CrystalDiT: A Diffusion Transformer for Crystal Generation

Xiaohan Yi[1,2], Guikun Xu[3], Zhong Zhang[2], Liu Liu[2], Yatao Bian[4], Xi Xiao[1]*, Peilin Zhao[3]†

[1]Shenzhen International Graduate School, Tsinghua University
[2]Tencent, Shenzhen, China
[3]School of Artificial Intelligence, Shanghai Jiao Tong University
[4]Department of Computer Science, National University of Singapore
xiaox@sz.tsinghua.edu.cn, peilinzhao@sjtu.edu.cn

Abstract

We present **CrystalDiT**, a diffusion transformer for crystal structure generation that achieves state-of-the-art performance by challenging the trend of architectural complexity. Instead of intricate, multi-stream designs, CrystalDiT employs a unified transformer that imposes a powerful inductive bias: treating lattice and atomic properties as a single, interdependent system. Combined with a periodic table-based atomic representation and a balanced training strategy, our approach achieves 8.78% SUN (Stable, Unique, Novel) rate on MP-20, substantially outperforming recent methods including FlowMM (4.21%) and MatterGen (3.66%). Notably, CrystalDiT generates 63.28% unique and novel structures while maintaining comparable stability rates, demonstrating that architectural simplicity can be more effective than complexity for materials discovery. Our results suggest that in data-limited scientific domains, carefully designed simple architectures outperform sophisticated alternatives that are prone to overfitting.

Code — https://github.com/hanyi2021/CrystalDiT.git

Introduction

Materials discovery limits technological advancement in energy and sustainability (Jain et al. 2013; Merchant et al. 2023). Traditional screening approaches constrain exploration to known structures (Curtarolo et al. 2012).

Generative models offer a transformative alternative by directly proposing novel crystal structures. Recent methods have achieved promising results through various sophisticated approaches: ADiT employs two-stage latent diffusion with separate VAE and DiT components (Joshi et al. 2025), FlowMM uses Riemannian flow matching to handle crystal symmetries (Miller et al. 2024), and MatterGen introduces joint diffusion processes with equivariant score networks and adapter modules (Zeni et al. 2025). While these approaches demonstrate the potential of generative modeling, their architectural complexity raises questions about necessity and effectiveness.

While transformer architectures have found success across diverse domains (Feng et al. 2024; Yi et al. 2023; Gong et al. 2024; Chen et al. 2024; Wu et al. 2020; Fang et al. 2024; Chen et al. 2025; Shibo et al. 2025; Wang et al. 2025; Xu et al. 2021), the design principles for data-limited scientific applications remain underexplored.

We identify two critical issues with current approaches. First, evaluation metrics emphasize structural similarity to training data, inadvertently penalizing the novelty essential for discovery (Xie et al. 2021). Models optimized for traditional validity metrics excel at reproducing known patterns but struggle with exploration. Second, the trend toward architectural complexity may be counterproductive in materials science, where datasets are small and biased toward known stable structures. Unlike vision or language tasks with massive datasets, complex models risk overfitting rather than enabling generalization.

We present CrystalDiT, exploring whether simplified architectures can outperform complex alternatives in crystal generation. To rigorously test this hypothesis, we develop both a simple unified architecture and a complex dual-stream variant for direct comparison. Our key contributions are:

Simplified Architecture: A unified diffusion transformer processes all crystal information through joint attention, contrasting with our dual-stream alternative that uses separate processing pathways with cross-attention mechanisms.

Chemical Representation: A two-dimensional atomic encoding using periodic table positions (period, group) naturally captures chemical relationships without architectural complexity.

Balanced Evaluation: A composite score explicitly optimizes the trade-off between generation quality and discovery potential, addressing the evaluation-objective mismatch.

Our findings suggest that in data-limited scientific domains, carefully designed simple architectures with domain-specific representations outperform complex alternatives, challenging assumptions about the necessity of architectural sophistication.

Related Work

Crystal generation has become a central focus in the field of materials informatics. To address the limitations of traditional computational materials discovery methods, such as their effectiveness and high computational cost (Pickard and Needs 2011; Curtarolo et al. 2012; Yamashita et al. 2018;

*Corresponding author.
†Corresponding author.

Wang, Botti, and Marques 2021), recent approaches have increasingly leveraged deep learning-based generative models.

CDVAE (Xie et al. 2021) was among the first to predict three key components (the number and type of atoms and the lattice structure) as an initial approximation of a material's structure, which is subsequently refined using a diffusion-based approach (Song and Ermon 2019; Ho, Jain, and Abbeel 2020). DiffCSP (Jiao et al. 2023) is the first to employ a jointly equivariant diffusion paradigm (i.e., jointly diffusing the lattices and fractional coordinates) for crystal structure prediction (CSP). This method can be further extended to handle ab initio crystal generation by incorporating an additional discrete diffusion (Austin et al. 2021) on atom types, with MatterGen (Zeni et al. 2025) further refining this joint diffusion paradigm.

Diffusion Transformers (DiT) (Peebles and Xie 2023) have shown remarkable capabilities in learning stable crystal structures through their powerful attention mechanisms and adaptive conditioning. However, this expressive power also introduces significant risks of overfitting in data-limited scientific domains, where models may memorize training patterns rather than learn generalizable principles for materials discovery. ADiT (Joshi et al. 2025) introduces a two-stage approach that first generates latent representations of crystal structures using an autoencoder, followed by applying the DiT architecture in the latent space for structure generation.

