MEANFLUX: FEW-STEP CRYSTAL GENERATION FOR ACCELERATED MATERIALS DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Current generative models for materials discovery produce high-quality candidates but suffer from massive sampling bottlenecks, often requiring dozens to thousands of network function evaluations (NFEs) per structure. We introduce MeanFlux, a diffusion transformer that accelerates generation by learning the average velocity across intervals of a generative trajectory. Operating in the per-atom latent space of a frozen autoencoder, MeanFlux generates viable crystals in just 1 to 16 steps. On the MP-20 benchmark, a 130M-parameter model matches the state-of-the-art yield of metastable, unique, and novel candidates while using one-sixth the inference compute. This efficiency extends to crystal structure prediction, where MeanFlux requires substantially less sampling compute than prior methods and scales predictably with model size. On LeMat-GenBench, MeanFlux produces an order of magnitude more viable crystals per compute unit than comparably pretrained models. With negative-aware fine-tuning (NFT) on a crystal reward, the MP-20 model reaches 22.5% relaxed LeMat-GenBench mSUN, twice the conditional base model, at the same sampling cost. These results demonstrate that few-step generation preserves structural quality while unlocking the extreme throughput required for large-scale materials screening.

1 INTRODUCTION

Generative models are becoming a practical tool for materials discovery. A typical computational materials discovery pipeline might involve the use of a generative model to propose candidate structures and subsequently filtering them using machine-learned interatomic potentials (MLIPs) and/or density functional theory (DFT) calculations (Yeo et al., 2026; Tao & Zhu, 2026). Sampling cost matters in these pipelines. Large-scale searches screen 10^5 to 10^9 candidates with ML filters before any DFT (Merchant et al., 2023; Schmidt et al., 2024), and diffusion-based generators such as MatterGen (Zeni et al., 2025) and DiffCSP (Jiao et al., 2023) spend 10^3 or more network function evaluations (NFE) on each one. Pipelines that denoise compact VAE latents (Park & Walsh, 2025; Joshi et al., 2025) make each step cheaper, since the network acts on a few latent channels per atom rather than directly on coordinates and lattice features, but still entail long denoising trajectories, typically 50 or more steps. Sampling cost therefore determines how many candidates a search can afford to screen.

We ask whether that step count can be reduced without degrading quality. Our starting point is MeanFlow (Geng et al., 2025). In standard flow matching (Lipman et al., 2023), the model learns the time-dependent vector field governing a probability-flow ODE, which requires iterative numerical integration to generate samples. MeanFlow instead learns the *average* velocity between two times $r < t$. One evaluation can therefore cover an entire interval of the trajectory, allowing the model to reach the data distribution in a few steps or even one.

We bring MeanFlow to crystals in three parts. First, we train a diffusion transformer (DiT) backbone (Peebles & Xie, 2023) in the per-atom latent space of a frozen VAE, so the generator never has to enforce periodicity in coordinate space. Second, we condition it on the crystal system and apply classifier-free guidance (CFG) (Ho & Salimans, 2022). Standard CFG runs the network twice per sampling step, with and without the label, so we distill the guided prediction into the regression target instead and keep a single evaluation per step. Guidance is applied only over an interval of times (Kynkäänniemi et al., 2024). Third, we anneal the objective from plain flow matching towards

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.

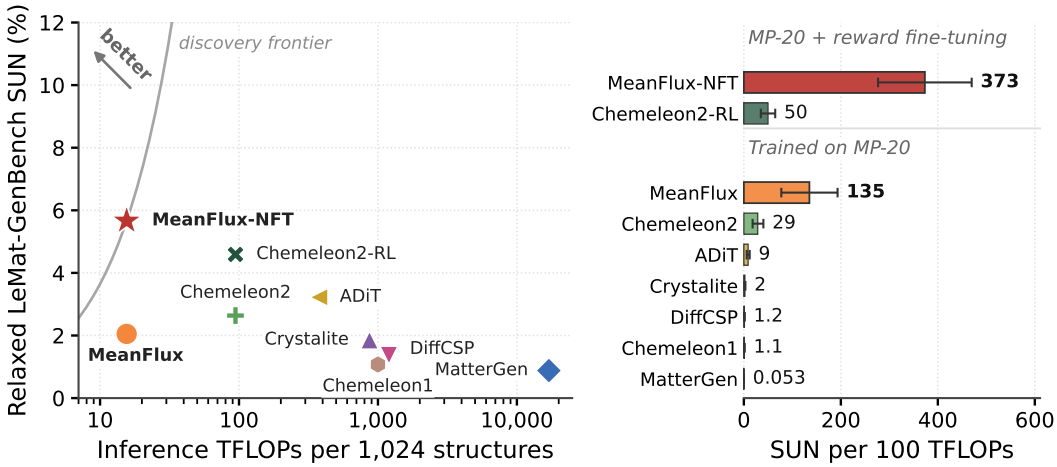


Figure 1: **Discovery efficiency under one sampling budget.** *Left:* LeMat-GenBench SUN rate, the share of all 1,024 generated structures that are stable, unique and novel after relaxation under the benchmark's three potentials (Table 4), against inference cost per 1,024 structures, for models trained on MP-20. Up and to the left is better. MeanFlux is the conditional 130M model at eight steps. The star is the unconditional 130M model after negative-aware reward fine-tuning (MF-NFT, §2.3), at the same cost, and the cross is Chemeleon2's released RL checkpoint. The grey curve is the discovery frontier, the rates and costs that match MF-NFT's yield per unit of compute; every other model lies below or to the right of it. *Right:* expected SUN crystals within a fixed 100-TFLOP budget; the two reward-finetuned models form their own group. Whiskers are 95% bootstrap confidence intervals of each measured rate, multiplied by the fixed number of draws the budget allows. We score every model under this one protocol, the five baselines of Table 1 on the structures that table scores.

MeanFlow over the first few thousand training steps. The resulting model, **MeanFlux**[1], samples a crystal in 1 to 16 NFE. In contrast, Chemeleon2's 50-step DDIM, run on the same backbone and decoder, can cost up to $40\times$ as much inference compute per structure, and MatterGen costs orders of magnitude more (Figures 1 and 2; the full estimates are included in Appendix B.3). The low step count also makes reinforcement learning (RL) cheap, so we fine-tune MeanFlux with negative-aware fine-tuning (NFT) (Zheng et al., 2025; Huang et al., 2026), using a reward that credits stability only for valid, unique and novel structures.

Contributions. Our main contributions are as follows:

- **Few-step crystal generation.** MeanFlux achieves the best mSUN at a given sampling cost on MP-20. Our 130M-parameter model matches Chemeleon2 in mSUN at one-sixth of its sampling cost and produces $5.9\times$ as many mSUN structures under a fixed compute budget (Table 1). Scored after relaxation on LeMat-GenBench, it yields almost five times as many SUN crystals per unit of compute as Chemeleon2 (Figure 1). Removing the average-velocity term (pure flow-matching) costs about 12pp of mSUN at eight steps (§3.6). Sampling quality also follows a scaling law in model size and keeps improving with training data (Figure 3; Appendices B.4 and B.8).
- **Structure prediction at a fraction of the cost.** A composition-conditioned variant outperforms the strongest published crystal structure prediction (CSP) model on MP-20 while using approximately $90\times$ less compute per candidate (Figure 4).
- **Fast reward fine-tuning.** Few-step sampling makes each rollout cheap, which speeds up the RL loop. Fine-tuning MeanFlux with NFT on its eight-step sampler doubles the relaxed LeMat-GenBench mSUN rate of the MP-20 model at the same sampling cost, and the fine-tuned model yields 7.5 times as many SUN crystals per unit of compute as Chemeleon2's RL checkpoint (§2.3, §3.4).

[1] *Mean* after MeanFlow's average-velocity objective; *flux* both as the Latin root of *flow* and after the flux method of crystal growth.

Related work is discussed in Appendix A.

2 METHOD

Following ADiT and Chemeleon2 (Joshi et al., 2025; Park & Walsh, 2025), MeanFlux separates representation from generation: a frozen variational autoencoder (VAE) maps each crystal to one latent per atom, and a diffusion transformer maps Gaussian noise to those latent sequences. The decoder runs once after sampling, leaving periodic coordinates and lattice geometry to the VAE. In one ablation we reuse Chemeleon2's released VAE checkpoint without modification, so that MeanFlux and Chemeleon2 share the same latent space and decoder and differ only in the generator. The binary-code variant (§2.1) is instead trained from scratch with a different objective.

2.1 LATENT CRYSTAL AUTOENCODER

A crystal with N atoms is $C = (A, F, L)$. Here $A = (a_1, \ldots, a_N)$ gives the atomic numbers, $F \in [0,1)^{N\times 3}$ the fractional coordinates, and $L = (a, b, c, \alpha, \beta, \gamma)$ the lattice lengths and angles. Following ADiT and Chemeleon2, the encoder represents atom i by $h_i = E_{\text{at}}(a_i) + W_{\text{lat}}\,\text{vec}(\mathbf{L}_{3\times 3}) + W_{\text{fc}} f_i + \text{PE}(i)$, combining its element, the shared lattice, its position in the cell, and its position in the token sequence. A pre-norm transformer maps these tokens to a diagonal Gaussian posterior $q_\phi(z_i \mid C) = \mathcal{N}(\mu_i, \text{diag}\,\sigma_i^2)$, with $z_i \in \mathbb{R}^{d_z}$ and $d_z = 8$.

