SYMMCD: SYMMETRY-PRESERVING CRYSTAL GENERATION WITH DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating novel crystalline materials has the potential to lead to advancements in fields such as electronics, energy storage, and catalysis. The defining characteristic of crystals is their symmetry, which plays a central role in determining their physical properties. However, existing crystal generation methods either fail to generate materials that display the symmetries of real-world crystals, or simply replicate the symmetry information from examples in a database. To address this limitation, we propose SymmCD[1], a novel diffusion-based generative model that explicitly incorporates crystallographic symmetry into the generative process. We decompose crystals into two components and learn their joint distribution through diffusion: 1) the asymmetric unit, the smallest subset of the crystal which can generate the whole crystal through symmetry transformations, and; 2) the symmetry transformations needed to be applied to each atom in the asymmetric unit. We also use a novel and interpretable representation for these transformations, enabling generalization across different crystallographic symmetry groups. We showcase the competitive performance of SymmCD on a subset of the Materials Project, obtaining diverse and valid crystals with realistic symmetries and predicted properties.

1 INTRODUCTION

Crystals serve as the fundamental building blocks of many materials, including most metals, ceramics, and rocks. The discovery of new crystalline materials is expected to lead to diverse technological breakthroughs in fields ranging from energy storage to computing hardware (Miret et al., 2024). Generative models have the potential to greatly accelerate this process by proposing new candidates materials, and possibly conditioning on desired properties or compositions.

The defining characteristic of crystals is their symmetry. These symmetries are Euclidean transformations that map the crystal structure back to itself. They can in general be some specific translations, rotations, reflections and combinations of these. The set of these operations is called the *space group* of the crystal. It is known that space groups in three dimensions fall into 230 distinct classes (Hahn et al., 1983). The symmetry of a crystal plays a crucial role in determining its stability along with its thermodynamic, electronic and mechanical properties (Nye, 1985). A classic example is given by piezoelectricity, the ability of a material to generate an electric dipole under mechanical stress, which can only be manifested in materials lacking inversion symmetry.

Importantly, many of the recently proposed generative models for crystals do not generate samples with non-trivial symmetry: for example, the most frequently generated crystals by DiffCSP (Jiao et al., 2023) and CDVAE (Xie et al., 2022) are in the low-symmetry P1 space group, which is very rare in nature. MatterGen (Zeni et al., 2023) can generate crystals conditioned on a desired space group for space groups that are highly represented in the dataset, but they only recover the target space group roughly 20% of the time, dropping to about 10% for more symmetric space groups. Cheetham & Seshadri (2024) analyse the space groups of the stable crystal structures proposed by the GNoME model of Merchant et al. (2023), finding that the top 4 most commonly generated space groups account for 34% of all generated crystals, even though each of those 4 space groups appears in less than 1% of crystals in the Inorganic Crystal Structure Database (Hellenbrandt, 2004).

---

[1] Our code is publicly available at `https://anonymous.4open.science/r/SymmCD-596C/`

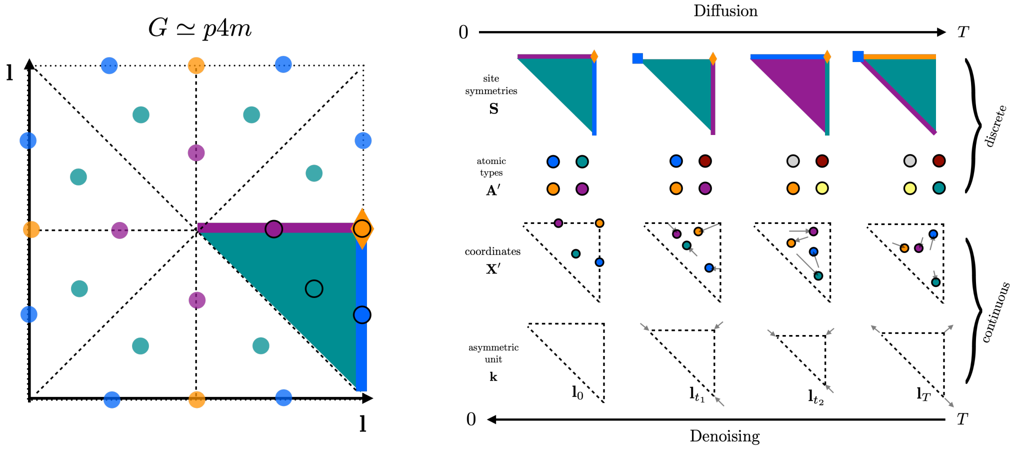


Figure 1: **Illustration of the SymmCD method.** *Left.* Representation of the unit cell of a 2D crystal with $p4m$ symmetry where the asymmetric unit and the site-symmetries of the atoms are highlighted. Leveraging symmetry results in a much more compact, yet complete representation. *Right.* Diffusion and denoising on the different components of the representation. For site symmetries and atom types, discrete diffusion is used. For the coordinate and asymmetric unit continuous diffusion is used. The diffusion and denoising processes preserve the space-group symmetry.

In this work, we propose a novel approach for generative modeling of inorganic crystals that ensures any desired distribution of space groups. The idea is similar to that of creating a paper snowflake, where we fold the paper to create an unconstrained space, and after an unconstrained cutting of the paper in this space, its unfolding creates an object with desired symmetries. In the context of crystals, the unconstrained space is called the *asymmetric unit*, which is a maximal subset of the unit cell with no redundancy. In order to be able to unfold the asymmetric unit, we need to generate the site symmetry of each atom inside the unit, i.e. the symmetry transformations that fix the atoms in place. In our generative process, the atomic positions are made consistent with generated site symmetries, enabling the unfolding of asymmetric unit into a symmetric crystal; see Figure 1.

Crystals and their individual atoms have many different types of symmetries, and so we need to address the issue of data-fragmentation. By representing symmetry information using standard crystallographic notations, such as Hermann–Mauguin notation or Wyckoff position labels (Hahn et al., 1983), we are faced with many crystals and site symmetries that have a low frequency in the training data. To address this problem, we introduce a novel representation of symmetries as binary matrices, which enables information-sharing and generalization across both crystal and site symmetries.

The main contributions of this work are as follows: **I)** We demonstrate a novel approach to generating crystals through the unconstrained generation of asymmetric units, along with their symmetry information. **II)** We introduce a physically-motivated representation for crystallographic site symmetries that generalizes across space groups. **(III)** We experimentally evaluate our method, finding that it performs on par with previous methods in terms of generating stable structures, while offering significantly improved computational efficiency due to our representation. **(IV)** We perform an in-depth analysis of the symmetry and diversity of crystal structures generated by existing generative models.

2 RELATED WORK

There has been a growing body of work in developing machine-learning methods for crystal structure modeling, including the development of datasets and benchmarks (Jain et al., 2013; Saal et al., 2013; Chanussot et al., 2021; Miret et al., 2023; Lee et al., 2023; Choudhary et al., 2024). Recent work has also focused on developing architectures that are equivariant to various symmetries Duval et al. (2023) or are specifically designed to include inductive biases useful for crystal structures (Xie & Grossman, 2018; Kaba & Ravanbakhsh, 2022; Goodall et al., 2022; Yan et al., 2022; 2024).

In addition to structure-based modeling, prior work has also generated full-atom crystal structures, in which all atoms of the three-dimensional structure are generated. A range of generation methods including variational autoencoders (Noh et al., 2019; Xie et al., 2022), GANs (Nouira et al., 2018; Kim et al., 2020), reinforcement learning (Govindarajan et al., 2023), diffusion models (Zeni et al., 2023; Yang et al., 2023; Jiao et al., 2023; Klipfel et al., 2024), flow-matching models (Miller et al., 2024), and active learning based discovery (Merchant et al., 2023) have been used. These follow similar works in 3-D molecule generation (Hoogeboom et al., 2022; Garcia Satorras et al., 2021; Xu et al., 2022), but extend them by incorporating crystal periodicity. In addition to full-atom crystal generation, prior work has also applied text-based methods to understand and generate crystals using language models (Gruver et al., 2024; Flam-Shepherd & Aspuru-Guzik, 2023; Alampara et al., 2024).

Other works have pointed out the importance of symmetry of the generated structures. DiffCSP++ (Jiao et al., 2024), does so by using predefined structural templates from the training data and learning atomic types and coordinates compatible with the templates. While this is an interesting solution, we show that predefining the templates in this way severely limits the diversity and novelty of the generated samples. CrystalGFN (AI4Science et al., 2023) incorporates constraints on the lattice parameters and composition based on space groups, but does not guarantee that the atomic positions respect the desired symmetry. Finally, the concurrent works CrystalFormer (Cao et al., 2024) and Wycryst (Zhu et al., 2024) generate symmetric crystals by predicting atom symmetries. However, they simply use the labels of Wyckoff positions to encode symmetries, which does not enable generalization across groups. The methods are therefore limited to generating from space groups that are common in the dataset. By contrast, our method generalizes across groups and can generate valid crystals even from groups that are rare in the dataset.

