FEW-STEP CRYSTAL GENERATION VIA COMPOSITION-CONDITIONED MOMENT MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep generative models offer a scalable route to explore the vast composition–structure search space underlying materials discovery. Existing diffusion- and flow-based crystal generators, however, still rely on long iterative sampling trajectories, limiting their use in high-throughput settings where large candidate libraries must be generated efficiently. We propose COMMAT (**Co**mposition-conditioned **M**oment matching for **Mat**erial generation), a few-step crystal generation framework that separates chemical composition modeling from geometric refinement. Given a composition, COMMAT obtains a composition-conditioned structural latent, predicts the lattice once, and refines only atomic coordinates while keeping species and lattice fixed, reducing iterative generation to a continuous geometric transport problem. To solve this problem in only a few network evaluations, we adapt inductive moment matching to periodic crystal geometry through a lattice-normalized Cartesian transport space coupled with periodic coordinate evaluation in wrapped fractional coordinates. Conditioning the structural prior on composition eliminates the need to encode target geometry, enabling a unified few-step generator for both crystal structure prediction and de novo generation. Across the MP-20, Perov-5, Carbon-24, and MPTS-52 benchmarks, COMMAT retains strong generation quality with only 2–8 neural function evaluations. Under matched few-step budgets, it achieves favorable CSP and de novo generation performance relative to recent diffusion- and flow-based baselines, with particularly strong results at the lowest sampling budgets. At its 32-NFE setting, COMMAT remains competitive with much higher-budget baselines while substantially reducing inference cost. These results demonstrate a practical route toward scalable few-step crystal generation.

1 INTRODUCTION

Crystalline materials are central to technologies in energy storage, catalysis, electronics, and quantum devices, but discovering new materials requires searching a vast space of possible compositions and structures (Schmidt et al., 2019). Generative models provide a practical way to explore this space by proposing candidate crystals that can later be relaxed, screened for stability, and evaluated for useful properties (Noh et al., 2019; Xie et al., 2021). For high-throughput screening, however, generation itself must be inexpensive.

Following prior works on diffusion models (Ho et al., 2020; Song et al., 2021) most crystal generators construct structures through a long sequence of iterative refinement. In practice, crystal generation must simultaneously generate three different components: atom types, lattice parameters, and fractional coordinates. These variables follow different distributions but are strongly coupled through chemical composition and crystal geometry (Jiao et al., 2023; Miller et al., 2024). This coupling favors iterative refinement, allowing composition and geometry to become progressively consistent over multiple updates (Zeni et al., 2025). As a result, many crystal generators rely on relatively long sampling trajectories. For example, CDVAE uses annealed Langevin dynamics (Xie et al., 2021), DiffCSP uses diffusion over periodic crystal variables (Jiao et al., 2023), and CrysBFN uses an iterative Bayesian flow process (Wu et al., 2025). These long sampling trajectories incur substantial inference cost, which becomes a practical bottleneck when generating large candidate pools for high-throughput materials screening (Ye et al., 2025; Metni et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recent crystal generators (Höllmer et al., 2025; Morehead et al., 2026; Joshi et al., 2025; Veljković et al., 2026) especially flow-based models (Miller et al., 2024; Luo et al., 2025) have reduced this sampling burden to hundreds of integration steps. In contrast, generative models in other domains, particularly computer vision, can produce high-quality samples in as few as one or two steps (Song & Dhariwal, 2024; Zhou et al., 2025; Geng et al., 2026). Yet such very-low-step generation remains largely unexplored for crystal generation. This motivates asking whether crystal generators can similarly operate in the *few-step sampling regime*. We therefore propose CoMMAT and evaluate both DNG and CSP in this regime, while also examining how existing sampling-efficient crystal generators behave when their inference trajectories are similarly compressed. Concretely, we focus on 2–8 network function evaluations (NFEs), testing how effectively each method operates when iterative refinement is severely constrained.

Figure 1 illustrates this very low-budget regime on MP-20. While several diffusion- and flow-based baselines experience substantial performance degradation when their sampling trajectories are shortened, CoMMat remains stable with only a few network evaluations.

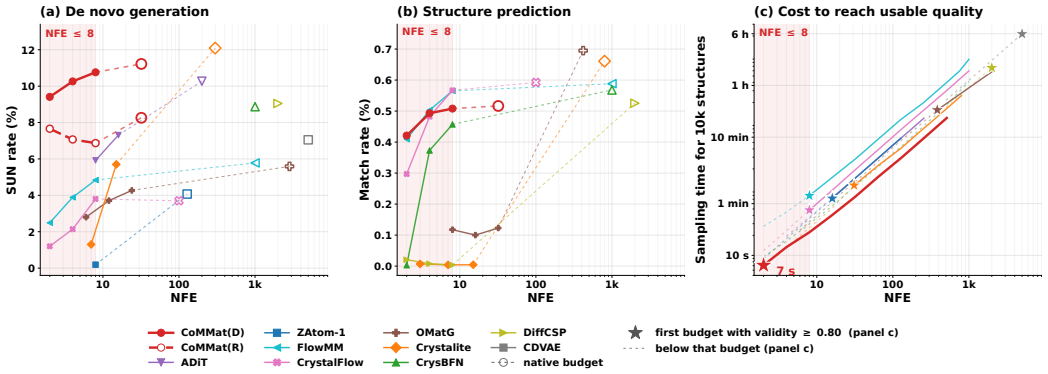

Figure 1: Few-step quality and efficiency on MP-20. (a) SUN rate for de novo generation and (b) CSP match rate as a function of NFE. The shaded region marks the few-step regime (NFE $\leq$ 8); solid curves show evaluated low-budget points, and dashed segments extend to each method's native sampling budget. (c) Sampling time for 10k structures versus the minimum NFE required to reach validity ≥ 0.80; stars mark the first budget satisfying this threshold. Timing is measured on a single NVIDIA A30 GPU (FP32).

Our proposed CoMMAT is a few-step crystal generator that separates composition modeling from structural refinement that we describe in Fig 2. A causal Transformer models crystal compositions and provides a composition-conditioned representation, while a periodic CSPNet decoder performs an MMD-based distributional moment-matching (Zhou et al., 2025) only over the atomic coordinates, with atom identities kept fixed during structural generation. The same model supports both crystal structure prediction (CSP) and de novo generation (DNG): CSP conditions directly on a supplied composition, whereas DNG first generates a composition. In both settings, the composition determines the lattice and conditions the same few-step coordinate sampler.

Across MP-20, MPTS-52, Perov-5 and, carbon-24 CoMMAT retains competitive crystal-generation quality at only 2–8 NFEs, while substantially reducing inference cost relative to longer diffusion- and flow-based sampling procedures.

Our main contributions are:

- We introduce CoMMAT, a composition-conditioned few-step crystal generator that decouples composition from geometry, predicts the lattice in one shot, and iteratively refines only atomic coordinates.
- We formulate coordinate moment matching for periodic crystal geometry through lattice-normalized Cartesian transport with coordinate-field evaluation in wrapped fractional space. in a lattice-normalized Cartesian transport space
- Across four crystal benchmarks, CoMMAT remains effective at only 2–8 NFEs under matched few-step evaluation, while its 32-NFE setting remains competitive with reference configurations requiring hundreds to thousands of evaluations.

2 Related Work

Generative models for crystalline materials. CDVAE (Xie et al., 2021) introduced a latent-variable framework with a periodic graph decoder that jointly refines atomic coordinates and species through diffusion, while SyMat (Luo et al., 2023) incorporated explicit symmetry-aware modeling into periodic material generation. DiffCSP (Jiao et al., 2023) subsequently formulated crystal structure prediction (CSP) as joint equivariant diffusion over fractional coordinates and lattice variables. More recent approaches increasingly model multiple crystal modalities include atomic species, fractional coordinates, and lattice variables within a unified generative process. CoMMAT retains a latent-conditioned periodic GNN decoder interface, allowing us to isolate the effect of the generative objective and sampling procedure without relying on a substantially larger backbone.

Few-step diffusion, flow, and moment matching. FlowMM (Miller et al., 2024) and CrystalFlow (Luo et al., 2025) replace long diffusion trajectories with flow-based transport over crystal variables, while OMatG (Höllmer et al., 2025) combines stochastic interpolants for continuous crystal variables with discrete flow matching for atomic species. Recent approaches such as Crystalite (Veljković et al., 2026), ADiT (Joshi et al., 2025), and Zatom-1 (Morehead et al., 2026) further improve sampling efficiency through lightweight diffusion, latent diffusion, and multimodal flow-matching formulations, respectively. In broader generative modeling, inductive moment matching (IMM) (Zhou et al., 2025) learns finite-step generative mappings by matching student and teacher pushforward distributions defined along stochastic-interpolant (Albergo et al., 2025) trajectories. CoMMAT applies this idea to periodic crystal coordinates in a scale-normalized representation. Consequently, it conditions on the known or generated composition and applies moment matching only to coordinate refinement.

3 Preliminaries

Crystal Representation and Generation Tasks. A crystal with N atoms in the unit cell is represented as $\mathcal{C} = (\mathbf{A}, \mathbf{X}, \mathbf{L})$, with species $\mathbf{A} = (A_1, \ldots, A_N)$, fractional coordinates $\mathbf{X} \in [0,1)^{N\times 3}$, and a nonsingular lattice matrix $\mathbf{L} \in \mathbb{R}^{3\times 3}$; Cartesian coordinates are $\mathbf{R} = \mathbf{X}\mathbf{L}$ under the row-vector convention. We call the multiset of species in $\mathbf{A}$ the *composition*. Fractional coordinates are defined modulo integer shifts (i.e., lattice translations), and a crystal distribution should be invariant to atom permutations, global orthogonal transformations, and periodic translations (Jiao et al., 2023). Both tasks target the conditional distribution $p(\mathbf{L}, \mathbf{X} \mid \mathbf{A})$: *crystal structure prediction* (CSP) generates structures for a given composition, whereas *de novo generation* (DNG) additionally samples the composition $\mathbf{A} \sim p(\mathbf{A})$.

Inductive Moment Matching. Inductive Moment Matching (IMM) (Zhou et al., 2025) learns one- or few-step generators without a pretrained model. An interpolant $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}$ between data $\mathbf{x}_0 \sim q_0$ and noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ defines marginals q_t, with larger t more corrupted. For $s \le t$, IMM learns a map $f_\theta(\mathbf{x}_t; t \to s)$ whose pushforward of q_t matches q_s. Training uses ordered times $s \le r \le t$, with r slightly below t: a student branch starts from $\mathbf{x}_t$ and a stop-gradient teacher branch from the less corrupted $\mathbf{x}_r$ on the same trajectory,

$$\mathbf{u}_{st} = f_\theta(\mathbf{x}_t; t \to s), \qquad \mathbf{u}_{sr} = \mathrm{sg}[f_\theta(\mathbf{x}_r; r \to s)]\,. \tag{1}$$

The empirical distributions P_{st} and P_{sr} of these outputs, computed over a group of particles (independent samples sharing the same (s, r, t)), are matched with maximum mean discrepancy,

$$\mathcal{L}_{\mathrm{IMM}} = \mathbb{E}_{t,r,s}\Big[w_{\mathrm{out}}(t,s)\ \mathrm{MMD}^2_{k_{t,s}}(P_{st}, P_{sr})\Big], \quad \mathrm{MMD}^2_k(P,Q) = \mathbb{E}_{P,P}[k] + \mathbb{E}_{Q,Q}[k] - 2\,\mathbb{E}_{P,Q}[k], \tag{2}$$

where w_{out} is a time-dependent weight and $k_{t,s}$ a time-dependent kernel. Using the shorter jump as a stop-gradient teacher progressively bootstraps longer mappings from shorter ones, enabling few-step generation at inference. The standard IMM formulation is developed for fixed-dimensional continuous Euclidean data, whereas crystals combine discrete species with periodic coordinates whose Euclidean geometry depends on $\mathbf{L}$ and have varying numbers of atoms; Section 4 addresses these differences.

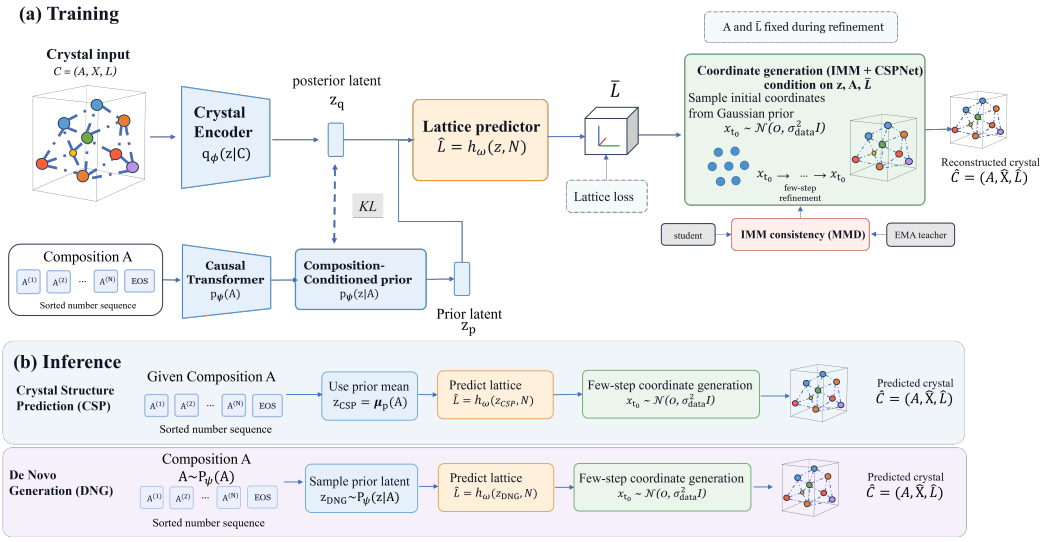

Figure 2: Overview of CoMMat. During training, a crystal encoder and a composition-conditioned prior provide posterior and prior structural latents that share a one-shot lattice predictor and a few-step periodic coordinate generator. At inference, CSP uses the prior mean for a given composition, while DNG samples both the composition and structural latent; both then predict the lattice once and refine atomic coordinates in only a few network evaluations.