The decoder reconstructs atom types $\hat{a}_i$, fractional coordinates $\hat{f}_i$, and the six lattice parameters $\hat{L}$; a pooled global token predicts the lattice. Training minimizes

$$\mathcal{L}_{\text{VAE}} = \lambda_{\text{at}}\,\text{CE}(\hat{a}, a) + \lambda_\ell \left\|\hat{L} - L\right\|^2_{\text{len+ang}} + \lambda_f \left\|\hat{F} - F\right\|^2 + \beta\,\text{KL}(q_\phi \,\|\, \mathcal{N}(0, I))\,, \qquad (1)$$

where the lattice term separately weights length and angle errors and the KL term regularizes the latent distribution. We then freeze the VAE. Before fitting MeanFlux, we standardize all encoded latents by one scalar latent standard deviation and reverse the transformation before decoding.

Binary-code variant (bcb). We also test a more compact representation, in the spirit of CliqueFlowmer's latent codes (Kuba et al., 2026) and Crystalite's compact subatomic atom-type encoding (Hadži Veljković et al., 2026). It replaces the 100-way atom-type head with a 7-bit binary encoding of the atomic number, trained with binary cross-entropy, and concatenates the token inputs before an MLP instead of adding them (d_z=16). This variant halves decoder inference FLOPs ($0.530 \to 0.265$ TFLOPs per 1,024 samples) at comparable quality (Figure 2, *MF-bcb*).

2.2 MEANFLOW DIT IN LATENT SPACE

Flow matching (Lipman et al., 2023) defines a linear trajectory $z_t = (1-t)x + te$ connecting a data latent $x \in \mathbb{R}^{N\times d_z}$ to standard Gaussian noise $e \sim \mathcal{N}(0, I)$, where $t = 0$ represents data, $t = 1$ represents noise, and the conditional velocity is $v = e - x$, whose expectation given z_t is the marginal velocity field $v(z_t, t)$. Standard flow models learn this instantaneous velocity at a single time t, requiring multi-step numerical integration backward from noise during sampling.

MeanFlow (Geng et al., 2025) instead predicts the time-averaged velocity between r and t,

$$u(z_t, r, t) = \frac{1}{t-r}\int_r^t v(z_s, s)\,ds. \qquad (2)$$

This enables a direct finite update $z_r = z_t - (t-r)\,u(z_t, r, t)$, and u reduces to the instantaneous velocity as $r \to t$. Because the integral is not available as a training target, MeanFlow instead uses the identity $u(z_t, r, t) = v(z_t, t) - (t-r)\,\frac{d}{dt}u(z_t, r, t)$. Here, $\frac{d}{dt}u = \partial_t u + (v \cdot \nabla_z)u$ is the total derivative along the trajectory z_t, and both terms are obtained with a single Jacobian-vector product (JVP). As in flow matching, training replaces the marginal $v(z_t, t)$ with the conditional v.

Conditioning and guidance distillation. We provide the seven-class crystal system c through adaptive layer normalization (adaLN) (Peebles & Xie, 2023) and replace it with a null token $\varnothing$ for 20% of training examples. This trains the conditional and unconditional branches required by classifier-free guidance (CFG) (Ho & Salimans, 2022). Standard CFG evaluates both branches at each sampling

step. To keep one evaluation per step, we instead build the guidance into the training target. The guided velocity is

$$\tilde{v} = w\, v + (1 - w)\, u_\theta(z_t, t, t \mid \varnothing), \tag{3}$$

with guidance scale w, applied only to rows that keep their label and only for t inside a guidance interval (Kynkäänniemi et al., 2024), $(0.2, 0.8)$ in our reference model. Elsewhere $\tilde{v} = v$.

Training objective. The model is trained on

$$\mathcal{L}_{\mathrm{MF}} = \mathbb{E}\Big[\mathrm{sg}\big[\omega(\delta^2)\big]\, \delta^2\Big], \qquad \delta^2 = \frac{1}{N d_z} \sum_{i=1}^{N} \big\| u_\theta(z_t, r, t \mid c)_i - u_i^{\mathrm{tgt}} \big\|_2^2, \tag{4}$$

with the bootstrapped target

$$u^{\mathrm{tgt}} = \mathrm{sg}\Big[\tilde{v} - \mathrm{clip}_\kappa\big((t - r)\,(\partial_t u_\theta + \tilde{v} \cdot \nabla_z u_\theta)\big)\Big], \qquad \mathrm{clip}_\kappa(y) = \frac{\kappa\, y}{\max(\kappa, \|y\|_2)}. \tag{5}$$

The sum runs over the N real atoms of each crystal, so padding tokens are excluded and δ^2 is the squared residual averaged over that crystal's atoms and latent channels. The bracketed JVP term is one Jacobian-vector product with tangent $(\tilde{v}, 1, 0)$ in (z_t, t, r). It is first multiplied by the interval $(t - r)$, and the resulting correction, taken as a single vector over the crystal's $N d_z$ entries, is then rescaled to Euclidean norm at most $\kappa = 1$. Without this cap the JVP term diverges in the first few hundred steps. On flow-matching rows ($r = t$) the correction vanishes and the target is $\tilde{v}$ itself. The stop-gradient $\mathrm{sg}[\cdot]$ fixes the target, and the unconditional branch inside $\tilde{v}$ is evaluated without gradient as well. The adaptive weight $\omega(\delta^2) = (\delta^2 + \epsilon)^{\gamma - 1}$, with $\gamma = 0.5$ and $\epsilon = 10^{-3}$, is also held under stop-gradient, so it rescales each crystal's gradient without redirecting it, and the loss is the plain mean of $\omega\, \delta^2$ over the crystals in a batch with no further normalization.

Time sampling and flow-ratio annealing. We draw r and t independently from $\mathrm{sigmoid}(\mathcal{N}(-0.4, 1.0^2))$ and sort them (Esser et al., 2024). The *flow ratio* ρ is the fraction of examples set to $r = t$, where the objective becomes ordinary flow matching. We reduce ρ from 1 to 0.25 over the first 3k steps, introducing the bootstrapped target only after the model has learned the local velocity. Section 3.6 ablates the final value.

Sampling. For a uniform schedule $t_0 = 1 > \cdots > t_K = 0$, we set $z_{t_0} = e$ and update $z_{t_{k+1}} = z_{t_k} - (t_k - t_{k+1})\, u_\theta(z_{t_k}, t_{k+1}, t_k \mid c)$. Thus K is the NFE; $K = 1$ gives $z_0 = e - u_\theta(e, 0, 1 \mid c)$. The VAE decodes the final latents once.

Backbone and recipe. u_θ is a DiT with two continuous time embeddings (t, r) and adaLN conditioning. The size ladder spans pico (1.3M) to large (307M), and base (130M) is our reference. The reference run trains on MP-20 and selects its checkpoint on validation loss. Appendix B.1 gives the full recipe.

2.3 Reward fine-tuning

We fine-tune MeanFlux with negative-aware fine-tuning (NFT) (Zheng et al., 2025) in the form given by MeanFlowNFT (Huang et al., 2026) for average-velocity models. The VAE remains frozen, and samples are drawn with the eight-step sampler of §2.2. Each iteration draws a batch of latents from the current policy, decodes and scores them with the reward below, and standardizes the rewards within groups of sixteen draws with the same atom count. The standardized reward is mapped to a weight $p \in [0, 1]$. Each sample is then re-noised along the training path, and the model is regressed toward it with weight p and away from it with weight $1 - p$. Both terms act on the instantaneous velocity that the MeanFlow identity assigns to u_θ, through two implicit predictors, and a squared-velocity penalty keeps the policy close to a frozen reference. The objective requires neither a trajectory likelihood nor a gradient through the MLIPs. Appendix B.2 gives the loss and all hyperparameters.

Crystal reward. The reward follows the LeMat-GenBench criteria. Let e be the single-point energy above the hull of the decoded crystal, averaged over the three LeMat-GenBench potentials. Let H, U and N be indicators that the crystal passes the benchmark's validity filter, is unique within its rollout batch, and is novel against LeMat-Bulk. The reward is

$$R = HUN\big(8M(e) + 8S(e) + 0.25\big) + G\big(0.35 + E(e) + 1.5D\big), \tag{6}$$

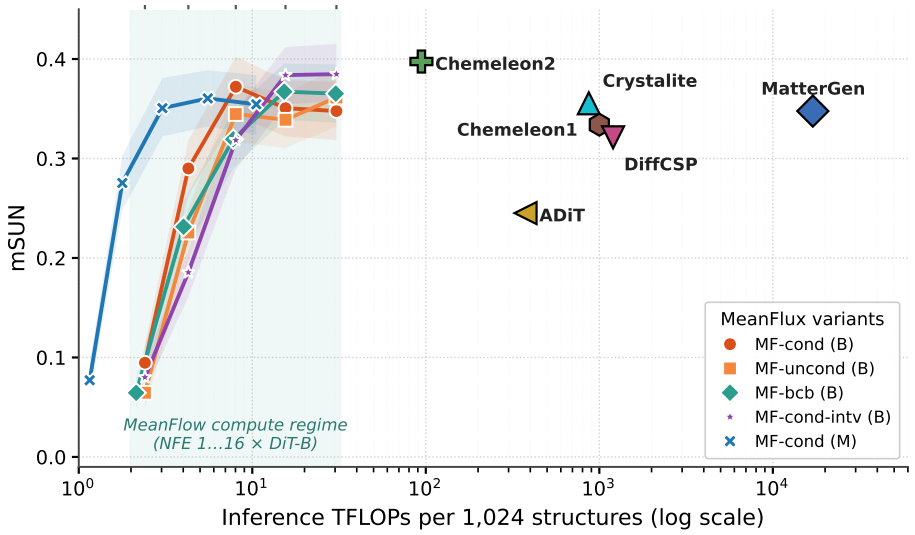


Figure 2: **Compute–quality frontier on MP-20** (in-distribution suite: n=1024, MACE relax-50, test prior; single-MACE E_{hull}). MeanFlux variants swept over NFE $\in \{1, 2, 4, 8, 16\}$ against baselines placed at their inference FLOPs per 1,024 structures. Variant labels: *cond/uncond*, with or without crystal-system conditioning; *intv*, guidance restricted to the interval of §2.2; *bcb*, the binary-code VAE of §2.1; (B)/(M), base 130M / medium 44M DiT. Every point is scored under the identical MACE relax-50 evaluation protocol, each baseline re-scored here from its released model's own structures (Table 1). Shading around each MeanFlux curve is the 95% bootstrap confidence interval over its n=1024 draw. The best 130M variant (*MF-cond-intv*) draws level with Chemeleon2 by 8 NFE at $\sim$6$\times$ less compute, and reaches mSUN 0.38 for $\sim$550$\times$ less compute than MatterGen.

where M and S are logistic scores of metastability and stability, E is a bounded energy score, $G \in [0, 1]$ is a soft gate on charge balance, interatomic distances and density, and $D \in [0, 1]$ is a leave-one-out diversity score over composition and geometry (Appendix B.2). The first term is nonzero only when H, U and N all hold, so a stable structure that is invalid, duplicated or already known receives no stability reward. The second term is smaller and provides a signal for samples that fail one of the three conditions. Energies are single-point, and no relaxation is run during training. The reported checkpoint, **MF-NFT**, is fine-tuned from the unconditional 130M model with 256 draws per iteration (Appendix B.2).

