Creator: Crystal Generation with Multivector-Enhanced Diffusion Transformers

Kairos Materials

research@kairosmaterials.com

Generative modeling is opening a new route to materials discovery. Instead of searching only among known compounds or manually proposed candidates, models can explore broad chemical and structural spaces and propose crystals that have not previously been considered. Diffusion Transformers provide a powerful and computationally efficient foundation for generation in images and videos, yet their generic token representations are poorly matched to crystals, where composition, atomic positions, lattice vectors and periodic structure must be learned jointly. We introduce Creator, a crystal generator built on a multivector-enhanced Diffusion Transformer. Atomic positions and lattice vectors are first embedded as vector elements of $\mathrm{Cl}_{3,0}$ and lifted through Clifford-parameterized linear maps into algebraically coupled features. The three lattice basis vectors are retained as global anchor tokens for an exchangeable set of atoms, while a multivector-inspired partition of query–key features structures their interactions. Value aggregation, feed-forward transformations and diffusion conditioning remain dense, providing geometric inductive bias without enforcing exact equivariance throughout the network. At matched model size, Creator outperforms a conventional crystal DiT, and increasing its capacity from 5 million to 60 million parameters under a fixed MP-20 training setting consistently raises the metastable fraction of generated crystals. On LeMat-GenBench, Creator trained on MP-20 reaches 99.64% validity, 97.84% uniqueness, 68.28% metastability and a 31.72% metastable–unique–novel rate after relaxation. The same backbone extends to composition-conditioned crystal structure prediction and property-conditioned *de novo* materials generation, with band-gap-guided generation spanning metallic to wide-gap regimes as one example. Together, these results establish multivector enhancement as a practical geometric interface for crystal Diffusion Transformers, combining structured three-dimensional representation with the flexibility and efficiency of dense generative models.

KAIROS MATERIALS

1 Introduction

Materials discovery relies on exploring an enormous space of chemical compositions, atomic structures and physical properties. The number of possible crystalline materials far exceeds those that have been experimentally synthesized or computationally characterized, making exhaustive exploration infeasible with conventional approaches. Recent advances in machine learning have created new opportunities to accelerate this process by learning distributions over known materials and generating candidate structures beyond previously explored chemical and structural spaces. In particular, generative models (Li et al., 2025) provide a framework for jointly navigating composition, lattice geometry and atomic arrangement, offering the possibility of discovering materials with prescribed structures, compositions or physical properties.

A variety of generative approaches have been developed for crystalline materials, demonstrating the ability of machine learning models to produce chemically plausible and structurally diverse candidates. Early methods based on variational autoencoders (Noh et al., 2019) and diffusion models (Ho et al., 2020) established the feasibility of learning distributions over periodic structures, while more recent approaches such as CDVAE (Xie et al., 2022b), DiffCSP (Jiao et al., 2023), MatterGen (Zeni et al., 2025), FlowMM (Miller et al., 2024) and CrystalFlow (Luo et al., 2025) have improved generation quality by incorporating physical constraints and geometric information. Many of these models rely on invariant or equivariant neural networks to encode the spatial relationships and symmetry properties of crystals. Such architectures provide strong geometric inductive biases, but they also require specialized feature representations and constrained transformations

throughout the network. This added structure can increase computational complexity and makes it less straightforward to exploit the dense operations and model-capacity increases that have driven progress in modern generative modeling.

An alternative direction has emerged from the success of Diffusion Transformers (Peebles and Xie, 2023) in image and video generation. By replacing domain-specific denoising architectures with standard Transformer blocks, DiTs combine global interaction, high model capacity and efficient dense computation. Despite the few pioneering work in this area(Yi et al., 2026; Joshi et al., 2025; Seong et al., 2026; Hadži Veljković et al., 2026), these properties make them an attractive foundation for generative modeling beyond visual domains. Directly transferring a conventional DiT to crystals, however, leaves much of the relevant structure unspecified. A crystal is neither a regular grid nor an ordinary sequence: it consists of an exchangeable set of atoms embedded within a global frame defined by three lattice basis vectors, while chemical identities, atomic positions and lattice geometry must remain mutually consistent. Standard coordinate embeddings treat the components of lattice and position vectors as generic numerical features, requiring the model to learn both an effective representation of individual three-dimensional variables and the interactions through which those variables collectively define a crystal.

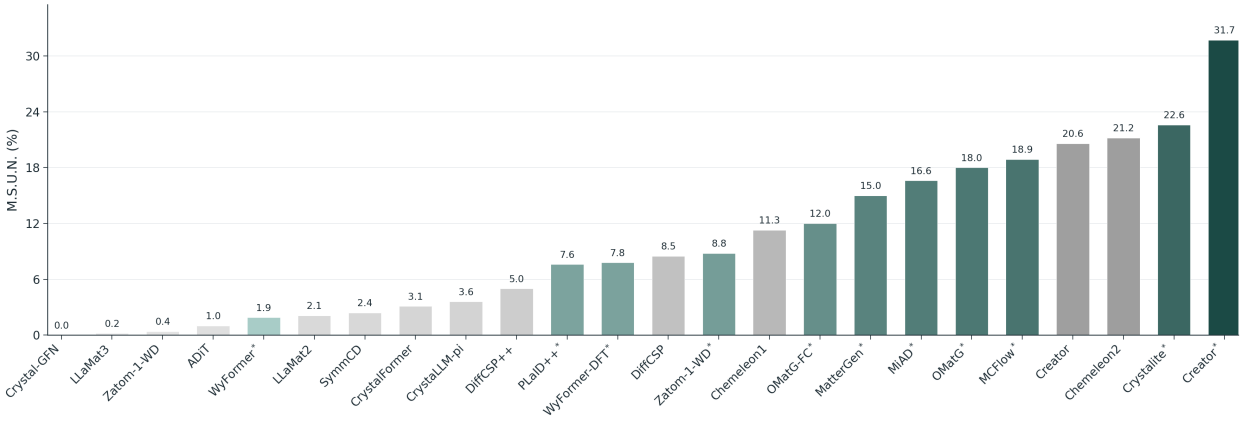


Figure 1 M.S.U.N. rate comparison across all models trained on MP-20 evaluated under LeMat-GenBench . Models ranked by M.S.U.N. in descending order. * Structures relaxed.

Here we introduce Creator, a crystal generator built on a multivector-enhanced Diffusion Transformer. Creator separates the representation of geometry within individual crystal elements from the modeling of interactions across the complete structure. Each lattice vector and atomic coordinate is first embedded as a vector element of the Clifford algebra $\mathrm{Cl}_{3,0}$ and lifted through a Clifford-parameterized linear map, producing algebraically coupled features rather than an unconstrained coordinate projection. The three lattice basis vectors are retained as global anchor tokens for the atomic set, allowing lattice–lattice, lattice–atom and atom–atom relationships to be modeled jointly through global self-attention. Query–key interactions additionally follow a partition inspired by the 1+3+3+1 organization of the multivector basis, whereas value aggregation, feed-forward transformations and diffusion conditioning remain dense. Creator therefore uses multivectors as a structured geometric interface rather than enforcing exact equivariance or preserving algebraic grades throughout the network. We evaluate this design against a conventional crystal DiT, examine its behavior as model capacity increases to 60 million parameters, and assess its unconditional generation performance using LeMat-GenBench (Duval et al., 2025), as shown in Fig.1. The same backbone is further extended to composition-conditioned crystal structure prediction and property-conditioned de novo materials generation, providing a unified framework for crystal generation, structural exploration and inverse materials design.

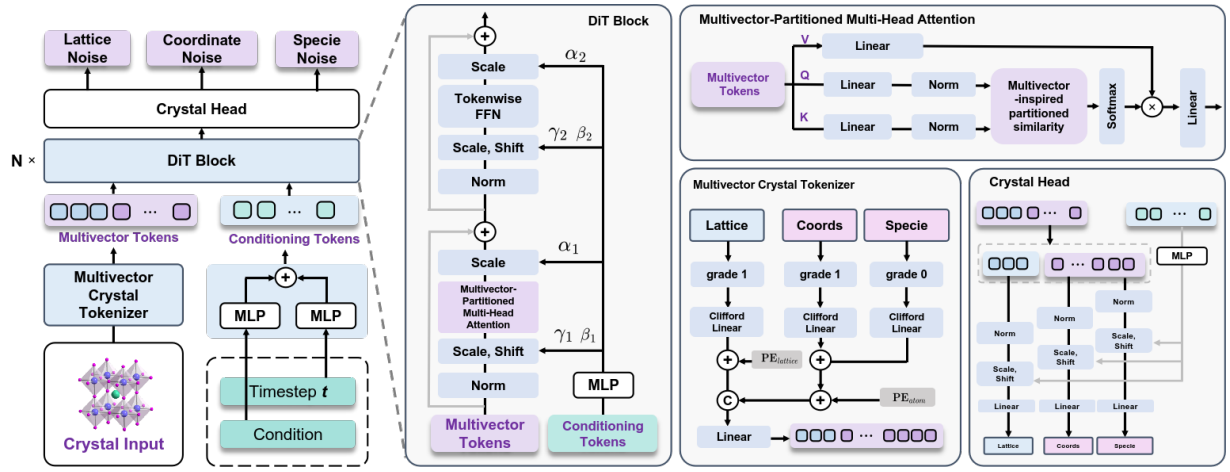


Figure 2 Model architecture

2 Creator: A Generative Model for Crystals with Multivector Enhancements

2.1 Overview

Creator is a diffusion transformer framework that introduces geometric inductive bias into crystal generation through multivector-based representations. The key idea is to represent crystal structures as geometric multivector tokens and perform token interactions in a unified geometric feature space.

