LET CSP BE YOUR ANCHOR: ADAPTIVE CRYSTAL SEARCH OVER FROZEN STRUCTURE PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

De novo crystal generation (DNG) models decide where to search in composition space and how to generate structures with one set of weights. We argue that discovery is better served by separating the two. A crystal structure prediction (CSP) model is a physical prior that should be improved by likelihood training, while rewards, including novelty measured against the search's own history, should act on a search over compositions. We introduce ANCHOR, a GRPO composition policy trained with multi-objective rewards around a frozen CSP model, and continuous adaptive novelty (CAN), a graded novelty score against known structures and a growing discovery history. Using the frozen CSP model as a fixed ruler under one evaluator, we test where adaptation should act. Replacing DNG compositions with ANCHOR's policy on the same CSP backbone raises MSUN from 11.4% to 47.6% and SUN from 1.1% to 22.1% at 99.9% formula uniqueness. Fine-tuning DNG models directly on the same rewards instead moves their composition marginal without raising their on-hull fraction. We show that KL-regularized fine-tuning of a DNG model can only reweight chemistry the pretrained model already supports by a bounded factor, while unregularized DNG fine-tunes move toward known or less stable chemistry. Even a stability-only reward routed into ANCHOR's CSP backbone roughly halves SUN relative to the frozen backbone, whereas likelihood training on structures found during search can improve a CSP backbone. Under MatterGen's evaluation pipeline, ANCHOR raises state-of-the-art MSUN from 29.2% to 41.3%, transfers without retraining to two further CSP backbones, and reaches 47.1% after distillation into Crystalite-CSP. As with any model optimised against a potential, its on-hull rate depends on that potential.

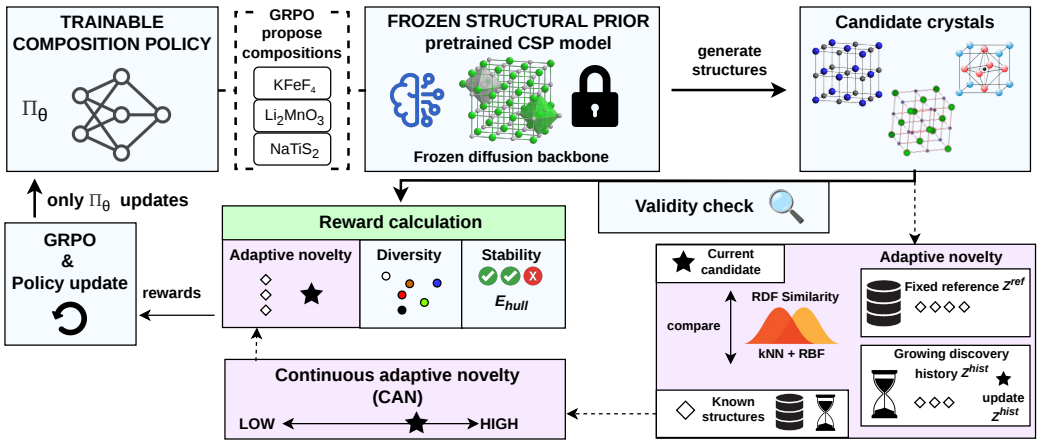

Figure 1: **ANCHOR.** A GRPO-trained composition policy queries a frozen crystal diffusion model to generate candidate structures. Candidates are scored by stability, diversity, and continuous adaptive novelty (CAN), a graded RDF-based score that takes the minimum novelty relative to a fixed reference index and the agent's growing discovery history. Only the composition policy is updated.

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.

1 INTRODUCTION & MOTIVATION

Crystalline materials underpin energy storage, catalysis, and electronics, but only a small fraction of possible compositions and structures are stable, and although many stable materials likely remain unexplored (Merchant et al., 2023), experiments and first-principles calculations are expensive. Most diffusion-based de novo crystal generation (DNG) models (Xie et al., 2021; Jiao et al., 2023; Cornet et al., 2025; Zeni et al., 2025) learn a joint distribution $p_0(c, x) = p_0(c)p_0(x \mid c)$ over crystal structures x with chemical composition c. The conditional distribution $p_0(x \mid c)$ describes the crystal structure prediction (CSP) problem at fixed composition c, while $p_0(c)$ reflects which compositions are probable under the training distribution. The latter is a useful chemical prior, but not the distribution that should be searched, since a CSP model may remain effective for compositions that $p_0(c)$ rarely queries. The marginal answers what chemistry resembles what has been observed, while discovery asks what to investigate next. Reward-guided DNG can move beyond this learned composition marginal (Park & Walsh, 2026; Chen et al., 2025), but does so by adapting the joint generator. A crystal-level reward can therefore alter both chemical exploration and the structural distribution, even when the useful adaptation is primarily compositional. Current models already work around this entanglement approximately, but use strong KL regularization that also restricts chemical movement (Park & Walsh, 2026), or a down-weighted atom-type loss because compositional memorization is hard to control independently of structural quality (Veljković et al., 2026). Novelty sharpens the problem. Conventional novelty is evaluated against a fixed reference collection (Zeni et al., 2025; Betala et al., 2025), so a rediscovered structure earns the same novelty reward the second time it is seen. Search history can discourage rediscovery (Turk et al., 2022; Oschinski et al., 2026; Bellemare et al., 2016), but its value changes while the crystal itself does not, so a structural generator optimized on it is asked to adapt to campaign history rather than to physics.

We therefore give each role its own model and signal. The CSP model is treated as a physical prior, trained on structures and improved with structural signals, and a separate composition policy absorbs every discovery signal, including history-dependent novelty. A frozen CSP model also acts as a fixed ruler, since querying it with any model's compositions measures what that model's composition marginal alone contributes. With this ruler, we find that the composition marginal and not the structural model limits DNG discovery, that reward fine-tuning a DNG model mostly moves its marginal toward compositions where the reward is easiest to collect, collapsing formula uniqueness to 40-50% on one backbone, and that even a stability-only reward routed into a CSP backbone roughly halves its SUN. Likelihood training on the structures the search found instead improves the backbone. Where a reward is applied therefore matters as much as what it measures. Rewards should guide the search over compositions, and likelihood should train the structural prior.

Contributions. We introduce ANCHOR, **A**daptive **N**ovelty for **C**rystals with **H**istory-aware **O**uter-loop **R**einforcement. Our contributions are:

1. **Composition search over a structure prior.** ANCHOR replaces the likelihood-trained composition marginal of DNG with a GRPO policy that queries a frozen CSP model, so discovery rewards change where to search, not how structures are generated.
2. **Continuous adaptive novelty (CAN).** A graded novelty score against known structures and the search's own history, so a crystal stops earning reward once it is found. We show that CAN alone prevents the search from collapsing onto repeated compositions, and that its graded and adaptive components together raise SUN from 8.9% with a binary novelty flag to 22.1%.
3. **Separation of composition and structure effects.** Using a frozen CSP model as a fixed ruler under one evaluator, we show that replacing DNG compositions with ANCHOR's policy raises MSUN more than 3-fold and SUN more than 10-fold on the same backbone. With our reward, fine-tuning through the denoising chain does not raise the on-hull fraction of DNG or CSP models, while distillation via likelihood training on discovered structures raises the SUN of Crystalite-CSP from 20.4% to 26.0% on the same compositions.
4. **State-of-the-art MatterGen results and transferability.** Under MatterGen's pipeline, ANCHOR reaches 41.3% MSUN against 29.2% for the previous best, rising to 47.1% after likelihood distillation into Crystalite-CSP, and 16.0% on-hull SUN against at most 2.5% for any DNG model, and transfers to two further CSP backbones without retraining.

2 BACKGROUND

Crystal generation and evaluation. Crystal structure prediction (CSP) generates structures for a fixed composition, while de novo crystal generation (DNG) generates composition and structure jointly (Jiao et al., 2023; Xie et al., 2021; Cornet et al., 2025). A composition c specifies elements from a vocabulary $\mathcal{E}$ and their stoichiometry, and a crystal structure $x = (L, R, \boldsymbol{t})$ adds the lattice $L \in \mathbb{R}^{3\times3}$, the fractional coordinates $R \in [0, 1)^{T\times3}$ of its T atoms, and their species $\boldsymbol{t} \in \mathcal{E}^T$, consistent with c. This gives a natural decomposition of DNG into composition selection and CSP via $p(c, x) = p(c)p(x \mid c)$, where $p(x \mid c)$ represents the CSP subproblem at fixed composition. This factorization is conceptual, since CSP and DNG models are typically trained separately rather than constructing DNG from a pretrained CSP backbone. Modern benchmarks evaluate validity, stability, novelty, uniqueness, and diversity (Betala et al., 2025), commonly summarized by SUN (Stable, unique, novel) or MSUN (metastable, unique, novel). Further details on prior DNG/CSP models and benchmarks are given in Appendix B.

Novelty and continued discovery. Novelty in crystal generation is conventionally evaluated against a fixed reference collection, while uniqueness measures repetition among generated samples (Zeni et al., 2025; Betala et al., 2025). Continuous structural metrics provide graded alternatives to binary structure matching (Valle & Oganov, 2010; Negishi et al., 2026), but are still evaluated against fixed reference data and so provide no memory of structures found during an ongoing search.

Reward-guided crystal generation. Composition-level RL has been used to search chemical space, with structure prediction performed downstream for selected compositions (Karpovich et al., 2024). More recent work applies rewards directly to crystal generators, such as MatInvent, which uses RL for single- and multi-objective diffusion-based crystal design (Chen et al., 2025). These approaches can move the generated composition distribution away from the likelihood-trained marginal, but they do so by adapting the joint crystal generator. Chemeleon2 provides a novel formulation by treating the reverse-denoising process as an RL policy (Park & Walsh, 2026). Starting from noise z_T, the denoising trajectory is $\tau = z_T \to z_{T-1} \to \cdots \to z_0$, and the decoder D_ϕ maps the final latent state to the generated crystal $x = D_\phi(z_0)$. The RL objective maximizes the expected terminal reward

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[R(x)\right]. \tag{1}$$

Because the reward is observed only for the final decoded crystal, it must be attributed across the preceding denoising decisions, creating a temporal credit-assignment problem.

However, the joint optimization also obscures whether adaptation should occur in composition or structure. Park & Walsh (2026) identify a mismatch between likelihood-based pretraining and their novelty- and uniqueness-based "creativity" objective and use strong KL regularization toward the pretrained denoiser to restrict reward-driven drift.

More generally, reward-driven fine-tuning is known to potentially reduce diversity and sample quality in diffusion models (Uehara et al., 2024). The optimized model still follows a reward-tilted endpoint distribution (Bergmeister et al., 2026), which is particularly relevant for novelty objectives that can favor regions assigned lower probability by the pretrained model. This motivates separating the distribution that decides where to search from the model that generates structures.

3 METHODS

The DNG composition marginal. Under the factorization of Section 2, discovery effort is allocated as the product of a composition distribution and the success rate of the conditional it queries. For any conditional structural model $g(x \mid c)$, we can define

$$s_g(c) = \Pr_{x\sim g(x|c)}\left[S(x, c) = 1\right]$$

as its discovery success rate at composition c, under a binary discovery criterion $S(x, c) \in \{0, 1\}$. Likelihood-trained DNG models collapse this design space into one generative model. However, its two factors can be chosen independently, and ANCHOR takes the one combination that a jointly trained generator cannot reach: a composition policy $\pi_\theta(c)$ trained by reinforcement learning against

the frozen conditional it queries. Successful candidates are then allocated as

$$\begin{aligned} \text{DNG} \qquad & p_0(c)\, s_{\mathrm{DNG}}(c), \\ \text{DNG compositions + CSP} \qquad & p_0(c)\, s_{\mathrm{CSP}}(c), \\ \text{ANCHOR} \qquad & \pi_\theta(c)\, s_{\mathrm{CSP}}(c). \end{aligned} \tag{2}$$

These substitutions give reciprocal controls: holding the composition list fixed isolates the structural generator, while holding the CSP model fixed isolates composition allocation. Replacing $p_0(c)$ with the marginal $q(c)$ of a reward fine-tuned generator gives the same control for any fine-tune, so a frozen CSP model acts as a fixed ruler for what fine-tuning did to the compositions. We expect $s_{\mathrm{CSP}} \geq s_{\mathrm{DNG}}$ because a CSP model is trained only to solve the structure problem at fixed composition whereas $p_{\mathrm{DNG}}(x \mid c)$ must also account for which compositions are probable. Separately, a composition can have high $s_{\mathrm{CSP}}(c)$ and still be rarely reached, because likelihood training optimizes $p_0(c)$ for reproducing the reference distribution rather than for discovery. Appendix C shows that the optimum is $p_0(c)$ reweighted by a per-composition partition function. Because the reward is bounded, so is this reweighting, since $q^*(c)/p_0(c) \leq e^{\Delta R/\beta}$ for a reward range ΔR (equation 22), and KL-regularized fine-tuning can therefore reweight chemistry the pretrained model already proposes, but cannot reach chemistry to which it assigns negligible probability. Removing the KL term lifts this bound, but it also removes the only constraint on the conditional. Reward fine-tuning therefore moves the marginal toward compositions where the reward is easy to collect, which need not be where discovery succeeds, and it also tilts the conditional, which is the coupling ANCHOR removes. Section 4 evaluates all three rows of equation 2 under one evaluator.

Furthermore, reward-guided fine-tuning can move the adapted marginal $q(c)$ away from $p_0(c)$, but a crystal-level reward does not specify whether adaptation should occur in $q(c)$ or in $q(x \mid c)$. This adds a semantic credit assignment problem on top of the unavoidable temporal credit assignment across denoising steps implicit in Eq. 1. A single KL coefficient does not separate them either:

$$D_{\mathrm{KL}}(q(c,x)\|p_0(c,x)) = D_{\mathrm{KL}}(q(c)\|p_0(c)) + \mathbb{E}_{c\sim q(c)} D_{\mathrm{KL}}(q(x \mid c)\|p_0(x \mid c))\,. \tag{3}$$

Raising the joint KL strength to restrict drift in $q(x \mid c)$ therefore also restricts the chemical adaptation needed to move beyond the pretrained marginal. Any reward that varies across structures at fixed composition can therefore change the conditional. This effect is amplified for history-dependent novelty. With $R_t(x) = R_{\mathrm{nov}}(x; Z^{\mathrm{ref}}, Z_t^{\mathrm{hist}})$ at search step t, the KL-regularized optimum is $q_t^*(x \mid c) \propto p_0(x \mid c)\exp(R_t(x)/\beta)$ (Bergmeister et al., 2026), so a structure the agent has already found loses probability although its physical plausibility is unchanged. However, a crystal does not become less physically plausible because it has been discovered. It just becomes less useful to sample again. The straightforward solution is to freeze $p_0(x \mid c)$ while leaving the composition policy free. A novelty- or history-driven reward carries no information about how to improve the structural model. The structures the search finds do, and can serve as likelihood training data for the prior. Appendix C develops these arguments in more detail.

Separating discovery from structure generation. ANCHOR makes the separation architectural by replacing the pretrained marginal $p_0(c)$ with an adaptive policy and freezing the conditional:

$$p_{\mathrm{ANCHOR}}(c,x) = \pi_\theta(c)\, p_{\mathrm{CSP}}(x \mid c). \tag{4}$$

The policy can now concentrate search where the frozen CSP is already productive, including compositions the original DNG marginal rarely reaches. Because $p_{\mathrm{ANCHOR}}(x \mid c) = p_{\mathrm{CSP}}(x \mid c)$ by construction, the structural term in Eq. 3 vanishes identically rather than being suppressed by a penalty that must be tuned. The reward can therefore act on π_θ without any route to the structural model, and can include history-dependent novelty without inducing undesirable reward tilting. We note that freezing the CSP does not freeze structural exploration, since repeated queries at the same composition can still sample different structures from $p_{\mathrm{CSP}}(x \mid c)$. Freezing is a property of the search rather than of the prior, which can be improved between searches by likelihood training on the structures they find (Section 4).

Discovery as composition-level RL. Each rollout samples a composition action

$$c = \big(k,\ \boldsymbol{e},\ T,\ \rho\big) \sim \pi_\theta(\cdot \mid \boldsymbol{s}), \tag{5}$$

which fixes the number of distinct elements k, their identities $\boldsymbol{e}$, the cell size T and the stoichiometry ρ, and hence the reduced formula $f(c)$ (Appendix D.1). The frozen CSP model (Cornet et al., 2025)

samples a structure for c, which is relaxed over atomic positions for at most 100 BFGS steps under MatterSim-1M (Yang et al., 2024) before it is scored (Figure 1). We optimize π_θ with GRPO, which standardizes each reward within a group of G actions sampled under the same state (Appendix D.6).

Continuous Adaptive Novelty. Fixed-reference novelty, whether binary or graded (Negishi et al., 2026), keeps rewarding a structure after it has been found. CAN instead scores each valid relaxed structure against both known structures and the search's own history. We describe a structure x by an ℓ_1-normalized radial distribution function (RDF) histogram $\boldsymbol{z}(x)$ and compare it only with structures of the same reduced formula f, which separates structural novelty within a formula family from the chemical exploration done by the policy. Against a set Z of such embeddings, novelty is the k-nearest-neighbour RBF distance

$$\nu(\boldsymbol{z}; Z) \;=\; \mathrm{clip}\left(1 - \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} \exp\Big(-\frac{\|\boldsymbol{z}-\boldsymbol{z}_j\|_2^2}{2\sigma^2}\Big),\; 0,\; 1\right), \qquad \mathcal{J} = \mathrm{kNN}_{k_{nn}}(\boldsymbol{z}, Z). \tag{6}$$

The agent keeps a growing history Z^{hist} of the RDFs it has produced alongside the fixed MP-20 reference Z^{ref}, and CAN takes the lower of the two scores,

$$\nu(x) = \min\big(\nu(\boldsymbol{z}; Z^{\mathrm{ref}}),\; \nu(\boldsymbol{z}; Z^{\mathrm{hist}})\big), \tag{7}$$

where an index with no structure of formula f is omitted and a formula found in neither receives the bootstrap score ν_0. After scoring, $\boldsymbol{z}(x)$ joins Z^{hist}, so regenerating a structure the policy has already found drives its novelty to zero and sustained reward requires continued discovery (Appendix D.4).

Reward and multi-objective GRPO. Each candidate first passes a validity gate. Valid structures are scored by stability $s_{\mathrm{stab}} = 1 - \mathrm{clip}(E_{\mathrm{hull}}, 0, 1)$, CAN ν, and structural and compositional diversity within the group, collected as $\boldsymbol{m}(c) = (s_{\mathrm{stab}}, \nu, d_{\mathrm{struct}}, d_{\mathrm{comp}}) \in [0,1]^4$ (Appendix D.3, Appendix D.4 and Appendix D.5). With a count bonus that decays as a reduced formula is revisited (Thiede et al., 2022), the reward is

$$r(c) \;=\; \begin{cases} \boldsymbol{m}(c)^\top \boldsymbol{w} \;+\; \dfrac{\lambda_c}{\sqrt{\max(n_f, 1)}}, & c \text{ valid}, \\ r_{\mathrm{inv}}, & \text{otherwise}, \end{cases} \tag{8}$$

where n_f counts valid generations of f in the run, so the count bonus prices repeated compositions while CAN prices repeated structures. Because these terms compete and a fixed weighting would need tuning, we sample $\boldsymbol{w} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$ and an element sub-group $\mathcal{S} \subseteq \mathcal{E}$ for each rollout, hold both fixed across the group, and supply them to the policy through $\boldsymbol{s}$ (equation 41). The policy is thereby conditioned on the trade-off it is asked to make, the group-relative advantage compares actions under one scalarization and one element set, and masking element draws to $\mathcal{S}$ spreads exploration across the periodic table. At evaluation, $\boldsymbol{w}$ is fixed to the Dirichlet mean (Appendix D.6).

4 Experiments

Setup and evaluation. The main ANCHOR model is trained with a frozen KLDM-CSP backbone, $G = 32$, $r_{\mathrm{inv}} = -0.2$ and MatterSim-1M as the training potential (Appendix F.1). Training uses the positions-only relaxation of Section 3 because relaxation dominates the cost of the reward (Appendix E.1.5), and evaluation is configured for fidelity instead. Every internal number is produced after training by sampling $N_{\mathrm{eval}} = 1000$ post-training generations across three seeds. Structures are relaxed with MatterSim-5M under BFGS for at most 400 steps with positions and lattice both free, and scored against one hull. Appendix E.1.4 repeats every internal table with the lattice fixed, with the same conclusions. We report **Valid** (the validity gate, see Appendix D.2), **Unique** (distinct reduced formulas), **Novel**, and **SUN / MSUN** (valid $\wedge$ unique $\wedge$ novel $\wedge$ $E_{\mathrm{hull}} \le 0$ / ≤ 0.1 eV/atom). We keep these names for the MatterGen and LeMat-GenBench pipelines, and write Stable and Meta for the stability rates at $E_{\mathrm{hull}} \le 0$ and ≤ 0.1 eV/atom without the uniqueness and novelty conditions. We selected the main configuration from eight runs (Appendix D.8, Appendix F.1 and Appendix F.2). Table 1 ablates the exploration mechanisms, Table 2 varies where adaptation acts under our internal evaluator, and Table 3 scores the same rows with MatterGen's pipeline next to published DNG results, with LeMat-GenBench in Appendix E.3. In Table 2, paired

Table 1: Exploration-mechanism ablation ($N_{\text{eval}} = 1000$, MatterSim-5M, BFGS-400, positions and lattice free; Appendix E.1). Rows give the full method, each mechanism removed (–), each mechanism alone, and the full method with CAN replaced by the binary novelty flag of Chemeleon2 (Appendix F.3). Removing adaptive novelty scores novelty against the frozen MP-20 index alone. All columns are percentages except *MP-20*, which counts valid samples whose reduced formula occurs in MP-20. Best in MSUN and SUN is bold, second-best underlined. Runs were matched on wall clock. Entries are mean ± std over three sampling seeds.

Configuration	Valid	Unique	Novel	Meta	Stable	MSUN	SUN	MP-20
Full method	$97.5_{\pm 0.4}$	$99.9_{\pm 0.1}$	$97.3_{\pm 0.5}$	$47.7_{\pm 2.0}$	$22.2_{\pm 1.2}$	$\underline{47.6}_{\pm 2.1}$	$\mathbf{22.1}_{\pm 1.2}$	$0_{\pm 0}$
– count bonus	$96.4_{\pm 0.2}$	$100.0_{\pm 0.0}$	$96.4_{\pm 0.3}$	$32.8_{\pm 1.5}$	$12.6_{\pm 0.4}$	$32.8_{\pm 1.5}$	$12.6_{\pm 0.4}$	$0_{\pm 0}$
– sub-group mask	$96.6_{\pm 0.3}$	$99.9_{\pm 0.1}$	$96.5_{\pm 0.1}$	$43.5_{\pm 2.5}$	$11.6_{\pm 0.8}$	$43.4_{\pm 2.4}$	$11.6_{\pm 0.9}$	$0_{\pm 0}$
– adaptive nov.	$97.0_{\pm 0.1}$	$99.8_{\pm 0.1}$	$96.9_{\pm 0.0}$	$39.7_{\pm 2.1}$	$14.7_{\pm 2.0}$	$39.7_{\pm 2.1}$	$14.7_{\pm 2.0}$	$0_{\pm 0}$
adaptive nov. only	$97.7_{\pm 0.6}$	$100.0_{\pm 0.0}$	$97.6_{\pm 0.6}$	$52.0_{\pm 0.4}$	$16.3_{\pm 2.0}$	$\mathbf{52.0}_{\pm 0.5}$	$\underline{16.3}_{\pm 2.0}$	$0_{\pm 0}$
count bonus only	$96.0_{\pm 0.6}$	$100.0_{\pm 0.0}$	$96.0_{\pm 0.6}$	$41.3_{\pm 1.4}$	$12.4_{\pm 0.8}$	$41.3_{\pm 1.4}$	$12.4_{\pm 0.8}$	$0_{\pm 0}$
sub-group mask only	$99.7_{\pm 0.2}$	$6.2_{\pm 0.1}$	$99.6_{\pm 0.2}$	$66.8_{\pm 0.6}$	$21.6_{\pm 0.8}$	$3.4_{\pm 0.2}$	$1.0_{\pm 0.1}$	$358_{\pm 1}$
none	$99.9_{\pm 0.1}$	$0.7_{\pm 0.5}$	$99.9_{\pm 0.1}$	$96.3_{\pm 0.8}$	$69.0_{\pm 0.4}$	$0.4_{\pm 0.3}$	$0.3_{\pm 0.4}$	$997_{\pm 1}$
binary novelty flag	$97.2_{\pm 0.0}$	$100.0_{\pm 0.1}$	$97.2_{\pm 0.1}$	$25.1_{\pm 2.3}$	$8.9_{\pm 0.8}$	$25.0_{\pm 2.3}$	$8.9_{\pm 0.8}$	$0_{\pm 0}$

comparisons use matched compositions wherever the structural generator is being isolated: the → CSP rows hold the preceding model's compositions fixed while changing only the structure generator, and the likelihood-tuned rows query each CSP model with the same ANCHOR compositions.

