COMPOSABLE CRYSTALS: CONTROLLABLE MATERIALS DISCOVERY VIA CONCEPT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing methods for de novo crystal generation largely produce complete crystal structures, either through stochastic sampling or under predefined structural constraints, providing limited mechanisms for explicitly structuring how the crystal space is explored. In this work, we introduce a concept-based compositional framework that represents crystal generation as the recombination of learned structural units. We learn a discrete vocabulary of local crystal concepts using a vector-quantized variational autoencoder, where each concept represents a recurring local atomic environment. Unlike predefined structural attributes, these concepts are automatically discovered from data, and multiple concepts can coexist and be recombined within a single crystal. Moreover, the learned concepts exhibit interpretable patterns in both local atomic environments and global crystal symmetries, and empirically transfer across crystal datasets. We further introduce a composition generator to produce novel concept combinations and refine it using high-quality samples. The resulting concept compositions guide downstream crystal generation toward structural combinations beyond those observed in the training data. Experiments on MP-20 and Alex-MP-20 show that concept composition improves the V.S.U.N. score of the base generative model by 53.2% and 51.7%, respectively, with the main gain arising from increased novelty. Our code is available at `https://anonymous.4open.science/r/Composition-Crystal-Generation/`.

1 INTRODUCTION

De novo crystal generation (DNG) focuses on discovering novel and stable crystal structures without relying on predefined templates. Recent techniques (Xie et al., 2021; Jiao et al., 2023; Joshi et al., 2025) generate crystals through stochastic sampling, while conditional approaches typically impose predefined properties (Zeni et al., 2023), symmetries (Levy et al., 2025), or structural patterns (Jiao et al., 2024). These paradigms are effective for sampling realistic crystals or satisfying known constraints, but provide limited mechanisms for explicitly structuring how new regions of the crystal space are explored.

In this work, we consider a complementary question: *how can a crystal generative model systematically explore beyond the training data?* A promising answer is motivated by compositional generation (Wu et al., 2024), where complex samples can be represented as combinations of reusable concepts and unseen combinations can be constructed from previously learned components. For crystal generation, such a mechanism is particularly appealing because a crystal naturally contains multiple recurring local atomic environments, while novel structures may emerge from new arrangements and combinations of such local patterns. Rather than generating an entire crystal only as an indivisible sample, we therefore seek an intermediate structural vocabulary through which crystal generation can be explicitly composed.

This raises a second question: *why is such a learned concept vocabulary fundamentally different from existing representations or structural conditioning mechanisms?* Continuous representation-learning methods (Joshi et al., 2025) can encode rich crystal structures into continuous latent variables, but these representations do not directly provide a finite vocabulary of structural units that can be individually identified and explicitly recombined. Common material conditions, such as space groups (Cao et al.), motifs (Wei et al., 2026) or physical properties (Zeni et al., 2023), provide

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.

useful global constraints or target outcomes, but are typically predefined or specified at a relatively coarse level, and do not directly provide a flexible vocabulary of fine-grained units that can be freely recombined within a single crystal. In contrast, we aim to learn fine-grained structural units directly from data while making them explicitly addressable, reusable, and composable within a single crystal.

These requirements motivate our formulation of learned discrete local crystal concepts. Locality allows multiple concepts to coexist within one crystal, discreteness provides an operational vocabulary for explicit recombination, and data-driven learning avoids restricting the representation to manually specified motifs. Importantly, the same concept encoder can re-extract concepts from generated structures, enabling direct verification of whether the intended concept composition is realized. We implement this formulation using a VQ-VAE (Van Den Oord et al., 2017), where each local atomic environment is mapped to one entry in a shared discrete codebook and each entry is treated as a reusable crystal concept. Empirically, the learned codes exhibit coherent local geometric and chemical-composition patterns, while representative concepts are also associated with recognizable global crystal symmetries. The codebook learned on MP-20 (Xie et al., 2021) further remains useful on crystals from Alex-MP-20 (Zeni et al., 2023), indicating that recurring local patterns can transfer across datasets.

Once these concepts are defined, the next challenge is how to compose them into new candidates. Because unrestricted composition may produce incompatible or low-quality combinations (Wu et al., 2024), we introduce a composition generator to sample new concept combinations, and refine it using candidates that satisfy V.S.U.N. (validity, stability, uniqueness, and novelty) (Zeni et al., 2023) criteria relative to the training set. The refined composition generator is then used to guide a downstream crystal generative model toward structural combinations beyond those directly observed in the training distribution. In this way, the learned concepts form an intermediate compositional interface between observed crystal structures and structural exploration in DNG.

Our contributions are summarized as follows: (1) We introduce a compositional formulation for DNG in which structural exploration is performed through the recombination of reusable local structural units. (2) We learn a discrete structural vocabulary from crystal data, providing an explicit interface for extracting local crystal concepts, recombining them, and recovering them from generated structures without requiring predefined motifs. (3) We develop a concept composition and refinement framework that improves the base generative model on MP-20 and Alex-MP-20 by 53.2% and 51.7% on the `V.S.U.N` metric, respectively, with the largest gains arising from improved novelty.

2 PRELIMINARIES

2.1 CRYSTAL REPRESENTATION

A crystal $\mathbf{C}$ is defined by the periodic arrangement of its fundamental repeating unit, the unit cell, throughout the three-dimensional space. The unit cell can be fully described by Cartesian coordinates $\boldsymbol{X} \in \mathbb{R}^{N\times 3}$, atomic types $\boldsymbol{A} \in \mathbb{R}^{N}$, and a lattice matrix $\boldsymbol{L} \in \mathbb{R}^{3\times 3}$, where N denotes the number of atoms in the unit cell. For $\boldsymbol{A}$, we consider the first 100 elements in the periodic table and encode them using one-hot vectors, denoted by one-hot$(\boldsymbol{A}) \in \mathbb{R}^{N\times 100}$. For $\boldsymbol{L}$, we follow prior work (Zeni et al., 2023) and decompose it via singular value decomposition (SVD), reparameterizing it as a symmetric matrix $\tilde{\boldsymbol{L}}$: $\boldsymbol{L} = \boldsymbol{U}\tilde{\boldsymbol{L}}, \quad \boldsymbol{U} = \boldsymbol{W}\boldsymbol{V}^{\top}, \tilde{\boldsymbol{L}} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$, where $\boldsymbol{W}$ and $\boldsymbol{V}$ are the left and right singular vectors of $\boldsymbol{L}$, respectively, and $\boldsymbol{\Sigma}$ is the diagonal matrix of singular values. Here, $\boldsymbol{U}$ is a rotation matrix, and $\tilde{\boldsymbol{L}}$ is symmetric positive definite. We extract the upper-triangular entries of $\tilde{\boldsymbol{L}}$ and reshape them into a vector $\widehat{\boldsymbol{L}} = \mathrm{vec}(\mathrm{triu}(\tilde{\boldsymbol{L}})) \in \mathbb{R}^{6}$. The information associated with atom i in $\mathbf{C}$ is summarized as an *atom vector*:

$$\boldsymbol{v}_i = [\boldsymbol{X}_i \,\|\, \text{one-hot}(A_i) \,\|\, \widehat{\boldsymbol{L}}] \in \mathbb{R}^{3+100+6}, \tag{1}$$

where $\|$ denotes the concatenation operator. The complete crystal is then represented by the matrix $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N]^{\top} \in \mathbb{R}^{N\times(3+100+6)}$, where each row corresponds to an individual atom vector.

Atom Local Environment Since there is no unique way to define edges in crystals (Xie et al., 2021), prior studies typically model the local pattern of each atom by considering either its k-nearest neighbors (Jiao et al., 2023) or all neighbors within a cutoff radius (Zeni et al., 2023). In this paper, we construct local atomic environments using the *minimum-distance* rule (Zimmermann & Jain,

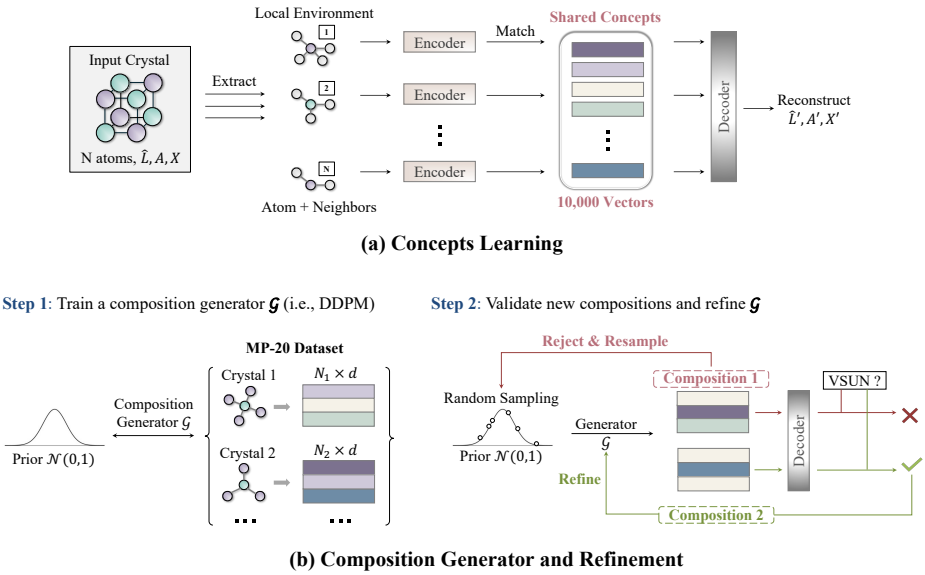


Figure 1: Pipelines for (a) learning a VQ-VAE to extract crystal concepts, and (b) training and refining a composition generator using qualified generated samples.

