A UNIFIED FRAMEWORK FOR JOINT CRYSTAL GENERATION AND PROPERTY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Crystal generation and property prediction are two fundamental tasks in the discovery of novel materials. Recently, AI-based frameworks have shown substantial promise for both tasks; however, they have largely evolved as separate paradigms, since each requires distinct representations and inductive biases. Predictive models typically rely on local, neighborhood-based multigraph representations, whereas generative models often operate on global three-dimensional crystal representations. In this work, we investigate whether these complementary representations can be unified to jointly support crystal generation and property prediction. To this end, we propose CrysUniNet, a unified framework that learns a shared embedding space by integrating local and global representations of crystal structures, enabling a single latent representation to support both generative and predictive tasks. We first train a VAE with a unified encoder and a multi-task decoder that maps crystals into this latent space, after which a latent diffusion model learns to generate new latent embeddings. Once trained, CrysUniNet can generate novel crystal materials via the latent diffusion model, while the VAE's property-prediction head enables property prediction for any target material. Empirically, CrysUniNet performs strongly on both tasks, generating $\mathbf{39.15}\%$ more SUN crystals than the strongest SUN baseline while also predicting the properties of existing structures more accurately.

1 INTRODUCTION

Crystal materials underpin a wide range of technologies, motivating the discovery of novel structures with properties tailored to specific applications. Materials discovery typically involves two complementary tasks: *generating novel three-dimensional periodic crystal structures* and *predicting their relevant chemical properties*. While first-principles methods such as density functional theory (DFT, Sholl & Steckel (2009)) provide physically grounded solutions, their high computational cost limits large-scale exploration. Recent advances in deep learning have enabled efficient data-driven approaches for both tasks. Generative models, including variational autoencoders (VAEs) (Xie et al., 2022), diffusion models (Jiao et al., 2023), flow-based models (Miller et al., 2024; Joshi et al., 2025), and language models (Antunes et al., 2024; Gruver et al., 2024), have demonstrated the ability to generate novel crystal structures with improved stability, uniqueness, and novelty. In parallel, graph neural networks (GNNs) (Xie & Grossman, 2018; Choudhary & DeCost, 2021; Yan et al., 2022) and transformer-based architectures (Liao et al., 2024; Yan et al., 2022; 2024; Shen et al., 2025b; Taniai et al., 2024) have achieved high accuracy in predicting material properties such as formation energy, band gap, and bulk modulus.

However, predictive and generative models for crystal materials have largely evolved as separate paradigms because the two tasks require different structural representations. Property prediction primarily benefits from local representations that capture the chemical and electronic environments of constituent atoms, and most existing approaches use radius- or k-nearest-neighbor graphs to model local coordination, including interactions across periodic boundaries. In contrast, crystal generation requires a geometry-aware representation of the complete unit cell, capturing the global arrangement of atoms, lattice geometry, and the fundamental symmetries of periodic crystals. This representational mismatch has led to task-specific latent spaces, limiting the transfer of complementary information between prediction and generation.

We posit that these two representations contain complementary information that can benefit both tasks. To bridge this gap, we propose CrysUniNet, a unified latent-diffusion framework for joint

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

crystal generation and property prediction under a multi-task learning paradigm. CrysUniNet employs a multi-encoder Variational Auto-Encoder (VAE, Kingma & Welling (2014)) that combines a local radius-graph encoder for atomic environments with a global lattice-aware encoder for the complete unit-cell geometry. These encoders map the complementary structural information into a shared latent space, which is jointly optimized through structure reconstruction and property prediction. A shared decoder with separate reconstruction and prediction heads enables the learned representation to capture both structural and property-relevant information.

To fully leverage CrysUniNet, we train a diffusion model in the learned latent space for crystal generation, while fine-tuning the VAE for downstream property prediction. The shared representation thus supports both generation and prediction, enabling complementary information from one task to benefit the other. Experiments show that this multi-task pretraining yields a structure- and property-aware latent space, improving the SUN rate by 39.15% for crystal generation and achieving the best results on most JARVIS-DFT properties.

Our contributions in this work are as follows:

1. **Common representation space.** We identify a representational mismatch between crystal property prediction, which relies on local coordination environments, and crystal generation, which requires a global, symmetry-aware view of the unit cell. We show that these local and global crystal representations carry complementary information, and joint training on reconstruction and property prediction improves both tasks, compared to single-task training.
2. **Unified framework.** To bridge the representation gap, we propose CrysUniNet[1], a unified latent-diffusion framework that jointly learns crystal generation and property prediction in a single shared latent space under a multi-task learning paradigm.
3. **Results.** With extensive experimentation, we observe better performance than existing baseline models on both the property prediction and generation tasks.

2 PRELIMINARIES

2.1 CRYSTAL REPRESENTATION

Crystal materials can be viewed as a 3D point cloud of atoms arranged in an orderly repeating pattern. For a material with N atoms in its unit cell, the structure can be defined as $\boldsymbol{M} = (\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$. ***Atom Type Matrix:*** $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N]^\top \in \mathbb{R}^{N \times k}$, where each $\boldsymbol{a}_i$ is a one-hot vector denoting the atomic type of the i-th atom, and k is the maximum number of possible atom types. ***Coordinate Matrix:*** $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times 3}$, where $\boldsymbol{x}_i \in \mathbb{R}^3$ represents the 3D coordinates of the i-th atom in the unit cell. ***Lattice Matrix:*** $\boldsymbol{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3]^\top \in \mathbb{R}^{3 \times 3}$, which specifies how the unit cell repeats itself in 3D space along directions $\boldsymbol{l}_1, \boldsymbol{l}_2$, and $\boldsymbol{l}_3$ to form the periodic crystal.

Symmetry in Crystal Structure. One of the key challenges in generative modeling of crystal representations is that crystal materials exhibit fundamental physical symmetries that any learned representation must respect. One such property is rotational invariance, which ensures that rotating the atom coordinates and lattice matrices by any orthogonal matrix $\mathbf{Q}$ results in an equivalent representation of the same material. Another key property is periodic translation invariance, meaning that translating the atom coordinates by any arbitrary vector and applying periodic wrapping does not alter the crystal structure. More details in Appendix A.

2.2 MULTI-GRAPH CONSTRUCTION FOR CRYSTALS

Although crystal materials are inherently three-dimensional structures, most prior state-of-the-art frameworks for crystal property prediction represent a crystal as a multi-graph, following CGCNN (Xie & Grossman, 2018). Building on this idea but incorporating richer geometric information, our 2D encoder in Section 3.1 adopts the SE(3)-invariant crystal graph of ComFormer (Yan et al., 2024), which yields a geometrically complete and symmetry-aware representation. We represent a crystal material $\boldsymbol{M}$ as an undirected multi-graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X}, \mathcal{F})$, where $\mathcal{V}$ is the set of nodes (atoms) in the unit cell, and $\mathcal{E} = \{(u, v, k_{uv})\}$ is a multi-set of node pairs in which k_{uv} denotes the number of edges between node pair (u, v). The node feature set $\mathcal{X} = \{\boldsymbol{c}_u \mid u \in \mathcal{V}\}$ uses the 92-dimensional CGCNN features, with each $\boldsymbol{c}_u$ encoding chemical properties. While CGCNN

[1]Source code is provided in the Supplementary Materials.