3 BACKGROUND

Lattices and unit cells Crystals are macroscopic atomic systems characterized by a periodic structure. A crystal can be described as an infinite 3-dimensional *lattice* of identical *unit cells*, each containing atoms in set positions. We can represent a crystal with the tuple $\mathcal{C} = (\mathbf{L}, \mathbf{X}, \mathbf{A})$, where $\mathbf{L} = (\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3) \in \mathbb{R}^{3\times3}$ is a matrix of *lattice vectors*, $\mathbf{X} \in [0, 1)^{3\times N}$ represents the *fractional* coordinates of N atoms within a unit cell, and $\mathbf{A} \in \{0, 1\}^{Z\times N}$ is a matrix of one-hot vectors of Z possible elements for each atom. The lattice describes the tiling of unit cells: the cartesian coordinates of atoms can be given by $\mathbf{X}^c = \mathbf{L}\mathbf{X}$, and if $\mathbf{x}^c_i$ is the cartesian coordinate of an atom in a unit cell, then the crystal will also contain an identical atom at $\mathbf{x}^c_i + \mathbf{L}\boldsymbol{j}, \forall \boldsymbol{j} \in \mathbb{Z}^3$.

Crystal symmetries In addition to the translational symmetry of the lattice, crystals typically have many other symmetries. Understanding these symmetries is fundamental in characterizing crystals and directly relates to many of the properties of these materials. The *space group* G of a crystal is the group of all Euclidean transformations that leave the crystal invariant, i.e., that simply permutes atoms of the same type. As space groups are subgroups of the Euclidean group, their elements can be represented as $(\mathbf{O}, \mathbf{t})$, where $\mathbf{O} \in O(n)$ and $\mathbf{t} \in \mathbb{R}^3$, with action on $\mathbf{x} \in \mathbb{R}^3$ defined as $(\mathbf{O}, \mathbf{t})\,\mathbf{x} = \mathbf{O}\mathbf{x} + \mathbf{t}$. The operations that are part of a space group can be generally understood as belonging to different types: translations, rotations, inversions, reflections, screw axes (combinations of rotations and translations), and glide planes (combinations of mirroring and translation). Different combinations of these symmetry operations are possible.

Two space-group belong to the same *type* if all their operations can be mapped to each other by an orientation-preserving Euclidean transformation (coordinate change). We denote the set of all space group types as $\mathcal{G}$. In 3 dimensions, there are only 230 unique space group types. By choosing a canonical coordinate system, we can in general work only with space group types. The *point group* P of a space group G is the image of the homomorphism $(\mathbf{O}, \mathbf{t}) \mapsto \mathbf{O}$, i.e the group obtained by keeping only the orthogonal parts of G. By contrast with space groups, any point group must at least preserve a single point, that is the origin. By a similar procedure to space groups, we can classify point groups and find that there are 32 crystallographic point groups types, consisting of inversions, rotations, and reflections. We denote the set of all point group types as $\mathcal{P}$.

Wyckoff positions Having classified symmetry groups, we can now also classify points of space using symmetry considerations. This will be important to our method, as we will seek to use these semantically meaningful classes to guide the generation process. Given a space group G, we say that two points $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^3$ are part of the same *crystallographic orbit* if there is a $(\mathbf{O}, \mathbf{t}) \in G$ such that

$(\mathbf{O},\mathbf{t})\,\mathbf{x}=\mathbf{x}'$. The orbits form a partition of $\mathbb{R}^3$; they can be understood as the finest level of classification under G. We define the *site-symmetry group* of a point $\mathbf{x}$, $S_{\mathbf{x}} = \{(\mathbf{O},\mathbf{t}) \in G \mid (\mathbf{O},\mathbf{t})\,\mathbf{x} = \mathbf{x}\}$ as the subgroup of G that leaves $\mathbf{x}$ invariant. It is clear that the site-symmetry must be a point group (since translations do not preserve any point), and is a subgroup of P. From the orbit-stabilizer theorem (see e.g. Dummit & Foote (2004)), we can find that the number of points in the orbit $\mathbf{x}$ and in the unit cell is given by $|P|/|S_{\mathbf{x}}|$. Points in highly symmetric positions, therefore, result in smaller orbits. A point is said to be in a *general position* if its site-symmetry group is trivial. In this case, there is a one-to-one correspondence between points in the orbit and group members. If the site-symmetry is non-trivial, a point is said to be in a *special position*.

Points in the same orbit have conjugate site-symmetry groups. Therefore, site-symmetry groups related by conjugation can be understood as equivalent. This motivates a coarser level of classification that will be very useful. Two points $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^3$ are part of the same *Wyckoff position* if their site-symmetry group is conjugate. Wyckoff positions have a clear meaning: they classify regions of space in terms of their type of symmetry. The *multiplicity* of a Wyckoff position is the number of equivalent atoms that must occupy that position and is equal to the $|P|/|S_{\mathbf{x}}|$ ratio introduced earlier.

Asymmetric Units The unit cell of a crystal can further be reduced into an *asymmetric unit*, which is a small part of the unit cell that contains no symmetry but can be used to generate the whole unit cell by applying the symmetry transformations of the space group. An asymmetric unit will only contain a single atom from each orbit.

4 METHOD: SYMMETRIC CRYSTAL DIFFUSION (SYMMCD)

4.1 REPRESENTATION OF CRYSTALS WITH WYCKOFF POSITIONS

As explained in the previous section, a crystal structure can, in general, be represented by the tuple $\mathcal{C} = (\mathbf{L}, \mathbf{X}, \mathbf{A})$. This representation has been used in previous generative models for crystals (Xie et al., 2022; Jiao et al., 2023; Luo et al., 2023; Zeni et al., 2023). However, a fundamental limitation of a model based on this representation is that it does not leverage the inductive bias of crystal symmetry and offers no guarantees for the generated positions $\mathbf{X}$ and lattice $\mathbf{L}$ to satisfy anything but a trivial space group.

We introduce an alternative representation that respects symmetry in addition to having many desirable properties. First, we explicitly specify the space group type of the crystal $G \in \mathcal{G}$ in the representation. Given the space group, instead of representing each of the N atoms individually with $\mathbf{X} \in \mathbb{R}^{3\times N}$ and $\mathbf{A} \in \mathbb{R}^{Z\times N}$, we represent the M crystallographic orbits; replicating the atoms within the orbit then creates the crystal. As explained in Section 3, the Wyckoff position identifies a set of orbits by site-symmetries. Therefore, specifying the site-symmetry and an arbitrary orbit representative is sufficient to identify a crystallographic orbit. This corresponds to a representation of an asymmetric unit within the unit cell. We thus define the set of orbit representatives with their Wyckoff positions as the tuple $\mathcal{C}' = (\mathbf{k}, \mathbf{X}', \mathbf{S}, \mathbf{A}')$, where $\mathbf{k}$ is a parametrization of the lattice (to be explained later), $\mathbf{X}' = [\mathbf{x}'_1, \ldots, \mathbf{x}'_M] \in \mathbb{R}^{3\times M}$ are the representative's fractional coordinates in the asymmetric unit, $\mathbf{S} = \left[S_{\mathbf{x}'_1}, \ldots, S_{\mathbf{x}'_M}\right] \in \mathcal{P}^M$ are the site-symmetry groups and $\mathbf{A}' = [\mathbf{a}'_1, \ldots, \mathbf{a}'_M] \in \mathbb{R}^{Z\times M}$ are the atomic types.

From the set of representatives, we can go back to the representation $\mathbf{X}$ and $\mathbf{A}$ in a unique way. This is done by generating the orbits using the replication operation that depends on the group G and the site symmetry $\mathbf{S}$. The replication operation essentially consists of applying all of the symmetry operations of the space group except for the ones included in the site symmetry group. The details of this operation are included in Appendix A. Finally, the lattice $\mathbf{L}$ can be constrained to be compatible with the space group in a convenient way using the vector $\mathbf{k} \in \mathbb{R}^6$ (Jiao et al., 2024): $\log(\mathbf{L}) = \sum_i^6 k_i \mathbf{B}_i$, where the $\mathbf{B}_i \in \mathbb{R}^{3\times 3}$ is a standard basis over symmetric matrices. This basis and the constraints on $\mathbf{k}$ for each space group are described in Appendix B.

Our representation of crystals that explicitly takes into account symmetry is therefore given by the tuple $\mathcal{C}' = (G, \mathbf{k}, \mathbf{X}', \mathbf{S}, \mathbf{A}')$. We convert crystal structures to this representation using the SPGLIB symmetry finding algorithm (Togo & Tanaka, 2018b) provided in the PYMATGEN Python package (Ong et al., 2013).

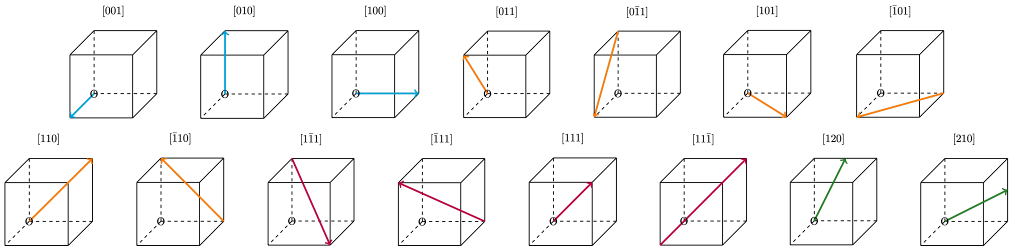


Figure 2: **Crystal symmetry axes.** The different axes describe the directions along which symmetry operations can occur. For each of the 15 axes, there are 13 possible symmetry operations.