2020): (1) For atom i, find its nearest neighbor j and denote their distance by $d_{\min}$, and (2) retain other neighboring atoms q (up to K neighbors) that satisfy $d_{iq} \leq \xi \cdot d_{\min}$, where ξ is set to 1.1 following standard practice.

2.2 Diffusion-Based Generation

The family of diffusion models (Sohl-Dickstein et al., 2015; Rombach et al., 2022) has become one of the most powerful paradigms in generative modeling. In this work, we adopt a standard diffusion framework, DDPM (Ho et al., 2020). In DDPM, the forward process gradually injects Gaussian noise into a sample $\boldsymbol{x}$ at each step s: $\boldsymbol{x}_s = \sqrt{1-\beta_s}\boldsymbol{x}_{s-1} + \sqrt{\beta_s}\boldsymbol{\epsilon}_s = \sqrt{\bar{\alpha}_s}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_s}\boldsymbol{\epsilon}_s, \boldsymbol{\epsilon}_s \sim \mathcal{N}(0, \boldsymbol{I})$, where $\{\beta_1, \ldots, \beta_S\}$ is a predefined variance schedule, typically satisfying $\beta_1 < \cdots < \beta_S$, $\alpha_s := 1 - \beta_s$, and $\bar{\alpha}_s := \prod_{i=1}^{s} \alpha_i$. The simplified DDPM objective is to predict the injected noise $\boldsymbol{\epsilon}_s$ using a denoising network $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_s, s)$: $L_{\text{DDPM}}(\theta) := \mathbb{E}_{s,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon}_s - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_s, s)\|^2\right]$.

Classifier-free Guidance In addition to the timestep s, one can condition the denoising network on an auxiliary signal c, written as $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_s, s, c)$, to guide generation. Classifier-free guidance (CFG) (Ho & Salimans, 2022) enables conditional generation without requiring gradients from an external classifier. By training the same denoising network in both conditional and unconditional modes, CFG estimates the noise at step s as $\tilde{\boldsymbol{\epsilon}}_s = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_s, s, \varnothing) + \lambda \cdot [\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_s, s, c) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_s, s, \varnothing)]$, where $\varnothing$ denotes the unconditional input and $\lambda > 0$ controls the guidance strength.

3 Crystal Concepts and Composition

3.1 Concept Extraction

In this section, we extract crystal concepts from the MP-20 training set (Xie et al., 2021), which contains nearly all experimentally stable materials from the Materials Project (Jain et al., 2013) with at most 20 atoms per unit cell. The training set contains 27,136 crystal structures. We train a VQ-VAE (Van Den Oord et al., 2017) on this dataset and treat the learned codebook entries as crystal concepts. The overall pipeline is illustrated in Fig. 1(a).

Given a crystal $\mathbf{C}$ with N atoms, we first construct local atomic environments using the minimum-distance rule and collect at most K neighbors for each atom $i \in \mathbf{C}$. We then concatenate the atom

vectors (Eq. (1)) of atom i and its K neighbors, yielding a matrix $\boldsymbol{T}_i \in \mathbb{R}^{(1+K)\times(3+100+6)}$. Next, we feed all such N local atom-neighbor groups $\{\boldsymbol{T}_i\}$ from crystal $\mathbf{C}$ into the VQ-VAE:

$$\{\boldsymbol{z}_i\} = \texttt{Encoder}(\{\boldsymbol{T}_i\}) \in \mathbb{R}^{N\times d}, \quad \{t_i\} = \arg\min_t ||\boldsymbol{z}_i - \boldsymbol{e}_t||_2^2, \quad \boldsymbol{A}', \boldsymbol{X}', \widehat{\boldsymbol{L}}' = \texttt{Decoder}(\{\boldsymbol{e}_{t_i}\}), \tag{2}$$

where d is the latent dimension, and $\{\boldsymbol{e}_t\}_{t=1}^T$ denotes the VQ-VAE codebook containing T codes, which we interpret as crystal concepts. In our experiments, we set $T = 10{,}000$.

Both `Encoder` and `Decoder` are implemented using vanilla Transformers (Vaswani, 2017), but with different interaction scopes. The `Encoder` operates on each $\boldsymbol{T}_i$ independently, such that each atom interacts only with its own local neighborhood. This design prevents the latent representations $\{\boldsymbol{z}_i\}$ from entangling distinct local patterns. In contrast, the `Decoder` processes all N atoms jointly to reconstruct the full crystal, i.e., $\boldsymbol{A}'$, $\boldsymbol{X}'$, and $\widehat{\boldsymbol{L}}'$. The quantization step in Eq. (2) matches each $\boldsymbol{z}_i$ to its nearest codebook entry $\boldsymbol{e}_{t_i}$, and replaces $\{\boldsymbol{z}_i\}$ with the corresponding concept vectors for reconstruction. The training objective consists of

$$\begin{aligned} &\mathcal{L}_{\boldsymbol{A}} = \frac{1}{N}\sum_i \text{CrossEntropy}(\boldsymbol{A}_i, \boldsymbol{A}'_i), \quad \mathcal{L}_{\boldsymbol{X}} = \frac{1}{N}\sum_i ||\boldsymbol{X}_i - \boldsymbol{X}'_i||_2^2, \\ &\mathcal{L}_{\boldsymbol{L}} = ||\widehat{\boldsymbol{L}} - \widehat{\boldsymbol{L}}'||_2^2, \quad \mathcal{L}_{reg} = \frac{1}{N}\sum_i ||\texttt{sg}[\boldsymbol{z}_i] - \boldsymbol{e}_{t_i}||^2 + ||\boldsymbol{z}_i - \texttt{sg}[\boldsymbol{e}_{t_i}]||^2, \end{aligned} \tag{3}$$

where $\texttt{sg}[\cdot]$ is stop-gradient operation. More training details are given in Appendix B.

3.2 Concept Composition

In this section, we describe how to generate new concept compositions, validate the generated outputs, and use the qualified samples to refine the composition generator. After refinement, the generator is used to sample multiple concept compositions, which then guide the downstream base generative model. These two stages are illustrated in Fig. 1(b).

Training a Composition Generator. To perform concept composition, we introduce a DDPM (Ho et al., 2020) based composition generator $\mathcal{G}$. Specifically, for each crystal $\mathbf{C}$ with N_C atoms in MP-20, we construct its N_C local environments and feed them into the encoder to obtain $\boldsymbol{Z}_C \in \mathbb{R}^{N_C\times d}$. The generator $\mathcal{G}$ is trained to learn a mapping from Gaussian noise to the set of pre-quantization latent matrices $\{\boldsymbol{Z}_C\}$ of all crystals in MP-20. After training, $\mathcal{G}$ generates new latent matrices $\{\boldsymbol{Z}^g\}$, which are then quantized using Eq. (2) to obtain $\{\boldsymbol{E}^g\}$. Each matrix $\boldsymbol{E}^g = [\boldsymbol{e}_{t_1}, \ldots, \boldsymbol{e}_{t_{N_g}}]^\top \in \mathbb{R}^{N_g\times d}$ represents a composition of crystal concepts, where N_g is the number of atoms.

Validating and Refining the Generator. Among the outputs from $\mathcal{G}$, only valid concept compositions can support meaningful downstream crystal generation. For example, Wu et al. (2024) presented an invalid composition in text-to-image generation such as "a triangle-shaped person." We therefore need criteria for filtering high-quality compositions from the generated candidates.

In materials discovery, four criteria are particularly important in practice: validity (V), stability (S), uniqueness (U), and novelty (N). Among them, validity and uniqueness can be checked directly from the generated crystals themselves, whereas stability and novelty require comparison against a set of reference crystals. However, during model training, we avoid relying on any external references for two reasons: (1) there is no single unified reference set across all experimental settings, and (2) using external data during training would lead to an unfair comparison with other baselines. Based on these considerations, we use only the training crystals as the reference set for evaluating stability and novelty during refinement. Our underlying assumption is that these training crystals are stable regardless of the final evaluation reference. Hence, a generated crystal with comparable formation energy under the same chemical system is likely to be stable as well.