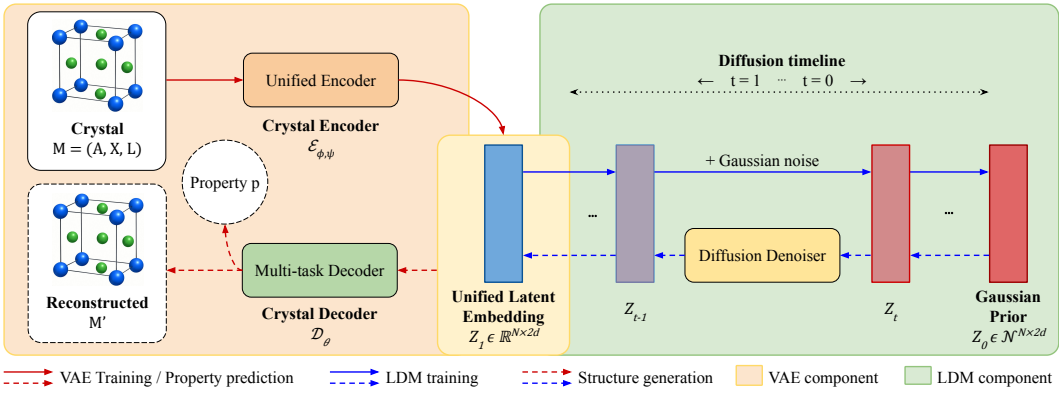

Figure 1: Overview of CrysUniNet. A unified encoder fuses 2D multi-graph and 3D geometry embeddings into a shared latent $\boldsymbol{Z} \in \mathbb{R}^{N\times 2d}$, from which a multi-task decoder reconstructs the crystal and predicts its property (red arrows). A flow-matching denoiser is then trained on the frozen latents (blue arrows); at generation time, sampled latents are decoded into new crystals by the structure head (dashed arrows).

represents edges using interatomic distances alone, which can map geometrically distinct crystals to the same representation, ComFormer augments each edge with richer geometric information. Specifically, for each node u it constructs a per-node lattice representation $\{\boldsymbol{e}_{u1}, \boldsymbol{e}_{u2}, \boldsymbol{e}_{u3}\}$ from the three shortest periodic duplicate vectors, built so as to be invariant to the choice of periodic unit cell. Each edge (u, v) is then described by the interatomic distance $\|\boldsymbol{e}_{vu}\|_2$ together with the three angles $\{\angle_{vu,1}, \angle_{vu,2}, \angle_{vu,3}\}$ that the edge vector forms with this lattice representation, defining the edge feature set $\mathcal{F}$. This distance–angle edge featurization makes the crystal graph invariant to translations and rotations of the unit cell as well as to periodic transformations, yielding a geometrically complete representation that our 2D encoder maps into the latent space.

3 Methodology: CrysUniNet

In this work, we aim to address two fundamental tasks in crystal materials modeling within a single unified framework: *de novo crystal generation* and *crystal property prediction*. For de novo generation, the goal is to learn the distribution of known valid materials and generate new crystal structures that are novel, valid, and stable. For property prediction, the model takes a crystal structure as input and estimates a target material property. We propose the **Cry**stal **Uni**fied **Net**work, CrysUniNet, a single unified framework capable of performing both crystal property prediction and de novo generation task with high precision. To support both tasks, our framework combines generative and predictive modeling within a shared latent representation: the material structure is first embedded into a rich and expressive latent space using a Variational Autoencoder (VAE), and a diffusion model then operates on this learned latent space via an N-step denoising process, as shown in Figure 1. Once trained, the generative component learns the underlying distribution of known crystal materials and samples new structures from it, while the predictive component leverages the same learned representation to estimate material properties. In the following, we elaborate on both stages in detail.

3.1 Variational Autoencoder (VAE)

The first component of our framework is a Variational Autoencoder (VAE) that encodes crystal structures into a rich latent space. Since we aim to achieve both crystal property prediction and de novo generation within a single unified framework, we want to learn a common latent space for both tasks. Crystal properties are strongly governed by the local chemical and electronic environments surrounding constituent atoms. Hence, the multi-graph representation (Section 2.2) is more effective for property prediction, as it efficiently captures the local chemical environment around each atom, including interactions across periodic boundaries. In contrast, crystal generation requires explicitly modeling the three-dimensional arrangement of atoms within the unit cell, along with the fundamental symmetries of crystal structures (Section 2.1).

To learn a common latent space that serves both tasks, we therefore fuse two complementary views: a 2D multi-graph and the 3D material structure of the unit cell. To accomplish this, our VAE consists

of two encoders, $\mathcal{E}_\phi$ and $\mathcal{E}_\psi$, which respectively encode the 2D multi-graph ($\mathcal{G}$) and the 3D geometry of a crystal. Formally, the VAE takes the crystal material $\boldsymbol{M} = (\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$ as input, processing it as both a 2D multi-graph and a 3D geometry using $\mathcal{E}_\phi$ and $\mathcal{E}_\psi$ respectively, and encodes it into two latent representations:

$$\boldsymbol{Z}_{\text{2D}} = \mathcal{E}_\phi(\mathcal{G}), \quad \boldsymbol{Z}_{\text{3D}} = \mathcal{E}_\psi(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L}) \tag{1}$$

where $\boldsymbol{Z}_{\text{2D}} = \{\boldsymbol{z}_h^{\text{2D}}\} \in \mathbb{R}^{N\times d}$ and $\boldsymbol{Z}_{\text{3D}} = \{\boldsymbol{z}_h^{\text{3D}}\} \in \mathbb{R}^{N\times d}$ are the respective sets of per-atom node embeddings, and N is the number of atoms in the crystal material. A key criterion in designing these encoders is that the learned latent space must preserve the physical symmetries of the crystal structure. Hence, for the 2D encoder we employ the ComFormer network (Yan et al., 2024), which operates on the crystal multi-graph $\mathcal{G}$ and guarantees that the learned representations are invariant to translations, rotations, and periodic transformations. For the 3D encoder, we adopt the standard Transformer (Vaswani et al., 2017) and learn the relevant symmetries via data augmentation, following Joshi et al. (2025). Further these two modality-specific embeddings are fused into a single shared latent representation by concatenating them: $\boldsymbol{Z} = [\,\boldsymbol{Z}_{\text{2D}} \,\|\, \boldsymbol{Z}_{\text{3D}}\,] \in \mathbb{R}^{N\times 2d}$, which jointly encodes the topological and geometric views of the crystal. Next, a decoder $\mathcal{D}_\theta$ is trained on $\boldsymbol{Z}$ in a multi-task learning setting, jointly reconstructing the 3D crystal structure and predicting a property of the crystal. It consists of two heads: a structure reconstruction head $\mathcal{D}_\theta^{struct}$ and a property prediction head $\mathcal{D}_\theta^{prop}$. The structure reconstruction head is trained as follows:

$$(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{L}}) = \mathcal{D}_\theta^{struct}(\boldsymbol{Z}), \tag{2}$$

where $\tilde{\boldsymbol{A}}$, $\tilde{\boldsymbol{X}}$, and $\tilde{\boldsymbol{L}}$ denote the reconstructed atomic types, coordinates, and lattice, respectively. For the property prediction head, we first aggregate the atom embeddings into a crystal-level representation via a symmetric readout that is invariant to atom ordering, and then predict the target property $\hat{y}$ as follows:

$$\boldsymbol{g} = \text{READOUT}(\boldsymbol{Z}), \quad \hat{y} = \mathcal{D}_\theta^{prop}(\boldsymbol{g}) \tag{3}$$

In this work, we choose formation energy as the target property, primarily because (1) popular databases such as GNoME (Merchant et al., 2023) provide substantially more training data for this property, and (2) a material's stability is correlated with its formation energy (Kirklin et al., 2015). As a result, the reconstruction and property-prediction heads share the same latent space and the encoder is driven to produce representations that are simultaneously informative for structure reconstruction and formation-energy prediction, enabling mutual transfer between the two tasks. The whole network is trained end-to-end using the following regularized objective function:

$$\mathcal{L}_{\text{VAE}} = \underbrace{\mathcal{L}_{\text{recon}}^{\boldsymbol{A}} + \mathcal{L}_{\text{recon}}^{\boldsymbol{X}} + \mathcal{L}_{\text{recon}}^{\boldsymbol{L}}}_{\text{Structure Reconstruction}} + \underbrace{\mathcal{L}_{\text{prop}}}_{\text{Property Prediction}} + \underbrace{\alpha\mathcal{L}_{\text{reg}}}_{\text{Regularizer}}, \tag{4}$$

Here, $\mathcal{L}_{\text{recon}}^{\boldsymbol{A}}$, $\mathcal{L}_{\text{recon}}^{\boldsymbol{X}}$, and $\mathcal{L}_{\text{recon}}^{\boldsymbol{L}}$ denote the reconstruction losses for the atom types, coordinates, and lattice, respectively. By design, we use a cross-entropy loss for $\boldsymbol{A}$ and an l_2 loss for $\boldsymbol{X}$ and $\boldsymbol{L}$. The property term $\mathcal{L}_{\text{prop}} = \|\hat{y} - y\|^2$ is the l_2 loss between the predicted and ground-truth formation energy Finally, $\mathcal{L}_{\text{reg}} = D_{\text{KL}}(q_{\phi,\psi}(\boldsymbol{Z} \mid \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L}) \,\|\, p(\boldsymbol{Z}))$ denotes the KL divergence, which measures how much the learned latent distribution $q_{\phi,\psi}(\boldsymbol{Z} \mid \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$ differs from the prior $p(\boldsymbol{Z})$ (commonly a standard Gaussian). This regularization term is weighed by α, and constrains the variance of the latent embeddings, making them more stable and better suited for learning the subsequent latent diffusion model.