Different from conventional Transformer architectures that treat atomic coordinates as independent scalar features, Creator initially organizes crystal geometry using a multivector-indexed parameterization. In three-dimensional space, geometric quantities can be naturally organized into different components: scalars capture invariant properties, vectors encode directional information, bivector-like components describe oriented planar structures, and higher-order components capture volumetric geometric information. By embedding crystal variables into this unified representation space, Creator enables the model to capture rich geometric relationships beyond conventional pairwise coordinate differences.

As illustrated in Fig. 2, Creator consists of three major components: (1) a geometric crystal tokenizer, which converts lattice vectors and atomic structures into geometry-aware multivector tokens; (2) a Diffusion Transformer backbone, which performs scalable denoising generation under timestep conditioning; and (3) a crystal head, which decodes multivector representations into lattice vectors, atomic coordinates, and atomic attributes.

2.2 Multivector Crystal Tokenization

We present a crystal tokenization scheme that encodes both geometric and chemical information into a unified sequence representation. Our approach combines lattice-anchored with Clifford multivector parameterization, enabling the model to reason about interatomic relationships.

2.2.1 Lattice-Anchored Atomic-Set Representation

We define the crystal token set as:

$$\mathcal{X} = [L_1, L_2, L_3, A_1, A_2, \ldots, A_N] \tag{1}$$

where $L_1, L_2, L_3 \in \mathbb{R}^D$ are three independent lattice anchor tokens corresponding to the three basis vectors of the unit cell, and $A_i \in \mathbb{R}^D$ are atom tokens. The three lattice tokens receive distinct learned identity embeddings, while all atoms share a common atom-token embedding.

This design ensures that each atom has direct access to the complete set of lattice basis vectors from the very first layer, eliminating the need for iterative local message passing to acquire global cell context. In terms of

symmetry properties, atom tokens are processed in a permutation-equivariant manner, the three lattice basis vectors serve as ordered anchors, and no explicit invariance to equivalent unit cell representations is enforced.

2.2.2 Clifford-Parameterized Multivector Lifting

The core of such tokenization is a multivector-valued linear lifting parameterized by the Clifford algebra structure. For each output channel c, we learn a multivector weight:

$$W_c = \sum_{a=0}^{7} w_{c,a}\, e_a \tag{2}$$

where e_a enumerates the 8 basis blades of $\mathrm{Cl}_{3,0}$. The lifting operation is then computed as:

$$M_c = XW_c + b_c \tag{3}$$

where multiplication is performed according to the complete $\mathrm{Cl}_{3,0}$ multiplication table, rather than using eight independent standard linear transformations.

This parameterization induces several important properties:

- Fixed blade coupling patterns determined by the Clifford multiplication table
- Fixed sign structure inherited from the algebra's anti-commutation relations
- Parameter sharing across different blade types
- Algebraic associations between scalar, vector, bivector, and trivector-indexed features

Both lattice vectors and atomic positions are lifted using this same form, though they employ separate `CliffordLinear` (Brandstetter et al., 2023; Ruhe et al., 2023) modules to capture their distinct geometric roles.

We offer a restrained interpretation of this lifting: it does not explicitly compute planes between lattice vectors or the volume spanned by the unit cell. Instead, it replaces an unconstrained coordinate embedding with a linear but algebraically coupled parameterization derived from the Clifford multiplication table. The geometric enhancement arises from the structured parameterization and feature coupling, not from explicit computation of derived geometric quantities such as crystal faces or volumes. The operation remains fundamentally linear in the input coordinates, merely structured by the algebra.

2.2.3 Chemical Integration and Dense Feature Fusion

Chemical embedding. For each atom i, its chemical identity is represented by two integers corresponding to its period and group in the periodic table, $T_i \in \mathbb{R}^2$. This is projected through a standard linear layer:

$$T_i \to t_i \in \mathbb{R}^{D/8} \tag{4}$$

where D is the embedding dimension and the factor of $1/8$ matches the channel count per multivector blade. The chemical embedding is then broadcast across all eight multivector-indexed slots and added to the position feature with a reduced weight:

$$A_i^{\mathrm{emb}} = P_i^{\mathrm{mv}} + \delta \cdot \mathrm{broadcast}(t_i) \tag{5}$$

Token identity embeddings. After geometric lifting, we add token identity embeddings: three distinct learned embeddings for the three lattice tokens, and one shared embedding for all atom tokens.

Dense fusion. Finally, lattice and atom features are concatenated along the sequence dimension and passed through a standard dense linear layer. This fusion layer marks the transition from explicitly multivector-valued features to a general dense hidden representation. Beyond this point, blade identity is no longer strictly maintained, and we make no claim that scalar/vector/bivector/trivector semantics are preserved. The multivector lifting has fulfilled its role as an input parameterization, providing a structured geometric foundation for subsequent processing.

2.3 Multivector-Partitioned Multi-Head Attention

We extend the standard multi-head attention mechanism with a multivector-inspired partitioned interaction metric. This design introduces structured inductive biases into the similarity computation while maintaining compatibility with standard transformer implementations.

Our attention block begins with standard dense linear projections for queries, keys, and values:

$$Q = W_Q X, \quad K = W_K X, \quad V = W_V X \tag{6}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{D \times D}$ are learned projection matrices. Following contemporary best practices, queries and keys are normalized with RMSNorm before similarity computation:

$$\tilde{Q} = \text{RMSNorm}(Q), \quad \tilde{K} = \text{RMSNorm}(K) \tag{7}$$

while values remain as standard dense features without additional normalization. The projected tensors are then reshaped into H parallel attention heads, each of dimension $d_h = D/H$.

Within each attention head, we partition the query and key representations into four subspace groups following the dimension pattern of the Clifford algebra $\text{Cl}_{3,0}$:

$$1 + 3 + 3 + 1 \tag{8}$$

corresponding to: 1 scalar, 3 vector components, 3 bivector components, and 1 trivector. Concretely, each head dimension d_h is further divided such that each blade subspace receives $d_b = d_h/8$ channels. For queries, this gives:

$$\tilde{Q}^{(g)} \in \mathbb{R}^{B \times H \times S \times d_b} \quad \text{for } g = 0, \ldots, 7 \tag{9}$$

where $g = 0$ indexes the scalar subspace, $g = 1, 2, 3$ index vector components, $g = 4, 5, 6$ index bivector components, and $g = 7$ indexes the trivector subspace. The same partitioning is applied to keys.

Because this partition is applied after dense feature fusion and dense Q/K projection, these groups are learned subspaces rather than exact Clifford grades. The partition does not enforce strict algebraic constraints; it merely provides a structural template that encourages the model to organize information in a manner reminiscent of multivector grades.

The attention score between token i and token j is computed as a weighted sum of dot products across the partitioned subspaces:

$$s_{ij} = q_i^{(0)} \cdot k_j^{(0)} + \sum_{r=1}^{3} q_i^{(r)} \cdot k_j^{(r)} + \lambda \sum_{r=4}^{6} q_i^{(r)} \cdot k_j^{(r)} \tag{10}$$

where λ is a configurable weight (we set 0.1 in this work) that downweights the contribution of bivector subspace interactions. Notably, the trivector group ($r = 7$) does not participate in the similarity score, serving only as additional representational capacity within the Q/V projections.

The attention weights are then obtained by applying softmax with appropriate scaling:

$$A_{ij} = \text{Softmax}_j \left(\frac{s_{ij}}{\sqrt{d_b}} \right) \tag{11}$$

where the scaling factor $\sqrt{d_b}$ is normalized by the dimension of each blade subspace rather than the full head dimension. The output follows standard attention aggregation:

$$O_i = \sum_j A_{ij} V_j \tag{12}$$

We characterize this attention mechanism as a multivector-inspired partitioned similarity . It is not a full geometric product, not an exact Clifford inner product, and does not provide equivariance guarantees. Instead, it introduces a mild structural prior that encourages the model to attend to different types of feature interactions with different strengths, loosely modeled after the grade structure of Clifford algebra.

2.4 Dense Diffusion Transformer Backbone

The core of our model follows a standard DiT block architecture with adaptive layer normalization (AdaLN-Zero) and residual gating. Each block processes the fused dense token sequence through alternating self-attention and feed-forward layers.

