CRYSLDM: LATENT DIFFUSION MODEL FOR CRYSTAL MATERIAL GENERATION

Subhojyoti Khastagir[1,*], **Kishalay Das**[1,*], **Pawan Goyal**[1], **Seung-Cheol Lee**[2], **Satadeep Bhattacharjee**[2], **Niloy Ganguly**[1]
[1] Indian Institute of Technology, Kharagpur, India
[2] Indo Korea Science and Technology Center, Bangalore, India
* Equal Contribution

ABSTRACT

Generating new crystal materials with desirable chemical properties has long been a challenging task. Existing diffusion models operate in feature space, requiring complex diffusion architectures to model the joint distribution of atom types, coordinates, and lattice structures. This complexity increases the number of diffusion steps, leading to higher training and sampling costs. In this work, we aim to generate novel crystal materials within a time- and resource-constrained setup, where existing models are not well-suited. To address this, we propose CrysLDM, a novel latent diffusion model for 3D crystal materials, which integrates a variational autoencoder (VAE) and a diffusion model. The VAE encoder maps 3D crystal structures into a latent space, where the diffusion model operates. Since CrysLDM leverages a smooth, lower-dimensional latent space, it simplifies the generative process and accelerates both training and inference. Through extensive experiments on benchmark datasets and tasks, we show that CrysLDM generates stable and valid materials with quality comparable to state-of-the-art methods, while being significantly more efficient in terms of computational resources and time.

1 INTRODUCTION

Discovering novel 3D crystal materials with desired chemical properties remains a long-standing challenge in the materials design community. These materials have been essential to major advancements, including the development of batteries, solar cells, and semiconductors. Butler et al. (2018); Desiraju (2002). Unlike molecules, which are typically represented as graphs, crystal materials consist of a fundamental unit cell that repeats itself regularly on a 3D lattice Schütt et al. (2014). Hence, the discovery of novel materials is highly challenging due to the vast search space of 3D crystal structures, which encompasses the number of constituent atoms and all their possible arrangements within 3D space. Historically, efforts to generate novel materials have relied highly on Density Functional Theory (DFT) Kohn & Sham (1965), which is both resource-intensive and time-consuming simulations. Most recent advancements in equivariant diffusion models Hoogeboom et al. (2022); Bao et al. (2022), have opened up a promising trajectory for the generation of novel three-dimensional periodic structures of crystal materials.

In this work, our aim is to generate novel crystal materials within a time- and resource-constrained setup. Given limited computational power and time budget, we focus on optimizing model capacity to maximize the generation of stable and valid materials. Existing diffusion models Xie et al. (2021); Luo et al. (2023); Jiao et al. (2023); Zeni et al. (2023); Jiao et al. (2024); Yang et al. (2023) have some fundamental limitations, making them suboptimal for the problem in this time- and resource-constrained setup. Current diffusion models for generating new crystals operate in the feature space, modeling the joint distribution of atom types, fractional coordinates, and lattice structures (unit cells). However, this distribution is highly multimodal, with each component exhibiting an independent structure and distinct modality. Atomic fractional coordinates, following a wrapped normal transition distribution, are typically modeled using score matching Song et al. (2020). In contrast, atom types, as discrete data, are modeled using discrete diffusion (D3PM)Austin et al. (2021), while the lattice structure is modeled using denoising diffusion (DDPM)Ho et al. (2020). As a result, a

complex diffusion architecture is employed to model such a multimodal joint distribution, typically involving a higher number of diffusion steps to generate realistic crystal structures. Consequently, these models require extensive training and more inference or sampling steps, making them computationally intensive and time-consuming.

In this work, we utilize a latent diffusion model to address the above limitations, enabling the rapid generation of stable and valid materials, within a given time and resource budget. We propose **Crys**tal **L**atent **D**iffusion **M**odel (CrysLDM), which consists of two modules: a Variational Auto Encoder(VAE) and a Diffusion Model(DM) that operates in a smoother latent space. The Variational Auto Encoder consists of an encoder and a decoder, where the encoder maps high-dimensional atom types and lattice structures into a lower-dimensional latent or embedding space, and the decoder reconstructs the original atom types and lattice structures from this latent representation. Additionally, a diffusion model is applied to learn and model the distribution of the latent representation of the crystal (encoded by the encoder), which is lower-dimensional and exhibits a much smoother distribution. A key challenge in developing generative models for 3D crystal structures is ensuring that the learned distribution satisfies periodic E(3) invariance, meaning it remains invariant to permutation, translation, rotation, and periodic transformations. To ensure this, we employ a periodic EGNN model as the encoder and decoder functions of the VAE, as well as the backbone denoising network in the diffusion model to guide the denoising process.

Operating in the latent space, CrysLDM offers unique advantages in generative modeling complexity over existing feature-domain diffusion models, making it more efficient in terms of time and resource consumption. Firstly, mapping atom types and lattice structures into a regularized latent space eliminates the need for separate diffusion processes to capture their distinct distributions. This allows the diffusion model to learn a much smoother distribution, simplifying the overall diffusion process and accelerating training. Moreover, mapping high dimensional atom type vectors into low dimensional space enables CrysLDM to conduct training and sampling with a lower dimensionality, which can also benefit the generative modeling complexity in terms of both time and resource consumption.

To the best of our knowledge, we are the first to leverage latent diffusion models for generating 3D crystal materials. We compare our proposed CrysLDM against two widely used state-of-the-art diffusion models for crystal material generation, using popular benchmark datasets. We find that CrysLDM generates novel materials with comparable validity and stability to other methods while being substantially faster during both training and inference. In specific, at sampling time, CrysLDM is **32x** and **11x** faster than CDVAE and DiffCSP, respectively, on the Perov-5 dataset, and **45x** and **6x** faster on the MP-20 dataset. These results make CrysLDM particularly well-suited for generating stable materials in time- and resource-constrained settings. Furthermore, we observe that the structures generated by CrysLDM are, on average, more stable than those produced by CDVAE and comparable in stability to those generated by DiffCSP. We provide the code base in the supplementary material.

2 Related Work: Crystal Material Generation

Earlier efforts in generating novel periodic materials primarily focused on atomic composition while largely ignoring 3D structures. With advancements in generative models, researchers began using VAEs and GANs to generate 3D periodic structures. However, these models either represented materials as voxel images (Court et al., 2020; Hoffmann et al., 2019; Long et al., 2021; Noh et al., 2019) or encoded structures as embedding vectors (Kim et al., 2020; Ren et al., 2020; Zhao et al., 2021), often neglecting stability and invariance to Euclidean and periodic transformations. Recent advancements in equivariant diffusion models have opened up a promising trajectory for the generation of novel three-dimensional periodic structures of crystal materials. CDVAE (Xie et al., 2021) was the first work that integrated a variational autoencoder (VAE) and powerful score-based decoder network, work directly with the atomic coordinates of the structures, and uses an equivariant graph neural network to ensure euclidean and periodic invariance. Subsequently, numerous studies (Luo et al., 2023; Jiao et al., 2023; Zeni et al., 2023; Jiao et al., 2024; Yang et al., 2023) have utilized diffusion models to learn the joint distribution of atom types, coordinates, and lattice structures, enabling the generation of stable periodic structures for novel materials.

More related works are given in Appendix A.

