REINFORCEMENT LEARNING ON THE DISCRETE COMPOSITION CHANNEL OF A CRYSTAL GENERATOR: VALIDATED GAINS AND REWARD HACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Inverse materials design is a long-standing goal of computational materials discovery. Generative models for crystalline materials are typically trained to match the distribution of a structure database, while nothing in their training objective points them at specific design goals such as targeted properties. We use group-relative policy optimization (GRPO) to align a generative model based on stochastic interpolants and discrete flow matching with general black-box reward functions through reinforcement learning. Atom types are generated by a discrete flow and the policy gradient of our generalization of GRPO directly acts on the likelihoods of the atom-type transitions, which differentiates our work from previous reinforcement-learning approaches for diffusion and flow-based generative models of crystalline materials. We introduce a reward function that raises the yield of metastable, unique and novel structures (mSUN) from 13.4% for the pretrained model to 45.5% for the reinforced model, as evaluated by a community benchmark. Our reward also improves the performance of a reinforcement learning framework for crystalline materials based on latent denoising diffusion models. At the same time, we find that directly reinforcing atom-type transition likelihoods enables reward exploitation that has to be prevented with explicit guards. The same analysis also exposes a gap in the community metric. Single-element structures in distinct packings are counted as metastable, unique and novel materials and inflate mSUN without yielding any new compounds. A stability claim is only as good as its reference hull. We report every result split by the number of reference phases behind it and argue that benchmarks should do the same.

1 INTRODUCTION

Discovering crystalline materials with target properties is a central goal of materials science, and machine learning now accelerates that search at every stage (Merchant et al., 2023; Zeni et al., 2025; Gibson et al., 2026; Albavera-Mata et al., 2025). Generative models trained on databases of experimental and computationally relaxed structures (Xie et al., 2022; Jiao et al., 2023; Miller et al., 2024; Zeni et al., 2025; Höllmer et al., 2025) learn the distribution of known materials, so they learn stability only implicitly. The field usually evaluates them by the fraction of generated samples that are stable, unique and novel (Zeni et al., 2025; Betala et al., 2025; Szymanski & Bartel, 2025), but the training objective itself rewards neither stability nor any other design goal. Conditional generation and classifier-free guidance (Ho & Salimans, 2022) can steer a model toward a property (Yang et al., 2024b; Zeni et al., 2025; Tangsongcharoen et al., 2026; Prakash et al., 2026b), but they require property labels for the training structures. Reinforcement learning, in contrast, requires only a reward that can be evaluated on a generated structure, such as an energy from a machine-learned interatomic potential (Batatia et al., 2025; Wood et al., 2025; Prakash et al., 2026a). It has become a common way to reinforce diffusion and flow models toward such rewards (Black et al., 2024; Fan et al., 2023; Liu et al., 2025), with group-relative policy optimization (GRPO) (Shao et al., 2024) as a particularly frequent choice in recent work. Crystal generators have been reinforced this way as well, but in diffusion and flow-based generators the atom type transitions have so far not been treated as separate, discrete actions of the policy. They are either frozen during reinforcement, relaxed to continuous variables, or decoded after the fact from a latent space (Höllmer & Martiniani, 2026; Su

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.

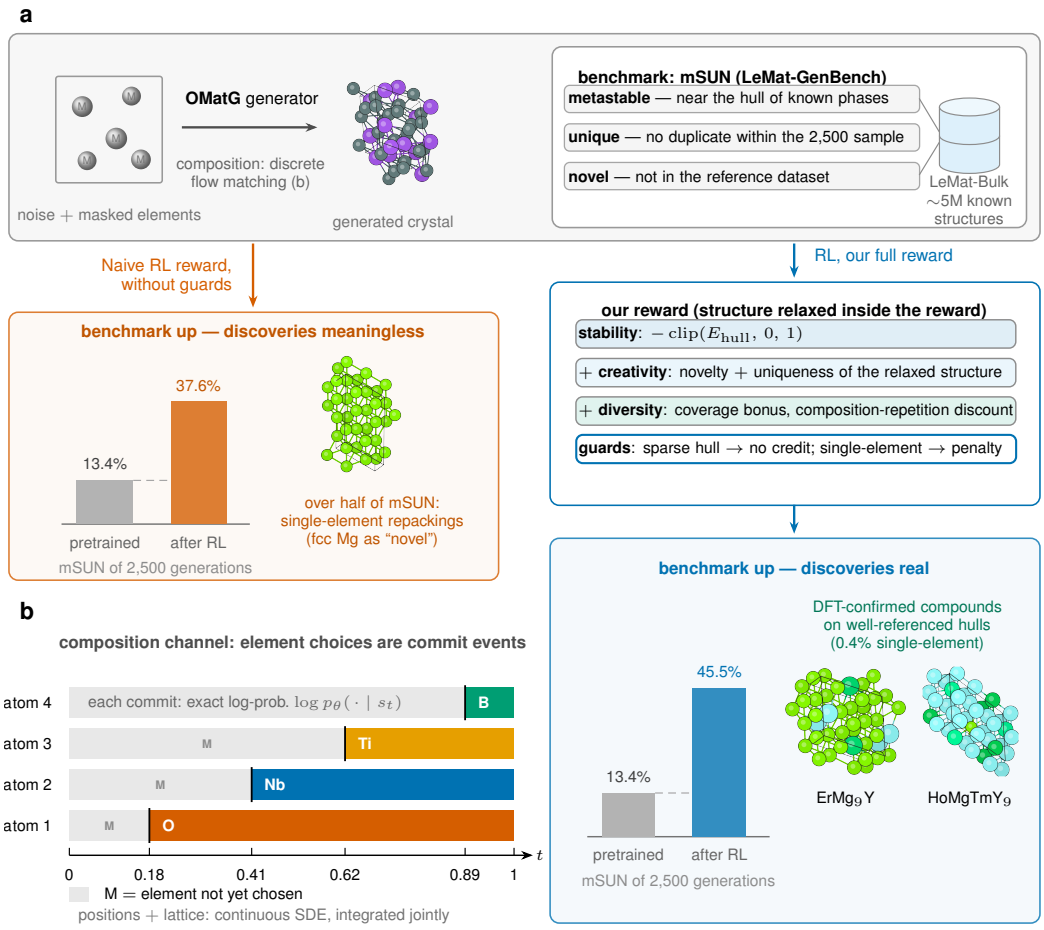

Figure 1: **Reinforcement learning on a crystal generator can satisfy the benchmark without real discoveries, but guards in the reward close the exploit. (a)** The OMatG generator turns noise with masked element identities into crystals and is scored by mSUN under LeMat-GenBench (Section 4.1). Without the single-element guard, RL inflates the pretrained model's 13.4% to 37.6%, but over half of that mSUN set is single-element repackings of known elements, such as face-centered-cubic Mg counted as a novel material. With the full reward (right, Section 3.3), which routes sparse-hull structures to abstention and single-element structures to the penalty reward, mSUN reaches 45.5% with 0.4% single-element structures and DFT-confirmed compounds on well-referenced hulls. **(b)** The composition channel. Under masked discrete flow matching, each atom's element identity is a discrete commit event with an exact log-probability, while positions and lattice evolve as a continuous SDE (Section 3.2).

et al., 2026a; Chen et al., 2025; Park & Walsh, 2026). This leaves a particularly important part of crystal generation, the elemental constituents of a structure that decide its chemistry, only indirectly controlled by the reward. A reward for stable new compounds is largely a reward for choosing favorable combinations of elements, but whether directly reinforcing these compositional decisions improves generation performance remains an open question.

We reinforce OMatG (Höllmer et al., 2025) with GRPO (Shao et al., 2024). OMatG treats the continuous positions and lattice through stochastic interpolants (Albergo et al., 2025) and the discrete atom types by masked discrete flow matching (Campbell et al., 2024). Each atom type is chosen in a single categorical draw at one time step during inference (Fig. 1b), and the only model-dependent factor in that draw is the predicted element distribution. The likelihood ratio has a closed form and the policy gradient reaches the composition channel exactly (Section 3.2). Our reward relaxes every generated structure with the machine-learned potential UMA (Wood et al., 2025) and scores its sta-

bility by the energy above the convex hull of known competing phases, its novelty and uniqueness, and its contribution to the diversity of the batch, with guards for hulls built from only a few known phases and for single-element structures (Section 3.3). Under the community benchmark LeMat-GenBench (Betala et al., 2025), the fraction of generated structures that is metastable, unique and novel (mSUN) rises from 13.4% for the pretrained model to 45.5% after reinforcement (Fig. 1a). In contrast, selecting the best of many samples from the pretrained model reaches 7.7%, and reinforcement learning with the composition generator frozen only reaches 9.8%. This shows that large improvements require reinforcement learning on the composition channel. Density functional theory agrees with the machine-learned potential on 63 of 67 spot-checked structures. We also consider Chemeleon2 (Park & Walsh, 2026) which reinforces a latent diffusion model whose atom types are decoded from the latent by a frozen decoder, so it changes the composition only indirectly. The mSUN metric of Chemeleon2's released model reaches 36.6%, and our reward raises it to 43.7% when it replaces that model's own reward inside its pipeline. Our reward thus appears generally favorable, but the best overall performance is achieved in OMatGRPO with separate, explicitly reinforced atom type transitions.

Each term and guard of our reward was added after the reinforced policy had exploited a previous version, scoring well without producing better materials (Skalse et al., 2022). For instance, an unbounded stability term drove the policy into compositions that the machine-learned potential artificially places far below a convex hull with almost no competing phases. In another exploit, the reinforced policy converged to repack single elements into new unit cells (Fig. 1a). Such a new packing of a known element is not a relevant new material. It still relaxes to an energy near the hull of that element, which holds only a few known competing phases, and it counts as unique and novel because no reference structure matches that packing. Such structures therefore pass every filter of the mSUN metric in LeMat-GenBench, in MatterGen's evaluation code and in Chemeleon2's definition, and they made up 53.5% of the unguarded run's mSUN set (Section 4.3). This also highlights the flaw in the community benchmark itself. Any stability label is only as good as the hull it is measured against, so we report the number of reference phases behind every generated structure and tier our results by it (Section 4.5).

In summary, this work makes three contributions. To our knowledge, it is the first to apply GRPO to the discrete flow-matching composition channel of a crystal generator, using exact per-event likelihoods so that the elements of a structure are actions of the policy in their native discrete form (Section 3.2). It introduces a reward that more than triples the benchmark yield of the generator, survives selection and frozen-composition controls as well as DFT spot-checks, and transfers to another published generator (Sections 3.3 and 4.2–4.4). Finally, it also records how the policy exploited each earlier version of the reward, and shows that single-element structures and hulls with few reference phases can inflate a community benchmark (Sections 4.3 and 4.5).

2 Related work

Most generative models for inorganic crystals learn the distribution of known materials (Jain et al., 2013; Schmidt et al., 2024) by score-based diffusion (Xie et al., 2022; Jiao et al., 2023; Zeni et al., 2025) or flow matching (Lipman et al., 2023; Miller et al., 2024). OMatG (Höllmer et al., 2025) builds on stochastic interpolants (Albergo et al., 2025) for the continuous atom positions and lattice vectors, and masked discrete flow matching for the atom types (Campbell et al., 2024; Gat et al., 2024). During inference, an element is chosen in a discrete jump rather than denoised as a continuous variable. Policy-gradient reinforcement learning, which is widely adopted for image diffusion and flow models (Black et al., 2024; Fan et al., 2023; Liu et al., 2025; Xue et al., 2025; Shao et al., 2024), was enabled in OMatG-IRL (Höllmer & Martiniani, 2026) for the stochastic interpolants of OMatG, which treat the atom positions and lattice vectors. OMatG-IRL, however, kept the species-generation frozen during reinforcement learning. A discrete jump needs its transition probability, and such policy gradients exist for masked diffusion language models (Zekri & Boullé, 2025; Zhao et al., 2025; Ma et al., 2025) and for discrete flow matching over graphs and sequences (Zhu et al., 2026; Su et al., 2026b; Wan et al., 2026).

