# Invariant Tokenization of Crystalline Materials for Language Model Enabled Generation

**Keqiang Yan**[1], **Xiner Li**[1], **Hongyi Ling**[1], **Kenna Ashen**[1], **Carl Edwards**[2],
**Raymundo Arróyave**[1], **Marinka Zitnik**[3], **Heng Ji**[2], **Xiaofeng Qian**[1], **Xiaoning Qian**[1],
**Shuiwang Ji**[1]
[1]Texas A&M University, College Station, TX 77843, USA
[2]University of Illinois Urbana-Champaign, Champaign, IL 61820, USA
[3]Harvard University, Boston, MA 02115, USA
{keqiangyan,lxe,hongyiling,kashen13,rarroyave,feng,xqian,sji}@tamu.edu
{cne2,hengji}@illinois.edu
marinka@hms.harvard.edu

## Abstract

We consider the problem of crystal materials generation using language models (LMs). A key step is to convert 3D crystal structures into 1D sequences to be processed by LMs. Prior studies used the crystallographic information framework (CIF) file stream, which fails to ensure $SE(3)$ and periodic invariance and may not lead to unique sequence representations for a given crystal structure. Here, we propose a novel method, known as Mat2Seq, to tackle this challenge. Mat2Seq converts 3D crystal structures into 1D sequences and ensures that different mathematical descriptions of the same crystal are represented in a single unique sequence, thereby provably achieving $SE(3)$ and periodic invariance. Experimental results show that, with language models, Mat2Seq achieves promising performance in crystal structure generation as compared with prior methods.

## 1 Introduction

Discovering crystalline materials with desired properties is valuable for advancing a variety of technological sectors [1, 2, 3, 4, 5, 6, 7, 8, 9]. However, current materials discovery relies on trial-and-error wet-lab experimental methods that are time-consuming and expensive. Computational methods based on advanced quantum mechanical approaches such as first-principles density functional theory (DFT) have sped up this process in the last two decades. Nevertheless, the high computational cost of these methods ranging from $O(n^3)$ to $O(n^7)$, where $n$ is number of electrons in a material system, poses a significant challenge for high-throughput screening of the infinite materials space.

Recent advances in large language models (LLMs) have showcased substantial capabilities in real-time question answering and image generation. However, the application of these models to atomistic systems, including molecular and material structures, remains relatively underexplored. Specifically, while recent studies have attempted to represent molecules and crystals using XYZ and CIF structure file formats [10, 11, 12], they have not adequately addressed the critical issues of uniqueness and invariance in sequence representations of 3D crystal structures, which play key roles for successful LLM-based discovery of novel crystalline materials with discussions provided in Sec. 2.2.

The concept of uniqueness [13, 14] in crystal sequences mandates a bijective relationship between crystal structures and their sequence representations. Current approaches using CIF files directly in crystal LLMs violate this uniqueness criterion, leading to significantly varied sequences for identical crystals, particularly in terms of atom ordering and fractional coordinates within the unit cell as illustrated in Figure 1. This lack of unique sequence representations means that an infinite number of

CIF files can describe the same crystal, necessitating extensive augmentation during LLM training to recognize these equivalences. Without such augmentation, previous CIF file based methods [10, 11] essentially train crystal LLMs using only a limited set of crystal structure snapshots.

In this work, we develop a framework for creating unique and complete crystal sequence representations, followed by the construction of a material LLM capable of generating novel crystal structures with desired properties of interest. To accomplish this, several challenges must be addressed. First, unlike molecular structures that contain a finite number of atoms, each crystal structure consists of an infinite number of atoms through a periodic translation of a finite set of atoms in a unit cell along the direction of three lattice vectors in three-dimensional (3D) space. Consequently, there are numerous different unit cells for the same crystal as shown in Figure 1. A unique and invariant unit cell must therefore be selected for each crystal. Second, it is crucial that this unit cell can be represented in a one-dimensional (1D) sequence that maintains invariance under arbitrary rotations and ensures completeness, allowing the full reconstruction of the crystal structure from its sequence representation. To tackle these complexities, we introduce Mat2Seq, a method that systematically transforms 3D crystal structures into 1D sequences. This is achieved by first identifying $SO(3)$ equivariant unit cells and subsequently converting these into $SE(3)$ invariant sequences. By integrating these unique and complete sequences into LLMs, we develop conditional generation capabilities for novel crystal structures. Our experimental results in crystal structure prediction and crystal discovery with desired properties validate the efficacy of Mat2Seq.

## 2 Preliminaries and related work

### 2.1 Crystal structure generation

In this work, we study the problem of generating 3D crystals from scratch. Unlike small molecules, crystalline materials are defined by a unit cell which contains a set of atoms repeated infinitely across three-dimensional space along three periodic lattice vectors. Following notations of ComFormer [14], each crystal structure is represented by $\mathbf{M} = (\mathbf{A}, \mathbf{P}, \mathbf{L})$, where $\mathbf{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_n] \in \mathbb{R}^{d_a \times n}$ represents the $d_a$-dimensional feature vectors of $n$ atoms within the unit cell, $\mathbf{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_n] \in \mathbb{R}^{3 \times n}$ represents the 3D Euclidean positions of these $n$ atoms, and $\mathbf{L} = [\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \boldsymbol{\ell}_3] \in \mathbb{R}^{3 \times 3}$ specifies three periodic lattice vectors, representing the repeating patterns of the unit cell in 3D space. The infinite structure of a given crystal $\mathbf{M} = (\mathbf{A}, \mathbf{P}, \mathbf{L})$ is formalized as $\hat{\mathbf{P}} = \{\hat{\boldsymbol{p}}_i | \hat{\boldsymbol{p}}_i = \boldsymbol{p}_i + k_1 \boldsymbol{\ell}_1 + k_2 \boldsymbol{\ell}_2 + k_3 \boldsymbol{\ell}_3, \ k_1, k_2, k_3 \in \mathbb{Z}, i \in \mathbb{Z}, 1 \leq i \leq n\}, \hat{\mathbf{A}} = \{\hat{\boldsymbol{a}}_i | \hat{\boldsymbol{a}}_i = \boldsymbol{a}_i, i \in \mathbb{Z}, 1 \leq i \leq n\}$, where $\hat{\mathbf{P}}$ denotes the positions of atoms and their infinite repeats in the 3D space, and $\hat{\mathbf{A}}$ denotes the corresponding feature vectors. This work targets two primary tasks:

**Crystal structure generation**: Using a dataset of 3D crystals $\{\mathbf{M}_j\}_{j=1}^m$, we aim to develop a structure generative model $p_\theta(\cdot | \mathbf{A}_j)$ to synthesize valid and stable 3D crystal structures for input compositions.

**Conditional generation**: With a dataset $\{(\mathbf{M}_j, s_j)\}_{j=1}^m$, where $s_j$ denotes a specific property of $\mathbf{M}_j$, we aim to establish a conditional generative model $p_\theta(\cdot | s)$ to generate 3D crystal structures possessing the property $s$.

Generative models for crystal structures must address complex geometric requirements, such as ensuring periodic and unit cell $SE(3)$ invariance [8, 15, 14, 16]. The ideal models should consider $\mathbf{M} = (\mathbf{A}, \mathbf{P}, \mathbf{L})$ and its equivalents under periodic transformations or rotations and translations as identically probable [15, 14]. CDVAE [15] achieves periodic and $SE(3)$ invariance by encoding crystal structures into $SE(3)$ invariant latent features but faces limitations in restricting the loss function to be periodic invariant. This limitation is later rectified in SyMat [16]. Beyond this, DiffCSP [17] is specifically optimized for crystal structure prediction that generates stable crystal structures for given compositions. In contrast, MatterGen [18] models $(\mathbf{A}, \mathbf{P}, \mathbf{L})$ in an equivariant manner by gradually corrupting them into known distributions.