Adaptive novelty sustains discovery. Table 1 ablates the three mechanisms that penalize sampling the same compositions: adaptive novelty, count bonus, and element sub-group mask, as a full factorial. With none of them enabled, the policy collapses to 0.7% unique compositions, so although 96.3% of structures are metastable, MSUN is 0.4%. Adaptive novelty alone restores 100% uniqueness and reaches the highest MSUN (52.0%, against 41.3% for the count bonus alone), while sub-group masking alone does not prevent collapse. The full method has the highest SUN (22.1% against 16.3% for adaptive novelty alone) at 4.4 points lower MSUN. Replacing the growing history with the frozen MP-20 index lowers SUN to 14.7% and MSUN to 39.7%, and replacing the graded score with binary novelty lowers them to 8.9% and 25.0%, so both components of the adaptive score matter. The graded score matters because a binary flag is constant on formulas outside MP-20, so it cancels in the group-relative advantage once the policy leaves the reference set (Appendix F.3). Over a full training run, the policy proposes $97{,}599$ distinct reduced formulas among $101{,}237$ valid structures, and its formula collisions with MP-20 fall from 346 in the first fifth of the run to 3 in the last.

The composition marginal limits DNG discovery. Replacing only the DNG generator with the corresponding CSP model at fixed DNG compositions changes MSUN and SUN by at most 1.0 point under either pipeline, whereas replacing the composition source with ANCHOR's policy at a fixed CSP backbone raises MSUN by 32-36 points and SUN by a factor of 11-20. Uniformly sampled charge-neutral formulas reach only 8.8% MSUN, against 1.5% for random formulas, so the policy learns more than charge neutrality. Under MatterGen, ANCHOR reaches 41.3% MSUN against 29.2% for the previous state-of-the-art and 18.5% for the strongest published KLDM variant, and 14.4-16.7% on-hull SUN, while every row sampling DNG compositions, including the fine-tuned ones, stays at or below 2.5%. Substituting Crystalite-CSP or OMatG-CSP under the KLDM-trained policy preserves the same high discovery rate, so the learned search transfers across structural priors without retraining. The pipelines differ in potential, optimiser, and novelty definition (Appendix E.2), but rank ANCHOR against the DNG models identically.

Reward fine-tuning mostly moves compositions. We fine-tune both DNG models on the same reward as ANCHOR using DDPO (Black et al., 2024), the standard policy gradient method for diffusion models. We do not use KL regularization so that the composition marginal is not constrained by the bounded reweighting in equation 22 (Appendix F.5). This doubles the validity and Meta of Crystalite-DNG, but does not improve the metastable fraction of valid structures and lowers the on-hull fraction, so validity drives the apparent MSUN gain. The fine-tune passes the gate by moving toward known chemistry: MP-20 formula overlap rises sharply, uniqueness falls to roughly

Table 2: **Separating composition and structure effects.** One validity gate, one relaxer (MatterSim-5M, BFGS-400, positions and lattice free), one reference hull, and one fixed MP-20 novelty index for every row; $N_{\text{eval}} = 1000$. All columns are percentages except *MP-20*, which counts valid samples whose reduced formula occurs in MP-20. → CSP queries the corresponding frozen CSP model with the preceding model's compositions; *FT* is DDPO reward fine-tuning at Dirichlet-mean (*mean*) or equal stability and novelty (*bal.*) objective weights. *Untrained composition sources* are defined in Appendix F.4. In the *likelihood-tuned CSP* block, each backbone is fine-tuned by likelihood on MP-20 alone (step-matched control) or with the structures found by ANCHOR added (+ ANCHOR data), and queried with the ANCHOR compositions. Entries are mean$_{\pm\text{std}}$ over three sampling seeds. Best in MSUN and SUN is bold, second-best underlined. Row definitions in Appendix E.1.1.

Configuration	Valid	Unique	Novel	Meta	Stable	MSUN	SUN	MP-20
Untrained composition sources (frozen KLDM-CSP)								
Random formulas	$9.8_{\pm 0.4}$	$100.0_{\pm 0.0}$	$8.3_{\pm 0.6}$	$2.4_{\pm 0.5}$	$0.4_{\pm 0.1}$	$1.5_{\pm 0.4}$	$0.2_{\pm 0.0}$	$17.5_{\pm 3.5}$
Enumerated charge-neutral	$68.9_{\pm 1.6}$	$100.0_{\pm 0.0}$	$68.2_{\pm 1.3}$	$9.4_{\pm 0.3}$	$1.2_{\pm 0.1}$	$8.8_{\pm 0.0}$	$1.2_{\pm 0.1}$	$9.5_{\pm 4.9}$
DNG and reward fine-tuned DNG (ours)								
Crystalite-DNG	$47.0_{\pm 2.7}$	$98.8_{\pm 0.1}$	$31.9_{\pm 1.8}$	$26.2_{\pm 2.1}$	$2.8_{\pm 0.3}$	$14.0_{\pm 0.7}$	$1.5_{\pm 0.4}$	$165.7_{\pm 17.2}$
→ CSP	$48.9_{\pm 2.6}$	$98.8_{\pm 0.1}$	$33.3_{\pm 1.4}$	$28.0_{\pm 3.0}$	$2.9_{\pm 0.6}$	$15.0_{\pm 1.0}$	$1.7_{\pm 0.3}$	$167.3_{\pm 16.6}$
+ FT (mean)	$96.7_{\pm 0.4}$	$40.5_{\pm 0.5}$	$50.4_{\pm 1.1}$	$54.0_{\pm 0.7}$	$1.9_{\pm 0.5}$	$20.2_{\pm 2.6}$	$0.9_{\pm 0.5}$	$631.5_{\pm 20.5}$
+ FT (bal.)	$97.0_{\pm 0.8}$	$49.5_{\pm 0.7}$	$61.1_{\pm 1.0}$	$51.4_{\pm 0.4}$	$2.6_{\pm 0.9}$	$22.9_{\pm 0.8}$	$1.5_{\pm 0.7}$	$612.0_{\pm 25.5}$
+ FT (mean) → CSP	$97.2_{\pm 0.6}$	$40.5_{\pm 0.5}$	$50.0_{\pm 2.0}$	$47.0_{\pm 1.6}$	$0.9_{\pm 0.1}$	$18.2_{\pm 2.3}$	$0.7_{\pm 0.2}$	$638.0_{\pm 17.0}$
+ FT (bal.) → CSP	$97.3_{\pm 0.5}$	$49.5_{\pm 0.7}$	$60.1_{\pm 1.8}$	$47.9_{\pm 0.7}$	$1.6_{\pm 0.5}$	$21.0_{\pm 2.0}$	$1.0_{\pm 0.5}$	$614.0_{\pm 22.6}$
KLDM-DNG	$44.6_{\pm 0.7}$	$99.2_{\pm 0.2}$	$35.6_{\pm 1.3}$	$18.0_{\pm 0.6}$	$2.4_{\pm 0.7}$	$11.1_{\pm 0.9}$	$1.4_{\pm 0.4}$	$113.3_{\pm 9.2}$
→ CSP	$43.8_{\pm 0.2}$	$99.2_{\pm 0.2}$	$34.4_{\pm 1.5}$	$19.0_{\pm 0.7}$	$2.2_{\pm 0.6}$	$11.4_{\pm 0.8}$	$1.1_{\pm 0.2}$	$112.3_{\pm 9.3}$
+ FT (mean)	$53.4_{\pm 0.8}$	$99.2_{\pm 0.4}$	$44.0_{\pm 1.1}$	$19.3_{\pm 0.5}$	$2.0_{\pm 0.4}$	$12.0_{\pm 0.6}$	$1.3_{\pm 0.3}$	$115.0_{\pm 13.0}$
+ FT (bal.)	$52.8_{\pm 2.4}$	$99.0_{\pm 0.5}$	$43.0_{\pm 1.9}$	$19.1_{\pm 0.9}$	$2.1_{\pm 0.4}$	$11.4_{\pm 1.7}$	$1.1_{\pm 0.4}$	$114.7_{\pm 18.2}$
+ FT (mean) → CSP	$52.5_{\pm 0.7}$	$99.5_{\pm 0.1}$	$44.6_{\pm 0.9}$	$19.1_{\pm 0.0}$	$2.1_{\pm 0.4}$	$13.3_{\pm 0.3}$	$1.1_{\pm 0.5}$	$109.5_{\pm 3.5}$
+ FT (bal.) → CSP	$50.5_{\pm 2.5}$	$98.8_{\pm 0.6}$	$41.7_{\pm 0.4}$	$18.1_{\pm 0.9}$	$1.8_{\pm 0.4}$	$11.1_{\pm 0.6}$	$0.9_{\pm 0.4}$	$116.0_{\pm 24.0}$
ANCHOR + frozen CSP								
Trained on KLDM	$97.5_{\pm 0.4}$	$99.9_{\pm 0.1}$	$97.3_{\pm 0.5}$	$47.7_{\pm 2.0}$	$22.2_{\pm 1.2}$	$47.6_{\pm 2.1}$	$22.1_{\pm 1.2}$	$0.0_{\pm 0.0}$
→ Crystalite	$93.1_{\pm 0.9}$	$99.9_{\pm 0.1}$	$93.0_{\pm 1.0}$	$46.8_{\pm 2.3}$	$19.5_{\pm 1.9}$	$46.7_{\pm 2.3}$	$19.5_{\pm 1.9}$	$0.0_{\pm 0.0}$
→ OMatG	$98.1_{\pm 0.2}$	$99.9_{\pm 0.1}$	$98.0_{\pm 0.3}$	$48.2_{\pm 2.3}$	$21.8_{\pm 0.6}$	$\underline{48.1_{\pm 2.3}}$	$21.7_{\pm 0.6}$	$0.0_{\pm 0.0}$
ANCHOR + reward-trained KLDM-CSP (GRPO)								
Shared reward	$97.6_{\pm 0.8}$	$99.9_{\pm 0.1}$	$97.5_{\pm 0.9}$	$22.7_{\pm 1.6}$	$6.3_{\pm 1.3}$	$22.6_{\pm 1.7}$	$6.3_{\pm 1.3}$	$0.0_{\pm 0.0}$
Stability-only CSP reward	$97.9_{\pm 0.5}$	$100.0_{\pm 0.0}$	$97.9_{\pm 0.5}$	$29.5_{\pm 0.7}$	$10.7_{\pm 0.2}$	$29.5_{\pm 0.7}$	$10.7_{\pm 0.2}$	$0.0_{\pm 0.0}$
ANCHOR + likelihood-tuned CSP								
KLDM-CSP	$96.8_{\pm 0.1}$	$99.9_{\pm 0.1}$	$96.7_{\pm 0.3}$	$48.0_{\pm 1.2}$	$22.1_{\pm 0.9}$	$48.0_{\pm 1.3}$	$22.1_{\pm 1.0}$	$0.0_{\pm 0.0}$
+ ANCHOR data	$97.5_{\pm 0.4}$	$99.9_{\pm 0.1}$	$97.3_{\pm 0.5}$	$48.2_{\pm 2.4}$	$22.8_{\pm 1.1}$	$\underline{48.1_{\pm 2.5}}$	$\underline{22.7_{\pm 1.1}}$	$0.0_{\pm 0.0}$
Crystalite-CSP	$93.6_{\pm 1.2}$	$99.9_{\pm 0.1}$	$93.5_{\pm 1.1}$	$46.8_{\pm 1.1}$	$20.5_{\pm 0.3}$	$46.7_{\pm 1.3}$	$20.4_{\pm 0.4}$	$0.0_{\pm 0.0}$
+ ANCHOR data	$98.8_{\pm 0.1}$	$99.9_{\pm 0.1}$	$98.7_{\pm 0.3}$	$52.9_{\pm 3.5}$	$26.1_{\pm 1.3}$	$\mathbf{52.8}_{\pm 3.7}$	$\mathbf{26.0}_{\pm 1.4}$	$0.0_{\pm 0.0}$

half, so the model increasingly repeats its own formulas. This is the compositional memorization that Veljković et al. (2026) limit during pretraining by down-weighting the atom-type loss, a constraint the reward gradient does not have. Novel rises only with validity, as the novel fraction of valid samples (Novel/Valid in Table 2) falls from 68% to 52-63%, so the fine-tune optimizes around the history-dependent term rather than following it, as expected for a target that changes with the search history (Section 3). MatterGen, which does not gate stability on validity, confirms that both fine-tunes lower Meta (51.9% against 57.0%) and Stable (4.3-4.4% against 7.2%). When realised by the frozen Crystalite-CSP, the fine-tuned compositions alone lower stability, while the fine-tuned structural model recovers only about one point, showing that the loss comes primarily from the compositions rather than structural drift. KLDM-DNG retains high uniqueness, but its fine-tuned structural model is likewise no better than the frozen CSP at the same compositions. When realised by the frozen CSP, its compositions become slightly more novel but less stable under MatterGen. Dirichlet-mean weights give an on-hull fraction of 1.9-2.0%, and raising the stability weight to equal novelty lifts it only to 2.1-2.6%, against 2.8% and 2.4% for the pretrained models, consistent with the reward having to be assigned across the denoising steps (Section 2). Even without a KL term, both fine-tunes behave as the reward tilting of Section 3 describes (Appendix C), and concentrate on compositions where the reward is easiest to collect, which for Crystalite-DNG lie within the chemistry the pretrained model already proposes. Park & Walsh (2026) report larger gains for a latent DNG model trained with GRPO and strong KL regularization under a fixed-reference novelty reward, raising metastability from 51.2% to 72.1% under the potential used for training. Much of this change is compositional, as their fine-tuned samples reach novelty through new element combinations on familiar prototype structures, and their policy collapses onto a single composition

Table 3: MP-20 under MatterGen's evaluation pipeline. MSUN is the quantity MatterGen and the cited works report as SUN; SUN applies the same conjunction at $E_{\text{hull}} \leq 0$. All columns are percentages except $\overline{E}_h$ (eV/atom) and RMSD (Å). Upper block: the strongest published variant of each model family, rounded to one decimal. All published results are in Table 13. Middle blocks: our own sampling of the released Crystalite-DNG checkpoint and of a KLDM-DNG model we retrained, with their DDPO fine-tunes, which reproduce the published MSUN of their families (23.6% and 18.0% against 24.3% and 18.5%). Our rows report mean$_{\pm\text{std}}$ over three sampling seeds, with row blocks as in Table 2. Best in each column is bold, second-best underlined.
[a]Cornet et al. (2025), [b]Veljković et al. (2026), [c]Zhang et al. (2026).

Model	Unique	Novel	Meta	Stable	$\overline{E}_h$	RMSD	MSUN	SUN
DNG baselines								
MatterGen-MP[a]	–	–	47.1	–	0.20	<u>0.15</u>	25.8	–
KLDM-x_0(D)[a]	–	–	59.2	–	0.16	0.28	18.5	–
Crystalite[b]	94.7	56.6	<u>64.5</u>	–	0.15	0.27	24.3	–
MatterGen+REPA[c]	99.7	63.5	**66.2**	–	<u>0.12</u>	**0.11**	29.2	–
Untrained composition sources (frozen KLDM-CSP)								
Random formulas	$100.0_{\pm 0.0}$	$94.8_{\pm 0.3}$	$14.2_{\pm 2.2}$	$1.5_{\pm 0.5}$	$0.329_{\pm 0.012}$	$0.85_{\pm 0.02}$	$10.3_{\pm 1.7}$	$0.8_{\pm 0.4}$
Enumerated charge-neutral	$100.0_{\pm 0.0}$	$99.2_{\pm 0.1}$	$10.7_{\pm 0.6}$	$1.8_{\pm 0.0}$	$0.325_{\pm 0.008}$	$1.08_{\pm 0.01}$	$10.1_{\pm 0.6}$	$1.8_{\pm 0.1}$
DNG and reward fine-tuned DNG (ours)								
Crystalite-DNG	$99.3_{\pm 0.3}$	$63.1_{\pm 0.9}$	$57.0_{\pm 1.9}$	$7.2_{\pm 1.1}$	$0.176_{\pm 0.017}$	$0.32_{\pm 0.01}$	$23.6_{\pm 0.6}$	$2.3_{\pm 0.4}$
→ CSP	$99.2_{\pm 0.4}$	$61.5_{\pm 1.3}$	$59.1_{\pm 1.7}$	$7.9_{\pm 0.7}$	$0.169_{\pm 0.017}$	$0.26_{\pm 0.01}$	$23.9_{\pm 0.5}$	$2.5_{\pm 0.4}$
+ FT (mean)	$61.9_{\pm 1.1}$	$64.8_{\pm 0.7}$	$51.9_{\pm 0.9}$	$4.4_{\pm 0.4}$	$0.154_{\pm 0.012}$	$0.46_{\pm 0.01}$	$17.2_{\pm 0.3}$	$1.5_{\pm 0.3}$
+ FT (bal.)	$68.8_{\pm 0.9}$	$64.8_{\pm 0.8}$	$51.9_{\pm 1.0}$	$4.3_{\pm 0.4}$	$0.175_{\pm 0.008}$	$0.48_{\pm 0.01}$	$19.4_{\pm 0.8}$	$1.6_{\pm 0.3}$
+ FT (mean) → CSP	$43.0_{\pm 0.8}$	$51.4_{\pm 1.6}$	$49.5_{\pm 0.1}$	$1.9_{\pm 0.2}$	$0.250_{\pm 0.008}$	$0.22_{\pm 0.02}$	$10.5_{\pm 0.8}$	$0.8_{\pm 0.1}$
+ FT (bal.) → CSP	$52.2_{\pm 1.2}$	$51.5_{\pm 3.9}$	$47.7_{\pm 0.8}$	$2.2_{\pm 0.4}$	$0.262_{\pm 0.011}$	$0.22_{\pm 0.01}$	$13.4_{\pm 0.3}$	$1.1_{\pm 0.3}$
KLDM-DNG	$99.8_{\pm 0.1}$	$76.2_{\pm 0.7}$	$38.3_{\pm 1.4}$	$6.2_{\pm 0.6}$	$0.212_{\pm 0.009}$	$0.58_{\pm 0.02}$	$18.0_{\pm 0.9}$	$2.3_{\pm 0.4}$
→ CSP	$99.8_{\pm 0.1}$	$75.1_{\pm 1.3}$	$40.6_{\pm 1.5}$	$6.1_{\pm 0.4}$	$0.200_{\pm 0.006}$	$0.59_{\pm 0.01}$	$18.8_{\pm 2.0}$	$2.1_{\pm 0.4}$
+ FT (mean)	$99.6_{\pm 0.3}$	$78.7_{\pm 1.0}$	$34.5_{\pm 1.2}$	$3.9_{\pm 0.4}$	$0.234_{\pm 0.007}$	$0.72_{\pm 0.02}$	$16.2_{\pm 0.8}$	$1.5_{\pm 0.5}$
+ FT (bal.)	$99.5_{\pm 0.4}$	$78.5_{\pm 1.6}$	$34.7_{\pm 2.4}$	$4.8_{\pm 0.6}$	$0.222_{\pm 0.013}$	$0.69_{\pm 0.02}$	$16.2_{\pm 1.1}$	$1.9_{\pm 0.2}$
+ FT (mean) → CSP	$99.5_{\pm 0.1}$	$77.5_{\pm 1.3}$	$36.0_{\pm 1.2}$	$3.9_{\pm 0.9}$	$0.225_{\pm 0.015}$	$0.65_{\pm 0.01}$	$17.0_{\pm 2.0}$	$1.7_{\pm 0.6}$
+ FT (bal.) → CSP	$99.0_{\pm 0.6}$	$77.2_{\pm 1.2}$	$34.7_{\pm 2.5}$	$4.2_{\pm 1.1}$	$0.228_{\pm 0.006}$	$0.61_{\pm 0.01}$	$16.3_{\pm 0.6}$	$1.6_{\pm 0.1}$
ANCHOR + frozen CSP (ours)								
Trained on KLDM	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$41.3_{\pm 1.0}$	$16.0_{\pm 1.0}$	$0.138_{\pm 0.006}$	$1.39_{\pm 0.01}$	$41.3_{\pm 1.0}$	$16.0_{\pm 1.0}$
→ Crystalite	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$41.1_{\pm 2.1}$	$14.4_{\pm 0.7}$	$0.142_{\pm 0.006}$	$1.27_{\pm 0.02}$	$41.1_{\pm 2.1}$	$14.4_{\pm 0.7}$
→ OMatG	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$43.1_{\pm 0.9}$	<u>$16.7_{\pm 0.6}$</u>	$0.136_{\pm 0.005}$	$1.27_{\pm 0.01}$	<u>$43.1_{\pm 0.9}$</u>	<u>$16.7_{\pm 0.6}$</u>
ANCHOR + reward-trained KLDM-CSP (GRPO)								
Shared reward	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$17.0_{\pm 1.9}$	$4.7_{\pm 0.4}$	$0.211_{\pm 0.008}$	$1.37_{\pm 0.01}$	$17.0_{\pm 1.9}$	$4.7_{\pm 0.4}$
Stability-only CSP reward	$100.0_{\pm 0.0}$	$100.0_{\pm 0.1}$	$22.6_{\pm 1.9}$	$8.1_{\pm 0.6}$	$0.187_{\pm 0.006}$	$1.27_{\pm 0.02}$	$22.6_{\pm 1.8}$	$8.1_{\pm 0.6}$
ANCHOR + likelihood-tuned CSP (ours)								
KLDM-CSP	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$41.2_{\pm 2.1}$	$16.3_{\pm 0.7}$	$0.138_{\pm 0.008}$	$1.40_{\pm 0.01}$	$41.2_{\pm 2.1}$	$16.3_{\pm 0.7}$
+ ANCHOR data	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$40.6_{\pm 1.4}$	$15.8_{\pm 1.0}$	$0.142_{\pm 0.006}$	$1.24_{\pm 0.02}$	$40.6_{\pm 1.4}$	$15.8_{\pm 1.0}$
Crystalite-CSP	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$39.5_{\pm 2.7}$	$14.1_{\pm 0.3}$	$0.148_{\pm 0.011}$	$1.27_{\pm 0.02}$	$39.5_{\pm 2.7}$	$14.1_{\pm 0.3}$
+ ANCHOR data	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$47.1_{\pm 2.4}$	**$22.0_{\pm 2.2}$**	**$0.114_{\pm 0.007}$**	$0.86_{\pm 0.02}$	**$47.1_{\pm 2.4}$**	**$22.0_{\pm 2.2}$**

when diversity reward is removed (Park & Walsh, 2026). Our fine-tunes differ in model, reward, and the KL penalty (Appendix F.5), but both results are consistent with reward fine-tuning acting mainly through the composition marginal, which querying a frozen CSP model with the fine-tuned compositions would separate. On the same reward, ANCHOR steers completely away from these rediscoveries, with no MP-20 formulas against 612-632 per 1000 for the Crystalite-DNG fine-tunes.

Structure improves through structural signals. Since MSUN ≈ Meta and SUN ≈ Stable for ANCHOR, its discovery rate is bounded by how often the CSP model generates stable structures, and its structures move further during relaxation than those of any DNG model (RMSD 1.27-1.39 Å against at most 0.72 Å), so the relaxer still does much of the structural work. Letting GRPO update the backbone with the full reward exposes it to the same tilting and reaches only 22.6% MSUN and 6.3% SUN. Because ANCHOR separates the two factors, we can instead route only the stability term to the backbone while the policy still receives the full reward. This keeps history-dependent novelty away from the structural model and raises MSUN to 29.5% and SUN to 10.7%, with the same ordering under MatterGen, but still roughly halves SUN against the frozen backbone (22.1%). Part of this gap may come from co-training against a moving backbone, but a terminal reward still does little to improve structures through the denoising chain, even when it measures only stability. To test whether the discovered structures have any useful structural signal, we fine-tune the backbone by likelihood on MP-20 together with 27,136 randomly drawn valid structures from one full

ANCHOR run, with no stability or novelty threshold, thereby doubling the training data. None of the structures used for the three-seed evaluation is part of this fine-tuning corpus. To isolate the structural model, we compare against a step-matched control on MP-20 alone and query both with the same ANCHOR compositions as the frozen backbone. On Crystalite-CSP, this raises internal MSUN/SUN from $46.7/20.4\%$ to $52.8/26.0\%$, the highest values in Table 2, and MatterGen MSUN/SUN from $39.5/14.1\%$ to $47.1/22.0\%$. KLDM-CSP is unchanged in MSUN and SUN despite the same doubled training set, so data volume alone does not explain the Crystalite gain. Distillation does lower the relaxation RMSD of both backbones (0.86 and 1.24 against 1.27 and 1.40 Å, Appendix F.6), so the prior takes back structural work from the relaxer. The reward reaches these structures only through the validity gate and the compositions the policy queries, so the same search improves the structural model as likelihood data but not as a reward gradient through the denoising chain. The search can therefore improve where it queries the prior while also generating useful training data for the prior, although the benefit of distillation is backbone-dependent.