Reinforcement learning has reached crystals at several levels of representation. The earliest agents acted on the composition, choosing elements for a formula or for the sites of a fixed template (Karpovich et al., 2024; Govindarajan et al., 2025; Hernandez-Garcia et al., 2023). Language models that

write a crystal as text were then reinforced with policy gradients or preference optimization (Cao & Wang, 2026; Hong et al., 2025; Wu et al., 2026; Xu et al., 2025; Yao et al., 2026), which reaches every token including the elements, although some of these works fix the formula in the prompt. The most recent step is reinforcement learning of the diffusion and flow generators that produce complete structures in three dimensions. In these models the composition has not entered the policy gradient as a discrete variable. It is frozen as a condition (Höllmer & Martiniani, 2026; Su et al., 2026a; Subramanian et al., 2026), relaxed to continuous variables (Chen et al., 2025), or decoded after the fact from a latent space (Park & Walsh, 2026). The recent work SPARC (Hsu et al., 2026) applies a reward-weighted denoising loss to the discrete atom types of the symmetry-constrained SymmCD generator (Levy et al., 2025), but without a likelihood ratio, a group baseline or clipping.

Reward hacking, where a policy scores well on a proxy without achieving the intended goal, is a known failure of reinforcement learning (Amodei et al., 2016; Skalse et al., 2022) that grows with optimization pressure (Gao et al., 2023) and is documented for diffusion fine-tuning (Zhang et al., 2024) and KL-regularized RL (Kwa et al., 2024; GX-Chen et al., 2025). In crystal generation the reported failure is collapse onto a few compositions, met with diversity terms (Chen et al., 2025; Park & Walsh, 2026), reweighted uniqueness (Negishi et al., 2026) or multiplicative aggregation of the objectives (Yao et al., 2026). On the evaluation side, the stable, unique and novel (S.U.N.) metric introduced with MatterGen (Zeni et al., 2025) is now standardized by LeMat-GenBench (Betala et al., 2025), and its limits are under study. Enumeration baselines match generative models on it (Szymanski & Bartel, 2025), many metastable generations are duplicates or substitution variants of training structures (Negishi & Walsh, 2026; Martirossyan et al., 2025), and every stability claim inherits the error of the potential behind it (Riebesell et al., 2025).

3 METHODS

3.1 BACKGROUND: STOCHASTIC INTERPOLANTS AND MASKED DISCRETE FLOW MATCHING

OMatG uses the unit-cell representation for crystalline materials where a unit cell with N atoms is described by its atomic numbers $A = (A^1, ..., A^N)$, its fractional coordinates $X = (x^1, ..., x^N)$ and its lattice matrix L (Höllmer et al., 2025), and we write $s = (A, X, L)$ for this state. The continuous variables X and L are modeled through the stochastic interpolants framework. In one possible choice in its flexible implementation, OMatG learns a velocity field $b_\theta(t, s_t)$ and a denoiser $z_\theta(t, s_t)$ to interpolate between samples from a base distribution ρ_0 at time 0 and crystal structures following the data distribution ρ_1 at time 1. During generation, a sample s_0 from the base distribution is transported to a sample s_1 by integrating, jointly for X and L, an SDE with drift $\mu_\theta(s_t) = b_\theta(t, s_t) - \frac{\epsilon(t)}{\gamma(t)} z_\theta(t, s_t)$ and noise amplitude $\sqrt{2\epsilon(t)}$, where $\gamma(t)$ was fixed during training and $\epsilon(t)$ can be chosen at inference (Höllmer et al., 2025; Höllmer & Martiniani, 2026). Euler–Maruyama steps of size Δt turn each step of each variable into a Gaussian transition kernel (Appendix A).

The species, atom types A, are modeled through masked discrete flow matching (Campbell et al., 2024). Each A_t^d takes a value in $\{M, 1, \ldots, S\}$, where M is the mask state and S the number of elements, and OMatG learns the denoising probability $p_\theta(A_1^d = j \mid s_t)$ of the final species of atom d given the current state. During generation all atoms start masked, and the denoising probabilities define the rate matrix $R_t^\star$ of a continuous-time Markov chain that progressively unmasks them, integrated in Euler steps on the same time grid as the SDE, so that every atom carries an element at time 1. A second rate matrix R_t^{DB} can add remasking with a noise parameter η.

3.2 GRPO ON THE NATIVE COMPOSITION CHANNEL

We use GRPO (Shao et al., 2024) as our RL algorithm. A group of K rollouts is drawn from one initial condition, the same noise and masks, and diverges throughout the stochastic generation dynamics. Each rollout i receives a scalar reward r^i that is computed based on the final structure. Our reward function is defined in Section 3.3. The advantage of a rollout is taken relative to its own group,

$$\hat{A}^i = \frac{r^i - \mathrm{mean}_j\, r^j}{\mathrm{std}_j\, r^j}. \tag{1}$$

The policy is the sampler's transition kernel $\pi_\theta(\cdot \mid s_t)$ over one integration step of size Δt. Positions, lattice and species are the three channels of the generator, indexed by f. Within a step the sampler first advances positions and lattice from s_t and then draws the species from the updated positions and lattice, so the probability of a step is a product of one factor per channel. The likelihood ratio between the current and the old policy is written per channel as

$$q_t^f(\theta) \;=\; \frac{\pi_\theta^f(s_{t+\Delta t}^f \mid s_t)}{\pi_{\theta_{\text{old}}}^f(s_{t+\Delta t}^f \mid s_t)}, \tag{2}$$

and the parameters are updated by maximizing $J(\theta) = \sum_f J^f(\theta)$, the sum over channels of the clipped surrogate objective (Schulman et al., 2017)

$$J^f(\theta) \;=\; \frac{\alpha_f}{|\mathcal{T}^f|} \sum_{(i,t)\in\mathcal{T}^f} \min\!\big(q_t^f(\theta)\,\hat{A}^i,\; \text{clip}(q_t^f(\theta),\,1-\epsilon_{\text{clip}},\,1+\epsilon_{\text{clip}})\,\hat{A}^i\big) \;-\; \beta_f\, D_{\text{KL}}^f(\theta), \tag{3}$$

where $\mathcal{T}^f$ collects the update points of channel f across the group. For the positions and lattice, these are the integration steps of every rollout, and for the composition they are the commit events of every atom. Dividing by $|\mathcal{T}^f|$ averages each channel over its own update points, and every update point of a trajectory carries the advantage $\hat{A}^i$ of the structure. The weights α_f compensate for the different scales of the channels, since one position step carries the log-density of all atoms while one commit event carries a single categorical log-probability, and are set so that the gradient magnitudes of the channels are comparable (Appendix F). The term D_{KL}^f is a KL penalty toward the frozen pretrained model with weight β_f, as detailed in Appendix A.

The ratio of Eq. 2 takes a simple form for the composition channel, which we derive next. Each step is a categorical transition that factorizes over atoms, $\pi_\theta^{\text{species}}(A_{t+\Delta t} \mid s_t) = \prod_d \pi_\theta^d(A_{t+\Delta t}^d \mid s_t)$. In discrete flow matching, a masked atom keeps its mask with probability $1 - \Delta t/(1-t)$ and otherwise commits to element j with probability $\frac{\Delta t}{1-t}\, p_\theta(A_1^d = j \mid s_t)$. This prediction is the only model-dependent part of the species channel. The factors $1-\Delta t/(1-t)$ and $\Delta t/(1-t)$ come from the schedule alone. They are the same under the current and the old policy and cancel in the ratio. Let U_t be the set of atoms that commit at step t. The ratio for one step is then

$$q_t^{\text{species}}(\theta) = \prod_{d\in U_t} \frac{p_\theta(A_1^d = A_{t+\Delta t}^d \mid s_t)}{p_{\theta_{\text{old}}}(A_1^d = A_{t+\Delta t}^d \mid s_t)}. \tag{4}$$

Nothing in Eq. 4 is approximated, and the policy gradient reaches the composition channel exactly. Each factor of this product is one update point of Eq. 3. At the last step the schedule factor equals one, so every atom still masked commits to a valid species and enters U_T. We train without species noise, so a committed atom keeps its element. Appendix A shows that the ratio keeps the form of Eq. 4 with species noise, where an atom can return to the mask state at a rate that carries no model dependence, and also gives the KL term of Eq. 3 for this channel.

3.3 Reward stack and guards

This section describes the reward function in its final form (Fig. 1a). Removing components of the reward opens specific exploits, as discussed in Section 4.3. Every final rollout structure is first relaxed inside the reward, with at most 100 FIRE steps including the cell under the UMA potential (Wood et al., 2025), and all terms are computed on the relaxed structure (Appendix F).

The stability term uses the energy above the convex hull, E_{hull}, measured against the UMA-consistent LeMat-Bulk hull released with LeMat-GenBench (Betala et al., 2025). The reward term is $r_{\text{stab}} = -\text{clip}(E_{\text{hull}}, 0, 1)$, the form used by Chemeleon2 (Park & Walsh, 2026), and a structure on or below the hull scores zero. Three rules guard the stability term. A structure whose chemical subsystem has fewer than 12 near-hull references abstains, which zeroes its advantage and removes it from the group baseline. We choose this abstention route because a stability claim against so few references is unreliable rather than wrong. A single-element structure receives the reward floor r_{pen}, because it is always sparsely referenced and would otherwise become an unpunished region that group-relative training drifts into (Section 4.3). A structure with more than 0.1 eV/atom below the hull receives the floor as well, because under the potential that defines the hull such a value marks an unphysical failure of the potential rather than a realistic discovery.

Table 1: Yield of mSUN under LeMat-GenBench. Valid is the share of the nominal 2,500 generations that passes the benchmark's structural checks. Unique, novel and metastable are rates over valid structures, the benchmark's own convention, and mSUN is reported over the nominal 2,500 so that methods are comparable (counts in parentheses). SUN and single-element mSUN are counts. Prior is the pretrained OMatG model. The zero single-element and sparse-hull entries of best-of-N are enforced by its selection score rather than learned, and the controls are defined in Section 4.2.

method	valid % ↑	unique % ↑	novel % ↑	metastable % ↑	SUN count ↑	**mSUN** % (count) ↑	single-element in mSUN ↓	sparse hull % of mSUN ↓
prior	94.1	99.5	82.1	27.2	8	13.4 (336)	7	21.7
best-of-N	97.9	77.9	11.5	65.6	72	7.7 (192)	0	0.0
frozen control	93.5	99.5	87.7	18.7	9	9.8 (245)	1	12.2
Chemeleon2	94.9	90.6	68.1	67.5	36	36.6 (915)	9	7.9
OMatGRPO (ours)	69.8	98.7	88.3	75.8	7	**45.5 (1138)**	4	23.5

The creativity term, adapted from Chemeleon2 (Park & Walsh, 2026), rewards a structure for being unique and novel. A relaxed structure is novel if no structure of the same reduced formula in the MP-20 training set (Xie et al., 2022; Jain et al., 2013) matches it under the pymatgen structure matcher (Ong et al., 2013), and unique if no earlier structure of the same formula in the rollout batch matches it structurally. A structure that is both scores 1, one that is neither scores 0, and the mixed cases receive partial credit defined in Appendix F. Two diversity terms discourage repetition at two scales. Within each rollout group, a composition-occurrence discount $c_{\mathrm{occ}} \in [0, 1]$, adapted from MatInvent (Chen et al., 2025), pulls the reward of a repeated composition toward the floor r_{pen}, because the group would otherwise reward local duplication. Across the entire batch consisting of several rollout groups, a leave-one-out maximum-mean-discrepancy bonus (MMD; Gretton et al., 2012) credits each structure by how much its composition reduces the discrepancy between the batch and the MP-20 composition distribution, because stability-led rewards could otherwise narrow the chemistry of the whole batch in a mode collapse. In summary, the combined reward is

$$r \;=\; r_{\mathrm{pen}} \;+\; c_{\mathrm{occ}} \left[r_{\mathrm{stab}} + r_{\mathrm{creat}} - r_{\mathrm{pen}} \right] \;+\; 0.4\, r_{\mathrm{mmd}}, \tag{5}$$

with $r_{\mathrm{pen}} = -1$. The kernel and per-batch rescaling of the MMD term, the partial-credit rule, and the routing logic are given in Appendix F.

4 Experiments

4.1 Setup

Training follows the GRPO algorithm defined in Section 3.2. Each rollout iteration draws 4 groups of 16 replicates with the training-time sampler (64 steps). All RL training runs start from an OMatG model pretrained on MP-20. Generated structures are scored with the reward function defined by Eq. 5. Each rollout batch is reused for three inner policy updates, and training runs for 750 rollout iterations. Full hyperparameters are in Appendix F.

Every score in Tables 1 and 2 comes from LeMat-GenBench (Betala et al., 2025), run unmodified. mSUN is the fraction of generated structures that are metastable, unique, and novel at once. Stability is scored by a three-MLIP ensemble (Orb, MACE, UMA) with agreement from at least two required and a metastability threshold of 0.1 eV/atom, and uniqueness and novelty from the benchmark's structure matcher against the roughly 5.3 million LeMat-Bulk structures. Each method submits a nominal 2,500 structures, and every mSUN rate in the paper is a fraction of 2,500, so methods are comparable regardless of how many structures survive the benchmark's validity checks.

