Optimal Crystal Flow for Fast and Reliable Crystal Generation

Anonymous authors
Paper under double-blind review

Abstract

De novo crystal generation aims to discover novel materials by exploring the vast combinatorial space of atomic species and lattice geometries. While recent models based on diffusion and flow matching have shown strong generative capabilities, they often require a large number of sampling steps. More importantly, they struggle to generate reliable structures in out-of-distribution regimes, especially for systems larger than those seen during training. To address these limitations, we introduce **Optimal Crystal Flow** (OCFlow), a flow matching framework that reduces unnecessarily long and curved coordinate paths by resolving *intra-sample* ambiguity within each crystal. Experiments on the MP-20 and MPTS-52 benchmarks show that OCFlow is superior to or competitive with state-of-the-art baselines in sample validity and stability, with particularly strong performance on larger out-of-distribution systems while remaining effective with as few as ten network function evaluations. Together, these results show that per-structure optimal couplings provide both efficient sampling and robust generalization in crystal generation.

1 Introduction

Designing novel crystal structures with desired functional properties has been a long-standing goal in materials science (Maddox, 1988), motivating large-scale efforts to catalog and screen candidate materials (Jain et al., 2013). Yet identifying geometrically valid and chemically plausible configurations requires navigating an enormous joint space of compositions and atomic geometries, where rule-based search is infeasible at scale. Generative models address this search bottleneck by learning from known materials a distribution from which plausible structures can be sampled directly.

Diffusion and flow matching (FM) (Ho et al., 2020; Lipman et al., 2023) have recently emerged as powerful frameworks for learning complex data distributions. In crystal generation, early models such as CDVAE (Xie et al., 2022) and DiffCSP (Jiao et al., 2023) extended deep generative modeling to periodic structures, enabling the joint generation of lattice, fractional coordinates, and atomic species. Recent studies formulate crystal generation with continuous-time dynamics for lattice and coordinates, together with discrete diffusion (Austin et al., 2021) or FM (Campbell et al., 2024) for atomic species (Miller et al., 2024; Zeni et al., 2025). Building upon these advances, subsequent studies explored more efficient representations of the atomic compositional space (Luo et al., 2025), the use of lattice symmetry or space-group information (Jiao et al., 2024; Kazeev et al., 2025; Levy et al., 2025), and latent-space diffusion to further improve generation quality (Joshi et al., 2025).

Despite this progress, two practical challenges remain for applying crystal generators to materials discovery: (i) the inference-time sampling cost, typically measured by the number of network function evaluations (NFEs), and (ii) generalization beyond the training distribution, particularly to larger structures. First, existing methods often require hundreds to thousands of network evaluations to generate high-quality samples, limiting the throughput required for large-scale materials screening. Because each additional sampling step requires further network evaluations, the computational cost grows with the number of steps and can become a bottleneck when generating large candidate sets for high-throughput materials screening. Flow-based approaches (Miller et al., 2024; Luo et al., 2025) can reduce this budget to around 100 NFEs. While this improves sampling efficiency, sample stability remains limited, and generation quality at substantially lower budgets, such as 10 NFEs, remains largely unexplored.

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.

Second, materials discovery often requires generating structures beyond the training distribution, particularly systems larger than those covered by the dataset. Although recent models (Jiao et al., 2024; Miller et al., 2024; Luo et al., 2025; Joshi et al., 2025) achieve high uniqueness and novelty, we observe a sharp drop in sample validity together with large structural changes during relaxation once the system size exceeds the training range. Based on these observations, we hypothesize that inefficient coordinate path construction contributes to both high sampling costs and limited extrapolation to larger systems.

In the FM framework, the coupling between prior and data endpoints determines the conditional paths used for training. Random prior-data pairings can produce crossing paths and curved marginal trajectories, making accurate integration with few NFEs more challenging. Rectified Flow (Liu et al., 2023) iteratively refines the prior-data coupling, while minibatch optimal transport (Tong et al., 2024) selects prior-data pairs of lower cost within each batch. Both approaches improve how prior and data samples are paired, but leave the atom-level correspondence within each sample unresolved.

For crystal structures, even within a given prior-data pair, periodic images, global translations, and atom orderings can represent the same physical crystal. Different representations can therefore induce different correspondences between atomic sites and different coordinate paths. We refer to this ambiguity within each crystal as *intra-sample ambiguity*. Arbitrary site correspondences can induce unnecessarily long and variable coordinate paths, making few-step integration more difficult and potentially limiting extrapolation towards larger systems. To resolve this issue, we introduce **Optimal Crystal Flow** (OCFlow), an FM framework that targets this intra-sample ambiguity through *local optimal couplings* (OCs). In each prior-data pair, we match coordinate sites by minimizing their total squared minimum-image displacement, then remove the common translation, which is performed independently within each structure.

Our contributions are threefold. First, we introduce torus-aware per-structure OC for coordinate flow matching, directly addressing the absence of a canonical atom order in periodic crystals. Second, we show that such coupling approach improves the speed-quality trade-off and is particularly useful for out-of-distribution (OOD) generation, where it achieves substantially improved validity, convergence, and hull-based stability of the generated structures. Third, we provide a systematic few-step benchmark for *de novo* crystal generation under a common validity and relaxation pipeline based on machine learning interatomic potentials (MLIPs), complemented by hull-based stability estimates and compositional analysis of the retained candidates.

2 PRELIMINARIES

This section first reviews the representation of crystal structures and the FM formulation used throughout this work. We then discuss the role of endpoint couplings and the crystal-specific intra-sample ambiguity that motivates our method.

2.1 CRYSTAL REPRESENTATION

A crystal structure is represented as $(\mathbf{L}, \mathbf{X}, S)$, where $\mathbf{L}$ defines the lattice basis, $\mathbf{X}$ gives the atomic positions in fractional coordinates relative to this basis, and S denotes the atomic species. The lattice $\mathbf{L}$ is represented either by a 3×3 matrix or by the six unit-cell parameters $(a, b, c, \alpha, \beta, \gamma)$. For a crystal with N atoms, the fractional coordinates $\mathbf{X} \in \mathbb{T}^{N\times 3}$ lie on the flat torus $\mathbb{T}^3 = \mathbb{R}^3/\mathbb{Z}^3$, while the species assignments $S \in [Z]^N$ encode discrete atomic labels, with $[Z] \equiv \{1, \ldots, Z\}$ and Z denoting the number of possible atomic species.

Accordingly, crystal generation operates on a joint space combining Euclidean continuous variables, periodic continuous variables, and categorical variables. Existing generators based on diffusion or FM differ in the dynamics assigned to each component while operating on the same joint space. In a general FM framework, a network ϕ_θ receives the noisy triple $(\mathbf{L}_t, \mathbf{X}_t, S_t)$ together with time t as input and predicts the corresponding outputs for the three channels:

$$\big(v_\theta^{\mathbf{L}},\ v_\theta^{\mathbf{X}},\ v_\theta^{S}\big)\ =\ \phi_\theta\big(\mathbf{L}_t, \mathbf{X}_t, S_t,\ t\big), \tag{1}$$

where $v_\theta^{\mathbf{L}}$ and $v_\theta^{\mathbf{X}}$ denote the marginal velocity fields on the Euclidean lattice space and on the torus, respectively. v_θ^{S} represents a per-atom predictive distribution over the species labels in $[Z]$.

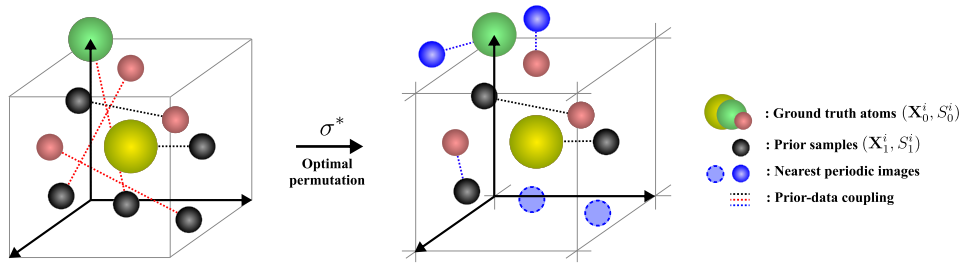


Figure 1: **Comparison of coupling schemes for crystal generation. (a) Random coupling. (b) Per-structure local OC (Ours).** Using an ABX_3 perovskite as an example, we illustrate how prior sites $(\mathbf{X}_1, \mathbf{S}_1)$ are matched to data sites $(\mathbf{X}_0, \mathbf{S}_0)$ by solving the local OC under the periodic minimum-image distance. The blue sites denote the nearest periodically equivalent representatives of the matched prior sites, and the blue dashed lines indicate the resulting optimal correspondences, replacing the random correspondences shown in red.

2.2 Flow Matching and Endpoint Couplings

Let $x_t \sim p_t$ denote a sample at time $t \in [0,1]$ from the marginal distribution p_t along the flow. Throughout the paper, we use p_0 for the data distribution and p_1 for a simple prior on $\mathbb{R}^d$. Given a coupling $\pi \in \Pi(p_0, p_1)$, FM learns a time-dependent velocity field $v_\theta : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ by regressing onto the velocity field induced by a prescribed conditional path between endpoint pairs $(x_0, x_1) \sim \pi$ (Lipman et al., 2023). For the linear path $x_t = (1-t)x_0 + tx_1$, the target velocity becomes $u_t(x_0, x_1) = x_1 - x_0$, and the training loss is given by:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t\sim U[0,1],\,(x_0,x_1)\sim\pi}\left[\|v_\theta(x_t,t) - u_t(x_0,x_1)\|_2^2\right]. \tag{2}$$

The coupling π determines which prior endpoint is connected to which data endpoint, therefore directly determines the conditional paths seen during training. If x_0 and x_1 are paired independently, conditional paths can cross even when each individual path is linear, making the marginal velocity field harder to approximate with few integration steps. Rectified Flow (Liu et al., 2023) refines endpoint couplings, while optimal-transport conditional flow matching (OT-CFM) (Tong et al., 2024) uses transport assignments to reduce this effect. In minibatch OT-CFM, data and prior batches $\{x_0^b\}_{b=1}^B$ and $\{x_1^b\}_{b=1}^B$ are paired by

$$\pi^\star = \arg\min_{\pi\in S_B} \sum_{b=1}^{B} \left\| x_0^b - x_1^{\pi(b)} \right\|_2^2, \tag{3}$$

where B is the batch size and S_B is the permutation group on $\{1,\dots,B\}$. This minibatch assignment solves an empirical optimal transport problem between the data and prior samples within each batch. As an *inter-sample* operation, it determines which data sample is paired with which prior sample.

2.3 Intra-sample Ambiguity in Crystals

Even after selecting a prior-data pair, crystals retain an additional *intra-sample* ambiguity. This ambiguity primarily affects the fractional coordinates $\mathbf{X}$, since the lattice has no degree of freedom on atom orderings. A crystal admits multiple equivalent coordinate representations: each fractional coordinate can be shifted by an integer lattice vector, all atoms can be translated by a common offset on the torus, and atom indices can be relabeled by applying the same permutation to the coordinates and their associated species labels. Consequently, physically equivalent representations can induce different correspondences between prior and data sites, leading to unnecessarily long conditional paths, as illustrated in Figure 1. Existing flow-based crystal generators (Miller et al., 2024; Luo et al., 2025) account for periodicity when computing coordinate displacements between paired sites, but the stored atom order, which has no physical significance, still determines which prior atom is connected to which data atom. Our method instead optimizes the site correspondence within each training pair under the periodic minimum-image cost, which we detail in the next section.

3 METHOD

OCFlow jointly models the lattice, fractional coordinates, and atomic species using a shared network ϕ_θ, while constructing a local OC for the coordinate channel of each training pair. The lattice and fractional coordinates share a structural time $t \in [0, 1]$, while the atomic species use an independent time $s \in [0, 1]$. We first describe the coordinate dynamics induced by the local OC, followed by the lattice and species dynamics and the sampling procedure.

3.1 COORDINATE DYNAMICS WITH LOCAL OPTIMAL COUPLING

To resolve the intra-sample ambiguity described in the previous section, we construct the coordinate target of each training pair in three steps: (1) we compute periodic minimum-image costs between all data and prior sites, (2) choose the site assignment that minimizes the total cost, and (3) remove the common translation of the matched displacements.

Fractional Coordinates. For a training pair with N atoms, let $\mathbf{X}_0, \mathbf{X}_1 \in \mathbb{T}^{N\times 3}$ denote the data and sampled prior fractional coordinates, respectively. For $z \in \mathbb{R}^3$, we define the wrapping operator on individual sites as $\Phi(z) \equiv z - \mathrm{round}(z)$ with $\Phi(z) \in [-1/2, 1/2]^3$. For an arbitrary data site i and prior site j, we define the minimum-image displacement and its squared cost as

$$\Delta_{ij} = \Phi(\mathbf{X}_{1,j} - \mathbf{X}_{0,i}), \qquad C_{ij} = \|\Delta_{ij}\|_2^2. \tag{4}$$

We choose the site correspondence using the Hungarian algorithm (Kuhn, 1955)

$$\sigma^\star = \arg\min_{\sigma\in S_N} \sum_{i=1}^{N} C_{i,\sigma(i)}, \tag{5}$$

where S_N is the permutation group on $\{1, \ldots, N\}$, and $\sigma^\star(i)$ is the prior site assigned to data site i. Following Miller et al. (2024), we center the matched minimum-image displacements by subtracting their mean:

$$\widehat{\mathbf{U}}_i = \Delta_{i,\sigma^\star(i)}, \qquad \overline{\mathbf{U}} = \frac{1}{N}\sum_{i=1}^{N} \widehat{\mathbf{U}}_i, \qquad \mathbf{U}_{\mathbf{X},i} = \widehat{\mathbf{U}}_i - \overline{\mathbf{U}}. \tag{6}$$

For this fixed assignment and choice of periodic images, $\tau^\star = -\overline{\mathbf{U}}$ minimizes $\sum_{i=1}^{N} \|\widehat{\mathbf{U}}_i + \tau\|_2^2$ over a common translation $\tau \in \mathbb{R}^3$. Then, the coordinate path used for training becomes

$$\widetilde{\mathbf{X}}_t = \mathbf{X}_0 + t\mathbf{U}_{\mathbf{X}}, \qquad \bar{\mathbf{X}}_t = \Phi(\widetilde{\mathbf{X}}_t), \tag{7}$$

where Φ is applied to each atom coordinate. The wrapped coordinates $\bar{\mathbf{X}}_t$ are provided to the network, and $\mathbf{U}_{\mathbf{X}}$ becomes the coordinate velocity target. Finally, we train the coordinate velocity using Huber loss (Huber, 1964) against the aligned velocity target,

$$\mathcal{L}_{\mathbf{X}}(\theta) = \mathbb{E}\left[\ell_{\mathrm{Huber}}\left(v_\theta^{\mathbf{X}}(\mathbf{L}_t, \bar{\mathbf{X}}_t, S_s, t, s) - \mathbf{U}_{\mathbf{X}}\right)\right], \tag{8}$$

where $v_\theta^{\mathbf{X}}$ denotes the predicted coordinate velocity of the joint network ϕ_θ from Eq. (1). At the prior endpoint, all species identities are masked, so the sampled prior coordinates do not carry species information for matching. Therefore, the coordinate matching is performed without species constraints in unconditional generation. The resulting permutation reindexes only the sampled prior coordinates $\mathbf{X}_1$, while the data coordinates $\mathbf{X}_0$ and species labels S_0 retain their original indexing.