Flow Matching techniques (Lipman et al. 2022, 2024) have recently emerged as a highly effective alternative to diffusion models, primarily owing to their enhanced inference efficiency and the flexibility in defining prior distributions. In this context, FlowMM (Miller et al. 2024) leverages joint Riemannian Flow Matching (Chen and Lipman 2023) within Riemannian manifolds, providing improved handling of crystal periodicities. Subsequently, FlowLLM (Sriram et al. 2024) extends this framework to large language models (LLMs), utilizing them as prior distributions for sampling the chemical formulas of meta-stable materials, with their corresponding structural configurations subsequently generated through Riemannian Flow Matching.

In another line of research, material symmetries have garnered significant attention in the context of crystal structure generation. Recently, DiffCSP++ (Jiao et al. 2024), SymmCD (Levy et al. 2025), and WyFormer (Kazeev et al. 2025) have made notable advances by incorporating Wyckoff positions (Wyckoff 1922) into the crystal generation task, enabling the generation of crystal structures with defined symmetries. Despite these technical advances, existing approaches exhibit increasing architectural complexity with specialized multi-component designs. This motivates our investigation of whether unified, simplified architectures can achieve superior performance through more effective learning of lattice-atom relationships.

Method

We propose CrystalDiT, a simple yet effective diffusion transformer architecture for crystal structure generation. Our approach consists of four key components: (1) a novel two-dimensional atomic representation that captures chemical relationships through periodic table positioning, (2) a streamlined diffusion transformer architecture that processes crystal structures through unified attention mechanisms, (3) a balanced model selection strategy that replaces traditional validation, and (4) a probabilistic atomic decoding procedure for inference. This section details each component of our methodology.

Crystal Representation

Effective crystal structure representation is crucial for training diffusion models that can generate stable and novel materials. Traditional approaches either use graph-based representations that scale quadratically with the number of atoms, or rely on atomic number encodings that ignore chemical relationships. While recent work like UniMat (Yang et al. 2023) proposed periodic table-based representations, their 4D tensor approach requires pre-allocating space for every possible element in the periodic table, leading to sparse representations where most positions remain unoccupied. Moreover, their method necessitates defining maximum atom counts per element type and specialized tensor operations across chemical dimensions. In contrast, our approach provides a more compact representation that only encodes 20 atoms and enables unified processing through standard transformer operations.

We introduce a simplified yet more chemically meaningful crystal representation that builds upon periodic table structure while incorporating domain-specific insights from materials science.

Two-Dimensional Atomic Representation Instead of representing atoms by their atomic numbers, we encode each atom using its position in the periodic table: the period (row) and group (column). This encoding is motivated by the fundamental principle that elements in the same period share similar electron shell configurations, while elements in the same group exhibit similar chemical properties.

Specifically, for an atom with atomic number Z, we map it to a tuple (r, c) where $r \in [0, 7]$ represents the period and $c \in [0, 18]$ represents the group. Here, $r = 0$ and $c = 0$ correspond to a special "null atom" representing empty positions in crystals with fewer than 20 atoms. For valid elements, $r \in [1, 7]$ and $c \in [1, 18]$, with lanthanides and actinides using fractional group numbers to preserve their unique positions. We then normalize these values to $[-1, 1]$:

$$r_{\text{norm}} = \frac{2r}{7} - 1 \tag{1}$$

$$c_{\text{norm}} = \frac{2c}{18} - 1 \tag{2}$$

This representation offers several advantages: (1) it naturally captures chemical similarity through spatial proximity in the periodic table, (2) it provides a continuous embedding space that facilitates diffusion modeling, (3) it reduces dimensionality while preserving chemical meaning, and (4) it seamlessly handles variable-sized crystal structures through the null atom representation at the origin $(0, 0)$.

Normalized Lattice Parameterization For lattice vectors, we adopt a normalization strategy that addresses the wide range of lattice parameter values in real materials. Given the lattice matrix $\boldsymbol{L} \in \mathbb{R}^{3\times3}$, we normalize by the maximum length scale observed in the MP-20 dataset:

$$\boldsymbol{L}_{\text{norm}} = \frac{\boldsymbol{L}}{L_{\max}} \tag{3}$$

where $L_{\max} = 46.7425$ Å from our analysis of the MP-20 dataset (Xie et al. 2021), which contains 45,231 metastable crystal structures with up to 20 atoms spanning 89 element types.

Complete Structure Representation Our complete crystal representation combines the normalized lattice vectors and atomic features. For a crystal with N atoms ($N \leq 20$ for MP-20):

$$\boldsymbol{L}_{\text{norm}} \in \mathbb{R}^{3\times3} \quad \text{(normalized lattice)} \tag{4}$$

$$\boldsymbol{A} \in \mathbb{R}^{20\times5} \quad \text{(atomic features)} \tag{5}$$

where each row of $\boldsymbol{A}$ contains $[r_{\text{norm}}, c_{\text{norm}}, x, y, z]$ representing the normalized period, normalized group, and fractional coordinates. For crystals with fewer than 20 atoms, we pad with "null" atoms using $[-1, -1, -1, -1, -1]$.

Architecture Design

Our CrystalDiT architecture is designed around the principle that simplicity leads to better generalization in crystal generation tasks. Unlike complex multi-stream architectures, we employ a unified approach that processes all crystal information through a single, streamlined transformer pathway.

Unified DiT Architecture The model consists of three main components: (1) Crystal structure embedding that maps lattice vectors and atomic features into a shared hidden space, (2) A sequence of 18 DiT blocks that process the combined representation through unified self-attention, and (3) Specialized output heads that generate noise predictions for both atomic and lattice components.