While the generation of 3D crystal structures as 3D point clouds is well-established, applying powerful language models to this task is novel and nontrivial as discussed in Sec. 2.2. Our approach uniquely transforms infinite 3D crystal structures into $SE(3)$ and periodic invariant sequences, demonstrating that language models can effectively generate crystal structures with high performance.
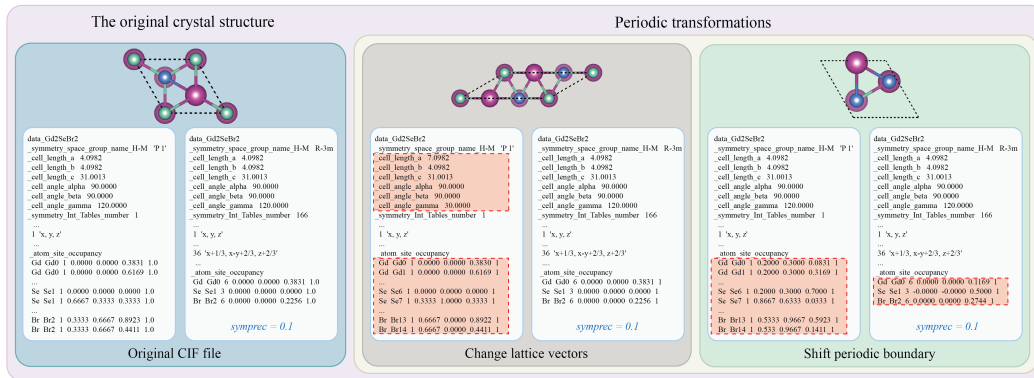
Figure 1: Limitations of directly using CIF files in achieving unique crystal sequence representations. This figure demonstrates variations in CIF files, either with or without symmetry control command denoted as "symprec", for the same crystal structure subjected to periodic transformations. Changes in the CIF contents are highlighted in red. Periodic transformations can significantly alter the unit cell structures, resulting in distinct CIF files *differed by fractional coordinates, atom ordering, and lattice parameters* for the same underlying crystal.

## 2.2 Language model for 3D crystal structures

Enabling language models to generate 3D crystal structures necessitates transforming these structures into sequence representations. Several studies have employed CIF files as sequence inputs for this purpose [10, 11, 12]. These approaches convert a set of 3D crystals $\{\mathbf{M}_j\}_{j=1}^m$ into a dataset of sequences, $\boldsymbol{C} = \{C_1, C_2, \cdots, C_m\}$, where each $C_j$ represents the CIF text of $\mathbf{M}_j$. The sequence $C_j = \{c_1, c_2, \cdots, c_{n_j}\}$ of length $n_j$ are composed of tokens $c_i$ from a predefined vocabulary $V$. Autoregressive language models are either trained [10, 12] or fine-tuned [11] to encode these sequences by maximizing the conditional probabilities of each token given its predecessors, $p(C_i|\theta) = \prod_{j=1}^{n_i} p(c_j|c_1 : c_{j-1}; \theta)$, with the objective to maximize the probabilities of the dataset $p(\boldsymbol{C}|\theta) = \prod_{i=1}^m p(C_i; \theta)$. During generation, novel crystal sequences are produced from $p(\theta)$ in an autoregressive manner.

However, ideal crystal generative models should consider $\mathbf{M} = (\mathbf{A}, \mathbf{P}, \mathbf{L})$ and its equivalents under periodic transformations or rotations and translations as identically probable, a criterion unmet by current models due to the variability in CIF text descriptions for the same crystal. For example, there are different sequence representations $C_i = \{c_1, c_2, \cdots, c_{n_i}\}$ and $C_i' = \{c_1', c_2', \cdots, c_{n_i'}'\}$ for the same crystal, yet ideally, $\prod_{j=1}^{n_i} p(c_j|c_1 : c_{j-1}; \theta) = \prod_{j=1}^{n_i'} p(c_j'|c_1' : c_{j-1}'; \theta)$ should hold. Despite being highly non-trivial, this is overlooked by previous studies. This issue is illustrated through the failure cases of CIF-based crystal language models [10, 11, 12] in Figure 1.

**Differences with previous works**. It is crucial to recognize that prior works [10, 11, 12] using CIF files as input violate invariance and generate a multitude of different sequence representations for the same crystal. Consequently, these models fail to consistently recognize $\mathbf{M} = (\mathbf{A}, \mathbf{P}, \mathbf{L})$ and its equivalents under periodic transformations including shifting periodic boundaries and changing lattice vectors as identically probable. While intensive data augmentation can mitigate this issue, it significantly burdens the training process of language models. In contrast, our Mat2Seq approach maps all equivalent crystals to a unique sequence, naturally ensuring that all equivalents are considered equally probable without the need for data augmentation.

## 3 Tokenization of 3D crystal materials

In this section, we demonstrate **Mat2Seq** for transforming 3D crystal structures into 1D sequences that adhere to the principles of $SE(3)$ invariance, periodic invariance, and completeness. We first discuss requirements for ideal crystal sequence representations in Sec. 3.1 and provide formal definitions of unit cell $SE(3)$ invariance, periodic invariance, and completeness. We then show how to design crystal sequence representations to achieve these requirements in Sec. 3.2 and 3.3, with the pipeline
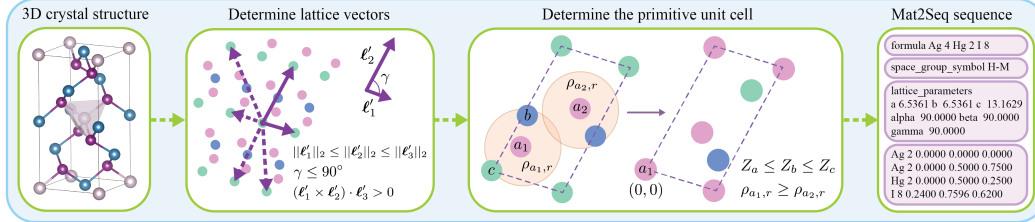
Figure 2: The pipeline of Mat2Seq that converts 3D crystal structures into unique crystal sequences. Mat2Seq first determines $SO(3)$ equivariant and periodic invariant lattice vectors using Niggli cell reduction [19], then determines the primitive unit cell. After that, Mat2Seq converts the determined $SO(3)$ equivariant and periodic invariant primitive cells into $SE(3)$ and periodic invariant sequences.

shown in Figure 2. We then provide property proofs of **Mat2Seq** in Sec. 3.4, and demonstrate the combination of **Mat2Seq** and language models for crystal structure generation in Sec. 3.5.

## 3.1 Requirements for ideal crystal sequence representations

**Uniqueness**. Ideal crystal sequences need to satisfy uniqueness which ensures a deterministic mapping of a crystal structure $\mathbf{M} = (\mathbf{A}, \mathbf{P}, \mathbf{L})$ and its equivalents, as transformed by periodic and $SE(3)$ operations, to a singular, unique sequence. However, to the best of our knowledge, none of previous studies [10, 11, 12] achieves uniqueness for crystal sequence representations.

When a sequence inherently satisfies this criterion, it guarantees that all structural equivalents are recognized as equally probable by language models, thus eliminating the need for any manual adjustments for language models or extensive augmentation during the training phase. Following ComFormer [14], we further define the necessary invariances for ideal sequence representations.

**Definition 1** (Unit Cell $SE(3)$ Invariance). *A function $f : (\mathbf{A}, \mathbf{P}, \mathbf{L}) \rightarrow \mathcal{X}$ is unit cell $SE(3)$ invariant if, for any rotation transformation $\mathbf{R} \in \mathbb{R}^{3 \times 3}, |\mathbf{R}| = 1$ and translation transformation $\mathbf{b} \in \mathbb{R}^3$ , we have $f(\mathbf{A}, \mathbf{P}, \mathbf{L}) = f(\mathbf{A}, \mathbf{RP} + b, \mathbf{RL})$.*

**Definition 2** (Periodic Invariance). *A function $f : (\mathbf{A}, \mathbf{P}, \mathbf{L}) \rightarrow \mathcal{X}$ is periodic invariant if, for any possible minimum unit cell representations $\mathbf{M}' = (\mathbf{A}', \mathbf{P}', \mathbf{L}')$ representing a given infinite crystal structure $(\hat{\mathbf{P}}, \hat{\mathbf{A}})$, we have $f(\mathbf{A}, \mathbf{P}, \mathbf{L}) = f(\mathbf{A}', \mathbf{P}', \mathbf{L}')$.*

This definition captures two periodic transformations that result in distinct minimum unit cell representations of the same crystal structure: (1) shifting periodic boundaries, and (2) altering periodic patterns while maintaining the same unit cell volume as shown in Figure 1.

**Completeness**. Ideal crystal sequences need to be complete, allowing the full reconstruction of 3D crystal structures from their 1D sequences.

## 3.2 Determination of $SO(3)$ equivariant unit cells

We propose Mat2Seq that converts 3D crystal materials into 1D sequences by (1) determining the $SO(3)$ equivariant unit cells and (2) converting $SO(3)$ equivariant unit cells into $SE(3)$ invariant sequences. In this section, we demonstrate how to determine the $SO(3)$ equivariant unit cell of a given crystal structure $\mathbf{M} = (\mathbf{A}, \mathbf{P}, \mathbf{L})$.

