THE MATERIALS SCIENTIST AGENT: REINFORCEMENT LEARNING FOR CRYSTAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Discovering new inorganic crystals can be seen as a search problem, where the evaluation reveals if a candidate is stable, but not how to improve it. For this purpose, we develop an LLM-based agent and train a 4B model to do targeted materials search. We give the model computational tools, have it propose candidates in a Wyckoff representation, check their novelty, and evaluate their stability and properties over multiple turns. To stop the agent from resubmitting one good material, we score a group of attempts together and split credit between repeated structures. We evaluate on in-domain compositions from MP-20, out-of-domain from WBM, and band-gap targets, judging every structure with MLIPs not used in training. Reinforcement learning nearly doubles the rate of stable, unique, and novel (S.U.N.) structures over supervised fine-tuning, on known and unknown compositions, and the resulting 4B agent outperforms GLM-5.3 Flash and nearly matches Gemini 3.1 Pro, the teacher model. For band-gap targets, it finds up to 7$\times$ more S.U.N. wide-gap materials than a state-of-the-art language model for crystal generation. We also show how the agent can be used to search for lead-free perovskites for solar cells.

1 INTRODUCTION

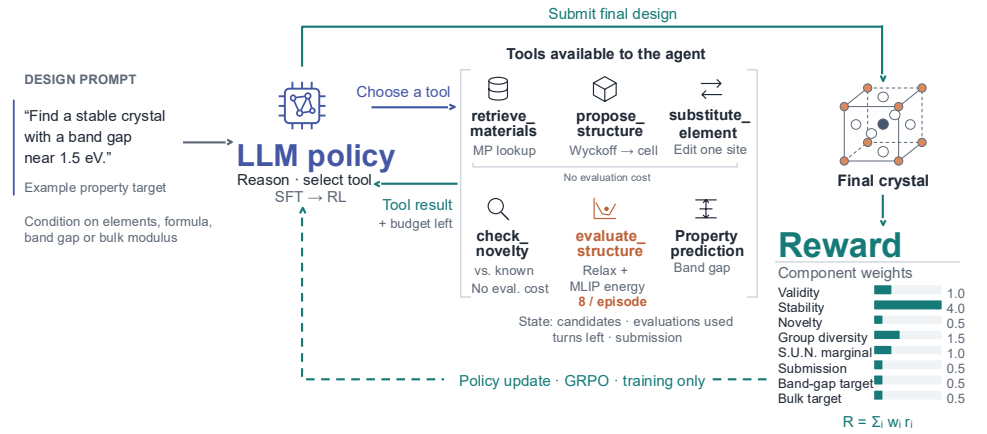
The discovery of new materials with certain properties is central to advances in energy conversion (Green et al., 2014; Zhao et al., 2020), electronics (Chhowalla et al., 2016), and structural engineering (George et al., 2019). For inorganic crystals, this requires navigating a large space of chemical compositions and atomic arrangements to identify structures that are both stable and suited to a desired application. This search is inherently iterative, where scientists start from known materials, propose candidates, evaluate their properties, and refine their hypotheses.

AI agents use language models to select and execute actions towards a goal, interacting with external tools and using their feedback to decide their next action (Yao et al., 2023). This loop makes them a great fit for scientific workflows that require iterative reasoning and evaluation. In the context of materials design, an agent can combine tools from information retrieval, crystal structure generation, or computational property prediction.

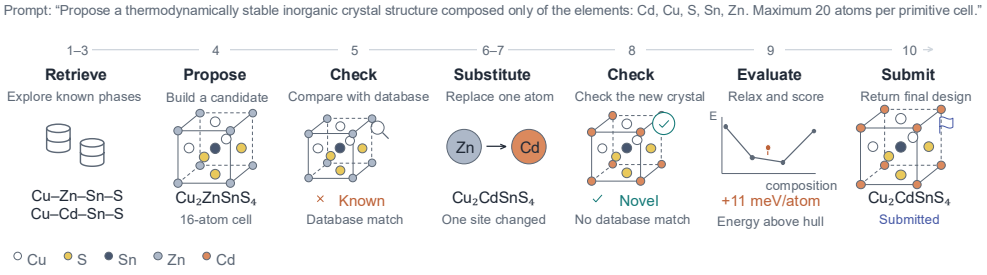
Computational approaches have accelerated materials discovery through several directions. Graph neural networks predict crystal stability and guide candidate selection through active learning, as in GNoME (Merchant et al., 2023). Diffusion-based generative models, such as MatterGen (Zeni et al., 2025), propose crystal structures conditioned on desired properties, while language-model approaches such as PLaID++ (Xu et al., 2025) improve crystal generation through post-training. More recently, MatAgent (Takahara et al., 2025) combines structure generation and property prediction tools in an iterative, feedback-driven workflow.

Building on these advances, effective materials discovery requires learning how to guide the model to search relevant areas of crystal space. While evaluating a candidate can indicate whether it is promising, it does not directly provide a signal on how to improve it or which structure to explore next. The underlying problem is therefore how to use feedback from individual evaluations to guide a search toward diverse, previously unknown materials. This raises our main questions: *Can an AI agent learn to turn computational feedback into an effective strategy for crystal discovery? And can reinforcement learning teach the language model to recursively improve its discovery?*

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.



(a) Training of the agent over multiple turns and tools until it submits a crystal, scored by the reward.



(b) Sample episode of multiple turns in the agent.

Figure 1: Framework of the LLM-based agent to discover crystals with computational tools.

In this work, we develop an LLM-based agent framework to discover inorganic crystal structures, and train a Qwen3.5-4B model (Bai et al., 2023) on it. The agent interacts with the environment through tools that allow it to retrieve known structures, propose candidates in a Wyckoff-based text representation, substitute elements, check novelty, evaluate stability and target properties, and submit a final structure. We evaluate this agent's capabilities across open-ended crystal discovery, composition-conditioned design, and property-conditioned band gap generation. The implemented reward combines structural validity, constraint satisfaction, stability, novelty, and diversity, encouraging the agent to discover promising materials rather than repeatedly generating the same candidate.