2.4 Composition-conditioned CSP variant

CSP asks for the structure given the composition. We cannot simply reuse the de-novo model, because its latent entangles atom types with geometry. We instead retrain the DiT on a frozen geometry-only VAE, obtained by removing the atom-type head. Each token receives an embedding of its known atomic number, and the decoder attaches the given species token by token. DiffCSP and Crystalite clamp the type channel in the same way. Evaluation follows the DiffCSP protocol (Jiao et al., 2023): 20 candidates per test composition, each compared with the ground-truth structure by pymatgen's `StructureMatcher` (Ong et al., 2013), at the tolerances of Appendix B.5.

3 Experiments

3.1 Setup

Datasets. We use three corpora of varying sizes. MP-20 is the standard benchmark of 45,229 Materials Project structures in total across its train, validation, and test splits, each with at most 20 atoms (Jain et al., 2013; Xie et al., 2022). MP-40 allows up to 40 atoms and tests temporal generalization: structures entered before 4 August 2018 form the training and validation sets, while later entries form the test set. Alexandria-MP-20 has 540k training structures (Schmidt et al., 2024; Zeni

Table 1: MP-20 de-novo generation, ordered by inference cost within each block. TFLOPs = inference FLOPs per 1,024 structures (DiT forwards + VAE decode). *Every row is scored on relaxed structures*, on one in-distribution suite (n=1024, test prior, MACE relax-50, single-MACE E_{hull} against MP-20). Every baseline is re-scored here, passing through the same protocol rather than being taken verbatim from the literature. Chemeleon2 is unconditional, without RL fine-tuning (Appendix B.6). Bold marks the best value in each column and, on mSUN, the two MeanFlux scores within measurement error of it, the eight-step one at a sixth of Chemeleon2's cost. Subscripts are 95% bootstrap confidence half-widths over the n=1024 draw.

Model	NFE ↓	TFLOPs ↓	Metastable ↑	Novel ↑	SUN ↑	mSUN ↑
Chemeleon2	50	94.3	$0.633_{\pm.028}$	$0.756_{\pm.027}$	$0.067_{\pm.016}$	$\mathbf{0.397}_{\pm.030}$
ADiT	200	375.7	$\mathbf{0.785}_{\pm.025}$	$0.415_{\pm.030}$	$0.043_{\pm.012}$	$0.245_{\pm.026}$
Crystalite (de novo)	299	869.8	$0.567_{\pm.030}$	$0.781_{\pm.025}$	$0.057_{\pm.016}$	$0.355_{\pm.029}$
Chemeleon1	512	1,000	$0.514_{\pm.031}$	$0.812_{\pm.024}$	$0.036_{\pm.011}$	$0.334_{\pm.028}$
DiffCSP	2000	1,200	$0.462_{\pm.030}$	$0.855_{\pm.022}$	$0.052_{\pm.014}$	$0.321_{\pm.029}$
MatterGen	2000	17,000	$0.481_{\pm.030}$	$0.862_{\pm.021}$	$0.034_{\pm.011}$	$0.348_{\pm.029}$
MeanFlux DiT-M (44M, cond.)	2	**1.8**	$0.328_{\pm.030}$	$\mathbf{0.947}_{\pm.013}$	$0.054_{\pm.014}$	$0.275_{\pm.029}$
MeanFlux DiT-B (130M, uncond.)	4	8.0	$0.480_{\pm.032}$	$0.859_{\pm.021}$	$0.067_{\pm.014}$	$0.345_{\pm.029}$
MeanFlux DiT-B (130M, cond.)	8	15.5	$0.639_{\pm.028}$	$0.738_{\pm.026}$	$\mathbf{0.089}_{\pm.018}$	$\mathbf{0.384}_{\pm.029}$
MeanFlux DiT-B (130M, cond.)	16	30.6	$0.673_{\pm.029}$	$0.704_{\pm.028}$	$0.078_{\pm.017}$	$\mathbf{0.385}_{\pm.029}$

et al., 2025). The structure-prediction experiments also use MPTS-52 (Jiao et al., 2023). It similarly trains on older Materials Project entries and tests on later additions, with cells of up to 52 atoms.

Metrics. We report SUN and mSUN, the fraction of generated structures that are at once stable (or metastable, within 0.1 eV/atom of the hull after MACE relaxation (Batatia et al., 2022)), unique, and novel. Uniqueness and novelty are decided by `StructureMatcher`, as in CSP (§2.4) but at pymatgen's default tolerances. Two evaluation protocols are used. In-distribution numbers come from a suite that draws 1,024 structures, relaxes each for 50 MACE steps, and scores against the MP-20 reference. Out-of-distribution numbers come from LeMat-GenBench (Betala et al., 2025), which applies a three-part validity filter, an ORB, MACE and UMA ensemble hull, and novelty against the five million structures of LeMat-Bulk. Its mSUN, unlike the in-distribution one, excludes the structures already counted in SUN. Inference compute is reported as FLOPs per 1,024 generated structures. Appendix B.5 gives the exact settings of both protocols and Appendix B.3 the compute estimates.

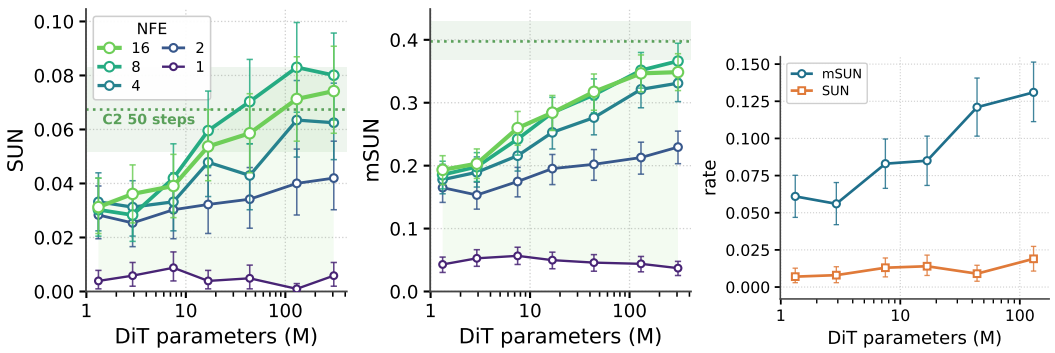
3.2 Matched quality on MP-20 at a sixth of the compute

Figure 2 and Table 1 summarize the main result. Quality climbs steeply between one and four steps and then levels off. At eight steps the conditional base model reaches an mSUN of 0.384 and Chemeleon2 reaches 0.397. The 1.3pp gap is smaller than the bootstrap half-widths of about 3pp, so the two are indistinguishable on this test suite. On the same comparison MeanFlux matches Chemeleon2's metastability (0.639 vs. 0.633) and returns a third more SUN crystals (0.089 vs. 0.067), at a sixth of the sampling compute. Drawing a thousand crystals takes MeanFlux under a second on an H100 and MatterGen about 45 minutes (Appendix B.3).

We further show how many mSUN crystals a fixed compute budget yields. At 100 TFLOPs Chemeleon2 draws about 1,100 structures and, at its 0.397 rate, yields about 430 mSUN crystals. The 44M model at two steps has a lower rate of 0.275, but the same budget lets it draw 57,000 structures. That works out to around 15,800 mSUN crystals, 37× more than Chemeleon2. The 130M model at eight steps returns 5.9× more. This is the in-distribution form of the discovery-efficiency argument of §3.4. Few-step sampling prevails by drawing far more structures per unit of compute, at a comparable hit rate per structure.

3.3 Scaling in model size and data

Parameters. Holding the recipe and the frozen VAE fixed, we trained a ladder of seven sizes from 1.3M to 307M parameters, spanning 2.4 orders of magnitude. Every model in the ladder is trained



(a) MP-20 parameter ladder (SUN left, mSUN right).

(b) MP-40 ladder (NFE 16).

Figure 3: **Model-size scaling** (unconditional ladder, in-distribution suite, n=1024, MACE relax-50). In (a), dotted lines mark the relaxed Chemeleon2 baseline of Table 1 at each metric. Panel (b) repeats the ladder on the harder MP-40 split, with sizes up to 130M. In (a), mSUN increases monotonically in parameters at NFE $\geq$ 4. NFE 1 lags behind at every size. The reducible gap follows the power law of equation 7 (α_N=0.12 at NFE 8, R^2=0.99); fit in Appendix B.4, Figure 7. Error bars and the shading around the Chemeleon2 lines are 95% bootstrap confidence intervals. Both ladders hold the data fixed; the isoFLOP sweep of Figure 6 varies size and data together on Alexandria-MP-20, whose larger corpus supports that stricter test.

on the same corpus for the same number of steps, so the training tokens D are held fixed and training compute $C \approx 6ND \propto N$ (Kaplan et al., 2020). The ladder therefore isolates the model-size term of a scaling law.