4.2 RL triples the yield of stable, new compounds

The first question is whether RL improves the yield of stable, new compounds at all. After 750 rollout iterations, mSUN rises from the prior's 13.4% (336 of 2,500) to 45.5% (1,138 of 2,500), a 3.4× gain (Fig. 2a, Table 1). Chemeleon2, the published work we compare against, reaches 36.6% (915 of 2,500) under the identical benchmark run.

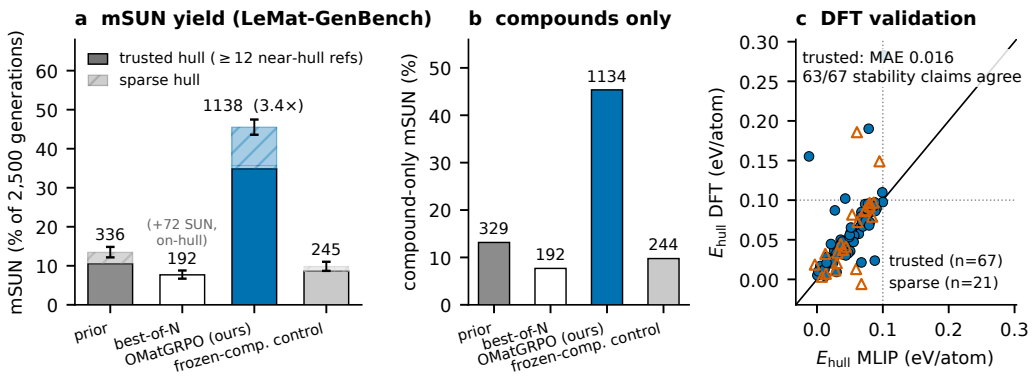

Figure 2: **Reinforcement learning gains, their controls, and DFT validation.** Every method submits 2,500 generated structures, relaxed and scored by LeMat-GenBench (Section 4.1). Counts are printed above the bars. **(a)** Solid segments sit on trusted hulls with at least 12 near-hull references, lighter segments of the same color on sparse hulls (Section 3.3), and error bars are 95% confidence intervals. The controls are defined in Section 4.2. **(b)** Compounds only, which removes the single-element structures of Section 4.3. **(c)** MLIP against DFT stability claims for the spot-checked structures. On trusted hulls (n=67) the mean absolute deviation is 0.016 eV/atom and 63 of 67 claims agree. On sparse hulls (n=21) the energies track but stability cannot be adjudicated. Appendix E lists every structure, the four disagreements and the two excluded cases.

Two control methods, frozen-composition channel and best-of-N, test two things. The frozen-composition control measures the contribution of species RL by turning it off, and best-of-N tests whether RL's gain can be achieved with oversampling. Best-of-N selection, which keeps the top 2,500 of 47,795 prior samples ranked by the stability reward, reaches an mSUN of 7.7%, a sixth of the RL-trained policy's rate. What it surfaces is mostly rediscovery, with a novelty rate of 11.5%. The frozen-composition control reaches 9.8%, below the prior itself. Under the same rollout budget, RL restricted to the position and lattice channels with the stability term alone does not reproduce the gain (Appendix F), and its mSUN set (244 of its 245 structures are compounds) stays close to the prior's composition profile. For a detailed study of RL restricted to the position and lattice channels of OMatG we refer to OMatG-IRL (Höllmer & Martiniani, 2026). This suggests that most of the full OMatGRPO gain in Table 1 comes from the discrete species RL. Fig. 2b restricts the comparison to compounds. The OMatGRPO mSUN set contains 1,134 compounds and 4 single-element structures (0.4% of the set), so it does not suffer from the single-element exploit that Section 4.3 discusses.

Every stability number to this point, in the reward and in the benchmark, comes from MLIPs, and a policy could raise its benchmark score by drifting toward compositions where the potentials err, so we check the metastability claims with DFT. Across the spot-checks accumulated in this work, 63 of 67 trusted-tier metastability claims are confirmed, with a mean absolute deviation of 0.016 eV/atom between UMA and DFT hull distances (Fig. 2c). For the OMatGRPO model alone the count is 20 of 22. Appendix E gives the DFT protocol and the per-structure results.

4.3 Reward hacking

Each term and guard of Eq. 5 was added in response to a failure of the reward that preceded it, where the policy found a way to score well without producing better materials. We describe that sequence here because it shows which parts of the reward are necessary and why, and Appendix C catalogs every variant with its per-run numbers.

Without the clip on E_{hull} in Eq. 5, the reward grows without limit as a structure falls below the hull, and within a group the highest reward goes to the lowest E_{hull}. The policy migrated into a narrow family of Li–O compositions that the MLIP places far below its own convex hull, an artifact of the energy model and of the few competing phases there rather than of true stability. Clipping the reward at the hull removed this incentive, but the clipped run drifted instead onto thinly referenced hulls, where $E_{\text{hull}} \approx 0$ is easy to reach because the reference contains little to compete against

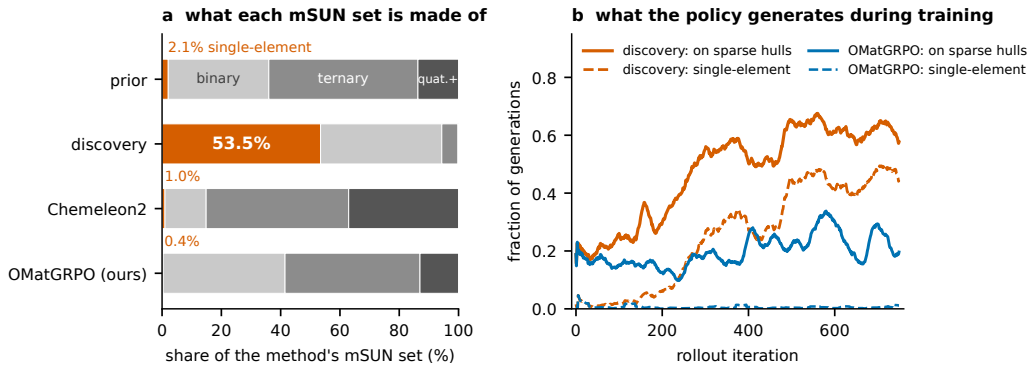

Figure 3: **Reward hacking via single-element structures.** Throughout, *discovery* is the reward of Eq. 5 with the single-element guard removed. **a**, Composition of each method's mSUN set, grouped by the number of distinct elements per structure, with the pretrained prior and Chemeleon2 for reference. **b**, Fraction of generations scored on sparse hulls and fraction that are single-element during training, smoothed over 25 rollout iterations, for the discovery run and OMatGRPO.

(Section 4.5). We responded by withholding reward on sparse hulls, so that a sparsely referenced E_{hull} earns abstention instead of credit.

Abstention created the next exploit, because a structure on a sparse hull is never punished, and the policy drifted into that region. In the ablation with the same reward without the single-element guard (the discovery run of Fig. 3), the fraction of generations routed to sparse hulls climbed from 0.19 to 0.58 over training (Fig. 3b), and the growth was almost entirely single-element structures. A repacked cell of a common metal passes every mSUN filter for the reasons given in Section 1, and fcc Mg in Fig. 1 is the clearest example. Under LeMat-GenBench 53.5% of the discovery run's mSUN set is single-element (Fig. 3a), and MatterGen's evaluation code and Chemeleon2's mSUN definition accept these structures as well.

The exploit is induced by the training. The MP-20 prior already emits a few single-element cells, but a second prior trained on Alexandria, roughly twenty times larger, places none in its mSUN set (0 in 1,000) and 104 per 1,000 after RL (Appendix C). We close the exploit by assigning every single-element generation the penalty reward r_{pen} of Eq. 5. With the guard, single-element structures make up 0.4% of the OMatGRPO mSUN set (Table 1), the guard holds across three seeds (Appendix C), and during training the elemental fraction remains near zero while the sparse-routed fraction holds near 0.2 (Fig. 3b).

Our reward and the benchmark share components, so the gain has to be read with that circularity in mind. Novelty and uniqueness are two of the three components of mSUN and our creativity term rewards both. Without the creativity term the reward reaches 39.3% mSUN against the prior's 13.4% (Appendix B), so about 80% of the gain comes from a reward that optimizes neither. We also note that novelty was rewarded against the MP-20 training set, while the benchmark tests it against roughly five million LeMat-Bulk structures that include essentially all of MP-20, a harder test. Stability is scored by one potential in the reward and by an ensemble of three in the benchmark, of which UMA is one, against hulls built from the same reference data, and DFT is the one check no reward used (Section 4.5).

4.4 Swapping rewards across generators

To separate what the reward contributes from what the training pipeline contributes, we exchanged rewards between OMatGRPO and Chemeleon2 (Park & Walsh, 2026) and trained a policy for each. The results are the off-diagonal cells of Table 2. Each reward was reimplemented inside the other pipeline with its terms and weights unchanged, and every run was evaluated by the protocol of Table 1. OMatGRPO relaxes each structure before computing the reward, while Chemeleon2's reward scores the decoded structure directly.

Table 2: Swapping rewards between the two pipelines. Each cell is a policy trained with the row's reward inside the column's pipeline and evaluated as in Table 1, given as mSUN in % of the nominal 2,500 with counts in parentheses. The diagonal holds OMatGRPO and the released Chemeleon2 checkpoint. Each cell keeps its pipeline's relaxation convention, ours relaxes before the reward and Chemeleon2's does not, and Appendix D reports both variants.

	trained in pipeline	
reward	**ours** (native-channel SDE)	Chemeleon2 (latent VAE)
ours	**45.5** (1138)	43.7 (1093)
Chemeleon2	18.6 (465)	36.6 (915)

Our reward produces the higher mSUN in both pipelines (Table 2). In our pipeline it reaches 45.5% against 18.6% for Chemeleon2's reward, and in Chemeleon2's pipeline it reaches 43.7% against the 36.6% of the released model trained on its native reward. All four cells sit above the prior's 13.4%. Under Chemeleon2's reward our pipeline also leaks single-element structures into its mSUN set, 4.1% with relaxation before scoring and 7.5% without, as no term in their reward accounts for the number of elements. Their own pipeline rarely meets the problem, since its frozen decoder trained on MP-20 seldom decodes a single-element cell (9 of 915 mSUN structures in Table 1), while our generator writes the composition directly. Appendix D interprets the two off-diagonal cells, marked as interpretation, with the supporting diagnostics and the open checks.

4.5 A STABILITY CLAIM IS ONLY AS GOOD AS ITS REFERENCE HULL

The metastability inside mSUN is a relative measure, because E_{hull} is a distance to the convex hull of known competing phases. On a hull with only a few references a value near zero says only that the structure is no worse than the few phases it was compared against, and every benchmark that scores stability against a reference set inherits this limit. In the pretrained prior, 21.7% of the mSUN set sits on hulls with fewer than 12 reference phases, while the discovery run of Section 4.3 has 63.4% of its mSUN set there (Table 1 and Appendix B), because RL moves the generator toward the regions where a metastable label is easiest to obtain. OMatGRPO holds the share at 23.5% (Fig. 2a), close to the prior's level.

The limiting factor is the reference hull rather than the potential. A second machine-learned potential reproduces the stability energies to a median of 0.015 eV/atom, and on the 21 spot-checked structures that fall on sparse hulls DFT tracks UMA as well as on the trusted tier yet cannot adjudicate the claim, because the DFT hull is missing the same phases (Appendix E). Sparse-tier mSUN is therefore reported in every result.

5 CONCLUSION

We reinforced the composition channel of a crystal generator directly. Under masked discrete flow matching every atom type is a commit event with an exact likelihood, so GRPO updates the elements of a structure as actions of the policy. The reward relaxes each structure, scores its distance to the convex hull, its novelty and uniqueness, and the diversity of the batch, and withholds credit where the reference hull is sparse. With it, the yield of metastable, unique and novel structures under LeMat-GenBench rises from 13.4% to 45.5%. DFT confirms 63 of 67 spot-checked stability claims, and the same reward raises Chemeleon2 from 36.6% to 43.7% inside its own pipeline. Each guard in the reward closed an exploit the policy had found, and the last of them, the repacking of single elements into new cells, is counted as discovery by every mSUN implementation we checked. A stability claim is only as good as its reference hull, so we report the number of reference phases behind every structure and read sparse-tier mSUN as candidates rather than discoveries. Benchmarks already compute these counts and should report them, and a promising sparse candidate can be settled by computing its missing competing phases with DFT.

