	62.36 [57.19,67.53]	16.35 [11.02,21.68]	17.53 [12.05,23.01]
Phys-MoE	Inorganic	51.88 [46.32,57.44]	57.61 [52.11,63.11]	63.55 [58.19,68.91]	21.79 [15.59,27.99]	21.40 [15.24,27.56]

Controlled perturbations. The controlled perturbation analysis applies five perturbation seeds per setting to the fixed CCDC evaluation split; Table 6 in the main text reports the medium-severity results. Mixtures are formed as $x' = (1-\alpha)x_{\text{main}} + \alpha x_{\text{impurity}}$, where $\alpha \in [0, 1]$ is the impurity mixing weight, and evaluated against the main component's space-group label.

Numerical results for the main figures and routing study. Table 9 gives the values for Figure 2; Tables 10 and 11 support Figure 3. Table 12 reports the controlled routing and multi-mask comparison.

Table 9: Incremental conversion from ICL-MoE to Phys-MoE on CCDC (accuracy, %). Each step adds to the preceding configuration; A′ adds RSCL directly to A.

Step	Configuration	Overall	Many	Medium	Tail	Extreme
0	ICL-MoE adaptation	33.13	43.69	5.23	0.00	0.00
A	Shared backbone + asymmetric DoL	72.80	73.47	60.19	27.85	14.29
B	A + HR	76.67	76.36	63.60	33.86	19.05
C	B + RGE	77.83	78.78	65.02	37.30	19.43
D	C + RSCL (Phys-MoE)	81.70	81.15	66.54	46.5	26.19
A′	A + RSCL (no HR or RGE)	74.62	81.36	59.56	21.52	9.52

Table 10: CCDC error severity. The column headings specify the denominator of each percentage. Low similarity is defined by $S(y, \hat{g}) < 0.2$.

Model	Low-similarity (% of errors)	Cross-system (% of examples)
LDAM-DRW	9.01	22.20
Phys-MoE	5.69	8.11

Table 11: Routing diagnostics on the original CCDC evaluation. Absolute accuracies are in %; changes and 95% CIs are in percentage points relative to predicted-system hard routing.

Inference setting	Overall	Extreme-Tail
Predicted system + hard mask	81.70	26.19
True system + hard mask	81.96	32.65
Oracle change	+0.26	+6.46
Top-2 change [95% CI]	$+0.04\ [-0.17, 0.26]$	$-2.38\ [-10.81, 5.88]$

Table 12: Controlled CCDC-trained models with single- or multi-mask training across repeated runs ($K = 8$). Each entry is mean Overall / class-balanced Extreme-Tail accuracy (%). Probability routing uses Equation 14. External models are evaluated with frozen parameters.

Training	Inference	CCDC	CoREMOF	Inorganic
Single mask	Hard	81.70 / 26.19	44.78 / 12.50	42.58 / 11.76
Single mask	Probability	72.64 / 11.90	48.07 / 12.50	41.61 / 11.76
Multi-mask	Hard	73.67 / 14.29	49.55 / 8.33	44.52 / 5.88
Multi-mask	Probability	74.11 / 14.29	50.45 / 8.33	45.48 / 5.88

Generic long-tailed baselines. Table 13 compares three class-frequency-based objectives and the Balanced-Softmax variant equipped with HR, RGE, and RSCL. All methods use the original CCDC evaluation setting.

Table 13: Generic long-tailed baselines and a Balanced-Softmax variant with the Phys-MoE domain modules on CCDC (Top-1, %).

Method	Domain modules	Overall	Extreme-Tail
Balanced Softmax	None	54.20	4.00
Logit Adjustment	None	48.80	5.50
LDAM-DRW	None	72.01	16.67
Balanced Softmax + Phys-MoE	HR+RGE+RSCL	74.71	14.29
Phys-MoE (DoL+Focal)	HR+RGE+RSCL	**81.70**	**26.19**

Error and crystal-system diagnostics. Table 14 reports class-averaged prediction outcomes by frequency bucket. Crystal-system accuracy by system is reported in main-text Table 4.

