MGB: The Material Generation Benchmark

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present MGB (Material Generation Benchmark), a comprehensive and standardized platform for evaluating deep generative models in materials science. MGB covers a diverse range of tasks—including crystal structure prediction, de novo material generation, MOF structure prediction, and out-of-distribution (OOD) generation—spanning datasets from inorganic crystals to complex MOFs. It integrates cutting-edge methodologies, from large language models (LLMs) to diffusion-based and hybrid approaches. A key feature of MGB is the construction of dedicated OOD test sets, enabling rigorous evaluation of generalization capabilities.

To ensure fair comparison, MGB employs multi-dimensional metrics that jointly assess structural accuracy, chemical validity, distributional coverage, physical plausibility, and computational efficiency. Extensive experiments highlight clear performance patterns: diffusion models excel in predicting complex crystalline systems, LLMs achieve competitive local accuracy, and MOF-specific flow models substantially outperform general-purpose approaches on MOF prediction. While most methods yield nearly perfect structural validity in de novo generation, their ability to balance accuracy, generalization, and efficiency varies considerably.

Importantly, we select LLMs for OOD case studies given their relatively state-of-the-art performance on in-distribution benchmarks. However, our results reveal a critical limitation: despite strong in-distribution accuracy, LLMs completely fail to generalize to unseen structural families. By establishing a unified framework and offering transparent comparative insights, MGB aims to drive the development of more robust and efficient generative models for materials discovery.

We are organizing all the code and model weights, and we are committed to making the cleanest open-source release possible.

1 Introduction

The discovery and design of novel materials are pivotal to addressing many of the world's most pressing challenges, ranging from energy storage [14, 49] to environmental sustainability [12, 62, 64]. Traditionally, material discovery has relied heavily on trial-and-error methods [62, 3] or computationally expensive first-principles simulations [45, 17]. However, these approaches face significant limitations. Trial-and-error experimentation is inherently slow and resource-intensive, while first-principles simulations often suffer from high computational cost, limited scalability when extending to large or complex systems, and low efficiency in exploring vast chemical design spaces [54, 38]. To overcome these obstacles, deep generative models [27, 18, 51] have recently emerged as promising tools to accelerate material discovery by generating candidate structures directly from data [13, 56, 24]. These models, including large language models (LLMs) [2, 20, 52, 59], diffusion models [24, 25, 31], and hybrid architectures [56, 35], have demonstrated the ability to predict diverse material structures and perform de novo generation [56, 37], ranging from molecular compounds [56, 35] to complex crystal lattices [24, 25]. Despite these advances, there is currently no unified platform to systematically evaluate and compare the performance of generative models in material discovery [55, 41, 63, 46].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

The diversity of model architectures and the breadth of tasks they can address—such as crystal structure prediction, de novo generation, and MOF structure prediction—further complicate objective assessment. The absence of standardized evaluation protocols has hindered direct comparisons between methods and limited their practical applicability to real-world material design problems.

To address this gap, we introduce the Material Generation Benchmark (MGB), a comprehensive and standardized evaluation platform for generative models in materials science. MGB aims to provide a rigorous framework for assessing generative models across multiple key tasks: (i) crystal structure prediction—predicting atomic arrangements given chemical compositions; (ii) de novo material generation—creating novel, valid materials beyond those observed in training datasets; (iii) MOF structure prediction—modeling the atomic configurations of metal–organic frameworks; and (iv) out-of-distribution (OOD) generation—evaluating the generalization ability of models when applied to novel compositions, structures, or property regimes that lie outside the training distribution. These tasks are fundamental to advancing materials design, especially in contexts where direct experimentation or first-principles simulations are prohibitively time-consuming or computationally expensive.

As shown in Figure 1, MGB includes a diverse set of benchmarking datasets, such as MP-20 [22], Perov-5 [5], Carbon-24 [43], MPTS-52 [22], and Boyd MOF [4], spanning materials from simple single-element structures to complex multi-element systems and metal-organic frameworks (MOFs). These datasets are carefully curated to ensure that they represent realistic, experimentally stable materials, and they provide a robust foundation for evaluating the accuracy, diversity, and generalization capabilities of generative models. The benchmark intergrates a variety of leading generative methods: large language models such as CrystalLLM (25M and 200M)[2] and Llama 3.1[19], diffusion models such as DiffCSP [24] and FlowMM [37], and hybrid models that combine variational autoencoders (VAEs) with diffusion, including CDVAE [56] and Cond-CDVAE [35]. Considering fair and meaningful evaluations, MGB adopts a suite of multi-dimensional metrics that go beyond prediction accuracy to assess generation quality, generalization, physical plausibility, symmetry awareness, and computational complexity. This holistic protocol enables standardized and balanced benchmarking of generative models, aligning performance assessments with the practical needs of real-world materials discovery.

Through extensive benchmarking, MGB provides key insights into the current landscape of generative models for materials science: (1) Diffusion-based models consistently perform well on challenging crystalline benchmarks. Notably, DiffCSP++ excels in large and high-symmetry systems due to its explicit modeling of space group features and physical constraints, such as SE(3)-equivariant architectures and crystal periodicity. This aligns well with the underlying physics of materials, offering advantages over VAEs, including better mode coverage and more stable training dynamics. Also, MOF-specific flow models like MOFFlow outperform general-purpose models on MOF prediction tasks. (2) Large language models (e.g., CrystalLLM) exhibit competitive local accuracy, achieving low coordinate errors once a correct match is identified, benefiting from their large model size and extensive pretraining data. (3) In de novo generation, most methods maintain near-perfect structural validity, with diffusion models demonstrating superior preservation of target property distributions. (4) Despite strong in-distribution performance, LLM-based models like CrystaLLM struggle with out-of-distribution (OOD) generation, failing to produce valid structures on diverse OOD datasets. This highlights the distribution-bound nature of these models and emphasizes the importance of evaluating OOD generalization for assessing robustness. (5) Physical plausibility remains a challenge for diffusion models, as atomic collision rates increase significantly on complex datasets. While advances like DiffCSP++ reduce collision rates, they do not eliminate failures, making it crucial to evaluate steric validity to ensure physically realizable materials. These findings highlight the trade-offs between accuracy, generalization, physical constraints, and computational efficiency, suggesting the need for more refined models that better incorporate physical constraints to enable robust material discovery.

Together, these findings underscore the need for a unified and transparent benchmarking framework to drive progress in generative materials modeling. Our primary goal with MGB is to establish a transparent and reproducible benchmarking suite that can catalyze the development of more robust and efficient generative models for materials science. By providing a unified platform, MGB seeks to accelerate the discovery of novel materials with tailored properties and promote fair comparisons across diverse methodological approaches.

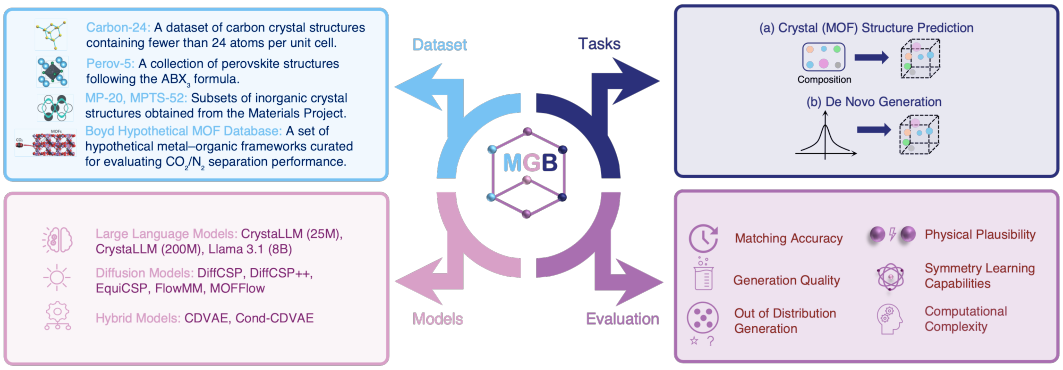


Figure 1: The overview of MGB.

2 Preliminaries

2.1 Crystal Structure

Representation of Crystal Structures and MOF Structures. A periodic material can be represented by a unit cell $\mathcal{M} = (A, F, L)$, where $A \in \mathbb{A}^N$ (or one-hot $A \in \mathbb{R}^{h\times N}$) encodes the atom types of N atoms, $F = [f_1, \ldots, f_N] \in [0, 1)^{3\times N}$ are fractional coordinates, and $L = [\ell_1, \ell_2, \ell_3] \in \mathbb{R}^{3\times 3}$ is the lattice matrix. Cartesian coordinates are given by $X = LF$. Periodicity induces an equivalence relation $(a_i', x_i') \equiv_L (a_i, x_i) \iff x_i' = x_i + Lk,\ k \in \mathbb{Z}^3$, so the infinite crystal is obtained by tiling the unit cell with L. For MOFs we can use either the same atomistic representation or a block-wise one. Let $\mathcal{B} = \{C_m\}_{m=1}^M$ be building blocks (metal nodes or organic linkers) with local coordinates Y_m and atom types a_m. Each block is placed by a roto-translation $(q_m, \tau_m) \in SE(3)$, giving $X_m = (q_m, \tau_m) \cdot Y_m$, and $X = \mathrm{Concat}(X_1, \ldots, X_M)$ forms the global atomic coordinates.

Space Group. Space-group symmetry is modeled as the action of $g = (O, t) \in E(3)$ on coordinates $g \cdot X = OX + t\mathbf{1}^\top$ (with $O \in O(3)$ and $t \in \mathbb{R}^3$). A crystal $\mathcal{M}$ is invariant to g if there exists a permutation matrix P_g such that

$$A = AP_g, \qquad g \cdot X \equiv_L XP_g.$$

The set of all such symmetries forms the space group $G(\mathcal{M})$; in 3D there are 230 distinct space groups. MOFs may realize a subset of symmetry operations depending on their topology and building blocks.

Equivariance. Learning algorithms should respect the physical symmetries. Given a model f acting on structures, f is $SE(3)$-equivariant if

$$f(OX + t\mathbf{1}^\top, OL) = \rho(O)\, f(X, L)$$

for a suitable representation ρ. For atom sets, outputs (e.g., per-atom vectors) should also be permutation-equivariant. In block-wise MOF models, the placement predictor over $\{(q_m, \tau_m)\}$ is $SE(3)$-equivariant.

Invariant Density. Generative models define a probability density on the *quotient space* induced by symmetries. Practically, most of benchmark models parameterize only invariants: (i) use fractional coordinates on the torus $\mathbb{T}^{3N} = [0, 1)^{3N}$ to factor out global translations; (ii) represent the lattice by its Gram matrix $G_L = L^\top L$ or by lattice parameters $(a, b, c, \alpha, \beta, \gamma)$ to factor out global rotations; and (iii) enforce permutation invariance by symmetrization or permutation-invariant architectures. Densities or scores can also be averaged over the space-group orbit to impose G-invariance.

Symmetries of Crystal. Key symmetries include: (1) atom index *permutation*; (2) *periodic translation* (choice of origin and integral lattice shifts); (3) global *rotation/reflection* of (X, L); (4) *lattice basis change* $L \mapsto LU$ with $U \in GL(3, \mathbb{Z})$ (e.g., supercells); and (5) *space-group* operations combining rotations with fractional translations (screws/glides).

2.2 Task Formulation

The generative modeling tasks for periodic crystals are formulated as follows. A crystal is represented by $\mathcal{M} = (L, F, A)$, where L is the lattice, F are the fractional atomic coordinates, and A denotes atom types or elemental fractions c. The first task, Crystal Structure Prediction (CSP), aims to recover a valid periodic structure given a composition (elemental fractions c or atom types A). This is modeled by the conditional distribution $p(L, F \mid A)$ or equivalently $p(\mathcal{M} \mid A)$. In De Novo Generation, both unconditional and conditional sampling of crystals are considered, represented by $p(\mathcal{M})$ and $p(\mathcal{M} \mid G)$, where G is a target space group. For MOF Structure Prediction,

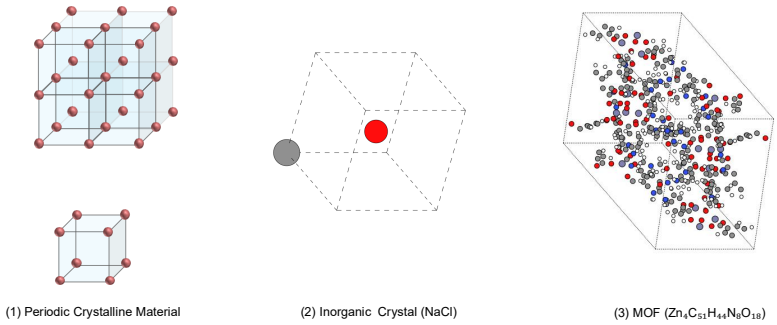


Figure 2: Examples of crystal data structures: (1) demo of periodic crystalline material, (2) inorganic crystal (NaCl), and (3) MOF ($Zn_4C_{51}H_{44}N_8O_{18}$).

two settings are studied: (i) atomistic CSP, which follows the same formulation as in the crystal structure prediction task with $p(L, F \mid A)$; and (ii) block-wise assembly, where we predict block placements and connectivity given a block library $\mathcal{B}$ and, optionally, a target topology T, modeled by $p\left(L, (q_m, \tau_m)_{m=1}^{M}, \text{connectivity} \mid \mathcal{B}, T\right)$, with (q_m, τ_m) representing the pose (orientation and translation) of block m.

Table 1: Algorithms are grouped by modeling paradigm, datasets by composition and experimental setting type, and evaluations by six standardized criteria: matching, generation quality, out-of-distribution generalization, physical plausibility, symmetry awareness, and computational complexity.

Models	
Large Language Models	CrystaLLM (25M) [2], CrystaLLM (200M) [2], Llama 3.1 (8B) [19]
Diffusion Flow Models	DiffCSP [24], DiffCSP++ [25], EquiCSP [31], FlowMM [37], MOFFlow [26]
Hybrid Models	CDVAE [56], Cond-CDVAE [35]
Datasets	
Single-element Composition	Carbon-24 [43]
Multi-element Composition	Perov-5 [5], MP-20 [22], MPTS-52 [22]
Multi-element (complex) Composition	Boyd MOF [4]
Curated OOD Test Sets	OOD-DPC, OOD-AC, OOD-LPC, OOD-HUCC, OOD-FDMOFs
Evaluations	
Matching Accuracy	Match Rate, RMSE
Generation Quality	Validity, coverage, property distribution alignment
Out-of-Distribution Generation	Generation on real world compostion
Physical Plausibility	Obey fundamental physical constraints arising from the balance of attractive and repulsive forces
Symmetry Awareness for LLMs	IPT (Increase in Perplexity under Translation)
Computational Complexity	Model Size, sampling efficiency

3 MGB: The Material Generation Benchmark

3.1 Benchmark Models

Table 1 summarizes the algorithms integrated in our benchmark, which are divided into three categories: diffusion and flow-based models, hybrid models and large language models. We briefly introduce each category and representative algorithms below, and more details are provided in Appendix B.

Diffusion-based Models. These models generate crystal structures by simulating continuous stochastic processes and modeling physical symmetries. DiffCSP [24] and DiffCSP++ [25] incorporate geometric constraints and space group in diffusion modeling. EquiCSP [31] focuses on equivariant diffusion with respect to permutation and periodicity. FlowMM [37] extends flow matching on riemannian manifold for crystal structure prediction and generation. MOFFlow [26] is a riemannian flow matching model for MOF structure prediction.

Hybrid Models. This class integrates variational autoencoders with diffusion processes to generate stable and diverse periodic materials. CDVAE [56] combines VAE and diffusion for periodic material generation. Cond-CDVAE [35] enables conditional generation under user-defined constraints such as composition and pressure.