3.2 LATTICE AND SPECIES DYNAMICS

Lattice. We parameterize the lattice by the upper-triangular Cholesky factor of its metric tensor, which encodes the six cell parameters $(a, b, c, \alpha, \beta, \gamma)$. We use a log-normal prior $p_1^{\mathbf{L}}$ matched to the dataset's empirical lattice statistics, following Miller et al. (2024). The conditional path is the linear interpolant

$$\mathbf{L}_t = (1-t)\mathbf{L}_0 + t\mathbf{L}_1, \qquad U_{\mathbf{L}} = \mathbf{L}_1 - \mathbf{L}_0, \tag{9}$$

where $\mathbf{L}_0$ and $\mathbf{L}_1$ denote the data and prior lattice representations, respectively, and the lattice loss is

$$\mathcal{L}_{\mathbf{L}}(\theta) = \mathbb{E}\left[\left\|v_\theta^{\mathbf{L}}(\mathbf{L}_t, \bar{\mathbf{X}}_t, S_s, t, s) - U_{\mathbf{L}}\right\|_2^2\right]. \tag{10}$$

Atomic Species. The composition evolves on the augmented state space $\{1, \ldots, Z\} \cup \{\mathbf{m}\}$, where $\mathbf{m}$ is an absorbing mask token. We use an absorbing-mask categorical corruption process (Austin et al., 2021) together with discrete FM (Campbell et al., 2024). At $s = 0$, each atom carries its data label $S_{0,i}$, while at $s = 1$ all atoms are masked. During training, each site retains its data label with probability $1 - s$ and is replaced by the mask token with probability s. During generation, masked sites are progressively unmasked as s decreases. When a masked site i is selected for unmasking, its species is sampled from the categorical distribution predicted by the species head,

$$\hat{S}_{0,i} \sim \mathrm{Cat}\big(v_\theta^S(\mathbf{L}_t, \bar{\mathbf{X}}_t, S_s, t, s)_i\big)\,. \tag{11}$$

Once a site is unmasked, its sampled species remains fixed for the rest of the generation process.

The species head v_θ^S returns a categorical distribution over $\{1, \ldots, Z\}$ for each masked atom, predicting its data label. Finally, the species loss is the masked cross-entropy

$$\mathcal{L}_S(\theta) = \mathbb{E}\left[\sum_{i\,:\,S_{s,i}=\mathbf{m}} -\log v_\theta^S(\mathbf{L}_t, \bar{\mathbf{X}}_t, S_s, t, s)_{i,\,S_{0,i}}\right]. \tag{12}$$

Total objective. The lattice and coordinate channels share the time variable t, while the species channel uses an independent time s. The full training loss is a weighted sum of the per-channel objectives,

$$\mathcal{L}(\theta) = \mathbb{E}_{t,s}\left[\lambda_{\mathbf{X}}\mathcal{L}_{\mathbf{X}}(\theta) + \lambda_{\mathbf{L}}\mathcal{L}_{\mathbf{L}}(\theta) + \lambda_S\mathcal{L}_S(\theta)\right], \tag{13}$$

where we use scalar loss weights $\lambda_{\mathbf{X}} : \lambda_{\mathbf{L}} : \lambda_S = 100 : 2 : 1$ across all datasets and experiments. A single network ϕ_θ is shared across the three channels, with the structural velocities and species posterior obtained as in Eq. (1). The training distributions for t and s are detailed in Appendix A.

3.3 Flow-Based Crystal Sampling

At inference time, we integrate the lattice and coordinate flows from $t = 1$ to 0 using a fixed-step Euler scheme. The species channel starts from the fully masked state at $s = 1$ and is progressively unmasked toward $s = 0$ using the discrete transition rule and the categorical predictions of the species head. Following the inference anti-annealing of FlowMM (Miller et al., 2024), we rescale the structural velocities during sampling by $c(1 + r\tau)$, where $\tau = 1 - t$. The coordinate and lattice channels use separate values of c and r (Appendix A.2). Each sampling step requires one network evaluation. The Hungarian assignment is used only during training and introduces a marginal training-time overhead, with no additional assignment cost during sampling. Detailed runtime measurements are provided in Tables 6 and 7.

4 Experiments and Results

We evaluate how the local optimal coordinate paths in OCFlow influence *de novo* crystal generation in two regimes: in-distribution (ID) generation with the number of atoms sampled from the training distribution, and OOD generation on systems larger than those seen during training. For each regime, we compare OCFlow against five state-of-the-art baselines with public implementations on MP-20 (Xie et al., 2022; Jain et al., 2013) and four on MPTS-52 (Baird et al., 2024) datasets. Details on training and sampling are provided in Appendix A, with additional results and complete tables in Appendix C.

4.1 Experimental Setup

Backbone architecture. We construct the denoiser ϕ_θ as a 3-layer GemNet (Gasteiger et al., 2021) with hidden and edge embedding dimensions of 512, totaling roughly 43M parameters. This backbone is shared across the lattice, coordinate, and species channels, and we use the same configuration for both benchmarks. Dataset-specific learning rates and training budgets are provided in Appendix A.

Datasets. We train on MP-20 (Xie et al., 2022; Jain et al., 2013), the Materials Project subset with at most 20 atoms per cell, and on MPTS-52 (Baird et al., 2024), which extends the same source to 52 atoms with a chronological train and test split. In the ID setting, models trained on either dataset

Table 1: **Results on ID generation.** We report validity, uniqueness and novelty, energetic stability, and relaxation metrics, with rates given in percentages, E_f in eV/atom, and ΔS in Å. In the NFE column, $\times 2$ indicates two network evaluations per sampling step due to predictor-corrector sampling. Values in parentheses denote the standard deviation across three independent runs. The best and second-best values are highlighted in **bold** and underline, respectively.

Model	NFE	Validity		Novelty	Stability	Relaxation	
		Struct. (↑)	Comp. (↑)	UN (845K) (↑)	E_f (↓)	ΔS (↓)	%conv. (↑)
Trained on MP-20							
MatterGen	1000×2	99.84(0.06)	84.02(0.52)	78.48(0.58)	-0.767(0.012)	0.252(0.007)	97.50(0.22)
MatterGen	100×2	99.88(0.05)	82.08(0.55)	84.52(0.51)	-0.683(0.012)	0.504(0.012)	96.04(0.27)
MatterGen	10×2	99.90(0.04)	80.72(0.57)	**91.78**(0.39)	-0.479(0.012)	1.094(0.012)	94.42(0.33)
DiffCSP++	1000	99.82(0.06)	83.56(0.52)	72.48(0.63)	-0.504(0.012)	0.316(0.012)	91.90(0.39)
ADiT	250	99.24(0.12)	**88.84**(0.44)	69.68(0.70)	-0.511(0.015)	0.435(0.011)	**98.62**(0.16)
FlowMM	250	99.44(0.11)	82.40(0.54)	82.88(0.53)	-0.392(0.013)	0.980(0.012)	95.50(0.29)
CrystalFlow	100	99.66(0.06)	81.55(0.39)	89.64(0.43)	0.073(0.014)	1.285(0.011)	95.62(0.29)
OCFlow (ours)	1000	**99.98**(0.03)	88.23(0.42)	73.47(0.25)	-1.053(0.020)	**0.171**(0.007)	97.83(0.45)
OCFlow (ours)	100	99.97(0.06)	88.11(0.75)	73.39(0.27)	**-1.060**(0.018)	0.234(0.001)	97.50(0.40)
OCFlow (ours)	**10**	99.97(0.03)	87.35(0.64)	84.25(0.52)	-0.970(0.007)	0.542(0.005)	96.19(0.12)
Trained on MPTS-52							
MatterGen	1000×2	99.42(0.11)	72.38(0.63)	87.12(0.47)	-0.603(0.011)	0.533(0.016)	83.90(0.52)
DiffCSP++	1000	99.58(0.09)	77.75(0.59)	86.58(0.48)	-0.272(0.012)	0.486(0.015)	80.60(0.56)
FlowMM	250	99.16(0.13)	69.90(0.65)	92.44(0.37)	-0.192(0.023)	1.335(0.012)	84.34(0.51)
CrystalFlow	100	99.16(0.13)	72.78(0.63)	**98.56**(0.17)	0.848(0.023)	1.735(0.012)	84.84(0.51)
OCFlow (ours)	1000	**99.99**(0.02)	**80.53**(0.77)	78.45(0.55)	**-1.131**(0.015)	0.465(0.004)	**90.56**(0.21)
OCFlow (ours)	100	99.98(0.02)	78.54(0.57)	84.60(0.33)	-1.091(0.020)	**0.421**(0.010)	89.94(0.37)
OCFlow (ours)	**10**	99.96(0.02)	76.78(0.66)	90.75(0.61)	-0.919(0.019)	0.786(0.013)	84.10(0.53)

draw the number of atoms N from the corresponding empirical training distribution, following FlowMM (Miller et al., 2024). To evaluate models trained on MP-20 in the OOD setting, we use the subset of MPTS-52 whose unit cells contain more than 20 atoms. For the MPTS-52 OOD setting, we instead sample N uniformly from $[53, 70]$, since the training distribution does not cover this range.

Evaluation metrics. We report structural and compositional validity (Struct., Comp.), the fraction of unique and novel (UN) samples, energetic stability (E_f), and structural changes under local relaxation (ΔS, %conv.). Structural validity requires all interatomic distances within a structure to exceed 0.5 Å, while compositional validity checks charge neutrality using SMACT (Davies et al., 2019). UN measures the fraction of distinct generated structures that do not match any of the 845,997 structures in the Alex-MP reference set (Zeni et al., 2025; Schmidt et al., 2022), as determined by Pymatgen's `StructureMatcher` (Ong et al., 2013).

To evaluate ΔS and %conv., we perform structural relaxation using `MACE-MPA-0` (Batatia et al., 2022; 2025), a widely used MLIP model, for up to 500 steps. The minimum-image root-mean-square deviation (RMSD) ΔS (Å) measures the structural deviation between each generated structure and its relaxed counterpart, while %conv. denotes the fraction of converged relaxations. A smaller ΔS indicates that the generated structure requires less adjustment during relaxation. For energetic stability, we report the mean formation energy E_f, following LeMat-GenBench (Betala et al., 2025). Additional details on evaluation metrics are provided in Appendix B.

4.2 In-distribution Generation

We first report the performance on in-distribution generation in Table 1. On both MP-20 and MPTS-52, OCFlow achieves the highest structural validity and the lowest ΔS, indicating that its generated structures are geometrically valid and require relatively little rearrangement to reach a relaxed local configuration. This trend is also reflected in formation energy, with OCFlow consistently achieving the lowest E_f across all evaluated sampling budgets. At 1,000 NFEs, its compositional validity and convergence further reach the best or second-best performance on both benchmarks, whereas its UN rate is lower than that of most baselines.

Table 2: **Results on OOD generation.** We report validity, energetic stability, and relaxation metrics, with rates given in percentages, E_f in eV/atom, and ΔS in Å. In the NFE column, $\times 2$ indicates two network evaluations per sampling step due to predictor-corrector sampling. Values in parentheses denote the standard deviation across three independent runs. The best and second-best values are highlighted in **bold** and underline, respectively.

Model	NFE	Validity		Stability	Relaxation	
		Struct. (↑)	Comp. (↑)	E_f (↓)	ΔS (↓)	%conv. (↑)
Trained on MP-20, evaluated on $N \in [21, 52]$						
MatterGen	1000×2	**100.00** (0.00)	66.08 (0.67)	-1.514 (0.010)	0.593 (0.007)	91.90 (0.39)
MatterGen	100×2	**100.00** (0.00)	63.04 (0.69)	-1.332 (0.010)	1.083 (0.008)	86.34 (0.49)
MatterGen	10×2	99.42 (0.11)	53.06 (0.71)	-0.814 (0.010)	1.616 (0.007)	75.80 (0.61)
DiffCSP++	1000	97.82 (0.21)	15.56 (0.51)	-0.580 (0.008)	1.653 (0.008)	76.20 (0.60)
ADiT	250	91.90 (0.39)	58.46 (0.70)	-0.023 (0.031)	2.095 (0.021)	63.00 (0.67)
FlowMM	250	74.30 (0.62)	20.90 (0.58)	1.168 (0.130)	3.392 (0.020)	69.04 (0.65)
CrystalFlow	100	92.54 (0.37)	42.48 (0.70)	0.373 (0.040)	2.168 (0.013)	70.24 (0.65)
OCFlow (ours)	1000	**100.00** (0.00)	**79.30** (0.61)	**-1.774** (0.016)	**0.544** (0.013)	93.60 (0.75)
OCFlow (ours)	100	**100.00** (0.00)	79.02 (0.69)	-1.716 (0.011)	0.576 (0.008)	**93.93** (0.36)
OCFlow (ours)	**10**	**100.00** (0.00)	78.75 (0.48)	-1.477 (0.006)	0.894 (0.007)	90.16 (0.48)
Trained on MPTS-52, evaluated on uniform $N \in [53, 70]$						
MatterGen	1000×2	98.87 (0.19)	57.27 (0.90)	-0.774 (0.012)	1.586 (0.011)	19.33 (0.72)
DiffCSP++	1000	99.73 (0.09)	19.33 (0.72)	-0.378 (0.011)	1.638 (0.016)	10.55 (0.56)
FlowMM	250	59.93 (0.89)	20.83 (0.74)	2.128 (0.499)	2.950 (0.024)	24.03 (0.78)
CrystalFlow	100	91.37 (0.51)	**66.80** (0.86)	0.500 (0.045)	2.261 (0.019)	31.97 (0.85)
OCFlow (ours)	1000	**100.00** (0.00)	63.76 (0.20)	**-1.130** (0.016)	**1.066** (0.006)	**50.37** (0.78)
OCFlow (ours)	100	99.98 (0.02)	64.43 (0.58)	-1.108 (0.018)	1.103 (0.013)	40.45 (1.20)
OCFlow (ours)	**10**	99.76 (0.09)	50.64 (0.87)	-1.017 (0.017)	1.412 (0.014)	23.86 (0.90)

More importantly, OCFlow largely retains this performance in the few-step regime. Even at only 10 NFEs, it maintains nearly perfect structural validity on both benchmarks while its compositional validity and convergence remain comparable to those of baselines using substantially larger sampling budgets. At this budget, OCFlow still attains the highest structural validity and the lowest E_f among the evaluated methods. These results show that OCFlow preserves strong generation quality even under a substantially reduced sampling budget.

