npj Computational Materials

https://doi.org/10.1038/s41524-026-02015-y

Article in Press

KAN-enhanced contrastive learning: the accelerator of crystal structure identification from XRD patterns

Received: 24 October 2025

Accepted: 11 February 2026

Cite this article as: Xu, C., Su, T., Xiong, J. *et al.* KAN-enhanced contrastive learning: the accelerator of crystal structure identification from XRD patterns. *npj Comput Mater* (2026). https://doi.org/10.1038/s41524-026-02015-y

Chenlei Xu, Tianhao Su, Jie Xiong, Yue Wu, Shuya Dong, Tian Jiang, Mengwei He, Shuai Chen & Tong-Yi Zhang

KAN-Enhanced Contrastive Learning Accelerating Crystal Structure Identification from XRD Patterns

Chenlei Xu[a,1], Tianhao Su[a,1], Jie Xiong[a,d]*, Yue Wu[a], Shuya Dong [a], Tian Jiang[a],

Mengwei He[b, c]*, Shuai Chen[a,d], Tong-Yi Zhang[a,e]*

[a] Materials Genome Institute, Shanghai University, Shanghai 200444 China

[b] Australian Centre for Microscopy and Microanalysis, School of Aerospace, Mechanical and Mechatronic Engineering, The University of Sydney, Sydney, New South Wales, 2006, Australia

[c] School of Computer Science, The University of Sydney, Sydney, New South Wales, 2006, Australia

[d] State Key Laboratory of Materials for Advanced Nuclear Energy, Shanghai University, Shanghai 20044, PR China

[e] Guangzhou Municipal Key Laboratory of Materials Informatics, Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511400, China

[1] These authors contributed equally to this work and should be considered co-first authors.

E-mail: xiongjie@shu.edu.cn (J. Xiong), mengwei.he@sydney.edu.au (M. He), zhangty@shu.edu.cn (T.-Y. Zhang)

Abstract Accurate crystal structure determination underpins materials discovery, yet powder X-ray diffraction (XRD) analysis still depends on expert-driven, iterative fitting that limits scalability for high-throughput and autonomous experiments. We introduce XRD-Crystal Contrastive Pretraining (XCCP), a physics-guided contrastive learning framework that aligns PXRD patterns with candidate crystal structures in a shared embedding space to enable efficient structure retrieval and symmetry inference. XCCP employs a dual-expert XRD encoder with a Kolmogorov-Arnold Network (KAN) projection head. A low-angle branch captures long-length-scale signatures, while a wide-angle branch encodes dense, symmetry-governed fingerprints. Attribution and perturbation analyses show that the KAN head concentrates evidence on physically meaningful Bragg reflections rather than background-dominated regions, improving robustness to peak-shape variations. We further introduce similarity-based confidence scores to flag potentially unreliable predictions in open-set settings. Without elemental priors, XCCP

achieves 46.42% top-1 accuracy for structure retrieval and 60.85% accuracy for space-group identification. When chemical composition is available for elemental pre-screening, performance increases to 88.98% and 93.39%, respectively. XCCP also generalizes to compositionally similar multi-principal element alloys and enables zero-shot transfer to experimental patterns. These results establish XCCP as an interpretable, confidence-aware, and scalable paradigm for XRD analysis, enabling high-throughput screening, rapid candidate shortlisting, and integration with autonomous laboratory workflows.

Keywords: Powder X-ray diffraction; crystal-structure retrieval; contrastive learning; Kolmogorov-Arnold Network (KAN); confidence-aware.

Introduction

Powder X-ray diffraction (PXRD) is a foundational analytical technique for characterizing crystalline materials. It provides diffraction signatures that reveal key crystallographic parameters such as lattice constants, space groups, and phase compositions[1-5]. These capabilities have made XRD indispensable for microstructural characterization across materials science and engineering. However, conventional analysis workflows face notable limitations. Manual peak assignment based on the Bragg equation followed by database matching remains standard practice, requiring substantial crystallographic expertise, hindering operational efficiency, and becoming particularly challenging when diffraction peak overlaps[6-8]. Rietveld refinement, although widely used, depends on accurate initial assumptions and reliable reference powder diffraction file (PDF) cards[9,10], and its iterative nature constrains rapid characterization in high-throughput materials discovery[11,12].

Recent advances in data-driven methods have shown potential to address these challenges[13-19]. Deep neural networks can capture complex, nonlinear relationships inherent in high-dimensional diffraction data. Convolutional neural networks (CNNs) can resolve overlapping XRD peaks [20,21] and attention mechanisms have shown strong capabilities in capturing polycrystalline textures[22-25]. These models have demonstrated superior performance in classification tasks, often exceeding traditional methods[26-33]. For example, Lee et al.[29] trained CNNs on approximated 1.7 million synthetic XRD patterns generated from the Sr-Li-Al-O quaternary and achieved near-perfect identification in multi-phase settings. Cao et al.[34] developed a XRD simulation method that incorporates comprehensive physical interactions, resulting in a high-fidelity database including 4,065,346 simulated powder XRD patterns, representing 119,569 unique crystal structures under 33 simulated conditions that reflect real-world variations. Salgado et al.[35] reported a non-pooling CNN trained on 170,000 inorganic crystal structures that achieved 67% accuracy for crystal systems and only 36% for space groups on 2,253 test samples. Nevertheless, most existing efforts treat XRD primarily as a symmetry-assignment task, whereas in practice the central challenge is often retrieval, namely matching an observed pattern to candidate entries in a reference database[36-38].

Cross-modal contrastive learning offers a natural framework for this problem, as it learns shared representations that bring each diffraction pattern closer to its corresponding structure. Xie et al.[39] proposed a contrastive learning system that couples a mass spectrometry spectral encoder with a molecular structural encoder and reports notable gains in retrieval. A similar strategy in crystallography can align diffraction patterns with candidate structures and enable fast, accurate retrieval. The success of such a framework depends on encoders for both crystals and XRD signals. Crystal encoding has been studied extensively, and several effective graph neural networks (GNNs) are available[20,40-43]. Representative examples include the crystal graph convolutional neural network (CGCNN)[20], the atomistic line GNN (ALIGNN)[41], the GNN with three-body interactions (M3GNet)[40] and related variants that process atomistic graphs with learned message passing. The remaining challenge is the PXRD side of the bridge, where the encoder must respect how diffraction features form and must remain robust to realistic perturbations in peak shape and background.

Several recent studies illustrate the promise of multimodal PXRD structure learning while also exposing open gaps[44-46]. XtalNet targets metal organic frameworks and couples contrastive PXRD crystal pretraining with an equivariant diffusion generator, leveraging the dominance of low-angle responses in MOF diffraction[44]. PXRDGen advances end-to-end structure solution by combining contrastive encoders with conditional generation and refinement[46]. These efforts mark important progress, but they largely inherit generic spectral backbones and projection heads from classification settings. As a result, they do not explicitly encode diffraction-specific inductive bias that emphasizes localized Bragg reflections while down weighting broad background degrees of freedom. In addition, diffraction profiles are typically processed as a single stream without explicitly disentangling low-angle and wide-angle cues, even though low-angle reflections can carry complementary information about larger length scales. Finally, uncertainty and reliability controls remain underexplored, despite their importance for real deployments where queries can be ambiguous and databases incomplete.

A key ingredient for diffraction-aware representation is therefore the projection module that converts learned features into an embedding suited for similarity comparison. Powder XRD is characterized by

sparse, sharp, and highly localized peaks superimposed on slowly varying backgrounds, and the discriminative evidence often lives in narrow 2θ neighborhoods rather than in global trends. Kolmogorov-Arnold Networks (KANs) [47-53] are well matched to this structure because their learnable spline-based nonlinearities can allocate capacity locally and adaptively. This flexibility enables the model to delineate sharper, non-linear decision boundaries in the latent space, distinguishing complex diffraction features with higher efficiency than fixed activation MLP heads. In other words, KAN provides a natural mechanism to model spectrum-like signals in which local support and smoothness are both essential, making it a promising choice for stable cross-modal alignment in diffraction retrieval.

Guided by these considerations, we propose an XRD-Crystal Contrastive Pretraining (XCCP) framework for automated crystal-structure analysis from PXRD. As illustrated in **Fig 1**, XCCP couples a dual-expert XRD encoder that separately models small-angle (SA) reflections and dense wide-angle (WA) peaks, with a KAN projection head that produces physically grounded embeddings for cross modal alignment. These XRD embeddings are contrastively aligned with crystal embeddings from a structure encoder, enabling direct structure retrieval from a query pattern. The same retrieval output can also be used to infer symmetry labels such as space group when needed. For practical deployment, XCCP supports optional elemental pre-screening that reduces the candidate pool to chemically plausible structures, while also providing confidence-aware outputs that facilitate conservative decision making while still returning actionable shortlists for refinement. These design choices establish XCCP as a physically grounded and scalable route to PXRD analysis, supporting high throughput screening, rapid validation, and integration with autonomous laboratory platforms.

Results

Performance of XCCP in structure retrieval on simulated dataset

We treat PXRD-based structure identification as a structure-retrieval task rather than a label-only classification problem. Given a query powder pattern, the goal is to retrieve its ground-truth crystal

structure from a candidate database of CIF entries. XCCP embeds the query pattern into v_{xrd} and each candidate structure into v_c. Candidates are ranked by cosine similarity $\boldsymbol{s}(v_{xrd}, v_c)$ and the model returns a top-k list of structures. A query is counted as a top-k hit if the ground-truth entry appears within the returned top-k list. Accordingly, the top-k accuracy is reported as the hit rate averaged over all test queries.

Fig.2a summarizes the retrieval performance under two practical settings. Without any elemental prior, XCCP equipped with the proposed DEN-KAN encoder achieves a 46.42% top-1 retrieval accuracy. When elemental composition is available and used for composition-based pre-screening, top-1 accuracy increases to 88.98%, and top-3 and top-5 accuracies further rise to 97.56% and 98.82%, respectively. In routine workflows, this means that the correct structure appears within the first three candidates for almost all queries, enabling rapid shortlisting and downstream refinement. Importantly, elemental information alone does not guarantee unambiguous identification. Under an element-only baseline that filters candidates by composition and then selects a structure uniformly at random from the matched set, the expected top-1 hit rate is only 17.93%. This gap highlights that diffraction fingerprints are essential to resolve structural degeneracy beyond chemical plausibility. Moreover, the element-filtered accuracy of XCCP exceeds the 67.8% identification accuracy reported for *JADE*[25] under a matched evaluation protocol (**Supplementary Materials**).

Architecture-level ablations without elemental filtering further clarify the sources of improvement, as shown in **Fig. 2b**. A single-path variant that retains only the WAXRD branch with a KAN head (WA-KAN) achieves 41.4% top-1, 64.2% top-3, 72.3% top-5, and 80.4% top-10 accuracy. Introducing SAXRD systematically increases early-rank precision. The DEN-KAN surpasses WA-KAN at small *k*, while a mild trade-off appears at larger *k* (*k* = 10). Fusion steepens the similarity landscape and concentrates scores on a few strong candidates. Near-miss structures that remain above the cutoff under WA-KAN can drop below it after fusion. The net effect is higher precision at small *k* with a modest reduction in recall at large *k*. With elemental constraints enabled, the top-3 accuracy of DEN-KAN is already near-saturated and supports shortlist-based identification.

Under the same backbone and training configuration, replacing an MLP projection head with KAN consistently improves retrieval, highlighting the projection module as a key representational bottleneck. Keeping the dual-expert design but substituting KAN with a conventional MLP (DEN-MLP) yields the weakest performance across ranks. Notably, even without SAXRD, WA-KAN outperforms DEN-MLP by around 6% in top-1 accuracy, underscoring that the projection head can dominate retrieval quality in this setting. The observed performance gain is likely driven by the spline-based nonlinearities in KANs. By offering a richer functional landscape, these non-linearities enable the network to construct sharper, highly non-linear decision boundaries that are essential for disentangling the complex, structured interactions inherent in diffraction features[50,51]. Together with symmetric InfoNCE training, these properties yield stable gradients and enhance retrieval robustness.

Embedding-space similarity analysis supports these observations. We define self-similarity as the cosine similarity between each query pattern and its ground-truth structure, and other-similarity as similarity between mismatched XRD-structure pairs. As shown in **Fig. 2c**, both DEN-KAN and WA-KAN separate these distributions. However, DEN-KAN shifts the self-similarity peak upward and narrows its spread, while reducing the intrusion of the other-similarity tail into the self-similarity region. This tighter separation lowers the false-positives rate and explains the improved precision at small *k*. In contrast, WA-KAN exhibits broader self-similarity tails, consistent with slightly higher recall at large *k*. The cosine-similarity heat map for DEN-KAN in **Fig. 2d** shows strong diagonal dominance, confirming that each query aligns most strongly with its ground-truth structure, while off-diagonal values remain low.

Interpretability and physical priors of the DEN-KAN encoder

Elemental pre-screening substantially reduces the candidate pool and therefore confound analyses intended to attribute gains to diffraction information and model design. Unless stated otherwise, experiments and interpretability analyses in this section therefore exclude elemental filtering. We examine whether the SAXRD branch adds physically meaningful constraints beyond WAXRD, and

whether the KAN projection head promotes decisions anchored in Bragg reflections rather than incidental background variation.

Across 27060 test patterns, incorporating SAXRD improves retrieval systematically. The SAXRD-aware DEN-KAN increases top-1 accuracy from 41.45% (WA-KAN) to 46.42%, indicating that low-angle cues provide complementary constraints beyond WAXRD-only fingerprints. This benefit persists across unit-cell volume ranges, with the largest improvement for intermediate volumes (200-400 $Å^3$) and a reduced but still positive gain for very large cells (> 800 $Å^3$). The volume-stratified trend in **Table 1** is consistent with the role of SAXRD in capturing information tied to large *d*-spacings and long-period motifs[54]. Such features can be weakly expressed, overlapped, or ambiguous in WAXRD-only patterns, thereby adding SAXRD reduces effective ambiguity when multiple structures share similar high-angle fingerprints.

