

# deCIFer: Crystal Structure Prediction from Powder Diffraction Data using Autoregressive Language Models

Frederik Lizak Johansen<sup>†</sup>, Ulrik Friis-Jensen<sup>†</sup>, Erik Bjørnager Dam<sup>‡</sup>,  
Kirsten Marie Ørnsbjerg Jensen<sup>†</sup>, Rocío Mercado<sup>‡</sup>, Raghavendra Selvan<sup>†</sup>

<sup>†</sup>Department of Computer Science, University of Copenhagen, Denmark

<sup>‡</sup>Department of Chemistry, University of Copenhagen, Denmark

<sup>‡</sup>Department of Computer Science & Engineering, Chalmers University of Technology, Sweden

{frjo,raghav}@di.ku.dk, kirsten@chem.ku.dk, rocio@mailab.bio

## Abstract

Novel materials drive progress across applications from energy storage to electronics. Automated characterization of material structures with machine learning methods offers a promising strategy for accelerating this key step in material design. In this work, we introduce an autoregressive language model that performs crystal structure prediction (CSP) from powder diffraction data. The presented model, deCIFer, generates crystal structures in the widely used Crystallographic Information File (CIF) format and can be conditioned on powder X-ray diffraction (PXRD) data. Unlike earlier works that primarily rely on high-level descriptors like composition, deCIFer performs CSP from diffraction data. We train deCIFer on nearly 2.3M unique crystal structures and validate on diverse sets of PXRD patterns for characterizing challenging inorganic crystal systems. Qualitative and quantitative assessments using the residual weighted profile and Wasserstein distance show that deCIFer produces structures that more accurately match the target diffraction data when conditioned, compared to the unconditioned case. Notably, deCIFer can achieve a 94% match rate on unseen data. deCIFer bridges experimental diffraction data with computational CSP, lending itself as a powerful tool for crystal structure characterization and accelerating materials discovery.<sup>1</sup>

## 1 Introduction

Characterizing the atomic structure in functional materials is fundamental for understanding and optimizing materials for e.g. new energy technologies. Such characterization can be done using X-ray diffraction (XRD), and in particular, powder X-ray diffraction (PXRD) is widely used in materials chemistry and related fields as a main characterization tool (Cheetham and Goodwin, 2014). PXRD patterns contain information on the arrangement of atoms in a sample. However, direct structure solution (i.e. determining all atomic coordinates) is challenging from PXRD data. Instead, PXRD data analysis is generally done based on a known structural model, while parameters can be refined against experimental data (Young, 1995). Identifying a model for such analysis is often referred to as *fingerprinting*, and can be a very challenging task, where chemical intuition as well as extensive database searches are required. Even then, model identification is not always successful, and this hinders further material development. Several advances in machine learning (ML) have shown promise in aiding analysis of diffraction data over the last decade (Tatlier, 2011; Bunn et al., 2016; Oviedo et al., 2019; Wang et al., 2020), including recent advances in generative models for crystal structure prediction (CSP) (Jiao et al., 2023; Mohanty et al., 2024; Antunes et al., 2024). While these conventional ML-aided CSP methods typically explore the structural phase space guided by energy evaluations or high-level descriptors like material composition, emerging *data-informed CSP* approaches have started to integrate diffraction data directly into the generative process (Kjær et al., 2023; Guo et al., 2024; Riesel et al., 2024; Lai et al., 2025). This shift represents a fundamental departure from conventional CSP.

In this work, we introduce deCIFer (Figure 1), a transformer-based model that directly generates Crystallographic Information Files (CIFs) – text files that encode for crystal structures – *conditioned* on the rich structural information explicitly represented in experimental data, such as PXRD

<sup>1</sup>Source code available at: <https://github.com/FrederikLizakJohansen/deCIFer>

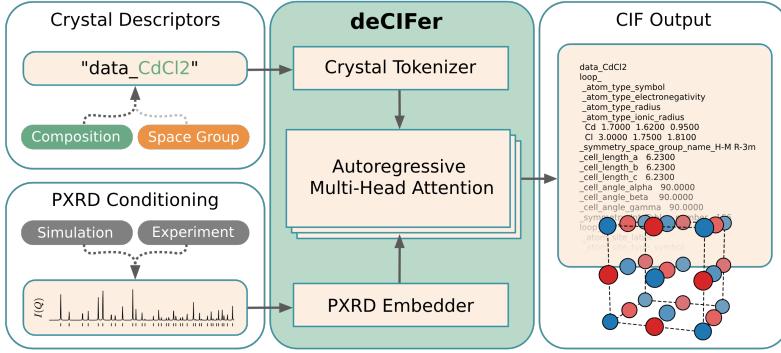


Figure 1: deCIFer performs autoregressive crystal structure prediction (CSP) from PXRD data, optionally guided by tokenized crystal descriptors. PXRD embeddings are prepended to the CIF token sequence, enabling the generation of structurally consistent CIFs directly from diffraction data.

patterns. We thus combine the strength of CSP with the possibility to include experimental data in the workflow. This approach can significantly advance the workflow used for atomic structure characterization based on PXRD methods. Instead of conditioning on the raw scattering data, in this work we learn PXRD embeddings based on a variety of data transformations. These transformations are designed to mimic noise due to experimental variations.

We evaluate the performance of deCIFer on a diverse set of PXRD patterns, demonstrating its robustness to noise and its ability to generate high-quality crystal structures. By using learned conditioning embeddings to extract fine-grained structural features from PXRD patterns, deCIFer is the first of its kind that produces syntactically correct and structurally meaningful CIFs while agreeing with reference PXRD patterns.

Our key contributions in this work are:

1. A novel autoregressive transformer model for CSP that incorporates materials science domain knowledge by conditioning on PXRD data to generate viable crystal structures in CIF format. This extends the use of CSP models, as experimental data can be incorporated in the material analysis.
2. An efficient conditioning mechanism for autoregressive models when dealing with variable length input.
3. High-fidelity simulation of experimental settings for PXRD data with Gaussian noise and peak broadening.
4. Systematic evaluation on two large-scale datasets: NOMA<sup>2</sup> & CHILI-100K ([Friis-Jensen et al., 2024](#)).

## Background and related work

PXRD is the most widely used method for structural characterization in materials chemistry. A PXRD pattern contains diffraction peaks, whose position and intensity contain information on the periodic structure in crystalline materials, i.e. their atomic positions and structure symmetry. Most quantitative analysis of PXRD data is done using structure refinement, where the parameters in a structural model is refined against experimental data. This requires a structural starting model of the correct structure type and symmetry. Identifying such a model is often a bottleneck for materials characterization.

The transformer architecture ([Vaswani et al., 2017](#)) with its capacity to capture long-range dependencies has inspired applications beyond natural language processing, including applications in materials science, where transformer-based large language models (LLMs) have been employed for CSP and property analysis. While LLMs have seen emerging use in automation of chemical syntheses ([Hocky and White, 2022](#); [Szymanski et al., 2023](#); [M. Bran et al., 2024](#)), data extraction ([Gupta et al., 2022](#); [Dagdelen et al., 2024](#); [Polak and Morgan, 2024](#); [Schilling-Wilhelmi et al., 2025](#)), and materials simulation and property prediction ([Zhang et al., 2024](#); [Rubungo et al., 2024](#); [Jablonka et al., 2024](#)), they are not yet as widely used in tasks like materials design. ChemCrow ([M. Bran et al., 2024](#))

<sup>2</sup>NOMA stands for NOMAD ([Draxl and Scheffler, 2019](#)) OQMD ([Kirklin et al., 2015](#)) & MP ([Jain et al., 2013](#)) Aggregation.

et al., 2024), for instance, is an LLM-powered chemistry search engine designed to automate reasoning tasks in materials design and other domains.

Recent works have used fine-tuning to adapt LLMs for CSP. Gruver et al. (2024) fine-tuned Llama-2 models (Touvron et al., 2023) on text-encoded atomistic data, enabling tasks like unconditional generation of stable materials. Similarly, Mohanty et al. (2024) fine-tuned LLaMA-3.1-8B (Dubey et al., 2024) using QLoRA (Dettmers et al., 2024) for efficient CIF generation conditioned on material composition and space group. In contrast, the work on CrystallLM (Antunes et al., 2024) relies on pre-training alone to generate CIFs. It is trained on an extensive corpus of CIFs representing millions of inorganic compounds.

CrystallLM relies on composition- and symmetry- level crystal descriptors. Its utility, as most other CSP methods, is thus to generate structures without direct reference to experimental data. deCIFer, on the other hand, conditions on PXRD signals, produces structures agreeing with observed diffraction patterns and address the need for models that bridge predictive power with experimental reality.

In parallel to LLM-based approaches, generative models using diffusion- or flow-based frameworks have also emerged for CSP (Jiao et al., 2023; Miller et al., 2024; Zeni et al., 2025). These methods rely on composition or partial structural constraints to guide generation, which requires domain knowledge and can limit direct incorporation of experimental data. While they show promise in producing stable crystal configurations, the gap between purely computational crystal structures and experimental data remains a challenge. The recent diffusion-based framework MatterGen (Zeni et al., 2025) addresses some of these challenges by enabling conditioning on a broad range of property constraints, and it can generate structures that are more likely to be synthesizable than previous methods; nevertheless, conditioning on PXRD has not been demonstrated in its current implementation, leaving room for further exploration in integrating experimental data more directly.

## 2 Methods

Consider a crystal structure in CIF (text input) format that is tokenized into a sequence of length  $T_i$ :  $\mathbf{x}^i = (x_1^i, x_2^i, \dots, x_{T_i}^i)$  (see Appendix A.3 for details on tokenization). The corresponding scattering pattern, which in our case is the PXRD pattern denoted by  $\mathbf{y}^i$ , is a continuous-valued vector representing the intensity profile of the scattering pattern. The dataset then consists of pairs of CIFs and PXRD patterns,  $\mathcal{D} = [(\mathbf{x}^i, \mathbf{y}^i)]_{i=1}^N$ .

Given this dataset, we want to minimize the negative conditional log-likelihood over the training data:

$$\mathcal{L}(\mathbf{X}|\mathbf{Y}; \Theta) = \frac{1}{N} \sum_{i=1}^N \left( - \sum_{t=1}^{T_i} \log P_\Theta(x_t^i | x_{<t}^i, \mathbf{y}^i) \right). \quad (1)$$

This is achieved using a conditional autoregressive model  $f_\Theta(\cdot)$  with trainable parameters  $\Theta$  based on the transformer architecture (Vaswani et al., 2017). Our autoregressive model, deCIFer, generates structures in CIF format when conditioned with an initial prompt and the PXRD data.

### 2.1 PXRD Conditioning

The PXRD data can be seen as the structural fingerprint of the structures in CIFs. We use this PXRD data to steer the CSP by using it as a conditioning input.

Following the standard procedure in materials chemistry, for each CIF, we generate the discrete diffraction peak data, given as the set  $\mathcal{P} = \{(q_k, i_k)\}_{k=1}^n$  using pymatgen (Ong et al., 2013). During the model training,  $\mathcal{P}$  is transformed into the PXRD data,  $\mathbf{y}$ , using different simulated experimental conditions. Formally, let  $\mathcal{T}$  be a set of transformations that can be applied to each  $\mathcal{P}$ .

