Synthesizability-Aware Materials Generation with Target Properties via Reinforcement Learning

Duc Dat Dang [*1] **Junwu Chen** [*1 2] **Philippe Schwaller** [1 2]

Abstract

Generative models have shown remarkable promise in accelerating materials discovery, yet most generated candidates remain synthetically inaccessible, limiting their practical impact. We address this critical gap by fine-tuning pretrained diffusion models through multi-objective reinforcement learning (RL) that explicitly incorporates multiple synthesis-related constraints for the generation of novel crystal structures with targeted properties and experimental synthesizability. A precursor set of commercially available, non-toxic, and low-cost compounds is constructed, and synthesis-planning models are employed to predict precursor availability, with the resulting score incorporated into the RL reward. Synthesis-related filters, encompassing material class, elemental complexity, and synthesizability score, are further integrated to ensure compatibility with solid-state synthesis routes. The results demonstrate that the proposed framework simultaneously optimizes different material properties, including bulk modulus and magnetic density, while satisfying synthesis constraints, effectively steering the generative model toward functionally promising and experimentally accessible crystal structures.

1. Introduction

Deep generative models have substantially broadened the scope of inorganic chemical space exploration, encompassing variational autoencoders (Xie et al., 2021; Luo et al., 2023), GAN-based approaches (Kim et al., 2020), diffusion-based frameworks (Jiao et al., 2023; 2024; Zeni et al., 2025), AI agentic systems (Du et al., 2025; Huang et al., 2025), and large-scale discovery pipelines (Merchant et al., 2023). For inverse materials design, these models are steered toward desired properties via condition generation (e.g., MatterGen (Zeni et al., 2025)) or reinforcement learning (RL) frameworks such as MatInvent (Chen et al., 2025a) and Chemeleon2 (Park & Walsh, 2025), which reframes the denoising process as a sequential decision problem and demonstrates effective property targeting across band gap, shear modulus, dielectric constant, and other functional properties.

Despite this progress, the novel and thermodynamically stable crystal structures proposed by the generative models remain difficult to realize experimentally. Stability-derived metrics, such as energy above the convex hull (E_{hull}) and formation energy, are not sufficiently predictive of whether a generated structure can be experimentally synthesized (Antoniuk et al., 2023). The synthesis of novel materials typically involves multiple steps, including choices of synthesis techniques, precursors availability and trial-and-error in the laboratory, where success often depends heavily on subjective experience (Zhu et al., 2023; Aykol et al., 2019; Raccuglia et al., 2016). For wet laboratory, commercially available, non-toxic, and inexpensive precursors are generally preferred for material synthesis (Cui et al., 2018; Olivetti et al., 2017). Therefore, incorporating precursor constraints into materials generation can substantially facilitate subsequent experimental validation. Solid-state synthesis is one of the most widely used techniques for inorganic materials preparation, particularly in automated laboratories, where it is well suited for the synthesis of metal oxides, phosphates and silicates that do not contain volatile elements (Szymanski et al., 2023). Introducing material-type constraints during the generation process can thus ensure the compatibility of the generated crystals with solid-state synthesis routes. Moreover, restricting the complexity of generated materials, for example by limiting the maximum number of elements, can further reduce the synthesis difficulty. In addition, several approaches have been developed to predict synthesizability scores (Jang et al., 2024) or structural similarity (He et al., 2023) to previously reported

[1]Laboratory of Artificial Chemical Intelligence (LIAC), Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland [2]National Centre of Competence in Research (NCCR) Catalysis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. Correspondence to: Junwu Chen <junwu.chen@epfl.ch>, Philippe Schwaller <philippe.schwaller@epfl.ch>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

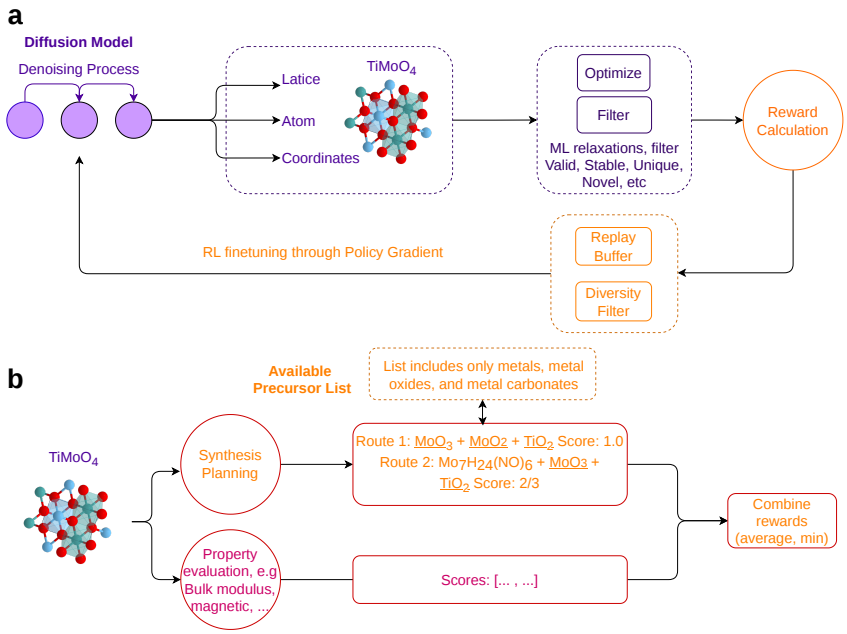

Figure 1. Schematic of the synthesizability-aware reinforcement learning framework. (a) The iterative generative pipeline. A diffusion model produces crystal structures that undergo structural relaxation and stability filtering (e.g., SUN filter). The policy is fine-tuned via policy gradient, stabilized by experience replay and diversity filters. (b) The multi-objective reward mechanism Generated candidates (e.g., $TiMoO_4$) are simultaneously evaluated for target physical properties and practical synthesizability, scored against a strictly constrained precursor inventory. These distinct scores are aggregated via minimum or average functions to compute the final reward signal.

materials, thereby providing complementary information beyond thermodynamic stability alone.

To address this critical gap, we fine-tune the pre-trained diffusion model using multi-objective RL framework and explicitly incorporate multiple synthesis-related constraints. Our contributions are as follows. **First**, we construct a precursor set $\mathcal{P}$ consisting of commercially accessible, non-toxic, and low-cost compounds. The synthesis-planning models are then employed to evaluate generated crystals by predicting whether they can be synthesized from precursors in $\mathcal{P}$. The resulting score is incorporated into the RL reward, thereby enabling precursor-constrained materials generation. **Second**, constraints on material class, number of elements, and synthesizability score are integrated into the RL filter to promote compatibility with solid-state synthesis routes and to reduce synthesis complexity. **Third**, the results demonstrate that our RL framework can simultaneously optimize different material properties (e.g., bulk modulus, magnetic density) under various synthesis-related constraints. These findings indicate that the proposed approach can effectively guide diffusion models toward the generation of novel crystal structures that are not only functionally promising but also amenable to experimental validation.