In our experiments, reinforcement learning nearly doubles the rate of stable, unique, and novel (S.U.N.) structures compared with SFT, mostly by learning to use more search budget. This gain holds for compositions from WBM that the agent never saw during training, and brings our 4B model close to Gemini 3.1 Pro, its teacher model. We judge all structures with held-out MLIPs not used in training, and validate a subset of the model's discoveries with DFT, where 23 out of 24 lie close to the hull. For band-gap targets, the agent finds several times more S.U.N. materials than PLaID++ (Xu et al., 2025). Finally, we demonstrate the agent's usefulness in real world materials' design, where our agent searches for lead-free perovskites for solar cells.

Our contributions are summarized as follows:

- **An LLM-based agent for crystal discovery.** We build an environment where an agent can look up known materials, propose new structures as text, check whether they are novel, and estimate their stability and properties before submission.
- **A small agent that competes with larger models.** We train Qwen3.5-4B with fine-tuning and then reinforcement learning, with a reward that scores attempts together so that repeating a material is penalized. It nearly matches Gemini 3.1 Pro on composition-conditioned discovery, and DFT confirms almost all the discoveries we check.
- **Conditioned materials discovery.** For band-gap targets, the agent finds up to seven times more stable, unique, and novel wide-gap materials than the previous best language model, and we use it to search for lead-free perovskites for solar cells.

2 Related Work

Generative Models for Crystal Structures Previous works, like CDVAE (Xie et al., 2022), used variational autoencoders to generate crystals by learning the distribution of known stable materials and sampling from it, later replaced by diffusion and flow matching (Jiao et al., 2023; Miller et al., 2024). MatterGen (Zeni et al., 2025) added conditioning on a target property. In all of them a structure is produced in one pass, so what a stability calculation says about it cannot change how the next one is generated.

Language Models for Crystal Generation With the growing capabilities of language models, they can now write the structure as text. To take advantage of this, practitioners fine-tune them to produce it either as a CIF file (Antunes et al., 2024; Gruver et al., 2024) or through Wyckoff representations, which encode the symmetry of the crystal and make the output shorter (Zhu et al., 2024; Kazeev et al., 2025). PLaID++ (Xu et al., 2025) is closer to our work as it post-trains a model on preference pairs ranked by an interatomic potential, so the model does learn from computational feedback. The feedback is collected in advance and at inference time the model writes a single structure.

LLM Agents for Science LLM Agents use tools and reason on their feedback (Yao et al., 2023). There is a great interest to connect them to scientific tools so they can run autonomous scientific workflows. ChemCrow (M. Bran et al., 2024) and Coscientist (Boiko et al., 2023) plan and run chemistry experiments, LLaMP answers materials questions by querying the Materials Project (Chiang et al., 2025), and MatAgent alternates between generating a structure and predicting its properties (Takahara et al., 2025). These works show how tool feedback helps. However, they all prompt fixed models that lose the information learned after every run.

Reinforcement Learning for Molecular Design and Tool Use Reinforcement learning has been proven effective in molecular design for some use cases. Olivecrona et al. (2017) optimizes a policy against a computed score and GFlowNets (Bengio et al., 2021) were introduced to find many different good candidates rather than a single best one, which is what our reward also aims for. Separately, reinforcement learning with verifiable rewards is now standard for post-training language models (Shao et al., 2024; Guo et al., 2025), including multi-turn settings where the model calls tools and is rewarded at the end (Jin et al., 2025). In our work, we bring the two together and train a language model to search for crystals with tools.

3 Background

Crystallography A crystal consists of atoms arranged in a repeating three-dimensional pattern. Its unit cell is described by three lengths (l_1, l_2, l_3), three angles $(\theta_1, \theta_2, \theta_3)$, and the elements and positions of its atoms. Each atom has an element e_i and fractional coordinates (x_i, y_i, z_i) relative to the cell. Symmetry makes this description more compact. A crystal's space group specifies operations that relate different atomic sites, while Wyckoff positions allow these sites to be described through representatives (Hahn et al., 1983). Applying the symmetry operations recovers the remaining atoms. This provides a compact text representation for language models, as used in PLaID++ (Xu et al., 2025).

A proposed structure must be evaluated to determine whether it is energetically favorable. First, its atomic positions and cell are adjusted to a local energy minimum, a process called relaxation. Its energy can then be compared with competing materials of the same overall composition. The energy above the convex hull, E_{hull}, measures the difference per atom from the lowest-energy combination of reference materials. A positive value means that decomposition could lower the energy, although some such structures can persist in a metastable state (Sun et al., 2016). Density functional theory (DFT) provides the energies, forces, and stresses needed for these calculations, but is expensive for large numbers of candidates. Machine-learned interatomic potentials (MLIPs) approximate these quantities using models trained on DFT data, making repeated evaluations practical (Yang et al., 2024; Fu et al., 2025; Rhodes et al., 2025).

Reinforcement Learning in LLMs Reinforcement learning (RL) trains a language model to generate responses that receive higher rewards. Given a prompt x, the model, or policy π_θ, generates a response y, and a reward function scores it. Unlike supervised fine-tuning (SFT), which teaches the model to imitate demonstrations, RL updates the model toward responses with higher scores. Group Relative Policy Optimization (GRPO) (Shao et al., 2024) samples G responses to the same prompt from the previous policy π_{old} and compares each reward R_i with the rest of the group through the normalized advantage $\hat{A}_i = (R_i - \bar{R})/\sigma_R$, where $\bar{R}$ and σ_R are the group's mean and standard deviation, so that responses better than average get a positive advantage. Its objective is

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{L_i}\sum_{t=1}^{L_i}\Big[\min\Big(\rho_{i,t}\hat{A}_i, \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)\hat{A}_i\Big) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})_{i,t}\Big]\right], \tag{1}$$

where the expectation is over prompts and sampled responses, L_i is the length of response i, and $\rho_{i,t} = \pi_\theta(y_{i,t} \mid x, y_{i,<t})/\pi_{\text{old}}(y_{i,t} \mid x, y_{i,<t})$ is the ratio of the current and previous policies' probabilities for token $y_{i,t}$. Maximizing this objective encourages tokens from higher-reward responses and discourages those from lower-reward responses. Clipping with ϵ limits the incentive for large probability changes. The Kullback–Leibler (KL) penalty, weighted by β, discourages the policy from moving too far from a reference model π_{ref}.