3 BACKGROUND

3.1 CRYSTAL STRUCTURE REPRESENTATION

The 3D structure of a crystal material can be modelled by a minimal *Unit Cell*, which gets repeated infinite times in three-dimensional space on a regular lattice to form the periodic crystal structure. Given a material with N number of atoms in its unit cell, we can describe the unit cell by two matrices: Atom Type matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N]^T \in \mathbb{R}^{N\times k}$, which denotes a set of atomic type in one hot representation (k denotes maximum number of possible atom types), and Coordinate Matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N]^T \in \mathbb{R}^{N\times 3}$ denotes fractional coordinate positions of atoms, where $\boldsymbol{x}_i \in \mathbb{R}^3$ corresponds to coordinates of i^{th} atom in the unit cell. Further, there is an additional *Lattice Matrix* $\boldsymbol{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3]^T \in \mathbb{R}^{3\times 3}$, which describes how a unit cell repeats itself in the 3D space towards $\boldsymbol{l}_1, \boldsymbol{l}_2$ and $\boldsymbol{l}_3$ direction to form the periodic 3D structure of the material. Formally, a given material can be defined as $\boldsymbol{M} = (\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$ and we can represent its infinite periodic structure as $\hat{\mathbf{X}} = \{\hat{x}_i | \hat{x}_i = x_i + \sum_{j=1}^{3} k_j l_j\}$; $\hat{\mathbf{A}} = \{\hat{a}_i | \hat{a}_i = a_i\}$ where $k_1, k_2, k_3, i \in Z, 1 \leq i \leq N$.

3.2 INVARIANCES IN CRYSTAL STRUCTURE

The basic idea of using generative models for crystal generation is to learn the underlying data distribution of material structure $f(\boldsymbol{M})$. Since crystal materials satisfy physical symmetry properties Dresselhaus et al. (2007); Zee (2016), one of the major challenges here is the learned distribution must satisfy periodic E(3) invariance i.e. invariance to permutation, translation, rotation, and periodic transformations. 1) ***Permutation Invariance :*** If we permute the indices of constituent atoms it will not change the material. 2) ***Translation Invariance :*** If we translate the atom coordinates by a random vector it will not change the structure of the material. 3) ***Rotational Invariance :*** If we rotate the atom coordinates and lattice matrix, the material remains unchanged. 4) ***Periodic Invariance :*** Finally, since the atoms in the unit cell can periodically repeat itself infinite times along the lattice vector, there can be many choices of unit cells and coordinate matrices representing the same material. (More details in Appendix B)

3.3 PROBLEM FORMULATION

In this work, we consider generative modeling of 3D crystal geometries from scratch, to discover new stable materials. Formally, given a dataset $\mathcal{M} = \{\boldsymbol{M}_i\}$, containing crystal structure $\boldsymbol{M}_i = (\boldsymbol{A}_i, \boldsymbol{X}_i, \boldsymbol{L}_i)$, the goal is to capture the underlying data distribution $f(\boldsymbol{M})$ via learning a generative model $p_\theta(\boldsymbol{M})$, where θ is a set of learnable parameters. While training, we need p_θ to ensure that the learned distribution is invariant to different symmetry transformations mentioned in Section 3.2. Once trained, the learned generative model can sample a valid and stable structure of the material, that is invariant to different symmetry transformations.

3.4 DIFFUSION MODELS FOR CRYSTAL MATERIALS

Diffusion models are popular generative models that are formulated using a T steps Markov Chain. Given a data point $\mathbf{d}_0$, the forward diffusion process gradually corrupts the data point over T steps, by adding a small amount of gaussian noise at each step:

$$q(\mathbf{d}_t|\mathbf{d}_0) = \mathcal{N}(\mathbf{d}_t|\sqrt{\bar{\alpha}_t}\mathbf{d}_0,\ (1\ -\ \bar{\alpha}_t)\mathbf{I}) \tag{1}$$

where, $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$, $\alpha_t = 1 - \beta_t$ and $\{\beta_t \in (0,1)\}_{t=1}^{T}$ controls the variance of diffusion step following certain noise scheduler. Further, the reverse denoising process, which is parameterized, begins with a gaussian noise input $\mathbf{d}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and incrementally denoises the intermediate noisy variables $\mathbf{d}_{T:1}$ to approximate the clean data $\mathbf{d}_0$ following the target data distribution:

$$p_\theta(\mathbf{d}_{t-1}|\mathbf{d}_t) = \mathcal{N}\big\{\mathbf{d}_{t-1}\ ;\ \mu_\theta(\mathbf{d}_{t-1}, t), \rho_t^2\mathbf{I}\big\} \tag{2}$$

where ρ_t is a predefined variance and mean μ_θ is typically modeled using some neural network (U-Net for images). However, leveraging diffusion models to generate new crystal materials is challenging due to the highly multi-modal nature of their joint distribution, where each component has an independent structure and a distinct modality. On one hand, atom types are discrete, while lattice parameters are continuous; additionally, atomic fractional coordinates are continuous but

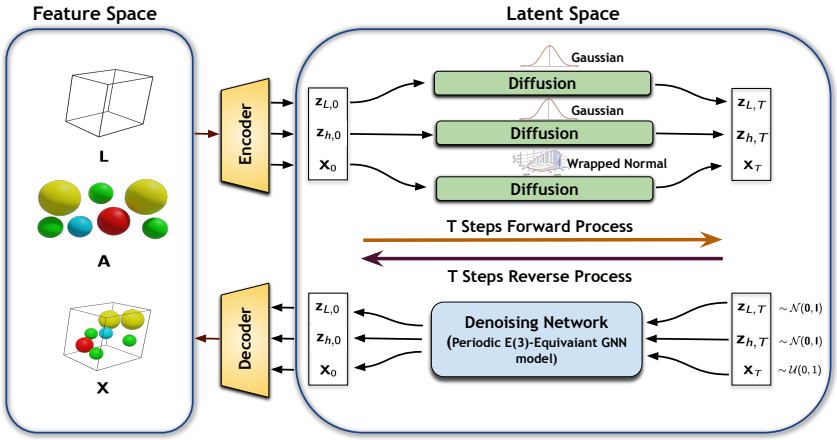

Figure 1: Model Architecture of our proposed CrysLDM.

exhibit periodicity. As a result, each variable necessitates a independent diffusion framework to accommodate its unique structure.

Diffusion on Atom Type ($\boldsymbol{A}$). Atom Type Matrix $\boldsymbol{A} \in \mathbb{R}^{N\times k}$ can be considered as N discrete variables belonging to k classes and discrete diffusion model (D3PM) (Austin et al., 2021) can be leveraged for diffusion on $\boldsymbol{A}$. In specific, with $\boldsymbol{a}$ as the one-hot representation of atom a, the transition probability for the forward process is $q(\boldsymbol{a}_t|\boldsymbol{a}_{t-1}) = Cat(\boldsymbol{a}_t; \boldsymbol{p} = \boldsymbol{a}_{t-1}\boldsymbol{Q}_t)$, where $Cat(\boldsymbol{a}; \boldsymbol{p})$ is a categorical distribution over $\boldsymbol{a}$ with probabilities $\boldsymbol{p}$ and $\boldsymbol{Q}_t$ is the Markov transition matrix at time step t, defined as $[\boldsymbol{Q}_t]_{i,j} = q(\boldsymbol{a}_t = i|\boldsymbol{a}_{t-1} = j)$. Different choices of $\boldsymbol{Q}_t$ and corresponding stationary distributions are proposed by (Austin et al., 2021) which provides flexibility to control the data corruption and denoising process.