2. Preliminaries

2.1. Diffusion Models for Crystal Generation

A crystal structure is represented as a unit cell $\mathcal{M} = (\boldsymbol{A}, \boldsymbol{F}, \boldsymbol{L})$, where $\boldsymbol{A} \in \mathbb{R}^{h \times N}$ encodes atom types, $\boldsymbol{F} \in [0,1)^{3\times N}$ are fractional coordinates, and $\boldsymbol{L} \in \mathbb{R}^{3\times 3}$ is the lattice matrix (Jiao et al., 2023; Hoogeboom et al., 2022; Chen et al., 2025b). Fractional coordinates preserve periodic boundary conditions under lattice translations; diffusion on $\boldsymbol{F}$ is therefore performed via score matching with a wrapped normal distribution to respect this periodicity (Jiao et al., 2023; Zeni et al., 2025).

We adopt MatterGen (Zeni et al., 2025) as our generative backbone. Diffusion models are trained on two coupled Markov chains: a **forward process** that gradually corrupts a clean structure $\mathcal{M}_0$ by injecting Gaussian noise over T timesteps, and a **reverse process** that learns to iteratively denoise it (Ho et al., 2020; Song et al., 2020). The forward process is defined as:

$$q(\mathcal{M}_t \mid \mathcal{M}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\,\mathcal{M}_{t-1},\, \beta_t \boldsymbol{I})$$

The reverse (denoising) process is parameterized by a graph neural network θ:

$$p_\theta(\mathcal{M}_{t-1} \mid \mathcal{M}_t) = \mathcal{N}(\mu_\theta(\mathcal{M}_t, t),\, \sigma_t^2 \boldsymbol{I})$$

and trained by minimizing the denoising objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\,\mathcal{M}_t\sim q}\left[\|\tilde{\mu}(\mathcal{M}_0,t) - \mu_\theta(\mathcal{M}_t,t)\|^2\right]$$

where $\tilde{\mu}(\mathcal{M}_0,t) = \frac{\sqrt{\bar{\alpha}_{t-1}}\,\beta_t}{1-\bar{\alpha}_t}\mathcal{M}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathcal{M}_t$ is the analytically tractable posterior mean of the forward process, with $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$ (see (Ho et al., 2020) for the full derivation).

During inference, novel crystal structures are generated by first sampling from the noise prior — $\boldsymbol{L}_T, \boldsymbol{A}_T \sim \mathcal{N}(\mathbf{0},\boldsymbol{I})$ for the lattice and atom types, and $\boldsymbol{F}_T \sim \mathcal{U}(0,1)$ for the fractional coordinates, and iteratively applying the learned reverse process to recover a valid crystal structure $\mathcal{M}_0$.

2.2. Reinforcement Learning for Crystal Generation

RL optimizes a policy π_θ by framing the problem as a Markov Decision Process (MDP) (Black et al., 2023; Fan et al., 2023), defined by the tuple (S, A, P, R, ρ_0), where S is the state space, A is the action space, $P(s_{t+1} \mid s_t, a_t)$ is the transition dynamics, $R(s_t, a_t)$ is the reward function, and ρ_0 is the initial state distribution. The objective is to find θ that maximizes the expected cumulative reward:

$$\mathcal{J}(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} R(s_t,a_t)\right]$$

MatInvent (Chen et al., 2025a) addresses this by treating the T-step denoising process as a T-step MDP, where at each step the agent observes a partially noisy crystal and acts to produce a less noisy one. The MDP components are defined as:

$$\begin{aligned} s_t &= \mathcal{M}_{T-t}, \quad a_t = \mathcal{M}_{T-t-1} \\ \rho_0(s_0) &= \mathcal{N}(\mathbf{0},\boldsymbol{I}) \times \mathcal{U}(0,1) \\ P(s_{t+1} \mid s_t, a_t) &= \delta_{a_t} \\ \pi_\theta(a_t \mid s_t) &= p_\theta(\mathcal{M}_{T-t-1} \mid \mathcal{M}_{T-t}) \end{aligned}$$

In this formulation, the policy π_θ coincides exactly with the denoising process, so optimizing the policy is equivalent to fine-tuning the diffusion model. Since crystal properties can only be evaluated on a fully denoised structure, the reward is sparse and assigned only at the final step when $\mathcal{M}_0$ is obtained:

$$R(s_t,a_t) = \begin{cases} r(\mathcal{M}_0) & \text{if } t = T-1 \\ 0 & \text{otherwise} \end{cases}$$

Training objective. Applying policy gradient to this formulation gives an objective that increases the log-likelihood of denoising trajectories proportional to their reward:

$$\mathcal{L}_{\text{PG}}(\theta) = -\,r(\mathcal{M}_0)\sum_{t=1}^{T}\log p_\theta(\mathcal{M}_{t-1} \mid \mathcal{M}_t)$$

However, since the reward is only observed at the end of each trajectory, optimizing $\mathcal{L}_{\text{PG}}$ alone risks the model drifting far from the pretrained distribution into physically invalid structures. MatInvent therefore adds a KL regularization term that penalizes deviation from the pretrained model p_{pre}, weighted by how much the current sample departs from the target reward:

$$\begin{aligned} \mathcal{L}_{\text{KL}}(\theta) = &\ (\lambda - r(\mathcal{M}_0)) \\ &\sum_{t=1}^{T}\text{KL}\Big(p_\theta(\mathcal{M}_{t-1} \mid \mathcal{M}_t)\,\Big\|\,p_{\text{pre}}(\mathcal{M}_{t-1} \mid \mathcal{M}_t)\Big) \end{aligned} \quad (1)$$

where λ is a constant set slightly above the maximum reward, ensuring the KL weight is always positive. The final training objective combines both terms:

$$\mathcal{L}(\theta) = \alpha\mathcal{L}_{\text{PG}}(\theta) + \beta\mathcal{L}_{\text{KL}}(\theta)$$

Filtering mechanisms and experience replay. To further stabilize training, MatInvent adopts several mechanisms. **SUN filter**(Chen et al., 2025a) is employed to constrain the generative action space. Only crystal structures that are thermodynamically Stable ($E_{\text{hull}} < 0.1$ eV/atom), Unique, and Novel (SUN) are retained after filtering. From these retained structures, n_{samples} are randomly selected for property evaluation and assigned corresponding rewards. Then, MatInvent applies **diversity filter** penalizes repeated elemental compositions to encourage the model to generate different chemical systems, and uses **experience replay** that reuses high-reward trajectories from previous iterations to improve sample efficiency (Chen et al., 2025a).