Reproducibility statement

The policy objective and the composition-channel ratio are given in Section 3.2, and Appendix A derives the ratio and the KL terms in the form the implementation evaluates them. The reward is defined in Section 3.3 and Eq. 5, and Appendix F lists its thresholds, the best-of-N selection rule and every training hyperparameter. All benchmark numbers come from LeMat-GenBench run unmodified on 2,500 generations per method (Section 4.1), and Appendix B maps the internal evaluation of Section 4.3 onto that convention. The reward-swap protocol and its recorded deviations are in Appendix D, and the DFT protocol and per-structure results in Appendix E. Code, run configurations and the generated structure sets will be released on publication.

Ethics statement

This work uses only public data (MP-20, Alexandria and the LeMat-Bulk reference set) and involves no human subjects or personal information. The reward targets thermodynamic stability, novelty and diversity of inorganic crystals and no property tied to a harmful use, so we do not foresee risks specific to this work beyond those of computational materials discovery in general. Because the paper shows that a reinforced policy can raise a community benchmark without producing new materials, we report every result split by the number of reference phases behind it, so that readers can judge which stability claims are supported.

LLM usage disclosure

We used Claude (Anthropic) in this work. It helped implement and review the training and analysis code, run experiments and make figures. It searched and summarized the related literature, and helped draft the derivation in Appendix A, which we verified ourselves. It also drafted parts of the text, which the authors rewrote or edited. It was not used to generate data. The authors reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims and artifacts produced with its aid.

References

Angel Albavera-Mata, Pawan Prakash, Jason B. Gibson, Eric Fonseca, Sijin Ren, Xiao-Guang Zhang, Hai-Ping Cheng, Michael Shatruk, S. B. Trickey, and Richard G. Hennig. Discovery of spin-crossover materials with equivariant graph neural networks and relevance-based classification. *Journal of Chemical Theory and Computation*, 21(8):3913–3921, 2025. doi: 10.1021/acs.jctc.4c01690. arXiv:2501.05341.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. arXiv:2303.08797.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

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, 11 2025. ISSN 0021-9606. doi: 10.1063/5.0297006.

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 preprint arXiv:2512.04562*, 2025.

Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch. Structural relaxation made simple. *Physical Review Letters*, 97:170201, 2006.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2305.13301.

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. arXiv:2402.04997.

Zhendong Cao and Lei Wang. Reinforcement fine-tuning for materials design. *Physical Review B*, 113(2):024106, 2026. doi: 10.1103/45zh-44bg. arXiv:2504.02367.

Junwu Chen, Jeff Guo, Edvin Fako, and Philippe Schwaller. Accelerating inverse materials design using generative diffusion models with reinforcement learning. *arXiv preprint arXiv:2511.03112*, 2025.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.16381.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2210.10760.

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. arXiv:2407.15595.

Jason B. Gibson, Ajinkya C. Hire, Pawan Prakash, Philip M. Dee, Benjamin Geisler, Jung Soo Kim, Zhongwei Li, James J. Hamlin, Gregory R. Stewart, P. J. Hirschfeld, and Richard G. Hennig. Developing a complete AI-accelerated workflow for superconductor discovery. *npj Computational Materials*, 12(1):95, 2026. doi: 10.1038/s41524-026-01964-8.

Prashant Govindarajan, Mathieu Reymond, Antoine Clavaud, Mariano Phielipp, Santiago Miret, and Sarath Chandar. CrystalGym: A new benchmark for materials discovery using reinforcement learning. *arXiv preprint arXiv:2509.23156*, 2025. ICLR 2025 AI4Mat workshop.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Anthony GX-Chen, Jatin Prakash, Jeff Guo, Rob Fergus, and Rajesh Ranganath. KL-regularized reinforcement learning is designed to mode collapse. *arXiv preprint arXiv:2510.20817*, 2025.

Alex Hernandez-Garcia, Alexandre Duval, Alexandra Volokhova, Yoshua Bengio, Divya Sharma, Pierre Luc Carrier, Yasmine Benabed, Michał Koziarski, Victor Schmidt, Gian-Marco Rignanese, Pierre-Paul De Breuck, and Paulette Clancy. Crystal-GFN: sampling crystals with desirable properties and constraints. *arXiv preprint arXiv:2310.04925*, 2023.

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

Philipp Höllmer and Stefano Martiniani. Open materials generation with inference-time reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.00424.

Philipp Höllmer, Thomas Egg, Maya M. Martirossyan, Eric Fuemmeler, Zeren Shui, Amit Gupta, Pawan Prakash, Adrian Roitberg, Mingjie Liu, George Karypis, Mark Transtrum, Richard G. Hennig, Ellad B. Tadmor, and Stefano Martiniani. Open materials generation with stochastic interpolants. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.02582.

Zhang-Wei Hong, Nofit Segal, Aviv Netanyahu, Rafael Gómez-Bombarelli, and Pulkit Agrawal. Generating stable materials with large language model reasoning and reinforcement learning. In *NeurIPS 2025 Workshop on AI for Science: The Reach and Limits of AI for Scientific Discovery*, 2025. OpenReview id Hkr2OfTjAc.

Ting-Wei Hsu, Arun Bansil, and Qimin Yan. Symmetry- and property-aware crystal generation with reinforcement learning for inverse materials design. *arXiv preprint arXiv:2609.13468*, 2026.

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. doi: 10.1063/1.4812323.

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)*, 2023. arXiv:2309.04475.

Christopher Karpovich, Elton Pan, and Elsa A. Olivetti. Deep reinforcement learning for inverse inorganic materials design. *npj Computational Materials*, 10(1):287, 2024. doi: 10.1038/s41524-024-01474-5. arXiv:2210.11931.

Georg Kresse and Jürgen Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54:11169–11186, 1996.

Thomas Kwa, Drake Thomas, and Adrià Garriga-Alonso. Catastrophic Goodhart: Regularizing RLHF with KL divergence does not mitigate heavy-tailed reward misspecification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2407.14503.

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. arXiv:2502.03638.

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 (ICLR)*, 2023. arXiv:2210.02747.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.05470.

Tianren Ma, Mu Zhang, Yibing Wang, and Qixiang Ye. Consolidating reinforcement learning for multimodal discrete diffusion models. *arXiv preprint arXiv:2510.02880*, 2025.

Maya M. Martirossyan, Thomas Egg, Philipp Höllmer, George Karypis, Mark Transtrum, Adrian Roitberg, Mingjie Liu, Richard G. Hennig, Ellad B. Tadmor, and Stefano Martiniani. All that structure matches does not glitter. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2509.12178.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023. doi: 10.1038/s41586-023-06735-9.

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. arXiv:2406.04713.

Masahiro Negishi and Aron Walsh. Substitution-based analysis of structural novelty for generative models of materials. *arXiv preprint arXiv:2606.23166*, 2026.

Masahiro Negishi, Hyunsoo Park, Kinga Oliwia Mastej, and Aron Walsh. Continuous SUN (stable, unique, and novel) metric for generative modeling of inorganic crystals. *Machine Learning: Science and Technology*, 7(3):035064, 2026. doi: 10.1088/2632-2153/ae7d85. arXiv:2510.12405.

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. doi: 10.1016/j.commatsci.2012.10.028.

Hyunsoo Park and Aron Walsh. Guiding generative models to uncover diverse and novel crystals via reinforcement learning. *Nature Machine Intelligence*, 8(7):1087–1099, 2026. doi: 10.1038/s42256-026-01262-4. arXiv:2511.07158.

John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77:3865–3868, 1996.

Pawan Prakash, Sam Dong, and Richard G. Hennig. Benchmarking of fast and interpretable UF^3 machine learning potentials. *arXiv preprint arXiv:2608.27277*, 2026a.

Pawan Prakash, Jason B. Gibson, Zhongwei Li, Gabriele Di Gianluca, Juan Esquivel, Eric Fuemmeler, Benjamin Geisler, Jung Soo Kim, Adrian Roitberg, Ellad B. Tadmor, Mingjie Liu, Stefano Martiniani, Gregory R. Stewart, James J. Hamlin, Peter J. Hirschfeld, and Richard G. Hennig. Guided diffusion for the discovery of new superconductors. *npj Computational Materials*, 12(1):286, 2026b. doi: 10.1038/s41524-026-02117-7.

Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, and Kristin A. Persson. A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence*, 7(6):836–847, 2025. doi: 10.1038/s42256-025-01055-1. arXiv:2308.14920.

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. doi: 10.1016/j.mtphys.2024.101560.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2209.13085.

Kaixiang Su, Hongfei Xue, and Qiang Zhu. CrystalGRPO: Target-aligned and coverage-preserving reinforcement learning for flow-based crystal structure prediction. *arXiv preprint arXiv:2608.06582*, 2026a.

Maojiang Su, Po-Chung Hsieh, Weimin Wu, Mingcheng Lu, Jiunhau Chen, Jerry Yao-Chieh Hu, and Han Liu. Discrete flow matching policy optimization. *arXiv preprint arXiv:2604.06491*, 2026b.

Akshay Subramanian, Elton Pan, Juno Nam, Maurice Weiler, Shuhui Qu, Cheol Woo Park, Tommi S. Jaakkola, Elsa Olivetti, and Rafael Gómez-Bombarelli. PackFlow: Generative molecular crystal structure prediction via reinforcement learning alignment. *arXiv preprint arXiv:2602.20140*, 2026.

Nathan J. Szymanski and Christopher J. Bartel. Establishing baselines for generative discovery of inorganic crystals. *Materials Horizons*, 12(19):8000–8011, 2025. doi: 10.1039/D5MH00010F. arXiv:2501.02144.

Krit Tangsongcharoen, Teerachote Pakornchote, Chayanon Atthapak, Natthaphon Choomphon-anomakhun, Annop Ektarawong, Björn Alling, Christopher Sutton, Thiti Bovornratanaraks, and Thiparat Chotibut. CrystalGRW: generative modeling of crystal structures with targeted crystallographic properties via geodesic random walks. *Scientific Reports*, 16:27276, 2026. doi: 10.1038/s41598-026-62470-x.

Zhengyan Wan, Yidong Ouyang, Panwen Hu, and Qiang Sun. dFlowGRPO: Rate-aware policy optimization for discrete flow models. *arXiv preprint arXiv:2605.09291*, 2026.

Amanda Wang, Ryan Kingsbury, Matthew McDermott, Matthew Horton, Anubhav Jain, Shyue Ping Ong, Shyam Dwaraknath, and Kristin A. Persson. A framework for quantifying uncertainty in DFT energy corrections. *Scientific Reports*, 11, 2021. doi:10.1038/s41598-021-94550-5; the MP2020 correction scheme.

Daniel Widdowson, Marco M. Mosca, Angeles Pulido, Andrew I. Cooper, and Vitaliy Kurlin. Average minimum distances of periodic point sets – foundational invariants for mapping periodic crystals. *MATCH Communications in Mathematical and in Computer Chemistry*, 87(3):529–559, 2022. doi: 10.46793/match.87-3.529W.

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 Cohen, Abhishek Das, Ammar Rizvi, Sushree Jagriti Sahoo, Zachary W. Ulissi, and C. Lawrence Zitnick. UMA: A family of universal models for atoms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.23971.

Yuyang Wu, Stefano Falletta, Delia McGrath, and Sherry Yang. CrystalReasoner: Reasoning and RL for property-conditioned crystal structure generation. *arXiv preprint arXiv:2605.14344*, 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 (ICLR)*, 2022. arXiv:2110.06197.

Andy Xu, Rohan Desai, Larry Wang, Ethan Ritz, and Gabriel Hope. PLaID++: A preference aligned language model for targeted inorganic materials design. *arXiv preprint arXiv:2509.07150*, 2025.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. DanceGRPO: Unleashing GRPO on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zügner, Tian Xie, Jake Smith, Lixin Sun, Qian Wang, Lingyu Kong, Chang Liu, Hongxia Hao, and Ziheng Lu. MatterSim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024a.

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)*, 2024b. arXiv:2311.09235.

Zhanao Yao, Jingyuan Shu, Boxuan Zhang, Xiaoyu Wu, Rongyan Wang, Tingwei Chen, Linjing Li, Daniel Zeng, Yu-Dong Yao, Xiaolin Zhao, Jiahui Shi, and Jianjun Liu. CRYSTAL: Coordinated multi-objective reinforcement learning for crystal generation. In *ICML 2026 Workshop on AI for Science: AI Scientists, Tools, Co-authors, or Founders?*, 2026.

