Crystalite: A Lightweight Transformer for Efficient Crystal Modeling

Anonymous Authors[1]

Abstract

Generative models for crystalline materials often rely on equivariant graph neural networks, which capture geometric structure well but are costly to train and slow to sample. We present Crystalite, a lightweight diffusion Transformer for crystal modeling built around two simple inductive biases. The first is *Chemically Structured Atom Tokens*, a compact atom representation that replaces high-dimensional one-hot encodings and is better suited to continuous diffusion. The second is the *Geometry Enhancement Module* (GEM), which injects periodic minimum-image pair geometry directly into attention through additive geometric biases. Together, these components preserve the simplicity and efficiency of a standard Transformer while making it better matched to the structure of crystalline materials. Crystalite achieves state-of-the-art results on crystal structure prediction benchmarks, and *de novo* generation, attaining the best Stable, Unique, and Novel (SUN) discovery score among the evaluated baselines while sampling substantially faster than geometry-heavy alternatives.

1. Introduction

The discovery of novel, synthesizable, and diverse crystalline materials with targeted properties remains a central goal of materials science (Merchant et al., 2023). The search space of possible compositions and structures is combinatorially vast, while only a small fraction of candidates is thermodynamically stable. Traditional computational approaches can explore this space systematically (Pickard & Needs, 2011; Oganov & Glass, 2006). However, even with large high-throughput infrastructures (Jain et al., 2013; Curtarolo et al., 2012; Kirklin et al., 2015), candidate evaluation still typically relies on density functional theory (DFT) (Kohn & Sham, 1965; Jones, 2015), which can be computationally expensive (Goedecker, 1999).

Deep generative models offer a promising alternative by learning to propose candidate materials directly from data (Xie et al., 2022; Zeni et al., 2025). In crystal generation, however, the geometric and symmetry structure of the problem has driven much of the literature toward equivariant graph neural networks (GNNs) and other specialized architectures (Luo et al., 2025; Jiao et al., 2023; Zeni et al., 2025; Miller et al., 2024). While highly effective, these approaches can be architecturally complex and computationally demanding, motivating the search for simpler approaches that still capture enough crystal geometry to remain competitive (Yang et al., 2023). This raises the question: can a lightweight Transformer recover enough geometric structure to compete without explicit equivariant message passing?

Recent work suggests that transformers can be competitive with GNN-based approaches for crystal generation. In particular, diffusion transformers have emerged as a promising lightweight alternative for atomistic and crystalline generation (Yi et al., 2025; Joshi et al., 2025; Jin et al., 2026). However, these approaches often incorporate crystal geometry only weakly or indirectly, leaving open whether a standard diffusion transformer can remain simple while benefiting from a more direct injection of periodic geometric structure.

In this work, we introduce **Crystalite**, a lightweight diffusion transformer for crystalline materials. Crystalite augments standard multi-head attention with periodic and geometric biases, and uses a compact chemically informed atom representation in place of high-dimensional one-hot type encodings. This preserves the simplicity and scalability of a standard transformer backbone while improving its suitability for crystal generation.

Our main contributions are as follows:

- We introduce the Geometric Enhancement Module (GEM), a lightweight attention-biasing mechanism that

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the AI4Physics Workshop at the International Conference on Machine Learning (AI4Physics@ICML 2026). Do not distribute.

injects periodic and pairwise geometry directly into standard Transformers, providing an efficient alternative to equivariant message passing.

- We replace one-hot atom types with a compact chemically informed representation that is better matched to continuous diffusion.
- We show that Crystalite achieves state-of-the-art crystal structure prediction (CSP) and *de novo* generation (DNG) performance, while sampling much faster than geometry-heavy baselines.
- We characterize the trade-off between novelty, validity, and stability, and show that MLIP-based stability estimates provide a practical signal for model selection.

2. Related Work

Crystal generators differ mainly in how they encode geometric structure. Some methods impose symmetry and periodicity through equivariant or geometry-aware architectures, while others use lighter transformer-style backbones with weaker inductive bias. Crystalite follows the latter direction, but injects periodic pairwise geometry directly into attention.

Equivariant and geometry-aware crystal generators. Diffusion models (Ho et al., 2020; Song & Ermon, 2019) are widely used for atomistic generation. In crystals, many methods pair diffusion with equivariant GNNs to handle graph structure, periodicity, and geometric symmetries (see Appendix A). MatterGen (Zeni et al., 2025), for example, uses a GemNet-based score network (Gasteiger et al., 2021) to jointly model atom types, fractional coordinates, and lattice parameters, and supports inverse. EGNN (Satorras et al., 2021), used in DiffCSP (Jiao et al., 2023), has also served as the backbone for several later methods (Miller et al., 2024; Höllmer et al., 2025; Cornet et al., 2025; Luo et al., 2025). Other work specializes the generative process itself: FlowMM (Miller et al., 2024) extends Riemannian flow matching (Chen & Lipman, 2024) to fractional coordinates, Höllmer et al. (2025) study stochastic interpolants (Albergo et al., 2025), and KLDM (Cornet et al., 2025) lifts periodic coordinates to an auxiliary flat space using the torus Lie group structure. These approaches demonstrate the value of strong geometric inductive bias, but increase architectural or algorithmic complexity.

Lightweight and transformer-based alternatives. Recent work also shows that full equivariance is not always necessary. UniMat (Yang et al., 2023) uses a 3D U-Net diffusion model that remains competitive with equivariant baselines and benefits from scale. Transformer-based models have been explored in autoregressive and hybrid settings (Mohanty et al., 2024; Kazeev et al., 2025; Gruver et al., 2024; Cao et al., 2025), including pipelines where language models provide crystal priors before geometric refinement (Khastagir et al., 2026; Sriram et al., 2024). Closest to Crystalite are diffusion-transformer approaches: ADiT (Joshi et al., 2025) uses latent diffusion transformers (Peebles & Xie, 2023; Rombach et al., 2022) for joint molecule and material generation, Morehead et al. (2026) simplify this formulation, OXtal (Jin et al., 2026) combines diffusion transformers with EDM-style preconditioning and sampling (Karras et al., 2022), and CrystalDiT (Yi et al., 2025) applies diffusion transformers to crystal generation. Crystalite belongs to this lightweight transformer family, but differs by injecting periodic minimum-image pair geometry as an additive attention bias.

3. Methodology

Crystalite is built around a simple idea: keep the denoising backbone close to a standard diffusion Transformer, and incorporate crystal-specific structure through the representation, attention mechanism, and sampling procedure. We begin from the standard unit-cell description of a crystal in terms of atom identities, fractional coordinates, and lattice geometry. On top of this representation, we replace one-hot atom identities with chemically structured tokens, define diffusion jointly over atom, coordinate, and lattice variables, and process the resulting state with a Transformer that uses one token per atom together with a single global lattice token. Periodic pairwise geometry can then be injected directly into attention through the Geometry Enhancement Module (GEM), while a channel-wise anti-annealing heuristic improves refinement at sampling time.

Concretely, throughout this section we represent a crystal with N atoms by the unit-cell tuple

$$\begin{aligned} &\mathcal{C} = (\mathbf{A}, \mathbf{F}, \mathbf{L}), \\ &\mathbf{A} \in \{0,1\}^{N \times N_Z},\ \mathbf{F} \in [0,1)^{N \times 3},\ \mathbf{L} \in \mathbb{R}^{3 \times 3}, \end{aligned} \tag{1}$$

where N_Z is the number of supported atom types and each row satisfies $\mathbf{A}_i = \mathrm{onehot}(a_i)$ for some label $a_i \in \{1, \ldots, N_Z\}$. Here, $\mathbf{A}$ is the atom-type matrix, $\mathbf{F}$ contains the fractional coordinates, and $\mathbf{L}$ defines the periodic unit cell. The corresponding Cartesian coordinates are given by $\mathbf{X} = \mathbf{F}\mathbf{L}$.

3.1. Chemically Structured Atom Tokens

A standard representation uses the one-hot atom-type matrix $\mathbf{A} \in \{0,1\}^{N \times N_Z}$. We found this choice suboptimal for diffusion over crystalline materials for two reasons. First, for realistic materials datasets N_Z can be large (e.g. $N_Z = 89$ on MP-20), making the atom-type channel unnecessarily high-dimensional relative to the underlying chemical vari-

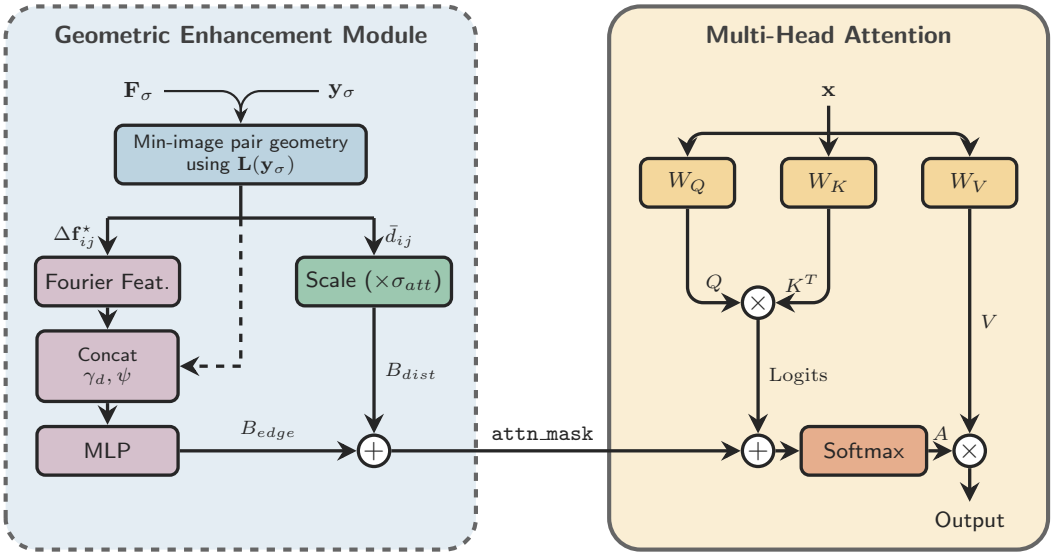


Figure 1. Overview of the proposed architecture. **Left:** The Geometric Enhancement Module (GEM) computes periodic minimum-image pair geometry from $\mathbf{F}_\sigma$ and the reconstructed lattice $\mathbf{L}(\mathbf{y}_\sigma)$. The resulting geometric features are converted into additive attention biases through a learned edge pathway and a distance-based pathway. **Right:** Biases are added to the standard multi-head attention logits before the softmax, allowing attention to depend on crystal geometry while retaining the usual dot-product attention structure.

able. Second, the one-hot geometry is chemically uninformative: all elements are mutually orthogonal, so for example, Li is as far from Na as it is from Xe. This can encourage the model to memorize recurring compositions, while providing no notion of smooth chemical similarity. A learned embedding could in principle reduce dimensionality, but would make the atom-type geometry itself a learned quantity. This adds another representation-learning problem to the denoising objective, and without explicit regularization does not guarantee that chemically similar elements are placed nearby. We therefore use fixed chemically structured descriptors, which provide a meaningful continuous geometry from the start of training.

To address this, we replace the one-hot channel by a low-dimensional continuous tokenization, which we refer to as *Chemically Structured Atom Tokens*. For each supported element $k \in \{1, \ldots, N_Z\}$, let r_k, g_k, and b_k denote its period, group, and block, and let (s_k, p_k, d_k, f_k) denote its ground-state valence-shell occupancies. The tokenized representation associated with element k is

$$\begin{aligned}\mathbf{h}_k = \big[&\mathsf{onehot}(r_k),\ \mathsf{onehot}(g_k),\\ &\mathsf{onehot}(b_k),\\ &s_k/2,\ p_k/6,\ d_k/10,\ f_k/14\big].\end{aligned} \tag{2}$$

Figure 2 illustrates representative chemically structured element tokens. Following the implementation used in our experiments, these element-wise descriptors are standardized across the supported elements, optionally projected with a fixed PCA basis, and finally ℓ_2-normalized. We continue to denote the resulting tokenized vectors by $\mathbf{h}_k$. The tokenized atom matrix is then

$$\mathbf{H} = \big[\mathbf{h}_{a_1}, \ldots, \mathbf{h}_{a_N}\big]^\top \in \mathbb{R}^{N \times d_H}, \tag{3}$$

where d_H denotes the token dimension after optional PCA compression. This design serves two purposes. First, it reduces the dimensionality of the atom-type channel, which makes denoising statistically easier and lowers the capacity of the model to memorize frequent compositional patterns. Second, it equips the diffusion process with a chemically meaningful geometry: errors in this continuous token space become structured, so that under noise the model is encouraged to confuse elements with plausible substitutions before unrelated species.

Chemically Structured Atom Tokens are especially natural in our EDM formulation, since atom types are treated as continuous diffusion variables jointly with fractional coordinates and lattice parameters. The denoiser therefore does not need to recover a sparse one-hot vector in a high-dimensional simplex-like space, but instead returns a low-dimensional chemical token. During sampling, the denoised token $\hat{\mathbf{h}}_i$ is mapped back to a discrete element by nearest-token decoding,

$$\hat{a}_i = \arg\max_{k \in \{1,\ldots,N_Z\}} \langle \hat{\mathbf{h}}_i, \mathbf{h}_k \rangle, \tag{4}$$

which is equivalent to cosine-similarity decoding because all token vectors are normalized, and keeps the training and

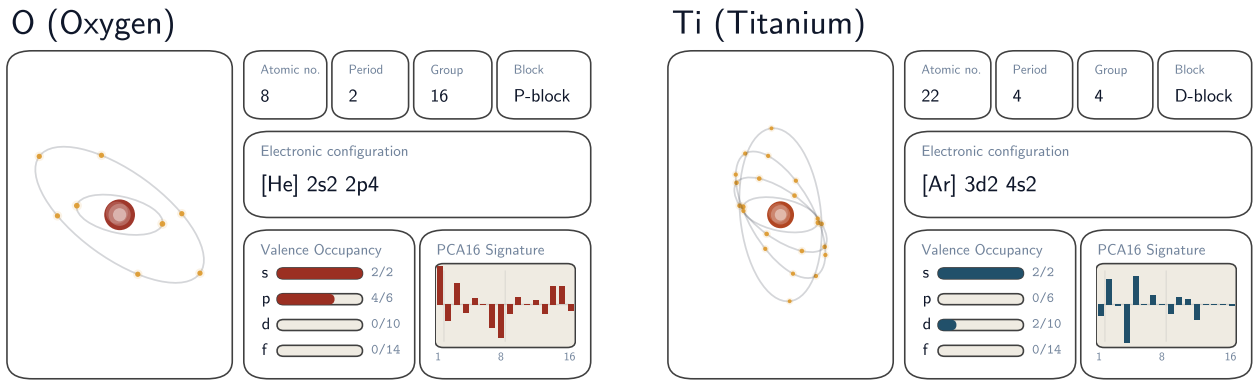


Figure 2. **Chemically structured tokenization of atomic species.** Instead of representing each chemical element by a one-hot identity vector, we assign each element a fixed 34-dimensional chemically structured descriptor built from its period, group, block, and valence-shell occupancies. These descriptors are compressed to a 16-dimensional token space using PCA, yielding a continuous atom-type representation for diffusion. Diffusion noise is applied in this 16-dimensional space, after which a learned embedding maps the noisy token to the Transformer hidden dimension. Representative examples are shown for oxygen (left) and titanium (right).

decoding geometries aligned. In the CSP setting, where the composition is known, the tokenized atom matrix is held fixed and only the coordinate and lattice channels are denoised. We provide additional information on this embedding in Appendix B.1, and we demonstrate its empirical advantage over standard one-hot encodings for generative discovery in Appendix F.1.

3.2. Diffusion formulation for crystals

Starting from a crystal $\mathcal{C} = (\mathbf{A}, \mathbf{F}, \mathbf{L})$, we define a continuous diffusion state

$$(\mathbf{H}, \mathbf{F}, \mathbf{y}),$$

where $\mathbf{H}$ is the chemically structured atom-type representation, $\mathbf{F} \in [0,1)^{N\times 3}$ contains the fractional coordinates, and $\mathbf{y} \in \mathbb{R}^6$ is an unconstrained parameterization of the lattice. Concretely, $\mathbf{h}_i \in \mathbb{R}^{d_H}$ denotes the token of atom i, and $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]^\top$. Likewise, $\mathbf{f}_i \in [0,1)^3$ denotes the fractional coordinate of atom i, and $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_N]^\top$.