4 METHOD

We introduce COMMAT, a composition-conditioned framework for few-step crystal generation. Our central design choice is to separate *discrete chemical composition* from *continuous crystal geometry*. A causal Transformer models the composition and parameterizes a composition-conditioned structural prior. A latent drawn from this prior determines the lattice through a one-shot predictor, while atomic coordinates are generated by adapting Inductive Moment Matching (IMM) to periodic crystal geometry. Throughout coordinate refinement, the composition and lattice remain fixed.

Using the notation of Section 3, the de novo generative model factorizes as

$$p(\mathcal{C}) = p_\psi(\mathbf{A}) \int p_\psi(\mathbf{z} \mid \mathbf{A})\, \delta(\mathbf{L} - h_\omega(\mathbf{z}, N))\, p_\theta(\mathbf{X} \mid \mathbf{A}, \mathbf{z}, \mathbf{L})\, d\mathbf{z}, \tag{3}$$

where $\delta(\cdot)$ denotes a deterministic lattice prediction and $N = |\mathbf{A}|$. For crystal structure prediction (CSP), the composition $\mathbf{A}$ is observed rather than sampled.

4.1 COMPOSITION-CONDITIONED STRUCTURAL LATENT

Posterior and conditional prior. During training, a crystal encoder infers a structural latent from the complete crystal, while a composition-conditioned prior predicts the corresponding latent distribution without access to the target geometry:

$$\begin{aligned} q_\phi(\mathbf{z} \mid \mathcal{C}) &= \mathcal{N}\big(\boldsymbol{\mu}_q(\mathcal{C}), \operatorname{diag}\left[\boldsymbol{\sigma}_q^2(\mathcal{C})\right]\big), \\ p_\psi(\mathbf{z} \mid \mathbf{A}) &= \mathcal{N}\big(\boldsymbol{\mu}_p(\mathbf{A}), \operatorname{diag}\left[\boldsymbol{\sigma}_p^2(\mathbf{A})\right]\big). \end{aligned} \tag{4}$$

We align the two distributions through $\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathcal{C}) \| p_\psi(\mathbf{z} \mid \mathbf{A}))$. Consequently, structural latents can be obtained from composition alone at inference, without encoding a target structure.

Autoregressive composition model. The same causal Transformer used to construct the conditional prior also models de novo compositions. We represent each composition canonically by sorting atomic numbers in non-decreasing order, $A_{(1)} \le \cdots \le A_{(N)}$, while retaining repeated elements to preserve stoichiometric multiplicity. The autoregressive model is

$$p_\psi(\mathbf{A}) = \left[\prod_{i=1}^{N} p_\psi\big(A_{(i)} \mid A_{(<i)}\big)\right] p_\psi\big(\mathrm{EOS} \mid A_{(1:N)}\big). \tag{5}$$

The canonical ordering removes redundant permutations of the same composition, while the End of Sequence (EOS) token determines the number of atoms N. The representation of the completed sequence additionally parameterizes $\boldsymbol{\mu}_p(\mathbf{A})$ and $\boldsymbol{\sigma}_p^2(\mathbf{A})$, coupling composition generation and structural conditioning through a shared chemical representation.

Prior-sampled structural training. KL alignment alone does not ensure that the structural generator remains accurate for prior-sampled latents encountered at inference. We therefore train two structural paths. The primary path uses $\mathbf{z}_q \sim q_\phi(\mathbf{z} \mid \mathcal{C})$, while A second path draws $\mathbf{z}_p \sim p_\psi(\mathbf{z} \mid \mathbf{A})$ and applies the same lattice and coordinate objectives using a prior-sampled latent. The two paths share the lattice and coordinate modules. This explicitly exposes the structural generator to the latent distribution used at inference.

4.2 One-Shot Lattice Prediction

Unlike the coordinates, the lattice is not iteratively refined. We represent it using the three lengths (a, b, c) and angles (α, β, γ), normalize each length by $N^{1/3}$, and standardize the resulting six-dimensional target. A deterministic head predicts

$$\hat{\mathbf{y}}_L = h_L(\mathbf{z}), \qquad \hat{\mathbf{L}} = h_\omega(\mathbf{z}, N), \tag{6}$$

where $h_\omega(\mathbf{z}, N)$ denotes inverse standardization, restoration of the $N^{1/3}$ length scale, and conversion of the predicted lengths and angles into a lattice matrix.

During an initial training warm-up, coordinate prediction is conditioned on the ground-truth lattice to stabilize structural learning. Thereafter, the posterior and prior-sampled paths use lattices predicted from $\mathbf{z}_q$ and $\mathbf{z}_p$, respectively. At inference, only the predicted lattice is used, and it remains fixed throughout coordinate generation.

4.3 Composition-Conditioned Coordinate Moment Matching

Given composition $\mathbf{A}$, structural latent $\mathbf{z}$, and conditioning lattice $\bar{\mathbf{L}}$ (ground truth during warm-up, predicted otherwise), the coordinate module generates atomic positions while treating species as fixed conditioning information. Atomic species are treated as fixed conditioning variables throughout both student and teacher branches; no categorical state is corrupted or refined. Coordinate generation is therefore decoupled from atom-type refinement.

Lattice-normalized Cartesian state. Euclidean distances in fractional coordinates depend on the lattice metric, whereas Cartesian coordinates retain absolute cell scale of the unit cell. We therefore define the IMM trajectory in a per-structure lattice-normalized Cartesian space.

We first define the centered periodic representative and characteristic lattice scale as

$$\mathbf{X}^c = \left(\mathbf{X} + \frac{1}{2}\right) \bmod 1 - \frac{1}{2}, \qquad \ell(\bar{\mathbf{L}}) = \frac{1}{3}\sum_{j=1}^{3} \left\|\bar{\mathbf{l}}_j\right\|_2, \tag{7}$$

where $\bar{\mathbf{l}}_j$ denotes the j-th lattice vector of $\bar{\mathbf{L}}$. The forward and inverse coordinate maps are

$$\begin{aligned} \mathbf{x}_0 &= \frac{\mathbf{X}^c \bar{\mathbf{L}}}{\ell(\bar{\mathbf{L}})}, \\ \mathbf{X}_u &= \left[\ell(\bar{\mathbf{L}})\, \mathbf{x}_u\, \bar{\mathbf{L}}^{-1}\right] \bmod 1. \end{aligned} \tag{8}$$

Centering fixes a periodic representative for the Euclidean corruption process and normalization makes the coordinate scale comparable across crystals with different cell sizes. The IMM state is maintained in this normalized Cartesian representation, but periodicity is retained at the network input: before each CSPNet evaluation, the state is mapped back to wrapped fractional coordinates using Eq. (8). CSPNet therefore constructs its periodic graph using the physical lattice $\bar{\mathbf{L}}$, even though corruption and moment matching are performed in the normalized Cartesian state space.

Coordinate corruption. We instantiate IMM using a linear flow-matching bridge. For an ordered training triplet $s \le r \le t$,

$$\mathbf{x}_u = (1-u)\mathbf{x}_0 + u\,\sigma_{\text{data}}\boldsymbol{\epsilon}, \qquad u \in \{r, t\}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{9}$$

The same Gaussian realization is used for $\mathbf{x}_t$ and $\mathbf{x}_r$, placing the student and teacher states on the same corruption trajectory.

Periodic coordinate field. At source time u, CSPNet receives the wrapped fractional coordinates $\mathbf{X}_u$, fixed species $\mathbf{A}$, conditioning lattice $\bar{\mathbf{L}}$, structural latent $\mathbf{z}$, and source time u. The decoder predicts a fractional-coordinate field $\Delta\mathbf{X}_\theta$, which we express in the normalized Cartesian transport space as

$$\mathbf{F}_\theta\big(\mathbf{x}_u, u \mid \mathbf{A}, \mathbf{z}, \bar{\mathbf{L}}\big) = \frac{\Delta\mathbf{X}_\theta\big(\mathbf{X}_u, u \mid \mathbf{A}, \mathbf{z}, \bar{\mathbf{L}}\big)\,\bar{\mathbf{L}}}{\ell(\bar{\mathbf{L}})}. \tag{10}$$

The corresponding finite-time Euler pushforward from source time u to target time $s \le u$ is

$$f_\theta(\mathbf{x}_u; u \to s) = \mathbf{x}_u - (u - s)\,\sigma_{\mathrm{data}}\,\mathbf{F}_\theta\big(\mathbf{x}_u, u \mid \mathbf{A}, \mathbf{z}, \bar{\mathbf{L}}\big)\,. \tag{11}$$

The student and stop-gradient teacher pushforwards are defined in Eq. (1) and optimized the moment-matching objective in Eq. (2). For crystals, the empirical distributions are constructed from groups of M structures sharing the same (s, r, t) triplet. We use a time-scaled Laplace kernel in the normalized Cartesian space with a V-statistic MMD estimator. Variable-size structures are padded within each group and augmented with an occupancy mask that distinguishes physical atoms from padded sites. Kernel scaling, time sampling, clipping, and group-size details are provided in Appendix D.

4.4 Training Objective

For a structural latent $\mathbf{z}$, we combine coordinate moment matching with one-shot lattice regression:

$$\mathcal{L}_{\mathrm{str}}(\mathbf{z}) = \lambda_x \mathcal{L}_{\mathrm{IMM}}(\mathbf{z}) + \lambda_L \mathcal{L}_L(\mathbf{z}). \tag{12}$$

Joint training of the posterior and prior-sampled paths uses

$$\mathcal{L} = \mathcal{L}_{\mathrm{str}}(\mathbf{z}_q) + \lambda_p \mathcal{L}_{\mathrm{str}}(\mathbf{z}_p) + \beta \mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{\mathrm{aux}}, \tag{13}$$

where $\mathcal{L}_{\mathrm{aux}}$ collects the autoregressive composition, atom-count, and latent-level composition objectives. The KL weight β is warmed up during training; full loss definitions and coefficients are given in the Appendix C.

4.5 Few-Step Crystal Generation

Once $(\mathbf{A}, \mathbf{z})$ are specified, CSP and de novo generation use the same geometric sampler. The lattice $\hat{\mathbf{L}}$ is predicted once using Eq. (6) and held fixed throughout coordinate refinement.

Coordinates are initialized from the terminal Gaussian prior and updated over a decreasing grid $1 = t_0 > t_1 > \cdots > t_K = t_{\min}$:

$$\begin{aligned} \mathbf{x}_{t_0} &\sim \mathcal{N}\big(\mathbf{0}, \sigma_{\mathrm{data}}^2 \mathbf{I}\big)\,, \\ \mathbf{x}_{t_{k+1}} &= \mathbf{x}_{t_k} - (t_k - t_{k+1})\sigma_{\mathrm{data}}\,\mathbf{F}_\theta\Big(\mathbf{x}_{t_k}, t_k \mid \mathbf{A}, \mathbf{z}, \hat{\mathbf{L}}\Big)\,. \end{aligned} \tag{14}$$

Atomic species remain fixed over the full trajectory. Each refinement step requires one coordinate-decoder evaluation, so K steps correspond exactly to K coordinate NFEs.

Crystal structure prediction. Given composition $\mathbf{A}$, CSP uses the conditional-prior mean $\mathbf{z}_{\mathrm{CSP}} = \boldsymbol{\mu}_p(\mathbf{A})$. The target geometry is never encoded; stochasticity arises only from the initial coordinate noise.

De novo generation. DNG first samples $\mathbf{A} \sim p_\psi(\mathbf{A})$ and $\mathbf{z}_{\mathrm{DNG}} \sim p_\psi(\mathbf{z} \mid \mathbf{A})$, then applies the same lattice predictor and coordinate sampler. DNG is therefore unconditional at the crystal level while internally decomposing generation into composition sampling and composition-conditioned geometric generation.

Table 1: Few-step and native-budget DNG on MP-20 and MPTS-52. Few-step results are ordered by the sampling budgets shown in parentheses after each model, which give the corresponding NFE ladder. Metric triplets follow the same order. The native block reports each model at its standard inference budget. Best and second-best results within each dataset and checkpoint are shown in **bold** and underlined, respectively.