**Determination of lattice vectors**. Let's start with an arbitrary minimum (primitive) cell with vectors $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \boldsymbol{\ell}_3$. A set of $SO(3)$ equivariant lattice vectors can be determined by Niggli cell reduction [19] as follows. The three lattice vectors $\boldsymbol{\ell}'_1, \boldsymbol{\ell}'_2, \boldsymbol{\ell}'_3$ of a new primitive cell can be represented in terms of the original ones as $\boldsymbol{\ell}'_i = k_{i,1}\boldsymbol{\ell}_1 + k_{i,2}\boldsymbol{\ell}_2 + k_{i,3}\boldsymbol{\ell}_3, \ k_{i,1}, k_{i,2}, k_{i,3} \in \mathbb{Z}$. Then, a unique and $SO(3)$ equivariant cell can be determined by choosing three shortest non-planar vectors $\boldsymbol{\ell}'_1, \boldsymbol{\ell}'_2, \boldsymbol{\ell}'_3$ that form a right-hand system using three Euclidean vector lengths and three relative angles (*i.e.* six lattice parameters) [14, 19].

**Determination of primitive unit cells**. After lattice vectors $\mathbf{L}_u = [\boldsymbol{\ell}'_1, \boldsymbol{\ell}'_2, \boldsymbol{\ell}'_3]$ have been determined, a unique unit cell is determined as follows. For different unit cells with the same lattice vectors $\boldsymbol{\ell}'_1, \boldsymbol{\ell}'_2, \boldsymbol{\ell}'_3$, they differ by the origin of the cell. We determine the origin of unit cell by: **1.** smallest atomic number; **2.** local densities of atoms in the cell that are E(3) invariant measurements defined

by $\rho_{i,r} = \sum_{j \in \mathcal{N}(i,r)} Z_j$, where $\mathcal{N}(i,r)$ denotes the neighbors of atom $i$ within the radius $r$, and $Z_j$ denotes the atomic number of neighboring atom $j$; and **3.** densities along three lattice vectors $\boldsymbol{\ell}_1', \boldsymbol{\ell}_2', \boldsymbol{\ell}_3'$ that are $SE(3)$ invariant measurements defined as the following,

$$\rho_{i,r}^{\boldsymbol{\ell}_1'} = \sum_{j \in \mathcal{N}^{\boldsymbol{\ell}_1'}(i,r)} Z_j, \rho_{i,r}^{\boldsymbol{\ell}_2'} = \sum_{j \in \mathcal{N}^{\boldsymbol{\ell}_2'}(i,r)} Z_j, \rho_{i,r}^{\boldsymbol{\ell}_3'} = \sum_{j \in \mathcal{N}^{\boldsymbol{\ell}_3'}(i,r)} Z_j, \tag{1}$$

where $\mathcal{N}^{\boldsymbol{\ell}_1'}(i,r), \mathcal{N}^{\boldsymbol{\ell}_2'}(i,r), \mathcal{N}^{\boldsymbol{\ell}_3'}(i,r)$ denote the neighbors of atom $i$ within the radius $r$ along $\boldsymbol{\ell}_1', \boldsymbol{\ell}_2', \boldsymbol{\ell}_3'$ that satisfy $\boldsymbol{p}_{\text{frac},j}^x > \boldsymbol{p}_{\text{frac},i}^x, \boldsymbol{p}_{\text{frac},j}^y > \boldsymbol{p}_{\text{frac},i}^y$, and $\boldsymbol{p}_{\text{frac},j}^z > \boldsymbol{p}_{\text{frac},i}^z$, correspondingly. Here, $\boldsymbol{p}_{\text{frac},j}$ denotes the fractional coordinate of atom $j$ as follows,

$$\boldsymbol{p}_j = \boldsymbol{p}_{\text{frac},j}^x \boldsymbol{\ell}_1' + \boldsymbol{p}_{\text{frac},j}^y \boldsymbol{\ell}_2' + \boldsymbol{p}_{\text{frac},j}^z \boldsymbol{\ell}_3'. \tag{2}$$

After the unique atom $i$ is chosen by **1.**, **2.**, and **3.**, the primitive unit cell is determined by moving this atom to the origin through the following transformation applied to all other fractional coordinates,

$$\boldsymbol{p}_{\text{frac},j}^{x}{}' = (\boldsymbol{p}_{\text{frac},j}^x - \boldsymbol{p}_{\text{frac},i}^x) \bmod 1, \tag{3}$$

$$\boldsymbol{p}_{\text{frac},j}^{y}{}' = (\boldsymbol{p}_{\text{frac},j}^y - \boldsymbol{p}_{\text{frac},i}^y) \bmod 1, \tag{4}$$

$$\boldsymbol{p}_{\text{frac},j}^{z}{}' = (\boldsymbol{p}_{\text{frac},j}^z - \boldsymbol{p}_{\text{frac},i}^z) \bmod 1, \tag{5}$$

where the resultant $\mathbf{P}_{\text{frac},u} = [\boldsymbol{p}_{\text{frac},1}', \boldsymbol{p}_{\text{frac},2}', \cdots, \boldsymbol{p}_{\text{frac},n}'] \in \mathbb{R}^{n \times 3}$, and $\mathbf{P}_u = \mathbf{L}_u \cdot \mathbf{P}_{\text{frac},u}$.

**Remarks**. By determining lattice vectors and primitive unit cells, a $SO(3)$ equivariant and periodic invariant unit cell $(\mathbf{A}_u, \mathbf{P}_u, \mathbf{L}_u)$ is determined for a given crystal material, which means after applying arbitrary $SO(3)$ transformations $\mathbf{R} \in \mathbb{R}^{3 \times 3}, |\mathbf{R}| = 1$ with $\mathbf{b} \in \mathbb{R}^3$ and periodic transformations, the resultant unit cell will transform to $(\mathbf{A}_u, \mathbf{RP}_u, \mathbf{RL}_u)$. We then demonstrate how to construct the $SE(3)$ invariant crystal sequence representation based on $(\mathbf{A}_u, \mathbf{P}_u, \mathbf{L}_u)$ in the following section.

### 3.3 $SE(3)$ **invariant crystal sequence representations**

Given determined $SO(3)$ equivariant and periodic invariant unit cells $\mathbf{M} = (\mathbf{A}_u, \mathbf{P}_u, \mathbf{L}_u)$, we aim to represent them by $SE(3)$ and periodic invariant sequences that are complete to guarantee the full reconstruction of crystal structures. As discussed in Sec. 2.2, a fundamental requirement for crystal sequence representations is uniqueness which stipulates that (1) different crystal structures must correspond to distinct sequences and (2) a given crystal should yield the same sequence representation across all possible structural descriptions $\mathbf{M} = (\mathbf{A}, \mathbf{P}, \mathbf{L})$. Previous works [10, 11, 12] directly use CIF files that

Figure 3: Converting determined unit cells into invariant crystal sequences.

breaks the uniqueness as shown in Figure 1 as the input sequence for LMs. Therefore, these methods require tremendous augmentation effort during the training process to make LMs aware that these different sequences refer to the same crystal structure. Additionally, due to the $O(n^2)$ computational cost of decoder-only language models where $n$ is the window length, the sequence representations of crystal structures need to be efficient to support shorter window length and ease the burden of training and inference processes.

Rather than relying on CIF files, our approach extracts complete geometric information from the uniquely determined $SO(3)$ equivariant unit cells $\mathbf{M} = (\mathbf{A}_u, \mathbf{P}_u, \mathbf{L}_u)$. We first follow CIF files and use six invariant lattice parameters including $a = ||\boldsymbol{\ell}_1||_2, b = ||\boldsymbol{\ell}_2||_2, c = ||\boldsymbol{\ell}_3||_2$ and $\alpha, \beta, \gamma$ that are three bond angles between $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \boldsymbol{\ell}_3$ to represent $\mathbf{L}_u$ as invariant sequences. Subsequently, we incorporate space group information of the crystal structures to reduce sequence length, documenting the **irreducible atom sets** alongside the corresponding symmetry transformations to facilitate the recovery of inner cell structures $(\mathbf{A}_u, \mathbf{P}_u)$. We then represent atom $i$ in the **irreducible atom sets** by the atom type and fractional coordinates $[Z_i, \boldsymbol{p}_{\text{frac},i}^x, \boldsymbol{p}_{\text{frac},i}^y, \boldsymbol{p}_{\text{frac},i}^z]$. It is worth noting that fractional coordinates of atom bases are invariant instead of equivariant, due to that $\mathbf{P} = \mathbf{L} \cdot \mathbf{P}_{\text{frac}}$ where $\cdot$ denotes matrix product and $\mathbf{RP} = \mathbf{RL} \cdot \mathbf{P}_{\text{frac}}$. The unique ordering of atoms within the irreducible set is determined by their atomic numbers, the number of duplicates recoverable from each atom, and their fractional coordinates, thereby ensuring consistency in the Mat2Seq sequence representations.
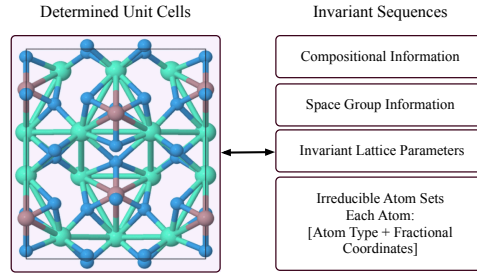
5

Table 1: Uniqueness verification compared with previous crystal language models [10, 11, 12].

| Uniqueness - MP20 [15] | CIF [10, 11] | CIF with symmetry [12] | Mat2Seq |
|---|---|---|---|
| Success rate ↑ | 0% | 30% | **100**% |

## 3.4 Properties and proofs

**Uniqueness**. Different from previous crystal sequence representations [10, 11, 12], Mat2Seq sequence representations guarantee that all possible mathematical descriptions $\mathbf{M} = (\mathbf{A}, \mathbf{P}, \mathbf{L})$ for a given crystal will have a unique sequence representation and the same probability in the follow-up generative modeling process. We prove the uniqueness of Mat2Seq sequences as follows.