3.2 Generative Extension via Latent Diffusion

Next, we aim to model the distribution of these latent representations $\boldsymbol{Z}$ to enable generation. Although the VAE itself admits generation by sampling from its Gaussian prior, the aggregated posterior over $\boldsymbol{Z}$ typically deviates from this prior, a well-known prior–posterior mismatch that leads to poor sample quality (Rosca et al., 2018; Dai & Wipf, 2019). Recent literature therefore adopts latent diffusion models (LDMs) for sampling rather than the VAE prior (Rombach et al., 2022), and we follow this paradigm by training a diffusion model over the latent representation $\boldsymbol{Z}$ to directly capture its true distribution. Following Joshi et al. (2025), we adopt a flow matching formulation (Lipman et al., 2023) (equivalently, Gaussian denoising diffusion (Gao et al., 2024)), in which the learned latent embeddings $\boldsymbol{Z}$ are gradually noised over T steps toward a base distribution, and a denoiser is trained to reverse this process and recover the original latent representations.

We formulate our approach using linear interpolation between a standard Gaussian base distribution and the target distribution defined by the VAE encoder's latent representations of the crystal

structure. Specifically, during training, given a material $\boldsymbol{M} = (\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$, we first encode it into a lower-dimensional fused latent representation $\boldsymbol{Z} = [\, \boldsymbol{Z}_{\text{2D}} \,\|\, \boldsymbol{Z}_{\text{3D}} \,]$ using the VAE encoders $\mathcal{E}_\phi$ and $\mathcal{E}_\psi$. For convenience, we denote $\boldsymbol{Z}$ as $\boldsymbol{Z}^{(1)}$, representing a clean training sample at time $t = 1$. Next, we sample a noisy latent variable $\boldsymbol{Z}^{(0)}$ at time $t = 0$ from a $2d$-dimensional standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})^{N \times 2d}$, followed by zero-centering through subtraction of its per-channel mean. Finally, we construct an interpolated noisy sample $\boldsymbol{Z}^{(t)}$ at a randomly sampled time step $t \sim \mathcal{U}(0, 1)$ using linear interpolation, $\boldsymbol{Z}^{(t)} = (1 - t)\, \boldsymbol{Z}^{(0)} + t\, \boldsymbol{Z}^{(1)}$. Therefore, along the trajectory from the noisy latent at time t to the clean latent, we define the ground-truth conditional vector field as: $u_t(\boldsymbol{Z}^{(t)} \mid \boldsymbol{Z}^{(1)}) = \frac{\boldsymbol{Z}^{(1)} - \boldsymbol{Z}^{(t)}}{1-t}$. By integrating this vector field over time, latent samples drawn from the Gaussian base distribution are transformed into true latent representations from the target distribution. We train a denoising network $\mathcal{F}_\Theta$ to approximate the conditional vector field $u_t(\boldsymbol{Z}^{(t)} \mid \boldsymbol{Z}^{(1)})$. The network takes as input the intermediate noisy latent $\boldsymbol{Z}^{(t)}$ together with the time step t, and predicts the corresponding clean latent representation, $\bar{\boldsymbol{Z}}^{(1)} = \mathcal{F}_\Theta(\boldsymbol{Z}^{(t)}, t)$. The denoiser is optimized by minimizing the mean squared error (MSE) between the predicted and ground-truth conditional vector fields:

$$\mathcal{L}_{\text{LDM}} = \frac{1}{N} \sum_{i=1}^{N} \left\| \frac{\boldsymbol{z}_i^{(1)} - \boldsymbol{z}_i^{(t)}}{1-t} - \frac{\bar{\boldsymbol{z}}_i^{(1)} - \boldsymbol{z}_i^{(t)}}{1-t} \right\|^2 = \frac{1}{(1-t)^2} \frac{1}{N} \sum_{i=1}^{N} \left\| \boldsymbol{z}_i^{(1)} - \bar{\boldsymbol{z}}_i^{(1)} \right\|^2. \tag{5}$$

For the denoising network $\mathcal{F}_\Theta$, we adopt the Diffusion Transformer (DiT) architecture (Peebles & Xie, 2023). During this latent diffusion stage, the pretrained VAE encoders are kept *frozen*, and only the parameters of $\mathcal{F}_\Theta$ are optimized using the objective in Eq. 5. Specifically, the frozen encoders map each crystal into its fused latent representation, which is subsequently perturbed with Gaussian noise according to the diffusion process. The denoising network $\mathcal{F}_\Theta$ is then trained to recover the corresponding clean latent representation from its noisy counterpart. Keeping the encoders fixed preserves the latent space learned during the first training stage and provides a stationary target distribution for the diffusion model. Consequently, $\mathcal{F}_\Theta$ learns to model the distribution of the pretrained crystal latents rather than jointly reshaping the representation space during diffusion training. This decoupling also cleanly separates representation learning from generative modeling, allowing the diffusion process to operate over a semantically structured and stable latent space. The complete two-stage training of the VAE and latent diffusion model is summarized in Appendix (Algorithm 2).

3.3 CrysUniNet Training

CrysUniNet follows a two-stage training procedure, consisting of VAE pretraining followed by latent diffusion training. In the first stage, we train the VAE under a multi-task learning objective using the loss defined in Eq. 4 on a large-scale dataset derived from GNoME (Merchant et al., 2023). Following prior work (Shen et al., 2025a), we use 380,740 crystal structures filtered from GNoME, excluding structures that overlap with the downstream datasets or lack physical or chemical significance (Detailed statistics in the Appendix, Table 3). In the second stage, we freeze the pretrained VAE encoders and train the latent diffusion model on the MP-20 dataset curated by Xie et al. (2022). The complete training procedure of CrysUniNet is summarized in the Appendix, Algorithm 2.

3.4 CrysUniNet Inference

After training, the learned components support two complementary inference tasks within the same shared latent space. For *crystal generation*, the latent diffusion model $\mathcal{F}_\Theta$ generates samples in the learned latent space, which are subsequently decoded by the structural head $\mathcal{D}_\theta^{struct}$ to produce novel crystal structures. For *property prediction*, the pretrained VAE encoders map an input crystal to latent representation $\boldsymbol{Z}$, from which the property prediction head $\mathcal{D}_\theta^{prop}$ estimates the target material property. Thus, the same learned latent space serves as a common representation for both generative and predictive modeling. Next, we describe the inference procedures for these two tasks in detail.

Crystal Generation. To generate a new crystal material, we leverage the trained latent diffusion model. We initialize the generative process from Gaussian noise $\boldsymbol{Z}^{(0)}$ and progressively transform it into a crystal latent representation by integrating the vector field learned by the denoising network $\mathcal{F}_\Theta$ over T steps. Specifically, at each step, we apply Euler integration to evolve the latent along the learned vector field, gradually transporting it from the noise distribution toward the learned latent data distribution. After the final denoising step, the generated latent $\boldsymbol{Z}^{(1)}$ is passed through

Algorithm 1 Inference Algorithm of CrysUniNet

Input: For generation: denoiser $\mathcal{F}_\Theta$, decoder $\mathcal{D}_\theta$ ($\mathcal{D}_\theta^{struct}, \mathcal{D}_\theta^{prop}$), integration steps T. For property prediction: crystal $\boldsymbol{M} = (\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$, fine-tuned encoders $\mathcal{E}_\phi, \mathcal{E}_\psi$.

Crystal Generation
Sample $\boldsymbol{Z}^{(0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})^{N \times 2d}$
$\Delta t \leftarrow 1/T$
for $t = 0$ **to** 1 **step** Δt **do**
  $\bar{\boldsymbol{Z}} \leftarrow \mathcal{F}_\Theta(\boldsymbol{Z}^{(t)}, t)$
  // Euler step along the vector field
  $\boldsymbol{Z}^{(t+\Delta t)} \leftarrow \boldsymbol{Z}^{(t)} + \Delta t \cdot \dfrac{\bar{\boldsymbol{Z}} - \boldsymbol{Z}^{(t)}}{1-t}$
end for
$\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{L}} \leftarrow \mathcal{D}_\theta^{struct}(\boldsymbol{Z}^{(1)})$
return crystal $(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{L}})$

Property Prediction
Construct multi-graph $\mathcal{G}$ from $\boldsymbol{M}$
$\boldsymbol{Z}_{\text{2D}} \leftarrow \mathcal{E}_\phi(\mathcal{G})$
$\boldsymbol{Z}_{\text{3D}} \leftarrow \mathcal{E}_\psi(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$
$\boldsymbol{Z} \leftarrow [\, \boldsymbol{Z}_{\text{2D}} \,\|\, \boldsymbol{Z}_{\text{3D}} \,]$
$\boldsymbol{g} \leftarrow \text{READOUT}(\boldsymbol{Z})$
$\hat{y}^* \leftarrow \mathcal{D}_\theta^{prop}(\boldsymbol{g})$
return predicted property $\hat{y}^*$

the decoder's structure reconstruction head $\mathcal{D}_\theta^{struct}$ to obtain the corresponding candidate crystal structure.