Diffusion on Atom Coordinates ($\boldsymbol{X}$). Coordinate Matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N]^T \in \mathbb{R}^{N\times 3}$ contains fractional coordinates of constituent atoms, that resides in quotient space $\mathbb{R}^{N\times 3}/\mathbb{Z}^{N\times 3}$ induced by the crystal periodicity. Hence, it is not suitable to apply DDPM to model $\boldsymbol{X}$, since the gaussian distribution used in DDPM is unable to model the cyclical and bounded domain of $\boldsymbol{X}$. Hence at each step of forward diffusion, noise sampled from Wrapped Normal (WN) distribution (De Bortoli et al., 2022) is added to $\boldsymbol{X}$ and during denoising Score Matching Network (Song & Ermon, 2019; 2020) is leveraged to model underlying transition probability.

Diffusion on Lattice ($\boldsymbol{L}$). Lattice Matrix $\boldsymbol{L} = [l_1, l_2, l_3]^T \in \mathbb{R}^{3\times 3}$ is a global feature of the material which determines the shape and symmetry of the unit cell structure. Since $\boldsymbol{L}$ is in continuous space, we leverage the idea of the Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) for diffusion on $\boldsymbol{L}$.

During the reverse denoising process, a key challenge is ensuring that the learned distribution of material structures adheres to periodic E(3) invariance. To address this, existing works have utilized variants of periodic-E(3)-equivariant GNN models, such as GemNet (Gasteiger et al., 2021), DimeNet (Gasteiger et al., 2020), or CSPNet (Jiao et al., 2023), as backbone denoising networks to guide the denoising process. Training the denoising network involves an aggregated objective function that combines cross-entropy loss, score matching loss, and l_2 for atom types, coordinates, and lattice parameters, respectively: Due to the complexity of the diffusion architecture required to model such a multimodal joint distribution, a higher number of diffusion steps is typically needed to generate realistic crystal structures. As a result, these models demand substantial training effort and involve more inference or sampling steps, making them both computationally intensive and time-consuming.

4 METHODOLOGY

In this section, we provide a detailed overview of our proposed methodology, **Crys**tal **L**atent **D**iffusion **M**odel (CrysLDM). The framework comprises two main components: a Variational Autoencoder (VAE) and a Diffusion Model (DM) that operates within a smoother latent space. We will first describe the detailed architecture of both modules, followed by their respective training and sampling processes.

4.1 VARIATIONAL AUTOENCODER (VAE)

Our first objective is to encode 3D crystal geometry into a lower-dimensional latent space that is meaningful and preserves all the physical symmetries of the crystal structure. To achieve this, we utilize a variational autoencoder framework comprising a GNN encoder, $\mathcal{E}_\phi$, and a GNN decoder, $\mathcal{D}_\psi$. The encoder takes the crystal material $\boldsymbol{M} = (\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$ as input and encodes the atomic types and lattice structure into the latent space:

$$\boldsymbol{z}_h, \boldsymbol{z}_L = \mathcal{E}_\phi(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L}) \tag{3}$$

where $\boldsymbol{z}_h$ and $\boldsymbol{z}_L$ are latent representation of constituent nodes(atoms) and lattice structure respectively. Note, we did not encode atomic fractional coordinates $\boldsymbol{X}$ into the latent space primarily for two main reasons: first, atomic coordinates are inherently low-dimensional, and second, empirical observations showed that encoding atomic coordinates into the latent space diminishes their physical significance, adversely affecting the message-passing process of GNNs and subsequently degrading the model's performance. Further, the decoder $\mathcal{D}_\psi$ is trained to reconstruct the atomic types and lattice structure from the latent representations:

$$\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{L}} = \mathcal{D}_\psi(\boldsymbol{z}_h, \boldsymbol{X}, \boldsymbol{z}_L) \tag{4}$$

A key criterion in designing this variational autoencoder is that the learned latent space must preserve the physical symmetries of the crystal structure and satisfy periodic E(3) invariance. To achieve this, we propose incorporating equivariance into the autoencoder design by using 3D equivariant graph neural networks (EGNNs) to implement both the encoder $\mathcal{E}_\phi$ and the decoder $\mathcal{D}_\psi$. The whole autoencoder network is trained end to end using a regularized reconstruction loss:

$$\begin{aligned} \mathcal{L}_{VAE} &= \mathcal{L}^{\boldsymbol{A}}_{recon} + \mathcal{L}^{\boldsymbol{L}}_{recon} + \lambda \mathcal{L}_{reg} \\ \mathcal{L}_{reg} &= d_{KL}\{q_\phi(\boldsymbol{z}_h, \boldsymbol{z}_L | \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L}) \,||\, p(\boldsymbol{z}_h, \boldsymbol{z}_L)\} \end{aligned} \tag{5}$$

Here, $\mathcal{L}^{\boldsymbol{A}}_{recon}$ and $\mathcal{L}^{\boldsymbol{L}}_{recon}$ represent the reconstruction losses for atom types and lattice structure, respectively. By design, we use cross-entropy loss for $\boldsymbol{A}$ and l_2 loss for $\boldsymbol{L}$. $\mathcal{L}_{reg}$ denotes KL divergence (Kullback–Leibler divergence) that measures how much the learned latent distribution $q_\phi(\boldsymbol{z}_h, \boldsymbol{z}_L | \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$ differs from the prior distribution $p(\boldsymbol{z}_h, \boldsymbol{z}_L)$ (commonly a standard gaussian distribution). This regularization term constrains the variance of latent embeddings, making them more stable and suitable for learning latent diffusion models (LDMs).

4.2 LATENT DIFFUSION MODEL(LDM)

The encoder function $\mathcal{E}_\phi$ of the crystal autoencoder allows us to encode crystal materials into a smoother and lower-dimensional latent space. Building on this, we leverage a latent diffusion model to capture the distribution of crystal latent or embedding space. Given latent representation of a crystal material as $\boldsymbol{M}_0 = (\boldsymbol{z}_{h,0}, \boldsymbol{X}_0, \boldsymbol{z}_{L,0})$, we define a forward diffusion process through a Markov chain over T steps to diffuse $\boldsymbol{z}_h, \boldsymbol{X}, \boldsymbol{z}_L$ independently as follows :

$$\begin{aligned} q(\boldsymbol{z}_{h,t}, \boldsymbol{X}_t, \boldsymbol{z}_{L,t} \mid \boldsymbol{z}_{h,t-1}, \boldsymbol{X}_{t-1}, \boldsymbol{z}_{L,t-1}) = \\ q(\boldsymbol{z}_{h,t} \mid \boldsymbol{z}_{h,t-1}) q(\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1}) q(\boldsymbol{z}_{L,t} \mid \boldsymbol{z}_{L,t-1}) \;\; t = 1, 2, ...T \end{aligned} \tag{6}$$