The family of transformations used in this work are designed to closely mimic the experimental conditions. We define a distribution of transformations  $\mathcal{T}$  such that each  $\tau \sim \mathcal{T}$  comprises 1) a random peak broadening with full width at half maximum (FWHM)  $\sim \mathcal{U}(0.001, 0.100)$  and 2) additive noise with variance  $\sigma_{\text{noise}}^2 \sim \mathcal{U}(0.001, 0.050)$ . Concretely, each  $\tau$  is the composition of these two steps, and new values for broadening and noise are sampled on each draw,  $\mathbf{y} = \tau(\mathcal{P})$ . Hence for each CIF, the PXRD pattern is transformed slightly differently every time it appears during training. Later for evaluation, we use  $\tau_{\text{fixed}}$  with manually specified parameters to test the robustness of the models. The *clean* transformation  $\tau_0$  fixes FWHM = 0.05 and  $\sigma_{\text{noise}}^2 = 0$ . Examples from  $\mathcal{T}$  on a PXRD are shown in Figure 10 (in Appendix).

**Conditioning model:** A multi-layered perceptron (MLP) with trainable parameters  $\Phi$  is used to embed the PXRD data into a learnable conditioning vector  $\mathbf{e} = f_\Phi(\mathbf{y}) \in \mathbb{R}^D$ . This embedding is

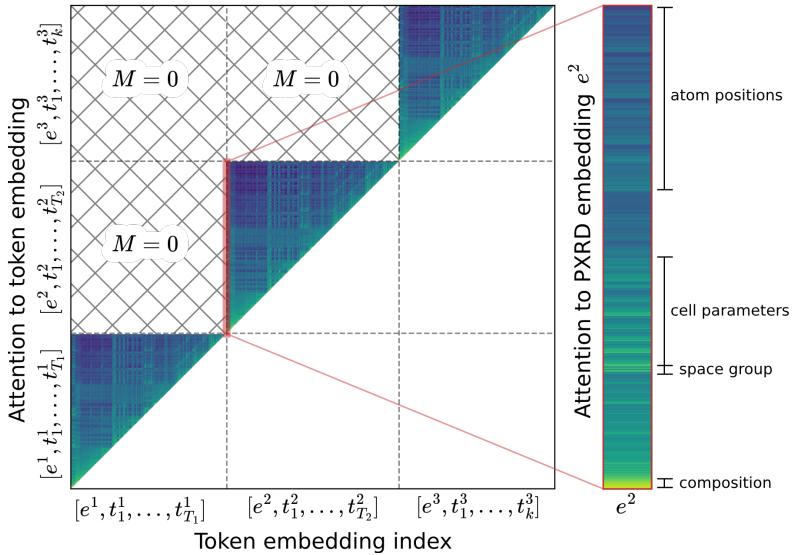


Figure 2: Visualization of the attention masking strategy, showing the log-mean attention weights (averaged over all heads) for an example sequence and highlighting how CIFs are isolated using the attention mask. The figure also illustrates how the embeddings of the second CIF attend to the conditioning PXRD embedding. Lighter shades indicate stronger attention.

inserted at the start of the tokenized CIF sequence as a conditioning signal. Jointly training the conditioning network,  $f_\Phi$ , and the autoregressive model,  $f_\Theta$ , enables the integration of structural information from the PXRD profiles in the autoregressive generation of CIFs. This joint training along with transformations to the PXRD data results in the final training objective  $\mathcal{L}(\mathbf{X}|\mathbf{Y}; \Theta, \Phi)$ .

## 2.2 Sequence packing and Isolating CIFs

To efficiently train on variable-length CIF sequences, we implement a batching strategy during training, inspired by recent sequence-packing methods that prevent cross-contamination and improve throughput (Kosec et al., 2021). Our approach concatenates multiple tokenized CIF sequences, each of length  $T_i$  (excluding the conditioning embedding) into segments of fixed context length,  $C$ . Formally, consider a sequence  $\mathbf{S} = [\mathbf{e}^1, \mathbf{t}_1^1, \dots, \mathbf{t}_{T_1}^1, \mathbf{e}^2, \mathbf{t}_1^2, \dots, \mathbf{t}_k^2]$ , where  $n$  is the last fully or partially included CIF. In this notation  $\mathbf{e}^i$  denotes a  $D$ -dimensional conditioning embedding for the  $i$ -th CIF, while  $\mathbf{t}_j^i$  is the  $D$ -dimensional input embedding for  $x_j^i$ , the  $j$ -th token of the  $i$ -th CIF. We choose  $k$  such that  $|\mathbf{S}| = C$ . If adding another CIF exceeds  $C$ , that CIF is split at the boundary and continued in the next segment. In practice, we set  $C = 3076$  based on the available GPU memory to balance throughput and memory constraints, which exceeds the length of the longest tokenized CIFs in the NOMA dataset. Although our method guarantees efficient batch utilization, it does occasionally split exceptionally long CIFs between batches ( $\approx 0.04\%$  of samples in the NOMA dataset; see Figure 11 in the Appendix). To mitigate this, we shuffle the training set at the start of each epoch so that previously split CIFs are more likely to appear in full in subsequent mini-batches, allowing the model to learn from complete sequences.

To isolate different CIFs in the same sequence, we employ an attention mask  $\mathbf{M}$ , defined such that  $M_{kl} = 1$  iff tokens  $k$  and  $l$  belong to the same CIF, and  $M_{kl} = 0$  otherwise. This results in a block-wise diagonal, upper-triangular, attention matrix as shown in Figure 2. To prevent positional information from leaking across CIF boundaries we also reset positional encodings at the start of each CIF by assigning positions from 0 to  $T_i - 1$  for the  $i$ -th sequence.

## 2.3 Evaluation Metrics

We evaluate generated structures using four metrics.

1) **Residual weighted profile** ( $R_{\text{wp}}$ ) is applied to convolved continuous PXRD profiles, measuring the discrepancy between a reference PXRD profile,  $\mathbf{y}$ , and a generated PXRD profile,  $\mathbf{y}^*$ . Formally:

$$R_{\text{wp}} = \sqrt{\frac{\sum_i w_i (y_i - y_i^*)^2}{\sum_i w_i y_i^2}}, \quad (2)$$

with all weights  $w_i = 1$  here, following convention.

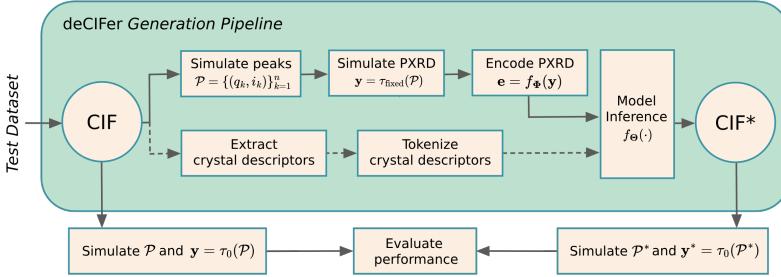


Figure 3: Evaluation pipeline. A reference CIF from the test set is used to generate discrete peaks,  $\mathcal{P} = \{(q_k, i_k)\}_{k=1}^n$ . These peaks are then transformed into a PXRD profile  $\mathbf{y} = \tau_{\text{fixed}}(\mathcal{P})$ , where  $\tau_{\text{fixed}}$  is selected according to the direc experimental setting. The resulting PXRD, along with any optional crystal descriptors, is tokenized and passed to deCIFer to produce a new CIF ( $\text{CIF}^*$ ). For evaluation, the generated structures are compared using a *clean* reference transformation  $\tau_0$  (FWHM= 0.05,  $\sigma_{\text{noise}^2} = 0$ ).

2) **Wasserstein distance** (WD) is used to compare the non-convolved, discrete intensity peaks of the PXRD patterns. WD measures the minimal transport cost of moving peaks in the reference set  $\mathcal{P} = \{(q_k, i_k)\}_{k=1}^n$  to peaks in the generated set  $\mathcal{P}^* = \{(q_j^*, i_j^*)\}_{j=1}^m$ . Letting  $\pi_{kj}$  be a coupling between intensities at positions  $q_k$  and  $q_j^*$ , we have in 1D:

$$\text{WD}(\mathcal{P}, \mathcal{P}^*) = \min_{\pi} \sum_{k=1}^n \sum_{j=1}^m \pi_{kj} \|q_k - q_j^*\|, \quad (3)$$

where  $\sum_{j=1}^m \pi_{kj} = i_k$  and  $\sum_{k=1}^n \pi_{kj} = i_j^*$ .

3) **Match rate** (MR) uses **StructureMatcher** (Ong et al., 2013) to assess structural similarity between reference and generated CIFs. Two structures are considered a match if their lattice parameters, atomic coordinates, and space group symmetries are within the defined tolerances. MR is the fraction of matching structures. See Section A.6 in the Appendix for more information.

4) **Validity** (Val.) ensures the internal consistency of each generated CIF by checking formula consistency, site multiplicity, bond lengths, and space group alignment. A CIF is deemed *overall valid* only if it passes all four checks. Detailed explanations of these metrics are provided in Section A.5 in the Appendix.

Figure 3 summarizes the evaluation pipeline.

### 3 Dataset and Experiments

**Dataset:** We use two datasets here. The first, NOMA, is a large-scale compilation of crystal structures curated in CrystaLLM (Antunes et al., 2024), which draws from the Materials Project (April 2022) (Jain et al., 2013), OQMD (v. 1.5, October 2023) (Kirklin et al., 2015), and NOMAD (April 2023) (Draxl and Scheffler, 2019) databases. The second, CHILI-100K (Friis-Jensen et al., 2024), is a large-scale dataset of experimentally determined crystal structures obtained from a curated subset of the Crystallography Open Database (COD) (Gražulis et al., 2009). We employ NOMA for both training and testing our models, whereas CHILI-100K is used exclusively for testing. Both datasets are open-source and available for download.<sup>3</sup>

**Preprocessing:** We begin by applying the standard data preprocessing steps outlined in CrystaLLM before performing additional curation to address complexities such as ionic charges and to ensure consistent structural representation between NOMA and CHILI-100K. The resulting NOMA dataset comprises approximately 2.3M CIFs, spanning 1–10 elements, including elements up to atomic number 94 (excluding polonium, astatine, radon, francium, and radium). Duplicate structures were filtered by selecting the configuration with the lowest volume per formula unit. All structures were converted to a standardized CIF format using the pymatgen library (Ong et al., 2013) and all floating point values were rounded to four decimal places. The resulting CHILI-100K dataset consists of  $\approx 8.2$ K CIFs, spanning 1–8 elements, including elements up to atomic number 85. Figures 11 and 12 (in Appendix) show the distribution of elemental compositions, space groups, and more for NOMA and CHILI-100K.

A challenge in NOMA is the disproportionate representation of high-symmetry structures. To account for the skew in the distribution, we derive the *space group*, a numerical label from 1 to

<sup>3</sup>NOMA: [github.com/lantunes/CrystaLLM](https://github.com/lantunes/CrystaLLM) (CC-BY 4.0 licence), CHILI-100K: [github.com/UlrikFriisJensen/CHILI](https://github.com/UlrikFriisJensen/CHILI) (Apache 2.0 licence).

230 that encodes crystal symmetry, from each CIF. We stratify the data by grouping these labels into bins of 10, e.g., space groups 1–10, 11–20, etc., and use these strata to split data into 90% training, 7.5% validation, and 2.5% testing subsets while preserving the overall distribution of crystal symmetries.