	Few-step regime				Native budget			
Model (NFE : $p_1/p_2/p_3$)	Val.↑	C-R↑	$\rho\downarrow$	SUN↑	NFE	Val.↑	SUN↑	Time↓
MP-20								
CoMMat(D) (2/4/8)	79.51/79.65/80.20	**98.75/99.26/99.57**	.251/.212/**.107**	**9.41/10.26/10.76**	32	81.23	11.22	**76 s**
CoMMat(R) (2/4/8)	**86.98/87.12/87.46**	98.70/99.20/99.50	**.235/.133**/.195	7.66/7.07/6.87	32	87.85	8.26	**76 s**
CrystalFlow (2/4/8)	30.93/76.15/81.58	42.10/69.94/94.52	1.83/1.87/1.18	1.20/2.14/3.80	100	81.24	3.70	591.8 s
ZAtom-1 (2/4/8)	22.4/25.5/29.2	43.0/30.2/32.0	4.85/4.92/5.18	NR/NR/0.19	128	**88.80**	4.07	579 s
ADiT (4/8/16)	2.20/73.15/85.53	9.05/83.78/97.21	NR/.353/.413	NR/5.92/7.31	200	88.71	10.27	908 s
Crystalite (3/7/15)	NR/79.00/81.10	NR/33.30/93.30	NR/2.325/1.312	NR/1.30/5.70	799	82.35	**12.09**	957 s
OMatG (6/12/24)	73.8/73.5/72.7	48.0/45.0/54.7	1.055/.876/1.069	2.80/3.71/4.27	2127	83.82	5.59	6073 s
FlowMM (2/4/8)	70.00/79.00/80.60	77.50/97.40/98.60	2.25/1.03/.606	2.49/3.89/4.84	1000	81.51	5.78	7801 s
CrysBFN (2/4/8)	NR/NR/NR	NR/NR/NR	NR/NR/NR	NR/NR/NR	1000	85.73	8.87	4160 s
DiffCSP (2/4/8)	NR/NR/NR	NR/NR/NR	NR/NR/NR	NR/NR/NR	2000	82.65	9.05	6643 s
CDVAE (2/4/8)	NR/NR/NR	NR/NR/NR	NR/NR/NR	NR/NR/NR	5000	84.70	7.05	21450 s
MPTS-52								
CoMMat(D) (2/4/8)	73.34/76.02/77.12	95.90/98.38/98.89	**.933**/.772/.613	4.55/**5.51/5.24**	32	78.89	**5.32**	**249 s**
CoMMat(R) (2/4/8)	**79.53/81.59/82.84**	**98.43/99.20/99.32**	1.250/.971/.821	**4.62**/4.80/4.85	32	**83.19**	4.85	**249 s**
FlowMM (2/4/8)	60.56/69.88/70.42	96.52/98.59/99.12	1.314/**.675/.385**	0.90/2.60/2.95	1000	71.58	3.40	18563 s
CrysBFN (2/4/8)	.036/33.44/45.31	NR/8.4/10.61	28.10/6.50/3.236	0.85/2.75/2.60	1000	65.43	4.01	14133 s
CDVAE (2/4/8)	NR/NR/NR	NR/NR/NR	NR/NR/NR	NR/NR/NR	5000	80.89	5.29	35940 s
SyMat (2/4/8)	NR/NR/NR	NR/NR/NR	NR/NR/NR	NR/NR/NR	5000	80.41	1.95	28670 s

Note: NR denotes "no result": at that sampling budget, the model did not produce valid/evaluable crystals under our protocol, so the metric was not computable. Detailed DNG metrics results for Perov-5 and Carbon-24 are reported in Appendix 11. Sampling steps k are converted to NFE p as follows: OMatG, $p = 3(k-1)$; Crystalite, $p = 2k-1$; ADiT, $p = 2k$; all others, $p = k$.

Table 2: Crystal structure prediction ($k = 1$). Match rate (MR, ↑) and RMSE (↓) are reported after 2, 4, and 8 sampling steps. NFE (2/4/8) gives the corresponding number of network evaluations at these checkpoints. Native or nearest available reference-budget results are shown for comparison. Best and second-best results per column within each dataset are shown in **bold** and underlined, respectively.

			p_1 steps		p_2 steps		p_3 steps		Native budget		
Dataset	Model	NFE ($p_1/p_2/p_3$)	MR	RMSE	MR	RMSE	MR	RMSE	NFE	MR	RMSE
MP-20	CoMMat	2/4/8	**42.1**	**.232**	49.3	**.186**	50.8	.160	32	51.65	.095
	CrystalFlow	2/4/8	29.7	.329	48.2	.255	**56.7**	.190	100	59.23	.128
	FlowMM	2/4/8	40.94	.269	**50.3**	.212	56.5	.162	1000	58.82	.117
	CrysBFN	2/4/8	.28	.361	37.3	.229	45.7	**.114**	1000	56.8	.061
	Crystalite	3/7/15	.7	.389	.4	.409	.44	.402	799	66.1	**.033**
	OMatG	8/16/32	11.74	.405	10.00	.406	12.28	.398	418	**69.5**	.096
	DiffCSP	2/4/8	2.06	.410	.77	.406	.35	.407	2000	52.54	.079
MPTS-52	CoMMat	2/4/8	**6.5**	**.376**	8.49	**.330**	9.12	.308	32	14.42	.281
	CrystalFlow	2/4/8	5.2	.411	**11.9**	.336	**15.8**	**.285**	100	23.2	.156
	FlowMM	2/4/8	.12	.635	2.19	.580	4.45	.456	1000	17.65	.243
	CrysBFN	2/4/8	NR	NR	NR	NR	NR	NR	1000	12.2	.139
	Crystalite	3/7/15	.11	.399	.06	.401	.10	.381	799	**31.9**	**.070**
	OMatG	2/6/14	1.26	.415	1.14	.414	1.12	.415	254	25.96	.187
	DiffCSP	2/4/8	NR	NR	NR	NR	.06	.469	2000	14.32	.139

Note: NR denotes "no result": at that sampling budget, the model did not produce valid/evaluable crystals under our protocol, so the metric was not computable. Perov-5 and Carbon-24 results are reported in Appendix 12. Sampling steps k are converted to NFE p as follows: OMatG, $p = 2(k-1)$; Crystalite, $p = 2k-1$; all others, $p = k$.

5 Experiments & Results

5.1 Experimental setup

Datasets and evaluation scope. We evaluate on the standard MP-20 (Jain et al., 2013), Perov-5 (Castelli et al., 2012b;a), and Carbon-24 (Pickard, 2020) benchmarks, using the same train/validation/test splits as (Xie et al., 2021; Luo et al., 2023). We also use MPTS-52 as a harder benchmark for larger and more diverse structures. In the main text, we report both DNG and CSP results on MP-20 and MPTS-52. Results for both tasks on Perov-5 and Carbon-24 are provided in Appendix G. Runtime, sampling efficiency, and most additional analyses focus on MP-20.

Baselines. We compare against two groups of baselines. For performance against long-budget iterative samplers we use **CDVAE** (Xie et al., 2021), the closest baseline, **DiffCSP** (Jiao et al., 2023), and **FlowMM** (Miller et al., 2024).

For robustness under reduced budgets we evaluate recent flow- and interpolant-based generators — **CrystalFlow** (Luo et al., 2025), **OMatG** (Höllmer et al., 2025), **Crystalite** (Veljković et al., 2026), **ADiT** (Joshi et al., 2025), and **Zatom-1** (Morehead et al., 2026) — as low NFE baselines. Where the relevant DNG or CSP model is publicly available, we use the released checkpoints and evaluate them at matched network-evaluation budgets on a single NVIDIA A30 under FP32 using a common scoring pipeline.

Tasks and metrics. Following prior work (Xie et al., 2021; Jiao et al., 2023; Miller et al., 2024; Das et al., 2025), we evaluate CoMMAT on DNG and CSP over both few-step and native budget regimes, with the latter corresponding to each baseline's originally adopted sampling budget. For DNG, we generate $10{,}000$ crystals and report overall validity, coverage, Wasserstein distances for density ρ and number of unique elements N_{el}, and S.U.N. rate, where S.U.N. denotes the fraction of generated crystals that are stable, unique, and novel. We also report DNG results for the realism (R) and discovery (D) operating modes of CoMMAT. See (Appendix. D.3, Table. 15). For CSP, we condition on held-out compositions and report match rate (MR) and RMSE. Following Martirossyan et al. (2026), we additionally use a polymorph-aware split in which all polymorphs of a given composition are assigned to the same subset, avoiding train–test leakage across composition-equivalent polymorphs; corresponding results are provided in Table 13. For both tasks, performance is evaluated across NFE budgets and compared with baselines at matched few-step budgets and at their native or nearest available inference budgets.

5.2 Results

De novo generation Table 1 shows that CoMMAT remains robust in the few-step regime on both MP-20 and the more challenging MPTS-52 benchmark. On MP-20, CoMMat(R) maintains about 87% validity with at least 98.7% coverage recall across 2–8 steps, while CoMMat(D) reaches up to 10.76% SUN and provides the strongest density matching at the lowest budgets. At its native setting, CoMMat(R) achieves 87.85% validity and CoMMat(D) 11.22% SUN using only 32 NFE, with a generation time about 87% lower than the fastest reported baseline.

The same behavior carries to MPTS-52. CoMMat(R) achieves 79.5–82.8% validity and above 98.4% coverage recall across 2–8 steps, exceeding FlowMM by roughly 12–19 percentage points in validity. The two CoMMat modes achieve 4.55–5.51% SUN, with CoMMat(D) giving the strongest discovery performance at 4 and 8 steps. At the native 32-NFE setting, CoMMat(R) attains the highest validity (83.19%), while CoMMat(D) achieves the highest SUN (5.32%). Generating 10,000 MPTS-52 crystals requires only 249 s, more than 98% less time than the fastest reported native-budget baseline. Overall, the results show that CoMMat retains strong validity, coverage, and discovery performance as the sampling trajectory is reduced to only a few steps.

Crystal structure prediction. Table 2 compares CSP performance under matched few-step budgets and at each model's native or nearest available reference budget. On MP-20, CoMMAT achieves the highest match rate and lowest RMSE at 2 NFE (42.1% MR, 0.232 RMSE), while retaining the lowest RMSE at 4 NFE (0.186). The low-budget advantage is particularly pronounced on MPTS-52, where CoMMAT attains 6.5%, 8.49%, and 9.12% MR at 2, 4, and 8 NFE, respectively; in contrast, DiffCSP produces no evaluable structures at 2 or 4 NFE. Results on Perov-5 and Carbon-24 (Appendix Table 12) further support this trend, with CoMMAT remaining strong under the same few-step constraints. Overall, CoMMAT preserves CSP quality under aggressive sampling compression, whereas the evaluated baselines generally require substantially larger reference budgets to reach their strongest performance.

Ablation analysis. Table 3 examines the contributions of the training design and lattice objective at a fixed 32-NFE budget. Removing the decoupled coordinate-training design substantially degrades CSP performance, reducing match rate from 51.65 to 31.87 and increasing RMSE from 0.0951 to 0.1923, while also slightly reducing DNG validity and coverage. Replacing the autoregressive composition prior with an MLP similarly lowers CSP performance (29.71 MR, 0.1585 RMSE); moreover, because this variant no longer models the generative composition distribution $p(\mathbf{A})$, DNG

Table 3: Ablation studies on MP-20. (a) Training and prior-design ablations. Each variant differs from Full CoMMat in a single design choice. **(b)** Lattice-objective ablations. Each variant differs from one-shot lattice prediction only in the lattice objective. All results are reported at 32 NFE. Higher is better except for RMSE and w_ρ.

(a) Training and prior design

	CSP		DNG	
Variant	MR ↑	RMSE ↓	Valid. ↑	COV-R ↑
Full CoMMat	**51.65**	**0.0951**	**87.85**	**99.61**
w/o decouple	31.87	0.1923	86.04	99.48
MLP comp-prior	29.71	0.1585	–	–

(b) Lattice objective

	CSP		DNG	
Lattice objective	MR ↑	RMSE ↓	Valid. ↑	w_ρ ↓
One-shot	**51.65**	**0.0951**	**87.85**	**0.1118**
Lattice-IMM	45.35	0.1581	86.44	0.1706
Lattice-Flow	45.50	0.1502	84.17	0.1193

Note. "–" denotes an undefined DNG evaluation rather than a missing measurement. The MLP variant parameterizes the conditional structural prior $p(\mathbf{z} \mid \mathbf{A})$ but does not model the generative composition distribution $p(\mathbf{A})$ required for de novo sampling.

is undefined under our sampling protocol. Together, these results support the importance of separating coordinate refinement from atom-type corruption and retaining an explicit generative model over composition.

For the lattice, iterative objectives provide no improvement over one-shot prediction at the matched 32-NFE budget. The one-shot formulation achieves the best result on every reported metric, with 51.65 MR, 0.0951 RMSE, 87.85% validity, and $w_\rho = 0.1118$, compared with both Lattice-IMM and Lattice-Flow. This indicates that, in our setting, iterative lattice refinement is not required to obtain strong structural quality and supports allocating the iterative sampling budget entirely to coordinate refinement.

6 Conclusion

In this work, we studied few-step crystal generation with the goal of reducing the sampling cost required to explore large spaces of candidate compositions and structures. We introduced CoMMAT, a few-step crystal generator that separates composition modeling from structural refinement. The model first determines the composition and a composition-conditioned latent representation, predicts the lattice once, and then refines only the atomic coordinates using an MMD-based distributional moment-matching objective. Keeping atom identities fixed during coordinate refinement avoids iterative atom-type denoising and allows the structural generator to focus on continuous crystal geometry.

Across MP-20, MPTS-52, Perov-5, and Carbon-24, CoMMAT remains competitive with recent crystal generators operating at substantially larger native sampling budgets, while maintaining strong performance in the severely constrained 2–8 NFE regime. The same framework supports both CSP and DNG: CSP conditions directly on a supplied composition, whereas DNG first generates a composition and then constructs a compatible lattice and atomic arrangement. The results show that high-quality crystal generation does not require composition and geometry to be iteratively updated together throughout sampling.