4.3 Out-of-distribution Generation

We next evaluate extrapolation to structures with atom counts beyond the training distribution in Table 2. Moving beyond the training range causes substantial degradation for most baselines, particularly in validity metrics and the structural changes required during relaxation. In contrast, OCFlow consistently shows strong performance across OOD settings on both benchmarks. At 100 NFEs, it retains nearly perfect structural validity on both OOD benchmarks while achieving lower E_f and ΔS than all evaluated baselines. On MP-20 OOD, it additionally maintains substantially higher compositional validity than the competing methods.

The advantage becomes particularly evident in the few-step regime. With only 10 NFEs, OCFlow still achieves 100% structural validity and 78.75% compositional validity on MP-20 OOD, while its ΔS remains lower than that of every baseline except MatterGen using 2,000 network evaluations. On the more challenging MPTS-52 OOD setting, OCFlow at 10 NFEs achieves 99.76% structural validity together with lower E_f and ΔS than all evaluated baselines, despite using $10\times$-$200\times$ fewer network evaluations. Overall, these results extend the few-step robustness observed in Section 4.2 to substantially larger out-of-distribution systems.

4.4 Few-step Robustness and Stability

Beyond the sampling budgets reported in Tables 1 and 2, we examine how generation quality evolves over a broader range of NFEs on MP-20 in Figure 2. In the ID setting, OCFlow achieves high validity

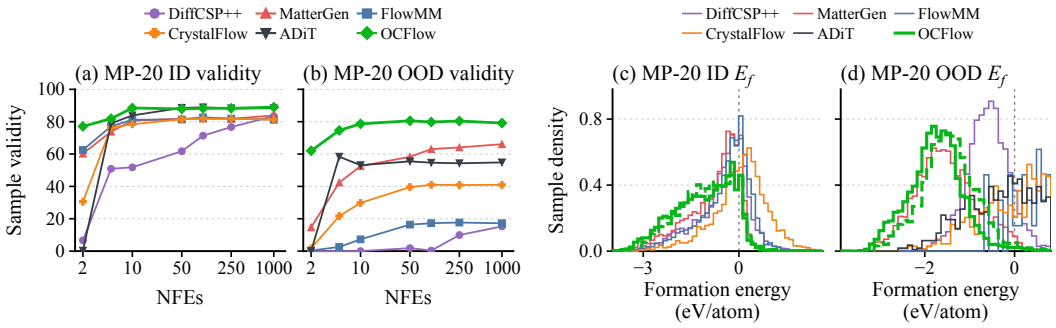


Figure 2: **Few-step robustness on MP-20.** (a, b) Fraction of both structurally and compositionally valid samples across sampling budgets in the ID and OOD settings. (c, d) Formation energy distributions of generated samples in the corresponding ID and OOD settings.

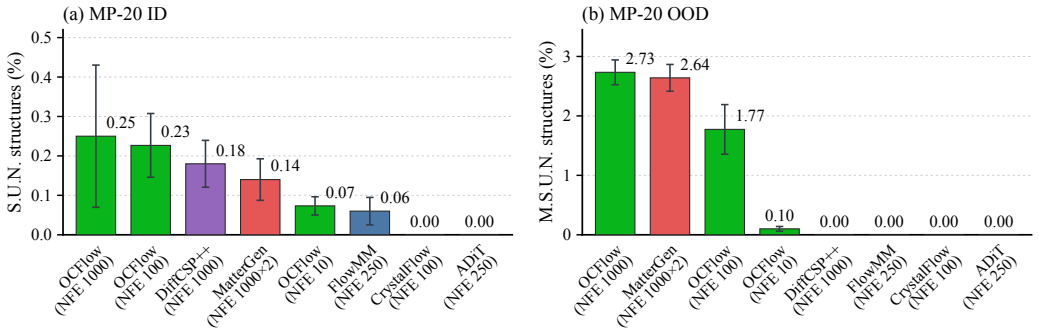


Figure 3: **Hull-based stability on MP-20.** (a) S.U.N. rates for ID generation and (b) M.S.U.N. rates for OOD generation. At 100 NFEs, OCFlow retains an ID S.U.N. rate of 0.23%, close to 0.25% at 1,000 NFEs. In the OOD setting, it still produces M.S.U.N. candidates with only 10 NFEs. Error bars indicate standard deviations.

with few NFEs and improves modestly as the sampling budget increases. The gap to the baselines becomes more evident under OOD generation, where OCFlow maintains high validity across the NFE range while several baselines degrade substantially at smaller sampling budgets.

The formation energy distributions in Figure 2 (c) and (d) show a consistent trend. In both ID and OOD settings, OCFlow generates samples concentrated at lower formation energies than the competing methods, indicating that its robustness at small sampling budgets is not limited to geometric validity.

We further examine whether these improvements translate into stable, unique, and novel (S.U.N.) candidates using the hull-based evaluation following LeMat-GenBench (Betala et al., 2025), and summarize the results in Figure 3. In the ID setting, OCFlow achieves the highest S.U.N. rates at both 1,000 and 100 NFEs. For OOD generation, where no method produces samples satisfying the strict S.U.N. criterion, OCFlow achieves the highest metastable S.U.N. (M.S.U.N.) rate at 1,000 NFEs and remains non-zero even at 10 NFEs. These results show that the gains in validity and formation energy persist under a stricter hull-based screening criterion, although the absolute yield of stable candidates remains small.

Effect of Optimal Coupling in Few-step Generation. To isolate the contribution of OC, we compare OCFlow with and without OC at 10 NFEs on MP-20. As shown in Table 5, OC consistently improves structural and compositional validity, formation energy, ΔS, and convergence in both ID and OOD settings, with larger gains under atom-count extrapolation. Additional analyses of the coupling geometry are provided in Appendix C.2. We also provide structure galleries in Appendix C.4, comparing OOD samples generated by OCFlow and the evaluated baselines on both MP-20 and MPTS-52.

5 Related Work

Joint Diffusion Models for Crystal Generation. Early modern *de novo* crystal generators adapt variational and diffusion-based models to periodic structures. CDVAE (Xie et al., 2022) combines a periodic-graph variational encoder (Gasteiger et al., 2020) with a Langevin denoising decoder, and DiffCSP (Jiao et al., 2023) formulates joint diffusion over lattice parameters and fractional coordinates. Subsequent larger-scale studies such as MatterGen (Zeni et al., 2025) and UniMat (Yang et al., 2024) extend this line by jointly modeling lattice, coordinates, and atomic species across broader chemical spaces. ADiT (Joshi et al., 2025) instead encodes atoms and lattices into a shared latent space, where a diffusion transformer (Peebles & Xie, 2023) operates.

Symmetry-Aware Crystal Generation. A complementary line builds crystallographic priors directly into generation. DiffCSP++ (Jiao et al., 2024) enforces Wyckoff position constraints (Wyckoff, 1922) during coordinate denoising of a target space group, Wyckoff Transformer (Kazeev et al., 2025) and CrystalFormer (Cao et al., 2025) generate Wyckoff- or space-group-aware token sequences, and SymmCD (Levy et al., 2025) preserves space-group symmetries throughout diffusion. These approaches reduce representation ambiguity at the cost of explicit crystallographic conditioning.

Discrete Species Dynamics. Another important axis is the treatment of atomic species and composition. Rather than treating species as a post-hoc classifier or as a continuous proxy, recent generators model them as a discrete stochastic channel generated jointly with the geometry. MatterGen (Zeni et al., 2025) pairs continuous-time dynamics on lattice and coordinates with D3PM-style discrete diffusion (Austin et al., 2021) on species, and discrete flow formulations (Campbell et al., 2024; Gat et al., 2024) provide continuous-time analogues for categorical transitions. Other generators, such as DiffCSP (Jiao et al., 2023) and CrysBFN (Wu et al., 2025), likewise give species their own dynamics, separate from those of the lattice and coordinates. OCFlow uses absorbing-mask discrete FM with an independent species time s (Section 3.2).

Flow Matching and Optimal Couplings. The most closely related lines of work are flow-based crystal generators and OC techniques for fast sampling. FlowMM (Miller et al., 2024) and CrystalFlow (Luo et al., 2025) apply FM directly on the torus, using minimum-image displacements and equivariant graph networks to respect crystal geometry, with subsequent extensions refining time scheduling and atom ordering (Seong et al., 2026). In Euclidean domains, Rectified Flow (Liu et al., 2023) and minibatch OT-CFM (Tong et al., 2024) show that better endpoint couplings can straighten marginal paths and reduce inference steps, but these couplings remain *inter-sample*, deciding only which source is paired with which data sample. Crystals introduce an additional *intra-sample* ambiguity, since periodic images, global translations, and same-species atom permutations can represent the same physical structure. OCFlow addresses the coordinate correspondence through a per-structure minimum-image assignment, followed by a fixed lift translation correction, rather than a joint optimization over the full symmetry orbit.

6 Conclusion

In this paper, we introduce OCFlow, an FM framework for *de novo* crystal generation built on local OC. OCFlow resolves site correspondence within each training pair using periodic distances and then removes the common translation component. Experimental results show that OCFlow retains near-perfect structural validity, low formation energy, and strong relaxation quality with only a few integration steps, while exhibiting particularly strong performance when extrapolating to larger OOD systems. Hull-based evaluation further demonstrates that these gains persist under a stricter stability criterion, although the absolute yield of stable, novel materials remains small. Taken together, these results suggest that local OCs provide a simple and effective approach to improving both sampling efficiency and reliability in crystal generation.

Limitations and Outlook. While modest in our current experiments, the per-structure Hungarian assignment introduces additional computational overhead during training, and this cost may become more significant for substantially larger structures containing hundreds of atoms. Improving the scalability of the assignment procedure through block-wise approximations or structure-aware matching schemes is therefore an important direction, particularly for extending OCFlow to larger and more complex crystalline systems such as metal-organic frameworks.

AI USE STATEMENT

We used large language models for polishing the grammar and plotting our experimental results.

REPRODUCIBILITY STATEMENT

We describe the coupling construction and sampling procedure in Section 3. Appendix A provides the training, inference, and baseline configurations, while Appendix B details the evaluation protocol and Appendix C reports the complete results.

ETHICS STATEMENT

This work focuses on generative modeling for inorganic crystal structures and does not involve human subjects, personal data, or other sensitive information. Accordingly, we do not identify any direct ethical concerns associated with the proposed methodology.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Sterling G. Baird, Hasan M. Sayeed, Joseph Montoya, and Taylor D. Sparks. matbench-genmetrics: A Python library for benchmarking crystal structure generative models using time-based splits of Materials Project structures. *Journal of Open Source Software*, 9(97):5618, 2024.

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Fabio Falcioni, Andrea C. Ferrari, John L. A. Gardner, Mikołaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Cheuk Hin Ho, Stephan Hofmann, Christian Holm, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Louise A. M. Rosset, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Christopher Sutton, Viktor Svahn, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163(18):184110, 2025.

Siddharth Betala, Samuel P. Gleason, Ali Ramlaoui, Andy Xu, Georgia Channing, Daniel Levy, Clémentine Fourrier, Nikita Kazeev, Chaitanya K. Joshi, Sékou-Oumar Kaba, Félix Therrien, Alex Hernandez-Garcia, Rocío Mercado, N. M. Anoop Krishnan, and Alexandre Duval. LeMat-GenBench: A unified evaluation framework for crystal generative models. *arXiv:2512.04562*, 2025.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *International Conference on Machine Learning (ICML)*, 2024.

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

Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 9974–10007, 2024.

Johannes Gasteiger, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. In *Machine Learning for Molecules Workshop, Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. GemNet: Universal directional graph neural networks for molecules. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964.

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

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

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. `https://kellerjordan.github.io/posts/muon/`, 2024.

Chaitanya K. Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary Ward Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. In *International Conference on Machine Learning (ICML)*, pp. 28393–28417, 2025.

Nikita Kazeev, Wei Nong, Ignat Romanov, Ruiming Zhu, Andrey E. Ustyuzhanin, Shuya Yamazaki, and Kedar Hippalgaonkar. Wyckoff transformer: Generation of symmetric crystals. In *International Conference on Machine Learning (ICML)*, pp. 29495–29526, 2025.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

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

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

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.

John Maddox. Crystals from first principles. *Nature*, 335(6187):201, 1988.

Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M. Wood. FlowMM: Generating materials with Riemannian flow matching. In *International Conference on Machine Learning (ICML)*, pp. 35664–35686, 2024.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. of the IEEE international conference on computer vision (ICCV)*, 2023.

Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: atomistic simulation at scale. *arXiv:2504.06231*, 2025.

Jonathan Schmidt, Noah Hoffmann, Hai-Chen Wang, Pedro Borlido, Pedro J. M. A. Carriço, Tiago F. T. Cerqueira, Silvana Botti, and Miguel A. L. Marques. Large-scale machine-learning-assisted exploration of the whole materials space, 2022.

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

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Brandon M. Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R. Kitchin, Daniel S. Levine, Kyle Michel, Anuroop Sriram, Taco S. Cohen, Abhishek Das, Sushree Jagriti Sahoo, Ammar Rizvi, Zachary W. Ulissi, and C. Lawrence Zitnick. UMA: A family of universal models for atoms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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 (ICLR)*, 2025.

Ralph W. G. Wyckoff. *The Analytical Expression of the Results of the Theory of Space-Groups*. Carnegie Institution of Washington, 1922.