The key insight is that by processing atomic and lattice features together in a single attention pathway, the model can naturally learn the complex interdependencies between atomic positions and lattice parameters without requiring explicit cross-attention mechanisms. This unified approach enforces a strong inductive bias: treating lattice and atomic properties as a single, interdependent system, which aligns with the physical reality that crystal properties emerge from the interplay between atomic composition and lattice geometry.

Crystal structures are embedded into a hidden space of dimension $d = 512$. The lattice vectors and atomic features are processed through separate linear embedding layers with positional and type encodings. The embedded features are concatenated to form a combined representation $\boldsymbol{H}_{\text{combined}} \in \mathbb{R}^{23\times d}$ (20 atoms + 3 lattice vectors).

This combined representation is processed through $L = 18$ identical DiT blocks, each incorporating time-conditional adaptive layer normalization (AdaLN) to modulate features based on the diffusion time step. Finally, specialized output heads generate noise predictions for atomic features (5D) and lattice vectors (3D per vector).

Architecture Comparison In contrast to our unified approach shown in Figure 1, we also implement a complex dual-stream variant for direct comparison. This architecture employs cascaded processing: 12 atom-only DiT blocks process atomic features independently, followed by 2 lattice-only blocks for lattice vectors, and finally 2 joint blocks with bidirectional cross-attention mechanisms to fuse information between streams. Unlike our simple unified approach that processes all features together from the start, this cascaded design separates atomic and lattice processing pathways with specialized modules. However, this architectural complexity leads to overfitting in the data-limited crystal generation domain, as evidenced by lower unique and novel generation rates despite achieving higher individual validity metrics.

Training Objective and Model Selection

Our training approach combines standard diffusion objectives with a novel model selection strategy that addresses the unique challenges of crystal generation.

Diffusion Loss Function We employ a Gaussian diffusion process with $T = 1000$ time steps and a linear noise schedule. Our model learns to predict the noise $\boldsymbol{\epsilon}$ added at each time step t using a weighted loss function:

$$\mathcal{L} = \mathcal{L}_{\text{lattice}} + \lambda \cdot \mathcal{L}_{\text{atoms}} \tag{6}$$

where:

$$\mathcal{L}_{\text{lattice}} = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon}_{\text{lattice}} - \boldsymbol{\epsilon}_\theta^{\text{lattice}}(\boldsymbol{L}_t, \boldsymbol{A}_t, t)\|^2\right] \tag{7}$$

$$\mathcal{L}_{\text{atoms}} = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\boldsymbol{w}^T \odot \|\boldsymbol{\epsilon}_{\text{atoms}} - \boldsymbol{\epsilon}_\theta^{\text{atoms}}(\boldsymbol{L}_t, \boldsymbol{A}_t, t)\|^2\right] \tag{8}$$

We set $\lambda = 100$ to balance different scales, and use feature-specific weights $\boldsymbol{w} = [1.5, 2.0, 1.0, 1.0, 1.0]$ to emphasize period and group predictions.

Balanced Model Selection Strategy Traditional checkpoint selection methods face fundamental limitations in crystal generation. Validation loss, while standard in machine learning, cannot capture the true stability and discovery potential of generated crystals, as it only measures reconstruction fidelity rather than physical plausibility. Alternative approaches employed by recent methods like ADiT (Joshi et al. 2025) generate 1000 structures at each checkpoint and select models with the highest validity rates. However, this strategy inadvertently promotes overfitting to known stable patterns in the training data, leading to high validity scores but poor novelty rates - exactly opposite to what materials discovery requires.

Our **Balance Score** provides a principled alternative that explicitly optimizes the trade-off between generation quality and discovery potential during checkpoint selection:

$$\text{Balance Score} = \text{UN Rate} \times (\text{Quality Composite})^\alpha \tag{9}$$

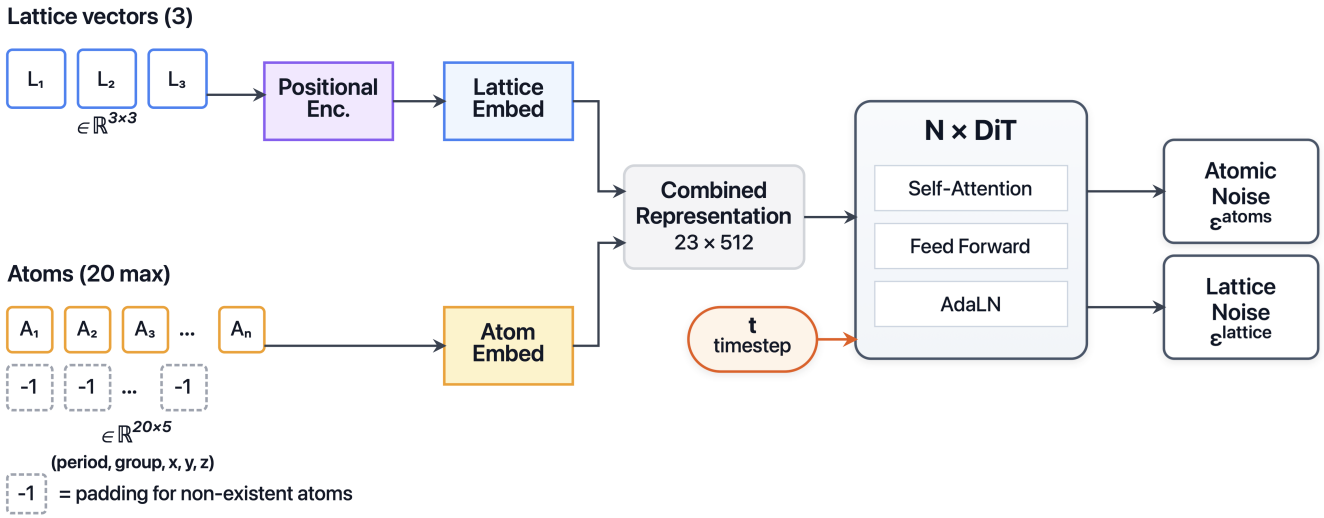

Figure 1: CrystalDiT unified architecture. Input crystal structures are embedded into a combined 23-token sequence (3 lattice vectors + 20 atoms), processed through N DiT blocks with unified self-attention, and decoded to atomic and lattice noise predictions. The architecture treats all crystal components as a single interdependent system.