Multi-property optimization. In most practical tasks, a single reward signal is insufficient to capture the full set of desired material properties. To handle this, the framework supports multi-property optimization (MPO) by combining individual property rewards into a single scalar using aggregation functions. Given a set of property rewards $\{R_1, R_2, \ldots, R_m\}$, the aggregated reward is computed as either their minimum or mean:

$$R_{agg} = \min(R_1,\ldots,R_m) \quad \text{or} \quad R_{agg} = \frac{1}{m}\sum_{i=1}^{m} R_i \quad (2)$$

The minimum encourages balanced optimization by penalizing any single underperforming objective, while the mean allows trade-offs across objectives. Before aggregation, each individual reward score is normalized to the range $[0,1]$.

3. Methodology

To bridge the gap between crystal generation and wet-lab synthesis, we add synthesizability-related signals into the RL framework. This section details the formulation of RL reward, the necessary construction of a constrained precursor space, and the integration of multiple objectives.

3.1. Synthesizability-Aware Reward Formulation

To evaluate the synthesizability of a generated crystal structure $\mathcal{M}_0$, we utilize inorganic retrosynthesis models (He et al., 2023; Noh et al., 2024) to suggest viable synthesis pathways. Given a target structure, these models output a vector of probabilities corresponding to a predefined vocabulary of materials, suggesting a top-k precursor set.

We quantify the synthesis planning reward, R_{synth}, by calculating the proportion of suggested precursors that are available in our pre-defined laboratory inventory. Mathematically, this reward is defined as the average availability across the k suggested sets:

$$R_{synth}(\mathcal{M}_0) = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{N_j}\sum_{i=1}^{N_j}\mathbb{1}[p_{j,i} \in \mathcal{P}]\right)$$

where k is the number of suggested precursor sets, N_j is the total number of precursors within the j-th set, $p_{j,i}$ is the i-th precursor in set j, and $\mathcal{P}$ denotes our pre-defined list of available precursors. The indicator function $\mathbb{1}$ returns 1 if the suggested precursor exists within our available list $\mathcal{P}$, and 0 otherwise.

Because retrosynthesis models draw from their own predefined material vocabularies,simply defining $\mathcal{P}$ similar to model's list, the calculated reward would be 1.0 to most materials (except those materials are too complex, models could hallucinate and suggest no valid precursors), meaning we are not teaching anything to the agent.

3.2. Precursor Space Construction

As mentioned above, to ensure that the reward signal accurately reflects real-world synthesizability and provides a meaningful learning signal, we constructed a constrained precursor space $\mathcal{P}$. Although the precursor dictionary derived from Synthesis Similarity (He et al., 2023) and A-Lab (Szymanski et al., 2023) encompasses over 400 materials, we systematically filtered these precursors to yield a curated set of 80 commercially available, non-toxic, and rare-earth-free compounds.

The finalized precursor list $\mathcal{P}$ is strictly limited to standard metals, metal oxides, and metal carbonates. To ensure laboratory safety and practical viability, all heavy metals (toxic metals), radioactive elements, and rare earth elements were explicitly excluded from the available inventory. By restricting $\mathcal{P}$ to this subset, the RL agent is forced to actively optimize its generative policy to utilize common, safe, and available precursors (more details in the Appendix A.1).

3.3. Material type constraints and solid-state synthesis compatibility

Current automated laboratories predominantly employ solid-state synthesis for material fabrication (Szymanski et al., 2023). This approach is best suited for binary and ternary oxides, phosphates, silicates, and carbides (Szymanski et al., 2023). Accordingly, a filtering step within the RL pipeline retains only generated structures belonging to these composition classes prior to reward computation. In addition, during this filtering stage, each generated structure is restricted to a maximum of four elements. Synthesizability scores are also computed using an existing ML-based predictive model (Jang et al., 2024), and only generated structures with a synthesizability score exceeding 0.5 are retained.

3.4. Multi-Objective Integration

To assess whether the synthesis planning reward can be jointly optimized alongside material properties, we integrate R_{synth} with other property rewards. The goal is to demonstrate that our framework can guide diffusion models to generate materials with the desired properties and that are synthetically feasible. Both aggregation methods described in the preliminaries are evaluated, and a comparison of their performance on these multi-objective tasks is presented in the experimental section.

4. Experiments

4.1. Experimental Setup

We adopt the same diffusion model as MatInvent (Chen et al., 2025a) as our agent backbone, pretrained on the MP-20 dataset consisting of 45,231 stable inorganic materials (Jain et al., 2013).

For synthesis planning, we implemented two models: Synthesis Similarity (He et al., 2023) and Retrieval-Retro (Noh et al., 2024) with $k = 3$ to suggest precursors for generated materials and then evaluate their effectiveness in RL fine-tuning. Synthesis Similarity (He et al., 2023), developed for the A-Lab autonomous laboratory (Szymanski et al., 2023), recommends precursors via similarity search over historical synthesis recipes. Retrieval-Retro (Noh et al., 2024) extends similarity-based retrieval by incorporating thermodynamic information between target and candidate precursors and applying a cross-attention mechanism to predict reaction-specific routes. We integrate both methods as reward signals into the MatInvent (Chen et al., 2025a) framework, scoring generated candidates by whether their predicted precursors fall within $\mathcal{P}$ and training the policy accordingly.

For multi-property optimization, we conducted four separate experiments, each pairing R_{synth} with a target property reward: bulk modulus, magnetic density, Minimal Co-Incident