Diffusion on $\boldsymbol{z}_h$ and $\boldsymbol{z}_L$. Since both atom types and lattice structure are projected into smooth latent space using The encoder function $\mathcal{E}_\phi$, we don't need separate sophisticated diffusion process to model them. Rather, as both $\boldsymbol{z}_h$ and $\boldsymbol{z}_L$ are in continuous space, we can use Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) to model them. Given $\boldsymbol{z}_{h,0} \sim p(\boldsymbol{z}_h)$ and $\boldsymbol{z}_{L,0} \sim p(\boldsymbol{z}_L)$,

the forward diffusion process iteratively diffuses both over T timesteps through transition probabilities $q(\boldsymbol{z}_{h,t}|\boldsymbol{z}_{h,0})$ and $q(\boldsymbol{z}_{L,t}|\boldsymbol{z}_{L,0})$ respectively. At each t^{th} step, we can derive these probabilities as follows :

$$\begin{aligned} q(\boldsymbol{z}_{h,t} \mid \boldsymbol{z}_{h,0}) &= \mathcal{N}(\boldsymbol{z}_{h,t} \mid \sqrt{\bar{\alpha}_t}\boldsymbol{z}_{h,0},\ (1\ -\ \bar{\alpha}_t)\mathbf{I}) \\ q(\boldsymbol{z}_{L,t} \mid \boldsymbol{z}_{L,0}) &= \mathcal{N}(\boldsymbol{z}_{L,t} \mid \sqrt{\bar{\alpha}_t}\boldsymbol{z}_{L,0},\ (1\ -\ \bar{\alpha}_t)\mathbf{I}) \end{aligned} \tag{7}$$

where, $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$, $\alpha_t = 1 - \beta_t$ and $\{\beta_t \in (0,1)\}_{t=1}^{T}$ controls the variance of diffusion step following certain variance scheduler. By reparameterization, we can rewrite equation 7 as:

$$\begin{aligned} \boldsymbol{z}_{h,t} &= \sqrt{\bar{\alpha}_t}\boldsymbol{z}_{h,0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{z_h} \\ \boldsymbol{z}_{L,t} &= \sqrt{\bar{\alpha}_t}\boldsymbol{z}_{L,0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{z_L} \end{aligned} \tag{8}$$

where, $\boldsymbol{\epsilon}_{z_h}, \boldsymbol{\epsilon}_{z_L} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are noises. After T such diffusion steps, noisy $\boldsymbol{z}_{h,T}, \boldsymbol{z}_{L,T}$ is generated, which follows prior noise distribution $\sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

In the reverse denoising process, given noisy $\boldsymbol{z}_{h,T}, \boldsymbol{z}_{L,T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ we reconstruct true latent representation of atom and lattice $\boldsymbol{z}_{h,0}, \boldsymbol{z}_{L,0}$ thorough iterative denoising step via learning reverse conditional distribution, which we formulate as follows :

$$\begin{aligned} p(\boldsymbol{z}_{h,t-1}|\boldsymbol{M}_t) &= \mathcal{N}\big\{\boldsymbol{z}_{h,t-1} \mid \mu^{z_h}(\boldsymbol{M}_t), \rho_t^2\mathbf{I}\big\} \\ p(\boldsymbol{z}_{L,t-1}|\boldsymbol{M}_t) &= \mathcal{N}\big\{\boldsymbol{z}_{L,t-1} \mid \mu^{z_L}(\boldsymbol{M}_t), \rho_t^2\mathbf{I}\big\} \end{aligned} \tag{9}$$

where $\mu^{z_h}(\boldsymbol{M}_t) = \frac{1}{\sqrt{\alpha_t}}\big(\boldsymbol{z}_{h,t} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\ \hat{\boldsymbol{\epsilon}}_{z_h}(\boldsymbol{M}_t, t)\big)$, $\mu^{z_L}(\boldsymbol{M}_t) = \frac{1}{\sqrt{\alpha_t}}\big(\boldsymbol{z}_{L,t} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\ \hat{\boldsymbol{\epsilon}}_{z_L}(\boldsymbol{M}_t, t)\big)$ and $\boldsymbol{M}_t = (\boldsymbol{z}_{h,t}, \boldsymbol{X}_t, \boldsymbol{z}_{L,t})$. Intuitively, $\hat{\boldsymbol{\epsilon}}_{z_h}, \hat{\boldsymbol{\epsilon}}_{z_L}$ are the denoising terms that needs to be subtracted from $\boldsymbol{z}_{h,t}$ and $\boldsymbol{z}_{L,t}$ to generate $\boldsymbol{z}_{h,t-1}$ and $\boldsymbol{z}_{L,t-1}$ respectively. We use a denoising network $f_\theta(\boldsymbol{M}_t, t)$ to model these noise terms. Following the simplified training objective proposed by Ho et al. (2020), we train the denoising network using following l_2 losses :

$$\begin{aligned} \mathcal{L}_{type} &= \mathbb{E}_{\boldsymbol{\epsilon}_{z_h}, t\sim\mathcal{U}(1,T)}\|\boldsymbol{\epsilon}_{z_h}\ -\ \hat{\boldsymbol{\epsilon}}_{z_h}\|_2^2 \\ \mathcal{L}_{lattice} &= \mathbb{E}_{\boldsymbol{\epsilon}_{z_L}, t\sim\mathcal{U}(1,T)}\|\boldsymbol{\epsilon}_{z_L}\ -\ \hat{\boldsymbol{\epsilon}}_{z_L}\|_2^2 \end{aligned} \tag{10}$$

Diffusion on $\boldsymbol{X}$. As discussed earlier in Section 4.1, atomic fractional coordinates are not being projected into the latent space. Therefore, we will utilize the conventional feature space diffusion process to model $\boldsymbol{X}$. Atomic fractional coordinates in crystal material live in quotient space $\mathbb{R}^{N\times3}/\mathbb{Z}^{N\times3}$ induced by the crystal periodicity. At each step of forward diffusion, we add noise sample from Wrapped Normal (WN) distribution De Bortoli et al. (2022) to $\boldsymbol{X}$ and during backward diffusion leverage Score Matching Diffusion Networks Song & Ermon (2019; 2020) to model underlying transition probability:

$$q(\boldsymbol{X}_t \mid \boldsymbol{X}_0) = \mathcal{N}_W(\boldsymbol{X}_t \mid \boldsymbol{X}_0, \sigma_t^2\mathbf{I}) \tag{11}$$

In specific, at each t^{th} step of diffusion, we derive $\boldsymbol{X}_t$ as : $\boldsymbol{X}_t = \Gamma_w(\boldsymbol{X}_0 + \boldsymbol{\sigma_t}\boldsymbol{\epsilon}_X)$ where, $\boldsymbol{\epsilon}_X$ is a noise, sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\sigma_t}$ is the noise scale following exponential scheduler and $\Gamma_w(.)$ is a truncation function. Given a fractional coordinate matrix X, truncation function $\Gamma_w(\boldsymbol{X}) = (\boldsymbol{X} - \lfloor\boldsymbol{X}\rfloor)$ returns the fractional part of each element of $\boldsymbol{X}$. As argued in Jiao et al. (2023), $q(X_t|X_0)$ is periodic translation equivariant, and approaches uniform distribution $\mathcal{U}(0,1)$ for sufficiently large values of σ_T. Hence during the backward denoising process, we first sample $\boldsymbol{X}_T \sim \mathcal{U}(0,1)$ and iteratively denoise via score network for T steps to recover back the true fractional coordinates $\boldsymbol{X}_0$. We use the denoising network $f_\theta(\boldsymbol{M}_t, t)$ to model the backward diffusion process, which is trained using the following score-matching objective function :

$$\mathcal{L}_{coord} = \mathbb{E}_{\substack{\boldsymbol{X}_t\sim q(\boldsymbol{X}_t|\boldsymbol{X}_0) \\ t\sim\mathcal{U}(1,T)}}\|\nabla_{\boldsymbol{X}_t}\log q(\boldsymbol{X}_t|\boldsymbol{X}_0) - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{X}}(\boldsymbol{M}_t, t)\|_2^2 \tag{12}$$

where $\nabla_{\boldsymbol{X}_t}\log q(\boldsymbol{X}_t|\boldsymbol{X}_0) \propto \sum_{\boldsymbol{K}\in\mathbb{Z}^{N\times3}} \exp(-\ \frac{\|\boldsymbol{X}_t - \boldsymbol{X}_0 + \boldsymbol{K}\|_F^2}{2\boldsymbol{\sigma}_t^2})$ is the score function of transitional distribution and $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{X}}(\boldsymbol{M}_t, t)$ denoising term.