**Lemma 1.** *A sequence mapping function $f : (\mathbf{A}, \mathbf{P}, \mathbf{L}) \to \mathcal{X}$ is unique, if function $f$ is periodic invariant and unit cell $SE(3)$ invariant.*

*Proof.* For a given crystal structure, a variety of structural descriptions $(\mathbf{A}, \mathbf{P}, \mathbf{L})$ can be obtained by applying $SO(3)$ rotations $(\mathbf{A}, \mathbf{RP} + b, \mathbf{RL})$ and periodic transformations $(\mathbf{A}', \mathbf{P}', \mathbf{L}')$ [14, 15] including (1) shifting periodic boundaries, and (2) altering periodic patterns while maintaining the same unit cell volume.

For a sequence mapping function $f : (\mathbf{A}, \mathbf{P}, \mathbf{L}) \to \mathcal{X}$ that satisfies periodic invariance and unit cell $SE(3)$ invariance, we have $f(\mathbf{A}, \mathbf{P}, \mathbf{L}) = f(\mathbf{A}, \mathbf{RP} + b, \mathbf{RL}) = f(\mathbf{A}', \mathbf{P}', \mathbf{L}')$ according to the definition provided in Sec. 3.1. Hence, a periodic invariant and unit cell $SE(3)$ invariant function is naturally unique for crystal structures. □

**Lemma 2.** *Mat2Seq : $(\mathbf{A}, \mathbf{P}, \mathbf{L}) \to \mathcal{X}$ is periodic invariant and unit cell $SE(3)$ invariant.*

*Proof.* To prove that Mat2Seq is periodic invariant and unit cell $SE(3)$ invariant, we first show that the unit cell $(\mathbf{A}_u, \mathbf{P}_u, \mathbf{L}_u)$ determined in Sec. 3.2 is $SO(3)$ equivariant and periodic invariant. In other words, after applying arbitrary $SO(3)$ transformations $\mathbf{R} \in \mathbb{R}^{3 \times 3}, |\mathbf{R}| = 1$ with $\mathbf{b} \in \mathbb{R}^3$ and periodic transformations, the resultant unit cell should change from $(\mathbf{A}_u, \mathbf{P}_u, \mathbf{L}_u)$ to $(\mathbf{A}_u, \mathbf{RP}_u, \mathbf{RL}_u)$, with detailed proof provided in Appendix A.1. We then prove that the obtained sequence representation is $SE(3)$ and periodic invariant in Appendix A.1. □

**Completeness**. Crystal sequence representations need to be complete to guarantee the full reconstruction of crystal structures from sequences.

**Lemma 3.** *Crystal structure $(\mathbf{A}, \mathbf{P}, \mathbf{L})$ can be fully recovered from Mat2Seq sequence $\mathcal{X}$.*

*Proof.* $\mathbf{A}$ can be directly recovered from atom types in Mat2Seq sequences, and $\mathbf{L}$ can be recovered by three lattice lengths $a, b, c$ and three lattice angles $\alpha, \beta, \gamma$. Then, fractional coordinates of all atoms in the unit cell can be fully recovered by applying recorded space group transformations $[(\mathbf{R}_1, \mathbf{b}_1), \cdots, (\mathbf{R}_s, \mathbf{b}_s)]$ to fractional coordinates of the irreducible atom set $\mathbf{P}_{\text{frac,irr}}$ as follows,

$$\mathbf{P}_{\text{frac}} = (\mathbf{P}_{\text{frac,irr}} \oplus_{1 \leq i \leq s} (\mathbf{R}_i \mathbf{P}_{\text{frac,irr}} + \mathbf{b}_i) \bmod 1)_{\text{set}}, \tag{6}$$

where $\oplus$ represents concatenation and $()_{\text{set}}$ represents the operation of removing duplicate entries. $\mathbf{P} = \mathbf{L} \cdot \mathbf{P}_{\text{frac}}$ can then be fully recovered. □

**Remarks**. Based on **Lemma 1**, **2**, and **3**, Mat2Seq sequences satisfy uniqueness for crystal structures and can fully reconstruct corresponding 3D crystal structures. We also provide experimental verification on the MP20 [15] training set for the uniqueness of Mat2Seq and previous crystal language models [10, 11, 12] in Table 1 by applying periodic transformation that shifts the periodic boundaries. The evaluation is conducted as follows: (1) for each crystal structure in the MP20 dataset, we apply a transformation that shifts the periodic boundaries, generating a different unit cell representation while keeping the crystal structure unchanged; (2) we then encode both the original and transformed structures into 1D sequence representations using different methods and compare the resulting sequences. If there is a mismatch (e.g., differences in the coordinates or atom type for the first atom), it is considered a failure; (3) finally, we compute the success rate across the entire dataset.

Table 2: Comparison of Match Rate (%) and RMSE across different models and crystal datasets.

| Model | Perov-5 | | Carbon-24 | | MP20 | | MPTS-52 | |
|---|---|---|---|---|---|---|---|---|
| | Match | RMSE | Match | RMSE | Match | RMSE | Match | RMSE |
| | One Shot | | | | | | | |
| CrystaLLM | 46.1% | 0.095 | 20.3% | 0.176 | 58.7% | 0.041 | 19.2% | 0.111 |
| CDVAE | 45.3% | 0.114 | 17.1% | 0.297 | 33.9% | 0.105 | 5.34% | 0.211 |
| DiffCSP | **52.0%** | **0.076** | 17.5% | 0.276 | 51.5% | 0.063 | 12.2% | 0.179 |
| Mat2Seq | 50.0% | 0.099 | **23.7%** | **0.169** | **61.3%** | **0.040** | **23.1%** | **0.109** |
| | 20 Shots | | | | | | | |
| CrystaLLM | 97.6% | 0.025 | 85.2% | 0.151 | 74.0% | **0.035** | 33.8% | 0.106 |
| CDVAE | 88.5% | 0.046 | 88.4% | 0.229 | 67.0% | 0.103 | 20.8% | 0.209 |
| DiffCSP | **98.6%** | **0.013** | **88.5%** | 0.219 | **77.9%** | 0.049 | 34.0% | 0.175 |
| Mat2Seq | 98.5% | 0.023 | 86.0% | **0.148** | 75.3% | 0.037 | **36.4%** | **0.088** |

## 3.5 Language model enabled 3D crystal structure generation

**Discretization**. To train a language model, crystal sequence representations obtained in Sec. 3.3 need to be discretized. Compositional and space group information including atom types, number of atoms in the cell, and space group symbol are discrete. For lattice parameters and fractional coordinates, we round real numbers to four decimal places to ensure consistency and efficiency following CrystaLLM [12]. We also include keywords and symbols like space_group_symbol, formula, lattice_parameters and so on in our Mat2Seq dictionary, with the full list provided in Appendix A.2.

**Training and sampling**. We enable language models to generate novel crystals using the proposed unique and complete sequence representations for 3D crystal structures. Specifically, we follow the previous work [12] and train a GPT [20] model $M$ with parameters $\theta$ to capture the distribution of Mat2Seq crystal sequences, $U$, constructed from corresponding crystal datasets with a standard next-token prediction cross-entropy loss $\ell$ for all elements in the sequence:

$$\min_\theta \mathbb{E}_{u \in U} \sum_{i=1}^{|u|-1} \ell(M_\theta(u_1, u_2, \cdots, u_i), u_{i+1}), \tag{7}$$

where $u$ represents a crystal structure within the dataset. During the sampling phase, we initiate the sequence with predetermined starting patterns, applying an autoregressive sampling technique to iteratively predict the next sequence component from the conditional model distribution $p_\theta(u_{i+1}|u_1, u_2, \cdots, u_i)$ until a stop pattern or a maximum length is achieved. Consistent with the referenced work [12], our sampling strategy incorporates top-$k$ [21] and temperature control [22] techniques. Importantly, the choice of language model $M$ is flexible and can be seamlessly substituted with more advanced or alternative powerful language models.