4 METHOD

We train a language model to search for new crystal structures using computational tools (Figure 1). The agent can propose a candidate, evaluate it, and use the result to decide what to try next. Training starts with examples of this process, followed by reinforcement learning that rewards the quality and diversity of the structures found.

4.1 PROBLEM FORMULATION

A task is a prompt x asking for a stable inorganic crystal. The prompt may add a constraint c: a set of elements or a chemical formula (composition-conditioned tasks), or a band-gap range (property-conditioned tasks). In unconditional tasks, the agent chooses both the composition and the structure. Appendix A.3 shows example prompts.

The agent π_θ solves a task in an *episode* of at most T turns. At each turn, it writes its reasoning and calls tools, and the tool results are added to the conversation. The episode ends when the agent submits a structure s. We count s as a *discovery* if it is (i) stable, $E_{\text{hull}}(s) \leq \delta$; (ii) novel, meaning it matches no known structure; and (iii) consistent with c. Finding the same material twice counts as one discovery, so the goal is for repeated episodes on the same prompt to find as many *different* discoveries as possible.

4.2 AGENT AND ENVIRONMENT

The agent decides which tools to call and in what order. A typical episode retrieves known materials related to the request, proposes a candidate, checks that it is new, evaluates its stability, and then revises the candidate or submits it. Candidates are stored under identifiers, so the agent can return to any structure proposed earlier in the episode. Appendix B.3 shows a complete episode. The tools support the following actions:

- **Retrieve:** look up Materials Project entries for a formula or set of elements, including their stability and available properties.
- **Propose and modify:** create a candidate in the Wyckoff representation described in Section 3, or replace an element at an atomic site in an existing candidate. Both actions check charge balance.
- **Check novelty:** compare a candidate with known Materials Project structures of the same composition.

- **Evaluate:** relax a candidate with the reward MLIP, EquFlashV2, and return its formation energy and energy above the hull. A separate tool predicts its band gap.
- **Submit:** select one candidate as the final answer and end the episode.

Only charge-balanced candidates can be evaluated or submitted. Since relaxation and property calculations are expensive, each episode has a limit on the number of turns T and on successful calls to each of these tools. The stability tool uses the same potential and relaxation settings as the reward calculation, so its feedback directly informs the agent about the quantity being optimized.

4.3 Reward Design

The goal is to reward the agent when finding structures that are stable, novel, and satisfy the request. Scoring each candidate independently, however, can reward repeated discoveries since an agent may receive a high score for submitting the same material across many episodes. Our reward therefore considers both the quality of a candidate and whether it adds a different structure to the discoveries made across episodes.

Candidate quality. At the end of each episode, we score the final candidate with four terms. The validity term r_{val} is 1 if the structure parses, is charge-balanced, and matches its space group, and 0 otherwise. The novelty term r_{nov} is one minus the structural similarity to the closest Materials Project entry (Jain et al., 2013). The stability term r_{stab} relaxes the candidate and decreases smoothly from 1 on the hull as E_{hull} grows, so the agent receives a signal before it finds stable structures. It is close to 0 for candidates that match a known structure or do not use exactly the requested elements or formula. For property-conditioned tasks, the property term r_{prop} is 1 when the predicted band gap meets the target and decreases with the distance to it.

Credit for distinct discoveries. We sample a group of G episodes for each prompt and compare their final structures with pymatgen's StructureMatcher (Ong et al., 2013). A candidate qualifies if it is novel, satisfies the request, and has $E_{\mathrm{hull}} \leq 0.1$ eV/atom. When several episodes find the same qualifying structure, they share the credit for that discovery. For a qualifying candidate i, the discovery term is

$$r_{\mathrm{disc}} = \frac{1}{k_i}, \tag{2}$$

where k_i is the number of matching structures in its group; candidates that do not qualify receive zero. For example, four occurrences of one qualifying structure receive $1/4$ each, while four different qualifying structures receive one each. Summed over the group, r_{disc} counts distinct qualifying structures.

Encouraging further exploration. The diversity term r_{div} is a fixed bonus for candidates with no structural match in their group, even before they are stable, which encourages exploration while stable discoveries remain rare. We also penalize repeats. Each candidate is identified by its formula and space group, and we keep a history of the last H submissions scored during training. If the identifier is shared with another candidate in the group or appears in the history, r_{stab} is multiplied by 0.05; if it appears in the history, r_{disc} is set to 0. If Materials Project already has a stable structure with the same composition, both r_{stab} and r_{disc} are multiplied by 0.3.

Finally, a submission term r_{sub} encourages an explicit final answer. The reward for episode i is the weighted sum of the terms described above, $R_i = \sum_j \lambda_j r_j$, with the weights λ_j and the exact form of each term given in Appendix A.1.

4.4 Training Procedure

Supervised fine-tuning. We first collect episodes from a teacher model using the same tools and environment. We retain trajectories that end with an explicit submission, pass the validity and novelty filters, and have a successfully computed stability score. The agent is trained to reproduce the teacher's reasoning and tool calls using the next-token prediction loss, with the teacher's reasoning placed in the agent's own thinking block. Tool responses provide context but are excluded from the loss. This stage teaches the agent how to use the tools and carry a search through to a final submission.

Reinforcement learning. Starting from the fine-tuned model, we sample G episodes for each prompt and score them using the reward in Section 4.3. Following GRPO (Shao et al., 2024), the advantage of episode i is its reward minus the group mean, $A_i = R_i - \frac{1}{G}\sum_{j=1}^{G} R_j$. Groups in which all episodes receive the same reward have zero advantage and are skipped. The advantage is assigned to the tokens generated by the agent in that episode; tool responses are excluded from the loss. We use the default loss of `prime-rl`, which differs from Equation 1 in two ways. It drops tokens whose probability has changed too much instead of clipping the ratio, and it keeps the model close to the policy that generated the episodes rather than to a reference model.

5 Experimental Setting

Training. We train Qwen3.5-4B (Bai et al., 2023) with `prime-rl` by Prime Intellect (Intellect, 2025). The SFT teacher is Gemini 3.1 Pro, and filtering leaves about 1,100 episodes collected on training prompts disjoint from the evaluation prompts. RL prompts are 50% element-set, 20% formula, 10% unconditional, and 20% property-conditioned tasks, with $G = 16$ episodes per prompt and a history of $H = 4{,}096$ submissions. Each episode allows at most $T = 10$ turns, 8 stability evaluations, and 8 band-gap predictions. Other hyperparameters are listed in Appendix A.2.