Property Prediction. For property prediction, we reuse the pretrained VAE encoders and property head and fine-tune them end-to-end on the labeled dataset of the downstream target property. Given a crystal $\boldsymbol{M}$, we obtain its crystal-level representation $\boldsymbol{g}$ as defined in Section 3.1 and regress the target property as $\hat{y}^* = \mathcal{D}_\theta^{\text{prop}}(\boldsymbol{g})$, where y^* denotes the ground-truth value of the downstream property (distinct from the formation energy y used in Section 3.1). The encoders and head are fine-tuned end-to-end by minimizing the mean squared error:

$$\min_{\phi,\psi,\theta} \mathcal{L}_{\text{MSE}} = \|\hat{y}^* - y^*\|^2. \tag{6}$$

Since the encoders are initialized from multi-task pretraining, the predictor adapts to each new target property with comparatively little labeled data, which is especially beneficial in the low-data regimes typical of materials property datasets.

4 Experiments

In this section, we provide a comprehensive evaluation of CrysUniNet for both tasks against existing state-of-the-art baselines. We evaluate the generative task in Section 4.1 and the predictive task in Section 4.2. For each task, we first describe the experimental setup, the selected baseline models, and the evaluation metrics, followed by a detailed analysis of the empirical results. Finally, we provide an ablation on input modality, pretraining objectives, and fusion strategy (Section 4.3).

4.1 Crystal Generation task

Objective. In the random material generation task, the objective is to generate novel crystal structures together with their compositions such that the resulting materials are both valid and stable. We train the latent diffusion stage using the widely adopted MP-20 dataset (Xie et al., 2022), which contains 45,231 crystal structures with up to 20 atoms in the unit cell. Following prior work, we split the dataset into training, validation, and test sets using a 60%-20%-20% ratio. Detailed specifications of the experimental setup are provided in the Appendix, Section D.

Baseline Models and Metrics. We benchmark CrysUniNet against a diverse set of representative crystal generation methods, including CDVAE (Xie et al., 2022), DiffCSP (Jiao et al., 2023), UniMat (Yang et al., 2024), FlowMM (Miller et al., 2024), FlowLLM (Sriram et al., 2024), and ADiT (Joshi et al., 2025). Our evaluation focuses on two complementary aspects of generation quality: *validity* and *thermodynamic stability* (Details in the Appendix, Section E).

Results. As shown in Table 1, CrysUniNet consistently outperforms the baseline methods across all four stability-related metrics: Stability, S.U.N., Metastability, and M.S.U.N. Specifically, compared with the next best-performing baseline, CrysUniNet achieves a **23.05%** improvement in the Stability rate and a **1.3%** improvement in the Metastability rate. More importantly, it yields **39.15%**

Method	Validity Metrics (↑)		Stability Metrics (↑)			
	Composition	Structural	Stable (%)	SUN (%)	Metastable (%)	MSUN (%)
CDVAE	86.70	**100.00**	5.00	3.30	23.58	21.99
DiffCSP	83.25	**100.00**	4.60	2.80	30.60	22.50
UniMat	89.40	97.20	13.90	<u>4.70</u>	66.90	<u>26.30</u>
FlowMM	83.19	96.85	4.60	2.80	30.60	22.50
FlowLLM	**90.84**	<u>99.94</u>	13.90	<u>4.70</u>	66.90	<u>26.30</u>
ADiT (MP20 only)	<u>90.46</u>	99.58	<u>14.10</u>	<u>4.70</u>	<u>81.60</u>	25.91
CrysUniNet	89.13	99.46	**17.35**	**6.54**	**82.66**	**31.59**

Table 1: Random structure generation performance of CrysUniNet against baselines on the MP-20 dataset. Best and second-best per metric are shown in **bold** and <u>underlined</u>, respectively.

and **20.11%** more S.U.N. and M.S.U.N. materials, respectively. In terms of validity, CrysUniNet maintains compositional and structural validity comparable to those of existing baseline models. Overall, these results demonstrate that CrysUniNet not only preserves high structural and compositional validity, but also substantially improves the generation of thermodynamically stable, unique, and novel crystal candidates, highlighting the effectiveness of the proposed unified latent representation for material generation.

4.2 Crystal Property Prediction Task

Objective. In the property-prediction task, given a crystal material as input, the objective is to accurately estimate its target property. We evaluate CrysUniNet on the JARVIS-DFT benchmark (Choudhary et al., 2020), which contains $55{,}722$ crystal structures with properties computed using DFT. Following prior works (Choudhary & DeCost, 2021; Yan et al., 2022), we consider nine target properties: formation energy, band gap (OPT), band gap (MBJ), total energy, energy above hull, bulk modulus, shear modulus, SLME, and spillage. For all properties, we adopt an $80/10/10$ train/validation/test split, consistent with Matformer (Yan et al., 2022). Detailed statistics for each property are provided in Table 4; experimental setup is described in the Appendix (Section D).

Baseline Models and Metrics. We compare CrysUniNet against a diverse set of state-of-the-art models for crystal property prediction, including CGCNN (Xie & Grossman, 2018), SchNet (Schütt et al., 2018), MEGNet (Chen et al., 2019), GATGNN (Louis et al., 2020), ALIGNN (Choudhary & DeCost, 2021), Matformer (Yan et al., 2022), PotNet (Lin et al., 2023), eComFormer (Yan et al., 2024), and iComFormer (Yan et al., 2024). Unlike CrysUniNet, these baselines are supervised models specifically designed for property prediction and are trained from scratch using property-labeled data for each target task. In contrast, CrysUniNet leverages the representations learned during its pretraining stage and subsequently fine-tunes the pretrained encoders and property prediction head for each target property. We evaluate all models using the Mean Absolute Error (MAE) between the predicted and ground-truth property values on the test set.

Results. The comparative results across all nine properties are reported in Table 2. We observe that CrysUniNet achieves lower MAE than the baseline models on the majority of the properties. In particular, CrysUniNet achieves the best performance on six of the nine properties, improving over the next-best baseline by **4.91%**, **1.92%**, **0.16%**, **0.04%**, **1.31%**, and **1.94%** on Bandgap (OPT), Bandgap (MBJ), Bulk Modulus, Shear Modulus, SLME, and Spillage, respectively. It further matches the best-performing baseline on Total Energy. For the remaining Formation Energy and E_{hull}, CrysUniNet still outperforms most competing methods and achieves results comparable to the strongest recent baselines. These results highlight the strong generalization capability of the learned representation, enabling CrysUniNet to achieve consistently competitive or state-of-the-art performance across a diverse range of crystal property prediction tasks.

Discussion. Overall, the results across both tasks highlight the effectiveness of our proposed unified framework for random material generation and property prediction. These findings empirically validate our hypothesis that jointly encoding the 2D multigraph and 3D crystal geometry into a shared latent space, while coupling structure reconstruction with formation-energy prediction during multi-task pretraining, yields a more informative and transferable representation. This design enables the latent space to jointly capture topological connectivity, geometric structure, and property-relevant information, making it both structure- and property-aware. As a result, the learned representation not

Model	Formation Energy	Bandgap (OPT)	Total Energy	E_{hull}	Bandgap (MBJ)	Bulk Modulus	Shear Modulus	SLME (%)	Spillage
CGCNN	0.063	0.200	0.078	0.170	0.410	14.47	11.75	8.022	0.454
SchNet	0.045	0.190	0.047	0.140	0.430	13.25	11.12	7.431	0.409
MEGNet	0.047	0.145	0.058	0.084	0.340	14.20	12.25	7.213	0.445
GATGNN	0.047	0.170	0.056	0.120	0.510	14.32	12.48	7.504	0.431
ALIGNN	0.033	0.142	0.037	0.076	0.310	10.40	9.481	5.146	0.389
Matformer	0.033	0.137	0.035	0.064	0.300	11.21	10.76	5.260	0.398
PotNet	0.029	0.127	<u>0.032</u>	0.055	0.270	10.11	9.232	4.570	0.361
eComFormer	<u>0.028</u>	0.124	<u>0.032</u>	0.047	0.282	10.79	9.826	4.610	0.373
iComFormer	**0.027**	<u>0.122</u>	**0.029**	**0.044**	<u>0.261</u>	<u>9.617</u>	<u>9.098</u>	<u>4.583</u>	<u>0.360</u>
CrysUniNet	0.029	**0.116**	**0.029**	<u>0.046</u>	**0.256**	**9.602**	**9.094**	**4.510**	**0.353**

Table 2: Property-prediction MAE (lower is better) of CrysUniNet against supervised baselines across nine JARVIS-DFT properties. Best and second-best per property are shown in **bold** and <u>underlined</u>, respectively.

only supports the generation of more thermodynamically stable crystal structures, but also transfers effectively to a diverse range of downstream crystal property prediction tasks.