Table 2: Summary of the benchmark crystalline and MOF datasets.

Dataset	Scope/Type	# Structures	Elements	Atoms/cell
PEROV-5 [5]	Perovskites ABX_3	18,928	56	5
Carbon-24 [43]	C allotropes (AIRSS, 10 GPa)	10,153	1 (C)	6–24
MP-20 [22]	Exp. grounded inorganic crystals	45,231	89	1–20
MPTS-52 [22]	Inorganic crystals (larger cells)	40,476	–	1–52
Boyd MOF [26]	Hypothetical MOFs (adsorption)	247,066	Multi-metal/organic	Variable

Large Language Models. This category covers approaches that leverage large language models for generative crystal design. CrystalLLM (25M) and CrystalLLM (200M) [2] employ autoregressive language modeling directly on CIF files. To further demonstrate the comprehensiveness of MGB, we also include the general-purpose large language models, Llama 3.1 (8B) [19]—to assess the performance on diverse materials design tasks. [1]

3.2 Benchmark Datasets

To comprehensively evaluate material generative models under diverse, realistic conditions, we benchmark on five datasets spanning crystalline solids and metal–organic frameworks (MOFs): PEROV-5 [5], Carbon-24 [43], MP-20 [22], MPTS-52 [22], and the Boyd MOF [26] for main tasks. Additionally, we benchmark on 5 out-of-distribution (OOD) evaluation datasets to assess the models' ability to generalize to unseen structures and functionalities, including shifts in composition, topology, and intended applications. These OOD datasets test the robustness of the generative models across different material classes and properties. Table 2 and 3 summarize their key statistics.

Axes of diversity and why they matter. Across these datasets, diversity arises along (i) **compositional** axes (number/types of elements and allowed chemistries), (ii) **structural** axes (unit-cell size, symmetry/space groups, dimensionality/topology), (iii) **thermodynamic** axes (stable vs. metastable distributions, pressure conditions), and for MOFs also (iv) **functional** axes (intended application such as adsorption, storage, catalysis). These orthogonal sources of variation stress different capabilities of generative models—from capturing composition–structure relationships to handling large, complex topologies and distributions with substantial metastability.

Crystalline datasets. PEROV-5 contains 18,928 perovskites with the nominal ABX_3 formula (A/B are nonradioactive metals; $X \in \{O, N, S, F\}$ and may be mixed). All structures are DFT-relaxed and many are not thermodynamically stable, making composition-to-structure mapping nontrivial. Carbon-24 comprises 10,153 carbon allotropes curated from ab initio random structure searching (AIRSS) at 10 GPa: following the previous work [56], we select the lowest-energy 10% from 101,529 candidates and relax all with DFT; diamond at 10 GPa is the most stable while most others are metastable. MP-20 gathers 45,231 experimentally grounded Materials Project entries (originally from ICSD) with ≤ 20 atoms/cell, filtered by energy-above-hull < 0.08 eV/atom and formation energy < 2 eV/atom; all are DFT-relaxed and largely synthesizable. MPTS-52 extends this regime to 40,476 structures with up to 52 atoms/cell, providing substantially larger search spaces and symmetry/topology variety.

MOF dataset. The Boyd MOF Database targets adsorption-driven carbon capture. It starts from 324,426 hypothetical MOFs generated by topology-based construction and evaluated for CO_2/N_2 uptake under dry/humid conditions. Following [26], we exclude structures with < 200 building blocks, retaining 247,066 MOFs. We adopt an 8:1:1 split (train/val/test) with approximately 197,653 / 24,707 / 24,707 structures. This dataset emphasizes functional and topological diversity at scale, complementing the crystalline benchmarks.

OOD evaluation datasets. To probe extrapolation beyond each training distribution, we design out-of-distribution (OOD) test suites that deliberately shift structure/composition/function while preserving related motifs (details in Appendix E). For PEROV-5 we test: (i) **OOD-DPC** (Double perovskites, $A_2BB'O_6$) with ordered B-site cations and rich magnetism/multiferroicity; (ii) **OOD-AC** (Antiperovskites, M_3AX) with inverted cation/anion roles and often $Pm\bar{3}m$ symmetry; and (iii) **OOD-LPC** (Layered Ruddlesden–Popper phases, $A_{n+1}B_nO_{3n+1}$) exhibiting tunable dimensionality (n). For Carbon-24, we use **OOD-HUCC** (huge unit cell carbon crystals) spanning 28–240

[1] We note that several other large language models have been developed for materials discovery—for example, CrystaltextLLM [20], FlowLLM [37], and Mat2Seq [59]. Their model weights are not publicly available, and no inference platforms exist; retraining them from scratch would require prohibitive computational resources. As is common in the LLM community, we therefore rely on existing checkpoints, making it infeasible to include these models in our benchmark.

Table 3: Out-of-distribution (OOD) test suites curated in this work.

OOD Suite	In-Distribution Target	Primary Shift Tested	Source
OOD-DPC (double perovskites)	PEROV-5	B-site ordering; magnetic/multiferroic variants	Materials Project
OOD-AC (antiperovskites)	PEROV-5	Inverted cation/anion topology; symmetry shift	Materials Project
OOD-LPC (RP phases)	PEROV-5	Reduced dimensionality ($n = 1$–3), layered stacking	Materials Project
OOD-HUCC (carbon)	Carbon-24	Large unit cells (28–240 atoms); symmetry variety	Materials Project/ICSD
OOD-FDMOFs (MOFs)	Boyd MOF	Function shift (delivery/storage/catalysis)	COD

atoms/cell with varied space groups and mixed synthesis status (experimental vs. hypothetical). For the Boyd MOF, we use **OOD-FDMOFs** (Function-Distinct MOFs) curated from COD, covering drug delivery, methane storage, and catalysis. These OOD suites challenge models along motif changes (perovskite→double/anti/layered), cell-size scaling (carbon), and function shift (MOFs), thereby directly testing generalization beyond in-distribution statistics.

Overall, the combination of (i) compositional/structural/thermodynamic/functional diversity (Table 2) and (ii) principled OOD shifts (Table 3) yields a robust testbed for assessing both *accuracy in-distribution* and *generalization out-of-distribution* in crystal/MOF generative modeling. Further dataset details and representative examples are provided in Appendix C and E.

3.3 Benchmark Evaluations

To rigorously assess generative models for materials discovery, we evaluate them in six categories: (1) Matching Accuracy – agreement between predicted and reference structures; (2) Generation Quality – validity, diversity, and property distribution alignment of generated materials; (3) Out-of-Distribution Generation – ability to generate valid, novel materials beyond training data; (4) Physical Plausibility – detection of atomic collisions to ensure physical realism; (5) Symmetry Awareness – capturing invariances such as translation symmetry; and (6) Computational Complexity – model size and inference time, indicating efficiency and scalability. These metrics provide a standardized protocol for fair, comprehensive benchmarking across models and tasks. More details are provided in Appendix D.

Matching Accuracy. For crystal and MOF structure prediction, we evaluate accuracy using the match rate (MR)—the fraction of generated structures that match ground truth via `StructureMatcher` [40], accounting for symmetries. We also report the root mean squared error (RMSE) of atomic coordinates, normalized by cell volume and atom count, to measure geometric fidelity.

Generation Quality. For de novo generation, we assess validity (structural: interatomic distances > 0.5 Å; compositional: charge neutrality via `SMACT` [10]), and diversity through coverage recall (COV-R) and precision (COV-P). Additional metrics include average minimum structure distance (AMSD), composition distance (AMCD), and Earth Mover's Distance (EMD) between generated and reference distributions of density (d_ρ) and number of unique elements (d_{elem}).

OOD Generation. We evaluate *out-of-distribution (OOD) generalization* by assessing a model's ability to generate meaningful and valid samples in regimes unseen during training on the crystal and MOF structure prediction task, particularly for complex and previously unknown structures. This evaluation covers performance on novel compositions and structures, and quantifies both the novelty and robustness of generated materials through targeted OOD benchmarks as well as real-world sampling tasks.

Physical Plausibility. Drawing inspiration from recent works [33, 36], we incorporate explicit *atomic collision* checks to ensure that generated crystals obey fundamental physical constraints arising from the balance of attractive and repulsive interatomic forces. We define an atomic collision as a case where atoms are unrealistically close in space, violating covalent-radius thresholds under explicit periodic boundary conditions (PBC). Given Cartesian coordinates $\mathbf{x}_i, \mathbf{x}_j$ in the unit cell and lattice matrix $\mathbf{L} = [\mathbf{a}, \mathbf{b}, \mathbf{c}]^\top$, all translations

$$\mathbf{n} \in \{-1, 0, 1\}^3, \quad \Delta\mathbf{r}_\mathbf{n} = \mathbf{x}_i - \left(\mathbf{x}_j + \mathbf{n}^\top \mathbf{L}\right)$$

are examined, and the minimum image distance is defined as

$$d_{\min}(i, j) = \min_{\mathbf{n}} \|\Delta\mathbf{r}_\mathbf{n}\|.$$

A collision is flagged if

$$d_{\min}(i, j) < r_i + r_j,$$

where r_i, r_j are triple-bond covalent radii, falling back to double-bond values if missing [8]. To quantify collision prevalence, we compute the periodic-aware pairwise collision ratio

$$\text{PLCR}_{\text{per}} = \frac{\sum_{\text{structures}} \sum_{i<j} \mathbb{I}\big(d_{\min}(i, j) < r_i + r_j\big)}{\sum_{\text{structures}} \binom{K}{2}},$$

Table 4: The benchmarking results (global seed) on the crystal structure prediction task for diffusion-based models.

Method	# of samples	Perov-5		Carbon-24		MP-20		MPTS-52	
		MR (↑)	RMSE (↓)	MR (↑)	RMSE (↓)	MR (↑)	RMSE (↓)	MR (↑)	RMSE (↓)
Cond-CDVAE [35]	1	42.31	0.1356	14.65	0.3216	29.91	0.1098	4.91	0.2387
DiffCSP [24]	1	51.81	0.0922	16.45	0.2865	47.07	0.0654	11.91	0.1493
DiffCSP++ [25] (w/ CSPML)	1	53.71	0.0880	-	-	70.94	0.0295	33.17	**0.0893**
DiffCSP++ [25] (w/ GT)	1	**98.47**	**0.0398**	-	-	**79.76**	0.0293	42.13	0.1134
EquiCSP [31]	1	51.89	0.0746	17.19	0.2751	52.33	0.0612	13.04	0.1293
FlowMM [37]	1	47.38	0.1183	15.53	0.2848	50.21	0.1192	8.20	0.2275
CrystaLLM-raw$_{(25M)}$	1	47.95	0.0966	21.13	**0.1687**	55.85	0.0437	17.47	0.1113
CrystaLLM$_{(25M)}$	1	45.65	0.0977	**21.87**	0.1734	56.58	0.0426	17.54	0.1028
CrystaLLM-raw$_{(200M)}$	1	46.10	0.0953	20.25	0.1761	58.70	0.0408	19.21	0.1110
CrystaLLM$_{(200M)}$	1	45.87	0.0970	20.64	0.1971	58.98	**0.0345**	18.97	0.1123
Cond-CDVAE [35]	20	91.35	0.0312	78.60	0.2657	66.12	0.0985	26.98	0.2250
DiffCSP [24]	20	**98.60**	**0.0118**	87.48	0.2102	**77.54**	0.0611	33.13	0.1843
EquiCSP [31]	20	97.38	0.0173	84.72	0.2278	72.65	0.0782	30.12	0.1985
FlowMM [37]	20	94.58	0.0231	81.45	0.2483	69.10	0.0904	28.34	0.2123
CrystaLLM-raw$_{(25M)}$	20	98.26	0.0236	83.60	0.1523	75.14	0.0395	32.98	0.1197
CrystaLLM$_{(25M)}$	20	98.34	0.0228	84.04	0.1518	75.36	0.0398	32.96	0.1206
CrystaLLM-raw$_{(200M)}$	20	97.60	0.0249	85.17	**0.1514**	73.97	0.0349	33.75	0.1059
CrystaLLM$_{(200M)}$	20	97.73	0.0261	**85.47**	0.1542	74.11	**0.0345**	**34.00**	**0.1076**

where $\mathbb{I}(\cdot)$ is the indicator function. Collisions are further classified as *same-cell* ($\mathbf{n} = \mathbf{0}$) or *cross-cell* ($\mathbf{n} \neq \mathbf{0}$), enabling a more detailed assessment of both intra- and inter-cell stability.

Symmetry Awareness for LLMs. Motivated by recent work [20], we evaluate a model's ability to capture invariances inherent to crystalline materials by assessing its *translation symmetry* using the Increase in Perplexity under Transformation (IPT) metric. For a transformation group G with elements g and group action t, the IPT for an input sequence s is defined as

$$\mathrm{IPT}(s) = \mathbb{E}_{g \in G} \left[\mathrm{PPL}(t_g(s)) - \mathrm{PPL}(t_{\hat{g}}(s)) \right],$$

where

$$\hat{g} = \arg\min_{g} \mathrm{PPL}(t_g(s))$$

is the translation yielding the lowest perplexity, and $\mathrm{PPL}(s) = 2^{\mathrm{CE}(s)/n}$ is the exponentiated length-normalized cross-entropy. In our setting, G is the group of lattice translations, and t_g performs coordinate translation with periodic boundary conditions before re-encoding the structure. IPT thus measures how much a model's compression ability (inverse perplexity) changes under continuous symmetry operations: smaller IPT indicates better invariance. We approximate IPT by sampling multiple translation offsets g (e.g., 20 uniformly spaced shifts), choosing $\hat{g}$ per sequence, and averaging over the test set. In addition to IPT, we compute the percent metastable metric—i.e., the fraction of generated crystal candidates with predicted formation energies below a stability threshold—on symmetry-augmented test inputs.

Computational Complexity. To assess the practical usability and scalability of various generative methods, we evaluate their computational complexity in terms of *model size and the inference time* required for structure generation or prediction. These metrics are especially critical for models designed for large-scale deployments or real-time applications.

4 Experiments and Analysis

4.1 Configurations

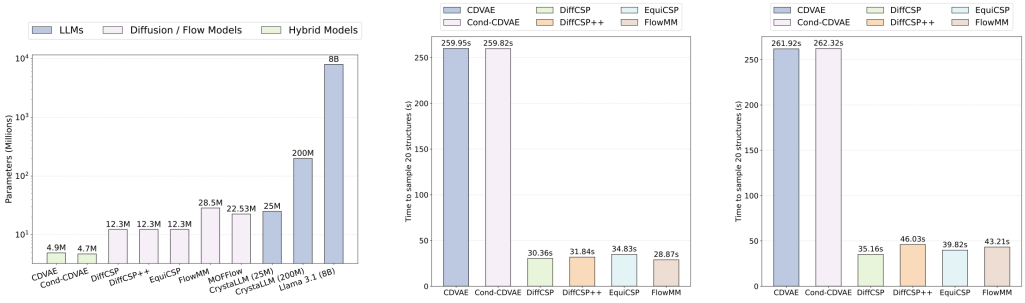
All algorithms and models were developed using Python 3.9.18, with PyTorch 2.2.0, PyTorch Geometric 2.2.0, and Transformers 4.55.0, under CUDA 12.1. For diffusion and hybrid models, experiments were conducted on a server equipped with 8 NVIDIA V100 GPUs (32 GB memory each) and an Intel® Xeon® Platinum 8255C CPU @ 2.50 GHz. For large language model experiments, we utilized NVIDIA A100 and 3090 GPUs.