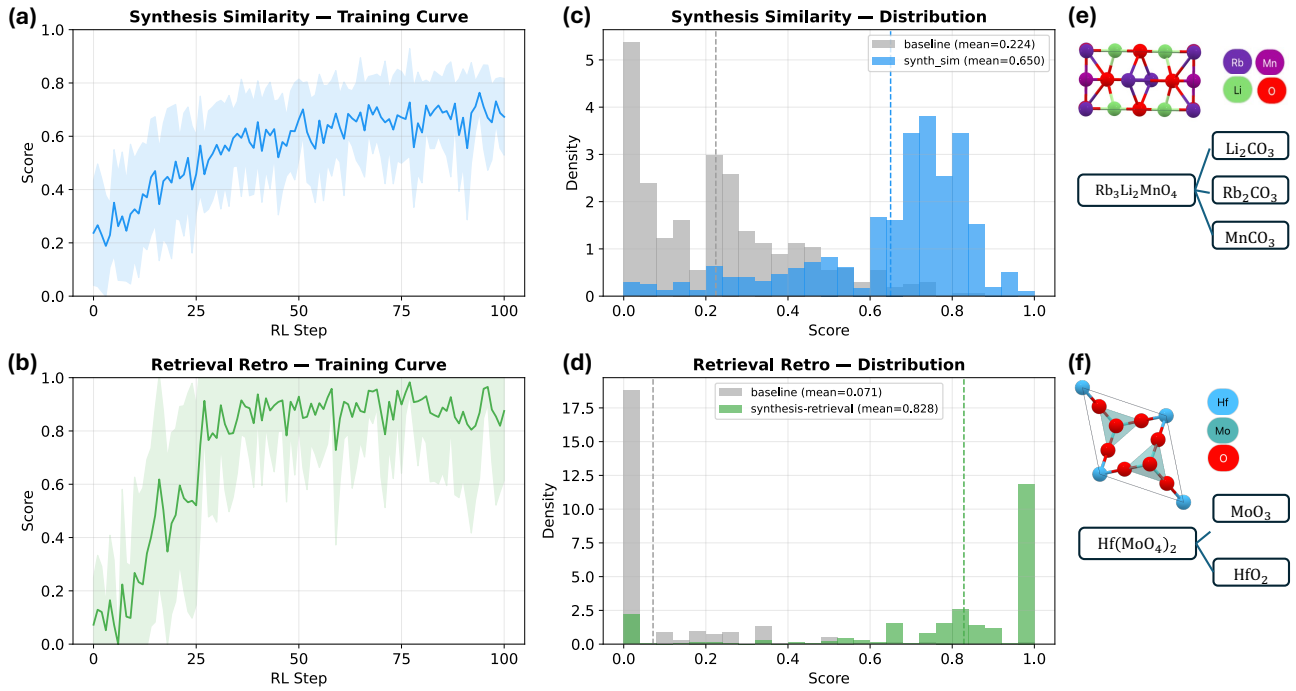

Figure 2. RL performance on synthesis planning optimization. (a) training curves for both models, (b) reward distribution of 1,024 samples generated from the finetuned models, (c–d) representative generated structures.

Area (MCIA) and Herfindahl–Hirschman Index (HHI) score. These properties are predicted using ALIGNN (Choudhary & DeCost, 2021) and PyMatGen (Ong et al., 2013)

The diffusion model is fine-tuned for 100 epochs on single objective tasks and 120 epochs on multi-objective tasks with a learning rate of 10^{-5}. For each RL iteration, 64 structures are generated, with a maximum of 16 structures used for reward calculation.

4.2. Synthesis Planning Optimization

To evaluate the efficacy of our synthesis planning reward R_{synth}, we conduct single-objective optimization comparing two retrosynthesis models as guidance signals: Synthesis Similarity (SS) (He et al., 2023) and Retrieval-Retro (RR) (Noh et al., 2024).

As illustrated in Figure 2a, SS improves steadily over the first 40 steps and converges to a mean reward of 0.65 after 100 iterations. While this exceeds the pretrained baseline of 0.224, it remains below our target threshold of 0.8. RR (Figure 2b) begins from a much lower baseline of 0.07, but converges within 40 iterations to a mean reward of 0.828, with 67.9% of finetuned samples scoring above 0.8, compared to only 9.4% under SS. We attribute these dynamics to two distinct training phases.

Early phase: the lower RR baseline reflects the pre-RL sample distribution Of 1,024 samples generated by the pretrained model, 48% contain rare-earth elements, which we exclude from desired precursor space $\mathcal{P}$. In addition, SS has a 417-precursor vocabulary, while RR draws from a broader 798-template space, yielding sparser precursor predictions with respect to $\mathcal{P}$ at initialization. Both factors penalize RR more heavily than SS in the early iterations, causing it to start substantially low.

Late phase: SS avoids rare-earth materials, while RR drives substantive learning. To isolate rare-earth avoidance from genuine synthesizability gains, we score the 1,024 baseline samples restricted to non-rare-earth materials. Under SS, this subset already scores 0.68 — essentially matching the SS-finetuned ceiling of 0.65: rare-earth filtering alone explains nearly all of SS's learning gain, and the reward signal saturates without driving further structural change. Under RR, the same non-rare-earth subset scores only 0.11, while RR-finetuned samples reach 0.828, a 7× improvement that cannot be attributed to filtering. This contrast is consistent with the mechanism of each model. SS retrieves the most similar known compound and inherits its synthesis route (He et al., 2023), mapping novel materials to standard oxide/carbonate precursors that lie within $\mathcal{P}$ but only loosely match the actual chemistry. RR uses neural reaction energy and cross-attention over thermodynamic features to predict reaction-specific routes (Noh et al., 2024), generalizing better to novel materials and producing reward signals that continue to scale as the agent's distribution shifts. F

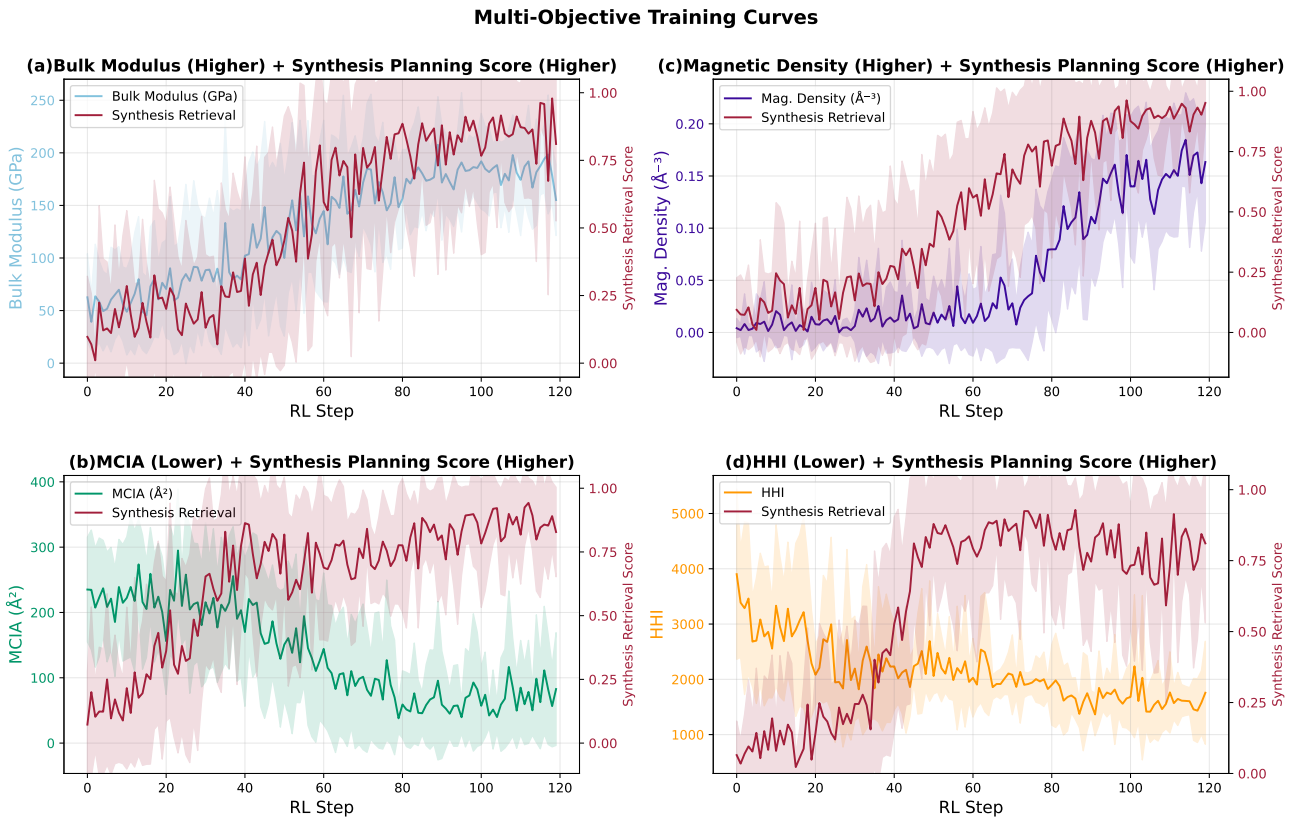

Figure 3. Multi-objective training curves for the four property–R_{synth} pairings: (a) bulk modulus (higher), (b) MCIA (lower), (c) magnetic density (higher), (d) HHI (lower). Each panel shows the property reward (left axis) and the synthesis retrieval score R_{synth} (right axis, dark red) over 120 RL fine-tuning steps. Solid lines are batch means; shaded regions denote one standard deviation across the batch.