Rather than diffusing the raw lattice matrix $\mathbf{L} \in \mathbb{R}^{3\times 3}$ directly, we represent it through a lower-triangular latent $\mathbf{y} \in \mathbb{R}^6$ and reconstruct

$$\mathbf{L}(\mathbf{y}) = \begin{bmatrix} e^{y_1} & 0 & 0 \\ y_2 & e^{y_3} & 0 \\ y_4 & y_5 & e^{y_6} \end{bmatrix}. \tag{5}$$

To reduce basis ambiguity we preprocess each structure into a Niggli-reduced cell, and during training the only explicit crystal augmentation is a random global translation of the fractional coordinates.

We adopt the EDM framework because it provides a well-optimized set of preconditioning, loss weighting, and sampling choices. Following EDM, at each training step we sample a noise level from

$$\log \sigma \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2),$$

and perturb all three channels jointly:

$$(\mathbf{H}_\sigma, \mathbf{F}_\sigma, \mathbf{y}_\sigma) = (\mathbf{H}, \mathbf{F}, \mathbf{y}) + \sigma\,\boldsymbol{\varepsilon}, \tag{6}$$

where $\boldsymbol{\varepsilon}$ denotes Gaussian noise with the appropriate channel-wise shapes. For the coordinate channel, noise is added in a centered Euclidean representation: fractional coordinates are first shifted to a centered cube, Gaussian noise is added in that space, and the resulting noisy coordinates are wrapped back into $[0,1)^3$ before being embedded by the Transformer. The training loss, however, is evaluated using a componentwise wrapped residual in fractional space. This respects periodicity on the torus, but unlike GEM it is not a metric-aware minimum-image search under the lattice metric. As in EDM, the noisy inputs and raw network outputs are combined through the standard channel-wise preconditioning coefficients $c_{\text{in}}(\sigma)$, $c_{\text{skip}}(\sigma)$, and $c_{\text{out}}(\sigma)$.

We train with separate denoising losses for the atom-type, coordinate, and lattice channels. Atom tokens and lattice latents are regressed in Euclidean space, while coordinates are compared through componentwise wrapped residuals on the torus, $\text{wrap}(\mathbf{u}) = \mathbf{u} - \text{round}(\mathbf{u})$. Using the standard EDM channel-wise weights $w_H(\sigma), w_F(\sigma), w_{\text{lat}}(\sigma)$, the total objective is

$$\mathcal{L} = \lambda_H \mathcal{L}_H + \lambda_F \mathcal{L}_F + \lambda_{\text{lat}} \mathcal{L}_{\text{lat}}. \tag{7}$$

Complete mathematical and implementation details for our diffusion formulation, including data preprocessing, exact

preconditioning formulas, and the explicit per-channel loss functions are provided in Appendix D.

We use the same diffusion formulation for both DNG and CSP. In DNG, Crystalite models the joint distribution $p_\theta(\mathbf{A}, \mathbf{F}, \mathbf{L})$. Because the number of atoms per unit cell varies across structures, we first sample $N \sim p(N)$ from the empirical training-set distribution. In CSP, it instead models the conditional distribution $p_\theta(\mathbf{F}, \mathbf{L} \mid \mathbf{A})$, treating structure prediction as conditional generation with the composition fixed.

3.3. Crystalite architecture

Crystalite operates on the continuous diffusion state $(\mathbf{H}, \mathbf{F}, \mathbf{y})$ using a standard Transformer backbone with one token per atom and one global token for the lattice. Each atom token is created by adding learned embeddings of the chemically structured token $\mathbf{h}_i$ and the fractional coordinate $\mathbf{f}_i$, so that chemical identity and geometric position are jointly represented. The latent lattice vector $\mathbf{y}$ is mapped to a separate global token. The diffusion noise level is encoded through the standard EDM coordinate $c_{\text{noise}}(\sigma) = \frac{1}{4} \log \sigma$ and injected into every block via adaptive layer normalization (AdaLN).

The sequence is processed by stacked self-attention and feed-forward blocks. Per-token output heads then decode atom and coordinate predictions from each of the first N tokens, while a separate lattice head decodes $\hat{\mathbf{y}}$ from the lattice token, yielding

$$(\hat{\mathbf{H}}, \hat{\mathbf{F}}, \hat{\mathbf{y}}) = \text{Crystalite}_\theta(\mathbf{H}_\sigma, \mathbf{F}_\sigma, \mathbf{y}_\sigma; \sigma), \tag{8}$$

which are interpreted as denoised predictions and combined with the noisy inputs via the standard EDM preconditioning rules. A schematic (Figure 9) and the full input/output parameterization equations are provided in Appendix C.

3.4. Geometry Enhancement Module (GEM)

Crystalite augments standard self-attention with a geometry-dependent additive bias, recomputed at each denoising step. The use of additive pairwise biases in attention has precedents in position encoding for language models (Press et al., 2022; Shaw et al., 2018) and graph transformers (Ying et al., 2021). GEM adapts this mechanism to periodic crystal geometry. This injects periodic pairwise structure into the attention mechanism without requiring equivariant message-passing, as shown in Figure 1.

Given the fractional coordinates $\mathbf{F}$ and lattice latent $\mathbf{y}$, we reconstruct the lattice matrix $\mathbf{L}(\mathbf{y})$. For each atom pair (i, j), we compute the minimum-image fractional displacement $\Delta\mathbf{f}^\star_{ij}$ under periodic boundary conditions and its normalized Cartesian distance:

$$\bar{d}_{ij} = \frac{\|\Delta\mathbf{f}^\star_{ij}\mathbf{L}(\mathbf{y})\|_2}{s(\mathbf{y})}, \tag{9}$$

where $s(y)$ is a characteristic cell scale; in our implementation we use the mean of the three lattice lengths. Unlike the wrapped fractional residual used in the coordinate loss, GEM selects the periodic image by minimizing the Cartesian quadratic form induced by the lattice metric $\mathbf{G} = \mathbf{L}\mathbf{L}^\top$.

From this geometry, GEM constructs a head-wise attention bias by combining a direct distance penalty with learned edge features. This combined bias is then modulated by a learned noise-dependent gate $g_h(\sigma)$ to form the final geometric bias:

$$B^{\text{geom}}_{hij} = g_h(\sigma)\left(B^{\text{dist}}_{hij} + B^{\text{edge}}_{hij}\right) \tag{10}$$

where the distance penalty $B^{\text{dist}}_{hij} = w_h \bar{d}_{ij}$ uses a learned, monotonically non-positive slope $w_h \leq 0$, and the edge bias models non-linear interactions through an MLP:

$$B^{\text{edge}}_{hij} = \text{MLP}_{\text{edge}}\left(\left[\gamma_\Delta(\Delta\mathbf{f}^\star_{ij}), \gamma_d(\bar{d}_{ij}), \psi(\mathbf{y})\right]\right)_h. \tag{11}$$

Here, γ_Δ applies Fourier features to the displacement, γ_d applies a Radial Basis Function (RBF) kernel to the distance, and $\psi(\mathbf{y})$ is a low-dimensional lattice descriptor.

This geometric bias is applied exclusively to atom–atom interactions. Padding $\mathbf{B}^{\text{geom}}$ with zeros for any interactions involving the global lattice token, the attention update becomes:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} + \mathbf{B}^{\text{geom}}\right)V. \tag{12}$$

This allows the model to emphasize geometrically compatible atom pairs directly in the attention logits while maintaining the simplicity and efficiency of a standard diffusion Transformer. We provide more details on the implementation in Appendix C.2.

3.5. Channel-wise anti-annealing during sampling

During EDM sampling, we optionally apply a *channel-wise anti-annealing* step that rescales the reverse-time update separately for the atom-token, coordinate, and lattice channels. Concretely, the standard Heun-style EDM update is multiplied by a channel-specific factor $\alpha_i^{(q)} \geq 1$ for $q \in \{H, F, \mathrm{y}\}$, derived from an auxiliary Karras schedule, with $\alpha_i^{(q)} = 1$ recovering the standard sampler. Intuitively, this acts as a channel-dependent time warp: if a particular channel denoises more slowly or dominates the remaining error, anti-annealing drives that channel more aggressively toward the denoised prediction while leaving the learned denoiser unchanged. We found this particularly useful for

improving geometric refinement at sampling time without modifying the training objective. Full details are given in Appendix D.1, with additional ablations of anti-annealing on DNG in Appendix F.5 and on CSP in Appendix F.6.

4. Experimental Setup

4.1. Datasets

We benchmark on three datasets. MP-20 (Xie et al., 2022), derived from the Materials Project (Jain et al., 2013), contains 45 231 crystals with up to 20 atoms per unit cell and 89 atom types. MPTS-52 (Baird et al., 2024) contains 40 476 Materials Project structures with up to 50 atoms per unit cell, split chronologically to induce distribution shift between train, validation, and test sets. Alex-MP-20 (Zeni et al., 2025) contains 675 204 structures with up to 20 atoms per unit cell, derived from Alexandria and MP-20. We use the splits from Höllmer et al. (2025).

4.2. Task setup

We evaluate Crystalite in two settings: DNG and CSP. In the DNG setting, the model generates atom types, fractional coordinates, and lattice parameters jointly from noise. In the CSP setting, the atomic composition is provided as input, and the model predicts only the crystal geometry, i.e. the fractional coordinates and lattice. This is implemented by fixing the chemically structured atom tokens to the known composition and masking the type loss during training and sampling.

Model settings. Unless otherwise stated, we use a single base Crystalite configuration for all datasets and for both DNG and CSP. The model has approximately 6.7×10^7 trainable parameters, with a 14-layer Transformer backbone of width $d = 512$ and 16 attention heads. Atom identities are represented using PCA-compressed Chemically Structured Atom Tokens with dimension $d_H = 16$, and GEM is enabled in all main experiments. We train in `bfloat16`, maintain an exponential moving average (EMA) of the model parameters, and use the EMA weights for all reported sampling and evaluation results. Training each model takes roughly one day on a single NVIDIA H100 GPU. Unless noted otherwise, sampling uses the same EDM configuration across benchmarks, including 150 sampling steps and the same channel-wise anti-annealing settings. The only task-specific modification is in CSP, where the composition is fixed and only the coordinate and lattice channels are denoised. Full architectural, training, and sampling details are given in Appendix C, Appendix D, and Table 3.

Sampling speed benchmarking. For a fair comparison of sampling speed, we measure the wall-clock time required to generate 1,000 crystals on a single NVIDIA H100 GPU. For each model, we use the largest sampling batch size that fits in memory, so that each method is evaluated at its highest feasible throughput. Unless otherwise noted, the reported timing corresponds to the standard inference setting used for cross-model comparison. For Crystalite, we additionally report a second timing, marked with † in Table 2, obtained with `xFormers` attention and `bfloat16` inference. We regard the primary timing as the main comparison across methods, and the daggered number as the optimized version.

5. Results and Discussion

5.1. CSP Results

Table 1 summarizes the results on the CSP benchmarks. Across all datasets, Crystalite outperforms prior methods. Using Match Rate to assess successful structure recovery and RMSE to measure geometric accuracy (see Appendix E.1), Crystalite achieves state-of-the-art results on both criteria. The improvement is especially pronounced in RMSE, indicating more accurate structural recovery even in settings where match-based performance is already strong.

The effect of GEM is examined in more detail in Appendix F.4. Compared to the best previous RMSE on each benchmark, Crystalite reduces RMSE by 34.8%, 44.3%, and 74.7% on MP-20, MPTS-52, and Alex-MP-20, respectively, for an average reduction of approximately 51%. The GEM ablation shows that this gain is partly due to the geometry-aware attention bias: GEM has little effect on Match Rate, but consistently improves RMSE by about 20%. This suggests that GEM mainly refines geometric accuracy rather than determining whether the correct structural mode is recovered.

5.2. DNG Results

Table 2 summarizes the main DNG results. Crystalite achieves the highest SUN rate and the fastest sampling speed among the compared methods. Since DNG is fundamentally governed by a trade-off between stability and diversity, we treat SUN as the primary summary metric; the remaining metrics group into *quality and diversity* and *stability and distribution* categories described in Appendix E.1. Figure 3 shows this trade-off explicitly during training: as the model fits the training distribution more closely, validity, stability, and distribution metrics improve, while novelty and uniqueness decrease. Since the ground-truth dataset itself lacks perfect compositional validity, very high scores on these metrics typically signal overfitting and lower novelty rates.

This trade-off is especially pronounced because atom types are modeled jointly with coordinates and lattice parameters, making compositional memorization difficult to control independently of structural quality. We mitigate this

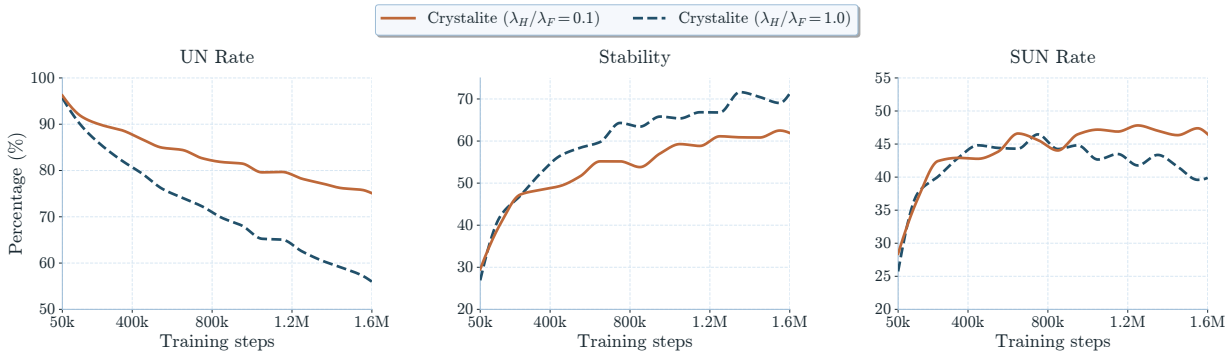


Figure 3. **Training-time trade-off in DNG.** UN rate, stability, and SUN rate over training for two atom-loss settings. The more stable run loses UN faster, while the diversity-preserving setting maintains a flatter SUN trajectory, illustrating the stability–diversity trade-off.

Table 1. CSP results across standard benchmarks. Best values are in bold.

Model	MP-20		MPTS-52		Alex-MP-20	
	MR (%) ↑	RMSE ↓	MR (%) ↑	RMSE ↓	MR (%) ↑	RMSE ↓
CDVAE	33.90	0.1045	5.34	0.2106	–	–
DiffCSP	51.49	0.0631	12.19	0.1786	–	–
FlowMM	61.39	0.0566	17.54	0.1726	–	–
CrystalFlow	62.02	0.0710	22.71	0.1548	–	–
KLDM	65.83	0.0517	23.93	0.1276	–	–
OMatG	63.75	0.0720	25.15	0.1931	64.71	0.1251
Crystalite	**66.09** ± 0.09	**0.0337** ± 0.005	**31.56** ± 0.06	**0.0711** ± 0.0010	**68.26** ± 0.12	**0.0317** ± 0.0004

by substantially downweighting the atom-type loss, which Figure 3 shows yields a flatter, more sustained SUN trajectory and more stable checkpoint selection compared to evenly balanced loss weights. As further analyzed in the GEM ablation (Appendix F.3), GEM mainly improves the stability side of this trade-off, contributing to consistently higher SUN throughout training.

Fairness and comparability between models. Our primary evaluation pipeline uses NequIP-based relaxation (Batzner et al., 2022) together with SUN-based checkpoint selection. For fairness, all main-table baselines were evaluated within this exact pipeline. However, to verify that Crystalite does not benefit disproportionately from this setup, we additionally evaluate it using the MatterGen evaluation pipeline and LeMat-GenBench (Betala et al., 2026). LeMat provides a strictly standardized evaluation against a reference set of roughly 5 million structures. Within this unified framework, Crystalite achieves state-of-the-art generative performance, attaining the highest SUN and MSUN scores among all evaluated pre-relaxed methods. The corresponding results for both external pipelines are reported in Appendix Tables 7 and 4.