4.2 Experimental Setup

For training, we trained all diffusion models and hybrid models from scratch. For large language models (LLMs), we used the official open-sourced models. To ensure fair and efficient comparison, all models were trained strictly following the parameter settings provided in their official github repository. For model saving, we retained the best checkpoint based on the minimum validation loss, as well as the final checkpoint at the end of the last training epoch. Both checkpoints were used for sampling and evaluation. For CrystaLLM and Llama 3.1 (8B), we employed prompt-based methods for crystal structure prediction tasks (The specific prompts used are provided in the Appendix J). More details are provided in Appendix H.

Table 5: The benchmarking results of MOF structure prediction task.

Method	# of samples	stol = 0.5		stol = 1.0	
		MR (↑)	RMSE (↓)	MR (↑)	RMSE (↓)
RS [58]	20	0.00	-	0.00	-
EA [58]	20	0.00	-	0.00	-
DiffCSP [24]	1	0.09	0.3961	23.12	0.8294
MOFFlow-raw [26]	1	31.69	0.2820	87.46	**0.5183**
MOFFlow-last [26]	1	**34.47**	**0.2712**	**89.59**	0.5037
MOFFlow-best [26]	1	34.45	**0.2712**	89.54	0.5036



(a) Parameter counts of benchmark models.

(b) Inference time comparison on Perov-5.

(b) Inference time comparison on Carbon-24.

Figure 3: Model scale and inference speed. (a) Parameter counts (log-scale) grouped by paradigm; colors denote families (LLMs, Diffusion/Flow, Hybrid). (b–c) Time to sample 20 structures on Perov-5 and Carbon-24: VAE models are slowest (~ 260 s), while Diffusion/Flow models are much faster (~ 30–46 s). LLM/Hybrid models have larger parameter counts and were not timed.

4.3 The Task of Crystal Structure Prediction

Experiment Analysis. We evaluate CSP on Perov-5, Carbon-24, MP-20, and MPTS-52. For each composition in the test sets, we draw either 1 or 20 random samples per model. Metrics follow the standard protocol: Match Rate (MR↑) via `pymatgen` `StructureMatcher` and coordinate RMSE↓ normalized by $\sqrt{V/N}$. Table 4 shows that multi-sample decoding is crucial: moving from 1 to 20 samples greatly increases MR on all datasets and models. Diffusion models (DiffCSP / DiffCSP++) obtain the highest MR on the challenging MP-20 and MPTS-52 splits—DiffCSP++ is especially strong on larger/complex cells—while LLMs (CrystaLLM-25M & 200M) are highly competitive in RMSE with 20 samples, indicating very accurate local coordinates once a match is found. FlowMM baselines improve with multi-sampling but generally trail diffusion models in MR on the hard splits. Dataset-wise, Perov-5 nearly saturates MR with 20 samples (differences appear mainly in RMSE), Carbon-24 is moderate, and MP-20/MPTS-52 remain the most discriminative. Practically, we recommend enabling multi-sample decoding by default; choose DiffCSP++ when maximizing MR is the priority, and CrystaLLM-200M when the lowest post-match RMSE is desired.

4.4 The Task of MOF Structure Prediction

Experiment Analysis. Table 5 shows that MOF-specific flow models dominate this task. Random search and EA fail to recover any structures (MR$=0$). DiffCSP improves but remains far from practical, while MOFFlow variants achieve large gains under both strict (`stol`=0.5) and loose (`stol`=1.0) matching. With just 5 samples, MOFFlow reaches $\sim$46% MR at `stol`=0.5 and $\geq$97% MR at `stol`=1.0, together with the lowest RMSE ($\approx$0.25–0.27). Multi-sample decoding consistently helps all methods (1$\rightarrow$5 samples), but the gap between MOFFlow and DiffCSP remains substantial, indicating the importance of a MOF-aware generator and the benefit of modeling SE(3) placements of building blocks.

4.5 The Task of De Novo Generation

Experiment Analysis. We report structural and composition validity, coverage (COV-R/P) and property alignment via Earth Mover's Distance (density, #elements) in Table 27. Across datasets, all models achieve near-100% structural validity, and most diffusion models reach $\sim$99% coverage, showing excellent compositional and structural diversity. Property distributions are best aligned by modern diffusion model families: Flow-based and equivariant baselines tend to give the lowest

Table 6: The benchmarking results (global seed) on de novo generation task. The best results are highlighted in bold.

Dataset	Method	Validity (↑)		Coverage (↑)		Property (↓)	
		Struc.	Comp.	COV-R	COV-P	d_ρ	d_{elem}
Perov-5	CDVAE [56]	100.00	96.73	97.33	96.25	0.1532	0.0842
	Cond-CDVAE [35]	100.00	96.32	96.43	95.56	0.1576	0.0932
	DiffCSP [24]	100.00	**98.66**	99.67	98.25	0.1370	0.0542
	DiffCSP++ [25]	100.00	98.65	**99.76**	98.76	**0.1331**	0.0407
	EquiCSP [31]	100.00	98.40	99.42	**98.44**	0.1350	**0.0120**
	FlowMM [37]	**100.00**	96.80	97.40	96.10	0.1520	0.0840
Carbon-24	CDVAE [56]	**100.00**	-	98.50	92.10	0.1450	-
	Cond-CDVAE [35]	99.98	-	96.32	89.90	0.2132	-
	DiffCSP [24]	**100.00**	-	99.90	93.61	0.1429	-
	DiffCSP++ [25]	**100.00**	-	99.90	47.51	**0.0562**	-
	EquiCSP [31]	99.99	-	**99.90**	**96.15**	0.2150	-
	FlowMM [37]	100.00	-	94.32	91.21	0.2390	
MP-20	CDVAE [56]	99.40	80.20	98.70	97.80	0.1600	0.7200
	Cond-CDVAE [35]	99.35	79.80	98.40	97.50	0.1650	0.7400
	DiffCSP [24]	99.78	83.86	99.61	99.47	**0.1027**	0.6129
	DiffCSP++ [25]	99.86	**84.92**	**99.76**	99.43	0.1386	0.4728
	EquiCSP [31]	**99.89**	81.67	99.57	**99.62**	0.6665	**0.3958**
	FlowMM [37]	99.50	80.80	98.90	98.20	0.1550	0.7000
MPTS-52	CDVAE [56]	99.20	63.00	98.50	85.00	1.0500	0.6400
	Cond-CDVAE [35]	99.10	62.50	98.20	84.50	1.0700	0.6600
	DiffCSP [24]	**99.78**	66.70	99.64	88.89	0.9409	0.5573
	DiffCSP++ [25]	99.20	64.50	99.10	85.50	1.0500	**0.6300**
	EquiCSP [31]	99.65	**69.48**	**99.78**	**96.27**	**0.8244**	0.5606

EMDs on small/medium sets, while DiffCSP++ and EquiCSP are highly competitive on the larger MP-20/MPTS-52 splits. Overall, distributional fidelity differences are modest compared with the strong across-the-board validity, suggesting that downstream metrics (e.g., stability or synthesis proxies) are needed to further separate methods.

4.6 Computational Complexity

Figure 3 contrasts model scale and inference speed. Parameter counts stratify by paradigm: VAEs are smallest (~ 4.7–4.9M), diffusion/flow models are mid-sized (~ 12–29M; e.g., DiffCSP/EquiCSP ~12.3M, FlowMM ~ 22.5M, MOFFlow ~ 28.5M), while LLM/hybrid models are much larger (CrystalLLM 25M/200M; Llama-3.1 8B). For sampling 20 structures, VAEs are slowest (~ 260 s on both Perov-5 and Carbon-24; ~ 13 s/structure). Diffusion/flow models are substantially faster: 28.9–34.8 s on Perov-5 and 35.2–46.0 s on Carbon-24 (~ 1.4–2.3 s/structure), a ~ 6–9x speed-up over VAEs with modest dataset-to-dataset variance. LLM/hybrid models were not timed due to their much larger parameter counts. Overall, diffusion/flow offers the most favorable latency–scale trade-off for practical generation workloads.

Due to space limitations, we provide detailed discussions of OOD Generation and Physical Plausibility evaluation in the Appendix I.2 and I.3.

5 Conclusion and Future Work

In this work, we introduce MGB, a unified and standardized platform for evaluating deep generative models in materials science. MGB covers diverse tasks—including crystal structure prediction, de novo generation, MOF structure prediction, and out-of-distribution generation—across representative datasets and models. Through multi-dimensional evaluation metrics, it enables fair, rigorous, and transparent comparisons among models, allowing robust, efficient, and generalizable solutions for material discovery. We hope MGB will serve as a catalyst for accelerating innovation in this field. Also, our future work includes the incorporation of benchmarking for material geometry modeling.

References

[1] Nawaf Alampara, Santiago Miret, and Kevin Maik Jablonka. Mattext: Do language models need more than text & scale for materials modeling? arXiv preprint arXiv:2406.17295, 2024.

[2] Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. Nature Communications, 15(1):1–16, 2024.

[3] Kihoon Bang, Jeongrae Kim, Doosun Hong, Donghun Kim, and Sang Soo Han. Inverse design for materials discovery from the multidimensional electronic density of states. Journal of Materials Chemistry A, 12(10):6004–6013, 2024.

[4] Peter G Boyd, Arunraj Chidambaram, Enrique García-Díez, Christopher P Ireland, Thomas D Daff, Richard Bounds, Andrzej Gładysiak, Pascal Schouwink, Seyed Mohamad Moosavi, M Mercedes Maroto-Valer, et al. Data-driven design of metal–organic frameworks for wet flue gas co2 capture. Nature, 576(7786):253–256, 2019.

[5] Ivano E Castelli, David D Landis, Kristian S Thygesen, Søren Dahl, Ib Chorkendorff, Thomas F Jaramillo, and Karsten W Jacobsen. New cubic perovskites for one-and two-photon water splitting using the computational materials repository. Energy & Environmental Science, 5(10):9034–9043, 2012.

[6] Ivano E Castelli, Thomas Olsen, Soumendu Datta, David D Landis, Søren Dahl, Kristian S Thygesen, and Karsten W Jacobsen. Computational screening of perovskite metal oxides for optimal solar light capture. Energy & Environmental Science, 5(2):5814–5819, 2012.

[7] Yan Chen, Xueru Wang, Xiaobin Deng, Yilun Liu, Xi Chen, Yunwei Zhang, Lei Wang, and Hang Xiao. Mattergpt: A generative transformer for multi-property inverse design of solid-state materials. arXiv preprint arXiv:2408.07608, 2024.

[8] Beatriz Cordero, Verónica Gómez, Ana E Platero-Prats, Marc Revés, Jorge Echeverría, Eduard Cremades, Flavia Barragán, and Santiago Alvarez. Covalent radii revisited. Dalton Transactions, (21):2832–2838, 2008.

[9] Callum J Court, Batuhan Yildirim, Apoorv Jain, and Jacqueline M Cole. 3-d inorganic crystal structure generation and property prediction via representation learning. Journal of Chemical Information and Modeling, 60(10):4518–4535, 2020.

[10] Daniel W Davies, Keith T Butler, Adam J Jackson, Jonathan M Skelton, Kazuki Morita, and Aron Walsh. Smact: Semiconducting materials by analogy and chemical theory. Journal of Open Source Software, 4(38):1361, 2019.

[11] Jyotirmoy Deb, Lakshi Saikia, Kripa Dristi Dihingia, and G Narahari Sastry. Chatgpt in the material design: Selected case studies to assess the potential of chatgpt. Journal of Chemical Information and Modeling, 64(3):799–811, 2024.

[12] Mohammad Mahdi Forootan, Iman Larki, Rahim Zahedi, and Abolfazl Ahmadi. Machine learning and deep learning in energy systems: A review. Sustainability, 14(8):4832, 2022.

[13] Xiang Fu, Tian Xie, Andrew S Rosen, Tommi Jaakkola, and Jake Smith. Mofdiff: Coarse-grained diffusion for metal-organic framework design. arXiv preprint arXiv:2310.10732, 2023.

[14] Addis S Fuhr and Bobby G Sumpter. Deep generative models for materials discovery and machine learning-accelerated innovation. Frontiers in Materials, 9:865270, 2022.

[15] Jingru Gan, Peichen Zhong, Yuanqi Du, Yanqiao Zhu, Chenru Duan, Haorui Wang, Carla P Gomes, Kristin A Persson, Daniel Schwalbe-Koda, and Wei Wang. Large language models are innate crystal structure generators. arXiv preprint arXiv:2502.20933, 2025.

[16] Octavian Ganea, Lagnajit Pattanaik, Connor Coley, Regina Barzilay, Klavs Jensen, William Green, and Tommi Jaakkola. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. Advances in Neural Information Processing Systems, 34:13757–13769, 2021.

[17] Colin W Glass, Artem R Oganov, and Nikolaus Hansen. Uspex—evolutionary crystal structure prediction. Computer physics communications, 175(11-12):713–720, 2006.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, pages 2672–2680, 2014.

[19] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[20] Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. arXiv preprint arXiv:2402.04379, 2024.

[21] Xiao-Qi Han, Zhenfeng Ouyang, Peng-Jie Guo, Hao Sun, Ze-Feng Gao, and Zhong-Yi Lu. Ai-accelerated discovery of high critical temperature superconductors. arXiv preprint arXiv:2409.08065, 2024.

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

[23] Shuyi Jia, Chao Zhang, and Victor Fung. Llmatdesign: Autonomous materials discovery with large language models. arXiv preprint arXiv:2406.13163, 2024.

[24] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. Advances in Neural Information Processing Systems, 36, 2024.

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

[26] Nayoung Kim, Seongsu Kim, Minsu Kim, Jinkyoo Park, and Sungsoo Ahn. Mofflow: Flow matching for structure prediction of metal-organic frameworks. arXiv preprint arXiv:2410.17270, 2024.

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. International Conference on Learning Representations, 2013.

[28] Bo Lei, Enze Chen, Hyuna Kwon, Tim Hsu, Babak Sadigh, Vincenzo Lordi, Timofey Frolov, and Fei Zhou. Grand canonical generative diffusion model for crystalline phases and grain boundaries. arXiv preprint arXiv:2408.15601, 2024.

[29] Ge Lei, Ronan Docherty, and Samuel J Cooper. Materials science in the era of large language models: a perspective. Digital Discovery, 2024.

[30] Qi Li, Rui Jiao, Liming Wu, Tiannian Zhu, Wenbing Huang, Shifeng Jin, Yang Liu, Hongming Weng, and Xiaolong Chen. Powder diffraction crystal structure determination using generative models. arXiv preprint arXiv:2409.04727, 2024.

[31] Peijia Lin, Pin Chen, Rui Jiao, Qing Mo, Cen Jianhuan, Wenbing Huang, Yang Liu, Dan Huang, and Yutong Lu. Equivariant diffusion for crystal structure prediction. In Forty-first International Conference on Machine Learning, 2024.

[32] Hongxuan Liu, Haoyu Yin, Zhiyao Luo, and Xiaonan Wang. Integrating chemistry knowledge in large language models via prompt engineering. Synthetic and Systems Biotechnology, 10(1):23–38, 2025.

[33] Shengchao Liu, Divin Yan, Weitao Du, Weiyang Liu, Zhuoxinran Li, Hongyu Guo, Christian Borgs, Jennifer Chayes, and Anima Anandkumar. Manifold-constrained nucleus-level denoising diffusion model for structure-based drug design. arXiv preprint arXiv:2409.10584, 2024.

[34] Shengchao Liu, Divin Yan, Hongyu Guo, and Anima Anandkumar. An equivariant flow matching framework for learning molecular crystallization. In ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling, 2024.