where α is a hyperparameter that controls the trade-off between generation quality and discovery potential. Higher α values emphasize quality for more reliable structures, while lower α values favor discovery of unique and novel crystals.

UN Rate measures the proportion of generated structures that are both unique and novel:

$$\text{UN Rate} = \frac{N_{\text{unique}} \cap N_{\text{novel}}}{N_{\text{total}}} \tag{10}$$

Quality Composite is the geometric mean of four normalized quality metrics:

$$\text{Quality Composite} = (S_{\text{struct}} \times S_{\text{chem}} \times S_{\text{density}} \times S_{\text{elements}})^{1/4} \tag{11}$$

Each component score is normalized to $[0, 1]$ based on empirically observed ranges:

$$S_{\text{struct}} = \max\left(0, \min\left(1, \frac{V_{\text{struct}} - 0.95}{0.05}\right)\right) \tag{12}$$

$$S_{\text{chem}} = \max\left(0, \min\left(1, \frac{V_{\text{chem}} - 0.8}{0.2}\right)\right) \tag{13}$$

$$S_{\text{density}} = \max\left(0, \min\left(1, \frac{1.0 - D_{\text{density}}}{0.9}\right)\right) \tag{14}$$

$$S_{\text{elements}} = \max\left(0, \min\left(1, \frac{1.0 - D_{\text{elements}}}{0.9}\right)\right) \tag{15}$$

We implement a multi-phase checkpoint selection strategy during training. We identify the models with the highest Balance Score from three distinct training phases: early (0-30%), middle (31-60%), and late (61-100%) stages. This approach recognizes that optimal trade-offs between quality and discovery potential may emerge at different training stages, with early models potentially favoring exploration and later models emphasizing quality. The detailed training protocol is provided in Appendix C.

Inference Procedure

During inference, our model generates crystal structures through the standard DDPM sampling process, followed by a probabilistic atomic decoding procedure to convert continuous predictions to discrete atomic types.

Probabilistic Atomic Decoding Our model predicts continuous values for atomic periods and groups, which must be mapped to discrete atomic numbers. We define the candidate atomic types as all valid elements plus the null atom (z=0) at position (0,0). Following the discrete decoder approach from DDPM (Ho, Jain, and Abbeel 2020), we employ a probabilistic mapping based on Gaussian distributions.

For predicted continuous values $(r_{\text{pred}}, c_{\text{pred}})$ and element z at position (r_z, c_z), the mapping probability is:

Probabilistic Atomic Decoding Our model predicts continuous values for atomic periods and groups, which must be mapped to discrete atomic numbers. We assign each candidate element a responsibility region and compute the probability using Gaussian integration.

For predicted continuous values $(r_{\text{pred}}, c_{\text{pred}})$ and candidate element z at normalized position (r_z, c_z), the mapping probability is:

$$P(z|r_{\text{pred}}, c_{\text{pred}}) = P_r(r_{\text{pred}}|r_z) \times P_c(c_{\text{pred}}|c_z) \tag{16}$$

$$P_r(r_{\text{pred}}|r_z) = \int_{r_{\text{lower}}}^{r_{\text{upper}}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - r_{\text{pred}})^2}{2\sigma^2}\right) dx \tag{17}$$

$$P_c(c_{\text{pred}}|c_z) = \int_{c_{\text{lower}}}^{c_{\text{upper}}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - c_{\text{pred}})^2}{2\sigma^2}\right) dx \tag{18}$$

where $\sigma = 0.1$ controls the mapping sharpness. The integration bounds define each element's responsibility region:

Algorithm 1 CrystalDiT Generation with Probabilistic Atomic Decoding

Input: Timesteps T, batch size B, model θ
Output: Crystal structures $\{\text{Structure}_i\}_{i=1}^{B}$

Sample initial noise: $\boldsymbol{Z}^{(0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, T-1, \ldots, 1$ **do**
  $\boldsymbol{\epsilon} = \text{CrystalDiT}_\theta(\boldsymbol{Z}^{(t)}, t)$
  $\boldsymbol{Z}^{(t-1)} = \text{DDPM\_step}(\boldsymbol{Z}^{(t)}, \boldsymbol{\epsilon}, t)$
end for
Extract lattice: $\boldsymbol{L} = \boldsymbol{Z}^{(0)}_{[:,:3,:]} \times L_{\max}$
Extract atoms: $\boldsymbol{A} = \boldsymbol{Z}^{(0)}_{[:,3:,:]}$
for each sample $i = 1, \ldots, B$ **do**
  Initialize: $atomic\_numbers_i = [\,]$, $coords_i = [\,]$
  for each atom $j = 1, \ldots, 20$ **do**
    $(r, c, x, y, z) = \boldsymbol{A}_{i,j,:5}$
    Compute $z^* = \arg\max_z P(z|r, c)$ using Eqs. (16)-(20)
    if $z^* > 0$ **then**
      $atomic\_numbers_i$.append(z^*)
      $coords_i$.append($[x, y, z] \bmod 1$)
    end if{Skip if $z^* = 0$ (null atom)}
  end for
  Construct Structure$_i$ from $atomic\_numbers_i$ and $coords_i$
end for