The dataset was tokenized into a vocabulary of 373 tokens, encompassing CIF tags, space group symbols, element symbols, numeric digits, punctuation marks, padding tokens and a special conditioning token. For a detailed explanation of the standardization and tokenization processes, refer to Section A.2 and A.3 in the Appendix.

**Hyperparameters:** deCIFer has two learnable components,  $f_{\Phi}$  and  $f_{\Theta}$ . The encoder  $f_{\Phi}$  is a 2-layer MLP that takes a PXRD profile of dimension 1000 and outputs a 512-dimensional embedding. This embedding is prepended as the conditioning-token to the sequence of CIF tokens, where each CIF token is also of  $D = 512$  dimensions.  $f_{\Theta}$  is a decoder-only transformer (Vaswani et al., 2017) with 8 layers, each containing 8 attention heads. We use a context length of 3076 and a batch size of 32. A linear warm-up of the learning rate is applied over the first 100 epochs, followed by a cosine decay schedule using AdamW (Loshchilov and Hutter, 2017) ( $\beta_1 = 0.9$ ,  $\beta_2 = 0.95$ , weight decay  $10^{-1}$ ) for 50K epochs with gradient accumulation of 40 steps on a single NVIDIA A100 GPU with mixed-precision acceleration. All aspects of the model architecture were implemented in Pytorch (Paszke et al., 2019).  $f_{\Phi}$  has  $\approx 0.78M$  trainable parameters and  $f_{\Theta}$  contains  $\approx 26.94M$ , resulting in a deCIFer model with 27.72M parameters. The code is open-source and additional details can be found in Section A.8 in the Appendix.

**Experiments:** We conducted a series of experiments to evaluate deCIFer’s ability to perform CSP when conditioned on PXRD profiles, space group, and composition. First, we compare the baseline performance of deCIFer against an unconditioned variant (U-deCIFer) to assess the direct impact of PXRD conditioning on the generated structures. Next, we introduce controlled noise and peak broadening into the input PXRD data, examining deCIFer’s robustness under more challenging scenarios resembling real-world PXRD data. Finally, we apply deCIFer (trained on NOMA) to CHILI-100K (Friis-Jensen et al., 2024) to demonstrate its scalability and performance on more complex crystal systems. In these experiments, we generate one CIF for each reference sample in the test sets. For consistent evaluation, each reference CIF is also processed through the fixed, *clean* transformation  $\tau_0$  (FWHM = 0.05,  $\sigma_{\text{noise}}^2 = 0$ ).

For some experiments we allow space group and composition to be specified as crystal descriptors. Since CIFs inherently encode space group and composition as text, we treat these descriptors as standard tokens in our vocabulary (see Section A.3 in the Appendix). The space group appears in the CIF header, while composition is stored as element–count pairs. During inference, these tokens are optionally inserted into the CIF.

## 4 Results

### 4.1 Importance of PXRD Conditioning

We compare deCIFer and U-deCIFer to study the effect of PXRD conditioning. We evaluate each model in three settings: without any crystal descriptors (“none”), with only compositional information (“comp.”), and with both compositional and space group information (“comp. + s.g.”). Both models share the same architectural backbone, but deCIFer receives structural guidance via the PXRD conditioning.

Table 1 shows that deCIFer achieves lower average  $R_{\text{wp}}$  and WD than U-deCIFer across all descriptor settings. Figure 4 visualizes the distributions of  $R_{\text{wp}}$  and WD for both models, highlighting the large performance improvement due to PXRD conditioning, and a smaller but still notable improvement due to composition conditioning. Figure 5 shows that structures of more common crystal systems lead to better outcomes, while under-represented, lower symmetry crystal systems remain challenging. Figure 6 further illustrates the effectiveness of PXRD conditioning with three examples from the NOMA test set, showcasing the range of generated structures, from near-perfect alignment to structural mismatch. While U-deCIFer benefits from the addition of crystal descriptors, it never reaches the accuracy of deCIFer. Both models achieve high rates of syntactic validity, with the unconditional model marginally surpassing deCIFer (Table 2). Overall, these results demonstrate that conditioning on PXRD data and crystal descriptors substantially enhances the model’s ability to generate CIFs with close alignment to desired PXRD profiles.

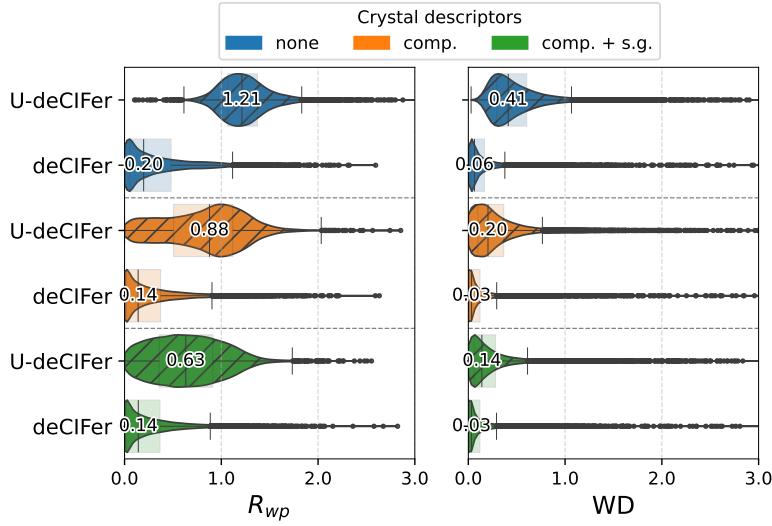


Figure 4: Distribution of  $R_{wp}$  and WD for deCIFer and U-deCIFer on the NOMA test set, presented as violin plots with overlaid boxplots; the median is shown for each distribution. Lower  $R_{wp}$  and WD values indicate a better alignment between the reference- and generated CIFs in terms of the reference PXRD profile. The models are evaluated in three settings: “none” (no crystal descriptors), “comp.” (composition only), and “comp. + s.g.” (composition plus space group).

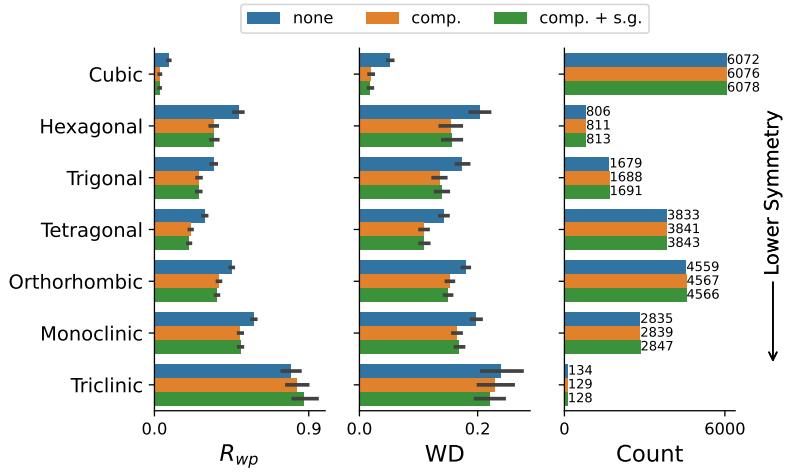


Figure 5: Average metric values by crystal systems for deCIFer on the NOMA test set show better performance for well-represented systems, while rarer, low-symmetry systems lead to worse performance ( $\uparrow R_{wp}, \uparrow WD$ ), reflecting their modelling complexity. The right-most plot shows crystal system distribution of the test set.

## 4.2 Robustness to Perturbations in PXRD Conditioning

We evaluated deCIFer’s ability to generate accurate CIFs under varying levels of additive noise and peak broadening in the PXRD conditioning input, while also providing the composition as a crystal descriptor. Building upon the baseline scenario of clean, noise-free PXRD data, we tested several increasingly challenging conditions: maximum noise, maximum peak broadening, combined noise and broadening, and out-of-distribution (OOD) broadening levels beyond the model’s training range.

Table 3 (and Figure 15 in Appendix) show that while additive noise results in slight performance degradation, the model remains robust to in-distribution noise and peak broadening. We also observe moderate performance degradation with OOD broadening or significantly higher noise levels. Unsurprisingly, we observe that lower-symmetry crystal systems remain more challenging to predict under perturbed conditions across all crystal systems. This is depicted in Figure 16 in the Appendix.

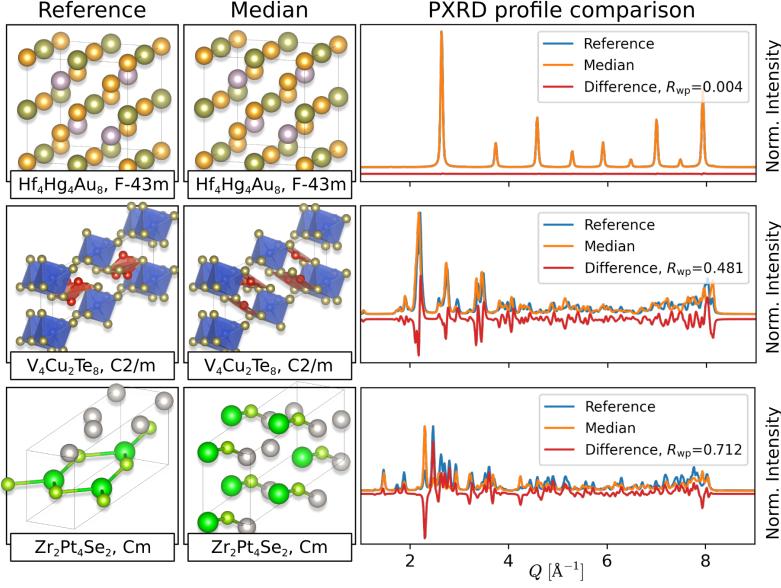


Figure 6: Three examples from the NOMA test set showcasing deCIFer generations. PXRD data is used for conditioning, with composition as a crystal descriptor. For each example, 100 CIFs are generated, and the median structure (based on  $R_{wp}$ ) is shown alongside the reference structure and PXRD profiles (noise-free, FWHM = 0.05). The examples illustrate: (top) near-perfect PXRD conformity, (middle) structural match with suboptimal PXRD alignment, and (bottom) structural mismatch.

Table 1: Performance comparison for 20K samples from the NOMA test set using deCIFer and U-deCIFer with descriptors. **none**: No crystal descriptors, **comp.**: compositional information only, and **comp. + s.g.**: both compositional and space group information.

Desc.	Model	$R_{wp}$ ( $\mu \pm \sigma$ ) ↓	WD ( $\mu \pm \sigma$ ) ↓
<b>none</b>	U-deCIFer	$1.24 \pm 0.26$	$0.49 \pm 0.31$
	deCIFer	$0.32 \pm 0.34$	$0.14 \pm 0.21$
<b>comp.</b>	U-deCIFer	$0.82 \pm 0.41$	$0.28 \pm 0.29$
	deCIFer	$0.25 \pm 0.29$	$0.10 \pm 0.20$
<b>comp. + s.g.</b>	U-deCIFer	$0.65 \pm 0.36$	$0.23 \pm 0.28$
	deCIFer	$0.24 \pm 0.29$	$0.10 \pm 0.20$

Table 2: Validity of generated CIFs for the NOMA test set using deCIFer and U-deCIFer. Abbreviations: Form = formula validity, SG = space group validity, BL = bond length validity, SM = site multiplicity validity. Overall validity (Val.) is calculated as the percentage of CIFs that satisfy all four validity metrics simultaneously. Match rate (MR) represents the percentage of generated CIFs that replicate the reference CIF.