Block structure. Formally, each block applies the following updates:

$$\begin{aligned} X_1 &= \mathrm{LN}(X) \odot (1 + \gamma_1) + \beta_1, \\ X' &= X + \alpha_1 \cdot \mathrm{Attention}(X_1), \\ X_2 &= \mathrm{LN}(X') \odot (1 + \gamma_2) + \beta_2, \\ X_{\mathrm{out}} &= X' + \alpha_2 \cdot \mathrm{FFN}(X_2), \end{aligned} \tag{13}$$

where $\beta_{1,2}, \gamma_{1,2}, \alpha_{1,2}$ are modulation parameters generated from a timestep embedding. The attention sublayer uses the multivector-partitioned similarity described in the previous section, with dense value aggregation. The feed-forward network is a standard two-layer MLP with GELU activation and an expansion ratio of $4\times$.

Conditioning. The timestep t is embedded via sinusoidal positional encoding and projected through a two-layer MLP with SiLU activations, then linearly mapped to six per-dimension modulation parameters $(\beta_1, \gamma_1, \alpha_1, \beta_2, \gamma_2, \alpha_2)$. A learnable task embedding is added to the timestep embedding before modulation projection.

Zero initialization. Following the AdaLN-Zero convention, the final linear layer of the timestep projection is initialized to zero, as are the per-block modulation parameters and all output head projections. This ensures that at initialization, each block behaves as an identity mapping, enabling stable training of deep architectures.

2.5 Joint Crystal Prediction Heads

After L DiT blocks, the final token sequence is split back into lattice and atom groups:

$$X_{\mathrm{lat}} = X[:, :3, :], \quad X_{\mathrm{atom}} = X[:, 3:, :] \tag{14}$$

Three separate prediction heads decode these features into the target crystal properties:

$$\begin{aligned} L &= \mathrm{Head}_L(X_{\mathrm{lat}}) \in \mathbb{R}^{B\times 3\times 3}, \\ P &= \mathrm{Head}_P(X_{\mathrm{atom}}) \in \mathbb{R}^{B\times N\times 3}, \\ T &= \mathrm{Head}_T(X_{\mathrm{atom}}) \in \mathbb{R}^{B\times N\times 2}, \end{aligned} \tag{15}$$

corresponding to the three lattice basis vectors, atomic positions, and atomic types (period and group), respectively. Each head applies AdaLN-modulated layer normalization followed by a linear projection.

3 Multivector-Enhanced DiT versus Vanilla DiT

To evaluate the whether structured geometric parameterization of input representation and attention interaction can improve crystal diffusion models without full equivariance constraints that complicate architecture design and limit scalability, we compare two variants, *i.e.*, a vanilla DiT baseline that uses standard linear embeddings and standard dot-product attention, and a multivector-enhanced DiT that introduces a lightweight geometric inductive bias through multivector tokenization and partitioned attention similarity. The two models share identical backbone architecture (DiT blocks with AdaLN-Zero, GELU FFN, zero initialization), training objective, and optimization setup, differing only in how input coordinates are embedded and how attention similarity is computed. The parameter count difference between the two models is negligible (31.60M vs 31.74M), since the multivector lifting is a structured linear parameterization rather than an expansion of channels. Both models are trained on MP-20 and evaluated against the LeMat-Bulk reference dataset with the same energy settings with LeMat-GenBench.

Table 1 Ablation study comparing vanilla DiT and multivector-enhanced DiT.

Model	Param (M)	Metastable (%)	Formation Energy (eV)
Vanilla DiT	31.60	24.2	−0.214
Multivector-enhanced DiT	31.74	30.6	−0.374

As shown in Table 1, the multivector-enhanced model consistently outperforms the vanilla DiT baseline on both the metastable fraction and mean formation energy metrics. Without futher relaxation, the metastability rate increases from 24.2% to 30.6%, and the formation energy improves by 0.16 eV relative to the convex hull. This improvement is achieved with virtually no parameter overhead (less than 0.5% additional parameters), indicating that the benefit comes from the structural inductive bias rather than increased model capacity.

These results support that structured geometric priors, even when introduced as lightweight parameterization rather than strict equivariance constraints, can meaningfully improve generative quality for crystal structures. The multivector formulation provides a principled way to organize geometric information within a standard Transformer framework, without sacrificing architectural simplicity or training stability.

To investigate whether multivector-enhanced DiT preserves the scaling properties of Diffusion Transformers, we conduct a model size ablation study by varying the number of trainable parameters while keeping the dataset, training objective, and optimization strategy fixed. We evaluate five model variants ranging from 5M to 60M parameters.

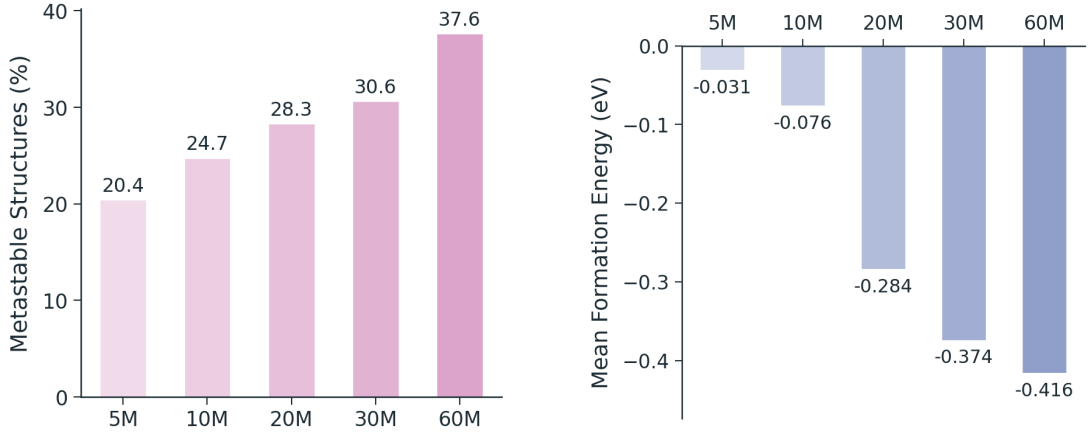


Figure 3 Scaling behavior of multivector-enhanced DiT across model sizes from 5M to 60M parameters. Left: fraction of metastable smean predicted formation energy per atom.

As shown in Fig. 3, increasing model capacity consistently improves both key generation metrics—the fraction of metastable structures and the mean formation energy relative to the convex hull, without further relaxtion. The fraction of metastable structures increases monotonically from 20.4% for the 5M model to 37.6% for the 60M model, indicating that larger models are increasingly capable of generating crystals that lie close to the thermodynamic stability hull. Concurrently, the mean formation energy improves from −0.031 eV to −0.416 eV, steadily moving toward energies matching DFT-relaxed reference structures. This approximately 0.385 eV improvement demonstrates that scaling the model directly translates to higher-fidelity energy predictions and thermodynamically more favorable generated structures.

These results verify that by restricting geometric structure to the input parameterization and attention interaction while maintaining standard dense Transformer components throughout the backbone, multivector-enhanced DiT inherits the favorable scaling behavior of modern Diffusion Transformers. The steady improvement across the entire parameter range confirms that the multivector formulation provides a scalable framework for large-scale crystal structure generation.

4 LeMat-GenBench

To evaluate Creator under a common and reproducible setting against other pioneering models, we conduct a full benchmark using LeMat-GenBench (Duval et al., 2025), a unified evaluation framework for generative models of crystalline materials. Crystal generation has previously been assessed using different validity filters, reference databases, structure-matching procedures, relaxation protocols and energy estimators, making numerical results difficult to compare across models. LeMat-GenBench fixes these choices within a common pipeline and evaluates generated structures along several complementary dimensions, including physical validity, structural uniqueness and novelty, thermodynamic stability, chemical and structural diversity, and distributional agreement with reference materials. Rather than considering these quantities independently, the benchmark also measures how frequently they are satisfied by the same generated structure through its S.U.N. and M.S.U.N. metrics. The evaluated model is trained exclusively on MP-20 and contains 60 million parameters. We sample 2,500 structures from the trained model and relax all candidates using MatterSim-v1.0.0-5M(Yang et al., 2024) to a force convergence threshold of $0.02\,\mathrm{eV}^{-1}$ before entering the LeMat–GenBench pipeline. The benchmark therefore evaluates the local energy minima reached from the generated candidates rather than residual geometric noise in the direct outputs. Validity is determined from compositional and structural checks, including charge balance, interatomic separation, density, and lattice constraints. Uniqueness and novelty are evaluated through symmetry-aware StructureMatcher comparisons within the generated set and against the LeMat-Bulk (Siron et al., 2025) reference collection. Thermodynamic stability is assessed using the standardized multi-MLIP pipeline, with $E_{\mathrm{hull}} \leq 0$ classified as stable and $0 < E_{\mathrm{hull}} \leq 0.1\,\mathrm{eV\,atom}^{-1}$ as metastable. S.U.N. and M.S.U.N. denote the fractions of all generated structures that are simultaneously stable or metastable, unique, and novel, respectively.