**Conditional generation.** For conditional generation, we use conditional tokens to represent desired compositions and properties. Specifically, for Mat2Seq crystal sequence representations, the compositional information is placed ahead of crystal structure information. Hence, by simply initialize sequences with desired compositions, the language model $M$ can generate corresponding crystal structures. To target specific properties, we introduce a property token to denote the desired attribute and place these tokens at the beginning of the Mat2Seq sequences. The training and sampling processes mirror those used in unconditional generation, with the initial input being the property token that corresponds to the desired attribute.

# 4 Experimental results

In this section, we evaluate the ability of Mat2Seq to discover stable crystal structure for interested crystal systems in Sec. 4.1 and 4.2 and generate novel crystal structures satisfying desired physical properties in Sec. 4.3. We also include experimental results for random crystal generation assessing validity, stability (measured by DFT), and the ratios of stability, unique, and novel (S.U.N.) crystals in Appendix A.3.

Table 3: Efficiency and model complexity comparisons.

| MP-20 test set | number of parameters | RMSE | 20 shots generation speed (sec./crystal) |
|---|---|---|---|
| CDVAE | 4.5 M | 0.103 | 37.9 s |
| DiffCSP | Similar to CDVAE | 0.049 | 7.3 s |
| Mat2Seq-small | 25 M | <u>0.039</u> | **2.1 s** |
| Mat2Seq-large | 200 M | **0.037** | <u>5.7 s</u> |

Table 4: Mat2Seq match rate (%) and RMSE for experimentally observed crystal structures in MP-20 test set.

| MP-20 | number of crystals | Match Rate (%) | RMSE |
|---|---|---|---|
| Mat2Seq - exp observed | 3,819 | **65.2** % | 0.042 |
| Mat2Seq - the whole test set | 9,046 | 61.3 % | **0.040** |

## 4.1 Generating stable crystal structures for given compositions

**Baselines**. We first evaluate the effectiveness of Mat2Seq on the widely used crystal structure prediction benchmark, comparing against diffusion-based methods including CDVAE [15] and DiffCSP [17], as well as a recent language model based CrystaLLM [12].

**Datasets**. We conduct experiments on four datasets following DiffCSP [17] and CrystaLLM [12], including Perov-5, Carbon-24, MP-20, and MPTS-52. **Perov-5** [23, 24] contains 18,928 perovskite materials with similar structures and 5 atoms in unit cells. **Carbon-24** [25] contains 10,153 carbon materials with at most 24 atoms in the unit cell. **MP-20** [26] includes 45,231 stable inorgainic materials from the Materials Project, covering the majority of experimentally-generated materials with at most 20 atoms in the unit cell. **MPTS-52** [17] is the most challenging one with 40,476 crystals and at most 52 atoms in the unit cell.

**Experimental setup**. For crystal structure prediction, machine learning methods are trained using crystal structures and corresponding compositions in the training set. During the inference phase, compositions of crystals in the test set are provided as input to methods, and the goal is to generate stable crystal structures that match the ground truth crystal structures. To assess the quality of generated crystal structures, two metrics are used, including match rate which measures the ratio of the generated structures that match with the ground truth structure determined by Pymatgen structure matcher [27], and RMSE [27] which measures the structural differences between the ground truth and matched generated structures. A single NVIDIA A100 GPU is used for computing for this task. We directly follow DiffCSP [17] to split corresponding datasets into training, evaluation, and test sets. We use the same language model settings following CrystaLLM [12] to make sure the comparison is fair. Detailed training and inference parameters are provided in Appendix A.2. We also provide additional experimental evaluations in Appendix A.3 for random crystal generation on the MP20 dataset, assessing validity, stability (measured by DFT), and the ratios of stable, unique, and novel crystals, compared with state-of-the-art methods, including FlowMM [28], DiffCSP [17], and CDVAE [15].

**Results**. There are several observations from the experimental results shown in Table. 2. **(1)** Compared with CrystaLLM [12] that directly uses CIF files to describe 3D crystal structures, Mat2Seq achieves better match rate for all eight tasks, demonstrating the effectiveness of Mat2Seq crystal sequence representations beyond CIF files. **(2)** Compared with DiffCSP [17] that is specifically designed for this task, Mat2Seq achieves comparable results, and achieves a significantly better match rate and smaller RMSE for the most challenging MPTS-52 dataset, with a **89%** improvement in match rate for one shot generation, and a **50%** decrease in RMSE for 20 shots generation. **(3)** Mat2Seq demonstrates excellent one shot generation performances across datasets with various difficulty levels and data scales.

**Results on experimentally observed crystal structures**. Comparing with experimentally observed crystal structures is essential to demonstrate the real-world applicability of the proposed method, beyond synthetic structures generated purely from DFT calculations. The MP20 dataset test set includes 3,819 crystal structures that have been experimentally observed (match entries in the ICSD [29]). To assess the accuracy of Mat2Seq on this data, we conducted additional evaluations on these 3,819 structures, with the results presented in Table 4. Notably, Mat2Seq achieved a 65.2% match rate with an RMSE of 0.042 for these experimentally observed structures.
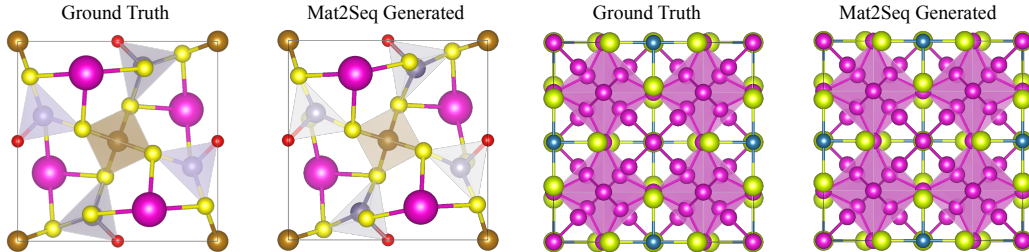
Figure 4: Mat2Seq can generate recently discovered novel crystals from literature. $Eu_2FeGe_2OS_6$ on the left, and $Ce_6Cd_{23}Te$ on the right. The structure generated by Mat2Seq for $Eu_2FeGe_2OS_6$ is the reflected version of the ground truth.

Table 5: Ability to generalize to recently discovered crystals from literature, measured by three seperate generation runs.

|  | CrystaLLM | Ours |
|---|---|---|
| Validity $\uparrow \pm$ std | $23.7\% \pm 2.5\%$ | $\mathbf{27.3}\% \pm 1.4\%$ |

**Efficiency**. It is well known that language models typically have higher parameter counts and greater computational demands. To address this, we provide a comparison of efficiency in terms of parameter count and generation speed, alongside specialized crystal models such as CDVAE and DiffCSP, in Table 3. While Mat2Seq, based on large language models, does have more parameters, using similar GPU resources (as done in DiffCSP [17]), Mat2Seq demonstrates significantly faster generation speed. Furthermore, we show that reducing the model size by a factor of eight does not significantly affect Mat2Seq's performance and improves efficiency. In this case, the smaller model achieved an RMSE of 0.039, which is notably better than DiffCSP's RMSE of 0.049, while also being more than three times faster in the generation process.

## 4.2 Ability to generalize to recently discovered crystals from literature

The promising results in crystal structure prediction shown in Sec. 4.1 have demonstrated the effectiveness of Mat2Seq crystal sequence representations. We then examine the ability of Mat2Seq with language model to discover novel crystal structures that are obtained from recent literature and are not seen during training. Specifically, we randomly choose 10 crystals from the **challenge set** [12] to evaluate Mat2Seq's ability to discover novel crystal structures that originate from recent literature, and compare with previous CrystaLLM [12].

**Dataset**. For this task, we follow CrystaLLM [12] and train the model using crystal structures from the Materials Project [26], OQMD [30] and NOMAD [31], which has been optimized using DFT calculations, containing 2,047,889 crystal structures for training in total. This dataset contains around 800K unique formulas and 94 chemical element types. The evaluation is conducted with 10 recently discovered crystal structures from literature, including $AlCu_2As(OH)_{12}$ [32], $Ba_2Fe_2F_9$ [33], $Ca_2Te_3O_8$ [34], $Ce_6Cd_{23}Te$ [35], $Cs_8Cu_3Si_{14}O_{35}$ [36], $Cu_4FeGe_2OS_6$ [37], $Eu_2FeGe_2OS_6$ [38], $K_2Sr_4(PO_3)_{10}$ [39], $La_4Ga_2S_8O_3$ [40], and $Li_9Al_4Sn_5$ [41].