Desc.	Model	Form (%) ↑	SG (%) ↑	BL (%) ↑	SM (%) ↑	Val. (%) ↑	MR (%) ↑
<b>none</b>	U-deCIFer	99.82	98.87	94.30	99.47	93.49	0.00
	deCIFer	99.42	98.85	93.69	99.46	92.66	5.01
<b>comp.</b>	U-deCIFer	99.87	99.09	94.40	99.46	93.78	49.30
	deCIFer	99.68	99.21	94.37	99.55	93.73	91.50
<b>comp.+s.g.</b>	U-deCIFer	99.85	98.88	94.51	99.47	93.72	87.07
	deCIFer	99.74	99.26	94.38	99.58	93.90	94.53

### 4.3 OOD Evaluation on CHILI-100K

To evaluate the generalization of deCIFer to more complex crystal systems, we tested its performance on the CHILI-100K dataset (Friis-Jensen et al., 2024) which had no overlap with the NOMA training data for deCIFer. Unlike synthetic datasets, CHILI-100K presents a closer approximation to real-world challenges, including complex structural motifs and a broader distribution of crystal symmetries. Additionally, CHILI-100K contains a significantly higher proportion of lower-symmetry structures compared to synthetic datasets like NOMA (see the sample distribution in Appendix Figure 17).

Table 3: Performance of deCIFer under varying levels of noise and peak broadening in PXRD inputs for NOMA and CHILI-100K test sets, evaluated with compositional descriptors. We evaluate two in-distribution (ID) scenarios and one out-of-distribution (OOD) scenario for the PXRD input. Reported metrics include residual weighted profile ( $R_{wp}$ ), Wasserstein distance (WD), overall validity, and match rate.

Dataset	$(\sigma_{noise}^2, \text{FWHM})$	$R_{wp} \downarrow$	$\text{WD} \downarrow$	$\text{Val. (\%)} \uparrow$	$\text{MR (\%)} \uparrow$
NOMA	ID: (0.00, 0.05)	0.25±0.29	0.10±0.20	93.73	91.50
	ID: (0.05, 0.10)	0.31±0.30	0.12±0.21	93.77	89.28
	OOD: (0.10, 0.20)	0.65±0.34	0.23±0.29	91.66	77.66
CHILI-100K	ID: (0.00, 0.05)	0.70±0.37	0.22±0.21	41.83	37.34
	ID: (0.05, 0.10)	0.73±0.36	0.22±0.21	40.95	35.97
	OOD: (0.10, 0.20)	0.87±0.33	0.24±0.20	33.62	26.09

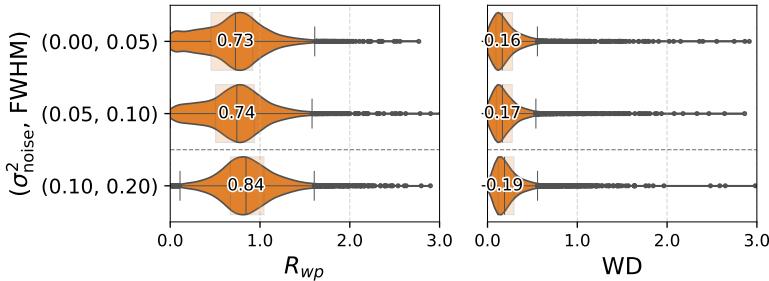


Figure 7: Distribution of  $R_{wp}$  and WD for the CHILI-100K test set, presented as violin plots with overlaid boxplots, for deCIFer using both PXRD conditionings and composition (**comp.**). The distributions show predictions under three PXRD conditioning scenarios: (top) clean, (middle) maximum noise and broadening, and (bottom) an out-of-distribution PXRD scenario.

The results on CHILI-100k dataset are summarized in Table 3 and Figure 7. These show that deCIFer maintains a reasonable level of structural accuracy on this challenging dataset. The relatively low validity score is partly due to challenges with bond length validity, which was notably lower than the other validity metrics. A full breakdown of the validity metrics for this experiment is available in Section A.10 in the Appendix.

Despite the performance drop, deCIFer demonstrated robustness to added noise and peak broadening in PXRD inputs, with stable  $R_{wp}$  and WD values across perturbed conditions as seen in Figure 7. This stability, along with its ability to generate a variety of structural features, suggests that deCIFer could be useful for practical applications involving experimentally derived PXRD data.

## 5 Discussion and Outlook

**PXRD-steered structure generation:** The experiments on NOMA and CHILI-100k in Section 4 clearly show the utility of including conditioning data such as PXRD when performing CSP. Unlike unconditional structure generation, generating crystal structures that are steered by scattering data such as PXRD is more useful in applications where a specific target property might be desirable.

By embedding the PXRD data,  $\mathbf{y}$ , into a learnable conditional embedding,  $\mathbf{e} = f_\Phi(\mathbf{y})$ , and prepending it onto prompts for deCIFer, we have demonstrated a general paradigm for integrating additional conditional data in materials design, an idea which can be extended to the incorporation of other material properties. In cases where multiple experimental data sources are available, they can be easily injected into the generative process using specific conditioning models. That is, if  $P$  properties are available such that  $\mathbf{y} = \{\mathbf{y}_1, \dots, \mathbf{y}_P\}$ , then deCIFer can be extended to incorporate these additional data using additional conditioning models,  $f_{\Phi_1}, \dots, f_{\Phi_P}$ , that can be jointly trained with the generative model,  $f_\Theta$ , via a training objective of the form  $\mathcal{L}(\mathbf{X}|\mathbf{Y}_1, \dots, \mathbf{Y}_P; \Theta, \Phi_1, \dots, \Phi_P)$ .

**Consistency in CIF generation:** To investigate the consistency and variability of deCIFer, we generated 16K CIFs for the same PXRD profile with different crystal descriptors (“none”, “comp.”, and “comp+s.g.”). Figure 8 illustrates our findings for a challenging monoclinic crystal system of  $\text{Sr}_2\text{Cd}_2\text{Se}_4$  from the NOMA test set. When unconstrained by crystal descriptors, the model generates a wide diversity of cell-parameters, compositions, and space groups, yet the  $R_{wp}$  values tend to cluster in a relatively narrow range. In contrast, imposing compositional and, especially, space

group constraints yields much tighter cell-parameter distributions and a broader  $R_{wp}$  distribution. The broader distributions highlight how the  $R_{wp}$  metric is highly sensitive to even small structural deviations and shows the importance of complementing less sensitive metrics like the WD with other validation methods. Despite these variations, the overall structural match rate to the reference CIF remains high. If the composition or space group is known with high confidence, incorporating these constraints can speed up convergence to a more accurate structural solution. Conversely, if the objective is to explore the material landscape more broadly, unconstrained generation with PXRD can be advantageous.

**Limitations:** In materials science, the concept of data leakage is nuanced due to the challenges in defining novelty and diversity, as emphasized in recent discussions on scaling AI for materials discovery (Cheetham and Seshadri, 2024). Although we stratify the NOMA dataset by space group bins and use CHILI-100K exclusively for testing, there remains the possibility of implicit overlaps, such as structurally similar crystal entries or shared compositional biases. However, the inherent redundancy in materials data, where many “new” compounds are variations of existing structures, might simply reflect the natural landscape of materials science. Furthermore, the vast size and diversity of the NOMA dataset as well as the thorough preprocessing steps (e.g., de-duplication, filtering, and standardization), coupled with the independent curation of CHILI-100K, significantly mitigate the impact of any potential overlaps.

We rely on two metrics to validate the performance of deCIFer,  $R_{wp}$  and WD. While  $R_{wp}$  is the standard in Rietveld refinement, and is highly sensitive to small deviations in atomic positions or cell parameters, it can overemphasize noise or strong peaks. Conversely, WD captures broad differences more robustly, but can be disproportionately sensitive to outlier peaks. Each reveal different aspects of pattern agreement, so employing them in tandem, along with comparing structural similarities is necessary for robust validation.

**Future Work:** This work opens several directions for further exploration and improvement.

One promising area for improvement lies in exploring more advanced decoding strategies, such as beam search, to enhance the generative model’s capabilities in downstream tasks. By maintaining multiple hypotheses during decoding, beam search could produce diverse candidate CIFs for a given PXRD profile, improving structure determination accuracy by ranking outputs based on metrics like  $R_{wp}$ . This method could also support optimization strategies that prioritize structural validity and relevance.

Another direction could be to integrate reinforcement learning from human feedback (RLHF) to guide the model more directly toward generating accurate and chemically valid structures (Ziegler et al., 2019). By defining a reward function tailored to properties such as low  $R_{wp}$  values, structural integrity, and adherence to chemical constraints, and interaction with a human expert, RLHF could further refine the model’s outputs.

## 6 Conclusion

In this work, we introduced deCIFer as a data-informed approach to CSP. deCIFer is an autoregressive transformer-based language model that generates CIFs directly from PXRD data. By conditioning on simulated PXRD profiles, deCIFer captures fine-grained structural information beyond what is possible with composition- or symmetry-based information alone. Evaluations on large synthetic and experimentally derived datasets highlight deCIFer’s robust performance in scenarios with varying levels of noise and peak broadening. deCIFer was developed using lab-scale resources with a single GPU. This holds potential for diverse experiments going forward. The conditional prompting paradigm introduced in deCIFer can be used to integrate other experimental data. This can be essential to accelerating material design processes.

## Acknowledgments

The authors thank Richard Michael and Adam F. Sapnik for useful feedback.

This work is part of a project that has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 Research and Innovation Programme (grant agreement No. 804066). We are grateful for funding from University of Copenhagen through the Data+ program. RM acknowledges funding provided by the Wallenberg AI, Autonomous Systems, and Software Program (WASP), supported by the Knut and Alice Wallenberg Foundation.