[35] Xiaoshan Luo, Zhenyu Wang, Pengyue Gao, Jian Lv, Yanchao Wang, Changfeng Chen, and Yanming Ma. Deep learning generative model for crystal structure prediction. arXiv preprint arXiv:2403.10846, 2024.

[36] Jian Ma, Peilin Zhao, Tingyang Xu, and Qifeng Bai. Reducing atomic clashes in geometric diffusion models for 3d structure-based drug design. 2023.

[37] Benjamin Kurt Miller, Ricky TQ Chen, Anuroop Sriram, and Brandon M Wood. Flowmm: Generating materials with riemannian flow matching. arXiv preprint arXiv:2406.04713, 2024.

[38] Md Hosne Mobarak, Mariam Akter Mimona, Md Aminul Islam, Nayem Hossain, Fatema Tuz Zohura, Ibnul Imtiaz, and Md Israfil Hossain Rimon. Scope of machine learning in materials research—a review. Applied Surface Science Advances, 18:100523, 2023.

[39] Viggo Moro, Charlotte Loh, Rumen Dangovski, Ali Ghorashi, Andrew Ma, Zhuo Chen, Peter Y Lu, Thomas Christensen, and Marin Soljačić. Multimodal learning for crystalline materials. arXiv preprint arXiv:2312.00111, 2023.

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

[41] Hyunsoo Park, Zhenzhu Li, and Aron Walsh. Has generative artificial intelligence solved inverse materials design? Matter, 7(7):2355–2367, 2024.

[42] Junkil Park, Youhan Lee, and Jihan Kim. Multi-modal conditioning for metal-organic frameworks generation using 3d modeling techniques. 2024.

[43] Chris J. Pickard. Airss data for carbon at 10gpa and the c+n+h+o system at 1gpa, 2020. URL: https://archive.materialscloud.org/record/2020.0026/v1.

[44] Chris J Pickard and RJ Needs. High-pressure phases of silane. Physical review letters, 97(4):045504, 2006.

[45] Chris J Pickard and RJ Needs. Ab initio random structure searching. Journal of Physics: Condensed Matter, 23(5):053201, 2011.

[46] Raffaele Pugliese, Silvia Badini, Emanuele Frontoni, and Stefano Regondi. Generative artificial intelligence for advancing discovery and design in biomateriomics. Intelligent Computing, 4:0117, 2025.

[47] Zekun Ren, Juhwan Noh, Siyu Tian, Felipe Oviedo, Guangzong Xing, Qiaohao Liang, Armin Aberle, Yi Liu, Qianxiao Li, Senthilnath Jayavelu, et al. Inverse design of crystals using generalized invertible crystallographic representation. arXiv preprint arXiv:2005.07609, 3(6):7, 2020.

[48] Mara Schilling-Wilhelmi, Martiño Ríos-García, Sherjeel Shabih, María Victoria Gil, Santiago Miret, Christoph T Koch, José A Márquez, and Kevin Maik Jablonka. From text to insight: Large language models for materials science data extraction. arXiv preprint arXiv:2407.16867, 2024.

[49] Zhong-Hui Shen, Han-Xing Liu, Yang Shen, Jia-Mian Hu, Long-Qing Chen, and Ce-Wen Nan. Machine learning in energy storage materials. Interdisciplinary Materials, 1(2):175–195, 2022.

[50] Naichen Shi, Hao Yan, Shenghan Guo, and Raed Al Kontar. Multi-physics simulation guided generative diffusion models with applications in fluid and heat dynamics. arXiv preprint arXiv:2407.17720, 2024.

[51] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, pages 2256–2265. PMLR, 2015.

[52] Anuroop Sriram, Benjamin Kurt Miller, Ricky TQ Chen, and Brandon M Wood. Flowllm: Flow matching for material generation with large language models as base distributions. arXiv preprint arXiv:2410.23405, 2024.

[53] Izumi Takahara, Kiyou Shibata, and Teruyasu Mizoguchi. Generative inverse design of crystal structures via diffusion models with transformers. arXiv preprint arXiv:2406.09263, 2024.

[54] Rama Vasudevan, Ghanshyam Pilania, and Prasanna V Balachandran. Machine learning for materials design and discovery. Journal of Applied Physics, 129(7), 2021.

[55] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. npj Computational Materials, 2(1):1–7, 2016.

[56] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. arXiv preprint arXiv:2110.06197, 2021.

[57] Minkai Xu, Shitong Luo, Yoshua Bengio, Jian Peng, and Jian Tang. Learning neural generative dynamics for molecular conformation generation. arXiv preprint arXiv:2102.10240, 2021.

[58] Tomoki Yamashita, Shinichi Kanehira, Nobuya Sato, Hiori Kino, Kei Terayama, Hikaru Sawahata, Takumi Sato, Futoshi Utsuno, Koji Tsuda, Takashi Miyake, et al. Cryspy: a crystal structure prediction tool accelerated by machine learning. Science and Technology of Advanced Materials: Methods, 1(1):87–97, 2021.

[59] Keqiang Yan, Xiner Li, Hongyi Ling, and Shuiwang Ji. Invariant tokenization for language model enabled crystal materials generation. arXiv preprint arXiv:2402.04320, 2024.

[60] Mengjiao Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable diffusion for materials generation. arXiv preprint arXiv:2311.09235, 2023.

[61] Sherry Yang, Simon Batzner, Ruiqi Gao, Muratahan Aykol, Alexander L Gaunt, Brendan McMorrow, Danilo J Rezende, Dale Schuurmans, Igor Mordatch, and Ekin D Cubuk. Generative hierarchical materials search. arXiv preprint arXiv:2409.06762, 2024.

[62] Zhenpeng Yao, Yanwei Lum, Andrew Johnston, Luis Martin Mejia-Mendoza, Xin Zhou, Yonggang Wen, Alán Aspuru-Guzik, Edward H Sargent, and Zhi Wei Seh. Machine learning for a sustainable energy future. Nature Reviews Materials, 8(3):202–215, 2023.

[63] Adrian Xiao Bin Yong, Tianyu Su, and Elif Ertekin. Dismai-bench: benchmarking and designing generative models using disordered materials and interfaces. Digital Discovery, 3(9):1889–1909, 2024.

[64] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. Mattergen: a generative model for inorganic materials design. arXiv preprint arXiv:2312.03687, 2023.

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

Appendix

Table of Contents

A Related Work

Beyond the models and datasets evaluated in our benchmark (MGB), there exists a range of related works that, while highly relevant, are not included due to factors such as lack of open-source implementation or differences in task scope. These studies provide complementary perspectives on generative approaches for materials design and help contextualize the contributions of MGB. In particular, recent advances in generative modeling have introduced a variety of methods and application settings that inform and inspire the design of future benchmarking efforts.

Such advances have shown promising potential for the inverse design of materials, particularly crystalline structures. Methods span diffusion models, GANs, large language models (LLMs), flow matching, and sequence-based encodings, as seen in IMD [41], UniMat [60], GenMS [61], PXRD-Gen [30], CDS&CDI [53], GRIP [28], AIAD [21], MatterGPT [7], CrystalFlow [34], FlowLLM [52], and Mat2Seq [59]. These works target objectives such as generating stable structures, optimizing material properties, and solving structure determination tasks, while facing common challenges in synthesizability, physical interpretability, and evaluation metrics.

In addition to individual generative approaches, specialized benchmarks and datasets have been developed to rigorously evaluate model performance. Dismai-Bench [63] focuses on disordered materials and interfaces, complementing benchmarks that target ordered crystalline systems, and provides structure-comparison-based metrics to reveal strengths, weaknesses, and failure modes in generative models.

Recent trends also highlight integration of multi-modal data, multi-physics simulations, and domain-specific constraints. MultiMat [39] aligns multiple material modalities for representation learning; MPDM [50] incorporates physics simulations into diffusion models; MOFFUSION [42] and MOFD-iff [13] target MOF generation; and MatterGen [64] demonstrates multi-property optimization across inorganic materials.

LLMs are increasingly applied in materials science for design, knowledge extraction, and scientific assistance. Works such as ChatGPTMG [11], LLMatDesign [23], DKPE [32], CrystaltextLLM [20], LLMMSDE [48], MatText [1], and MicroGPT [29] explore applications from structure generation and property prediction to literature mining and autonomous research agents, while also noting limitations in data availability, controllability, and factual accuracy.

B The Details of the Benchmarking Algorithms

B.1 Large Language Models

CrystalLLM (25M & 200M) [2]. CrystaLLM proposes crystal structure generation by autoregressive large language modeling directly on CIF files. Treating crystal structures as token sequences, CrystaLLM is trained on millions of inorganic crystals and can generate valid, diverse structures for unseen compositions and symmetries. It achieves competitive or superior match rates and geometric accuracy compared to diffusion-based models. The model supports conditional generation on space group or composition, and integrates MCTS and energy predictors for low-energy structure search. Code and web app: https://github.com/lantunes/CrystaLLM, https://crystallm.com.

Llama 3.1 (8B) [19]. Llama 3.1 (8B) is an open-source large language model released by Meta AI in 2024. With 8 billion parameters, it supports context windows up to 128k tokens and incorporates Grouped-Query Attention for efficient long-context reasoning. The model is pretrained on large-scale text corpora and further instruction tuned, enabling strong performance on multilingual dialogue, text generation, and classification tasks. Compared with larger variants, Llama 3.1 (8B) achieves a balance between capability and computational efficiency, making it suitable for research and deployment in resource-constrained environments. The model is open sourced at https://huggingface.co/meta-llama/Llama-3.1-8B.

We note that Table 7 summarizes all currently available large language models for material design, including both open-source and closed-source efforts. In particular, models such as CrystaltextLLM [20], FlowLLM [37], and Mat2Seq [59] have been proposed for materials discovery. However, their model weights are not publicly released, and no inference platforms exist to support them. As indicated in the table, these models are either fine-tuned from proprietary checkpoints (e.g., CrystaltextLLM, FlowLLM) or trained from scratch without accessible artifacts (e.g., Mat2Seq). Retraining them independently would require prohibitive computational resources, making it infeasible to include them in our benchmark. Consistent with common practice in the LLM community, we therefore rely on existing checkpoints with accessible weights and APIs, which ensures reproducibility and fair comparison across models.

Table 7: Summary of LLMs for material design.

Model	Open Source	#Params	Base LLM	Training Type	Benchmarked in MGB
CrystaLLM [2]	✓	25M/200M	None	Trained from scratch	✓
MatLLMSearch [15]	✓	70B	Llama 3.1	No training	✗
CrystaltextLLM [20]	✗	7B/13B/70B	Llama-2	Fine-tuned	✗
FlowLLM [37]	✗	70B	Llama-2	Fine-tuned	✗
Mat2Seq [59]	✗	25M/200M	None	Trained from scratch	✗

B.2 Diffusion Models

DiffCSP [24]. DiffCSP is a novel diffusion-based generative model designed for Crystal Structure Prediction (CSP), addressing the challenges posed by the geometric symmetries of crystals, such as translation, rotation, and periodicity. By leveraging fractional coordinates and a periodic-E(3)-equivariant denoising model, DiffCSP jointly generates both lattice vectors and atom positions, effectively capturing the intrinsic periodicity and symmetries of crystal structures. Unlike conventional methods that rely on computationally expensive DFT or Cartesian coordinate-based generative models, DiffCSP provides a more accurate and computationally efficient solution. The code for DiffCSP is publicly available at: https://github.com/jiaor17/DiffCSP.

DiffCSP++ [25]. DiffCSP++ is a novel diffusion-based model designed for crystal generation that incorporates space group constraints, which are crucial for capturing the geometric and symmetry properties of crystals. It translates the space group constraint into two tractable components: the basis constraint of the O(3)-invariant logarithmic space of the lattice matrix and the Wyckoff position constraint of the fractional coordinates of atoms. These constraints are seamlessly integrated into the diffusion process, allowing DiffCSP++ to generate lattices, atomic coordinates, and atom compositions while maintaining the symmetry of the crystal. By explicitly considering these constraints, DiffCSP++ improves upon the previous DiffCSP model and achieves superior performance in tasks such as crystal structure prediction, ab initio crystal generation, and controllable generation with

specific space groups across various datasets. The code for DiffCSP++ is publicly available at: https://github.com/jiaor17/DiffCSP-PP.

EquiCSP [31]. EquiCSP is a novel equivariant diffusion-based generative model developed to tackle the challenges of Crystal Structure Prediction (CSP). It ensures both lattice permutation and periodic translation equivariance, addressing limitations in previous models that overlooked these critical symmetries. To achieve this, EquiCSP introduces a specialized diffusion noising algorithm called Periodic CoM-free Noising, which maintains periodic translation equivariance throughout both training and generation. Additionally, it employs simple loss functions to enforce lattice permutation equivariance without embedding it directly into the neural network architecture, thus improving computational efficiency. The code for EquiCSP is publicly available at: https://github.com/EmperorJia/EquiCSP.

FlowMM [37]. FlowMM is a generative model framework developed to predict and generate stable crystalline materials by extending the Riemannian Flow Matching method. It is specifically designed to handle the unique symmetries of periodic crystals, including translation, rotation, permutation invariances, and periodic boundary conditions. By modeling the joint distribution over lattice parameters, atomic coordinates, and atom types, FlowMM provides a unified solution for both Crystal Structure Prediction (CSP) and De Novo Generation (DNG). The code for FlowMM is publicly available at: https://github.com/facebookresearch/flowmm.

MOFFlow [26]. MOFFlow is a generative framework designed for the discovery and design of Metal–Organic Frameworks (MOFs). Built upon an equivariant flow-based model, it captures the inherent symmetries of crystalline MOF structures, including translation, rotation, and periodic boundary conditions. The model jointly generates atom types, coordinates, and lattice parameters, enabling both crystal structure prediction and de novo MOF generation. Trained on large-scale MOF datasets, MOFFlow demonstrates strong capability in generating stable and diverse MOF structures while preserving chemical validity. The code for MOFFlow is publicly available at: https://github.com/nayoung10/MOFFlow.

B.3 Hybrid Models

CDVAE [56]. CDVAE is a Crystal Diffusion Variational Autoencoder designed to generate stable periodic materials by addressing the challenge of material generation, where stability is dictated by quantum mechanical energy minima and specific atomic bonding preferences. CDVAE employs a diffusion process in its decoder that iteratively refines atomic coordinates and atom types, pushing them towards stable configurations. Built upon SE(3) equivariant graph neural networks, CDVAE respects critical physical invariances, including permutation, translation, rotation, and periodic boundary conditions. The model outperforms previous methods in tasks such as input structure reconstruction, generating diverse and realistic materials, and optimizing materials for specific properties. Additionally, CDVAE contributes standard datasets and evaluation metrics to facilitate a consistent comparison of generative models in material science. The code for CDVAE is publicly available at: https://github.com/txie-93/cdvae.

Cond-CDVAE [35]. The Cond-CDVAE is a deep learning-based generative model developed for crystal structure prediction (CSP) under user-defined conditions such as chemical composition and pressure. Trained on a vast dataset of 670,979 stable crystal structures from the Materials Project and CALYPSO databases, it can generate valid and diverse crystal structures with high accuracy. The Cond-CDVAE outperforms conventional CSP methods in both efficiency and fidelity, particularly for structures with fewer than 20 atoms per unit cell. Conditioning on physical parameters enables the exploration of crystal structures across a wide range of pressures, facilitating materials discovery without the need for computationally expensive local optimization. The code for Cond-CDVAE is publicly available at:https://github.com/ixsluo/cond-cdvae.