4.3 Ablation on Input Modality, Pretraining Objectives, and Fusion Strategy

The encoder of CrysUniNet fuses a 2D multi-graph and a 3D geometric representation of the crystal by concatenation, and is pretrained with a multi-task objective combining structure reconstruction and property prediction. We ablate the input modality, the pretraining objectives, and the fusion strategy, and evaluate each variant on structure generation (stable, SUN, metastable, and MSUN rates) and property prediction (MAE on four JARVIS-DFT properties). Results are shown in Figure 2, with one column per ablation axis.

Input modality. We compare CrysUniNet against *2D only* and *3D only*, which pretrain the VAE on a single modality. For generation (Figure 2a), the 3D geometry is indispensable: *2D only* attains a structural validity of only 16.91% (compositional 86.69%), and only 0.80% of its samples are stable, as the multi-graph lacks the explicit geometry needed to decode plausible lattices and coordinates. *3D only* recovers high validity (99.31%) but trails CrysUniNet in stable (13.53% vs. 17.35%) and metastable (71.02% vs. 82.66%) rates, while achieving higher SUN and MSUN rates. For property prediction (Figure 2d), the roles reverse: *2D only* consistently outperforms *3D only*, indicating that the multi-graph captures the local chemical environment more effectively than raw 3D geometry. Both single-modality encoders remain inferior to the fused encoder on all four properties. The modalities are thus complementary: 3D is essential for generation, 2D is the stronger signal for property prediction, and their fusion yields both the highest stable rate and the lowest MAE.

Pretraining objectives. *w/o Prop* removes the property-prediction term from the pretraining loss, and *w/o Struct* removes the structure-reconstruction term; as the latter no longer trains the decoder, it is evaluated on property prediction only. Removing the property objective reduces the stable rate from 17.35% to 13.49% and the metastable rate from 82.66% to 70.01% (Figure 2b), suggesting that property supervision steers the latent space toward thermodynamically favorable regions. Conversely, removing reconstruction increases MAE on all four properties (Figure 2e), indicating that reconstruction shapes a latent space better suited for property prediction. The two objectives are therefore mutually beneficial: each improves the task primarily served by the other.

Fusion strategy. We replace concatenation with *Self-attention*, *Cross-attention (3D query)*, and *Cross-attention (2D query)*. For generation (Figure 2c), *Self-attention* matches the stable rate of CrysUniNet (17.16% vs. 17.35%) and has a higher metastable rate, but much lower SUN (4.21%) and MSUN (20.05%) rates. *Cross-attention (3D query)* slightly exceeds CrysUniNet in SUN and MSUN, but yields fewer stable structures (13.37%). *Cross-attention (2D query)* is the weakest: its structural validity drops to 92.68%, and it falls below CrysUniNet on all four stability metrics (e.g., 8.28% stable), consistent with the modality ablation, since the geometry that generation relies on is accessed only indirectly when 2D serves as the query. For property prediction (Figure 2f), all attention variants yield higher MAE than concatenation, with *Cross-attention (2D query)* again the weakest, showing that simple concatenation better preserves the complementary signals of the two modalities.

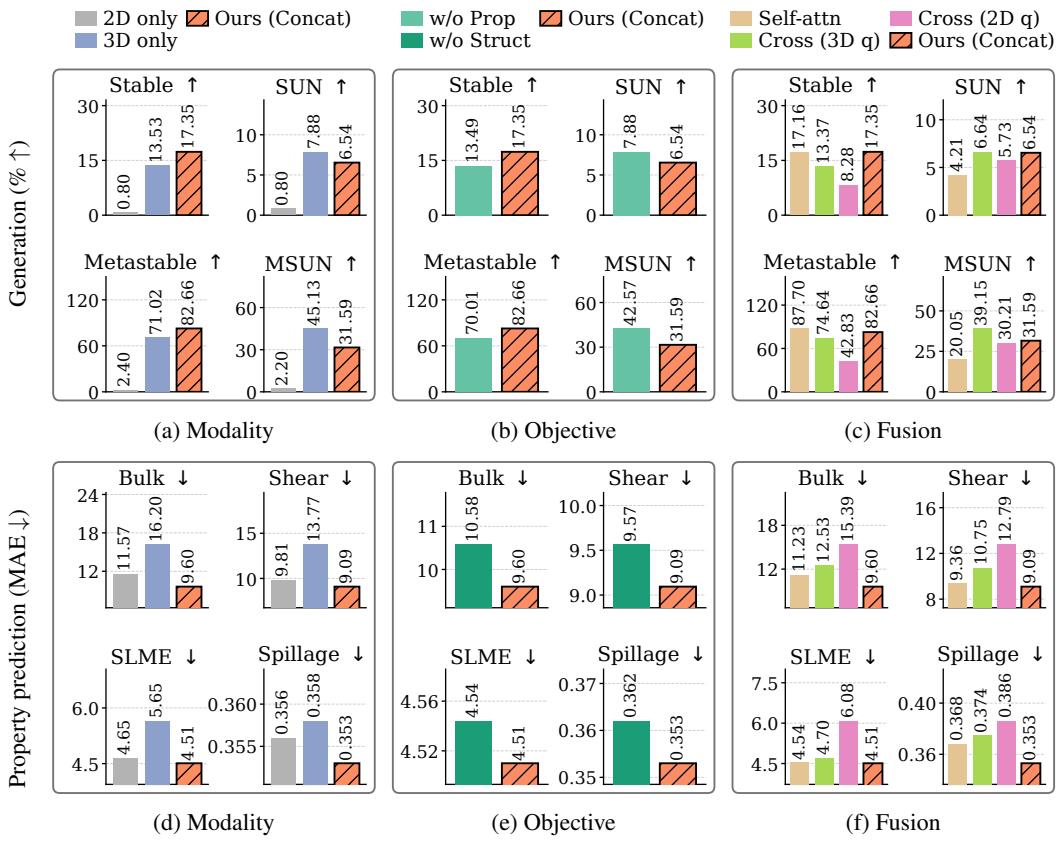

Figure 2: Ablation of CrysUniNet on input modality, pretraining objectives, and fusion strategy. Top (a–c): stable, SUN, metastable, and MSUN rates of generated structures in %. Bottom (d–f): MAE on four JARVIS-DFT properties (bulk and shear modulus in GPa, SLME in %). Legends are shared within each column. Cross (3D q) and Cross (2D q) denote cross-attention fusion with the 3D and 2D representation as the query, respectively.

Discussion. We observe a trade-off between stability and novelty in generation. Variants with fewer stable structures (*3D only*, *w/o Prop*, *Cross-attention (3D query)*) achieve higher SUN and MSUN rates, as a larger fraction of their stable samples are unique and novel, whereas fusion and property supervision concentrate samples in stable regions at a moderate cost in novelty. CrysUniNet is the only variant that achieves both the lowest MAE on all four properties and the highest stable rate. We therefore adopt concatenation-based fusion with both pretraining objectives as our final model, as it offers the best balance between predictive accuracy and the stability of generated structures.

5 Conclusion

In this work, we presented CrysUniNet, a unified framework that integrates property prediction and crystal structure generation into a common latent space of periodic materials. A unified encoder combines two different material representations, one with a local 2D view and the other a global 3D structure view, into the latent space that a diffusion model is trained upon to generate new material structures. The multi-task decoder is then able to decode the structure as well as predict the property of the generated material. This unified design achieves higher stability and SUN rates than baseline models trained on individual tasks, as well as lower prediction error in predicting properties like bulk modulus, band gap, etc. of material structures. This suggests that the two tasks benefit from each other rather than competing for representational capacity. Our model enables a common latent space for multiple tasks as well as multiple tasks (generation and prediction) with the same model architecture.