Extensive and intensive metrics. DNG metrics differ in how they scale with the number of generated samples. Some are sample-*intensive*, reflecting properties of individual draws and therefore estimable from random subsets. Others are sample-*extensive*, characterizing the generated set as a whole and varying with sampling budget. Uniqueness and derived quantities such as the UN rate are strongly sample-extensive: as more crystals are generated, duplicates accumulate and these metrics typically decrease.

This matters because useful generators should continue to discover distinct, novel candidates as sampling scales. We therefore compare Crystalite and ADiT in Figure 4 as a function of the number of generated crystals, showing that Crystalite preserves diversity more effectively at larger budgets. More generally, sample-extensive metrics should be reported with the total number of generated samples, since values are not directly comparable across budgets. Appendix E.3 formalizes this distinction and identifies which metrics can be reliably estimated from subsets.

6. Conclusion

We introduced Crystalite, a lightweight diffusion Transformer for CSP and DNG. By combining chemically structured atom tokens with the Geometry Enhancement Module

Table 2. Generative quality, diversity, stability, distribution, and sampling speed metrics. Metrics are computed from 10,000 generated crystals per model. Stability is evaluated using the same NequIP relaxation pipeline. Time is seconds per 1k samples; † denotes optimized Crystalite sampling.

Model	Quality / Diversity					Stability / Distribution / Speed				
	Str. (%)↑	Comp. (%)↑	Uniq. (%)↑	Nov. (%)↑	U.N. (%)↑	Stab. (%)↑	S.U.N. (%)↑	W_ρ ↓	W_n ↓	Time (s)↓
FlowMM	93.03	83.15	97.44	85.00	83.99	46.05	31.64	1.389	**0.075**	1560
CrystalDiT	77.82	67.28	90.88	59.33	56.86	**83.41**	41.70	0.202	0.171	73.72
DiffCSP	**99.93**	82.10	96.90	89.53	87.89	50.28	38.60	0.192	0.344	237
MatterGen	99.78	83.72	**98.10**	**91.14**	**90.26**	51.70	42.29	0.088	0.184	2639
ADiT	99.52	**90.15**	90.25	59.80	56.91	76.90	36.76	0.231	0.089	84.81
Crystalite	99.61 ± 0.06	81.72 ± 0.24	95.24 ± 0.19	79.30 ± 0.12	77.33 ± 0.21	69.72 ± 0.85	**47.49** ± 0.77	**0.051** ± 0.010	0.127 ± 0.006	**22.36/5.14**†

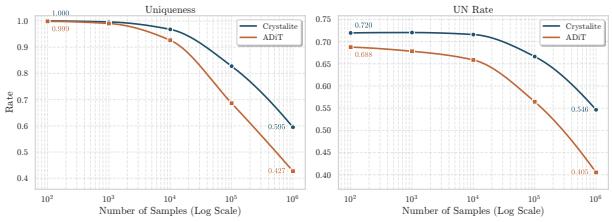


Figure 4. **Large-scale generation.** Uniqueness and unique-and-novel (UN) rate are shown as a function of the number of generated crystals for Crystalite and ADiT. Crystalite consistently preserves more diversity at scale, reaching a higher UN rate at 10^6 samples and higher uniqueness.

(GEM), Crystalite injects crystal-specific inductive bias into a standard Transformer without relying on expensive equivariant message passing.

Across benchmarks, Crystalite achieves state-of-the-art CSP performance and strong DNG, attaining the best SUN score among evaluated baselines while sampling substantially faster than geometry-heavy alternatives. These results suggest that strong crystal modeling does not require full equivariance when periodic geometry and chemical structure are incorporated effectively. Overall, Crystalite offers a simple and efficient approach to scalable materials discovery.

Limitations. Crystalite uses MLIP-based relaxation for scalable evaluation, but the resulting hull-distance estimates are screening proxies rather than substitutes for DFT-level validation.

Future Work. Building on our scalable Transformer backbone, future work will focus on training Crystalite on massive datasets like Alex-MP-20 and extending it to property-conditioned generation for targeted inverse design.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning for materials discovery. Crystalite may accelerate materials discovery for green technologies and improve accessibility through efficient sampling. However, generative materials AI carries dual-use risks, including proposing hazardous compounds, so automated deployment should include human oversight. Beyond these considerations, there are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Albergo, M., Boffi, N. M., and Vanden-Eijnden, E. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. URL http://jmlr.org/papers/v26/23-1605.html.

Baird, S. G., Sayeed, H. M., Montoya, J., and Sparks, T. D. matbench-genmetrics: A python library for benchmarking crystal structure generative models using time-based splits of materials project structures. *Journal of Open Source Software*, 9(97):5618, 2024. doi: 10.21105/joss.05618. URL https://doi.org/10.21105/joss.05618.

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., and Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

Betala, S., Gleason, S. P., Ramlaoui, A., Xu, A., Channing, G., Levy, D., Fourrier, C., Kazeev, N., Joshi, C. K., Kaba, S.-O., Therrien, F., Hernandez-Garcia, A., Mercado, R., Krishnan, N. M. A., and Duval, A. Lemat-genbench: A unified evaluation framework for crystal generative mod-

els, 2026. URL https://arxiv.org/abs/2512.04562.

Cao, Z., Luo, X., Lv, J., and Wang, L. Space group informed transformer for crystalline materials generation. *Science Bulletin*, 70(21):3522–3533, November 2025. ISSN 2095-9273. doi: 10.1016/j.scib.2025.09.035. URL http://dx.doi.org/10.1016/j.scib.2025.09.035.

Chen, R. T. Q. and Lipman, Y. Flow matching on general geometries. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=g7ohDlTITL.

Cornet, F. R. J., Bergamin, F., Bhowmik, A., Garcia-Lastra, J. M., Frellsen, J., and Schmidt, M. N. Kinetic langevin diffusion for crystalline materials generation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=7J1kwZY72h.

Curtarolo, S., Setyawan, W., Hart, G. L. W., Jahnatek, M., Chepulskii, R. V., Taylor, R. H., Wang, S., Xue, J., Yang, K., Levy, O., Mehl, M. J., Stokes, H. T., Demchenko, D. O., and Morgan, D. AFLOW: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226, 2012. doi: 10.1016/j.commatsci.2012.02.005.

Davies, D. W., Butler, K. T., Jackson, A. J., Skelton, J. M., Morita, K., and Walsh, A. Smact: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software*, 4(38):1361, 2019. doi: 10.21105/joss.01361. URL https://doi.org/10.21105/joss.01361.

Gasteiger, J., Becker, F., and Günnemann, S. Gemnet: Universal directional graph neural networks for molecules. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Goedecker, S. Linear scaling electronic structure methods. *Reviews of Modern Physics*, 71(4):1085–1123, 1999. ISSN 1539-0756. doi: 10.1103/revmodphys.71.1085. URL http://dx.doi.org/10.1103/RevModPhys.71.1085.

Gruver, N., Sriram, A., Madotto, A., Wilson, A. G., Zitnick, C. L., and Ulissi, Z. W. Fine-tuned language models generate stable inorganic materials as text. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vN9fpfqoP1.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Höllmer, P., Egg, T., Martirossyan, M., Fuemmeler, E., Shui, Z., Gupta, A., Prakash, P., Roitberg, A., Liu, M., Karypis, G., Transtrum, M., Hennig, R., Tadmor, E. B., and Martiniani, S. Open materials generation with stochastic interpolants. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=gHGrzxFujU.

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K. A. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1), 2013. ISSN 2166-532X. doi: 10.1063/1.4812323. URL http://dx.doi.org/10.1063/1.4812323.

Jiao, R., Huang, W., Lin, P., Han, J., Chen, P., Lu, Y., and Liu, Y. Crystal structure prediction by joint equivariant diffusion on lattices and fractional coordinates. In *Workshop on "Machine Learning for Materials" ICLR 2023*, 2023. URL https://openreview.net/forum?id=VPByphdu24j.

Jiao, R., Huang, W., Liu, Y., Zhao, D., and Liu, Y. Space group constrained crystal generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jkvZ7v4OmP.

Jin, E., Nica, A. C., Galkin, M., Rector-Brooks, J., Lee, K. L. K., Miret, S., Arnold, F. H., Bronstein, M. M., Bose, J., Tong, A., and Liu, C.-H. OXtal: An all-atom diffusion model for organic crystal structure prediction. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6Jd5aBml0y.

Jones, R. Density functional theory: Its origins, rise to prominence, and future. *Reviews of Modern Physics*, 87(3):897–923, August 2015. ISSN 1539-0756. doi: 10.1103/revmodphys.87.897. URL http://dx.doi.org/10.1103/RevModPhys.87.897.

Joshi, C. K., Fu, X., Liao, Y.-L., Gharakhanyan, V., Miller, B. K., Sriram, A., and Ulissi, Z. W. All-atom diffusion transformers: Unified generative modelling of molecules and materials. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=89QPmZjIhv.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a98846e9d9cc01cfb87eb694d946ce6b-Paper-Conference.pdf.

Kazeev, N., Nong, W., Romanov, I., Zhu, R., Ustyuzhanin, A., Yamazaki, S., and Hippalgaonkar, K. Wyckoff transformer: Generation of symmetric crystals, 2025. URL https://arxiv.org/abs/2503.02407.

Khastagir, S., DAS, K., Goyal, P., Lee, S.-C., Bhattacharjee, S., and Ganguly, N. LLM meets diffusion: A hybrid framework for crystal material generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=E6gwPtWjb1.

Kirklin, S., Saal, J. E., Meredig, B., Thompson, A., Doak, J. W., Aykol, M., Rühl, S., and Wolverton, C. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Computational Materials*, 1(1), December 2015. ISSN 2057-3960. doi: 10.1038/npjcompumats.2015.10. URL http://dx.doi.org/10.1038/npjcompumats.2015.10.

Kohn, W. and Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140: A1133–A1138, Nov 1965. doi: 10.1103/PhysRev.140.A1133. URL https://link.aps.org/doi/10.1103/PhysRev.140.A1133.

Levy, D., Panigrahi, S. S., Kaba, S.-O., Zhu, Q., Lee, K. L. K., Galkin, M., Miret, S., and Ravanbakhsh, S. SymmCD: Symmetry-preserving crystal generation with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xnssGv9rpW.

Luo, X., Wang, Z., Wang, Q., Shao, X., Lv, J., Wang, L., Wang, Y., and Ma, Y. Crystalflow: a flow-based generative model for crystalline materials. *Nature Communications*, 16(1), October 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-64364-4. URL http://dx.doi.org/10.1038/s41467-025-64364-4.

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, November 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06735-9. URL http://dx.doi.org/10.1038/s41586-023-06735-9.

Miller, B. K., Chen, R. T. Q., Sriram, A., and Wood, B. M. Flowmm: generating materials with riemannian flow matching. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Mistal, Hernández-García, A., Volokhova, A., Duval, A. A., Bengio, Y., Sharma, D., Carrier, P. L., Koziarski, M., and Schmidt, V. Crystal-GFN: sampling materials with desirable properties and constraints. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023. URL https://openreview.net/forum?id=l167FjdPOv.

Mohanty, T., Mehta, M., Sayeed, H. M., Srikumar, V., and Sparks, T. D. Crystext: A generative ai approach for text-conditioned crystal structure generation using llm. December 2024. doi: 10.26434/chemrxiv-2024-gjhpq. URL http://dx.doi.org/10.26434/chemrxiv-2024-gjhpq.

Morehead, A., Cretu, M., Panescu, A., Anand, R., Weiler, M., Perez, T., Blau, S., Farrell, S., Bhimji, W., Jain, A., Sahasrabuddhe, H., Lio, P., Jaakkola, T., Gomez-Bombarelli, R., Ying, R., Erichson, N. B., and Mahoney, M. W. Zatom-1: A multimodal flow foundation model for 3d molecules and materials, 2026. URL https://arxiv.org/abs/2602.22251.

Oganov, A. R. and Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *The Journal of Chemical Physics*, 124 (24), June 2006. ISSN 1089-7690. doi: 10.1063/1.2210932. URL http://dx.doi.org/10.1063/1.2210932.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Pickard, C. J. and Needs, R. J. Ab initio random structure searching. *Journal of Physics: Condensed Matter*, 23(5): 053201, jan 2011. doi: 10.1088/0953-8984/23/5/053201. URL https://doi.org/10.1088/0953-8984/23/5/053201.

Press, O., Smith, N., and Lewis, M. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=R8sQPpGCv0.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR), pp. 10674–10685. IEEE, 2022. doi: 10.1109/cvpr52688.2022.01042. URL http://dx.doi.org/10.1109/CVPR52688.2022.01042.

Satorras, V. G., Hoogeboom, E., and Welling, M. E(n) equivariant graph neural networks. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Shaw, P., Uszkoreit, J., and Vaswani, A. Self-attention with relative position representations. In Walker, M., Ji, H., and Stent, A. (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2074. URL https://aclanthology.org/N18-2074/.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf.

Sriram, A., Miller, B. K., Chen, R. T. Q., and Wood, B. M. FlowLLM: Flow matching for material generation with large language models as base distributions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0bFXbEMz8e.

Xie, T., Fu, X., Ganea, O.-E., Barzilay, R., and Jaakkola, T. S. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=03RLpj-tc_.

Xu, A., Desai, R., Wang, L., Hope, G., and Ritz, E. Plaid++: A preference aligned language model for targeted inorganic materials design, 2025. URL https://arxiv.org/abs/2509.07150.

Yang, S., Cho, K., Merchant, A., Abbeel, P., Schuurmans, D., Mordatch, I., and Cubuk, E. D. Scalable diffusion for materials generation, 2023. URL https://arxiv.org/abs/2311.09235.

Yi, X., Xu, G., Xiao, X., Zhang, Z., Liu, L., Bian, Y., and Zhao, P. Crystaldit: A diffusion transformer for crystal generation, 2025. URL https://arxiv.org/abs/2508.16614.

Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T.-Y. Do transformers really perform bad for graph representation?, 2021. URL https://arxiv.org/abs/2106.05234.

Zeni, C., Pinsler, R., Zügner, D., Fowler, A., Horton, M., Fu, X., Wang, Z., Shysheya, A., Crabbé, J., Ueda, S., Sordillo, R., Sun, L., Smith, J., Nguyen, B., Schulz, H., Lewis, S., Huang, C.-W., Lu, Z., Zhou, Y., Yang, H., Hao, H., Li, J., Yang, C., Li, W., Tomioka, R., and Xie, T. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, January 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08628-5. URL http://dx.doi.org/10.1038/s41586-025-08628-5.

A. Introduction to Materials

A.1. Unit-cell representation of crystals

A crystalline material is, ideally, an infinite periodic arrangement of atoms in three-dimensional space, as shown in Figure 5. Rather than describing the full solid atom by atom, it suffices to specify a single *unit cell* together with the rule that this cell repeats under integer translations of the lattice. This is the standard representation used throughout the paper.

Concretely, we represent a crystal with N atoms by the triple

$$\mathcal{C} = (\mathbf{A}, \mathbf{F}, \mathbf{L}), \tag{13}$$

where $\mathbf{A} \in \{0,1\}^{N \times N_Z}$ is the atom-type matrix, $\mathbf{F} = [\mathbf{f}_1^\top; \ldots; \mathbf{f}_N^\top] \in [0,1)^{N \times 3}$ contains the fractional coordinates, and $\mathbf{L} \in \mathbb{R}^{3 \times 3}$ is the lattice matrix. Each row of $\mathbf{A}$ satisfies $\mathbf{A}_i = \mathrm{onehot}(a_i)$ for some atomic species $a_i \in \{1, \ldots, N_Z\}$. The pair $(\mathbf{A}, \mathbf{F})$ specifies the basis atoms inside the cell, while $\mathbf{L}$ determines the geometry of the cell itself.