Strong search exploits the scorer. Our pipeline, MatterGen, and the training reward all use MatterSim potentials. Following Veljković et al. (2026), we also test pre-relaxing structures with NequIP and submitting them to LeMat-GenBench, which computes E_{hull} as the mean of ORB-v3, MACE-MP, and UMA. With the LeMat validity gate held fixed, ANCHOR→KLDM-CSP falls from 13.6% SUN in our pipeline to 0.4% under LeMat-GenBench, while Crystalite-DNG rises from 1.7% to 9.8% (Table 16). The on-hull ranking therefore reverses when the potential, relaxation, and hull change, and each model is strongest under the potential it was optimised against, since the released Crystalite-DNG checkpoint was selected for SUN under this NequIP relaxation (Veljković et al., 2026). ANCHOR's MSUN also falls (34.1% to 7.8%) while that of Crystalite-DNG does not (27.1% to 30.0%), so a policy trained against a potential exploits it more strongly than checkpoint selection does. The policy exploits a validity gate in the same way: when trained against LeMat-GenBench's own gate, ANCHOR moves to roughly 80% all-metal formulas and reaches 32.5% LeMat-GenBench MSUN, above every DNG model we evaluate, but has almost no on-hull structures (Appendix E.3.1, Figure 2). We therefore treat on-hull SUN as the headline metric. The gate also changes what an evaluator rewards, since under MatterGen, which does not gate MSUN on validity, even random formulas reach 10.3% MSUN although only 9.8% pass our gate.

5 Discussion

Our results suggest separating search from structural learning in reward-driven crystal discovery, with rewards guiding the search over compositions and likelihood training the structural prior. On the search side, discovery rewards, including history-dependent novelty, belong in the composition policy, where ANCHOR keeps 99.9% uniqueness and proposes no MP-20 formulas, whereas the same signals propagated through the denoising chain of a DNG model mostly move its composition marginal toward what is easiest to reward. CAN is therefore a search objective rather than an evaluation metric, and fixed-reference novelty and on-hull SUN remain the benchmark quantities. On the structural side, the structures the search finds carry useful signal, which in our experiments improves the prior when distilled by likelihood but not when delivered as reward gradients, even when gradients measure only stability. Search over a frozen CSP model can therefore be alternated with distillation of what it finds. Querying the frozen CSP model with a fine-tuned model's compositions tells the two kinds of change apart. The same separation could be applied beyond crystals, for example to searching protein sequences over a frozen structure predictor. The strength that makes ANCHOR effective also makes it exploit whatever it is scored with. This includes the validity gate and potential, so its gains are gains under the reward it was trained against, consistent with reward hacking observed in other reward-fine-tuned crystal generators (Park & Walsh, 2026). This is Goodhart's law applied to materials search (Gao et al., 2023), i.e., a limit of the reward rather than of the search. Our stability estimates rely on MatterSim potentials, and scoring under the LeMat-GenBench potential ensemble and validity gate, neither used in training, is only a partial check. Training against potential ensembles and DFT validation are natural extensions. However, because ANCHOR keeps all structural knowledge in transferable and replaceable components, its gains should translate more directly into discoveries as potentials and CSP models improve. Better structural priors can be substituted directly rather than rediscovered through reward optimization. In the limit of a DFT-accurate potential and a physically faithful CSP model, discovery reduces to learning where to query the prior, and any modification of the prior can only make it less faithful.

AI USE STATEMENT

We used large language models to provide feedback on the presentation of our experiments, to cross-check results already obtained, to check numerical consistency between the main text and appendix tables, and to help connect a hypothesis about reward credit assignment developed by the authors to an existing theoretical result and work through its application to our setting, as well as to draft and revise prose that was subsequently edited by the authors. All experimental design, code, training runs, and evaluations are the authors' own, and all AI-assisted theoretical reasoning was independently checked by the authors against the cited literature and derivations. No text, number, or claim appears in this paper without author verification against the underlying results or sources. We did not use generative AI to generate data or implement methods. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work is methodological and concerns generative models for inorganic crystal structures. We see no direct path from this work to harmful application.

REPRODUCIBILITY STATEMENT

Section 3 specifies the composition action space, reward, and optimisation objective, with full policy parameterisation, reward details, hyperparameters, and the training algorithm in Appendix D. The post-training evaluation protocol, including the relaxation settings, reference hull, and validity gate, is given in Appendix E.1, with the corresponding MatterGen and LeMat-GenBench protocols in Appendix E.2 and Appendix E.3. All backbones, reference datasets, and interatomic potentials are publicly available and cited. Code will be released with the camera-ready version.

REFERENCES

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

Andreas Bergmeister, Stefanie Jegelka, Nikolas Nüsken, Carles Domingo-Enrich, and Jakiw Pidstrigach. Reinforce adjoint matching: Scaling rl post-training of diffusion and flow-matching models. *arXiv preprint arXiv:2605.10759*, 2026.

Siddharth Betala, Samuel P Gleason, Ali Ramlaoui, Andy Xu, Georgia Channing, Daniel Levy, Clémentine Fourrier, Nikita Kazeev, Chaitanya K Joshi, Sékou-Oumar Kaba, et al. Lematgenbench: A unified evaluation framework for crystal generative models. *arXiv preprint arXiv:2512.04562*, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

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.

François Cornet, Federico Bergamin, Arghya Bhowmik, Juan Maria Garcia Lastra, Jes Frellsen, and Mikkel N Schmidt. Kinetic langevin diffusion for crystalline materials generation. *arXiv preprint arXiv:2507.03602*, 2025.

Yabo Dan, Yong Zhao, Xiang Li, Shaobo Li, Ming Hu, and Jianjun Hu. Generative adversarial networks (gan) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Computational Materials*, 6(1):84, 2020.

Daniel W. Davies, Keith T. Butler, Adam J. Jackson, Jonathan M. Skelton, Kazuki Morita, and Aron Walsh. SMACT: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software*, 4(38):1361, 2019. doi: 10.21105/joss.01361.

Jonas Elsborg and Arghya Bhowmik. Equivariant graph-representation-based actor–critic reinforcement learning for nanoparticle design. *Journal of chemical information and modeling*, 63(12): 3731–3741, 2023.

Jonas Elsborg and Arghya Bhowmik. Artisan: navigating the complexity of material structures with deep reinforcement learning. *Machine Learning: Science and Technology*, 5(3):035043, 2024.

Jonas Elsborg, Emma Lei Hovmand, and Arghya Bhowmik. Reinforcement learning for chemical ordering in alloy nanoparticles. *ACS Materials Au*, 6(4):696, 2026.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pp. 10835–10866. PMLR, 2023.

Philipp Höllmer and Stefano Martiniani. Open materials generation with inference-time reinforcement learning. *arXiv preprint arXiv:2602.00424*, 2026.

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

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

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

Mathias S Jørgensen, Henrik L Mortensen, Søren A Meldgaard, Esben L Kolsbjerg, Thomas L Jacobsen, Knud H Sørensen, and Bjørk Hammer. Atomistic structure learning. *The Journal of Chemical Physics*, 151(5), 2019.

Christopher Karpovich, Elton Pan, and Elsa A Olivetti. Deep reinforcement learning for inverse inorganic materials design. *npj Computational Materials*, 10(1):287, 2024.

Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017. URL `http://stacks.iop.org/0953-8984/29/i=27/a=273002`.

Nian Liu, Yuwei Zeng, Ryoji Kubo, Nikita Kazeev, Stephen Gregory Dale, Artem Maevskiy, Pengru Huang, Thomas Laurent, Kostya S Novoselov, and Xavier Bresson. Composable crystals: Controllable materials discovery via concept learning. *arXiv preprint arXiv:2605.14769*, 2026.

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

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.

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

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.

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

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.

Hedda Oschinski, Maximilian L Ach, Konstantin S Jakob, Christian Carbogno, and Karsten Reuter. General-purpose llms as constrained crystal composition generators. *arXiv preprint arXiv:2605.31495*, 2026.

Hyunsoo Park and Aron Walsh. Guiding generative models to uncover diverse and novel crystals via reinforcement learning. *Nature Machine Intelligence*, pp. 1–13, 2026.

Hyunsoo Park, Anthony Onwuli, and Aron Walsh. Exploration of crystal chemical space using text-guided generative artificial intelligence. *Nature Communications*, 16(1):4379, 2025.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

Zekun Ren, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G Aberle, Shijing Sun, et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 5(1): 314–335, 2022.

Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. *arXiv preprint arXiv:1810.02274*, 2018.

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

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, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Gregor Simm, Robert Pinsler, and José Miguel Hernández-Lobato. Reinforcement learning for molecular design guided by quantum mechanics. In *International Conference on Machine Learning*, pp. 8959–8969. PMLR, 2020.

Martin Siron, Inel Djafar, Ali Ramlaoui, Etienne du Fayette, Amandine Rossello, Edvin Fako, Matthew McDermott, Felix Therrien, Luis Barroso-Luque, Flaviu Cipcigan, et al. Lematbulk: aggregating, and de-duplicating quantum chemistry materials databases. *arXiv preprint arXiv:2511.05178*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Luca A Thiede, Mario Krenn, AkshatKumar Nigam, and Alán Aspuru-Guzik. Curiosity in exploring chemical spaces: intrinsic rewards for molecular reinforcement learning. *Machine Learning: Science and Technology*, 3(3):035008, 2022.

Hanna Turk, Elisabetta Landini, Christian Kunkel, Johannes T Margraf, and Karsten Reuter. Assessing deep generative models in chemical composition space. *Chemistry of Materials*, 34(21): 9455–9467, 2022.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Tommaso Biancalani, and Sergey Levine. Fine-tuning of continuous-time diffusion models as entropy-regularized control. *arXiv preprint arXiv:2402.15194*, 2024.

Mario Valle and Artem R Oganov. Crystal fingerprint space–a novel paradigm for studying crystal-structure sets. *Foundations of Crystallography*, 66(5):507–517, 2010.

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

Daniel Widdowson, Marco Mosca, Angeles Pulido, Vitaliy Kurlin, and Andrew I Cooper. Average minimum distances of periodic point sets are fundamental invariants for mapping all periodic crystals. *MATCH Communications in Mathematical and in Computer Chemistry*, 87(3):529–559, 2022.

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

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, et al. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024.

Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in neural information processing systems*, 31, 2018.

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

Chengqian Zhang, Yucheng Jin, Duo Zhang, Tiejun Li, and Han Wang. Crystalrepa: Transferring physical priors from universal mlips to crystal generative models. *arXiv preprint arXiv:2605.08960*, 2026.

Yong Zhao, Mohammed Al-Fahdi, Ming Hu, Edirisuriya MD Siriwardane, Yuqi Song, Alireza Nasiri, and Jianjun Hu. High-throughput discovery of novel cubic crystal materials using deep generative neural networks. *Advanced Science*, 8(20):2100566, 2021.

APPENDIX

TABLE OF CONTENTS

A Notation **16**

B Additional Related Work **17**

C Composition Search, Novelty, and the Structural Prior **19**

D Method and Training Details **24**

D.1 Element Vocabulary and Action Space 24

D.2 Structure Generation, Relaxation, and the Validity Gate 24

D.3 Stability Score 25

D.4 Continuous Adaptive Novelty and the Count Bonus 26

D.5 Within-group Diversity Kernels 27

D.6 Policy Parametrization and Optimization Details 28

D.7 Training Algorithm 30

D.8 Training Setup and Hyperparameters 30

E Evaluation Protocols **33**

E.1 Internal Evaluation Protocol 33

E.2 MatterGen Evaluation Pipeline 37

E.3 LeMat-GenBench Evaluation 39

F Additional Experiments and Baselines **43**

F.1 Base-model Selection 43

F.2 Invalid-structure Penalty Sweep 43

F.3 Exploration-mechanism Ablations 45

F.4 Learned Composition Preferences 47

F.5 Reward Fine-tuning Baselines 47

F.6 Training the CSP Backbone on ANCHOR Structures 49

A NOTATION

Table 4: Symbols for the composition policy, reward, and evaluation. Rates are fractions of an evaluation batch of $N_{\text{eval}} = 1000$ structures, unless stated. Generative-model notation is defined in section 2.

Symbol	Type	Meaning
Composition action space		
$\mathcal{C}$	set	discrete chemical space searched by the policy
$\mathcal{E}$	set	element vocabulary; $E = \lvert\mathcal{E}\rvert = 84$
$\mathcal{K}$	set	admissible element counts, $\mathcal{K} = \{2, 3, 4\}$; $k_{\max} = \max \mathcal{K} = 4$
k	integer	number of distinct elements in an action, $k \in \mathcal{K}$
$\boldsymbol{e}$	$\mathcal{E}^k$	the k distinct chosen elements
ρ	$\mathbb{Z}^k_{>0}$	atom stoichiometry
T	integer	total atoms in the cell, $T = \sum_i \rho_i$, $k \le T \le N$
N	integer	maximum atoms per unit cell, $N = 20$
$c = (k, \boldsymbol{e}, T, \rho)$	action	one composition action
$f(c)$	string	reduced chemical formula of action c
$\mathcal{S} \subseteq \mathcal{E}$	set	element sub-group active for a rollout (mask)
$n_{\mathcal{S}} = \lvert\mathcal{S}\rvert$	integer	sub-group size, $n_{\mathcal{S}} \sim \text{Uniform}\{k_{\max}, \dots, E\}$
Structures and novelty		
x	structure	crystal returned by the frozen backbone
$\boldsymbol{z}(x)$	$\mathbb{R}^{n_{\text{bins}}}$	ℓ_1-normalized RDF histogram of x over n_{bins} radial bins
Z^{ref}	set	frozen MP-20 reference embeddings for formula $f(c)$
Z^{hist}	set	growing agent-history embeddings for formula $f(c)$
$\mathcal{F}_{\text{ref}}$	set	reduced formulas present in MP-20
$\nu(\boldsymbol{z}; Z)$	$[0, 1]$	kNN–RBF novelty of $\boldsymbol{z}$ against reference set Z
$\nu(x)$	$[0, 1]$	adaptive novelty (min over reference and history)
ν_0	scalar	bootstrap novelty for an unseen off-MP-20 formula
k_{nn}	integer	kNN neighbourhood size used by $\nu(\boldsymbol{z}; Z)$
σ	scalar	RBF bandwidth (per formula, with global floor)
Reward		
s_{stab}	$[0, 1]$	stability score from E_{hull}
E_{p}	eV	relaxed potential energy under the interatomic potential
E_{hull}	eV/atom	relaxed energy above the convex hull
$d_{\text{struct}}, d_{\text{comp}}$	$[0, 1]$	within-group structural / compositional diversity
$\boldsymbol{m}(c)$	$[0, 1]^4$	objective vector $(s_{\text{stab}}, \nu, d_{\text{struct}}, d_{\text{comp}})$
$\boldsymbol{w}$	simplex	objective weights, $\boldsymbol{w} \sim \text{Dirichlet}(\boldsymbol{\alpha})$
$\boldsymbol{\alpha}$	$\mathbb{R}^4_{>0}$	Dirichlet concentration over the four objectives
n_f	integer	times formula f has been generated in the run
λ_{c}	scalar	count-bonus scale
r_{inv}	scalar < 0	fixed penalty for invalid structures
$r(c)$	scalar	scalar reward for action c
Policy and optimization		
π_θ	policy	autoregressive composition policy
$\boldsymbol{s}$	vector	policy state, $[\, \boldsymbol{W}_f \bar{\boldsymbol{f}}_{\mathcal{S}} \,;\, \boldsymbol{w} \,]$
$\boldsymbol{f}_e$	vector	fixed chemistry-aware feature vector of element e
$\bar{\boldsymbol{f}}_{\mathcal{S}}$	vector	mean of $\boldsymbol{f}_e$ over the active sub-group $\mathcal{S}$
$\boldsymbol{W}_f$	matrix	learned projection of $\bar{\boldsymbol{f}}_{\mathcal{S}}$ into the state
G	integer	group size (samples per rollout)
A_i	scalar	group-relative advantage of sample i
η_i	scalar	importance ratio $\pi_\theta / \pi_{\theta_{\text{old}}}$
ϵ	scalar	PPO clipping parameter
ϵ_A	scalar	floor on the group reward standard deviation
β_H	scalar	entropy-bonus coefficient (scheduled)
Evaluation		
Valid	rate	fraction passing the validity gate of Appendix E.1
Unique	rate	distinct reduced formulas, as a fraction of the batch
Stable	rate	valid $\wedge$ $E_{\text{hull}} \le 0$
Metastable	rate	valid $\wedge$ $E_{\text{hull}} \le 0.1$ eV/atom
SUN	rate	valid $\wedge$ unique $\wedge$ novel $\wedge$ $E_{\text{hull}} \le 0$
MSUN	rate	valid $\wedge$ unique $\wedge$ novel $\wedge$ $E_{\text{hull}} \le 0.1$ eV/atom

B Additional Related Work

Crystal-generation tasks and benchmark datasets. A crystalline material is specified by its chemical composition and periodically repeated unit cell, typically represented by atomic species, fractional coordinates, and lattice vectors. This gives rise to two closely related generative tasks. Crystal structure prediction (CSP) conditions on a fixed composition and searches over compatible structures, while de novo crystal generation explores composition and structure jointly (Jiao et al., 2023; Xie et al., 2021; Cornet et al., 2025).

Early crystal-generation benchmarks commonly used datasets derived from the Materials Project, including MP-20 (Jain et al., 2013; Xie et al., 2021). Evaluation has since expanded toward larger reference collections such as Alexandria, Alex-MP, and LeMat-Bulk (Schmidt et al., 2024; Zeni et al., 2025; Siron et al., 2025). This expansion matters directly for novelty evaluation because the reference corpus forms part of the definition of whether a generated candidate is considered known. A structure judged novel with respect to a smaller reference may no longer be novel when evaluated against a larger collection.

LeMat-GenBench provides a broader evaluation framework spanning validity, stability, novelty, uniqueness, diversity, and computational efficiency (Betala et al., 2025). Its stable, unique, and novel metric can be written as

$$\mathrm{SUN}(x) = \mathbb{1}[\mathrm{valid}(x) \wedge \mathrm{unique}(x) \wedge \mathrm{novel}(x) \wedge E_{\mathrm{hull}}(x) \leq 0]\,, \tag{9}$$

with mSUN (MSUN in this paper) relaxing the stability threshold to $E_{\mathrm{hull}} \leq 0.1\,\mathrm{eV/atom}$. MatterGen (Zeni et al., 2025) instead reports the 0.1 quantity as SUN. Benchmark results expose a persistent interaction between stability, novelty, and diversity, motivating evaluation of all three rather than stability in isolation.

Generative models for crystalline materials. Early latent-variable and adversarial approaches established learned generative design for inorganic crystals (Noh et al., 2019; Ren et al., 2022; Dan et al., 2020; Zhao et al., 2021). Subsequent models increasingly incorporated the periodic geometry of crystals directly into their generative parameterization. CDVAE introduced diffusion-based decoding for periodic crystal structures, while DiffCSP jointly models lattice parameters and fractional coordinates for composition-conditioned crystal structure prediction (Xie et al., 2021; Jiao et al., 2023). KLDM models fractional coordinates on the hypertorus through kinetic Langevin dynamics (Cornet et al., 2025) and is used as the primary structural backbone in this work.

More recent models have expanded conditioning and generative scale. MatterGen targets property-conditioned inorganic materials generation, while Chemeleon extends diffusion-based crystal generation toward broader controllable exploration (Zeni et al., 2025; Park et al., 2025). Flow- and Transformer-based approaches have also been proposed to simplify or accelerate generative sampling (Miller et al., 2024; Luo et al., 2025; Veljković et al., 2026). These developments motivate treating a pretrained crystal generator as a reusable source of structural knowledge rather than necessarily retraining it for each downstream discovery objective.

Novelty and structural similarity. Novelty and uniqueness answer different questions in crystal-generation evaluation. Novelty conventionally measures whether a generated structure is absent from a fixed external reference collection, while uniqueness measures repetition among generated samples (Zeni et al., 2025; Betala et al., 2025). Stability is evaluated separately using formation energy or energy above the convex hull (Ong et al., 2013). This separation makes fixed-reference novelty well suited to reproducible comparison between generative models, but it does not encode what has already been discovered during a continuing search.

Binary structure matching further discards information about the degree of structural similarity. Continuous structural metrics provide a graded alternative (Valle & Oganov, 2010; Negishi et al., 2026). In particular, continuous formulations can distinguish near-repeats from structures that are substantially different even when both would receive the same binary novel/known label. The continuous component of CAN builds on this general idea, while its adaptive component introduces persistent search history.

A concurrent compositional approach makes the same trade visible from a different direction. Liu et al. (2026) learn a discrete codebook of local atomic environments with a VQ-VAE and condition

generation on recombined codes, reporting large novelty gains (78.0% to 90.0% on MP-20, 61.8% to 85.9% on Alex-MP-20) alongside a drop in stability (48.4% to 46.9% and 69.0% to 59.1%). Their intervention is a structural conditioning signal rather than a reward, but the signature is the one Appendix C predicts for any pressure that moves the conditional distribution toward novelty: candidates further from the reference set are also further from the region the pretrained model assigns high probability. Their headline scores are computed against MP-2023 for both novelty and the hull, so they are not directly comparable to the MatterGen-pipeline numbers in Table 3.

Repeated proposals are not unique to crystal-generation benchmarks. Prior materials-search work has observed repeated sampling during prolonged exploration and has used search history to discourage revisiting previously explored candidates (Turk et al., 2022; Oschinski et al., 2026). More broadly, reinforcement learning has a long history of history-dependent exploration mechanisms, including count-based bonuses, curiosity, and related intrinsic rewards (Bellemare et al., 2016; Pathak et al., 2017; Burda et al., 2018; Thiede et al., 2022). These approaches establish precedent for allowing an exploration objective to depend on the state of the search rather than only on a fixed external dataset.

Reinforcement learning for molecular and materials discovery. Reinforcement learning has been applied across several levels of molecular and materials design. Molecular graph-generation methods use sequential actions to construct molecules optimized for target properties (You et al., 2018). Other work applies RL directly to atomistic or molecular construction and assembly (Jørgensen et al., 2019; Simm et al., 2020). Materials-focused approaches have used reinforcement learning for composition search, atomic ordering, and experimentally or computationally evaluated design problems (Turk et al., 2022; Elsborg & Bhowmik, 2023; 2024; Karpovich et al., 2024; Elsborg et al., 2026).

These studies are important precedent for the composition-level policy used in ANCHOR. Composition-level reinforcement learning itself is therefore not the central distinction of the present work. The distinction is the role assigned to that policy: it acts as an adaptive outer search over a separately pretrained conditional crystal generator, which remains frozen throughout discovery.

Reward-guided crystal generation. A separate line of work applies optimization directly to generative crystal models. MatInvent uses reinforcement learning for single- and multi-objective diffusion-based crystal design (Chen et al., 2025). OMatG-IRL applies policy-gradient optimization to continuous-time crystal generators at inference (Höllmer & Martiniani, 2026). These methods use downstream rewards to modify the behavior of the model responsible for generating crystal structures.

Chemeleon2 gives an explicit reinforcement-learning formulation of latent diffusion (Park & Walsh, 2026). Starting from noise z_T, generation follows the reverse-denoising trajectory

$$\tau = z_T \to z_{T-1} \to \cdots \to z_0.$$

The decoder D_ϕ maps the final latent state to a generated crystal $x = D_\phi(z_0)$, and the RL objective maximizes the expected terminal crystal-level reward,

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[R(x) \right]. \tag{10}$$

A terminal crystal reward must therefore be attributed across the denoising trajectory.

Chemeleon2 explicitly discusses an objective mismatch between likelihood-based pretraining and discovery. Pretraining concentrates probability in regions represented by the training distribution, while its "creativity" reward promotes novelty and uniqueness and therefore encourages exploration of less represented regions (Park & Walsh, 2026). The resulting GRPO objective uses strong KL regularization to the pretrained denoiser to restrict drift from the learned prior.