Since mSUN is a bounded success rate instead of a loss, we adapt the standard scaling-law form (Hoffmann et al., 2022) to the reducible gap between mSUN and an irreducible ceiling,

$$\mathrm{mSUN}_\infty - \mathrm{mSUN}(N) = A\,N^{-\alpha_N}, \tag{7}$$

and fit it per NFE by least squares (Appendix B.4, Figure 7). The fit is tight at the full-quality budgets, $\alpha_N = 0.12$ at eight NFEs and 0.23 at sixteen, both with $R^2 \geq 0.98$ and jackknife-stable to within ± 0.02. The small exponent quantifies a steep diminishing return: halving the gap to the ceiling costs a factor $2^{1/\alpha_N} \approx 20$–$400\times$ in compute. Past some point, growing the model at fixed data is therefore a poor use of compute, and the data has to grow with it.

Data. Out of distribution, the training data matter more than the sampler. On LeMat-GenBench, every base model trained only on MP-20, ours and Chemeleon2's, reaches 2–3% mSUN. Continuing training on Alexandria-MP-20 triples this and lowers the mean hull distance by a third (Appendix B.8). Over the course of Alexandria training, validity and metastability keep improving while novelty falls gradually (Appendix Figure 9). This experiment grows the data D at a fixed model size N. Its gains are limited by the corpus, so we do not fit an exponent to it.

To vary N and D together, we run an isoFLOP sweep on Alexandria-MP-20. For each training budget $C \approx 6ND$, it finds the model size that gives the best mSUN. Between 10^{16} and 3×10^{17} FLOPs, this size grows as $N_{\mathrm{opt}} \propto C^{0.80\pm0.08}$, so the number of training tokens grows as $D_{\mathrm{opt}} = C/6N_{\mathrm{opt}} \propto C^{0.20}$. A ten times larger budget is therefore best spent on a model about six times larger, trained on about 1.6 times as many tokens (Appendix B.4, Figure 6).

3.4 Discovery efficiency under a fixed sampling budget

Discovery efficiency is best quantified by how many useful new materials a generator yields per unit of compute. That yield is the product of two quantities that often go against each other: how often a generator produces something worth keeping, and how many candidates it can produce at all. Counting qualified candidates (e.g., SUN crystals) under a fixed budget, a model with a high hit rate that takes 800 steps per structure can easily lose to a decent one that takes four.

Figure 1 scores generators by their yield of SUN crystals per unit of compute. Before relaxation no model has more than 5 SUN structures in 1,024 draws, so we relax every valid structure under each

of the three LeMat-GenBench potentials and score all models after relaxation, with rates over all 1,024 draws (Appendix B.8). We run every model ourselves under this one protocol: the five MP-20 baselines of Table 1 on the structures that table scores, and MeanFlux, MF-NFT and Chemeleon2's released base and RL checkpoints as separate cohorts (Appendix B.2). Costs are the inference estimates of Appendix B.3.

We fix the budget at 100 TFLOPs. At the effective H100 throughput assumed in Appendix B.3, that takes about two seconds of sampling. MeanFlux spends it on around 6,600 structures, Chemeleon2 on about 1,100, ADiT on 270, Crystalite on 118 and MatterGen on 6. The SUN rate of MeanFlux, 2.1%, is below Chemeleon2's 2.6% and ADiT's 3.2%, yet it returns about 135 expected SUN crystals, almost five times Chemeleon2's 29 and fifteen times ADiT's 9.

Reward fine-tuning moves the frontier. At the same eight-step cost, MF-NFT reaches 5.7% SUN, almost three times the conditional base model, and returns about 370 SUN crystals per 100 TFLOPs. Chemeleon2's RL checkpoint raises its base model's rate from 2.6% to 4.6%, but it costs six times as much per draw and returns about 50, so MF-NFT yields 7.5 times as many. Scored by relaxed mSUN instead (Table 4), MF-NFT reaches 22.5%, twice the conditional base model, and Chemeleon2 with RL reaches 44.0%.

3.5 Crystal structure prediction

Figure 4 plots match rate against inference cost for the composition-conditioned ladder of §2.4, with Crystalite and others as the reference point. On MP-20 the large MeanFlux model recovers $\mathbf{69.1}\%$ of test structures at sixteen steps, above Crystalite's 66.1% and for around $90\times$ less compute per candidate, and the ladder as a whole traces a Pareto frontier. The flow-matching baselines FlowMM and CrystalFlow sit below that frontier. They need 10^2–10^3 ODE steps to reach match rates that our base model crosses at four (Appendix Table 3).

MPTS-52 is harder because it tests on later Materials Project entries and includes cells with up to 52 atoms. The same method reaches 25.0%, ahead of the flow-matching baselines but behind Crystalite's 31.7%. We attribute the shortfall to the model scale. Match@1 follows the power law of equation 7 in N on both splits, but the fitted exponent is far smaller on MPTS-52, where the ladder is still well short of saturation at 307M (Appendix B.4). Extrapolating that fit reaches Crystalite's parity near a projected 2B parameters at sixteen steps (Figure 4b, dashed). Even after such an upscale, MeanFlux would sample each candidate at roughly $14\times$ lower inference cost, since it would use sixteen network evaluations instead of Crystalite's 799 (best-fit projection; range in Appendix B.4).

3.6 Ablations

Flow ratio. The annealing endpoint ρ is the fraction of training rows with $r{=}t$. At $\rho{=}1$ every row has $r{=}t$, the MeanFlow term drops out, and the objective is plain flow matching. Sweeping ρ therefore measures what the MeanFlow term contributes. All runs with $\rho < 1$ reach similar mSUN, while the $\rho{=}1$ run falls 10–12pp below them and loses more than half of its SUN (Figure 5). The backbone (an unconditional DiT-M) and autoencoder are the same in every run, so we attribute the few-step quality to the average-velocity objective itself.

VAE decoder. The binary-code autoencoder halves decoder FLOPs and tracks the default within noise on the frontier (Figure 2).

Guidance. Dropping either the guidance interval or the conditioning was uniformly worse in development runs.

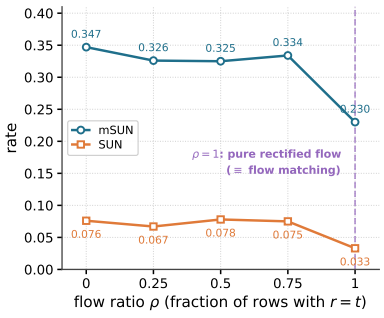
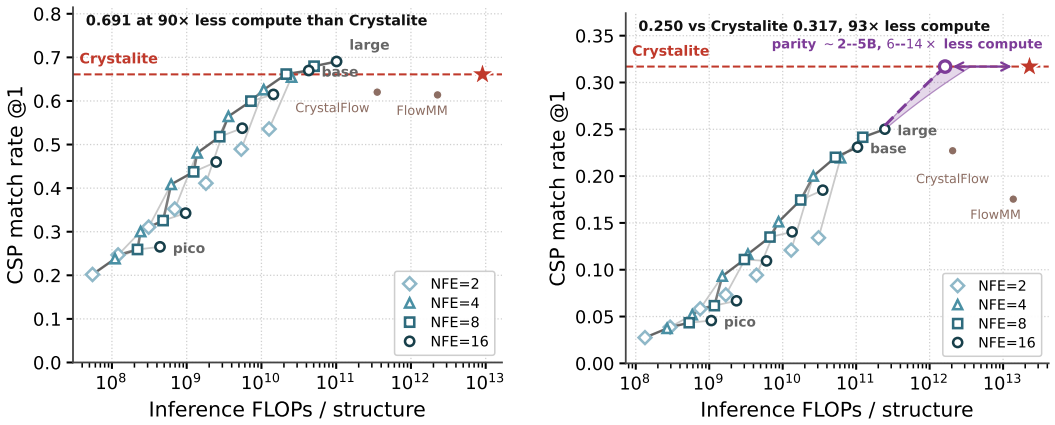


Figure 5: **Flow-ratio ablation** (NFE 8, relaxed). Pure rectified flow ($\rho{=}1$) loses $\sim$12pp mSUN.

4 Discussion

We attribute the success of mean flow matching here to the regularity of the latent velocity field. Per-atom VAE latents have only eight or sixteen channels, the KL term pulls them toward an isotropic



(a) MP-20. (b) MPTS-52.

Figure 4: **CSP scaling in inference FLOPs** (unrelaxed, full test splits). Markers are NFE $\in \{2, 4, 8, 16\}$; gray lines connect one model size across NFE and the black Pareto frontier tracks the best match@1 per budget. The red star/line is Crystalite's match@1 at its inference cost, counted over its transformer as $\approx 2N \cdot \text{tokens} \cdot \text{NFE}$ at its *padded* length ($n_{\max}+1$ tokens), since it runs a fixed-length batch while our models pack to the mean atom count. Brown dots are the flow-matching baselines CrystalFlow (100 ODE steps) and FlowMM (1000), incurring an explicit per-structure matmul count of their released CSP networks, including edge messages. Both are situated off the frontier. *(b)* The purple curve extrapolates the NFE-16 fit of equation 7 to Crystalite's parity, the shaded band sweeping the unconstrained ceiling over 0.5–1.0; a projected 2–5B-parameter model reaches 31.7% at 6–14× lower inference cost (Appendix B.4).