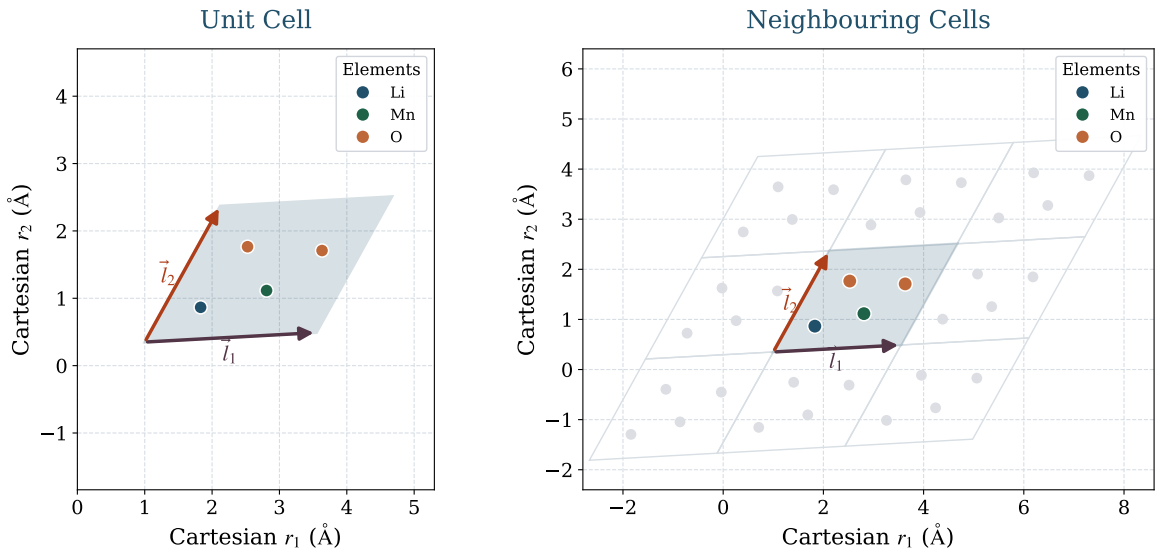


Figure 5. A crystal can be represented by a unit cell together with its periodic repetition under lattice translations.

Fractional and Cartesian coordinates. We use fractional coordinates because they make periodicity explicit. Each row $\mathbf{f}_i \in [0,1)^3$ gives the position of atom i relative to the lattice basis. Under the row-vector convention used in this paper, Cartesian coordinates are obtained by

$$\mathbf{X} = \mathbf{F}\mathbf{L} \in \mathbb{R}^{N \times 3}, \tag{14}$$

so that the Cartesian coordinate of atom i is the i-th row

$$\mathbf{x}_i = \mathbf{f}_i \mathbf{L}. \tag{15}$$

Thus, $\mathbf{L}$ controls the size and shape of the cell, while $\mathbf{F}$ determines where atoms are placed inside it. Figure 6 visualizes this transformation.

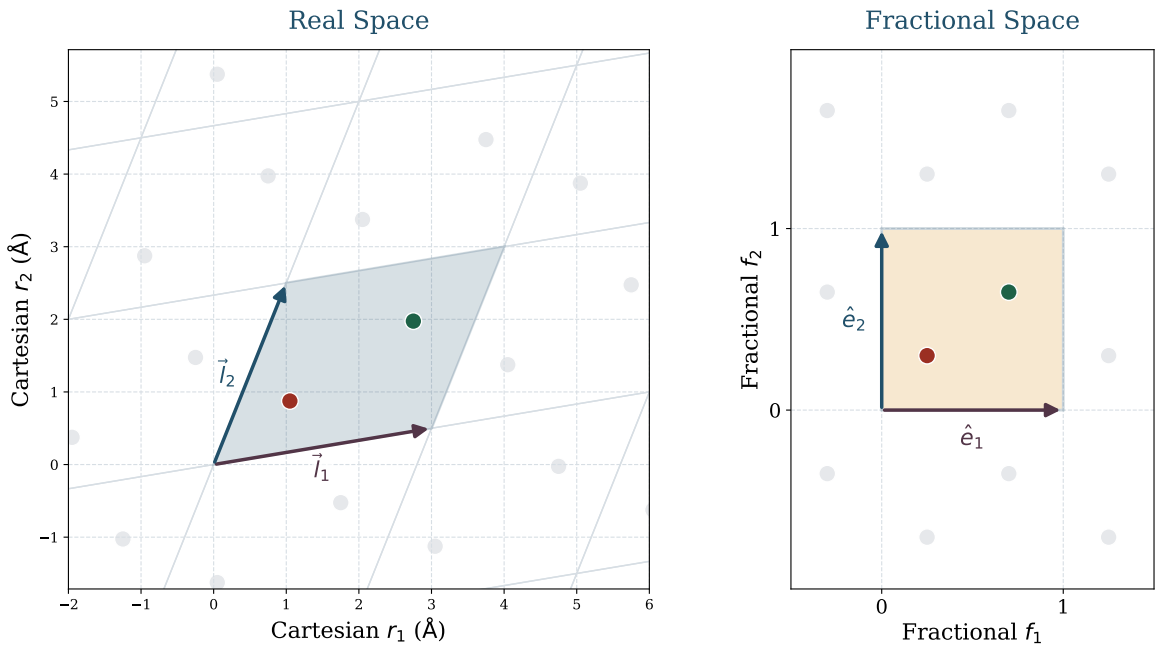


Figure 6. Fractional coordinates are defined in the unit cube and mapped to Cartesian space by the lattice matrix. Integer shifts in fractional coordinates correspond to lattice translations in real space.

Periodic boundary conditions. Fractional coordinates live on the flat torus

$$\mathbb{T}^3 \cong (\mathbb{R}/\mathbb{Z})^3, \tag{16}$$

meaning that $\mathbf{f}$ and $\mathbf{f} + \mathbf{n}$ represent the same physical position for any $\mathbf{n} \in \mathbb{Z}^3$. This is precisely the periodic boundary condition: atoms leaving one face of the unit cell re-enter through the opposite face.

The full infinite crystal is therefore generated by translating each basis atom by all integer lattice shifts:

$$\mathbf{x}_{i,\mathbf{n}} = (\mathbf{f}_i + \mathbf{n})\mathbf{L}, \qquad \mathbf{n} \in \mathbb{Z}^3. \tag{17}$$

A finite unit-cell description thus implicitly defines the entire periodic material.

Wrapped residuals and metric-aware minimum-image geometry. Because fractional coordinates are periodic, geometric quantities must respect the torus structure. In the coordinate loss, we use the componentwise wrapped residual in fractional space,

$$\boldsymbol{\delta}_{ij}^{\text{wrap}} = \text{wrap}(\mathbf{f}_i - \mathbf{f}_j), \qquad \text{wrap}(\mathbf{u}) = \mathbf{u} - \text{round}(\mathbf{u}), \tag{18}$$

so that each component of $\boldsymbol{\delta}_{ij}^{\text{wrap}}$ lies in $[-\frac{1}{2}, \frac{1}{2})$. The associated Cartesian displacement and distance are

$$\mathbf{r}_{ij}^{\text{wrap}} = \boldsymbol{\delta}_{ij}^{\text{wrap}}\mathbf{L}, \qquad d_{ij}^{\text{wrap}} = \|\mathbf{r}_{ij}^{\text{wrap}}\|_2. \tag{19}$$

In the Geometry Enhancement Module (GEM), however, we do not use componentwise wrapping. Instead, we use a metric-aware periodic-image search under the lattice metric. Writing

$$\mathbf{G} = \mathbf{L}\mathbf{L}^\top, \tag{20}$$

and restricting the search to a finite set of lattice offsets $\Omega_R = \{-R, \ldots, R\}^3$, we define

$$\Delta\mathbf{f}_{ij}^\star = \arg\min_{\mathbf{r}\in\Omega_R} (\mathbf{f}_i - \mathbf{f}_j + \mathbf{r})\, \mathbf{G}\, (\mathbf{f}_i - \mathbf{f}_j + \mathbf{r})^\top, \tag{21}$$

with corresponding Cartesian displacement and distance

$$\mathbf{r}_{ij}^\star = \Delta\mathbf{f}_{ij}^\star\mathbf{L}, \qquad d_{ij}^\star = \|\mathbf{r}_{ij}^\star\|_2. \tag{22}$$

For orthogonal cells these two constructions coincide, but for general non-orthogonal cells they need not be equivalent. Throughout the paper, we therefore distinguish between the wrapped fractional residual used in the coordinate loss and the metric-aware minimum-image geometry used in GEM. When we refer to minimum-image geometry, we mean the latter construction.

A.2. Symmetries and representation non-uniqueness

The same physical crystal can admit multiple equivalent representations. As a result, the target distribution over crystals should respect several symmetries. In the notation of the main paper, these can be expressed directly in terms of $(\mathbf{A}, \mathbf{F}, \mathbf{L})$.

Permutation of atom indices. The ordering of atoms inside the unit cell is arbitrary. For any permutation matrix $P \in \mathcal{P}_N$,

$$p(\mathbf{A}, \mathbf{F}, \mathbf{L}) = p(P\mathbf{A},\ P\mathbf{F},\ \mathbf{L}). \tag{23}$$

Global rotation in Cartesian space. A rigid rotation of the entire crystal changes only the Cartesian frame, not the underlying material. Under our row-vector convention, this corresponds to right multiplication of the lattice matrix. For any rotation $R \in SO(3)$,

$$p(\mathbf{A}, \mathbf{F}, \mathbf{L}) = p(\mathbf{A},\ \mathbf{F},\ \mathbf{L}R). \tag{24}$$

Permutation of the lattice basis. The choice of lattice basis vectors is not unique. Permuting the lattice basis while applying the inverse permutation to the fractional coordinates leaves the Cartesian crystal unchanged. For any $S \in \mathcal{P}_3$,

$$p(\mathbf{A}, \mathbf{F}, \mathbf{L}) = p(\mathbf{A},\ \mathbf{F}S^\top,\ S\mathbf{L}). \tag{25}$$

Global translation on the torus. Shifting all fractional coordinates by the same torus element does not change the crystal. For any $\mathbf{t} \in \mathbb{T}^3$,

$$p(\mathbf{A}, \mathbf{F}, \mathbf{L}) = p\big(\mathbf{A},\ \mathrm{wrap}(\mathbf{F} + \mathbf{1}\mathbf{t}^\top),\ \mathbf{L}\big), \tag{26}$$

where $\mathbf{1} \in \mathbb{R}^N$ denotes the all-ones vector.

These symmetries motivate several of the design choices in Crystalite. In particular, we represent positions in fractional coordinates, use wrapped periodic residuals for coordinate denoising, use metric-aware minimum-image geometry in GEM, and apply random global translations during training to encourage approximate translation equivariance.

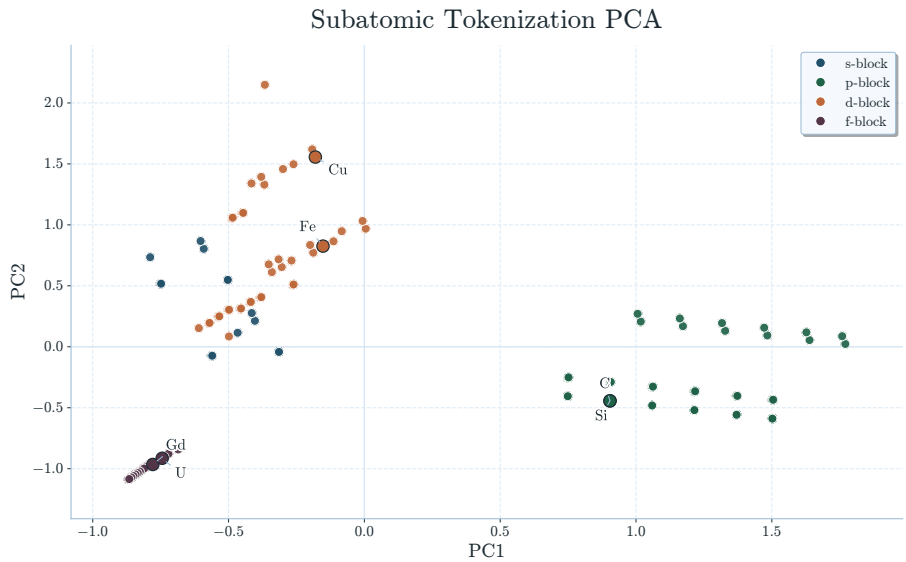


Figure 7. **Two-dimensional PCA projection of the chemically structured atom tokens.** Each point corresponds to one supported element after projecting the balanced descriptors onto the first two principal components. The projection shows that the tokenization preserves meaningful chemical organization even after strong dimensionality reduction.

B. Chemically Structured Atom Tokens

B.1. Chemically Structured Atom Tokens

We replace the usual one-hot atom identity with a continuous token that encodes basic chemical structure while still allowing deterministic decoding back to a valid element. The construction starts from simple periodic-table information and valence-shell occupancies, then standardizes and balances these features before optionally compressing them with PCA.

Let $a_i \in \{1, \dots, N_Z\}$ denote the atomic number at site i. For each supported element $z \in \{1, \dots, N_Z\}$, we build a descriptor from four ingredients: its period, its group, its block, and its ground-state valence-shell occupancies. Concretely, let $r(z) \in \{1, \dots, 7\}$ be the period, $g(z) \in \{0, \dots, 18\}$ the group, where $g(z) = 0$ is reserved for f-block elements, and $b(z) \in \{s, p, d, f\}$ the block. Let (s_z, p_z, d_z, f_z) denote the corresponding valence occupancies from a fixed lookup table. We then define the raw descriptor

$$\mathbf{d}_z = \Big[\mathrm{onehot}_7\big(r(z)-1\big),\ \mathrm{onehot}_{19}\big(g(z)\big),\ \mathrm{onehot}_4\big(b(z)\big),\ s_z/2,\ p_z/6,\ d_z/10,\ f_z/14\Big]. \tag{27}$$

In our implementation this gives a 34-dimensional vector, since

$$7 + 19 + 4 + 4 = 34.$$

Because these feature groups have different dimensionalities, we standardize each coordinate across the supported elements and then rebalance the groups so that large one-hot blocks do not dominate purely because they contain more entries. Let

$$D = \begin{bmatrix} \mathbf{d}_1^\top \\ \vdots \\ \mathbf{d}_{N_Z}^\top \end{bmatrix} \in \mathbb{R}^{N_Z \times 34}$$

collect the raw descriptors for all elements. We compute the featurewise mean and standard deviation,

$$\boldsymbol{\mu} = \frac{1}{N_Z} \sum_{z=1}^{N_Z} \mathbf{d}_z, \qquad \boldsymbol{\sigma} = \mathrm{std}(D),$$

and form the standardized descriptor

$$\tilde{\mathbf{d}}_z = (\mathbf{d}_z - \boldsymbol{\mu}) \oslash \boldsymbol{\sigma}, \tag{28}$$

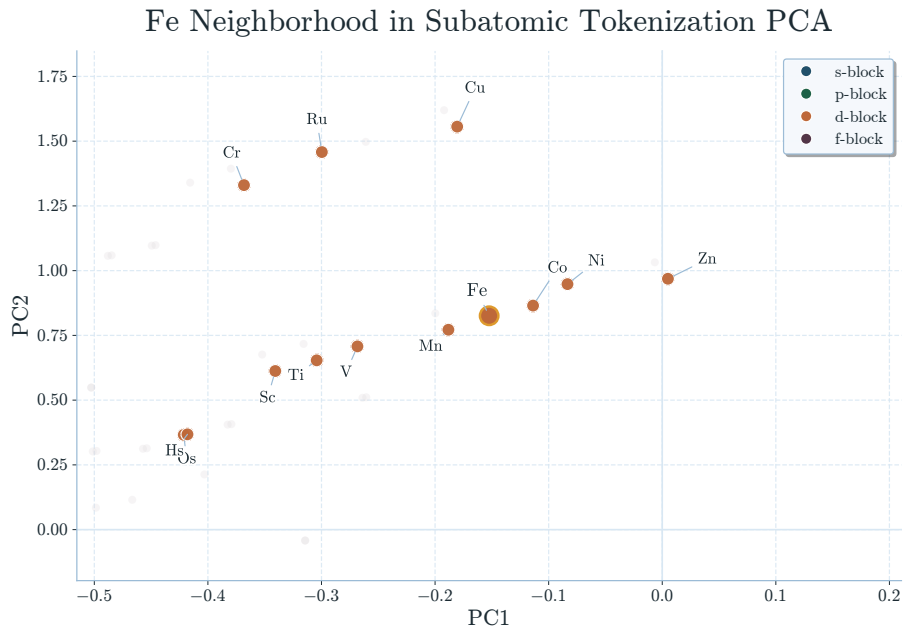


Figure 8. **Local neighborhood of Fe in the two-dimensional PCA space.** The plot highlights Fe together with its nearest elements in the projected representation, illustrating how the learned token geometry places chemically related species close to one another.

where $\oslash$ denotes elementwise division. Any near-zero entry of $\boldsymbol{\sigma}$ is replaced by 1 for numerical stability.