$$r_{\text{upper}} = \begin{cases} r_z + \Delta_r & \text{if } r_z < 1 \\ +\infty & \text{if } r_z = 1 \end{cases} \tag{19}$$

$$r_{\text{lower}} = \begin{cases} -\infty & \text{if } r_z = -1 \\ r_z - \Delta_r & \text{if } r_z > -1 \end{cases} \tag{20}$$

with similar bounds for groups, where $\Delta_r = 1/7$ and $\Delta_c = 1/18$ are the discretization intervals. The final atomic number is selected as:

$$z^* = \arg\max_z P(z|r_{\text{pred}}, c_{\text{pred}})$$

For the null atom at $(-1, -1)$, its region extends to $(-\infty, -1+\Delta_r] \times (-\infty, -1+\Delta_c]$, allowing natural handling of empty positions.

Experiments and Results

We conduct comprehensive experiments to evaluate our CrystalDiT approach against state-of-the-art crystal generation methods. Our evaluation uses established metrics for cross-method comparison, while our Balance Score is specifically employed for checkpoint selection during our model training to optimize the trade-off between generation quality and discovery potential.

Experimental Setup

Unified Evaluation Protocol: To ensure fair comparison, we re-evaluate all baseline methods using a unified testing protocol with identical evaluation metrics, DFT calculation parameters, and statistical sampling procedures. This approach eliminates potential discrepancies arising from different evaluation implementations across original papers.

Dataset and Preprocessing We use the MP-20 dataset (Xie et al. 2021), which contains 45,231 metastable crystal structures from the Materials Project with up to 20 atoms spanning 89 element types. Following standard practice, we use the established train/test split and preprocess structures using our two-dimensional atomic representation and lattice normalization.

Model Configurations We evaluate three model variants:

CrystalDiT (Simple): Our main model with unified attention processing. Architecture: $d = 512$, $L = 18$ layers, 8 attention heads. Model size: 330MB.

CrystalDiT (Complex): A dual-stream architecture for comparison, featuring separate atom and lattice processing streams with cross-attention mechanisms. This architecture uses cascaded processing: 12 atom-only DiT blocks $\rightarrow$ 2 lattice-only DiT blocks $\rightarrow$ 2 joint DiT blocks with bidirectional cross-attention. Model size: 370MB (parameter count controlled to be similar to the simple version).

Both models are trained for 50,000 epochs with batch size 256, learning rate 1×10^{-4} using 8 V100 GPUs over 4 days.

Baseline Methods We compare against five state-of-the-art methods: DiffCSP (Jiao et al. 2023), FlowMM (Miller et al. 2024), DiffCSP++ (Jiao et al. 2024), and MatterGen (Zeni et al. 2025) using pretrained checkpoints or official implementations; ADiT (Joshi et al. 2025) using 10,000 pregenerated structures from their official repository.

Evaluation Protocol Following the evaluation framework established by FlowMM (Miller et al. 2024), we generate 10,000 structures from each method and compute comprehensive metrics:

Validity Metrics:

- *Structural validity*: Percentage of crystals with all pairwise atomic distances ≥ 0.5 Å and crystal volume ≥ 0.1 Å^3
- *Compositional validity*: Percentage satisfying charge neutrality and electronegativity balance via SMACT (Davies et al. 2019), using oxidation state enumeration and Pauling electronegativity rules

Distribution Metrics:

- *Density distance* (d_ρ): Wasserstein distance between generated and test set density distributions
- *Elements distance* (d_{elem}): Wasserstein distance between generated and test set element occurrence frequency distributions

Discovery Metrics:

- *Uniqueness*: Structures deemed distinct by PyMatGen's StructureMatcher (Ong et al. 2013)
- *Novelty*: Structures not matching any MP-20 training set crystal via StructureMatcher
- *UN Rate*: Fraction of structures that are simultaneously unique and novel

Stability Assessment: We adapt the protocol established by FlowMM (Miller et al. 2024) for DFT evaluation. Due to computational resource limitations, we randomly sample 500 UN structures for stability assessment (compared

to FlowMM's evaluation of all structures). To quantify sampling uncertainty, we repeat this sampling three times independently for CrystalDiT, FlowMM, and MatterGen, reporting mean±std across samples. Each sampled structure undergoes:

1. Pre-relaxation using CHGNet (Deng et al. 2023) ML potential
2. DFT relaxation using VASP with MPRelaxSet parameters (Jain et al. 2013)
3. Energy above hull calculation against Matbench Discovery convex hull (Riebesell et al. 2023)
4. Classification: Stable ($E_{\text{hull}} < 0.0$ eV/atom), Metastable ($E_{\text{hull}} < 0.1$ eV/atom)

Final Discovery Metrics:

- *SUN Rate*: UN Rate × Stable Rate among UN structures
- *MSUN Rate*: UN Rate × Metastable Rate among UN structures

Detailed evaluation parameters and implementation specifics are provided in Appendix A.

Main Results

Table 1 presents our comprehensive comparison against state-of-the-art methods. The results reveal several key insights about the effectiveness of different approaches for crystal generation. **Note on Baseline Results:** Our re-evaluation using unified protocols yields results that differ from some originally reported values in the literature. Detailed analysis of these differences is provided in Appendix A.