Gaussian, and they carry none of the periodicity and symmetry constraints of coordinate space. We therefore expect the paths from the prior to the latent manifold to be short and close to straight, a regime in which a few large steps approximate the average velocity well. Separating chemistry from geometry may help too. Crystalite (Hadži Veljković et al., 2026) gives atom types their own channel next to continuous coordinates and lattice, and our binary atom-type codes quantize chemistry while geometry stays continuous. Finally, mSUN is still rising along its power law at the top of our ladder (Figure 3, Appendix B.4), so we expect further gains from larger models trained on broader corpora. The low step count also makes reward fine-tuning inexpensive. NFT trains the eight-step sampler directly and gives almost three times the SUN yield of the conditional model at the same cost (Figure 1). Like Chemeleon2's RL checkpoint, MF-NFT is trained on the benchmark's potentials and novelty reference. Its gain nonetheless holds for relaxed energies, which the reward does not use, and on the MP-20 suite, where SUN rises from 0.058 to 0.110 (Appendix B.6).

5 Conclusion

MeanFlux shows that crystal generation does not need long sampling trajectories. Learning the average velocity in a frozen per-atom latent space lets a 130M-parameter model match Chemeleon2's mSUN on MP-20 in eight steps, with a denoiser of the same size and one sixth of the sampling compute. The same recipe gives the best CSP match rate on MP-20 at about 90× less compute per candidate than Crystalite, and quality improves predictably with model size. Because each sample is cheap, a fixed budget affords many more candidates. Scored after relaxation on LeMat-GenBench, it yields almost five times as many stable, unique and novel crystals per unit of compute as Chemeleon2. Reward fine-tuning with NFT doubles the relaxed LeMat-GenBench mSUN rate of the MP-20 model at the same sampling cost. Sampling cost therefore no longer has to limit how many candidates a generator can pass to MLIP and DFT screening. The remaining limits are the training corpus and the conditioning, which point to broader pretraining and symmetry-aware conditioning, combined with reward fine-tuning, as next steps.

Reproducibility statement

All models, evaluation pipelines, and figures will be available in the MeanFlux code release. Appendix B.1 lists the full reference training recipe, Appendix B.2 the fine-tuning recipe and checkpoint selection, and Appendix B.5 the exact settings of both evaluation protocols.

AI use statement

We used generative AI tools as coding and writing assistants. They helped implement and refactor training, evaluation and plotting code, and they assisted in drafting and editing the text of this paper. The research questions, experimental design, and the interpretation of the results are the authors' own. All AI-assisted code was reviewed by the authors and its outputs were verified against the evaluation artifacts reported here, and all AI-assisted text was reviewed and edited by the authors. We take responsibility for the final content of this work, including its text, claims and artifacts.

References

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. 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*, 2022.

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. Lematgenbench: A unified evaluation framework for crystal generative models. *arXiv preprint arXiv:2512.04562*, 2025.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. *arXiv preprint arXiv:2410.12557*, 2024.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

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.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Yushi Huang, Xiangxin Zhou, Jun Zhang, Liefeng Bo, and Tianyu Pang. MeanFlowNFT: Bringing forward-process RL to average-velocity generators. *arXiv preprint arXiv:2607.15273*, 2026. URL `https://arxiv.org/abs/2607.15273`.

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

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*, 2023.

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

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. In *International Conference on Machine Learning*, 2025.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Jakub Grudzien Kuba, Ben Kurt Miller, Sergey Levine, and Pieter Abbeel. Solving offline materials optimization with cliqueflowmer. *arXiv preprint arXiv:2603.06082*, 2026.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

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

Xiaoshan Luo, Zhenyu Wang, Pengyue Gao, Jian Lv, Yanchao Wang, Changfeng Chen, and Yanming Ma. CrystalFlow: a flow-based generative model for crystalline materials. *Nature Communications*, 16, 2025. doi: 10.1038/s41467-025-64364-4.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624:80–85, 2023.

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

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.

Hyunsoo Park and Aron Walsh. Guiding generative models to uncover diverse and novel crystals via reinforcement learning. *arXiv preprint arXiv:2511.07158*, 2025.

Hyunsoo Park, Anthony Onwuli, and Aron Walsh. Exploration of crystal chemical space using text-guided generative artificial intelligence. *Nature Communications*, 16:4379, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Jonathan Schmidt, Tiago F. T. Cerqueira, Aldo H. Romero, Antoine Loew, Fabian Jäger, Hai-Chen Wang, Silvana Botti, and Miguel A. L. Marques. Improving machine-learning models in materials science through large datasets. *Materials Today Physics*, 48:101560, 2024.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

Shuo Tao and Qiang Zhu. Accelerated inorganic electrides discovery by generative models and hierarchical screening. *PRX Intelligence*, 1:013016, 2026.

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*, 2022.

Byung Chul Yeo, Hyun-Jae Lee, Sungwoo Kang, and Jung-Hoon Lee. Diffusion-model-driven discovery of ferroelectrics for photocurrent applications. *Advanced Science*, 2026.

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*, 639:624–632, 2025.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025. URL https://arxiv.org/abs/2509.16117.

A Related work

Crystal generative models. CDVAE (Xie et al., 2022) and DiffCSP (Jiao et al., 2023; 2024) established score-based generation of periodic structures; MatterGen (Zeni et al., 2025) scaled it with broader data; FlowMM (Miller et al., 2024) and CrystalFlow (Luo et al., 2025) replaced the score with flow matching, still integrating 10^2–10^3 ODE steps. ADiT (Joshi et al., 2025) moved generation into a transformer-VAE latent space, followed by Chemeleon (Park et al., 2025) and Chemeleon2 (Park & Walsh, 2025), the latter adding GRPO-based RL fine-tuning on MLIP rewards. Crystalite (Hadži Veljković et al., 2026) also targets CSP. It is a lightweight diffusion transformer that denoises padded per-atom tokens of atom type, fractional coordinates and lattice directly in coordinate space, with no autoencoder or latent codes, using an EDM-style Heun sampler at NFE 299 for de novo generation and 799 for CSP. All of these spend 50–5000 NFE per sample. Our focus is the few-NFE regime.

Few-step generative models. Rectified flow and flow matching (Liu et al., 2023; Lipman et al., 2023) straighten probability-flow trajectories. Consistency and shortcut models (Song et al., 2023; Frans et al., 2024) learn direct jumps along them. MeanFlow (Geng et al., 2025) regresses the *average velocity* between two times, using an identity that it evaluates with a single Jacobian-vector product (JVP), and drifting (Deng et al., 2026) trains a one-step generator on kernel-estimated displacement targets. We adapt the former to crystals. MeanFlux is therefore not a consistency model. Consistency training anchors every jump at the fixed data endpoint, which is $r = 0$ in our notation. MeanFlow instead models the average velocity over pairs (r, t) drawn freely along the trajectory (§2.2).

Reinforcement learning for generators. Chemeleon2 (Park & Walsh, 2025) fine-tunes its latent diffusion model with GRPO on MLIP rewards, which requires likelihood ratios along a stochastic reverse trajectory. DiffusionNFT (Zheng et al., 2025) instead defines the objective on the forward process. It re-noises samples from the policy and regresses toward implicit positive and negative predictors, so it places no constraint on the sampler. MeanFlowNFT (Huang et al., 2026) extends this objective to average-velocity models. We use the MeanFlowNFT objective without modification and adapt the representation, reward and evaluation to crystals.

Latent representations. CliqueFlowmer (Kuba et al., 2026) encodes crystals into a latent space, optimizes latents, and decodes candidates, arguing that a well-shaped latent manifold is the right substrate for search. We share that view, and generate entirely in frozen per-atom VAE latents, including the cheaper binary-code variant of §2.1. Crystalite shows that the opposite choice, a transformer over raw atom tokens, can also lead the leaderboard, at a far higher step count.

B Training and evaluation details

B.1 Reference recipe

The headline conditional base model is trained on MP-20 with: frozen Chemeleon2 VAE (`dng_m4owq4i5_v0`: d_{model}=512, 8 heads, FF 2048, 8 layers, d_z=8, 100 elements); DiT-B (depth=12, d=768, 12 heads, 130.3M params); crystal-system conditioning (7 classes, `cfg-ratio` 0.2, `cfg-scale` w=7, interval $(0.2, 0.8)$, unconditional rows keep v); 1000 epochs, batch 256, AdamW lr 10^{-4} (2k warmup, cosine to 10^{-6}), weight decay 0.05, dropout 0.1, grad clip 1.0; EMA 0.9999; flow ratio 0.25 with 3k-step linear anneal from 1.0; lognormal (r, t) ($\mu = -0.4$, σ=1.0); JVP correction norm cap κ=1.0 and adaptive weight $(\gamma, \epsilon) = (0.5, 10^{-3})$, Eqs. 4–5; fractional-translation augmentation 0.05; checkpoint selection on validation loss over the EMA weights. The medium conditional model replaces the backbone with depth=9, d=512 (43.9M) and comes from an earlier recipe: `cfg-scale` w=2 at every t (no interval), `cfg-ratio` 0.1, dropout 0.15, weight decay 0.1 and a 5k-step flow-ratio anneal. The Alexandria-MP-20 continuation of Figure 9 warm-starts a conditional MP-20 DiT-B on Chemeleon2's released Alexandria-MP-20 VAE (`dng_j1jgz9t0_v1`) and trains for 1,024 epochs; its cosine schedule, first set for 32 epochs, warm-restarts at epoch 32 when training was extended. VAE (§2.1): 50 epochs, batch 128, lr 5×10^{-5}, loss weights $(\lambda_{\text{at}}, \lambda_{\ell}, \lambda_{\text{ang}}, \lambda_f, \beta) = (1, 1, 1, 10, 10^{-3})$ for the bcb variant; the Chemeleon2 VAE uses its released configuration.