The symmetry-resolved analysis supports this interpretation. DEN-KAN outperforms WA-KAN across all seven crystal systems (**Fig. 3a**), with gains varying by symmetry. The largest improvement occurs for triclinic structures, whereas hexagonal and cubic systems show smaller but still substantial gains. This non-uniformity is expected because symmetry controls both lattice degrees of freedom and diffraction redundancy. In low-symmetry systems, the set of distinct candidate structures compatible with a given pattern is typically larger, and additional low-angle constraints can therefore stabilize early-rank discrimination more strongly. Even in high-symmetry systems, reflection overlap and systematic constraints can still produce near-degenerate WAXRD signatures[54,55]. SAXRD then acts as an additional, scale-sensitive fingerprint that separates otherwise confusable candidates.

Because unit-cell sizes differ across crystal systems in our test split, we further verify that symmetry-dependent gains cannot be explained by size effects alone. Mean unit-cell volumes vary markedly across systems. Triclinic, monoclinic and orthorhombic structures have large means of around 600 $Å^3$, whereas cubic structures have a much smaller mean of around 168 $Å^3$. Tetragonal and trigonal lie in between (~240 and 310 $Å^3$), and hexagonal remaining relatively small (~195 $Å^3$). However, the DEN-KAN advantage does not simply track these volume trends. Systems with similar mean volumes do not

necessarily show similar gains, and cubic structures still benefit appreciably from SAXRD-aware fusion. Coupled with the improvements under explicit volume-stratified results, this indicates that the cross-system gains are not only driven by unit-cell size, but also reflect how symmetry controls peak degeneracy and how much additional constraint low-angle diffraction can contribute.

A more stringent test then targets WAXRD-induced ambiguity directly. We construct a WAXRD-hard subset by computing, for each test sample, the maximum cosine similarity to the training set using only the WAXRD representation and selecting samples whose nearest-neighbor WAXRD similarity exceeds 0.9. These patterns lie in crowded regions of the high-angle similarity landscape, where WAXRD fingerprints are highly similar and thus less discriminative. On this subset, retrieval is performed without elemental filtering using the full test split as the candidate pool. The dual-branch model achieves 57.17% top-1 accuracy, compared with 51.00% for the single-branch model, yielding a larger-than-average improvement relative to the full test set. This provides direct evidence that when high-angle fingerprints alone cannot cleanly separate candidates, SAXRD supplies complementary cues that improve early-rank retrieval.

We then test whether the projection head bases retrieval decisions on physically meaningful diffraction features. A global, perturbation-based view of what angular regions drive the retrieval decisions was obtained using a sliding-window masking evaluation. We replace local 2θ segments with Gaussian noise and record whether the prediction changes, yielding a sensitivity heat map over 2θ. The KAN head exhibits concentrated sensitivity hotspots aligned with physically meaningful main-reflections, whereas the MLP head shows more diffuse sensitivity with appreciable response in background-dominated regions. **Fig. 4** indicates that KAN relies more directly on localized diffraction signatures, while the MLP is more susceptible to broad, low-frequency cues that need not correspond to specific Bragg reflections. Details of perturbation-based analysis and the following SHAP analysis are provided in the **Supplementary Materials**.

We further perform SHAP analysis[56,57] on four representative patterns spanning SAXRD and WAXRD regimes. For clarity, we visualize only the most salient attribution locations (top 5 SHAP-

ranked 2θ locations in SAXRD and top 15 in WAXRD). As shown in **Fig. 5**, the KAN head assigns high attributions sparsely and co-localizes them with genuine Bragg reflections, especially dominant peaks. For $Li_{12}Ca_{34}Hg_{18}$, KAN highlights the strongest low-angle reflection together with a small set of major WAXRD peaks, indicating reliance on high signal-to-noise ratio cues rather than diffuse background trends. For $Na_8Be_8H_8C_8O_{40}$, KAN again anchors attribution to the prominent SAXRD peak and the strongest mid-angle reflection, while placing limited weight on secondary features. Similar peak-anchored behavior is observed for $Rb_8Ti_4V_{12}P_{12}O_{72}$ and $Na_{18}La_{12}Y_{12}Si_9S_{27}O_{144}$, where KAN preferentially selects the tallest reflections and a small number of discriminative secondary peaks rather than distributing importance across baseline regions. In stark contrast, the MLP head exhibits a lower peak-occupancy with physical diffraction features. Many high-SHAP locations cluster in regions with near-zero intensity, suggesting sensitivity to broad baseline variations rather than crystalline signatures. Even when it captures strong peaks, the MLP simultaneously distributes substantial attribution to weak or off-peak regions, indicating a susceptibility to spurious background correlations.

These qualitative differences are consistent with the distinct functional forms of the two heads. In KAN, the projection is built from learnable spline functions with local support, enabling selective capacity allocation to narrow 2θ neighborhoods around diagnostic reflections. Once a localized region is informative, a spline can generate a steep, localized response while remaining nearly flat elsewhere, naturally yielding sparse, peak-anchored attributions. In contrast, MLP projections rely on dense linear mixing followed by global activations. Individual units therefore tend to aggregate information across broader input regions. Capturing narrow peak signatures may require coordinated behavior across many units, whereas modeling smooth low-frequency trends such as baseline curvature can be easier. Under contrastive training, such globally correlated cues can become shortcut features, leading MLP heads to over-weight background-dominated regions rather than anchoring decisions on reflections. These mean that KAN aligns more closely with a basic physical criterion of diffraction interpretation, prioritizing strong reflections and is therefore less prone to noise-driven shortcuts than an MLP. Overall, these findings indicate that the SAXRD branch adds physically meaningful constraints and that the KAN

projection head better respects localized diffraction signatures, improving both interpretability and retrieval reliability.

Application I: Space-group identification with XCCP

The space group encodes the discrete symmetry of a crystal and governs systematic absences, reflection multiplicities, and selection rules in powder diffraction. Because the XCCP framework is trained to align diffraction patterns with crystal structures, its ability to recover symmetry labels is a stringent test of physical fidelity and practical utility. We benchmark space-group identification on MP-SXRD against a representative spectrum of supervised classifiers, including a fully convolutional neural network (FCN)[38], a residual network with an MLP classification head (ResNet-MLP)[58], a vision transformer (ViT)[59], and a hybrid that combine ResNet and ViT (ResViT-MLP). KAN classification heads were also tested, replacing MLP heads in ResNet and ResViT to form ResNet-KAN and ResViT-KAN. All baselines are trained and evaluated on intensity-normalized XRD profiles using the same split and training protocol. For XCCP, the predicted label is taken as the space group of the top-1 retrieved structure. Unless noted otherwise, **Table 2** reports results averaged over multiple runs with different random seeds for the supervised baselines, and XCCP results are reported without elemental priors to ensure comparability.

Overall, stronger backbones based on residual learning and self-attention outperform plain convolutional stacks, and KAN heads provide additional gains over MLP heads in most cases. The run-to-run variability is modest, indicating that the ranking is reproducible rather than driven by favorable initialization. Most prior reports have emphasized the challenge of space-group classification using WAXRD-only inputs, with representative accuracies of 49.8% on SIMXRD-4M[34] and 37.48% on SIMPOD[60]. On MP-SXRD, performance improves from convolutional baselines to residual/attention-based models, while XCCP remains competitive with the strongest supervised predictors, consistent with

retrieval-guided prediction benefiting from structure-level alignment rather than relying solely on a direct classifier.

Including SAXRD yields consistent accuracy gains across architectures, and XCCP improves from 59.66% (WAXRD-only) to 60.85% (SAXRD-aware). The average gain from adding SAXRD is smaller here than in structure retrieval, which is expected given the different information bottlenecks. Retrieval must separate one specific structure from a large candidate pool, so even a few additional low-angle reflections can reduce ambiguity and reshuffle nearest-neighbor rankings. In contrast, space-group classification compresses each pattern into one of many discrete labels, and the discriminative evidence is dominated by systematic absences and intensity relationships across the WAXRD. The SAXRD typically contains fewer reflections and can be less diagnostic because many space groups share similar low-angle behavior. Consequently, SAXRD provides a consistent but modest improvement, likely concentrated in subsets where WAXRD fingerprints are least discriminative.

Application II: candidate-pool scaling and confidence-aware retrieval

We next evaluate XCCP under a full-database candidate pool and conduct a stratified analysis based on proximity of test samples to the training distribution. For each test pattern, we compute cosine similarities to all training patterns and take the maximum value as its nearest-neighbor similarity. The test set is then partitioned into low-, medium- and high-similarity subsets using fixed thresholds. We stratify by PXRD similarity because retrieval operates directly on diffraction signatures, and this proximity most directly reflects how well a query is represented by the learned embedding. A structure-level proximity check using *pymatgen* [61] confirms that structurally near-duplicate train-test pairs are rare under our tolerance setting, suggesting that stratification is not dominated by memorization. More details for similarity analysis please see the **Supplementary Materials**.

As shown in **Table 3**, with composition-based pre-screening enabled, the overall top-1 accuracy under the full-database pool is 72.62%, and the low/medium/high similarity subsets achieve

69.69%/73.00%/75.45%, respectively. The gaps are modest, and the low-similarity subset still maintains a high hit rate, indicating that performance is not primarily driven by near-duplicate test instances. Compared with the test-only candidate pool (top-1 = 88.98%, also with elemental pre-screening), the reduced accuracy under the full-database pool quantifies the increased difficulty induced by many more distractors. Candidate-pool scale is therefore a first-order driver of retrieval accuracy. Expanding the pool introduces more near-tie candidates and increases the chance that a confusable structure ranks above the ground-truth even when the model representation is unchanged. This also motivates elemental pre-screening in practice, as it can sharply shrink the candidate-pool and mitigate the difficulty introduced by large-scale distractors.

The second deployment challenge is open-set retrieval, where the correct structure is absent from the candidate database. Elemental pre-screening can obscure this failure mode by shrinking the pool and inflating apparent confidence. We therefore conduct an extreme missing-candidate stress test in the most challenging setting, no elemental priors and the ground-truth is entry explicitly removed from the candidate pool. To isolate the effect of missing ground-truth from ordinary retrieval errors, we restrict analysis to queries that are correctly retrieved in the nominal no-element setting under the test-only pool. This yields a paired comparison in which the same queries are evaluated with and without the correct candidate available.

Let the ranked cosine similarities be $s_1 \geq s_2 \geq \cdots$, where s_k is the similarity score of the k-th candidate. Intuitively, a smaller s_k indicates a weaker global match, increasing the risk of false acceptance. For deployment, we attach a confidence score to each returned candidate in the top-k list by mapping similarity through a calibrated sigmoid,

$$\boldsymbol{C}(s_k) = \frac{1}{1 + \exp\left(-\frac{s_k - \mu}{\beta}\right)} \tag{1}$$

We set $\mu = 0.85$ as the low-confidence threshold and choose $\beta = 0.025$ to obtain a sharp transition such that only very high similarities yield high confidence. In practice, because the presence of the ground-

truth structure is unknown, XCCP always returns the top-k candidates together with their confidence scores $\boldsymbol{C}(s_k)$. We additionally issue a discrete warning when $\boldsymbol{C}(s_1) < 0.5$ (equivalently, $s_1 < \mu$), indicating a potentially unreliable prediction and encouraging downstream refinement and expert inspection rather than hard acceptance of the top-1 assignment.

This confidence behavior is validated by comparing the normal condition to the missing-candidate test. When the ground truth is present, the mean top-1 similarity is 0.861 ± 0.07, while it decreases to 0.813 ± 0.08 after removing the correct structure. The corresponding mean top-1 confidence drops from 0.6133 to 0.3745, consistent with increased open-set risk. Two representative cases further illustrate this effect. For La_3AlN (Pm-3m), the top-1 confidence is 0.9726 when the correct structure is present but drops sharply to 0.072 when it is removed. For $Pr_2Cr_2Te_2O_{12}$ (C2/c), retrieval yields s_1= 0.9801 and $\boldsymbol{C}(s_1)$ = 0.9945 in the normal setting, whereas removing the ground-truth reduces confidence to 0.4889. This degradation confirms that the proposed confidence provides a retrieval-native indicator of open-set risk, enabling XCCP to caution users against over-interpreting a top-1 result while still returning a short list of plausible candidates for subsequent analysis.

Application III: Robustness from controlled simulations to experimental data

Practical PXRD retrieval must remain stable when composition cues are weak, peak shapes vary, and experimental artefacts break idealized simulation assumptions. We therefore evaluate robustness in three increasingly realistic settings, starting from controlled simulation-based tests and ending with a real experimental benchmark.

We first consider an elemental-similar case motivated by multi-principal element alloys (MPEAs), where many candidates share nearly identical elemental sets and composition priors become less informative. In such systems, subtle composition changes can induce local lattice distortions or symmetry-breaking that produce only slight peak shifts and modest intensity changes, making phase discrimination challenging[14]. FeCrAl-based and TaNbMo-based MPEA were selected as representative

systems, both adopting BCC structures. Crystal structures were generated using the special quasi-random structure (SQS) method[62], and supercells were constructed with the Alloy Theoretic Automated Toolkit (ATAT) software[63]. **Fig. 6** shows representative crystal structures and their simulated XRD patterns. For alloys such as $Fe_{20}Cr_{10}Al_2$ and $Fe_{20}Cr_9Al_3$, only slight peak shifts and intensity changes are observed. A similar effect is seen for $Ta_{14}Nb_{14}Mo_{14}V_{12}$ and $Ta_{15}Nb_8Mo_{15}V_{16}$, where patterns are almost identical although atomic arrangements differ. Despite these difficulties, the XCCP framework achieves 66.67% top-1 accuracy and 95.87% top-3 accuracy across 22 crystal structures (**Supplementary Materials**), indicating that the learned embedding can recover the correct structure when composition cues are intentionally weakened.