C The Details of the Benchmarking Datasets

C.1 PEROV-5

Perovskite is a class of materials that share a similar structure and have the general chemical formula ABX_3. The ideal perovskites have a cubic structure, where the site A atom sits at a corner position, the site B atom sits at a body-centered position and site X atoms sit at face centered positions. Perovskite materials are known for their wide applications. We curate the Perov-5 dataset from an open database that was originally developed for water splitting [5, 6]. All 18928 materials in the original database are included. In the database, A, B can be any nonradioactive metal, and X can be one or several elements from O, N, S, and F. Note that there can be multiple different X atoms in the same material. All materials in Perov-5 are relaxed using density functional theory (DFT), and their relaxed structure can deviate significantly from the ideal structures. A significant portion of the materials is not thermodynamically stable, i.e., they will decompose to nearby phases and cannot be synthesized. PEROV-5 [5] includes 18928 perovskite materials that share the same structure but differ in composition. There are 56 elements, and all materials have 5 atoms in the unit cell.

C.2 Carbon-24

Carbon-24 includes various carbon structures obtained via ab initio random structure searching (AIRSS) [44, 45] performed at 10 GPa. The original dataset includes 101529 carbon structures, and we selected 10% of the carbon structures with the lowest energy per atom to create Carbon-24. All 10153 structures in Carbon-24 are relaxed using DFT. The most stable structure is diamond at 10 GPa. All remaining structures are thermodynamically unstable but may be kinetically stable. Most of the structures cannot be synthesized. Carbon-24 [43] includes 10153 materials that are all made up of carbon atoms but differ in structures. There is 1 element, and the materials have 6 - 24 atoms in the unit cells.

C.3 MP-20

MP-20 includes almost all experimentally stable materials from the Materials Project [22] with unit cells including at most 20 atoms. We only include materials that are originally from ICSD [] to ensure the experimental stability, and these materials represent the majority of experimentally known materials with at most 20 atoms in unit cells. To ensure stability, we only select materials with energy above the hull smaller than 0.08 eV/atom and formation energy smaller than 2 eV/atom, following [47]. Differing from [47], we do not constrain the number of unique elements per material. All materials in MP-20 are relaxed using DFT. Most materials are thermodynamically stable and have been synthesized. MP-20 [22] includes 45231 materials that differ in both structure and composition. There are 89 elements, and the materials have 1 - 20 atoms in the unit cells.

C.4 MPTS-52

MPTS-52 [22] is a more challenging extension of MP-20, consisting of 40,476 structures up to 52 atoms per cell, sorted according to the earliest published year in literature.

C.5 Boyd MOF Database

The Boyd MOF Database originates from the work [4], focusing on the data-driven design of metal–organic frameworks (MOFs) for wet flue gas CO_2 capture. The original dataset consists of 324,426 hypothetical MOF structures generated by high-throughput topology-based construction. Each structure was evaluated for CO_2 and N_2 adsorption properties under both dry and humid conditions, aiming to identify robust adsorbent materials capable of selective CO_2 capture in industrial flue gas streams. In the benchmark setting, following prior work [26], structures with fewer than 200 building blocks were excluded, resulting in 247,066 MOFs retained. The dataset is randomly split into training, validation, and test sets with an 8:1:1 ratio, yielding approximately 197,653 / 24,707 / 24,707 structures, respectively. This dataset is particularly challenging due to its large scale and diversity. MOFs in the database span a wide range of compositions, topologies, and pore characteristics. While many structures are hypothetical, they provide a rich testbed for machine learning algorithms in materials discovery, particularly in the context of adsorption-based carbon capture.

D The Details of the Benchmarking Evaluations

D.1 Matching Accuracy for Crystal (MOF) Structure Prediction

Match Rate. The Match Rate is the proportion of the matched structures over the test set. We evaluate the match rate performance by matching the generated structure and the input structure for all materials in the test set. We use `StructureMatcher` from `pymatgen` [40], , which finds the best match between two structures considering all invariances of materials. The match rate is the percentage of materials satisfying the criteria `stol=0.5, angle tol=10, ltol=0.3`.

RMSE. RMSE is calculated between the ground truth and the best matching candidate, normalized by $\sqrt{V/N}$ where V is the volume of the lattice, N is the number of atoms in the unit cell, and averaged over the matched structures.

D.2 Quality for De Novo Generation

Structural Validity. Following [9], a structure is valid as long as the shortest distance between any pair of atoms is larger than 0.5 Å, which is a relative weak criterion.

Compositional Validity. The composition is valid if the overall charge is neutral as computed by SMACT [10].

Coverage Recall (COV-R) and Coverage Precision (COV-P). Inspired by [57, 16], we define two coverage metrics, COV-R (Recall) and COV-P (Precision), to measure the similarity between ensembles of generated materials and ground truth materials in the test set. Intuitively, COV-R measures the percentage of ground truth materials being correctly predicted. COV-P measures the percentage of predicted materials having high quality.

Inspired by [57, 16], we define six metrics to compare two ensembles of materials: materials generated by a method $\{\boldsymbol{M}_k\}_{k\in[1..K]}$, and ground truth materials in test data $\{\boldsymbol{M}_l^*\}_{\in[1..L]}$. We use the Euclidean distance of the CrystalNN fingerprint [65] and normalized Magpie fingerprint [55] to define the structure distance and composition distance between generated and ground truth materials, respectively. They can be written as $D_{\text{struc.}}(\boldsymbol{M}_k, \boldsymbol{M}_l^*)$ and $D_{\text{comp.}}(\boldsymbol{M}_k, \boldsymbol{M}_l^*)$. We further define the thresholds for the structure and composition distance as $\delta_{\text{struc.}}$ and $\delta_{\text{comp.}}$, respectively. Following the established classification metrics of Precision and Recall, we define the coverage metrics as:

$$\text{COV-R (Recall)} = \frac{1}{L}|\{l \in [1..L] : \exists k \in [1..K], D_{\text{struc.}}(\boldsymbol{M}_k, \boldsymbol{M}_l^*) < \delta_{\text{struc.}}, D_{\text{comp.}}(\boldsymbol{M}_k, \boldsymbol{M}_l^*) < \delta_{\text{comp.}}\}| \tag{1}$$

$$\text{AMSD-R (Recall)} = \frac{1}{L}\sum_{l\in[1..L]} \min_{k\in[1..K]} D_{\text{struc.}}(\boldsymbol{M}_k, \boldsymbol{M}_l^*) \tag{2}$$

$$\text{AMCD-R (Recall)} = \frac{1}{L}\sum_{l\in[1..L]} \min_{k\in[1..K]} D_{\text{comp.}}(\boldsymbol{M}_k, \boldsymbol{M}_l^*), \tag{3}$$

where COV is "Coverage", AMSD is "Average Minimum Structure Distance", AMCD is "Average Minimum Composition Distance", and COV-P (precision), AMSD-P (precision), AMCD-P (precision) are defined as in above equations, but with the generated and ground truth material sets swapped. The recall metrics measure how many ground truth materials are correctly predicted, while the precision metrics measure how many generated materials are of high quality (more discussions can be found in [16]).

We note several points on why we define the metrics in their current forms. 1) COV requires *both* structure and composition distances to be within the thresholds, because generating materials that are structurally close to one ground truth material and compositionally close to another is not meaningful. As a result, AMSD and AMCD are less useful than COV. 2) We use fingerprint distance, rather than RMSE from `StructureMatcher` [40], because the material space is too large for the models to generate enough materials to *exactly* match the ground truth materials. `StructureMatcher` first requires the compositions of two materials to exactly match, which will cause all models to have close-to-zero coverage. For Perov-5 and Carbon-24, we choose $\delta_{\text{struc.}} = 0.2, \delta_{\text{comp.}} = 4$. For MP-20 and MPTS-52, we choose $\delta_{\text{struc.}} = 0.4, \delta_{\text{comp.}} = 10$.

Property Statistics d_ρ, d_E and d_{elem}. To quantitatively evaluate the similarity between the generated and test material property distributions, we compute the Earth Mover's Distance (EMD, i.e., Wasserstein distance) for three representative properties: (1) density (ρ, unit: $\mathrm{g/cm^3}$), (2) formation energy per atom (E, unit: $\mathrm{eV/atom}$), and (3) the number of unique elements (# elem.). The formation energy is predicted using an independent graph neural network (GNN) trained on an external dataset, ensuring unbiased property evaluation. For each property, the Wasserstein distance is calculated between the distributions of generated structures and those of the test set. Unless otherwise specified, the property metrics are evaluated on a subset of 1,000 valid generated samples. Validity and coverage are computed over $N(N \in [5000, 10000, 15000, 20000])$ materials randomly sampled from $\mathcal{N}(0, 1)$. Property statistics is computed over 1,000 valid materials randomly sampled from those that pass the validity test.

E Out-of-distribution Test

In this section, we evaluate the generalization ability of models beyond their training distributions through out-of-distribution (OOD) tests. We construct representative OOD datasets for three material classes: (**i**) materials with perovskite-related motifs but outside the Perov-5 distribution, including double perovskites, antiperovskites, and layered perovskites (Ruddlesden-Popper phases) as shown in Tables 8, 9, and 10; (**ii**) carbon allotropes with large unit cells for testing the extrapolation of Carbon-24 models (Table 11); and (**iii**) MOFs with distinct functions from CO_2 adsorption, used to assess the Boyd hypothetical MOF database (Table 12). Each dataset is curated from publicly available sources (Materials Project, ICSD, COD) and follows a consistent selection criterion. The OOD evaluation metrics include effectiveness, uniqueness/diversity, stability (e.g., E_{hull} or phase consistency), and property performance, facilitating cross-model and material class comparisons.

E.1 Out of Distribution Datasets for Perov-5

To comprehensively evaluate the generalization ability of generative models beyond the Perov-5 dataset, we construct several out-of-distribution (OOD) test sets. These datasets are designed to include material families that share structural motifs with perovskites but are not part of the Perov-5 training distribution. Specifically, we consider three representative categories: double perovskites, antiperovskites, and layered perovskites (Ruddlesden-Popper phases). All materials are collected from the Materials Project, and they provide diverse structural and functional characteristics that challenge models to extrapolate beyond the standard perovskite composition space.

Double Perovskite Crystals (OOD-DPC). Double perovskites ($A_2BB'O_6$) are an important class of materials in which two distinct cations occupy alternating lattice sites. Their compositional tunability leads to diverse functionalities, ranging from magnetism and multiferroicity to ferroelectricity and catalysis. Owing to this structural and functional diversity, double perovskites provide a strong out-of-distribution (OOD) benchmark for evaluating models trained on Perov-5. In Table 8, we present representative double perovskite crystals collected from the Materials Project, covering categories such as magnetic & spin-polarized, multiferroic, dielectric & ferroelectric, photocatalytic & photoelectric, oxygen reduction & catalytic, and other representative compounds. Each entry reports the chemical formula, Materials Project ID, and space group, highlighting the broad coverage of double perovskites beyond the Perov-5 dataset.

Antiperovskite Crystals (OOD-AC). Antiperovskites (M_3AX) are structural analogues of perovskites in which anion and cation positions are inverted. They exhibit unique physical properties such as metallic conductivity, mechanical robustness, and unconventional magnetism, making them distinct from the perovskite family while still sharing related motifs. Their structural differences render them a suitable OOD test set for Perov-5–based generative models. Table 9 summarizes representative nitride-type (M_3AN), carbide-type (M_3AC), and other common antiperovskites, all collected from the Materials Project. For each material, we provide the chemical formula, Materials Project ID, and space group. The majority belong to the high-symmetry group Pm-3m (221), yielding a simple yet clearly out-of-domain evaluation set.

Layered Perovskite Crystals (Ruddlesden-Popper Phase) (OOD-LPC). Layered perovskites, or Ruddlesden-Popper (RP) phases, consist of perovskite layers separated by rock-salt layers, following the general formula $A_{n+1}B_nO_{3n+1}$. Their tunable dimensionality, controlled by the stacking parameter n, gives rise to rich electronic and optical behaviors, particularly in reduced-dimensional systems. Since RP phases extend beyond the standard perovskites of Perov-5, they provide a challenging benchmark for OOD evaluation. Table 10 reports representative RP phases collected from the Materials Project, grouped into $n = 1$, $n = 2$, and $n = 3$ categories, together with 2D organic–inorganic RP compounds and other RP variants. Each entry lists the chemical formula, Materials Project ID, and space group, illustrating the structural diversity of layered perovskites outside the Perov-5 distribution.

Table 8: Selected double perovskite crystals for out-of-distribution test.

Category	Pretty Formula	Material ID	Space Group	Source
Magnetic& Spin-polarized Materials	Sr2CrReO6	mp-1205958	Fm-3m, 225	The Materials Project
	Sr2CrOsO6	mp-1078354	R-3, 148	The Materials Project
	Ca2FeMoO6	mp-18783	P2_1/c, 14	The Materials Project
	Ba2FeReO6	mp-31756	Fm-3m, 225	The Materials Project
	La2VMnO6	mp-560369	P2_1/c, 14	The Materials Project
	La2CoMnO6	mp-19208	P2_1/c, 14	The Materials Project
	La2CrMnO6	mp-1223342	P2_1/c, 14	The Materials Project
Multiferroic Materials	Bi2FeCrO6	mp-551086	R3, 146	The Materials Project
	La2NiMnO6	mp-1079517	Fm-3m, 225	The Materials Project
	Y2CoMnO6	mp-1189894	P2_1/c, 14	The Materials Project
	Pb2CoWO6	mp-20069	C2/m, 12	The Materials Project
Dielectric & Ferroelectric Materials	Sr2LaTaO6	mp-1205692	Fm-3m, 225	The Materials Project
	Sr2GdNbO6	mp-1518774	Pn-3, 201	The Materials Project
	Sr2ScSbO6	mp-1106218	P2_1/c, 14	The Materials Project
	Ba2LaNbO6	mp-553281	C2/m, 12	The Materials Project
Photocatalytic & Photoelectric Materials	Sr2AlTaO6	mp-1147547	P4/mmm, 123	The Materials Project
	Sr2FeTiO6	mp-1094048	Fm-3m, 225	The Materials Project
	Ba2BiSbO6	mp-23091	R-3, 148	The Materials Project
	Sr2MgMoO6	mp-1078539	I4/m, 87	The Materials Project
Oxygen Reduction & Catalytic Materials	Pr2NiMnO6	mp-1209751	P2_1/c, 14	The Materials Project
	La2FeCoO6	mp-1223373	P2_1/c, 14	The Materials Project
	La2MnCoO6	mp-19208	P2_1/c, 14	The Materials Project
	La2NiCoO6	mp-1223259	R-3, 148	The Materials Project
Other Representative Double Perovskites	Sr2GaSbO6	mp-6065	Fm-3m, 225	The Materials Project
	Ba2ScSbO6	mp-20709	Fm-3m, 225	The Materials Project
	Ba2HoTaO6	mp-13000	I4/m, 87	The Materials Project
	Sr2MgWO6	mp-18848	Fm-3m, 225	The Materials Project
	Sr2CoWO6	mp-18771	I4/m, 87	The Materials Project
	Ba2ErNbO6	mp-6653	Fm-3m, 225	The Materials Project

Table 9: Selected antiperovskite crystals for out-of-distribution test.