After training the composition generator $\mathcal{G}$, we use it to produce a pool of new compositions $\{\boldsymbol{E}^g\}$, and retain only those whose decoded crystals satisfy the V.S.U.N. criteria with respect to the training set. We then fine-tune $\mathcal{G}$ on these validated compositions as a refinement step, following the same training protocol as the previous stage. This refinement explicitly optimizes proxies aligned with the final V.S.U.N. criteria but only against the training set. The external reference used for final

evaluation is never accessed at this step. Thus, the procedure is metric-aligned self-training, but not evaluation-set leakage. The ablation study for refinement is given in Appendix A.2.

Composition-based Generative Model. Once refined, the generator $\mathcal{G}$ produces new compositions $\{\boldsymbol{E}^g\}$ that enable the downstream generative model to explore novel crystal structures in a controllable manner. In this work, we implement the composition-based model using a Denoising Diffusion Probabilistic Model (DDPM) with a vanilla Transformer architecture conditioned on these compositions. To validate the utility of the extracted concepts, we evaluate whether conditioning on the final $\{\boldsymbol{E}^g\}$ enhances generation quality relative to both unconditional models and baselines.

3.3 Transformers as Backbones

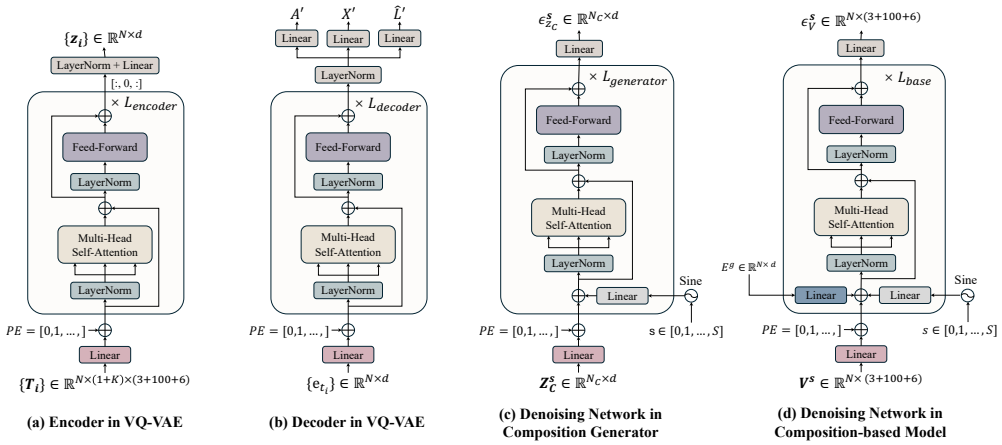


Figure 2: Transformer architectures used in (a) the encoder and (b) the decoder of the VQ-VAE, as well as the denoising networks in (c) the composition generator and (d) the composition-based model.

Throughout concept extraction and composition, we use backbone models as the encoder and decoder in the VQ-VAE, and as the denoising networks in both the composition generator and the base generative model. For all four cases, we adopt the vanilla Transformer architecture (Vaswani, 2017) as our backbone, incorporating several key modifications. The detailed designs are shown in Fig. 2, and a more detailed introduction can be found in Appendix C.2. The first modification is the use of pre-normalization before multi-head attention and feed-forward layers (Xiong et al., 2020). We also introduce a simple positional encoding scheme and a condition incorporation strategy.

Positional Encoding (PE). There is no widely accepted canonical ordering of atoms in a crystal. However, as sequence-based models, Transformers require an ordering scheme to function effectively. In this paper, we order atoms according to two principles: (1) atoms with smaller atomic numbers are placed before those with larger atomic numbers, and (2) among atoms of the same type, we compare their coordinates sequentially along the x-, y-, and z-axes, placing atoms with smaller coordinates first. After this ordering procedure, each atom is assigned a learnable position vector according to its order. These position vectors are shared across all crystals.

Condition Incorporation. In the denoising networks, additional conditions are incorporated as part of the input, including the diffusion timestep s and the generated concept composition $\boldsymbol{E}^g$. For timestep s, we apply a linear layer after sinusoidal encoding to obtain a time embedding. For composition $\boldsymbol{E}^g$, we use another linear layer to obtain a composition embedding. These condition embeddings are then added to the input at the beginning of each Transformer layer.

3.4 Why learned discrete local concepts for composition?

Existing representations and conditioning signals provide useful structural information, but are less suited to the operations required by concept composition.

• *Continuous latent representations*. Continuous representations (e.g., VAEs (Kingma & Welling, 2013)) can encode rich structural information, but do not directly form a finite vocabulary of reusable units. Quantization assigns each local atomic environment to a discrete code, enabling explicit identification and recombination. We therefore use discreteness as an operational interface for composition rather than claiming that discrete representations are universally superior.

• *Physical-property conditioning*. Physical properties such as band gap specify desired material outcomes rather than reusable structural units that can be explicitly recombined. Some properties may also require expensive calculations such as density functional theory for verification. In contrast, our concepts represent recurring local structural patterns, and their presence in generated structures can be assessed by re-encoding those structures.

• *Explicit structural constraints*. Predefined patterns such as space groups (Levy et al., 2025), Archimedean lattices (Okabe et al., 2026), or motifs (Wei et al., 2026) are generally global constraints and are therefore difficult to freely combine within a single crystal. For example, a crystal cannot generally belong to multiple incompatible space groups. Our concepts are instead learned at the level of local atomic environments, allowing multiple concepts to coexist within a structure and enabling a larger compositional space.

• *Thermodynamic and configurational representations*. Cluster expansion (Sanchez et al., 1984) and CALPHAD (Saunders & Miodownik, 1998) are designed to model configurational energetics and thermodynamic behavior. Our objective is different, where we learn recurring local structural patterns and use them as reusable conditions for de novo generation.

Overall, our learned discrete local concepts provide a data-driven interface for automatically discovering structural units, explicitly reusing them, and composing them within a crystal.

4 UNDERSTANDING LEARNED CRYSTAL CONCEPTS

After training the VQ-VAE, we analyze the optimized 10,000 discrete codes as crystal concepts, since each code represents a cluster of similar local atomic environments. In this section, we analyze these learned concepts from both local and global perspectives.

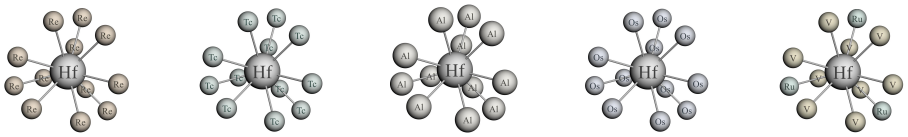


Figure 3: Five representative local atomic environments assigned to the same VQ-VAE code.

Local view. We first examine what local information is encoded by the learned concepts. Specifically, for a given code $\boldsymbol{e}_t$, we collect all local environments $\{\boldsymbol{T}_i\} \in \mathbb{R}^{(1+K)\times(3+100+6)}$ whose latent representations $\{\boldsymbol{z}_i\}$ are assigned to $\boldsymbol{e}_t$ during the quantization step in Eq. (2). We then compute the distances $\{d_i^{\boldsymbol{e}_t} = \|\boldsymbol{z}_i - \boldsymbol{e}_t\|_2^2\}$ and retrieve the top-5 local environments with the smallest distances, i.e., those that best match concept $\boldsymbol{e}_t$. Finally, we visualize their atomic coordinates and species. One example is shown in Fig. 3, from which we draw the following observations:

- The most representative local environments of the same concept can originate from different chemical systems. However, they exhibit similar 3D structures, suggesting that the learned concepts primarily capture geometric patterns.
- Elements with similar physical properties tend to share the same concepts. In Fig. 3, Re and Tc belong to the same group in the periodic table, as do Ru and Os; these elements appear as neighbors of the central atom in local environments assigned to the same concept.

It should be noted that not all of the 10,000 learned concepts exhibit such visually distinct patterns. This is anticipated, as the concepts are learned in an unsupervised manner without the imposition of human-defined chemical or structural priors. Consequently, the interpretable patterns that emerge from the learned codebook serve as compelling evidence that the model captures non-trivial regularities within local crystal environments. We provide additional local-view visualizations in Appendix A.1.

Global view. Space groups are commonly used to describe the 230 possible types of global symmetry in crystals. These groups can be further categorized into six crystal families. Table 1 summarizes the relationship between space groups and crystal families. In this part, we investigate how the learned local concepts relate to these global structural properties.

Table 1: Lattice shape and included space groups of crystal families (Jiao et al., 2024), where a, b, c and α, β, γ denote the lengths and angles of the lattice base, respectively.