This setup differs from ANCHOR in where reinforcement learning acts. Reward-guided DNG adapts the joint generator and can therefore change both $p(c)$ and $p(x \mid c)$, whereas ANCHOR adapts only the composition-level search while keeping the conditional structural distribution fixed. The novelty-prior consequences of modifying $p(x \mid c)$ are developed separately in Appendix C.

C Composition Search, Novelty, and the Structural Prior

The main text separates two questions that are combined in conventional de novo crystal generation. The first is where in chemical space to search. The second is which structures are plausible for a queried composition. We develop this distinction further here and examine how it interacts with reward-guided generator fine-tuning and history-dependent novelty.

DNG, CSP, and the composition marginal. De novo crystal generation concerns a joint distribution over composition c and structure x, which admits the conceptual factorization

$$p(c, x) = p(c)\, p(x \mid c). \tag{11}$$

The conditional factor corresponds to the CSP problem at fixed composition. This factorization does not require the underlying DNG model to contain separately parameterized composition and CSP components. In practice, DNG and CSP models are typically trained separately.

The factorization nevertheless exposes two distinct roles during discovery. A likelihood-trained DNG model learns a composition marginal that reflects which chemistries are probable under its training distribution. A CSP model instead answers which structures are plausible after a composition has already been specified. These need not have the same domain of usefulness. In particular, a pretrained CSP model may generate useful structures for compositions that receive little probability under the composition marginal learned by DNG.

For a frozen CSP model $p_0(x \mid c)$, define the probability that querying composition c produces a successful discovery as

$$s_0(c) = \Pr_{x \sim p_0(x|c)} \left[S(x, c) = 1 \right], \tag{12}$$

where S denotes a chosen discovery criterion. If compositions are sampled from a distribution $r(c)$, successful candidates from composition c occur with probability proportional to

$$r(c)\, s_0(c). \tag{13}$$

Direct DNG therefore allocates discovery effort according to $p_0(c)s_0(c)$. A composition can have high discovery potential under the structural model while remaining rarely observed because $p_0(c)$ is small. Replacing the composition marginal with an adaptive search policy instead gives $\pi_\theta(c)s_0(c)$ and allows search effort to move toward compositions where the frozen CSP model is empirically productive.

The distinction is therefore between reproducing the chemical statistics of the training distribution and allocating future discovery effort. The composition marginal $p_0(c)$ describes which chemistries resemble those already represented in the data. A discovery policy $\pi_\theta(c)$ instead learns which chemistries should be investigated next. The former can provide a useful prior without being the optimal search distribution.

KL-regularized reward tilting. Reward-guided fine-tuning provides one way to move beyond the pretrained DNG distribution. It can modify the composition marginal, but can simultaneously modify the conditional structural distribution. We use the standard KL-regularized reward-tilting identity to examine the latter effect, in the form stated by Bergmeister et al. (2026).

For the remainder of this analysis, $p_0(x \mid c)$ denotes a pretrained conditional structural distribution, $q(x \mid c)$ a reward-adapted endpoint distribution, and $q^*(x \mid c)$ the reward-optimal endpoint distribution under a KL-regularized objective. For a pretrained endpoint distribution and scalar reward, Bergmeister et al. (2026) consider

$$\max_q \left\{ \mathbb{E}_{x \sim q(\cdot|c)} \left[R(x, c) \right] - \beta D_{\mathrm{KL}} \left(q(\cdot \mid c) \,\|\, p_0(\cdot \mid c) \right) \right\}, \qquad \beta > 0.$$

The corresponding reward-optimal endpoint distribution is

$$q^*(x \mid c) = \frac{p_0(x \mid c) \exp(R(x, c)/\beta)}{Z(c)}, \tag{14}$$

with

$$Z(c) = \mathbb{E}_{x \sim p_0(\cdot|c)} \left[\exp(R(x, c)/\beta) \right].$$

This is the Gibbs solution of the KL-regularized objective, written with an explicit regularization coefficient β and conditioned on composition, which Bergmeister et al. (2026) derive from the Donsker-Varadhan variational principle.

Application to novelty. We first consider the illustrative novelty reward

$$R_{\text{nov}}(x, c) = -\alpha \log p_0(x \mid c), \qquad \alpha > 0, \tag{15}$$

which assigns higher reward to structures receiving lower probability under the pretrained structural distribution. Because composition is fixed, this example isolates the structural effect of novelty and does not itself encourage exploration of new compositions. Low conditional probability also does not distinguish a useful new polymorph from an implausible structure that departs from the structural preferences learned during pretraining.

Substitution into Eq. 14 gives

$$q^*(x \mid c) \propto p_0(x \mid c)^{\,1-\alpha/\beta}. \tag{16}$$

Increasing novelty pressure relative to KL regularization reduces the exponent $1 - \alpha/\beta$, weakening the preference for high-probability structures and flattening the pretrained structural distribution. As α/β approaches one, differences inherited from p_0 are progressively suppressed. For $\alpha/\beta > 1$, the probability ordering can reverse wherever the resulting distribution remains normalizable.

Equation 16 is an illustrative specialization of Eq. 14. The novelty measure used in ANCHOR is not defined through model likelihood. Structural diversity within $p_0(x \mid c)$ remains useful for discovery. The point is that novelty alone does not identify whether an unusual structure is a useful new polymorph or an implausible departure from the pretrained CSP distribution.

General reward objectives. The inverse-density example gives a simple closed-form case, while the direction of reward-driven shift can be characterized more generally. Consider the infinitesimal tilt

$$q_\epsilon(x \mid c) \propto p_0(x \mid c) \exp(\epsilon R(x, c)) \,. \tag{17}$$

The change in expected pretrained log-likelihood at $\epsilon = 0$ is

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} \mathbb{E}_{x \sim q_\epsilon(\cdot \mid c)} \left[\log p_0(x \mid c)\right] \Big|_{\epsilon=0} = \mathrm{Cov}_{p_0(\cdot \mid c)} \left(\log p_0(x \mid c), R(x, c)\right). \tag{18}$$

A negative covariance therefore means that increasing reward pressure decreases expected pretrained log-likelihood. For the illustrative novelty reward above,

$$\mathrm{Cov}_{p_0} \left(\log p_0, -\alpha \log p_0\right) = -\alpha \,\mathrm{Var}_{p_0} \left[\log p_0\right] \leq 0.$$

For crystal discovery, the relevant reward component is novelty. A negative association between novelty and pretrained likelihood is plausible because structural novelty is usually measured against collections of known crystals drawn from the same materials space, and often from the same or closely related databases used to train the structural model. Structures farther from this reference set may therefore also lie in regions assigned lower probability by $p_0(x \mid c)$. This association is not guaranteed because a generative model can generalize beyond its training examples.

Semantic credit assignment. Reward-guided DNG can partly correct the composition-marginal problem because the reward-adapted joint distribution

$$q(c, x) = q(c)\, q(x \mid c) \tag{19}$$

is free to move its composition marginal away from the likelihood-trained distribution. The same optimization can also change the conditional structural distribution.

A joint crystal-level objective does not enforce how adaptation is allocated between these factors. A high reward may arise from selecting promising chemistry, from generating a useful structure within that chemistry, or from both. Updating the full DNG generator can therefore modify both the composition distribution and the conditional structural distribution.

This creates a semantic credit-assignment problem distinct from assigning a terminal reward across the steps of a denoising trajectory. Reward obtained after moving into productive chemistry can still change $q(x \mid c)$ even when no change to the structural distribution was required. Structural exploration within $p_0(x \mid c)$ can remain useful without requiring the conditional distribution itself to adapt.

This distinction does not imply that structural adaptation is inherently harmful. A reward that measures structural quality can provide useful information for improving the conditional model. Stability is one example. The concern is that a joint crystal-level objective also permits structural adaptation when reward originates from chemical exploration or from discovery signals that do not imply any corresponding change in structural plausibility.

The reward-optimal composition marginal. Applying equation 14 to the joint rather than the conditional makes the effect on composition explicit. For a crystal-level reward $R(x, c)$ and the pretrained joint $p_0(c, x) = p_0(c)p_0(x \mid c)$,

$$q^*(c, x) \propto p_0(c)\, p_0(x \mid c) \exp(R(x, c)/\beta)\,, \tag{20}$$

and marginalizing over x gives

$$q^*(c) \propto p_0(c)\, \mathbb{E}_{x\sim p_0(\cdot|c)}[\exp(R(x, c)/\beta)] = p_0(c)\, Z(c). \tag{21}$$

The reward-optimal composition marginal is therefore the pretrained marginal reweighted by a per-composition partition function. For a binary discovery reward $R = \mathbb{1}[S(x, c) = 1]$, $Z(c) = 1 + (e^{1/\beta} - 1)s_0(c)$, which is monotone in $s_0(c)$: equation 21 is the soft analogue of the allocation $p_0(c)s_0(c)$ in equation 13, and recovers the same ordering over compositions. The conditional is tilted in parallel, $q^*(x \mid c) \propto p_0(x \mid c) \exp(R(x, c)/\beta)$, which is exactly the structural adaptation that the frozen construction removes.

The tilt is bounded. If the reward is bounded, $R_{\min} \le R(x, c) \le R_{\max}$, then $e^{R_{\min}/\beta} \le Z(c) \le e^{R_{\max}/\beta}$ for every composition, and the normalizer $\sum_{c'} p_0(c')Z(c')$, a convex combination of these values, lies in the same interval. With $\Delta R = R_{\max} - R_{\min}$,

$$e^{-\Delta R/\beta} \;\le\; \frac{q^*(c)}{p_0(c)} \;\le\; e^{\Delta R/\beta}. \tag{22}$$

The reward-optimal marginal therefore keeps the support of $p_0(c)$, and a composition to which the pretrained model assigns probability $p_0(c)$ receives at most $e^{\Delta R/\beta}p_0(c)$. KL-regularized fine-tuning can reweight chemistry the pretrained model already proposes, but it cannot reach chemistry to which that model assigns negligible probability unless β becomes small, which is also the regime in which the structural tilt of equation 14 is least constrained. Our DDPO baselines use no KL term, so equation 22 does not constrain them, but neither does anything constrain their conditional, which is then free to adapt to the full reward. The reward of ANCHOR is bounded as well, with $r(c) \in [r_{\text{inv}}, 1 + \lambda_c]$ (equation 8), but its composition policy is not regularized toward $p_0(c)$ and its conditional is frozen, so it escapes the bound without the structural drift.

KL regularization and a frozen CSP model. The same distinction appears in the KL penalty. Writing the pretrained and reward-adapted joint endpoint distributions as

$$p_0(c, x) = p_0(c)p_0(x \mid c) \qquad \text{and} \qquad q(c, x) = q(c)q(x \mid c),$$

the chain rule for KL divergence gives

$$D_{\text{KL}}(q(c, x)\|p_0(c, x)) = D_{\text{KL}}(q(c)\|p_0(c)) + \mathbb{E}_{c\sim q(c)} D_{\text{KL}}(q(x \mid c)\|p_0(x \mid c))\,. \tag{23}$$

A single KL coefficient therefore penalizes changes in both chemical exploration through $q(c)$ and structural adaptation through $q(x \mid c)$. Increasing its strength to protect the conditional structural distribution also restricts movement of the composition distribution away from its pretrained marginal.

Freezing the CSP model imposes the hard constraint

$$q(x \mid c) = p_0(x \mid c), \tag{24}$$

which gives

$$\mathbb{E}_{c\sim q(c)} D_{\text{KL}}(q(x \mid c)\|p_0(x \mid c)) = 0. \tag{25}$$

The conditional structural distribution is then fixed exactly while a separately parameterized composition policy remains free to adapt.

One could instead introduce separate regularization strengths

$$\beta_c D_{\text{KL}}(q(c)\|p_0(c)) + \beta_x \mathbb{E}_{c\sim q(c)} D_{\text{KL}}(q(x \mid c)\|p_0(x \mid c))\,. \tag{26}$$

which makes the distinction explicit. A single joint KL coefficient couples chemical and structural regularization. Freezing the CSP model fixes the second factor directly while leaving chemical adaptation separate.

The frozen construction therefore differs qualitatively from simply increasing the strength of joint KL regularization. Stronger joint regularization can reduce structural drift, but it simultaneously restricts the same compositional adaptation required to move beyond the likelihood-trained DNG marginal. ANCHOR instead removes structural adaptation as a route for optimizing the discovery reward while leaving the composition policy unconstrained by the structural KL term.

Application to history-dependent novelty. History-dependent novelty follows a standard idea from reinforcement learning, where exploration bonuses decrease as states become familiar through visitation counts or episodic memory (Bellemare et al., 2016; Savinov et al., 2018). Applying this principle to crystal discovery makes novelty depend on the state of the search. At search step t,

$$R_t(x) = R_{\text{nov}}\left(x, \mathcal{D}_{\text{ref}}, \mathcal{H}_t\right), \tag{27}$$

where $\mathcal{D}_{\text{ref}}$ is the fixed novelty reference set and $\mathcal{H}_t$ contains the structures discovered so far.

Under the same endpoint-distribution analysis, Eq. 14 gives

$$q_t^*(x \mid c) = \frac{p_0(x \mid c)\exp(R_t(x)/\beta)}{Z_t(c)}. \tag{28}$$

The mathematics follows directly from the same tilting identity. As $\mathcal{H}_t$ grows, a previously discovered structure can receive lower novelty reward and therefore lower relative probability under $q_t^*(x \mid c)$ even though the structure itself, its physical properties, and its probability under $p_0(x \mid c)$ have not changed.

This exposes the distinction between search utility and structural quality particularly clearly. A crystal does not become less structurally plausible because it has already been discovered. It only becomes less useful to sample again. A structural generator optimized directly with history-dependent novelty is therefore asked to change because the search has already encountered a candidate, not because new information has changed the candidate's physical plausibility. The preferred reward-optimal structural distribution consequently evolves with campaign history.

A search policy should respond to this information. The structural prior need not. Campaign history can change which compositions should be queried and how often they should be revisited without requiring a corresponding change in the conditional structural distribution. Stronger KL regularization can suppress this effect when fine-tuning the generator, but again reduces the influence of the discovery reward on the joint model.

Connection to diffusion fine-tuning. The analysis above concerns endpoint distributions and does not describe the exact parameter updates of a particular diffusion-RL algorithm. Diffusion models generate samples through reverse dynamics associated with a target distribution and its noisy marginals (Song et al., 2020). Chemeleon2, for example, regularizes its learned denoising policy against a pretrained reference throughout the generative process (Park & Walsh, 2026). The endpoint analysis isolates the corresponding distribution-level effect of reward optimization. Changing the reward changes the endpoint distribution favored by the regularized objective.

This distinction is important because successful reward fine-tuning does not require structural quality to deteriorate. A joint reward containing stability can improve the structural distribution, and recent reward-guided crystal generators demonstrate that such optimization can substantially improve discovery metrics (Park & Walsh, 2026; Chen et al., 2025). The argument here is narrower. Joint fine-tuning permits both $q(c)$ and $q(x \mid c)$ to adapt, while the reward does not generally specify which factor should change. Reward-driven diffusion fine-tuning can also exhibit loss of diversity or sample quality when optimization moves sufficiently far from pretrained behavior (Uehara et al., 2024), providing an additional practical reason to control unnecessary structural adaptation. Bergmeister et al. (2026) additionally show that the KL-regularized optimum tilts only the clean-endpoint distribution and leaves the conditional law of noisy states given that endpoint unchanged. Reward fine-tuning of a diffusion model therefore changes which endpoints are likely, not how a fixed endpoint is noised. The separation ANCHOR makes architecturally is the composition-space analogue: the reward changes which compositions are queried, not how a structure is generated once one is fixed.

Implication for ANCHOR. ANCHOR instantiates the composition–structure separation directly as

$$p_{\text{ANCHOR}}(c, x) = \pi_\theta(c)\, p_0(x \mid c), \tag{29}$$

where $\pi_\theta(c)$ is an adaptive composition policy and $p_0(x \mid c)$ is a separately pretrained and frozen CSP model. Unlike the conceptual factorization of a conventional DNG distribution, Eq. 29 is an architectural decomposition. The two factors are represented by separate models and only the composition policy is updated during discovery.

The composition policy can therefore increase the probability of querying regions where the CSP model produces useful candidates even if those regions receive little probability under a likelihood-trained DNG composition marginal. At the same time,

$$p_{\text{ANCHOR}}(x \mid c) = p_0(x \mid c) \tag{30}$$

by construction. This removes reward-driven changes to the conditional structural distribution, including both the semantic credit-assignment effects of joint generator fine-tuning and the non-stationarity introduced by history-dependent novelty.

Freezing the CSP model does not freeze structural exploration. Repeated queries to the same composition can still sample different structures from $p_0(x \mid c)$. ANCHOR therefore changes where and how often the structural model is queried without restricting the structural diversity already represented by the pretrained CSP distribution.

This separation becomes increasingly natural as CSP models improve. In the limiting case of a CSP model that perfectly captures the physically plausible structural distribution for any queried composition, a separately trained DNG model becomes unnecessary for discovery. Discovery reduces to learning where and how often to query the CSP model. A learned DNG composition marginal may remain useful as an initialization or chemical prior, but there is no reason for the discovery distribution itself to reproduce the composition frequencies of the training data or to modify a conditional structural distribution that is already correct.

D METHOD AND TRAINING DETAILS

D.1 ELEMENT VOCABULARY AND ACTION SPACE

Vocabulary. The policy draws elements from the $E = 84$ symbols listed in Table 5. These are the 89 elements appearing in MP-20, minus the five noble gases (He, Ne, Ar, Kr, Xe). The noble gases are removed because they have no tabulated ICSD oxidation states in pymatgen (Ong et al., 2013), so any composition containing one of these 5 elements will fail the charge-neutrality half of the validity gate by construction (Appendix D.2). Keeping these elements in the vocabulary would spend needless rollouts on actions whose reward is known in advance to be r_{inv}, which adds variance to the group baseline (equation 45) without adding information.

Table 5: The $E = 84$ elements available to the policy, grouped by block and listed by ascending atomic number within each row. Removed from MP-20: He, Ne, Ar, Kr, Xe.

Block	Elements
Period 1–3	H, Li, Be, B, C, N, O, F, Na, Mg, Al, Si, P, S, Cl
Period 4	K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Ge, As, Se, Br
Period 5	Rb, Sr, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, In, Sn, Sb, Te, I
Period 6	Cs, Ba, Hf, Ta, W, Re, Os, Ir, Pt, Au, Hg, Tl, Pb, Bi
Lanthanides	La, Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu
Actinides	Ac, Th, Pa, U, Np, Pu

Action-space bounds. A composition action (equation 5) uses $k \in \mathcal{K} = \{2, 3, 4\}$ distinct elements and a cell size $T \in \{k, \ldots, N\}$ with $N = 20$, with at most 12 atoms of any single element. We exclude $k = 1$ because unary compositions carry no compositional-discovery signal, and cap k at 4.

Size of the search space. Enumerating all charge-neutral compositions consistent with the tabulated ICSD states (Ong et al., 2013) gives roughly 4.4×10^4 binary, 1.1×10^7 ternary and 1.1×10^9 quaternary candidates. The policy therefore searches a space of order 10^9 compositions while a training run evaluates on the order of 10^5 of them. This is the regime that makes the exploration mechanisms, described in section 3 and ablated in section 4, load-bearing rather than cosmetic.

D.2 STRUCTURE GENERATION, RELAXATION, AND THE VALIDITY GATE

Frozen CSP backbone and relaxation. Given a reduced formula $f(c)$, the frozen KLDM CSP model (Cornet et al., 2025) samples a structural realization. The CSP model is queried anew for every composition action, so repeated queries of the same reduced formula may produce different structures. The sampled crystal is then relaxed with BFGS under the MatterSim interatomic potential (Yang et al., 2024), over atomic positions only and with the lattice fixed at the backbone output. Relaxation and all subsequent reward calculations form part of the environment observed by the composition policy. Neither the CSP backbone nor the interatomic potential receives gradient updates. The complete path for a sampled action is therefore

$$c \longrightarrow f(c) \longrightarrow x \sim p_0(x \mid f(c)) \longrightarrow x_{\text{relaxed}} \longrightarrow r(c),$$

and only the mapping from the rollout state to the composition action is learned.

Evaluation order. Each relaxed structure is evaluated in a fixed order: **(i)** validity gate, **(ii)** stability (equation 32, valid only), **(iii)** novelty (equation 7, valid only), **(iv)** within-group diversity (equation 34, over all G samples in the rollout, after per-structure scoring).

Validity gate. A structure is *valid* only if it passes both of the following checks.

1. **Bond lengths.** No two atoms may sit closer than an element-pair minimum. For a site pair of elements (a, b) at minimum-image distance d, the structure is rejected if $d < \theta_{ab}$, with

$$\theta_{ab} = \begin{cases} (1 - \tau)\, d_{ab}^{\min}, & (a, b) \text{ in the reference table,} \\ d_{\text{floor}}, & \text{otherwise.} \end{cases} \tag{31}$$

Here $d_{ab}^{\min}$ is the shortest a–b separation observed in MP-20, and $\tau = 0.2$ is a fractional tolerance that admits bonds up to 20% below that reference. Pairs never seen in MP-20 fall back to a global floor $d_{\text{floor}} = 1.2\,\text{Å}$, applied without tolerance. A single too-close pair fails the check.

2. **Charge neutrality.** The composition must admit a charge-neutral oxidation-state assignment under pymatgen's `oxi_state_guesses`, using ICSD oxidation states (Ong et al., 2013). The ICSD list only restricts which integer states may be combined; the test itself is a global neutrality search, and any non-empty set of assignments is accepted (unweighted). Per-element oxidation bounds are available in the implementation but disabled in all reported runs.

Invalid structures receive the fixed penalty r_{inv} (equation 8) and skip stability and novelty evaluation. They are still embedded and kept in the group when within-group diversity (equation 34) is computed: an invalid structure can lower a valid sibling's diversity score, even though its own reward never reflects that term. Treating validity as a gate rather than as a fifth weighted objective is deliberate: as a weighted term it could be traded away. The magnitude of r_{inv} still controls how strongly invalid actions are discouraged; we sweep it in Appendix F.2.

D.3 Stability Score

For valid structures we score thermodynamic viability from the energy of the relaxed cell above the convex hull (Jain et al., 2013),

$$s_{\text{stab}} \;=\; 1 - \text{clip}(E_{\text{hull}},\, 0,\, 1). \tag{32}$$

E_{hull} is smaller for more stable crystals, so the score inverts it and places stability on the same higher-is-better $[0, 1]$ scale as novelty and diversity. Thus $s_{\text{stab}} = 1$ on the hull and falls to 0 for $E_{\text{hull}} \geq 1$ eV/atom. Sharing one bounded scale between the objectives is what makes the Dirichlet weights $\boldsymbol{w}$ meaningful: the scalarization $\boldsymbol{m}^\top \boldsymbol{w}$ is then a convex combination of comparable terms, so no objective can dominate through magnitude alone and the group-relative advantages stay on a common scale. Given the relaxed structure, equation 32 is deterministic; this section is the path from that cell to E_{hull}.

From a generated cell to E_{hull}. Each valid structure passes through four steps, all using the MatterSim potential of Appendix D.8.

1. **Relaxation.** The atomic positions are relaxed with the ASE BFGS optimizer (Larsen et al., 2017) to a force tolerance $f_{\max} = 0.02$ eV/Å, capped at a step budget of 100. The lattice is held fixed at the backbone output. Relaxations that reach the cap are not discarded, and the potential energy at the final step is taken as the structure's energy E_p. At this budget, most relaxations of ANCHOR structures are truncated (Appendix E.1.5), so the stability term is a ranking signal within a group rather than a converged E_{hull}.
2. **MP2020-compatible entry.** The relaxed energy E_{p} and the cell's composition are wrapped as a computed entry and passed through `MaterialsProject2020Compatibility` (Ong et al., 2013). This places the MatterSim energy on the same corrected scale as the reference entries, so structure and hull are directly comparable.
3. **Reference convex hull.** For the structure's chemical system we gather the reference entries of that system and all of its subsystems from MatterGen's Alex-MP reference dataset (Zeni et al., 2025), which combines Materials Project and Alexandria (Schmidt et al., 2024) entries, and build a pymatgen `PhaseDiagram`. The reference entries carry the same MP2020 compatibility correction. Note that this is a different and far larger corpus than the MP-20 index used for novelty (Appendix D.4). Phase diagrams are cached per chemical system and reused across the run.
4. **Energy above convex hull.** E_{hull} (eV/atom) signifies the distance of the structure's E_{p} above the reference convex hull. We retain its sign, so a structure on or below the hull is not clipped to zero before equation 32.