Our simple CrystalDiT achieves the highest SUN rate (8.78%) and MSUN rate (25.94%), substantially outperforming recent methods including FlowMM (4.21%) and MatterGen (3.66%). Notably, CrystalDiT generates 63.28% unique and novel structures while maintaining comparable stability rates, demonstrating that architectural simplicity can be more effective than complexity for materials discovery. Different methods exhibit distinct trade-offs: ADiT achieves high validity scores but suffers from poor novelty (37.08% UN rate) due to overfitting, while FlowMM and MatterGen generate novel structures with lower stability rates.

Architecture Comparison Analysis

The comparison between our simple and complex architectures reveals fundamental insights about generative modeling for scientific applications. Despite sophisticated cross-attention mechanisms, the complex dual-stream architecture underperforms the simple version across discovery metrics (6.36% vs 8.78% SUN rate). The implementation and analysis of the dual-stream architecture are provided in Appendix B.

This result challenges prevailing assumptions about architectural sophistication in machine learning. The complex model achieves better individual quality metrics but significantly lower UN rates, indicating that architectural complexity promotes overfitting in crystal generation tasks. The simple unified attention mechanism appears more effective at learning generalizable patterns rather than memorizing training data distributions.

Similarly, ADiT's two-stage approach (autoencoder + latent DiT) achieves excellent validity scores but poor novelty (37.08% UN rate), demonstrating that architectural sophistication without careful consideration of the discovery objective can be counterproductive in materials science applications.

Training Dynamics and Model Selection

Table 2 shows the evolution of our simple CrystalDiT model across different training epochs, demonstrating the importance of our balance score for checkpoint selection during training. Note that our Balance Score is used exclusively for selecting the best checkpoint during training of our CrystalDiT models, not for comparing different methods. All cross-method comparisons use the standard SUN/MSUN discovery metrics.

The simple model shows a clear pattern: as training progresses, validity metrics improve, but the UN rate steadily decreases from 80.72% to 57.37%. Traditional validation methods focusing on validity would select the final checkpoint, but our balance score correctly identifies earlier checkpoints with better discovery potential. This demonstrates the critical importance of balanced evaluation for materials discovery applications.

Component Analysis

We evaluate the contribution of our two-dimensional atomic representation by comparing against traditional one-dimensional atomic number encoding on the simple CrystalDiT architecture. Detailed implementation of the 1D atomic representation is provided in Appendix B.

As shown in Table 1, our two-dimensional periodic table-based representation (CrystalDiT Simple: 8.78% SUN rate) significantly outperforms the one-dimensional encoding (CrystalDiT 1D atomic: 6.28% SUN rate). While the one-dimensional approach achieves higher UN rate (78.47% vs 63.28%), the generated structures exhibit lower stability rates (8.00% vs 13.87% stable rate), resulting in inferior final discovery performance. This demonstrates that chemical knowledge embedded in the periodic table structure enhances the quality of generated crystals for materials discovery.

We additionally conduct ablation studies on architecture depth, finding that 18 layers provide the optimal balance between model capacity and generalization. Detailed results are provided in Appendix C.

Energy Distribution Analysis

Figure 2 presents energy distribution comparison between CrystalDiT and FlowMM as a representative baseline method. Our analysis reveals that CrystalDiT demonstrates superior ability to generate thermodynamically favorable structures compared to FlowMM. CrystalDiT shows a pronounced peak in the stable region ($E^{\text{hull}} < 0$), indicating that our simplified architecture effectively learns to generate

Method	Struct. Valid (%)	Chem. Valid (%)	$d_\rho \downarrow$	$d_{\text{elem}} \downarrow$	UN Rate (%)	Stable in UN (%)	Metastable in UN (%)	SUN (%)	MSUN (%)
MP-20(train)*	100.0	90.55	0.214	0.049	-	44.07*	100.0*	-	-
DiffCSP	99.90	82.52	0.347	0.369	87.17	4.00	23.80	3.49	20.75
FlowMM	99.22	82.09	0.185	0.128	87.66	$4.80_{\pm 0.20}$	$23.69_{\pm 0.10}$	$4.21_{\pm 0.18}$	$20.77_{\pm 0.09}$
DiffCSP++	99.96	84.74	**0.135**	0.453	87.62	3.80	21.80	3.33	19.10
MatterGen	**99.99**	83.62	0.393	0.207	**89.89**	$4.07_{\pm 0.32}$	$26.90_{\pm 0.71}$	$3.66_{\pm 0.28}$	$24.18_{\pm 0.63}$
ADiT	99.58	**90.83**	0.179	**0.082**	37.08	7.40	36.40	2.74	13.50
CrystalDiT (Complex)	98.39	89.44	0.271	0.115	40.28	**15.80**	**55.20**	6.36	22.24
CrystalDiT (Simple)	97.79	87.02	0.459	0.211	63.28	$13.87_{\pm 1.21}$	$40.93_{\pm 1.51}$	$\mathbf{8.78_{\pm 0.74}}$	$\mathbf{25.90_{\pm 0.95}}$
CrystalDiT (1D atomic)	97.34	86.82	0.499	0.276	78.47	8.00	31.00	6.28	24.33

Table 1: Comprehensive comparison on MP-20. Best results in bold. For methods with bootstrap sampling (indicated by $\pm$), we report mean$\pm$std over 3 independent samples of 500 UN structures each for DFT evaluation. *Rates for MP-20(train) represent proportions across all training structures.