Denoising Network. The goal of the denoising network during the reverse denoising process is to iteratively remove noise from the noisy crystal representation (sampled from gaussian noise), transforming it into a realistic crystal latent representation. A crucial requirement for this process is that

Dataset	Method	Steps	Validity(%) ↑		Property ↓		Stability ↑	Cost ↓
			Compositional	Structural	# Elements	Density	Rate (%)	Steps / Stable
Perov	CDVAE	5000	98.59	**100**	0.0264	0.1258	**2.5**	200
	DiffCSP	1000	**98.85**	**100**	**0.0263**	**0.111**	1.5	66.67
	CrysLDM	100	97.81	**100**	0.045	1.21	1.6	**6.25**
MP-20	CDVAE	5000	**86.7**	**100**	0.2778	0.6875	33.2	15.06
	DiffCSP	1000	83.25	**100**	**0.1247**	0.3502	**47.6**	2.1
	CrysLDM	100	83.02	99.87	0.5149	**0.3115**	30.3	**0.33**

Table 1: Summary of results on Ab Initio Crystal Generation task of CrysLDM and different baseline models on Perov-5 and MP-20 dataset. The best performances are highlighted in bold.

the learned distribution of material structures must adhere to periodic E(3) invariance. As established in the literature Xu et al. (2022), the learned distribution $p(\boldsymbol{M}_0)$ of the denoising model will satisfy periodic E(3) invariance, provided the prior distribution $p(\boldsymbol{M}_T)$ is invariant and the neural network parameterizing the transition probability $q(\boldsymbol{M}_{t-1}|\boldsymbol{M}_t)$ is equivariant to permutation, translation, rotation, and periodic transformations. To fulfill this requirement, we employed 3D equivariant graph neural networks (EGNNs) to implement the denoising network. In practice, we extend the CSPNet architecture Jiao et al. (2023), which was originally designed for the crystal structure prediction (CSP) task. CSPNet is based on the EGNN and satisfies the periodic E(3) invariance condition for periodic crystal structures(Details in Appendix C). The denoising network is trained using a combined loss: $\mathcal{L}_{LDM} = \lambda_L \mathcal{L}_{lattice} + \lambda_A \mathcal{L}_{type} + \lambda_X \mathcal{L}_{coord}$ where, the hyperparameters λ_L, λ_A, and λ_X control the relative weighting of these loss components.

Training and Sampling. Next, we outline the training and sampling process of CrysLDM. Following previous works on latent diffusion models in other domains (Sinha et al., 2021; Rombach et al., 2022; Xu et al., 2023), we have adopted a two-stage training strategy. We first train the VAE with a regularized reconstruction loss (5), followed by training the Latent Diffusion Model using the combined diffusion loss $\mathcal{L}_{LDM}$.
During sampling, our setup first requires determining the number of constituent atoms (N) in the material. Following common practice in the literature (Jiao et al., 2023), we begin by estimating the distribution of atom counts, $p(N)$, across different materials in the training set. We then sample $N \sim p(N)$ and generate atom latent features and coordinates of size N. Next, we generate atomic fractional coordinates along with latent embeddings of atoms and lattices using the latent diffusion models. These embeddings are then fed into the decoder $\mathcal{D}_\psi$ of the VAE to generate 3D structure of a realistic material. The training and sampling algorithms are given in Appendix D.

Advantage of CrysLDM. By design, operating a diffusion model in the latent space offers several inherent advantages. First, since we utilize a variational autoencoder (VAE) trained with a regularized reconstruction loss, the latent space becomes more compact and smooth, enhancing the training efficiency of the diffusion model. Second, in feature space, atom types and lattice structures belong to different modalities, requiring separate diffusion processes to accommodate their distinct distributions. However, in CrysLDM, both are encoded into a unified smooth latent space, simplifying the overall diffusion process and accelerating training. Finally, mapping high-dimensional atom-type vectors into a lower-dimensional space allows CrysLDM to perform training and sampling with reduced dimensionality. This not only improves generative modeling efficiency but also reduces computational costs in terms of both time and resource consumption.

5 EXPERIMENTS

Evaluation Metric. Following prior works, we evaluate the performance of CrysLDM and baseline models in generating novel material structures using a diverse set of metrics, categorized under **Validity**, **Property Statistics**, and **Stability** measure. Under **Validity**, in line with previous studies Court et al. (2020); Xie et al. (2021), we assess both structural and compositional validity. Structural validity represents the percentage of generated crystals with valid periodic structures, while compositional validity refers to the percentage of structures with correct atom types. A structure is

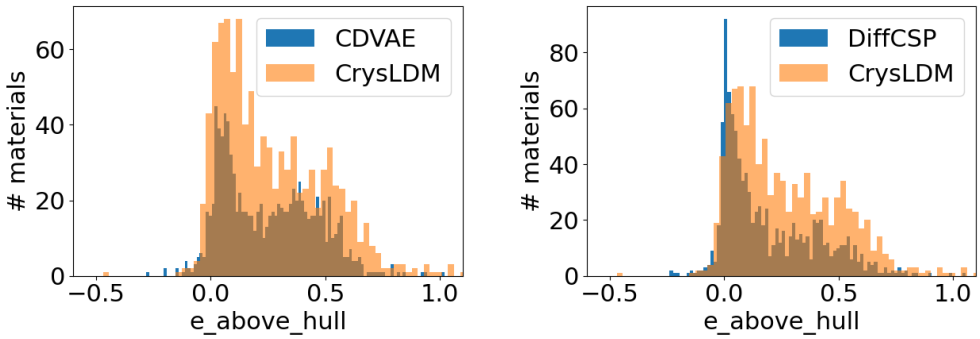

Figure 2: Histogram of E^{hull} distribution for relaxed structures generated by different models.

considered valid if the shortest distance between any pair of atoms exceeds 0.5 Å, and its composition is deemed valid if the overall charge remains neutral, as determined by SMACT Davies et al. (2019). Additionally, we evaluate the similarity between the generated materials and those in the test set using various **Property Statistics**, where we compute the earth mover's distance (EMD) between the distributions in element number (# Elem) and density (ρ, unit g/cm3). Finally, we evaluate **Stability** of our generated materials, which is based on energy above hull (E^{hull}) calculations. To evaluate the stability, we first sample 1000 materials and use M3GNet () to relax the structures of the generated materials and approximate force, energy, and stress within the crystal unit. We then classify final relaxed structures with a predicted energy above hull of less than 0.1 eV/atom as stable materials. We report **Stability Rate** as % of stable materials out of 1000 samples.