D.4 CONTINUOUS ADAPTIVE NOVELTY AND THE COUNT BONUS

Our novelty reward, which we refer to throughout as continuous adaptive novelty (CAN), makes two design choices, each isolated by an ablation in Appendix F.3. It is *continuous*: instead of a binary novel/not-novel flag, every structure receives a graded score in $[0, 1]$ from a kNN–RBF comparison in descriptor space (equation 6–equation 7). And it is *adaptive*: the reference is not the frozen MP-20 corpus alone but a growing index of the agent's own discoveries, so rediscovering a structure the policy already found is penalized. This section specifies the pieces that implement both; the RDF descriptor, the frozen MP-20 reference index and the online agent-history index. Novelty is computed only for valid relaxed structures (Appendix D.2).

RDF embedding. Given a relaxed structure x, let d_{ab} denote the minimum-image distance between sites a and b. We collect all pairwise distances with $0 < d_{ab} \leq r_{\max}$, histogram them into n_{bins} equal-width bins on $(0, r_{\max}]$, and ℓ_1-normalize, writing h_b for the number of pairs that fall in bin b:

$$z_b = \frac{h_b}{\sum_{b'} h_{b'}}, \qquad \boldsymbol{z}(x) = (z_1, \ldots, z_{n_{\text{bins}}}) \in \mathbb{R}^{n_{\text{bins}}}. \tag{33}$$

Normalization reduces sensitivity to unit-cell size. The same $\boldsymbol{z}$ is reused for within-group structural diversity (equation 34).

MP-20 reference index. Before training, every MP-20 structure is bucketed by reduced formula f and converted to $\boldsymbol{z}(x)$. We write Z^{ref} for the embeddings of the formula currently being scored, and $\mathcal{F}_{\text{ref}}$ for the reduced formulas that have at least one such embedding. At runtime the index is loaded from a prebuilt cache; it is never updated during RL.

kNN–RBF scoring. Novelty comparisons are *within-formula*: a candidate $\boldsymbol{z}$ is scored only against reference embeddings sharing its reduced formula f. This isolates structural novelty within a composition family; exploration *across* families is handled instead by the composition policy and the count bonus. In equation 6, $\text{kNN}_{k_{nn}}(\boldsymbol{z}, Z)$ is the set of indices of the $\min(k_{nn}, |Z|)$ embeddings in Z with smallest ℓ_2 distance to $\boldsymbol{z}$, so $\mathcal{J}$ is truncated when a formula has fewer than k_{nn} references and the average is over that (possibly smaller) set. The RBF kernel converts distances in RDF space into graded similarity, and averaging over the neighbourhood reduces sensitivity to a single nearest reference while remaining a local measure of structural similarity.

Bandwidth σ. For the MP-20 branch we use a precomputed per-formula bandwidth σ_f: the median pairwise distance among embeddings in Z^{ref} (median heuristic). For the agent-history branch, σ is computed on-the-fly from the current Z^{hist}. Both paths then apply a global floor $\sigma \leftarrow \max(\sigma, \sigma_{\text{floor}})$. The floor is the 10th percentile of leave-one-out, within-formula k-NN distances on MP-20 itself (how close known polymorphs of the same formula already sit in RDF space), computed once offline. Without it, a family with two near-identical references would drive σ to zero and the RBF would collapse to a δ: novelty 0 on an exact match and 1 otherwise.

Agent-history index. The two indices ask different questions. The fixed index asks whether a structure differs from known data; the growing index asks whether it differs from what the search itself has already found. During training the agent maintains Z^{hist}, a FIFO buffer of RDF embeddings keyed by reduced formula. After each novelty evaluation, $\boldsymbol{z}(x)$ is appended, so a candidate is always scored against the history as it stood before that candidate was generated and never against itself; when the buffer exceeds a per-formula cap C_{hist} the oldest entry is evicted.

Scoring procedure. Algorithmically, novelty evaluation follows the rule in equation 7:

1. Compute $\boldsymbol{z}(x)$ and look up Z^{ref} and Z^{hist}.
2. If $f \in \mathcal{F}_{\text{ref}}$: score against MP-20; if $|Z^{\text{hist}}| > 0$, also score against agent history and take the minimum (anti-rediscovery for known chemistries).
3. If $f \notin \mathcal{F}_{\text{ref}}$ and $|Z^{\text{hist}}| = 0$: return the bootstrap default $\nu_0 = 0.75$.
4. If $f \notin \mathcal{F}_{\text{ref}}$ and $|Z^{\text{hist}}| > 0$: score by kNN against Z^{hist}.
5. Append $\boldsymbol{z}(x)$ to Z^{hist}.

The bootstrap $\nu_0 = 0.75$ is set below the ceiling of 1 on purpose. It is a placeholder returned before any same-formula reference exists, not a claim that the structure is maximally novel; leaving headroom above it means that once references for that formula accumulate, a genuinely distinct structure can be scored *higher* than the first-seen bootstrap by the kNN–RBF metric. A value of 1 would saturate the term on the first sample and flatten within-formula novelty thereafter.

Hyperparameters. Table 6 summarizes the novelty configuration used in the main runs.

Table 6: Novelty hyperparameters (main runs).

Symbol / parameter	Value	Role
$r_{\max}$	7 Å	RDF distance cutoff
n_{bins}	80	RDF histogram resolution
k_{cfg}	10	kNN neighbourhood size
ν_0	0.75	Bootstrap for first off-MP-20 formula
C_{hist}	20	FIFO cap per reduced formula
σ_{floor}	10th pct. calibration	Minimum RBF bandwidth

Count Bonus and the Absence of a Uniqueness Term The count bonus in equation 8 adds

$$\frac{\lambda_c}{\sqrt{\max(n_f,\,1)}}$$

to the reward of every valid action, where n_f is the number of times reduced formula f has been generated as a valid structure so far in the current run, including the current group. The inverse-square-root dependence gives a large incentive for early visits to a formula and progressively reduces the bonus as that formula becomes familiar.

We deliberately do not add a fifth, explicit uniqueness term. The count bonus only sees how often a formula has been *proposed*; whether the crystal itself is new is the job of CAN, so a separate uniqueness objective would charge the same repeat twice. The two mechanisms also act at different levels: the count bonus discourages repeatedly allocating search effort to the same reduced formula, while CAN discourages rediscovery of the same or closely related structures *within* a formula family. Both also persist for the whole run, whereas uniqueness is conventionally measured as repetition within a single generated group (Zeni et al., 2025; Betala et al., 2025).

D.5 Within-group Diversity Kernels

A group that contains only one structural or chemical motif provides little useful ranking signal for a group-relative update. We therefore reward candidates that differ from the other $G-1$ samples in the same rollout,

$$d_\bullet(x_i) = \text{clip}\Big(1 - \tfrac{1}{G-1}\sum_{j\neq i} K_\bullet(x_i, x_j),\ 0,\ 1\Big), \qquad \bullet \in \{\text{struct}, \text{comp}\}, \tag{34}$$

where K_{struct} is an RBF kernel on the RDF embeddings $\boldsymbol{z}$ and K_{comp} is the Jaccard similarity of chemical-role sets. The clip is a numerical guard, since both kernels already return similarities in $[0, 1]$, which places the two terms on the same scale as the other objectives. Writing roles_i for the set of chemical families present in action c_i,

$$K_{\text{comp}}(x_i, x_j) \;=\; \frac{|\text{roles}_i \cap \text{roles}_j|}{|\text{roles}_i \cup \text{roles}_j|}.$$

These families partition the vocabulary of Appendix D.1 into 11 classes; alkali, alkaline earth, early transition metal, late transition metal, lanthanide, actinide, p-block metal/metalloid, chalcogen, halogen, pnictogen, and a residual class. The partition is deliberately coarse enough that a $k \in \{2, 3, 4\}$ action maps to 2–4 families, which keeps this similarity discriminative at every admissible k. Comparing role sets rather than raw element sets means two actions that differ only by substituting Na

for K are treated as compositionally similar, which is the behaviour we want from a term whose job is to prevent a group from collapsing onto one chemical motif.

Both terms are computed once per rollout, after every sample in the group has been individually scored for stability and novelty. Unlike novelty (equation 7), they compare a structure only against its $G-1$ siblings and carry no memory across steps. The diversity objectives therefore serve to limit within-rollout redundancy rather than the adaptive novelty's long-horizon role to prevent rediscovery.

D.6 Policy Parametrization and Optimization Details

Composition action space and policy parameterization Throughout this section $\mathcal{E}$ is the element vocabulary of size $E = |\mathcal{E}|$, N the maximum number of atoms per cell, and $\mathcal{K} = \{2, 3, 4\}$ the admissible number of distinct elements; their values are given in Appendix D.1 and all symbols in Appendix A. The composition action $c = (k, \boldsymbol{e}, T, \rho)$ of equation 5 deliberately carries no structural degrees of freedom: once the reduced formula $f(c)$ is fixed, atomic coordinates and lattice parameters are generated entirely by the frozen CSP backbone, so the reinforcement-learning problem acts on the chemical query supplied to that model.

The policy factorizes the composition action autoregressively in the order in which its components are sampled,

$$\pi_\theta(c \mid \boldsymbol{s}) \;=\; \underbrace{p_k(k \mid \boldsymbol{s})}_{\text{cardinality}} \prod_{i=1}^{k} \underbrace{p_e(e_i \mid e_{<i}, \boldsymbol{s})}_{\text{elements}} \underbrace{p_T(T \mid \boldsymbol{e}, \boldsymbol{s})}_{\text{cell size}} \underbrace{p_\rho(\rho \mid T, \boldsymbol{e}, \boldsymbol{s})}_{\text{stoichiometry}}. \tag{35}$$

Each factor is implemented as a masked categorical head, with the corresponding multinomial construction used for ρ. The cell-size head chooses T before the stoichiometry is written, so the policy knows how many atoms must be distributed across the selected elements. The stoichiometry mask ensures that every selected element appears at least once.

The rollout state $\boldsymbol{s}$ carries both the objective weighting $\boldsymbol{w}$ and a size-invariant summary of the element sub-group $\mathcal{S}$ active for the rollout; its concrete form is given in equation 41. Element draws are restricted to $\mathcal{S}$ by masking, so $p_e(e_i \mid e_{<i}, \boldsymbol{s}) = 0$ for every $e_i \notin \mathcal{S}$ (equation 38). The element vocabulary, maximum cell size, categorical masks, and resulting size of the discrete action space are fixed before training.

Element sub-group sampling. Each rollout restricts element choice to a random sub-group $\mathcal{S} \subseteq \mathcal{E}$, drawn by first sampling its size and then the sub-group uniformly among all sets of that size,

$$n_{\mathcal{S}} \sim \mathrm{Uniform}\{\, k_{\max}, \ldots, E \,\}, \qquad \mathcal{S} \sim \mathrm{Uniform}\big(\{\, \mathcal{S}' \subseteq \mathcal{E} : |\mathcal{S}'| = n_{\mathcal{S}} \,\}\big), \tag{36}$$

after which all element draws in equation 35 are masked to $\mathcal{S}$. The lower bound $k_{\max} = 4$ guarantees that every sub-group can support the largest admissible number of distinct elements.

Per-head distributions. Each factor of the policy in equation 35 is a small MLP head conditioned on the state vector $\boldsymbol{s}$ (equation 41). Logits outside an admissible set A are set to $-\infty$ before the softmax; we write this $\mathrm{mask}(g, A)$.

$$k \sim \mathrm{Categorical}\big(\mathrm{softmax}\, g_k(\boldsymbol{s})\big), \tag{37}$$

$$e_i \mid e_{<i} \sim \mathrm{Categorical}\big(\mathrm{softmax}\, \mathrm{mask}(g_e(\boldsymbol{s}, e_{<i}),\, \mathcal{E}_i)\big), \qquad \mathcal{E}_i = \mathcal{S} \setminus \{e_{<i}\}, \tag{38}$$

$$T \sim \mathrm{Categorical}\big(\mathrm{softmax}\, \mathrm{mask}(g_T(\boldsymbol{s}, \boldsymbol{e}),\, \{k, \ldots, N\})\big), \tag{39}$$

$$\tilde{\rho} \sim \mathrm{Multinomial}\big(T - k,\ \mathrm{softmax}\, g_\rho(\boldsymbol{s}, \boldsymbol{e})\big), \qquad \rho = \mathbf{1}_k + \tilde{\rho}. \tag{40}$$

These four draws are the components of a composition action c in equation 5, in that order. Elements are drawn without replacement from the active sub-group $\mathcal{S}$; $T \in \{k, \ldots, N\}$ is sampled as its own categorical; and equation 40 distributes the $T - k$ atoms over the k chosen elements. Only $T - k$ atoms are to be distributed since we pre-distribute one atom for each element; ensuring each sampled element appears at least once in the composition.

State vector. The head inputs concatenate a size-invariant summary of the active sub-group $\mathcal{S}$ with the objective weights $\boldsymbol{w}$,

$$\boldsymbol{s} \;=\; \big[\, \boldsymbol{W}_f\, \bar{\boldsymbol{f}}_{\mathcal{S}} \;;\; \boldsymbol{w} \,\big], \qquad \bar{\boldsymbol{f}}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{e \in \mathcal{S}} \boldsymbol{f}_e, \tag{41}$$

where $\boldsymbol{W}_f$ is a learned linear projection and $\boldsymbol{f}_e \in \mathbb{R}^{38}$ is a fixed (non-trained) descriptor of element e,

$$\begin{aligned} \boldsymbol{f}_e = \big[&\mathrm{onehot}(\mathrm{period}_e);\ \mathrm{onehot}(\mathrm{group}_e);\ \mathrm{onehot}(\mathrm{block}_e); \\ &\tfrac{\phi_s}{2},\ \tfrac{\phi_p}{6},\ \tfrac{\phi_d}{10},\ \tfrac{\phi_f}{14};\ \tilde{\chi}_e,\ \tilde{r}_e,\ \tilde{o}_e^{\min},\ \tilde{o}_e^{\max},\ \tilde{o}_e^{\mathrm{mean}} \big], \end{aligned} \tag{42}$$

which concatenates one-hot period ($\mathbb{R}^7$), group ($\mathbb{R}^{18}$) and block ($\mathbb{R}^4$) encodings, the capacity-normalized valence-shell occupancies $(\phi_s, \phi_p, \phi_d, \phi_f)$, and the min–max-normalized Pauling electronegativity $\tilde{\chi}_e$, atomic radius $\tilde{r}_e$, and common-oxidation-state min, max and mean. The choice of period, group, block and valence-shell occupancies as the element descriptor follows the subatomic token of Crystalite (Veljković et al., 2026), though we keep these as one-hot and normalized channels rather than their PCA-compressed continuous token; the five scalar channels and the mean-pooling over $\mathcal{S}$ in equation 41 are ours.

Log-probability and entropy. The log-probability of an action and the policy entropy each decompose as a sum over the four heads,

$$\log \pi_\theta(c \mid \boldsymbol{s}) = \log p_k + \textstyle\sum_{i=1}^{k} \log p_e^{(i)} + \log p_T + \log p_\rho, \tag{43}$$

$$H[\pi_\theta](\boldsymbol{s}) = H_k + \textstyle\sum_{i=1}^{k} H_e^{(i)} + H_T + H_\rho, \qquad H_\rho = (T-k)\, H\big(\mathrm{softmax}\, g_\rho\big). \tag{44}$$

The stoichiometry head draws the $T-k$ surplus atoms from a single categorical $\mathrm{softmax}\, g_\rho$ (equation 40). Its log-probability (equation 43) is the exact multinomial log-probability, and its entropy contribution is the summed entropy of those $T-k$ independent trials, $H_\rho = (T-k)\, H(\mathrm{softmax}\, g_\rho)$, a closed-form upper bound on the exact multinomial entropy.

Multi-objective rollout conditioning. Each rollout samples objective weights $\boldsymbol{w} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$ together with an element sub-group $\mathcal{S}$ (equation 36). Both are held fixed across all G actions of the rollout and are supplied to the policy through the state $\boldsymbol{s}$ (equation 41). Sampling $\boldsymbol{w}$ avoids committing to one scalarization in advance, which would otherwise be a weighting to tune. Because $\boldsymbol{w}$ enters the conditioning state, a single set of parameters covers the whole weight family rather than one policy per scalarization. All evaluations fix $\boldsymbol{w}$ to the Dirichlet mean. Sampling $\mathcal{S}$ restricts each rollout to a random chemical subspace, forcing the policy to keep practicing chemistries that could otherwise disappear from the sampled distribution as training becomes more exploitative. Sharing both variables within a group matters for the update: every action entering a group-relative comparison is then evaluated under the same objective and the same available element set.

Group-relative advantages. Rewards are converted into advantages using the statistics of the sampled group as the baseline, so no learned value function is required (Shao et al., 2024),

$$A_i \;=\; \frac{r_i - \bar{r}}{\max\big(\mathrm{std}(r),\ \epsilon_A\big)}, \qquad \bar{r} = \frac{1}{G} \sum_{j=1}^{G} r_j. \tag{45}$$

The advantage measures how well an action performs relative to the alternative compositions sampled under the same rollout context. Sharing $\boldsymbol{w}$ and $\mathcal{S}$ within the group is what makes this meaningful: it prevents the normalization from comparing candidates evaluated under different objective trade-offs or different chemical constraints.

Clipped policy update. The policy is updated over K epochs using the clipped PPO surrogate (Schulman et al., 2017)

$$\mathcal{L}^{\mathrm{clip}}(\theta) = \mathbb{E}_i\big[\min\big(\eta_i A_i,\ \mathrm{clip}(\eta_i, 1-\epsilon, 1+\epsilon) A_i\big)\big], \qquad \eta_i = \frac{\pi_\theta(c_i \mid \boldsymbol{s})}{\pi_{\theta_{\mathrm{old}}}(c_i \mid \boldsymbol{s})}. \tag{46}$$

Only the composition policy is updated. No gradient is propagated through the frozen CSP model (unless stated otherwise), the relaxation procedure, or the reward evaluator.

Full objective. The loss we minimize is the negative clipped surrogate (equation 46) with an entropy bonus:

$$\mathcal{L}(\theta) = -\,\mathcal{L}^{\text{clip}}(\theta) \;-\; \beta_H\, \mathbb{E}_i\big[H[\pi_\theta]\big]. \tag{47}$$

Entropy schedule. The coefficient β_H follows a warmup/plateau/cosine schedule, modulated by a proportional controller that nudges the normalized entropy $H[\pi_\theta]/H_{\max}$ toward a target fraction of its maximum. We take $H_{\max}$ as the largest per-sample entropy the chain equation 44 can attain, with every head at its uniform maximum,

$$H_{\max} = \log|\mathcal{K}| + \max_{k\in\mathcal{K}}\Big[\textstyle\sum_{i=0}^{k-1}\log(E-i) \;+\; \log(N-k+1) \;+\; (N-k)\log k\Big], \tag{48}$$

where $\log|\mathcal{K}|$ bounds the cardinality head and the three bracketed terms bound the k without-replacement element draws, the cell-size head H_T, and the count head H_c (at $T = N$). The element bound uses the full vocabulary E rather than a per-rollout sub-group $\mathcal{S}$, so $H_{\max}$ is a single constant for the whole run and the controller targets the same absolute entropy scale at every step, whichever sub-group was sampled. The schedule is keyed to cumulative structures generated (with G per step) rather than to optimizer steps, so runs with different group sizes G see the same entropy pressure at the same sample budget; this is what makes the comparison across G in Appendix F.1 a fair one.

D.7 Training Algorithm

Algorithm 1 gives the full ANCHOR training procedure. Each optimization step has three phases. The *rollout* phase draws one context (a preference vector $\boldsymbol{w}$ and an element sub-group $\mathcal{S}$) and samples G compositions from the current policy under it. Each composition is decoded by the frozen backbone $\mathcal{G}$ and relaxed. The *scoring* phase computes the reward of each structure and converts the group's rewards into group-relative advantages. The *update* phase performs K clipped policy-gradient epochs on the stored rollout. Only the update phase changes θ; no gradient passes through $\mathcal{G}$, the relaxer, or the reward.

D.8 Training Setup and Hyperparameters

Backbone and training potential. Every policy trained with KLDM uses a single KLDM CSP checkpoint as the frozen structure-prediction backbone, sampled with a 50-step Euler–Maruyama solver, one structure per formula. Generated cells are relaxed under BFGS with the training potential MatterSim-1M, over atomic positions only and for at most 100 steps (Appendix D.3).

Two MatterSim variants appear in the paper and it matters which is which: the *training potential*, which supplies the stability term of the reward inside the training loop, and the *evaluation potential*, which scores final structures after training (Appendix E.1). The training potential is a base-model-selection variable; the evaluation potential is fixed to MatterSim-5M for every internal table. LeMat-GenBench and MatterGen evaluation submissions are scored by their respective default pipelines, not by our MatterSim-5M (Appendix E.2 and E.3).

Hyperparameters. Table 7 lists every value needed to reproduce a main model run. Entries marked "swept" are varied in the corresponding appendix experiment and kept at the listed default elsewhere.

Training budget. Because the group size G is one of the variables we compare, wall-clock and optimizer steps are not comparable across runs; the number of structures generated is. The entropy schedule is keyed to a horizon of 10^5 structures (Appendix D.6). Jobs were stopped by a common seven-day wall-clock, so the number actually consumed is $G\times$ (final step) and is not exactly 10^5. Table 8 is that count for every trained policy in the paper. Collapsed cells consumed more, not less, because repeating small cells makes each BFGS relaxation cheap (Appendix F.3).

Compute. Each reinforcement-learning run occupies a single NVIDIA A100-40GB under a seven-day wall, so one run is 168 GPU-hours. Every such run hit that wall, which is why runs are compared here by structures generated (Table 8). Within that wall the selected configuration generates $\sim$123k structures and most other cells 102–131k. The collapsed and the LeMat-GenBench

Algorithm 1 ANCHOR: group-relative policy optimization over compositions, with a frozen CSP backbone.

Require: frozen backbone $\mathcal{G}$ and relaxer; frozen MP-20 index Z^{ref}; hyperparameters (Table 7)

$\theta \leftarrow$ initial policy parameters; $\quad Z^{\mathrm{hist}} \leftarrow \emptyset$; $\quad n_f \leftarrow 0$ for all formulas f
for step $= 1, 2, \ldots$ **do**
  $\theta_{\mathrm{old}} \leftarrow \theta$
  Rollout (no gradients)
  $\boldsymbol{w} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$; $\quad \mathcal{S} \leftarrow$ SAMPLESUBGROUP($\mathcal{E}$); $\quad \boldsymbol{s} \leftarrow [\, \boldsymbol{W}_f \bar{\boldsymbol{f}}_{\mathcal{S}} \,;\, \boldsymbol{w} \,]$ ▷ (36, 41)
  for $i = 1$ **to** G **do**
    $c_i \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid \boldsymbol{s})$; $\quad x_i \leftarrow$ RELAX($\mathcal{G}(f(c_i))$) ▷ (5, 35)
  end for
  Scoring
  $(r_1, \ldots, r_G) \leftarrow$ SCOREGROUP($\{c_i\}, \{x_i\}, \boldsymbol{w}$)
  $A_i \leftarrow \big(r_i - \bar{r}\big) \big/ \max\big(\mathrm{std}(r), \epsilon_A\big)$ for $i = 1, \ldots, G$ ▷ (45)
  Update
  for epoch $= 1$ **to** K **do**
    $\mathcal{L} \leftarrow -\,\mathcal{L}^{\mathrm{clip}}(\theta) \;-\; \beta_H\, \mathbb{E}_i\big[H[\pi_\theta(\cdot \mid \boldsymbol{s})]\big]$ ▷ (46, 47)
    $\theta \leftarrow$ ADAMSTEP(θ, $\nabla_\theta \mathcal{L}$) ▷ gradients reach π_θ only
  end for
end for
return trained policy π_θ

procedure SCOREGROUP($\{c_i\}_{i=1}^{G}$, $\{x_i\}_{i=1}^{G}$, $\boldsymbol{w}$)
  for $i = 1$ **to** G **do** ▷ in sampling order
    if x_i fails the validity gate **then** ▷ (Appendix D.2)
      continue
    end if
    $s_{\mathrm{stab},i} \leftarrow 1 - \mathrm{clip}\big(E_{\mathrm{hull}}(x_i),\, 0,\, 1\big)$ ▷ (32)
    $n_{f(c_i)} \leftarrow n_{f(c_i)} + 1$
    $\nu_i \leftarrow \nu(x_i)$; $\quad$ append $\boldsymbol{z}(x_i)$ to Z^{hist} ▷ (7)
  end for
  $\big(d_{\mathrm{struct},i},\, d_{\mathrm{comp},i}\big)_{i=1}^{G} \leftarrow$ diversity over all G structures, valid or not ▷ (34)
  for $i = 1$ **to** G **do** ▷ scalarize
    if x_i valid **then**
      $\boldsymbol{m}_i \leftarrow \big(s_{\mathrm{stab},i},\, \nu_i,\, d_{\mathrm{struct},i},\, d_{\mathrm{comp},i}\big)$
      $r_i \leftarrow \boldsymbol{m}_i^{\top} \boldsymbol{w} + \lambda_c \big/ \sqrt{n_{f(c_i)}}$ ▷ (8)
    else
      $r_i \leftarrow r_{\mathrm{inv}}$
    end if
  end for
  return $(r_1, \ldots, r_G)$
end procedure

gate cells generate more structures because repeated small cells make each BFGS cheap. The cost is backbone sampling plus BFGS of every generated cell; the policy update is a few small MLP heads and is negligible beside that.