Tasks. We evaluate on four tasks with 8 episodes per prompt, sampled at temperature 1.0: 100 prompts ($100 \times 8 = 800$ episodes) for in-domain and out-of-domain, 125 prompts ($125 \times 8 = 1{,}000$ episodes) in band-gap so that its counts per 1,000 submissions are measured and not extrapolated, and 10,000 structures on the unconditional task to match PLaID++. Composition-conditioned tasks come in two versions: *in-domain*, with compositions from the MP-20 validation split (Xie et al., 2022), and *out-of-domain*, with compositions from WBM (Wang et al., 2021), which are not seen during training. Unconditional prompts leave the composition free. The band-gap task uses the targets from PLaID++ (Xu et al., 2025). Its prompts ask for a wide (≥ 3 eV), moderate (1–3 eV), or low (0.1–1 eV) gap.

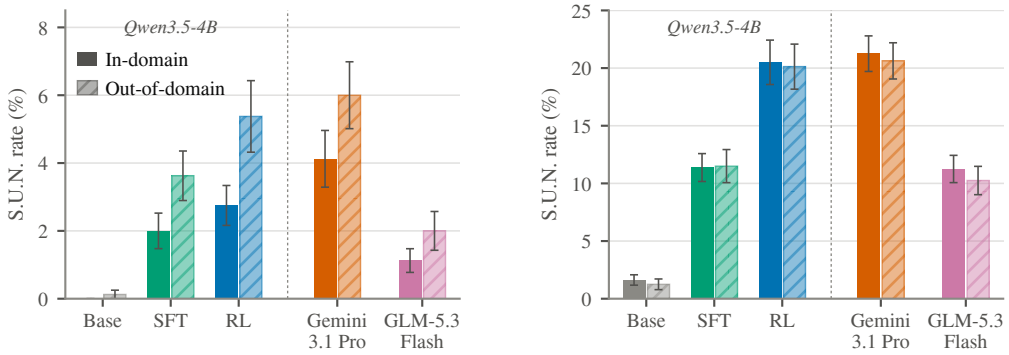
Evaluation. We relax each submitted structure for up to 500 steps with two MLIPs not used in training, Orb-v3 (Rhodes et al., 2025) and eSEN (Fu et al., 2025), and compute E_{hull} against the Materials Project hull (Jain et al., 2013). We report the S.U.N. rate (Zeni et al., 2025), the fraction of submissions that are stable, unique, and novel, where a structure that matches another submission counts only once. A structure is stable only if both MLIPs agree, and we report two thresholds: $E_{\text{hull}} \leq 0$, as in PLaID++, and $E_{\text{hull}} < 0.1$ eV/atom, the metastable threshold used by the reward. Novelty is checked against all of Materials Project, except in the unconditional task, which follows the protocol of PLaID++ with eSEN alone and novelty against the MP-20. For the band-gap task, we predict the band gap of each structure with the same model as the agent's tool, a band-gap head on MatterSim (Yang et al., 2024), and with MEGNet (Chen et al., 2019), which the agent never sees.

Baselines. On the composition-conditioned tasks, we compare the base, SFT, and RL versions of the same model, which isolates what each training stage contributes, together with two frontier models, Gemini 3.1 Pro and GLM-5.3 Flash, prompted with the same tools and the same turn budget. Gemini 3.1 Pro is also the SFT teacher, so it indicates what the agent imitates before RL. The remaining comparison depends on the task. PLaID++ (Xu et al., 2025) writes a structure in a single pass from a prompt that fixes a space group or a property target, so it applies to the unconditional and band-gap tasks but not to the composition-conditioned ones. For these two tasks we compare against its published numbers, as percentages over 10,000 samples for the unconditional task and as counts per 1,000 submissions for the band-gap task.

6 Results

6.1 In-domain/Out-of-domain

We start with composition-conditioned tasks, where the prompt gives the elements or the formula and the agent has to find a stable, novel structure for it. Figure 2 compares the base, SFT, and RL versions of Qwen3.5-4B with two frontier models that use the same tools, and Table 4 in the appendix gives the exact numbers.



(a) S.U.N. at $E_{\text{hull}} \leq 0$

(b) S.U.N. at $E_{\text{hull}} < 0.1$

Figure 2: S.U.N. rates on composition-conditioned tasks, with in-domain compositions from MP-20 and out-of-domain compositions from WBM, under the held-out consensus. Solid bars are in-domain and hatched bars out-of-domain, and error bars are bootstrap standard errors over prompts. Exact values, with submission rates, are in Table 4.

The base model rarely submits a structure, and after SFT the agent submits almost always but finds few discoveries. RL nearly doubles the S.U.N. rate at the metastable threshold, from about 11% to 20%, both in-domain and on the WBM compositions it never saw during training, with smaller gains at $E_{\text{hull}} \leq 0$. Much of this gain comes from a larger search budget, which we examine in Figure 5. The reward MLIP gives rates close to the held-out consensus (Appendix B), so the agent is not simply exploiting the model it was trained on. The 4B agent ends up close to Gemini 3.1 Pro, the model it learned from during SFT. The two are level at the metastable threshold, Gemini is still ahead at $E_{\text{hull}} \leq 0$, and both do better than GLM-5.3 Flash.

6.2 Unconstrained

In unconditional tasks the prompt does not fix the composition, so the agent has to choose both the elements and the structure. We follow the evaluation from Xu et al. (2025). We relax the structures with eSEN, compute stability against the same Materials Project hull with $E_{\text{hull}} \leq 0$, check novelty against the MP-20 training set, and count each group of matching structures once. Table 1 compares 10,000 structures from our RL agent with the numbers reported in PLaID++ (Xu et al., 2025).

Table 1: Unconditional generation over 10,000 samples, as a percentage of all samples, where our model leads on stability.