B.2 REWARD FINE-TUNING

Objective. We follow MeanFlowNFT (Huang et al., 2026). For a rollout latent x_0, $\epsilon \sim \mathcal{N}(0, I)$ and $z_t = (1-t)x_0 + t\epsilon$, rewards are standardized within the atom-count groups g of §2.3, $a_i = (R_i - \overline{R}_g)/(\mathrm{std}(R_g)+10^{-6})$, and mapped to $p_i = \frac{1}{2} + \mathrm{clip}(a_i, -A, A)/(2A)$ with $A = 2$. The current, old and reference policies share the correction $(t-r)\,\mathrm{sg}[D_t u_{\mathrm{old}}(z_t, r, t)]$ with $D_t = \partial_t + (\epsilon - x_0)\cdot\nabla_{z_t}$, which turns each average velocity u_j into an instantaneous velocity v_j. The implicit predictors $v_i^{\pm} = v_{\mathrm{old},i} \pm \beta(v_{\theta,i} - v_{\mathrm{old},i})$ reconstruct $\widehat{x}_{0,i}^{\pm} = z_{t,i} - t_i v_i^{\pm}$, and the loss is

$$\mathcal{L}_{\mathrm{NFT}} = \mathbb{E}_i\Big[\tfrac{A}{\beta}\big(p_i \ell_i^{+} + (1-p_i)\ell_i^{-}\big) + \lambda\big\langle (v_{\theta,i} - v_{\mathrm{ref},i})^2\big\rangle\Big], \qquad \beta = 0.1, \quad \lambda = 0.15, \tag{8}$$

where $\ell_i^{\pm}$ is the mean squared error of $\widehat{x}_{0,i}^{\pm}$ over the crystal's non-padding latent entries, divided by its stop-gradient mean absolute error. $D_t u_{\mathrm{old}}$ is a finite difference of width 0.005, and time pairs are 25% with $r = t$, 25% with $r = 0$ and 50% general.

Reward. In equation 6, with e in eV/atom and σ the logistic function, $M(e) = \sigma((0.10 - e)/0.05)$, $S(e) = \sigma((-0.03 - e)/0.025)$ and $E(e) = 1 - \mathrm{clip}(e, 0, 2)/2$. N uses the `compatible_pbe` configuration of LeMat-Bulk. The gate is $G = \mathrm{clip}(q\,d\,\rho, 0, 1)$, with smooth scores q for charge balance, d for the minimum interatomic distance relative to elemental radii and ρ for atomic and mass density. D is the batch-normalized leave-one-out log-determinant gain of a Gaussian kernel over composition fractions and six geometric features.

Recipe. The policy starts from the unconditional DiT-B of Table 1. Each iteration draws 256 latents and takes two AdamW updates (learning rate 10^{-5}, betas $(0.9, 0.95)$, no weight decay, gradient clipping at 1, microbatches of 128). Rollouts use an exponential moving average of the weights with decay $\min(0.001k, 0.5)$ at step k, and the reference policy is frozen at the start. We keep the checkpoint with the highest mean reward on a fixed 256-draw development set. Decoding uses atom-type temperature 1.5 and a push-apart step to a minimum interatomic distance of 0.55 Å.

Evaluation. The LeMat-GenBench cohort has 1,024 draws at eight NFE, decoded with atom types restricted to the 89 elements that all three potentials support; 928 pass the validity filter. Its relaxation run stopped at a repeatability check on repeated single-point energies after all relaxations had finished (largest discrepancy 4.9×10^{-4} eV/atom, no stability bin changed), so its relaxed counts come from the saved outputs. The cohort is a single sample from a checkpoint selected on development reward, and the reward uses the benchmark's novelty reference.

B.3 COMPUTE ACCOUNTING

Sampling cost. We measure the per-forward cost of our own networks directly. At batch 1,024, one DiT-B forward pass costs 1.876 TFLOPs and one DiT-M forward 0.632. The VAE decoder runs once per sample rather than once per step, and costs 0.530 TFLOPs for the Chemeleon2 decoder or 0.265 for the binary-code variant. For DiT-B with the Chemeleon2 decoder, generating 1,024 structures therefore costs 2.41, 8.04, 15.54 and 30.55 TFLOPs at one, four, eight and sixteen steps. Chemeleon2's 50-step DDIM sampler, run through the identical backbone and decoder, costs 94.33 TFLOPs for the same 1,024 structures.

Other baselines. We have no measured forwards for ADiT, DiffCSP or MatterGen, so we estimate them from their architectures and step counts: roughly 376 TFLOPs per 1,024 structures for ADiT, 0.6 to 2×10^3 for DiffCSP, and 1 to 3×10^4 for MatterGen. Crystalite's de-novo transformer (67.6M parameters) is counted at its padded length of 21 tokens over 299 NFE, $2 \cdot 67.6\mathrm{M} \cdot 21 \cdot 299 \cdot 1{,}024 \approx 870$ TFLOPs.

Wall-clock. At an effective H100 throughput of 50 TFLOP/s, generating 1,000 crystals takes about 0.05 s at NFE 1 and 0.16 s at NFE 4. Hadži Veljković et al. (2026) report 22.4 s and 5.1 s for Crystalite's two implementations and 2,639 s for MatterGen.

Training FLOPs. The training-compute axes use $C \approx 6N_{\mathrm{params}} \cdot \mathrm{tokens}$ (Kaplan et al., 2020). The MeanFlow JVP adds a constant factor of about $1.3\times$ to training cost. We do not apply it; it would shift our curves by a constant and change the shape of none.

Flow-matching CSP baselines. The CSP baselines of Figure 4 are accounted differently. CrystalFlow and FlowMM are fully connected graph networks, which spend most of their arithmetic on

edge messages, so a parameters-times-tokens estimate would undercount them. Instead we count the matmuls of their released CSP networks per structure. Graph sizes come from the atom counts of the same test splits, $\langle N \rangle$=10.3 on MP-20 and 25.0 on MPTS-52. CrystalFlow's graph carries self-edges, so E=$\langle N^2 \rangle$; FlowMM's does not, so E=$\langle N(N-1) \rangle$. This gives 3.54 and 20.6 GFLOPs per CrystalFlow forward on MP-20 and MPTS-52, and 2.27 and 13.8 GFLOPs for FlowMM, around 8 to 20× above the parameter-only figure.

B.4 Scaling-law fits

Sections 3.3 and 3.5 interpret the model-size ladders through the parametric scaling laws of Kaplan et al. (2020) and Hoffmann et al. (2022). For a loss, this form is $L(N, D) = E + A\,N^{-\alpha} + B\,D^{-\beta}$, an irreducible floor E plus reducible power laws in parameters N and training tokens D. Our targets are bounded success rates (mSUN, CSP match@1), so we model instead the reducible *deficit* from an irreducible ceiling,

$$S_\infty - S(N, D) = A\,N^{-\alpha_N} + B\,D^{-\beta_D}, \tag{9}$$

which collapses to equation 7 along a ladder where one factor is held fixed.

Model size (Figure 7, Table 2). Along the MP-20 ladder D is fixed. Every model trains on the same corpus for the same 1,000 epochs, so $C \approx 6ND \propto N$ and the ladder isolates the N-term. For each NFE we fit (S_∞, A, α_N) to the seven sizes by profiling the ceiling and regressing $\log(S_\infty - S)$ on $\log N$ (Figure 7). The exponent is jackknife-stable (leave-one-size-out), and the power law is tight at the full-quality budgets ($R^2 = 0.99$ at NFE 8, 0.98 at NFE 16). We report α_N as the robust quantity and treat S_∞ as a soft asymptote. Inverting equation 7 at fixed D shows that halving the reducible gap needs a factor $2^{1/\alpha_N}$ in N, and hence in C. That is ≈21× at NFE 16 and ≈4×10^2 at NFE 8, the steep diminishing return that motivates growing the data along with the model (see the isoFLOP sweep below).

CSP match@1. The composition-conditioned ladders obey the same law (Table 2). Fitting equation 7 to NFE-16 match@1 gives $\alpha_N = 0.31$ on MP-20 (ceiling 0.80, $R^2 = 0.99$) and $\alpha_N = 0.05$ on MPTS-52 ($R^2 = 0.99$), the smaller exponent reflecting a ladder still far from saturation. On MPTS-52 our largest model reaches 25.0% against Crystalite's 31.7%, and the fit projects parity near $N^\star \approx 2.0$B parameters at sixteen steps. A leave-one-out check (fit pico–base, predict large) errs by < 1pp, so the extrapolation of ≈0.8 orders of magnitude to $N^\star$ stays on the same curve. The unconstrained ceiling is the larger uncertainty. Sweeping it from 0.5 to 1.0 moves the crossover to $N^\star \approx$ 2–5B, the shaded band in Figure 4b. At $N^\star$ MeanFlux's inference cost is $2N^\star \cdot \text{tokens} \cdot 16$ against Crystalite's $2 \cdot 270\text{M} \cdot \text{tokens} \cdot 799$, a 6–14× reduction (≈ 14× at the best-fit ceiling). Sampling at lower NFE widens it further. These parity figures extrapolate the fitted power law beyond our largest trained model.

Data axis. The Alexandria-MP-20 continuation (Figure 9) traces the D-term of equation 9 at fixed $N = 130$M against measured training compute $C = 6ND$. Under the OOD protocol, validity and ensemble metastability rise with training FLOPs while novelty decays from one, and mSUN climbs from zero to a noisy plateau near 11% once the model has fit the broader corpus. Because novelty decay caps mSUN before a clean power-law regime is established, we do not fit a single β_D.

isoFLOP sweep (Figure 6). The ladders above change one of N and D at a time. The isoFLOP sweep changes both while holding the training budget fixed. We run it on Alexandria-MP-20 only. Its 540k training structures are twenty times MP-20's, so each budget amounts to tens of passes over the corpus, and the optimum trades parameters against fresh data rather than against repeated data.