In addition to accounting for the symmetry, this representation of a crystal provides two important advantages compared to existing methods. First, it provides the generative model with a powerful physically-motivated inductive bias. It is known from crystallography that atoms are typically not located in arbitrary positions in the unit cell (Aroyo, 2013). Rather, it is energetically more favourable for atoms to occupy positions of high symmetry, e.g. special Wyckoff positions. The representation in terms of positions $\mathbf{X}$ does not make this explicit. The representation using Wyckoff positions $(\mathbf{X}', \mathbf{S}')$ provides explicit supervision to the model and guides the generation process: the model decides in which *type* of high-symmetry position an atom should be located and generates a position compatible with that type. Second, the representation in terms of Wyckoff positions is much more compact than the representation that operates on individual atoms. M is often significantly smaller than N. In the MP-20 dataset (a subset of the Materials Project dataset (Jain et al., 2013)) for example, the average number of orbits is $\bar{M} = 4.7$ whereas the average number of atoms per unit cells is $\bar{N} = 18.9$, representing a fourfold difference [2]. We therefore eliminate the redundant information from the representation and increase the computational efficiency of our method.

4.2 SYMMETRY REPRESENTATION

A key component of our representation of crystals with Wyckoff positions is the encoding of the space group G and site-symmetry groups $\mathbf{S}'$. While there are many existing methods to encode these symmetries, they generally do not make explicit the commonalities between the site-symmetries of Wyckoff positions in the same space group, and the commonalities between different space groups across crystal systems. This is an important limitation: because there are 230 space groups, not having a representation that is common across space groups results in dividing the effective amount of data the model is trained on by a large amount. For example, in the MP-20 dataset (Jain et al., 2013; Xie & Grossman, 2018) 113 space groups out of 169 in the training set have fewer than a hundred samples associated with them, and specific Wyckoff positions in each group have even fewer samples. We propose a method to represent the site-symmetries of different Wyckoff positions and to encode the symmetries of space groups to address this shortcoming.

We represent atom site-symmetries using a binary representation based on the oriented site-symmetry symbol used by the International Tables for Crystallography to describe Wyckoff positions (Hahn et al., 1983; Donnay & Turrell, 1974). The oriented site-symmetry symbols denote generators of the site-symmetry group along different possible axes, illustrated in Figure 2. In total, there are 15 possible axes of symmetry in a crystal, corresponding to each of the Cartesian axes, along with body and face diagonals.

Examples of possible symmetry operations along each axis include rotations and roto-inversions, as well as mirror symmetry along a plane perpendicular to the axis. There are 13 possible symmetries along each axis. Listing out the site symmetry operation along each axis yields a 15×13 binary matrix, or equivalently 15 different one-hot vectors. There is an injective mapping between site symmetries and site symmetry matrix representations, so a representative atom can be replicated to produce a full orbit using this representation.

The space group G can also be encoded into a binary representation using a similar scheme, by listing out the 15 possible axes of symmetry and listing out the possible symmetry operations along each axis. Unlike the point group symmetries of atoms, these space group symmetry operations may

---

[2] This is using the conventional unit cell, not the primitive unit cell. A conventional cell may be twice or four times as big as a primitive cell.

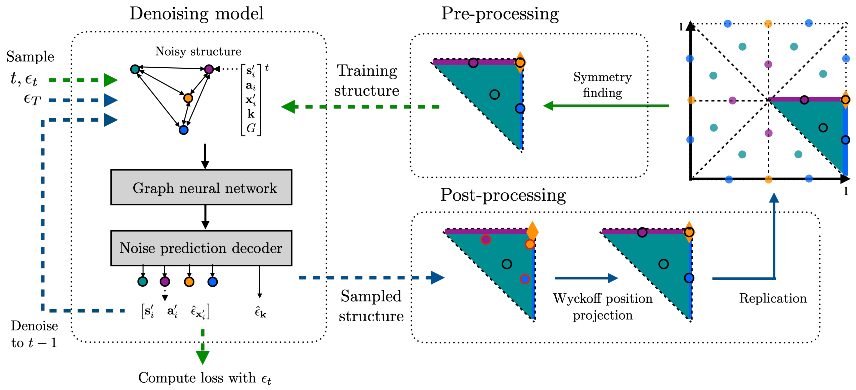


Figure 3: **SymmCD training and sampling pipeline.** For training (green), the crystal structures are pre-processed to find the space group G along with the site-symmetries $\mathbf{S}$ and a set of orbit representatives inside the asymmetric unit. The denoising model is a GNN with fully connected graphs, followed by a decoder. For sampling (blue), the positions are projected to the closest one compatible with their site-symmetry. Then, the asymmetric unit is replicated to obtain the unit cell.

involve translations and so include screw and glide transformations, leading to 26 possible symmetry operations. Further details are included in Appendix C.

4.3 DIFFUSION MODEL

We can now describe the generative model and training process. In SymmCD, the space group and the number of orbit representatives are first sampled from separate distributions obtained from data, such that the distribution over crystal structures is $p(\mathcal{C}) = p(\mathbf{k}, \mathbf{X}', \mathbf{S}, \mathbf{A}' \mid M, G)\, p(M \mid G)\, p(G)$. We will seek to model the conditional distribution $p(\mathbf{k}, \mathbf{X}', \mathbf{S}, \mathbf{A}' \mid M, G)$ with a denoising diffusion model (Sohl-Dickstein et al., 2015; Ho et al., 2020).

We leverage our binary representation for incorporating crystal symmetry information (described in Section 4.2) and perform joint diffusion over lattice representation ($\mathbf{k}$), fractional coordinates of atoms ($\mathbf{X}'$), their types ($\mathbf{A}'$), and the associated binary representation of site symmetry ($\mathbf{S}$).

Diffusion process We consider a separate diffusion process over the different components of the crystal representation. We apply discrete diffusion from Austin et al. (2021) for site-symmetries and atom types. Rather than adding Gaussian noise as in conventional diffusion, we add noise to categorical features by multiplying probability vectors by a transition matrix and sampling from the new probabilities. Inspired by Vignac et al. (2023), the transition matrices are parameterized so that the process converges to the marginals from the data distribution for atom types and site-symmetries. The loss function used for discrete diffusion on atomic types is

$$\mathcal{L}_{\mathbf{A}'} = \mathbb{E}_{\mathbf{a}_t \sim \mathrm{Cat}(\mathbf{a}_0^\top \bar{\mathbf{Q}}_t), t \sim \mathrm{U}(1,T)} \sum_{i=1}^{M} \mathrm{CrossEntropy}(\mathbf{a}_i, \hat{\mathbf{a}}_i), \tag{1}$$

where $\mathbf{a}_0$ is the initial one-hot encoding of the atom types for a single representative and $\bar{\mathbf{Q}}_t = \prod_{i=1}^{t} \mathbf{Q}_i \in \mathbb{R}^{Z \times Z}$ is the cumulative product of transition matrices between timesteps, and $\hat{\mathbf{a}}_i$ are the predicted denoised probabilities. The same loss function is used for site-symmetries.

Continuous diffusion is used for fractional coordinates and lattice parameters, similar to Jiao et al. (2023). The loss function for the continuous diffusion on lattice parameters is

$$\mathcal{L}_{\mathbf{k}} = \mathbb{E}_{\epsilon_{\mathbf{k}} \sim \mathcal{N}(0, \mathbf{I}), t \sim \mathrm{U}(1,T)} [||m \odot \epsilon_{\mathbf{k}} - \hat{\epsilon}_{\mathbf{k}}(\mathcal{C}'_t, t)||_2^2],$$

where m is a space group-dependent mask, and $\hat{\epsilon}_{\mathbf{k}}$ is the predicted denoising vector. The same loss function is used for the fractional coordinates, except that to capture their periodic nature, we use a wrapped normal distribution $\mathcal{WN}(0, 1)^{3 \times M}$. We provide more details in Appendix D.

Denoising network The architecture of the denoiser is a message-passing graph neural network that operates on a fully connected graph of representatives, based on Jiao et al. (2023). Features for each representative $\mathbf{h}_i$ are initialized using an embedding of their atom types $\mathbf{a}_i$ and their site symmetries $\mathbf{S}_i$, along with the graph-level features of the diffusion timestep t, the lattice features $\mathbf{k}$, and an embedding of the space group G. At each layer, messages $\mathbf{m}_{ij}$ are computed between representatives i and j by applying an MLP to $\mathbf{h}_i, \mathbf{h}_j$, and a Fourier basis embedding of the vector $\mathbf{x}_i - \mathbf{x}_j$ to respect periodic invariance. These messages are then used to update $\mathbf{h}_i$. More details on the architecture are included in Appendix E.1. Note that this denoising network is not equivariant. It is not necessary since the unit cell axes provide a canonical reference system (Kaba et al., 2023). We also found that using an equivariant denoising network like E(n)-GNN did not work well in part due to the fact that since we use periodic encodings, the crystal structure input has a translational symmetry. An equivariant model is not be able to break that symmetry (Kaba & Ravanbakhsh, 2023) resulting in a inability to output correct positions in the asymmetric unit (or unit cell).