Results. We present the material generation results for both datasets in Table (1). We observe that across all metrics, CrysLDM demonstrates competitive performance compared to baseline models, highlighting the strong generative capabilities of our proposed latent diffusion framework. Furthermore, given the time- and resource-constrained setup, our goal is to minimize the overall computational cost of material generation. To assess this, we report the **Stability Cost**, defined as the number of integration steps per stability rate (# Integration Steps /# Stable Materials), which quantifies the average number of sampling steps required to generate a stable material. We observe that, compared to baseline models, CrysLDM requires fewer sampling steps, resulting in a significantly lower sampling cost. Specifically, at inference time, CrysLDM is **32x** and **11x** more faster than CDVAE and DiffCSP, respectively, on the Perov-5 dataset, and **45x** and **6x** more faster on the MP-20 dataset. Overall, these results highlight the effectiveness of CrysLDM in time- and resource-constrained scenarios. Like other baseline diffusion models, CrysLDM is capable of generating stable and valid materials. However, by significantly reducing the sampling time per stable material, it enables users to generate more stable materials within a limited time budget compared to other models.
To further investigate how well CrysLDM generates low-energy structures compared to baseline models, we conducted an experiment using the MP-20 dataset. We plotted the histogram of the computed E^{hull} distribution for relaxed structures across different methods in Figure 2. Our observations indicate that, on average, CrysLDM produces a higher proportion of low-energy structures than CDVAE while being competitive with DiffCSP. Thus, the structures generated by CrysLDM are, on average, more stable than those generated by CDVAE and comparable in stability to those generated by DiffCSP.

6 Conclusion

In this work, we focus on generating stable crystal materials within a time- and resource-constrained setup by exploring latent diffusion models. We introduce CrysLDM, a novel latent diffusion model that operates in a lower-dimensional, smooth latent space, making it more efficient in terms of time and resource consumption. Extensive experiments on benchmark generative tasks using two popular datasets demonstrate that CrysLDM produces novel materials with comparable validity and stability to existing methods while being significantly faster during both training and inference. Additionally, CrysLDM generates a higher proportion of lower-energy structures compared to baseline models.

REFERENCES

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.

Fan Bao, Min Zhao, Zhongkai Hao, Peiyao Li, Chongxuan Li, and Jun Zhu. Equivariant energy-guided sde for inverse molecular design. *arXiv preprint arXiv:2209.15408*, 2022.

Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.*, 31(9):3564–3572, 2019.

Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):1–8, 2021.

Callum J Court, Batuhan Yildirim, Apoorv Jain, and Jacqueline M Cole. 3-d inorganic crystal structure generation and property prediction via representation learning. *Journal of Chemical Information and Modeling*, 60(10):4518–4535, 2020.

Kishalay Das, Bidisha Samanta, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Crysxpp: An explainable property predictor for crystalline materials. *npj Computational Materials*, 8(1):43, 2022.

Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Crys-mmnet: multimodal representation for crystal property prediction. In *Uncertainty in Artificial Intelligence*, pp. 507–517. PMLR, 2023a.

Kishalay Das, Bidisha Samanta, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Crysgnn: Distilling pre-trained knowledge to enhance property prediction for crystalline materials. *arXiv preprint arXiv:2301.05852*, 2023b.

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.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. *Advances in Neural Information Processing Systems*, 35:2406–2422, 2022.

Gautam R Desiraju. Cryptic crystallography. *Nature materials*, 1(2):77–79, 2002.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Mildred S Dresselhaus, Gene Dresselhaus, and Ado Jorio. *Group theory: application to the physics of condensed matter*. Springer Science & Business Media, 2007.

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

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34:6790–6802, 2021.

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

Jordan Hoffmann, Louis Maestrati, Yoshihide Sawada, Jian Tang, Jean Michel Sellier, and Yoshua Bengio. Data-driven approach to encoding and decoding 3-d crystal structures. *arXiv preprint arXiv:1909.00949*, 2019.

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.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. *arXiv preprint arXiv:2309.04475*, 2023.

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

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.

Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.

Meng Liu, Keqiang Yan, Bora Oztekin, and Shuiwang Ji. Graphebm: Molecular graph generation with energy-based models. *arXiv preprint arXiv:2102.00546*, 2021.

Teng Long, Nuno M Fortunato, Ingo Opahle, Yixuan Zhang, Ilias Samathrakis, Chen Shen, Oliver Gutfleisch, and Hongbin Zhang. Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Computational Materials*, 7(1):66, 2021.

Steph-Yves Louis, Yong Zhao, Alireza Nasiri, Xiran Wang, Yuqi Song, Fei Liu, and Jianjun Hu. Graph convolutional neural networks with global attention for improved materials property prediction. *Physical Chemistry Chemical Physics*, 22(32):18141–18148, 2020.

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems*, 35:9754–9767, 2022.

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.

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.

Cheol Woo Park and Chris Wolverton. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Physical Review Materials*, 4(6), Jun 2020. ISSN 2475-9953. doi: 10.1103/physrevmaterials.4.063801. URL http://dx.doi.org/10.1103/PhysRevMaterials.4.063801.

Zekun Ren, Juhwan Noh, Siyu Tian, Felipe Oviedo, Guangzong Xing, Qiaohao Liang, Armin Aberle, Yi Liu, Qianxiao Li, Senthilnath Jayavelu, et al. Inverse design of crystals using generalized invertible crystallographic representation. *arXiv preprint arXiv:2005.07609*, 3(6):7, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Jonathan Schmidt, Love Pettersson, Claudio Verdozzi, Silvana Botti, and Miguel AL Marques. Crystal graph attention networks for the prediction of stable materials. *Science Advances*, 7(49): eabi7948, 2021.

Kristof T Schütt, Henning Glawe, Felix Brockherde, Antonio Sanna, Klaus-Robert Müller, and Eberhard KU Gross. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Physical Review B*, 89(20):205118, 2014.

Chence Shi, Shitong Luo, Minkai Xu, and Jian Tang. Learning gradient fields for molecular conformation generation. In *International conference on machine learning*, pp. 9558–9568. PMLR, 2021.

Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. *Advances in Neural Information Processing Systems*, 34: 12533–12548, 2021.

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.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in neural information processing systems*, 34:11287–11302, 2021.

Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022.

Jiaxiang Wu, Tao Shen, Haidong Lan, Yatao Bian, and Junzhou Huang. Se (3)-equivariant energy-based models for end-to-end protein folding. *bioRxiv*, pp. 2021–06, 2021.

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

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

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.

Minkai Xu, Alexander S Powers, Ron O Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3d molecule generation. In *International Conference on Machine Learning*, pp. 38592–38610. PMLR, 2023.

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*, 2022.

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.

Anthony Zee. *Group theory in a nutshell for physicists*, volume 17. Princeton University Press, 2016.

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.

Yong Zhao, Mohammed Al-Fahdi, Ming Hu, Edirisuriya MD Siriwardane, Yuqi Song, Alireza Nasiri, and Jianjun Hu. High-throughput discovery of novel cubic crystal materials using deep generative neural networks. *Advanced Science*, 8(20):2100566, 2021.

APPENDIX

A MORE RELATED WORK

A.1 CRYSTAL REPRESENTATION LEARNING

In recent times, graph neural network (GNN) based approaches have emerged as a powerful model in learning robust representation of crystal materials, which enhance fast and accurate property prediction. CGCNN Xie & Grossman (2018) is the first proposed model, which represents a 3D crystal structure as an undirected weighted multi-edge graph and builds a graph convolution neural network directly on the graph. Following CGCNN, there are a lot of subsequent studies Chen et al. (2019); Choudhary & DeCost (2021); Das et al. (2023a); Louis et al. (2020); Park & Wolverton (2020); Schmidt et al. (2021), where authors proposed different variants of GNN architectures for effective crystal representation learning. Recently, graph transformer-based architecture Matformer Yan et al. (2022) is proposed to learn the periodic graph representation of the material, which marginally improves the performance, however, is much faster than the prior SOTA model. Moreover, scarcity of labeled data makes these models difficult to train for all the properties, and recently, some key studies Das et al. (2022; 2023b) have shown promising results to mitigate this issue using transfer learning, pre-training, and knowledge distillation respectively.