We next test robustness to two common experimental perturbations, including peak broadening and systematic 2θ offsets. We construct additional test sets by varying FWHM and by applying uniform shifts to the entire pattern. Evaluation is performed at test time without retraining. Unless stated otherwise, results correspond to DEN-KAN with elemental pre-screening enabled and a test-only candidate pool. As summarized in **Table 4**, accuracy degrades gradually as broadening increases, top-1 decreases from 88.98% at FWHM = 0.1^o to 79.67% at FWHM = 0.5^o. Importantly, even under the strongest broadening tested, the correct structure remains within the top-10 list for most queries (top-10 = 96.76%), and top-3/top-5 accuracies remain above 92%/94%, respectively, indicating that the representation retains strong shortlist-level utility under substantial peak-shape distortion. We further emulate calibration offsets by shifting the entire profile by $+0.1^o$ and $+0.5^o$ at fixed FWHM = 0.1^o. The resulting top-1 accuracies are 88.53% and 84.40%, respectively, with correspondingly high top-*k* metrics (top-10 ≥ 97.95%), suggesting moderate tolerance to global peak-position drift. Consistent with the ablations, DEN-KAN outperforms its MLP-head counterpart under these perturbations (**Supplementary Table S3**), suggesting that the KAN advantage persists under peak-shape variations.

Finally, we benchmark a zero-shot transfer on 773 experimental scans from a subset of the opXRD database[64]. These files typically lack reliable small-angle coverage, so we use the single-path WA-KAN variant to match the available 2θ range. To ensure a well-defined ground-truth structure for each query,

we restrict evaluation to samples that can be uniquely linked to a structure entry, yielding 773 paired experimental PXRD-structure instances. The candidate pool is restricted to the corresponding 773 structures, and elemental filtering is enabled to reflect standard laboratory practice. As summarized in **Table 5**, the XCCP framework demonstrates strong retrieval capabilities on the real-world dataset. The top-1 accuracy reaches 57.18%, and retrieval performance improves significantly as the retrieval scope expands. Top-3, top-5, and top-10 accuracies reach 84.73%, 93.66%, and 99.61%, respectively. The near-saturated top-10 accuracy indicates that the correct structure almost always appears within a short list, supporting efficient human-in-the-loop inspection and downstream refinement. The remaining gap to simulated performance is expected, our simulated training patterns are generated by applying Gaussian broadening to ideal reflections, whereas experimental scans often contain additional non-idealities that preferentially degrade early-rank precision. The strong high-k recall nonetheless suggests that the learned representation captures transferable crystallographic signal and that incorporating more realistic forward models or domain-adaptive augmentation could further narrow the simulation-to-experiment gap.

Discussion

PXRD retrieval is intrinsically confusable because distinct crystal structures can produce very similar diffraction profiles. This ambiguity is evident in our train-test proximity analyses on MP-SXRD. Although near-duplicates structures are rare, nearest-neighbor PXRD similarities can still be high, reflecting genuine degeneracy in diffraction fingerprints rather than data leakage. XCCP is designed to operate in this regime. By learning a physically grounded metric space in which diffraction evidence and structural hypotheses are embedded into a shared representation, XCCP enables rapid candidate ranking that can be inspected, refined, and validated, rather than forcing a brittle single-shot assignment.

Our results indicate that retrieval quality is driven not only by backbone capacity but also by whether the learned representation reflects how diffraction features arise and how ambiguity forms. Multi-scale encoding provides one such physically motivated prior. Low-angle information is sparse, yet it often

carries constraints tied to larger *d*-spacings and longer-period motifs that are underrepresented at high angles. This explains why adding SAXRD improves early-rank retrieval when WAXRD fingerprints are more degenerate. The smaller average gain for space-group identification is also informative. Space-group prediction compresses patterns into discrete symmetry labels and relies heavily on systematic absences and broad intensity relationships already expressed across WAXRD. By contrast, retrieval benefits directly from any additional constraint that removes distractors and reshapes neighbor ranking. The task dependence clarifies when SAXRD is most valuable, namely in hard cases where high-angle signatures are least discriminative.

A second influential prior lies in the projection head. Our ablations show that it can be a dominant bottleneck even with a strong feature extractor. The KAN head, with locally supported spline parameterization, encourages capacity allocation to narrow angular neighborhoods around diagnostic reflections and promotes peak-anchored decision rules. This attenuates sensitivity to broad background trends that vary across instruments and samples, consistent with interpretability results and observed robustness to peak broadening and modest shifts. More broadly, it suggests a design principle for scientific encoders. When evidence is localized in 2θ, the model should allocate capacity locally and in a physically aligned manner.

Two deployment realities shape retrieval performance. First, task difficulty is largely dictated by the size of the candidate pool. Larger databases contain more distractors with high similarity scores. This reduces the tolerance for error and raises the likelihood of mis-ranking, regardless of the embedding quality. Second, the correct structure may be absent from the database. A confident but incorrect top-1 answer is more harmful than exposing uncertainty in such open-set conditions. Similarity scores provide retrieval-native confidence indicators that degrade predictably when the ground-truth candidate is removed, enabling conservative interfaces that still return shortlists while warning against over-interpretation. This preserves the speed advantage of retrieval and matches practical workflows, where shortlists are refined and validated rather than accepted blindly.

Experimental evaluation highlights a boundary between what our current simulation engine captures and what remains domain-specific. High-*k* recall on experiment patterns suggests that the embedding transfers meaningful crystallographic signal, yet early-rank accuracy is lower than in simulation. This gap is expected because simulated profiles rely on ideal reflections with simple Gaussian broadening, whereas experimental scans include structured backgrounds and noise, sample-dependent peak widths, preferred orientation, impurity peaks and calibration drift. Narrowing this simulation-to-experiment gap is therefore an important direction. Incorporating more realistic forward models and domain-targeted augmentation should improve early-rank precision under such non-idealities. In addition, the present formulation implicitly assumes a single dominant structure for each query, while real samples can be multiphase with severe peak overlap and impurity reflections. Extending the retrieval interface from single-structure matching to mixture-aware shortlisting (e.g., returning sets of mutually compatible candidates and uncertainty-aware residuals) would better reflect practical phase-analysis workflows. Finally, while elemental pre-screening is highly effective, uncertain or noisy composition estimates could exclude the true structure. Coupling retrieval with chemistry-uncertainty modeling could allow soft composition constraints and confidence-aware fallback strategies.

Beyond robustness, XCCP can be extended to incorporate richer evidence, continual improvement, and broader downstream utility. Additional encoders could incorporate complementary measurements such as electron diffraction, real-space imaging features and spectroscopy-derived priors, aligned via multi-view contrastive learning to reduce degeneracy in hard cases. Incremental learning can continually update the contrastive library and the model as new structures and experimental scans accumulate, using rehearsal and confidence-driven curation to expand coverage without catastrophic forgetting. Downstream tasks could further leverage retrieval outputs by using shortlisted candidates and embeddings to initialize or guide refinement, support multiphase identification, and provide transferable descriptors for anomaly detection and property prediction. Overall, by combining multi-scale diffraction evidence with a physics-aligned contrastive metric space and confidence-aware retrieval, XCCP provides a practical and extensible foundation for large-scale, human-in-the-loop structure solution from PXRD.

Methods

Data Preparation

Our dataset contains 155,003 crystallographic information files (CIFs) from the Materials Project[65] database, spanning a wide range of crystal structures and chemical compositions. **Fig. 7a** summarizes the distribution across the seven crystal systems. **Fig. 7b** reports the ten most frequent space groups among 220 reported in the database. The histogram of space groups displays a pronounced long tail, which can bias classification toward majority space groups. This imbalance reflects the natural prevalence of crystalline symmetries and is therefore retained. No filtering or artificial rebalancing was applied so that class proportions remain representative of real materials problems.

PXRD patterns were simulated from CIFs using *PyXtal*[66], yielding 155,003 synthetic diffraction profiles. For each structure, crystallographic planes hkl were identified and their interplanar spacings calculated from reciprocal lattice geometry. Bragg's law is enforced to identify planes that satisfy the angular and wavelength conditions,

$$n\lambda = 2d_{hkl}sin\theta \tag{2}$$

where λ is the X-ray wavelength and θ is the diffraction angle. Throughout this work, we used Cu Kα radiation with λ = 1.54184 Å. For each eligible plane, the relative diffraction intensity I_{hkl} is evaluated from the structure factor F_{hkl} and the Lorentz-polarization corrections factor L_p,

$$I_{hkl} \propto |F_{hkl}|^2 \cdot L_p \tag{3}$$

with

$$F_{hkl} = \sum_j f_j exp\left[2\pi i\left(hx_j + ky_j + lz_j\right)\right] \tag{4a}$$

$$L_p = \frac{1 + cos^2 2\theta}{sin^2 \theta cos\,\theta} \tag{4b}$$

Here f_j is the atomic scattering factor and (x_j, y_j, z_j) are the atomic fractional coordinates. Each pattern was intensity-normalized by the maximal peak intensity before model training to ensure stable optimization. Peak broadening was modeled using a constant full width at half maximum (FWHM) of 0.1°.

Many simulated powder XRD studies[34,64] only consider the range $10^o \leq 2\theta \leq 80^o$. The present work follows this convention and refer to it as the WA profile. XRD patterns at $2\theta < 10^o$ were also simulated here, denoted as SA XRD. The statistical analysis (see **Supplementary Materials**) of simulated patterns shows that the dominate signal lies in the WA region, while non-negligible reflections remain in the SA range. These SA peaks correspond to large *d*-spacings from Bragg's law. They often report interlayer distances, superlattice ordering, guest induced expansion, and other long period motifs. Such simulation can improve discrimination among structures with similar high angle fingerprints.

The resulting dataset, termed MP-SXRD, contains 155,003 paired crystal structures and simulated diffraction patterns profiles with $0^o < 2\theta \leq 80^o$. It was randomly partitioned into training (70%), validation (10%), and test (20%) subsets. Stratified sampling preserved the distribution of space group across splits so that evaluation remains faithful to the long-tailed character of the source set.

XRD-Crystal Contrastive Pretraining and inference

The proposed framework aligns powder XRD patterns and crystal structures in a shared latent space, as illustrated in **Fig. 1**. Two modality specific encoders are trained jointly to produce comparable embeddings that support cross-modal retrieval and symmetry inference. The crystal encoder is a modified CGCNN that is widely used and easy to reproduce. Each structure is represented as a graph of atoms and bonds. Message passing aggregated local chemical environments, and a global pooling layer yields a 64-dimensional crystal embedding v_c.

Powder diffraction patterns are encoded with a dual-expert network incorporating a KAN projection head (DEN-KAN). The DEN-KAN encoder follows the SAXRD and WAXRD ranges defined before.

Two parallel branches process the two measurement ranges in tandem, and the KAN-based projection head fuses their outputs. Each branch adopts a ResNet backbone with residual connections and hierarchical feature extraction capabilities[58]. The two branches use the same residual blocks and ReLU activation functions. The WAXRD pathway applies max pooling with stride two and 1×3 kernels after each residual layer to enlarge the receptive field for closely spaced peaks. The outputs are concatenated and passed to the KAN-based projection head, producing a 64-dimensional XRD embedding v_{xrd}.

For settings where small-angle coverage is unavailable, a single-path variant that ingests only the conventional WAXRD range starting at$2\theta = 10^{o}$ was also developed. This variant keeps the WAXRD branch and the same KAN-based projection head (referred to as WA-KAN), with implementation details provided in **Supplementary Materials**. Both DEN-KAN and WA-KAN encoders are trained end to end with the same contrastive loss so that crystal and diffraction embeddings remain comparable.

Training maximizes the agreement and reduces the disagreement for XRD pattern-structure pairs with a symmetric InfoNCE loss $\mathcal{L}$,

$$\mathcal{L}(\theta_{\mathrm{xrd}}, \theta_{\mathrm{c}}) = \frac{1}{2}\left[\mathbb{E}_{(i,j)}\left[-\log \frac{e^{s(\mathbf{v}_{\mathrm{xrd}}^{i}, \mathbf{v}_{\mathrm{c}}^{j})/\tau}}{\sum_{k=1}^{N} e^{s(\mathbf{v}_{\mathrm{xrd}}^{i}, \mathbf{v}_{\mathrm{c}}^{k})/\tau}}\right] + \mathbb{E}_{(i,j)}\left[-\log \frac{e^{s(\mathbf{v}_{\mathrm{c}}^{j}, \mathbf{v}_{\mathrm{xrd}}^{i})/\tau}}{\sum_{k=1}^{N} e^{s(\mathbf{v}_{\mathrm{c}}^{j}, \mathbf{v}_{\mathrm{xrd}}^{k})/\tau}}\right]\right] \quad (5)$$

where $s(\cdot,\cdot)$ is cosine similarity, hyperparameter $\tau = 0.07$ is used here to appropriately control the similarity distribution sharpness, and $N = 128$ is the batch size. We employ a MultiStepLR learning rate schedule and batch normalization for stable training and regularization. Trainable parameters of both encoders θ_{xrd} and θ_{c} are optimized jointly.

Inference follows a two-stage pipeline. Candidate structures can first be filtered by chemical composition (elemental pre-screening). Cosine similarity between the query embedding v_{xrd} and the candidate structure embeddings v_{c} then yields a ranked list, and the model returns the top-k ($k = 1,2,\cdots$) candidates. Cross-modal retrieval performance is evaluated by top-k accuracy, defined as the probability that the correct structure appears among the first k candidates retrieved candidates. Based on similarity

scores within the returned top-*k* list, we further provide a confidence estimate for each query (Section 3.4).