Creator reaches 99.64% validity after relaxation, corresponding to 2,491 valid structures among the 2,500 generated samples. Its uniqueness rate remains 97.84%, yielding 2,446 non-redundant structures. The proximity of these two values is notable. The near-complete validity shows that the generated compositions, lattices, and atomic coordinates reliably relax into physically admissible crystals, while the similarly high uniqueness shows that this reliability is not achieved by repeatedly sampling a narrow set of structures. Relative to the models listed in Table 2, Creator improves validity by 2.44 percentage points over the largest preceding value while preserving broad structural coverage.

The most pronounced improvement appears in the thermodynamic distribution. Creator achieves stable and metastable fractions of 13.28% and 68.28%, respectively, placing 81.56% of all generated structures on or within $0.1\,\mathrm{eV\,atom}^{-1}$ of the predicted convex hull. This corresponds to 332 stable and 1,707 metastable candidates. The stable fraction slightly exceeds the largest comparator value of 12.70%, while the metastable fraction is 6.28 percentage points above the next largest value of 62.00%. The result indicates that Creator concentrates a substantially larger fraction of its generated distribution in physically relevant low-energy regions rather than producing structures that are merely geometrically valid.

This energetic advantage remains after uniqueness and novelty filtering. Creator reaches a S.U.N. rate of 5.20% and an M.S.U.N. rate of 31.72%, corresponding to 130 stable–unique–novel and 793 metastable–unique–novel structures. Compared with the largest values among the listed models, S.U.N. increases from 1.50% to 5.20%, while M.S.U.N. increases from 22.60% to 31.72%, a relative improvement of approximately 40%. These gains do not arise from maximizing novelty alone. Creator's novelty rate of 53.48% is comparable to several baselines and lower than that of Chemeleon2 (Park and Walsh, 2026). The higher M.S.U.N. instead shows that a larger fraction of its novel outputs also remains non-redundant and thermodynamically competitive after relaxation. Creator therefore improves the joint balance among structural validity, diversity, novelty, and low-energy stability rather than optimizing any single metric in isolation.

5 Crystal Structure Prediction

Crystal structure prediction (CSP) seeks to explore the polymorphic structural landscape of a prescribed chemical composition by identifying energetically favorable crystal structures. As a central problem in computational materials science, a single chemical composition may support multiple competing crystal structures with distinct thermodynamic stability and physical properties (Oganov et al., 2019). The ability to

Table 2 Performance comparison of crystal generation models on MP-20 dataset.

Model	Training Set	Valid (%)	Unique (%)	Novel (%)	Stable (%)	Metastable (%)	SUN (%)	MSUN (%)
Crystal-GFN (Mistal et al., 2023)	MP-20	51.70	51.70	51.70	0.00	0.00	0.00	0.00
LLaMat3 (Ahlawat et al., 2026)	MP-20	15.40	15.20	10.50	0.10	2.10	0.00	0.20
WyFormer* (Kazeev et al., 2025)	MP-20	93.40	93.00	66.40	0.50	15.70	0.10	1.90
SymmCD (Levy et al., 2025)	MP-20	73.40	73.00	47.00	1.40	18.60	0.10	2.40
Zatom-1-WD (Morehead et al., 2026)	MP-20, QM9	67.20	65.40	24.20	0.20	22.80	0.00	0.40
WyFormer-DFT* (Kazeev et al., 2025)	MP-20	95.20	95.00	66.40	3.70	24.80	0.40	7.80
DiffCSP++ (Jiao et al., 2024)	MP-20	95.30	95.10	62.00	1.00	26.40	0.20	5.00
CrystalFormer (Cao et al., 2025)	MP-20	69.90	69.40	31.80	1.40	28.80	0.00	3.10
DiffCSP (Jiao et al., 2023)	MP-20	95.70	94.80	66.20	2.30	29.80	0.10	8.50
MatterGen* (Zeni et al., 2025)	MP-20	95.70	95.10	70.50	2.00	33.40	0.20	15.00
Chemeleon1 (Park et al., 2025)	MP-20	95.40	94.80	64.00	3.60	34.40	0.30	11.30
LLaMat2 (Ahlawat et al., 2026)	MP-20	84.40	81.40	30.00	0.70	34.70	0.10	2.10
ADiT (Joshi et al., 2025)	MP-20	90.60	87.80	26.00	0.40	36.50	0.00	1.00
Chemeleon2 (Park and Walsh, 2026)	MP-20	95.20	88.10	71.60	0.00	39.80	0.00	21.20
CrystaLLM-pi (Antunes et al., 2024)	MP-20	86.80	84.90	24.90	2.90	49.00	0.30	3.60
MCFlow* (Seong et al., 2026)	MP-20	97.20	96.30	52.20	11.90	49.30	0.70	18.90
OMatG* (Höllmer et al., 2025)	MP-20	96.40	95.20	51.20	11.60	49.80	1.00	18.00
Zatom-1-WD* (Morehead et al., 2026)	MP-20, QM9	88.30	86.00	37.90	4.70	50.50	0.30	8.80
Crystalite* (Hadži Veljković et al., 2026)	MP-20	97.20	95.80	53.20	12.70	51.60	1.50	22.60
OMatG-FC* (Martirossyan et al., 2026)	MP-20	97.20	92.80	28.90	18.40	57.90	1.70	12.00
PLaID++* (Xu et al., 2026)	MP-20	96.00	77.80	24.20	12.40	60.70	1.00	7.60
MiAD* (Okhotin et al., 2026)	MP-20	96.20	94.30	40.20	6.40	62.00	1.00	16.60
Creator	MP-20	97.40	96.16	**82.00**	0.80	34.00	0.28	20.60
Creator*	MP-20	**99.64**	**97.84**	53.48	**13.28**	**68.28**	**5.20**	**31.72**

*Structures relaxed.# Sorted by Metastable rate ascending.

reliably identify such low-energy polymorphs has enabled experimentally validated materials discovery across diverse systems, including high-temperature superconducting hydrides, porous frameworks, and molecular crystals (Ma et al., 2022; O'Shaughnessy et al., 2024; Zhu et al., 2024). Unlike unconditional crystal generation, CSP fixes the chemical composition while exploring only the lattice vectors and fractional atomic coordinates, enabling diverse polymorphs to be sampled for a given chemistry (Xie et al., 2022a; Jiao et al., 2023; Zeni et al., 2025). This formulation naturally aligns with conditional generative modeling, where the composition serves as an external conditioning signal while the crystal geometry remains the generation target. The pre-trained Creator backbone can be readily adapted to composition-conditioned CSP by augmenting it with a lightweight composition-conditioning module, while leaving the diffusion backbone unchanged.

Each target composition is encoded as a 7×18 period–group matrix following the layout of the periodic table. The resulting composition embedding is projected to the hidden dimension and combined with the timestep and task embeddings, $\mathbf{e} = \mathbf{e}_{\text{task}} + \mathbf{e}_{\text{time}} + \mathbf{e}_{\text{comp}}$, before being injected into every DiT block through the existing adaLN-Zero conditioning pathway. Classifier-free guidance is employed during sampling to further strengthen composition conditioning (Ho and Salimans, 2022). Starting from the pre-trained unconditional Creator checkpoint, the model is fine-tuned on the MP-20 dataset (Xie et al., 2022a). For each target composition, structures containing one to four formula units are generated, and subsequently relaxed using the pre-trained MatterSim-v1-5M machine-learning potential before structural analysis (Yang et al., 2024).