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

Sherry Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable diffusion for materials generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025.

A Experimental Details

A.1 Training details

We use a batch size of 64 graphs per step, weight decay of 10^{-8}, gradient clipping at norm 5, and an EMA decay of 0.995; all reported numbers use the EMA-averaged weights. The loss weights in Eq. (13) are set in the ratio $\lambda_{\mathbf{X}} : \lambda_{\mathbf{L}} : \lambda_S = 100 : 2 : 1$ for both datasets. We use the standard splits provided by Xie et al. (2022) for MP-20 and Baird et al. (2024) for MPTS-52. MP-20 contains 45,229 structures (training 27,136) with at most 20 atoms per cell. MPTS-52 contains 40,476 structures (training 27,380) with up to 52 atoms per cell, split chronologically.

For training the model on the MP-20 dataset, we use a hybrid Muon/AdamW optimizer (Jordan et al., 2024) with a learning rate 3×10^{-4}, Muon momentum 0.95, and five Newton-Schulz iterations. All reported MP-20 results use the checkpoint at iteration 450,000. For the MPTS-52 dataset, the model is trained with schedule-free AdamW (Loshchilov & Hutter, 2019; Defazio et al., 2024) with learning rate 2.5×10^{-4}, and all reported MPTS-52 results use the checkpoint at iteration 450,000. During training, the structural time is drawn as $t = u^{0.9}$ with $u \sim U(0, 1)$, and the species time s is drawn independently and uniformly on $[0, 1]$. Across all experiments, the models are trained on a single NVIDIA B200 GPU.

A.2 Sampling Details

We integrate the joint flow with a fixed-step Euler scheme. The species channel is integrated by sampling, at each Euler step, a Bernoulli unmask decision per still-masked atom with probability $\Delta s/s$, and drawing the realized label from the network's categorical prediction whenever the atom is selected. Here, $\Delta s > 0$ denotes the decrease in s over one step, with $0 < \Delta s \leq s \leq 1$. We use a sampling batch size of 64 for MP-20 and 32 for MPTS-52.

The structural velocities are rescaled by $c(1 + r\tau)$ with $\tau = 1 - t$ (Section 3.3), with separate scale and slope for the coordinate channel, $(c_{\mathbf{X}}, r_{\mathbf{X}})$, and the lattice channel, $(c_{\mathbf{L}}, r_{\mathbf{L}})$. The species channel is unmasked over $K_S \leq K$ steps, where K is the total sampling steps. Also, we ramp the sampling temperature of the species channel following a schedule from T_{start} to T_{end}. The standard settings behind each MP-20 experiments at 10 and 100 NFEs are $(c_{\mathbf{X}}, r_{\mathbf{X}}) = (1.0, 9.0)$ and $(c_{\mathbf{X}}, r_{\mathbf{X}}) = (2.5, 7.0)$. For MPTS-52, we use $(c_{\mathbf{X}}, r_{\mathbf{X}}) = (1.25, 5.0)$ at both budgets and $(c_{\mathbf{L}}, r_{\mathbf{L}}) = (1.0, 0.5)$ at 10 NFEs and $(2.0, 1.0)$ at 100 NFEs, in both the ID and OOD settings. For all experiments, we generate 5,000 structures per seed on MP-20 and 2,500 structures per seed on MPTS-52, using three different seeds for each setting.

A.3 Details of training baseline methods

We compare against five publicly available baselines on MP-20—DiffCSP++ (Jiao et al., 2024), MatterGen (Zeni et al., 2025), FlowMM (Miller et al., 2024), CrystalFlow (Luo et al., 2025), and ADiT (Joshi et al., 2025)—and the same set excluding ADiT on MPTS-52. For MP-20, we use the official checkpoints for DiffCSP++, MatterGen, CrystalFlow, and ADiT, and retrain FlowMM ourselves on the public release's default `abits_params` configuration to match the numbers reported in Miller et al. (2024). For MPTS-52, no official checkpoints are released for any of the four remaining baselines, so we retrain all four on the same chronological split used to train OCFlow, each on a single NVIDIA B200 GPU. Below we list only the hyperparameters that we set during retraining. All other hyperparameters follow each work's official release.

MatterGen on MPTS-52. We retrain MatterGen from scratch on MPTS-52, using the default optimizer and noise schedule of the MatterGen release with `accumulate_grad_batches` $= 32$ and validation and test batch size 16. The latest available checkpoint is used for sampling.

DiffCSP++ on MPTS-52. `diffusion_w_type` configuration, 500 epochs at batch size 128, Adam with learning rate 10^{-3}, cosine schedule with $\eta_{\min} = 10^{-5}$, and weight decay 0. The checkpoint with the lowest test loss is used at sampling time.

FlowMM on MPTS-52. `abits_params` configuration, 1,000 epochs at batch size 64, Adam with learning rate 3×10^{-4}, cosine schedule with $T_{\max} = 1000$ and $\eta_{\min} = 10^{-5}$, and weight decay 0.

CrystalFlow on MPTS-52. `flow_polar_w_type` configuration with the `type_encoding=table` option and $\sigma_{\text{lattice-polar}} = 0.1$, 500 epochs at batch size 128, Adam with learning rate 10^{-3} and weight decay 0, and ReduceLROnPlateau with factor 0.6. The cost weights are set to $(\text{type}, \text{coord}, \text{lattice}) = (10, 10, 1)$ and the decoder uses 6 layers of hidden width 512, 256 radial Fourier features, and 8 Miller-index components.

B Additional Evaluation Details

B.1 Experiment settings

For each dataset, regime, model, and sampling budget, we generate three independent sample sets. Each seed set contains 5,000 structures for MP-20 and 2,500 structures for MPTS-52. We report the mean over the three sampling seeds, with the standard deviation in parentheses. Sampling budgets are reported in terms of network function evaluations (NFEs), and for MatterGen, we note '$\times 2$' due to its predictor-corrector sampling.

In the MP-20 ID setting, atom counts are drawn from the empirical MP-20 training distribution, while OOD atom counts follow the MPTS-52 training distribution restricted to $N \geq 21$. For MPTS-52, ID atom counts follow the corresponding training distribution, and OOD atom counts are sampled uniformly from $[53, 70]$. The evaluation metrics and their detailed definitions are described in the following sections.

B.2 Evaluation metrics

Validity and novelty. Structural validity requires all interatomic distances to exceed 0.5 Å, while compositional validity evaluates charge neutrality using SMACT (Davies et al., 2019). UN measures the fraction of distinct generated structures that do not match the Alex-MP reference set of 845,997 structures using Pymatgen's StructureMatcher (Ong et al., 2013).

Relaxation quality. Generated structures are relaxed with `MACE-MPA-0` (Batatia et al., 2022; 2025) for up to 500 steps, using a stopping criterion based on the mean force magnitude over all atoms in each structure, with a threshold of $|F| = 0.005\text{eV}/\text{ Å}$. For each generated set, ΔS is the median minimum-image RMSD between the generated and relaxed structures, and %conv. is the fraction of relaxations that converge within 500 steps. Smaller ΔS indicate that the generated structures lie closer to their relaxed local minima.

Hull-based evaluation. We evaluate thermodynamic stability using LeMat-GenBench (Betala et al., 2025) with MACE-MP (Batatia et al., 2025), UMA (Wood et al., 2025), and Orb-v3 (Rhodes et al., 2025) and construct a self-consistent convex hull for each potential. A structure is considered stable when $E_{\text{hull}} \leq 0$ and metastable when $E_{\text{hull}} \leq 0.1$ eV/atom, where E_{hull} is computed against the hull of each potential and then averaged over the three potentials. S.U.N. denotes the fraction of generated structures that are stable, unique, and novel, while M.S.U.N. uses the metastability threshold. We additionally report the mean formation energy E_f. For the set of structures $\mathcal{C}_{\text{MSUN}}$ which passes the M.S.U.N. threshold, we characterize the composition of the retained candidates by

$$H_{\text{MSUN}} = \frac{1}{|\mathcal{C}_{\text{MSUN}}|} \sum_{c \in \mathcal{C}_{\text{MSUN}}} \left(-\sum_i x_{c,i} \ln x_{c,i} \right), \qquad K_{\text{MSUN}} = \frac{1}{|\mathcal{C}_{\text{MSUN}}|} \sum_{c \in \mathcal{C}_{\text{MSUN}}} K_c, \quad (14)$$

where $x_{c,i}$ is the atomic fraction of element i in candidate c, and K_c is the number of distinct elements. These quantities measure the composition diversity within thermodynamically stable structures.

C Additional Results

This appendix collects the complete results behind the main tables: the MP-20 sampling-budget sweeps (Appendix C.1), the ablation experiments on OC (Appendix C.2), and training and sampling runtime measurements (Appendix C.3). NFE counts network function evaluations, and parentheses follow the conventions of Appendix B.1.

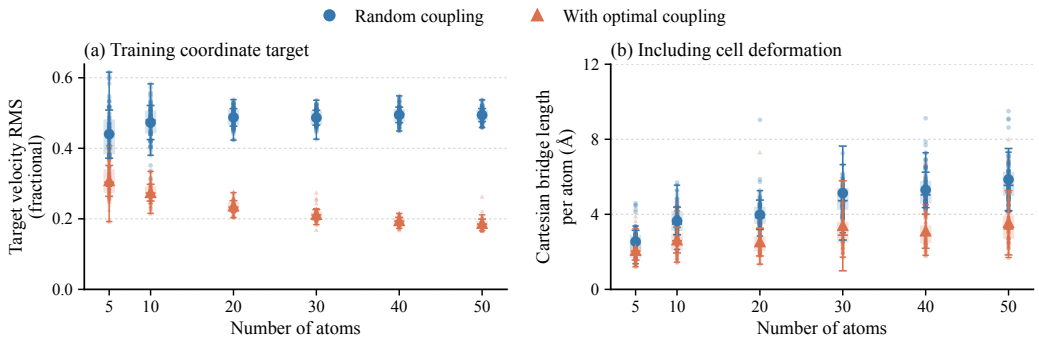


Figure 4: **Effect of OC across system sizes.** For each atom count, we compare random index-wise coupling with OC over 100 sampled structures. (a) RMS magnitude of the training coordinate target in fractional coordinates. (b) Cartesian bridge length per atom after accounting for cell deformation. OC produces shorter transport targets under both measures, with the gap generally increasing for larger systems. Error bars indicate the variation across sampled structures.

C.1 Sampling-budget sweeps on MP-20

Tables 3 and 4 report validity, novelty, hull-based, and relaxation metrics of all MP-20 models at 10, 50, 100, 250, and 1,000 NFEs (MatterGen: sampler steps times two) for the in-distribution and atom-count OOD settings, wherever a model was evaluated at that budget. The 10, 100, and 1,000 NFE rows of OCFlow, the 10×2, 100×2, and 1000×2 rows of MatterGen, the 1,000 NFE row of DiffCSP++, the 250 NFE rows of ADiT and FlowMM, and the 100 NFE row of CrystalFlow repeat the entries of Tables 1 and 2; the remaining rows are additional sampling budget evaluations, for which only the metrics that were computed are shown.

C.2 Additional Observations on Optimal Coupling

To better understand this behavior, we directly examine the geometry of the coordinate targets constructed during training. For each atom count $N \in \{5,\ 10,\ 20,\ 30,\ 40,\ 50\}$, we sample 100 structures and compare random index-wise coupling with OC. Figure 4(a) reports the RMS magnitude of the target coordinate velocity in fractional space, while Figure 4(b) measures the corresponding Cartesian bridge length per atom after accounting for cell deformation.

OC substantially shortens the training targets under both measures. More importantly, the gap between random coupling and OC generally widens as the number of atoms increases. Under random coupling, the fractional-coordinate target remains large as the system size grows, whereas OC progressively reduces its magnitude. The same trend is observed in Cartesian space, where OC produces consistently shorter bridge lengths, with a larger separation for larger structures.

These observations are consistent with the OOD improvements in Table 5. As the number of atoms increases, the ambiguity over possible site correspondences becomes more severe, while OC explicitly selects a shorter correspondence within each structure. This results in shorter and more consistent transport targets and may yield a simpler velocity field to learn and integrate. We therefore hypothesize that this reduction in correspondence ambiguity contributes to the stronger few-step performance of OCFlow on larger out-of-distribution systems.

C.3 Computational Cost

Training overhead. The Hungarian assignment is used only to construct the coordinate training targets. Table 6 isolates its cost in an otherwise fixed training loop: for each dataset, two workers each trained with and without OC in opposite orders, using batch size 64, one NVIDIA B200, and 100 warm-up followed by 1,000 measured steps per mode. Epoch times are projected from the measured step times, and the increase is the mean of the two paired overheads (5.91-8.93% on MP-20 and 2.93-5.15% on MPTS-52). The assignment block itself, including the per-structure cost matrices and

Table 3: **MP-20 in-distribution generation across sampling budgets.** Columns and units follow Table 1. E_f, H_{MSUN}, and K_{MSUN} are defined in Appendix B.2, and values in parentheses follow the reporting convention in Appendix B.1. Dashes indicate metrics that are unavailable for a given model. All measurements use 5,000 samples per seed, and novelty is reported as UN (845K). Best and second-best values in each column are highlighted in **bold** and underline, respectively.