Data Availability

The opXRD-related experimental dataset used in the present work can be found in https://huggingface.co/datasets/caobin/opxrd_hkust_expdata, and the CIFs for 22 MPEAs can be found in https://github.com/George-JieXIONG/Materials-Dataset/tree/main/XRD-Files

Acknowledgement

This work was financially supported by the Advanced Materials-National Science and Technology Major Project (Grant No. 2025ZD0620102), National Natural Science Foundation of China (Grant No. 52401015), Shanghai Pujiang Program (Grant No. 23PJ1403500) and Shanghai Artificial Intelligence Open Source Award Project. We also sincerely thank Mr. Bin Cao from The Hong Kong University of Science and Technology (Guangzhou) for his assistance in organizing and openly releasing the experimental database.

Author Contribution

C. Xu performed model training and drafted the manuscript. **T. Su** contributed to data preprocessing and statistical analysis, and revised the manuscript. **J. Xiong** conceived the research idea, performed data analysis, reviewed and revised the manuscript, and supervised the project. **Y. Wu** participated in the model design and training. **S. Chen**, **S. Dong**, and **T. Jiang** participated in the interpretation and discussion of the results. **M. He** analyzed the results, co-conceived the research idea, and revised the manuscript. **T.Y. Zhang** reviewed the manuscript, and provided overall supervision.

Competing Interests

The authors declare no competing financial or non-financial interests.

Reference

1 Altomare, A. *et al.* EXPO2009: structure solution by powder data in direct and reciprocal space. *Journal of Applied Crystallography* **42**, 1197-1202 (2009). https://doi.org/10.1107/s0021889809042915

2 Huang, T. C. Powder Diffraction: Theory and Practice—a new and comprehensive book on the methods and applications of powder diffraction. *Powder Diffraction* **24**, 2-3 (2012). https://doi.org/10.1154/1.3076390

3 Long, C. J., Bunker, D., Li, X., Karen, V. L. & Takeuchi, I. Rapid identification of structural phases in combinatorial thin-film libraries using x-ray diffraction and non-negative matrix factorization. *Review of Scientific Instruments* **80** (2009). https://doi.org/10.1063/1.3216809

4 Leonardi, A. & Bish, D. L. Interactions of lattice distortion fields in nano polycrystalline materials revealed by molecular dynamics and X-ray powder diffraction. *Acta Mater* **133**, 380-392 (2017). https://doi.org/10.1016/j.actamat.2017.04.072

5 Vishina, A., Eriksson, O. & Herper, H. C. Stable and metastable rare-earth-free permanent magnets from a database of predicted crystal structures. *Acta Mater* **261** (2023). https://doi.org/10.1016/j.actamat.2023.119348

6 Albesa-Jové, D. *et al.* Challenges in Direct-Space Structure Determination from Powder Diffraction Data: A Molecular Material with Four Independent Molecules in the Asymmetric Unit. *ChemPhysChem* **5**, 414-418 (2004). https://doi.org/10.1002/cphc.200301078

7 Stanev, V. *et al.* Unsupervised phase mapping of X-ray diffraction data by nonnegative matrix factorization integrated with custom clustering. *npj Computational Materials* **4** (2018). https://doi.org/10.1038/s41524-018-0099-2

8 Hua, Y. *et al.* Outstanding comprehensive piezoelectric properties in KNN-based ceramics via co-optimization of crystal structure and grain orientation. *Acta Mater* **297** (2025). https://doi.org/10.1016/j.actamat.2025.121378

9 McCusker, L. B., Von Dreele, R. B., Cox, D. E., Louër, D. & Scardi, P. Rietveld refinement guidelines. *Journal of Applied Crystallography* **32**, 36-50 (1999). https://doi.org/10.1107/S0021889898009856

10 Rietveld, H. M. The Rietveld Method: A Retrospection. *Zeitschrift für Kristallographie* **225**, 545-547 (2010). https://doi.org/10.1524/zkri.2010.1356

11 Marks, L. D. A standard format for reporting atomic positions in measured or calculated surface structures: The CIF file. *Surface Science* **604**, 878-881 (2010). https://doi.org/10.1016/j.susc.2010.02.019

12 Brown, I. D. CIF (Crystallographic Information File). A standard for crystallographic data interchange. *J Res Natl Inst Stan* **101**, 341-346 (1996). https://doi.org/10.6028/jres.101.035

13 Liu, Y. *et al.* Machine learning in materials genome initiative: A review. *Journal of Materials Science & Technology* **57**, 113-122 (2020). https://doi.org/10.1016/j.jmst.2020.01.067

14 Long, T. *et al.* Inverse design of crystal structures for multicomponent systems. *Acta Mater* **231** (2022). https://doi.org/10.1016/j.actamat.2022.117898

15 Guo, G. B. *et al.* Ab initio structure solutions from nanocrystalline powder diffraction data via diffusion models. *Nat Mater* (2025). https://doi.org/10.1038/s41563-025-02301-y

16 Agrawal, A. & Choudhary, A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Materials* **4** (2016). https://doi.org/10.1063/1.4946894

17 Pei, C. H. *et al.* A novel model to predict oxidation behavior of superalloys based on machine learning. *Journal of Materials Science & Technology* **235**, 232-243 (2025). https://doi.org/10.1016/j.jmst.2025.01.071

18 Su, T. H., Cao, B., Hu, S. B., Li, M. S. & Zhang, T. Y. CGWGAN: crystal generative framework based on Wyckoff generative adversarial network. *J Mater Inform* **4** (2024). https://doi.org/10.20517/jmi.2024.24

19 Cao, B. *et al.* XQueryer: an intelligent crystal structure identifier for powder X-ray diffraction. *Natl Sci Rev* **12** (2025). https://doi.org/10.1093/nsr/nwaf421

20 Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters* **120** (2018). https://doi.org/10.1103/PhysRevLett.120.145301

21 Liu, Z. *et al.* BraggNN: fast X-ray Bragg peak analysis using deep learning. *IUCrJ* **9**, 104-113 (2021). https://doi.org/10.1107/s2052252521011258

22 Zheng, X., Zhang, X., Chen, T. T. & Watanabe, I. Deep Learning in Mechanical Metamaterials: From Prediction and Generation to Inverse Design. *Advanced Materials* **35** (2023). https://doi.org/10.1002/adma.202302530

23 Schleder, G. R., Padilha, A. C. M., Acosta, C. M., Costa, M. & Fazzio, A. From DFT to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials* **2** (2019). https://doi.org/10.1088/2515-7639/ab084b

24 Choudhary, K. *et al.* Recent advances and applications of deep learning methods in materials science. *npj Computational Materials* **8** (2022). https://doi.org/10.1038/s41524-022-00734-6

25 Zhang, S. Y. *et al.* Crystallographic phase identifier of a convolutional self-attention neural network (CPICANN) on powder diffraction patterns. *IUCrJ* **11**, 634-642 (2024). https://doi.org/10.1107/S2052252524005323

26 Dong, H. *et al.* A deep convolutional neural network for real-time full profile analysis of big powder diffraction data. *npj Computational Materials* **7** (2021). https://doi.org/10.1038/s41524-021-00542-4

27 Lee, B. D. *et al.* A Deep Learning Approach to Powder X-Ray Diffraction Pattern Analysis: Addressing Generalizability and Perturbation Issues Simultaneously. *Advanced Intelligent Systems* **5** (2023). https://doi.org/10.1002/aisy.202300140

28 Maffettone, P. M. *et al.* Crystallography companion agent for high-throughput materials discovery. *Nat Comput Sci* **1**, 290-297 (2021). https://doi.org/10.1038/s43588-021-00059-2

29 Lee, J.-W., Park, W. B., Lee, J. H., Singh, S. P. & Sohn, K.-S. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns. *Nature Communications* **11** (2020). https://doi.org/10.1038/s41467-019-13749-3

30 Pan, T., Jin, S., Miller, M. D., Kyrillidis, A. & Phillips, G. N. A deep learning solution for crystallographic structure determination. *IUCrJ* **10**, 487-496 (2023). https://doi.org/10.1107/s2052252523004293

31 Szymanski, N. J. *et al.* Adaptively driven X-ray diffraction guided by machine learning for autonomous phase identification. *npj Computational Materials* **9** (2023). https://doi.org/10.1038/s41524-023-00984-y

32 Wang, H. *et al.* Rapid Identification of X-ray Diffraction Patterns Based on Very Limited Data by Interpretable Convolutional Neural Networks. *Journal of Chemical Information and Modeling* **60**, 2004-2011 (2020). https://doi.org/10.1021/acs.jcim.0c00020

33 Chen, L. *et al.* Crystal Structure Assignment for Unknown Compounds from X-ray Diffraction Patterns with Deep Learning. *Journal of the American Chemical Society* **146**, 8098-8109 (2024). https://doi.org/10.1021/jacs.3c11852

34 Bin, C. *et al.* in *The Thirteenth International Conference on Learning Representations* (2025).

35 Salgado, J. E., Lerman, S., Du, Z., Xu, C. & Abdolrahim, N. Automated classification of big X-ray diffraction data using deep learning models. *npj Computational Materials* **9** (2023). https://doi.org/10.1038/s41524-023-01164-8

36 Park, W. B. *et al.* Classification of crystal structure using a convolutional neural network. *IUCrJ* **4**, 486-494 (2017). https://doi.org/10.1107/s205225251700714x

37 Vecsei, P. M., Choo, K., Chang, J. & Neupert, T. Neural network based classification of crystal symmetries from x-ray diffraction patterns. *Physical Review B* **99** (2019). https://doi.org/10.1103/PhysRevB.99.245120

38 Lee, B. D. *et al.* Powder X-Ray Diffraction Pattern Is All You Need for Machine-Learning-Based Symmetry Identification and Property Prediction. *Advanced Intelligent Systems* **4** (2022). https://doi.org/10.1002/aisy.202200042

39 Xie, T. *et al.* CSU-MS2: A Contrastive Learning Framework for Cross-Modal Compound Identification from MS/MS Spectra to Molecular Structures. *Analytical Chemistry* **97**, 13350-13360 (2025). https://doi.org/10.1021/acs.analchem.5c01594

40 Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat Comput Sci* **2**, 718-+ (2022). https://doi.org/10.1038/s43588-022-00349-3

41 Choudhary, K. & DeCost, B. Atomistic Line Graph Neural Network for improved materials property predictions. *Npj Computational Materials* **7** (2021). https://doi.org/10.1038/s41524-021-00650-1

42 Deng, B. W. *et al.* CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat Mach Intell* **5**, 1031-1041 (2023). https://doi.org/10.1038/s42256-023-00716-3

43 Zeni, C. *et al.* A generative model for inorganic materials design. *Nature* **639** (2025). https://doi.org/10.1038/s41586-025-08628-5

44 Lai, Q. S. *et al.* End-to-End Crystal Structure Prediction from Powder X-Ray Diffraction. *Advanced Science* **12** (2025). https://doi.org/10.1002/advs.202410722

45 Zhang, X., Wang, X. T. & Hu, S. B. Structural Fingerprinting of Crystalline Materials from XRD Patterns Using Atomic Cluster Expansion Neural Network and Atomic Cluster Expansion. *Appl Sci-Basel* **15** (2025). https://doi.org/10.3390/app15115851

46 Li, Q. *et al.* Powder diffraction crystal structure determination using generative models. *Nature Communications* **16** (2025). https://doi.org/10.1038/s41467-025-62708-8

47 Jia, Z., Li, Z., Wang, Z., Liu, Z. & Vong, C.-M. Joint Prediction of SOH and RUL for Lithium Batteries Considering Capacity Self Recovery and Model Drift. *IEEE Internet of Things Journal* **12**, 22187-22196 (2025). https://doi.org/10.1109/jiot.2025.3550270

48 Bandyopadhyay, Y., Avlani, H. & Zhuang, H. L. Kolmogorov–Arnold neural networks for high-entropy alloys design. *Modelling and Simulation in Materials Science and Engineering* **33** (2025). https://doi.org/10.1088/1361-651X/adbb83

49 Wu, X. Y., Song, X. Y., Yue, Y. F., Zheng, R. & Jiang, J. W. MOF-KAN: Kolmogorov-Arnold Networks for Digital Discovery of Metal-Organic Frameworks. *J Phys Chem Lett* **16**, 2452-2459 (2025). https://doi.org/10.1021/acs.jpclett.5c00211

50 Ziming, L. *et al.* in *The Thirteenth International Conference on Learning Representations* (2025).

51 Thakolkaran, P. *et al.* Can KAN CANs? Input-convex Kolmogorov-Arnold Networks (KANs) as hyperelastic constitutive artificial neural networks (CANs). *Computer Methods in Applied Mechanics and Engineering* **443** (2025). https://doi.org/10.1016/j.cma.2025.118089

52 Wang, Y. *et al.* Kolmogorov–Arnold-Informed neural network: A physics-informed deep learning framework for solving forward and inverse problems based on Kolmogorov–Arnold Networks. *Computer Methods in Applied Mechanics and Engineering* **433** (2025). https://doi.org/10.1016/j.cma.2024.117518

53 Wu, Y. *et al.* Kolmogorov-Arnold Network Made Learning Physics Laws Simple. *J Phys Chem Lett* **15**, 12393-12400 (2024). https://doi.org/10.1021/acs.jpclett.4c02589

54 Graewert, M. A. & Svergun, D. I. Impact and progress in small and wide angle X-ray scattering (SAXS and WAXS). *Current Opinion in Structural Biology* **23**, 748-754 (2013). https://doi.org/10.1016/j.sbi.2013.06.007

55 Ge, J. *et al.* Correlations of multiscale structural evolution and homogeneous flows in metallic glass ribbons. *Materials Research Letters* **11**, 547-555 (2023). https://doi.org/10.1080/21663831.2023.2187264