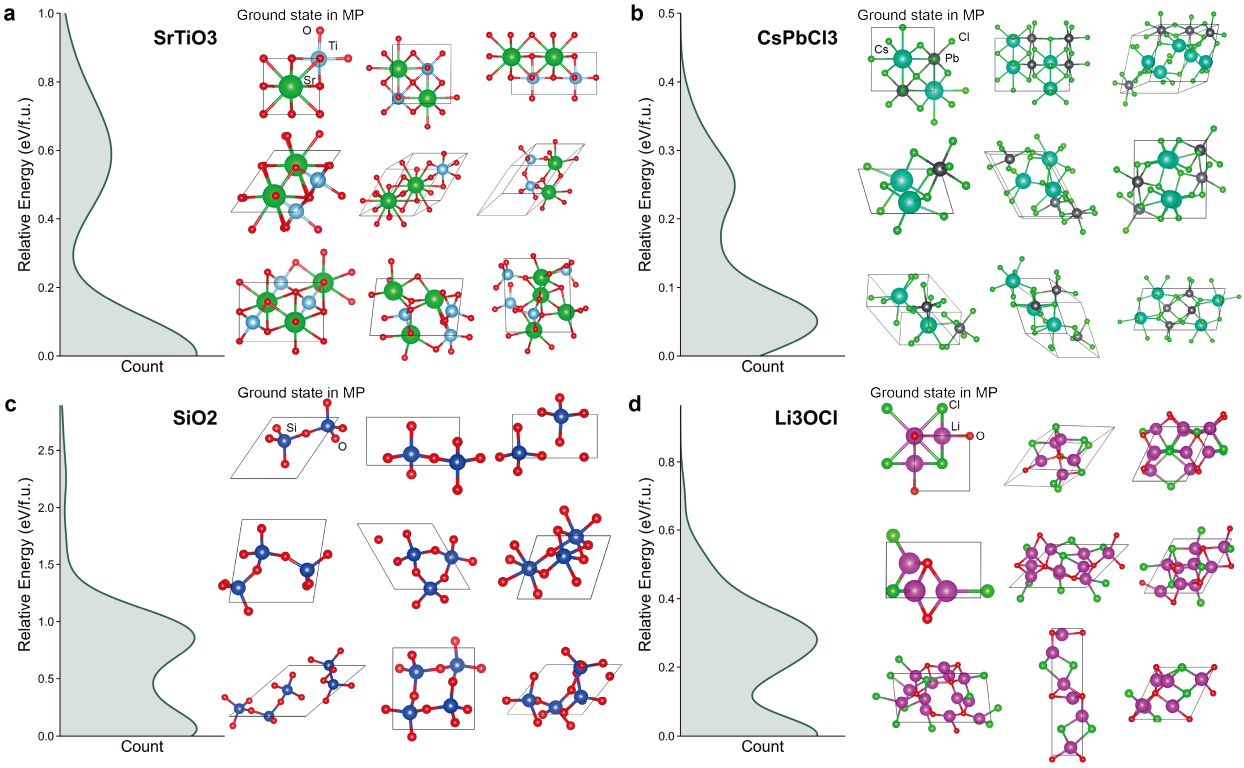


Figure 4 Crystal structure prediction (CSP) results for four representative compositions. **a–d**, Smoothed distributions of the relaxed energies for $SrTiO_3$ (**a**), $CsPbCl_3$ (**b**), SiO_2 (**c**), and Li_3OCl (**d**), together with nine representative predicted structures for each composition. The MP reference ground-state structure is indicated in each panel. Relative energies are reported per formula unit and referenced to the MP reference ground-state for each composition. For each composition, 200 structures are generated, comprising 50 samples for each of the one- to four-formula-unit systems. The generated structures are retained without structural deduplication and subsequently relaxed using MatterSim-v1-5M to $f_{\max} < 0.01$ eV^{-1}.

To assess the ability of composition-conditioned Creator to generate physically relevant crystal structures, we generated 200 candidates for each of four representative compositions ($SrTiO_3$, $CsPbCl_3$, SiO_2, and Li_3OCl), followed by MLFF relaxation and structural analysis. For all four systems, the corresponding MP reference structures (Jain et al., 2013) are recovered as the lowest-energy generated candidates, demonstrating that

composition-conditioned Creator consistently concentrates sampling within the physically relevant low-energy regions of the structural energy landscape across chemically distinct materials.

The generated energy landscapes exhibit marked chemistry-dependent characteristics (Fig. 4). $CsPbCl_3$ shows a compact low-energy basin, indicating that only a limited number of polymorphs compete energetically, whereas SiO_2 displays a broad multimodal distribution spanning a much wider energy range, consistent with its substantially more rugged polymorphic landscape. $SrTiO_3$ and Li_3OCl exhibit intermediate behaviour, with dense sampling around the lowest-energy basin together with smaller populations of higher-energy polymorphs. Consistent with these distributions, 72%, 100%, 34%, and 85% of the generated candidates for $SrTiO_3$, $CsPbCl_3$, SiO_2, and Li_3OCl, respectively, lie within 0.1 eV/atom of the MLFF-estimated convex hull. The substantially lower near-hull fraction for SiO_2 further reflects its broader polymorphic energy landscape, whereas the other three systems are sampled predominantly within thermodynamically competitive regions. Together, the successful recovery of the MP reference structures, the high proportion of near-hull candidates, and the chemistry-dependent energy landscapes demonstrate that composition-conditioned Creator efficiently explores low-energy structural basins while preserving the diversity required to sample competing polymorphs. These representative examples illustrate how a unified composition-conditioned generative framework can simultaneously recover known low-energy structures and capture the distinct polymorphic energy landscapes characteristic of different chemistries.

6 De Novo Materials Design

In addition to crystal structure prediction, Creator can be naturally extended to inverse materials design through continuous property conditioning (Zeni et al., 2025). Here we consider the electronic band gap as a representative target property because it fundamentally governs the electronic response of crystalline materials, determining whether they exhibit metallic, semiconducting or insulating behavior. Band gap also plays a central role in the design of semiconductors, photovoltaics, photocatalysts and optoelectronic devices, making it one of the most widely optimized quantities in computational materials discovery (Wang et al., 2024; Lin et al., 2023; Gong et al., 2026). Unlike CSP, where the chemical composition is specified a priori, band-gap inverse design requires neither predefined elemental species nor stoichiometry. Instead, the target band gap serves as the sole external condition, while the composition, lattice parameters and atomic coordinates are generated jointly during the reverse diffusion process.

To enable property-conditioned generation, lightweight band-gap adapters are introduced into the pretrained Creator model while the original generative backbone remains frozen. The target band gap is encoded as a continuous sinusoidal embedding and injected into every DiT block through the existing adaLN conditioning pathway (Peebles and Xie, 2023; Ho and Salimans, 2022), allowing the property condition to modulate the denoising of atom types, lattice parameters and atomic coordinates. Classifier-free guidance is further employed during sampling to enhance the influence of the prescribed band gap. The adapters are fine-tuned on the MatBench `mp_gap` dataset (Dunn et al., 2020). Because the dataset is strongly biased toward metallic and small-gap materials, a tempered inverse-frequency sampling strategy is adopted to increase the representation of sparsely populated large-band-gap regions without discarding the abundant metallic examples. During inference, structures are generated without specifying their composition, subsequently relaxed using MatterSim-v1-5M with variable-cell optimization, and finally evaluated using the pretrained `mp_gappbe_alignn` band-gap predictor (Choudhary and DeCost, 2021).

Figure 5 summarizes the predicted band-gap distributions for four representative targets spanning metallic (0 eV), narrow-gap (0.5 eV), semiconducting (2 eV), and wide-gap insulating (6 eV) regimes. The predicted distributions shift progressively toward higher band-gap values as the prescribed target increases, with their centers remaining close to the target values across the entire 0–6 eV range, demonstrating effective conditional control over the generated electronic properties. Quantitatively, the mean absolute deviation from the prescribed target increases from 0.327 eV for the metallic target to 2.282 eV for the 6 eV target, while the corresponding success rate within the predefined ± 0.45 eV tolerance decreases from 82.6% to 19.6%, indicating that wide-gap materials remain more challenging to generate accurately. Nevertheless, representative candidates satisfying the success criterion are obtained for all four target regimes. Despite sharing similar target band gaps, the generated candidates exhibit substantial diversity in composition, stoichiometry, and crystal structure. Metallic and narrow-gap candidates are enriched in intermetallic and