Putting it all together The algorithm for training our diffusion model is outlined in Algorithm 1. We use different loss coefficients $\lambda_{\mathbf{k}}, \lambda_{\mathbf{X}'}, \lambda_{\mathbf{A}'}$ and $\lambda_{\mathbf{S}}$ to weigh the importance of the different components of the model. The algorithm for sampling from the diffusion model is shown in Algorithm 2. The full pipeline is summarized in Figure 3. Since both the diffusion and the denoising process both operate only on the asymmetric unit, they fully preserve the symmetry of the crystal.

Algorithm 1 Training the Crystal Generation Diffusion Model

```
Input: Dataset of crystals $\mathcal{D}$
while not converged do
    Sample a crystal $\mathcal{C} = (\mathbf{L}, \mathbf{X}, \mathbf{A})$ from dataset $\mathcal{D}$, and a timestep $t \sim \text{Uniform}(1, T)$
    Derive the asymmetric representation $\mathcal{C}' = (G, \mathbf{k}, \mathbf{X}', \mathbf{A}', \mathbf{S})$ from $\mathcal{C}$
    Add noise to $\mathbf{k}$, $\mathbf{X}'$, $\mathbf{A}'$, and $\mathbf{S}'$:
        $\mathbf{k}_t = \sqrt{\bar{\alpha}_t}\mathbf{k}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_{\mathbf{k}}, \quad \epsilon_{\mathbf{k}} \sim \mathcal{N}(0, \mathbf{I})$
        $\mathbf{X}'_t = \sqrt{\bar{\alpha}_t}\mathbf{X}'_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_{\mathbf{X}'}, \quad \epsilon_{\mathbf{X}'} \sim \mathcal{WN}(0, \mathbf{I})$
        $\mathbf{A}'_t \sim \text{Cat}(\mathbf{A}\bar{\mathbf{Q}}_{a,t})$
        $\mathbf{S}_{u,t} \sim \text{Cat}(\mathbf{S}\bar{\mathbf{Q}}_{u,G,t})$
    Use denoising network $\phi$ to predict $\hat{\epsilon}_{\mathbf{k}}, \hat{\epsilon}_{\mathbf{X}'}, \hat{\mathbf{A}}', \hat{\mathbf{S}}$ from noisy $\mathcal{C}_t = (G, \mathbf{k}_t, \mathbf{X}'_t, \mathbf{A}'_t, \mathbf{S}_t)$, $t$
    Compute losses $\mathcal{L}_{\mathbf{k}}, \mathcal{L}_{\mathbf{X}'}, \mathcal{L}_{\mathbf{A}'}, \mathcal{L}_{\mathbf{S}'}$
    Update the denoising network $\phi$ using total loss:
        $\mathcal{L} = \lambda_{\mathbf{k}}\mathcal{L}_{\mathbf{k}} + \lambda_{\mathbf{X}'}\mathcal{L}_{\mathbf{X}'} + \lambda_{\mathbf{A}'}\mathcal{L}_{\mathbf{A}'} + \lambda_{\mathbf{S}}\mathcal{L}_{\mathbf{S}'}$
end while
```

Algorithm 2 Sampling from Crystal Generation Diffusion Model

```
Input: Target space group $G$, Number of representatives $M$
Initialize:
    Sample $\mathbf{k}_T \sim \mathcal{N}(0, \mathbf{I})$
    Sample $\mathbf{X}'_T \sim \mathcal{U}(0, 1)^{3 \times M}$
    Sample $\mathbf{A}'_T \sim p_{\text{marginal}}(\mathbf{A}')$
    Sample $\mathbf{S}'_T \sim p_{\text{marginal}}(\mathbf{S}'|G)$ (site symmetries)
for $t = T$ to 1 do
    Compute $\hat{\epsilon}_{\mathbf{k}}, \hat{\epsilon}_{\mathbf{X}'}, \hat{\mathbf{A}}', \hat{\mathbf{S}}$ using denoising network $\phi(\cdot)$
    Sample $\mathbf{k}_{t-1}$, $\mathbf{X}'_{t-1}$, $\mathbf{A}'_{t-1}$, $\mathbf{S}'_{t-1}$ using $\hat{\epsilon}_{\mathbf{k}}, \hat{\epsilon}_{\mathbf{X}'}, \hat{\mathbf{A}}', \hat{\mathbf{S}}$.
end for
Project $\mathbf{S}'_0$ onto nearest valid point group
Project $\mathbf{X}'_0$ onto nearest Wyckoff position with that site symmetry
Replicate representative atoms $\mathbf{X}'_0$ using site symmetries $\mathbf{S}'_0$ to generate full crystal $\mathbf{X}_0$
Output: Crystal structure $\mathbf{X}_0$, Atom types $\mathbf{A}_0$, lattice $\mathbf{L}_0$
```

5 EXPERIMENTS

We test our model on *de novo* crystal generation using the MP-20 dataset (Xie et al., 2022), a subset of the Materials Project (Jain et al., 2013) consisting of 40,476 crystals, each with up to 20 atoms per primitive unit cell. The data is preprocessed to use the conventional unit cell rather than the

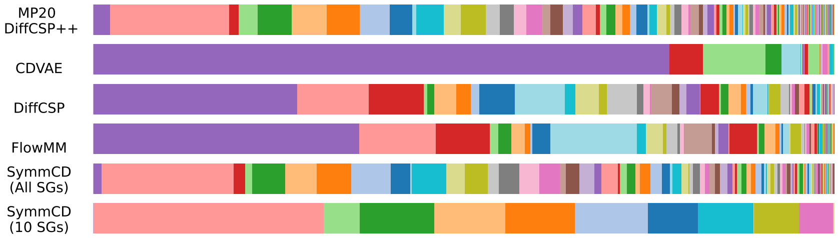


Figure 4: Proportion of space group symmetries of the dataset, and each method. The width of each color segment represents the proportion of crystals with that symmetry. From left to right, the first few space groups are: P1, Fm$\bar{3}$m, Cm, P$\bar{1}$, C2/m, I4/mmm, Pm$\bar{3}$m, P6_3/mmc, and Pm.

primitive unit cell, as the former has more conveniently expressed symmetries and constraints. A conventional unit cell may be larger than a primitive unit cell, which results in up to 80 atoms in the unit cell. We withhold 20% of the dataset as a validation set, and 20% as a test set.

We empirically demonstrate our contributions, particularly in ensuring we generate crystals with desired symmetries while being competitive with existing baselines. In other words, we show that SymmCD generates symmetric, stable, and valid crystals. We compare our proposed method with four recent strong baselines: CDVAE (Xie et al., 2022), DiffCSP (Jiao et al., 2023), DiffCSP++ (Jiao et al., 2024) and FlowMM (Miller et al., 2024).

5.1 SYMMETRY AND STRUCTURAL DIVERSITY

First, we evaluate the different methods on their ability to generate crystals with diverse structures and space groups. This aspect has not been investigated yet for the considered baselines, yet it is significant in understanding if they generate realistic structures.

Table 1: Template statistics for various models.

Method	# Unique	% in Train	# New
Training Set	3318	100	-
CDVAE	797	**28.7**	568
DiffCSP	1347	43.2	764
DiffCSP++	1905	94.2	110
FlowMM	1291	41.7	753
SymmCD	**2794**	40.8	**1654**

Space groups To detect the space group of the generated structures, we use spglib's symmetry finding method (Togo & Tanaka, 2018a; Ong et al., 2013) with a tolerance of 0.1Å. This is applied to 10,000 crystals sampled from each model. The distribution of space groups of the generated structures is shown in Figure 4. It can be observed that while SymmCD matches the highly diverse data distribution, CDVAE mostly generates crystals with trivial $P1$ symmetry, and DiffCSP and FlowMM generate many crystals with low symmetry and generally have lower diversity of space groups. We also consider a new quantitative metric to characterize the space group distribution, d_{sg}, which is calculated as the Jensen-Shannon distance between the distribution of space groups of the generated structures and the test set. We report it for the different methods in the rightmost column of Table 2. The results confirm that SymmCD and DiffCSP++ are the only methods that accurately match the distribution of space groups in the dataset.

Unique Templates We also evaluate the ability of the different methods to generate diverse crystal structures. We define a structural *template* to be a combination of a space group and a multiset of occupied Wyckoff positions, regardless of the atomic types in the Wyckoff position. Templates, also known as Wyckoff sequences, are used in practice to classify crystals by their symmetry. They have the advantage of providing a notion of a structure that is highly flexible, while being robust to perturbations of coordinates that do not change the position of atoms with respect to symmetry elements. Most potential templates have not yet been experimentally observed, motivating the development of methods that can discover materials with new templates (Hornfeck, 2022).