	Validity (%)			
Method	Structural	Composition	Stable (%)	S.U.N. (%)
CDVAE (Xie et al., 2022)	**100.0**	86.7	1.6	–
DiffCSP (Jiao et al., 2023)	**100.0**	83.3	5.0	3.3
FlowMM (Miller et al., 2024)	96.9	83.2	4.6	2.8
FlowLLM (Sriram et al., 2024)	99.9	90.8	13.9	4.7
MatterGen-MP (Zeni et al., 2025)	–	–	13.0	–
CrystaLLM-70B (Antunes et al., 2024)	99.6	95.4	5.3	–
ADiT (Joshi et al., 2025)	99.7	92.1	15.4	5.3
PLaID++, SFT only (Xu et al., 2025)	99.8	91.2	7.2	3.6
PLaID++, DPO (Xu et al., 2025)	98.8	92.6	15.6	6.3
PLaID++ (Xu et al., 2025)	99.8	**97.3**	22.3	**7.7**
Qwen3.5-4B RL (ours)	99.8	91.1	**32.4**	3.0

Nearly all of the agent's structures are structurally valid, and it has the highest stability rate in the table, 32.4% under eSEN, but its S.U.N. rate of 3.0% is below PLaID++ (7.7%) and its DPO variant (6.3%), while close to DiffCSP and FlowMM. What holds it back is uniqueness. Its 3,237 stable structures reduce to only 318 distinct ones, because the agent keeps coming back to the same

structures when nothing in the prompt pushes it elsewhere. We leave this as a matter to future work on how to embed open-ended diversity in an agent without specific constraints.

6.3 Property-conditioned: Band-gap

The band gap decides whether a material behaves as a metal, a semiconductor, or an insulator, and asking for a range of band gaps is how one looks for materials for uses such as solar cells or high-voltage electronics. This task is harder for the agent than a composition constraint. The band gap cannot be read off a structure, so the agent has to pick the elements, estimate the gap with a tool, and try again when the estimate is off. Following PLaID++ (Xu et al., 2025), which we use as a baseline, Table 2 counts the submissions that are S.U.N. and have a band gap above each threshold. The agent finds more such materials than PLaID++ at every threshold, and the difference grows with the threshold, from more than twice as many above 1 eV to more than seven times as many above 3 eV. The counts are about the same with MEGNet. Part of the difference comes from the search, as our agent can check the band gap up to 8 times per episode while PLaID++ writes each structure in one pass. Under MEGNet, 33.6% of the submissions fall in the requested range, against about 13% of MP-20 structures. The moderate range is the hardest (13.9%), and in the low range the agent mostly finds structures that its tool, but not MEGNet, places in the range.

Table 2: Band-gap discoveries per 1,000 submissions, counting the structures that are S.U.N. at $E_{\text{hull}} \leq 0$ and have a band gap above each threshold, $\pm$ bootstrap standard error over prompts.

Model	Band gap	> 1 eV	> 2 eV	> 3 eV
PLaID++, no preference tuning	MEGNet	4.0	1.0	0.0
PLaID++, one iteration	MEGNet	15.0	9.0	3.0
Qwen3.5-4B RL (ours)	MatterSim	$36.0_{\pm 6.3}$	$29.0_{\pm 5.7}$	$23.0_{\pm 5.2}$
	MEGNet	$35.0_{\pm 6.1}$	$30.0_{\pm 5.8}$	$25.0_{\pm 5.3}$

6.4 Application: Lead-free perovskites

As an application, we showcase how the agent can search for lead-free perovskites. Halide perovskites such as $CsPbI_3$ are promising materials for solar cells and LEDs, and their band gap can be tuned by changing the elements they are made of. The best-performing ones contain lead, which is toxic, and finding lead-free alternatives with a similar band gap is still an open problem. We prompt the agent for a stable and novel perovskite (formula ABX_3) that contains no lead and has a band gap within 0.2 eV of $CsPbI_3$, $CsPbBr_3$, or $CsPbCl_3$, and run 512 episodes for each of the three targets. Almost all submitted structures (98.6%) are lead-free, and 24 episodes (1.6%) satisfy every condition, which gives 7 different materials, one of which, $CsMgI_3$, matches two targets (Table 8 in the appendix). All the matches for $CsPbI_3$ replace lead with tin or germanium, the elements researchers have studied most as lead substitutes. To see whether these compositions can actually form perovskites, we use the usual check based on the tolerance and octahedral factors, which compare the sizes of the ions. Three of the seven compositions, including both tin compounds, fall in the region where halide perovskites form, next to the lead compounds, and $CsBaI_3$ sits on its edge (Figure 3). The other three, $KGeI_3$ and the two magnesium iodides, have a B cation too small for the octahedron of iodide atoms and are unlikely to be perovskites. Self-consistent DFT places the two new compositions, $KGeI_3$ and $RbMgI_3$, within 0.1 eV/atom of the hull. More details, the structures, and an example episode are in Appendix D.

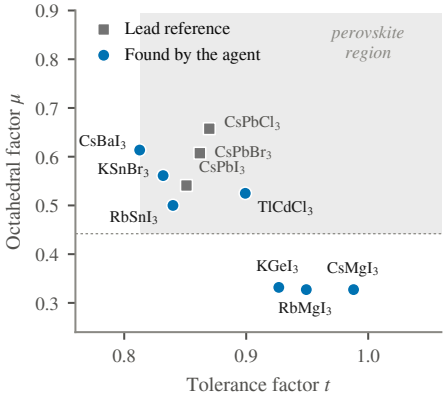


Figure 3: Tolerance factor t and octahedral factor μ of the lead compounds and of the lead-free materials found by the agent. The shaded area is where halide perovskites usually form.

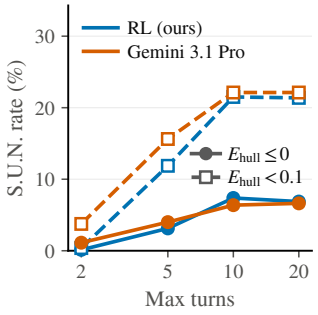


Figure 4: S.U.N. rate against the turn limit.

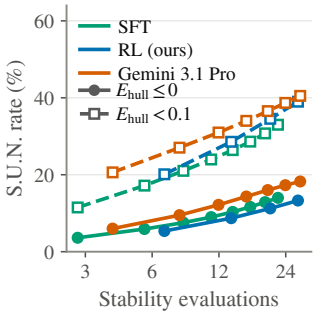


Figure 5: S.U.N. rate of the best of k episodes.