Counting unique training jobs, Table 8 holds 19 reinforcement-learning runs: six base-model cells, two further invalid-penalty cells, eight further exploration cells (the selected G=32 / MatterSim-1M cell is also the $r_{\mathrm{inv}} = -0.2$ run and the full method), one validity-gate cell, and two trainable KLDM-CSP runs, or $19 \times 168 = 3{,}192$ GPU-hours.

Likelihood training sits outside that loop and adds about 350 GPU-hours. KLDM releases CSP weights only, with no DNG checkpoint, so the KLDM-DNG baseline is one we trained ourselves (46). On top of that come four CSP fine-tunes, which continue training KLDM-CSP and Crystalite-CSP on MP-20 alone and on MP-20 augmented with ANCHOR structures (44); four DNG reward fine-tunes, KLDM-DNG and Crystalite-DNG under a mean and a balanced weighting ($4 \times 48 = 192$); and generating the ANCHOR corpus that the augmented arms train on (61).

Table 7: Hyperparameters for main runs.

Parameter	Value	Note
Action space		
E	84	Element vocabulary (Appendix D.1)
$\mathcal{K}$	$\{2, 3, 4\}$	Distinct elements per action
N	20	Max atoms per cell
max atoms / element	12	
Policy network		
state dim	32	Projected sub-group summary
hidden dim	256	Per-head MLP width
element embedding	32	
GRPO		
G	32	Group size; swept in Appendix F.1
K	2	Update epochs per step
ϵ	0.2	PPO clip
ϵ_A	0.05	Advantage std floor (45)
optimizer	Adam, learning rate 10^{-4}	
budget	10^5 structures	Entropy-schedule horizon; see below
Entropy schedule		
β_H base / min / max	7.5×10^{-3} / 2×10^{-3} / 2.5×10^{-2}	
target fraction	0.85	Of $H_{\max}$
deadband / gain	0.03 / 3.0	Controller
warmup / warm start	0.02 / 0.5	Fraction of horizon; fraction of base at step 0
decay start	1.0	No cosine decay within the budget
Reward		
$\boldsymbol{\alpha}$ (Dirichlet)	$(1, 4, 1, 1)$	Order: stability, novelty, d_{struct}, d_{comp}
λ_{c}	1.0	Count bonus; switched on/off in Appendix F.3
r_{inv}	-0.2	-0.2 in base-model selection; $\{-0.1, -0.5\}$ in the invalid-penalty sweep (Appendix F.2)
Evaluator		
bond-length tolerance	0.2	Fractional
fallback min distance	1.2 Å	Unlisted element pairs
oxidation states	ICSD	pymatgen oxidation-state mode (Appendix D.2)

Evaluating three seeds of N_{eval}=1000 for every main-text row adds about 540. The single-seed appendix protocols add about 80 A100-hours, plus about 70 RTX 3090-24GB GPU-hours for LeMat-GenBench. The total is roughly 4,200 A100 GPU-hours, excluding unreported exploratory runs.

Table 8: Structures generated by each trained policy, $G\times$ (final checkpoint step). All jobs are a single A100 and a seven-day wall. The entropy-schedule horizon is 10^5; this table is what was actually consumed. The selected base model is G=32 / MatterSim-1M.

Experiment	Configuration	G	Step	Structures
Base-model sweep	G=16 / 1M	16	7950	127,200
	G=16 / 5M	16	7800	124,800
	G=32 / 1M	32	3850	123,200
	G=32 / 5M	32	3850	123,200
	G=64 / 1M	64	2000	128,000
	G=64 / 5M	64	1750	112,000
Invalid penalty	$r_{\text{inv}} = -0.1$	32	3950	126,400
	$r_{\text{inv}} = -0.2$	32	3850	123,200
	$r_{\text{inv}} = -0.5$	32	3850	123,200
Exploration factorial	full method	32	3850	123,200
	– count bonus	32	4100	131,200
	– sub-group mask	32	3200	102,400
	– adaptive nov.	32	4000	128,000
	adaptive nov. only	32	3300	105,600
	count bonus only	32	3250	104,000
	sub-group mask only	32	7400	236,800
	none	32	20250	648,000
	binary novelty	32	4050	129,600
Validity gate	KLDM, LeMat-GenBench gate	32	6750	216,000
Trainable KLDM-CSP	ANCHOR shared rewards	32	3800	121,600
	Stability only	32	3850	123,200

E Evaluation Protocols

E.1 Internal Evaluation Protocol

Two relaxation protocols. Every internal number in this paper uses MatterSim-5M and BFGS for at most 400 steps. The protocols differ only in whether the lattice is relaxed. All internal numbers in the main text (Tables 1–2) relax positions and lattice. The lattice-fixed appendix tables (Tables 9, 10, 17,18, 19, 20 and 21) hold the lattice fixed at the backbone output, and their captions are marked *Lattice fixed*. Neither is the training reward, which uses MatterSim-1M and 100 steps with the lattice fixed (section D.2 and D.3). The same checkpoint therefore has different values in the main text and in this appendix: the selected policy reaches 38.5% MSUN with the lattice fixed and 47.6% with it free. Tables 1 and 20 give both. The MatterGen and LeMat-GenBench tables in Appendix E.2 and Appendix E.3 use those pipelines' own relaxation and are unaffected. Statistics logged during training measure a policy along the trajectory it happened to take, which makes them path-dependent and unsuitable for comparing final policies. Every headline number in this paper is therefore produced after training, from fresh samples. We load a checkpoint, sample new actions, generate and relax the corresponding structures, and score them with a potential that is the same for every policy being compared.

Sampling. Each policy generates $N_{\text{eval}} = 1000$ structures per sampling seed. The lattice-free tables of the main text report mean and standard deviation over three sampling seeds, and the MatterGen and LeMat-GenBench appendix tables use seed 0. The lattice-fixed appendix tables use seed 0 alone and serve as a check on the protocol. Actions are sampled stochastically rather than by `argmax`, since a deterministic policy would collapse to one composition and the quantity of interest is the distribution the policy induces. The objective weights are fixed to the mean of the training Dirichlet, $\boldsymbol{w} = \mathbb{E}[\text{Dirichlet}(\boldsymbol{\alpha})]$, and the sub-group mask is disabled so the policy may use the full 84-element vocabulary.

Backbone transfer. Unless stated otherwise, sampled compositions are realized by the same frozen KLDM backbone used during training. In the backbone-transfer rows of section 4 the pol-

icy checkpoint and the composition set it produces are held fixed and only the structure model is replaced, querying a different frozen CSP model through the same composition-to-structure interface. Neither component is fine-tuned. In the internal tables, these rows are scored by the same MatterSim-5M protocol as every other row. Under MatterGen and LeMat-GenBench they are scored by those pipelines.

Evaluation potential. Structures are relaxed with MatterSim-5M under BFGS for at most 400 steps, and E_{hull} is computed from the resulting energies against the Alex-MP hull (Appendix D.3). We distinguish two roles for the machine-learned interatomic potential throughout the paper: the *training potential* supplies the stability term of the reward during a run, whereas the *evaluation potential* used here scores the final structures. The evaluation potential is fixed to MatterSim-5M for every policy, including policies trained against the smaller 1M variant. Without this, a policy trained against the weaker potential would be graded by a weaker judge than one trained against the stronger, and the base-model comparison in Appendix F.1 would confound model quality with grader leniency. The relaxation budget is larger than the budget used in training because stability is being reported here rather than used as a learning signal. Main-text internal tables additionally relax the lattice. The internal appendix tables reported below hold it fixed (Appendix E.1.4). These choices define the internal tables. Structures submitted to MatterGen's evaluation pipeline or to LeMat-GenBench are scored by those pipelines (Appendix E.2, Appendix E.3).

Metric definitions. We report the following, all as percentages of the N_{eval} generated structures:

- **Valid**: passes the gate of Appendix D.2.
- **Unique**: distinct reduced formulas, $|\{f(a_i)\}|/N$. This is deliberately composition-level rather than structure-level and therefore does not depend on the choice of unit-cell representation. Structural novelty within a formula is handled separately by the continuous CAN score.
- **Novel**: valid and $\nu(x) > 0.5$. The reported rates of ANCHOR variants are insensitive to this threshold (Appendix F.1).
- **Stable**: valid and $E_{\text{hull}} \leq 0$.
- **Meta**: valid and $E_{\text{hull}} \leq 0.1$ eV/atom.
- **SUN**: stable $\wedge$ unique $\wedge$ novel.
- **MSUN**: meta $\wedge$ unique $\wedge$ novel.
- **MP-20 hits**: number of structures whose reduced formula occurs in MP-20.

The validity gate and the novelty score are applied to the relaxed structure under both relaxation protocols.

Novelty at evaluation time. At evaluation, we switch off the adaptive part of CAN. The agent-history index is neither scored against nor updated, so novelty is the graded kNN-RBF score against the frozen MP-20 index alone. Keeping the history live would usually change little, because 1000 samples from such a large action space rarely repeat. It matters for policies that collapse, however. On the ablation arm without any exploration mechanism, which repeats a handful of formulas, the same structures are 99.9% novel against MP-20 alone but only 0.9% novel once the search history is included. A growing reference is a training signal, so a metric computed against it would depend on draw order and would charge a second time for repetition that Unique already reports. Across the $5{,}737$ valid structures of the base-model selection grid, $5{,}731$ (99.9%) received the bootstrap $\nu_0 = 0.75$ because their reduced formula is absent from MP-20; only 6 were scored by the kNN–RBF machinery. Novel% in our tables is therefore almost always a statement that the formula does not occur in MP-20. The graded score and the growing index are exercised during training (Appendix D.4) and isolated by the ablation (Appendix F.3); they are not what Novel% at model inference reports.

E.1.1 Shared rows and reading conventions

The internal, MatterGen, and LeMat-GenBench tables in this appendix score the *same* set of rows with different evaluators, so any row can be read across pipelines.

Table 9: **Lattice fixed.** Internal evaluation pipeline (MatterSim-5M, BFGS-400). Valid = bonds + ICSD oxidation. Unique = distinct formulas / N. Novel = $\mathrm{novelty_score} > 0.5$. Stable / Meta: $E_{\mathrm{hull}} \leq 0$ / 0.1. The last column counts valid samples whose reduced formula occurs in the frozen MP-20 reference index (not ranked). Every ANCHOR row is 0 in the MP-20 column, the DNG rows are not: the novelty reward moves the policy off the reference set, whereas a finetuned generator re-emits it. Table 2 gives the lattice-free values of the headline rows; main-text tables relax the lattice.

Configuration	N	Valid	Unique	Novel	Meta	Stable	MSUN	SUN	MP-20
ANCHOR, inference-time backbone									
ANCHOR → KLDM-CSP	1000	97.9	**100.0**	97.9	**38.5**	**18.1**	**38.5**	**18.1**	0
ANCHOR → Crystalite-CSP	1000	94.1	**100.0**	94.1	37.0	13.7	37.0	13.7	0
ANCHOR → OMatG-CSP	998	**98.2**	**100.0**	**98.2**	<u>37.9</u>	<u>14.9</u>	<u>37.9</u>	<u>14.9</u>	0
ANCHOR trained with LeMat validity									
KLDM backbone	1000	5.6	**100.0**	5.6	0.7	0.0	0.7	0.0	1
Trainable KLDM-CSP									
ANCHOR reward	1000	<u>98.1</u>	**100.0**	<u>98.1</u>	15.0	4.2	15.0	4.2	0
Stability only	1000	<u>98.1</u>	**100.0**	<u>98.1</u>	19.4	7.4	19.4	7.4	0
DNG									
Crystalite-DNG	1000	49.0	98.7	33.9	22.5	1.5	11.4	0.6	163
Crystalite-DNG → CSP	1000	50.5	98.7	34.9	24.5	2.3	12.4	1.0	165
Crystalite-DNG FT (mean)	1000	96.4	40.1	32.3	37.2	0.9	8.0	0.4	617
Crystalite-DNG FT (bal.)	1000	96.4	50.0	37.4	36.9	0.6	10.3	0.2	594
KLDM-DNG	1000	44.7	98.9	34.5	11.4	0.9	5.6	0.5	124
KLDM-DNG → CSP	1000	43.7	98.9	32.6	12.4	0.9	5.1	0.4	123
KLDM-DNG FT (mean)	1000	53.9	99.4	45.0	11.1	0.3	6.3	0.1	107
KLDM-DNG FT (bal.)	1000	50.0	99.3	41.4	10.5	0.6	4.7	0.3	100

Reading the rows. ANCHOR → *X* is a single composition policy, trained against a frozen KLDM-CSP backbone and queried at inference through generator *X*. The *trained with LeMat validity* block instead changes the validity gate entering the reward during GRPO, with the same KLDM-CSP backbone, and is not a post-hoc swap. *Trainable KLDM-CSP* lets GRPO update the backbone as well, under two reward variants: the shared ANCHOR reward and the stability term alone. *DNG* models choose composition and structure themselves with no policy. FT marks DDPO reward fine-tuning of the generator, at either Dirichlet-mean objective weights (mean, novelty-heavy) or equal stability-novelty weights (bal.). *FT → CSP* rows query the frozen CSP model of the same family with the fine-tuned model's compositions, so they differ from the *FT* row only in the structure generator.

Reading the numbers. All entries are percentages of N unless stated otherwise. A `"–"` is a missing or non-final run, never a zero, and is never ranked. Best value in each column is **bold**, second-best <u>underlined</u>.

E.1.2 Internal results

Table 9 applies the protocol above to every row. The lattice-fixed evaluator preserves the same separation as the main-text cell-relaxed results. ANCHOR's composition policy transfers across all three frozen CSP backbones with 37.0-38.5% MSUN, whereas allowing the KLDM-CSP backbone to receive the reward lowers MSUN to 15.0% under the shared reward and 19.4% under stability alone. The DNG fine-tunes show the opposite failure mode: Crystalite validity rises to 96.4%, but formula uniqueness falls to 40.1-50.0% and 594-617 samples return to MP-20 formulas, leaving SUN at 0.2-0.4%. The conclusions therefore do not depend on lattice relaxation: freezing the structural prior preserves the part of the generator that already works, while reward adaptation of the joint generator can raise validity without improving discovery.

E.1.3 Sensitivity to the validity gate

Validity is the definition our protocol and the LeMat-GenBench leaderboard protocol disagree on most (Appendix E.3), and it gates every downstream count. Table 10 isolates that one choice: the

Table 10: **Lattice fixed.** Internal evaluation pipeline with the LeMat-GenBench validity gate (charge + distance + plausibility) substituted for our own. Relaxer, hull, novelty indices and column definitions are unchanged from Table 9; the alternative gate changes which chemistries are admitted rather than acting as a uniformly looser filter. The *trained with LeMat validity* block was trained against this gate, so its Valid and Novel entries are graded by their own training signal and are not comparable across blocks. Table 2 gives the lattice-free values of the headline rows. Tables in the main text relax the lattice.

Configuration	N	Valid	Unique	Novel	Meta	Stable	MSUN	SUN	MP-20
ANCHOR, inference-time backbone									
ANCHOR → KLDM-CSP	1000	72.9	**100.0**	72.9	25.0	**9.4**	25.0	**9.4**	0
ANCHOR → Crystalite-CSP	1000	72.7	**100.0**	72.7	26.0	8.4	26.0	8.4	0
ANCHOR → OMatG-CSP	998	72.8	**100.0**	72.8	26.2	7.9	26.2	7.9	0
ANCHOR trained with LeMat validity									
KLDM backbone	1000	98.3	**100.0**	97.5	9.2	0.0	8.7	0.0	10
Trainable KLDM-CSP									
ANCHOR reward	1000	86.5	**100.0**	86.5	17.7	4.8	17.7	4.8	0
Stability only	1000	81.5	**100.0**	81.5	16.1	4.6	16.1	4.6	0
DNG									
Crystalite-DNG	1000	93.2	98.7	64.4	43.6	2.0	21.6	0.8	311
Crystalite-DNG → CSP	1000	94.4	98.7	64.8	**47.3**	3.0	23.6	1.2	311
Crystalite-DNG FT (mean)	1000	96.8	40.1	31.7	37.5	1.1	8.2	0.5	625
Crystalite-DNG FT (bal.)	1000	96.4	50.0	36.8	37.8	0.8	11.0	0.3	596
KLDM-DNG	1000	93.2	98.9	73.9	26.0	1.1	13.5	0.5	232
KLDM-DNG → CSP	1000	90.4	98.9	70.1	24.9	1.0	10.5	0.4	231
KLDM-DNG FT (mean)	1000	91.7	99.4	75.9	19.0	0.6	9.9	0.3	189
KLDM-DNG FT (bal.)	1000	91.0	99.3	75.7	19.3	1.0	9.2	0.7	177

same samples, relaxer, hull and novelty index as Table 9, with the LeMat-GenBench validity gate substituted for ours. Differences between the two tables are attributable to the gate alone, not to the energy model or the reference hull.

This substitution is not a uniformly looser or stricter filter. Relative to Table 9, ANCHOR→KLDM-CSP validity falls from 97.9% to 72.9%, while Crystalite-DNG rises from 49.0% to 93.2%; the gate changes which chemistry is admitted rather than applying a common severity shift. The downstream metrics move accordingly: Crystalite-DNG→CSP rises from 12.4% to 23.6% MSUN, while ANCHOR→KLDM-CSP falls from 38.5% to 25.0%. The policy trained directly against the LeMat gate reaches 98.3% validity under that same gate but 0% SUN under the unchanged MatterSim hull, separating optimisation of the validity rule from discovery of on-hull structures.

E.1.4 Sensitivity to lattice relaxation

The main text protocol relaxes the lattice together with the atomic positions, following the cell-relaxing convention of the MatterGen pipeline (Appendix E.2). The training reward instead relaxes positions only, under MatterSim-1M and a 100-step budget (Appendix D.3). Tables 9 and 10 hold the lattice fixed and leave everything else unchanged (MatterSim-5M, BFGS, 400 steps), so they are not the training reward, which also differs in potential and budget. Holding the lattice fixed raises E_{hull} and lowers every stability-derived column, because a lattice taken directly from the backbone leaves structures strained. We report the free-cell numbers as the headline for that reason, and because free-cell relaxation is not the signal the policies were optimised against. Tables 2 and 9 place the two protocols side by side.

Conclusions across protocols. Freeing the lattice raises every stability-derived number but leaves the conclusions of the main text unchanged. In the exploration ablation (Tables 1 and 20), the full method has the highest SUN under both protocols (18.1% fixed, 22.1% free) and the collapsed cells are last under both. The full method and adaptive novelty alone swap order on MSUN (38.5% against 37.1% fixed, 47.6% against 52.0% free), while their SUN order holds under both. In Table 2 and its lattice-fixed counterpart Table 9, the frozen backbone outperforms both reward-trained backbones

and the stability-only reward outperforms the shared reward under both protocols ($38.5 > 19.4 > 15.0\%$ MSUN fixed, $47.6 > 29.5 > 22.6\%$ free), every reward fine-tune lowers the on-hull fraction of its DNG model, and replacing the composition source moves MSUN far more than replacing the structure generator. The one comparison that changes sign is the MSUN of the KLDM-DNG fine-tunes, slightly below or above the pretrained model with the lattice fixed (4.7% and 6.3% against 5.6%) and slightly above it with the lattice free (12.0% and 11.4% against 11.1%). The main text therefore bases its fine-tuning claims on the on-hull fraction, which falls under both.

E.1.5 Relaxation budget

The training relaxation is chosen for throughput, since the reward relaxes G structures per policy step and relaxation is essentially its whole cost. We measure the trade-off on the four structure sets relaxed under both protocols: $2{,}820$ paired structures, identical CIFs, MatterSim-5M. The evaluator records the BFGS step count and wall-clock time of every structure, so the 100-step budget can be read exactly from the 400-step runs. A relaxation that converged within 100 steps under the 400-step cap would also have converged under a 100-step cap, and its time is rescaled as the per-step cost times the truncated step count. These rates are measured under the evaluation potential rather than MatterSim-1M, so they indicate the relative cost of the two protocols rather than the exact convergence of the training runs.

Table 11: Relaxation convergence and cost at the training (100-step) and evaluation (400-step) budgets, on $2{,}820$ structures relaxed under both protocols with MatterSim-5M. 100-step values are read from the 400-step runs (see text). Training uses positions only at 100 steps; the main-text evaluation frees the lattice at 400 steps. "–": timing unavailable for that set.

		100-step budget		400-step budget	
Structures	Relaxation	Conv. (%)	s/struct	Conv. (%)	s/struct
ANCHOR→KLDM-CSP	positions	16.4	–	98.8	–
	cell free	6.5	7.18	99.4	14.00
ANCHOR→Crystalite-CSP	positions	27.7	5.47	99.5	8.00
	cell free	9.4	7.08	99.5	12.86
Crystalite-DNG	positions	90.2	2.10	100.0	2.28
	cell free	82.5	3.63	100.0	4.27
KLDM-DNG	positions	68.4	3.58	99.8	4.55
	cell free	57.0	5.18	99.8	7.25

Pooled over the four sets, freeing the lattice costs $1.28\times$ the force calls at the 400-step budget and 1.6–$1.9\times$ the wall clock, more than the step ratio because each cell-free step also updates the nine lattice degrees of freedom and evaluates the stress. At the 100-step training budget, freeing the lattice lowers convergence from 16.4% to 6.5% on ANCHOR→KLDM structures and from 27.7% to 9.4% on ANCHOR→Crystalite. Restoring convergence requires the 400-step budget, at which 99.4–99.5% of cell-free ANCHOR relaxations converge. On ANCHOR→Crystalite structures, where both timings are available, that costs $2.4\times$ the per-structure wall clock of the positions-only relaxation used in training (12.86 s against 5.47 s). Because the multiplier applies to the dominant cost of every policy step, a cell-free reward at matched convergence would have reduced the number of structures generated within the seven-day wall (Table 8) by roughly the same factor.

The positions-only reward is itself truncated: at 100 steps, only 16.4% of ANCHOR→KLDM relaxations converge. ANCHOR structures converge far less often than DNG structures under either protocol (68.4–90.2% positions-only at 100 steps for DNG), consistent with their larger relaxation displacement in Table 12 (RMSD 1.27-1.40 Å against 0.31–0.56 Å). The stability term of the training reward is therefore an energy along a truncated relaxation, used to rank candidates within a group. The reported numbers use the evaluation relaxation, at which at least 98.8% of structures converge.

E.2 MatterGen Evaluation Pipeline

MatterGen gives the same qualitative decomposition under a different evaluator. The Crystalite-DNG rows have higher raw metastability than ANCHOR (54.8–57.3% Meta versus 42.3–43.2%),

Table 12: MatterGen evaluation pipeline (unmodified), which supplies its own relaxer, reference hull and novelty definition. MSUN is unique $\wedge$ novel $\wedge$ $E_{\mathrm{hull}} \leq 0.1$, which the pipeline reports as S.U.N.; SUN uses $E_{\mathrm{hull}} \leq 0$. Meta is the stability rate at $E_{\mathrm{hull}} \leq 0.1$; Stable is the raw stability rate at $E_{\mathrm{hull}} \leq 0$, unfiltered by uniqueness or novelty. $\langle E_{\mathrm{hull}} \rangle$ (eV/atom) and RMSD (Å, against the pre-relaxation cell): lower is better. N is the number of structures the pipeline scored, below 1000 where it dropped structures of its own accord. We cannot report RMSD for Crystalite-DNG $\rightarrow$ CSP for its particular structures: the official pipeline fails while computing it. The other columns on that row use the same relaxed cells. Rows are defined in Appendix E.1. The pipeline relaxes positions and lattice (FIRE with `ExpCellFilter`). Entries are from sampling seed 0, and Table 3 reports the three-seed means.