Cross-validation between models. To verify that gains under RR reflect genuine synthesizability rather than model-specific overfitting, we score the RR-finetuned samples with SS. They reach a mean SS reward of 0.57, approaching the 0.65 ceiling obtained by training directly on SS, despite never having seen SS during finetuning. Consistently, 54.3% of RR-finetuned samples qualify as Stable, Unique, and Novel (SUN), compared to 50.8% under SS (Figure 5), indicating that RR steers generation toward materials that are synthesizable from both retrosynthesis perspectives.

We note that R_{synth} measures the average fraction of model-predicted precursors contained in $\mathcal{P}$ rather than chemical synthesizability directly: a converged mean of 0.828 indicates that, on average, 82.8% of the precursors required to synthesize a generated material are available in our available precursor list. Based on these results, we adopt Retrieval-Retro as the syhthesis planning predictions for all subsequent experiments.

4.3. Multi-property Optimization

Having established Retrieval-Retro as our synthesis planning model in Section 4.2, we now test whether R_{synth} can be jointly optimized with target material properties. We pair R_{synth} with four objectives spanning different categories of practical relevance: (1) bulk modulus (higher, GPa), measuring mechanical stiffness for structural materials; (2) magnetic density (higher, Å^{-3}), relevant to permanent magnets and spintronics; (3) minimal coincident interfacial area (MCIA, lower, Å^2) (Ding et al., 2016), which measures lattice mismatch between a generated material and a substrate for epitaxial growth; and (4) Herfindahl–Hirschman Index (HHI, lower) (Gaultois et al., 2013), a measure of supply chain risks. All four pairings use the mean aggregator from Equation 2; we explored the minimum aggregator on a subset of pairings without observing improvement, and leave a systematic comparison to future work.

The training curves in Figure 3 show that joint optimization works across all four tasks. In every case, R_{synth} rises within the first 30–40 steps and finally stabilizes near or above the 0.8 threshold, while the property reward improves more gradually over the 120-step horizon. The magnetic density pairing shows the slowest property takeoff, beginning to climb only after roughly 70 steps, while R_{synth} remains stable throughout. We observe no interference between the two signals: R_{synth} does not regress as the prop-

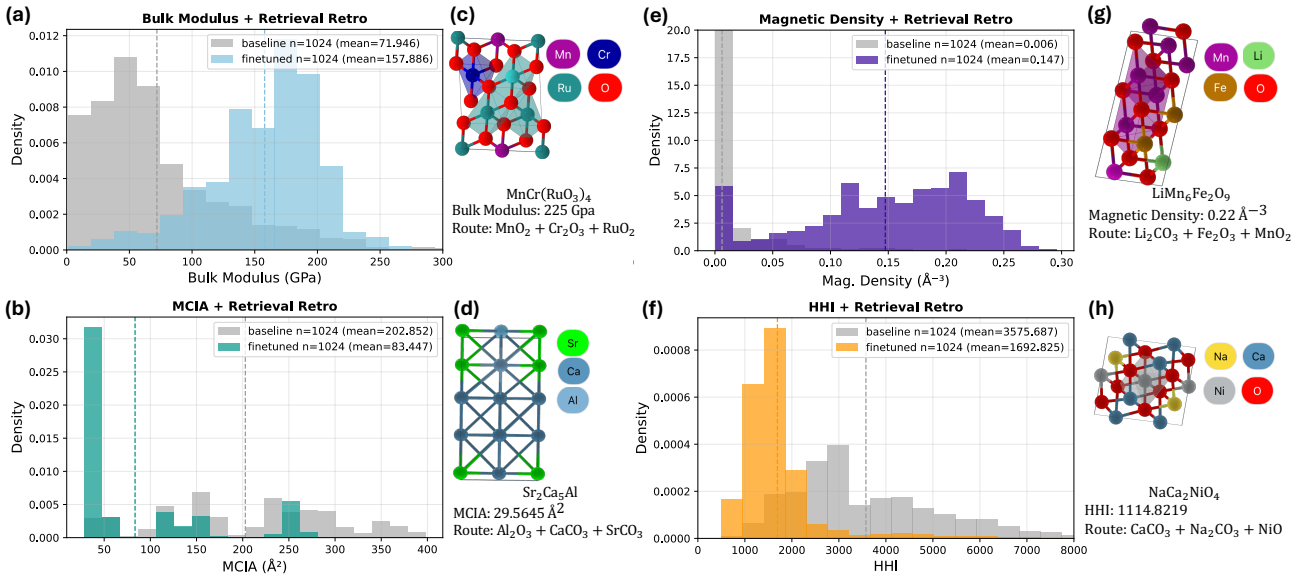

Figure 4. Property distributions for 1024 samples generated by the pretrained diffusion model (gray) and by each multi-objective finetuned model (colored), across the four pairings: (a) bulk modulus, (b) MCIA, (c) magnetic density, (d) HHI. Dashed lines mark distribution means; mean values are annotated in each panel legend.

Table 1. Multi-objective results across four property–R_{synth} pairings, evaluated on 1,024 generated samples per setting. S, U, and N denote Stable, Unique, and Novel ratios; SUN is the joint ratio. Pretrained and single-objective RR rows are included for reference.

	Synthesizability		Structure quality				Property mean	
Setting	R_{synth}	≥ 0.8	S	U	N	SUN	Pretrained	Finetuned
Pretrained	0.071	1.2%	76.4%	61.6%	100%	39.6%	—	—
Single-obj RR	0.828	67.9%	88.4%	73.8%	84.5%	54.3%	—	—
+ Bulk modulus (GPa) ↑	0.783	68.1%	64.3%	76.7%	84.7%	33.5%	71.9	157.9
+ Mag. density (Å^{-3}) ↑	0.816	76.1%	70.0%	63.1%	85.7%	28.9%	0.006	0.147
+ MCIA (Å^2) ↓	0.780	59.8%	88.0%	62.1%	83.6%	42.9%	202.9	83.4
+ HHI ↓	0.723	41.3%	88.7%	66.2%	81.7%	47.0%	3,575.7	1,692.8

erty reward climbs in any of the four pairings, indicating the diffusion model finds materials where both objectives are simultaneously satisfied.