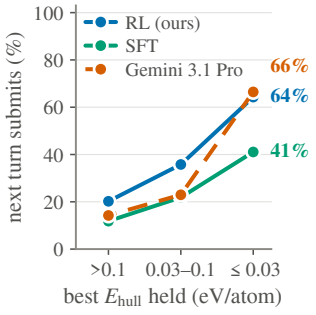


Figure 6: Submission rate against the best E_{hull} found.

7 Analysis and Ablations

Turn limit. We evaluate the agent with Qwen-3.5 4B-RL and Gemini 3.1 Pro on the out-of-domain task with limits of 2, 5, 10, and 20 turns, judging the structures with the reward MLIP (Figure 4). We see the agent requires multiple turns to start making discoveries. Its S.U.N. rate is close to zero with 2 turns as it uses 1 turn to submit. Then it grows up to 10 turns, where both reach the maximum of their discoveries. Gemini does better when the budget is short, which suggests that the RL agent has adapted its search to the budget it was trained with.

Search budget. After RL, the agent makes about 27 tool calls per episode, against 16 after SFT, and runs 6.8 stability evaluations instead of 2.8 (Appendix B.2). To see how much of its gain comes from this larger budget, we give the other models the same number of evaluations by sampling k episodes per prompt and keeping the submission with the lowest E_{hull} under the reward MLIP (Figure 5). At the same budget, SFT with this selection matches the RL agent at the metastable threshold and does better at $E_{\text{hull}} \leq 0$. RL thus teaches the agent to do in one episode what SFT needs several episodes and an external selection step for.

Reacting to tool feedback. To see how the agent reacts to what its tools return, we look at every turn where it still has evaluations left and check whether it submits on the next turn, given the best E_{hull} it has found so far (Figure 6). After RL, the agent submits in 64% of these turns when it holds a stable candidate and in only 20% when all its candidates are unstable, a larger difference than after SFT (41% against 12%).

DFT validation. All the results above rely on MLIPs, which are approximations. We therefore recomputed E_{hull} with DFT (Quantum ESPRESSO) for structures that the final RL agent submitted on the in-domain and out-of-domain tasks, leaving out those with f-electron elements, for which standard DFT is slow to converge and less reliable. Of the 24 structures that both held-out MLIPs call stable, 23 lie within 0.1 eV/atom of the hull in DFT, against only 2 of the 11 that only the reward MLIP calls stable. Our reported discoveries are confirmed by DFT, showing the value of using multiple MLIPs (Appendix C).

8 Conclusion

We built an agent environment for crystal discovery, in which a language model uses computational tools over several turns, as a materials scientist would. We trained Qwen3.5-4B in this environment, first with SFT and then RL. After RL, the agent finds almost twice as many stable, unique, and novel structures, also for compositions it never saw in training, and performs about as well as Gemini 3.1 Pro, the model it learned from. Its discoveries also hold up when checked with other MLIPs and with DFT. This model can be used as a cheap way to search for candidates with a given composition or property, as in our search for lead-free perovskites. Without constraints, the agent often returns to the same structures, which keeps its S.U.N. rate below that of the best generative models, and making its search more diverse is an important next step. In the long term, we see this work as a step toward agents that take a plain description of the material a scientist needs and keep working until they find one that meets every requirement.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data sets (teacher model traces), implement methods, clean and reformat datasets, and interpret some of the results. We have not used generative AI tools to help develop conceptual frameworks, formulate claims, or provide the critical elements of our claims and hypotheses.

Additionally, we used generative AI tools to create and edit software code, draft parts of this paper and edit it for readability, create and modify scientific figures, and identify and summarize existing literature. We have reviewed all AI-assisted work. The synthetic training data was generated by Gemini 3.1 Pro in our own environment and filtered as described in Section 4.4; the AI-written code was reviewed by the authors and exercised in the experiments reported here; every claim we draw from the evaluations was checked by the authors against the raw evaluation outputs; AI-drafted text was revised by the authors prior to the submission; and the cited references were referenced and added manually. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The use of autonomous tools in science has enormous potential to accelerate discovery, for example of new materials for renewable energy, electronics, or carbon capture. However, making generative tools publicly available could also make it easier to design dangerous compounds, so researchers should address these problems and find ways to mitigate these risks. This work does not involve human subjects or personal data, and all data were used under their licenses and terms of use.

REPRODUCIBILITY STATEMENT

We aim to make every part of our pipeline reproducible, from the environment to the evaluation. Section 4.2 describes the environment and its tools, and Section 4.3 and Appendix A.1 define the reward. Section 4.4, Section 5, and Appendix A.2 give the training recipe and hyperparameters, and Appendix A.3 lists the prompts for each task. RL training takes 32 hours on 8 A100 GPUs. The evaluation protocol, including the MLIPs, relaxation settings, S.U.N. criteria, and number of episodes per task, is in Section 5, standard errors are bootstrapped over prompts, and the DFT setup is in Appendix C. An official public release of the source code, trained model and the generated structures will be made after de-anonymization.

REFERENCES

Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):10570, 2024.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in neural information processing systems*, 34:27381–27394, 2021.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019.

Manish Chhowalla, Debdeep Jena, and Hua Zhang. Two-dimensional semiconductors for transistors. *Nature Reviews Materials*, 1(11):16052, 2016.

Yuan Chiang, Elvis Hsieh, Chia-Hong Chou, and Janosh Riebesell. LLaMP: Large language model made powerful for high-fidelity materials knowledge retrieval. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25189–25221, Suzhou,

China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1280. URL https://aclanthology.org/2025.emnlp-main.1280/.

Xiang Fu, Brandon M Wood, Luis Barroso-Luque, Daniel S. Levine, Meng Gao, Misko Dzamba, and C. Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=R0PBjxIbgm.

Easo P George, Dierk Raabe, and Robert O Ritchie. High-entropy alloys. *Nature reviews materials*, 4(8):515–534, 2019.

Martin A Green, Anita Ho-Baillie, and Henry J Snaith. The emergence of perovskite solar cells. *Nature photonics*, 8(7):506–514, 2014.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, Larry Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *International conference on learning representations*, volume 2024, pp. 21671–21690, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Theo Hahn, Uri Shmueli, and JC Wilson Arthur. *International tables for crystallography*, volume 1. Reidel Dordrecht, 1983.

Prime Intellect. prime-rl, 2025. URL https://github.com/PrimeIntellect-ai/prime-rl.