We next split $\tilde{\mathbf{d}}_z$ into the four groups

$$\text{period } (7), \qquad \text{group } (19), \qquad \text{block } (4), \qquad \text{valence } (4),$$

and rescale each group by the inverse square root of its dimensionality. If $\tilde{\mathbf{d}}_z^{(G)}$ denotes the subvector corresponding to group G, we define

$$\bar{\mathbf{d}}_z^{(G)} = |G|^{-1/2}\, \tilde{\mathbf{d}}_z^{(G)}. \tag{29}$$

Concatenating the reweighted groups gives the balanced descriptor $\bar{\mathbf{d}}_z$. The final raw token is then obtained by ℓ_2-normalization,

$$\mathbf{h}_z = \frac{\bar{\mathbf{d}}_z}{\|\bar{\mathbf{d}}_z\|_2}. \tag{30}$$

For a crystal with atomic numbers $(a_1, \ldots, a_N)$, the atom-type channel becomes

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_{a_1}^\top \\ \vdots \\ \mathbf{h}_{a_N}^\top \end{bmatrix} \in \mathbb{R}^{N \times d_H}, \tag{31}$$

with $d_H = 34$ in the raw representation.

When a lower-dimensional token is preferred, we apply PCA to the balanced descriptors. Let

$$\bar{D} = \begin{bmatrix} \bar{\mathbf{d}}_1^\top \\ \vdots \\ \bar{\mathbf{d}}_{N_Z}^\top \end{bmatrix} \in \mathbb{R}^{N_Z \times 34},$$

and let $U_d \in \mathbb{R}^{34 \times d}$ contain the top d principal directions. Each element is then represented by

$$\mathbf{p}_z = \bar{\mathbf{d}}_z U_d \in \mathbb{R}^d, \qquad \mathbf{h}_z^{\text{PCA}} = \frac{\mathbf{p}_z}{\|\mathbf{p}_z\|_2}. \tag{32}$$

This gives a compressed tokenization with $d_H = d$. A two-dimensional PCA projection of the element tokens is shown in Figure 7. Even in two dimensions, the representation retains visible chemical structure. Figure 8 shows the local neighborhood of Fe in this projected space, which provides an intuitive view of how chemically related elements cluster around it.

Finally, both the raw and PCA-compressed tokens can be decoded deterministically by nearest-prototype matching. Given a predicted continuous token $\hat{\mathbf{h}}_i$, we assign the atomic species as

$$\hat{a}_i = \arg\max_{z \in \{1,\ldots,N_Z\}} \langle \hat{\mathbf{h}}_i, \mathbf{h}_z^\star \rangle, \tag{33}$$

where $\mathbf{h}_z^\star$ is either the raw prototype $\mathbf{h}_z$ or the PCA-compressed prototype $\mathbf{h}_z^{\mathrm{PCA}}$. Since all prototypes are normalized, this is equivalent to cosine-similarity decoding.

C. Crystalite Architecture

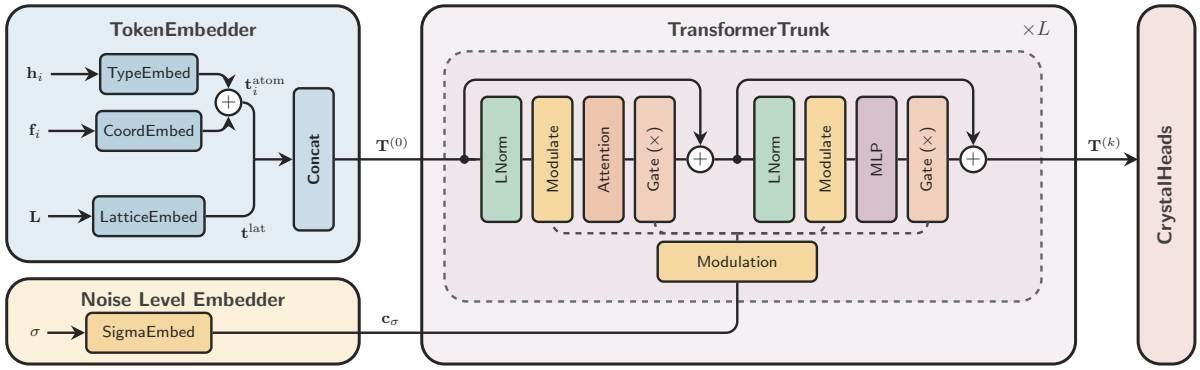


Figure 9. Overview of the Crystalite architecture. The model operates on the continuous crystal state $(\mathbf{H}, \mathbf{F}, \mathbf{y})$. Atom-type and coordinate embeddings are added to form one token per atom, while the lattice embedding produces a single global lattice token. The resulting sequence is processed by an AdaLN-conditioned Transformer trunk, and output heads predict $\hat{\mathbf{H}}$, $\hat{\mathbf{F}}$, and $\hat{\mathbf{y}}$.

This appendix provides a more detailed description of Crystalite using the notation of the main text. Recall that a crystal is represented as

$$\mathcal{C} = (\mathbf{A}, \mathbf{F}, \mathbf{L}),$$

and that the diffusion model operates on the continuous state

$$(\mathbf{H}, \mathbf{F}, \mathbf{y}),$$

where $\mathbf{H}$ denotes the chemically structured atom tokens obtained from $\mathbf{A}$, and $\mathbf{y} \in \mathbb{R}^6$ is the lower-triangular lattice parameterization satisfying $\mathbf{L} = \mathbf{L}(\mathbf{y})$. Figure 9 gives an overview of the full architecture, while Figure 10 illustrates the Geometry Enhancement Module (GEM).

C.1. Tokenization and input embeddings

Each atomic site i contributes one token to the Transformer sequence. The chemically structured atom token $\mathbf{H}_i \in \mathbb{R}^{d_H}$ is first mapped to the model dimension through a learned embedder E_H,

$$\mathbf{h}_i^H = E_H(\mathbf{H}_i), \tag{34}$$

where E_H is implemented as a two-layer MLP with SiLU activation acting directly on the continuous atom token:

$$E_H : \mathbb{R}^{d_H} \to \mathbb{R}^d, \qquad \text{Linear}(d_H, d) \;\to\; \text{SiLU} \;\to\; \text{Linear}(d, d).$$

The corresponding fractional coordinate $\mathbf{f}_i \in [0, 1)^3$ is embedded separately through

$$\mathbf{h}_i^F = E_F(\mathbf{f}_i) = \text{MLP}_F\big(\gamma_F(\mathbf{f}_i)\big), \tag{35}$$

where γ_F denotes a deterministic Fourier feature map. Concretely, we use sinusoidal features at multiple frequencies,

$$\gamma_F(\mathbf{f}_i) = \big[\sin(2\pi\ell\mathbf{f}_i),\, \cos(2\pi\ell\mathbf{f}_i)\big]_{\ell=1}^{n_F},$$

followed by a two-layer MLP with SiLU activation. Thus E_F has the form

$$E_F : \mathbb{R}^{6n_F} \to \mathbb{R}^d, \qquad \text{Linear}(6n_F, d) \;\to\; \text{SiLU} \;\to\; \text{Linear}(d, d),$$

with $n_F = 32$ in the base configuration. The resulting atom token is

$$\mathbf{t}_i^{\text{atom}} = E_H(\mathbf{H}_i) + E_F(\mathbf{f}_i). \tag{36}$$

The lattice is represented by a single global token. The lattice latent $\mathbf{y} \in \mathbb{R}^6$ is the lower-triangular parameterization introduced in Eq. (5), and is embedded through

$$\mathbf{t}^{\text{lat}} = E_{\text{lat}}(\mathbf{y}), \tag{37}$$

where E_{lat} is implemented as a two-layer MLP with SiLU activation acting directly on $\mathbf{y}$:

$$E_{\text{lat}} : \mathbb{R}^6 \to \mathbb{R}^d, \qquad \text{Linear}(6, d) \;\to\; \text{SiLU} \;\to\; \text{Linear}(d, d).$$

For a crystal with N atoms, the initial Transformer sequence is therefore

$$\mathbf{T}^{(0)} = \left[\mathbf{t}_1^{\text{atom}}, \ldots, \mathbf{t}_N^{\text{atom}}, \mathbf{t}^{\text{lat}}\right] \in \mathbb{R}^{(N+1)\times d}. \tag{38}$$

Thus Crystalite uses one token per atom, together with one additional token that summarizes the global unit-cell geometry.

The diffusion noise level is embedded through the standard EDM noise coordinate

$$c_{\text{noise}}(\sigma) = \tfrac{1}{4} \log \sigma, \tag{39}$$

followed by a learned embedder E_σ, giving a conditioning vector

$$\mathbf{c}_\sigma = E_\sigma\big(c_{\text{noise}}(\sigma)\big). \tag{40}$$

This conditioning is injected into every Transformer block through adaptive layer normalization (AdaLN).

The token sequence is then processed by a standard Transformer trunk with stacked self-attention and feed-forward blocks. Writing $\mathbf{T}^{(k)}$ for the token sequence entering block k, the update can be written schematically as

$$\mathbf{T}^{(k+\frac{1}{2})} = \mathbf{T}^{(k)} + \text{MHA}^{(k)}\big(\mathbf{T}^{(k)}; \mathbf{c}_\sigma, \widetilde{\mathbf{B}}^{(k)}\big), \tag{41}$$

$$\mathbf{T}^{(k+1)} = \mathbf{T}^{(k+\frac{1}{2})} + \text{MLP}^{(k)}\big(\mathbf{T}^{(k+\frac{1}{2})}; \mathbf{c}_\sigma\big), \tag{42}$$

where $\widetilde{\mathbf{B}}^{(k)}$ denotes the optional additive attention bias produced by GEM. When GEM is disabled, $\widetilde{\mathbf{B}}^{(k)} = 0$ and the model reduces to a standard AdaLN-conditioned diffusion Transformer.

After the final block, shallow output heads map the updated atom tokens to denoised atom-type and coordinate predictions, and the lattice token to the denoised lattice latent:

$$\hat{\mathbf{H}}_i = D_H(\mathbf{t}_i^{(K)}), \qquad \hat{\mathbf{f}}_i = D_F(\mathbf{t}_i^{(K)}), \qquad \hat{\mathbf{y}} = D_{\text{lat}}(\mathbf{t}_{\text{lat}}^{(K)}). \tag{43}$$

Thus atom-wise quantities are predicted from the site tokens, while the global lattice parameters are predicted from the lattice token.

C.2. Geometry Enhancement Module (GEM)

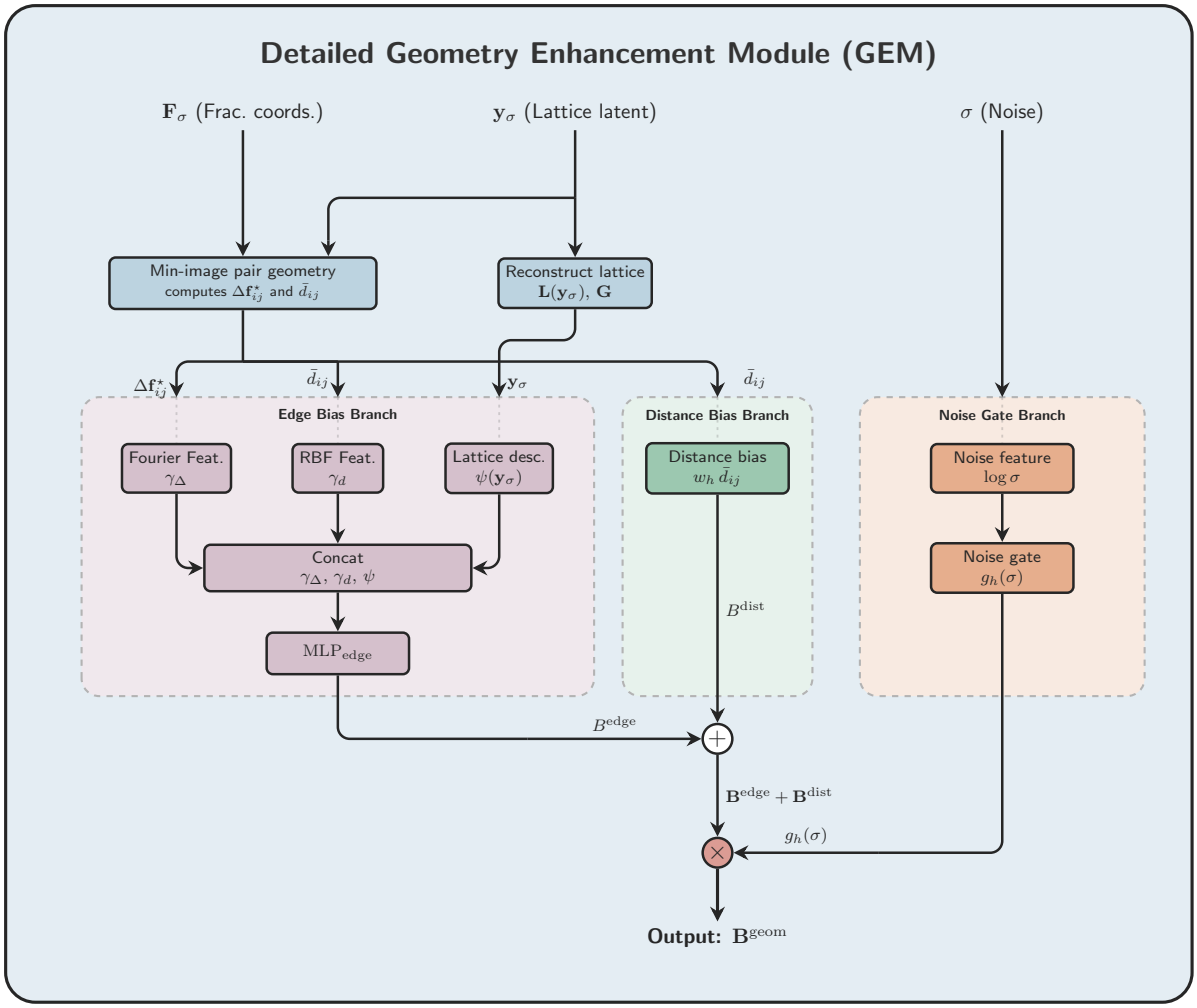


Figure 10. Detailed view of the Geometry Enhancement Module (GEM). Starting from the current fractional coordinates and lattice parameters, GEM computes periodic pairwise geometry under minimum-image conventions, converts it into distance and edge-aware attention biases, and injects the resulting signal additively into the attention logits.

GEM augments self-attention with pairwise geometric biases derived from the current crystal geometry. It does not change the tokenization or prediction heads; instead, it modifies the attention logits through an additive bias tensor.

Given the current fractional coordinates $\mathbf{F}$ and lattice latent $\mathbf{y}$, GEM first reconstructs the lattice matrix $\mathbf{L}(\mathbf{y})$ and computes pairwise minimum-image geometry under periodic boundary conditions. Let

$$\mathbf{G}(\mathbf{y}) = \mathbf{L}(\mathbf{y})\mathbf{L}(\mathbf{y})^\top \tag{44}$$

denote the corresponding metric tensor. For each pair of atoms (i, j), we consider periodic offsets $\mathbf{r} \in \Omega_R = \{-R, \ldots, R\}^3$ and define

$$\Delta\mathbf{f}_{ij}(\mathbf{r}) = \mathbf{f}_i - \mathbf{f}_j + \mathbf{r}. \tag{45}$$

The minimum-image displacement is then chosen as

$$\Delta\mathbf{f}^\star_{ij} = \arg\min_{\mathbf{r}\in\Omega_R} \Delta\mathbf{f}_{ij}(\mathbf{r})\,\mathbf{G}(\mathbf{y})\,\Delta\mathbf{f}_{ij}(\mathbf{r})^\top, \tag{46}$$

with corresponding Cartesian distance

$$d_{ij} = \left\|\Delta \mathbf{f}_{ij}^{\star} \mathbf{L}(\mathbf{y})\right\|_2. \tag{47}$$

In practice, this distance is normalized by a characteristic cell scale $s(\mathbf{y})$, yielding $\bar{d}_{ij} = d_{ij}/s(\mathbf{y})$.