Configuration	N	Unique	Novel	Meta	Stable	MSUN	SUN	$\langle E_{\mathrm{h}} \rangle$	RMSD
ANCHOR, inference-time backbone									
ANCHOR $\rightarrow$ KLDM-CSP	993	**100.0**	**100.0**	42.3	17.1	42.3	17.1	**0.133**	1.40
ANCHOR $\rightarrow$ Crystalite-CSP	993	**100.0**	**100.0**	42.8	14.5	42.8	14.5	0.138	1.27
ANCHOR $\rightarrow$ OMatG-CSP	991	**100.0**	**100.0**	43.2	17.3	**43.2**	**17.3**	0.134	1.28
ANCHOR trained with LeMat validity									
KLDM backbone	998	**100.0**	97.2	22.0	0.3	19.4	0.0	0.189	0.60
Trainable KLDM-CSP									
ANCHOR reward	997	**100.0**	**100.0**	18.3	5.0	18.3	5.0	0.202	1.37
Stability only	995	**100.0**	99.9	23.8	8.6	23.7	8.6	0.181	1.27
DNG									
Crystalite-DNG	922	99.1	63.7	54.8	7.2	22.9	2.3	0.195	0.31
Crystalite-DNG $\rightarrow$ CSP	922	99.1	62.8	**57.3**	8.0	23.6	2.8	0.187	–
Crystalite-DNG FT (mean)	989	60.8	64.5	52.5	4.8	17.2	1.9	0.158	0.45
Crystalite-DNG FT (bal.)	989	67.9	63.9	51.6	4.7	19.0	1.9	0.184	0.49
KLDM-DNG	943	99.8	75.9	39.8	6.3	18.6	2.7	0.213	0.56
KLDM-DNG $\rightarrow$ CSP	943	99.8	73.7	41.0	6.5	17.8	2.5	0.201	0.59
KLDM-DNG FT (mean)	943	99.8	78.5	33.6	3.4	15.2	1.2	0.239	0.70
KLDM-DNG FT (bal.)	948	99.8	80.0	32.0	4.2	15.4	2.1	0.223	0.69

but only 22.9–23.6% survives the unique-and-novel conjunction. For ANCHOR, Unique and Novel are both 100%, so MSUN is numerically the raw metastable rate: its advantage comes from allocating compositions to regions where the frozen CSP succeeds, not from a higher raw metastable fraction. Reward fine-tuning the DNGs again fails to convert adaptation into discovery, lowering both Stable and MSUN, while replacing only DNG structure generation with the corresponding frozen CSP changes MSUN by at most 0.8 points.

What the pipeline measures. MatterGen treats validity as two independent quantities rather than one (Zeni et al., 2025). *Composition validity* is a SMACT test (Davies et al., 2019): the reduced formula must admit a charge-neutral assignment of SMACT oxidation states that also respects the Pauling electronegativity ordering, with unary cells and all-metal alloys exempt from that check. *Structure validity* is far weaker, requiring only a minimum pair distance of 0.5 Å and a cell volume above 0.1 Å^3. Neither quantity coincides with LeMat Valid (Appendix E.3) or with our own gate (Appendix D.2).

Its columns are not nested. Unique, Novel and S.U.N. (MSUN in our tables) do not filter through either validity column. Unique and Novel are `StructureMatcher` flags (Ong et al., 2013) evaluated on every scored cell, and S.U.N. is unique $\wedge$ novel $\wedge$ $E_{\mathrm{hull}} \leq 0.1$ eV/atom. A model can therefore report Unique and Novel near 100% alongside a much lower composition validity, an ordering our own columns do not admit. Table 12 uses the pipeline's definitions throughout, in its default configuration: MatterSim-v1-1M, FIRE with `ExpCellFilter`, `DefaultDisorderedStructureMatcher`, and the Alex-MP `reference_MP2020correction` hull. Table 13 lists every published DNG result on this pipeline. The main text keeps the strongest variant of each model family.

Element coverage. Our 84-element vocabulary is broader than the element coverage of that reference hull, and a cell cannot be placed on the hull unless a terminal entry exists for every element it

Table 13: Representative published DNG results on MP-20 under MatterGen's evaluation pipeline, rounded to one decimal, with the frozen-backbone ANCHOR rows of Table 3 for reference (mean$_{\pm\text{std}}$ over three sampling seeds). MSUN is the quantity MatterGen and the cited works report as SUN. Table 3 keeps the strongest variant of each model family. Best in each column is bold, second-best underlined. [a]Cornet et al. (2025), [b]Höllmer et al. (2025), [c]Veljković et al. (2026), [d]Zhang et al. (2026).

Model	Unique	Novel	Meta	$\overline{E}_h$	RMSD	MSUN
Published DNG baselines						
MatterGen-MP[a]	–	–	47.1	0.20	<u>0.15</u>	25.8
DiffCSP[a]	–	–	41.3	0.19	0.41	20.1
KLDM-x_0(C)[a]	–	–	38.6	0.27	0.37	16.7
KLDM-x_0(C-AB)[a]	–	–	49.8	0.19	0.30	17.9
KLDM-x_0(D)[a]	–	–	59.2	0.16	0.28	18.5
FlowMM[b]	–	–	37.5	0.25	0.65	13.9
OMatG-SBD[b]	–	–	–	–	–	22.1
ADiT[c]	89.6	43.2	**70.0**	0.15	0.49	17.2
Crystalite[c]	94.7	56.6	64.5	0.15	0.27	24.3
CrystalFlow[d]	99.6	<u>73.1</u>	41.2	0.20	0.60	17.2
CrystalFlow+REPA[d]	<u>99.9</u>	69.0	55.0	0.17	0.42	20.0
DiffCSP+REPA[d]	99.5	72.0	56.5	0.15	0.32	24.8
MatterGen+REPA[d]	99.7	63.5	<u>66.2</u>	**0.12**	**0.11**	29.2
ANCHOR + frozen CSP (ours)						
Trained on KLDM	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	41.3$_{\pm 1.0}$	<u>0.14</u>$_{\pm 0.01}$	1.39$_{\pm 0.01}$	<u>41.3</u>$_{\pm 1.0}$
→ Crystalite	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	41.1$_{\pm 2.1}$	<u>0.14</u>$_{\pm 0.01}$	1.27$_{\pm 0.02}$	41.1$_{\pm 2.1}$
→ OMatG	**100.0**$_{\pm 0.0}$	**100.0**$_{\pm 0.0}$	43.1$_{\pm 0.9}$	<u>0.14</u>$_{\pm 0.01}$	1.27$_{\pm 0.01}$	**43.1**$_{\pm 0.9}$

contains. A single unsupported element voids the energy columns for the entire submitted batch, so such cells are removed beforehand, screened against the same reference object the pipeline is then given. This is why N falls below 1000 in Table 12; the surviving cells are scored exactly as the pipeline defines them. The same discard precedes published results on this pipeline (Cornet et al., 2025).

E.3 LeMat-GenBench Evaluation

Bulk-hull results. Against the Bulk hull, all three frozen-backbone ANCHOR variants retain roughly 73% Valid, Unique, and Novel samples but yield 0% SUN and only 2.9–3.1% MSUN. The Crystalite-DNG variants retain 1.4–2.3% SUN and 15.8–20.4% MSUN. Training against the LeMat validity gate raises MSUN to 26.0% while still producing 0% SUN, so a high metastable count under this evaluator need not imply any on-hull yield. Because the Bulk hull is not the reference used by the published Table 11 leaderboard, we treat this table as evaluator sensitivity and use the MP-20 hull below for direct comparison.

MP-20-hull results. Changing only the reference hull leaves Valid, Unique, and Novel unchanged but changes the stability picture substantially. ANCHOR→KLDM-CSP moves from 2.9% to 7.8% MSUN and from 0% to 0.4% SUN, while Crystalite-DNG moves from 19.6% to 30.0% MSUN and from 1.4% to 9.8% SUN. The reference hull alone therefore accounts for a large shift in absolute stability rates within the same LeMat pipeline. Under the published MP-20 protocol, the DNG rows remain much stronger on-hull than ANCHOR trained under our internal evaluator (9.8–10.9% versus 0.4–0.8% SUN), while the LeMat-validity-trained policy reaches 32.5% MSUN but only 1.0% SUN. This is why we keep MSUN and on-hull SUN separate when comparing evaluators.

What the pipeline measures. Structures submitted to LeMat-GenBench (Betala et al., 2025) are scored entirely by that pipeline, which defines every column differently from our internal tables. A structure is valid only if it passes three independent checks, and only valid structures are eligible for the downstream counts. Charge neutrality is a cascade rather than a single SMACT screen: compositions with a high SMACT metallicity score are exempted and pass, and the rest are tried

Table 14: LeMat-GenBench evaluation against the Bulk ensemble hull, NequIP pre-relaxation unless noted. All rates are fractions of the submitted N, with validity, uniqueness and novelty applied in that order. $\langle E_{\text{hull}} \rangle$ (eV/atom): lower is better. The *trained with LeMat validity* block optimised a LeMat validity filter during training, so its Valid, Unique and Novel entries are not comparable across blocks. Rows are defined in Appendix E.1.

Configuration	N	Valid	Unique	Novel	Meta	Stable	$\langle E_{\text{h}} \rangle$	MSUN	SUN
ANCHOR, inference-time backbone									
ANCHOR → KLDM-CSP	1000	73.0	73.0	73.0	2.9	0.0	0.20	2.9	0.0
ANCHOR → Crystalite-CSP	1000	73.0	73.0	73.0	3.1	0.0	0.20	3.1	0.0
ANCHOR → OMatG-CSP	998	72.9	72.9	72.9	3.1	0.0	0.20	3.1	0.0
ANCHOR trained with LeMat validity									
KLDM backbone	1000	99.4	99.4	96.2	28.1	0.1	0.17	26.0	0.0
Trainable KLDM-CSP									
ANCHOR reward	1000	86.5	86.5	86.5	6.9	0.0	0.19	6.9	0.0
Stability only	1000	81.5	81.5	81.4	5.3	0.0	0.20	5.2	0.0
DNG									
Crystalite-DNG	1000	95.6	94.9	57.5	45.4	8.9	**0.11**	19.6	1.4
Crystalite-DNG → CSP	1000	95.6	95.0	55.0	**48.4**	9.7	**0.11**	20.4	1.5
Crystalite-DNG FT (mean)	1000	97.8	59.3	53.8	45.3	10.5	0.16	15.8	1.2
Crystalite-DNG FT (bal.)	1000	97.0	66.4	57.1	41.9	**14.3**	0.17	17.5	**2.3**
KLDM-DNG	999	95.4	95.3	68.1	35.0	5.7	0.16	16.8	0.6
KLDM-DNG → CSP	999	95.4	95.1	67.1	35.1	6.9	0.15	17.4	0.5
KLDM-DNG FT (mean)	997	93.6	93.5	69.6	28.7	4.4	0.18	13.2	0.6
KLDM-DNG FT (bal.)	999	92.8	92.7	69.8	27.5	3.8	0.18	13.4	0.2

Table 15: LeMat-GenBench evaluation against the MP-20 ensemble hull, following the published Table 11 protocol so the rows can be read against that leaderboard. Columns as in Table 14; only the reference hull changes. The same caveat applies to the *trained with LeMat validity* block.

Configuration	N	Valid	Unique	Novel	Meta	Stable	$\langle E_{\text{h}} \rangle$	MSUN	SUN
ANCHOR, inference-time backbone									
ANCHOR → KLDM-CSP	1000	73.0	73.0	73.0	8.2	0.4	0.170	7.8	0.4
ANCHOR → Crystalite-CSP	1000	73.0	73.0	73.0	8.6	0.8	0.172	7.8	0.8
ANCHOR → OMatG-CSP	998	72.9	72.9	72.9	9.1	0.6	0.170	8.5	0.6
ANCHOR trained with LeMat validity									
KLDM backbone	1000	99.4	99.4	98.9	34.0	1.1	0.154	32.5	1.0
Trainable KLDM-CSP									
ANCHOR reward	1000	86.5	86.5	86.5	12.9	0.4	0.169	12.5	0.4
Stability only	1000	81.5	81.5	81.5	11.4	0.5	0.172	10.9	0.5
DNG									
Crystalite-DNG	1000	95.6	94.9	73.8	61.3	19.8	0.083	30.0	9.8
Crystalite-DNG → CSP	1000	95.6	95.0	70.7	**64.4**	**21.6**	**0.076**	29.0	**10.9**
Crystalite-DNG FT (mean)	1000	97.8	59.3	67.0	57.6	13.9	0.150	19.1	2.7
Crystalite-DNG FT (bal.)	1000	97.0	66.4	65.6	58.6	18.7	0.160	20.8	3.9
KLDM-DNG	999	95.4	95.3	80.3	45.8	11.7	0.133	26.2	4.5
KLDM-DNG → CSP	999	95.4	95.1	79.3	48.2	13.9	0.120	26.0	6.2
KLDM-DNG FT (mean)	997	93.6	93.5	83.0	38.3	9.9	0.153	23.0	5.1
KLDM-DNG FT (bal.)	999	92.8	92.7	82.5	37.2	9.0	0.157	23.0	4.1

against bond-valence sums, then `BVAnalyzer` decoration, then a frequency-weighted guess using LeMat's ICSD oxidation-state mapping, and finally all tabulated states. This cascade is the dominant filter, accounting for essentially all of the structures the benchmark rejects in our submissions. The minimum-interatomic-distance check sets its threshold per atom pair from the two atomic radii rather than as one global cutoff. The plausibility check screens bulk and cell quantities — density, atoms per unit volume, lattice lengths and angles, and space-group assignment — rather than local coordination environments.

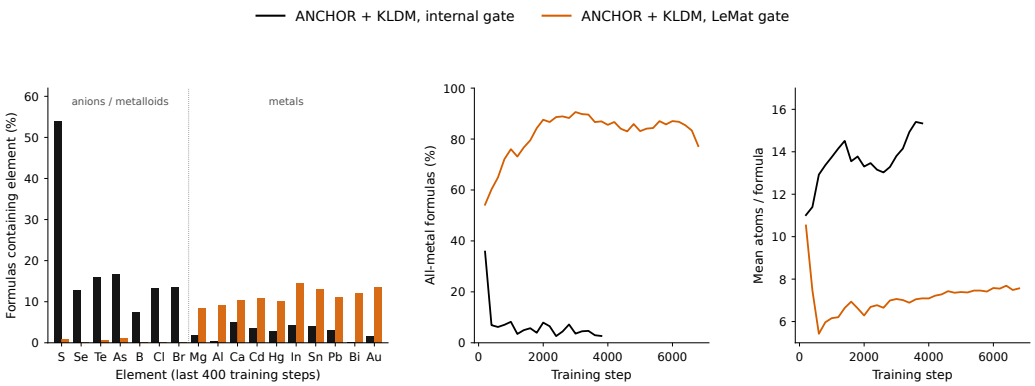

Figure 2: **Training gate sensitivity.** Policies trained with KLDM-CSP against our validity gate and against the LeMat-GenBench gate. Left: elements appearing in formulas over the last 400 training steps. Middle: fraction of all-metal compositions during training. Right: mean atoms per formula.

Table 16: **The gate and the threshold decide the ranking.** Identical structures scored three ways. *Internal* and *LeMat gate* differ only in the validity definition; relaxer (MatterSim-5M, BFGS-400, lattice free), hull and novelty index are held fixed (Table 2). *LeMat-GenBench* is that benchmark's own pipeline against its MP-20 ensemble hull, so its columns follow the published leaderboard. SUN is on-hull throughout, MSUN is $E_{\text{hull}} \leq 0.1$. Single sampling seed (seed 0), so the Internal-gate columns differ slightly from the three-seed means of Table 2. Policies in the second block were trained against the LeMat validity filter.

	Internal gate			LeMat gate, our pipeline			LeMat-GenBench		
Configuration	Valid	SUN	MSUN	Valid	SUN	MSUN	Valid	SUN	MSUN
ANCHOR trained against the internal validity gate									
ANCHOR → KLDM-CSP	97.9	23.2	49.9	72.9	13.6	34.1	73.0	0.4	7.8
ANCHOR trained against the LeMat validity gate									
KLDM backbone	5.6	0.0	2.1	98.3	0.1	28.1	99.4	1.0	32.5
DNG reference									
Crystalite-DNG	49.0	1.4	14.8	93.2	1.7	27.1	95.6	9.8	30.0

Where it diverges from our protocol. The pipeline deduplicates by structure rather than by reduced formula, so two polymorphs of one formula count twice there and once for us. It computes E_{hull} as the mean of a three-potential ensemble (ORB-v3, MACE-MP, UMA) against its own reference hull, where ours comes from a single MatterSim-5M relaxation, and its novelty is a structure-level match against that same reference corpus, where ours during training is the graded CAN score of Appendix D.4, computed per reduced formula against the frozen MP-20 index and against the history the training run itself accumulates. Two consequences matter for Valid%. Our gate accepts any pymatgen ICSD assignment whereas theirs can reject compositions for which `oxi_state_guesses` still returns a solution, so the two columns measure different charge searches, not "neutrality versus no neutrality," and should not be equated; Table 10 isolates that substitution inside our own pipeline. And the metallicity exemption leaves the charge check silent on compositions the heuristic treats as metallic, a potentially hackable trait under RL that could bias training toward the exempt set rather than toward charge-neutral formulas.

Reference hull. The benchmark admits two reference hulls and they are not interchangeable, so both are reported: the LeMat-Bulk hull (Siron et al., 2025) in Table 14 and the MP-20 hull in Table 15.

E.3.1 Training against the LeMat validity gate

The validity gate in ANCHOR's reward determines what the policy learns. Table 16 trains the same architecture, with the same KLDM-CSP backbone, against the LeMat-GenBench validity gate

and scores the result in three ways. Using the same gate during training and evaluation gives an essentially perfect policy, at 98.3% validity. Under our validity gate, however, only 5.6% of the same structures survive. The mechanism is documented in the benchmark: its charge cascade exempts compositions classified as metallic by the SMACT metallicity heuristic (Betala et al., 2025), which assigns every atom an oxidation state of zero rather than testing neutrality. The LeMat-trained policy is already about 54% all-metal within its first 200 steps and saturates near 85% by step 2000, ending at 80.0% over the last 400 steps against 3.1% for the policy trained with our gate (Figure 2). Anions largely disappear. Sulfur, present in 53.9% of formulas under our gate, appears in 0.8% under the LeMat gate, and the mean cell shrinks from 15.2 to 7.6 atoms per formula.

The stability threshold is exploitable in the same way. The LeMat-trained policy reaches 28.1% MSUN under the LeMat gate in our pipeline and 32.5% on LeMat-GenBench itself, above every DNG model, while producing almost no on-hull structures under either of our evaluators. Metallic alloys near the hull therefore satisfy every criterion the benchmark applies except on-hull stability, which is why we report SUN at $E_{\text{hull}} \leq 0$ as the headline and treat MSUN as secondary.

The dependence also cuts against ANCHOR. On LeMat-GenBench, Crystalite-DNG reaches 9.8% SUN and 30.0% MSUN against the MP-20 hull, while ANCHOR, trained against a different gate, relaxer and hull in our pipeline, reaches 0.4% and 7.8% at 73.0% validity. Training against the LeMat gate recovers metastability and beats every DNG model, but still loses on on-hull SUN. Which model leads therefore depends on the gate and threshold a benchmark uses, including our own. Full results against both LeMat-GenBench reference hulls are in Tables 14 and 15.

F Additional Experiments and Baselines

F.1 Base-model Selection

We sweep group size $G \in \{16, 32, 64\}$ against MatterSim calculators $\{1\mathrm{M}, 5\mathrm{M}\}$ under a fixed model setup including most significantly the exploration mechanisms; CAN, element sub-grouping and count bonus. Each job ran until a 7-day wall-clock limit with a structure-keyed entropy schedule without decay.

Reported Metrics. We report validity and uniqueness as the fraction of all generated samples whose formula is unique within the generated set i.e. not conditioned on validity. We additionally report SUN and MSUN rates (SUN = Valid $\wedge$ Unique $\wedge$ Novel $\wedge$ ($E_{\mathrm{hull}} \leq 0$); MSUN = Valid $\wedge$ Unique $\wedge$ Novel $\wedge$ ($E_{\mathrm{hull}} \leq 0.1$ eV/atom)). Novel uses our continuous score with a threshold of $\nu(x) > 0.5$, the midpoint of its $[0, 1]$ range. The reported rates are insensitive to where this cut is placed. A candidate whose reduced formula is absent from MP-20 receives the bootstrap value $\nu_0 = 0.75$, and 5,731 of the 5,737 valid structures in this grid did so; only 6 were scored by the kNN–RBF comparison at all. Any threshold below ν_0 therefore moves Novel% by at most those 6 structures (0.1%), and the choice of 0.5 over any other value in that range does not affect a single number we report. Appendix E.1 gives the full breakdown.

Selection protocol. We select the base configuration from the post-training evaluation of each final checkpoint (Table 17): N=1000 samples, seed 0, full element vocabulary, frozen KLDM prior, and a *shared* MatterSim-5M evaluation potential (BFGS, 400 steps, lattice held fixed) for every structure. This selects G=32 with MatterSim-1M. Selection therefore used the lattice-fixed protocol, and we did not re-select under the lattice-free protocol that Section 4 reports. Appendix E.1.4 shows that the conclusions drawn from the two protocols agree.

Table 18 reports late-window metrics over the final 10k training structures and ranks the same cell first, but it is not independent corroboration: each run is scored with its own training-time MatterSim potential, so the 1M and 5M rows are graded by different potentials and grader leniency is confounded with one of the two axes under comparison. We include it as a descriptive summary of where each run ended and rest the selection on the shared-potential evaluation alone. Subsequent invalid-penalty ablations use this base model setup, with this experiment run providing the $\mathrm{invalid_penalty} = -0.2$ reference model for the investigation in Appendix F.2.

Table 17: **Lattice fixed.** Post-training evaluation. Unique% is computed over all structures. SUN: $E_{\mathrm{hull}} \leq 0$; MSUN: $E_{\mathrm{hull}} \leq 0.1\,\mathrm{eV/atom}$.

Rank	Config	Valid%	Unique%	SUN%	MSUN%
1	**G32 / 1M**	**97.9**	100.0	**18.1**	**38.5**
2	G16 / 5M	94.3	100.0	<u>8.6</u>	<u>20.8</u>
3	G16 / 1M	93.6	100.0	6.0	13.5
4	G64 / 1M	<u>96.5</u>	100.0	3.5	13.4
5	G32 / 5M	96.0	100.0	3.9	11.5
6	G64 / 5M	95.4	100.0	1.5	8.5

F.2 Invalid-structure Penalty Sweep

Fixing the group size and training potential (Appendix F.1) leaves another important reward parameter undetermined: the penalty r_{inv} assigned to a structure that fails the validity gate. It is tempting to read this as a knob on the validity rate, and we test that reading below, but r_{inv} is better understood as the exchange rate between the gate and the discovery objectives, and it enters the update through three distinct channels.

It sets the dynamic range of the group-relative advantage. Valid structures score in $[0, 1]$ by construction (Appendix D.2), while an invalid one scores $r_{\mathrm{inv}} < 0$, so a group containing both spans a reward range wider than 1. Because the advantage is standardised within the group (equation 45), a

Table 18: Late-window comparison on the final 10k training structures. Unique% is computed over all structures. SUN/MSUN use training-time MatterSim scores (stability $=1.0$ / ≥ 0.9 as proxies for $E_{\text{hull}} \leq 0$ / ≤ 0.1).

Rank	Config	Valid%	Unique%	SUN%	MSUN%
1	**G32 / 1M**	87.0	99.8	**7.2**	**22.1**
2	G16 / 5M	**88.0**	99.7	3.8	11.7
3	G16 / 1M	85.1	99.8	3.3	11.2
4	G64 / 5M	85.2	99.8	1.2	7.5
5	G32 / 5M	84.8	99.8	2.1	7.5
6	G64 / 1M	84.9	99.8	1.4	7.0

Table 19: **Lattice fixed.** Invalid-penalty sweep on the selected base model (G=32, MatterSim-1M), evaluated post-training under Appendix E.1 (N=1000, seed 0). *Bond* and *Charge* are the pass rates of the two components of the validity gate (Appendix D.2); *Valid* is the conjunction.

r_{inv}	Valid%	Bond%	Charge%	Unique%	SUN%	MSUN%
-0.1	96.6	96.7	**99.9**	100.0	8.7	22.0
-0.2 (selected)	**97.9**	**98.4**	99.5	100.0	**18.1**	**38.5**
-0.5	97.4	97.6	99.8	100.0	12.0	28.4

more negative r_{inv} lowers the group mean but also inflates the group standard deviation, and the second effect *compresses* the normalised differences among the valid members. Enlarging the penalty does not simply add pressure against invalidity; it spends the group's available signal on a valid-versus-invalid contrast that the policy has largely already resolved, leaving less to separate a good valid structure from a mediocre one. This is the channel we care about most, because it acts on discovery rather than on validity.