**Experimental setup**. We first train our model using around 2M crystal structures in the training set following CrystaLLM with broken files removed. After training, we evaluate the model's ability to generate valid crystal structures for 10 recently discovered crystal structures from literature by taking the compositional information as conditions and generating 20 structures for each composition. To compare with CrystaLLM, we directly use their pre-trained model for the same training set and generate 20 structures for each composition. We follow CrystaLLM and evaluate the validity of generated structures by whether the generated formula is consistent with the given composition, whether the space group is consistent with the structure, and whether the generated crystal has reasonable bond lengths. 4 A100 GPUs are used for computing for this task. We also provide additional experimental results in Appendix A.3, evaluating the Hit rate (i.e., whether the generated structure matches recently observed experimental crystals from the literature) and RMSE across 10 challenge crystal systems, in comparison with the previous state-of-the-art method, CrystaLLM.

Table 6: Ability to generate crystal structures with desired band gap properties. We measure the success rate of generating crystal structures with band gap value < 0.5 eV and band gap value > 3.0 eV, measured by the state-of-the-art band gap predictor ComFormer [14]. We also evaluate the validity, uniqueness, and novelty of the generated crystals.

| Generation Codition | band gap < 0.5 eV | band gap > 3.0 eV | Validity | Uniqueness | Novelty |
|---|---|---|---|---|---|
| Towards band gap ↓ | **83.6**% | 12.0% | 88.0% | **98.0**% | 86.2% |
| Towards band gap ↑ | 6.4% | **90.7**% | **89.8**% | 92.2% | **98.6**% |

**Results**. As shown in Table 5, for the $20 \times 10$ generated crystals, our method achieved **27.3%** validity rate for novel crystal structures from recent literature, better than CrystaLLM with 23.7%. We visualize the generation results for $Eu_2FeGe_2OS_6$ and $Ce_6Cd_{23}Te$ in Figure 4 which shows Mat2Seq actually can discover novel crystals reported in recent literature with nearly identical 3D structures. It is worth noting that the structure generated by Mat2Seq for $Eu_2FeGe_2OS_6$ is the reflected version of the founded one, and reflected crystal structure will have the identical energy, which on the other hand demonstrating the reflection sensitivity of Mat2Seq.

### 4.3 Discovering crystal structures with desired properties

**Experimental details**. We further evaluate Mat2Seq's ability to discover crystals with desired properties. Specifically, we have put 10 placeholders marked as unknown_prop before Mat2Seq crystal sequences when training our model on around 2M crystal structures. After pre-training, we fine-tune the model on JARVIS-DFT dataset [42] with 61,541 crystal structures and corresponding band gap values in the training set. We use intervals to tokenize band gap values, where values from 0 to 0.5 are marked as 0, values from 0.5 to 1 are marked as 1, and so on, and replace one unknown_prop token in the sequence. For future users, this design can enable the discovery of physical properties of their interest by simply replacing the unknown_prop token with corresponding properties and fine-tuning the pre-trained model from Sec. 4.2. A single A100 GPU is used for this task.

**Evaluation metrics**. We evaluate our model in two settings, including (1) generating crystals with low band gap ($< 0.5$ eV) that could potentially be used for thermoelectrics, thermophotovoltaics, infrared sensing and vision, cryogenic cooling, as well as discovering novel catalysts and topological quantum materials, and (2) generating crystals with high band gap ($> 3.0$ eV) that could potentially be used for solid-state LEDs and lasers, nonlinear optics, ferroelectrics, multiferroics, power electronics, and high-temperature applications. We measure the success rate of 500 randomly generated crystals that satisfy (1) and (2) when our model is conditioned towards generating crystals with small band gap values and large band gap values. We measure the band gap value of generated crystals using the state-of-the-art method ComFormer [14] for band gap prediction with mean absolute error of 0.122 eV. We also provide measurements of validity, uniqueness, and novelty for the generated crystals.

**Results**. Table 6 demonstrates Mat2Seq's ability to significantly alter the original property distribution and discover crystals towards large or small band gap values. Specifically, when conditioned towards small band gap values with interval 0 to 0.5 eV, **83.6%** generated crystals have band gap values $< 0.5$ eV, and when conditioned towards large band gap values with interval 3.5 to 4.0 eV, **90.7%** generated crystals have band gap values $> 3.0$ eV.

## 5 Conclusion, limitation, and discussion

In this work, we propose Mat2Seq, which converts 3D crystals into unique and complete 1D sequences for language model-enabled crystal generation, naturally ensuring that all equivalent crystals are considered identically probable without the need for data augmentation. Mat2Seq demonstrates promising performance in crystal structure generation and discovery, and the trained model can be used to discover crystal structures with desired properties of interest to users. The limitations of the current Mat2Seq include: (1) it cannot be directly used for other atomic systems, like molecules and proteins; (2) the extension to model disordered materials remains a challenging frontier; and (3) large-scale training with more stable crystal structures can potentially enhance the robustness and performance when more computational resources are available. Positive and negative societal impacts of discovering novel materials with desired properties may apply to this work.

## Acknowledgments and Disclosure of Funding

## References

[1] Ioannis Petousis, Wei Chen, Geoffroy Hautier, Tanja Graf, Thomas D Schladt, Kristin A Persson, and Fritz B Prinz. Benchmarking density functional perturbation theory to enable high-throughput screening of materials for dielectric constant and refractive index. *Physical Review B*, 93(11):115151, 2016.

[2] Susmriti Das Mahapatra, Preetam Chandan Mohapatra, Adrianus Indrat Aria, Graham Christie, Yogendra Kumar Mishra, Stephan Hofmann, and Vijay Kumar Thakur. Piezoelectric Materials for Energy Harvesting and Sensing Applications: Roadmap for Future Smart Materials. *Advanced Science*, 8(17):2100864, 2021.

[3] Hua Wang and Xiaofeng Qian. Ferroelectric nonlinear anomalous Hall effect in few-layer $WTe_2$. *npj Computational Materials*, 5(1):119, 2019.

[4] Jun Xiao, Ying Wang, Hua Wang, C. D. Pemmaraju, Siqi Wang, Philipp Muscher, Edbert J. Sie, Clara M. Nyby, Thomas P. Devereaux, Xiaofeng Qian, Xiang Zhang, and Aaron M. Lindenberg. Berry curvature memory through electrically driven stacking transitions. *Nature Physics*, 16(10):1028–1034, 2020.

[5] Hua Wang and Xiaofeng Qian. Electrically and magnetically switchable nonlinear photocurrent in $PT$-symmetric magnetic topological quantum materials. *npj Computational Materials*, 6:199, 2020.

[6] Kamal Choudhary, Daniel Wines, Kangming Li, Kevin F Garrity, Vishu Gupta, Aldo H Romero, Jaron T Krogel, Kayahan Saritas, Addis Fuhr, Panchapakesan Ganesh, et al. JARVIS-Leaderboard: a large scale benchmark of materials design methods. *npj Computational Materials*, 10(1):93, 2024.

[7] Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, Keir Adams, Maurice Weiler, Xiner Li, Tianfan Fu, Yucheng Wang, Haiyang Yu, YuQing Xie, Xiang Fu, Alex Strasser, Shenglong Xu, Yi Liu, Yuanqi Du, Alexandra Saxton, Hongyi Ling, Hannah Lawrence, Hannes Stärk, Shurui Gui, Carl Edwards, Nicholas Gao, Adriana Ladera, Tailin Wu, Elyssa F. Hofgard, Aria Mansouri Tehrani, Rui Wang, Ameya Daigavane, Montgomery Bohde, Jerry Kurtin, Qian Huang, Tuong Phung,

Minkai Xu, Chaitanya K. Joshi, Simon V. Mathis, Kamyar Azizzadenesheli, Ada Fang, Alán Aspuru-Guzik, Erik Bekkers, Michael Bronstein, Marinka Zitnik, Anima Anandkumar, Stefano Ermon, Pietro Liò, Rose Yu, Stephan Günnemann, Jure Leskovec, Heng Ji, Jimeng Sun, Regina Barzilay, Tommi Jaakkola, Connor W. Coley, Xiaoning Qian, Xiaofeng Qian, Tess Smidt, and Shuiwang Ji. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*, 2023.

[8] Keqiang Yan, Yi Liu, Yuchao Lin, and Shuiwang Ji. Periodic graph transformers for crystal material property prediction. In *The 36th Annual Conference on Neural Information Processing Systems*, pages 15066–15080, 2022.

[9] Yuchao Lin, Keqiang Yan, Youzhi Luo, Yi Liu, Xiaoning Qian, and Shuiwang Ji. Efficient approximations of complete interatomic potentials for crystal property prediction. In *ICML*, pages 21260–21287, 2023.

[10] Daniel Flam-Shepherd and Alán Aspuru-Guzik. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as XYZ, CIF, and PDB files. *arXiv preprint arXiv:2305.05708*, 2023.

[11] Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379*, 2024.

[12] Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *arXiv preprint arXiv:2307.04340*, 2023.