A limitation of CoMMAT is that it does not explicitly enforce crystallographic symmetry. Future work could incorporate space-group or Wyckoff information into its composition-conditioned generation process and extend the framework to larger and more structurally diverse crystals.

Overall, CoMMAT demonstrates that effective crystal generation is possible within a few-step sampling regime, providing a more efficient foundation for exploring large compositional and structural search spaces.

AI use statement

For this work, generative AI tools were used to assist with language editing, manuscript organization, and stylistic refinement. All scientific content, methodology, experiments, results, and conclusions were reviewed and verified by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

We have considered the ethical implications of this work in accordance with the ICLR Code of Ethics. COMMAT is developed for general computational crystal generation and structure-prediction research using publicly available materials datasets. The study does not involve human subjects, personal information, or other sensitive data.

We encourage users to apply the model responsibly, comply with relevant regulations, and critically assess generated outputs before using them in downstream scientific applications. As with other generative models, the method could potentially be misused outside its intended research setting, including for the design of harmful materials.

With these considerations, we do not anticipate ethical concerns arising from the development or evaluation presented in this work, provided that COMMAT is used for responsible scientific purposes.

REPRODUCIBILITY STATEMENT

We have taken several steps to support reproducibility. Section 4 gives the complete model formulation: the composition-conditioned structural prior and autoregressive composition model (Sec. 4.1), one-shot lattice prediction (Sec. 4.2), the lattice-normalized Cartesian coordinate representation and corruption process (Sec. 4.3), the training objective (Sec. 4.4), and the few-step sampling procedure for both CSP and DNG (Sec. 4.5). Section 5.1 describes the datasets, tasks, baselines, metrics, sampling budgets, and hardware; all timings are measured on a single NVIDIA A30 GPU in FP32, and baselines are evaluated from publicly released checkpoints through a common scoring pipeline. The appendix provides dataset splits and preprocessing (Appendix A), full architecture hyperparameters (Appendix B), and training details including time sampling, teacher-gap construction, particle grouping, kernel scaling and clipping, outer weighting, KL warm-up, lattice teacher forcing, loss coefficients, and optimizer settings (Appendix C). Sampling configurations and pseudocode for training, coordinate moment matching, CSP, and DNG are given in Appendices D and E, and exact metric definitions, thresholds, and `StructureMatcher` tolerances are given in Appendix F. All reported CoMMat experiments use random seed 42. We will release the implementation, configuration files, trained checkpoints, evaluation scripts, and the polymorph-aware MPTS-52 split as anonymized supplementary material.

REFERENCES

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

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, 2012a.

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, 2012b.

Kishalay Das, Subhojyoti Khastagir, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Periodic materials generation using text-guided joint diffusion model. *arXiv preprint arXiv:2503.00522*, 2025.

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.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023.

Johannes Gasteiger, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *Advances in Neural Information Processing Systems*, 38:75460–75482, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Philipp Höllmer, Thomas Egg, Maya M Martirossyan, Eric Fuemmeler, Zeren Shui, Amit Gupta, Pawan Prakash, Adrian Roitberg, Mingjie Liu, George Karypis, et al. Open materials generation with stochastic interpolants. *arXiv preprint arXiv:2502.02582*, 2025.

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

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

Chaitanya K Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary W Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. *arXiv preprint arXiv:2503.03965*, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Xiaoshan Luo, Zhenyu Wang, Qingchang Wang, Xuechen Shao, Jian Lv, Lei Wang, Yanchao Wang, and Yanming Ma. Crystalflow: a flow-based generative model for crystalline materials. *Nature Communications*, 16(1):9267, 2025.

Youzhi Luo, Chengkai Liu, and Shuiwang Ji. Towards symmetry-aware generation of periodic materials. *Advances in Neural Information Processing Systems*, 36:53308–53329, 2023.

Maya Martirossyan, Thomas Egg, Philipp Höllmer, George Karypis, Mark Transtrum, Adrian Roitberg, Mingjie Liu, Richard Hennig, Ellad Tadmor, and Stefano Martiniani. All that structure matches does not glitter. *Advances in Neural Information Processing Systems*, 38, 2026.

Houssam Metni, Laura Ruple, Lauren N Walters, Luca Torresi, Jonas Teufel, Henrik Schopmans, Jona Östreicher, Yumeng Zhang, Marlen Neubert, Yuri Koide, et al. Generative models for crystalline materials. *Advanced Materials*, 38(18):e23620, 2026.

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.

Alex Morehead, Miruna Cretu, Antonia Panescu, Rishabh Anand, Maurice Weiler, Ty Perez, Samuel M Blau, Steven Farrell, Wahid Bhimji, Anubhav Jain, et al. Zatom-1: A multimodal flow foundation model for 3d molecules and materials. In *ICLR 2026 Workshop on Foundation Models for Science: Real-World Impact and Science-First Design*, 2026.

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

Chris J Pickard. Airss data for carbon at 10gpa and the c+ n+ h+ o system at 1gpa. *(No Title)*, 2020.

Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj computational materials*, 5 (1):83, 2019.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, volume 2024, pp. 15078–15097, 2024.

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

Tin Hadži Veljković, Joshua Rosenthal, Ivor Lončarić, and Jan-Willem van de Meent. Crystalite: A lightweight transformer for efficient crystal modeling. *arXiv preprint arXiv:2604.02270*, 2026.

Hanlin Wu, Yuxuan Song, Jingjing Gong, Ziyao Cao, Yawen Ouyang, Jianbing Zhang, Hao Zhou, Wei-Ying Ma, and Jingjing Liu. A periodic bayesian flow for material generation. In *International Conference on Learning Representations*, volume 2025, pp. 11741–11770, 2025.

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.

Caiyuan Ye, Yuzhi Wang, Xintian Xie, Tiannian Zhu, Jiaxuan Liu, Yuqing He, Lili Zhang, Junwei Zhang, Zhong Fang, Lei Wang, et al. Materials discovery acceleration by using conditional generative methodology. *npj Computational Materials*, 2025.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025.

Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive moment matching. *arXiv preprint arXiv:2503.07565*, 2025.

7 APPENDIX

A DATASETS AND PREPROCESSING

We evaluate COMMAT on four periodic-material benchmarks: Perov-5 (Castelli et al., 2012b;a), Carbon-24 (Pickard, 2020), MP-20 (Jain et al., 2013), and MPTS-52. Perov-5, Carbon-24, and MP-20 follow the standard benchmark partitions used in prior crystal-generation work (Xie et al., 2021; Luo et al., 2023). MPTS-52 provides a more challenging setting containing larger and more structurally diverse crystals, with up to 52 atoms per unit cell.

Both de novo generation (DNG) and crystal structure prediction (CSP) are evaluated across these benchmarks. The main text reports DNG on MP-20 and CSP on MP-20 and MPTS-52, while the remaining dataset-level results are provided in the Appendix 11.

Table 4: Datasets used in the COMMAT experiments.

Dataset	Train	Val	Test	Max. atoms	Tasks
Perov-5	11,356	3,787	3,785	5	DNG, CSP
Carbon-24	6,091	2,032	2,030	24	DNG, CSP
MP-20	27,136	9,047	9,046	20	DNG, CSP
MPTS-52	27,380	5,000	8,096	52	DNG, CSP

Structure preprocessing. Before graph construction, each structure is reduced to its Niggli cell. Primitive-cell conversion is disabled, so the Niggli-reduced structure is retained without an additional primitive-cell transformation.

Periodic graph construction. Periodic graphs for the training encoder are constructed using `CrystalNN`. Neighbor relations retain their periodic image offsets, allowing message passing to account for interactions crossing unit-cell boundaries.

Lattice representation. We use the `scale_length` representation inherited from the CDVAE preprocessing pipeline. For a crystal with N atoms and lattice lengths (a, b, c), the six-dimensional lattice regression target is

$$\mathbf{y}_L = \left(\frac{a}{N^{1/3}}, \frac{b}{N^{1/3}}, \frac{c}{N^{1/3}}, \alpha, \beta, \gamma \right). \tag{15}$$

Only the lattice lengths are normalized by $N^{1/3}$; the lattice angles remain unchanged.

Coordinate representation. Atomic positions are stored as fractional coordinates $\mathbf{X}$; the normalized Cartesian states used for coordinate matching are constructed on-the-fly during training as described in Appendix C.

B MODEL ARCHITECTURE

COMMAT consists of a training-time crystal encoder, a composition-conditioned latent prior, a one-shot lattice predictor, and a time-conditioned CSPNet coordinate decoder. This section provides the architectural details and hyperparameters of these components.

B.1 TRAINING ENCODER

The structure-aware posterior $q_\phi(\mathbf{z} \mid \mathcal{C})$ introduced in Eq. 4 is parameterized by a DimeNet++ encoder (Gasteiger et al., 2020). The encoder produces a 256-dimensional latent representation and predicts the mean and diagonal log-variance of the posterior distribution.

The encoder is used only during training to construct the posterior latent $\mathbf{z}_q$. It is not used during CSP or DNG inference, where structural latents are obtained from the composition-conditioned prior.

The DimeNet++ configuration is summarized in Table 5.

Table 5: DimeNet++ training-encoder configuration.

Parameter	Value
Latent dimension	256
`hidden_channels`	128
`num_blocks`	4
`int_emb_size`	64
`basis_emb_size`	8
`out_emb_channels`	256
`num_spherical` / `num_radial`	7 / 6
Cutoff radius	7.0 Å
`max_num_neighbors`	20
`envelope_exponent`	5
`num_before_skip` / `num_after_skip`	1 / 2
`num_output_layers`	3
Readout	Mean pooling

B.2 Composition-Conditioned Structural Prior

The composition-conditioned prior $p_\psi(\mathbf{z} \mid \mathbf{A})$ and autoregressive composition model $p_\psi(\mathbf{A})$ introduced in Sec. 4.1 are implemented with a shared causal Transformer. Compositions are represented as nondecreasing sequences of atomic numbers with explicit beginning-of-sequence (BOS) and end-of-sequence (EOS) tokens, preserving stoichiometric multiplicity through repeated element tokens.

Token and positional embeddings are processed by three causal Transformer layers. The representation of the completed composition sequence is projected to the mean $\boldsymbol{\mu}_p(\mathbf{A})$ and diagonal log-variance $\log \boldsymbol{\sigma}_p^2(\mathbf{A})$ of the structural prior. A separate output head on the same Transformer parameterizes the autoregressive next-token distribution used for composition generation. During sampling, tokens that violate the nondecreasing atomic-number ordering are masked.

The Transformer configuration is summarized in Table 6.

Table 6: Composition-prior Transformer configuration.

Parameter	Value
Latent dimension	256
Transformer width d_{model}	256
Attention heads	4
Transformer layers	3
Feed-forward dimension	512
Activation	GELU
Dropout	0
Maximum sequence length	64
Atomic-number vocabulary	1–100
Ordering	Nondecreasing atomic number
Special tokens	BOS, EOS

B.3 Global Prediction Heads

The latent representation is passed to lightweight multilayer perceptron (MLP) heads, each with a single hidden layer of width 256 and ReLU activation.

The one-shot lattice predictor is the h_L/h_L^{dec} module defined in Sec. 4.2. The optional lattice output internal to CSPNet is disabled in the reported CoMMat configuration and is considered only in the corresponding ablation study. Thus, in the main model, the lattice is predicted once from the structural latent and is not updated during coordinate refinement.

Auxiliary atom-count and composition heads provide the training objectives $\mathcal{L}_N$ and $\mathcal{L}_{\text{comp}}$, respectively. These heads do not determine the composition or atom count during reported inference: for

CSP, $\mathbf{A}$ and N are supplied by the task, whereas for DNG, $\mathbf{A}$ is sampled from the autoregressive composition model and $N = |\mathbf{A}|$.

B.4 TIME-CONDITIONED CSPNET DECODER

Coordinate refinement is performed with a time-conditioned CSPNet decoder based on DiffCSP (Jiao et al., 2023). Following the notation of Sec. 4.3, at source time u the decoder receives the wrapped fractional coordinates $\mathbf{X}_u$, fixed atomic species $\mathbf{A}$, conditioning lattice $\bar{\mathbf{L}}$, structural latent $\mathbf{z}$, and source time u.

The source time is embedded as

$$\mathbf{e}_u = \mathrm{MLP}_t(u) \in \mathbb{R}^{256}. \tag{16}$$

For atom i, the initial node representation combines its learned element embedding, the time embedding, and the structural latent:

$$\mathbf{h}_i^{(0)} = W_{\mathrm{in}} \left[\mathbf{e}_{A_i} \,\|\, \mathbf{e}_u \,\|\, \mathbf{z} \right]. \tag{17}$$

CSPNet constructs a fully connected graph within each crystal. For edge (i, j), the message function uses the two node states, a periodic sinusoidal encoding of the fractional displacement, and the lattice Gram matrix:

$$\mathbf{m}_{ij}^{(\ell)} = \mathrm{MLP}_e^{(\ell)} \left(\mathbf{h}_i^{(\ell)}, \mathbf{h}_j^{(\ell)}, \Phi(\Delta\mathbf{X}_{ij}), \mathrm{vec}\left[\bar{\mathbf{L}}\bar{\mathbf{L}}^{\mathsf{T}}\right] \right), \tag{18}$$

where Φ denotes a periodic sinusoidal encoding with ten frequencies. Incoming edge features are mean aggregated and incorporated through a residual node update,

$$\mathbf{h}_i^{(\ell+1)} = \mathbf{h}_i^{(\ell)} + \mathrm{MLP}_v^{(\ell)} \left[\mathbf{h}_i^{(\ell)} \,\|\, \mathrm{Mean}_j\, \mathbf{m}_{ij}^{(\ell)} \right]. \tag{19}$$

After six message-passing layers, a linear coordinate head predicts the three-dimensional fractional-coordinate field

$$\Delta\mathbf{X}_{\theta,i} = W_x \mathbf{h}_i^{(6)}. \tag{20}$$

The resulting fractional field is mapped to the normalized Cartesian field $\mathbf{F}_\theta$ according to Eq. (18) in the main text before entering the coordinate-matching objective.