Crystal Family	Triclinic	Monoclinic	Orthorhombic	Tetragonal	Cubic	Hexagonal
Space Group No.	1∼2	3∼15	16∼74	75∼142	195∼230	143∼194
Lattice Shape	No Constraint	$\alpha = \gamma = 90°$	$\alpha = \beta = \gamma = 90°$	$\alpha = \beta = \gamma = 90°$ $a = b$	$\alpha = \beta = \gamma = 90°$ $a = b = c$	$\alpha = \beta = 90°$, $\gamma = 120°, a = b$

a. Crystal Families Select Concepts. We first categorize all MP-20 crystals according to their crystal families. For each family $\mathcal{F}$, we construct a probability vector $\boldsymbol{p}^{\mathcal{F}} = (p^{\mathcal{F}}_{e_1}, \cdots, p^{\mathcal{F}}_{e_T}) \in \mathbb{R}^T$, where $p^{\mathcal{F}}_{e_j}$ denotes the frequency with which the j-th code is selected among all crystals in family $\mathcal{F}$. We then compute pairwise cosine similarities among the vectors $\{\boldsymbol{p}^{\mathcal{F}}\}$, as shown in Fig. 4. In the figure, we exclude the comparison against `Triclinic` family, since the Triclinic family does not exhibit an explicit lattice constraint in Table 1. From the figure, we make two observations:

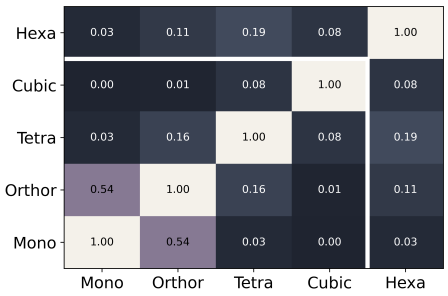


Figure 4: Pairwise cosine similarities between concept-usage distributions across five crystal families.

(1) Table 1 shows that `Cubic` → `Tetragonal` → `Orthorhombic` → `Monoclinic` forms a symmetry-reduction hierarchy, where moving downward corresponds to progressively relaxed lattice constraints. This hierarchy is well aligned with Fig. 4, in which adjacent families exhibit higher similarity in concept selection. For example, `Orthorhombic` has higher similarity to `Monoclinic` and `Tetragonal`, and these three families are adjacent in the hierarchy.

(2) Although `Hexagonal` is not part of this chain, its lattice differs from `Tetragonal` mainly in the γ angle. Consequently, `Hexagonal` crystals exhibit concept preference distributions that are more similar to those of `Tetragonal` crystals than to those of other families in Fig. 4.

b. Concepts Predict Groups And Families. Given only the selected concepts of a crystal, we predict its space group and crystal family to further strengthen their connection. Specifically, for a crystal $\mathcal{C}$ with N selected concepts $\{\boldsymbol{e}_{t_1}, \ldots, \boldsymbol{e}_{t_N}\}$, we concatenate these concept embeddings into a matrix $\boldsymbol{E} = [\boldsymbol{e}_{t_1}, \cdots, \boldsymbol{e}_{t_N}]^\top$. We then feed E into an eight-layer Transformer with hidden dimension 512. The output representations are aggregated by mean pooling and passed through a two-layer multi-layer perceptron to separately predict 230 space groups and 6 crystal families. As shown in Table 2, the selected concepts alone achieve 67.11% accuracy for space-group prediction and 77.37% accuracy for crystal-family prediction. These results indicate that although the concepts are learned from local atomic environments, their compositions retain substantial information about global crystal symmetry.

Table 2: Prediction on space groups and crystal families based on concepts.

%	Space Group	Crystal Family
Accuracy	67.11	77.37

5 Concept Generalization

In this section, we evaluate the extent to which concept composition enhances generative performance, with a specific focus on structural novelty. We further assess the cross-dataset transferability of concepts learned from MP-20 to Alex-MP-20 (Zeni et al., 2023), a dataset characterized by a distinct distribution of crystal structures. Implementation details are provided in Appendix C.2.

5.1 BENCHMARK ON MP-20

Table 3: Evaluation of 10,000 generated crystals after training on MP-20. Reference is MP-2023. († means the authors did not provide model checkpoints. We reproduce their results.)

%	Validity (V)	Stability (S)	Uniqueness (U)	Novelty (N)	S.U.N (MatterSim)	V.S.U.N (MatterSim)	S.U.N (DFT)	V.S.U.N (DFT)
CDVAE† (Xie et al., 2021)	85.3±1.3	29.9±1.2	99.8±0.2	96.5±0.6	27.0±1.4	22.8±1.1	-	-
DiffCSP (Jiao et al., 2023)	83.1±1.3	45.9±1.8	99.5±0.1	83.6±0.6	30.9±1.9	25.6±1.3	-	-
DiffCSP++ (Jiao et al., 2024)	85.0±0.9	39.7±2.0	99.8±0.1	82.8±1.0	23.9±1.3	20.3±1.2	-	-
FlowMM† (Miller et al., 2024)	81.8±0.8	40.8±2.0	99.8±0.0	83.1±1.1	25.3±1.7	21.0±1.3	-	-
FlowLLM† (Sriram et al., 2024)	83.8±0.7	36.5±1.5	98.2±0.5	86.4±1.5	25.1±0.6	21.3±0.8	-	-
SymmCD† (Levy et al., 2025)	75.5±1.5	34.7±1.4	99.7±0.1	85.1±1.5	19.0±1.0	15.7±0.8	-	-
ADiT (Joshi et al., 2025)	89.7±1.0	69.5±1.1	99.1±0.2	58.9±0.9	30.3±0.9	27.2±1.3	-	-
CrysLLMGen (7B)† (Khastagir et al., 2025)	88.1±0.9	35.1±1.9	98.4±0.3	87.4±1.1	22.9±0.9	20.4±0.9	-	-
SGEquiDiff (Chang et al., 2025)	85.1±1.0	46.5±1.9	98.9±0.4	74.8±1.2	23.6±0.9	20.0±1.0	-	-
MatterGen-MP (Zeni et al., 2023)	84.2±1.2	47.0±1.1	**99.8±0.1**	86.6±1.0	34.7±1.3	29.2±0.9	30.8	26.4
Base model	82.1±1.5	**48.4±1.4**	99.6±0.0	78.0±1.6	27.8±1.1	23.3±1.3	-	-
w/ composition	**93.9±0.4**	46.9±1.0	98.8±0.3	**90.0±0.5**	**36.4±1.5**	**35.7±1.7**	**31.8**	**31.1**

Evaluation Setting. We repeat sampling 10 times and collect 1,000 crystals in each run. We then report the mean and standard deviation of `validity (V)`, `stability (S)`, `uniqueness (U)`, and `novelty (N)`. We also report `S.U.N` and `V.S.U.N`. During evaluation, structure relaxation and energy prediction are performed using MatterSim-v1-1M (Yang et al., 2024). Following prior work (Miller et al., 2024), we use MP-2023 as the reference dataset and regard crystals with energy above hull below 0.1 eV/atom as stable. We further perform density functional theory (DFT) calculations (Perdew et al., 1996) to provide a more rigorous stability assessment, with details provided in Appendix C.5. Due to the high computational cost, we only perform DFT on 1,000 crystals sampled from both our model and the strongest baseline.

Baselines. We train our composition-based generative model (DDPM + Transformer) on MP-20, and further condition it on the compositions generated by $\mathcal{G}$. We compare these two variants against state-of-the-art baselines, including CDVAE, DiffCSP, DiffCSP++, FlowMM, FlowLLM, SymmCD, ADiT, CrysLLMGen (7B), SGEquiDiff and MatterGen.

Results. As shown in Table 3, incorporating concept compositions makes our composition-based model the best-performing method, achieving improvements of 17.8% in `V.S.U.N` using DFT over MatterGen, the strongest baseline. In particular, concept composition leads to a modest decrease in stability while substantially improving novelty, resulting in a higher joint V.S.U.N. rate. These results show compositionality can effectively guide structured exploration beyond the training distribution.

5.2 GENERALIZATION TO ALEX-MP-20

VQ-VAE reconstruction. After training the VQ-VAE, we freeze the learned codebook and pass the crystals from Alex-MP-20 through the model, then evaluate whether each reconstructed crystal matches the input. We find that the concepts learned on MP-20 can successfully reconstruct 61% of Alex-MP-20 crystals, indicating that the concepts can represent unseen crystals from another distribution to a meaningful extent.

Generation on Alex-MP-20. We repeat the same experimental setting for MatterGen and for our composition-based model trained on Alex-MP-20. Notably, the compositions are still formed by recombining concepts learned from MP-20, while their validation is performed using the Alex-MP-20 training set as the reference. As shown in Table 4, incorporating concept compositions leads to a decrease in stability. This reduction may suggest that concepts learned from MP-20 do not fully capture the newly emerging patterns in Alex-MP-20. Nevertheless, our primary metrics of interest, `V.S.U.N.`, still improve by 14.5% using DFT, due to the substantial gain in novelty brought by compositional generation. We further discuss this novelty–stability tradeoff in Appendix A.3.

Table 4: Evaluation of 10,000 generated crystals after training on Alex-MP-20. Reference is MP2020correction provided by MatterGen (Zeni et al., 2023).