Epoch	Struct. Valid (%)	Chem. Valid (%)	d_ρ	d_{elem}	UN Rate (%)
10k	96.77	83.09	0.551	0.178	80.72
20k	97.51	87.07	0.170	0.249	73.31
30k	97.99	86.74	0.308	0.231	63.30
40k	98.44	89.33	0.592	0.208	62.32
50k	98.28	89.40	0.166	0.266	57.37

Table 2: Training progression of CrystalDiT (Simple) showing deteriorating UN rate despite improving validity metrics.

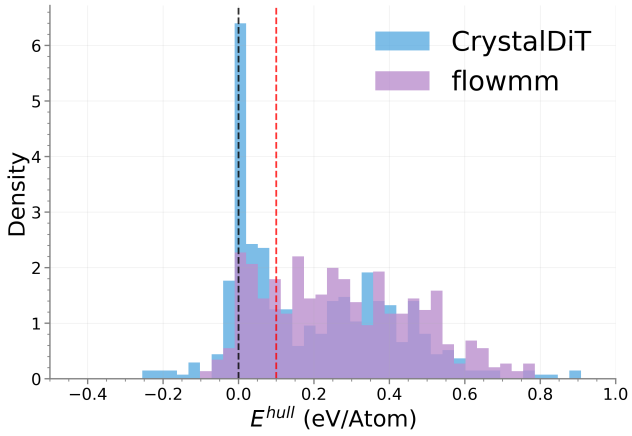

Figure 2: Energy distribution comparison. Black and red dashed lines mark stability ($E^{\text{hull}} = 0$) and metastability ($E^{\text{hull}} = 0.1$ eV/atom) thresholds. CrystalDiT generates more stable and metastable structures. Full comparisons in Appendix C.

energetically favorable crystal structures. The energy distribution provides crucial validation that CrystalDiT not only generates more unique and novel structures but also ensures these structures are more likely to be thermodynamically viable for practical materials applications.

Similar patterns are observed when comparing against other baseline methods (ADiT, MatterGen, DiffCSP++), with CrystalDiT consistently producing more stable and metastable structures. These comprehensive energy distribution analyses are provided in Appendix C, reinforcing our main findings that simple, well-designed architectures with appropriate domain knowledge can outperform complex alternatives in crystal generation tasks.

Scaling to Larger Structures

CrystalDiT achieves 6.73% SUN rate on MPTS-52 (up to 52 atoms), demonstrating effective generalization with only 2% degradation compared to MP-20. Detailed results are in Appendix C.

Conclusion

We presented CrystalDiT, demonstrating that simplified architectures can substantially outperform complex alternatives for crystal generation. Our unified diffusion transformer achieves an 8.78% SUN rate on MP-20, outperforming existing methods through three key contributions: (1) a simplified architecture using unified attention mechanisms; (2) a two-dimensional atomic representation using periodic table positions; and (3) a balanced model selection strategy optimizing generation quality and discovery potential.

Beyond generating isolated structures, important directions for future research include extending our framework to material-molecule interaction systems (Bian, Wu, and Yan 2026; Bian et al. 2026), which could enable discovery of functional materials for catalysis, drug delivery, and energy applications. Additionally, incorporating target property constraints into the generation process would allow direct design of materials with desired characteristics, further advancing AI-driven materials discovery.

Acknowledgments

This work was supported by the Natural Science Foundation of Guangdong Province (grant no. 2025A1515011946) and the National University of Singapore School of Computing (grant no. A-0010308-00-00 for YB). Part of this work was conducted when authors Xiaohan Yi and Guikun Xu were at

Tencent AI Lab. We acknowledge computational resources from Tencent, thank Tao Chen for insightful discussions, and the anonymous reviewers for their constructive feedback.

References

Austin, J.; Johnson, D. D.; Ho, J.; Tarlow, D.; and Van Den Berg, R. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34: 17981–17993.

Bian, Y.; Wu, H.; and Yan, J. 2026. Deep learning for affinity prediction and interface prediction in molecular interactions. In *Deep Learning in Drug Design*, 283–296. Elsevier.

Bian, Y.; Yang, N.; Wu, J.; and Yan, J. 2026. Deep learning for complex structure prediction in molecular interactions. In *Deep Learning in Drug Design*, 297–308. Elsevier.

Chen, R. T.; and Lipman, Y. 2023. Riemannian flow matching on general geometries. *arXiv preprint arXiv:2303.03660*.

Chen, Z.; Feng, S.; Xiao, X.; Zhang, Z.; Li, Q.; Gao, X.; and Zhao, P. 2025. MSDformer: Multi-scale Discrete Transformer For Time Series Generation. *arXiv preprint arXiv:2505.14202*.

Chen, Z.; Feng, S.; Zhang, Z.; Xiao, X.; Gao, X.; and Zhao, P. 2024. Sdformer: Similarity-driven discrete transformer for time series generation. In *Advances in Neural Information Processing Systems*, volume 37, 132179–132207.

Curtarolo, S.; Setyawan, W.; Hart, G. L.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; et al. 2012. AFLOW: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58: 218–226.

Davies, D. W.; Butler, K. T.; Jackson, A. J.; Morris, A.; Frost, J. M.; Skelton, J. M.; and Walsh, A. 2019. SMACT: Semiconducting Materials by Analogy and Chemical Theory. *Journal of Open Source Software*, 4(38): 1361.

Deng, B.; Zhong, P.; Jun, K.; Riebesell, J.; Han, K.; Bartel, C. J.; and Ceder, G. 2023. CHGNet: A pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5: 1031–1041.

Fang, Y.; Liu, R.; Huang, H.; Zhao, P.; and Wu, Q. 2024. A spatio-temporal diffusion model for missing and real-time financial data inference. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 602–611.