56 Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**, 56-67 (2020). https://doi.org/10.1038/s42256-019-0138-9

57 Sundararajan, M., Taly, A. & Yan, Q. Q. Axiomatic Attribution for Deep Networks. *Pr Mach Learn Res* **70** (2017).

58 He, K., Zhang, X., Ren, S. & Sun, J. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770-778 (2016).

59 Alexey, D. *et al.* in *International Conference on Learning Representations* (2021).

60 Rincón, S., González, G., Macías, M. A. & Arbeláez, P. A new benchmark for machine learning applied to powder X-ray diffraction. *Scientific Data* **12** (2025). https://doi.org/10.1038/s41597-025-05534-3

61 Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68**, 314-319 (2013). https://doi.org/10.1016/j.commatsci.2012.10.028

62 Zunger, A., Wei, S. H., Ferreira, L. G. & Bernard, J. E. Special quasirandom structures. *Physical Review Letters* **65**, 353-356 (1990). https://doi.org/10.1103/PhysRevLett.65.353

63 van de Walle, A., Asta, M. & Ceder, G. The Alloy Theoretic Automated Toolkit: A user guide. *Calphad-Computer Coupling of Phase Diagrams and Thermochemistry* **26**, 539-553 (2002). https://doi.org/10.1016/S0364-5916(02)80006-2

64 Hollarek, D. *et al.* opXRD: Open Experimental Powder X-Ray Diffraction Database. *Advanced Intelligent Discovery* (2025). https://doi.org/10.1002/aidi.202500044

65 Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1** (2013). https://doi.org/10.1063/1.4812323

66 Fredericks, S., Parrish, K., Sayre, D. & Zhu, Q. PyXtal: A Python library for crystal structure generation and symmetry analysis. *Computer Physics Communications* **261** (2021). https://doi.org/10.1016/j.cpc.2020.107810

Figures and Tables

Fig. 1. **XCCP framework for crystal–XRD alignment and retrieval.** Contrastive training aligns powder XRD patterns with their crystal structures in a shared embedding space. A CGCNN-based crystal encoder produces structure embeddings. The dual-

expert XRD encoder (DEN-KAN) encodes SAXRD and WAXRD inputs using two branches, and a KAN head fuses branch features to produce the XRD embedding. When SA data are unavailable, only the WA branch alone is used (WA-KAN).

Fig. 2. Retrieval performance and embedding-space similarity. (a) Top-*k* retrieval accuracy of the XCCP framework with and without elemental filtering. (b) Ablations across different XRD encoders. (c) Distributions of self-similarity (matched pairs) and other-similarity (mismatched pairs). (d) Representative cosine-similarity heat map. All accuracy values refer to structure-retrieval hit rates.

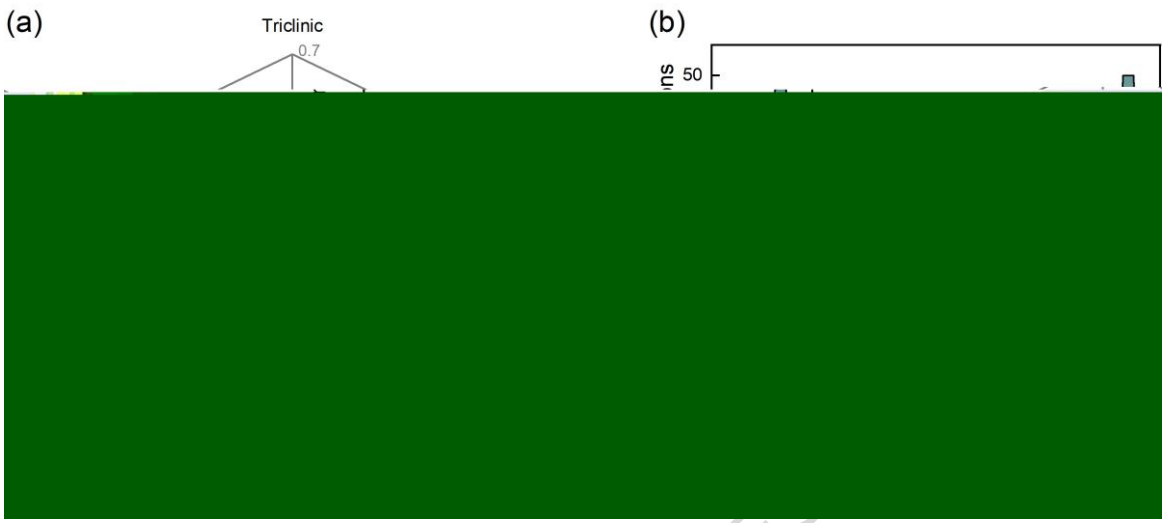


Fig. 3. SAXRD-driven gains by symmetry. (a) Accuracy gain of DEN-KAN encoder over WA-KAN encoder across the seven crystal systems. (b) Schematic of symmetry-operations richness across crystal systems.

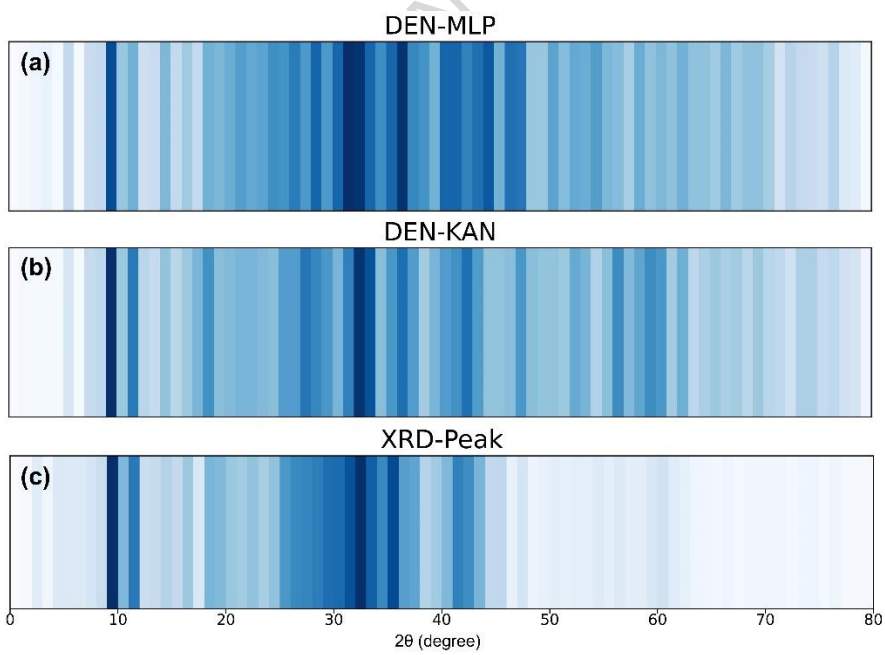


Fig. 4. Masking-based sensitivity maps. Sliding-window masking sensitivity for (a) the DEN-MLP and (b) the DEN–KAN encoder. (c) Reference peak profile for aligning sensitivity hotspots with reflections.

Fig. 5. **SHAP attributions reveal peak-anchored decision rules**. Representative PXRD patterns for (a) $Li_{12}Ca_{34}Hg_{18}$, (b) $Na_8Be_8H_8C_8O_{40}$, (c) $Rb_8Ti_4V_{12}P_{12}O_{72}$ and (d) $Na_{18}La_{12}Y_{12}Si_9S_{27}O_{144}$. Blue circles and orange triangles mark the top-ranked SHAP locations for the MLP head and the KAN head. The dashed line denotes the SAXRD/WAXRD boundary. For visualization, we highlight the top 5 attribution bins in the SAXRD region and the top 15 attribution bins in the WAXRD region for each pattern.

Fig. 6. Representative crystal structures and simulated patterns for (a) FeCrAl-based and (b) TaNbMo-based MPEAs.

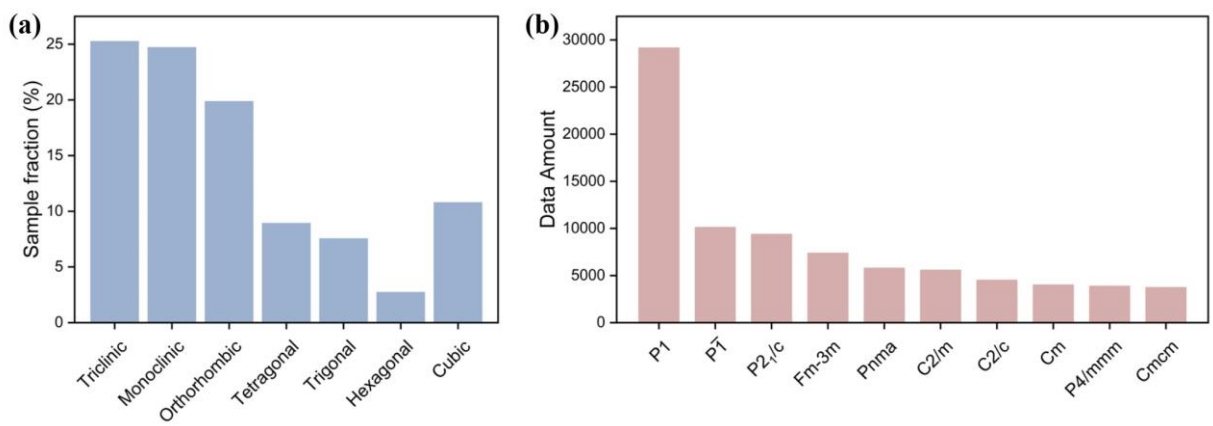


Fig. 7. **Crystallographic data analysis.** (a) Statistical distribution of the seven crystal systems present in the dataset. (b) Frequency distribution of the ten most prevalent space groups among 220 labels.

Table 1. Structure-retrieval top-1 hit rates grouped by cell-volume ranges.

Data	Accuracy (SAXRD-aware)	Accuracy (WAXRD-only)	Data Count
All Test Data	46.42%	41.45%	27060
<100 Å³	55.54%	50.45%	2989
100-200 Å³	52.57%	47.64%	5199
200-400 Å³	48.23%	41.19%	7825
400-800 Å³	40.61%	35.85%	6203
>800 Å³	38.81%	36.83%	4844

Table 2. Space-group prediction accuracy (no elemental information)

Model	Accuracy (SAXRD-aware)	Accuracy (WAXRD-only)
FCN	48.23±0.22%	47.62±0.42%
ViT	48.16±1.12%	48.02±0.64%
ResNet-MLP	58.08±0.26%	57.72±0.11%
ResViT-MLP	59.06±0.44%	58.11±0.55%
ResNet-KAN	58.95±0.32%	57.93±0.28%
ResViT-KAN	59.22±0.42%	59.05±0.37%
XCCP	**60.85%**	**59.66%**

Table 3. Top-1 retrieval accuracy under a full-database candidate pool (with elemental pre-screening), stratified by PXRD-similarity

All test data	Low-similarity	Medium-similarity	High-similarity
72.62%	69.69%	73.00%	75.45%

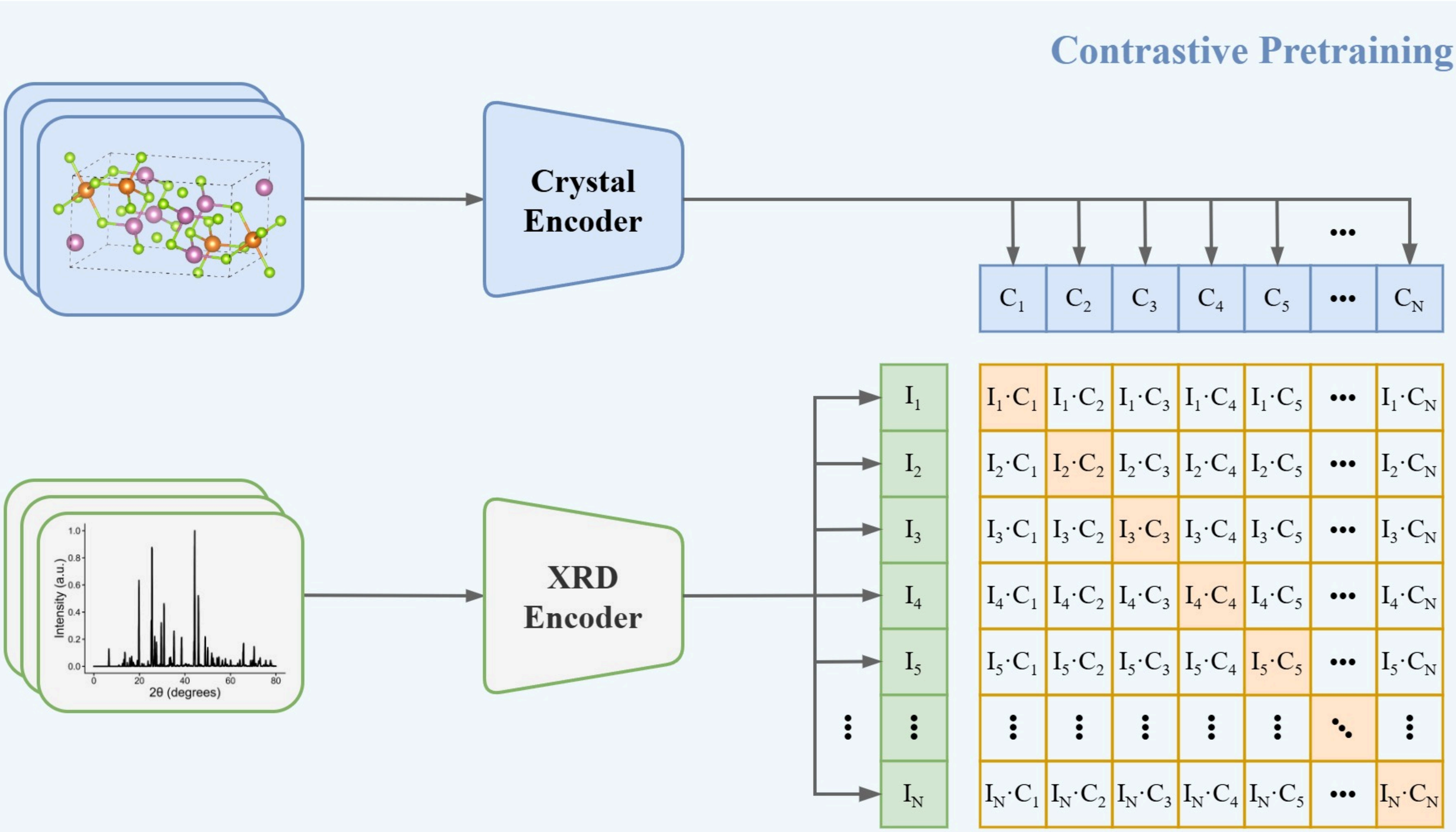
Table 4. Top-*k* accuracy under simulated perturbations (with elemental information).