Oussama Zekri and Nicolas Boullé. Fine-tuning discrete diffusion models with policy gradient methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2502.01384.

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. doi: 10.1038/s41586-025-08628-5.

Ziyi Zhang, Sen Zhang, Yibing Zhan, Yong Luo, Yonggang Wen, and Dacheng Tao. Confronting reward overoptimization for diffusion models: A perspective of inductive and primacy biases. In *International Conference on Machine Learning (ICML)*, pp. 60396–60413, 2024. arXiv:2402.08552.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. d1: Scaling reasoning in diffusion large language models via reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2504.12216.

Baoheng Zhu, Deyu Bo, Delvin Ce Zhang, and Xiao Wang. Graph-GRPO: Training graph flow models with reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2603.10395.

A Derivation of the policy ratio and the KL terms

This appendix derives the composition-channel ratio of Eq. 4 for any species noise $\eta \geq 0$ and gives the KL terms of Eq. 3 in the form the implementation evaluates them. It rests on the continuous-time Markov chain construction of masked discrete flow matching by Campbell et al. (2024), the species model of OMatG (Höllmer et al., 2025). We use the notation of Section 3.2. A structure has N atoms, $A_t^d \in \{M, 1, \ldots, S\}$ is the species of atom d at time t with M the mask state, generation starts fully masked and ends with no atom masked, and the sampler takes T Euler steps of size Δt. The model output $p_\theta(A_1^d = j \mid s_t)$, abbreviated $p_\theta^d(\cdot \mid s_t)$, is the predicted final species of atom d. For the composition channel s_t holds the species at time t and the positions and lattice already advanced to $t + \Delta t$, since the sampler moves the continuous channels first within a step. The reference policy π_{ref} is the frozen pretrained checkpoint with predictions p_{ref}, run with the same η and the same time grid.

A.1 Masked DFM as a continuous-time Markov chain

Masked DFM connects the fully masked state to the data through the conditional interpolant

$$p_{t|1}(A_t \mid A_1) \;=\; \prod_{d=1}^{N} \big[\, t\, \delta_{A_t^d,\, A_1^d} + (1-t)\, \delta_{A_t^d,\, M} \,\big], \tag{6}$$

so at time t each atom independently shows its final species with probability t and the mask otherwise (Campbell et al., 2024, Eq. 6). Generation simulates a Markov chain with rate matrix $R_t^\star + \eta R_t^{\mathrm{DB}}$, which generates this interpolant for every $\eta \geq 0$ (Campbell et al., 2024, Prop. 3.3 and App. F.1), with the unknown final species replaced by the model's prediction. For atom d the only nonzero off-diagonal rates are

$$R_t^d(M, j) \;=\; \frac{1+\eta t}{1-t}\, p_\theta(A_1^d = j \mid s_t) \quad \text{for } j \neq M, \qquad R_t^d(k, M) \;=\; \eta \quad \text{for } k \neq M. \tag{7}$$

The model enters only through the unmask rate. The remask rate is a constant of the sampler, and there is no direct move between two species.

A.2 The Euler step kernel

Over one step every atom moves independently given s_t, so the step kernel factorizes over atoms, $\pi_\theta^{\mathrm{species}}(A_{t+\Delta t} \mid s_t) = \prod_d \pi_\theta^d(A_{t+\Delta t}^d \mid s_t)$. With $p_{\mathrm{unmask}}(t) = \min\{1, \frac{1+\eta t}{1-t}\, \Delta t\}$ and $p_{\mathrm{remask}} =$

$\eta\,\Delta t$ the per-atom kernel is

$$\pi_\theta^d(A_{t+\Delta t}^d \mid s_t) \;=\; \begin{cases} 1 - p_{\text{unmask}}(t) & \text{if } A_t^d = M,\ A_{t+\Delta t}^d = M, \\ p_{\text{unmask}}(t)\, p_\theta(A_1^d = j \mid s_t) & \text{if } A_t^d = M,\ A_{t+\Delta t}^d = j \neq M, \\ p_{\text{remask}} & \text{if } A_t^d = k \neq M,\ A_{t+\Delta t}^d = M, \\ 1 - p_{\text{remask}} & \text{if } A_t^d = k \neq M,\ A_{t+\Delta t}^d = k, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The implementation draws the unmask and remask decisions as Bernoulli variables with these probabilities and the species of an unmasking atom from $p_\theta^d(\cdot \mid s_t)$. At $\eta = 0$ an unmasked atom keeps its species with probability one, the setting of every run in this paper. Writing $I_{\text{um}}^d = \mathbf{1}[A_t^d = M,\ A_{t+\Delta t}^d \neq M]$ for an unmask transition, the log kernel is

$$\log \pi_\theta^d(A_{t+\Delta t}^d \mid s_t) = \log c_t^d \;+\; I_{\text{um}}^d \,\log p_\theta\big(A_1^d = A_{t+\Delta t}^d \mid s_t\big), \tag{9}$$

where $\log c_t^d$ collects the schedule factors of the four cases and depends only on t, η and Δt. The model appears in a single term, and only at atoms that unmask in this step. Transitions outside the four cases have probability zero under π_θ, $\pi_{\theta_{\text{old}}}$ and π_{ref} alike, so every quantity below is evaluated on their common support.

A.3 The policy ratio

The composition-channel ratio of Eq. 2 is the product over atoms of $\pi_\theta^d / \pi_{\theta_{\text{old}}}^d$. By Eq. 9 the factors c_t^d cancel and every atom that does not unmask contributes a factor of one. With U_t the set of atoms that commit at step t,

$$q_t^{\text{species}}(\theta) \;=\; \prod_{d \in U_t} \frac{p_\theta\big(A_1^d = A_{t+\Delta t}^d \mid s_t\big)}{p_{\theta_{\text{old}}}\big(A_1^d = A_{t+\Delta t}^d \mid s_t\big)}, \tag{10}$$

which is Eq. 4. Nothing in this step used $\eta = 0$. Remasking adds transitions with θ-independent factors that drop out in the same way. What changes with η is the set of update points. At $\eta = 0$ every atom commits exactly once and a rollout with N atoms contributes N update points to $\mathcal{T}^{\text{species}}$. With $\eta > 0$ a remasked atom commits again later and each commit is a further update point, evaluated at the state of its own step. The surrogate of Eq. 3 clips every factor of Eq. 10 separately. At the last integration step the implementation unmasks every atom that is still masked and draws no remask, so $p_{\text{unmask}} = 1$, the forced commits enter U_t with weight one, and Eq. 10 keeps its form.

A.4 KL term for the composition channel

Both step kernels factorize over atoms, so the KL divergence between the current and the reference kernel is a sum of per-atom divergences. For an unmasked atom the kernel does not depend on θ and the divergence vanishes. A masked atom keeps its mask with the same probability under both policies or unmasks to species j with probability $p_{\text{unmask}}(t)$ times the respective prediction, so the shared factor cancels inside the logarithm and

$$D_{\text{KL}}\big[\pi_\theta^d(\cdot \mid s_t) \,\big\|\, \pi_{\text{ref}}^d(\cdot \mid s_t)\big] \;=\; \mathbf{1}[A_t^d = M]\; p_{\text{unmask}}(t)\; D_{\text{KL}}\big[p_\theta^d(\cdot \mid s_t) \,\big\|\, p_{\text{ref}}^d(\cdot \mid s_t)\big], \tag{11}$$

with the ordinary categorical KL between the two species predictions on the right. This holds for any $\eta \geq 0$. The weight $p_{\text{unmask}}(t)$ grows toward the end of generation and equals one at the last step. Writing $\mathcal{B}$ for the rollouts of a batch and N_i for the atoms of rollout i, the term entering Eq. 3 is this divergence summed over the masked atoms of every step and averaged per atom, per step and over the batch,

$$D_{\text{KL}}^{\text{species}}(\theta) \;=\; \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{1}{T N_i} \sum_t \sum_{d:\, A_t^{i,d} = M} p_{\text{unmask}}(t)\; D_{\text{KL}}\big[p_\theta^d(\cdot \mid s_t^i) \,\big\|\, p_{\text{ref}}^d(\cdot \mid s_t^i)\big]. \tag{12}$$

It is evaluated at every masked atom of every step, whether or not the atom commits, so no Bernoulli draw enters it. The implementation evaluates it for $\eta = 0$, where $p_{\text{unmask}}(t) = \min\{1, \Delta t/(1-t)\}$, with $\beta_{\text{species}} = 0.05$ (Appendix F).

A.5 CONTINUOUS CHANNELS

Positions and lattice follow an Euler–Maruyama discretization of the OMatG SDE sampler with drift $\mu_\theta(s_t)$ and noise schedule $\epsilon(t)$ (Höllmer et al., 2025; Höllmer & Martiniani, 2026). One step of the positions is $x^d_{t+\Delta t} = x^d_t + \mu^d_\theta(s_t)\,\Delta t + \sqrt{2\epsilon(t)\Delta t}\,\xi^d$ with ξ^d standard normal, so the step kernel of a structure's positions is a Gaussian with log-density

$$\log \pi^{\mathrm{pos}}_\theta(X_{t+\Delta t} \mid s_t) \;=\; -\sum_{d=1}^{N} \frac{\left\| x^d_{t+\Delta t} - x^d_t - \mu^d_\theta(s_t)\,\Delta t \right\|^2}{4\,\epsilon(t)\,\Delta t} \;+\; \mathrm{const}, \tag{13}$$

where the displacement is taken to the nearest periodic image and the constant does not depend on θ. The implementation floors the variance $2\epsilon(t)\Delta t$ at 10^{-4}. The ratio $q^{\mathrm{pos}}_t(\theta)$ of Eq. 2 is the ratio of two such densities of the realized step, one per structure and step. The lattice channel has the same form with the nine lattice components in place of the atoms and its own noise schedule. The reference policy shares the noise schedule, so the KL divergence between the two position kernels is a squared difference of drifts, and averaged as in Eq. 12 the term entering Eq. 3 is

$$D^{\mathrm{pos}}_{\mathrm{KL}}(\theta) \;=\; \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \frac{1}{T N_i} \sum_t \sum_{d=1}^{N_i} \frac{\left\| \mu^d_\theta(s^i_t) - \mu^d_{\mathrm{ref}}(s^i_t) \right\|^2 \Delta t}{4\,\epsilon(t)}, \tag{14}$$

with $\beta_{\mathrm{pos}} = 0.01$. No KL term acts on the lattice channel, so $\beta_{\mathrm{cell}} = 0$ and only the clipped surrogate of Eq. 3 acts on it.

B INTERNAL EVALUATION AND CONVENTION MAP

Every number in the main text comes from LeMat-GenBench (Betala et al., 2025). During development we adjudicated runs with a faster evaluation of our own, and Section 4.3 and Appendix C quote counts per $1{,}000$ generations from it. This appendix defines that evaluation, sets it against the benchmark, and lists the benchmark results of every run the paper names.

The internal evaluation. For one checkpoint, $1{,}000$ structures are generated de novo with the training-time sampler, taking only the number of atoms from a fixed set of MP-20 validation inputs, with the noise seeded identically for every checkpoint. Each structure is relaxed with FIRE under UMA, including the cell, to a force tolerance of 0.05 eV/Å within at most 500 steps, and its E_{hull} is measured against the UMA-consistent LeMat-Bulk hull together with the number of reference phases within 1 meV/atom of the hull in its chemical subsystem. A structure is metastable if $E_{\mathrm{hull}} \leq 0.1$ eV/atom, unique if it is the first representative of its structure-match cluster within the run, and novel if no structure with the same reduced formula in the MP-20 training split matches it under the pymatgen structure matcher with default tolerances, the convention of MatterGen (Zeni et al., 2025). mSUN counts the structures that are all three, over all $1{,}000$ generations. For the model pretrained on Alexandria, novelty is measured against the Alex-MP-20 set of Zeni et al. (2025), training and validation splits together, about $600{,}000$ structures.

Convention map. Table 3 sets the two conventions side by side. The benchmark is run unmodified at commit `58e6eae3` with the preset `comprehensive_multi_mlip_hull`. Counts under the two conventions are not interchangeable, and every count the main text quotes from the internal evaluation has a benchmark counterpart in Table 4 or Table 1. The sparse-hull share of a benchmark mSUN set is obtained by joining the benchmark's per-structure mSUN list with the internal reference count of the same structure. The join was checked structure by structure against the benchmark's own per-structure stability values for every run, with no disagreement.