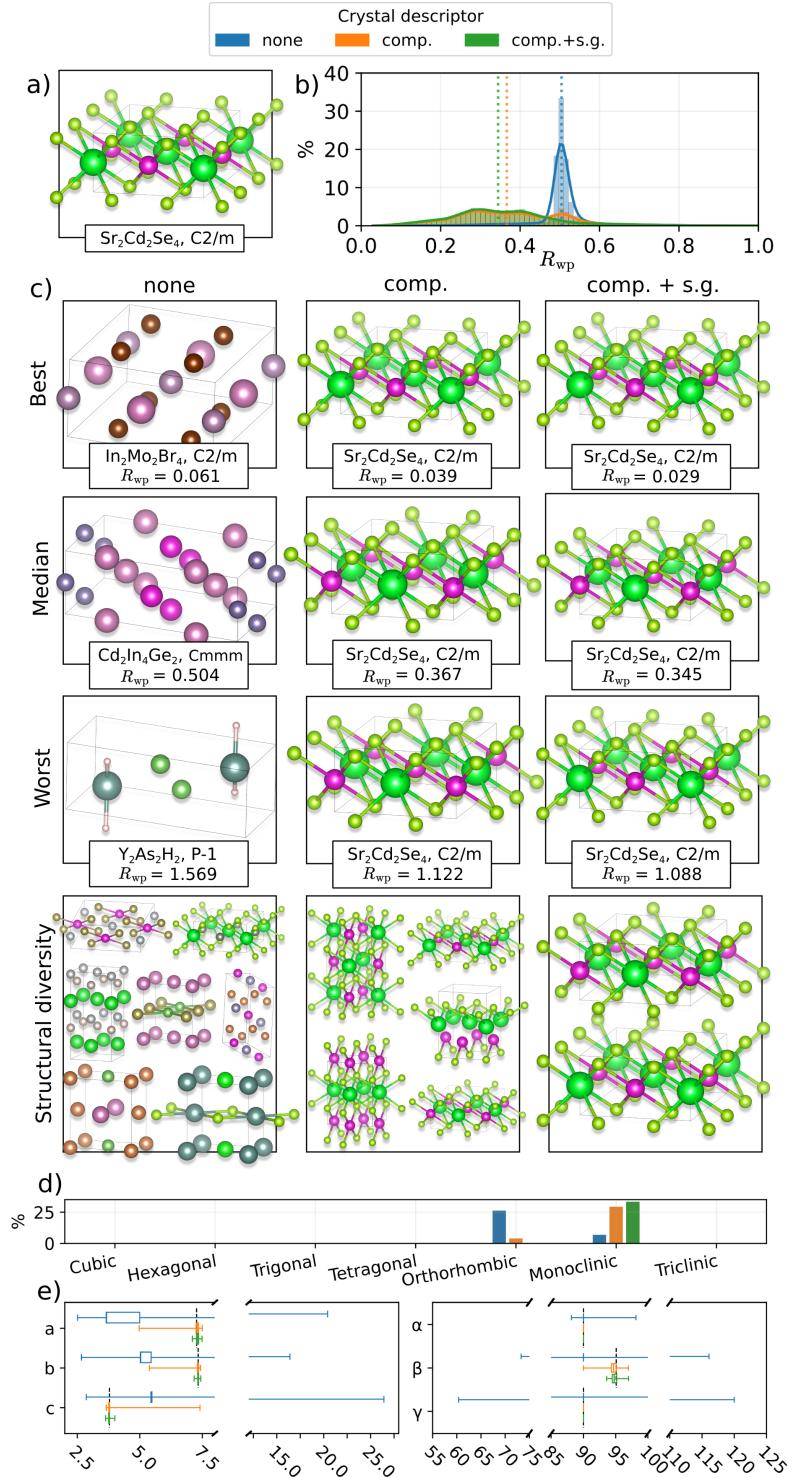


Figure 8: Illustration of deCIFer-sampled structures for a single PXRD profile of the monoclinic crystal system of  $\text{Sr}_2\text{Cd}_2\text{Se}_4$  (16K samples).  $R_{\text{wp}}$  comparisons were made with transformation  $\tau_0$ , void of noise and with  $\text{FWHM} = 0.05$ , and with three crystal descriptor scenarios. a) Reference structure. b) Empirical probability distribution of generating a CIF with a given  $R_{\text{wp}}$  to the reference structure. c) Representative examples of generated structures showing best (lowest  $R_{\text{wp}}$ ), median, worst (highest  $R_{\text{wp}}$ ), and additional diverse examples. d) Distribution of the sampled crystal systems. e) Histograms of cell- lengths (a,b,c) and angles ( $\alpha, \beta, \gamma$ ) for the 16K sampled structures, with reference cell parameters shown as dotted lines.

## References

- L. M. Antunes, K. T. Butler, and R. Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):10570, 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54639-7.

- J. K. Bunn, J. Hu, and J. R. Hattrick-Simpers. Semi-supervised approach to phase identification from combinatorial sample diffraction patterns. *JOM*, 68(8):2116–2125, Aug 2016. ISSN 1543-1851. doi: 10.1007/s11837-016-2033-8.
- A. K. Cheetham and A. L. Goodwin. Crystallography with powders. *Nature Materials*, 13(8):760–762, Aug 2014.
- A. K. Cheetham and R. Seshadri. Artificial intelligence driving materials discovery? perspective on the article: Scaling deep learning for materials discovery. *Chemistry of Materials*, 36(8):3490–3495, 2024. doi: 10.1021/acs.chemmater.4c00643.
- J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, and A. Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418, 2024.
- T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.
- C. Draxl and M. Scheffler. The nomad laboratory: From data sharing to artificial intelligence. *Journal of Physics: Materials*, 2, 05 2019. doi: 10.1088/2515-7639/ab13bb.
- A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- U. Friis-Jensen, F. L. Johansen, A. S. Anker, E. B. Dam, K. M. O. Jensen, and R. Selvan. Chili: Chemically-informed large-scale inorganic nanomaterials dataset for advancing graph machine learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 4962–4973, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671538.
- S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, and A. Le Bail. Crystallography Open Database – an open-access collection of crystal structures. *Journal of Applied Crystallography*, 42(4):726–729, Aug 2009. doi: 10.1107/S0021889809016690.
- N. Gruver, A. Sriram, A. Madotto, A. G. Wilson, C. L. Zitnick, and Z. W. Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.
- G. Guo, T. Saidi, M. Terban, M. Valsecchi, S. J. Billinge, and H. Lipson. Ab initio structure solutions from nanocrystalline powder diffraction data, 2024.
- T. Gupta, M. Zaki, N. A. Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102, 2022.
- G. M. Hocky and A. D. White. Natural language processing models that automate programming will transform chemistry research and teaching. *Digital discovery*, 1(2):79–83, 2022.
- K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero, and B. Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024.
- A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 07 2013. ISSN 2166-532X. doi: 10.1063/1.4812323.
- R. Jiao, W. Huang, P. Lin, J. Han, P. Chen, Y. Lu, and Y. Liu. Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information Processing Systems*, 36:17464–17497, 2023.
- S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Computational Materials*, 1(1):1–15, 2015.
- E. T. S. Kjær, A. S. Anker, M. N. Weng, S. J. L. Billinge, R. Selvan, and K. M. Ø. Jensen. Deepstruc: towards structure solution from pair distribution function data using deep generative models. *Digital Discovery*, 2:69–80, 2023. doi: 10.1039/D2DD00086E.

- M. Kosec, S. Fu, and M. M. Krell. Packing: Towards 2x NLP BERT acceleration. *CoRR*, abs/2107.02027, 2021.
- Q. Lai, F. Xu, L. Yao, Z. Gao, S. Liu, H. Wang, S. Lu, D. He, L. Wang, L. Zhang, C. Wang, and G. Ke. End-to-end crystal structure prediction from powder x-ray diffraction. *Advanced Science*, page 2410722, 2025.
- I. Loshchilov and F. Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017.
- A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11, 2024.
- B. K. Miller, R. T. Chen, A. Sriram, and B. M. Wood. Flowmm: Generating materials with riemannian flow matching. In *Forty-first International Conference on Machine Learning*, 2024.
- T. Mohanty, M. Mehta, H. M. Sayeed, V. Srikanth, and T. D. Sparks. Crystext: A generative ai approach for text-conditioned crystal structure generation using llm. *ChemRxiv*, 2024. doi: 10.26434/chemrxiv-2024-gjhpq. This content is a preprint and has not been peer-reviewed.
- K. Momma and F. Izumi. VESTA: a three-dimensional visualization system for electronic and structural analysis. *Journal of Applied Crystallography*, 41(3):653–658, Jun 2008. doi: 10.1107/S0021889808012016.
- S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. Chevrier, K. A. Persson, and G. Ceder. Python materials genomics ( pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013. doi: 10.1016/j.commatsci.2012.10.028.
- F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. P. Tian, G. Romano, A. Gilad Kusne, and T. Buonassisi. Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks. *npj Computational Materials*, 5(1):60, May 2019. ISSN 2057-3960. doi: 10.1038/s41524-019-0196-x.
- A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- M. P. Polak and D. Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569, 2024.
- E. A. Riesel, T. Mackey, H. Nilforoshan, M. Xu, C. K. Badding, A. B. Altman, J. Leskovec, and D. E. Freedman. Crystal structure determination from powder diffraction patterns with generative machine learning. *Journal of the American Chemical Society*, 146(44):30340–30348, 2024. doi: 10.1021/jacs.4c10244. PMID: 39298266.
- A. N. Rubungo, K. Li, J. Hattrick-Simpers, and A. B. Dieng. LLM4mat-bench: Benchmarking large language models for materials property prediction. In *AI for Accelerated Materials Design - NeurIPS 2024*, 2024.
- M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez, and K. M. Jablonka. From text to insight: large language models for chemical data extraction. *Chemical Society Reviews*, 2025.
- N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91, 2023.
- M. Tatlier. Artificial neural network methods for the prediction of framework crystal structures of zeolites from xrd data. *Neural Computing and Applications*, 20(3):365–371, Apr 2011. ISSN 1433-3058. doi: 10.1007/s00521-010-0386-4.
- A. Togo and I. Tanaka. Spglib: a software library for crystal symmetry search, 2018.

- H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- H. Wang, Y. Xie, D. Li, H. Deng, Y. Zhao, M. Xin, and J. Lin. Rapid identification of x-ray diffraction patterns based on very limited data by interpretable convolutional neural networks. *Journal of Chemical Information and Modeling*, 60(4):2004–2011, Apr 2020. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c00020.
- R. Young. *The Rietveld Method*. IUCr monographs on crystallography. Oxford University Press, 1995. ISBN 9780198559122.
- C. Zeni, R. Pinsler, D. Zügner, A. Fowler, M. Horton, X. Fu, Z. Wang, A. Shysheya, J. Crabbé, S. Ueda, et al. A generative model for inorganic materials design. *Nature*, pages 1–3, 2025.
- D. Zhang, X. Liu, X. Zhang, C. Zhang, C. Cai, H. Bi, Y. Du, X. Qin, A. Peng, J. Huang, B. Li, Y. Shan, J. Zeng, Y. Zhang, S. Liu, Y. Li, J. Chang, X. Wang, S. Zhou, J. Liu, X. Luo, Z. Wang, W. Jiang, J. Wu, Y. Yang, J. Yang, M. Yang, F.-Q. Gong, L. Zhang, M. Shi, F.-Z. Dai, D. M. York, S. Liu, T. Zhu, Z. Zhong, J. Lv, J. Cheng, W. Jia, M. Chen, G. Ke, W. E, L. Zhang, and H. Wang. DPA-2: a large atomic model as a multi-task learner. *npj Computational Materials*, 10 (1):293, Dec. 2024. ISSN 2057-3960. doi: 10.1038/s41524-024-01493-2.
- D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

## A Appendix

### A.1 Code and Data Availability

The code for training and using the deCIFer model is open source and released under the MIT License. Source code is available here: <https://github.com/FrederikLizakJohansen/deCIFer>.