Category	Antiperovskite	Material ID	Space Group	Source
Nitride (M3AN type)	Mn3GaN	mp-627439	Pm-3m, 221	The Materials Project
	Mn3ZnN	mp-15805	Pm-3m, 221	The Materials Project
	Mn3CuN	mp-510380	Pm-3m, 221	The Materials Project
	Mn3NiN	mp-20362	Pm-3m, 221	The Materials Project
	Fe3Mo3N	mp-510619	Fd-3m, 227	The Materials Project
	Co3InN	mp-1068786	Pm-3m, 221	The Materials Project
	Ni3ZnN	mp-1069270	Pm-3m, 221	The Materials Project
Carbide (M3AC type)	Fe3SnC	mp-21850	Pm-3m, 221	The Materials Project
	Co3SnC	mp-20679	Pm-3m, 221	The Materials Project
	Mn3AlC	mp-4593	Pm-3m, 221	The Materials Project
	Ni3AlC	mp-1207084	Pm-3m, 221	The Materials Project
	Fe3ZnC	mp-10266	Pm-3m, 221	The Materials Project
Other Common Antiperovskites	Ni3InN	mp-1070713	Pm-3m, 221	The Materials Project
	Fe3SbN	mp-1246554	Imma, 74	The Materials Project
	Mn3GeN	mp-1205588	I4/mcm, 140	The Materials Project
	Mn3SbN	mp-1206805	Pm-3m, 221	The Materials Project
	Mn3SnN	mp-505571	Pm-3m, 221	The Materials Project

Table 10: Selected layered perovskite crystals for out of distribution test (Ruddlesden-Popper phase).

Category	Layered Perovskite	Material ID	Space Group	Source
n = 1 RP phase	Sr2TiO4	mp-5532	I4/mmm, 139	The Materials Project
	La2CuO4	mp-19735	I4/mmm, 139	The Materials Project
	K2NiF4	mp-556546	I4/mmm, 139	The Materials Project
	Ca2MnO4	mp-19050	I4_1/acd, 142	The Materials Project
	Ba2CuO4	mp-1147762	I4/mmm, 139	The Materials Project
	Sr2RuO4	mp-4596	I4/mmm, 139	The Materials Project
n = 2 RP phase	Sr3Ti2O7	mp-3349	I4/mmm, 139	The Materials Project
	Ca3Ti2O7	mp-4163	Cmc2_1, 36	The Materials Project
	La3Ni2O7	mp-18926	Cmcm, 63	The Materials Project
	Sr3Fe2O7	mp-18820	I4/mmm, 139	The Materials Project
	Sr3Ru2O7	mp-5868	I4/mmm, 139	The Materials Project
n = 3 RP phase	Sr4Ti3O10	mp-31213	I4/mmm, 139	The Materials Project
	La4Ni3O10	mp-19298	I4/mmm, 139	The Materials Project
	Sr4Ru3O10	mp-680680	Cmce, 64	The Materials Project
2D Organic-Inorganic RP Perovskite	(BA)2PbI4	mp-6280	Pnma, 62	The Materials Project
	(PEA)2PbI4	mp-550306	I4/mmm, 139	The Materials Project
	(BA)2MAPb2I7	mp-720710	P-1, 2	The Materials Project
Other RP perovskites	Sr2FeO4	mp-19102	I4/mmm, 139	The Materials Project
	La2NiO4	mp-20143	Cmce, 64	The Materials Project
	Sr2CoO4	mp-18724	I4/mmm, 139	The Materials Project
	K2MgF4	mp-31212	I4/mmm, 139	The Materials Project

E.2 Out of Distribution Datasets for Carbon-24

Huge Unit Cell Carbon Crystals (OOD-HUCC). The Carbon-24 dataset contains diverse carbon allotropes generated via ab initio random structure searching (AIRSS) at 10 GPa, from which over 10,000 low-energy structures were curated and relaxed using DFT. While diamond remains the most stable phase, most of these structures are metastable and not experimentally synthesizable, thereby offering a wide structural diversity beyond well-known carbon forms such as diamond and graphite. To construct a meaningful out-of-distribution (OOD) benchmark for Carbon-24, we further collected representative carbon crystals from the Materials Project, as shown in Table 11. These crystals exhibit varied huge unit cell sizes, space groups, and stability profiles, with some experimentally observed and others hypothetical. Each entry reports the atom number, Materials Project ID, space group, and whether it has been experimentally realized. This dataset highlights both the diversity of carbon structures and their suitability for OOD evaluation beyond the Carbon-24 training distribution.

Table 11: Selected huge unit cell carbon crystals for the out-of-distribution test.

Atom Numbers	Material ID	Space Group	Synthesis Status	Source
240	mp-1196583	Pa-3, 205	✓	The Materials Project
140	mp-683919	Cmcm, 63	✓	The Materials Project
120	mp-1147718	Pnnm, 58	✗	The Materials Project
120	mp-568028	Pnnm, 58	✓	The Materials Project
120	mp-1205283	Pnnm, 58	✗	The Materials Project
100	mp-1245190	P1, 1	✗	The Materials Project
100	mp-1244913	P1, 1	✗	The Materials Project
100	mp-1244964	P1, 1	✗	The Materials Project
80	mp-1197903	P1, 1	✗	The Materials Project
80	mp-1182684	P2_12_12_1, 19	✗	The Materials Project
71	mp-1096869	Cm, 8	✗	The Materials Project
60	mp-680372	R-3m, 166	✓	The Materials Project
60	mp-667273	Fm-3, 202	✓	The Materials Project
60	mp-630227	Immm, 71	✓	The Materials Project
52	mp-1196857	Pnma, 62	✓	The Materials Project
48	mp-723638	P2_1/c, 14	✓	The Materials Project
29	mp-1192619	I-43m, 217	✗	The Materials Project
28	mp-731594	P2_1, 4	✓	The Materials Project
32	icsd-673340	P2_1/c	✗	ICSD
32	icsd-673342	P21c	✗	ICSD
96	icsd-671853	Pm-3m	✗	ICSD

E.3 Out of Distribution Datasets for Boyd MOF Database

Out of Distribution Datasets for Function-Distinct MOFs (OOD-FDMOFs). The Boyd MOF Database consists of over 300,000 hypothetical MOF structures generated by topology-based construction and primarily evaluated for CO_2 capture. While this dataset is valuable for adsorption studies, it does not fully represent the diversity of experimentally realized MOFs used in other applications. To construct a meaningful out-of-distribution (OOD) benchmark, we deliberately collected MOFs whose primary functions are distinct from CO_2 adsorption, such as drug delivery, methane storage, and heterogeneous catalysis. Table 12 summarizes representative MOFs from the Crystallography Open Database, covering well-known families such as MIL, UiO, ZIF, and HKUST. For each MOF, we provide the metal center, organic linker, space group, and data source, along with CIF availability and reference links. This curated dataset highlights structural and functional diversity outside the CO_2-focused Boyd database, making it a suitable OOD benchmark for evaluating model generalization.

Table 12: Representative MOFs with primary functions distinct from CO_2 adsorption for the out-of-distribution test.

Category	MOF	Metal Center	Organic Linker	Space Group	CIF	Source	Link
Drug Delivery	MIL-100(Fe)	Fe(III), Cr(III)	1,3,5-benzenetricarboxylate (BTC)	-	✓	COD	link
	MIL-100(Cr)	Fe(III), Cr(III)	1,3,5-benzenetricarboxylate (BTC)	-	✓	COD	link
	UiO-66	Zr(IV)	Terephthalic acid (BDC), Biphenyldicarboxylate	-	✓	COD	link
	UiO-67	Zr(IV)	Terephthalic acid (BDC), Biphenyldicarboxylate	-	✓	COD	link
	ZIF-8	Zn(II)	2-methylimidazolate	-	✓	COD	link
	BioMOF-100	Zn(II)	Adenine, BTC	-	✗	COD	link
	BioMOF-1	Various (e.g., Cu)	Biomolecule, peptide or aromatic carboxylates	-	✓	COD	link
	BioMOF-11	Various (e.g., Cu)	Biomolecule, peptide or aromatic carboxylates	-	✓	COD	link
Methane Storage	HKUST-1 (Cu-BTC)	Cu(II)	1,3,5-benzenetricarboxylate (BTC)	-	✓	COD	link
	MOF-177	Zn(II)	1,3,5-tris(4-carboxyphenyl)benzene	-	✓	COD	link
Catalysis	MIL-101(Cr)	Cr, Fe	Terephthalic acid (BDC)	-	✓	COD	link
	MIL-53(Fe)	Cr, Fe	Terephthalic acid (BDC)	-	✓	COD	link
	ZIF-67	Co(II)	Imidazolate	-	✓	COD	link
	MOF-5 (IRMOF-1)	Zn_4O	Terephthalic acid (BDC)	-	✓	COD	link
	MIL-68	In(III), Ga(III), Al(III)	Terephthalic acid (BDC)	-	✓	COD	link

F Atomic Collision Problem in Crystal Structure

Motivation. An important failure mode in generative crystal modeling is the atomic collision problem, where two atoms are placed unrealistically close, violating basic physical constraints (Pauli exclusion and electrostatic repulsion) [33, 36]. Such collisions typically render structures nonphysical or unstable and thus unsuitable for downstream use.

F.1 Problem Definition (with PBC)

Let $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^3$ be the Cartesian coordinates of atoms i and j in a unit cell with lattice matrix $\mathbf{L} = [\mathbf{a}, \mathbf{b}, \mathbf{c}] \in \mathbb{R}^{3\times3}$. Under periodic boundary conditions (PBC), we define the minimum-image distance between atoms i and j as

$$d_{\min}(i,j) \;=\; \min_{\mathbf{n}\in\{-1,0,1\}^3} \left\|\mathbf{x}_i - \left(\mathbf{x}_j + \mathbf{n}^\top \mathbf{L}\right)\right\|_2. \tag{4}$$

Let r_i, r_j be the effective allowable radii (see below). We flag a collision iff

$$d_{\min}(i,j) \;<\; r_i + r_j. \tag{5}$$

In practice, Eq. equation 4 enumerates the $3\times3\times3 = 27$ images, which is sufficient whenever $\max_{i,j}(r_i + r_j) < \min(\|\mathbf{a}\|, \|\mathbf{b}\|, \|\mathbf{c}\|)$.

Effective radii. Following our implementation, we prioritize tabulated covalent radii for triple bonds and fall back to double-bond values only when the triple-bond entry for an element is unavailable. If neither is available for an atomic number, the structure is marked as invalid for collision checking (no heuristic imputation).

F.2 Implementation Details

Given fractional coordinates, species (atomic numbers), and lattice parameters $(a, b, c, \alpha, \beta, \gamma)$, we first build a `pymatgen` `Structure` to obtain $\mathbf{L}$ and Cartesian $\{\mathbf{x}_k\}$. For each unordered pair (i, j) we:

1. enumerate $\mathbf{n} \in \{-1, 0, 1\}^3$ and compute $d_{\mathbf{n}} = \left\|\mathbf{x}_i - \left(\mathbf{x}_j + \mathbf{n}^\top \mathbf{L}\right)\right\|_2$;
2. take $d_{\min}(i, j) = \min_{\mathbf{n}} d_{\mathbf{n}}$ and the corresponding $\mathbf{n}^\star$;
3. compare $d_{\min}(i, j)$ with $r_i + r_j$ per Eq. equation 5.

We also classify collision pairs by the minimizing image: *same-cell* if $\mathbf{n}^\star = \mathbf{0}$, and *cross-cell* otherwise. This distinction is reported in all summaries.

F.3 Metrics

We report both structure-level summaries (for practitioners) and dataset-level rates.

Structure-level (per crystal). For a crystal with K atoms and $\binom{K}{2}$ unordered pairs, define

$$\#\text{CollPairs} \;=\; \sum_{1\le i<j\le K} \mathbb{I}\big(d_{\min}(i,j) < r_i + r_j\big), \tag{6}$$

$$\#\text{CrossCell} \;=\; \sum_{1\le i<j\le K} \mathbb{I}\big(d_{\min}(i,j) < r_i + r_j,\; \mathbf{n}^\star \neq \mathbf{0}\big), \tag{7}$$

$$\#\text{SameCell} \;=\; \#\text{CollPairs} - \#\text{CrossCell}, \tag{8}$$

$$\text{PLCR}_{\text{per}} \;=\; \frac{\#\text{CollPairs}}{\binom{K}{2}}. \tag{9}$$

We also return a boolean `HasCollision` $= \mathbb{I}(\#\text{CollPairs} > 0)$ and a list of collision details (pair of atomic numbers, $d_{\min}$, (i, j), and $\mathbf{n}^\star$).

Dataset-level (over N crystals). Let $\mathcal{S} = \{1, \ldots, N\}$ index crystals and let $\#\mathrm{Pairs}^{(s)} = \binom{K^{(s)}}{2}$. We aggregate:

$$\mathrm{MLCR} \;=\; \frac{1}{N} \sum_{s \in \mathcal{S}} \mathbb{I}\Big(\#\mathrm{CollPairs}^{(s)} > 0\Big) \quad (\%\text{ structures with any collision}), \tag{10}$$

$$\mathrm{PLCR}_{\mathrm{per}} \;=\; \frac{\sum_{s \in \mathcal{S}} \#\mathrm{CollPairs}^{(s)}}{\sum_{s \in \mathcal{S}} \#\mathrm{Pairs}^{(s)}} \quad (\text{pairwise collision ratio under PBC}), \tag{11}$$

$$\mathrm{CrossCell}\% \;=\; \frac{\sum_{s \in \mathcal{S}} \#\mathrm{CrossCell}^{(s)}}{\sum_{s \in \mathcal{S}} \#\mathrm{CollPairs}^{(s)}}, \quad \mathrm{SameCell}\% \;=\; 1 - \mathrm{CrossCell}\%. \tag{12}$$

We also report the absolute counts: total crystals, # with collisions, total collision pairs, and the cross-/same-cell breakdown.

Remarks. (i) Our $\mathrm{PLCR}_{\mathrm{per}}$ is a minimum-image metric—operationally equivalent to averaging over the 27 lattice images but counting each pair at most once using its minimizing image. (ii) Prioritizing triple-bond covalent radii makes the criterion conservative; falling back to double-bond values avoids undefined entries while keeping a consistent lower bound on allowable separations.

G Measuring Symmetry Learning Capabilities in LLMs

Large language models (LLMs) designed for materials science tasks should ideally respect fundamental physical invariances, such as translational symmetry in crystalline structures. To quantitatively evaluate this capability, we adopt the Increase in Perplexity under Transformation (IPT) metric, inspired by recent insights from CrystalTextLLM [20]. IPT measures how much a model's sequence likelihood changes under continuous group transformations, with smaller values indicating stronger invariance.

G.1 Definition of IPT

For a transformation group G with group elements g and group action t, the IPT for an input sequence s is defined as

$$\mathrm{IPT}(s) = \mathbb{E}_{g \in G}\left[\mathrm{PPL}(t_g(s))\right] - \mathrm{PPL}(t_{g^*}(s)),$$

where

$$\hat{g} = \arg\min_g \mathrm{PPL}(t_g(s)).$$

Here, $\mathrm{PPL}(s) = 2^{\mathrm{CE}(s)/n}$ is the exponentiated length-normalized cross-entropy loss, $\mathrm{CE}(s)$ is the cross-entropy, and n is the sequence length. The element g^* corresponds to the translation that yields the minimum perplexity for the given input.

G.2 Transformation Group and Implementation

In our setting, G represents the group of lattice translations in fractional coordinates. Each transformation t_g decodes the string representation of a crystal structure, translates its atomic coordinates by a fractional vector g (wrapping around under periodic boundary conditions), and re-encodes it back into the input format. The transformations are implemented using `pymatgen` [40], ensuring strict adherence to periodic boundary conditions.

G.3 Experimental Procedure

We compute IPT for each model as follows:

1. **Test set selection:** Randomly sample 500 crystal structures from the held-out test set.
2. **Transformation sampling:** For each structure, generate 20 random translation vectors g, each sampled uniformly from $[0, 1)$ per dimension in fractional coordinates.
3. **Perplexity computation:** For each g, apply t_g to obtain a transformed structure, and compute its perplexity $\mathrm{PPL}(t_g(s))$ using the target LLM.
4. **Normalization:** To prevent datapoints with inherently high perplexity from dominating the metric, we normalize IPT values by the mean perplexity over the sampled transformations for each structure.
5. **Aggregation:** Compute $\hat{g}$ as the translation yielding the lowest perplexity per structure, evaluate $\mathrm{IPT}(s)$, and then average over all test structures to obtain the model's final IPT score.