Benchmark results of every named run. Table 4 lists the LeMat-GenBench results of every run the paper names. The pre-creativity runs are the development runs before the creativity term was added and differ from their counterparts only by that term. The penalty-routing runs route every sparsely referenced structure to the floor instead of to abstention, which removes the single-element exploit as well, at a lower yield than the guard.

Table 3: The internal evaluation and LeMat-GenBench side by side.

	internal evaluation	LeMat-GenBench
sample per method	1,000 generations	2,500 generations
rate basis	count per 1,000 generations	fraction of the nominal 2,500 (mSUN) or of the valid structures (other rates)
relaxation before scoring	FIRE under UMA with cell, at most 500 steps	our relaxed geometry, then the benchmark's own 50-step relaxation
stability model	UMA	Orb, MACE and UMA, each against its own hull, two of three must agree
reference hull	UMA-consistent LeMat-Bulk hull	the benchmark's per-potential hulls
metastability threshold	$E_{\mathrm{hull}} \leq 0.1$ eV/atom	$E_{\mathrm{hull}} \leq 0.1$ eV/atom
reference count	phases within 1 meV/atom of the hull in the chemical subsystem	not reported, joined from the internal evaluation
uniqueness and novelty	pymatgen structure matcher, MP-20 training split	the benchmark's structure-matcher fingerprint, LeMat-Bulk (5.3 million structures)

Table 4: LeMat-GenBench results of every run named in the paper, on the nominal 2,500 basis. Single-element and sparse-hull entries are shares of the run's mSUN set, and compounds is the mSUN count without the single-element structures. The identifier is the run name in the released code and data, without the suffix `_r750`. The five runs of Table 1 are repeated for reference.

run	identifier	mSUN	%	single-element	sparse hull	compounds
prior	`R0`	336	13.4	7 (2.1%)	21.7%	329
best-of-N	`bestofn48k_top2500`	192	7.7	0 (0.0%)	0.0%	192
frozen-composition control	`frozen_control`	245	9.8	1 (0.4%)	12.2%	244
Chemeleon2, released model	`chemeleon2_mp20rl`	915	36.6	9 (1.0%)	7.9%	906
discovery, pre-creativity	`canonical`	1063	42.5	528 (49.7%)	65.8%	535
discovery (Section 4.3)	`canonical_creatrelax`	939	37.6	502 (53.5%)	63.4%	437
penalty routing, pre-creativity	`sparseworst`	789	31.6	1 (0.1%)	0.9%	788
penalty routing	`sparseworst_creatrelax`	841	33.6	1 (0.1%)	3.6%	840
guarded, pre-creativity	`arityguard`	982	39.3	20 (2.0%)	14.1%	962
OMatGRPO	`arityguard_creatrelax`	1138	45.5	4 (0.4%)	23.5%	1134

The creativity ablation. The guarded pre-creativity run bounds how much the overlap between the creativity term and the benchmark matters, as discussed in Section 4.3. Without the creativity term the reward reaches 39.3% mSUN against the prior's 13.4%, and adding the term raises it to 45.5%, so about 80% of the gain comes from a reward that optimizes neither novelty nor uniqueness. Of the two, uniqueness is the same measurement on both sides, duplicate suppression within a set of generations, and we do not treat the uniqueness column of Table 1 as independent evidence. Novelty was rewarded against the MP-20 training split and is tested by the benchmark against LeMat-Bulk, which includes essentially all of MP-20.

C Reward-hacking catalog with per-run numbers

Five reward designs produced five distinct exploits before the reward of Eq. 5 held. Table 5 lists them, and the paragraphs below give the numbers behind each row. Modes 1 to 4 arose during development and are documented from training-time logs and targeted closure runs. Mode 5 and the final-recipe runs are adjudicated by the internal evaluation of Appendix B, and counts quoted per 1,000 refer to it. The discovery run is the reward of Eq. 5 without the single-element guard, as in Fig. 3, and the pre-creativity runs are the development runs that predate the creativity term.

Mode 1, displacement. The earliest reward scored how little a structure moves under relaxation. With a loose relaxation the policy produced near-stationary structures with implausible energetics.

Table 5: The reward-hacking taxonomy. Five specification-gaming modes met under displacement, energy and E_{hull} rewards, each closed by one targeted change.

#	reward optimized	what the policy did	why it paid	closure
1	relaxation displacement (RMSD)	near-stationary but energetically implausible structures	displacement carries no energy information	energy-based reward
2	formation energy	collapse onto deeply bound, far-above-hull fluorides	strong bonding is a poor proxy for phase stability when more stable competitors exist	switch objective to E_{hull}
3	unfloored E_{hull}	below-hull depth chasing in sparsely referenced regions	with few references E_{hull} is unbounded below, and depth pays linearly	floor at zero. $-\mathrm{clip}(E_{\mathrm{hull}}, 0, 1)$ removes the payout by construction
4	floored E_{hull} + residual geometry term	sparse above-hull O/F binaries gaming the geometry term	thinly referenced hulls score leniently	route sparse-reference scores away from credit
5	full recipe, abstention routing	distinct single-element packings flooding the mSUN set	single-element subsystems are sparsely charted, and the packings pass the metastability, uniqueness and novelty checks of the three community implementations we checked at source	single-element guard. A single-element composition receives the penalty reward, and no relocation followed

At strict force tolerance the within-group reward spread collapsed to 0.045 Å, noise level against the 1.63 eV/atom spread of the energy-family rewards, and learning stopped.

Mode 2, formation energy. A 350-rollout formation-energy run drove the median formation energy to -2.31 eV/atom while the same checkpoint had the worst hull distance of the runs compared in that study (median E_{hull} of $+1.08$ eV/atom). The policy collapsed onto deeply bound fluorides such as $DyMgF_2$ and Cs_2LiYbF_6. The fluorine-containing fraction of generations rose from 14% to 91% and the Yb-containing fraction from 1% to 73% between rollouts 0 and 350, and the per-checkpoint correlation between formation energy and hull distance stayed weak (Pearson $+0.09$ to $+0.36$).

Mode 3, unfloored E_{hull}. With the objective moved to E_{hull} but not floored, the policy chased below-hull depth where references are thin, drifting toward Li–O compositions whose hull distances sit far below zero only because the reference hull contains little to compete against. Flooring the term at zero makes on-hull and below-hull structures tie at the best value, so fictitious depth earns nothing.

Mode 4, sparse above-hull binaries. Floored and combined with a residual geometry term, the recipe drifted into sparse above-hull O/F binaries, and the fluorine-containing fraction of generations rose from 11% to 59% over 750 rollouts. A KL term anchoring the composition channel to the pretrained model at $\beta = 5 \times 10^{-3}$ did not arrest the drift (held-out fluorine share 12% to 60%). Routing structures with fewer than 12 near-hull references away from credit closed it. The held-out fluorine share fell from 12% to 7%, the sparse O/F binaries were replaced by dense Cu- and Ag-rich oxides, and the median reference count of the checkpoint held near 32 where the ungated run's fell to 3. This is why the guards live in the reward rather than in a trust region.

Mode 5, single-element packings. The abstention form of that routing created the unpunished region described in Section 4.3, and Table 6 and Fig. 4 collect the counts. The pretrained prior places 6 single-element structures per $1{,}000$ generations in the mSUN set, the pre-creativity discovery recipe 253 and the discovery run of the main text 242. The guard brings this to 12 in the pre-creativity run, 2 to 15 across three seeds, and to 2 for OMatGRPO. The mode is induced from zero

on the roughly 20× larger Alexandria prior, which places no single-element structure in its mSUN set on its own and 104 per 1,000 under the pre-creativity recipe, and it grows with group size at matched rollout budget (K=16: 253, K=32: 307). Under the guard the single-element fraction of generations holds at a mean of 0.008, its largest excursion (0.27) is punished and self-corrects, and a pre-registered relocation watch (an alert if the largest sparse-compound element family crossed 15) did not fire. The residual sparse-compound mSUN set (103 structures) is led by alkali-halide and chalcogenide binaries, its largest family (Br–Cs) reaches 11, and its novelty against the Alexandria reference is 0.884 (pre-creativity discovery run: 0.760).

Table 6: Single-element structures in the mSUN set. Upper block, per 1,000 generations under the internal evaluation. Lower block, raw counts in released sample sets of generators that were sampled rather than trained against the metric.

run	single-element in mSUN	note
pretrained prior (MP-20)	6	
discovery recipe, pre-creativity, K=16	253	abstention routing, no single-element guard
discovery recipe, pre-creativity, K=32	307	matched rollout budget
+ single-element guard, pre-creativity	12	2 to 15 across three seeds
discovery run (Eq. 5 without the guard)	242	
OMatGRPO (Eq. 5)	2	
penalty routing, pre-creativity	1 to 2	whole sparse tier penalized
Alexandria prior	0	1 single-element structure per 1,000 generations, outside the mSUN set
Alexandria prior + discovery recipe, pre-creativity	104	induced from zero
MatterGen, released base model	0 of 1,024	passive sampling
Chemeleon2, released MP-20 set	87 of 10,000	latent-space RL, 28 copies of one K_{20} packing
Chemeleon2, released Alexandria set	24 of 10,000	latent-space RL

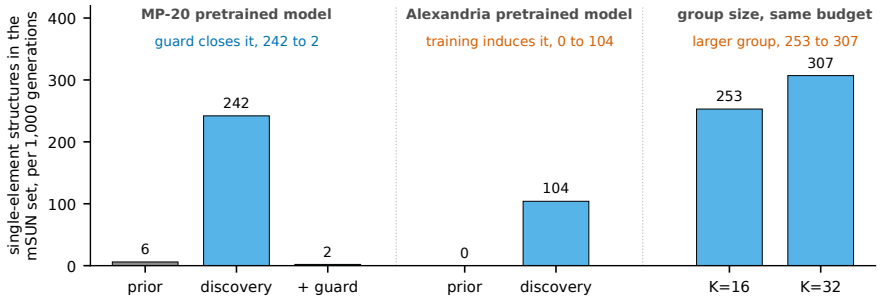

Figure 4: **Where the single-element structures come from.** Single-element structures in the mSUN set per 1,000 generations under the internal evaluation. On the MP-20 prior the guard reduces the count from 242 to 2. The same training induces these structures from zero on a second model pretrained on Alexandria, and on the MP-20 model a larger group size at the same rollout budget produces more of them. The Alexandria and group-size comparisons were run before the creativity term was added to the reward.

Cross-protocol confirmation. The first twenty offending packings, pushed through MatterGen's released evaluation pipeline unmodified, come back 16 of 20 stable, unique and novel. Under LeMat-GenBench, 53.5% of the discovery run's mSUN set is single-element (Fig. 3a) and 49.7% for the pre-creativity recipe (Table 4), and Chemeleon2's mSUN definition accepts the packings as well.

Table 7: Full per-run LeMat-GenBench results. Rate columns are on the valid set (n_{val}), and *valid* itself is on the scored set. mSUN is given as count, rate on valid, and rate on the $2{,}500$ nominal draw.

		rate on valid					mSUN		
run	n_{val}	valid	uniq	novel	meta	SUN	count	/val	/2500
prior (no RL)	2352	0.948	0.995	0.821	0.272	8	336	0.143	0.134
ours $\times$ ours	1745	0.703	0.987	0.883	0.758	7	1138	0.652	0.455
theirs $\times$ theirs	2372	0.949	0.906	0.681	0.675	36	915	0.386	0.366
B (theirs $\times$ ours)	2305	0.928	0.990	0.812	0.321	3	401	0.174	0.160
B$'$ (B + relax)	2322	0.933	0.992	0.817	0.341	6	465	0.200	0.186
A (ours $\times$ theirs)	2268	0.907	0.997	0.843	0.338	6	587	0.259	0.235
A$'$ (A $-$ relax)	2468	0.987	0.943	0.717	0.693	52	1093	0.443	0.437

One collapse that is not gaming. Removing the coverage bonus produces collapse rather than an exploit. Without it the element palette narrows to a Shannon entropy of roughly 1.7 over emitted elements, against at least 3 held across all 750 rollouts with the bonus active, which is what motivates the diversity terms of Eq. 5.

D REWARD-SWAP DETAILS

Chemeleon2's reward in our pipeline (B, B$'$). The four published Chemeleon2 components are imported unchanged and summed. These are the creativity term (weight 1), the energy term (weight 1, single-point MACE-MP energy), the structure-diversity term (weight 0.1) and the composition-diversity term (weight 1), each min–max normalized over the scored batch with the MP-20 reference. Their per-batch standardization is dropped because our GRPO advantage step supplies it, and our own creativity and diversity terms are disabled, so the transplanted reward is their full term set and nothing else. B scores the decoded geometry, as their reward does natively. B$'$ relaxes each batch before scoring (FIRE, cell included, force tolerance 0.05 eV/Å, 100 steps).