Table 14: Class-averaged CCDC prediction outcomes by frequency bucket (%). Outcome shares are computed for each true space group and averaged equally across groups in the bucket. Each row sums to 100%.

Bucket	Correct	Intra-system (SPC)	Cross-system
Many-shot	81.2	10.3	8.5
Medium-shot	66.5	28.2	5.3
Tail	46.5	40.1	13.4
Extreme-Tail	26.2	**47.1**	26.7

Controlled training and frozen evaluation. The controlled comparisons use 108,407 CCDC training examples, 8,968 validation examples, and 3,213 test examples. CoREMOF and Inorganic use validation/test sizes of 594/337 and 611/310, respectively. Model weights are fitted on CCDC. The repeated runs use eight experts, the same backbone and losses, a 230-by-230 Jaccard matrix, and a 200-epoch training schedule. The model has a trainable 7-by-8 rule-bias matrix initialized at ± 0.5. Model checkpoints are selected by CCDC validation hard-route Top-1.

Single- and multi-mask models are initialized separately under the same training configuration in each run. Reported accuracies are means across runs. For a set of crystal systems $\mathcal{C} \subseteq \{1, \ldots, 7\}$, define the union mask $M_{\mathcal{C}}(g) = 0$ when $T(g) \in \mathcal{C}$ and $M_{\mathcal{C}}(g) = -\infty$ otherwise. The batch-averaged space-group loss is

$$\mathcal{L}_{\mathrm{sg}}^{\mathrm{multi}} = \frac{1}{m} \sum_{i \in \mathcal{I}} \Big[0.5 \ \mathrm{CE}(\ell_i + M_{s_i}, y_i) + 0.25 \ \mathrm{CE}(\ell_i + M_{\{s_i, s_i^\star\}}, y_i) + 0.25 \ \mathrm{CE}(\ell_i, y_i) \Big],$$

where $s_i = T(y_i)$ and $s_i^\star = \arg\max_{s \neq s_i} q_{i,s}$. This replaces $\mathcal{L}_{\mathrm{sg}}$ in Equation 13; all other training losses are unchanged.

An exact-byte comparison of preprocessed spectra finds one training/test overlap in CCDC and zero overlaps between CCDC training and either external test set. The validation/test pairs have zero

exact-spectrum overlap. Five space-group labels in six CoREMOF test examples and three labels in three Inorganic examples are absent from CCDC training. All test examples are included in the reported results.

Additional routing alternatives for multi-mask training. Hard Top-k_{sys} retains the union of the $k_{\mathrm{sys}} \in \{1, 2, 3\}$ most probable crystal systems under q_i. Adaptive Top-2 retains the top system when $\max_s q_{i,s} \geq t$ and the top two otherwise. Here t is a system-confidence threshold. Writing $\ell_{i,g}$ for the g-th component of ℓ_i, the probability route used in the main comparison is

$$p_{\mathrm{prob}}(g \mid x_i) = q_{i,T(g)} \frac{\exp(\ell_{i,g})}{\sum_{g':T(g')=T(g)} \exp(\ell_{i,g'})}. \tag{14}$$

The probability-routing sensitivity study ranks groups by

$$u_{i,\beta}(g) = \beta \log q_{i,T(g)} + \ell_{i,g} - \log \sum_{g':T(g')=T(g)} \exp(\ell_{i,g'}),$$

where $\beta \in \{0.25, 0.5, 1, 2, 4\}$ scales the system log-probability. Confidence fallback uses hard Top-1 when $\max_s q_{i,s} \geq t$ and probability routing with $\beta = 1$ otherwise. Both threshold families use $t \in \{0.5, 0.7, 0.85, 0.95, 0.99\}$. Hard Top-1 is included in the fallback selection set. The multi-mask routing grid is reported in Table 15.