%	Validity (V)	Stability (S)	Uniqueness (U)	Novelty (N)	S.U.N (MatterSim)	V.S.U.N (MatterSim)	S.U.N (DFT)	V.S.U.N (DFT)
MatterGen	85.7±1.2	**73.1±1.0**	**99.9±0.0**	67.9±1.4	43.6±1.2	37.4±1.4	40.1	35.1
Base model	87.4±0.9	69.0±0.9	**99.9±0.0**	61.8±1.2	33.3±1.4	28.8±1.0	-	-
w/ composition	**94.8±0.8**	59.1±1.6	99.4±0.1	**85.9±0.9**	**46.9±1.4**	**43.7±1.5**	**40.6**	**40.2**

6 Concept Verification

Finally, we assess whether the learned concepts capture coherent local patterns and whether generated structures recover their conditioning concepts through two complementary automated analyses.

a. Concept-to-structure coherence. We analyze the 1,000 most frequently activated concepts. For each concept, we collect all assigned local atomic environments, compute CrystalNN (Zimmermann & Jain, 2020) and normalized Magpie (Ward et al., 2016) fingerprints, and measure the mean distance to the concept-specific centroid. We compare this dispersion with 1,000 equally sized random groups. With CrystalNN fingerprints, 80.9% of the concepts are more compact than random groups and 65.3% achieve nominal significance at $p < 0.05$. With Magpie fingerprints, the corresponding values are 83.6% and 69.1%. Thus, among frequently used codes, most group coherent local structural and compositional patterns rather than arbitrary environments.

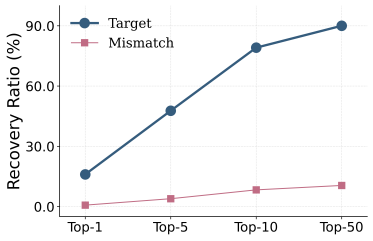


Figure 5: Concept recovery under target and mismatched conditioning compositions.

b. Structure-to-concept recovery. We re-encode generated crystals and compare the recovered codes with their conditioning concepts. As a baseline, we randomly exchange target compositions between crystals with the same number of atoms. The large gap in Fig. 5 shows that the generated local environments preferentially recover the intended concepts.

More details can be found in Appendix C.4. Fig. 6 further shows one generated crystal together with its supporting concepts. The results demonstrate that the base model can effectively respond to the extracted concepts and incorporate the corresponding local patterns into generated structures. This verification provides two complementary perspectives on concept interpretability. One is *coherence* that local environments assigned to the same code should be similar in the aspect of structure and composition. Another is *faithfulness* that conditioning on a code should measurably increase the probability that the corresponding local pattern appears in the generated crystal. However, it does not mean that the code provides a complete physical explanation by itself.

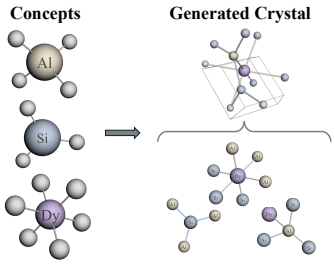


Figure 6: One generated crystal together with its concepts.

7 Conclusion

We introduced a compositional framework for de novo crystal generation based on learned discrete local concepts. These concepts can be automatically extracted, recombined within a crystal, and re-extracted from generated structures, providing a reusable interface for structured exploration beyond the training data. Experiments on MP-20 and Alex-MP-20 show consistent improvements in V.S.U.N., mainly driven by increased novelty, while concept analysis and cross-dataset transfer further support the interpretability and generalizability of the learned vocabulary. The observed novelty–stability tradeoff also suggests that better modeling concept compatibility is an important direction for future work.

AI USE STATEMENT

Generative AI tools were used only to improve the readability, grammar, and wording of the manuscript. They were not used for research ideation, methodology design, experiments, data analysis, result interpretation, or code generation. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the datasets, model architectures, training procedures, hyperparameters, evaluation metrics, and computational environment in Appendices B and C. The settings for concept verification, stability evaluation, and DFT calculations are also documented, and the implementation used in our experiments is publicly available.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Deyu Bo, Chuan Shi, Lele Wang, and Renjie Liao. Specformer: Spectral graph neural networks meet transformers. *arXiv preprint arXiv:2303.01028*, 2023.

Zhendong Cao, Xiaoshan Luo, Jian Lv, and Lei Wang. Space group informed transformer for crystalline materials generation, 2024. *URL https://arxiv. org/abs/2403.15734*, 11.

Rees Chang, Angela Pak, Alex Guerra, Ni Zhan, Nick Richardson, Elif Ertekin, and Ryan P Adams. Space group equivariant crystal diffusion. *arXiv preprint arXiv:2505.10994*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*, 2020.

Futoshi Futami and Masahiro Fujisawa. Information-theoretic generalization analysis for vq-vaes: A role of latent variables. *Advances in Neural Information Processing Systems*, 38:49653–49704, 2026.

Alex M. Ganose, Hrushikesh Sahasrabuddhe, Mark Asta, Kevin Beck, Tathagata Biswas, Alexander Bonkowski, Joana Bustamante, Xin Chen, Yuan Chiang, Daryl Chrzan, Jacob Clary, Orion Cohen, Christina Ertural, Max Gallant, Janine George, Sophie Gerits, Rhys Goodall, Rishabh Guha, Geoffroy Hautier, Matthew Horton, Aaron Kaplan, Ryan Kingsbury, Matthew Kuner, Bryant Li, Xavier Linn, Matthew McDermott, Rohith Srinivaas Mohanakrishnan, Aakash Naik, Jeffrey Neaton, Kristin Persson, Guido Petretto, Thomas Purcell, Francesco Ricci, Benjamin Rich, Janosh Riebesell, Gian-Marco Rignanese, Andrew Rosen, Matthias Scheffler, Jonathan Schmidt, Jimmy-Xuan Shen, Andrei Sobolev, Ravishankar Sundararaman, Cooper Tezak, Victor Trinquet, Joel Varley, Derek Vigil-Fowler, Duo Wang, David Waroquiers, Mingjian Wen, Han Yang, Hui Zheng, Jiongzhi Zheng, Zhuoying Zhu, and Anubhav Jain. Atomate2: Modular Workflows for Materials Science. *ChemRxiv*, 2025. URL `https://chemrxiv.org/engage/chemrxiv/article-details/678e76a16dde43c9085c75e9`.

N Gruver, A Sriram, A Madotto, AG Wilson, CL Zitnick, and Z Ulissi. Fine-tuned language models generate stable inorganic materials as text, arxiv, 2024. *arXiv preprint arXiv:2402.04379*, 10.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

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

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information Processing Systems*, 36:17464–17497, 2023.

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

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. *arXiv preprint arXiv:2503.03965*, 2025.

Subhojyoti Khastagir, Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Llm meets diffusion: A hybrid framework for crystal material generation. *arXiv preprint arXiv:2510.23040*, 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54:11169–11186, Oct 1996. doi: 10.1103/PhysRevB.54.11169. URL https://link.aps.org/doi/10.1103/PhysRevB.54.11169.

G. Kresse and D. Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B*, 59:1758–1775, Jan 1999. doi: 10.1103/PhysRevB.59.1758. URL https://link.aps.org/doi/10.1103/PhysRevB.59.1758.

Daniel Levy, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Qiang Zhu, Kin Long Kelvin Lee, Mikhail Galkin, Santiago Miret, and Siamak Ravanbakhsh. Symmcd: Symmetry-preserving crystal generation with diffusion models. *arXiv preprint arXiv:2502.03638*, 2025.

Benjamin Kurt Miller, Ricky TQ Chen, Anuroop Sriram, and Brandon M Wood. Flowmm: Generating materials with riemannian flow matching. In *Forty-first International Conference on Machine Learning*, 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Ryotaro Okabe, Mouyang Cheng, Abhijatmedhi Chotrattanapituk, Manasi Mandal, Kiran Mak, Denisse Córdova Carrizales, Nguyen Tuan Hung, Xiang Fu, Bowen Han, Yao Wang, et al. Structural constraint integration in a generative model for the discovery of quantum materials. *Nature Materials*, 25(2):223–230, 2026.

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

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.

John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868, Oct 1996. doi: 10.1103/PhysRevLett.77.3865. URL https://link.aps.org/doi/10.1103/PhysRevLett.77.3865.

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.

Juan M Sanchez, Francois Ducastelle, and Denis Gratias. Generalized cluster description of multi-component systems. *Physica A: Statistical Mechanics and its Applications*, 128(1-2):334–350, 1984.

Nigel Saunders and A Peter Miodownik. *CALPHAD (calculation of phase diagrams): a comprehensive guide*, volume 1. Elsevier, 1998.

Jonathan Schmidt, Tiago FT Cerqueira, Aldo H Romero, Antoine Loew, Fabian Jäger, Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Improving machine-learning models in materials science through large datasets. *Materials Today Physics*, 48:101560, 2024.

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.

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

Yuan Sui, Yanming Zhang, Yi Liao, Yu Gu, Guohua Tang, Zhongqian Sun, Wei Yang, and Bryan Hooi. What-if analysis of large language models: Explore the game world using proactive thinking. *arXiv preprint arXiv:2509.04791*, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):16028, 2016.