### A.2 CIF Syntax Standardization

To enhance the transformer model to process CIFs effectively, we standardized all CIFs in the dataset. Inspired by CrystaLLM (Antunes et al., 2024), we employed similar pre-processing and tokenization strategies, incorporating additional steps to ensure that CHILI-100K (Friis-Jensen et al., 2024) was aligned to the standardized format of NOMA, by the removal certain details such as oxidation states and partial occupancies. We employ the following steps:

1. **Uniform Structure Conversion:** CIFs were converted to `pymatgen.Structure` (Ong et al., 2013) objects to provide a consistent base representation.
2. **Standardized CIF Regeneration:** Using `pymatgen.CifWriter` (Ong et al., 2013), CIFs were regenerated to ensure uniform formatting, eliminate customs headers, etc.
3. **Data Tag Normalization:** The reduced formula, following the `data_` tag was replaced with the full cell composition, sorted by atomic number for consistency.
4. **Symmetry Operator Removal:** Symmetry operators were excluded during pre-processing to simplify the data, but reintroduced during evaluation for validating structural matches. This can easily be done because the reintroduction process uses the space group information retained in the pre-processed files, ensuring compatibility and accuracy.
5. **Incorporation of Extra Information:** Custom properties that are easily derived from the composition of each CIF, such as electronegativity, atomic radius, and covalent radius, were appended to each CIF to maximize the readily available information within each CIF.
6. **Oxidation State and Occupancy Filtering:** Oxidation state refers to the charge of an atom within a compound, which can vary depending on chemical bonding. Occupancy indicates the fraction of a particular atomic site that is occupied in the crystal structure (e.g., a value of 1.0 represents a fully occupied site, while 0.5 indicates partial occupancy). All traces of oxidation states were removed, and only crystal structures with full occupancy were retained. This ensures consistency by aligning CHILI-100K (Friis-Jensen et al., 2024) with the standardized format of NOMA (Antunes et al., 2024).
7. **Numerical Value Normalization:** Numerical values were rounded to four decimal places.

Figure 9 shows a pre-processed and standardized CIF from the NOMA dataset alongside its corresponding unit cell representation and a realisation of its corresponding PXRD profile, as could be input into deCIFer.

### A.3 CIF Tokenization

To process CIF files effectively, we tokenized each file into a sequence of tokens using a custom vocabulary tailored to crystallographic data in the CIF format. Each CIF was parsed to extract key structural and chemical information, such as lattice parameters, atomic positions, and space group symbols. Numerical values were tokenized digit-by-digit, including decimal points and special characters as separate tokens. Table 4 shows all supported tokens.

### A.4 PXRD Simulation

**What do the axes in PXRD mean?** In a typical PXRD experiment, the **x-axis** corresponds to the magnitude of the scattering vector, commonly denoted by  $Q$  (in units of  $\text{\AA}^{-1}$ ), or sometimes the diffraction angle  $2\theta$ . In this work, we use  $Q = \frac{4\pi \sin \theta}{\lambda}$  where  $\lambda$  is the radiation wavelength and  $\theta$  is the scattering angle. The **y-axis** represents the scattered intensity observed at each  $Q$ -value, sometimes normalized to have a maximum intensity of 1.

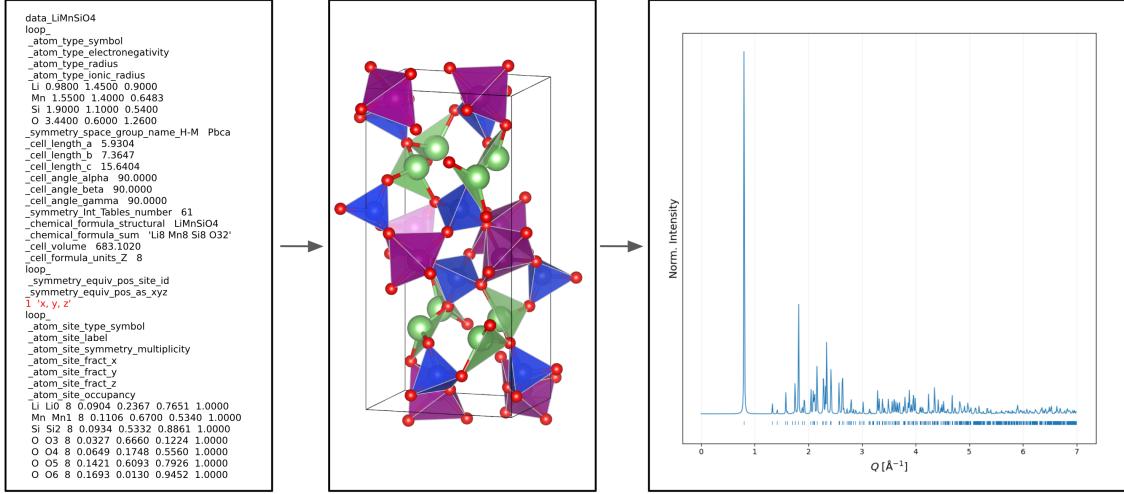


Figure 9: Illustration of a CIF after applying the pre-processing and standardization steps described. Also shown are the corresponding unit cell representation using VESTA (Momma and Izumi, 2008) for visualization and the simulated PXRD profile (with  $\sigma^2 = 0.00$  and FWHM=0.01). The red highlight in the CIF indicates where the original symmetry operators were replaced during pre-processing and would be restored for evaluation.

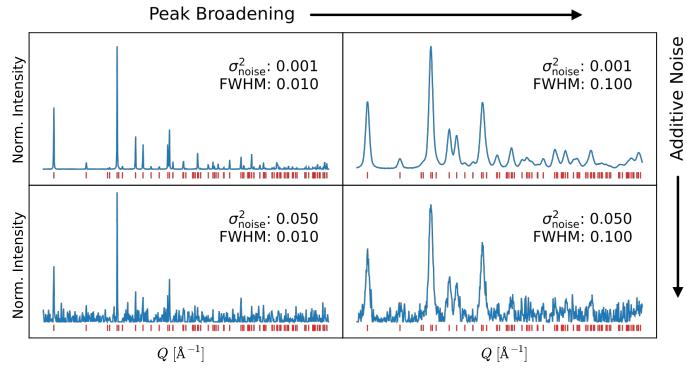


Figure 10: Simulated PXRD profiles with fixed transformation of FWHM and  $\sigma_{\text{noise}}^2$  as indicated. Discrete peaks,  $\mathcal{P} = \{(q_k, i_k)\}_{k=1}^n$ , are shown in red, while the convolved PXRD profiles,  $\mathbf{y}$ , are shown in blue. Examples with minimal and maximal noise and broadening levels are shown for a compound with composition CdRhBr<sub>2</sub> and space group R3m.

**Peak data and transformations.** Following Section 2.1, we start with the discrete diffraction peak data:  $\mathcal{P} = (q_k, i_k)_{k=1}^n$ , where each  $q_k$  is the center of a reflection peak, and  $i_k$  is the associated peak intensity. To simulate experimental effects, we apply transformation  $\tau \sim \mathcal{T}$ , which includes **peak broadening** and **additive noise**.

**Peak broadening.** For each peak  $k$ , centered at  $q_k$ , we convolve an idealized delta function peak with a pseudo-Voigt profile. At the continuous variable  $Q$ , the psuedo-Voigt profile is the mixture of a Lorentzian  $L$  and a Gaussian  $G$ , such that

$$PV_k(Q - q_k) = \eta L(Q - q_k) + (1 - \eta)G(Q - q_k), \quad (4)$$

where  $0 \leq \eta \leq 1$  is fixed at  $\eta = 0.5$  in this work.

Let FWHM denote the full width at half maximum. The Lorentzian half-width is then  $\gamma = \frac{\text{FWHM}}{2}$ , making

$$L(Q - q_k) = \frac{1}{1 + \left(\frac{Q - q_k}{\gamma}\right)}. \quad (5)$$

The Gaussian standard deviation is  $\sigma = \frac{\text{FWHM}}{2\sqrt{2 \ln 2}}$ , making

$$G(Q - q_k) = \exp\left(-\frac{1}{2} \left(\frac{Q - q_k}{\sigma}\right)^2\right). \quad (6)$$

**Convolved PXRD.** Given the peak centers  $q_k$ , intensities  $i_k$ , and a choice of FWHM, we obtain the convolved PXRD profile

$$I_{\text{conv}}(Q) = \sum_{k=1}^n i_k \text{PV}_k(Q - q_k). \quad (7)$$

Afterwards, we normalize  $I_{\text{conv}}(Q)$  so that its maximum intensity is 1.

**Noise addition.** Let  $\epsilon(Q)$  be drawn from a zero-mean Gaussian distribution with variance  $\sigma_{\text{noise}}^2$ . This yields the final transformed intensity PXRD profile:

$$I(Q) = I_{\text{conv}}(Q) + \epsilon(Q). \quad (8)$$

**Implementation details.** In practice, we use the `XRDCalculator` from the `pymatgen` library ([Ong et al., 2013](#)) for generating the initial discrete peak data  $\mathcal{P}$ . For training, we sample  $Q$ -values in  $[Q_{\min}, Q_{\max}]$  at increments of  $Q_{\text{step}}$ . We then apply random transformations  $\tau$  during model training. Specific parameters for FWHM and  $\sigma_{\text{noise}}$  are listed in Table 5.

## A.5 Validity Metrics

To evaluate consistency and chemical sensibility of the generated CIFs, we conduct a series of validation checks. The methodology is described below.

### Formula Consistency

We check for consistency in the chemical formula printed in different locations within the CIF. Specifically, we ensure that:

- The chemical formula in the `_chemical_formula_sum` tag matches the reduced chemical formula derived from the atomic sites.
- The chemical formula in the `_chemical_formula_structural` tag is consistent with the composition derived from the CIF file.

### Site Multiplicity Consistency

We validate that the total multiplicity of atomic sites is consistent with the stoichiometry derived from the composition. Specifically, we ensure:

- The atom types are specified under the `_atom_site_type_symbol` tag.
- The multiplicity of each atom is provided in the `_atom_site_symmetry_multiplicity` tag.
- The total number of atoms derived from these tags matches the stoichiometry derived from the `_chemical_formula_sum` tag.

### Bond Length Reasonability

To check the reasonableness of bond lengths:

- We use a Voronoi-based nearest-neighbour algorithm implemented in the `CrystalNN` module of `pymatgen` ([Ong et al., 2013](#)) to identify bonded atoms.
- For each bond, the expected bond length is calculated based on the atomic radii and the electronegativity difference between the bonded atoms:
  - If the electronegativity difference is greater than or equal to 1.7, the bond is treated as ionic, and the bond length is based on the cationic and anionic radii.
  - Otherwise, the bond is treated as covalent, and the bond length is based on the atomic radii.
- A bond length reasonableness score  $B$  is computed as the fraction of bonds whose lengths are within  $\pm 30\%$  of the expected lengths.
- A structure passes this test if  $B \geq c_{\text{bond}}$ , where  $c_{\text{bond}} = 1.0$ .

### Space Group Consistency

We validate the space group by:

- Extracting the stated space group from the `_symmetry_space_group_name_H-M` tag.

- Analyzing the space group symmetry using the `SpacegroupAnalyzer` class in `pymatgen` ([Ong et al., 2013](#)), which employs the `spglib` ([Togo and Tanaka, 2018](#)) library.
- Comparing the stated space group with the one determined by the symmetry analysis.

### Overall Validity

A CIF file is deemed valid if all the above checks are satisfied:

- Formula consistency (FM).
- Site multiplicity consistency (SM).
- Bond length reasonableness  $B \geq c_{\text{bond}}$ , where  $c_{\text{bond}} = 1.0$  (BL).
- Space group consistency (SG).