To assess the population-level effect of finetuning, we generated 1,024 samples from both the pretrained and finetuned models and scored each set against the corresponding property predictor and against R_{synth}. The resulting property distributions are shown in Figure 4. All four tasks produce large shifts in the intended direction: (1) bulk modulus more than doubles (71.9 $\rightarrow$ 157.9 GPa); (2) magnetic density rises from a near-zero baseline by over an order of magnitude (0.006 $\rightarrow$ 0.147 Å^{-3}), reflecting a shift from a predominantly non-magnetic prior toward magnetic materials; (3) MCIA drops to roughly 40% of its pretrained mean (202.9 $\rightarrow$ 83.4 Å^2); and (4) HHI more than halves (3,575.7 $\rightarrow$ 1,692.8). Figure 6 shows the corresponding R_{synth} distributions: magnetic density and bulk modulus concentrate most of their mass above the 0.8 threshold, while HHI retains a visibly heavier left tail — consistent with HHI being the hardest pairing on the synthesizability axis.

Table 1 summarizes the synthesizability and Figure 5 SUN outcomes across the four pairings. The mean R_{synth} ranges from 0.723 (HHI) to 0.816 (magnetic density), with the ≥ 0.8 fraction spanning 41.3% to 76.1% — all well above the pretrained baseline of 0.071. HHI is the hardest pairing on this axis, despite its compositional-economic motivation overlapping a priori with the cheap-precursor inventory. The $R_{synth} \geq 0.8$ fraction, however, captures only the precursor axis. Decomposing SUN (Figure 5) shows that bulk modulus (33.5%) and magnetic density (28.9%) fall below the pretrained SUN baseline of 39.6%, while MCIA (42.9%) and HHI (47.0%) exceed it — and the two underperforming pairings fail on different components. In details Figure 7, Magnetic density holds Stable at 70.0% but its Unique fraction drops to 63.1%, the lowest of any setting and well

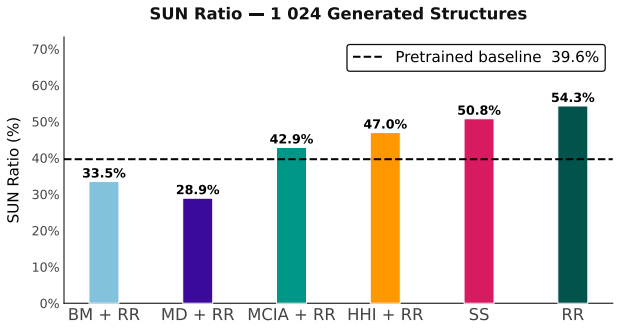

Figure 5. SUN ratio of Finetuned Models, from left to right respectively: Bulk Modulus, Magnetic Density, MCIA, HHI, Synthesis Similarity and Retrieval-Retro

below the 73.8% from single-objective RR, consistent with mode collapse onto a narrow chemical family. Bulk modulus shows the opposite pattern: Unique stays at 76.7% while Stable drops to 64.3%, indicating that high-modulus structures favored by the policy are less thermodynamically favorable. MCIA and HHI depend only on lattice geometry and on element identity respectively, and pull less directly against either stability or diversity. The synthesizability constraint thus composes with diverse property objectives, but the SUN decomposition is necessary to surface the structural cost that each pairing imposes.

5. Limitation

Our method has proved it can ensure both satisfying physical demands and precursors constraints. However, these properties predictors (SS, RR or ALIGNN) are machine learning models trained on historical and in-distribution compounds, which can struggle with OOD since diffusion model are generating novel structures. This leads to the reward signal can be uncertain and noisy. Moreover, SUN decomposition in Figure 5 and Table 1 shows that bulk modulus and magnetic density, while generating physically demanding samples, are exploiting "reward hacking" with low Stable and Unique ratio. The agent satisfies the explicit reward by exploiting regions of chemical space that score well on each individual objective while compromising the implicit structural-quality objective that SUN captures. Finally, the synthesis method focuses on solid-state only, which can limit other options of choosing precursors and type of target materials.

6. Conclusion

We presented a synthesizability-aware RL framework that fine-tunes a pretrained diffusion model against a fixed inventory of 80 cheap, lab-accessible precursors. The choice of retrosynthesis model used to score generated structures matters: thermodynamics-grounded retrieval (Retrieval-Retro) substantially outperforms similarity-based retrieval (Synthesis Similarity), reaching a mean R_{synth} of 0.828 versus 0.65, with 67.9% versus 9.4% of finetuned samples meeting the 0.8 threshold. Cross-validation across the two scorers indicates that these gains reflect genuine synthesizability rather than scorer-specific overfitting. Joint optimization of R_{synth} with bulk modulus, magnetic density, MCIA, and HHI keeps mean R_{synth} well above the pretrained baseline in all four pairings, though the SUN decomposition reveals pairing-specific structural costs — diversity loss for magnetic density, stability loss for bulk modulus, with cleaner composition for MCIA and HHI.

By scoring generated structures against a fixed, cheap-precursor inventory, our framework biases generation toward materials that can plausibly be synthesized in a real laboratory — a constraint we view as essential for practical experimental access. Several directions remain open: stronger diversity regularization for pairings prone to mode collapse; retrosynthesis models calibrated on out-of-distribution generated structures rather than historically synthesized compounds; and richer reward signals incorporating estimated reaction yield, energetics, and process complexity.

References

Antoniuk, E. R., Cheon, G., Wang, G., Bernstein, D., Cai, W., and Reed, E. J. Predicting the synthesizability of crystalline inorganic materials from the data of known material compositions. *npj Computational Materials*, 9 (1):155, 2023.

Aykol, M., Hegde, V. I., Hung, L., Suram, S., Herring, P., Wolverton, C., and Hummelshøj, J. S. Network analysis of synthesizable materials discovery. *Nature communications*, 10(1):2018, 2019.

Black, K., Janner, M., Du, Y., Kostrikov, I., and Levine, S. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Chen, J., Guo, J., Fako, E., and Schwaller, P. Accelerating inverse materials design using generative diffusion models with reinforcement learning. *arXiv preprint arXiv:2511.03112*, 2025a.

Chen, J., Huang, X., Hua, C., He, Y., and Schwaller, P. A multi-modal transformer for predicting global minimum adsorption energy. *Nature Communications*, 16(1):3232, 2025b.

Choudhary, K. and DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021.

Cui, J., Kramer, M., Zhou, L., Liu, F., Gabay, A., Hadjipanayis, G., Balasubramanian, B., and Sellmyer, D. Cur-

rent progress and future challenges in rare-earth-free permanent magnets. *Acta Materialia*, 158:118–137, 2018.