A.2 DIFFUSION MODELS

The fundamental idea of the diffusion model, as initially proposed by (Sohl-Dickstein et al., 2015), is to gradually corrupt data with diffusion noise and learn a neural model to recover back data from noise. Idea of diffusion further developed in two broad categories - 1) *Score Matching Network* (Song & Ermon, 2019; 2020) and 2) *Denoising Diffusion Probabilistic Models (DDPM)* (Ho et al., 2020). In recent times diffusion models have emerged as a powerful new family of deep generative models, achieving remarkable performance records across numerous applications such as image synthesis (Dhariwal & Nichol, 2021), molecular conformer generation (Shi et al., 2021; Xu et al., 2022), molecular graph generation (Liu et al., 2021), protein folding (Wu et al., 2021; Luo et al., 2022) etc. Recently, several studies have successfully developed latent diffusion models (LDMs) with promising results across various applications, including image generation (Vahdat et al., 2021), point clouds (Vahdat et al., 2022), and text generation (Li et al., 2022). One of the most remarkable successes among them is the Stable Diffusion (Rombach et al., 2022) models, which demonstrate surprisingly realistic text-guided image generation results.

A.3 CRYSTAL MATERIAL GENERATION

In the past, there were limited efforts in creating novel periodic materials, with researchers concentrating on generating the atomic composition of periodic materials while largely neglecting the 3D structure. With the advancement of generative models, the majority of the research focuses on using popular generative models like VAEs or GANs to generate 3D periodic structures of materials, however, they either represent materials as three-dimensional voxel images (Court et al., 2020; Hoffmann et al., 2019; Long et al., 2021; Noh et al., 2019) and generate images to depict material structures (atom types, coordinates, and lattices), or they directly encode material structures as embedding vectors (Kim et al., 2020; Ren et al., 2020; Zhao et al., 2021). However, these models neither incorporate stability in the generated structure nor are invariant to any Euclidean and periodic transformations. Recent advancements in equivariant diffusion models have opened up a promising trajectory for the generation of novel three-dimensional periodic structures of crystal materials. CDVAE (Xie et al., 2021) was the first work that integrated a variational autoencoder (VAE) and powerful score-based decoder network, work directly with the atomic coordinates of the structures and uses an equivariant graph neural network to ensure euclidean and periodic invariance. Subsequently, numerous studies (Luo et al., 2023; Jiao et al., 2023; Zeni et al., 2023; Jiao et al.,

2024; Yang et al., 2023) have utilized diffusion models to learn the joint distribution of atom types, coordinates, and lattice structures, enabling the generation of stable periodic structures for novel materials.

B INVARIANCES IN CRYSTAL STRUCTURE

The basic idea of using generative models for crystal generation is to learn the underlying data distribution of material structure $p(\boldsymbol{M})$. Since crystal materials satisfy physical symmetry properties Dresselhaus et al. (2007); Zee (2016), one of the major challenges here is the learned distribution must satisfy periodic E(3) invariance i.e. invariance to permutation, translation, rotation, and periodic transformations.

- ***Permutation Invariance :*** If we permute the indices of constituent atoms it will not change the material. Formally, given any material $\boldsymbol{M} = (\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$, using any permutation matrix $\mathbf{P}$ if we permute $\boldsymbol{A}$ and $\boldsymbol{X}$ as $\mathbf{P}(\boldsymbol{A})$ and $\mathbf{P}(\boldsymbol{X})$, then new material $\boldsymbol{M_P} = (\mathbf{P}(\boldsymbol{A}), \mathbf{P}(\boldsymbol{X}), \boldsymbol{L})$ will remains unchanged. Hence the underlying distribution is also the same i.e $p(\boldsymbol{M}) = p(\boldsymbol{M_P})$.
- ***Translation Invariance :*** If we translate the atom coordinates by a random vector it will not change the structure of the material. Formally, given any material $\boldsymbol{M} = (\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$, if we translate $\boldsymbol{X}$ by an arbitrary translation vector $\mathbf{u} \in \mathbb{R}^3$, new generated material $\boldsymbol{M_P} = (\boldsymbol{A}, \boldsymbol{X} + \mathbf{u}\mathbf{1}^T, \boldsymbol{L})$ will be the same as M. Hence $p(\boldsymbol{M}) = p(\boldsymbol{M_T})$ must satisfy.
- ***Rotational Invariance :*** If we rotate the atom coordinates and lattice matrix, the material remains unchanged. Formally, using any orthogonal rotational matrix $\mathbf{Q} \in R^{3\times3}$ (satisfying $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$), if we rotate $\boldsymbol{X}$ and $\boldsymbol{L}$ of any material $\boldsymbol{M}$ and generate new $\boldsymbol{M_R} = (\boldsymbol{A}, \boldsymbol{QX}, \boldsymbol{QL})$, then actually different representations of the same material. Hence $p(\boldsymbol{M}) = p(\boldsymbol{M_R})$ must satisfy.
- ***Periodic Invariance :*** Finally, since the atoms in the unit cell can periodically repeat itself infinite times along the lattice vector, there can be many choices of unit cells and coordinate matrices representing the same material. Formally, given coordinates $\boldsymbol{X}$, after applying periodic transformation using random matrix $\boldsymbol{K} \in R^{n\times3}$, new coordinates $\mathbf{X'} = \boldsymbol{X} + \boldsymbol{KL}$ are periodically equivalent. Hence $M = (\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$ and $\mathbf{M'} = (\boldsymbol{A}, \mathbf{X}', \boldsymbol{L})$ are same material and $p(\boldsymbol{M}) = p(\mathbf{M'})$ must hold.

C DENOISING NETWORK ARCHITECTURE: CSPNET

For the backbone network in the backward diffusion process of CrysLDM, we extend the CSPNet architecture Jiao et al. (2023). CSPNet is based on EGNN Satorras et al. (2021), ensuring periodic E(3) invariance for periodic crystal structures. At the k^{th} layer of message passing, the Equivariant Graph Convolutional Layer (EGCL) takes as input the atom embeddings $\boldsymbol{h}^k = [\boldsymbol{h}_1^k, \boldsymbol{h}_2^k, ..., \boldsymbol{h}_N^k]$, atom coordinates $\boldsymbol{x}^k = [\boldsymbol{x}_1^k, \boldsymbol{x}_2^k, ..., \boldsymbol{x}_N^k]$, and the lattice matrix $\boldsymbol{L}$, and outputs a transformed set of atom embeddings $\boldsymbol{h}^{k+1}$. Formally, the message passing operation at the k^{th} layer is defined as follows:

$$\boldsymbol{m}_{i,j} = \rho_m\{\boldsymbol{h}_i^k,\ \boldsymbol{h}_j^k,\ \boldsymbol{L}^T\boldsymbol{L},\ \psi_{FT}(\boldsymbol{x}_i^k - \boldsymbol{x}_j^k)\}; \tag{13}$$

$$\boldsymbol{h}_i^{k+1} = \boldsymbol{h}_i^k + \rho_h\{\boldsymbol{h}_i^k, \boldsymbol{m}_i\} \tag{14}$$