G.4 Additional Metrics: Percent Metastable

Alongside IPT, we also measure the Percent Metastable—the proportion of generated or transformed crystal candidates predicted to have formation energies below a given metastability threshold, as estimated by an independent property predictor. This serves as a complementary measure of physical plausibility.

G.5 Interpretation

Lower IPT values indicate that the model's likelihood estimates are more invariant under physically valid transformations, reflecting better internalization of translational symmetry.

H The Details of Experimental Setup

We observed that several public implementations of diffusion and hybrid models do not fix random seeds during sampling, leading to non-reproducible crystal structures. To ensure fair and reproducible evaluation, we fix the global seed to 42 in both training and sampling and set `torch.backends.cudnn.deterministic = True`. This forces cuDNN to use deterministic kernels for convolutions (forward/backward), certain reductions/normalizations (e.g., BatchNorm), and cuDNN RNNs—eliminating their non-determinism—but it does not cover cuBLAS/GEMM or other non-cuDNN operators, and may modestly slow down sampling.

H.1 CDVAE

Hyperparameters and Training Details. The total loss of CDVAE can be written as

$$\mathcal{L} = \underbrace{\lambda_c \mathcal{L}_c + \lambda_L \mathcal{L}_L + \lambda_N \mathcal{L}_N}_{\mathcal{L}_{AGG}} + \underbrace{\lambda_X \mathcal{L}_X + \lambda_A \mathcal{L}_A}_{\mathcal{L}_{DEC}} + \underbrace{\beta \mathcal{L}_{KL}}_{\mathcal{L}_{KL}}, \tag{13}$$

where $\mathcal{L}_c$ is the composition prediction loss, $\mathcal{L}_L$ is the lattice parameter prediction loss, $\mathcal{L}_N$ is the number-of-atoms prediction loss, $\mathcal{L}_X$ is the coordinate denoising loss, $\mathcal{L}_A$ is the atom-type denoising loss, and $\mathcal{L}_{KL}$ is the variational KL divergence regularization.

To keep each loss term at a similar scale, the coefficients are set as $\lambda_c = 1$, $\lambda_L = 10$, $\lambda_N = 1$, $\lambda_X = 10$, and $\lambda_A = 1$. The KL weight β is tuned among $\{0.01, 0.03, 0.1\}$, with $\beta = 0.01$ for Perov-5 and MP-20, and $\beta = 0.03$ for Carbon-24.

For noise scheduling, the number of noise levels is set to $L = 50$; atom-type noise standard deviation is sampled in the range $\sigma_A \in [0.01, 5]$, and coordinate noise standard deviation in $\sigma_X \in [0.01, 10]$.

During training, the initial learning rate is 0.001, decayed by a factor of 0.6 if the validation loss does not improve after 30 epochs, with a minimum learning rate of 0.0001. During generation, the step size is fixed at $\epsilon = 0.0001$, and Langevin dynamics is run for 100 steps at each noise level.

H.2 DiffCSP

Hyperparameters and Training Details. For DiffCSP, we adopt the following experimental setup. We use 4 layers and 256 hidden states for the Perov-5 dataset, and 6 layers with 512 hidden states for other datasets. The dimension of the Fourier embedding is set to $k = 256$. We apply a cosine scheduler with $s = 0.008$ to control the variance of the DDPM process on $\mathcal{L}_t$, and an exponential scheduler with $\sigma_1 = 0.005$ and $\sigma_T = 0.5$ to control the noise scale in the score matching process on $\mathcal{L}_f$. The diffusion step is set to $T = 1000$. Our model is trained for 3500, 4000, 1000, and 1000 epochs on Perov-5, Carbon-24, MP-20, and MPTS-52, respectively, with the same optimizer and learning rate schedule as CDVAE. For the step size γ in Langevin dynamics for the structure prediction task, we apply $\gamma = 5 \times 10^{-7}$ for Perov-5, 1×10^{-5} for MP-20 and MPSTS-52, and $\gamma = 5 \times 10^{-6}$ for Carbon-24 to predict a single sample. For the ab initio generation and optimization tasks on Perov-5, Carbon-24, and MP-20, we apply $\gamma = 1 \times 10^{-6}$, 1×10^{-5}, and 5×10^{-6}, respectively.

Table 13: The hyperparameters for training of DiffCSP in different datasets.

Datasets	Training Epochs	Number of Layers	Hidden Dimension
Perov-5	3500	4	256
Carbon-24	4000	6	512
MP-20	1000	6	512
MPTS-52	1000	6	512

Table 14: The step size γ in Langevin dynamics for different datasets.

Datasets	CSP	CSP-multi	De Novo
Perov-5	5×10^{-7}	5×10^{-7}	1×10^{-6}
Carbon-24	5×10^{-6}	5×10^{-7}	1×10^{-5}
MP-20	1×10^{-5}	1×10^{-5}	5×10^{-6}
MPTS-52	1×10^{-5}	1×10^{-5}	1×10^{-5}

H.3 DiffCSP++

For DiffCSP++, we follow the same data split as proposed in CDVAE [56] and DiffCSP [24]. For the implementation of the CSPML ranking models, we construct 100,000 positive and 100,000 negative pairs from the training set for each dataset to train a 3-layer MLP with 100 epochs and a 1×10^{-3} learning rate. To train the DiffCSP++ models, we train a denoising model with 6 layers, 512 hidden states, and 128 Fourier embeddings for each task and the training epochs are set to 3500, 4000, 1000, 1000, 1000 for Perov-5, Carbon-24, MP-20, and MPTS-52. The diffusion step is set to $T = 1000$. We utilize the cosine scheduler with $s = 0.008$ to control the variance of the DDPM process on k and A, and an exponential scheduler with $\sigma_1 = 0.005, \sigma_T = 0.5$ to control the noise scale on F. The loss coefficients are set as $\lambda_k = \lambda_F = 1, \lambda_A = 20$. We apply $\gamma = 2 \times 10^{-5}$ for Carbon-24, 1×10^{-5} for MPTS-52 and 5×10^{-6} for other datasets for the corrector steps during generation.

Table 15: The hyperparameters for training of DiffCSP in different datasets.

Datasets	Training Epochs	Number of Layers	Hidden Dimension
Perov-5	3500	6	512
Carbon-24	4000	6	512
MP-20	1000	6	512
MPTS-52	1000	6	512

Table 16: The step size γ in Langevin dynamics for different datasets.

Datasets	CSP	CSP-multi	De Novo
Perov-5	5×10^{-7}	-	1×10^{-6}
Carbon-24	5×10^{-6}	-	1×10^{-5}
MP-20	1×10^{-5}	-	5×10^{-6}
MPTS-52	1×10^{-5}	-	-

Table 17: The updated step size γ in Langevin dynamics for different datasets in original paper.

Datasets	CSP	CSP-multi	De Novo
Perov-5	5×10^{-6}	-	5×10^{-6}
Carbon-24	2×10^{-5}	-	2×10^{-5}
MP-20	5×10^{-6}	-	5×10^{-6}
MPTS-52	1×10^{-5}	-	1×10^{-5}

H.4 EquiCSP

For EquiCSP, we employ a 4-layer setting with 256 hidden states for Perov-5 and a 6-layer setting with 512 hidden states for other datasets. The dimension of the Fourier embedding is set to $k = 256$. We utilize the cosine scheduler with $s = 0.008$ to regulate the variance of the DDPM process on C_t, and an exponential scheduler with $\sigma_1 = 0.005, \sigma_T = 0.5$ to control the noise scale of the score matching process on F_t. The diffusion step is set to $T = 1000$. Our model undergoes training for 3500, 4000, 1000, and 1000 epochs respectively for Perov-5, Carbon-24, MP-20, and MPTS-52 using the same optimizer and learning rate scheduler as CDVAE. For Langevin dynamics' step size γ, we apply values of $\gamma = 5 \times 10^{-7}$ for Perov-5, $\gamma = 5 \times 10^{-6}$ for MP-20, $\gamma = 1 \times 10^{-5}$ for MPTS-52; while for *ab initio* generation in Carbon-24 case we use $\gamma = 1 \times 10^{-5}$.

Table 18: The hyperparameters for training of DiffCSP in different datasets.

Datasets	Training Epochs	Number of Layers	Hidden Dimension
Perov-5	3500	4	256
Carbon-24	4000	6	512
MP-20	1000	6	512
MPTS-52	1000	6	512

Table 19: The step size γ in Langevin dynamics for different datasets.

Datasets	CSP	CSP-multi	De Novo
Perov-5	5×10^{-7}	5×10^{-7}	1×10^{-6}
Carbon-24	5×10^{-6}	5×10^{-7}	1×10^{-5}
MP-20	5×10^{-6}	5×10^{-6}	5×10^{-6}
MPTS-52	1×10^{-5}	1×10^{-5}	1×10^{-5}

H.5 FlowMM

In this benchmark, the hyperparameter configuration of **FlowMM** is divided into four parts: general settings, network architecture, crystal structure prediction (CSP) task settings, and *de novo* generation (DNG) task settings.

General Hyperparameters. As shown in Table 1, the maximum number of atoms and training epochs vary across datasets to accommodate differences in data scale and structural complexity. Carbon-24 and Perov-5 are trained for 8000 and 6000 epochs, respectively, while the larger MP-20 and MPTS-52 datasets require only 2000 and 1000 epochs to avoid overfitting. The batch size is also dataset-dependent, e.g., Perov-5 employs a large batch size of 1024, whereas MPTS-52 is limited to 64 due to the larger unit cells.

Network Hyperparameters. As summarized in Table 2, FlowMM employs a six-layer architecture with a hidden dimension of 512 and a time embedding dimension of 256. The `silu` activation function is used throughout the network, and layer normalization is applied to improve training stability.

CSP Hyperparameters. For crystal structure prediction (Table 3), the learning rate decreases with increasing dataset complexity (0.001 for Carbon, 0.0001 for MP-20/MPTS-52). Weight decay is enabled for all datasets except Carbon to improve generalization. In the loss function, the fractional coordinate loss weight $\tilde{\lambda}_f$ is dataset-specific, with Perov-5 assigned the highest value (1500) to emphasize structural accuracy. The lattice loss weight λ_l is fixed to 1.0, while the anti-annealing slope s' is tuned per dataset to balance the optimization schedule.

DNG Hyperparameters. For the *de novo* generation task (Table 4), FlowMM is trained with a learning rate of 0.0005 and weight decay of 0.005 to encourage generative diversity. The loss function includes contributions from atom type ($\tilde{\lambda}_a = 300$), fractional coordinates ($\tilde{\lambda}_f = 600$), lattice ($\lambda_l = 1.0$), and cross-entropy ($\tilde{\lambda}_{ce} = 20$). To improve stability, annealing is enabled for fractional coordinates and lattice but not for atom types.

Table 20: General Hyperparameters

	Carbon	Perov	MP-20	MPTS-52
Max Atoms	24	20	20	52
Max Epochs	8000	6000	2000	1000
Total Number of Samples	10153	18928	45231	40476
Batch Size	256	1024	256	64

Table 21: Network Hyperparameters

	Value
Hidden Dimension	512
Time Embedding Dimension	256
Number of Layers	6
Activation Function	silu
Layer Norm	True

Table 22: CSP Hyperparameters

	Carbon	Perov	MP-20	MPTS-52
Learning Rate	0.001	0.0003	0.0001	0.0001
Weight Decay	0.0	0.001	0.001	0.001
$\tilde{\lambda}_f$ (Frac Coords)	400	1500	300	300
λ_l (Lattice)	1.0	1.0	1.0	1.0
s' (Anti-Anneal Slope)	2.0	1.0	10.0	5.0
Anneal f	False	False	True	True
Anneal l	False	False	False	False

Table 23: DNG Hyperparameters

	Value
Learning Rate	0.0005
Weight Decay	0.005
$\tilde{\lambda}_a$ (Atom Type)	300
$\tilde{\lambda}_f$ (Frac Coords)	600
λ_l (Lattice)	1.0
$\tilde{\lambda}_{\text{ce}}$ (Cross Entropy)	20
s' (Anti-Annealing Slope)	5.0
Anneal a	False
Anneal f	True
Anneal l	True

I More Comprehensive Results

I.1 More Comprehensive Results about CSP&DNG

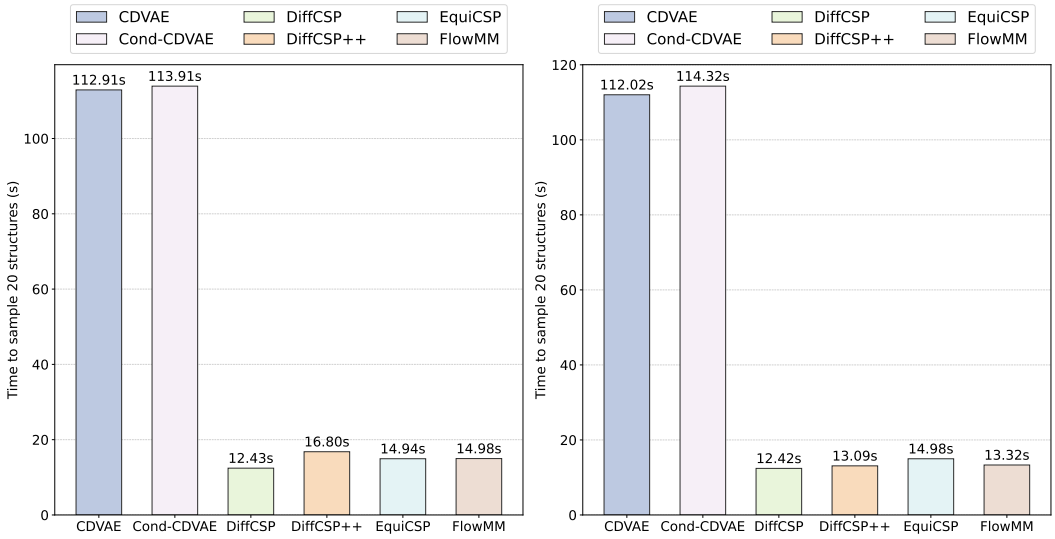
Building upon the official implementations of most benchmark models, we further report additional results. Random sampling is employed during both generation and evaluation, and the outcomes are presented for crystal structure prediction (CSP), *de novo* generation (DNG), and inference efficiency across various generative models.

Crystal Structure Prediction. Table 24 summarizes CSP performance under both single-sample and multi-sample (20) settings. Across datasets, diffusion-based approaches (e.g., DiffCSP, DiffCSP++) generally outperform VAE-based models in terms of match rate (MR). DiffCSP++ in particular achieves the highest MR across multiple datasets, while maintaining very low RMSE values. Increasing the number of samples consistently improves performance for all models, demonstrating the benefit of multiple candidate generations. CrystaLLM variants also exhibit competitive results, especially in terms of low RMSE, indicating strong local structural accuracy.

De Novo Generation. As shown in Table 25, nearly all models achieve close to 100% structural validity, confirming their ability to generate physically plausible materials. In terms of coverage (COV-R and COV-P), diffusion-based models maintain high scores above 97% across datasets. Property alignment metrics (d_ρ, d_{elem}) further highlight differences: some models, such as DiffCSP++ and EquiCSP, achieve particularly low deviations on specific datasets, indicating strong capability to preserve realistic material properties. On more complex datasets (e.g., MP-20), model performance varies more widely, reflecting challenges in generalization.