From this pairwise geometry, GEM builds two additive bias terms. The first is a distance bias,

$$B_{hij}^{\text{dist}} = \alpha_h \, \bar{d}_{ij}, \qquad \alpha_h \leq 0, \tag{48}$$

which acts as a learnable locality prior for each attention head h. The second is an edge-aware bias produced by a small MLP acting on periodic pairwise features,

$$\phi_{ij} = \left[\gamma_\Delta(\Delta \mathbf{f}_{ij}^{\star}), \gamma_d(\bar{d}_{ij}), \psi(\mathbf{y})\right], \qquad B_{hij}^{\text{edge}} = \text{MLP}_{\text{edge}}(\phi_{ij})_h, \tag{49}$$

where γ_Δ and γ_d denote Fourier/RBF feature maps and $\psi(\mathbf{y})$ is a low-dimensional lattice descriptor.

The two branches are combined, optionally modulated by a noise-dependent gate,

$$B_{hij}^{(k)} = g_h(\sigma)\Big(B_{hij}^{\text{dist}} + B_{hij}^{\text{edge}}\Big), \tag{50}$$

and then expanded from atom pairs to the full token sequence by leaving lattice-token interactions unbiased:

$$\widetilde{\mathbf{B}}_h^{(k)} = \begin{bmatrix} \mathbf{B}_h^{(k)} & \mathbf{0} \\ \mathbf{0}^\top & 0 \end{bmatrix}. \tag{51}$$

Finally, this bias is added directly to the attention logits,

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} + \widetilde{\mathbf{B}}^{(k)}\right) V. \tag{52}$$

This construction lets Crystalite inject periodic geometric information directly into attention while preserving the simplicity of a standard Transformer backbone. When GEM is disabled, the model uses the same tokenization, diffusion objective, and output heads, but with $\widetilde{\mathbf{B}}^{(k)} = 0$.

C.3. Base model configuration

Unless otherwise stated, the main MP-20 DNG results in the paper use the base Crystalite configuration summarized in Table 3. This instantiation contains approximately 6.7×10^7 trainable parameters. It uses a 14-layer Transformer trunk with model width $d = 512$ and 16 attention heads, together with PCA-compressed Chemically Structured Atom Tokens with token dimension $d_H = 16$.

Table 3. Base Crystalite configuration used for the main MP-20 DNG results.

(a) Architecture

Component	Setting
Trainable parameters	$\sim$ 67M
Transformer width d	512
Transformer layers	14
Attention heads	16
Dropout / attn. dropout	0 / 0
Atom tokenization	Chemically Structured, PCA $d_H = 16$
Coordinate embedding	Fourier, 32 freqs.
Coordinate head	direct fractional head
GEM	enabled
GEM sharing	shared across layers
PBC search radius R	1
Distance bias	enabled
Edge-aware bias	enabled
Edge-bias hidden dim.	256
Edge-bias Fourier freqs.	12
Edge-bias RBF features	32
Noise-dependent gate	enabled

(b) Training and sampling

Component	Setting
Batch size	128
Learning rate	10^{-4}
Weight decay	0
EMA decay	0.9999
LR warmup	1000 steps
Training steps	2.5×10^6
Precision	bfloat16
EDM $(P_{\mathrm{mean}}, P_{\mathrm{std}})$	$(-1.2,\ 1.2)$
σ_{data} for all channels	0.3
Loss weights $(\lambda_H, \lambda_F, \lambda_{\mathrm{lat}})$	$(1,\ 50,\ 5)$
Sampling steps	150
$[\sigma_{\min}, \sigma_{\max}]$	$[0.002,\ 80]$
$(S_{\mathrm{churn}}, S_{\mathrm{noise}})$	$(60,\ 1.003)$
$(S_{\min}, S_{\max})$	$(0,\ 999)$
Atom-count strategy	empirical
Max atoms per cell	20
Sampling weights	EMA

These settings define the base model used throughout the main experiments. The broader implementation supports alternative tokenizations, embedding variants, and GEM configurations, but the specification above corresponds to the principal model reported in the paper.

D. EDM Training Details

EDM noising and preconditioning. At each training step, we sample a noise level according to

$$\log \sigma \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2). \tag{53}$$

Following the notation of the main text, the diffusion model operates on the continuous crystal state

$$(\mathbf{H}, \mathbf{F}, \mathbf{y}),$$

where $\mathbf{H}$ denotes the chemically structured atom tokens, $\mathbf{F} \in [0,1)^{N\times 3}$ the fractional coordinates, and $\mathbf{y} \in \mathbb{R}^6$ the lattice latent.

The atom-token and lattice channels are noised directly in Euclidean space, while the coordinate channel is noised in a centered representation. Concretely, we define

$$\mathbf{F}_{\text{c}} = \mathbf{F} - \tfrac{1}{2}, \tag{54}$$

and then sample

$$\widetilde{\mathbf{H}} = \mathbf{H} + \sigma \boldsymbol{\varepsilon}_H, \qquad \widetilde{\mathbf{F}}_{\text{c}} = \mathbf{F}_{\text{c}} + \sigma \boldsymbol{\varepsilon}_F, \qquad \widetilde{\mathbf{y}} = \mathbf{y} + \sigma \boldsymbol{\varepsilon}_{\text{lat}}, \tag{55}$$

with independent Gaussian noise terms. Before the coordinate embedder, the noisy centered coordinates are shifted back and wrapped into the unit cube,

$$\widetilde{\mathbf{F}}_{\text{in}} = \text{mod1}\left(\widetilde{\mathbf{F}}_{\text{c}} + \tfrac{1}{2}\right), \qquad \text{mod1}(\mathbf{u}) = \mathbf{u} - \lfloor \mathbf{u} \rfloor. \tag{56}$$

The noise level is provided to the Transformer through the usual EDM conditioning scalar

$$c_{\text{noise}}(\sigma) = \tfrac{1}{4} \log \sigma. \tag{57}$$

For each channel $u \in \{H, F, \text{lat}\}$, we use the standard EDM preconditioning coefficients

$$c_{\text{skip},u}(\sigma) = \frac{\sigma_{\text{data},u}^2}{\sigma^2 + \sigma_{\text{data},u}^2}, \qquad c_{\text{out},u}(\sigma) = \frac{\sigma\, \sigma_{\text{data},u}}{\sqrt{\sigma^2 + \sigma_{\text{data},u}^2}}, \qquad c_{\text{in},u}(\sigma) = \frac{1}{\sqrt{\sigma^2 + \sigma_{\text{data},u}^2}}. \tag{58}$$

In our implementation, the atom-token and lattice channels are scaled by $c_{\text{in},u}(\sigma)$ before being passed to the network, whereas the coordinate channel is passed as wrapped fractional coordinates $\widetilde{\mathbf{F}}_{\text{in}}$. Denoting the raw network outputs by $\mathbf{R}_H$, $\mathbf{R}_F$, and $\mathbf{R}_{\text{lat}}$, the corresponding denoised predictions are

$$\hat{\mathbf{H}} = c_{\text{skip},H}(\sigma)\, \widetilde{\mathbf{H}} + c_{\text{out},H}(\sigma)\, \mathbf{R}_H, \tag{59}$$

$$\hat{\mathbf{F}}_{\text{c}} = c_{\text{skip},F}(\sigma)\, \widetilde{\mathbf{F}}_{\text{c}} + c_{\text{out},F}(\sigma)\, \mathbf{R}_F, \tag{60}$$

$$\hat{\mathbf{y}} = c_{\text{skip,lat}}(\sigma)\, \widetilde{\mathbf{y}} + c_{\text{out,lat}}(\sigma)\, \mathbf{R}_{\text{lat}}. \tag{61}$$

For the coordinate loss, we map the centered prediction back to fractional coordinates,

$$\hat{\mathbf{F}} = \text{mod1}\left(\hat{\mathbf{F}}_{\text{c}} + \tfrac{1}{2}\right), \tag{62}$$

and then compute the wrapped fractional residual

$$\Delta_i = \text{wrap}(\hat{\mathbf{f}}_i - \mathbf{f}_i), \qquad \text{wrap}(\mathbf{u}) = \mathbf{u} - \text{round}(\mathbf{u}),$$

so that each component lies in $[-\frac{1}{2}, \frac{1}{2})$. This is a torus-aware residual in fractional space, not the metric-aware minimum-image displacement used in GEM.

Finally, the EDM loss weights are

$$w_u(\sigma) = \frac{\sigma^2 + \sigma_{\text{data},u}^2}{(\sigma\, \sigma_{\text{data},u})^2}, \qquad u \in \{H, F, \text{lat}\}. \tag{63}$$

These are the weights used in the channel-wise training objective described in the main text.

D.1. Channel-wise anti-annealing during sampling

We write the sampler state at step i as

$$\mathbf{z}_i = (\mathbf{H}_i, \mathbf{F}_i, \mathbf{y}_i), \qquad i = 0, \ldots, N,$$

along a decreasing EDM noise schedule

$$\sigma_0 > \sigma_1 > \cdots > \sigma_{N-1} > \sigma_N = 0,$$

with

$$\sigma_i = \left(\sigma_{\max}^{1/\rho} + \frac{i}{N-1}\left(\sigma_{\min}^{1/\rho} - \sigma_{\max}^{1/\rho}\right)\right)^{\rho}, \qquad i = 0, \ldots, N-1. \tag{64}$$

As in EDM, we optionally apply churn at step i, defining

$$\bar{\sigma}_i = (1 + \gamma_i)\sigma_i, \tag{65}$$

and the corresponding perturbed state

$$(\bar{\mathbf{H}}_i, \bar{\mathbf{F}}_i, \bar{\mathbf{y}}_i) = (\mathbf{H}_i, \mathbf{F}_i, \mathbf{y}_i) + \sqrt{\bar{\sigma}_i^2 - \sigma_i^2}\,(\boldsymbol{\varepsilon}_i^H, \boldsymbol{\varepsilon}_i^F, \boldsymbol{\varepsilon}_i^y), \tag{66}$$

where the noise tensors have the appropriate shapes.

We then evaluate the denoiser at $\bar{\sigma}_i$,

$$(\mathbf{H}_i^{\text{den}}, \mathbf{F}_i^{\text{den}}, \mathbf{y}_i^{\text{den}}) = D_\theta(\bar{\mathbf{H}}_i, \bar{\mathbf{F}}_i, \bar{\mathbf{y}}_i, \bar{\sigma}_i). \tag{67}$$

The corresponding EDM drifts are

$$\mathbf{d}_i^H = \frac{\bar{\mathbf{H}}_i - \mathbf{H}_i^{\text{den}}}{\bar{\sigma}_i}, \tag{68}$$

$$\mathbf{d}_i^F = \frac{\text{wrap}(\bar{\mathbf{F}}_i - \mathbf{F}_i^{\text{den}})}{\bar{\sigma}_i}, \tag{69}$$

$$\mathbf{d}_i^y = \frac{\bar{\mathbf{y}}_i - \mathbf{y}_i^{\text{den}}}{\bar{\sigma}_i}, \tag{70}$$

where $\text{wrap}(\mathbf{u}) = \mathbf{u} - \text{round}(\mathbf{u})$ is applied elementwise to respect periodicity in fractional coordinates.

To anti-anneal a selected channel $q \in \{H, F, y\}$, we introduce an auxiliary Karras schedule

$$\tilde{\sigma}_i^{(q)} = \left(\sigma_{\max}^{1/\rho_q^{\text{AA}}} + \frac{i}{N-1}\left(\sigma_{\min}^{1/\rho_q^{\text{AA}}} - \sigma_{\max}^{1/\rho_q^{\text{AA}}}\right)\right)^{\rho_q^{\text{AA}}}, \qquad i = 0, \ldots, N-1. \tag{71}$$

Writing

$$\Delta_i = \sigma_i - \sigma_{i+1}, \qquad \tilde{\Delta}_i^{(q)} = \tilde{\sigma}_i^{(q)} - \tilde{\sigma}_{i+1}^{(q)},$$

we define the anti-annealing factor

$$\alpha_i^{(q)} = \max\left(1,\ \frac{\tilde{\Delta}_i^{(q)}}{\Delta_i}\right). \tag{72}$$

If anti-annealing is disabled for channel q, we set $\alpha_i^{(q)} = 1$. For fractional coordinates, we may additionally cap this factor,

$$\alpha_i^{(F)} \leftarrow \min\!\big(\alpha_i^{(F)}, \alpha_{\max}\big). \tag{73}$$

Let

$$\mathbf{z}^{(H)} = \mathbf{H}, \qquad \mathbf{z}^{(F)} = \mathbf{F}, \qquad \mathbf{z}^{(y)} = \mathbf{y}.$$

The Euler predictor step is then

$$\mathbf{z}_{i+1}^{(q),\mathrm{pred}} = \bar{\mathbf{z}}_i^{(q)} + (\sigma_{i+1} - \bar{\sigma}_i)\, \alpha_i^{(q)}\, \mathbf{d}_i^{(q)}, \qquad q \in \{H, F, y\}. \tag{74}$$

When $\sigma_{i+1} > 0$, we apply the usual Heun correction. We first evaluate the denoiser at the predicted state,

$$(\mathbf{H}_{i+1}^{\mathrm{den}}, \mathbf{F}_{i+1}^{\mathrm{den}}, \mathbf{y}_{i+1}^{\mathrm{den}}) = D_\theta(\mathbf{H}_{i+1}^{\mathrm{pred}}, \mathbf{F}_{i+1}^{\mathrm{pred}}, \mathbf{y}_{i+1}^{\mathrm{pred}}, \sigma_{i+1}), \tag{75}$$

and define corrected drifts

$$\mathbf{d}_{i+1}^{H} = \frac{\mathbf{H}_{i+1}^{\mathrm{pred}} - \mathbf{H}_{i+1}^{\mathrm{den}}}{\sigma_{i+1}}, \tag{76}$$

$$\mathbf{d}_{i+1}^{F} = \frac{\mathrm{wrap}(\mathbf{F}_{i+1}^{\mathrm{pred}} - \mathbf{F}_{i+1}^{\mathrm{den}})}{\sigma_{i+1}}, \tag{77}$$

$$\mathbf{d}_{i+1}^{y} = \frac{\mathbf{y}_{i+1}^{\mathrm{pred}} - \mathbf{y}_{i+1}^{\mathrm{den}}}{\sigma_{i+1}}. \tag{78}$$

The final Heun update becomes

$$\mathbf{z}_{i+1}^{(q)} = \bar{\mathbf{z}}_i^{(q)} + (\sigma_{i+1} - \bar{\sigma}_i)\, \alpha_i^{(q)}\, \frac{\mathbf{d}_i^{(q)} + \mathbf{d}_{i+1}^{(q)}}{2}, \qquad q \in \{H, F, y\}. \tag{79}$$

At the terminal step, where $\sigma_{i+1} = 0$, we simply use the predictor:

$$\mathbf{z}_{i+1}^{(q)} = \mathbf{z}_{i+1}^{(q),\mathrm{pred}}. \tag{80}$$

In this form, anti-annealing is a channel-wise rescaling of the EDM drift. Equivalently, it introduces a channel-dependent time warp: channels with $\alpha_i^{(q)} > 1$ are driven more aggressively toward their denoised predictions, while the denoiser itself and the underlying EDM schedule remain unchanged.

E. Evaluation Details

E.1. De novo generation (DNG)

For *de novo* generation, we sample

$$N_{\text{gen}} = 10{,}000$$

crystals and decode them into periodic structures

$$\mathcal{G} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{N_{\text{gen}}}\}.$$

We report four groups of metrics: validity, uniqueness and novelty, distribution matching, and thermodynamic competitiveness.

Validity. We report composition validity, structure validity, and overall validity separately.

Composition validity is evaluated with SMACT (Davies et al., 2019). For each generated crystal, the stoichiometry is reduced to its primitive integer ratio, after which oxidation-state assignments, charge neutrality, and the Pauling electronegativity criterion are checked. Unary systems and all-metal alloys are handled in the standard way used in prior crystal-generation work.