Yihao Wei, Ivan Savochkin, Artem Mishchenko, Xiangwen Wang, and Qian Yang. Symmetry-aware generative design of flat-band materials beyond known crystal-net prototypes. *arXiv preprint arXiv:2606.15321*, 2026.

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty. *Advances in Neural Information Processing Systems*, 37:86004–86047, 2024.

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.

Yujie Xing, Xiao Wang, Yibo Li, Hai Huang, and Chuan Shi. Less is more: on the over-globalizing problem in graph transformers. *arXiv preprint arXiv:2405.01102*, 2024.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International conference on machine learning*, pp. 10524–10533. PMLR, 2020.

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.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, et al. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024.

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.

Zhongjian Zhang, Xiao Wang, Huichi Zhou, Yue Yu, Mengmei Zhang, Cheng Yang, and Chuan Shi. Can large language models improve the adversarial robustness of graph neural networks? In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2008–2019, 2025.

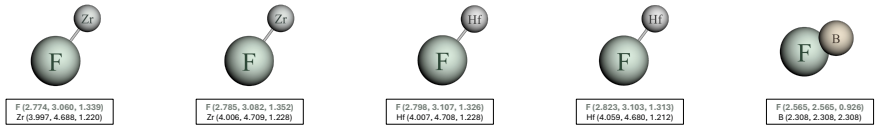
Baoheng Zhu, Deyu Bo, Delvin Ce Zhang, and Xiao Wang. Graph-grpo: Training graph flow models with reinforcement learning. *arXiv preprint arXiv:2603.10395*, 2026.

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

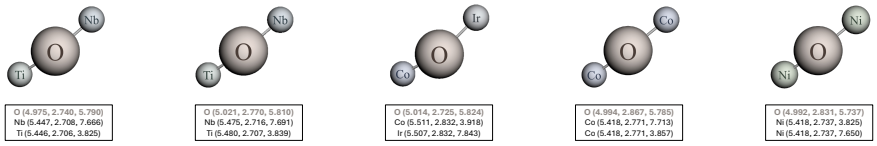
A SUPPLEMENTARY EXPERIMENTS

A.1 VISUALIZING MORE LOCAL PATTERNS ENCODED IN CONCEPTS

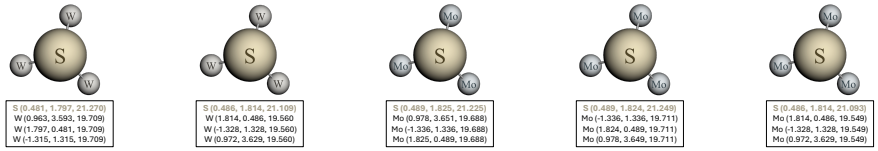
In Section 4, we visualize the five most representative local atomic environments associated with one learned concept. Here, we provide additional examples with different numbers of neighboring atoms.



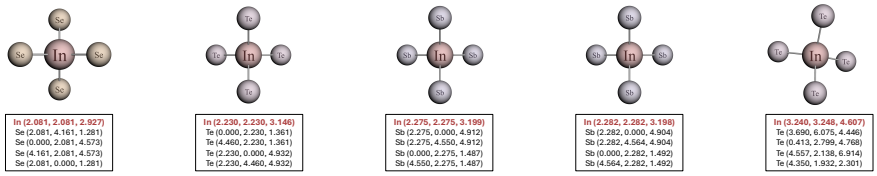
`Zr` and `Hf` are group-IV transition metals.



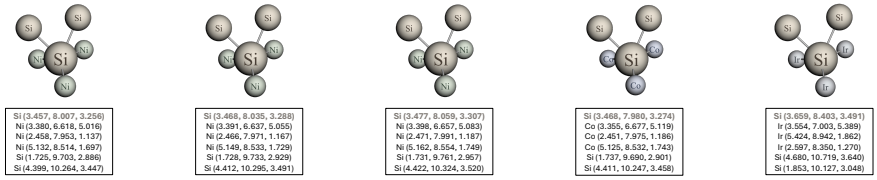
`Co, Ir, Ni, Nb`, and `Ti` are transition metals, with `Co–Ir–Ni` and `Ti–Nb` reflecting chemically related groups.



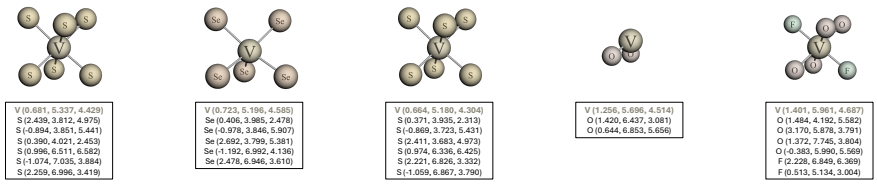
`W` and `Mo` are group-VI transition metals.



`Se` and `Te` are group-VI chalcogens, while `Sb` is adjacent in the periodic table.

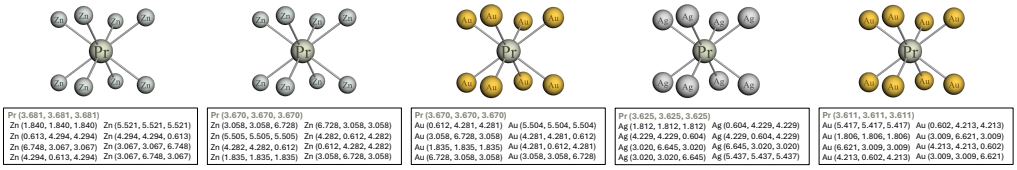


Element `Co` and `Ir` belong to the same group. `Co, Ir`, and `Ni` are transition metals from neighboring groups.

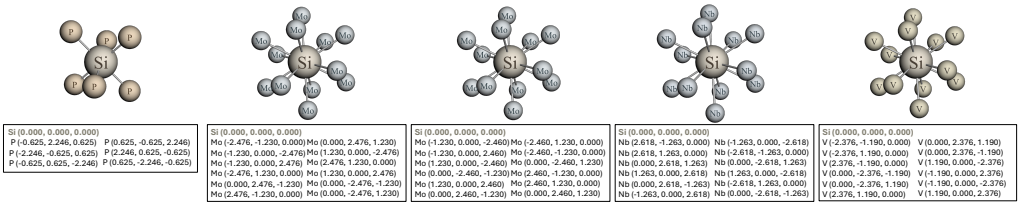


`O, S`, and `Se` are group-VI chalcogens, while `F` belongs to the neighboring halogen group.

These examples show that the learned concepts encode meaningful patterns that combine both structural and elemental similarities. In particular, local environments assigned to the same concept



Au and Ag are group-XI transition metals, while Zn belongs to the adjacent group.



Nb and V are group-V transition metals, while Mo belongs to the adjacent group.

often exhibit similar geometric arrangements, even when they originate from different chemical systems. Moreover, chemically related elements, such as elements from the same group or adjacent groups in the periodic table, frequently appear within the same concept clusters.

A.2 Ablation Study of the Refinement Step

In this section, we isolate the contribution of the refinement step. In Table 5, we report the proposed concept-composition model without refinement, as well as DiffCSP with the same refinement budget as our method. We choose DiffCSP since its original V.S.U.N. score is close to that of our base model.

Table 5: Ablation study of the refinement step on MP-20, evaluated using MatterSim.

%	V	S	U	N	S.U.N.	V.S.U.N.
DiffCSP	83.1	45.9	99.5	83.6	30.9	25.6
DiffCSP w/ refinement	93.0	50.5	95.8	84.3	35.2	34.4
Base model w/o composition	82.1	48.4	99.6	78.0	27.8	23.3
Composition w/o refinement	82.7	43.1	99.8	88.1	28.3	24.0
Composition w/ refinement	93.9	46.9	98.8	90.0	36.4	35.7

The ablation separates the two effects. Composition without refinement increases novelty from 78.0% to 88.1%, but stability decreases from 48.4% to 43.1%, so the V.S.U.N. gain is small. Refinement then raises validity from 82.7% to 93.9% and stability from 43.1% to 46.9%, producing a final V.S.U.N. rate of 35.7%. Applying V.S.U.N.-based filtering to DiffCSP also improves its score, from 25.6% to 34.4%, confirming that selection itself is useful. However, it changes DiffCSP novelty only from 83.6% to 84.3%, whereas composition alone reaches 88.1%. This supports our interpretation: compositional conditioning drives the novelty gain, while refinement mainly improves candidate quality.

A.3 Stability-novelty tradeoff shown in Alex-MP-20

As shown in Table. 4, the Alex-MP-20 results suggest a clear novelty-stability tradeoff: novelty increases from 61.8 to 85.9, while stability drops from 69.0 to 59.1 under MatterSim. The filtering threshold δ in refinement step and guidance strength λ in CFG control this balance. A larger δ admits more potentially unstable samples into refinement, while a smaller λ weakens guidance toward the validated concept compositions. In this section, we quantify how the tradeoff changes with these two hyper-parameters. We sweep the filtering threshold δ while fixing $\lambda = 2.0$, and sweep the guidance strength λ while fixing $\delta = 0.05$. The results are reported in Table 6.