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.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Chaitanya K Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary W Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. *arXiv preprint arXiv:2503.03965*, 2025.

Nikita Kazeev, Wei Nong, Ignat Romanov, Ruiming Zhu, Andrey E Ustyuzhanin, Shuya Yamazaki, and Kedar Hippalgaonkar. Wyckoff transformer: Generation of symmetric crystals. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=eFHfRQRjJo.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature machine intelligence*, 6(5):525–535, 2024.

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.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):48, 2017.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013. doi: 10.1016/j.commatsci.2012.10.028.

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

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.

Anuroop Sriram, Benjamin K Miller, Ricky T Chen, and Brandon Wood. Flowllm: Flow matching for material generation with large language models as base distributions. *Advances in Neural Information Processing Systems*, 37:46025–46046, 2024.

Wenhao Sun, Stephen T Dacek, Shyue Ping Ong, Geoffroy Hautier, Anubhav Jain, William D Richards, Anthony C Gamst, Kristin A Persson, and Gerbrand Ceder. The thermodynamic scale of inorganic crystalline metastability. *Science advances*, 2(11):e1600225, 2016.

Izumi Takahara, Teruyasu Mizoguchi, and Bang Liu. Accelerated inorganic materials design with generative ai agents. *Cell Reports Physical Science*, 6(12), 2025.

Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Predicting stable crystalline compounds using chemical similarity. *npj Computational Materials*, 7(1):12, 2021.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=03RLpj-tc_.

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

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.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

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.

Qing Zhao, Sanjuna Stalin, Chen-Zi Zhao, and Lynden A Archer. Designing solid-state electrolytes for safe, energy-dense batteries. *Nature Reviews Materials*, 5(3):229–252, 2020.

Ruiming Zhu, Wei Nong, Shuya Yamazaki, and Kedar Hippalgaonkar. Wycryst: Wyckoff inorganic crystal generator framework. *Matter*, 7(10):3469–3488, 2024.

A Method Details

A.1 Reward Details

Table 3 gives the value, range, and weight of each reward term described in Section 4.3.

Table 3: Reward terms and weights. The property term is only active for property-conditioned tasks.

Term	Value	Range	λ_j
r_{val}	1 if the structure parses, is charge-balanced, and matches its space group; 0 otherwise	$\{0, 1\}$	1
r_{nov}	One minus the structural similarity to the closest Materials Project entry	$[0, 1]$	0.5
r_{stab}	$s(E_{\text{hull}})$ (Eq. 3) times the novelty factor; $\times 0$ if the constraint is violated, $\times 0.05$ for repeats in the group or history, $\times 0.3$ for compositions with a known stable structure	$[0, 1]$	4
r_{prop}	Band gap: 1 inside the requested range, linear to 0 at 1 eV outside it.	$[0, 1]$	0.5
r_{disc}	$1/k_i$ for a qualifying candidate (Eq. 2); 0 for repeats in the history, $\times 0.3$ for compositions with a known stable structure	$[0, 1]$	1
r_{div}	0.5 if no other candidate in the group matches the structure; 0 otherwise	$\{0, 0.5\}$	1.5
r_{sub}	$+1$ for an explicit submission, 0 if the best evaluated candidate is used instead, -0.5 if there is no valid candidate	$\{-0.5, 0, 1\}$	0.5

Stability score. The stability term r_{stab} starts from a score of the candidate's energy above the hull E (in eV/atom),

$$s(E) = \frac{1}{2} \exp\left(-\frac{E}{0.05}\right) + \frac{1}{2} \cdot \frac{1}{1 + E/0.3}. \tag{3}$$

The first term peaks sharply at the hull, and the second decays slowly so that structures far from the hull still receive a signal ($s = 0.44$ at 0.1 eV/atom and 0.19 at 0.5 eV/atom). This score is multiplied by the novelty factor $\sigma\big((r_{\text{nov}} - 0.55)/0.1\big)$, where σ is the logistic function, which is close to 0 for a match to a known structure and close to 1 otherwise.

A.2 Training Details

This section gives the hyperparameters not listed in Section 5.

Supervised fine-tuning. We train for three epochs with a batch size of 8 and a learning rate of 2×10^{-5}.

Reinforcement learning. Each step uses 8 prompts with $G = 16$ episodes each, sampled at temperature 1.15 with at most 2,048 generated tokens per model response. All other loss settings follow the prime-rl defaults.

A.3 Task Prompts

Each task fills a fixed template. The boxes below show one example prompt per task.

Element set

```
Propose a thermodynamically stable inorganic crystal structure composed
only of the elements:  Li, Mn, O. Maximum 20 atoms per primitive cell.
```

Formula

```
Propose a thermodynamically stable inorganic crystal structure with the
formula LiMnO2.
```

Unconditional

```
Propose a thermodynamically stable, novel inorganic crystal structure.
Maximum 20 atoms per primitive cell.
```

Band gap

```
Propose a thermodynamically stable inorganic crystal structure with
a moderate electronic band gap between 1.0 and 3.0 eV (PBE scale).
Maximum 20 atoms per primitive cell.
```

B Composition-Conditioned Results

B.1 Additional Tables

Table 4 gives the submission and S.U.N. rates plotted in Figure 2.

Table 4: Submission and S.U.N. rates (%) on composition-conditioned tasks, with in-domain (ID) compositions from MP-20 and out-of-domain (OOD) compositions from WBM. Subscripts are bootstrap standard errors over prompts.