It is the only weight-invariant term in the reward. An invalid structure is assigned r_{inv} in every objective coordinate, so its scalarised reward is $r_{\text{inv}} \sum_j w_j = r_{\text{inv}}$ for any $\boldsymbol{w}$ on the simplex. The penalty is therefore the one component of equation 8 that does not move with the sampled trade-off: it is a common floor shared by every policy in the weight-conditioned family, and any distortion it introduces is introduced everywhere at once rather than in one corner of the trade-off surface.

It prices exploration risk. The mechanisms of Appendix F.3 deliberately push the policy into unfamiliar chemistries, where the invalid rate is higher than in the well-covered regions it would otherwise exploit. r_{inv} is what that risk costs, so a penalty set too high has a plausible failure mode that is invisible in the validity rate: a policy that retreats to safe, familiar composition space and looks well behaved precisely because it has stopped exploring.

We therefore sweep $r_{\text{inv}} \in \{-0.1, -0.5\}$ on the selected base model, holding everything else fixed; the selected base model supplies the $r_{\text{inv}} = -0.2$ reference. All three arms are independent runs from the same frozen configuration under the same 7-day wall-clock, and they are budget-matched to within 3% (123–126k generated structures), so they can be compared directly. The outcome, together with Appendix F.1, defines the final configuration used everywhere else in the paper. Results of the sweep are in Table 19.

Validity is insensitive to the penalty. Over a five-fold range of r_{inv} the validity rate moves by 1.3 percentage points (96.6–97.9%), and not monotonically. At this point in training the gate is already satisfied by ∼97% of samples whatever the penalty, so the marginal structure that r_{inv} could discourage is rare enough that its contribution to the objective is small: what sets the validity rate is the frozen backbone and the gate itself, not the price attached to failing it. This also bounds the first channel above, since a small and near-constant invalid fraction leaves the group standard deviation dominated by the spread *among* valid structures. The gate's two components are not equally binding: charge neutrality passes at 99.5–99.9% while bond-length plausibility passes at 96.7–98.4%, so essentially all of our validity failures are geometric.

Table 20: **Lattice fixed.** Exploration mechanism ablation, evaluated post-training ($N_{\text{eval}} = 1000$, full chemical space, shared MatterSim-5M evaluation potential, BFGS relaxation of positions only for at most 400 steps, see Appendix E.1).
• marks that the mechanism is enabled, – means disabled. *Adapt.* = novelty scored against the growing agent-history index or frozen MP-20 index; *bin.* = the binary Chemeleon2 flag of Appendix F.3. *CB* = count bonus, *SG* = element sub-group mask. Runs were matched on wall clock.

Configuration	Adapt.	CB	SG	Valid (%)	Unique (%)	Stable (%)	Meta (%)	MSUN (%)	SUN (%)
Full method	•	•	•	97.9	**100.0**	18.1	38.5	**38.5**	**18.1**
– count bonus	•	–	•	96.3	**100.0**	8.7	24.6	24.6	8.7
– sub-group mask	•	•	–	96.4	**100.0**	7.6	27.9	27.9	7.6
– adaptive nov.	–	•	•	96.9	99.9	8.4	26.4	26.4	8.4
adaptive nov. only	•	–	–	98.1	**100.0**	12.3	37.2	37.1	12.3
count bonus only	–	•	–	96.4	**100.0**	7.3	28.3	28.3	7.3
sub-group mask only	–	–	•	**99.8**	6.3	19.3	46.0	2.0	0.6
none	–	–	–	**99.8**	1.0	**60.0**	**68.7**	0.6	0.6
binary novelty	bin.	•	•	97.2	99.9	5.7	16.6	16.6	5.7

No cell becomes conservative. The failure mode that Valid% would hide, a policy buying its 97% by retreating into safe, familiar chemistry, does not appear at any of the swept penalties: compositional uniqueness is 100.0% in all three arms. The measure is sensitive enough to have caught it, since removing a genuinely load-bearing exploration mechanism collapses uniqueness to 1.0% (Appendix F.3). Over $[-0.5, -0.1]$ the penalty fixes the ordering of valid above invalid without changing how much the policy explores. SUN and MSUN move more than validity or uniqueness across the sweep (8.7–18.1% and 22.0–38.5%), peaking at the default -0.2. We use -0.2 on that basis but do not read three single-seed points as an interior optimum.

F.3 Exploration-mechanism Ablations

Table 20 repeats Table 1 with the lattice held fixed, and the numbers below refer to it. It varies the three exploration mechanisms as a full 2^3 factorial rather than one ablation at a time. Because all three act on the same failure mode, the question is not only whether each helps but whether any one of them is *sufficient*, and a leave-one-out design alone cannot separate redundancy from complementarity. All cells share the configuration of Table 7 fixed by the two selection studies (G=32 with MatterSim-1M, Appendix F.1; $r_{\text{inv}} = -0.2$, Appendix F.2) and are evaluated under Appendix E.1 with the lattice held fixed. Every cell is a mechanism switched on or off, and nothing here re-selects a hyperparameter. Disabling the adaptive axis of CAN means novelty is scored by the same continuous kNN-RBF metric against the frozen MP-20 index alone. The novelty term is never removed from the reward. All cells are single-seed, so we read the ordering of widely separated rows and not differences of a few points.

Reading the factorial. The collapse of the all-off cell is the reward being optimized, not a failure to optimize it. That cell attains the highest per-structure stability in the whole study (68.7% of samples at $E_{\text{hull}} \leq 0.1$ against 38.5% for the full method, and 60.0% at $E_{\text{hull}} \leq 0$ against 18.1%) while proposing only 10 distinct reduced formulas in 1000 samples. The policy locates a composition the potential scores favorably and re-proposes it, which is exactly the degenerate solution a stability-weighted objective admits when nothing prices repetition. Per-structure stability is therefore not a safe headline metric for a generative policy; it has to be read together with a diversity measure.

The two halves of the factorial rank the mechanisms differently, and the disagreement is the point. In isolation, adaptive novelty is clearly the most valuable (MSUN $= 37.1\%$, against 28.3% for the count bonus and 2.0% for the mask alone). The leave-one-out costs are instead tightly clustered (24.6–27.9%, each roughly a third of MSUN) and put the count bonus first. That is what redundancy looks like: with two of the three still active, the survivors partly cover for the one removed. Either of the two depth mechanisms alone restores full compositional diversity (100.0% unique), so their anti-repetition role is genuinely interchangeable in kind, though not in quality.

One caveat runs in our favor. The runs were stopped by a common wall clock rather than a common structure budget, and the collapsed cells consumed *more* structures than the full method (648k for the all-off cell and 237k for sub-group-only, against 102–131k elsewhere), because repeatedly proposing the same small cells makes each relaxation cheap (Table 8). Those cells had more training and still produced two orders of magnitude fewer distinct formulas.

Adaptive versus fixed-reference novelty The control disables the agent-history index so that novelty is scored against the frozen MP-20 corpus alone, with everything else held fixed. Beyond the post-training row, a single full-method run shows the adaptive branch is exercised as intended. We instrument the selected base model (Appendix F.1) by recording, for every scored structure, which branch of equation 7 produced its score. Over the run ($123{,}968$ generated structures, $101{,}237$ valid) the agent produced $97{,}599$ distinct reduced formulas, a formula-repeat rate of 3.6%. Every repeat is routed through an anti-rediscovery branch rather than silently collecting the bootstrap value ν_0: the $3{,}638$ repeated formulas split into $3{,}550$ scored against agent history and 88 against both indices. Splitting the run into fifths shows the reference moving. Collisions with MP-20 fall from 346 in the first fifth to 3 in the last, so the policy leaves the corpus the backbone was trained on. More tellingly, collisions with the agent's *own* history fall from $1{,}078$ to 635 over the same span, even as Z^{hist} grows by orders of magnitude. At a fixed sampling distribution a larger history can only raise the collision probability, so a falling count means the sampling distribution itself moved away from already-visited chemistries, the behavior the $\min$ over history in equation 7 is meant to induce. These counts are descriptive and come from one run: they show the mechanism is active, not that it *causes* the downstream gains. The fixed-reference row is what isolates the effect.

Continuous versus binary novelty This ablation replaces our graded novelty score with a static, binary one and changes nothing else. For each generated structure we compute a flag $v \in \{0, 1\}$ from two same-formula tests:

- **Novel** if no reference of the *same reduced formula* matches the candidate under pymatgen's `StructureMatcher` (tolerances $l_{\mathrm{tol}} = 0.2$, $s_{\mathrm{tol}} = 0.3$, $\theta_{\mathrm{tol}} = 5°$). The reference is our frozen MP-20 index ($45{,}229$ structures, $37{,}217$ reduced formulas).
- **Unique** if the candidate is the first member of its `StructureMatcher` group within the current batch of generated structures.

The policy is trained on v alone: the novelty entry of every valid structure is overwritten with 1 when v holds and 0 otherwise, replacing equation 6. Our continuous RDF novelty and the agent-history index still run, but only as logged diagnostics. Binary novel/known tests of this kind are the standard evaluation convention in crystal generation (Zeni et al., 2025; Betala et al., 2025), and these two flags are the ones used in Chemeleon2 (Park & Walsh, 2026); rather than design our own, we adopt them here, with uniqueness evaluated within one GRPO group of $G = 32$. Chemeleon2 combines the flags differently: its "creativity" reward is 1 when a candidate is both unique and novel, 0 when it is neither, and the Chebyshev distance between average-minimum-distance (AMD) descriptors (Widdowson et al., 2022), a continuous isometry invariant of periodic crystals, only when the two disagree. We log that reward but do not train on it, so this cell is a binary variant of *our* novelty term rather than a reproduction of their objective.

Why a binary signal saturates in our setting. The tests are defined relative to structures of the same reduced formula in MP-20. Our composition policy is explicitly rewarded for proposing formulas that lie outside MP-20, and once it does so the same-formula reference list is empty, the tests cannot fail, and $v = 1$ identically. Being constant across a GRPO group, the flag then contributes nothing to the group-relative advantage (equation 45): the novelty term cancels and the policy is trained on validity and the count bonus alone. The resulting loss is therefore one of *gradient* rather than the collapse of the all-off cell (validity and uniqueness stay at 97.2% and 99.9%) yet the cell still scores below every configuration that retains a continuous novelty signal, including single-mechanism cells with *fewer* active mechanisms. What the comparison isolates is the value of a graded signal to a group-relative update, in a search whose objective is to leave the reference corpus; it is not a statement about binary novelty tests as evaluation metrics.

Count bonus and sub-group mask. The informative cell for the count bonus is its isolation row rather than its removal. On its own it holds compositional uniqueness at 100.0%, so it is sufficient

to prevent the all-off collapse, but it converts that diversity into far fewer stable discoveries than adaptive novelty does (MSUN 28.3% against 37.1%). The count bonus prices *repetition of a formula*; adaptive novelty prices *proximity to what has already been made*, which additionally pushes the policy away from near-duplicates that a formula counter cannot see. The sub-group mask is the only one of the three that cannot stand alone: by itself it leaves uniqueness at 6.3%, barely better than the 1.0% of the all-off cell, and its removal is the cheapest of the three ($38.5 \rightarrow 27.9\%$). Both are the expected signature of a breadth mechanism. Restricting the element pool per rollout changes which region of the periodic table a group explores, but nothing in it penalises proposing the same composition twice inside that region, so it redistributes exploration without bounding repetition. It is valuable in combination and inert in isolation.

F.4 Learned Composition Preferences

To isolate whether ANCHOR simply learns elementary compositional validity, we perform an experiment keeping the complete generation, relaxation, and scoring fixed and compare uniformly random formulas, uniformly sampled charge-neutral formulas, and the trained ANCHOR policy (Table 21). Random formulas are drawn uniformly from the untrained action space of Appendix D.1, and enumerated formulas uniformly from its charge-neutral enumeration. Both are queried through the same frozen KLDM-CSP backbone and scored by the same evaluator as the policy, so only the formula source changes. With the lattice held fixed (Appendix E.1.4), charge-neutral enumeration raises Valid from 9.5% to 70.1%, but MSUN increases only from 0.4% to 3.1%. ANCHOR reaches 97.9% Valid and 38.5% MSUN. All three composition sources remain 100% unique, showing that the improvement cannot be explained by collapse or by learning charge neutrality alone. The policy learns to allocate queries to compositions for which the frozen CSP model produces substantially more successful discovery candidates.

Table 21: **Lattice fixed.** Effect of composition selection with the structural pipeline held fixed. All rows use the same frozen KLDM-CSP backbone and post-training evaluation protocol (N=1000, seed 0, full chemical space, shared MatterSim-5M, BFGS=400). Unique% is distinct reduced formulas over all N.

Composition source	Valid%	Unique%	SUN%	MSUN%
KLDM + random formulas	9.5	100.0	0.2	0.4
KLDM + enumerated charge-neutral	70.1	100.0	0.8	3.1
ANCHOR (selected policy)	**97.9**	100.0	**18.1**	**38.5**

F.5 Reward Fine-tuning Baselines

Two families of baseline optimize the same reward as ANCHOR equation 8 but place the gradient on a structure model instead of, or in addition to, a composition policy. Together they measure what reward fine-tuning of the generative model buys on its own, and so isolate the contribution of keeping the backbone frozen and searching over compositions.

DDPO fine-tuning of DNG models. We fine-tune the released Crystalite-DNG checkpoint and a KLDM-DNG model retrained on MP-20 with DDPO (Black et al., 2024), treating each denoising step as an action and assigning the terminal reward of equation 8 to every step of the trajectory with no discounting. The implementation is the clipped surrogate of equation 46 averaged over denoising steps rather than over a group of compositions,

$$\mathcal{L}^{\text{clip}}(\theta) = \mathbb{E}\Big[\tfrac{1}{|\mathcal{T}|}\textstyle\sum_{t\in\mathcal{T}} \min\big(\eta_t A,\ \text{clip}(\eta_t, 1-\epsilon, 1+\epsilon)\, A\big)\Big], \qquad \eta_t = \tfrac{p_\theta(x_{t-1}|x_t)}{p_{\theta_{\text{old}}}(x_{t-1}|x_t)}, \tag{49}$$

where A is the sample's terminal reward standardized over the sampling batch, shared by every step of its trajectory, and $\mathcal{T}$ is the set of scored steps. For Crystalite-DNG $\mathcal{T}$ is restricted to steps whose churn noise lies in $[0.2, 2.0]$, where that sampler's stochasticity makes the step log-probability informative; for KLDM-DNG all EM steps are scored.

Both runs take a single inner epoch per batch, which is DDPO's score-function form: the update is $-A\,\nabla_\theta \log p_\theta(x_{t-1} \mid x_t)$ averaged over the scored steps, with the clip at $\epsilon = 0.05$ and a trust-region

stop on the mean $|\log \eta_t|$ retained for configurations that take more epochs. Reward fine-tuning costs uniqueness on Crystalite-DNG, which falls from 98.8% to 40.5% (*mean*) and 49.5% (*bal.*) in our pipeline (Table 2), so part of the reward there is collected by re-emitting structures, while KLDM-DNG stays above 99%.

The reward uses the same validity gate, stability term, adaptive novelty index, both diversity terms and count bonus as ANCHOR, with $r_{\text{inv}} = -0.2$ and $\lambda_c = 1$, evaluated by the same scorer. Two details differ from the evaluation protocol of Appendix E.1 and are properties of training, not of the reported metrics: relaxation during fine-tuning is BFGS for at most 100 steps against 400 at evaluation, and the training potential is MatterSim-1M against MatterSim-5M at evaluation. Objective weights are fixed rather than resampled per rollout, either at the mean of the training Dirichlet (*mean*, $\boldsymbol{w} = (0.143, 0.571, 0.143, 0.143)$) or with stability lifted to novelty's level (*bal.*, $\boldsymbol{w} = (0.4, 0.4, 0.1, 0.1)$), both in the order $(s_{\text{stab}}, \nu, d_{\text{struct}}, d_{\text{comp}})$. Hyperparameters are in Table 22. Both reported rows sample from a pinned snapshot taken after about 4.1k episodes, roughly 66k structures generated, about half the budget of an ANCHOR run (Table 8). By then Crystalite-DNG had already doubled its validity and lost more than half its uniqueness, so further steps would extend the same shift rather than reverse it.

Table 22: DDPO fine-tuning of the two DNG models. One Adam step per episode, taken after accumulating gradients over the scored denoising steps of the batch.

	Crystalite-DNG	KLDM-DNG
Initialization	released checkpoint	MP-20 retrain (below)
Learning rate	5×10^{-6}	5×10^{-6}
Batch size (structures / episode)	16	16
Denoising steps per sample	150	50
Scored steps $\mathcal{T}$	churn $\in [0.2, 2.0]$	all
Inner epochs per batch	1	1
Clip ϵ	0.05	0.05
Checkpoint	rolling, every 50 episodes	rolling, every 50 episodes

Why DDPO rather than GRPO on the generator. In this setting, the two use the same advantage estimator, so the choice is not a substantive one. GRPO replaces a learned value baseline with the mean reward of a group of samples drawn under a shared context. For ANCHOR's composition policy that context is a preference vector and an element sub-group, and the group is the G compositions sampled under it (equation 45). Our DNG baselines are sampled unconditionally, so the only group available is the sampling batch, and the group-relative advantage becomes exactly the batch-standardized terminal reward A of equation 49. Park & Walsh (2026) instead group rollouts by a conditioning input such as the number of atoms. What separates these baselines from Chemeleon2 is therefore not the estimator, which differs only in how the group is formed, but the treatment of drift from the pretrained denoiser, the axis this comparison is designed around, and the reward, since Chemeleon2 scores novelty against a fixed reference rather than a search history. Chemeleon2 restrains drift with a KL penalty, the DNG baselines leave it unrestrained, and ANCHOR removes it exactly by freezing the backbone (equation 25).

MP-20 retraining of KLDM-DNG. The KLDM checkpoint used elsewhere in this paper is a CSP model, so the DNG baseline is a separate KLDM-DNG model we train from scratch on MP-20 with KLDM's own `dng` recipe, on the same split as the CSP backbone. Fine-tuning initializes from its final checkpoint.

Reward-trained CSP backbone. In the *reward-trained KLDM-CSP* rows, GRPO updates the backbone alongside the composition policy. All 14.9M parameters of the KLDM network are unfrozen; there are no adapters. The gradient does not flow through the sampling chain: rollouts are drawn under `no_grad`, and the stored denoising trajectory is then replayed to give the backbone the same clipped per-step objective as equation 49, so the backbone is trained by DDPO nested inside the GRPO update of the policy. The backbone is held fixed for the first 50 policy steps, then updated with learning rate 5×10^{-5} against 10^{-4} for the policy, and clip $\epsilon = 0.05$ against 0.2. Unlike the DNG runs above, these use $K = 2$ epochs per batch, so on the second epoch the ratio departs from

one and both the clip and a trust-region guard are live: an epoch whose mean $|\log \eta_t|$ exceeds 0.05 is discarded rather than applied. There is still no divergence penalty to the pretrained weights in the loss.

With the *shared* reward the backbone receives the full reward of equation 8. With the *stability-only* reward it receives s_{stab} alone while the composition policy still receives the full reward, which tests whether the backbone should be asked to chase novelty and diversity that only the composition policy can act on. The two arms are otherwise identical: their configurations differ by this one setting and the run name.

F.6 Training the CSP Backbone on ANCHOR Structures

We fine-tune each CSP backbone with its own training objective on MP-20 together with $27{,}136$ structures from a single ANCHOR training run, the base model selected in Appendix F.1. None of the added ANCHOR structures occur in the original training set of either backbone. A run does not persist geometry, only a log of the compositions it proposed and their scores, so the corpus is built in two stages: we sample logged candidates, then replay each composition through the frozen backbone to obtain a cell. What is preserved exactly is the composition distribution the trained policy produced, not the individual cells seen during training. Each backbone is fine-tuned on KLDMs rendering of that composition list unless otherwise stated.

Of the run's $123{,}968$ logged candidates we remove $22{,}731$ that fail the validity gate and $3{,}224$ duplicates, then draw $27{,}136$ of the remaining $98{,}013$ uniformly at random without replacement. We apply no stability or novelty threshold and no window over training steps, so the sample represents the run as a whole rather than its converged phase. Validity rises from 68.6% in the first quartile of the run to 87.3% in the last, so this choice dilutes the sample relative to drawing from late steps only.

The sample size equals that of the MP-20 training split, so each + ANCHOR arm trains on $54{,}272$ structures against its control's $27{,}136$. Each + ANCHOR arm is paired with a *ctl* arm trained on pristine MP-20 for the same number of optimizer steps at the same batch size, which equalizes compute and total structure exposures and leaves the corpus as the only difference between them. Matching steps rather than epochs means the augmented arm sees each individual structure about half as often as its control; that is the intended trade, since the question is whether a fixed budget is better spent on more diverse data or on more repetitions of less of it. The budget is matched within a backbone, not across the two, which have different published recipes.

FT resumes from the released checkpoint for a further $\sim$150k steps, reaching global step $849{,}600$ for KLDM, and *from scratch* retrains for the same 699.6k (KLDM) or 1.4M (Crystalite) steps as the released model. Two further arms isolate artifacts of how the corpus was built rather than properties of the chemistry in it. *Unrelaxed cells* trains on the ANCHOR cells as generated, before MatterSim relaxation, which tests whether a change is an artifact of the relaxer. *Crystalite-rendered* keeps the ANCHOR compositions but realises them with Crystalite-CSP instead of KLDM-CSP, so that KLDM is not fine-tuned on its own output; it is the control against self-distillation. At each of three sampling seeds, every CSP checkpoint is evaluated on the same 1000 held-out ANCHOR compositions, none of which are part of the $27{,}136$-composition fine-tuning corpus. Table 23 reports every arm.

The corpus helps Crystalite-CSP and does not help KLDM-CSP. Against its step-matched control, Crystalite-CSP gains 10.1 points of SUN when finetuned ($14.4 \rightarrow 24.5$) and 7.6 when retrained from scratch ($13.4 \rightarrow 21.0$), which makes the finetuned arm the strongest CSP checkpoint we report. On KLDM-CSP the same corpus is flat to slightly negative ($17.1 \rightarrow 16.5$ finetuned, $17.9 \rightarrow 17.1$ from scratch). The two diagnostic arms bear on why. Holding the compositions fixed and rendering them with Crystalite instead of KLDM moves KLDM-CSP from below its control to above it ($16.5 \rightarrow 18.9$), whereas removing the relaxation step changes almost nothing ($16.5 \rightarrow 16.7$): the geometries matter, and which model produced them matters more than whether they were relaxed. The one arm that fails to improve is also the only one trained on cells from the backbone with which it was trained, which is consistent with a backbone gaining little from its own output, though the two backbones differ enough in starting quality that we do not read the size of the Crystalite gain as attributable to that alone.

Table 23: Training the CSP backbone on ANCHOR structures, under the MatterGen evaluation protocol. ANCHOR contributed 27,136 structures from one of its training runs; each + ANCHOR arm adds them to MP-20 and each *ctl* arm is the step-matched control on pristine MP-20, so the pair isolates the corpus from the extra optimisation. Finetuning resumes from the released checkpoint; *from scratch* retrains it for the same 699.6k (KLDM) or 1.4M (Crystalite) steps. Every row is scored on the same 1000 ANCHOR compositions as the inference-time backbone rows of Appendix E.1; only the CSP checkpoint changes. Entries are from sampling seed 0; Tables 2 and 3 report seed-averaged values for the FT arms. Composition validity and uniqueness are fixed by the shared formula list and are reported per protocol in Appendix E.1.

Configuration	MSUN	SUN
KLDM-CSP		
Released checkpoint	42.3	17.1
FT: MP-20 only (ctl)	41.9	17.1
FT: MP-20 + ANCHOR	41.6	16.5
FT: MP-20 + ANCHOR, unrelaxed cells	42.5	16.7
FT: MP-20 + Crystalite-rendered ANCHOR	43.9	18.9
From scratch: MP-20 only (ctl)	42.2	17.9
From scratch: MP-20 + ANCHOR	42.5	17.1
Crystalite-CSP		
Released checkpoint	42.8	14.5
FT: MP-20 only (ctl)	41.9	14.4
FT: MP-20 + ANCHOR	**49.8**	**24.5**
From scratch: MP-20 only (ctl)	37.6	13.4
From scratch: MP-20 + ANCHOR	<u>47.8</u>	<u>21.0</u>