Structure validity is implemented as a small pipeline rather than as a single geometric test. Before constructing a pymatgen `Structure`, the evaluator applies a safe-wrapper prefilter that rejects malformed decoded samples, including invalid atomic numbers and implausible lattice angles. The code then attempts to construct a periodic structure and marks the sample as structurally invalid if this fails, if lattice parameters or coordinates are non-finite, if lattice lengths are negative, or if the resulting cell volume is smaller than $0.1\,\text{Å}^3$. Only samples that survive these checks reach the final geometric validity test, which requires both

$$\text{vol}(\mathcal{C}) \geq 0.1\,\text{Å}^3 \qquad \text{and} \qquad d_{\min}(\mathcal{C}) \geq 0.5\,\text{Å}, \tag{81}$$

where $d_{\min}(\mathcal{C})$ is the minimum non-self interatomic distance in the constructed periodic structure.

Thus, the familiar condition $d_{\min} \geq 0.5\,\text{Å}$ together with $V \geq 0.1\,\text{Å}^3$ is the final structural-validity gate, but malformed samples may already be rejected earlier by wrapper- or construction-stage checks.

Let $\mathcal{G}_{\text{comp}}$, $\mathcal{G}_{\text{struct}}$, and $\mathcal{G}_{\text{val}}$ denote the subsets of generated crystals that pass the composition check, the structure check, and both checks, respectively. We then report

$$\text{CompVal} = \frac{|\mathcal{G}_{\text{comp}}|}{N_{\text{gen}}}, \qquad \text{StructVal} = \frac{|\mathcal{G}_{\text{struct}}|}{N_{\text{gen}}}, \qquad \text{Val} = \frac{|\mathcal{G}_{\text{val}}|}{N_{\text{gen}}}. \tag{82}$$

These validity metrics are reported for interpretability, but they are not the eligibility filter used for the main uniqueness, novelty, and UN metrics.

Uniqueness, novelty, and UN. For the main DNG metrics, we first construct filtered generated and reference sets,

$$\mathcal{G}_{\text{eval}} \subseteq \mathcal{G}, \qquad \mathcal{T}_{\text{eval}} \subseteq \mathcal{T},$$

by retaining only structures with finite geometry that satisfy the implemented N-ary threshold. In the current DNG code path, this threshold is `minimum_nary` $= 1$, so unary structures are retained.

Structure comparisons are performed with `pymatgen`'s `StructureMatcher` using

$$\text{stol} = 0.5, \qquad \text{ltol} = 0.3, \qquad \text{angle_tol} = 10.$$

A pair of structures is treated as matching whenever the matcher returns a valid alignment.

Let

$$N_{\text{eval}} = |\mathcal{G}_{\text{eval}}|$$

denote the number of generated structures that enter this evaluation stage. Uniqueness is computed by greedily deduplicating $\mathcal{G}_{\text{eval}}$, keeping only the first representative of each duplicate cluster. If N_{unique} denotes the number of retained representatives, then

$$\text{Unique} = \frac{N_{\text{unique}}}{N_{\text{eval}}}. \tag{83}$$

Novelty is evaluated relative to the filtered reference set $\mathcal{T}_{\text{eval}}$, after the usual chemistry-system filtering used by the benchmark. Let $N_{\text{novel_cand}}$ denote the number of generated structures that enter this novelty comparison, and let N_{novel} denote the number of these structures that do not match any structure in $\mathcal{T}_{\text{eval}}$. We report

$$\text{Novel} = \frac{N_{\text{novel}}}{N_{\text{novel_cand}}}. \tag{84}$$

The unique-and-novel set is not obtained by intersecting separately computed uniqueness and novelty flags. Instead, the code first restricts to the novel subset and then greedily deduplicates within that subset using the same first-occurrence rule as above. If N_{UN} denotes the number of resulting representatives, then

$$\text{UN} = \frac{N_{\text{UN}}}{N_{\text{novel_cand}}}. \tag{85}$$

In the usual non-degenerate case, $N_{\text{novel_cand}} = N_{\text{eval}}$, but we keep the notation separate here to reflect the implementation more faithfully.

Distribution matching. Distribution metrics are computed on the validity-filtered generated set $\mathcal{G}_{\text{val}}$. For any scalar crystal statistic $x(\mathcal{C})$, let P_x^{gen} and P_x^{ref} denote its empirical distributions over the generated and reference sets, respectively. We compare these distributions using the one-dimensional Wasserstein-1 distance

$$W_1(P, Q) = \int_{\mathbb{R}} \big|F_P(t) - F_Q(t)\big|\, dt, \tag{86}$$

where F_P and F_Q are the corresponding cumulative distribution functions.

In the main text we report two such metrics. The first is based on mass density,

$$\rho(\mathcal{C}) = \frac{\text{mass}(\mathcal{C})}{\text{vol}(\mathcal{C})}, \tag{87}$$

and the second is based on the N-ary statistic,

$$n_{\text{ary}}(\mathcal{C}) = \big|\{\text{elements present in } \mathcal{C}\}\big|. \tag{88}$$

We therefore report

$$\text{wdist-}\rho = W_1\big(P_\rho^{\text{gen}}, P_\rho^{\text{ref}}\big), \qquad \text{wdist-}N\text{-ary} = W_1\big(P_{n_{\text{ary}}}^{\text{gen}}, P_{n_{\text{ary}}}^{\text{ref}}\big). \tag{89}$$

Thermodynamic stabilities. For offline evaluation, we generate 10,000 crystals and perform thermodynamic post-processing on all 10,000 generated structures. During training, we use a lighter version of this procedure, in which thermodynamic evaluation may be restricted to a smaller subset for efficiency.

Relaxation is performed with a compiled NequIP model using the batched TorchSim backend on CUDA, together with FIRE optimization and a Fréchet cell filter, so that both atomic positions and lattice degrees of freedom are optimized jointly. In this batched code path, relaxation is run for a fixed 200 FIRE steps; no force-threshold early stopping is used.

After relaxation, the implementation does not compute energy above hull via a hand-written subtraction formula. Instead, for each relaxed crystal $\widetilde{\mathcal{C}}$ with final MLIP-predicted total energy $E^{\text{MLIP}}(\widetilde{\mathcal{C}})$, the code constructs a `ComputedStructureEntry`, attaches synthetic VASP-style metadata needed by `MaterialsProject2020Compatibility`, applies

```
MaterialsProject2020Compatibility(check_potcar=False),
```

and then evaluates the corrected entry against the patched Materials Project phase diagram through `get_e_above_hull(...)`. The reported quantity is therefore the hull distance of the corrected entry produced by this compatibility-processing pipeline.

Equivalently, one may view this as applying an MP2020-style correction to the relaxed MLIP energy before evaluating the distance to the reference convex hull, but the literal implementation is entry-based rather than an explicit subtraction against a separately written $E_{\text{hull}}^{\text{ref}}$ term. If compatibility processing fails, returns no corrected entry, or produces a non-finite hull distance, the sample is recorded as a thermodynamic failure.

Internally, the thermo logger records two thresholds:

$$\text{Stable} = \frac{1}{N_{\text{thermo}}} \left| \{ \widetilde{\mathcal{C}} : e_{\text{hull}}(\widetilde{\mathcal{C}}) \leq 0.0 \text{ eV/atom} \} \right|, \tag{90}$$

and

$$\text{Meta} = \frac{1}{N_{\text{thermo}}} \left| \{ \widetilde{\mathcal{C}} : e_{\text{hull}}(\widetilde{\mathcal{C}}) \leq 0.1 \text{ eV/atom} \} \right|, \tag{91}$$

where N_{thermo} is the number of crystals submitted to the thermodynamic pipeline. Relaxation and thermodynamic-processing failures count against these rates.

Thus, the implementation logs 0.0 eV/atom as *stable* and 0.1 eV/atom as *metastable*. In the main results, however, we often follow the common convention that the 0.1 eV/atom threshold is referred to simply as *stable*. The appendix keeps the stricter logger terminology to match the implementation more closely.

Finally, we combine thermodynamic competitiveness with the unique-and-novel rate. Let $\text{Stable}_{\text{UN}}$ and Meta_{UN} denote the fractions of unique-and-novel structures that satisfy the 0.0 and 0.1 eV/atom thresholds, respectively. We then define

$$\text{SUN} = \text{UN} \times \text{Stable}_{\text{UN}}, \qquad \text{MSUN} = \text{UN} \times \text{Meta}_{\text{UN}}. \tag{92}$$

Accordingly, when the main text informally treats the 0.1 eV/atom threshold as stability, it is this latter quantity that is being referred to.

E.2. Crystal structure prediction (CSP)

Crystal structure prediction is a conditional task. For each test composition, the model generates a crystal conditioned on that composition, and the prediction is compared with the corresponding ground-truth structure $\mathcal{C}_i^{\text{gt}}$ using `pymatgen`'s `StructureMatcher`. Unless noted otherwise, we use the same matcher tolerances as in the DNG evaluation:

$$\text{stol} = 0.5, \qquad \text{ltol} = 0.3, \qquad \text{angle_tol} = 10.$$

A prediction $\widehat{\mathcal{C}}_i$ is counted as correct if `StructureMatcher` finds a valid match to $\mathcal{C}_i^{\text{gt}}$ under these tolerances. The match rate is therefore

$$\text{MR} = \frac{1}{N_{\text{test}}} \left| \{ i : \widehat{\mathcal{C}}_i \text{ matches } \mathcal{C}_i^{\text{gt}} \} \right|, \tag{93}$$

where N_{test} is the number of test compositions.

For matched pairs, we additionally report the RMS displacement returned by the matcher after alignment. Let

$$\mathcal{M} = \{ i : \widehat{\mathcal{C}}_i \text{ matches } \mathcal{C}_i^{\text{gt}} \}$$

denote the set of matched test cases, and let r_i be the corresponding matcher RMS displacement for pair i. We report

$$\text{RMSD} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} r_i. \tag{94}$$

All CSP results in the main text use this standard single-sample setting.

E.3. Sample-size intensive and extensive metrics

An important practical point in *de novo* generation is that not all metrics behave the same way when the number of generated samples changes. Some metrics describe the quality of a *typical generated crystal*. Others describe the discovery yield of the *entire generated set*. We refer to these two cases, by analogy with physics, as *sample-intensive* and *sample-extensive* metrics.

Sample-intensive metrics. A metric is sample-intensive if its target does not depend strongly on the total generation budget n. These are quantities that can be estimated from a random subset of generated crystals without changing their meaning. In our setting, this includes:

- compositional validity and structural validity,
- per-sample stability rates,
- average hull distance or other per-sample property means,
- distribution metrics such as Wasserstein distances on density or N-ary statistics.

For such quantities, a random subset gives an approximation to the same underlying target. In the simplest case, if $g(\mathcal{C}_i)$ is a per-sample score or indicator, then

$$\widehat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} g(\mathcal{C}_i) \tag{95}$$

is the natural estimator from a subset of size m.

Sample-extensive metrics. A metric is sample-extensive if it depends directly on how many samples were generated. In crystal generation, this happens whenever duplicates matter. As the generation budget grows, duplicate collisions become more common, so the same model can look more or less diverse depending only on how many structures were sampled. In our setting, this includes:

- uniqueness,
- the number of distinct discovered structures,
- novelty when reported as a discovery yield over the generated set,
- UN,
- SUN.

For example, if $N_{\text{unique}}(n)$ is the number of unique generated structures after drawing n samples, then

$$\text{Unique}_n = \frac{N_{\text{unique}}(n)}{n}, \qquad \text{UN}_n = \frac{N_{\text{UN}}(n)}{n} \tag{96}$$

are explicitly functions of n. Evaluating these quantities on a smaller subset does *not* estimate their value at the full budget. It simply computes the same metric at a different budget. In practice, this usually makes uniqueness and related discovery metrics look artificially better on small subsets.

This distinction explains why some metrics can be estimated on subsets and others cannot. Validity, stability, and average property metrics can be approximated on random subsets. By contrast, uniqueness, UN, and SUN should be reported together with the number of generated samples and compared only at matched sample budgets.

A small caveat is that novelty can be defined in two different ways. If novelty is tested *per sample* against a fixed reference set, then it behaves like an intensive quantity. In our setting, however, novelty is used as part of the deduplicated discovery pipeline, so it is more natural to treat it together with UN and SUN as a sample-extensive quantity.

Implication for SUN. This viewpoint also clarifies why it is reasonable to compute UN on the full generated batch, but estimate stability only on a subset of the UN structures. If $\widehat{p}(\text{stable} \mid \text{UN})$ denotes the estimated stable fraction within the UN set, then the natural estimator is

$$\widehat{\text{SUN}}_n = \text{UN}_n \times \widehat{p}(\text{stable} \mid \text{UN}). \tag{97}$$

Here the first factor is a full-batch discovery statistic, while the second factor is a subset-based estimate of thermodynamic quality inside that discovered set.

Practical recommendation. For DNG evaluation, sample-extensive metrics such as uniqueness, UN, and SUN should always be reported together with the generation budget n. Sample-intensive metrics, such as validity, stability, and Wasserstein distances, can be estimated from random subsets when needed. This makes it easier to separate two different questions: whether the model generates *good individual crystals*, and whether it continues to produce *many distinct discoveries* as sampling is scaled up.

F. Additional Results

F.1. Ablation of Chemically Structured Atom Tokens

To isolate the impact of our proposed chemically structured representation, we train an identical Crystalite model using standard one-hot element encodings in place of Chemically Structured Atom Tokens. Because the two embedding types have different dimensionalities, we carefully balanced the atom-type loss weight for the one-hot baseline so that its stability curve aligned with the chemically structured model. Once the stability trajectories were matched, we observed the resulting unique and novel (UN) rate and the combined SUN score to evaluate the trade-off.

Figure 11 illustrates this comparison. While both models follow the expected trajectory of increasing stability at the cost of diversity as training progresses, the model equipped with chemically structured tokenization of atomic species consistently maintains a superior balance. Specifically, it achieves a higher overall stability ceiling and sustains a strictly higher SUN rate across the later stages of training. This indicates that a continuous, chemically structured tokenization provides a more effective inductive bias for diffusion over atomic species, translating directly into an advantage for generative discovery performance.

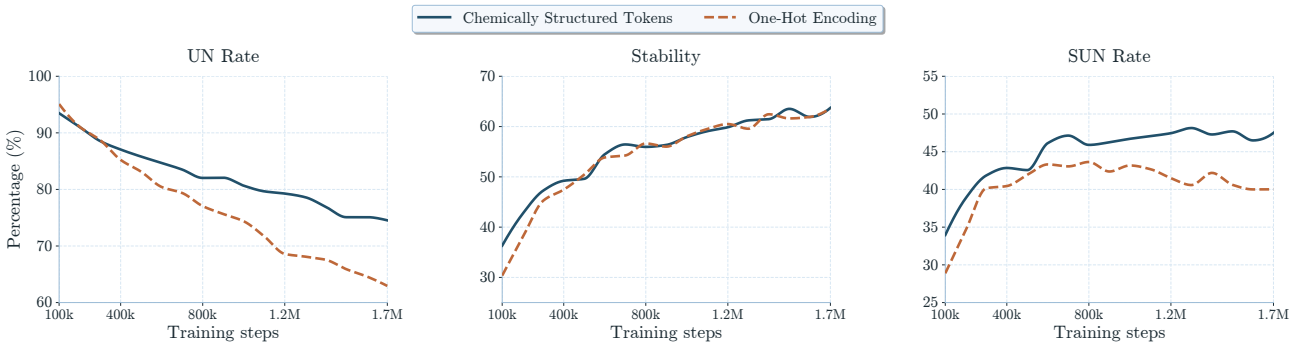


Figure 11. **Effect of chemically structured tokenization of atomic species. on training dynamics.** The evolution of UN rate (left), stability (middle), and SUN rate (right) is compared between models using Chemically Structured Atom Tokens and standard one-hot encoding. After balancing the loss weights to align the stability curves, the chemically structured representation achieves consistently higher UN and SUN scores throughout the latter half of training.

F.2. DNG MatterGen evaluation pipeline results

Here we present our DNG metrics when evaluated using Mattergen evaluation pipeline, so that we can compare different models against ours on a setup that is not designed by us.

Table 4. Validity, uniqueness, novelty, stability, and relaxation metrics using the MatterGen (Zeni et al., 2025) evaluation pipeline for MP-20.