Table 15: Frozen-routing grid after multi-mask training. Each entry is mean Overall / class-balanced Extreme-Tail accuracy (%).

Routing	CCDC	CoREMOF	Inorganic
Hard Top-1	73.67 / 14.29	49.55 / 8.33	44.52 / 5.88
Hard Top-2	74.11 / 11.90	48.96 / 8.33	45.48 / 0.00
Hard Top-3	74.01 / 9.52	48.96 / 8.33	46.13 / 0.00
Adaptive Top-2 ($t = 0.5$)	73.70 / 14.29	49.85 / 8.33	44.52 / 5.88
Adaptive Top-2 ($t = 0.7$)	74.11 / 11.90	49.55 / 8.33	45.48 / 5.88
Adaptive Top-2 ($t = 0.85$)	74.14 / 11.90	49.26 / 8.33	45.16 / 5.88
Adaptive Top-2 ($t = 0.95$)	74.20 / 11.90	49.26 / 8.33	45.48 / 0.00
Adaptive Top-2 ($t = 0.99$)	74.11 / 11.90	49.26 / 8.33	45.48 / 0.00
Probability ($\beta = 0.25$)	74.14 / 11.90	48.96 / 8.33	44.84 / 0.00
Probability ($\beta = 0.5$)	74.29 / 14.29	49.85 / 8.33	45.81 / 5.88
Probability ($\beta = 1$)	74.11 / 14.29	50.45 / 8.33	45.48 / 5.88
Probability ($\beta = 2$)	73.92 / 14.29	50.45 / 8.33	45.16 / 5.88
Probability ($\beta = 4$)	73.86 / 14.29	49.85 / 8.33	44.84 / 5.88
Fallback ($t = 0.5$)	73.70 / 14.29	50.15 / 8.33	44.52 / 5.88
Fallback ($t = 0.7$)	74.11 / 14.29	50.45 / 8.33	45.48 / 5.88
Fallback ($t = 0.85$)	74.11 / 14.29	50.45 / 8.33	45.48 / 5.88
Fallback ($t = 0.95$)	74.11 / 14.29	50.45 / 8.33	45.48 / 5.88
Fallback ($t = 0.99$)	74.11 / 14.29	50.45 / 8.33	45.48 / 5.88

Multi-mask training comparisons. Table 16 compares training strategies using probability routing. Table 17 reports both sample-weighted and class-balanced Extreme-Tail results.

Table 16: Multi-mask minus single-mask training under probability routing. Accuracies are means across runs (%); gains are in percentage points and use unrounded accuracies. Bracketed paired-bootstrap 95% CIs are from the seed-42 paired evaluation.

Test set	Route	Single	Multi	Gain [95% CI]
CCDC	Probability	72.64	74.11	1.46 [0.22, 2.71]
CoREMOF	Probability	48.07	50.45	2.37 [-1.48, 6.53]
Inorganic	Probability	41.61	45.48	3.87 [-0.01, 8.06]

Table 17: Mean Extreme-Tail accuracy across runs under probability routing ($\beta = 1$), with CCDC training-frequency buckets.

Test set	Training	Samples	Classes	Sample-weighted	Class-balanced
CCDC	Single	42	42	11.90	11.90
CCDC	Multi	42	42	14.29	14.29
CoREMOF	Single	36	24	8.33	12.50
CoREMOF	Multi	36	24	5.56	8.33
Inorganic	Single	30	17	10.00	11.76
Inorganic	Multi	30	17	3.33	5.88

C Architectural Details

C.1 Physics Rule Matrix

For space groups g, g', let $\mathcal{O}_g$ and $\mathcal{O}_{g'}$ be their symmetry-operation sets from the *International Tables* (Brock et al., 2016). The fixed class-level similarity is

$$S(g, g') = \frac{|\mathcal{O}_g \cap \mathcal{O}_{g'}|}{|\mathcal{O}_g \cup \mathcal{O}_{g'}|} \in [0, 1].$$

For samples i, j, RSCL uses $S(y_i, y_j)$ to form the positive set $\mathcal{P}(i)$ in Equation 9 and the non-self pair weight $\omega_{ij} = 1 - S(y_i, y_j)$ in Equation 10. The similarity matrix $S \in [0, 1]^{230\times 230}$ is distinct from the hierarchical logit mask M_s and the trainable expert-bias matrix B.