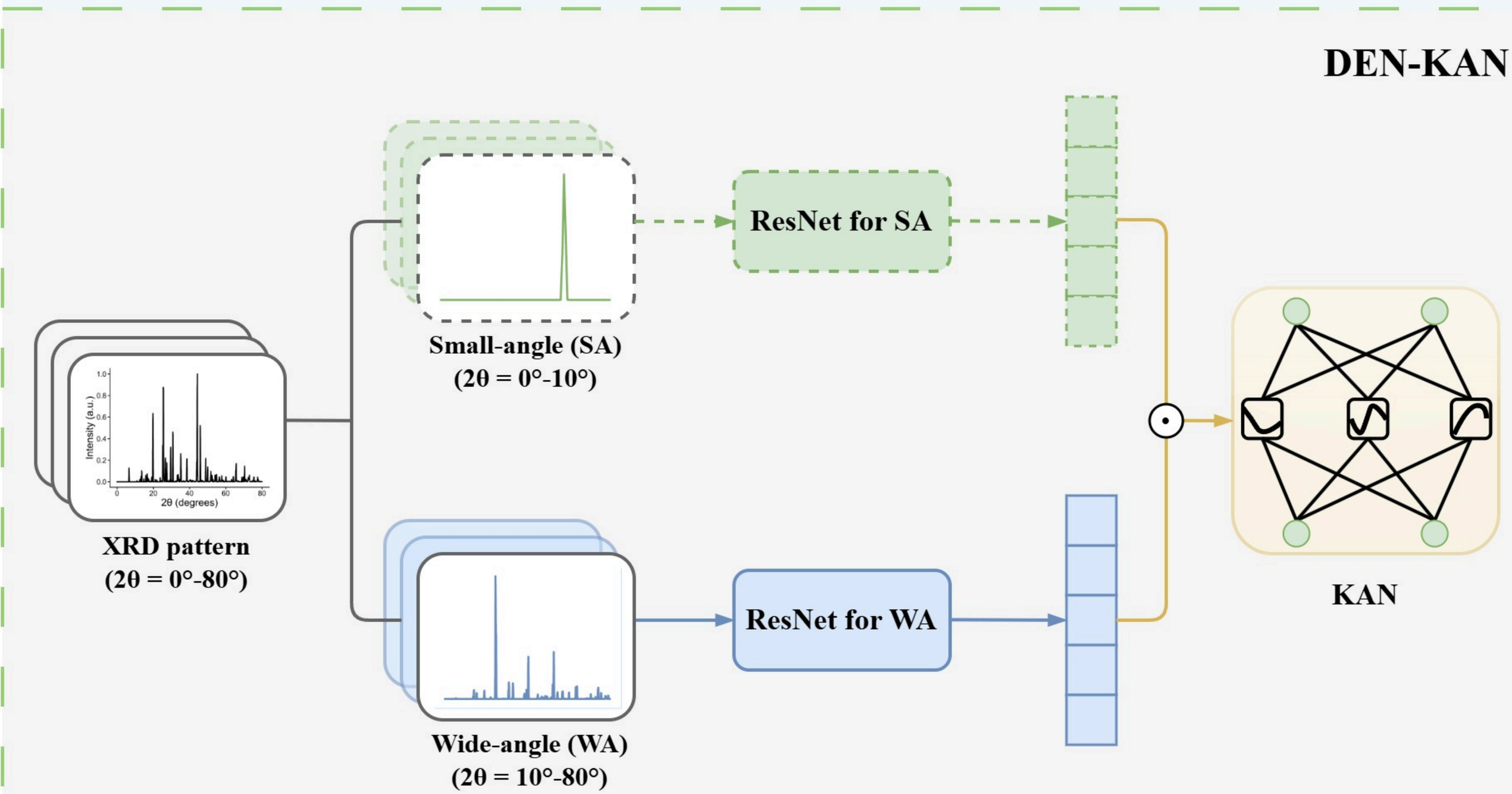
Test Data	Top-1	Top-3	Top-5	Top-10
0.1° FWHM, 0 ° shift	88.98%	97.56%	98.82%	99.66%
0.2° FWHM, 0 ° shift	85.70%	95.93%	97.68%	99.01%
0.3° FWHM, 0 ° shift	83.30%	94.46%	96.47%	98.15%
0.5° FWHM, 0 ° shift	79.67%	92.31%	94.68%	96.76%
0.1° FWHM, + 0.1 ° shift	88.53%	97.25%	98.62%	99.45%
0.1° FWHM, + 0.5 ° shift	84.40%	94.71%	96.45%	97.95%

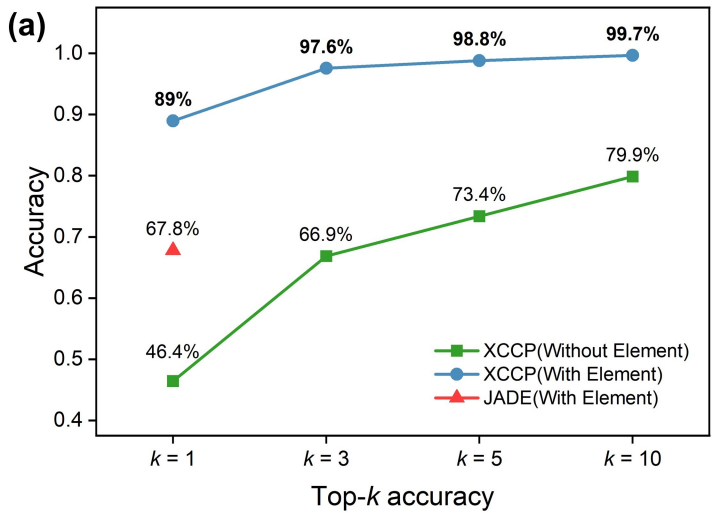
Table 5. Top-*k* retrieval accuracy on experimental data (with elemental information), stratified by crystal system.

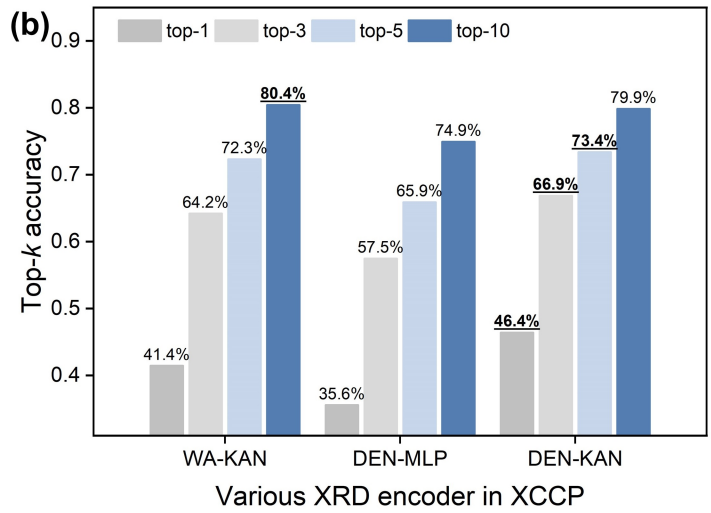
Crystal system	top-1	top-3	top-5	top-10	Data Count
Triclinic	60.27%	85.48%	94.16%	99.67%	599
Monoclinic	46.81%	81.91%	90.43%	98.94%	94

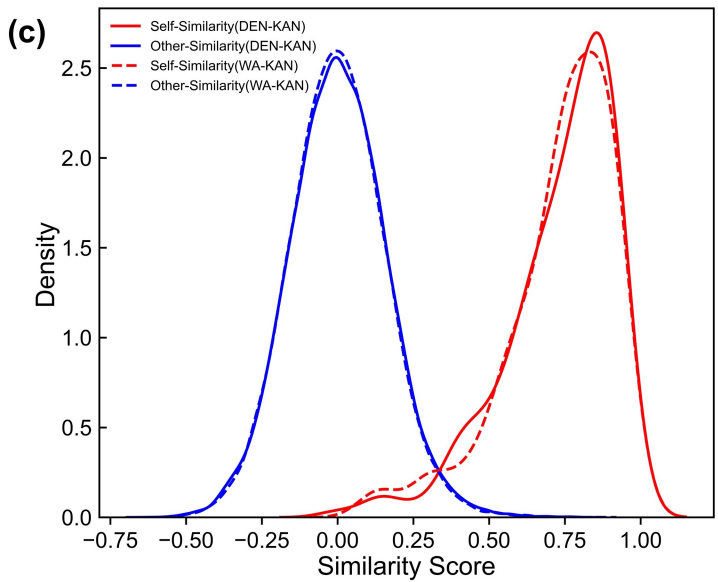
Orthorhombic	26.32%	84.21%	89.47%	100.00%	19
Tetragonal	33.33%	50.00%	66.67%	100.00%	6
Trigonal	45.95%	78.38%	97.30%	100.00%	37
Hexagonal	58.33%	100.00%	100.00%	100.00%	12
Cubic	100.00%	100.00%	100.00%	100.00%	6
Average	**57.18%**	**84.73%**	**93.66%**	**99.61%**	**773**

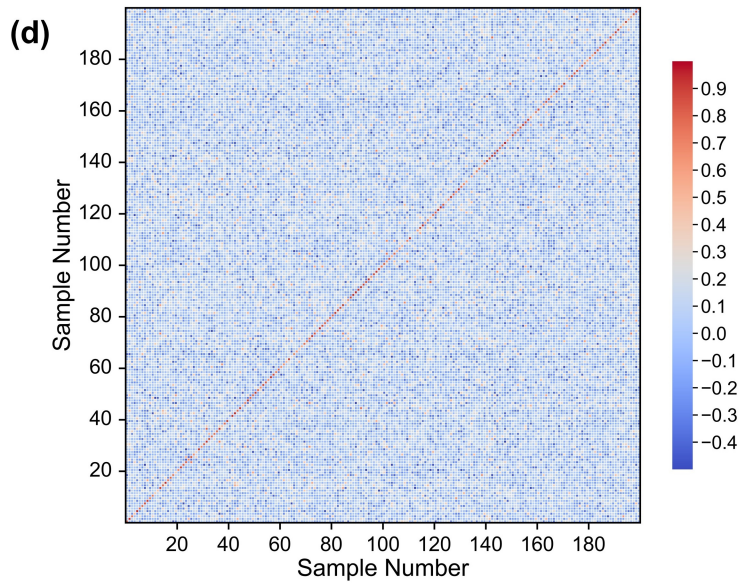

Contrastive Pretraining
Crystal Encoder
XRD Encoder
Intensity (a.u.)
2θ (degrees)
C1 C2 C3 C4 C5 ••• CN
I1 I2 I3 I4 I5 ⋮ IN
I1·C1 I1·C2 I1·C3 I1·C4 I1·C5 ••• I1·CN
I2·C1 I2·C2 I2·C3 I2·C4 I2·C5 ••• I2·CN
I3·C1 I3·C2 I3·C3 I3·C4 I3·C5 ••• I3·CN
I4·C1 I4·C2 I4·C3 I4·C4 I4·C5 ••• I4·CN
I5·C1 I5·C2 I5·C3 I5·C4 I5·C5 ••• I5·CN
IN·C1 IN·C2 IN·C3 IN·C4 IN·C5 ••• IN·CN

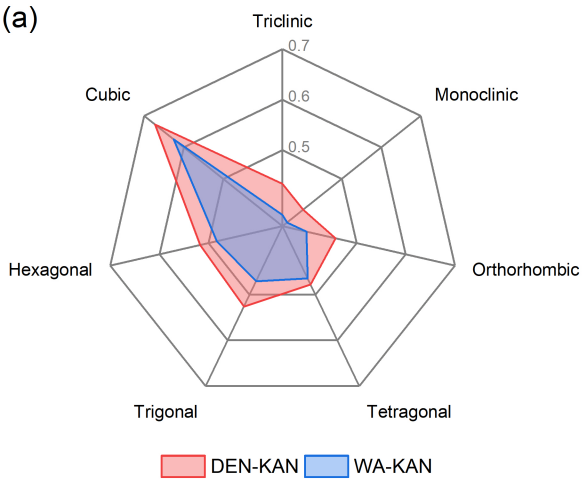

DEN-KAN
ResNet for SA
Small-angle (SA)
(2θ = 0°-10°)
XRD pattern
(2θ = 0°-80°)
ResNet for WA
Wide-angle (WA)
(2θ = 10°-80°)
KAN