Table 6: Quantify the tradeoff changes with the repsect of δ and λ.

%	V	S	U	N	S.U.N.	V.S.U.N.
$\delta = 0.01$	92.9	63.2	97.8	76.6	38.8	37.1
$\delta = 0.03$	92.7	61.9	99.4	83.5	46.5	42.9
$\delta = 0.05$	94.8	59.1	99.4	85.9	46.9	43.7
$\delta = 0.08$	93.1	54.9	99.6	90.0	45.1	42.9
$\lambda = 1.0$	95.0	58.6	99.4	86.5	44.9	43.6
$\lambda = 2.0$	94.8	59.1	99.4	85.9	46.9	43.7
$\lambda = 3.0$	92.8	60.2	99.5	82.5	42.5	40.7

The δ sweep shows the trade-off clearly: tightening the threshold from 0.08 to 0.01 raises stability from 54.9% to 63.2%, but lowers novelty from 90.0% to 76.6%. The best V.S.U.N. rate occurs at the intermediate value $\delta = 0.05$. Guidance has a similar effect at high strength: increasing λ from 2.0 to 3.0 raises stability slightly, from 59.1% to 60.2%, but reduces novelty from 85.9% to 82.5% and lowers V.S.U.N. from 43.7% to 40.7%. These results show that the exploration–stability balance is tunable, while also confirming that distribution shift remains a real limitation rather than being fully eliminated by guidance.

B VQ-VAE Training

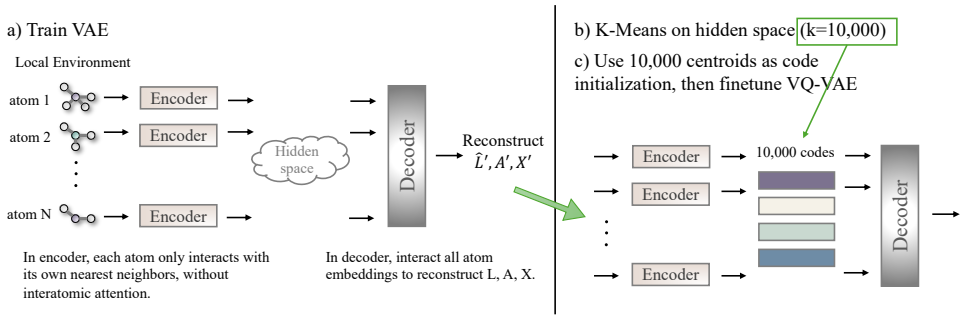


Figure 7: Three-stage training pipeline for the VQ-VAE.

Directly training a VQ-VAE on crystal structures led to limited reconstruction performance in our preliminary experiments. Therefore, we adopt a three-stage training strategy. We first train a variational autoencoder (VAE) to learn a continuous latent space for local atomic environments. We then apply K-Means clustering to the learned latent embeddings to obtain discrete centroids. Finally, we initialize the VQ-VAE codebook with these centroids and fine-tune the full VQ-VAE model. The whole pipeline is shown in Fig. 7.

Training the VAE. As described in Section 3.1, we extract a local atomic environment $\boldsymbol{T}_i$ for each atom i and feed these environments into the VAE:

$$\begin{aligned} \{\mu_{\boldsymbol{z}_i}\}, \{\log \sigma_{\boldsymbol{z}_i}\} &= \texttt{Encoder}(\{\boldsymbol{T}_i\}) \in \mathbb{R}^{N\times d}, \mathbb{R}^{N\times d}, \\ \boldsymbol{z}_i &= \mu_{\boldsymbol{z}_i} + \sigma_{\boldsymbol{z}_i} \odot \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(0,1)^{N\times d}, \\ \boldsymbol{A}', \boldsymbol{X}', \widehat{\boldsymbol{L}}' &= \texttt{Decoder}(\{\boldsymbol{z}_i\}). \end{aligned} \tag{4}$$

Both `Encoder` and `Decoder` use the Transformer architectures shown in Fig. 2(a) and (b), respectively. The training objective consists of the following terms:

$$\begin{aligned} &\mathcal{L}_{\boldsymbol{A}} = \frac{1}{N}\sum_i \text{CrossEntropy}(\boldsymbol{A}_i, \boldsymbol{A}'_i), \quad \mathcal{L}_{\boldsymbol{X}} = \frac{1}{N}\sum_i ||\boldsymbol{X}_i - \boldsymbol{X}'_i||_2^2, \\ &\mathcal{L}_{\boldsymbol{L}} = ||\widehat{\boldsymbol{L}} - \widehat{\boldsymbol{L}}'||_2^2, \quad \mathcal{L}_{reg} = \texttt{KL}(\mathcal{N}(\boldsymbol{z}_i; \mu_{\boldsymbol{z}_i}, \sigma_{\boldsymbol{z}_i}))||\mathcal{N}(0,1)) \end{aligned} \tag{5}$$

K-Means Initialization. After training the VAE, we collect the latent embeddings of all local atomic environments from the MP-20 training set, resulting in 279,608 latent vectors. We then apply K-Means clustering using the Faiss library (Douze et al., 2024), with the number of clusters set to $T = 10{,}000$. The resulting cluster centroids are used to initialize the VQ-VAE codebook.

Fine-tuning the VQ-VAE. We then fine-tune a VQ-VAE by replacing the continuous latent variables of the VAE with a discrete learnable codebook. During quantization, each local atomic environment is assigned to its nearest codebook entry, and the selected code is used as a reusable structural concept. The `Encoder` and `Decoder` are initialized from the pretrained VAE, while the codebook is initialized using the K-Means centroids. The VQ-VAE pipeline and training objective are given in Section 3.1.

The role of VQ-VAE. Some studies (Futami & Fujisawa, 2026) proved that VQ-VAE has irreducible generalization errors from the perspective of information theory. However, the role of our VQ-VAE is to produce concepts, which are not expected to predict or uniquely determine an unseen crystal. Instead, they serve as reusable structural condition. This conditioning is inherently one-to-many: the same concept may correspond to multiple valid crystals with different compositions, lattices, and atomic coordinates. The downstream generative model is still responsible for producing a complete and globally consistent structure. Therefore, the relevant question is not whether the concept representation can perfectly reconstruct every unseen sample, but whether it provides a transferable structural bias for generating diverse valid crystals.

C Experimental Details

C.1 Datasets Descriptions

We conduct experiments on two datasets, MP-20 and Alex-MP-20.

- **MP-20** is a standard benchmark dataset curated from the Materials Project (Jain et al., 2013) and introduced by CDVAE (Xie et al., 2021). It contains 45,231 inorganic crystals with no more than 20 atoms per unit cell and spans 89 chemical elements. We follow the standard 60/20/20 train/validation/test split used in CDVAE. Since most MP-20 structures are experimentally known and globally stable, the dataset serves as a realistic benchmark for de novo crystal generation.
- **Alex-MP-20** is a larger inorganic crystal dataset introduced by MatterGen (Zeni et al., 2023), constructed from the Alexandria database (Schmidt et al., 2024) and the Materials Project (Jain et al., 2013). It contains 607,684 structures with at most 20 atoms per unit cell. The dataset is further filtered to retain stable or near-stable materials, typically with DFT-computed energy above hull below 0.1 eV/atom, and structures containing radioactive elements are excluded.

C.2 Implementation Details

The encoder and decoder in the VAE/VQ-VAE are implemented as 8-layer Transformers with #hidden_dimension=512, #latent_dimension=8, #attention_heads=8, #dropout=0. The maximum number of neighboring atoms is set to $\#K = 12$. The regularization weights are set to 10^{-5} for the VAE and 10^{-2} for the VQ-VAE.

The composition generator $\mathcal{G}$ uses a DDPM whose denoising network is a Transformer with 12-layer=12, #hidden_dimension=768 and #attention_heads=12. The diffusion process uses #timesteps=100 with a cosine noise schedule (Nichol & Dhariwal, 2021). After training, the $\mathcal{G}$ produces 100,000 candidates before filtering for both MP-20 and Alex-MP-20. The V.S.U.N. filter retains 8,086 candidates on MP-20 and 12,408 on Alex-MP-20, corresponding to acceptance rates of 8.1% and 12.4%, respectively. Filtering uses only the training-set reference. We also use a stricter stability threshold during filtering – energy above hull below 0.05 eV/atom, compared with 0.1 eV/atom in the final evaluation – to reduce the number of weak candidates passed to refinement.

The base generative model also adopts a DDPM, where the denoising network is an Transformer with #layer=8, #hidden_dimension=512 and #attention_heads=8. We use diffusion #timesteps=256 and

the same cosine noise schedule (Nichol & Dhariwal, 2021). For classifier-free guidance, the guidance strength is set to $\#\omega = 2$, and the condition is randomly masked during training with probability 0.2.