The training dataset contains 3318 such unique templates. We examine the templates for the 10,000 crystals generated by each method, and report results in Table 1. We find that SymmCD performs best out of all models, proposing unique and novel templates. This highlights an important limitation of DiffCSP++. While it is able to produce diverse space groups and to a certain extent diverse

Table 2: Results for comparing the validity, coverage, and property distribution metrics. Best results in each category are bolded.

	Validity (%) (↑)		Coverage (%) (↑)		Property Distribution (↓)			
	Struct.	Comp.	Recall	Precision	d_ρ	d_{E}	d_{elem}	d_{sg}
CDVAE	99.93	**86.93**	98.31	99.35	0.9144	0.1645	1.6538	0.7263
DiffCSP	99.61	82.23	99.53	99.35	0.2565	0.1402	0.4027	0.4446
DiffCSP++	**99.99**	85.81	99.48	**99.66**	0.2779	**0.0872**	0.4079	**0.0771**
FlowMM	96.43	83.37	99.47	99.71	0.2905	0.1072	**0.0788**	0.5137
SymmCD (All SGs)	94.32	85.85	**99.64**	98.87	**0.0901**	0.1166	0.3990	0.0899
SymmCD (10 SGs)	97.31	87.10	97.21	99.42	0.2829	0.1510	0.1769	0.4737

templates, since it uses pre-defined templates it fails to generate structures with *novel* templates. Our method does not suffer from this problem since it learns to generate templates.

5.2 PROXY METRICS

We compare the different methods using the metrics established by Xie et al. (2022), measuring the validity, coverage, and property statistics of the generated crystals. We measure the validity by checking structural validity, defined as whether no two atoms are closer than 0.5 Å apart, and compositional validity, defined as whether the charges are balanced as determined by SMACT (Davies et al., 2019)[3]. To determine coverage, we examine the CrystalNN structural fingerprints (Zimmermann & Jain, 2020) and Magpie compositional fingerprints (Ward et al., 2016) of the generated crystals, and look at their distances to the fingerprints of the crystals in the test set. This gives us recall and precision metrics. We look at the distances between the properties of the generated crystals and the crystals from the test set to compare the ability of each model to match the data distribution. We specifically compare the Wasserstein distances between the atomic densities d_ρ, number of unique elements d_{elem}, and predicted formation energy d_{E}. The results are shown in Table 2. We observe that SymmCD performs on par with other methods across different metrics, and that sampling from a smaller set of space groups improves the validity of crystals while trading off diversity and matching the data distribution. These results also show that SymmCD can generalize to generate valid structures even for groups which are rarely represented in the training data. We also consider a variant of the model (10 SGs) where we sample space groups from the MP-20 distribution, restricted to the 10 most common space groups, similar to (Cao et al., 2024) [4] . This is to provide a more nuanced comparison with other methods, which are not constrained in matching the space group distribution. This choice still captures a large portion of the data distribution, since these are the most prevalent space groups.

5.3 STABLE, UNIQUE AND NOVEL (S.U.N.) STRUCTURES

Regardless of their target application, generative models for crystals should produce sets of crystals that are thermodynamically stable, unique (not duplicated within the predicted set), and novel (not already in the training data), or S.U.N. To this end, we adapt the evaluation procedure of Miller et al. (2024) to assess the capability of our model to generate S.U.N. materials. Thermodynamic stability is determined by estimating the energy of a material with respect to a *convex hull*. The convex hull gives linear combinations of known phases that represent the lowest-energy mixtures of materials; if a material has an energy above the hull, it is energetically favorable for it decompose into a com-

Table 3: Number of stable and S.U.N. samples produced from an initial set of 1000 generated crystals for each method.

	Initial Stable	Relaxed Stable	Relaxed S.U.N.
CDVAE	0.1%	3.6%	3.5%
DiffCSP	**8.9%**	12.5%	9.7%
DiffCSP++	**8.9%**	**13.2%**	9.1%
FlowMM	4.1%	9.3%	6.3%
SymmCD (all SGs)	5.0%	9.4%	7.0%
SymmCD (10 SGs)	7.9%	11.7%	**9.9%**

---

[3] It should be noted that the compositional validity of the MP-20 dataset is only 92%.

[4] These space groups are numbered: 2, 12, 14, 62, 63, 139, 166, 194, 221, 225

bination of these stable phases and is therefore thermodynamically unstable. We assess the stability of generated crystals by estimating their energies using a pretrained CHGNet model (Deng et al., 2023), and comparing that to a convex hull computed for Materials Project (Riebesell et al., 2024).

For each method, we randomly sub-sample 1000 crystals of the 10,000 generated samples and predict their stability. We also use CHGNet to compute relaxed structures for each crystal, which results in higher stability. Finally, we check whether the stable relaxed crystals are also unique and novel. Details of this procedure are included in Miller et al. (2024). Note that we use a machine learning potential instead of a full Density Functional Theory (DFT) calculation, as DFT relaxation would be orders of magnitude more expensive to compute.

The results are shown in Table 3. SymmCD (all SGs) performs slightly better than FlowMM in generating stable structures, but worse than DiffCSP and DiffCSP++. The version sampling from a smaller number of space groups however obtains a larger proportion of S.U.N. structures than all baselines. Note that, while DiffCSP++ has a larger proportion of relaxed stable structures, filtering for unique and novel structures gives SymmCD the advantage, providing more evidence that it generates more diverse structures.

5.4 COMPUTATIONAL EFFICIENCY

Finally, we demonstrate significant computational efficiency gains and reduced memory footprint due to using a more compact representation based on crystallographic orbits. We compare our model to an equivalent model that looks at a full unit cell, rather than just the asymmetric unit. It also uses a fully connected graph to represent the atoms in the unit cell, but unlike SymmCD, it does not use site symmetry representations as they are not necessary. This makes the model essentially equivalent to DiffCSP (Jiao et al., 2023), but with the same architecture and hyperparameters as SymmCD for consistent comparison. We compare the two representations for an epoch of training using 40GB of RAM and a single NVIDIA MIG A100 instance and report the results in Table 4. These results highlight SymmCD's memory efficiency and faster training capabilities.

Table 4: Computational efficiency of our compact representation with a 40 GB NVIDIA MIG A100 instance.

	Asymmetric Unit (ours)	Conventional Unit Cell
Maximum batch size (↑)	**8192**	512
Memory for 512 batch size (↓)	**3.6 GB**	31 GB
Time for one training epoch (↓)	**27 sec.**	52 sec.

CONCLUSION

In this paper, we introduced a novel approach for generating crystals with precise symmetry properties. We proposed to leverage asymmetric units and site-symmetry representations within a diffusion model framework. This approach ensures that the generated crystals inherently preserve desired symmetries while allowing greater diversity, computational efficiency and flexibility in the generation process. To encode crystal and site symmetries we introduced a new representation of crystal symmetries that enables information sharing across space groups, improving generalization when learning with a diverse set of crystal symmetries. Our results indicate that this method produces valid, stable, novel, and structurally diverse crystals, and improves computational efficiency of the generative model showing promise for discovery in materials science. In this work, we focused on inorganic crystals, but SymmCD could potentially be promising for applications on molecular crystals and co-crystals, which also have non-trivial symmetries. Beyond materials, other data types such as molecules and graphs often exhibit complex symmetries. Future work could investigate if symmetry constraints could also be useful in these modalities to improve generative models.

One limitation of our framework is that it makes it more challenging to perform crystal structure prediction given a composition, since it relies on sampling a space group first, and then a composition conditioned on the space group. Finally, an important area of future work in generative models for crystals is also to go beyond single crystals, and consider generation of polycrystalline materials. These types of materials are common in applications, yet not suited to generation using single unit cells or asymmetric units.

REFERENCES

Mila AI4Science, Alex Hernandez-Garcia, Alexandre Duval, Alexandra Volokhova, Yoshua Bengio, Divya Sharma, Pierre Luc Carrier, Michał Koziarski, and Victor Schmidt. Crystal-gfn: sampling crystals with desirable properties and constraints. *arXiv preprint arXiv:2310.04925*, 2023.

Nawaf Alampara, Santiago Miret, and Kevin Maik Jablonka. Mattext: Do language models need more than text & scale for materials modeling? In *AI for Accelerated Materials Design-Vienna 2024*, 2024.

Mois Ilia Aroyo. *International Tables for Crystallography*. John Wiley and Sons Limited, 2013.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Zhendong Cao, Xiaoshan Luo, Jian Lv, and Lei Wang. Space group informed transformer for crystalline materials generation. *arXiv preprint arXiv:2403.15734*, 2024.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

Anthony K Cheetham and Ram Seshadri. Artificial intelligence driving materials discovery? perspective on the article: Scaling deep learning for materials discovery. *Chemistry of Materials*, 36 (8):3490–3495, 2024.

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.

Daniel W Davies, Keith T Butler, Adam J Jackson, Jonathan M Skelton, Kazuki Morita, and Aron Walsh. Smact: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software*, 4(38):1361, 2019.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023.