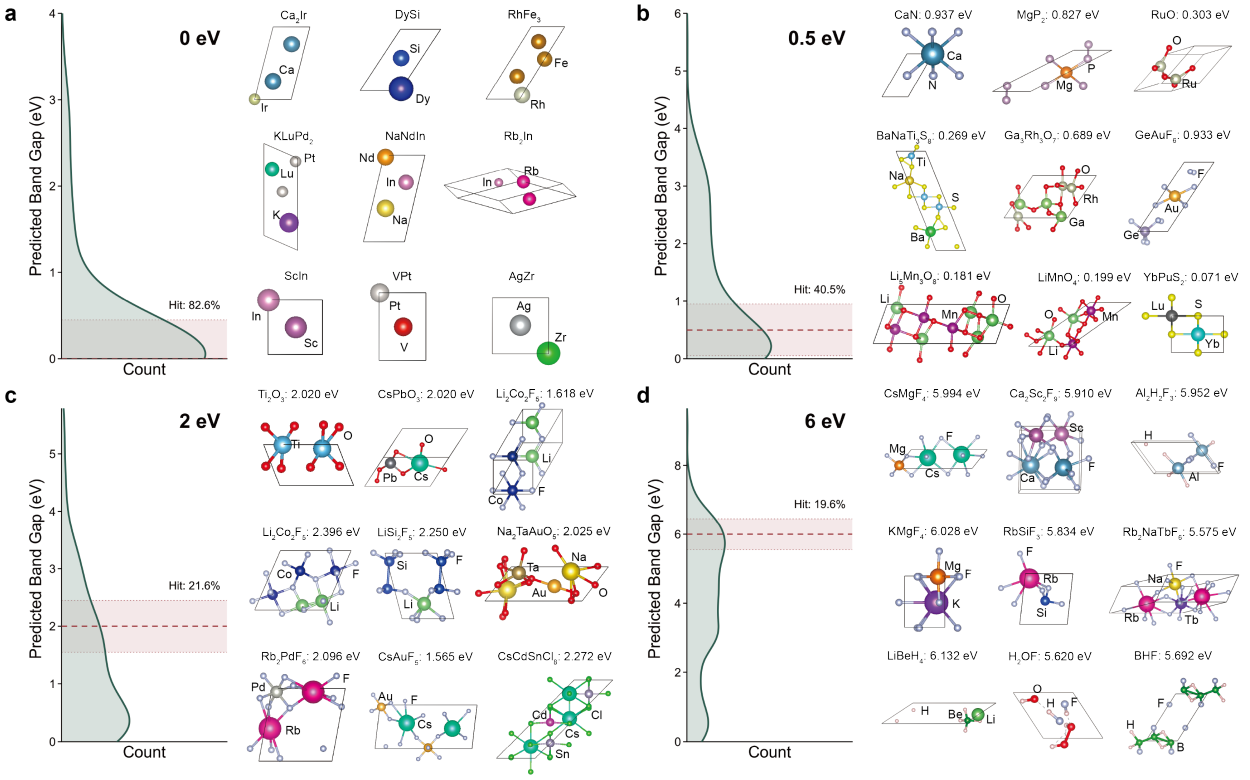


Figure 5 Band-gap inverse design conditioned on target band gaps. **a–d**, Results for target band gaps of 0 eV (**a**), 0.5 eV (**b**), 2 eV (**c**), and 6 eV (**d**). The green shaded curves show the smoothed distributions of the predicted band gaps, the red dashed lines indicate the prescribed target values, and the light-red regions denote the ±0.45 eV success windows. The annotated hit percentages indicate the fractions of valid predictions falling within the corresponding success windows. Nine representative structures are shown for each target, selected from generated candidates satisfying $|E_{\mathrm{g}}^{\mathrm{pred}} - E_{\mathrm{g}}^{\mathrm{target}}| \leq 0.45$ eV. Predicted band gaps are annotated above the representative structures for the nonzero targets. For each target, 300 structures were generated, relaxed using MatterSim-v1-5M with variable-cell optimization until $f_{\max} < 0.01$ eV $^{-1}$, and evaluated using `mp_gappbe_alignn` (Choudhary and DeCost, 2021).

transition-metal-containing compounds, whereas wide-gap candidates are predominantly ionic materials containing highly electronegative elements such as F and O, consistent with established chemical principles governing electronic band gaps. Together, these results demonstrate that the same conditional generative framework can generate chemically diverse candidate materials across a broad range of target electronic properties while maintaining effective control over the prescribed band-gap objective.

Together, these results demonstrate that Creator provides a unified conditional generation framework for materials discovery. By conditioning on either chemical composition or target material properties, the same generative model supports both crystal structure prediction and inverse materials design, enabling flexible exploration of diverse materials design problems within a single architecture.

7 Limitations and Future Directions

Despite its strong generation performance, Creator remains limited by both its geometric formulation and the scale of the present study. Multivector enhancement provides a structured representation of lattice vectors and atomic coordinates, but exact algebraic grades are not preserved after dense feature fusion and the resulting architecture is not strictly equivariant to three-dimensional rotations or reflections. Our previous effort in implementation of an equivariant architecture which fully respects the algebraic structure of Clifford algebra results in large training overhead and needs to be further explored. Periodicity is represented through lattice-anchor tokens and atomic coordinates rather than enforced throughout the attention mechanism, so equivalent

choices of unit cell, lattice basis or periodic atomic image may be treated differently by the model. The current formulation further assumes ideal, fully occupied periodic crystals and does not explicitly represent defects, disorder, partial occupancies, surfaces, interfaces or non-stoichiometric structures. Finally, the conditional studies demonstrate composition-guided structure generation and band-gap-guided de novo design, but do not yet establish generality across multiple coupled properties, synthesis constraints or experimentally defined objectives.

Future work should incorporate periodicity more directly into both the representation and interaction mechanism, making the model less sensitive to equivalent unit-cell choices and explicitly accounting for relationships across periodic images. More extensive component ablations are also needed to separate the contributions of Clifford-parameterized multivector lifting, lattice-anchored tokenization and multivector-partitioned attention, and to determine which aspects of the geometric interface are responsible for the observed improvement over a conventional DiT. Training on larger and more heterogeneous datasets would provide a stronger test of model-capacity scaling and transfer beyond MP-20, particularly for larger cells, higher-order compositions, metastable structures and chemically underrepresented regions. Extending the representation to defects, disorder, variable occupancy, low-dimensional materials and interfaces would move Creator closer to experimentally realized materials. On the application side, the conditional framework could be expanded from individual targets to multiple, potentially competing objectives involving stability, electronic, mechanical, magnetic and transport properties, together with practical constraints on composition, cost and synthesizability.

References

Dhruv Ahlawat, Vaibhav Mishra, Somaditya Singh, Mohd Zaki, Vaibhav Bihani, Hargun Singh Grover, Biswajit Mishra, Santiago Miret, Mausam, and N. M. Anoop Krishnan. A family of large language models for materials research with insights into model adaptability in continued pretraining. *Nature Machine Intelligence*, 8(3):435–448, 2026. doi: 10.1038/s42256-026-01199-8.

Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15:10570, 2024. doi: 10.1038/s41467-024-54639-7.

Johannes Brandstetter, Rianne van den Berg, Max Welling, and Jayesh K. Gupta. Clifford neural layers for PDE modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhendong Cao, Xiaoshan Luo, Jian Lv, and Lei Wang. Space group informed transformer for crystalline materials generation. *Science Bulletin*, 70(21):3522–3533, 2025. doi: 10.1016/j.scib.2025.09.035.

Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021. doi: 10.1038/s41524-021-00650-1. URL https://doi.org/10.1038/s41524-021-00650-1.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020. doi: 10.1038/s41524-020-00406-3. URL https://doi.org/10.1038/s41524-020-00406-3.

Alexandre Duval, Siddharth Betala, Samuel P. Gleason, Andy Xu, Georgia Channing, Daniel Levy, Ali Ramlaoui, Clémentine Fourrier, Chaitanya K. Joshi, Nikita Kazeev, Sékou-Oumar Kaba, Félix Therrien, Alex Hernández-García, Rocío Mercado, and N M Anoop Krishnan. Lemat-genbench: Bridging the gap between crystal generation and materials discovery. In *AI for Accelerated Materials Design – NeurIPS 2025*, 2025.

Hehe Gong, Xin Yang, Boyan Wang, Zichen Zhang, Qingrui Yuchi, Zineng Yang, Matthew Porter, Hongchang Cui, Yuan Qin, Rong Zhang, Han Wang, Dong Dong, Jiandong Ye, Guo-Quan Lu, and Yuhao Zhang. A megawatt ultra-wide bandgap semiconductor module for pulsed power electronics. *Nature Communications*, 17(1):4783, 2026. doi: 10.1038/s41467-026-71274-6. URL https://doi.org/10.1038/s41467-026-71274-6.

Tin Hadži Veljković, Joshua Rosenthal, Ivor Lončarić, and Jan-Willem van de Meent. Crystalite: A lightweight transformer for efficient crystal modeling, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

Philipp Höllmer, Thomas Egg, Maya Martirossyan, Eric Fuemmeler, Zeren Shui, Amit Gupta, Pawan Prakash, Adrian Roitberg, Mingjie Liu, George Karypis, Mark Transtrum, Richard Hennig, Ellad B. Tadmor, and Stefano Martiniani. Open materials generation with stochastic interpolants. In *Forty-second International Conference on Machine Learning*, 2025.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. doi: 10.1063/1.4812323. URL https://doi.org/10.1063/1.4812323.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. In *International Conference on Learning Representations*, pages 6836–6853, 2024.

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

Nikita Kazeev, Wei Nong, Ignat Romanov, Ruiming Zhu, Andrey E. Ustyuzhanin, Shuya Yamazaki, and Kedar Hippalgaonkar. Wyckoff transformer: Generation of symmetric crystals. In *Forty-second International Conference on Machine Learning*, 2025.

Daniel Levy, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Qiang Zhu, Kin Long Kelvin Lee, Mikhail Galkin, Santiago Miret, and Siamak Ravanbakhsh. Symmcd: Symmetry-preserving crystal generation with diffusion models. In *International Conference on Learning Representations*, pages 23166–23189, 2025.

Zhixun Li, Bin Cao, Rui Jiao, Liang Wang, Ding Wang, Yang Liu, Dingshuo Chen, Jia Li, Qiang Liu, Yu Rong, Liang Wang, Tong-yi Zhang, and Jeffrey Xu Yu. Materials generation in the era of artificial intelligence: A comprehensive survey. *CoRR*, abs/2505.16379, 2025.

Renxing Lin, Yurui Wang, Qianwen Lu, Beibei Tang, Jiayi Li, Han Gao, Yuan Gao, Hongjiang Li, Changzeng Ding, Jin Wen, Pu Wu, Chenshuaiyu Liu, Siyang Zhao, Ke Xiao, Zhou Liu, Changqi Ma, Yu Deng, Ludong Li, Fengjia Fan, and Hairen Tan. All-perovskite tandem solar cells with 3D/3D bilayer perovskite heterojunction. *Nature*, 620 (7976):994–1000, 2023. doi: 10.1038/s41586-023-06278-z. URL https://doi.org/10.1038/s41586-023-06278-z.

X. Luo, Z. Wang, Q. Wang, et al. CrystalFlow: A flow-based generative model for crystalline materials. *Nature Communications*, 16:9267, 2025. doi: 10.1038/s41467-025-64364-4.