Model	Validity and Novelty				Stability and Relaxation			
	Struct. Val. (%) ↑	Comp. Val. (%) ↑	Unique (%) ↑	Novel (%) ↑	Stable (%) ↑	S.U.N. (%) ↑	Avg. Hull (eV/atom) ↓	Avg. RMSD (Å) ↓
MatterGen	**100.00**	83.48	**97.94**	**75.02**	45.74	23.75	0.182	**0.153**
ADiT	**100.00**	**90.24**	89.62	43.15	**69.96**	17.17	0.148	0.493
Crystalite	**100.00**	84.62	94.73	56.63	64.52	**24.26**	**0.145**	0.274

F.3. GEM effect on DNG Results

Figure 12 compares the training dynamics of Crystalite with and without the Geometry Enhancement Module (GEM) in the *de novo* generation setting. Both configurations exhibit the expected decline in the unique-and-novel (UN) rate as training progresses, reflecting the general trade-off between diversity and stability. However, the model with GEM learns substantially faster on the stability axis and maintains higher stability throughout training. As a result, it also achieves a consistently higher Stable, Unique, and Novel (SUN) rate across the full training trajectory. This suggests that injecting periodic pairwise

geometry into attention improves the structural quality of generated crystals without causing a disproportionate loss in generative diversity.

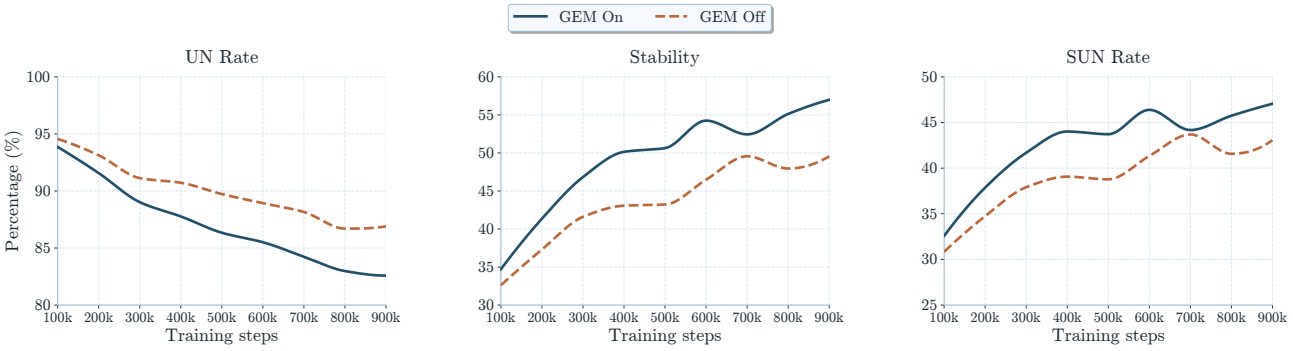


Figure 12. **Effect of GEM on *de novo* generation training dynamics.** UN rate (left), stability (middle), and SUN rate (right) as a function of training steps, with and without GEM.

F.4. GEM effect on CSP Results

Figure 13 shows the corresponding ablation for crystal structure prediction (CSP). Here, GEM has only a modest effect on Match Rate (MR), but leads to a clearer and more consistent improvement in RMSE throughout training. In other words, GEM appears to have a limited effect on whether the model recovers the correct structural mode, but a stronger effect on how accurately that structure is refined once recovered. This is consistent with the interpretation that the geometric biases introduced by GEM primarily improve local atomic placement and overall structural fidelity during denoising.

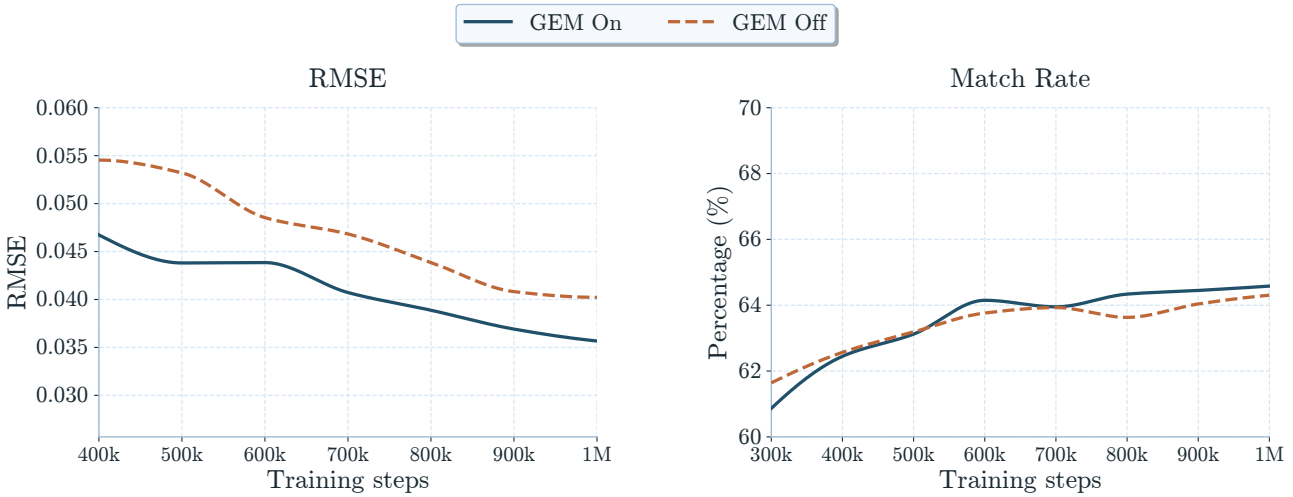


Figure 13. **Effect of GEM on crystal structure prediction.** RMSE (left) and Match Rate (right) as a function of training steps, with and without GEM.

F.5. DNG Sensitivity to anti-annealing

We also ablate the channel-wise anti-annealing settings used at sampling time, varying the strength of anti-annealing for the coordinate and lattice channels while keeping the trained model fixed.

Table 5. Generative quality, diversity, stability, and distribution metrics for Crystalite across the aa grid.

AA settings			Quality and Diversity					Stability and Distribution			
$\mathrm{aa}_{\mathrm{coords}}$	$\mathrm{aa}_{\mathrm{types}}$	$\mathrm{aa}_{\mathrm{lattice}}$	Struct. Val. (%) ↑	Comp. Val. (%) ↑	Unique (%) ↑	Novel (%) ↑	U.N. (%) ↑	Stable (%) ↑	S.U.N. (%) ↑	wdist-ρ ↓	wdist N-ary ↓
0	0	0	99.76	81.49	98.78	86.04	85.55	63.28	49.07	0.125	0.200
0	0	4	99.71	83.30	98.58	83.15	82.37	66.75	49.37	0.428	0.221
0	0	10	99.71	81.25	99.02	86.28	85.94	62.74	48.93	0.131	**0.191**
4	0	0	99.76	80.91	99.22	86.18	85.99	61.18	47.22	0.179	0.249
4	0	4	99.85	81.69	98.44	83.94	83.20	**68.12**	**51.42**	0.500	0.226
4	0	10	99.71	80.57	99.02	85.94	85.55	62.21	47.90	0.176	0.248
10	0	0	**99.90**	81.59	98.83	86.47	86.04	62.89	49.12	**0.111**	0.205
10	0	4	99.66	83.15	98.58	82.91	82.13	66.80	49.12	0.421	0.205
10	0	10	99.76	80.81	98.88	85.79	85.40	63.62	49.22	0.125	0.198
0	10	0	99.76	81.49	98.93	86.43	85.99	62.06	48.29	0.117	0.199
0	10	4	99.80	82.91	98.54	83.20	82.47	67.04	49.66	0.401	0.210
0	10	10	99.66	80.91	98.97	86.33	85.89	62.65	48.78	0.126	0.196
4	10	0	99.71	80.03	99.27	86.38	86.13	61.04	47.27	0.168	0.247
4	10	4	99.76	81.93	98.63	83.84	83.25	67.33	50.73	0.482	0.231
4	10	10	99.80	79.88	**99.32**	85.94	85.74	60.64	46.48	0.155	0.228
10	10	0	99.85	81.15	98.93	86.43	86.04	63.13	49.41	0.123	0.212
10	10	4	99.71	**83.35**	98.54	82.81	82.03	67.04	49.27	0.416	0.209
10	10	10	99.85	81.79	98.93	86.28	85.84	62.55	48.63	0.125	0.210
0	20	0	99.76	81.15	98.93	**86.62**	**86.18**	62.35	48.78	0.131	0.214
0	20	4	99.80	82.96	98.44	82.86	81.98	66.80	48.97	0.415	0.215
0	20	10	99.85	81.69	98.97	86.38	85.94	63.04	49.22	0.120	0.197
4	20	0	99.85	80.91	99.17	86.18	85.89	60.94	46.92	0.180	0.240
4	20	4	99.85	82.13	98.39	83.94	83.11	67.58	50.78	0.477	0.226
4	20	10	99.76	80.27	99.07	85.84	85.50	60.94	46.48	0.187	0.247
10	20	0	**99.90**	81.20	98.93	86.33	85.89	62.89	49.02	0.133	0.210
10	20	4	99.66	**83.35**	98.63	83.15	82.47	67.04	49.71	0.412	0.220
10	20	10	99.85	81.64	98.88	86.38	85.99	62.89	49.12	0.130	0.202

Overall, the results are fairly insensitive to this choice: across a reasonable range of settings, the main conclusions remain unchanged and Crystalite performs consistently well. Although one anti-annealing configuration achieved the highest SUN score, it also produced noticeably worse Wasserstein distances, indicating poorer distributional alignment. For this reason, we do not report the single best-SUN configuration, but instead select a more balanced setting that preserves strong discovery performance while maintaining better agreement with the reference distribution. This suggests that anti-annealing is a useful but non-fragile sampling heuristic, and that the reported results do not depend critically on a finely tuned choice of anti-annealing parameters.

F.6. CSP Sensitivity to anti-annealing

We also ablate the channel-wise anti-annealing settings used at sampling time for the CSP model, varying the strength of anti-annealing for the coordinate and lattice channels while keeping the trained model fixed.

Table 6. Match rate and RMSE metrics for varying anti-annealing settings on coordinate and lattice channels (CSP results).

AA settings		Metrics	
$\mathrm{aa}_{\mathrm{coords}}$	$\mathrm{aa}_{\mathrm{lattice}}$	Match Rate (%) ↑	RMSE ($\times 10^{-2}$) ↓
0	0	65.04 ± 0.23	4.11 ± 0.05
0	4	65.22 ± 0.16	4.14 ± 0.05
0	10	65.01 ± 0.22	4.12 ± 0.05
0	20	65.02 ± 0.22	4.11 ± 0.06
4	0	65.96 ± 0.17	3.41 ± 0.07
4	4	$\mathbf{66.09 \pm 0.09}$	$\mathbf{3.37 \pm 0.05}$
4	10	65.94 ± 0.15	3.39 ± 0.07
4	20	65.96 ± 0.16	3.41 ± 0.08
10	0	65.15 ± 0.19	4.19 ± 0.04
10	4	65.22 ± 0.08	4.12 ± 0.04
10	10	65.13 ± 0.14	4.15 ± 0.05
10	20	65.17 ± 0.17	4.20 ± 0.05
20	0	65.19 ± 0.15	4.19 ± 0.08
20	4	65.26 ± 0.11	4.16 ± 0.08
20	10	65.17 ± 0.16	4.21 ± 0.07
20	20	65.13 ± 0.14	4.20 ± 0.06
27	36	65.13 ± 0.00	4.07 ± 0.00

Similar to the DNG results, CSP performance exhibits broad stability across a range of anti-annealing settings, confirming that the sampling process is generally robust. However, unlike the metric trade-offs observed in the DNG ablations, we find a clear and consistent alignment between our predictive metrics for CSP: the configuration with modest anti-annealing on both channels ($\mathrm{aa}_{\mathrm{coords}} = 4, \mathrm{aa}_{\mathrm{lattice}} = 4$) simultaneously achieves the highest match rate and the lowest RMSE. This optimal pocket provides a distinct performance advantage without inducing instability. Therefore, we highlight this balanced configuration, demonstrating that while the model does not depend critically on highly tuned parameters to perform well, targeted anti-annealing still yields a measurable and synergistic boost to structural prediction accuracy.

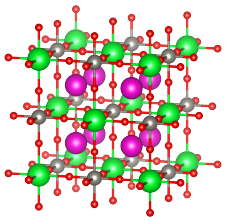
G. LeMat-GenBench Results

We additionally evaluate Crystalite under the LeMat-GenBench external benchmarking pipeline (Betala et al., 2026); results for MP-20 models are reported in Table 7. Crystalite achieves the best SUN and MSUN scores under this pipeline among the pre-relaxed methods, consistent with the conclusions drawn from our primary evaluation.

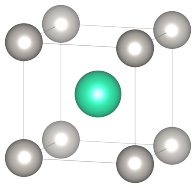
Table 7. LeMat-GenBench results for MP-20 models, split by whether generations are pre-relaxed. Higher is better except E_{hull} and RMSD.

Model	Val. (%)	Uniq. (%)	Nov. (%)	Stab. (%)	Meta. (%)	SUN (%)	MSUN (%)	E_{hull} (eV)	RMSD (Å)
Pre-relaxed									
WyFormer (Kazeev et al., 2025)	93.40	93.00	66.40	0.50	15.70	0.10	1.90	0.4988	0.8121
WyFormer-DFT (Kazeev et al., 2025)	95.20	95.00	66.40	3.70	24.80	0.40	7.80	0.2708	0.4173
PLaID++ (Xu et al., 2025)	96.00	77.80	24.20	12.40	**60.70**	1.00	7.60	**0.0854**	0.1286
MatterGen (Zeni et al., 2025)	95.70	95.10	**70.50**	2.00	33.40	0.20	15.00	0.1834	0.3878
OMatG (Höllmer et al., 2025)	96.40	95.20	51.20	11.60	49.80	1.00	18.00	0.0956	**0.0759**
Crystalite	**97.20**	**95.80**	53.20	**12.70**	51.60	**1.50**	**22.60**	0.0905	0.1320
Non-pre-relaxed									
Crystal-GFN (Mistal et al., 2023)	51.70	51.70	51.70	0.00	0.00	0.00	0.00	2.0858	1.8665
ADiT (Joshi et al., 2025)	90.60	87.80	26.00	0.40	**36.50**	0.00	1.00	0.3333	0.3794
CrystalFormer (Cao et al., 2025)	69.90	69.40	31.80	1.40	28.80	0.00	3.10	0.7039	0.6585
SymmCD (Levy et al., 2025)	73.40	73.00	47.00	1.40	18.60	0.10	2.40	0.8761	0.8720
DiffCSP++ (Jiao et al., 2024)	95.30	**95.10**	62.00	1.00	26.40	**0.20**	5.00	0.4093	0.6933
DiffCSP (Jiao et al., 2023)	**95.70**	94.80	**66.20**	**2.30**	29.80	0.10	**8.50**	**0.2747**	**0.3794**

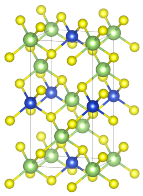
H. Crystalite S.U.N. Crystals



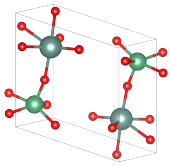
(a) $Sr_4Eu_8W_4O_{24}$



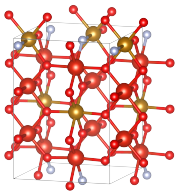
(b) LuPt



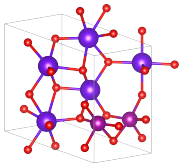
(c) $Ga_4Cu_2S_8$



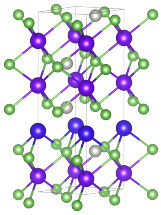
(d) $Y_2Nb_2O_8$



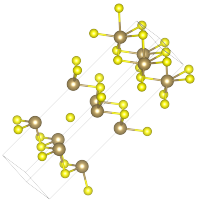
(e) $V_8Fe_4O_{22}F_2$



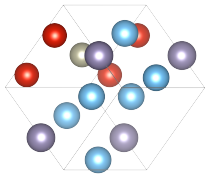
(f) $Tb_5Mn_2O_{11}$



(g) $Tb_3DyAs_4Pd_4$



(h) Ta_3S_5



(i) Ti_4V_2ReSn

Figure 14. A set of stable, unique and novel crystals generated by Crystalite trained on MP-20.