C.2 Expert Partitioning and Rule-bias Initialization

With $K = 8$, the expert-index sets are $\mathcal{K}_{\text{tail}} = \{1, 2, 3\}$, $\mathcal{K}_{\text{head}} = \{7, 8\}$, and $\mathcal{K}_{\text{trans}} = \{4, 5, 6\}$. The initial allocation associates the tail experts with Cubic, Hexagonal, and Trigonal systems, the head experts with Monoclinic and Triclinic systems, and the transition experts with Orthorhombic and Tetragonal systems. The expert heads use the asymmetric objective $\mathcal{L}_{\text{DoL}}$ defined in Equation 12. Let $\mathcal{K}_s$ denote the designated expert-index set for crystal system s. The trainable matrix $B \in \mathbb{R}^{7\times K}$ is initialized as

$$B^{(0)}_{s,k} = \begin{cases} +0.5, & k \in \mathcal{K}_s, \\ -0.5, & k \notin \mathcal{K}_s. \end{cases}$$

This initialization biases the gate toward the designated experts; B is subsequently optimized jointly with the other model parameters.

C.3 Forward Pass

Algorithm 1 summarizes one Phys-MoE training step using the notation of Section 4. The encoder and projection jointly produce $z_i = \phi_\theta(x_i) \in \mathbb{R}^d$. The system probabilities q_i enter the gate through $q_i B$; the expert logits $h_k(z_i)$ are aggregated into ℓ_i using $w_{i,k}$. Training applies the true-system mask M_{s_i}, where $s_i = T(y_i)$. The mini-batch index set is $\mathcal{I}$ with $m = |\mathcal{I}|$; η denotes the current learning rate.

D Scope of PXRD comparisons

Tables 1 and 2 compare the listed methods using PXRD-only inputs. XQueryer (Cao et al., 2025b) combines PXRD with elemental information for crystal-type identification and also reports an element-free variant, XQueryer-NoElem. Its prediction target and data protocol differ from the space-group benchmarks used here. SimXRD-4M (Cao et al., 2025a) provides simulated diffraction patterns and evaluations of sequence models for crystal-symmetry classification. Comparing these approaches under a common data split and prediction target remains future work.

Algorithm 1 Phys-MoE training step

Input: mini-batch $\{(x_i, y_i)\}_{i\in\mathcal{I}}$, fixed T, S, parameters θ (including B)
for $i \in \mathcal{I}$ **do**
  $z_i \leftarrow \phi_\theta(x_i)$; $s_i \leftarrow T(y_i)$
  $q_i \leftarrow \mathrm{Softmax}(W_{\mathrm{sys}}(z_i))$
  $a_i \leftarrow W_g(z_i) + q_i B$; $w_i \leftarrow \mathrm{Softmax}(a_i)$
  $h_k(z_i) \leftarrow z_i W_k + b_k$ for $k = 1, \ldots, K$
  $\ell_i \leftarrow \sum_{k=1}^{K} w_{i,k} h_k(z_i)$
  Form M_{s_i} using Equation 3
end for
Compute $\mathcal{L}_{\mathrm{sys}}$ and $\mathcal{L}_{\mathrm{sg}}$ using Equation 4
Compute $\mathcal{L}_{\mathrm{DoL}}$ using Equation 12
Form $\mathcal{P}(i)$ and ω_{ij} using Equations 9 and 10
Compute $\mathcal{L}_{\mathrm{RSCL}}$ using Equation 11
$\mathcal{L} \leftarrow \lambda_{\mathrm{DoL}}\mathcal{L}_{\mathrm{DoL}} + \lambda_{\mathrm{hier}}(\mathcal{L}_{\mathrm{sys}} + \mathcal{L}_{\mathrm{sg}}) + \lambda_{\mathrm{RSCL}}\mathcal{L}_{\mathrm{RSCL}}$
$\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}$