Ding, H., Dwaraknath, S. S., Garten, L., Ndione, P., Ginley, D., and Persson, K. A. Computational approach for epitaxial polymorph stabilization through substrate selection. *ACS applied materials & interfaces*, 8(20):13086–13093, 2016.

Du, Y., Yu, B., Liu, T., Shen, T., Chen, J., Rittig, J. G., Sun, K., Zhang, Y., Song, Z., Zhou, B., et al. Accelerating scientific discovery with autonomous goal-evolving agents. *arXiv preprint arXiv:2512.21782*, 2025.

Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., and Lee, K. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

Gaultois, M. W., Sparks, T. D., Borg, C. K., Seshadri, R., Bonificio, W. D., and Clarke, D. R. Data-driven review of thermoelectric materials: performance and resource considerations. *Chemistry of Materials*, 25(15):2911–2920, 2013.

He, T., Huo, H., Bartel, C. J., Wang, Z., Cruse, K., and Ceder, G. Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature. *Science advances*, 9(23):eadg8180, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Huang, X., Chen, J., Fei, Y., Li, Z., Schwaller, P., and Ceder, G. Cascade: Cumulative agentic skill creation through autonomous development and evolution. *arXiv preprint arXiv:2512.23880*, 2025.

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

Jang, J., Noh, J., Zhou, L., Gu, G. H., Gregoire, J. M., and Jung, Y. Synthesizability of materials stoichiometry using semi-supervised learning. *Matter*, 7(6):2294–2312, 2024.

Jiao, R., Huang, W., Lin, P., Han, J., Chen, P., Lu, Y., and Liu, Y. Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information Processing Systems*, 36:17464–17497, 2023.

Jiao, R., Huang, W., Liu, Y., Zhao, D., and Liu, Y. Space group constrained crystal generation. *arXiv preprint arXiv:2402.03992*, 2024.

Kim, S., Noh, J., Gu, G. H., Aspuru-Guzik, A., and Jung, Y. Generative adversarial networks for crystal structure prediction. *ACS central science*, 6(8):1412–1420, 2020.

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

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

Noh, H., Lee, N., Na, G. S., and Park, C. Retrieval-retro: retrieval-based inorganic retrosynthesis with expert knowledge. *Advances in Neural Information Processing Systems*, 37:25375–25400, 2024.

Olivetti, E. A., Ceder, G., Gaustad, G. G., and Fu, X. Lithium-ion battery supply chain considerations: analysis of potential bottlenecks in critical metals. *Joule*, 1(2): 229–243, 2017.

Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A., and Ceder, G. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

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

Raccuglia, P., Elbert, K. C., Adler, P. D., Falk, C., Wenny, M. B., Mollo, A., Zeller, M., Friedler, S. A., Schrier, J., and Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601): 73–76, 2016.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624 (7990):86, 2023.

Xie, T., Fu, X., Ganea, O.-E., Barzilay, R., and Jaakkola, T. Crystal diffusion variational autoencoder for periodic

material generation. *arXiv preprint arXiv:2110.06197*, 2021.

Zeni, C., Pinsler, R., Zügner, D., Fowler, A., Horton, M., Fu, X., Wang, Z., Shysheya, A., Crabbé, J., Ueda, S., et al. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025.

Zhu, R., Tian, S. I. P., Ren, Z., Li, J., Buonassisi, T., and Hippalgaonkar, K. Predicting synthesizability using machine learning on databases of existing inorganic materials. *ACS omega*, 8(9):8210–8218, 2023.

A. Experimental details

A.1. Precursor Vocabulary Used in Retrieval-Retro

We use a filtered subset of the Synthesis Similarity precursor vocabulary (He et al., 2023) consisting of **80 compounds** (57 metal oxides and 23 metal carbonates) drawn from 29,900 training reactions spanning 83 elements. Compounds containing rare-earth elements (Ce, Dy, Er, Eu, Gd, Ho, La, Lu, Nd, Pm, Pr, Sc, Sm, Tb, Tm, Y, Yb), radioactive elements (Ac, Am, At, Bk, Cf, Cm, Es, Fm, Fr, Lr, Md, No, Np, Pa, Po, Pu, Ra, Rn, Tc, Th, U), and heavy/toxic metals (As, Be, Bi, Cd, Hg, Pb, Sb, Tl) are excluded. Table 2 lists all retained precursors sorted by training-set frequency within each category.

Table 2. Filtered precursor vocabulary used by the Retrieval-Retro scorer. *Count* is the weighted occurrence frequency in the training reaction dataset. Precursors are grouped by chemical class and sorted by count (descending).

Metal Oxides (57)									**Metal Carbonates** (23)		
#	Formula	Count	#	Formula	Count	#	Formula	Count	#	Formula	Count
1	TiO_2	5541.8	24	CoO	297.3	47	CrO_3	15.3	1	$SrCO_3$	3989.5
2	Fe_2O_3	3179.4	25	CaO	292.2	48	MoO_2	14.5	2	$BaCO_3$	3977.6
3	Nb_2O_5	2409.6	26	MnO	219.7	49	Nb_2O_3	9.0	3	$CaCO_3$	2896.2
4	Al_2O_3	1760.2	27	Cu_2O	157.1	50	Re_2O_7	9.0	4	Li_2CO_3	2404.5
5	ZrO_2	1676.3	28	Mn_3O_4	144.8	51	CrO_2	8.0	5	Na_2CO_3	1494.6
6	ZnO	1622.9	29	SrO	127.8	52	OsO_2	7.0	6	$MnCO_3$	573.8
7	CuO	1520.2	30	Ag_2O	118.5	53	Re_2O_3	6.5	7	$Fe(CO_2)_2$	164.8
8	MgO	1373.3	31	BaO	98.6	54	ReO_3	6.0	8	$MgCO_3$	125.0
9	MnO_2	1328.5	32	Li_2O	97.0	55	Ta_2O_3	6.0	9	$CoCO_3$	90.2
10	Co_3O_4	1201.4	33	HfO_2	90.2	56	CuO_2	6.0	10	Cs_2CO_3	83.8
11	NiO	874.6	34	BaO_2	73.6	57	Mo_2O_3	6.0	11	Rb_2CO_3	64.8
12	Ga_2O_3	846.2	35	IrO_2	66.0				12	K_2CO_3	35.0
13	Ta_2O_5	763.4	36	Rh_2O_3	60.2				13	$NiCO_3$	32.8
14	V_2O_5	750.9	37	Fe_3O_4	57.1				14	$Co(CO_2)_2$	18.0
15	MoO_3	724.8	38	SrO_2	52.0				15	Sr_2CO_3	13.0
16	Mn_2O_3	670.7	39	V_2O_3	51.0				16	$Mn(CO_2)_2$	10.8
17	WO_3	628.3	40	Na_2O_2	37.2				17	Ag_2CO_3	6.9
18	Cr_2O_3	525.5	41	Na_2O	32.0				18	$Ni(CO_2)_2$	6.7
19	SnO_2	470.8	42	PdO	31.0				19	$Sr(CO_3)_2$	6.0
20	In_2O_3	416.1	43	SnO	26.2				20	$ZnCO_3$	6.0
21	RuO_2	341.6	44	VO_2	20.5				21	$Sr(CO_2)_2$	6.0
22	GeO_2	338.5	45	FeO	19.0				22	$Co_2(CO_3)_3$	5.8
23	Co_2O_3	305.5	46	Li_2O_2	16.8				23	$Sn(CO_2)_2$	5.0