Each of the seven ladder sizes is trained once from scratch with the ladder recipe, at a constant learning rate after warm-up, so that any intermediate checkpoint stands as a legitimate model for its budget (Hoffmann et al., 2022). We snapshot and evaluate each run as its training compute $C = 6ND$ (tokens counted as atoms) crosses 10^{15}, 3×10^{15}, 10^{16}, 3×10^{16}, 10^{17} and 3×10^{17} FLOPs. Four sizes fall within each budget up to 10^{17} and three at 3×10^{17}, twenty-three cells in all. Each cell draws 1,024 structures at NFE 8. Stability is scored with single-point MACE energies against the MP hull, and novelty against the Alexandria-MP-20 training split, so absolute mSUN here is not comparable to Table 1 or to LeMat-GenBench.

For each budget we fit a quadratic in $\log_{10} N$ and, when its vertex lies inside the sampled range, take the vertex as N_{opt}. The two smallest budgets are flat: mSUN stays at or below 0.3% at every size.

The four larger budgets have optima at 6.5M, 11M, 31M and 94M parameters, which gives $N_{\text{opt}} \propto C^{0.80\pm0.08}$. The held-out flow-matching error gives $C^{0.77\pm0.01}$ over four budgets; at 3×10^{17} it is still falling at the smallest of the three sizes, so that budget contributes no interior optimum for the error. Since $D_{\text{opt}} = C/6N_{\text{opt}}$, these imply $D_{\text{opt}} \propto C^{0.20}$ and $C^{0.23}$. Both parameters and data should grow with the budget, and parameters should grow faster.

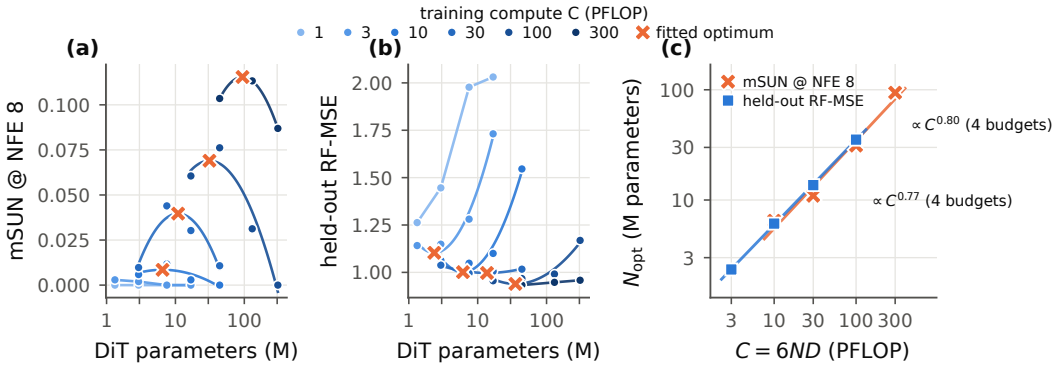


Figure 6: **isoFLOP sweep on Alexandria-MP-20** (unconditional, trained from scratch, n=1024; mSUN at NFE 8 on *unrelaxed* structures with single-point MACE against the MP hull and novelty against the Alexandria-MP-20 training split, so not comparable to Table 1). Each size trains once at a constant learning rate and every budget is read off that run at its crossing epoch; colour is the budget in PFLOP. *(a)* mSUN at NFE 8 and *(b)* held-out flow-matching error against parameters, with a quadratic in $\log N$ where it has an interior optimum (crosses). *(c)* Fitted optima against budget, $N_{\text{opt}} \propto C^{0.80\pm0.08}$ for mSUN (four budgets) and $C^{0.77\pm0.01}$ for the held-out error (four).

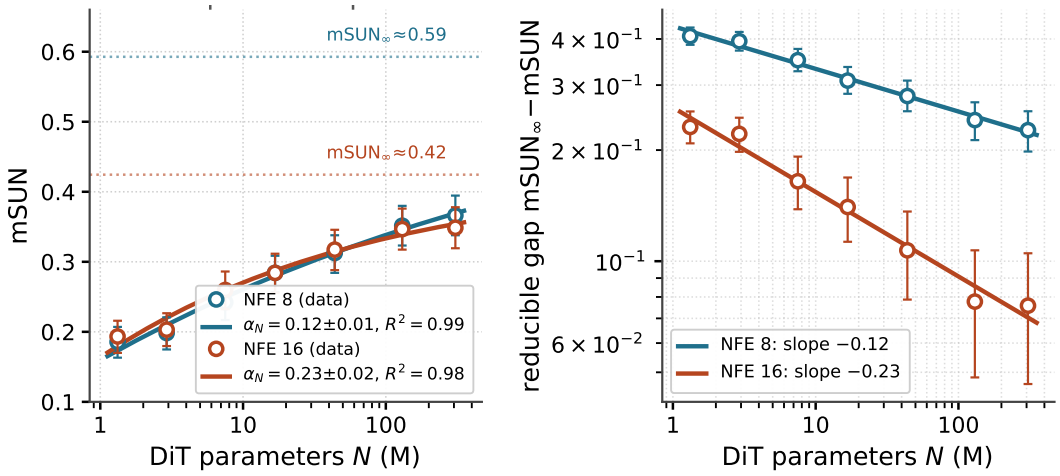


Figure 7: **Model-size scaling law on MP-20** (in-distribution mSUN, seven sizes 1.3M–307M, at NFE 8 and 16). *Left:* the fitted saturating power law of equation 7, $\text{mSUN}(N) = \text{mSUN}_\infty - A\,N^{-\alpha_N}$, through the data (soft ceilings dotted). The top axis shows the actual training compute $C = 6ND$ (PFLOP), which is proportional to N because every model sees the same corpus for the same epochs. *Right:* the reducible gap $\text{mSUN}_\infty - \text{mSUN}$ on log–log axes, a straight line of slope $-\alpha_N$. Error bars are 95% confidence intervals of each measured point, widening on the logarithmic right panel as the gap approaches zero. The fit statistics are presented in Table 2.

B.5 Evaluation protocols

Corpora. MP-20 is the standard benchmark of 45,229 Materials Project structures in total across its train, validation, and test splits, each with at most 20 atoms (Jain et al., 2013; Xie et al., 2022). MP-40 contains cells with up to 40 atoms and is split by database entry date: entries before 4 August

Table 2: Power-law fits $S_\infty - S = A\,N^{-\alpha_N}$ to the in-distribution ladders (seven sizes, 1.3M–307M). α_N uncertainty is jackknife (leave-one-size-out). Last column: R^2 of the log-linear fit $S = a + b\log_{10} N$ over the same points, for reference.

Target	NFE	α_N	S_∞	A	R^2	$R^2_{\text{log-lin}}$
mSUN (MP-20)	8	0.12 ± 0.01	0.59	2.1	0.99	0.98
mSUN (MP-20)	16	0.23 ± 0.02	0.42	6.2	0.98	0.96
match@1 (MP-20)	16	0.31 ± 0.01	0.80	44	0.99	—
match@1 (MPTS-52)	16	0.047	$\sim$1.0	1.9	0.99	—

2018 are used for training and validation, and later entries for testing. Alexandria-MP-20's training split holds 540,162 structures (Schmidt et al., 2024; Zeni et al., 2025) averaging 9.6 atoms, and our Alexandria models warm-start from an MP-20 model (Appendix B.1). Structure prediction adds MPTS-52 (Jiao et al., 2023), which likewise trains on older Materials Project entries and tests on later additions. Its cells are larger ($\langle N\rangle$=25.0 against 10.3 on MP-20) and may contain up to 52 atoms.

In-distribution suite. This draws 1,024 structures at each of one, two, four, eight and sixteen steps, optionally taking atom counts and crystal-system labels from the test split. Candidates are relaxed for 50 MACE steps (BFGSLineSearch with a FrechetCellFilter, $f_{\max}$=0.05) and scored by a single-MACE hull energy against the MP phase diagram, with SUN and mSUN thresholds at 0 and 0.1 eV/atom of the hull and the full MP-20 reference used for uniqueness and novelty. As a population-level composition check, Appendix Figure 8 compares the NFE-4 atom-type distribution with MP20-all and matched baseline outputs: MeanFlux broadly tracks the reference chemistry while sampling several transition-metal and rare-earth species more often. The Chemeleon2 baseline passes through this same suite at 50 DDIM steps, relaxation included, so Table 1 compares like with like. Relaxation does matter here. Scoring the same Chemeleon2 samples without relaxation would understate the baseline by 28pp of mSUN.

LeMat-GenBench. We use the official `comprehensive_multi_mlip_hull` configuration at n=1024, which combines an ORB, MACE and UMA ensemble hull with a three-part validity filter and novelty against the five million structures of LeMat-Bulk, and we report the rates on the structures that pass it, as the leaderboard does.

CSP. Evaluation is performed over the full test splits, 9,046 targets on MP-20 and 8,096 on MPTS-52, with 20 candidates per composition, matched by `StructureMatcher` at the DiffCSP tolerances (ltol=0.3, stol=0.5, angle=10°).

B.6 Comparison with Chemeleon2 RL fine-tuning

Table 1, Figure 2 and the MP-20 group of Figure 1 compare base models only, because RL on an MLIP reward optimizes the quantity these suites measure. The two fine-tuned models, MF-NFT and Chemeleon2's GRPO checkpoint, are reported together in Figure 1 and in the last block of Table 4. On the MP-20 suite of Table 1, MF-NFT reaches 0.372 mSUN and 0.110 SUN at eight NFE, against 0.339 and 0.058 for the unconditional base model. MF-NFT and Chemeleon2's GRPO checkpoint differ in sampler, RL budget and reward, so the comparison between them is not controlled.