[13] Limei Wang, Yi Liu, Yuchao Lin, Haoran Liu, and Shuiwang Ji. ComENet: Towards complete and efficient message passing for 3D molecular graphs. In *The 36th Annual Conference on Neural Information Processing Systems*, pages 650–664, 2022.

[14] Keqiang Yan, Cong Fu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Complete and efficient graph transformers for crystal material property prediction. In *The Twelfth International Conference on Learning Representations*, 2024.

[15] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal Diffusion Variational Autoencoder for Periodic Material Generation. In *International Conference on Learning Representations*, 2022.

[16] Youzhi Luo, Chengkai Liu, and Shuiwang Ji. Towards symmetry-aware generation of periodic materials. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. MatterGen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

[19] Hong-Long Shi and Zi-An Li. Niggli reduction and bravais lattice determination. *Journal of Applied Crystallography*, 55(1):204–210, 2022.

[20] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[21] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

[22] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

[23] Ivano E Castelli, David D Landis, Kristian S Thygesen, Søren Dahl, Ib Chorkendorff, Thomas F Jaramillo, and Karsten W Jacobsen. New cubic perovskites for one-and two-photon water splitting using the computational materials repository. *Energy & Environmental Science*, 5(10):9034–9043, 2012.

[24] Ivano E Castelli, Thomas Olsen, Soumendu Datta, David D Landis, Søren Dahl, Kristian S Thygesen, and Karsten W Jacobsen. Computational screening of perovskite metal oxides for optimal solar light capture. *Energy & Environmental Science*, 5(2):5814–5819, 2012.

[25] Chris J Pickard, Ashkan Salamat, Michael J Bojdys, Richard J Needs, and Paul F McMillan. Carbon nitride frameworks and dense crystalline polymorphs. *Physical Review B*, 94(9):094104, 2016.

[26] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

[27] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

[28] Benjamin Kurt Miller, Ricky TQ Chen, Anuroop Sriram, and Brandon M Wood. FlowMM: Generating Materials with Riemannian Flow Matching. In *Forty-first International Conference on Machine Learning*, 2024.

[29] Dejan Zagorac, H Müller, S Ruehl, J Zagorac, and Silke Rehme. Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. *Journal of Applied Crystallography*, 52(5):918–925, 2019.

[30] James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM*, 65:1501–1509, 2013.

[31] Markus Scheidgen, Lauri Himanen, Alvin Noe Ladines, David Sikter, Mohammad Nakhaee, Ádám Fekete, Theodore Chang, Amir Golparvar, José A Márquez, Sandor Brockhauser, et al. NOMAD: A distributed web-based platform for managing materials science research data. *Journal of Open Source Software*, 8(90):5388, 2023.

[32] Alexandra Plumhoff. *Thermodynamic properties, crystal structures, phase relations and isotopic studies of selected copper oxysalts*. PhD thesis, 2020.

[33] Elizabeth A Pogue, Jack Bond, Cassandra Imperato, John BS Abraham, Natalia Drichko, and Tyrel M McQueen. A Gold(I) Oxide Double Perovskite: $Ba_2AuIO_6$. *Journal of the American Chemical Society*, 143(45):19033–19042, 2021.

[34] Matthias Weil. $Ca_2Te_3O_8$, a new phase in the $CaO–TeO_2$ system. *Acta Crystallographica Section E: Crystallographic Communications*, 75(1):26–29, 2019.

[35] Griffen Desroches and Svilen Bobev. Synthesis and structure determination of $Ce_6Cd_{23}Te$: a new chalcogen-containing member of the $RE_6Cd_{23}T$ family ($RE$ is a rare-earth metal and T is a late group 14, 15 and 16 element). *Acta Crystallographica Section C: Structural Chemistry*, 73(2):121–125, 2017.

[36] Gregory Morrison, Virginia G Jones, K Pilar Zamorano, John E Greedan, and Hans-Conrad Zur Loye. Flux Synthesis, UV–vis Absorbance, and Magnetism of Cesium Copper Silicates with an Isolated Super-Super Exchange Spin Dimer in $Cs_6Cu_2Si_9O_{23}$. *Inorganic Chemistry*, 62(29):11682–11689, 2023.

[37] Andrew J Craig, Stanislav S Stoyko, Allyson Bonnoni, and Jennifer A Aitken. Syntheses and crystal structures of the quaternary thiogermanates $Cu_4FeGe_2S_7$ and $Cu_4CoGe_2S_7$. *Acta Crystallographica Section E: Crystallographic Communications*, 76(7):1117–1121, 2020.

[38] Nan Zhang, Xiao Huang, Wen-Dong Yao, Yao Chen, Zheng-Rui Pan, Bingxuan Li, Wenlong Liu, and Sheng-Ping Guo. $Eu_2MGe_2OS_6$ (M= Mn, Fe, Co): Three Melilite-Type Rare-Earth Oxythiogermanates Exhibiting Balanced Nonlinear-Optical Behaviors. *Inorganic Chemistry*, 62(40):16299–16303, 2023.

[39] Weimin Dong, Yingjie Sun, Henghao Feng, Dazheng Deng, Jun Jiang, Jin Yang, Wei Guo, Libin Tang, Jincheng Kong, and Jun Zhao. $K_2Sr_4(PO_3)_{10}$: A Polyphosphate with Deep-UV Cutoff Edge and Enlarged Birefringence. *Inorganic Chemistry*, 62(39):16215–16221, 2023.

[40] Hong Yan, Kotaro Fujii, Houria Kabbour, Akira Chikamatsu, Yu Meng, Yoshitaka Matsushita, Masatomo Yashima, Kazunari Yamaura, and Yoshihiro Tsujimoto. $La_4Ga_2S_{83}$: A Rare-Earth Gallium Oxysulfide with Disulfide Ions. *Inorganic Chemistry*, 62(26):10481–10489, 2023.

[41] Volodymyr Pavlyuk, Grygoriy Dmytriv, Ivan Tarasiuk, and Helmut Ehrenberg. $Li_9Al_4Sn_5$ as a new ordered superstructure of the $Li_{13}Sn_5$ type. *Acta Crystallographica Section C: Structural Chemistry*, 73(4):337–342, 2017.

[42] Kamal Choudhary, Kevin F Garrity, Andrew CE Reid, Brian DeCost, Adam J Biacchi, Angela R Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A Gilad Kusne, Andrea Centrone, et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Computational Materials*, 6:173, 2020.

# A   Appendix

## A.1   Proofs

**Lemma 4.** *Mat2Seq* $: (\mathbf{A}, \mathbf{P}, \mathbf{L}) \to \mathcal{X}$ *is periodic invariant and unit cell* $SE(3)$ *invariant.*

*Proof.* To prove that Mat2Seq is periodic invariant and unit cell $SE(3)$ invariant, we first show that the unit cell $(\mathbf{A}_u, \mathbf{P}_u, \mathbf{L}_u)$ determined in Sec. 3.2 is $SO(3)$ equivariant and periodic invariant. In other words, after applying arbitrary $SO(3)$ transformations $\mathbf{R} \in \mathbb{R}^{3 \times 3}, |\mathbf{R}| = 1$ with $\mathbf{b} \in \mathbb{R}^3$ and periodic transformations, the resultant unit cell should change from $(\mathbf{A}_u, \mathbf{P}_u, \mathbf{L}_u)$ to $(\mathbf{A}_u, \mathbf{R}\mathbf{P}_u, \mathbf{R}\mathbf{L}_u)$. To begin with, because Euclidean distances $a, b, c$, bond angles $\alpha, \beta, \gamma$, and whether three vectors form a right-hand system are $SE(3)$ and periodic invariant, which means after arbitrary $SO(3)$ transformation $\mathbf{R} \in \mathbb{R}^{3 \times 3}, |\mathbf{R}| = 1$ with $\mathbf{b} \in \mathbb{R}^3$ and periodic transformation, the same but rotated lattice matrix $\mathbf{R}\mathbf{L}_u$ will be determined. Additionally, because local density $\rho_{i,r}$, atom type $Z_i$, and densities along three lattice vectors $\rho_{i,r}^{\boldsymbol{\ell}_1'}, \rho_{i,r}^{\boldsymbol{\ell}_2'}, \rho_{i,r}^{\boldsymbol{\ell}_3'}$ are $SE(3)$ and periodic invariant, the same atom $i$ will be determined to serve as the origin of the unit cell after arbitrary $SO(3)$ transformation and periodic transformation. Hence, when applying arbitrary $SO(3)$ transformation $\mathbf{R} \in \mathbb{R}^{3 \times 3}, |\mathbf{R}| = 1, \mathbf{b} \in \mathbb{R}^3$ and periodic transformation to $(\mathbf{A}, \mathbf{P}, \mathbf{L})$, the resultant primitive unit cell will be $(\mathbf{A}_u, \mathbf{R}\mathbf{P}_u, \mathbf{R}\mathbf{L}_u)$. Hence, the unit cell $(\mathbf{A}_u, \mathbf{P}_u, \mathbf{L}_u)$ determined by Mat2Seq is $SO(3)$ equivariant and periodic invariant.