A.2. R_{synth} distribution of each multi-property optimization

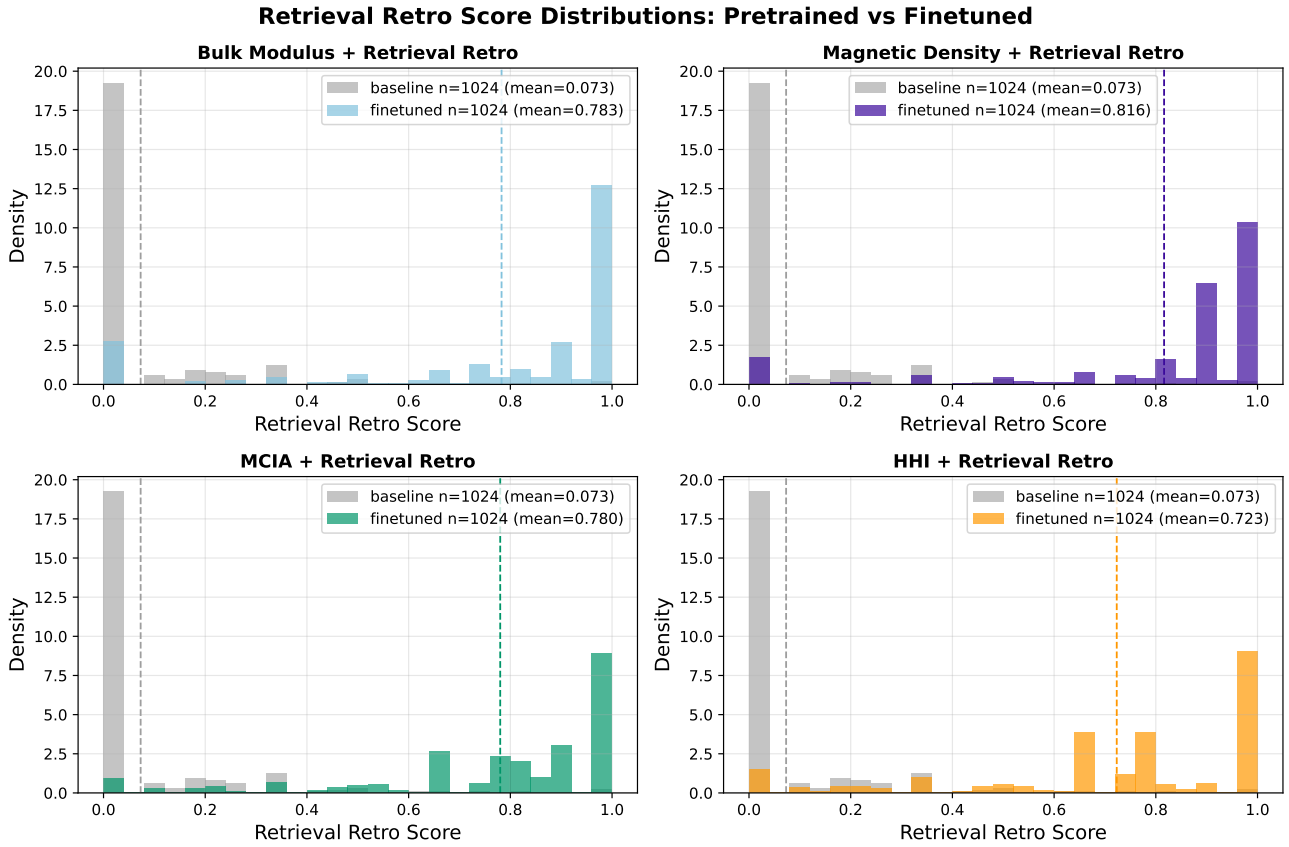

Figure 6. Distribution of Retrieval-Retro scores for 1,024 samples generated by the pretrained baseline (gray) and by each multi-objective fine-tuned model (colored), across the four property pairings: bulk modulus, magnetic density, MCIA, and HHI

A.3. Details Stable, Unique, Novel components of SUN Ratio

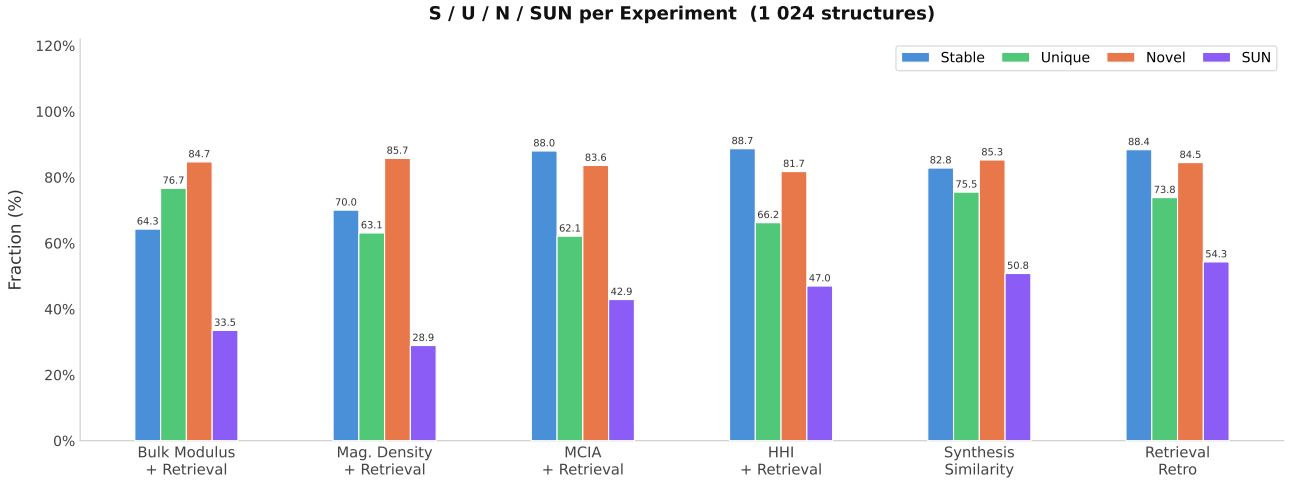

Figure 7. Details of Each Component: Stable, Unique and Novel in SUN Ratio of 1,024 Generated Samples from Finetuned Models

A.4. Implementation Details

The code will be published under MIT license upon paper accepted