B.7 CSP results

B.8 OOD results (LeMat-GenBench)

Table 4 collects the multi-MLIP rows, in which the SUN and mSUN denominators are the structures that pass the validity filter. Under the identical protocol, MP-20-trained MeanFlux models are competitive with the Chemeleon2 DDIM baseline, and six to seventeen times cheaper to sample. As generated, no model in the first two blocks exceeds 0.2% SUN, so the last two blocks are scored after relaxation: first the five MP-20 baselines of Table 1, on the same 1,024 structures per model that table scores, then the reward-finetuned models and their base models of §3.4. The five baselines reach 0.9–3.2% SUN of all draws after relaxation, against at most 0.5% as generated. The condi-

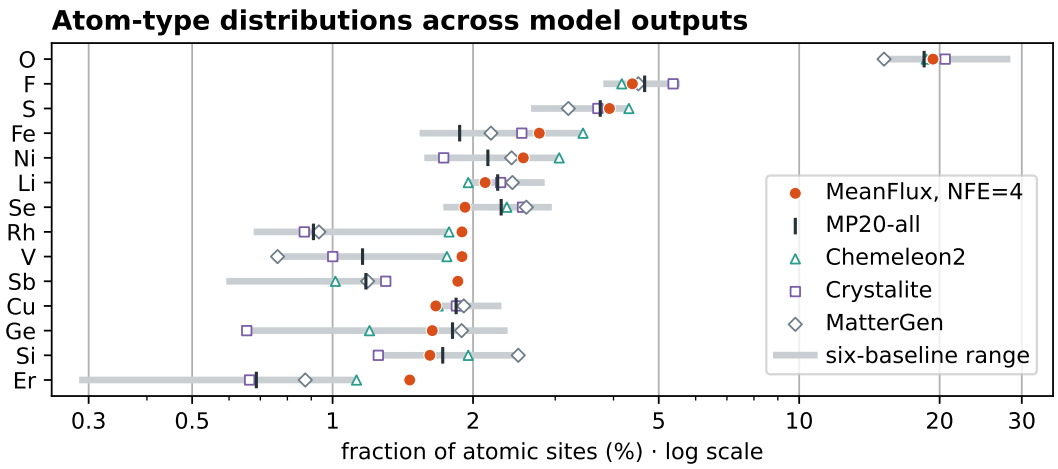


Figure 8: **Atom-type population diagnostic for the NFE-4 unconditional draw.** Site fractions for the 14 most frequent MeanFlux species, compared with all 45,229 local MP-20 structures and matched 1,024-structure outputs from Chemeleon2, Crystalite, and MatterGen. The gray segment spans all six evaluated baseline cohorts, including Chemeleon1, ADiT, and DiffCSP. The logarithmic axis exposes differences among less frequent species without allowing oxygen to dominate the scale.

Table 3: CSP match rate (%) and RMSE under the DiffCSP `StructureMatcher` protocol (Jiao et al., 2023; Ong et al., 2013) (§2.4, Appendix B.5). Baselines as reported by Jiao et al. (2023), Miller et al. (2024), Luo et al. (2025), and Hadži Veljković et al. (2026). Every MeanFlux match rate and RMSE here is measured on the raw generated structures, as the baselines report theirs. Match@20 counts a target as recovered if any of 20 sampled candidates matches, the convention of Jiao et al. (2023) and Luo et al. (2025); CDVAE, FlowMM and Crystalite report single-sample rates only, hence the dashes.

Dataset	Model	NFE ↓	Match@1 ↑	Match@20 ↑	RMSE ↓
MP-20	CDVAE	∼5000	33.9	—	0.105
MP-20	DiffCSP	∼1000	51.5	77.9	0.063
MP-20	FlowMM	∼1000	61.4	—	0.057
MP-20	Crystalite	799	66.1	—	0.033
MP-20	CrystalFlow	100	62.0	78.3	0.071
MP-20	MeanFlux-CSP large	16	**69.1**	83.9	0.094
MP-20	MeanFlux-CSP base	16	63.1	74.3	0.107
MPTS-52	DiffCSP	∼1000	12.2	34.0	0.179
MPTS-52	FlowMM	∼1000	17.5	—	0.173
MPTS-52	Crystalite	799	**31.7**	—	0.072
MPTS-52	CrystalFlow	100	22.7	40.4	0.155
MPTS-52	MeanFlux-CSP large	16	25.0	39.5	0.180
MPTS-52	MeanFlux-CSP base	16	22.6	37.3	0.182

tional MeanFlux model reaches 2.1%, within their range, and MF-NFT reaches 5.7%, the highest SUN in the table.

Table 4: LeMat-GenBench (n=1024). The first two blocks use the official single-point configuration, with percentages on valid structures and $\langle E_{\text{hull}} \rangle$ in eV/atom. The last two blocks rescore after relaxing each valid structure for up to 50 steps under each of ORB, MACE and UMA: the MP-20 baselines of Table 1 on the same structures that table scores, and the reward-finetuned models and their baselines of §3.4, drawn as separate cohorts (Appendix B.2). Their percentages are of all 1,024 draws, SUN and mSUN are disjoint bins, the mSUN rate before relaxation is given in parentheses, novelty and uniqueness are judged on the generated structures, and $\langle E_{\text{hull}} \rangle$ is the mean relaxed hull distance of the valid structures. Bold marks the lowest cost in each block and the highest relaxed SUN. TFLOPs = inference FLOPs per 1,024 structures (DiT forwards + VAE decode), costed as in Table 1 and Appendix B.3. † counts from an analysis of saved relaxation outputs (Appendix B.2). ‡ one structure left unscored because its ORB relaxation failed.

Model	NFE ↓	TFLOPs ↓	Valid ↑	M-Stable ↑	SUN ↑	mSUN ↑	Novel ↑	Unique ↑	$\langle E_{\text{hull}} \rangle$ ↓
Trained on MP-20									
Chemeleon2 DDIM-50	50	94.3	83.3	39.0	0.1	2.0	52.6	99.1	0.48
MeanFlux DiT-B cond.	8	15.5	66.4	32.0	0.1	2.6	41.6	98.8	0.49
MeanFlux DiT-M cond.	8	**5.6**	60.7	38.3	0.0	2.3	44.7	99.0	0.52
MeanFlux DiT-B bcb	8	15.3	65.1	36.4	0.0	2.4	50.8	99.7	0.56
Trained on Alexandria-MP-20									
Chemeleon2 DDIM-50	50	94.3	91.7	53.2	0.2	13.4	54.1	100.0	0.22
MeanFlux DiT-B cond.	8	15.5	78.7	34.7	0.0	6.8	53.2	99.9	0.34
MeanFlux DiT-B bcb	8	**15.3**	72.4	26.2	0.0	7.4	65.5	99.9	0.48
MP-20 baselines of Table 1, relaxed energies (% of all draws; as generated in parentheses)									
ADiT	200	**375.7**	90.8	56.2	3.2	11.9 (1.5)	26.5	89.3	0.05
Crystalite	299	869.8	92.9	43.7	1.9	19.9 (9.5)	55.2	92.4	0.11
Chemeleon1	512	1,000	94.4	42.0	1.1	20.2 (10.4)	62.9	94.1	0.12
DiffCSP	2000	1,200	95.6	38.4	1.4	19.7 (9.0)	66.0	95.3	0.14
MatterGen‡	2000	17,000	96.4	40.2	0.9	22.7 (15.3)	69.3	96.3	0.13
Reward fine-tuning, relaxed energies (% of all draws; as generated in parentheses)									
MeanFlux DiT-B cond.	8	**15.5**	62.4	33.0	2.1	11.2 (1.2)	27.9	62.0	0.09
MF-NFT (uncond.)†	8	**15.5**	90.6	54.3	**5.7**	22.5 (7.7)	58.8	71.5	0.06
Chemeleon2 DDIM-50	50	94.3	83.5	42.9	2.6	19.1 (1.6)	46.0	83.1	0.09
Chemeleon2 + RL	50	94.3	94.7	63.6	4.6	44.0 (20.4)	73.8	91.4	0.07

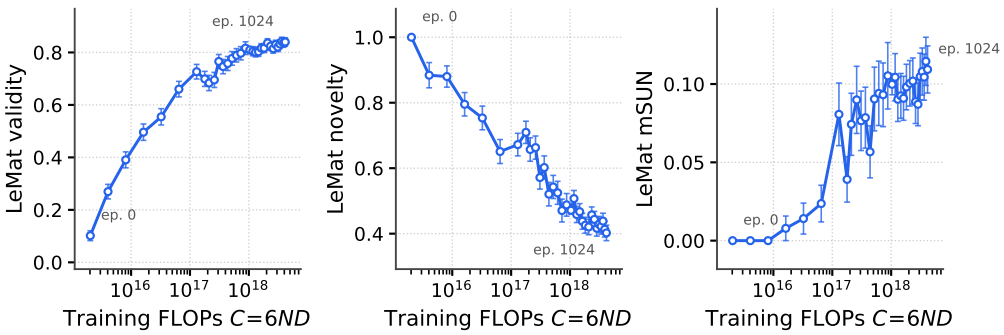


Figure 9: **Alexandria-MP-20 continuation vs. actual training compute.** LeMat-GenBench validity, novelty, and mSUN of the base model (fixed N=130M) against measured cumulative training FLOPs C=$6ND$. This is the D-axis companion to the in-distribution N-ladder of Figure 7 and shares neither its protocol nor its axes. One point per snapshot (NFE 8, push-apart + type-temperature decode; draws of 256–2,500 structures; error bars are ±1 standard error, whose varying width reflects the draw sizes). The cosine schedule warm-restarts at epoch 32 when training was extended (Appendix B.1), and the FLOPs axis remains comparable throughout. Validity and metastability rise with compute while mSUN is corpus-limited, climbing to ∼11% as novelty decays from one.