## A.6 Match Rate

The Match Rate (MR) quantifies how many generated structures successfully match their corresponding reference structures, as determined by `StructureMatcher` from the `pymatgen` library ([Ong et al., 2013](#)). Two structures are considered a match if their compositions, lattice parameters, atomic coordinates, and symmetry are sufficiently similar, according to the tolerances set in `StructureMatcher`. For the implementation of deCIFer, we follow the example set by CrystaLLM ([Antunes et al., 2024](#)), using the parameter values:

- `stol` = 0.5: site tolerance, defined as a fraction of the average free length per atom.
- `angle_tol` =  $10^\circ$ : maximum angular deviation tolerance.
- `ltol` = 0.3: fractional length tolerance, meaning the lattice parameters can differ by up to 30% relative to the reference lattice.

`StructureMatcher` compares two structures by:

- Optionally reducing them to primitive (Niggli) cells.
- Verifying that the lattice parameters are within the fractional length tolerance (`ltol`).
- Checking that the angles are within the angle tolerance (`angle_tol`).
- Ensuring that atomic coordinates align within the site tolerance (`stol`), normalized by the average free length per atom.

With these parameters, each generated CIF is compared against its reference CIF\*. If the two structures are deemed structurally equivalent, we count that as a successful match. MR is computed as the fraction of structures in the dataset for which a match is found:

$$\text{MR} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(\text{match}(\text{CIF}_i, \text{CIF}^*_i)), \quad (9)$$

where  $N$  is the total number of structures and  $\mathbf{1}(\cdot)$  is an indicator function that returns 1 if two structures match (according to `StructureMatcher`) and 0 otherwise.

## A.7 Datasets Statistics

Figure 11 illustrates the NOMA dataset. Figure 12 illustrates the statistics of the curated CHILLI-100K ([Friis-Jensen et al., 2024](#)) dataset.

## A.8 Model Architecture- and Training Details

Table 5 provides a concise overview of all hyperparameters and data augmentation settings used for training deCIFer (and its variant U-deCIFer). Below, we describe additional implementation details.

**Hardware Setup** All experiments were conducted on GPUs with sufficient memory to accommodate a batch size of 32 tokenized sequences, each truncated or padded to a context length of 3076. We employed half-precision (float16) to reduce memory usage and improve throughput, ensuring that gradient updates remain numerically stable via built-in automatic mixed-precision.

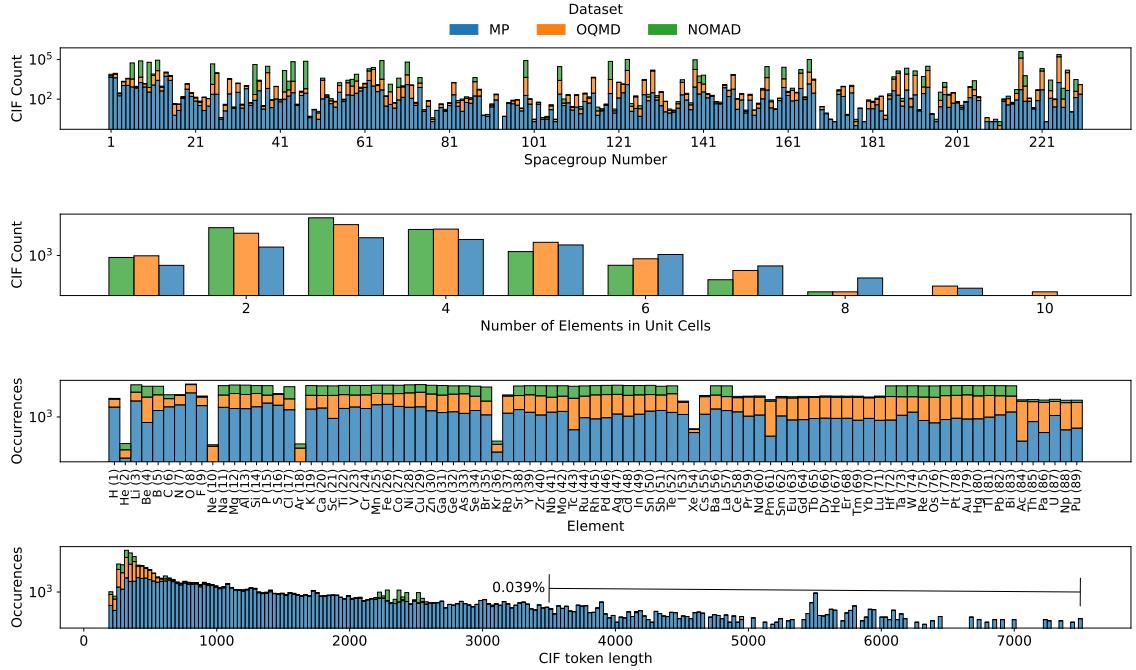


Figure 11: Statistical overview of the NOMA (Antunes et al., 2024) dataset (2,283,346 total samples), showing the distribution of space group frequencies, the number of elements per unit cell, elemental occurrences and CIF token lengths (indicating the percentage of CIFs with larger token sequences than the context length of 3076)

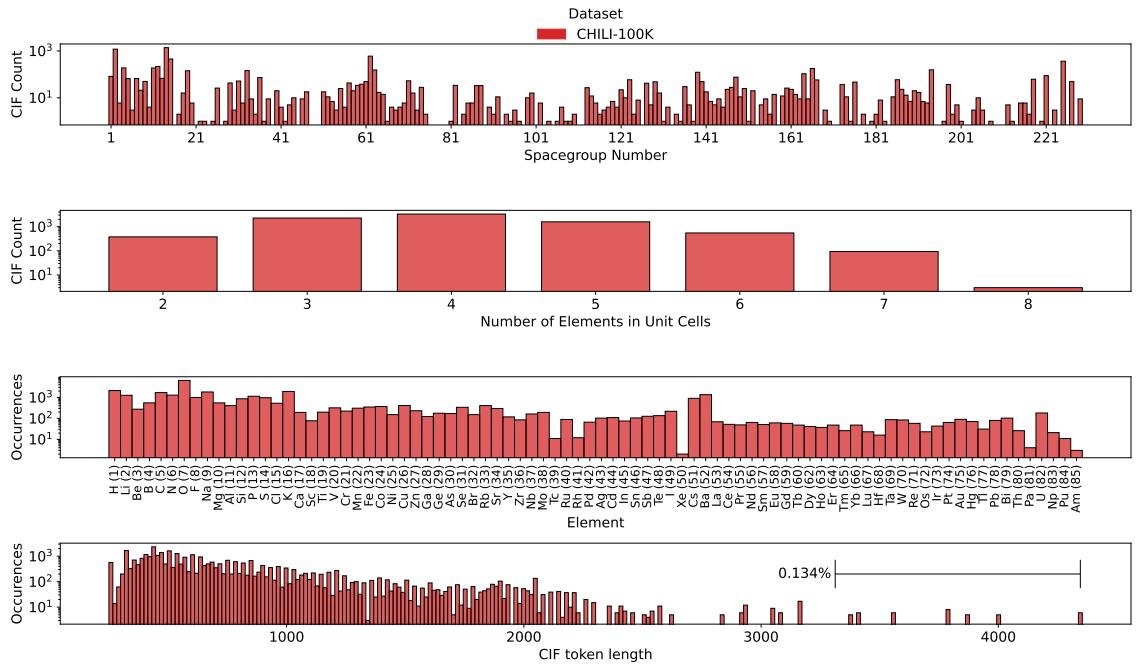


Figure 12: Statistical overview of the curated CHILI-100K (Friis-Jensen et al., 2024) dataset (8201 total samples), showing the distribution of space group frequencies, the number of elements per unit cell, elemental occurrences, and CIF token lengths (indicating the percentage of CIFs with larger token sequences than the context length of 3076).

**Optimizer and Learning Rate Schedule.** We adopt AdamW with a base learning rate of  $1 \times 10^{-3}$ , which is warmed up for 100 steps and then gradually decayed to  $1 \times 10^{-6}$  over 50,000 steps (Table 5). Weight decay is set to 0.1 to regularize model parameters, and we employ gradient

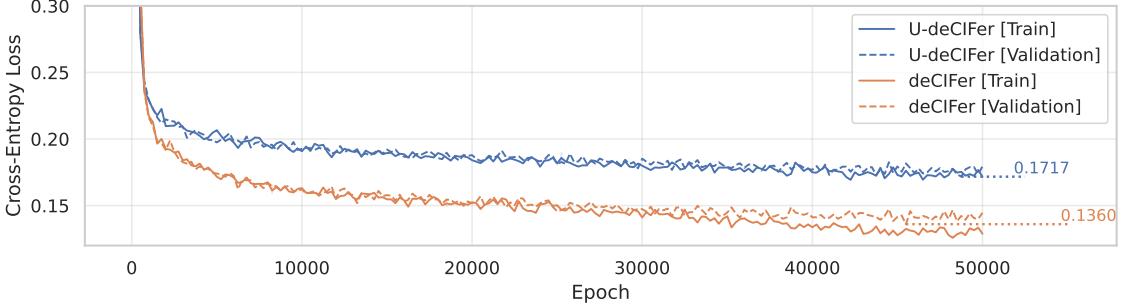


Figure 13: Cross-entropy loss curves for U-deCIFer and deCIFer over 50,000 training iterations, showing progressive reduction in the training and validation losses.

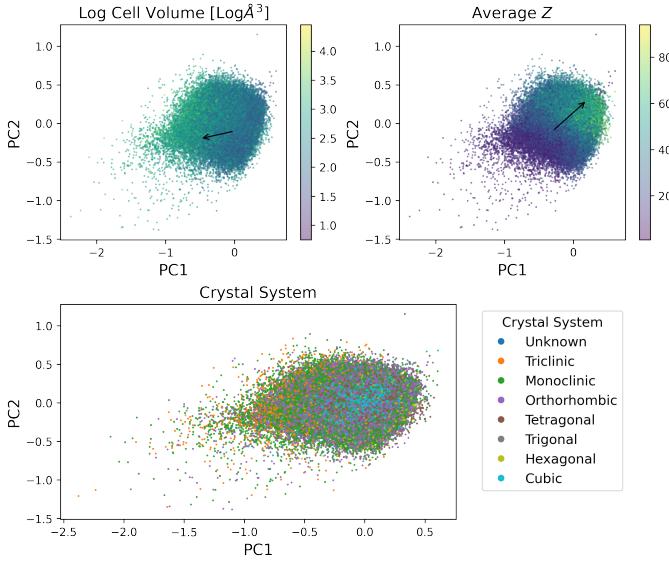


Figure 14: 2D PCA projection of learned PXRD embeddings for 500K training-set samples from NOMA. The three subplots are colored by crystal system, log(cell volume), and average atomic number  $Z$ , illustrating clear gradients that correspond to structural and compositional features as indicated by the arrows.

accumulation (40 steps) to effectively increase the total number of tokens processed per update.

**Transformer Architectural Notes.** The final transformer stack has 8 layers, each with 8 attention heads, and a model dimension of 512 (embedding dimension). The feed-forward blocks inside each layer use a dimension of  $4 \times 512$ , and dropout is set to 0.0 to minimize underfitting. We continue to observe stable convergence in practice despite using no dropout.

**Maximum Iterations and Convergence.** We train for 50,000 iterations, at which point the model’s cross-entropy loss stabilizes, as illustrated in Figure 13. Beyond this range, no significant improvements were observed on validation metrics.

## A.9 PXRD Embedding Space

For completeness, we examined the learned embeddings for 50K random training-set PXRD profiles and applied principle component analysis (PCA) for visualization. As shown in Figure 14, the embeddings form distinct gradients when colored by crystal system, cell-volumes, and constituent atomic numbers  $Z$ , indicating that the model’s PXRD embedding captures relevant structural characteristics, such as symmetry, scale, and elemental composition. These patterns highlight the effectiveness of the conditioning mechanism in encoding meaningful structural information directly from the PXRD input.