Feng, S.; Miao, C.; Zhang, Z.; and Zhao, P. 2024. Latent diffusion transformer for probabilistic time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 11979–11987.

Gong, S.; Agarwal, S.; Zhang, Y.; Ye, J.; Zheng, L.; Li, M.; An, C.; Zhao, P.; Bi, W.; Han, J.; et al. 2024. Scaling diffusion language models via adaptation from autoregressive models. *arXiv preprint arXiv:2410.17891*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. 2013. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1).

Jiao, R.; Huang, W.; Han, J.; and Liu, Y. 2024. Space Group Constrained Crystal Generation. In *International Conference on Learning Representations*.

Jiao, R.; Huang, W.; Lin, P.; Han, J.; Chen, P.; Lu, Y.; and Liu, Y. 2023. Crystal Structure Prediction by Joint Equivariant Diffusion. In *Advances in Neural Information Processing Systems*, volume 36.

Joshi, C. K.; Fu, X.; Liao, Y.-L.; Gharakhanyan, V.; Miller, B. K.; Sriram, A.; and Ulissi, Z. W. 2025. All-atom diffusion transformers: Unified generative modelling of molecules and materials. *arXiv preprint arXiv:2503.03965*.

Kazeev, N.; Nong, W.; Romanov, I.; Zhu, R.; Ustyuzhanin, A. E.; Yamazaki, S.; and Hippalgaonkar, K. 2025. Wyckoff Transformer: Generation of Symmetric Crystals. In *Forty-second International Conference on Machine Learning*.

Levy, D.; Panigrahi, S. S.; Kaba, S.-O.; Zhu, Q.; Lee, K. L. K.; Galkin, M.; Miret, S.; and Ravanbakhsh, S. 2025. SymmCD: Symmetry-Preserving crystal generation with diffusion models. *arXiv preprint arXiv:2502.03638*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Lipman, Y.; Havasi, M.; Holderrieth, P.; Shaul, N.; Le, M.; Karrer, B.; Chen, R. T.; Lopez-Paz, D.; Ben-Hamu, H.; and Gat, I. 2024. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*.

Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; and Cubuk, E. D. 2023. Scaling deep learning for materials discovery. *Nature*, 624(7990): 80–85.

Miller, B. K.; Chen, R. T.; Sriram, A.; and Wood, B. M. 2024. Flowmm: Generating materials with riemannian flow matching. *arXiv preprint arXiv:2406.04713*.

Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; and Ceder, G. 2013. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68: 314–319.

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

Pickard, C. J.; and Needs, R. 2011. Ab initio random structure searching. *Journal of Physics: Condensed Matter*, 23(5): 053201.

Riebesell, J.; Goodall, R. E. A.; Baird, S. G.; Sparks, T. D.; and Jain, A. 2023. Matbench Discovery: An evaluation framework for machine learning crystal stability prediction. *arXiv preprint arXiv:2308.14920*.

Shibo, F.; Zhao, P.; Liu, L.; Wu, P.; and Shen, Z. 2025. Hdt: Hierarchical discrete transformer for multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 746–754.

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

Sriram, A.; Miller, B.; Chen, R. T.; and Wood, B. 2024. Flowllm: Flow matching for material generation with large language models as base distributions. In *Advances in Neural Information Processing Systems*, volume 37, 46025–46046.

Wang, H.-C.; Botti, S.; and Marques, M. A. 2021. Predicting stable crystalline compounds using chemical similarity. *npj Computational Materials*, 7(1): 12.

Wang, Q.; Meng, Z.; Cui, W.; Zhang, Y.; Wu, P.; Wu, B.; King, I.; Chen, L.; and Zhao, P. 2025. NTPP: Generative Speech Language Modeling for Dual-Channel Spoken Dialogue via Next-Token-Pair Prediction. *arXiv preprint arXiv:2506.00975*.

Wu, S.; Xiao, X.; Ding, Q.; Zhao, P.; Wei, Y.; and Huang, J. 2020. Adversarial sparse transformer for time series forecasting. In *Advances in Neural Information Processing Systems*, volume 33, 17105–17115.

Wyckoff, R. W. G. 1922. *The Analytical Expression of the Results of the Theory of Space-groups*. 318. Carnegie institution of Washington.

Xie, T.; Fu, X.; Ganea, O.-E.; Barzilay, R.; and Jaakkola, T. 2021. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*.

Xu, K.; Zhang, Y.; Ye, D.; Zhao, P.; and Tan, M. 2021. Relation-aware transformer for portfolio policy learning. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, 4647–4653.

Yamashita, T.; Sato, N.; Kino, H.; Miyake, T.; Tsuda, K.; and Oguchi, T. 2018. Crystal structure prediction accelerated by Bayesian optimization. *Physical Review Materials*, 2(1): 013803.

Yang, S.; Cho, K.; Merchant, A.; Abbeel, P.; Schuurmans, D.; Mordatch, I.; and Cubuk, E. D. 2023. Scalable diffusion for materials generation. *arXiv preprint arXiv:2311.09235*.

Yi, Y.; Wan, X.; Bian, Y.; Ou-Yang, L.; and Zhao, P. 2023. Etdock: A novel equivariant transformer for protein-ligand docking. *arXiv preprint arXiv:2310.08061*.

Zeni, C.; Pinsler, R.; Zügner, D.; Fowler, A.; Horton, M.; Fu, X.; Wang, Z.; Shysheya, A.; Crabbé, J.; Ueda, S.; et al. 2025. A generative model for inorganic materials design. *Nature*, 639(8055): 624–632.