JDH Donnay and G Turrell. Tables of oriented site symmetries in space groups. *Chemical Physics*, 6(1):1–18, 1974.

David Steven Dummit and Richard M Foote. *Abstract algebra*, volume 3. Wiley Hoboken, 2004.

Alexandre Duval, Simon V Mathis, Chaitanya K Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D Malliaros, Taco Cohen, Pietro Liò, Yoshua Bengio, and Michael Bronstein. A hitchhiker's guide to geometric gnns for 3d atomic systems. *arXiv preprint arXiv:2312.07511*, 2023.

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.

Scott Fredericks, Kevin Parrish, Dean Sayre, and Qiang Zhu. Pyxtal: A python library for crystal structure generation and symmetry analysis. *Computer Physics Communications*, 261:107810, 2021. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2020.107810. URL `http://www.sciencedirect.com/science/article/pii/S0010465520304057`.

Victor Garcia Satorras, Emiel Hoogeboom, Fabian Fuchs, Ingmar Posner, and Max Welling. E (n) equivariant normalizing flows. *Advances in Neural Information Processing Systems*, 34:4181–4192, 2021.

Johannes Gasteiger, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020a.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020b.

Rhys E. A. Goodall, Abhijith S. Parackal, Felix A. Faber, Rickard Armiento, and Alpha A. Lee. Rapid discovery of stable materials by coordinate-free coarse graining. *Science Advances*, 8(30): eabn4117, 2022.

Prashant Govindarajan, Santiago Miret, Jarrid Rector-Brooks, Mariano Phielipp, Janarthanan Rajendran, and Sarath Chandar. Learning conditional policies for crystal design using offline reinforcement learning. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023. URL `https://openreview.net/forum?id=VbjD8w2ctG`.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ward Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *The Twelfth International Conference on Learning Representations*, 2024.

Theo Hahn, Uri Shmueli, and JC Wilson Arthur. *International tables for crystallography*, volume 1. Reidel Dordrecht, 1983.

Mariette Hellenbrandt. The inorganic crystal structure database (icsd)—present and future. *Crystallography Reviews*, 10(1):17–22, 2004.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Wolfgang Hornfeck. On the combinatorics of crystal structures: number of wyckoff sequences of given length. *Acta Crystallographica Section A: Foundations and Advances*, 78(2):149–154, 2022.

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.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=DNdN26m2Jk`.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. *arXiv preprint arXiv:2402.03992*, 2024.

Oumar Kaba and Siamak Ravanbakhsh. Equivariant networks for crystal structures. *Advances in Neural Information Processing Systems*, 35:4150–4164, 2022.

Sékou-Oumar Kaba and Siamak Ravanbakhsh. Symmetry breaking and equivariant neural networks. *arXiv preprint arXiv:2312.09016*, 2023.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning*, pp. 15546–15566. PMLR, 2023.

Sungwon Kim, Juhwan Noh, Geun Ho Gu, Alan Aspuru-Guzik, and Yousung Jung. Generative adversarial networks for crystal structure prediction. *ACS central science*, 6(8):1412–1420, 2020.

Astrid Klipfel, Yaël Fregier, Adlane Sayede, and Zied Bouraoui. Vector field oriented diffusion model for crystal material generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22193–22201, 2024.

Kin Long Kelvin Lee, Carmelo Gonzales, Marcel Nassar, Matthew Spellings, Mikhail Galkin, and Santiago Miret. Matsciml: A broad, multi-task benchmark for solid-state materials modeling. *arXiv preprint arXiv:2309.05934*, 2023.

Youzhi Luo, Chengkai Liu, and Shuiwang Ji. Towards symmetry-aware generation of periodic materials. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=Jkc74vn1aZ`.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, pp. 1–6, 2023.

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.

Santiago Miret, Kin Long Kelvin Lee, Carmelo Gonzales, Marcel Nassar, and Matthew Spellings. The open matsci ML toolkit: A flexible framework for machine learning in materials science. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=QBMyDZsPMd`.

Santiago Miret, NM Anoop Krishnan, Benjamin Sanchez-Lengeling, Marta Skreta, Vineeth Venugopal, and Jennifer N Wei. Perspective on ai for accelerated materials design at the ai4mat-2023 workshop at neurips 2023. *Digital Discovery*, 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Juhwan Noh, Jaehoon Kim, Helge S Stein, Benjamin Sanchez-Lengeling, John M Gregoire, Alan Aspuru-Guzik, and Yousung Jung. Inverse design of solid-state materials via a continuous representation. *Matter*, 1(5):1370–1384, 2019.

Asma Nouira, Nataliya Sokolovska, and Jean-Claude Crivello. Crystalgan: learning to discover crystallographic structures with generative adversarial networks. *arXiv preprint arXiv:1810.11203*, 2018.

John Frederick Nye. *Physical properties of crystals: their representation by tensors and matrices*. Oxford university press, 1985.

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, June 2013. URL `https://github.com/materialsproject/pymatgen`.

Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Alpha A. Lee, Anubhav Jain, and Kristin A. Persson. Matbench discovery – a framework to evaluate machine learning crystal stability predictions, 2024. URL `https://arxiv.org/abs/2308.14920`.

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.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Atsushi Togo and Isao Tanaka. `Spglib`: a software library for crystal symmetry search. `https://github.com/spglib/spglib`, 2018a.

Atsushi Togo and Isao Tanaka. Spglib: a software library for crystal symmetry search. *arXiv preprint arXiv:1808.01590*, 5, 2018b.

Clément Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):1–7, 2016.

Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=03RLpj-tc_`.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

Keqiang Yan, Yi Liu, Yuchao Lin, and Shuiwang Ji. Periodic graph transformers for crystal material property prediction. *Advances in Neural Information Processing Systems*, 35:15066–15080, 2022.

Keqiang Yan, Alexandra Saxton, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. A space group symmetry informed network for o (3) equivariant crystal tensor prediction. *arXiv preprint arXiv:2406.12888*, 2024.

Mengjiao Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable diffusion for materials generation. *arXiv preprint arXiv:2311.09235*, 2023.

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.

Ruiming Zhu, Wei Nong, Shuya Yamazaki, and Kedar Hippalgaonkar. Wycryst: Wyckoff inorganic crystal generator framework. *Matter*, 7(10):3469–3488, 2024.

Nils ER Zimmermann and Anubhav Jain. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC advances*, 10 (10):6063–6081, 2020.

A REPLICATION

We define the replication operator as $\mathsf{R} : \mathcal{G} \times \mathcal{P} \times \mathbb{R}^3 \to 2^{\mathbb{R}^3}$. This operation is defined by considering the group $S_{\mathbf{x}} \ltimes T_{\mathbf{S}}$, with $T_{\mathbf{S}}$ being the group of translations defined by the lattice. $S_{\mathbf{x}} \ltimes T_{\mathbf{S}}$ is the set of operations that preserve the position of $\mathbf{x}$ *within* the unit cell as opposed to within the crystal. We can then consider the coset decomposition of the space group with respect to that group $G/\left(S_{\mathbf{x}} \ltimes T_{\mathbf{S}}\right)$. Then, we denote by $\left[G/\left(S_{\mathbf{x}} \ltimes T_{\mathbf{S}}\right)\right]_{\mathbf{0}}$ a system of coset representatives where the translation parts are chosen to move only within the unit cell. This defines the set of operations that move a position $\mathbf{x}$ within its orbit and the unit cell. The replication operation then simply consists of applying all these operations:

$$\mathsf{R}\left(G, S_{\mathbf{x}}, \mathbf{x}\right) = \left(\left(\mathbf{O}, \mathbf{t}\right)\mathbf{x} \mid \left(\mathbf{O}, \mathbf{t}\right) \in \left[G/\left(S_{\mathbf{x}} \ltimes T_{\mathbf{S}}\right)\right]_{\mathbf{0}}\right)$$

The representation in terms of individual atoms is then:

$$\mathbf{X} = \bigoplus_{i}^{M} \mathsf{R}\left(G, S_{\mathbf{x}'_i}, \mathbf{x}'_i\right) \tag{2}$$

$$\mathbf{A} = \bigoplus_{i}^{M} \mathrm{repeat}\left(\mathbf{a}_i, \left[G : \left(S_{\mathbf{x}} \ltimes T_{\mathbf{S}}\right)\right]\right) \tag{3}$$

where $\mathrm{repeat}\left(\mathbf{a}, n\right)$ repeats the vector $\mathbf{a}$ n times and $\left[G : \left(S_{\mathbf{x}} \ltimes T_{\mathbf{S}}\right)\right]$ is the multiplicity of the orbit.

In our diffusion model, our predicted site symmetries $\hat{\mathbf{S}}$ do not always necessarily correspond to a valid crystallographic point group. To get around this, we project $\hat{\mathbf{S}}$ to the nearest point group that is a subgroup of the given space group, as measured by the Frobenius Norm of their difference. Once a point group is chosen, the PyXtal `search_closest_wp` function is used to get the nearest coordinates to $\mathbf{X}'$ that correspond to a Wyckoff position with the given site symmetry, and $\mathbf{X}'$ is updated to be placed on those coordinates (Fredericks et al., 2021). Finally, the representative atoms at the Wyckoff position are replicated, using operations implemented in PyXtal.