C.3 Evaluation Metric

Our goal in de novo generation task it to generate valid, stable, unique and novel materials. These four basic metrics are defined as follows:

- `Validity` (V). We consider the validity of a crystal from both structure and composition (Xie et al., 2021). For structure, a valid crystal should have volume larger than 0.1, and the minimal distance among all atom pairs should larger than 0.5. For element composition, we check charge neutrality and electronegativity difference. If one crystal satisfies these two validity simultaneously, it is overall valid.
- `Stability` (S). Stability measures whether a generated crystal is thermodynamically feasible. For each generated structure, we compute its energy above the convex hull, denoted as ΔE, using a surrogate model (e.g., MLFF) or DFT when available. A structure is considered stable if $\Delta E < \epsilon$, where ϵ is a small threshold.
- `Uniqueness` (U). Uniqueness measures the diversity of generated structures by removing duplicates within the generated set. Two generated structures are considered identical if they match under the same structural equivalence criterion.
- `Novelty` (N). Novelty evaluates whether generated crystals are distinct from the reference dataset. A generated structure is considered non-novel if it matches any structure in the reference set under a structural equivalence criterion (e.g., using `StructureMatcher`). Otherwise, it is regarded as novel.

We discuss more on stability and novelty in this paper. To further evaluate generation quality, we further report compound metrics that measure the joint satisfaction of multiple criteria. Let $\{\mathbf{C}_{gen}\}_{i=1}^{N}$ denote N generated crystals. For each crystal, we define indicator functions:

$$S_i = \mathbb{I}(E_{\text{hull}}^{(i)} \leq \epsilon), \quad N_i = \mathbb{I}(\mathbf{C}_i \notin \mathcal{D}_{\text{ref}}), \quad U_i = \mathbb{I}(\mathbf{C}_i \text{ is unique in } \mathcal{C}), \quad V_i = \mathbb{I}(\mathbf{C}_i \text{ is valid}). \tag{6}$$

Then, we define

- `Stable & Unique & Novel` (S.U.N) measures the fraction of generated crystals that are simultaneously stable, unique, and novel:

$$\texttt{S.U.N} = \frac{1}{N}\sum_{i=1}^{N} S_i \cdot U_i \cdot N_i. \tag{7}$$

- `Valid & Stable & Unique & Novel` (V.S.U.N) further incorporates validity, measuring the fraction of samples that satisfy all four criteria:

$$\texttt{V.S.U.N} = \frac{1}{N}\sum_{i=1}^{N} V_i \cdot S_i \cdot U_i \cdot N_i. \tag{8}$$

C.4 Concept Verification Settings

In Section 6, we conduct automated analysis to quantify the consistency between concepts and structural patterns in two directions. The corresponding experimental settings are as follows.

a. Concept-to-structure coherence. We select the 1,000 most frequently activated concepts. For each concept, we collect all atomic local environments assigned to that concept and computed CrystalNN (Zimmermann & Jain, 2020) (local structural descriptor) and normalized Magpie (Ward et al., 2016) (composition-based descriptor) fingerprints for them. Next, we obey the following steps using two-type fingerprints, respectively:

1. Measure the within-concept dispersion as the average distance between the fingerprints and their concept-specific centroid. A lower dispersion indicates that the same concept is consistently assigned to structurally similar local environments.

2. To construct a null distribution, we randomly sample equally sized groups of local environments and calculate their dispersion. This random sampling was repeated 1,000 times for each concept.
3. Compare each within-concept dispersion with its 1,000 null distribution dispersion.
4. Additionally conduct a one-sided permutation-based significance test. The empirical (p)-value measures how frequently a randomly sampled group achieves a dispersion equal to or lower than the observed within-concept dispersion.

b. Structure-to-concept recovery. After generation, we re-encode the generated crystals with the VQ-VAE encoder and compare the recovered codes with their corresponding target codes given in conditions. We report exact Top-1 and Top-k recovery. Top-k counts a match to either the target or one of its k-1 nearest neighbours in the 10,000-entry codebook. As a control, we exchange target compositions between crystals with the same number of atoms.

A representative failure in this recovery occurs when a complex local environment contains a simpler motif as a partial substructure. For example, a distorted eight-fold environment may be encoded as a nearby six-fold, octahedral-like concept. This reflects ambiguity from deformation and structural overlap rather than a complete absence of the intended pattern.

C.5 Density Functional Theory Settings

We use DFT settings from Materials Project `https://docs.materialsproject.org/methodology/materials-methodology/calculation-details/gga+u-calculations/parameters-and-convergence` for structure relaxation and energy computation. In particular, we do GGA and GGA+U calculations with `atomate2.vasp.flows.mp.MPGGADoubleRelaxStaticMaker` (Ganose et al., 2025), which in turn relies on `pymatgen.io.vasp.sets.MPRelaxSet` and `pymatgen.io.vasp.sets.MPStaticSet` (Ong et al., 2013). Computations themselves were done with VASP (Kresse & Furthmüller, 1996) version 5.4.4. with the plane-wave basis set (Kresse & Furthmüller, 1996). The electron-ion interaction is described by the projector augmented wave (PAW) pseudo-potentials (Kresse & Joubert, 1999). The exchange-correlation of valence electrons is treated with the Perdew-Burke-Ernzerhof (PBE) functional within the generalized gradient approximation (GGA) (Perdew et al., 1996). The raw total energies computed by DFT were corrected with `MaterialsProject2020Compatibility` before putting into the `PhaseDiagram` to obtain the DFT E_{hull}.

We do DFT relaxation firstly using the generated crystal structures. Then, for the crystals failed to DFT, we follow previous studies (Miller et al., 2024; Joshi et al., 2025) and use MatterSim-v1-1M (Yang et al., 2024) to do pre-relaxation, and then redo DFT.

C.6 Operating Environment

The environment where our code runs is shown as follows:

- Operating system: Linux version 6.8.0-63-generic
- CPU information: AMD EPYC 9554 64-Core Processor
- GPU information: NVIDIA Corporation AD102GL [L40S]

D Related Work

Generation on Scientific Data Recent progress in deep generative modeling, particularly diffusion-based methods, has greatly advanced scientific discovery tasks such as molecule and material design. AlphaFold3 (Abramson et al., 2024) employs diffusion to achieve accurate all-atom biomolecular complex generation. GeoLDM (Xu et al., 2023) highlights the capability of diffusion models in scientific data by modeling 3D geometric structures and generating physically consistent molecular samples. Graph-GRPO (Zhu et al., 2026) further enhances molecule generation by aligning graph flow models with task-specific objectives through reinforcement learning (Sui et al., 2025).

For crystal generation, CDVAE (Xie et al., 2021), FlowMM (Miller et al., 2024), and ADiT (Joshi et al., 2025) introduce variational, flow matching, and latent diffusion frameworks, respectively. DiffCSP (Jiao et al., 2023) proposes CSPNet, which has become a standard equivariant denoising backbone. Follow-up works, including DiffCSP++ (Jiao et al., 2024), SGEquiDiff (Chang et al., 2025), and SymmCD (Levy et al., 2025), incorporate domain-specific physical priors such as space group constraints and crystallographic symmetry. MatterGen (Zeni et al., 2023) further extends diffusion models to conditional generation for inverse material design guided by target properties and symmetry. In addition, recent studies explore integrating large language models with crystal generation (Gruver et al.; Sriram et al., 2024; Khastagir et al., 2025).

Existing approaches develop generative architectures, training objectives, and conditioning mechanisms for crystal generation. Our work focuses on learning a discrete vocabulary of local crystal concepts and using their compositions to guide structural exploration. Within this framework, refinement improves the quality of sampled concept compositions, while concept-conditioned generation supports the exploration of novel structures.

Transformer Transformers (Vaswani, 2017) have emerged as a versatile architecture for modeling structured data, owing to their ability to capture long-range dependencies via self-attention. Representative developments include LLaMA (Touvron et al., 2023), which advances large-scale autoregressive modeling, and Vision Transformer (ViT) (Dosovitskiy et al., 2020), which extends Transformers to image modeling and facilitates progress in multimodal learning. Transformers have also been incorporated into generative modeling, such as Diffusion Transformers (DiT) (Peebles & Xie, 2023), which serve as strong backbones for diffusion models. In graph domains (Zhang et al., 2025), Graph Transformer (GT) (Dwivedi & Bresson, 2020) introduces self-attention mechanisms to graph data, while subsequent studies aim to address its limitations. For instance, CoBFormer (Xing et al., 2024) firstly reveals the over-globalization issue and enhances local inductive biases with theoretical guarantees. Specformer (Bo et al., 2023) makes the first attempt to learn eigenvalues interaction via Transformer, triggering the fusion of attention and spectral domain.

In this work, we employ a standard Transformer as the backbone. Investigating more advanced Transformer architectures remains an important direction for future research.