Model	NFE	Validity (↑)		Novelty (↑)	Stability and Diversity			Relaxation	
		Struct.	Comp.	UN (845K)	E_f (↓)	H_{MSUN}	K_{MSUN}	ΔS (↓)	%conv. (↑)
MatterGen	1000×2	99.84 (0.06)	84.02 (0.52)	78.48 (0.58)	-0.767 (0.012)	0.92 (0.01)	3.04 (0.03)	0.252 (0.007)	97.50 (0.22)
	250×2	99.92 (0.04)	81.84 (0.56)	81.38 (0.54)	-0.716 (0.011)	0.89 (0.01)	2.93 (0.03)	0.365 (0.008)	96.64 (0.26)
	100×2	99.88 (0.05)	82.08 (0.55)	84.52 (0.51)	-0.683 (0.012)	0.87 (0.01)	2.89 (0.04)	0.504 (0.011)	96.04 (0.27)
	50×2	99.92 (0.04)	81.94 (0.55)	86.36 (0.48)	-0.632 (0.011)	0.83 (0.02)	2.77 (0.04)	0.661 (0.013)	96.14 (0.27)
	10×2	99.90 (0.04)	80.72 (0.57)	91.78 (0.39)	-0.479 (0.012)	0.76 (0.03)	2.52 (0.09)	1.094 (0.012)	94.42 (0.33)
DiffCSP++	1000	99.82 (0.06)	83.56 (0.52)	72.48 (0.63)	-0.504 (0.012)	0.94 (0.02)	3.17 (0.06)	0.316 (0.012)	91.90 (0.39)
	250	99.64 (0.09)	76.90 (0.61)	84.46 (0.51)	-0.261 (0.013)	0.91 (0.03)	3.05 (0.09)	0.627 (0.019)	84.90 (0.51)
	100	99.86 (0.06)	71.44 (0.63)	89.24 (0.44)	0.125 (0.014)	0.77 (0.05)	2.61 (0.14)	0.800 (0.020)	84.22 (0.52)
	50	99.94 (0.03)	61.76 (0.66)	89.32 (0.44)	1.295 (0.023)	0.60 (0.26)	2.00 (0.47)	1.231 (0.025)	66.78 (0.67)
	10	99.76 (0.07)	46.10 (0.70)	**98.56** (0.17)	3.514 (0.136)	–	–	2.099 (0.037)	50.18 (0.71)
ADiT	1000	99.32 (0.11)	88.90 (0.42)	55.96 (0.68)	-0.486 (0.015)	0.76 (0.09)	2.70 (0.24)	0.435 (0.009)	98.36 (0.17)
	250	99.24 (0.12)	88.84 (0.44)	69.68 (0.70)	-0.511 (0.015)	0.78 (0.06)	2.77 (0.19)	0.435 (0.011)	98.62 (0.16)
	100	99.48 (0.10)	**89.22** (0.44)	57.44 (0.62)	-0.506 (0.015)	0.85 (0.06)	2.98 (0.17)	0.444 (0.012)	**98.72** (0.16)
	50	99.54 (0.10)	88.98 (0.47)	60.32 (0.69)	-0.504 (0.015)	0.82 (0.06)	2.87 (0.17)	0.486 (0.009)	98.18 (0.19)
	10	99.38 (0.11)	84.44 (0.50)	78.84 (0.63)	-0.257 (0.018)	0.80 (0.07)	2.79 (0.21)	0.936 (0.009)	97.56 (0.22)
FlowMM	1000	99.50 (0.09)	81.62 (0.53)	82.44 (0.61)	-0.415 (0.016)	0.88 (0.03)	2.83 (0.08)	1.001 (0.012)	95.96 (0.28)
	250	99.44 (0.11)	82.40 (0.54)	82.88 (0.53)	-0.392 (0.013)	0.82 (0.04)	2.69 (0.10)	0.980 (0.011)	95.50 (0.29)
	100	99.52 (0.09)	83.06 (0.55)	82.62 (0.55)	-0.397 (0.019)	0.80 (0.03)	2.65 (0.09)	0.991 (0.012)	95.44 (0.28)
	50	99.66 (0.07)	81.64 (0.52)	84.14 (0.52)	-0.388 (0.014)	0.87 (0.04)	2.82 (0.11)	1.021 (0.012)	95.52 (0.28)
	10	99.66 (0.09)	81.48 (0.51)	86.88 (0.48)	-0.193 (0.015)	0.88 (0.05)	2.73 (0.12)	1.168 (0.011)	94.64 (0.31)
CrystalFlow	1000	99.62 (0.08)	81.68 (0.57)	90.06 (0.43)	0.061 (0.014)	0.69 (0.08)	2.22 (0.15)	1.286 (0.010)	95.74 (0.31)
	250	99.56 (0.09)	82.12 (0.49)	90.20 (0.42)	0.060 (0.015)	0.66 (0.13)	2.17 (0.26)	1.315 (0.012)	95.56 (0.29)
	100	99.66 (0.06)	81.55 (0.39)	89.64 (0.43)	0.073 (0.014)	0.71 (0.09)	2.33 (0.20)	1.285 (0.011)	95.62 (0.29)
	50	99.62 (0.09)	81.58 (0.56)	90.74 (0.41)	0.075 (0.014)	0.61 (0.11)	2.18 (0.27)	1.306 (0.011)	95.94 (0.28)
	10	99.16 (0.11)	80.32 (0.57)	94.56 (0.32)	0.229 (0.015)	0.51 (0.10)	1.75 (0.15)	1.399 (0.009)	95.04 (0.29)
OCFlow (ours)	1000	**99.98** (0.03)	88.23 (0.42)	73.47 (0.25)	-1.053 (0.020)	1.08 (0.01)	3.97 (0.06)	**0.171** (0.007)	97.83 (0.45)
	250	99.96 (0.03)	88.30 (0.44)	74.94 (0.61)	**-1.070** (0.014)	1.08 (0.01)	4.02 (0.07)	0.240 (0.010)	97.12 (0.24)
	100	99.97 (0.06)	88.11 (0.75)	73.39 (0.27)	-1.060 (0.018)	1.07 (0.01)	3.91 (0.03)	0.234 (0.001)	97.50 (0.40)
	50	99.96 (0.02)	87.91 (0.26)	78.34 (0.16)	-1.065 (0.013)	1.08 (0.01)	3.78 (0.06)	0.322 (0.006)	96.71 (0.53)
	10	99.97 (0.03)	87.35 (0.64)	84.25 (0.52)	-0.970 (0.007)	1.05 (0.02)	3.47 (0.06)	0.542 (0.005)	96.19 (0.12)

Table 4: **MP-20 atom-count OOD generation across sampling budgets.** Models trained on MP-20 are evaluated on structures with $N \in [21, 52]$ atoms, following the OOD setting in Table 2. Columns and conventions follow Table 3. No generated structure satisfies the strict S.U.N. criterion in this setting; therefore, H_{MSUN} and K_{MSUN} are computed over the M.S.U.N. subset and are left undefined when no candidate satisfies the M.S.U.N. criterion. Best and second-best values in each column are highlighted in **bold** and underline, respectively.

Model	NFE	Validity (↑)		Stability and Diversity			Relaxation	
		Struct.	Comp.	E_f (↓)	H_{MSUN}	K_{MSUN}	ΔS (↓)	%conv. (↑)
MatterGen	1000×2	**100.00** (0.00)	66.08 (0.67)	-1.514 (0.010)	1.02 (0.02)	4.09 (0.10)	0.593 (0.007)	91.90 (0.39)
	250×2	**100.00** (0.00)	64.10 (0.68)	-1.422 (0.010)	0.87 (0.04)	3.29 (0.16)	0.863 (0.008)	88.16 (0.46)
	100×2	**100.00** (0.00)	63.04 (0.69)	-1.332 (0.010)	0.86 (0.11)	3.50 (0.34)	1.083 (0.008)	86.34 (0.49)
	50×2	**100.00** (0.00)	58.36 (0.70)	-1.227 (0.010)	0.95 (0.02)	3.50 (0.35)	1.254 (0.008)	83.30 (0.53)
	10×2	99.42 (0.11)	53.06 (0.71)	-0.814 (0.010)	–	–	1.616 (0.007)	75.80 (0.61)
DiffCSP++	1000	97.82 (0.21)	15.56 (0.51)	-0.580 (0.008)	–	–	1.653 (0.008)	76.20 (0.60)
	250	96.20 (0.29)	10.40 (0.44)	0.044 (0.010)	–	–	2.042 (0.009)	52.96 (0.71)
	100	98.76 (0.18)	0.32 (0.08)	0.521 (0.017)	–	–	2.276 (0.009)	48.24 (0.71)
	50	99.94 (0.04)	1.84 (0.21)	2.806 (0.043)	–	–	2.850 (0.014)	20.66 (0.57)
	10	97.50 (0.21)	0.00 (0.00)	–	–	–	3.768 (0.022)	18.60 (0.55)
ADiT	1000	91.00 (0.35)	59.62 (0.60)	-0.036 (0.033)	–	–	2.108 (0.023)	62.32 (0.70)
	250	91.90 (0.39)	58.46 (0.70)	-0.023 (0.031)	–	–	2.095 (0.021)	63.00 (0.67)
	100	92.22 (0.38)	58.46 (0.74)	0.018 (0.045)	–	–	2.027 (0.019)	62.24 (0.68)
	50	92.86 (0.40)	59.00 (0.79)	-0.139 (0.031)	–	–	1.989 (0.019)	63.38 (0.68)
	10	93.08 (0.36)	56.58 (0.72)	-0.212 (0.029)	–	–	1.833 (0.011)	69.26 (0.65)
FlowMM	1000	73.92 (0.61)	20.26 (0.57)	1.326 (0.113)	–	–	3.418 (0.018)	70.28 (0.65)
	250	74.30 (0.62)	20.90 (0.58)	1.168 (0.130)	–	–	3.392 (0.020)	69.04 (0.65)
	100	73.54 (0.62)	19.90 (0.57)	1.242 (0.101)	–	–	3.411 (0.019)	69.26 (0.65)
	50	73.90 (0.60)	19.00 (0.54)	1.653 (0.136)	–	–	3.373 (0.019)	68.52 (0.66)
	10	72.00 (0.62)	8.44 (0.39)	1.442 (0.168)	–	–	3.325 (0.019)	70.00 (0.63)
CrystalFlow	1000	92.22 (0.40)	42.74 (0.69)	0.477 (0.040)	0.00 (0.00)	1.00 (0.00)	2.206 (0.014)	70.86 (0.59)
	250	92.04 (0.37)	42.26 (0.70)	0.431 (0.043)	–	–	2.187 (0.013)	70.28 (0.65)
	100	92.54 (0.37)	42.48 (0.70)	0.373 (0.040)	–	–	2.168 (0.013)	70.24 (0.65)
	50	93.08 (0.36)	40.02 (0.70)	0.468 (0.040)	0.00 (0.00)	1.00 (0.00)	2.124 (0.011)	69.78 (0.66)
	10	92.46 (0.37)	29.92 (0.65)	0.586 (0.038)	–	–	2.036 (0.011)	70.31 (0.65)
OCFlow (ours)	1000	**100.00** (0.00)	79.30 (0.61)	**-1.774** (0.016)	1.13 (0.01)	4.78 (0.18)	**0.544** (0.013)	93.60 (0.75)
	250	**100.00** (0.00)	**80.40** (0.56)	-1.739 (0.009)	1.12 (0.01)	4.61 (0.11)	0.605 (0.005)	92.72 (0.37)
	100	**100.00** (0.00)	79.02 (0.69)	-1.716 (0.011)	1.09 (0.01)	4.43 (0.13)	0.576 (0.008)	**93.93** (0.36)
	50	**100.00** (0.00)	80.20 (0.57)	-1.759 (0.013)	1.11 (0.02)	4.61 (0.18)	0.598 (0.014)	93.13 (0.56)
	10	**100.00** (0.00)	78.75 (0.48)	-1.477 (0.006)	1.07 (0.07)	4.06 (0.46)	0.894 (0.007)	90.16 (0.48)

Table 5: **Ablation of OC** on MP-20 at 10 NFEs. Bold indicates the better results within each split.

	In-distribution, $N \in [1, 20]$					*Out-of-distribution*, $N \in [21, 52]$				
OC	Struct.↑	Comp.↑	$E_f \downarrow$	$\Delta S \downarrow$	Conv.↑	Struct.↑	Comp.↑	$E_f \downarrow$	$\Delta S \downarrow$	Conv.↑
×	99.22	86.10	-0.866	0.634	95.88	99.82	77.16	-1.243	1.067	89.20
✓	**99.97**	**87.35**	**-0.970**	**0.542**	**96.19**	**100.00**	**78.75**	**-1.477**	**0.894**	**90.16**

the Hungarian solves, takes 20.8 ms per batch on MP-20 and 19.5 ms on MPTS-52, i.e., 6.7% and 5.3% of a training step with OC.

Sampling time. Table 7 reports the sampling throughput of all MP-20 models on the same hardware and the same sequence of atom counts; OCFlow is timed at 100 NFEs with species unmasking spread over all 100 steps. Since the assignment is absent at inference, the differences reflect the architectures, implementations, and configured numbers of network evaluations, and equal NFE budgets does not imply equal wall-clock cost.

Table 6: **Training overhead of the coordinate coupling.** Projected wall-clock time per training epoch with and without OC, holding the rest of the training loop fixed (one NVIDIA B200, batch size 64, 100 warm-up and 1,000 measured steps per mode). Times are the mean over two same-worker replicates with swapped execution order, and the increase is the mean of the two paired overheads.

Dataset	Without OC (s/epoch)	With OC (s/epoch)	Increase (%)
MP-20	122.7	131.8	7.42
MPTS-52	150.08	156.10	4.04

Table 7: **Sampling time on MP-20.** Seconds per generated structure on one NVIDIA B200 with inference batch size 64, using the same fixed sequence of 832 MP-20 atom counts for every model; the first 64 structures are discarded as warm-up and the remaining 768 are timed. Parentheses give the standard deviation over five independent repeats. Timers cover the sampler only and exclude checkpoint loading, metric evaluation, relaxation, and file output; ADiT includes its required VAE decoding. NFE counts network function evaluations, with $\times 2$ denoting predictor-corrector sampling.

Model	NFE	Seconds/structure (↓)
MatterGen	50×2	0.11412 (0.00093)
	100×2	0.22835 (0.00155)
	1000×2	2.27808 (0.01620)
CrystalFlow	100	0.02111 (0.00120)
FlowMM	250	0.13049 (0.00792)
ADiT	250	0.04962 (0.00018)
OCFlow (ours)	100	0.10345 (0.00235)

C.4 Examples of Generated and Relaxed Structures

We provide qualitative comparisons of samples generated by OCFlow and the evaluated baselines. Figures 5-10 compare generated structures before and after structure optimization (relaxation with `MACE-MPA-0`) for MP-20-trained models at $N = 30$, 40, and 50, beyond the training range of $N \leq 20$. Figures 11-14 show the same comparison for MPTS-52-trained models at $N = 60$ and 70.