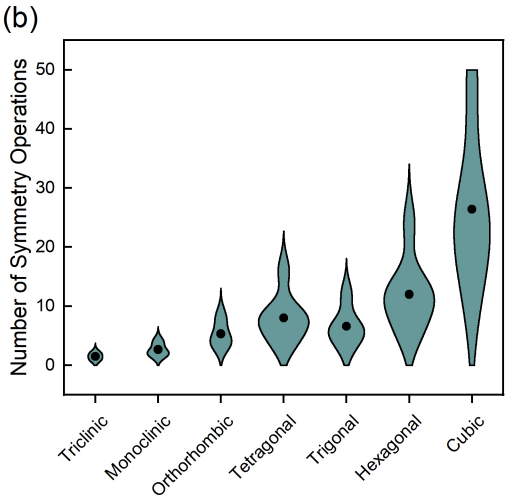

(a)
Accuracy
Top-k accuracy
k = 1
k = 3
k = 5
k = 10
89%
97.6%
98.8%
99.7%
67.8%
66.9%
73.4%
79.9%
46.4%
XCCP(Without Element)
XCCP(With Element)
JADE(With Element)

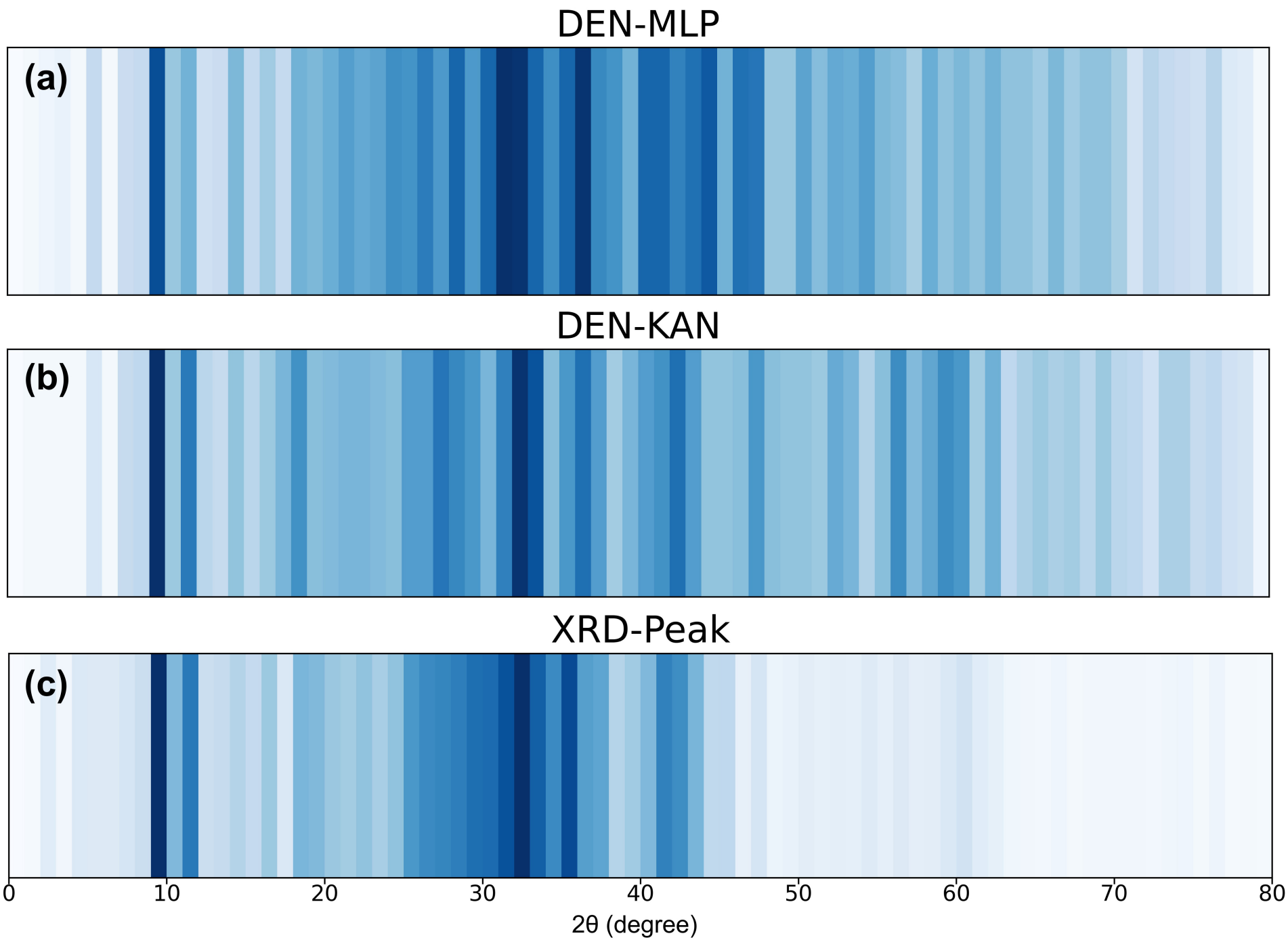

(b)
top-1
top-3
top-5
top-10
Top-k accuracy
41.4%
64.2%
72.3%
80.4%
35.6%
57.5%
65.9%
74.9%
46.4%
66.9%
73.4%
79.9%
WA-KAN
DEN-MLP
DEN-KAN
Various XRD encoder in XCCP

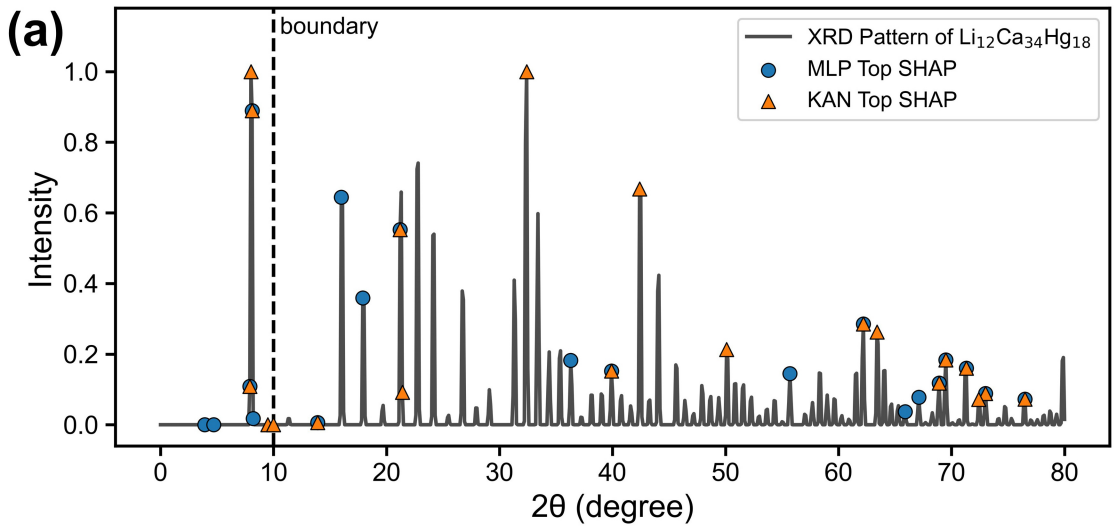

(c)
Self-Similarity(DEN-KAN)
Other-Similarity(DEN-KAN)
Self-Similarity(WA-KAN)
Other-Similarity(WA-KAN)
Density
Similarity Score