AI USE STATEMENT

No AI tools were used in conducting the research presented in this paper. However, we employed large language models (ChatGPT and Claude) solely for editorial purposes, including refining grammar, spelling, word choice, and overall clarity of the manuscript.

ETHICS STATEMENT

This work introduces CrysUniNet, a unified framework for crystal property prediction and de novo generation. By accelerating the screening and design of novel crystalline materials, the proposed approach may benefit applications in catalysis, energy storage, semiconductors, and other areas of materials science and chemistry. As with any method that facilitates targeted materials discovery, there is potential for misuse in developing materials with unintended or harmful applications; we therefore encourage responsible use and appropriate oversight when applying such models in practice.

REPRODUCIBILITY STATEMENT

The full source code for CrysUniNet, including model implementation, training scripts, and configuration files, is provided as anonymous supplementary material. Model architecture and fusion variants are described in Section 3, and training details and hyperparameters are provided in Appendix D.

REFERENCES

Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):10570, 2024.

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.

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.

Kamal Choudhary, Kevin F Garrity, Andrew CE Reid, Brian DeCost, Adam J Biacchi, Angela R Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A Gilad Kusne, Andrea Centrone, et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj computational materials*, 6(1):173, 2020.

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.

Bin Dai and David Wipf. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019.

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

Kishalay Das, Subhojyoti Khastagir, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Periodic materials generation using text-guided joint diffusion model. In *The Thirteenth International Conference on Learning Representations*, 2025.

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.

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.

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.

Ruiqi Gao, Emiel Hoogeboom, Jonathan Heek, Valentin De Bortoli, Kevin P Murphy, and Tim Salimans. Diffusion meets flow matching: Two sides of the same coin. 2024. *URL https://diffusionflow. github. io*, 2024.

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

Prashant Govindarajan, Kishalay Das, and Sarath Chandar. CrysTune: Crystal generation via fine-tuning of large language models on wyckoff representations. In *ICML 2026 AI for Science Workshop*, 2026.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, Larry Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21671–21690, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5dc6d1c81754c32008c9667339b3fdd3-Paper-Conference.pdf.

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.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 17464–17497. Curran Associates, Inc., 2023. doi: 10.52202/075280-0767. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/38b787fc530d0b31825827e2cc306656-Paper-Conference.pdf.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 6836–6853, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/1ba1b2133ac3d3dda0ce5cd67dab8456-Paper-Conference.pdf.

Chaitanya K. Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary W. Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. In *International Conference on Machine Learning*, 2025.

Subhojyoti Khastagir, Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. CrysLDM: Latent diffusion model for crystal material generation. In *AI for Accelerated Materials Design-ICLR 2025*, 2025a.

Subhojyoti Khastagir, KISHALAY DAS, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. LLM meets diffusion: A hybrid framework for crystal material generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=E6gwPtWjb1.

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.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1(1):1–15, 2015.

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.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=mCOBKZmrzD.

Yuchao Lin, Keqiang Yan, Youzhi Luo, Yi Liu, Xiaoning Qian, and Shuiwang Ji. Efficient approximations of complete interatomic potentials for crystal property prediction. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21260–21287. PMLR, 23–29 Jul 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M Wood. FlowMM: Generating materials with Riemannian flow matching. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35664–35686. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/miller24a.html.

Shrimon Mukherjee, KISHALAY DAS, Partha Basuchowdhuri, Pawan Goyal, and Niloy Ganguly. Latent diffusion pretraining for crystal property prediction. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=Q5uj05ynu5.

Shrimon Mukherjee, Kishalay Das, Partha Basuchowdhuri, Pawan Goyal, and Niloy Ganguly. Model agnostic graph prompt learning for crystal property prediction. *arXiv preprint arXiv:2607.08996*, 2026b.

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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Zekun Ren, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G. Aberle, Shijing Sun, Xiaonan Wang, Yi Liu, Qianxiao Li, Senthilnath Jayavelu, Kedar Hippalgaonkar, Yousung Jung, and Tonio Buonassisi. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 5(1):314–335, 2022. ISSN 2590-2385. doi: https://doi.org/10.1016/j.matt.2021.11.032. URL https://www.sciencedirect.com/science/article/pii/S2590238521006251.

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.

Mihaela Rosca, Balaji Lakshminarayanan, and Shakir Mohamed. Distribution matching in variational inference. *arXiv preprint arXiv:1802.06847*, 2018.

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, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. SchNet–a deep learning architecture for molecules and materials. *The Journal of chemical physics*, 148(24), 2018.

Shuaike Shen, Ke Liu, Muzhi Zhu, and Hao Chen. A denoising pre-training framework for accelerating novel material discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28368–28376, 2025a.

Xiangxiang Shen, Zheng Wan, Lingfeng Wen, Licheng Sun, Jian Yang, Xuan Tang, Shing-Ho J. Lin, Xiao He, Mingsong Chen, and Xian Wei. PDDFormer: Pairwise distance distribution graph transformer for crystal material property prediction. In *IJCAI*, pp. 7724–7732, 2025b. URL https://doi.org/10.24963/ijcai.2025/859.

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.

David S. Sholl and Janice A. Steckel. *Density Functional Theory: A Practical Introduction*. Wiley, March 2009. ISBN 9780470447710. doi: 10.1002/9780470447710. URL http://dx.doi.org/10.1002/9780470447710.

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.

Zixing Song, Ziqiao Meng, and Irwin King. A diffusion-based pre-training framework for crystal property prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8993–9001, 2024.

Anuroop Sriram, Benjamin Miller, Ricky TQ Chen, and Brandon Wood. FlowLLM: Flow matching for material generation with large language models as base distributions. *Advances in Neural Information Processing Systems*, 37:46025–46046, 2024.

Tatsunori Taniai, Ryo Igarashi, Yuta Suzuki, Naoya Chiba, Kotaro Saito, Yoshitaka Ushiku, and Kanta Ono. Crystalformer: Infinitely connected attention for periodic structure encoding. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. URL https://openreview.net/forum?id=fxQiecl9HB.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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 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. In *International Conference on Learning Representations*, 2022. URL https://mlanthology.org/iclr/2022/xu2022iclr-geodiff/.

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.

Keqiang Yan, Cong Fu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Complete and efficient graph transformers for crystal material property prediction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BnQY9XiRAS.

Sherry Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable Diffusion for Materials Generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wm4WlHoXpC.

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, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, Mar 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08628-5. URL https://doi.org/10.1038/s41586-025-08628-5.

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.

A Symmetry in Crystal Structure

Crystal materials obey fundamental physical symmetries (Dresselhaus et al., 2007; Zee, 2016), and a central challenge in crystal representation learning is that the learned embedding must remain invariant under translation, rotation, and periodic transformations.

- ***Rotational Invariance:*** A crystal is unchanged when its atomic coordinates and lattice are rotated. Formally, for any orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{3\times3}$ satisfying $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$, rotating the coordinates and lattice of a material $\boldsymbol{M}$ yields $\boldsymbol{M}_R = (\boldsymbol{A}, \boldsymbol{QX}, \boldsymbol{QL})$, which is merely a different representation of the same underlying material. A faithful encoder must therefore map $\boldsymbol{M}$ and $\boldsymbol{M}_R$ to identical representations.
- ***Periodic Translation Invariance:*** A crystal is also unchanged when its atomic coordinates are translated, because the periodic unit cell can be shifted arbitrarily along the lattice vectors, giving many equivalent choices of unit cell and coordinate matrix for the same material. Formally, for a material $\boldsymbol{M} = (\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$ and an arbitrary translation vector $\boldsymbol{u} \in \mathbb{R}^3$, applying the periodic wrapping operator $w[\boldsymbol{X}] := \boldsymbol{X} - \lfloor \boldsymbol{X} \rfloor$ produces $\boldsymbol{M}_P = (\boldsymbol{A},\ w(\boldsymbol{X}+\boldsymbol{u}\boldsymbol{1}^\top),\ \boldsymbol{L})$, which represents the same crystal as $\boldsymbol{M}$. The learned representation should thus be invariant to such periodic translations of $\boldsymbol{X}$.