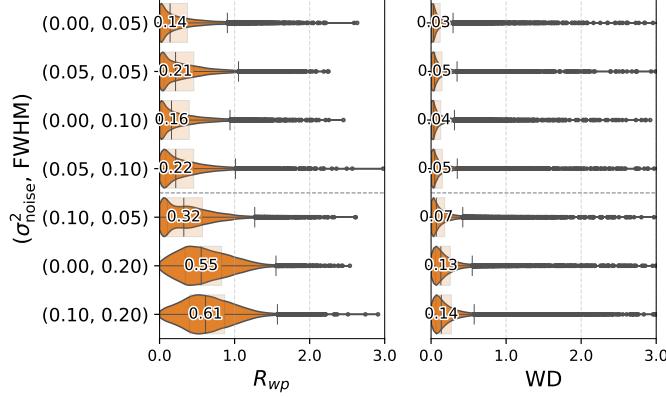


Figure 15: Distribution of  $R_{wp}$  and WD for deCIFer on the NOMA test set, presented as violin plots with overlaid boxplots; the median is shown for each distribution. Presented are four in-distribution transformations of the input PXRD profiles and three out-of-distribution transformations.

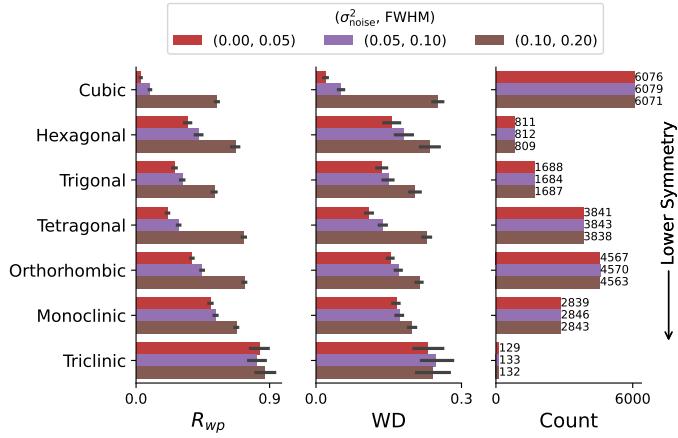


Figure 16: Average metric values by crystal systems for deCIFer on the NOMA test set under two in-distribution transformations of the input PXRD profiles and one out-of-distribution transformation. deCIFer shows better performance for well-represented systems, while rarer, low-symmetry systems lead to worse performance. The right-most plot shows crystal system distribution of the NOMA test set.

## A.10 Additional Results

Figures 15, 16, and 17 provide additional insights into deCIFer’s performance. Table 6 shows a detailed breakdown of validity metrics for the NOMA test set and CHILI-100K test set evaluated on two in-distribution (ID) scenarios and one out-of-distribution (OOD) scenario for the PXRD input.

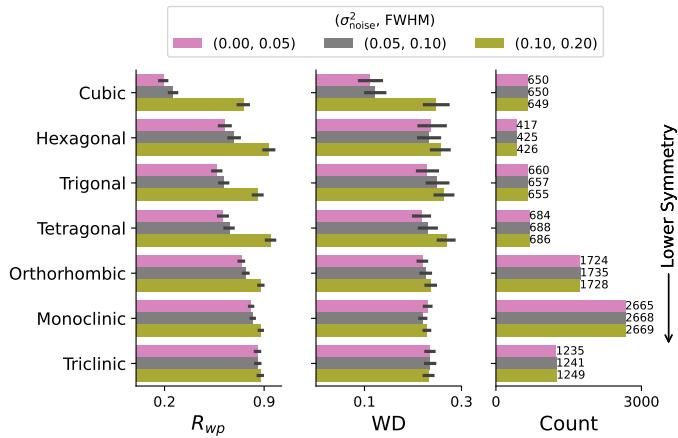


Figure 17: Average metric values by crystal systems for deCIFer on the CHILI-100K test set show better performance for well-represented systems in the training data (NOMA), while low-symmetry systems lead to worse performance. The right-most plot shows crystal system distribution of the CHILI-100K test set, highlighting that CHILI-100K contains a significantly higher proportion of lower-symmetry structures compared to synthetic datasets like NOMA.

Table 4: Supported atoms, CIF tags, space groups, numbers, and special tokens.

Category	Num.	Tokens
Atoms	89	Si C Pb I Br Cl Eu O Fe Sb In S N U Mn Lu Se Tl Hf Ir Ca Ta Cr K Pm Mg Zn Cu Sn Ti B W P H Pd As Co Np Tc Hg Pu Al Tm Tb Ho Nb Ge Zr Cd V Sr Ni Rh Th Na Ru La Re Y Er Ce Pt Ga Li Cs F Ba Te Mo Gd Pr Bi Sc Ag Rb Dy Yb Nd Au Os Pa Sm Be Ac Xe Kr He Ne Ar
CIF Tags	31	data_ loop_ _symmetry_space_group_name_H-M _symmetry_Int_Tables_number _cell_length_a _cell_length_b _cell_length_c _cell_angle_alpha _cell_angle_beta _cell_angle_gamma _cell_volume _atom_site_fract_x _atom_site_fract_y _atom_site_fract_z _atom_site_occupancy _symmetry_equiv_pos_as_xyz _chemical_formula_structural _cell_formula_units_Z _chemical_name_systematic _chemical_formula_sum _atom_site_symmetry_multiplicity _atom_site_attached_hydrogens _atom_site_label _atom_site_type_symbol _atom_site_B_iso_or_equiv _symmetry_equiv_pos_site_id _atom_type_symbol _atom_type_electronegativity _atom_type_radius _atom_type_ionic_radius _atom_type_oxidation_number
Space Groups	230	P6/mmm Imma P4_32_12 P4_2/mnm Fd-3m P3m1 P-3 P4mm P4_332 P4/nnc P2_12_12 Pnn2 Pbca P4_2/n Cm R3m Cmce Aea2 P-42_1m P-42m P2_13 R-3 Fm-3 Cmm2 Pn-3n P6/mcc P-6m2 P3_2 P-3m1 P3_212 I23 P-62m P4_2nm Pma2 Pmma I-42m P-31c Pa-3 Pmmm Pmmm P4_2/nem I4/mcm I-4m2 P3_1 Pcc2 Cmcm I222 Fddd P312 Cccm P6_1 F-43c P6_322 Pm-3 P3_121 P6_4 Ia-3d Pm-3m P2_1/c C222_1 Pc P4/n Pba2 Ama2 Pbcm P31m Pcca P222 P-43n Pccm P6_422 F23 P42_12 C222 Pnnn P6_3cm P4_12_12 P6/m Fmm2 I4_1/a P4/mbm Pmn2_1 P4_2bc P4_22_12 I-43d I4/m P4bm Fdd2 P3 P6_122 Pnc2 P4_2/mcm P4_122 Cmc2_1 P-6c2 R32 P4_1 P4_232 Pnna P422 Pban Cc I4_122 P6_3/m P6_3mc I4_1/amd P4_2 P4/nmm Pmna P4/m Fm-3m P4/mmm Imm2 P4/ncc P-62c Ima2 P6_5 P2/c P4/nbm Ibam P6_522 P6_3/mmc I4/mmm Fmmm P2/m P-4b2 I-4 C2/m P4_2/mmc P4 Fd-3c P4_3 P2_1/m I-43m P-42c F4_132 Pm Pccn P-4n2 P4_132 P23 I4cm R3c Amm2 Imm2 Iba2 I4 Fd-3 P1 Pbam P4_2/nbc Im-3 P4_2/nmm Pmc2_1 P-31m R-3m Ia-3 P622 F222 P2 P-1 Pmm2 P-4 Aem2 P6_222 P-3c1 P4_322 I422 Pnma P6_3 P3c1 Pn-3 P4nc P-6 P4/mcc I2_12_12_1 P4_2/mbc P31c Ccc2 P4_2/nmc P6_3/mcm C2 Pbc2 P-4c2 I4_1cd P2_1 P3_112 P4_2mc Pn-3m C2/c R3 P-43m I432 P222_1 I-42d I-4c2 P6cc P6_2 P3_221 P321 Pca2_1 I4_1/acd I4_132 F432 Pna2_1 Ccce Ibca P4/mnc I4_1md P2_12_12_1 R-3c I2_13 P-4m2 Pm-3n I4mm F-43m Pnnm P-42_1c Cmmm P6mm P4_2cm P4_2/m Im-3m Fm-3c I4_1 P4cc Cmme
Numbers	10	1 2 3 4 5 6 7 8 9 0
Special	13	x y z . ( ) ' , <space> <newline> <unk> <pad> <cond>

Table 5: Training configuration for deCIFer and U-deCIFer.

PXRD Transformation Training Parameters	Value
Wavelength ( $\lambda$ )	Cu-K $\alpha$ (1.5406 Å)
Q-grid ( $Q_{\min}$ , $Q_{\max}$ , $Q_{\text{step}}$ )	(0.0, 10.0, 0.01)
FWHM	$\mathcal{U} \sim (0.001, 0.10)$
Mixing Factor ( $\eta$ )	0.5
Noise Magnitude	$\mathcal{U} \sim (0.001, 0.05)$
Model / Training Parameters	Value
Optimizer	AdamW
Learning Rate	$1 \times 10^{-3}$
Warmup Steps	100
Decay Steps	50,000
Minimum Learning Rate	$1 \times 10^{-6}$
Weight Decay	0.1
Batch Size	32
Gradient Accumulation Steps	40
Maximum Iterations	50,000
Embedding Dimension ( $n_{\text{embd}}$ )	512
Layers ( $n_{\text{layer}}$ )	8
Attention Heads ( $n_{\text{head}}$ )	8
Conditioning Model Layers ( $n_{\text{c-layers}}$ )	2
Conditioning Model Hidden Size	512
Sequence Length (block_size)	3076
Precision	float16
Dropout	0.0

Table 6: Validity of generated CIFs for the CHILI-100K test set using deCIFer. Abbreviations: Form = formula validity, SG = space group validity, BL = bond length validity, SM = site multiplicity validity. Overall validity (Val.) is calculated as the percentage of CIFs that satisfy all four validity metrics simultaneously. Match rate (MR) represents the percentage of generated CIFs that replicate the reference CIF.

Dataset	$(\sigma_{\text{noise}}^2, \text{FWHM})$	FORM (%) $\uparrow$	SG (%) $\uparrow$	BL (%) $\uparrow$	SM (%) $\uparrow$		Val. (%) $\uparrow$		MR (%) $\uparrow$
NOMA	ID: (0.00, 0.05)	99.68	99.21	94.37	99.55		93.73		91.50
	ID: (0.05, 0.10)	99.64	99.18	94.39	99.55		93.77		89.28
	OOD: (0.10, 0.20)	99.60	99.87	92.60	99.49		91.66		77.66
CHILI-100K	ID: (0.00, 0.05)	95.98	97.88	42.61	94.58		41.83		37.34
	ID: (0.05, 0.10)	96.17	98.22	41.50	94.47		40.95		35.97
	OOD: (0.10, 0.20)	95.80	98.42	34.11	93.91		33.62		26.09