E FEATURE VISUALIZATION

Figure 4 compares two-dimensional feature projections for $Pnma$ and $Pn\bar{3}n$ under the baseline and Phys-MoE. The two panels show the head- and Extreme-Tail-class samples in each model's feature projection.

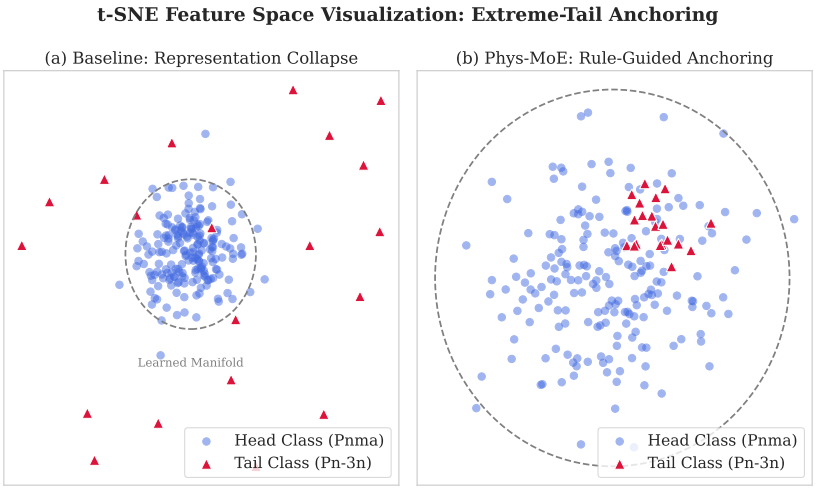

Figure 4: t-SNE feature projections for (a) the baseline and (b) Phys-MoE. Head-class $Pnma$ samples are shown as blue circles and Extreme-Tail $Pn\bar{3}n$ samples as red triangles.

F MODEL SIZE

Table 18 reports model-size counts. The Phys-MoE count includes floating-point tensors in a saved state dictionary; comparison-model counts come from separate implementations. These figures use implementation-specific counting conventions.

Table 18: Reported model-size figures and CCDC Extreme-Tail Top-1. Parameter counts require confirmation under a common counting convention before an efficiency claim is made. Unmeasured FLOPs are omitted.

Method	Reported params (M)	Extreme-Tail (%)
CNN	30.36	0.00
XRDMamba	68.60	2.00
XRDecoupler	82.23	10.00
ICL-MoE	75.42	0.00
Phys-MoE	67.76	**26.19**

G LIMITATIONS

(L1) Hard-mask cascading failure. A wrong crystal system at test time eliminates the correct space group from the candidate set. Replacing system predictions with ground truth raises Extreme-Tail Top-1 by 6.46 points in the original evaluation. Alternative routing and multi-mask configurations are evaluated in Appendix Table 12. Improving rare-group routing across material distributions remains an open problem.

(L2) Static rule matrix and measured spectra. S remains fixed during training, including when patterns are affected by experimental artifacts. Table 6 evaluates controlled distortions and simulated mixtures. Evaluation on a large labeled collection of measured spectra and a wider range of mixed phases remains future work.

(L3) Distribution-tied expert partition. Expert allocation and bias initialization use the CCDC class distribution. Broader material distributions, alternative expert allocations, and expert counts remain to be evaluated (Zhang et al., 2023).

(L4) Single-modality input. Phys-MoE uses only the PXRD pattern. XQueryer (Cao et al., 2025b) also uses elemental information for crystal-type identification. Extending system-conditioned routing to incorporate such information would require a multimodal evaluation.