B Diffusion Models for Crystal Materials

Diffusion models are generative models formulated as a T-step Markov chain. Given a data point $\boldsymbol{d}_0$, the forward process gradually corrupts it over T steps by injecting a small amount of Gaussian noise at each step:

$$q(\boldsymbol{d}_t \mid \boldsymbol{d}_0) = \mathcal{N}\big(\boldsymbol{d}_t \mid \sqrt{\bar{\alpha}_t}\,\boldsymbol{d}_0,\ (1-\bar{\alpha}_t)\boldsymbol{I}\big)\,, \tag{7}$$

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 noise variance at each step according to a chosen scheduler. The parameterized reverse process then starts from Gaussian noise $\boldsymbol{d}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and iteratively denoises the intermediate variables $\boldsymbol{d}_{T:1}$ to recover clean data $\boldsymbol{d}_0$ drawn from the target distribution:

$$p_\theta(\boldsymbol{d}_{t-1} \mid \boldsymbol{d}_t) = \mathcal{N}\big(\boldsymbol{d}_{t-1};\ \mu_\theta(\boldsymbol{d}_t, t),\ \rho_t^2\boldsymbol{I}\big)\,, \tag{8}$$

where ρ_t is a predefined variance and the mean μ_θ is typically modeled by a neural network (e.g., a U-Net for images). Applying diffusion models to crystal generation is difficult, however, because a crystal's joint distribution is highly multi-modal, with each component following a distinct modality: atom types are discrete, lattice parameters are continuous, and fractional coordinates are continuous yet periodic. Consequently, each variable calls for a diffusion formulation tailored to its own structure.

Diffusion on Atom Type ($\boldsymbol{A}$). The atom-type matrix $\boldsymbol{A} \in \mathbb{R}^{N\times k}$ can be viewed as N discrete variables over k classes, for which a discrete diffusion model (D3PM) (Austin et al., 2021) is well suited. With $\boldsymbol{a}$ denoting the one-hot encoding of an atom, the forward transition is $q(\boldsymbol{a}_t \mid \boldsymbol{a}_{t-1}) = \mathrm{Cat}(\boldsymbol{a}_t;\ \boldsymbol{p} = \boldsymbol{a}_{t-1}\boldsymbol{Q}_t)$, where $\mathrm{Cat}(\boldsymbol{a};\boldsymbol{p})$ is a categorical distribution over $\boldsymbol{a}$ with probabilities $\boldsymbol{p}$, and $\boldsymbol{Q}_t$ is the Markov transition matrix at step t, with $[\boldsymbol{Q}_t]_{i,j} = q(\boldsymbol{a}_t = i \mid \boldsymbol{a}_{t-1} = j)$. Different choices of $\boldsymbol{Q}_t$ and their stationary distributions (Austin et al., 2021) offer flexibility in controlling the corruption and denoising process.

Diffusion on Atom Coordinates ($\boldsymbol{X}$). The coordinate matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N\times3}$ holds the fractional coordinates of the atoms, which live in the quotient space $\mathbb{R}^{N\times3}/\mathbb{Z}^{N\times3}$ induced

by crystal periodicity. A standard DDPM is therefore unsuitable for $\boldsymbol{X}$, since its Gaussian noise cannot capture the cyclical, bounded domain of fractional coordinates. Instead, each forward step adds noise drawn from a Wrapped Normal (WN) distribution (De Bortoli et al., 2022), and a score-matching network (Song & Ermon, 2019; 2020) is used during denoising to model the transition probability.

Diffusion on Lattice ($\boldsymbol{L}$). The lattice matrix $\boldsymbol{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3]^\top \in \mathbb{R}^{3\times3}$ is a global feature that determines the shape and symmetry of the unit cell. Since $\boldsymbol{L}$ lies in a continuous space, a standard Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) is applied for its diffusion.

A key challenge in the reverse process is ensuring that the learned distribution of crystal structures respects periodic $E(3)$ invariance. To this end, prior work employs periodic-$E(3)$-equivariant GNNs, such as GemNet (Gasteiger et al., 2021) or CSPNet (Jiao et al., 2023), as the backbone denoising network. Training combines a cross-entropy loss, a score-matching loss, and an ℓ_2 loss for atom types, coordinates, and lattice parameters, respectively. Existing pretraining approaches (Song et al., 2024; Shen et al., 2025a) operate directly in high-dimensional feature spaces, applying diffusion-based pretraining on unlabeled data. This heterogeneous modeling demands complex denoising architectures and many diffusion steps to yield high-quality representations. To overcome this, we instead perform pretraining in a smoother latent space, which streamlines learning while preserving the essential structural and chemical information.

C Related Work

C.1 Diffusion Models

Diffusion models, first introduced by Sohl-Dickstein et al. (2015), operate by progressively corrupting data with noise and training a neural network to reverse this process and recover the original data. The idea subsequently branched into two main families: *Score Matching Networks* (Song & Ermon, 2019; 2020) and *Denoising Diffusion Probabilistic Models (DDPM)* (Ho et al., 2020). Diffusion models have since become a dominant class of deep generative models, delivering strong results across a wide range of domains, including image synthesis (Dhariwal & Nichol, 2021), molecular conformer generation (Shi et al., 2021; Xu et al., 2022), molecular graph generation (Liu et al., 2021), and protein folding (Wu et al., 2021; Luo et al., 2022). More recently, latent diffusion models (LDMs) have shown promising results across several applications such as image generation (Vahdat et al., 2021), point cloud synthesis (Vahdat et al., 2022), and text generation (Li et al., 2022). Among these, Stable Diffusion (Rombach et al., 2022) stands out for its highly realistic text-guided image generation.

C.2 Crystal Material Generation

Early efforts in generating novel periodic materials focused largely on atomic composition while neglecting the 3D structure. As generative modeling advanced, much of the work turned to VAEs and GANs for producing 3D periodic structures, either by representing materials as three-dimensional voxel images (Court et al., 2020; Hoffmann et al., 2019; Long et al., 2021; Noh et al., 2019) that depict atom types, coordinates, and lattices, or by directly encoding structures as embedding vectors (Kim et al., 2020; Ren et al., 2022; Zhao et al., 2021). However, these models neither enforce stability in the generated structures nor guarantee invariance to Euclidean and periodic transformations. Equivariant diffusion models opened a more promising direction for generating three-dimensional periodic crystals. CDVAE (Xie et al., 2022) was the first to combine a VAE with a score-based decoder that operates directly on atomic coordinates, using an equivariant GNN to ensure Euclidean and periodic invariance. A number of subsequent works (Luo et al., 2023; Jiao et al., 2023; Zeni et al., 2025; Jiao et al., 2024; Yang et al., 2024; Das et al., 2025) employed diffusion models to learn the joint distribution over atom types, coordinates, and lattices, enabling the generation of stable periodic structures. A key drawback of these approaches is that they operate directly in a high-dimensional feature space, jointly modeling heterogeneous components (atom types, fractional coordinates, and lattice parameters), which demands complex denoising architectures and many diffusion steps to produce high-quality structures. To alleviate this, CrysLDM (Khastagir et al., 2025a) and ADiT (Joshi et al., 2025) introduced latent diffusion models that operate in a smoother, lower-dimensional latent space, improving the stability and validity of generated materials. CrysLLMGen (Khastagir et al., 2025b) adopts a hybrid two-stage strategy, using an LLM to predict atomic compositions and a diffusion model to refine continuous structural variables, outperforming

Algorithm 2 Training Algorithm of CrysUniNet

Input: Crystal Material $\boldsymbol{M} = (\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$, Property Value y, Encoders $\mathcal{E}_\phi, \mathcal{E}_\psi$, Decoder $\mathcal{D}_\theta^{struct}$, Property Head $\mathcal{D}_\theta^{prop}$, and Denoising Network $\mathcal{F}_\Theta$.