The CSPNet coordinate-decoder configuration used in the reported model is summarized in Table 7.

Table 7: CSPNet coordinate-decoder configuration for COMMAT.

Parameter	Value
Hidden dimension	512
Message-passing layers	6
Structural latent dimension	256
Time-embedding dimension	256
Graph connectivity	Fully connected per crystal
Displacement encoding	Periodic sinusoidal
Number of frequencies	10
Activation	SiLU
Layer normalization	Yes
Lattice conditioning	$\bar{\mathbf{L}}\bar{\mathbf{L}}^{\mathsf{T}}$
Coordinate output dimension	3 per atom

C TRAINING

COMMAT is trained jointly through a structure-aware posterior pathway and a composition-conditioned prior pathway. The coordinate module is optimized by matching student and stop-gradient teacher pushforward populations at a shared target time, while the composition model, one-shot lattice predictor, and auxiliary global heads are trained jointly.

C.1 Coordinate Corruption Settings

Coordinate corruption follows the normalized Cartesian representation and linear flow-matching bridge defined in Sec. 4.3. In particular, the centered fractional coordinates, lattice-normalized state $\mathbf{x}_0$, wrapped fractional state $\mathbf{X}_u$, and corrupted state $\mathbf{x}_u$ are constructed according to Eqs. 8–9.

For all reported experiments, the corruption interval and coordinate noise scale are

$$\epsilon = 0.006, \qquad T = 0.994, \qquad \sigma_{\text{data}} = 0.20. \tag{21}$$

The linear schedule is $\alpha_u = 1 - u$ and $\sigma_u = u$. The conditioning lattice $\bar{\mathbf{L}}$ follows the convention of Sec. 4.2: the ground-truth lattice is used while lattice teacher forcing is active, and the corresponding predicted lattice is used thereafter.

Throughout coordinate training, the clean atomic identities $\mathbf{A}$ are used as fixed conditioning for both student and teacher CSPNet evaluations; no categorical state is corrupted or refined.

C.2 Student–Teacher Time Sampling

Training uses the ordered triplet

$$s \le r \le t, \tag{22}$$

where t is the student source time, r is the nearby teacher source time, and s is the common target time.

We parameterize time through the noise-to-signal ratio

$$n_u = \frac{\sigma_u}{\alpha_u} = \frac{u}{1-u}. \tag{23}$$

For each particle group, the student ratio is sampled from a truncated log-normal distribution,

$$\log n_t \sim \mathcal{N}\left(p_{\text{mean}}, p_{\text{std}}^2\right), \qquad p_{\text{mean}} = -1.1, \qquad p_{\text{std}} = 2.0, \tag{24}$$

with $n_t \in [n_\epsilon, n_T]$. The target ratio n_s is sampled from the same distribution with its upper bound restricted to n_t.

The teacher is placed a fixed distance below the student in noise-to-signal space:

$$\Delta n = (n_T - n_\epsilon)2^{-k}, \qquad k = 12, \tag{25}$$

$$n_r = \max\left(n_t - \Delta n, n_\epsilon\right), \tag{26}$$

with $n_s \le n_r \le n_t$. The three ratios are mapped back to time using

$$u = \frac{n_u}{1+n_u}, \qquad u \in \{s, r, t\}. \tag{27}$$

A single triplet (s, r, t) is shared by all $M = 4$ members of a particle group. Each group contains four distinct crystals rather than repeated noise realizations of the same structure. The student and teacher states are constructed using the same Gaussian realization $\boldsymbol{\epsilon}$, as in Eq. 9, so that they lie on the same corruption trajectory and differ only in source time.

C.3 Student–Teacher Pushforward Matching

For each sampled triplet $s \le r \le t$, the student and stop-gradient teacher pushforwards are constructed according to Eqs. 10–11 and Eqs. 1–2 of the main text. Both branches use the same composition $\mathbf{A}$, structural latent $\mathbf{z}$, and conditioning lattice $\bar{\mathbf{L}}$; they differ only in their source states and source times. The resulting pushforward populations are compared at the common target time s using the kernel-MMD objective described below.

C.4 Population-Matching Objective

For each group of $M = 4$ crystals, the student and teacher pushforwards form two empirical populations. We compare them using the biased V-statistic estimator of the squared maximum mean

discrepancy (MMD),

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\Big[k_{t,s}\left(\mathbf{u}_{st}^{(i)},\mathbf{u}_{st}^{(j)}\right) + k_{t,s}\left(\mathbf{u}_{sr}^{(i)},\mathbf{u}_{sr}^{(j)}\right) - 2k_{t,s}\left(\mathbf{u}_{st}^{(i)},\mathbf{u}_{sr}^{(j)}\right)\Big], \tag{28}$$

including the diagonal terms.

Variable-size structures are padded to a common atom dimension and augmented with an occupancy-mask channel that distinguishes physical atoms from padded sites. For the reported configuration, the kernel distance is

$$d(\mathbf{u},\mathbf{v}) = \frac{\|\widetilde{\mathbf{u}} - \widetilde{\mathbf{v}}\|_2}{D}, \tag{29}$$

where $\widetilde{\mathbf{u}}$ and $\widetilde{\mathbf{v}}$ denote the flattened padded representations including the occupancy channel, and D is their flattened feature dimension.

We use the time-scaled Laplace kernel

$$k_{t,s}(\mathbf{u},\mathbf{v}) = \exp\left[-\frac{w_k(t,s)\,d(\mathbf{u},\mathbf{v})}{\sigma_\kappa}\right], \qquad \sigma_\kappa = 1, \tag{30}$$

with

$$w_k(t,s) = \min\left(\frac{1}{|t-s|\sigma_{\mathrm{data}}}, 50\right). \tag{31}$$

The clipping prevents near-degenerate source–target intervals from producing arbitrarily sharp kernels.

The generic outer weight $w_{\mathrm{out}}(t,s)$ in Eq. (2) is implemented here as a function of the student time t,

$$\omega(t) = \frac{\alpha_t^a}{\alpha_t^2+\sigma_t^2}\,\frac{1}{2}\left(-\frac{d\lambda(t)}{dt}\right)\mathrm{sigmoid}\left(b-\lambda(t)\right), \qquad a=2, \quad b=4, \tag{32}$$

where

$$\lambda(t) = \log\frac{\alpha_t^2}{\sigma_t^2}. \tag{33}$$

The resulting coordinate-matching loss is

$$\mathcal{L}_{\mathrm{IMM}}(\mathbf{z}) = \frac{1}{G}\sum_{g=1}^{G}\omega(t_g)\,\widehat{\mathrm{MMD}}_g^2, \tag{34}$$

where G is the number of particle groups in the minibatch.

C.5 Latent Regularization and Prior-Sampled Training

The conditional KL term in sub section 4.1 is evaluated in closed form for the diagonal Gaussian posterior and composition-conditioned prior. For numerical stability, the predicted prior log-variance is clipped elementwise to

$$-4 \le \log\boldsymbol{\sigma}_p^2 \le 8. \tag{35}$$

The KL coefficient is linearly warmed up over the first 200 epochs:

$$\beta_{\mathrm{KL}}(e) = 0.01\min\left(1, \frac{e+1}{200}\right). \tag{36}$$

In addition to the posterior structural pathway, each training step draws an independent latent $\mathbf{z}_p \sim p_\psi(\mathbf{z} \mid \mathbf{A})$ and passes it through the same one-shot lattice predictor and coordinate decoder. The resulting coordinate and lattice objectives are denoted $\mathcal{L}_{\mathrm{IMM}}(\mathbf{z}_p)$ and $\mathcal{L}_L(\mathbf{z}_p)$, respectively. This prior-sampled pathway exposes the structural decoder directly to the latent distribution used during generation.

Table 8: Loss coefficients for the reported CoMMAT configuration.

Term	Coefficient
Coordinate loss $\mathcal{L}_{\mathrm{IMM}}$	$\lambda_x = 10$
Lattice loss $\mathcal{L}_L$	$\lambda_L = 10$
Prior structural pathway	$\lambda_p = 1$
Conditional KL $\mathcal{L}_{\mathrm{KL}}$	$\beta_{\mathrm{KL}} = 0.01$ after warmup
Composition LM loss $\mathcal{L}_{\mathrm{LM}}$	$\lambda_{\mathrm{LM}} = 0.5$
Atom-count loss $\mathcal{L}_N$	$\lambda_N = 1$
Auxiliary composition loss $\mathcal{L}_{\mathrm{comp}}$	$\lambda_{\mathrm{comp}} = 1$
Decoder atom-type CE	0

C.6 Training Loss Weights

The complete training objective follows Eq. 13 in the main text. Table 8 lists the coefficients used in all reported CoMMAT experiments. The conditional KL coefficient follows the 200-epoch warmup schedule defined in Eq. 36.

C.7 Lattice Teacher Forcing

Lattice teacher forcing follows the procedure described in Sec. 4.2. For the reported MP-20 training configuration, the coordinate decoder is conditioned on the ground-truth lattice through epoch 1000 of the 2000-epoch training run. After this cutoff, the posterior and prior structural pathways use their respective predicted lattices.

The one-shot lattice predictor is optimized throughout training; in particular, the lattice regression loss is always computed from the predicted lattice, including during the teacher-forcing stage.

C.8 Optimization

For MP-20, we optimize CoMMAT with Adam (Kingma & Ba, 2014) using a learning rate of 10^{-3}, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\varepsilon_{\mathrm{Adam}} = 10^{-8}$, and weight decay 10^{-5}. The learning rate is linearly warmed up from zero over the first 300 optimizer steps and remains fixed thereafter.

Gradients are clipped by norm at 1.0. Clipping is applied separately to three parameter groups: the CSPNet coordinate decoder, the global prediction heads, and all remaining parameters, including the training encoder and composition Transformer.

Training uses FP32 precision, batch size 256, and random seed 42. The MP-20 training budget is 2000 epochs, with early stopping effectively disabled; the checkpoint with the lowest validation loss is used for evaluation.

D Sampling Configuration

This section specifies the numerical sampling settings used for the few-step coordinate-generation procedure described in Sec. 4.5.

D.1 Shared Coordinate Sampler

Sampling follows the coordinate-generation procedure of Sec. 4.5 in the normalized Cartesian state space defined in Sec. 4.3. For all reported experiments, coordinates are initialized as

$$\mathbf{x}_{t_0} \sim \mathcal{N}\left(\mathbf{0}, \sigma_{\mathrm{data}}^2 \mathbf{I}\right), \qquad \sigma_{\mathrm{data}} = 0.20, \tag{37}$$

and refined on a uniformly spaced decreasing time grid

$$t_k = T - \frac{k}{K}(T - \epsilon), \qquad k = 0, \ldots, K, \qquad T = 0.994, \quad \epsilon = 0.006. \tag{38}$$

The predicted lattice is held fixed throughout sampling, and atomic species remain unchanged during coordinate refinement. At each step, the current normalized Cartesian state is converted to wrapped

Table 9: Sampling configuration for COMMAT. CSP and DNG share the same coordinate sampler after their task-specific composition and latent initialization.

Setting	Value
Coordinate schedule	$\alpha_t = 1 - t,\ \sigma_t = t$
Time interval	$[\epsilon, T] = [0.006, 0.994]$
Sampling discretization	Uniform in t
Coordinate noise scale σ_{data}	0.20
Coordinate initialization	$\mathcal{N}(\mathbf{0}, 0.20^2\mathbf{I})$
Integrator	Euler flow-matching update
Decoder evaluations per sample	K
Lattice prediction	One-shot before coordinate loop
Atom types during refinement	Fixed
CSP composition	Given $\mathbf{A}$
CSP latent	$\mathbf{z}_{\text{CSP}} = \boldsymbol{\mu}_p(\mathbf{A})$
CSP atom count	$N = \lvert\mathbf{A}\rvert$
CSP lattice	$\hat{\mathbf{L}} = h_L^{\text{dec}}(\mathbf{z}_{\text{CSP}}, N)$
DNG composition	$\mathbf{A} \sim p_\psi(\mathbf{A})$
DNG latent	$\mathbf{z}_{\text{DNG}} \sim p_\psi(\mathbf{z} \mid \mathbf{A})$
DNG atom count	$N = \lvert\mathbf{A}\rvert$
DNG lattice	$\hat{\mathbf{L}} = h_L^{\text{dec}}(\mathbf{z}_{\text{DNG}}, N)$

fractional coordinates, evaluated once by CSPNet, and updated according to Eq. 14. Consequently, a K-step trajectory requires exactly K coordinate-decoder neural function evaluations (NFE).