Our reward in Chemeleon2's pipeline (A, A$'$). Our full reward of Eq. 5, with the routing of Appendix F, is installed as a single custom component in Chemeleon2's plug-in reward slot. A flag-gated patch ports our abstention semantics into their per-group standardization, and with the flag off their code is byte-identical to the release. A relaxes before scoring. A$'$ removes the relaxation, matching their pipeline's native convention.

Held fixed. Each transplant keeps the host pipeline's recipe unchanged. Our pipeline uses the configuration of Table 12. Their pipeline uses their released MP-20 RL configuration, with clip 0.001, KL weight 1.0, entropy weight 10^{-5}, AdamW at 10^{-5}, up to $5{,}000$ steps and the checkpoint with the best validation reward. A$'$ ran the full horizon without early stopping. Every run generates $2{,}500$ structures, relaxes them, and scores them with the pinned LeMat-GenBench install of Table 1. Every run is single-seed, and A and A$'$ generation used Chemeleon2's sampler, which is unseeded and not bit-reproducible.

Results. Table 7 reports every run and Table 8 the compositional and energetic companions. The relaxation convention flips sign across pipelines. Adding relaxation before scoring raises Chemeleon2's reward in our pipeline from 0.160 to 0.186 (B to B$'$) and lowers our reward in their pipeline from 0.437 to 0.235 (A$'$ to A), which is why Table 2 holds each host pipeline's convention fixed so that a swap changes only the reward. A$'$ is the one run whose mSUN mass shifts to quaternary and quinary chemistry, with 4.8% binary, 21.5% ternary, 43.0% quaternary, 22.6% quinary and 8.1% six or more elements among its $1{,}093$ mSUN structures, and no single-element structures.

Interpretation of the two off-diagonal cells. This paragraph is interpretation, and the tables above are the measurements. Chemeleon2's reward normalizes every term onto $[0, 1]$ per batch although the raw spread of its energy term is roughly $60\times$ that of its composition-diversity term,

Table 8: Compositional and energetic companions per run. Single-element and sparse-hull shares are over each run's mSUN set, and dominant arity is the most common element count and its share. Means over the valid set, energies in eV/atom, relaxation RMSE in Å.

run	single-element in mSUN	sparse in mSUN	dominant arity	$\overline{E_{\mathrm{hull}}}$	$\overline{E_{\mathrm{form}}}$	relax RMSE
prior (no RL)	2.1%	21.7%	3 (50%)	0.192	-0.733	0.441
ours $\times$ ours	0.4%	23.5%	3 (46%)	0.070	-0.055	0.278
theirs $\times$ theirs	1.0%	7.9%	3 (48%)	0.078	-0.804	0.168
B	7.5%	26.7%	3 (44%)	0.178	-0.735	0.456
B′	4.1%	22.2%	3 (45%)	0.161	-0.689	0.448
A	0.0%	15.3%	3 (49%)	0.143	-0.779	0.379
A′	0.0%	2.1%	4 (43%)	0.070	-0.837	0.139

and a replay diagnostic shows that only about 30% of its within-group learnable signal sits in the energy term. Most of its gradient therefore pushes novelty and coverage rather than hull proximity, and a metric gated on stability penalizes that allocation, which is consistent with its weak showing in our pipeline. Our reward keeps the stability and creativity terms on their absolute scales and rescales only the diversity bonus. In the frozen-VAE pipeline the decoder already guarantees geometric sanity, since A′'s outputs move only 0.139 Å under evaluation relaxation against 0.278 Å for OMatGRPO, so a reward's entire steering budget goes to composition, and ours spends it on the stability channel that mSUN gates on. The traits of A′ agree with this account. Its sparse-hull share falls to 2.1%, its mean formation energy is the deepest of any run (-0.837 eV/atom), and it has the highest strict-SUN count in the study (52). The same reward produces none of these traits in our own pipeline, so the effect is an interaction between reward and decoder rather than a property of the reward alone. Replaying the A and A′ checkpoints through the same diagnostic with and without the relaxation step would test the account, and a seed replicate of A′ would bound the sensitivity of its value.

Did our pipeline hack the transplanted reward? We audited the B and B′ runs for reward rising while the quality it proxies stagnates. Table 9 compares the reward's composition under the untrained and the trained policy. Total reward climbed about 17%, mostly through the composition-diversity term rather than the energy term that mSUN gates on, and metastability moved from 0.272 to 0.321, in proportion to the small energy-term movement. Creativity sat at its ceiling from the first step (uniqueness 1.000, novelty 0.984 at batch scale), so it contributed no within-group variance and steered nothing, and advantage statistics stayed well conditioned throughout. The transplanted reward was therefore optimized without an exploit, but weakly, and with one mispricing. B's mSUN set is 7.5% single-element structures, the highest share of any run, because Chemeleon2's reward carries no term for the number of elements. Its frozen decoder rarely decodes single-element cells (1.0% of its own mSUN set), while our generator writes the composition directly and can reach them. Relaxing before scoring roughly halves the leak (B′, 4.1%). This is the counterpart, in the other direction, of the failure mode of Section 4.3, where a reward does not specify what its new host is free to vary.

Table 9: Reward composition of the B run under the untrained and trained policy (post-normalization means).

term	untrained policy	trained policy	Δ
creativity	1.000	0.994	≈ 0
composition diversity	0.473	0.648	$+0.175$
energy	0.043	0.125	$+0.082$
structure diversity	0.035	0.044	$+0.009$
total	1.551	1.811	$+0.260$

E DFT PROTOCOL AND SPOT-CHECK TABLES

Protocol. Structures were drawn from the relaxed generation sets of the internal evaluation (Appendix B) in three batches of 28, 32 and 30, with the sampling frame written down before any calculation ran. The batches cover the pre-creativity discovery run, anchors from the prior, the two sparse oxygen–fluorine binaries of the mode 4 run of Appendix C, the guarded pre-creativity run, probes of fluorine- and oxygen-bearing compositions, sparse-hull compounds, five single-element packings and, in the third batch, OMatGRPO with its trusted tier stratified by the number of elements and its sparse tier sampled by family and at random. Within each stratum the structures were drawn at random with a fixed seed, except for the sparse-family anchors of the third batch, which were chosen by hand to probe the largest sparse families. Each calculation starts from the UMA-relaxed geometry and runs in VASP 6.4.1 (Kresse & Furthmüller, 1996) with the projector augmented-wave method, a relaxation of positions and cell with pymatgen's `MPRelaxSet` (PBE (Perdew et al., 1996), with the Materials Project Hubbard U values where they apply) followed by a static calculation with `MPStaticSet`. The DFT energy above the hull is measured against the DFT split of the LeMat-Bulk hull released with LeMat-GenBench, with the reference count defined as in the reward. The Materials Project 2020 corrections (Wang et al., 2021) are applied to neither side. Applying them to our entries alone shifts fluorine- and oxygen-bearing structures by 0.3 to 0.5 eV/atom and leaves intermetallics untouched, and the correction status of the LeMat-Bulk energies cannot be verified from the released files, so leaving both sides uncorrected is the only choice that cannot double-correct. The UMA comparator is the reward's E_{hull} on the geometry the DFT calculation starts from, and a structure counts as trusted or sparse by the UMA reference count with the threshold of 12 of Section 3.3. Two calculations were excluded, both on Gd-bearing structures. $B_2Gd_2Ru_2$, a prior anchor, did not converge electronically in the static calculation, and $Gd_4Lu_1Sm_1$, a sparse OMatGRPO structure, diverged from the second ionic step with a final energy of $+1.7 \times 10^4$ eV although the relaxation log reported convergence. The Materials Project settings map Gd to the f-in-valence pseudopotential while every other lanthanide in the set uses the f-in-core variant. The two rows enter no statistic and were not repeated with another pseudopotential, which would break the protocol identity across batches.

Summary. Table 10 gives the agreement between UMA and DFT hull distances by tier and by run. On the trusted tier the mean absolute deviation is 0.016 eV/atom with a bias of $+0.010$ eV/atom, DFT being the higher, and 63 of 67 metastability claims agree at the 0.1 eV/atom threshold. On the sparse tier the energies track as well (0.020 eV/atom, 19 of 21), but the DFT hull is built from the same reference data and is missing the same competing phases, so this agreement validates the energies and not the stability of those compositions. Four trusted-tier claims flip. Mg_2Yb_4 (guarded pre-creativity run) moves from -0.012 eV/atom under UMA to $+0.155$ eV/atom under DFT, the largest deviation among the four, and Yb is an element whose pseudopotential differs between the databases behind the reference hull. $Cu_2S_4Sn_2$ (prior, 0.099 to 0.110 eV/atom) and $Lu_1Pr_2Sm_5$ (OMatGRPO, 0.043 to 0.102 eV/atom) straddle the threshold, and the latter has only four reference phases on the DFT hull although its subsystem counts as trusted on the UMA hull. $F_3O_2Rb_1$ (OMatGRPO, fluorine and oxygen probe) moves from 0.079 to 0.190 eV/atom. The two sparse-tier flips are $Ge_{14}Yb_6$ (0.095 to 0.149 eV/atom), again a Yb case, and the single-element packing Rb_6 (0.061 to 0.186 eV/atom), whose subsystem has one reference phase. One agreement is nominal. $Co_{10}Ho$ (OMatGRPO) counts as agreeing only because its UMA value of 0.101 eV/atom sits just above the threshold, while DFT gives 0.283 eV/atom, the largest deviation in the set.

Second potential. The 0.015 eV/atom of Section 4.5 is the median absolute difference per atom between UMA and MatterSim (Yang et al., 2024a) energies on the UMA-relaxed structures of the mode 4 run of Appendix C under the internal evaluation (990 of its $1{,}000$ generations), after a composition-linear offset fitted by least squares is removed, so that differences in the elemental reference energies of the two potentials do not enter. It was computed on that development run and not repeated for OMatGRPO.

Per-structure values. Table 11 lists every calculation that entered a statistic. The run codes are P for the prior, D^- and G^- for the pre-creativity discovery and guarded runs, M4 for the mode 4 run and O for OMatGRPO.

Table 10: UMA against DFT hull distances by tier and by run. Agreement counts metastability claims on the same side of 0.1 eV/atom under both methods.

tier	run	n	MAE (eV/atom)	agreement
trusted	prior	4	0.005	3 of 4
trusted	discovery, pre-creativity	28	0.009	28 of 28
trusted	guarded, pre-creativity	13	0.019	12 of 13
trusted	OMatGRPO	22	0.026	20 of 22
trusted	**all**	**67**	**0.016**	**63 of 67**
sparse	discovery, pre-creativity	10	0.033	8 of 10
sparse	mode 4 run of Appendix C	2	0.014	2 of 2
sparse	guarded, pre-creativity	2	0.009	2 of 2
sparse	OMatGRPO	7	0.007	7 of 7
sparse	**all**	**21**	**0.020**	**19 of 21**

Table 11: Per-structure UMA and DFT hull distances (eV/atom) for the 88 spot-checked structures. Tier T is trusted and S is sparse by the UMA hull. Refs is the number of reference phases within 1 meV/atom of the DFT hull in the structure's chemical subsystem. Agree (y or n) marks metastability claims on the same side of 0.1 eV/atom.