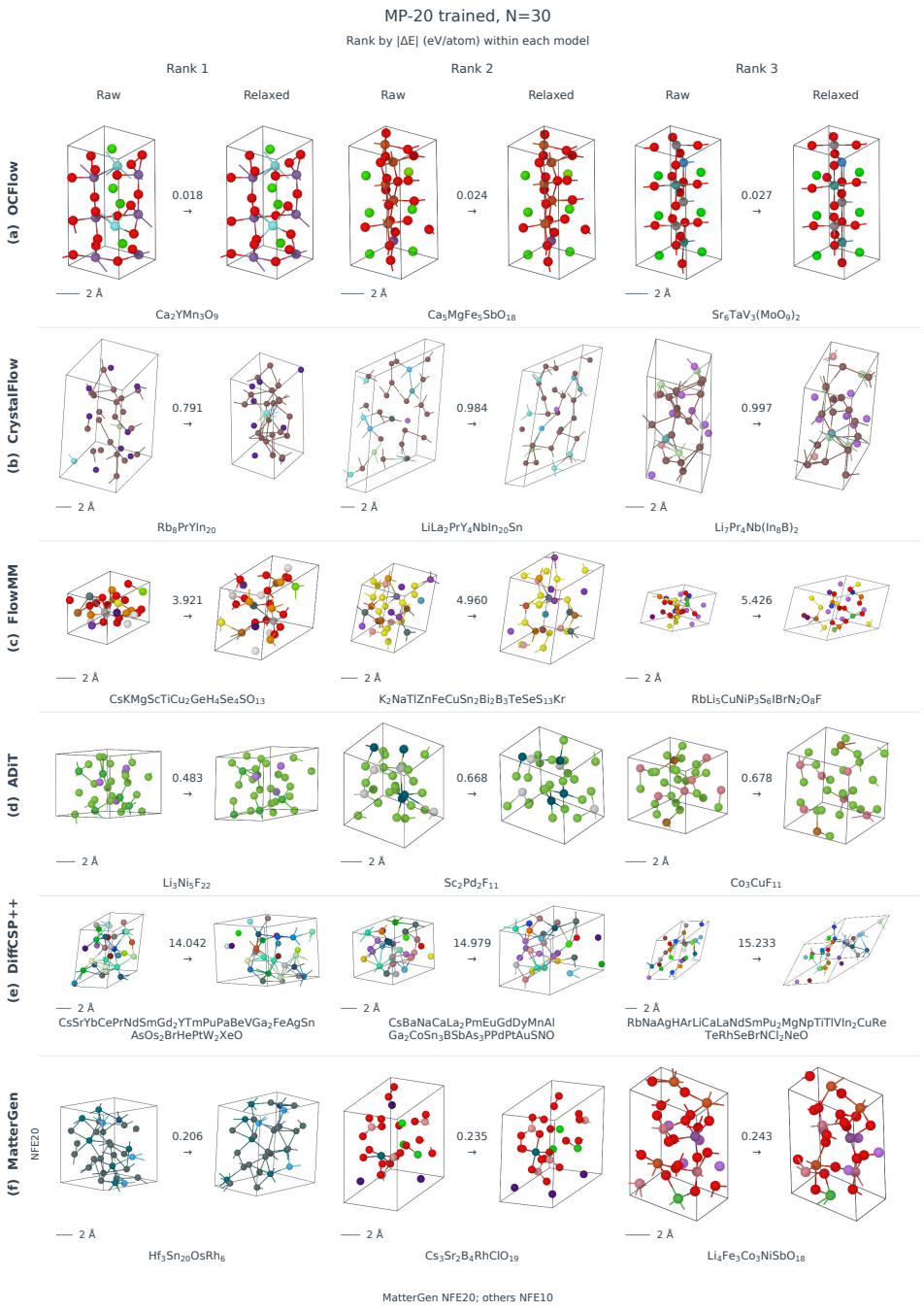


Figure 5: **Generated and relaxed structures from MP-20-trained models at $N = 30$ (ranks 1-3).** Rows show (a) OCFlow, (b) CrystalFlow, (c) FlowMM, (d) ADiT, (e) DiffCSP++, and (f) MatterGen. Samples are ranked within each model by increasing absolute per-atom energy change upon relaxation, $|\Delta E|$ (eV/atom), with raw and relaxed structures shown side by side. MatterGen uses 10×2 NFEs; the other models use 10 NFEs. Scale bars indicate 2 Å.

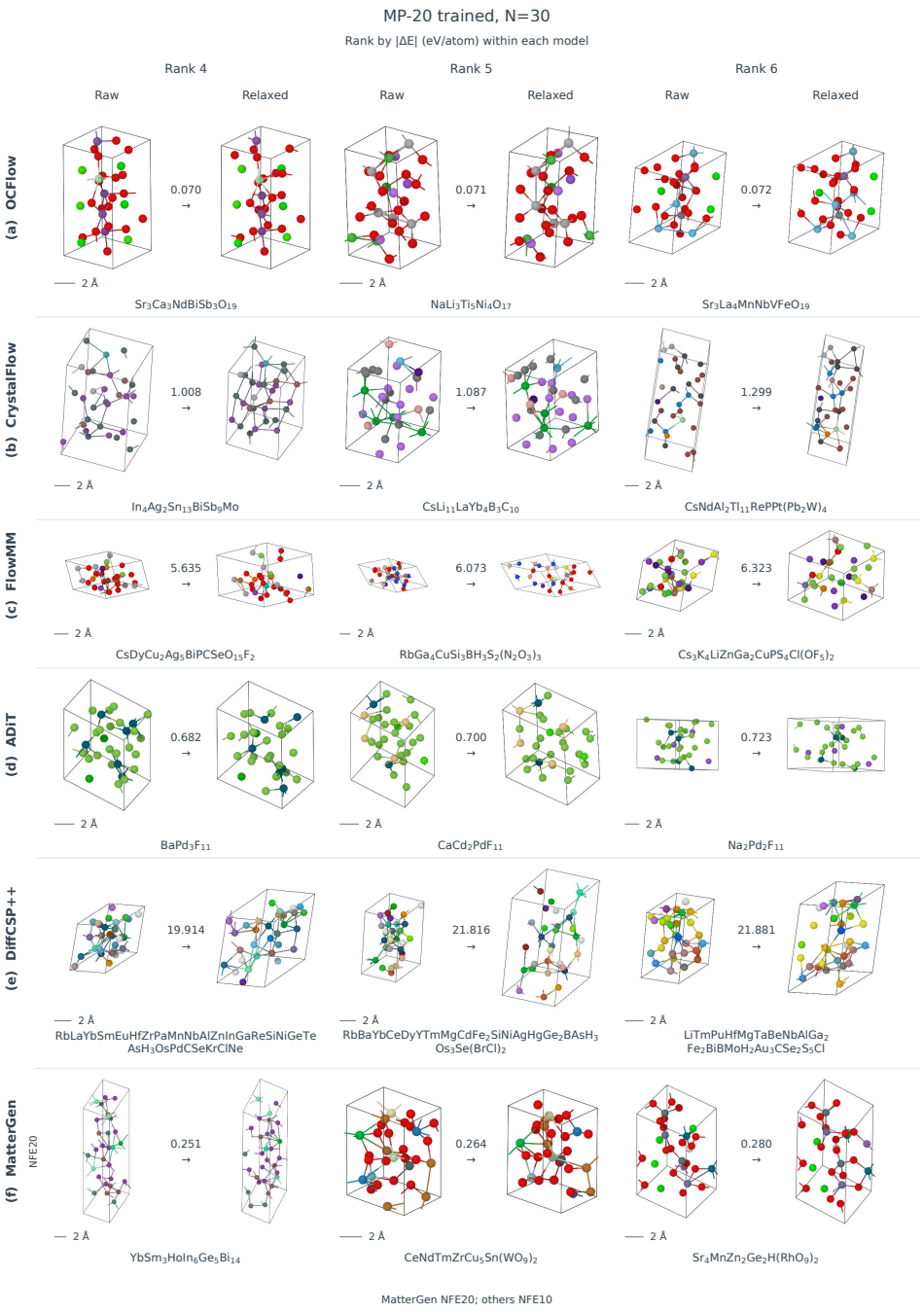


Figure 6: **Generated and relaxed structures from MP-20-trained models at $N = 30$ (ranks 4-6).** Rows show (a) OCFlow, (b) CrystalFlow, (c) FlowMM, (d) ADiT, (e) DiffCSP++, and (f) MatterGen. Samples are ranked within each model by increasing absolute per-atom energy change upon relaxation, $|\Delta E|$ (eV/atom), with raw and relaxed structures shown side by side. MatterGen uses 10×2 NFEs; the other models use 10 NFEs. Scale bars indicate 2 Å.

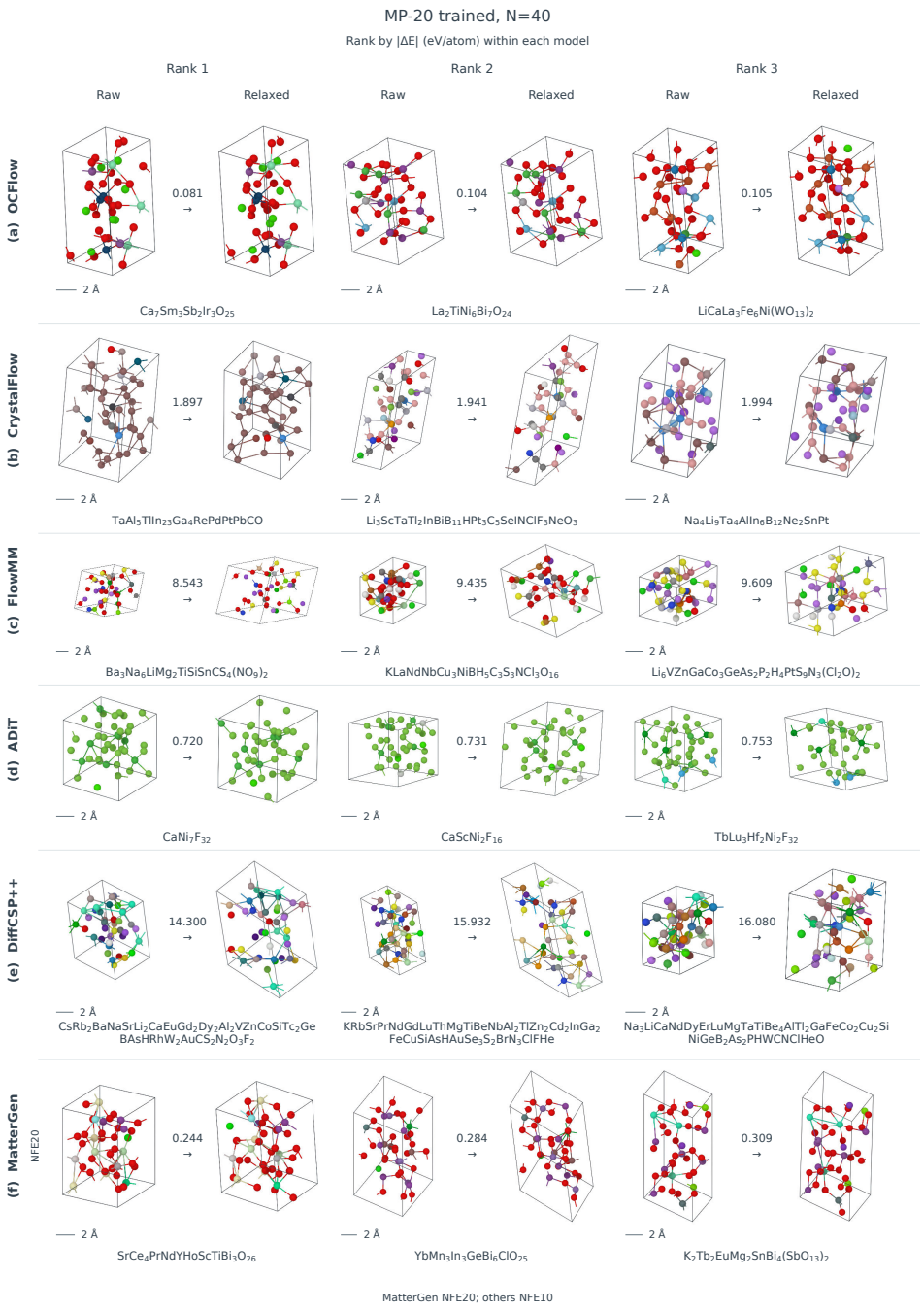


Figure 7: **Generated and relaxed structures from MP-20-trained models at $N = 40$ (ranks 1-3).** Rows show (a) OCFlow, (b) CrystalFlow, (c) FlowMM, (d) ADiT, (e) DiffCSP++, and (f) MatterGen. Samples are ranked within each model by increasing absolute per-atom energy change upon relaxation, $|\Delta E|$ (eV/atom), with raw and relaxed structures shown side by side. MatterGen uses 10×2 NFEs; the other models use 10 NFEs. Scale bars indicate 2 Å.