D.2 Crystal Structure Prediction

CSP inference follows Sec. 4.5: the supplied composition determines $\mathbf{z}_{\text{CSP}} = \boldsymbol{\mu}_p(\mathbf{A})$, and the corresponding lattice is predicted once before coordinate refinement. Neither the target coordinates nor the target lattice are used during inference.

For repeated samples of the same composition, the latent and predicted lattice are held fixed, while a new Gaussian coordinate initialization is drawn for each sample. Thus, CSP sampling diversity arises from the coordinate initialization.

D.3 De Novo Generation

DNG follows Sec. 4.5 of the main text. The composition is sampled autoregressively in nondecreasing atomic-number order until an EOS token is generated or the maximum sequence length of 64 is reached.

We report two DNG operating modes that differ only in the autoregressive composition-sampling temperature τ_A. The realism mode COMMAT(R) uses $\tau_A = 1$, whereas the discovery mode COMMAT(D) uses $\tau_A = 1.2$. The latent-sampling temperature is fixed at $\tau_z = 1$ in both modes.

The resulting sequence defines $N = |\mathbf{A}|$, after which $\mathbf{z}_{\text{DNG}} \sim p_\psi(\mathbf{z} \mid \mathbf{A})$ is sampled and used for one-shot lattice prediction and the shared coordinate sampler of Appendix D.1.

D.4 Sampling Hyperparameters

Table 9 summarizes the sampling configuration used by the reported model.

D.5 Sampling cost.

Let $C_{\text{dec}}(N)$ denote the cost of one coordinate-decoder evaluation for a crystal containing N atoms. COMMAT uses a fully connected atomic graph, so $|\mathcal{E}| = \mathcal{O}(N^2)$. With L message-passing layers and hidden width d, the dominant coordinate-decoder cost scales as

$$C_{\text{dec}}(N) = \mathcal{O}\big(LN^2d^2\big)\,, \qquad T_{\text{coord}}(K, N) = K\,C_{\text{dec}}(N) = \mathcal{O}\big(KLN^2d^2\big)\,. \tag{39}$$

Our proposed configuration uses $L = 6$ and $d = 512$. Thus, K exactly counts coordinate-decoder evaluations, but it should not be interpreted as the total number of neural-network forward passes in the complete generation pipeline.

For CSP, the end-to-end inference cost can be decomposed as

$$T_{\mathrm{CSP}}(K, N) = C_{\mathbf{z}|\mathbf{A}}(N) + C_{\mathrm{lat}}(N) + K\, C_{\mathrm{dec}}(N), \tag{40}$$

where $C_{\mathbf{z}|\mathbf{A}}$ denotes construction of the composition-conditioned structural latent and C_{lat} denotes the one-shot lattice prediction. De novo generation additionally incurs autoregressive composition-generation cost.

$$T_{\mathrm{DNG}}(K, N) = C_{\text{AR-comp}}(N) + C_{\mathbf{z}|\mathbf{A}}(N) + C_{\mathrm{lat}}(N) + K\, C_{\mathrm{dec}}(N). \tag{41}$$

The auxiliary terms may depend on crystal size N but are independent of the coordinate sampling budget K. Consequently, reducing K directly reduces the repeatedly incurred coordinate-refinement cost, while the fixed composition, conditioning, and lattice stages impose an end-to-end runtime floor. We therefore report measured wall-clock time alongside coordinate NFE when evaluating sampling efficiency.

E Algorithms

This section summarizes the training and inference procedures used by CoMMAT. Algorithm 1 gives one training step, Algorithm 2 details the coordinate moment-matching objective, and Algorithms 3 and 4 describe crystal structure prediction (CSP) and de novo generation (DNG), respectively. The two inference procedures differ only in how the composition $\mathbf{A}$ and structural latent $\mathbf{z}$ are obtained; both subsequently use the same few-step coordinate sampler.

Neural function evaluations. The iterative portion of both CSP and DNG consists only of the coordinate decoder evaluations in Algorithms 3 and 4. Accordingly, a trajectory with K coordinate refinement steps requires exactly K CSPNet neural function evaluations. The composition Transformer, composition-conditioned latent prior, and one-shot lattice predictor are evaluated outside the iterative coordinate loop.

F Evaluation Protocol

This section provides the exact evaluation settings and metric definitions used for the DNG and CSP results reported in Sec. 5

F.1 De Novo Generation

F.1.1 Validity

We report structural validity, compositional validity, and overall validity. A generated crystal is structurally valid if its periodic structure can be constructed, its unit-cell volume is at least $0.1\ \text{Å}^3$, and its minimum interatomic distance under periodic boundary conditions is at least $0.5\ \text{Å}$.

Compositional validity is evaluated with SMACT (Davies et al., 2019) using charge-neutrality and Pauling-electronegativity checks. Single-element and all-metal compositions are treated as valid, consistent with the CDVAE evaluation implementation.

Overall validity is the fraction of generated crystals satisfying both the structural and compositional validity criteria:

$$V_{\mathrm{overall}} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathbb{1}\left[\mathrm{Valid}_{\mathrm{struct}}(g) \wedge \mathrm{Valid}_{\mathrm{comp}}(g)\right]. \tag{42}$$

F.1.2 Coverage

Coverage follows the fingerprint-based CDVAE protocol. Each crystal $\mathcal{C}$ is represented by a structural fingerprint $\mathbf{f}_{\mathrm{s}}(\mathcal{C})$, obtained by averaging CrystalNNFingerprint site features over the crystal, and a composition fingerprint $\mathbf{f}_{\mathrm{c}}(\mathcal{C})$ given by the standardized Magpie elemental-property representation.

Algorithm 1 CoMMAT training step (one minibatch).

Require: Crystal batch $\{(\mathbf{A}^{(i)}, \mathbf{X}^{(i)}, \mathbf{L}^{(i)})\}_{i=1}^{B}$
Require: Group size $M = 4$, $\epsilon = 0.006$, $T = 0.994$, $\sigma_{\text{data}} = 0.20$, teacher-gap exponent $k = 12$

Structure-aware posterior:
$(\boldsymbol{\mu}_q, \log \boldsymbol{\sigma}_q^2) \leftarrow q_\phi(\mathbf{z} \mid \mathcal{C})$
$\mathbf{z}_q \leftarrow \boldsymbol{\mu}_q + \boldsymbol{\sigma}_q \odot \boldsymbol{\epsilon}_q, \quad \boldsymbol{\epsilon}_q \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

Composition-conditioned prior:
$(\boldsymbol{\mu}_p, \log \boldsymbol{\sigma}_p^2) \leftarrow p_\psi(\mathbf{z} \mid \mathbf{A})$
$\mathcal{L}_{\text{KL}} \leftarrow D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathcal{C}) \,\|\, p_\psi(\mathbf{z} \mid \mathbf{A}))$
$\mathcal{L}_{\text{LM}} \leftarrow -\log p_\psi(\mathbf{A})$

Posterior structural pathway:
Predict atom-count logits and auxiliary composition logits from $\mathbf{z}_q$
$\hat{\mathbf{L}}_q \leftarrow h_L^{\text{dec}}(\mathbf{z}_q, N)$
Compute $\mathcal{L}_N$, $\mathcal{L}_{\text{comp}}$, and $\mathcal{L}_L(\mathbf{z}_q)$
if lattice teacher forcing is active **then**
    $\bar{\mathbf{L}}_q \leftarrow \mathbf{L}$
else
    $\bar{\mathbf{L}}_q \leftarrow \hat{\mathbf{L}}_q$
end if
$\mathcal{L}_{\text{IMM}}(\mathbf{z}_q) \leftarrow \textsc{PopulationMatch}(\mathbf{X}, \mathbf{A}, \bar{\mathbf{L}}_q, \mathbf{z}_q)$

Composition-prior structural pathway:
$\mathbf{z}_p \leftarrow \boldsymbol{\mu}_p + \boldsymbol{\sigma}_p \odot \boldsymbol{\epsilon}_p, \quad \boldsymbol{\epsilon}_p \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\hat{\mathbf{L}}_p \leftarrow h_L^{\text{dec}}(\mathbf{z}_p, N)$
Compute $\mathcal{L}_L(\mathbf{z}_p)$
if lattice teacher forcing is active **then**
    $\bar{\mathbf{L}}_p \leftarrow \mathbf{L}$
else
    $\bar{\mathbf{L}}_p \leftarrow \hat{\mathbf{L}}_p$
end if
$\mathcal{L}_{\text{IMM}}(\mathbf{z}_p) \leftarrow \textsc{PopulationMatch}(\mathbf{X}, \mathbf{A}, \bar{\mathbf{L}}_p, \mathbf{z}_p)$

Structural objectives:
$\mathcal{L}_{\text{str}}(\mathbf{z}_q) \leftarrow \lambda_x \mathcal{L}_{\text{IMM}}(\mathbf{z}_q) + \lambda_L \mathcal{L}_L(\mathbf{z}_q)$
$\mathcal{L}_{\text{str}}(\mathbf{z}_p) \leftarrow \lambda_x \mathcal{L}_{\text{IMM}}(\mathbf{z}_p) + \lambda_L \mathcal{L}_L(\mathbf{z}_p)$
$\beta(e) \leftarrow 0.01 \min\left(1, \frac{e+1}{200}\right)$
$\mathcal{L} \leftarrow \mathcal{L}_{\text{str}}(\mathbf{z}_q) + \lambda_p \mathcal{L}_{\text{str}}(\mathbf{z}_p) + \beta(e)\mathcal{L}_{\text{KL}} + \lambda_{\text{LM}}\mathcal{L}_{\text{LM}} + \lambda_N \mathcal{L}_N + \lambda_{\text{comp}}\mathcal{L}_{\text{comp}}$

where $\lambda_x = 10$, $\lambda_L = 10$, $\lambda_p = 1$, $\lambda_{\text{LM}} = 0.5$, $\lambda_N = 1$, and $\lambda_{\text{comp}} = 1$.
Backpropagate $\mathcal{L}$, clip gradient norms at 1.0, and take an Adam step.

For a generated crystal g and reference crystal t, we use Euclidean distances in the two fingerprint spaces,

$$d_{\text{s}}(g,t) = \|\mathbf{f}_{\text{s}}(g) - \mathbf{f}_{\text{s}}(t)\|_2, \qquad d_{\text{c}}(g,t) = \|\mathbf{f}_{\text{c}}(g) - \mathbf{f}_{\text{c}}(t)\|_2. \tag{43}$$

Coverage recall uses the nearest generated sample for each reference, whereas coverage precision uses the nearest reference for each generated sample. Defining

$$d_{\text{s}}^R(t) = \min_{g \in \mathcal{G}} d_{\text{s}}(g,t), \qquad d_{\text{c}}^R(t) = \min_{g \in \mathcal{G}} d_{\text{c}}(g,t), \tag{44}$$

and

$$d_{\text{s}}^P(g) = \min_{t \in \mathcal{T}} d_{\text{s}}(g,t), \qquad d_{\text{c}}^P(g) = \min_{t \in \mathcal{T}} d_{\text{c}}(g,t), \tag{45}$$

we compute

$$\text{COV-R} = \frac{1}{|\mathcal{T}_{\text{fp}}|} \sum_{t \in \mathcal{T}_{\text{fp}}} \mathbb{1}\left[d_{\text{s}}^{R}(t) \le \delta_{\text{s}} \wedge d_{\text{c}}^{R}(t) \le \delta_{\text{c}}\right], \tag{46}$$

and

$$\text{COV-P} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}_{\text{fp}}} \mathbb{1}\left[d_{\text{s}}^{P}(g) \le \delta_{\text{s}} \wedge d_{\text{c}}^{P}(g) \le \delta_{\text{c}}\right]. \tag{47}$$

Here, $\mathcal{T}_{\text{fp}}$ and $\mathcal{G}_{\text{fp}}$ contain structures for which both fingerprints can be constructed. The precision denominator remains the full generated set, so samples without valid fingerprints receive no coverage credit.

The dataset-specific thresholds are summarized in Table 10.

Table 10: Structural and compositional fingerprint thresholds used for DNG coverage evaluation.

Dataset	δ_{s}	δ_{c}
Perov-5	0.2	4.0
Carbon-24	0.2	4.0
MP-20	0.4	10.0

F.1.3 Distributional Wasserstein Distances

We compare generated and held-out reference distributions using the one-dimensional Wasserstein-1 distance,

$$W_1(P, Q) = \int_{\mathbb{R}} |F_P(x) - F_Q(x)| \, dx. \tag{48}$$

Consistent with the metrics reported in the main text, we evaluate the Wasserstein distances for mass density ρ and the number of distinct elements per crystal N_{el}:

$$W_{\rho}, \qquad W_{N_{\text{el}}}. \tag{49}$$

The standard evaluation uses a deterministic subset of $1{,}000$ overall-valid generated crystals sampled without replacement using random seed 42 and compares this subset with the complete held-out reference set.

F.1.4 Thermodynamic Stability and S.U.N.