Stage-1: Training VAE
repeat
  $\boldsymbol{Z}_{\text{2D}} \leftarrow \mathcal{E}_\phi(\mathcal{G})$
  $\boldsymbol{Z}_{\text{3D}} \leftarrow \mathcal{E}_\psi(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$
  $\mathbf{Z} \leftarrow [\, \boldsymbol{Z}_{\text{2D}} \,\|\, \boldsymbol{Z}_{\text{3D}} \,]$
  $\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{L}} \leftarrow \mathcal{D}_\theta^{struct}(\mathbf{Z})$
  $\hat{y} \leftarrow \mathcal{D}_\theta^{prop}(\text{READOUT}(\mathbf{Z}))$
  $\mathcal{L}^{\boldsymbol{A}} = \text{CrossEntropyLoss}(\tilde{\boldsymbol{A}}, \boldsymbol{A})$
  $\mathcal{L}^{\boldsymbol{L}} = \|\tilde{\boldsymbol{L}} - \boldsymbol{L}\|_2^2$
  $\mathcal{L}^{\boldsymbol{X}} = \|\tilde{\boldsymbol{X}} - \boldsymbol{X}\|_2^2$
  $\mathcal{L}_{\text{prop}} = \|\hat{y} - y\|^2$
  Minimize $\mathcal{L}^{\boldsymbol{A}}_{\text{recon}} + \mathcal{L}^{\boldsymbol{X}}_{\text{recon}} + \mathcal{L}^{\boldsymbol{L}}_{\text{recon}} + \mathcal{L}_{\text{prop}} + \alpha\mathcal{L}_{reg}$
  Update $\mathcal{E}_\phi, \mathcal{E}_\psi, \mathcal{D}_\theta^{struct}, \mathcal{D}_\theta^{prop}$
until Converged

Stage-2: Training LDM
repeat
  $\mathbf{Z}^{(1)} \leftarrow [\, \mathcal{E}_\phi(\mathcal{G}) \,\|\, \mathcal{E}_\psi(\boldsymbol{M}) \,]$ // Clean Latent
  Sample $t \sim \mathcal{U}(0, 1)$
  Sample Noise $\mathbf{Z}^{(0)} \sim N(0, 1)^{N \times 2d}$
  $\boldsymbol{Z}^{(t)} = (1 - t)\, \boldsymbol{Z}^{(0)} + t\, \boldsymbol{Z}^{(1)}$
  $\bar{\boldsymbol{Z}}^{(1)} = \mathcal{F}_\Theta(\boldsymbol{Z}^{(t)}, t)$
  $\mathcal{L}_{LDM} = \frac{1}{(1-t)^2} \frac{1}{N} \sum_{i=1}^{N} \left\| \boldsymbol{z}_i^{(1)} - \bar{\boldsymbol{z}}_i^{(1)} \right\|^2$
  Minimize $\mathcal{L}_{LDM}$ and Update only $\mathcal{F}_\Theta$
until Converged

both standalone LLM-based and denoising-based methods. More recently, CrysTune (Govindarajan et al., 2026) fine-tunes LLaMA-2 on Wyckoff-based representations with symmetry-aware auxiliary tasks and GRPO reinforcement learning to produce more valid, stable, and property-controllable crystals, albeit at some cost to novelty and diversity.

C.3 Crystal Property Prediction

Graph neural networks (GNNs) (Xie & Grossman, 2018; Chen et al., 2019; Louis et al., 2020; Park & Wolverton, 2020; Schmidt et al., 2021; Choudhary & DeCost, 2021; Yan et al., 2022; Lin et al., 2023) have become the standard tool for crystal property prediction. Early methods (Xie & Grossman, 2018; Chen et al., 2019; Louis et al., 2020; Park & Wolverton, 2020; Schmidt et al., 2021) build a multi-edge graph from the 3D crystal structure and use a GNN to encode the local structural environment around each atom. Building on this foundation, subsequent work introduced architectural variants that embed domain knowledge into the encoder to improve representation quality. ALIGNN (Choudhary & DeCost, 2021) incorporates bond-angle information to model many-body interactions, while Matformer (Yan et al., 2022) is designed to be invariant to periodicity so that it can capture repeating patterns explicitly. CrysMMNet (Das et al., 2023) takes a multi-modal approach, fusing textual descriptions with the crystal graph to enhance prediction. A known limitation of Matformer is that it can occasionally represent structurally distinct crystals as identical graphs (Yan et al., 2024). To resolve this, ComFormer (Yan et al., 2024) introduces an SE(3)-invariant and an SO(3)-equivariant graph representation that capture both local and global geometry through interatomic distances and angular features, encoding them with node-wise and edge-wise transformer layers to represent diverse structures faithfully. More recently, CrysLDNet (Mukherjee et al., 2026a) shifts pretraining away from the raw feature space and into a compact latent space. A variational autoencoder first encodes each crystal into a smooth latent representation, and a diffusion model is trained to denoise within this latent space rather than directly over atom types, coordinates, and lattice parameters. The pretrained encoder is then fine-tuned with a lightweight head for downstream property prediction. Lately, Mukherjee et al. (Mukherjee et al., 2026b) proposed a lightweight, model-agnostic soft-prompt framework that adds node-level (local chemistry) and crystal-system-aware graph-level (global symmetry) prompts to any GNN-based crystal encoder, boosting crystal property prediction accuracy.

D Experimental Setup and Hyperparameters

All experiments were performed on an NVIDIA L40 GPU server with 46 GB of GPU memory and 128 GB of RAM. We pretrain the VAE component of CrysUniNet for 100 epochs on GNoME (Merchant et al., 2023) and train the latent diffusion component for 5000 epochs on MP-20, both using

Metric	$\|C\|$	$\|A\|$	$\|T\|$	Volume	Density
Max		40	6	9291.69	24.10
Min	380740	2	2	25.84	0.18
Mean		4.10	4	436.96	8.34
Var		0.46	0	62283.63	7.38

Table 3: Statistics of the GNoME dataset. Max, Min, Mean, and Var denote the maximum, minimum, average, and variance, respectively. $|A|$, $|T|$, and $|C|$ represent the number of atoms per crystal, the number of atom types per crystal, and the total number of crystal structures.

Property	Unit	Data-size	Train/Valid/Test
Formation Energy	eV/atom	55722	44578/5572/5572
Bandgap (OPT)	eV	55722	44578/5572/5572
Total Energy	eV/atom	55722	44578/5572/5572
Ehull	eV	55370	44296/5537/5537
Bandgap (MBJ)	eV	18171	14537/1817/1817
Bulk Modulus (Kv)	GPa	19680	15744/1968/1968
Shear Modulus (Gv)	GPa	19680	15744/1968/1968
SLME	(%)	9066	7254/906/906
Spillage	No unit	11377	9101/1138/1138

Table 4: Summary of different crystal properties in JARVIS-DFT datasets. Here we present total count and train/valid/test split for each property in the JARVIS-DFT dataset.

the AdamW optimizer with a batch size of 256 and a learning rate of 10^{-3}. For property-prediction fine-tuning, we adopt the same optimizer with a weight decay of 10^{-5} and a one-cycle scheduler, fine-tuning the pretrained VAE for 1000 epochs with a batch size of 64.

Denoiser for generation task The denoiser in the latent diffusion stage is a Diffusion Transformer model with a hidden dimension of 768, 12 attention heads, 12 layers, a feedforward dimension of 768, and a GELU activation.

VAE for joint structure decoding and prediction tasks The VAE comprises a ComFormer encoder that captures local 2D structural information and a standard Transformer encoder that captures global 3D geometry. The ComFormer encoder uses 256-dimensional node and edge features across 4 convolutional layers, while the Transformer encoder adopts a model dimension of 512 with 8 attention heads, 8 layers, a feedforward dimension of $2{,}048$, and GELU activations.

E Evaluation Metrics

Following the evaluation protocol of Xie et al. (2022); Miller et al. (2024), we sample 10,000 material structures and compute the validity metrics, followed by stability metrics after relaxing the structures with CHGNet (Deng et al., 2023).

Validity metrics Validity is a proxy measure (as noted by Miller et al. (2024)) designed to act as a cheap filter to judge the model's quality, and is measured in terms of both structural and compositional validity. Structural validity is computed as the percentage of generated crystals with valid periodic structures. A structure is defined to be valid if the shortest distance between any pair of atoms exceeds 0.5 Å and the unit cell has a volume more than 0.1 Å^3. Compositional validity is computed as the percentage of structures with correct combination of atom types. A composition is considered valid if the overall charge remains neutral as determined by SMACT (Davies et al., 2019).

Stability metrics Stability metrics consist of stable rate, SUN rate, metastable rate, and MSUN rate. Stable rate is defined as the percentage of materials with $E^{hull} \leq 0$ eV/atom after relaxation. Similarly, metastable rate is defined as the percentage of materials with $E^{hull} \leq 0.1$ eV/atom. SUN rate is defined as the percentage of materials that are stable as well as unique (not repeated in the batch of generated materials) and novel (not present in the training dataset), while MSUN rate is defined as the percentage of materials that are metastable as well as unique and novel.

Property Prediction Metrics Following prior works (Choudhary & DeCost, 2021; Xie & Grossman, 2018), we use the Mean Absolute Error (MAE) between predicted and true property values to evaluate the performance of all baseline models on the crystal property prediction task.

$$\text{MAE}(\mathcal{M}, f) = \frac{1}{m}\sum_{i=1}^{m}|f(\mathcal{M}_i) - y_i|$$