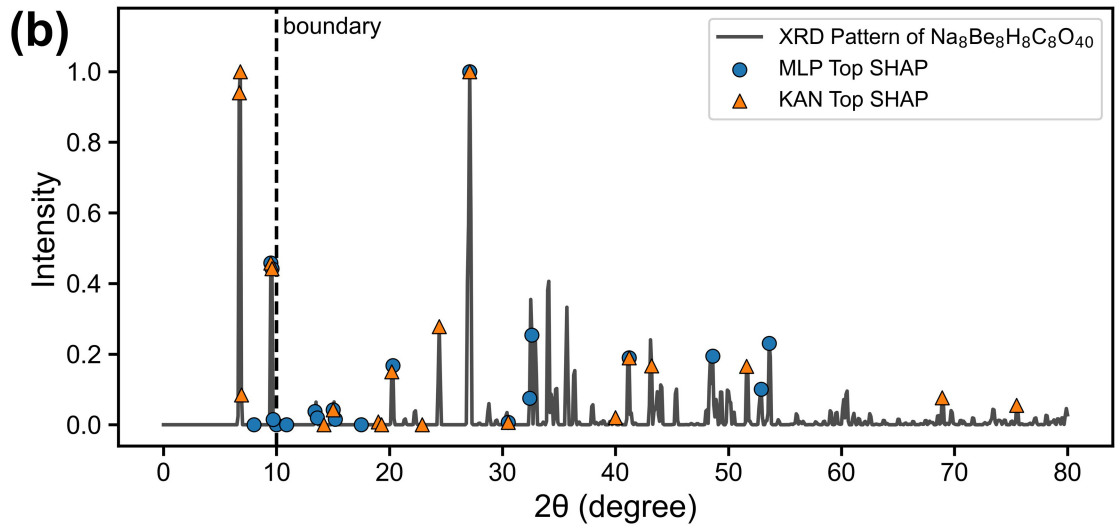

(d)
Sample Number
Sample Number

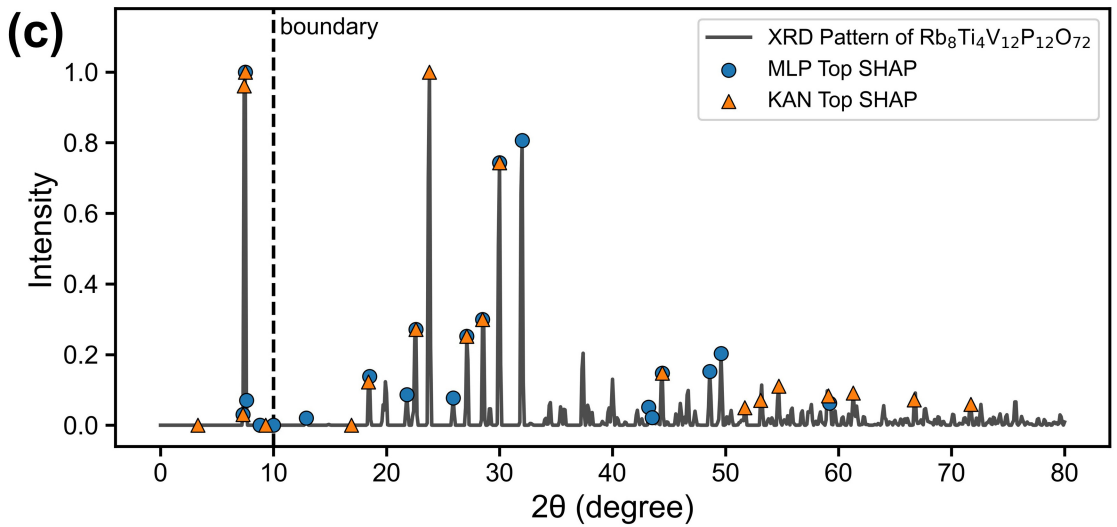


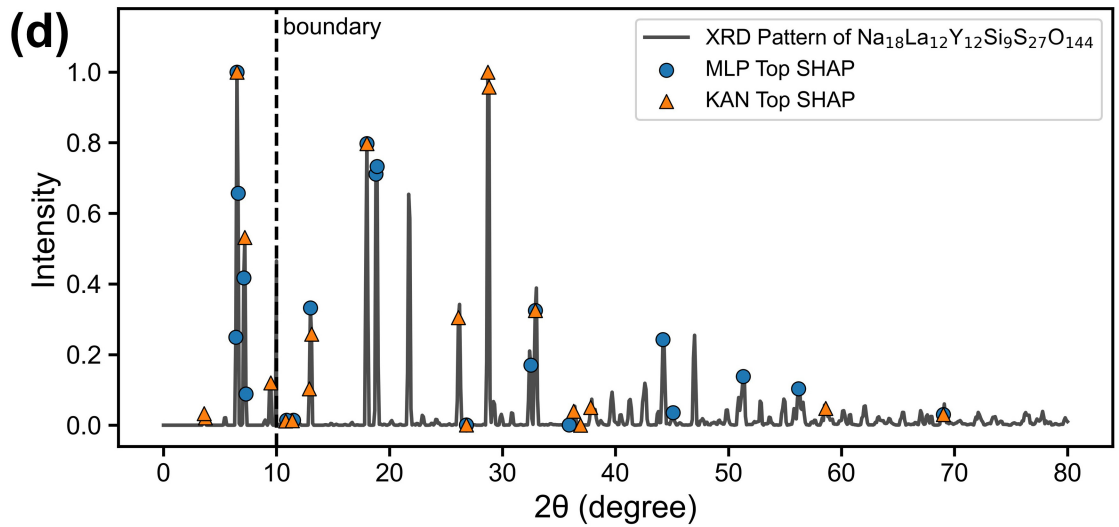


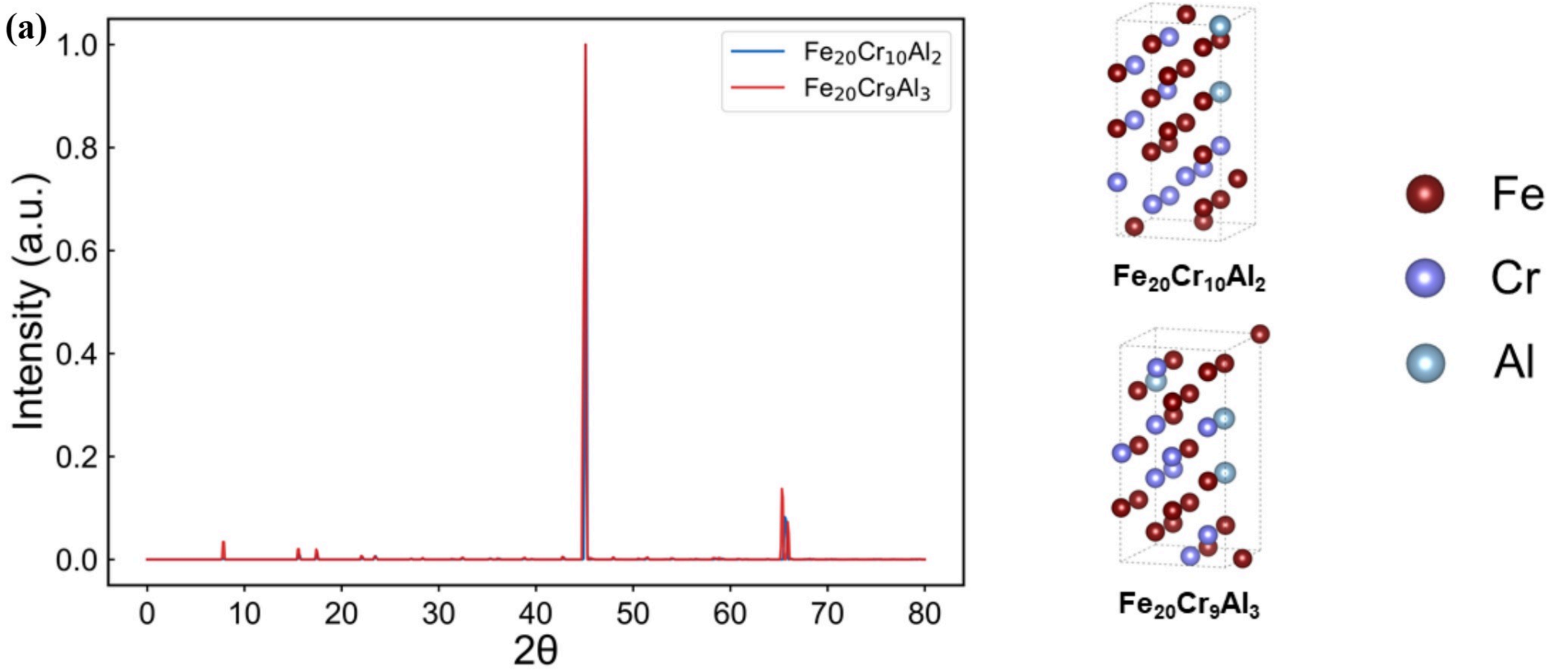

DEN-MLP
(a)
DEN-KAN
(b)
XRD-Peak
(c)
0
10
20
30
40
50
60
70
80
2θ (degree)

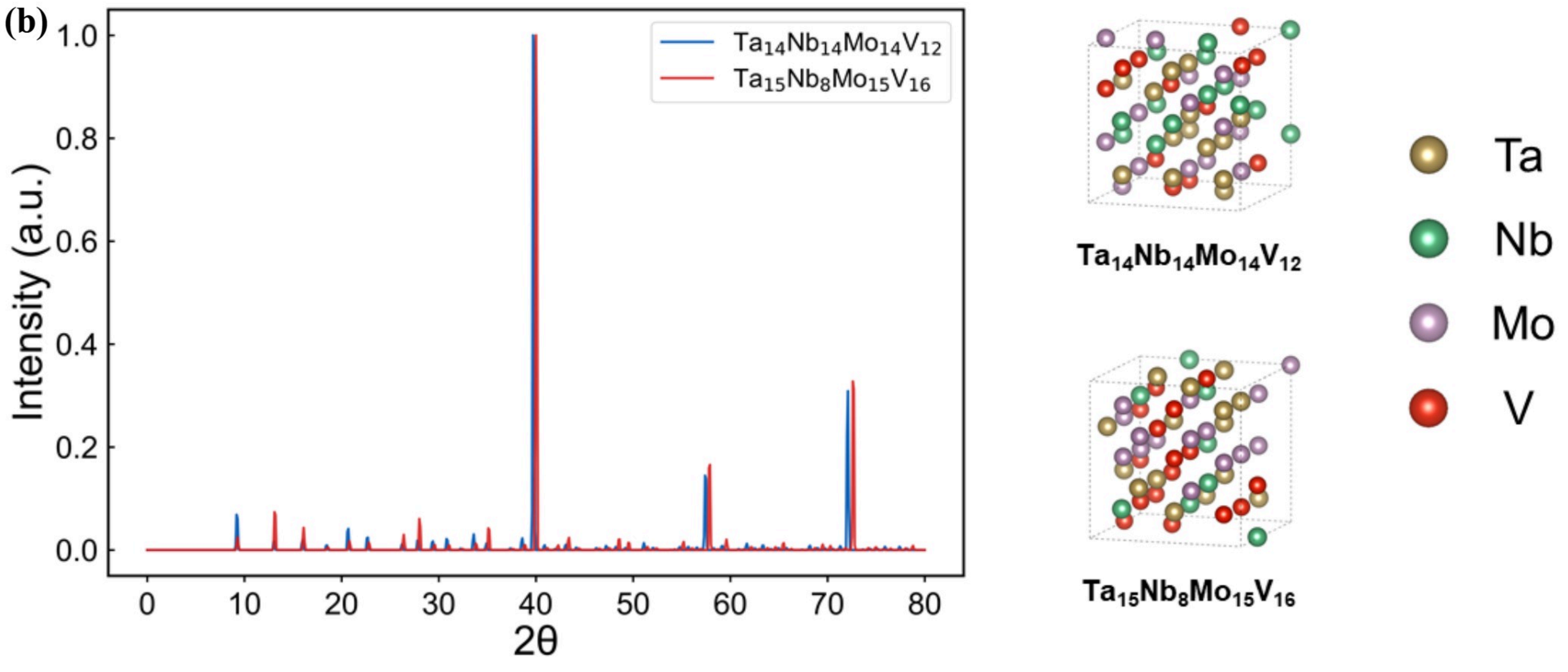

(a)
boundary
XRD Pattern of $Li_{12}Ca_{34}Hg_{18}$
MLP Top SHAP
KAN Top SHAP
Intensity
2θ (degree)

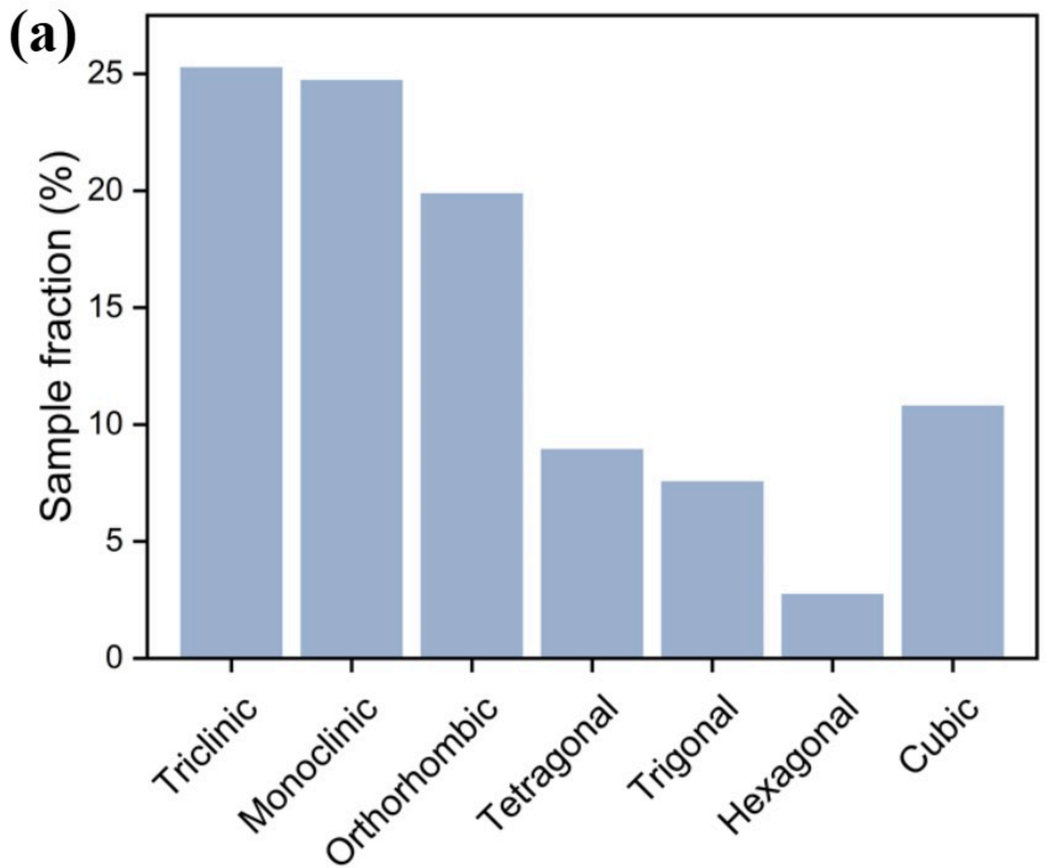

(b)
boundary
XRD Pattern of $Na_8Be_8H_8C_8O_{40}$
MLP Top SHAP
KAN Top SHAP
Intensity
2θ (degree)

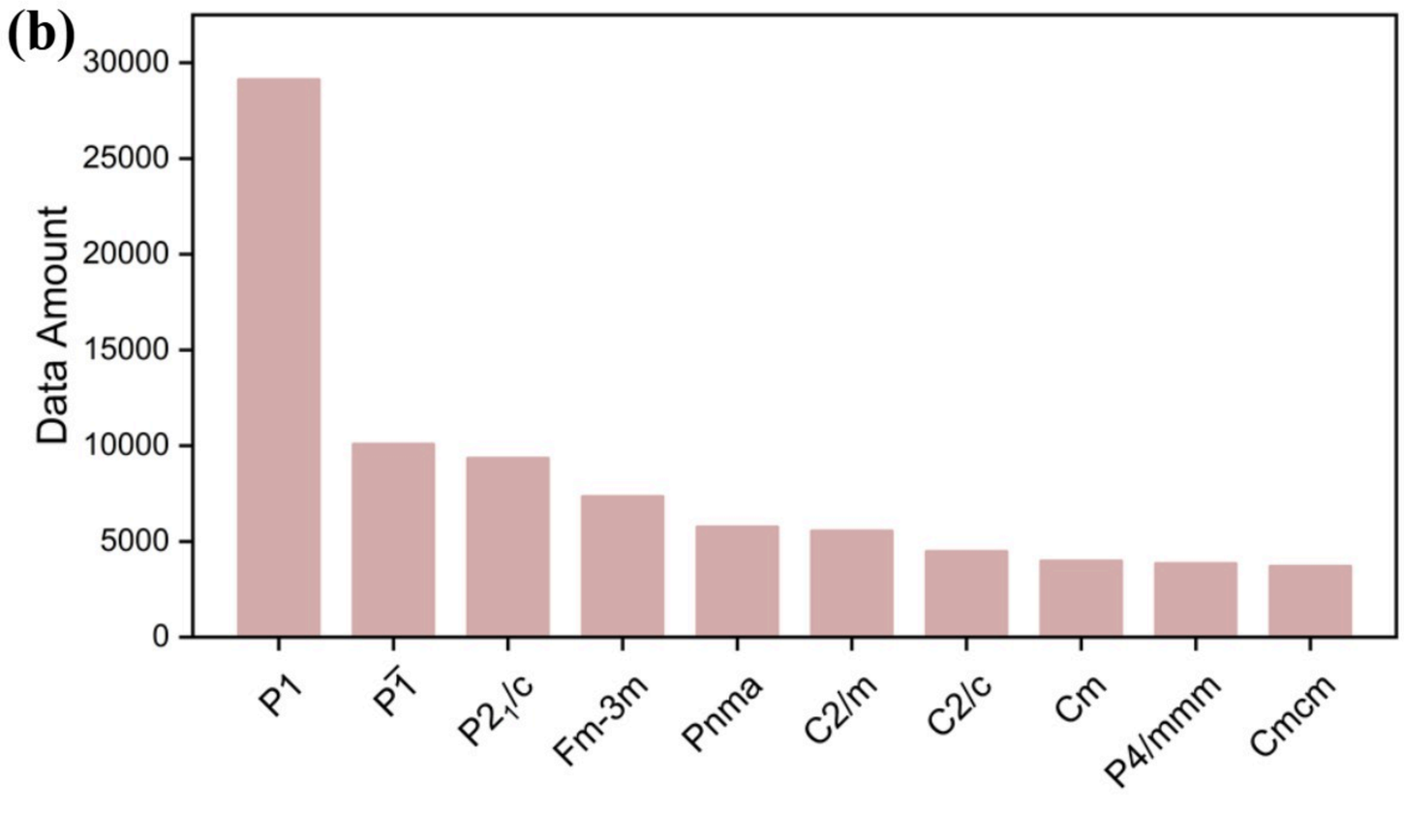

(c)
boundary
XRD Pattern of $Rb_8Ti_4V_{12}P_{12}O_{72}$
MLP Top SHAP
KAN Top SHAP
Intensity
2θ (degree)


(d)
boundary
XRD Pattern of $Na_{18}La_{12}Y_{12}Si_9S_{27}O_{144}$
MLP Top SHAP
KAN Top SHAP
Intensity
2θ (degree)


(a)
Intensity (a.u.)
2θ
Fe20Cr10Al2
Fe20Cr9Al3
Fe
Cr
Al


(b)
Intensity (a.u.)
2θ
Ta14Nb14Mo14V12
Ta15Nb8Mo15V16
Ta
Nb
Mo
V


(a)
Sample fraction (%)
0
5
10
15
20
25
Triclinic
Monoclinic
Orthorhombic
Tetragonal
Trigonal
Hexagonal
Cubic


(b)
Data Amount
0
5000
10000
15000
20000
25000
30000
P1
P$\bar{1}$
P2$_1$/c
Fm-3m
Pnma
C2/m
C2/c
Cm
P4/mmm
Cmcm