Where reported, thermodynamic stability is evaluated after CHGNet relaxation (Deng et al., 2023). Generated structures are relaxed for up to 1500 optimization steps with a force-convergence threshold of $0.05\,\text{eV/Å}$. The relaxed energy is used to compute the energy above the composition-specific convex hull,

$$E_{\text{hull}}(g). \tag{50}$$

Under the strict stability criterion, a generated crystal is stable when

$$\text{Stable}(g) = \mathbb{1}\left[E_{\text{hull}}(g) \le 0 \ \wedge \ \text{arity}(g) \ge 2\right]. \tag{51}$$

Uniqueness and novelty are evaluated on relaxed structures using `StructureMatcher` with

$$\texttt{ltol} = 0.2, \qquad \texttt{stol} = 0.3, \qquad \texttt{angle_tol} = 5^{\circ}. \tag{52}$$

The stable, unique, and novel rate, denoted S.U.N. in the text and SUN in the result tables, is

$$\text{SUN} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \text{Stable}(g)\,\text{Unique}(g)\,\text{Novel}(g). \tag{53}$$

The evaluation implementation additionally computes a metastable analogue using $E_{\text{hull}} \le 0.10\,\text{eV/atom}$; this is kept distinct from the strict SUN metric and is reported only where explicitly indicated.

F.2 Crystal Structure Prediction

For CSP, the composition $\mathbf{A}$ and atom count N of each held-out reference crystal are provided to the model, while the lattice $\mathbf{L}$ and fractional coordinates $\mathbf{X}$ are generated. The target geometry is not used to construct the composition-conditioned structural latent.

Let $\mathcal{G}_t^{(k)}$ denote the set of k independently generated candidates associated with reference crystal $t \in \mathcal{T}$. Candidate structures are compared with their corresponding references using pymatgen `StructureMatcher` with

$$\texttt{ltol} = 0.3, \qquad \texttt{stol} = 0.5, \qquad \texttt{angle_tol} = 10^\circ. \tag{54}$$

A reference crystal is counted as matched if at least one of its generated candidates matches the target:

$$m_t^{(k)} = \mathbb{1}\left[\exists\, g \in \mathcal{G}_t^{(k)} : \mathrm{SM}(g, t) = 1\right]. \tag{55}$$

The Match Rate is

$$\mathrm{MR@}k = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} m_t^{(k)}. \tag{56}$$

For the $k = 1$ results reported in the main tables, this reduces to the fraction of held-out compositions for which the single generated structure matches the corresponding reference.

For matched references, we report the normalized structural root-mean-square error. Let $\mathcal{T}_{\mathrm{match}} \subseteq \mathcal{T}$ denote references with at least one successful match. Then

$$\mathrm{RMSE@}k = \frac{1}{|\mathcal{T}_{\mathrm{match}}|} \sum_{t \in \mathcal{T}_{\mathrm{match}}} \min_{g \in \mathcal{G}_t^{(k)}} \frac{1}{V_t^{1/3}} \sqrt{\frac{1}{N_t} \sum_{i=1}^{N_t} \left\| \mathbf{R}_i^{(g)} - \mathbf{R}_i^{(t)} \right\|_2^2}, \tag{57}$$

where the site correspondence and periodic displacements are determined by `StructureMatcher`, V_t is the reference unit-cell volume, and N_t is the number of atoms in the reference crystal. For the main-paper CSP results, $k = 1$, and we therefore report simply MR and RMSE. When best-of-k results are additionally reported, the same definitions are used with $|\mathcal{G}_t^{(k)}| = k$ and the best matching candidate selected for each reference.

G Ablation Results

Full de novo generation results. Table 11 reports the complete DNG metrics under our unified evaluation protocol. On MP-20, CoMMat(R) emphasizes validity and distributional fidelity, while CoMMat(D) favors discovery, reaching an SUN rate of 11.22%. Despite using only 32 NFE, both modes remain competitive with baselines operating at substantially larger native budgets. In particular, CoMMat achieves the lowest element-distribution Wasserstein distance on MP-20.

On Perov-5, CoMMat reaches 98.83% overall validity and performs particularly well on distributional metrics, obtaining the lowest Wasserstein distances for both density and elemental composition, together with strong coverage. On Carbon-24, CoMMat achieves 100% structural and compositional validity, although its coverage precision and density distribution matching remain below the strongest native-budget baselines. Overall, the results show that the reduction in sampling budget does not come at the cost of a general collapse in generation quality across datasets.

Few-step sampling efficiency and structural fidelity. As summarized in Table 12, conventional diffusion- and flow-based baselines degrade substantially when their sampling trajectories are truncated to only a few function evaluations. On Perov-5, DiffCSP and OMatG achieve below 5% match rate throughout the evaluated few-step regime, while continuous-flow baselines are more robust but exhibit a clear loss in structural accuracy at aggressive truncation. For example, CrystalFlow's RMSE increases from 0.105 Å at its 100-NFE reference budget to 0.381 Å at 2 NFE.

In contrast, CoMMat retains strong predictive quality even under extreme sampling compression. At only 2 NFE on Perov-5, it achieves a 32.80% match rate with an RMSE of 0.122 Å, giving the

Table 11: De novo crystal generation on MP-20, Perov-5, and Carbon-24. All metrics are recomputed using our unified evaluation protocol. CoMMat is reported at 32 neural function evaluations (NFE), and baselines are evaluated at their native inference budgets. **R** and **D** denote the realism and discovery modes, respectively. The best and second-best results within each dataset are shown in **bold** and underlined, respectively. Higher is better for validity, coverage, stability, and SUN; lower is better for Wasserstein distances.

		Validity ↑			Coverage ↑		W-dist ↓		Discovery ↑	
Model	NFE	Overall	Struct.	Comp.	R	P	ρ	El.	Stab.	SUN
MP-20										
CoMMat (R)	32	87.85	99.97	88.45	99.61	**99.90**	.221	**.034**	10.67	8.26
CoMMat (D)	32	81.23	99.19	82.40	99.25	99.58	.194	.176	11.75	11.22
CrystalFlow	100	81.24	99.70	81.38	98.33	99.83	.181	.292	6.54	3.70
ZAtom-1	128	**88.80**	97.80	**89.80**	99.00	99.80	.248	.041	14.54	4.07
ADiT	200	88.71	99.56	89.09	99.04	99.81	**.1214**	.106	**17.38**	10.27
Crystalite	799	82.35	99.75	82.51	99.61	99.86	.199	.183	16.14	**12.09**
OMatG	2127	83.82	98.81	84.71	99.58	99.70	.2912	.0468	9.56	5.59
CrysBFN	1000	85.73	**99.99**	85.73	99.16	99.78	.203	.274	17.28	8.87
FlowMM	1000	81.51	99.29	82.05	99.54	99.66	.149	.191	7.63	5.78
DiffCSP	2000	82.65	99.90	82.74	**99.73**	99.71	.133	.3638	11.46	9.05
CDVAE	5000	84.70	99.96	84.71	99.24	99.56	.7347	1.111	7.36	7.05
SyMat	5000	80.51	99.86	88.64	91.20	97.51	.939	.340	1.70	1.44
Perov-5										
CoMMat	32	98.83	99.98	**98.90**	99.78	**98.64**	**.071**	**.020**	–	–
CrystalFlow	100	98.20	99.51	98.67	99.63	92.71	.656	.035	–	–
CrysBFN	1000	**98.90**	**100.00**	**98.90**	99.10	98.10	.159	.025	–	–
FlowMM	1000	98.60	99.90	98.70	**99.80**	98.10	.077	.035	–	–
DiffCSP	2000	98.19	**100.00**	98.23	99.39	98.19	.120	.021	–	–
CDVAE	5000	98.57	**100.00**	98.57	99.23	98.41	.153	.127	–	–
SyMat	5000	96.98	**100.00**	96.98	96.64	83.58	.199	.045	–	–
Carbon-24										
CoMMat	32	**100.00**	**100.00**	**100.00**	99.76	78.71	.103	–	–	–
CrystalFlow	100	99.87	99.87	**100.00**	99.61	30.91	.113	–	–	–
CrysBFN	1000	**100.00**	**100.00**	**100.00**	99.60	**98.90**	.087	–	–	–
FlowMM	1000	**100.00**	**100.00**	**100.00**	**99.98**	84.90	**.043**	–	–	–
DiffCSP	2000	**100.00**	**100.00**	**100.00**	99.90	98.47	.058	–	–	–
CDVAE	5000	**100.00**	**100.00**	**100.00**	99.80	86.24	.136	–	–	–
SyMat	5000	99.79	99.79	**100.00**	51.63	1.97	.165	–	–	–

Note: "–" indicates that the corresponding metric is not reported in the results used here. Element-distribution Wasserstein distance is not reported for Carbon-24. Stability and SUN are shown for MP-20 only. When multiple models attain the same best value, all tied results are shown in bold.

Table 12: Crystal structure prediction on Perov-5 and Carbon-24 ($k = 1$). Match rate (MR, ↑, %) and RMSE (↓, Å) are reported at three few-step operating points and at each model's native reference budget. The NFE column gives the actual number of function evaluations corresponding to the Low/Mid/High operating points. Best and second-best results within each dataset and column are shown in **bold** and underline, respectively.

			Few-step operating points						**Native reference**		
Dataset	**Model**	**NFE**	**Low**		**Mid**		**High**		**NFE**	**MR**	**RMSE**
			MR	RMSE	MR	RMSE	MR	RMSE		↑	↓
Perov-5	**CoMMat (Ours)**	2/4/8	**32.80**	**0.122**	**31.80**	**0.142**	33.10	**0.133**	32	42.50	0.125
	CrystalFlow	2/4/8	25.37	0.381	30.10	0.293	**47.31**	0.182	100	53.80	0.105
	FlowMM	2/4/8	24.45	0.321	25.60	0.228	30.90	0.153	1000	52.30	0.097
	OMatG	2/6/14	2.75	0.390	2.85	0.329	1.14	0.403	958	**62.67**	0.247
	DiffCSP	2/4/8	0.74	0.445	1.77	0.452	4.39	0.441	2000	49.33	**0.081**
Carbon-24	**CoMMat (Ours)**	2/4/8	**23.90**	**0.411**	**25.00**	**0.282**	**25.50**	**0.140**	32	**27.84**	**0.215**
	FlowMM	2/4/8	22.12	0.415	24.19	0.380	20.54	0.356	1000	23.20	0.242
	DiffCSP	2/4/8	7.93	0.445	6.80	0.439	5.62	0.435	2000	17.14	0.290

highest MR and lowest RMSE among the evaluated methods at this operating point. This advantage persists across the few-step regime: CoMMat attains the lowest RMSE at 2, 4, and 8 NFE, while remaining competitive in match rate. The effect is particularly pronounced on Carbon-24, where CoMMat achieves the best MR and RMSE at every evaluated few-step operating point. At its 32-NFE reference budget, it further reaches 27.84% MR, exceeding FlowMM's 23.20% at 1000 NFE and DiffCSP's 17.14% at 2000 NFE, corresponding to $31.25\times$ and $62.5\times$ fewer function evaluations,

Table 13: Robustness to polymorph- and duplicate-aware evaluation. We evaluate CoMMat at 32 NFE using the dataset corrections and metrics motivated by (Martirossyan et al., 2026). MR denotes the conventional one-to-one match rate, while METRe evaluates reference coverage when predictions are pooled by composition. cRMSE accounts for unmatched references when measuring structural error. For Carbon-24, *Unique* denotes the duplicate-pruned dataset rather than a polymorph split.

Dataset	Evaluation	MR ↑	METRe ↑	RMSE ↓	cRMSE ↓	N_{ref}
MP-20	Standard	51.65	52.15	0.095	0.305	9046
	Polymorph-aware	52.23	54.03	0.114	0.312	9046
Perov-5	Standard	42.50	44.59	0.125	0.306	3785
	Polymorph-aware	78.66	89.76	0.0834	0.229	3785
Carbon-24	Standard	27.84	92.96	0.215	0.203	2030
	Unique	29.78	98.27	0.3516	0.4257	850
MPTS-52	Standard	14.42	14.82	0.281	0.3808	8096
	Polymorph-aware[†]	36.96	38.21	0.1289	0.2395	8095

[†]Polymorph-aware MPTS-52 follows the composition-grouping principle discussed in (Martirossyan et al., 2026); it is not an officially released split from that work. MR and METRe are reported in percentage points.

Table 14: Capacity and decoder ablations on MP-20. (a) Capacity scaling of CoMMat with CSPNet while keeping the training and sampling protocol fixed. **(b)** Decoder architecture comparison under the corresponding matched setting. MR is evaluated at 32 NFE; t_{10k} is wall-clock sampling time for $10k$ structures. Higher MR and validity are better; lower time is better.

(a) Capacity scaling

Hidden dim.	Params ↓	MR@32 ↑	Valid. ↑	t_{10k} ↓
128	5.18 M	42.86	86.44	**24.33**
256	7.15 M	**51.75**	87.60	36.75
512	14.73 M	51.65	**87.85**	76.84
1024	44.43 M	51.40	86.25	237.12

(b) Decoder architecture (256)

Decoder	Params ↓	MR@32 ↑	Valid. ↑	t_{10k} ↓
GemNet-T	**6.77 M**	43.07	87.51	151.8
CSPNet	14.73 M	**51.75**	**87.60**	**36.75**

respectively. Overall, these results show that CoMMAT preserves structural fidelity under aggressive reduction of the sampling trajectory, enabling effective crystal structure prediction with substantially fewer network evaluations than iterative diffusion and flow-based alternatives.