B LATTICE REPRESENTATION

We use the lattice representations derived by Jiao et al. (2024), as they are useful for constraining lattices to respect the symmetries of a given space group. The authors found that any lattice matrix $\mathbf{L}$ can be written as $\mathbf{L} = \mathbf{Q}\exp(\mathbf{S})$ for some orthogonal $\mathbf{Q}$ (which we can ignore, as orthogonal transformations do not change the lattice), and symmetric $\mathbf{S}$. The matrix $\mathbf{S}$ can then be decomposed into a sum of the following basis lattices:

$$\mathbf{B}_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{B}_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \mathbf{B}_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

$$\mathbf{B}_4 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{B}_5 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}, \quad \mathbf{B}_6 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

with $\mathbf{S} = \sum_{i=1}^{6} k_i \mathbf{B}_i$. They derive constraints on k_i depending on the space groups that a crystal belongs to:

- Triclinic: $\mathbf{k} = (k_1, k_2, k_3, k_4, k_5, k_6)$
- Monoclinic: $\mathbf{k} = (0, k_2, 0, k_4, k_5, k_6)$
- Orthorhombic: $\mathbf{k} = (0, 0, 0, k_4, k_5, k_6)$
- Tetragonal: $\mathbf{k} = (0, 0, 0, 0, k_5, k_6)$
- Hexagonal: $\mathbf{k} = (-\log(3)/4, 0, 0, 0, k_5, k_6)$
- Cubic: $\mathbf{k} = (0, 0, 0, 0, 0, k_6)$

C SITE SYMMETRY REPRESENTATION

The 15 possible symmetry axes of crystals are: [001], [010], [100], [111], [$1\bar{1}1$], [$\bar{1}11$], [$\bar{1}\bar{1}1$], [110], [$1\bar{1}0$], [101], [$10\bar{1}$], [011], [$01\bar{1}$], [210], [120], [$1\bar{1}0$]. These axes are written in short form: for example, [$1\bar{1}0$] denotes the direction of the vector $(1, -1, 0)$. The axes depend on the symmetries of the crystal system: for example, in an orthorhombic crystal (a rectangular prism whose side lengths are not necessarily equal), a crystal may have different site-symmetries oriented around the x, y, or z-axes. Conversely, in a tetragonal crystal (a rectangular prism with a square base), any site-symmetry oriented along the x-axis must also be along the y-axis, there may be additional symmetries along the diagonal of the x-y plane.

The possible set of symmetry elements along each axis for a site symmetry group correspond to the identity 1; an inversion $\bar{1}$; rotations of different orders 2, 3, 4, and 6; rotoinversions $\bar{2}$ (equivalent to a mirror symmetry m across a plane perpendicular to the axis), $\bar{3}$, $\bar{4}$, and $\bar{6}$; and combinations of rotations and mirror reflections $2/m$, $4/m$, and $6/m$. This enumeration yields 13 possible symmetries along each axis.

The possible symmetry elements along each axis for a space group correspond to the identity 1; an inversion $\bar{1}$; rotations of different orders 2, 3, 4, and 6; rotoinversions $\bar{2}$ (equivalent to a mirror symmetry m across a plane perpendicular to the axis), $\bar{3}$, $\bar{4}$, and $\bar{6}$; screws 2_1, 3_1, 3_2, 4_1, 4_2, 4_3, 6_1, 6_2, 6_3, 6_4, 6_5, and glides a, b, c, n, d, e.

To encode a space group, an additional 7-dimensional one-hot encoding is used to denote the Bravais lattice to which the space group belongs. This yields a $(26 \times 15) + 7 = 397$ dimensional binary representation of space group.

D DIFFUSION AND DENOISING PROCESS DETAILS

Diffusion on lattice parameters $\mathbf{k}$ Inspired by Jiao et al. (2024), we perform diffusion over $\mathbf{k}$, the $O(3)$-invariant lattice representation. The forward noising process is given by $q(\mathbf{k}_t|\mathbf{k}_0) \sim \mathcal{N}(\mathbf{k}_t|\sqrt{\bar{\alpha}_t}\mathbf{k}_0, (1-\bar{\alpha}_t)\mathbf{I})$, where $\mathbf{k}_t$ is the noised version of $\mathbf{k}_0$ at timestep t. Here, similar to Nichol & Dhariwal (2021), $\bar{\alpha}_t = \Pi_{j=1}^{t}(1-\beta_j)$, where $\beta_j \in (0,1)$ determines variance in each step controlled by the cosine scheduler. During the generation process, we start with $\mathbf{k}_T \sim \mathcal{N}(0, \mathbf{I})$ and use learned denoising network to generate $\mathbf{k}_{t-1}$ from $\mathbf{k}_t$:

$$p_\theta(\mathbf{k}_{t-1}|\mathcal{C}'_t) = \mathcal{N}\Big(\mathbf{k}_{t-1}|\mu_{\mathbf{k}}(t), \sigma(t)\mathbf{I}\Big),$$

$$\mu_{\mathbf{k}}(t) = \frac{1}{\sqrt{\bar{\alpha}_t}}\Big(\mathbf{k}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}_{\mathbf{k}}(\mathcal{C}'_t, t)\Big), \sigma(t) = \beta_t\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}.$$

Here, $\mathcal{C}'_t$ is the noised crystal and $\hat{\epsilon}_k(\mathcal{C}'_t, t)$ is the predicted denoising term predicted from a denoising network $\phi(\mathcal{C}'_t, t)$. We also use a mask m to only implement diffusion over unconstrained dimensions of $\mathbf{k}_t$, since depending upon space groups, certain dimensions have fixed values (Appendix B). The mask can be represented as $m \in \{0,1\}^6$ and $m_i = 1$ indicates that i^{th} index of $\mathbf{k}$ is unconstrained. The corresponding loss used to train the denoising network is:

$$\mathcal{L}_{\mathbf{k}} = \mathbb{E}_{\epsilon_{\mathbf{k}} \sim \mathcal{N}(0,\mathbf{I}), t \sim \mathrm{U}(1,T)}[||m \odot \epsilon_{\mathbf{k}} - \hat{\epsilon}_{\mathbf{k}}(\mathcal{C}'_t, t)||_2^2]$$

where $\odot$ is the elementwise product and $U(1,T)$ is a uniform distribution over timesteps.

Diffusion over representative fractional coordinates $\mathbf{X}'$ We perform diffusion over the fractional coordinates using the same method as (Jiao et al., 2023). Due to the periodicity of fractional coordinates, the noising process $q(\mathbf{X}_t|\mathbf{X}_0)$ is determined by a Wrapped Normal distribution rather than a Gaussian distribution, and we initialize the fractional coordinates $\mathbf{X}_T$ with the uniform distribution $U(0,1)$ when sampling.

Diffusion on atom types $\mathbf{A}'$ We use discrete diffusion from Austin et al. (2021) to sample the atom types of each representative. If $\mathbf{a}_0 \in \{0,1\}^Z$ is the one-hot encoding of atom types for a single representative, then we can noise it as: $q(\mathbf{a}_t|\mathbf{a}_0) = \mathrm{Cat}(\mathbf{a}_t; \mathbf{p} = \mathbf{a}_0^\top\bar{\mathbf{Q}}_t)$, where $\bar{\mathbf{Q}}_t = \prod_{i=1}^{t}\mathbf{Q}_i \in$

$\mathbb{R}^{Z\times Z}$ is the cumulative product of transition matrices between timesteps. Inspired by Vignac et al. (2023), the transition matrix can be parametrized quite simply as $\mathbf{Q}_t = \alpha_t \mathbf{I} + \beta_t \mathbf{m}_a$, where $\mathbf{m}_a$ are the marginals over the atom types in the data, and α_t and β_t are scheduling parameters. The effect of this noising scheme is that regardless of $\mathbf{a}_0$, the fully noised $\mathbf{a}_T = \mathbf{a}_0^\top \mathbf{Q}_T = \mathbf{m}_a$, so we can sample from the prior distribution $\mathbf{m}_a$, which is close to the data distribution. The discrete diffusion model is trained using a cross-entropy loss:

$$\mathcal{L}_{\mathbf{A}'} = \mathbb{E}_{\mathbf{a}_t \sim \mathrm{Cat}(\mathbf{a}_0^\top \bar{\mathbf{Q}}_t), t\sim \mathrm{U}(1,T)} \sum_{i=1}^{M} \mathrm{CrossEntropy}(\mathbf{a}_i', \hat{\mathbf{a}}_i'), \tag{4}$$

where $\hat{\mathbf{a}}_i$ are the probabilities predicted by the model $\phi(\mathcal{C}_t', t)$. To sample from the discrete diffusion model, we sample from the marginal distribution over atom types $\mathbf{m}_a$, then progressively denoise using:

$$q(\mathbf{a}_{t-1}|\mathbf{a}_t, \mathbf{a}_0) = \mathrm{Cat}\left(\mathbf{a}_{t-1}; \mathbf{p} = \frac{\mathbf{a}_t^\top \mathbf{Q}_t^\top \odot \mathbf{a}_0^\top \bar{\mathbf{Q}}_{t-1}}{\mathbf{a}_0^\top \bar{\mathbf{Q}}_t \mathbf{a}_t}\right) \tag{5}$$

More details of this implementation can be seen in Vignac et al. (2023).