run	formula	tier	refs	UMA	DFT	agree
D^-	Ac_3Hg_5	T	9	0.054	0.050	y
D^-	Ac_9In_3	T	6	0.080	0.084	y
D^-	$AgLi_8$	T	11	0.007	0.013	y
D^-	$AgLi_9$	T	11	0.040	0.043	y
D^-	$AgMg_9$	T	7	0.068	0.085	y
D^-	$AgSr_9$	T	12	0.088	0.024	y
D^-	$AgZn_{19}$	T	7	0.070	0.079	y
D^-	Ag_3Li_5	T	11	0.007	0.011	y
D^-	Ag_4Li	T	11	0.001	0.005	y
D^-	Ba_7Na_3	T	4	0.053	0.052	y
D^-	Cd_4Li_2	T	12	0.015	0.028	y
D^-	ClCu	T	5	0.100	0.098	y
D^-	Hg_6Na	T	13	0.032	0.034	y
D^-	La_9Pb_3	T	13	0.078	0.096	y
D^-	$Li_{11}Mg_3$	T	5	0.005	0.017	y
D^-	$Li_{12}Mg_4$	T	5	0.051	0.060	y
D^-	Mg_8Sc_2	T	7	0.031	0.032	y
D^-	Na_3Pb_7	T	8	0.074	0.084	y
D^-	Ni_4Zn	T	8	0.066	0.067	y
D^-	Pt_2Sn_{16}	T	12	0.073	0.095	y
P	BaF_6Si	T	15	0.026	0.025	y
P	$CuSe_2Tm$	T	24	0.064	0.058	y
D^-	Ag_3Li_4	T	11	0.016	0.020	y
D^-	$CdHg_{16}Ho$	T	31	0.056	0.055	y
D^-	$CdLi_9$	T	12	0.045	0.046	y
D^-	Hg_7Li	T	12	0.025	0.021	y
D^-	Li_4Mg	T	5	0.002	0.010	y
D^-	NaO_4	T	8	0.030	0.009	y
D^-	Na_2O_7	T	8	0.044	0.030	y
D^-	$O_{11}Rb$	T	8	0.056	0.047	y
G^-	$AgGa_{15}$	T	4	0.056	0.054	y
G^-	Ag_4Ga_{16}	T	4	0.053	0.061	y
G^-	$Al_3Sm_8Y_2$	T	15	0.089	0.086	y
G^-	AuMnPd	T	27	0.050	0.036	y
G^-	$Cd_5Hg_2Li_2$	T	30	0.045	0.048	y
G^-	$Cu_{12}MnPd$	T	20	0.038	0.032	y
G^-	$K_2Rb_2Sb_2$	T	18	0.021	0.020	y
G^-	Mg_2Yb_4	T	5	-0.012	0.155	n
G^-	Na_3Sb_3	T	6	0.090	0.089	y
G^-	$O_{16}RbSr$	T	12	0.080	0.072	y
G^-	O_7Rb	T	8	0.075	0.078	y
G^-	O_9Rb	T	8	0.067	0.079	y
G^-	Pd_2Zr_2	T	15	0.047	0.033	y
P	As_5BaCu_7	T	21	0.070	0.067	y
P	$Cu_2S_4Sn_2$	T	17	0.099	0.110	n
O	AcLaNdTm	T	6	0.050	0.056	y
O	$AlErPr_9Sm$	T	31	0.052	0.065	y
O	$CdMgTl_{16}$	T	17	0.021	0.045	y
O	$Co_{10}Ho$	T	15	0.101	0.283	y
O	$CsCu_4F_{15}$	T	23	0.079	0.068	y
O	CsO_2Rb	T	17	0.012	0.004	y
O	Cu_3Zn_{17}	T	8	0.036	0.050	y
O	$ErMg_9Y$	T	15	0.035	0.042	y
O	$ErNd_7SmTm$	T	15	0.015	0.021	y
O	Er_2Sc_8SmTm	T	10	0.027	0.087	y
O	F_3O_2Rb	T	17	0.079	0.190	n
O	$GaLa_3Pr$	T	29	0.077	0.088	y
O	$GaTm_4Y$	T	22	0.074	0.073	y
O	HgK_{19}	T	14	0.026	0.037	y
O	$HoMgTmY_9$	T	25	0.049	0.054	y
O	$HoNdPrTb_5$	T	17	0.086	0.092	y
O	Li_5Zn	T	8	0.011	0.013	y
O	$LuPr_2Sm_5$	T	4	0.043	0.102	n
O	$LuSmY_3$	T	6	0.030	0.033	y
O	Na_3Sn_8	T	9	0.069	0.073	y
O	$Nd_{11}Ni_5$	T	11	0.088	0.097	y
O	Nd_8TmY	T	8	0.038	0.048	y
D^-	Cs_9Na_3	S	4	0.053	0.082	y
D^-	Cu_{12}	S	2	0.007	0.003	y
D^-	O_4	S	1	0.028	0.011	y
M4	$F_{11}O_3$	S	6	0.073	0.087	y
M4	$F_{17}O_3$	S	6	0.073	0.086	y
D^-	As_4	S	1	0.030	0.021	y
D^-	F_8	S	1	-0.003	0.019	y
D^-	$Ge_{14}Yb_6$	S	5	0.095	0.149	n
D^-	Ho_2Nd_{12}	S	4	0.013	0.032	y
D^-	Li_3	S	1	0.059	0.013	y
D^-	Rb_6	S	1	0.061	0.186	n
D^-	Sr_{16}	S	2	0.013	0.007	y
G^-	Ca_7Cd_2	S	9	0.031	0.034	y
G^-	$Ce_{13}Rh_2$	S	6	0.080	0.096	y
O	Au_2Cd_{10}	S	8	0.039	0.037	y
O	$BaLuSm_8$	S	5	0.086	0.097	y
O	Ca_3Mg_{11}	S	4	0.080	0.092	y
O	CdK_5	S	7	0.036	0.045	y
O	$K_{11}O_9$	S	7	0.084	0.079	y
O	$K_{12}Pb$	S	6	0.044	0.041	y
O	$LiNd_7Tm$	S	5	0.062	0.072	y

F TRAINING CONFIGURATION AND BEST-OF-N DEFINITION

Model, sampler and optimization. All runs start from the OMatG checkpoint pretrained on MP-20 for de novo generation (Höllmer et al., 2025), which is also the reference policy of the KL terms. Rollouts use a time grid of 64 points on the unit interval, so 63 Euler steps, with species noise $\eta = 0$ and the velocity annealing of the pretrained sampler switched off. Each iteration draws $B = 4$ groups of $K = 16$ structures, each group from one draw of the initial noise and masks, and scores them with Eq. 5. Within a group the rewards are clipped to three standard deviations around the group mean, the mean and standard deviation are taken over the trusted members only, that is over the structures not routed to abstention, and every member is normalized with these two numbers. Members routed to abstention receive advantage zero, and so does every member of a group with fewer than two trusted members or a standard deviation below 10^{-4}. The batch is then used for three inner epochs of Eq. 3 with AdamW, learning rate 10^{-4} and gradient-norm clipping at 0.5, so 750 rollout iterations are 2,250 optimizer steps. Table 12 lists the remaining values, and one run takes about eight hours on one NVIDIA B200 GPU. The channel weights α_f were set from the logged magnitudes of the per-channel surrogate terms in short calibration runs. At equal weights the species term is far smaller than the position term, since it averages single categorical log-probabilities over commit events while the position term averages the log-density of all atoms over integration steps, so the species channel carries weight 1 and the two continuous channels 0.1.

Table 12: Hyperparameters of the OMatGRPO run. The discovery run of Section 4.3 uses the same values with the single-element guard removed.

setting	value
groups per iteration B, group size K	4, 16
time grid, species noise η	64 points, 0
rollout iterations, inner epochs	750, 3
optimizer, learning rate, gradient-norm clip	AdamW, 10^{-4}, 0.5
clip range ϵ_{clip} of Eq. 3	0.2
channel weights $\alpha_{\text{species}},\ \alpha_{\text{pos}},\ \alpha_{\text{cell}}$	1, 0.1, 0.1
KL weights $\beta_{\text{species}},\ \beta_{\text{pos}},\ \beta_{\text{cell}}$	0.05, 0.01, 0
reward clip within a group	± 3 standard deviations
relaxation inside the reward	FIRE under UMA, cell included, at most 100 steps, force tolerance 0.05 eV/Å
stability term r_{stab}	$-\text{clip}(E_{\text{hull}}, 0, 1)$
sparse-hull abstention	fewer than 12 reference phases within 1 meV/atom of the hull
floor r_{pen}	-1, for single-element structures and for $E_{\text{hull}} < -0.1$ eV/atom
creativity weight, partial credit	1, clamped AMD distance ($k = 100$)
occurrence discount c_{occ}	1 up to 3 occurrences, 0 from 6
MMD bonus	weight 0.4, polynomial kernel $(x \cdot y + 1)^3$, per-batch min–max scaling

Relaxation before scoring. The sampler's raw geometries move by about 1 Å in RMSD when they are relaxed. Scored on the raw geometry, the stability term would measure that distance rather than the energy of the phase, and the novelty term would pay full credit to almost every structure, since at most 0.2% of raw geometries match a known structure against 4 to 13% after relaxation. A term that is nearly constant within a group carries no gradient under group-relative advantages. Every rollout structure is therefore relaxed with at most 100 FIRE (Bitzek et al., 2006) steps, including the cell, under UMA (UMA-s-1p2, task omat) before any term is computed.

Stability term and routing. E_{hull} is measured against the UMA-consistent LeMat-Bulk hull released with LeMat-GenBench, and the reference count of a structure is the number of hull entries within 1 meV/atom of the hull in its chemical subsystem. A structure with fewer than 12 reference phases is routed to abstention. It keeps its reward value, but its advantage is set to zero and it is left out of the group statistics. A single-element structure receives the floor $r_{\text{pen}} = -1$ as

its stability term r_{stab}, and so does a structure more than 0.1 eV/atom below the hull, because the hull is built from the same potential and a value that far below it marks a failure of the potential. The creativity term is evaluated for these structures as for any other, so their reward is at most $r_{\mathrm{pen}} + c_{\mathrm{occ}}\, r_{\mathrm{creat}} + 0.4\, r_{\mathrm{mmd}}$, one unit of r_{stab} below a structure on the hull with the same creativity credit. The floor takes precedence over abstention, since a single-element structure is always sparsely referenced and would otherwise abstain instead of being punished, which is the exploit of Section 4.3. A structure whose relaxation fails or whose cell fails the sanity checks of the reward also receives the floor.

Creativity term. The term follows Chemeleon2 (Park & Walsh, 2026). A relaxed structure is novel if no structure of the same reduced formula in the MP-20 training split matches it under the pymatgen structure matcher with default tolerances, and unique if no earlier structure of the same reduced formula in the rollout batch matches it. A structure that is both scores 1, one that is neither scores 0, and one that is only novel or only unique receives the minimum average-minimum-distance (Widdowson et al., 2022) between its relaxed geometry and the matching structures ($k = 100$), clamped to $[0, 1]$. Chemeleon2 scores novelty against the whole MP-20 set and leaves the partial credit unclamped. A structure whose matcher call exceeds three seconds scores 0.

Diversity terms. The occurrence discount acts within each group. With n the number of structures in the group that share a composition, taken as the multiset of species, $c_{\mathrm{occ}} = 1$ for $n \leq 3$, falls linearly to 0 at $n = 6$, and is applied around the floor as in Eq. 5, so a repeated composition is pulled toward r_{pen} whatever the sign of its reward. The MMD bonus acts across the whole batch of $BK = 64$ structures. Each structure is represented by its fractional composition vector over 119 element slots, the squared maximum mean discrepancy (Gretton et al., 2012) between the batch and $10{,}000$ compositions drawn at random from the MP-20 training split is computed with the polynomial kernel $k(x, y) = (x \cdot y + 1)^3$, and the credit of a structure is the increase of that discrepancy when the structure is left out of the batch. The credits of a batch are rescaled to $[0, 1]$ by their minimum and maximum, following Chemeleon2, and enter Eq. 5 with weight 0.4. This rescaling is the only normalization in the reward.

Frozen-composition control. The control of Section 4.2 trains only the position and lattice channels. The composition of each group is drawn once per rollout iteration from the pretrained model and is shared by the 16 members of the group. The reward is the stability term of Eq. 5 with the same floor, cap, sparse abstention and relaxation step limit, with the relaxation moving the positions only. The creativity, occurrence and MMD terms are omitted, because every member of a group shares one composition, so the occurrence discount would penalize every structure and the MMD bonus has nothing to steer, and the single-element guard is not needed because the compositions are the pretrained model's. The channel weights are $\alpha_{\mathrm{pos}} = \alpha_{\mathrm{cell}} = 1$ with $\beta_{\mathrm{pos}} = 0.01$, and the rollout budget, sampler, optimizer and number of iterations are those of Table 12.

Best-of-N definition. We drew $47{,}795$ structures from the pretrained model with the training-time sampler and scored each with the stability pathway of the training reward, a FIRE relaxation of the positions of at most 100 steps followed by $-\mathrm{clip}(E_{\mathrm{hull}}, 0, 1)$, with structures on sparse hulls, which includes every single-element structure, and structures more than 0.1 eV/atom below the hull routed to the floor. The structures were ranked by this score, ties were broken by the lower raw relaxed E_{hull} and then by sample index, and the top $2{,}500$ were relaxed with the cell included and submitted to LeMat-GenBench like every other method. Because the routing acts at scoring time, the selected set contains no sparse-hull and no single-element structure by construction, which is why those entries of Table 1 are zero. The same ranking cut at $1{,}000$ gives an mSUN rate of 4.4%.