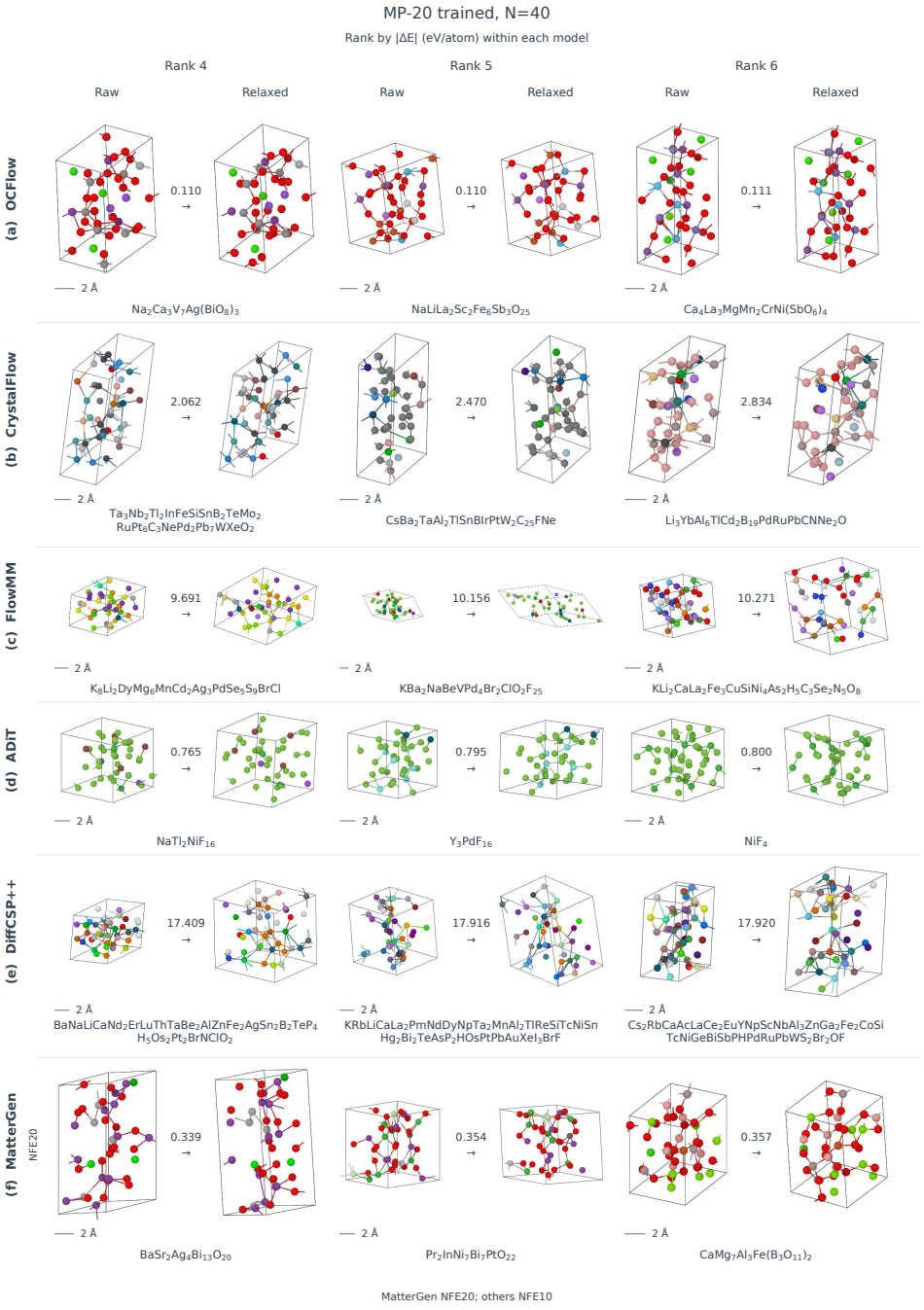


Figure 8: **Generated and relaxed structures from MP-20-trained models at $N = 40$ (ranks 4-6).** Rows show (a) OCFlow, (b) CrystalFlow, (c) FlowMM, (d) ADiT, (e) DiffCSP++, and (f) MatterGen. Samples are ranked within each model by increasing absolute per-atom energy change upon relaxation, $|\Delta E|$ (eV/atom), with raw and relaxed structures shown side by side. MatterGen uses 10×2 NFEs; the other models use 10 NFEs. Scale bars indicate 2 Å.

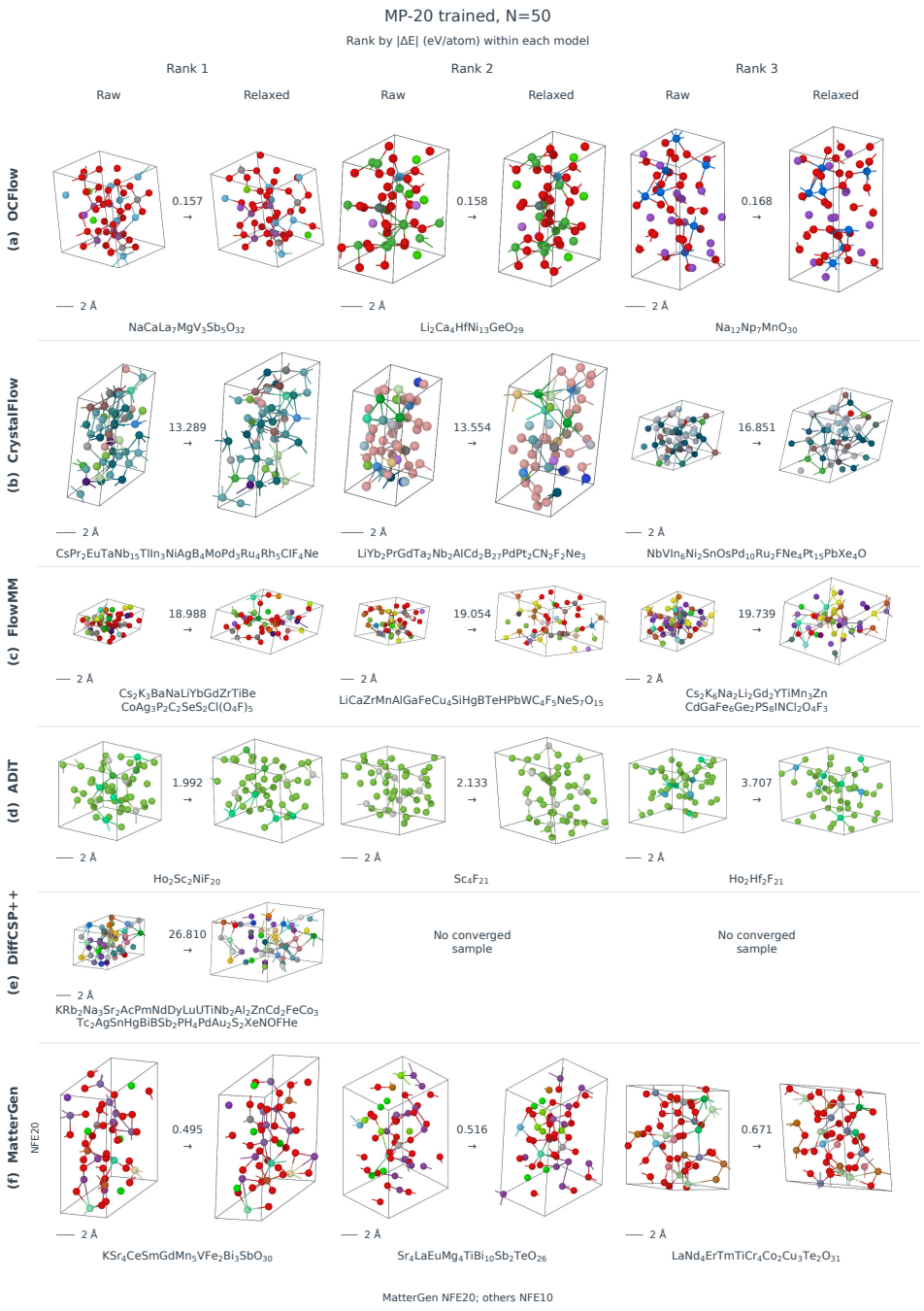


Figure 9: **Generated and relaxed structures from MP-20-trained models at $N = 50$ (ranks 1-3).** Rows show (a) OCFlow, (b) CrystalFlow, (c) FlowMM, (d) ADiT, (e) DiffCSP++, and (f) MatterGen. Samples are ranked within each model by increasing absolute per-atom energy change upon relaxation, $|\Delta E|$ (eV/atom), with raw and relaxed structures shown side by side. MatterGen uses 10×2 NFEs; the other models use 10 NFEs. Scale bars indicate 2 Å.

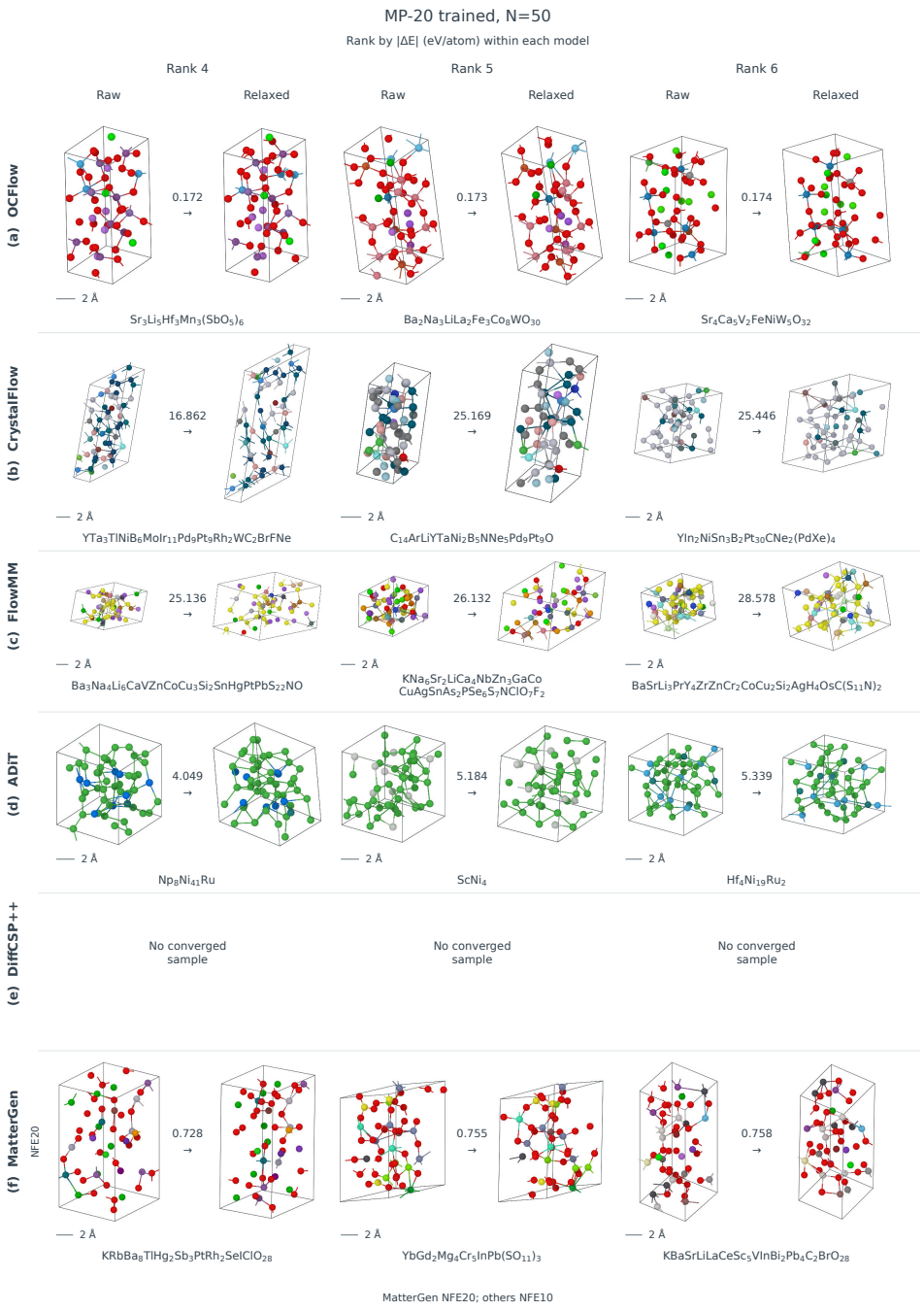


Figure 10: **Generated and relaxed structures from MP-20-trained models at $N = 50$ (ranks 4-6).** Rows show (a) OCFlow, (b) CrystalFlow, (c) FlowMM, (d) ADiT, (e) DiffCSP++, and (f) MatterGen. Samples are ranked within each model by increasing absolute per-atom energy change upon relaxation, $|\Delta E|$ (eV/atom), with raw and relaxed structures shown side by side. MatterGen uses 10×2 NFEs; the other models use 10 NFEs. Scale bars indicate 2 Å.

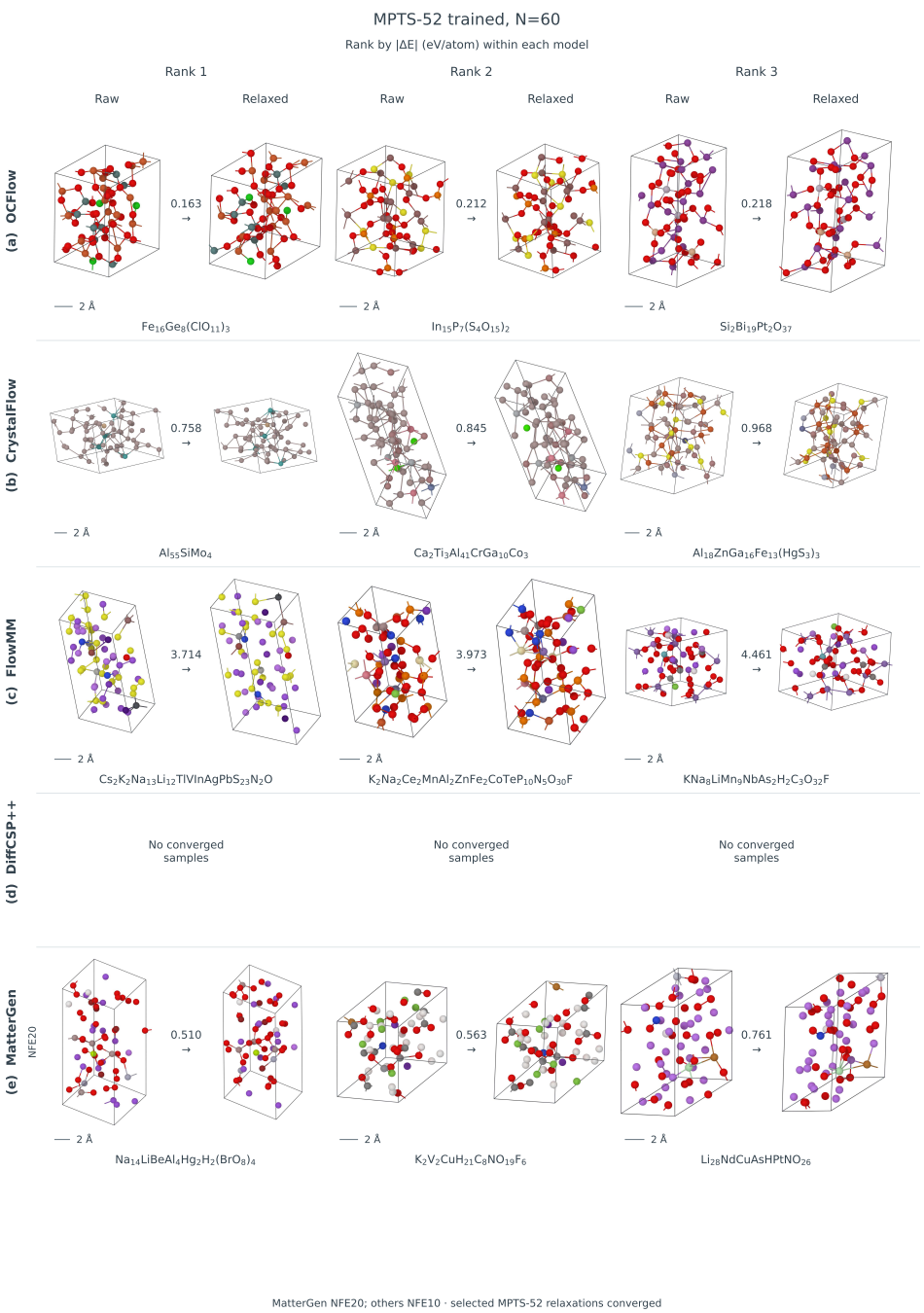


Figure 11: **Generated and relaxed structures from MPTS-52-trained models at $N = 60$ (ranks 1-3).** Rows show (a) OCFlow, (b) CrystalFlow, (c) FlowMM, (d) DiffCSP++, and (e) MatterGen. Selected converged samples are ranked within each model by increasing absolute per-atom energy change upon relaxation, $|\Delta E|$ (eV/atom), with raw and relaxed structures shown side by side. Values beside the arrows denote $|\Delta E|$. MatterGen uses 10×2 NFEs; the other models use 10 NFEs. Scale bars indicate 2 Å. The DiffCSP++ panels are marked "No converged samples" for this comparison.