Diffusion for site-symmetries S The site-symmetry representation matrices described in Section 4.2 can be thought of as 15 separate 13-dimensional categorical variables: one site-symmetry operation per axis. Our diffusion model over site-symmetries is almost identical to the method for atom types, applying discrete diffusion separately over each of the axes. Because the site-symmetries depend strongly on the space group, we use transition matrices that are different for each space group: $\mathbf{Q}_{t,i,G} = \alpha_t \mathbf{I} + \beta_t \mathbf{m}_{S_\alpha,G}$, where $\mathbf{m}_{S_u,G}$ denotes the marginals over site-symmetry operations for axis S_u given space group G. For each representative node, we average the cross-entropy loss over each of the axes.

E ARCHITECTURE DETAILS

E.1 DENOISING MODEL

We use a graph neural network based on the architecture of Jiao et al. (2023). We embed the timestep t using sinusoidal embeddings, $\psi_t(t)$. We embed our space group representation from Section 4.2 using an MLP, $\phi_G(G)$. We embed our site symmetries by separately embedding each axis using the same network, and feeding the resulting embeddings into a secondary MLP: $\phi_S(\bigoplus_{u=1}^{15} \phi_U(S_u))$. These are all used to initialize the node embeddings $\mathbf{h}_i$.

$$\mathbf{h}_i \leftarrow \phi_h(\mathbf{a}_i, \mathbf{x}_i, \phi_S\left(\bigoplus_{u=1}^{15} \phi_U(S_u)\right), \phi_G(G), \psi_t(t)).$$

As noted earlier, we directly use coordinates $\mathbf{x}$, because we are working a conventional or canonical lattice, and so Euclidean symmetries are not necessarily useful here.

At each layer we compute messages and use them to update node embeddings:

$$\mathbf{m}_{ij} \leftarrow \phi_m(\mathbf{h}_i, \mathbf{h}_j, \mathbf{k}, \psi(\mathbf{x}_i - \mathbf{x}_j))$$

$$\mathbf{h}_i \leftarrow \mathbf{h}_i + \phi_h(\mathbf{h}_i, \sum_j^M \mathbf{m}_{ij})$$

Here, ψ is a Fourier embedding, ϕ_m and ϕ_h are MLPs acting on edges and nodes respectively. We use a SiLU activation function for each MLP. Finally, we output predicted $\hat{\epsilon}_{\mathbf{X}'}$, $\hat{\mathbf{A}}'$ and $\hat{\mathbf{S}}$ using the node embeddings $\mathbf{h}_i$, and $\hat{\epsilon}_{\mathbf{k}}$ using $\sum_i^M \mathbf{h}_i$.

E.2 MODEL HYPERPARAMETERS

The graph neural network has 8 layers, and we use a representation dimension of 1024 for $\mathbf{h}_i$. We encode distances between nodes using a sinusoidal embedding, with 128 different frequencies. We encode the timestep t into a 10 dimensional vector. We apply layer normalization at each layer of the GNN. The loss coefficients selected were $\lambda_{\mathbf{k}} = 5, \lambda_{\mathbf{X}'} = 1, \lambda_{\mathbf{A}'} = 0.1$ and $\lambda_{\mathbf{S}} = 10$.

We performed two hyperparameter sweeps: we first tested each combination of $\lambda_{\mathbf{k}}$, $\lambda_{\mathbf{A}'}$ and $\lambda_{\mathbf{S}}$ set to values in $\{0.1, 0.5, 1, 5, 10\}$, while keeping $\lambda_{\mathbf{X}'}$ fixed at 1, and then selected the loss coefficients that lead to the highest structural validity. Next, we performed a random sweep of other architecture parameters, running 150 different hyperparameter combinations and choosing a model that had high performance on structural validity, compositional validity, and d_E. We varied the number of GNN layers in $\{6, 8, 12, 16\}$, representation dimension in $\{256, 512, 1024\}$, time embedding dimension in $\{10, 64, 256\}$, and varied whether layer normalization was used.

F ADDITIONAL RESULTS

F.1 NUMBER OF ATOMS

To demonstrate that SymmCD is able to correctly predict reasonable site symmetries, we show here that the distribution of number of atoms per crystal matches the dataset it is trained on. This is not a trivial task, as the model needs to learn the multiplicity of different possible site symmetries, which depends on both the different symmetry elements of the site symmetry and the space group that it belongs to.

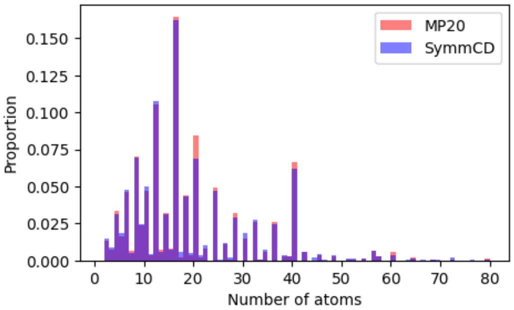


Figure 5: Histogram of number of atoms in crystals from MP-20 and generated by SymmCD.

F.2 PROPERTY PREDICTION TASK

We test the usefulness of our site symmetry representation using a regression experiment. We selected formation energy per atom as the target property to predict. We use DimeNet++ (Gasteiger et al., 2020b;a) as a base model to perform ablation over the type of input graph and encoding site symmetry information per node.

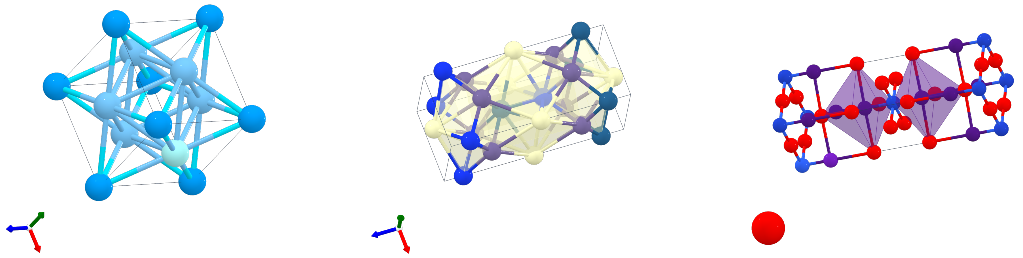
One input format is a multi-graph (Xie et al., 2022), which describes the unit cell as a graph with nodes as atoms and edges between them according to a cutoff radius. These edges could potentially span to neighbouring unit cells. the other input format is the asymmetric unit that we use in SymmCD. Under these two inputs, we test the effects of including a site symmetry encoding for each node. We report the Mean Absolute Error (MAE) for the test set in Table 5. We see that the effect of including site symmetry information is minimal when we have access to the full graph. However, we see that when we are restricted to only using the asymmetric unit, having access to the site symmetry info greatly helps, showing that we can recover some geometric information lost when using just an asymmetric unit by also including symmetry.

Table 5: Mean average error when predicting crystal formation energy. The input could be the asymmetric unit or a multi-graph, and the site symmetry information can be encoded or ignored. We observe that our encoding of site symmetry helps predict the target property.

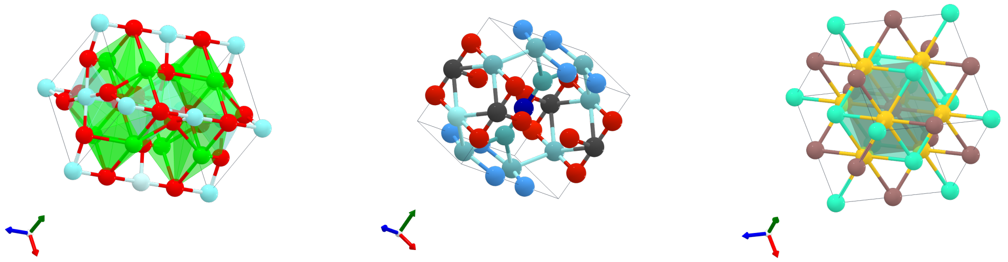
	Multigraph	Asymm. Unit
W/out **S**	0.0214	0.0711
With **S**	0.0212	0.0490

F.3 EXAMPLES

In Figure 6, we include 6 randomly sampled crystals generated by SymmCD along with their respective space groups.



(a) $PaTi_3$ $Pm\bar{3}m$ (b) $CeSiGe_2Os$ $I4mm$ (c) Cs_2CuO_4 $I4_{m}mm$



(d) $SrYO_3$ $I4_{m}cm$ (e) $Ta_2Nb_4V_4CoMo_2C$ $Pmmm$ (f) $TbInAu_2$ $Fm\bar{3}$

Figure 6: Example materials generated by SymmCD, along with their chemical formulate and space groups symmetries.