Liang Ma, Kui Wang, Yu Xie, Xin Yang, Yingying Wang, Mi Zhou, Hanyu Liu, Xiaohui Yu, Yongsheng Zhao, Hongbo Wang, Guangtao Liu, and Yanming Ma. High-temperature superconducting phase in clathrate calcium hydride CaH_6 up to 215 K at a pressure of 172 GPa. *Physical Review Letters*, 128(16):167001, 2022. doi: 10.1103/PhysRevLett.128.167001. URL https://doi.org/10.1103/PhysRevLett.128.167001.

Maya M. Martirossyan, Hillary Pan, Philipp Höllmer, and Stefano Martiniani. Framework-constrained materials generation. In *AI for Accelerated Materials Design – ICLR 2026*, 2026.

Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M. Wood. FlowMM: Generating materials with Riemannian flow matching. In *International Conference on Machine Learning (ICML)*, 2024.

Mistal, Alex Hernández-García, Alexandra Volokhova, Alexandre A. G. M. Duval, Yoshua Bengio, Divya Sharma, Pierre Luc Carrier, Michał Koziarski, and Victor Schmidt. Crystal-GFN: Sampling materials with desirable properties and constraints. In *AI for Accelerated Materials Design – NeurIPS 2023 Workshop*, 2023.

Alex Morehead, Miruna Cretu, Antonia Panescu, Rishabh Anand, Maurice Weiler, Ty Perez, Samuel M. Blau, Steven Farrell, Wahid Bhimji, Anubhav Jain, Hrushikesh Sahasrabuddhe, Pietro Lio, Tommi Jaakkola, Rafael Gomez-Bombarelli, Rex Ying, N. Benjamin Erichson, and Michael W. Mahoney. Zatom-1: A multimodal flow foundation model for 3D molecules and materials. In *ICLR 2026 Workshop on Foundation Models for Science*, 2026.

Juhwan Noh, Jaehoon Kim, Helge S. Stein, Benjamin Sanchez-Lengeling, John M. Gregoire, Alan Aspuru-Guzik, and Yousung Jung. Inverse design of solid-state materials via a continuous representation. *Matter*, 1(5):1370–1384, 2019. doi: 10.1016/j.matt.2019.08.017.

Artem R. Oganov, Chris J. Pickard, Qiang Zhu, and Richard J. Needs. Structure prediction drives materials discovery. *Nature Reviews Materials*, 4:331–348, 2019. doi: 10.1038/s41578-019-0101-8. URL https://doi.org/10.1038/s41578-019-0101-8.

Andrey Okhotin, Maksim Nakhodnov, Nikita Kazeev, Mikhail Lazarev, Andrey E. Ustyuzhanin, and Dmitry Vetrov. Miad: Mirage atom diffusion for de novo crystal generation, 2026.

Megan O'Shaughnessy, Joseph Glover, Roohollah Hafizi, Mounib Barhi, Rob Clowes, Samantha Y. Chong, Stephen P. Argent, Graeme M. Day, and Andrew I. Cooper. Porous isoreticular non-metal organic frameworks. *Nature*, 630 (8015):102–108, 2024. doi: 10.1038/s41586-024-07353-9. URL https://doi.org/10.1038/s41586-024-07353-9.

H. Park and A. Walsh. Guiding generative models to uncover diverse and novel crystals via reinforcement learning. *Nature Machine Intelligence*, 2026. doi: 10.1038/s42256-026-01262-4.

H. Park, A. Onwuli, and A. Walsh. Exploration of crystal chemical space using text-guided generative artificial intelligence. *Nature Communications*, 16:4379, 2025. doi: 10.1038/s41467-025-59636-y.

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

David Ruhe, Jayesh K. Gupta, Steven De Keninck, Max Welling, and Johannes Brandstetter. Geometric clifford algebra networks. In *International Conference on Machine Learning*, 2023.

Kiyoung Seong, Sungsoo Ahn, Sehui Han, and Changyoung Park. Multimodal crystal flow: Any-to-any modality generation for unified crystal modeling. In *Forty-fourth International Conference on Machine Learning (ICML)*, 2026. Poster.

Martin Siron, Inel Djafar, Etienne du Fayet, Amandine Rossello, Ali Ramlaoui, and Alexandre Duval. Lemat-bulk: Aggregating, and de-duplicating quantum chemistry materials databases. In *AI for Accelerated Materials Design – ICLR 2025*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yongjie Wang, Lucheng Peng, Julien Schreier, Yu Bi, Andres Black, Aditya Malla, Stijn Goossens, and Gerasimos Konstantatos. Silver telluride colloidal quantum dot infrared photodetectors and image sensors. *Nature Photonics*, 18(3):236–242, 2024. doi: 10.1038/s41566-023-01345-3. URL https://doi.org/10.1038/s41566-023-01345-3.

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

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations*, 2022b.

Andy Xu, Rohan Desai, Larry Wang, Ethan T. Ritz, and Gabriel Hope. PLaID++: A preference aligned language model for targeted inorganic materials design. In *Forty-third International Conference on Machine Learning*, 2026.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zügner, Tian Xie, Jake Smith, Lixin Sun, Qian Wang, Lingyu Kong, Chang Liu, Hongxia Hao, and Ziheng Lu. MatterSim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024. URL https://arxiv.org/abs/2405.04967.

Xiaohan Yi, Guikun Xu, Zhong Zhang, Liu Liu, Yatao Bian, Xi Xiao, and Peilin Zhao. CrystalDiT: Simple diffusion transformers for crystal generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(2):1462–1470, 2026. doi: 10.1609/aaai.v40i2.37121.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639:624–632, 2025. doi: 10.1038/s41586-025-08628-5.

Qiang Zhu, Hang Qu, Gokay Avci, Roohollah Hafizi, Chengxi Zhao, Graeme M. Day, Kim E. Jelfs, Marc A. Little, and Andrew I. Cooper. Computationally guided synthesis of a hierarchical [4[2+3]+6] porous organic "cage of cages". *Nature Synthesis*, 3:825–834, 2024. doi: 10.1038/s44160-024-00531-7. URL https://doi.org/10.1038/s44160-024-00531-7.

Appendices

A Clifford Algebra

In this appendix we provide a concise introduction to Clifford algebras, establishing the notation and definitions used throughout the paper.

A.1 Definitions

Algebras over a field. An algebra over a field $\mathbb{F}$ consists of a vector space V over $\mathbb{F}$ together with a bilinear product $V \times V \to V$. That is, for any $a, b \in V$, the product ab is also in V, and the product satisfies left and right distributivity:

$$a(\lambda_1 b + \lambda_2 c) = \lambda_1 ab + \lambda_2 ac, \quad (\lambda_1 a + \lambda_2 b)c = \lambda_1 ac + \lambda_2 bc \tag{16}$$

for all $\lambda_1, \lambda_2 \in \mathbb{F}$ and $a, b, c \in V$.

Real Clifford algebras. A real Clifford algebra is generated by an n-dimensional vector space $\mathbb{R}^n$ through a set of relations imposed on the basis elements. Let $\{e_1, e_2, \ldots, e_n\}$ be an orthonormal basis of $\mathbb{R}^n$. Given nonnegative integers p, q with $p + q = n$, the Clifford algebra $\mathrm{Cl}_{p,q}(\mathbb{R})$ of signature (p, q) is defined by the relations:

$$e_i^2 = +1 \quad \text{for } 1 \le i \le p, \tag{17}$$

$$e_j^2 = -1 \quad \text{for } p < j \le p + q, \tag{18}$$

$$e_i e_j = -e_j e_i \quad \text{for } i \ne j. \tag{19}$$

The vector space of the Clifford algebra, denoted $\mathcal{G}^n$, is spanned by all products of the generating basis elements. Since $e_i e_j = -e_j e_i$, any product can be reordered (up to sign) so that indices are strictly increasing. The basis is therefore given by:

$$\{1,\ e_1, e_2, \ldots, e_n,\ e_1e_2, e_1e_3, \ldots, e_{n-1}e_n,\ \ldots,\ e_1e_2\cdots e_n\} \tag{20}$$

The dimension of $\mathcal{G}^n$ is $2^n = 2^{p+q}$.

A.2 Grade Structure

An important structural property of Clifford algebras is their grading by the number of basis vectors in each element.

Grade. The *grade* of a basis element is the number of generating vectors it contains. The basis element 1 has grade 0 (a scalar), e_i have grade 1 (vectors), $e_i e_j$ $(i < j)$ have grade 2 (bivectors), and so on. The vector space $\mathcal{G}^n$ decomposes as a direct sum of grade subspaces:

$$\mathcal{G}^n = \mathcal{M}_0 \oplus \mathcal{M}_1 \oplus \cdots \oplus \mathcal{M}_n \tag{21}$$

where $\mathcal{M}_k$ is the subspace of all grade-k elements.

Multivectors. A general element of the Clifford algebra, called a *multivector*, may contain components of multiple grades:

$$\mathbf{X} = x_0 + \sum_i x_i e_i + \sum_{i<j} x_{ij} e_i e_j + \cdots + x_{1\cdots n} e_1 \cdots e_n \tag{22}$$

We write $\langle \mathbf{X} \rangle_k$ to denote the grade-k component of $\mathbf{X}$.