Polymorph-aware evaluation. Table 13 shows that the apparent performance of CoMMAT depends strongly on how structurally equivalent references are represented in the evaluation set. The effect is most pronounced on Perov-5, where polymorph-aware evaluation increases MR from 42.50% to 78.66% and METRe from 44.59% to 89.76%, while simultaneously reducing RMSE from 0.125 to 0.0834 and cRMSE from 0.306 to 0.229. MPTS-52 exhibits a similar trend, with MR increasing from 14.42% to 36.96% and cRMSE decreasing from 0.3808 to 0.2395. In contrast, MP-20 changes only marginally, suggesting that conventional matching already captures most of the recoverable references in this dataset. Carbon-24 behaves differently: duplicate pruning substantially reduces the reference set ($2030 \rightarrow 850$) and increases METRe to 98.27%, indicating broad composition-level coverage, although the larger RMSE and cRMSE show that this improvement does not translate into lower structural error. Overall, these results indicate that one-to-one matching can substantially understate recovered structural diversity on datasets with many polymorphic or redundant references, while the magnitude of this effect is highly dataset dependent.

Capacity. Table 14(a) shows that increasing the hidden dimension from 128 to 256 substantially improves MR from 42.86% to 51.75%, while further scaling yields no meaningful improvement. The 512-dimensional model achieves the highest validity (87.85%) but nearly identical MR (51.65%) at more than twice the sampling time of the 256-dimensional variant (76.84 s vs. 36.75 s). Scaling to

Table 15: Composition-prior temperature ablation on MP-20. We vary the sampling temperature τ under the Discovery and Realism modes while keeping the coordinate-refinement budget fixed at CoMMat's native setting of **32 NFE**. Higher SUN is better. Best results within each mode are shown in **bold**.

τ	SUN (Discovery) ↑	SUN (Realism) ↑
0.6	10.24	6.87
0.8	10.34	7.91
1.2	**11.22**	**11.00**
2.0	10.50	8.94

1024 dimensions further increases the parameter count and sampling time to 44.43M and 237.12 s, respectively, without improving MR or validity. These results suggest that CoMMat reaches its best accuracy–efficiency trade-off at moderate capacity, with little benefit from further parameter scaling.

Decoder architecture. Table 14(b) shows that GemNet-T uses fewer parameters than CSPNet (6.77M vs. 14.73M), yet requires nearly twice the sampling time (151.8 s vs. 36.75 s per 10^4 structures). CSPNet also achieves substantially higher MR (51.75% vs. 43.07%) with comparable validity (87.60% vs. 87.51%). Thus, parameter count alone does not determine inference efficiency: the computational structure of the decoder has a substantial effect on both sampling cost and predictive performance.

Composition-prior temperature. Table 15 shows that sampling temperature has a clear but non-monotonic effect on discovery quality. Performance improves as τ increases from 0.6 to 1.2, where CoMMat attains the best SUN in both Discovery (11.22%) and Realism (11.00%) modes. Increasing the temperature further to 2.0 degrades performance in both settings, indicating that overly diffuse composition sampling reduces the likelihood of generating stable, unique, and novel structures. The effect is particularly pronounced in Realism mode, where SUN increases from 6.87% at $\tau = 0.6$ to 11.00% at $\tau = 1.2$. These results suggest that a moderately elevated composition-prior temperature improves exploratory diversity without sacrificing generation quality, with $\tau = 1.2$ providing the best operating point under the fixed 32-NFE refinement budget.

Algorithm 2 POPULATIONMATCH: coordinate moment matching.

Require: Fractional coordinates $\mathbf{X}$, fixed species $\mathbf{A}$, conditioning lattice $\bar{\mathbf{L}}$, structural latent $\mathbf{z}$
Require: $M = 4$, $\sigma_{\text{data}} = 0.20$, $p_{\text{mean}} = -1.1$, $p_{\text{std}} = 2.0$, $k = 12$, kernel-scale clip 50

Construct the centered periodic representative

$$\mathbf{X}_c \leftarrow \left(\mathbf{X} + \frac{1}{2}\right) \bmod 1 - \frac{1}{2}$$

Compute the characteristic lattice scale

$$\ell(\bar{\mathbf{L}}) \leftarrow \frac{1}{3}\sum_{j=1}^{3} \|\bar{\mathbf{l}}_j\|_2$$

Construct the normalized Cartesian clean state

$$\mathbf{x}_0 \leftarrow \frac{\mathbf{X}_c \bar{\mathbf{L}}}{\ell(\bar{\mathbf{L}})}$$

Partition the minibatch into groups of M crystals
for each group g **do**
  Sample n_t from the truncated log-normal distribution in Eq. 24
  Sample n_s from the same distribution subject to $n_s \le n_t$
  $\Delta n \leftarrow (n_T - n_\epsilon)2^{-k}$
  $n_r \leftarrow \max(n_t - \Delta n, n_\epsilon)$
  Enforce $n_s \le n_r \le n_t$
  Convert $n_u \mapsto u = n_u/(1+n_u)$ for $u \in \{s, r, t\}$
  Draw shared Gaussian noise

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

  Construct student and teacher states

$$\mathbf{x}_t \leftarrow (1-t)\mathbf{x}_0 + t\,\sigma_{\text{data}}\boldsymbol{\epsilon}$$

$$\mathbf{x}_r \leftarrow (1-r)\mathbf{x}_0 + r\,\sigma_{\text{data}}\boldsymbol{\epsilon}$$

  Convert each normalized Cartesian state to fractional coordinates

$$\mathbf{X}_u \leftarrow \left[\ell(\bar{\mathbf{L}})\mathbf{x}_u\bar{\mathbf{L}}^{-1}\right] \bmod 1, \qquad u \in \{r, t\}$$

  Predict the fractional coordinate fields

$$\Delta\mathbf{X}_{\theta,t} \leftarrow \Delta\mathbf{X}_\theta\left(\mathbf{X}_t, t \mid \mathbf{A}, \mathbf{z}, \bar{\mathbf{L}}\right)$$

$$\Delta\mathbf{X}_{\theta,r} \leftarrow \text{sg}\left[\Delta\mathbf{X}_\theta\left(\mathbf{X}_r, r \mid \mathbf{A}, \mathbf{z}, \bar{\mathbf{L}}\right)\right]$$

  Convert to normalized Cartesian fields

$$\mathbf{F}_{\theta,t} \leftarrow \frac{\Delta\mathbf{X}_{\theta,t}\bar{\mathbf{L}}}{\ell(\bar{\mathbf{L}})}$$

$$\mathbf{F}_{\theta,r} \leftarrow \frac{\Delta\mathbf{X}_{\theta,r}\bar{\mathbf{L}}}{\ell(\bar{\mathbf{L}})}$$

  Form student and stop-gradient teacher pushforwards

$$\mathbf{u}_{st} \leftarrow \mathbf{x}_t - (t-s)\sigma_{\text{data}}\mathbf{F}_{\theta,t}$$

$$\mathbf{u}_{sr} \leftarrow \text{sg}\left[\mathbf{x}_r - (r-s)\sigma_{\text{data}}\mathbf{F}_{\theta,r}\right]$$

  Construct empirical populations $P_{st}^{(g)}$ and $P_{sr}^{(g)}$ from the M student and teacher pushforwards
  Compute

$$w_k(t,s) \leftarrow \min\left(\frac{1}{|t-s|\sigma_{\text{data}}}, 50\right)$$

  Compute the V-statistic Laplace-kernel $\widehat{\text{MMD}}_g^2(P_{st}^{(g)}, P_{sr}^{(g)})$
end for
return

$$\mathcal{L}_{\text{IMM}}(\mathbf{z}) = \frac{1}{G}\sum_{g=1}^{G} w_{\text{out}}(t_g, s_g)\widehat{\text{MMD}}_g^2\left(P_{st}^{(g)}, P_{sr}^{(g)}\right)$$

Algorithm 3 COMMAT crystal structure prediction (CSP).

Require: Composition $\mathbf{A}$, atom count N, step count K
Require: $\epsilon = 0.006$, $T = 0.994$, $\sigma_{\text{data}} = 0.20$

Composition-conditioned initialization:
$(\boldsymbol{\mu}_p, \log \boldsymbol{\sigma}_p^2) \leftarrow p_\psi(\mathbf{z} \mid \mathbf{A})$
$\mathbf{z}_{\text{CSP}} \leftarrow \boldsymbol{\mu}_p(\mathbf{A})$ ▷ no target geometry encoded
$\hat{\mathbf{L}} \leftarrow h_L^{\text{dec}}(\mathbf{z}_{\text{CSP}}, N)$
$\mathbf{x}_{t_0} \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{data}}^2 \mathbf{I})$
Construct uniform grid $T = t_0 > t_1 > \cdots > t_K = \epsilon$

Few-step coordinate refinement:
for $k = 0, \ldots, K-1$ **do**
Compute

$$\ell(\hat{\mathbf{L}}) \leftarrow \frac{1}{3} \sum_{j=1}^{3} \|\hat{\mathbf{l}}_j\|_2$$

Convert to fractional coordinates

$$\mathbf{X}_{t_k} \leftarrow \left[\ell(\hat{\mathbf{L}}) \mathbf{x}_{t_k} \hat{\mathbf{L}}^{-1} \right] \bmod 1$$

Predict the normalized Cartesian field

$$\mathbf{F}_\theta \left(\mathbf{x}_{t_k}, t_k \mid \mathbf{A}, \mathbf{z}_{\text{CSP}}, \hat{\mathbf{L}} \right) \leftarrow \frac{\Delta \mathbf{X}_\theta \left(\mathbf{X}_{t_k}, t_k \mid \mathbf{A}, \mathbf{z}_{\text{CSP}}, \hat{\mathbf{L}} \right) \hat{\mathbf{L}}}{\ell(\hat{\mathbf{L}})}$$

Update

$$\mathbf{x}_{t_{k+1}} \leftarrow \mathbf{x}_{t_k} - (t_k - t_{k+1}) \sigma_{\text{data}} \mathbf{F}_\theta \left(\mathbf{x}_{t_k}, t_k \mid \mathbf{A}, \mathbf{z}_{\text{CSP}}, \hat{\mathbf{L}} \right)$$

end for
$\mathbf{X}_{\text{out}} \leftarrow \left[\ell(\hat{\mathbf{L}}) \mathbf{x}_{t_K} \hat{\mathbf{L}}^{-1} \right] \bmod 1$
return $(\mathbf{A}, \mathbf{X}_{\text{out}}, \hat{\mathbf{L}})$

Algorithm 4 CoMMAT de novo generation (DNG).

Require: Step count K, maximum composition length $N_{\max}$
Require: Formula temperature $\tau_A = 1$, latent temperature $\tau_z = 1$
Require: $\epsilon = 0.006$, $T = 0.994$, $\sigma_{\text{data}} = 0.20$

Generate composition:
Initialize the autoregressive sequence with BOS
while EOS has not been generated and sequence length $< N_{\max}$ **do**
    Predict
$$p_\psi\left(A_{(i)} \mid A_{(<i)}\right)$$
    Mask tokens violating the nondecreasing atomic-number ordering
    Sample $A_{(i)}$ at temperature τ_A
end while
Remove BOS/EOS and obtain
$$\mathbf{A} = (A_{(1)}, \ldots, A_{(N)})$$
$N \leftarrow |\mathbf{A}|$

Sample composition-conditioned structural latent:
$(\boldsymbol{\mu}_p, \log \boldsymbol{\sigma}_p^2) \leftarrow p_\psi(\mathbf{z} \mid \mathbf{A})$
$\mathbf{z}_{\text{DNG}} \leftarrow \boldsymbol{\mu}_p + \tau_z \boldsymbol{\sigma}_p \odot \boldsymbol{\epsilon}_z, \quad \boldsymbol{\epsilon}_z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\hat{\mathbf{L}} \leftarrow h_L^{\text{dec}}(\mathbf{z}_{\text{DNG}}, N)$

Generate coordinates:
$\mathbf{x}_{t_0} \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{data}}^2 \mathbf{I})$
Construct uniform grid $T = t_0 > t_1 > \cdots > t_K = \epsilon$
for $k = 0, \ldots, K-1$ **do**
    Compute
$$\ell(\hat{\mathbf{L}}) \leftarrow \frac{1}{3} \sum_{j=1}^{3} \|\hat{\mathbf{l}}_j\|_2$$
    Convert to fractional coordinates
$$\mathbf{X}_{t_k} \leftarrow \left[\ell(\hat{\mathbf{L}}) \mathbf{x}_{t_k} \hat{\mathbf{L}}^{-1}\right] \bmod 1$$
    Predict the normalized Cartesian field
$$\mathbf{F}_\theta\left(\mathbf{x}_{t_k}, t_k \mid \mathbf{A}, \mathbf{z}_{\text{DNG}}, \hat{\mathbf{L}}\right) \leftarrow \frac{\Delta \mathbf{X}_\theta\left(\mathbf{X}_{t_k}, t_k \mid \mathbf{A}, \mathbf{z}_{\text{DNG}}, \hat{\mathbf{L}}\right) \hat{\mathbf{L}}}{\ell(\hat{\mathbf{L}})}$$
    Update
$$\mathbf{x}_{t_{k+1}} \leftarrow \mathbf{x}_{t_k} - (t_k - t_{k+1}) \sigma_{\text{data}} \mathbf{F}_\theta\left(\mathbf{x}_{t_k}, t_k \mid \mathbf{A}, \mathbf{z}_{\text{DNG}}, \hat{\mathbf{L}}\right)$$
end for
$\mathbf{X}_{\text{out}} \leftarrow \left[\ell(\hat{\mathbf{L}}) \mathbf{x}_{t_K} \hat{\mathbf{L}}^{-1}\right] \bmod 1$
return $(\mathbf{A}, \mathbf{X}_{\text{out}}, \hat{\mathbf{L}})$