Based on the $SO(3)$ equivariant and periodic invariant primitive unit cell $(\mathbf{A}_u, \mathbf{P}_u, \mathbf{L}_u)$, we then prove that the obtained sequence representation is $SE(3)$ and periodic invariant. To begin with, it can be seen that lattice lengths, lattice angles, and atomic numbers are naturally invariant. Additionally, due to that $\mathbf{P}_u, \mathbf{L}_u$ in $(\mathbf{A}_u, \mathbf{P}_u, \mathbf{L}_u)$ are $SO(3)$ equivariant, periodic invariant, and are connected by fractional coordinates $\boldsymbol{p}_i = \boldsymbol{p}_{\text{frac},i} \cdot \mathbf{L} = x * \boldsymbol{\ell}_1 + y * \boldsymbol{\ell}_2 + z * \boldsymbol{\ell}_3$, when the crystal is rotated by $\mathbf{R}$ we can have

$$\mathbf{R}\boldsymbol{p}_i = \mathbf{R}(x * \boldsymbol{\ell}_1 + y * \boldsymbol{\ell}_2 + z * \boldsymbol{\ell}_3) = x' * \mathbf{R}\boldsymbol{\ell}_1 + y' * \mathbf{R}\boldsymbol{\ell}_2 + z' * \mathbf{R}\boldsymbol{\ell}_3,$$

which means $x = x', y = y', z = z'$. Hence fractional coordinates are $SE(3)$ invariant. Additionally, because $\mathbf{P}_u, \mathbf{L}_u$ are periodic invariant, the corresponding fractional coordinates are periodic invariant. To summarize, it can be seen that all components in Mat2Seq sequence, including (1) compositional information, (2) space group information, (3) lattice parameters obtained by lattice lengths and angles, and (4) atom types and fractional coordinates are $SE(3)$ and periodic invariant. $\qquad \square$

## A.2   Experimental details

Following CrystaLLM [12], we use GPT-2 with 16 layers, 16 heads, and embedding size of 1024 for all tasks. We show the detailed training parameters including window size, batch size, learning rate, drop out ratio, number of training iterations for different tasks in Table. 7.

Table 7: Training parameters of Mat2Seq for crystal structure prediction benchmark.

| Task | Window size | Batch size | Learning rate | Num. iterations | Drop out |
|---|---|---|---|---|---|
| Perov-5 | 768 | 32 | 0.0005 | 40k | 0.1 |
| Carbon-24 | 768 | 48 | 0.0003 | 40k | 0.1 |
| MP20 | 768 | 32 | 0.0003 | 50k | 0.1 |
| MPTS52 | 1,385 | 24 | 0.0003 | 40k | 0.1 |
| CrystaLLM 2.3 M | 1,500 | 128 | 0.0005 | 500k | 0.1 |
| JARVIS bandgap | 1,500 | 20 | 0.0001 | 10k | 0.1 |

During sampling phase, for Perov-5 dataset, we use temperature=0.7 and top-$k$=10 one shot generation, and temperature=1.0, top-$k$=10 for 20 shots generation. For Carbon-24 dataset, we use temperature=0.7 and top-$k$=10 one shot generation, and temperature=1.0, top-$k$=10 for 20 shots generation. For MP20 dataset, we use temperature=0.4 and top-$k$=5 one shot generation, and temperature=1.5, top-$k$=10 for 20 shots generation. for MPTS52 dataset, we use temperature=1.0 and top-$k$=10 one shot generation, and temperature=1.0, top-$k$=10 for 20 shots generation. For discovering novel crystal structures obtained from recent literature, we use temperature=0.7 and

top-$k$=10 during sampling. For discovering crystal structures with desired band gap properties, we use temperature=0.8 and top-$k$=10 during sampling phases.

**Mat2Seq tokens**. We provide comprehensive list of tokens used by Mat2Seq, including atom types listed in Table 8, integer numbers from 0 to 300, digits including 0 to 9 with "." to represent real values, special tokens listed in Table 9, and 227 space group symbols [12].

Table 8: List of atom types.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Ac | Ag | Al | Ar | As | Au | B | Ba | Be | Bi | Br | C |
| Ca | Cd | Ce | Cl | Co | Cr | Cs | Cu | Dy | Er | Eu | F |
| Fe | Ga | Gd | Ge | H | He | Hf | Hg | Ho | I | In | Ir |
| K | Kr | La | Li | Lu | Mg | Mn | Mo | N | Na | Nb | Nd |
| Ne | Ni | Np | O | Os | P | Pa | Pb | Pd | Pm | Pr | Pt |
| Pu | Rb | Re | Rh | Ru | S | Sb | Sc | Se | Si | Sm | Sn |
| Sr | Ta | Tb | Tc | Te | Th | Ti | Tl | Tm | U | V | W |
| Xe | Y | Yb | Zn | Zr | | | | | | | |

Table 9: List of special tokens.

| | | | | | | |
|---|---|---|---|---|---|---|
| space_group_symbol | formula | atoms | lattice_parameters | a | b | c |
| alpha | beta | gamma | unknown_prop | , | " " | : |
| \n | <pad> | | | | | |

## A.3  Additional Experimental Results

We provide additional experimental results of Mat2Seq in this section.

In Table 10, we compare Mat2Seq with the recent state-of-the-art method FlowMM [28], assessing Validity, Stability, and S.U.N. (stable, unique, and novel) metrics. To ensure a fair comparison, we follow the FlowMM pipeline and generate 1000 materials to obtain these results.

Table 10: Comparison of Validity (%), Stability (%). and Stable, Unique, and Novel (S.U.N) (%) on MP20 dataset. The unit of $E_{hull}$ is eV/atom.

| Model | Validity (%) ↑ | | Stability Rate - DFT (%) ↑ | | S.U.N. Rate (%) ↑ |
|---|---|---|---|---|---|
| | Composition | Structural | $E_{hull} < 0.0$ | $E_{hull} < 0.1$ | MP |
| CDVAE | <u>86.7</u> % | **100** % | 1.57% | - | 1.4% |
| DiffCSP | 83.3 % | **100** % | **5.06**% | - | **3.3**% |
| FlowMM (ICML 24) | 83.2 % | 96.9 % | 4.19% | - | 2.5% |
| Mat2Seq (temp=1.35) | **88.5** % | 94.2 % | 4.10% | **49.2**% | 2.0% |
| Mat2Seq (temp=1.65) | 81.7 % | 88.6 % | <u>4.50</u>% | <u>46.6</u>% | <u>3.2</u>% |

In Table 11, we present the Hit Rate (whether the generated structures match the ground truth structure) and RMSE for the 10 challenge crystal systems discussed in Section 4.2, compared to the previous state-of-the-art method, CrystaLLM.

Table 11: Comparison of Hit Rate (%) and RMSE for ten recently discovered crystal structures from literature.

| Model | Hit Rate (%) ↑ | RMSE ↓ |
|---|---|---|
| CrystaLLM | 0 | nan |
| Mat2Seq | **10.0** % | **0.0388** |

## A.4  Licenses for existing assets

We have used datasets including Perov-5, Carbon-24, and MP20 curated by CDVAE [15] with MIT License, MPTS-52 curated by DiffCSP [17] with MIT License, JARVIS-DFT [42] with NIST

License, CrystaLLM [12] with MIT License, the Materials Project [26] with Creative Commons Attribution 4.0 License, OQMD [30] with Creative Commons Attribution 4.0 International License, and NOMAD [31] with Apache License Version 2.0, January 2004. We have used CrystaLLM [12] model with MIT License.

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: As clearly discussed in Sec. 3 and Sec. 4, the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of this work are discussed in Sec. 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Detailed proofs are provided in Sec. 3.4 and Appendix A.1.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Detailed experimental settings, datasets, and computations needed are shared in Sec. 4 and Appendix A.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code will be release after the paper is publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Detailed experimental settings are provided in Sec. 4 and Appendix A.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide the validity of randomly generated crystal structures for interested crystal compositions from recent literature, including 20 generation runs for each composition. We also repeat this generation process for our method and the baseline method three times and show the std in Sec. 4. For crystal structure prediction tasks, we follow previous benchmark settings and include 20 shots generation results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Detailed compute resources needed are provided in Sec. 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We conform that we follow NeurIPS Code of Ethics in every respect.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We provide Broader Impacts in Sec. 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: We will consider this when we release the model in the future.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the existing assets in the main paper and provide licenses of them in Appendix A.4.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We discuss details of the dataset and model in Sec. 4 and Appendix A.2.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.