The Clifford algebra $Cl_{3,0}(\mathbb{R})$. In this work we work exclusively with the three-dimensional Euclidean Clifford algebra $\mathrm{Cl}_{3,0}(\mathbb{R})$. It has dimension $2^3 = 8$, with basis elements:

Grade	0	1	2	3
Elements	1	e_1, e_2, e_3	e_{12}, e_{13}, e_{23}	e_{123}
Name	scalar	vector	bivector	trivector / pseudoscalar
Count	1	3	3	1

where we use the shorthand $e_{ij} = e_i e_j$ and $e_{123} = e_1 e_2 e_3$. The highest-grade element e_{123} is also called the *pseudoscalar*.

A.3 Geometric Product and Inner Product

The product of the Clifford algebra is called the *geometric product*. For two vectors $\mathbf{u}, \mathbf{v} \in \mathcal{M}_1$, their geometric product decomposes into a scalar (grade 0) and a bivector (grade 2):

$$\mathbf{uv} = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \wedge \mathbf{v} \tag{23}$$

where $\mathbf{u} \cdot \mathbf{v} = \langle \mathbf{uv} \rangle_0$ is the standard inner product (symmetric) and $\mathbf{u} \wedge \mathbf{v} = \langle \mathbf{uv} \rangle_2$ is the outer product (anti-symmetric).

More generally, the geometric product of two multivectors of grades k and l produces components of grades $|k-l|, |k-l|+2, \ldots, k+l$. The inner product between multivectors is the lowest-grade component of their geometric product:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \langle \mathbf{AB} \rangle_0 \tag{24}$$

which for vectors reduces to the standard dot product.

B Diffusion Models for Crystal Generation

Diffusion Models. Diffusion models (Ho et al., 2020) generate samples by progressively perturbing data with noise and learning the corresponding reverse denoising process. Given a clean sample $\mathbf{x}_0 \sim p_{\text{data}}$, the forward process produces a noisy sample at timestep t according to

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{25}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\{\beta_t\}_{t=1}^{T}$ is a predefined noise schedule. Equivalently, $\mathbf{x}_t$ can be sampled directly as

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{26}$$

A commonly used parameterization trains a denoising network $\boldsymbol{\epsilon}_\theta$ to predict the injected noise by minimizing

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0, t, \boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|_2^2\right]. \tag{27}$$

Alternative formulations may instead predict the clean sample, the score function, or a velocity variable. During generation, the reverse process starts from Gaussian noise and progressively transforms it into a sample from the learned data distribution through a sequence of denoising transitions. In the continuous-time formulation introduced by score-based generative modeling (Song et al., 2021), the forward perturbation process is described by a stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \tag{28}$$

where $\mathbf{f}(\mathbf{x}, t)$ and $g(t)$ denote the drift and diffusion coefficients, respectively, and $\mathbf{w}$ is a standard Wiener process. The corresponding reverse-time SDE is

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}, \tag{29}$$

where time is integrated from the terminal time back to zero, $\bar{\mathbf{w}}$ denotes a Wiener process in reverse time, and the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is estimated by a neural network trained through denoising score matching. Applying diffusion models to crystals introduces additional geometric constraints compared with conventional data domains. Atomic coordinates obey periodic boundary conditions, scalar physical properties are invariant under global rotations and translations, and lattice parameters are strongly coupled with atomic configurations. Crystal diffusion models must therefore account for periodicity, permutation symmetry, Euclidean invariance or equivariance, and spatial correlations throughout the denoising process. Existing approaches, including CDVAE (Xie et al., 2022b), DiffCSP (Jiao et al., 2023), and MatterGen (Zeni et al., 2025), address these requirements using invariant or equivariant geometric neural networks.

Diffusion Transformers. A Diffusion Transformer (DiT) (Peebles and Xie, 2023) is a diffusion model that employs a Transformer as its denoising network instead of the convolutional U-Net commonly used in image diffusion models. Given a noisy sample $\mathbf{x}_t$, DiT first projects the input into a sequence of token representations $\mathbf{H} \in \mathbb{R}^{N \times d}$ and then processes them using a stack of Transformer blocks. Each block employs multi-head self-attention to model interactions among tokens. For a single attention head, this operation is defined as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathsf{T}}}{\sqrt{d_h}}\right)\mathbf{V}, \tag{30}$$

where

$$\mathbf{Q} = \mathbf{H}\mathbf{W}_Q, \qquad \mathbf{K} = \mathbf{H}\mathbf{W}_K, \qquad \mathbf{V} = \mathbf{H}\mathbf{W}_V, \tag{31}$$

and d_h is the feature dimension of each attention head.

Since the appropriate denoising operation depends on the current noise level, DiT encodes the diffusion timestep into a conditioning vector. When additional conditioning information is available, the combined conditioning representation can be written as

$$\mathbf{c}_t = \text{TimeEmbed}(t) + \text{CondEmbed}(\mathbf{c}), \tag{32}$$

where $\mathbf{c}$ may represent information such as a class label, chemical composition, or target material property; the second term is omitted for unconditional generation. In the standard DiT architecture, a modulation network maps $\mathbf{c}_t$ to the scale and shift parameters used in adaptive layer normalization:

$$[\mathbf{s}_t, \mathbf{b}_t] = \text{MLP}_{\text{mod}}(\mathbf{c}_t), \tag{33}$$

$$\text{AdaLN}(\mathbf{H}; \mathbf{c}_t) = (\mathbf{1} + \mathbf{s}_t) \odot \text{LN}(\mathbf{H}) + \mathbf{b}_t, \tag{34}$$

where $\mathbf{s}_t$ and $\mathbf{b}_t$ denote the conditioning-dependent scale and shift, respectively. The original DiT further employs AdaLN-Zero, in which additional conditioning-dependent gates modulate the self-attention and feed-forward residual branches and are initialized to zero. This initialization allows each Transformer block to begin as an identity mapping and facilitates stable optimization. Other conditioning mechanisms may also be used depending on the model architecture.

After the Transformer blocks, an output projection predicts the noise, score, velocity, or clean sample required by the chosen reverse diffusion parameterization. For crystal generation, atomic species, fractional coordinates, and lattice-related features can be represented as tokens, while periodic relative positions and geometric relations can be incorporated into their interactions. Existing crystal diffusion models commonly introduce invariant or equivariant geometric operations to capture these structural constraints. Although these methods significantly improve crystal generation quality, their geometric operations often rely on computationally expensive equivariant layers. In contrast, Clifford-DiT adopts a scalable Transformer backbone and introduces geometry directly into the attention mechanism, providing geometric awareness while preserving the efficiency of standard Diffusion Transformers.

C Training Configuration

To ensure a rigorous and reproducible comparison, we follow the standardized evaluation protocol of LEMAT-GENBENCH, a unified benchmark for crystal generative models.

We train Clifford-DiT on the MP-20 benchmark dataset, which contains experimentally and computationally characterized inorganic crystal structures with at most 20 atoms per unit cell. Following previous crystal generation studies, we randomly sample 2,500 structures from the trained model for evaluation. All generated samples are processed through the LEMAT-GENBENCH evaluation pipeline, including validity filtering, stability estimation, uniqueness and novelty analysis, and distributional evaluation.

All experiments are conducted on a single node equipped with 8 NVIDIA A800 GPUs (80 GB HBM each). Training is distributed using data parallelism across all 8 GPUs. We use the AdamW optimizer with a cosine learning rate schedule and linear warmup.

Table 3 Training configuration for the main experiments.

Parameter	Value
Model architecture	
Embedding dimension	512
Number of DiT blocks	18
Number of attention heads	8
MLP expansion ratio	4.0×
Timestep embedding dimension	256
Clifford signature	$(+,+,+)$ / $\mathrm{Cl}(3,0)$
Max atoms per crystal	20
Optimization	
Optimizer	AdamW
Batch size (total)	256
Learning rate (peak)	3×10^{-4}
Minimum learning rate	10^{-5}
Weight decay	0.01
Gradient clipping	1.0
LR schedule	Cosine with warmup
Random seed	42
Diffusion	
Timesteps	1000
Noise schedule	Linear
Lattice loss weight	1.0
Position loss weight	100.0
Atom type loss weight	100.0
Infrastructure	
Hardware	8× NVIDIA A800 (80 GB)
Parallelism	Data parallel
Data workers per GPU	8

Dataset. We use the Materials Project 20 (MP-20) subset, which contains crystal structures with up to 20 atoms in the unit cell. Coordinates and lattice vectors are normalized to $[-1, 1]$ before training. The dataset is stored in LMDB format for efficient I/O.

Loss weighting. The total training loss is a weighted combination of three mean squared error terms: lattice prediction, atomic position prediction, and atom-type prediction. Position and type losses receive a weight of 100, while the lattice loss is weighted by 1.0, reflecting the smaller absolute scale of normalized lattice vectors relative to the other targets.

Gradient clipping. Gradients are clipped to a global norm of 1.0 to stabilize training, particularly during early iterations when the multivector-partitioned attention scores can exhibit larger variance.