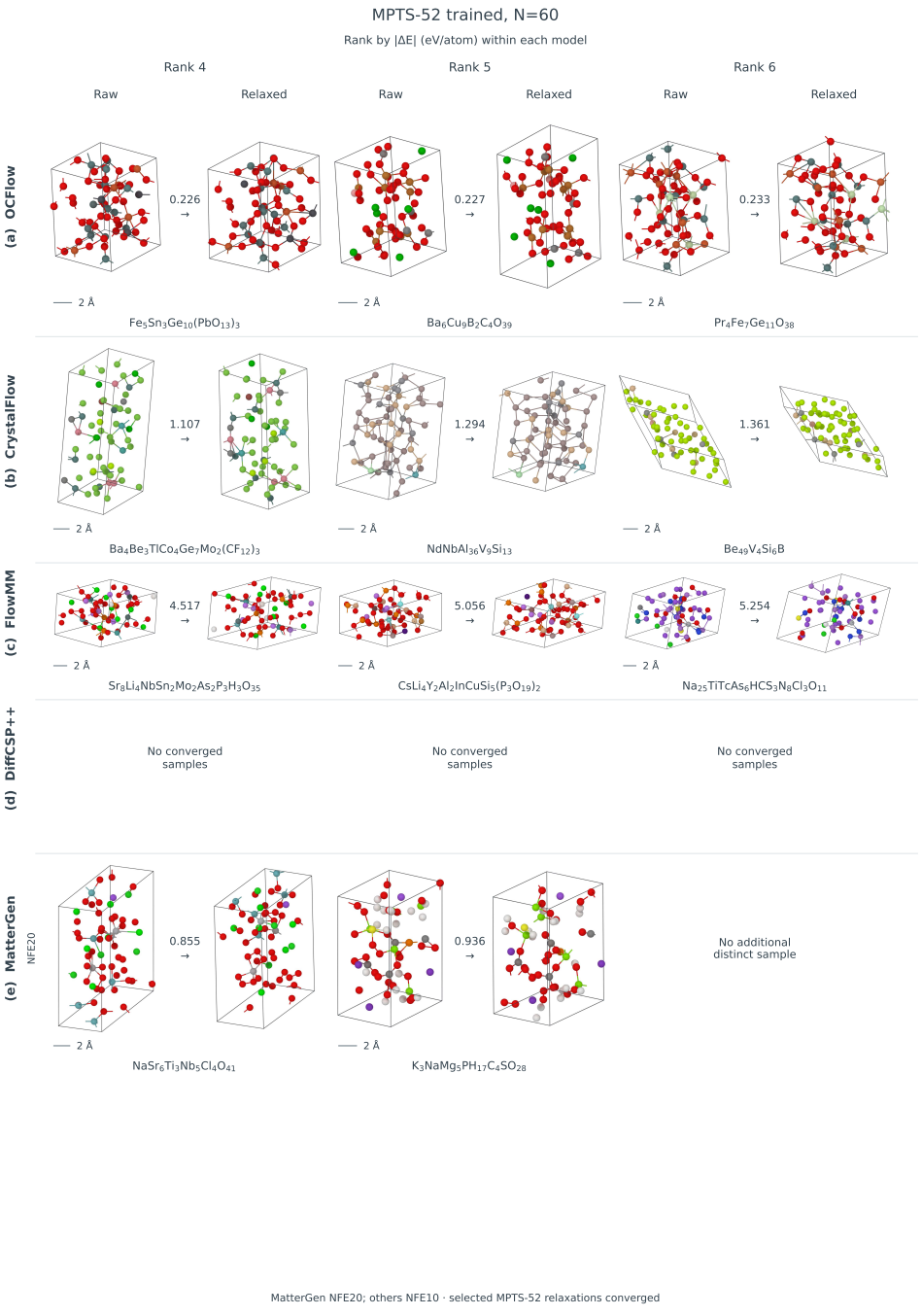


Figure 12: **Generated and relaxed structures from MPTS-52-trained models at $N = 60$ (ranks 4-6).** Rows show (a) OCFlow, (b) CrystalFlow, (c) FlowMM, (d) DiffCSP++, and (e) MatterGen. Selected converged samples are ranked within each model by increasing absolute per-atom energy change upon relaxation, $|\Delta E|$ (eV/atom), with raw and relaxed structures shown side by side. Values beside the arrows denote $|\Delta E|$. MatterGen uses 10×2 NFEs; the other models use 10 NFEs. Scale bars indicate 2 Å. For this comparison, the DiffCSP++ panels are marked "No converged samples", while the MatterGen rank-6 panel is marked "No additional distinct sample".

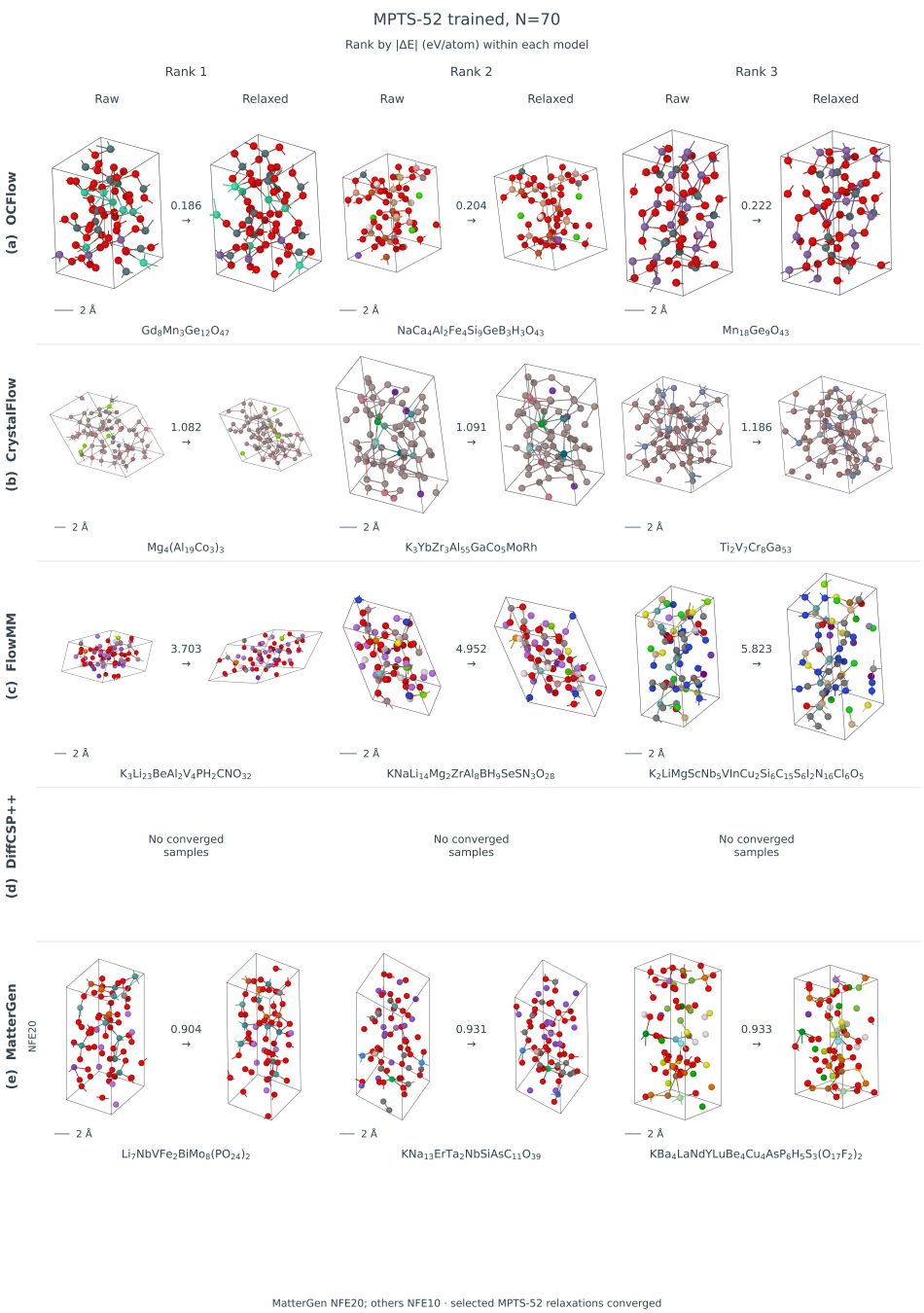


Figure 13: **Generated and relaxed structures from MPTS-52-trained models at $N = 70$ (ranks 1-3).** Rows show (a) OCFlow, (b) CrystalFlow, (c) FlowMM, (d) DiffCSP++, and (e) MatterGen. Selected converged samples are ranked within each model by increasing absolute per-atom energy change upon relaxation, $|\Delta E|$ (eV/atom), with raw and relaxed structures shown side by side. Values beside the arrows denote $|\Delta E|$. MatterGen uses 10×2 NFEs; the other models use 10 NFEs. Scale bars indicate 2 Å. The DiffCSP++ panels are marked "No converged samples" for this comparison.

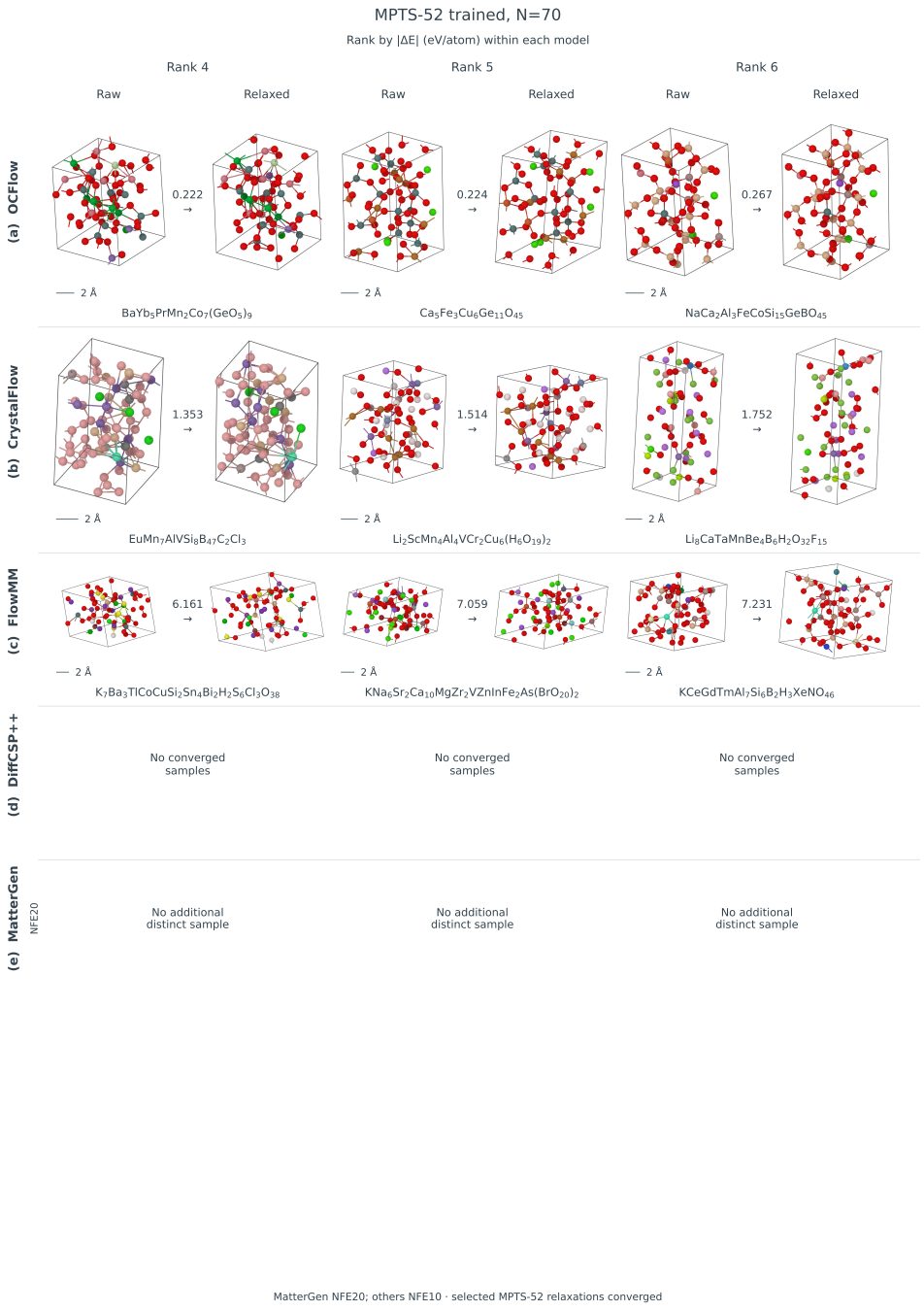


Figure 14: **Generated and relaxed structures from MPTS-52-trained models at $N = 70$ (ranks 4-6).** Rows show (a) OCFlow, (b) CrystalFlow, (c) FlowMM, (d) DiffCSP++, and (e) MatterGen. Selected converged samples are ranked within each model by increasing absolute per-atom energy change upon relaxation, $|\Delta E|$ (eV/atom), with raw and relaxed structures shown side by side. Values beside the arrows denote $|\Delta E|$. MatterGen uses 10×2 NFEs; the other models use 10 NFEs. Scale bars indicate 2 Å. For this comparison, the DiffCSP++ panels are marked "No converged samples", while the MatterGen panels are marked "No additional distinct sample".