Where $\boldsymbol{m}_i = \sum_{j=1}^{N} \boldsymbol{m}_{i,j}$, ρ_m and ρ_h are multi-layer perceptrons, and ψ_{FT} is a Fourier Transformation function applied to the relative difference between fractional coordinates $\boldsymbol{x}_i^k$ and $\boldsymbol{x}_j^k$. The Fourier Transformation is utilized as it remains invariant to periodic translation and captures various frequency components of relative fractional distances, which are essential for accurate crystal structure modeling. Input atom features $\boldsymbol{h}^0$ and coordinates $\boldsymbol{x}^0$ are fed through $\mathcal{K}$ layers of EGCL to produce $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{z}_L}$, $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{z}_h}$ and $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{X}}$ as follows :

$$\begin{aligned} \hat{\boldsymbol{\epsilon}}_{\boldsymbol{z}_L} &= \boldsymbol{L}\rho_L\Big(\frac{1}{N}\sum_{N}^{i=1}\boldsymbol{h}^{\mathcal{K}}\Big); \\ \hat{\boldsymbol{\epsilon}}_{\boldsymbol{z}_h} &= \rho_A(\boldsymbol{h}^{\mathcal{K}}); \\ \hat{\boldsymbol{\epsilon}}_{\boldsymbol{X}} &= \rho_X(\boldsymbol{h}^{\mathcal{K}}) \end{aligned} \tag{15}$$

where ρ_L, ρ_A, ρ_X are multi-layer perceptrons on the final layer embeddings.

D TRAINING AND SAMPLING ALGORITHM FOR CRYSLDM

Algorithm 1 Training Algorithm of CrysLDM

Input: Crystal Material $\boldsymbol{M} = (\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$, Encoder $\mathcal{E}_\psi$, Decoder $\mathcal{D}_\phi$, and Denosing Network f_θ.
Stage-1: Training VAE
repeat
  $\boldsymbol{\mu_h}, \boldsymbol{\mu_L} \leftarrow \mathcal{E}_\phi(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$
  Sample $\boldsymbol{\epsilon^h}, \boldsymbol{\epsilon^L} \sim N(\boldsymbol{0}, \boldsymbol{I})$
  $\boldsymbol{z}_h \leftarrow \boldsymbol{\mu_h} + \boldsymbol{\epsilon^h} \odot \sigma_{\mathbf{h}}$
  $\boldsymbol{z}_L \leftarrow \boldsymbol{\mu_L} + \boldsymbol{\epsilon^L} \odot \sigma_{\mathbf{L}}$
  $\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{L}} \leftarrow \mathcal{D}_\psi(\boldsymbol{z}_h, \boldsymbol{X}, \boldsymbol{z}_L)$
  $\mathcal{L}^{\boldsymbol{A}}_{recon} = CrossEntropyLoss(\tilde{\boldsymbol{A}}, \boldsymbol{A})$
  $\mathcal{L}^{\boldsymbol{L}}_{recon} = \|\tilde{\boldsymbol{L}} - \boldsymbol{L}\|_2^2$
  Minimize $\mathcal{L}_{VAE} = \mathcal{L}^{\boldsymbol{A}}_{recon} + \mathcal{L}^{\boldsymbol{L}}_{recon} + \lambda \mathcal{L}_{reg}$ and update parameters of $\mathcal{E}_\psi$ and $\mathcal{D}_\phi$
until Converged
Stage-2: Training LDM
repeat
  Sample $t \sim \mathcal{U}(\mathbf{0}, \mathbf{T})$
  Sample Noise $\epsilon_{\mathbf{X}}, \boldsymbol{\epsilon}_{z_h}, \boldsymbol{\epsilon}_{z_L} \sim N(0, I)$
  $\boldsymbol{z}_{h,t} = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_{h,0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{z_h}$
  $\boldsymbol{z}_{L,t} = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_{L,0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{z_L}$
  $\boldsymbol{X}_t = f_w(\boldsymbol{X}_0 + \boldsymbol{\sigma_t}\boldsymbol{\epsilon}^x)$
  $\hat{\boldsymbol{\epsilon}}_{z_h}, \hat{\boldsymbol{\epsilon}}_{\mathbf{X}}, \hat{\boldsymbol{\epsilon}}_{z_L} \leftarrow f_\theta(\boldsymbol{z}_{h,t}, \mathbf{X}_t, \boldsymbol{z}_{L,t}, t)$
  $\mathcal{L}_{type} = \|\boldsymbol{\epsilon}_{z_h} - \hat{\boldsymbol{\epsilon}}_{z_h}\|_2^2$
  $\mathcal{L}_{lattice} = \|\boldsymbol{\epsilon}_{z_L} - \hat{\boldsymbol{\epsilon}}_{z_L}\|_2^2$
  $\mathcal{L}_{coord} = \|\nabla_{\boldsymbol{X}_t} \log q(\boldsymbol{X}_t|\boldsymbol{X}_0) - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{X}}\|_2^2$
  Minimize $\mathcal{L}_{LDM} = \lambda_L \mathcal{L}_{lattice} + \lambda_A \mathcal{L}_{type} + \lambda_X \mathcal{L}_{coord}$ and update parameters of f_θ
until Converged

Algorithm 2 Sampling Algorithm of CrysLDM

$N \sim p(N)$
Sample $\boldsymbol{z}_{h,T}, \boldsymbol{z}_{L,T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \boldsymbol{X}_T \sim \mathcal{U}(0, 1)$
for $t \leftarrow T$ to 1 **do**
  $\boldsymbol{\epsilon}_{z_h}, \epsilon_{\mathbf{X}}, \boldsymbol{\epsilon}_{z_L} \sim N(0, I) / * \, Sample \, * /$
  $\hat{\boldsymbol{\epsilon}}_{z_h}, \hat{\boldsymbol{\epsilon}}_{\mathbf{X}}, \hat{\boldsymbol{\epsilon}}_{z_L} \leftarrow f_\theta(\boldsymbol{z}_{h,t}, \mathbf{X}_t, \boldsymbol{z}_{L,t}, t)$
  $\boldsymbol{z}_{h,t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{z}_{h,t} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\epsilon}}_{z_h}) + \sqrt{\beta_t \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{z_h}$
  $\boldsymbol{z}_{L,t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{z}_{L,t} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\epsilon}}_{z_L}) + \sqrt{\beta_t \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{z_L}$
  $\mathbf{X}_{t-\frac{1}{2}} \leftarrow w(\mathbf{X}_t + (\sigma_t^2 - \sigma_{t-1}^2)\hat{\epsilon}^{\mathbf{X}} + \frac{\sigma_{t-1}\sqrt{\sigma_t^2 - \sigma_{t-1}^2}}{\sigma_t}\boldsymbol{\epsilon}^{\mathbf{X}})$
  $,\hat{\boldsymbol{\epsilon}}_{\mathbf{X}} \leftarrow f_\theta(\boldsymbol{z}_{h,t}, \mathbf{X}_{t-\frac{1}{2}}, \boldsymbol{z}_{L,t}, t)$
  $\eta_t \leftarrow step\_size * \frac{\sigma_{t-1}}{\sigma_t}$
  $\mathbf{X}_{t-1} \leftarrow w(\mathbf{X}_{t-\frac{1}{2}} + \eta_t\hat{\boldsymbol{\epsilon}}^{\mathbf{X}} + \sqrt{2\eta_t}\boldsymbol{\epsilon}^{\mathbf{X}})$
end for
$\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L} = \mathcal{D}_\psi(\boldsymbol{z}_{h,0}, \boldsymbol{X}_0, \boldsymbol{z}_{L,0})$
return $\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L}$