Inference Efficiency. Figure 4 compares inference time for generating 20 structures. Diffusion- and flow-based methods (DiffCSP, DiffCSP++, EquiCSP, FlowMM) complete sampling within ∼12–17 seconds, showing clear efficiency advantages. In contrast, VAE-based approaches (CDVAE, Cond-CDVAE) require over 110 seconds, making them significantly slower for large-scale generation. These results suggest that diffusion-based architectures are better suited for high-throughput applications where both speed and quality are critical.

Overall, the comprehensive experiments show that diffusion-based methods consistently provide strong performance in CSP and DNG tasks, with competitive accuracy, property preservation, and substantially faster inference compared to VAE baselines. Language-model approaches demonstrate promising structural precision, while traditional architectures still face trade-offs between accuracy, efficiency, and generalization across datasets.



(a) Inference time comparison on Carbon-24. (b) Inference time compartion on Perov-5.

Figure 4: Inference time comparition across different GNN architectures (random sampling).

Table 24: The benchmarking results (random sampling) on the crystal structure prediction task for diffusion-based models.

Method	# of samples	Perov-5		Carbon-24		MP-20		MPTS-52	
		MR (↑)	RMSE (↓)	MR (↑)	RMSE (↓)	MR (↑)	RMSE (↓)	MR (↑)	RMSE (↓)
Cond-CDVAE [35]	1	44.82	0.1178	16.89	0.2896	33.91	0.1098	5.29	0.2071
DiffCSP [24]	1	51.26	**0.0798**	17.93	0.2842	50.42	0.0650	12.46	0.1832
DiffCSP++ [25] (w/ CSPML)	1	**52.19**	0.0819	15.78	0.3276	**71.18**	**0.0276**	**35.98**	**0.0687**
EquiCSP [31]	1	43.45	0.1254	12.32	0.3212	43.21	0.1254	9.43	0.2340
FlowMM [37]	1	47.63	0.1087	15.92	0.2784	50.37	0.1168	8.32	0.2187
CrystaLLM-raw$_{(25M)}$	1	47.95	0.0966	21.13	**0.1687**	55.85	0.0437	17.47	0.1113
CrystaLLM$_{(25M)}$	1	45.65	0.0977	**21.87**	0.1734	56.58	0.0426	17.54	0.1028
CrystaLLM-raw$_{(200M)}$	1	46.10	0.0953	20.25	0.1761	58.70	0.0408	19.21	0.1110
CrystaLLM$_{(200M)}$	1	45.87	0.0970	20.64	0.1971	58.98	0.0416	18.97	0.1123
Cond-CDVAE [35]	20	88.25	0.0513	88.71	0.2252	67.08	0.0994	22.16	0.2107
DiffCSP [24]	20	98.24	**0.0127**	**89.00**	0.2207	**77.45**	0.0495	34.26	0.1741
DiffCSP++ [25] (w/ CSPML)	20	97.54	0.0132	85.43	0.2304	73.34	0.0576	**37.21**	0.1465
EquiCSP [31] (ours)	20	89.54	0.0543	82.31	0.2564	69.43	0.0853	28.76	0.2135
FlowMM [37] (ours)	20	88.15	0.0502	87.92	0.2325	68.01	0.1009	22.11	0.2052
CrystaLLM-raw$_{(25M)}$	20	98.26	0.0236	83.60	0.1523	75.14	0.0395	32.98	0.1197
CrystaLLM$_{(25M)}$	20	**98.34**	0.0228	84.04	0.1518	75.36	0.0398	32.96	0.1206
CrystaLLM-raw$_{(200M)}$	20	97.60	0.0249	85.17	**0.1514**	73.97	0.0349	33.75	0.1059
CrystaLLM$_{(200M)}$	20	97.73	0.0261	85.47	0.1542	74.11	**0.0345**	34.00	**0.1076**

Table 25: The benchmarking results (random sampling) on de novo generation task.

Dataset	Method	Validity (↑)		Coverage (↑)		Property (↓)	
		Struc.	Comp.	COV-R	COV-P	d_ρ	d_{elem}
Perov-5	CDVAE [56]	100.00	97.45	98.32	97.46	0.1500	0.0698
	Cond-CDVAE [35]	100.00	98.73	99.59	98.73	0.1412	0.0620
	DiffCSP [24]	100.00	98.85	99.74	98.27	0.1110	0.0128
	DiffCSP++ [25]	99.98	98.69	99.55	98.73	0.0674	0.0043
	EquiCSP [31]	100.0	98.72	99.74	**98.83**	0.1095	0.0489
	FlowMM [37]	100.00	98.85	99.62	98.81	**0.0659**	**0.0040**
Carbon-24	CDVAE [56]	100.00	–	99.86	83.12	0.1421	–
	Cond-CDVAE [35]	100.00	–	**99.92**	83.21	0.1418	–
	DiffCSP [24]	100.00	–	99.90	97.27	0.0805	–
	DiffCSP++ [25]	99.95	–	99.58	98.76	0.0312	–
	EquiCSP [31]	100.0	–	99.78	97.25	0.0721	–
	FlowMM [37]	99.98	–	99.66	**98.89**	**0.0298**	–
MP-20	CDVAE [56]	100.00	86.75	99.23	99.53	0.6832	1.4210
	Cond-CDVAE [35]	100.00	**86.82**	99.29	99.58	0.6838	1.4218
	DiffCSP [24]	100.00	83.25	**99.71**	**99.76**	0.350	-
	DiffCSP++ [25]	99.88	85.27	99.62	99.63	0.2389	**0.3721**
	EquiCSP [31]	100.0	82.45	99.70	99.74	**0.1278**	0.3942
	FlowMM [37]	96.85	83.19	99.49	99.58	0.239	-

I.2 Case Studies on OOD Generation

Table 26 reports the OOD evaluation results for CrystaLLM (25M) and CrystaLLM (200M) on four representative out-of-distribution datasets: double perovskites (OOD-DPC), antiperovskites (OOD-AC), layered perovskites (OOD-LPC), and huge unit cell carbon crystals (OOD-HUCC). *We focus on these two variants of CrystaLLM because they achieved state-of-the-art performance on the in-distribution CSP benchmarks, making them strong candidates for testing whether high in-distribution accuracy translates into robust generalization.* Surprisingly, despite their superior in-distribution results, both models completely failed to generate valid structures on all four OOD datasets. In every case, the match rate (MR) drops to 0.00 and the RMSE values are undefined (NaN), regardless of whether single or multiple samples (20) are generated.

This observation highlights a critical limitation of LLM-based approaches for crystallographic generation. While CrystaLLM is highly effective at learning the statistical patterns present in the training distribution (e.g., Perov-5, Carbon-24), it struggles to extrapolate beyond these domains to unseen structural families. The OOD datasets were intentionally constructed to probe such generalization: double perovskites introduce cation ordering complexity, antiperovskites invert the canonical anion–cation arrangement, layered perovskites (Ruddlesden–Popper phases) impose dimensional reduction and stacking variability, and huge unit cell carbons challenge the model with drastically larger structural scales.

The complete failure of CrystaLLM on these datasets suggests that its generative capability remains strongly distribution-bound, in contrast to diffusion-based models which often demonstrate partial transferability to related material families. This finding underscores the importance of explicitly evaluating OOD performance when assessing generative models for materials discovery, as in-distribution accuracy alone does not guarantee broader scientific utility. In future work, we plan to comprehensively evaluate all benchmarked models on their OOD generalization ability to provide a more complete understanding of their robustness.

Table 26: The OOD evluation results on the crystal structure prediction task for CrystaLLM (25M) and CrystaLLM (200M).

Method	**# of samples**	**OOD-DPC**		**OOD-AC**		**OOD-LPC**		**OOD-HUCC**	
		MR (↑)	RMSE (↓)	MR (↑)	RMSE (↓)	MR (↑)	RMSE (↓)	MR (↑)	RMSE (↓)
$\text{CrystaLLM}_{(25M)}$	1	0.00	Nan	0.00	Nan	0.00	Nan	0.00	Nan
$\text{CrystaLLM}_{(200M)}$	1	0.00	Nan	0.00	Nan	0.00	Nan	0.00	Nan
$\text{CrystaLLM}_{(25M)}$	20	0.00	Nan	0.00	Nan	0.00	Nan	0.00	Nan
$\text{CrystaLLM}_{(200M)}$	20	0.00	Nan	0.00	Nan	0.00	Nan	0.00	Nan

I.3 Evaluations on the Physical Plausibility Problem

Experiment Analysis. We benchmarked representative diffusion-based models under a fixed global seed (42) to assess the prevalence of atomic collisions during crystal structure prediction. Several key observations emerge:

(1) *Dataset complexity effect.* Collision rates increase markedly with dataset difficulty. On Perov-5 and MP-20, the proportion of collided structures remains below $\sim$12%, whereas on MPTS-52 more than one-third of generated crystals contain overlapping atoms. This confirms that large and chemically diverse unit cells exacerbate steric violations.

(2) *Model family differences.* On Perov-5, all tested models yield comparable collision rates ($\sim$9–10%), but differences become clearer on MP-20: DiffCSP++ variants achieve lower collision incidence ($\sim$7.3–7.4%) relative to DiffCSP (10.5%) and EquiCSP (11.9%). Notably, DiffCSP++ (with CSPML) produces fewer collided structures but accumulates the largest number of total collision pairs, suggesting that when collisions occur, they can be more severe. On MPTS-52, both DiffCSP and EquiCSP show very high collision rates ($\sim$35–37%), underscoring the challenges of complex systems.

(3) *Cross-cell vs. same-cell breakdown.* Across all datasets, collision pairs are split relatively evenly between same-cell and cross-cell cases on Perov-5 and MP-20 (roughly 45–55%). However, on MPTS-52, same-cell collisions dominate (70% or more), indicating that resolving local steric clashes within the unit cell is the primary bottleneck at scale.

Overall, these results demonstrate that the atomic collision problem is a persistent failure mode in generative crystal modeling, with severity strongly dependent on dataset complexity and architectural choices. Evaluating collision metrics alongside traditional accuracy measures provides an essential complementary perspective on the physical plausibility of generated materials. For completeness, we note that DiffCSP++ (w/ GT) results on MPTS-52 are omitted due to prohibitive resource requirements during sampling, which made large-scale evaluation impractical.

Table 27: The atomic collision benchmarking results on crystal structure prediction task (global seed).

Dataset	Method	# Crystals	Collided (↓)	Collision Rate (↓)	# Collision Pairs (↓)	Cross-cell (↓)	Same-cell (↓)
Perov-5	DiffCSP [24]	3785	350	9.25%	769	364 (47.33%)	405 (52.67%)
	DiffCSP++ [25] (w/ GT)	3785	375	9.91%	860	198 (23.02%)	662 (76.98%)
	DiffCSP++ [25] (w/ CSPML)	3785	376	9.93%	944	251 (25.59%)	693 (73.41%)
	EquiCSP [31]	3785	367	9.70%	835	444 (53.17%)	391 (46.83%)
MP-20	DiffCSP [24]	9046	945	10.45%	2912	1434 (49.24%)	1478 (50.76%)
	DiffCSP++ [25] (w/ GT)	9046	664	7.34%	4848	2103 (43.38%)	2745 (56.62%)
	DiffCSP++ [25] (w/ CSPML)	9046	669	7.40%	9103	4089 (44.92%)	5014 (55.08%)
	EquiCSP [31]	9046	1079	11.93%	3036	1494 (48.73%)	1572 (51.27%)
MPTS-52	DiffCSP [24]	8096	3008	37.15%	13818	4426 (32.03%)	9392 (67.97%)
	DiffCSP++ [25] (w/ GT)	-	-	-	-	-	-
	DiffCSP++ [25] (w/ CSPML)	-	-	-	-	-	-
	EquiCSP [31]	8096	2875	35.51%	12138	3628 (29.89%)	8510 (70.11%)

J Prompts of LLMs

Large language models (LLMs) require carefully designed prompts to ensure consistent and structured outputs for scientific applications. In our experiments, we employed Llama 3.1 (8B) to generate crystallographic information files (CIFs) directly from chemical formulas. To achieve reliable results, the prompts explicitly instruct the model to adhere to the standard CIF format, fill in necessary structural fields, and avoid producing any extraneous text. This section provides the exact prompts used in our study, which were crafted to enforce strict formatting rules and to guarantee that the generated outputs are both syntactically valid and physically meaningful.

```
You are an expert in generating crystallographic data in structured text format.
↪  Your task is to output a single, clean CIF block for the given formula:
↪  {formula}.

Use the exact format below. Fill in all fields once only-**do not repeat any
↪  section**. Output **only** the CIF block. No markdown, no explanations, no
↪  formatting, no comments, no dash line, no extra text of any kind.

Strictly follow this structure:

_symmetry_space_group_name_H-M   ?
_cell_length_a   ?
_cell_length_b   ?
_cell_length_c   ?
_cell_angle_alpha   ?
_cell_angle_beta   ?
_cell_angle_gamma   ?
_symmetry_Int_Tables_number   ?
_chemical_formula_structural   ?
_chemical_formula_sum   ?
_cell_volume   ?
_cell_formula_units_Z   ?

loop_
_symmetry_equiv_pos_site_id
_symmetry_equiv_pos_as_xyz
[id]  '[x, y, z]'
...

loop_
_atom_site_type_symbol
_atom_site_label
_atom_site_symmetry_multiplicity
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_occupancy
[symbol]  [label]  [multiplicity]  [x]  [y]  [z]  [occupancy]
[symbol]  [label]  [multiplicity]  [x]  [y]  [z]  [occupancy]
...

Instructions:
- Fill in all question marks (`?`) with reasonable, physically consistent values
↪  inferred from the given chemical formula.
- The first `loop_` section must contain symmetry equivalent position IDs and
↪  operations (in xyz format).
- The second `loop_` section must list all atom sites present in the formula,
↪  including their element symbol, label, symmetry multiplicity, fractional
↪  coordinates (x, y, z), and occupancy.
- Output a clean CIF block only, with no duplication or extra content.
- All output must follow this structure precisely. Do **not** include notes,
↪  hints, explanations, or any formatting outside the CIF structure block.

Notes: Only output the CIF block, no any other reply.
```

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are consistent with the theoretical contributions, method design, and extensive experimental results presented throughout the paper, including the development of the Material Generation Benchmark (MGB) for evaluating deep generative models in materials science. The claims regarding the evaluation of models across various tasks like crystal structure prediction and MOF prediction are well-supported by the experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations and outlines potential future work of the proposed benchmark, including the incorporation of methods for material geometry modeling.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient details on datasets, model architectures, evaluation metrics, and experimental setups to allow reproduction of the main results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper states that the detailed code implementation will be open sourced recently.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All necessary training and test details, including data splits, hyperparameters, optimizer choices, and evaluation metrics, are specified in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The results are reported with using global seed, and the method for calculating them is described in the experimental section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The type of compute resources (e.g., GPU) and experimental settings are discussed in the appendix.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. All data used are publicly available and cited appropriately. No personally identifiable or sensitive data is involved.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The work focuses on foundational algorithmic research for materials generation and does not have a direct societal impact. The paper does not focus on any specific application scenarios.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The models and data used in this work pose no particular risk for misuse; no high-risk assets are released.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and codebases used are publicly available, properly cited, and used according to their respective licenses. Details are included in Section 5.1 and the references.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new datasets or code assets are introduced beyond the model implementation; no new dataset is released.

    Guidelines:

    - The answer NA means that the paper does not release new assets.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This research does not involve human subjects or crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Not applicable; there are no experiments involving human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

Answer: [NA]

Justification: No large language model is used as an important or original component of the core methodology; LLMs may only have been used for minor writing/editing assistance.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.