	Submission		S.U.N. (≤ 0)		S.U.N. (< 0.1)	
Model	ID	OOD	ID	OOD	ID	OOD
Gemini 3.1 Pro	$99.2_{\pm 0.3}$	$\mathbf{99.8}_{\pm 0.2}$	$\mathbf{4.1}_{\pm 0.8}$	$\mathbf{6.0}_{\pm 1.0}$	$\mathbf{21.2}_{\pm 1.5}$	$\mathbf{20.6}_{\pm 1.6}$
GLM-5.3 Flash	$98.4_{\pm 0.6}$	$99.5_{\pm 0.4}$	$1.1_{\pm 0.3}$	$2.0_{\pm 0.6}$	$11.2_{\pm 1.2}$	$10.2_{\pm 1.2}$
Qwen3.5-4B						
Base	$32.6_{\pm 1.8}$	$31.1_{\pm 1.7}$	$0.0_{\pm 0.0}$	$0.1_{\pm 0.1}$	$1.6_{\pm 0.5}$	$1.2_{\pm 0.4}$
SFT	$94.0_{\pm 0.9}$	$95.2_{\pm 0.7}$	$2.0_{\pm 0.5}$	$3.6_{\pm 0.7}$	$11.4_{\pm 1.2}$	$11.5_{\pm 1.4}$
RL	$\mathbf{99.4}_{\pm 0.3}$	$\mathbf{99.8}_{\pm 0.2}$	$2.8_{\pm 0.6}$	$5.4_{\pm 1.1}$	$20.5_{\pm 1.9}$	$20.1_{\pm 2.0}$

Table 5 reports S.U.N. rates for each MLIP separately, over 800 episodes per task. EquFlashV2 is the reward MLIP and is shown for reference only; the consensus counts a structure as stable only if both held-out MLIPs, Orb-v3 and eSEN, agree. Table 6 breaks these results down into counts. At $E_{\text{hull}} \leq 0$, each column keeps the structures from the previous one that are also unique, and then novel. Converged is the percentage of relaxations that reach $f_{\max} = 0.02$ eV/Å within 500 steps. In both tables, the best result for each MLIP is in bold.

Table 5: S.U.N. rates (%) per MLIP on in-domain (ID) and out-of-domain (OOD) composition-conditioned tasks.

		S.U.N. (≤ 0)		S.U.N. (< 0.1)	
Model	MLIP	ID	OOD	ID	OOD
Gemini 3.1 Pro	EquFlashV2 (reward)	3.9	**6.4**	**22.2**	**22.1**
	Orb-v3	**4.8**	**6.9**	25.6	22.4
	eSEN	**4.8**	**7.0**	23.5	22.9
	Consensus	**4.1**	**6.0**	**21.2**	**20.6**
GLM-5.3 Flash	EquFlashV2 (reward)	2.2	2.2	12.5	11.5
	Orb-v3	2.0	3.2	16.5	16.4
	eSEN	1.8	3.1	14.2	13.0
	Consensus	1.1	2.0	11.2	10.2
Qwen3.5-4B					
Base	EquFlashV2 (reward)	0.4	0.8	1.6	2.4
	Orb-v3	0.8	0.4	4.2	3.1
	eSEN	0.6	0.2	2.9	1.9
	Consensus	0.0	0.1	1.6	1.2
SFT	EquFlashV2 (reward)	3.4	3.9	14.6	13.4
	Orb-v3	3.0	4.1	18.2	16.6
	eSEN	3.0	4.2	15.4	14.2
	Consensus	2.0	3.6	11.4	11.5
RL	EquFlashV2 (reward)	**4.4**	6.0	21.8	20.4
	Orb-v3	3.5	6.5	**25.8**	**23.5**
	eSEN	3.9	6.9	**24.4**	**23.0**
	Consensus	2.8	5.4	20.5	20.1

Table 6: Breakdown of discoveries per MLIP on in-domain and out-of-domain composition-conditioned tasks (counts out of 800 episodes).

Model	MLIP	$E_{\text{hull}} \leq 0$			$E_{\text{hull}} < 0.1$		
		Stable	Unique	Novel	Stable	S.U.N.	Converged
In-domain							
Gemini 3.1 Pro	EquFlashV2 (reward)	**81**	34	31	**446**	**178**	100
	Orb-v3	**97**	**41**	**38**	**492**	205	100
	eSEN	**103**	**42**	**38**	**487**	188	91
GLM-5.3 Flash	EquFlashV2 (reward)	28	20	18	198	100	97
	Orb-v3	27	20	16	252	132	98
	eSEN	29	21	14	244	114	77
Qwen3.5-4B							
Base	EquFlashV2 (reward)	5	3	3	18	13	97
	Orb-v3	9	7	6	40	34	98
	eSEN	13	8	5	39	23	67
SFT	EquFlashV2 (reward)	42	29	27	186	117	97
	Orb-v3	40	27	24	237	146	98
	eSEN	45	29	24	220	123	75
RL	EquFlashV2 (reward)	65	**38**	**35**	278	174	98
	Orb-v3	55	30	28	337	**206**	99
	eSEN	67	34	31	333	**195**	81
Out-of-domain							
Gemini 3.1 Pro	EquFlashV2 (reward)	**133**	**55**	**51**	**419**	**177**	100
	Orb-v3	**148**	**58**	**55**	**427**	179	100
	eSEN	**146**	**58**	**56**	**436**	183	97
GLM-5.3 Flash	EquFlashV2 (reward)	27	20	18	157	92	99
	Orb-v3	37	28	26	200	131	99
	eSEN	35	26	25	194	104	86
Qwen3.5-4B							
Base	EquFlashV2 (reward)	6	6	6	23	19	97
	Orb-v3	3	3	3	32	25	97
	eSEN	3	3	2	26	15	65
SFT	EquFlashV2 (reward)	60	34	31	163	107	99
	Orb-v3	64	35	33	199	133	99
	eSEN	61	34	34	187	114	81
RL	EquFlashV2 (reward)	76	49	48	259	163	99
	Orb-v3	89	53	52	290	**188**	99
	eSEN	91	56	55	296	**184**	88

B.2 Tool Usage

Figure 7 shows how many times each model calls each tool per episode on the out-of-domain task.

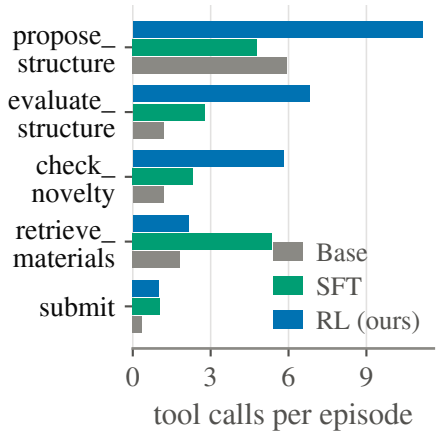


Figure 7: Tool calls per episode, by tool and model, on the out-of-domain task.

B.3 Example Trajectory

Figure 8 shows one complete episode of the RL agent on an in-domain composition-conditioned prompt.

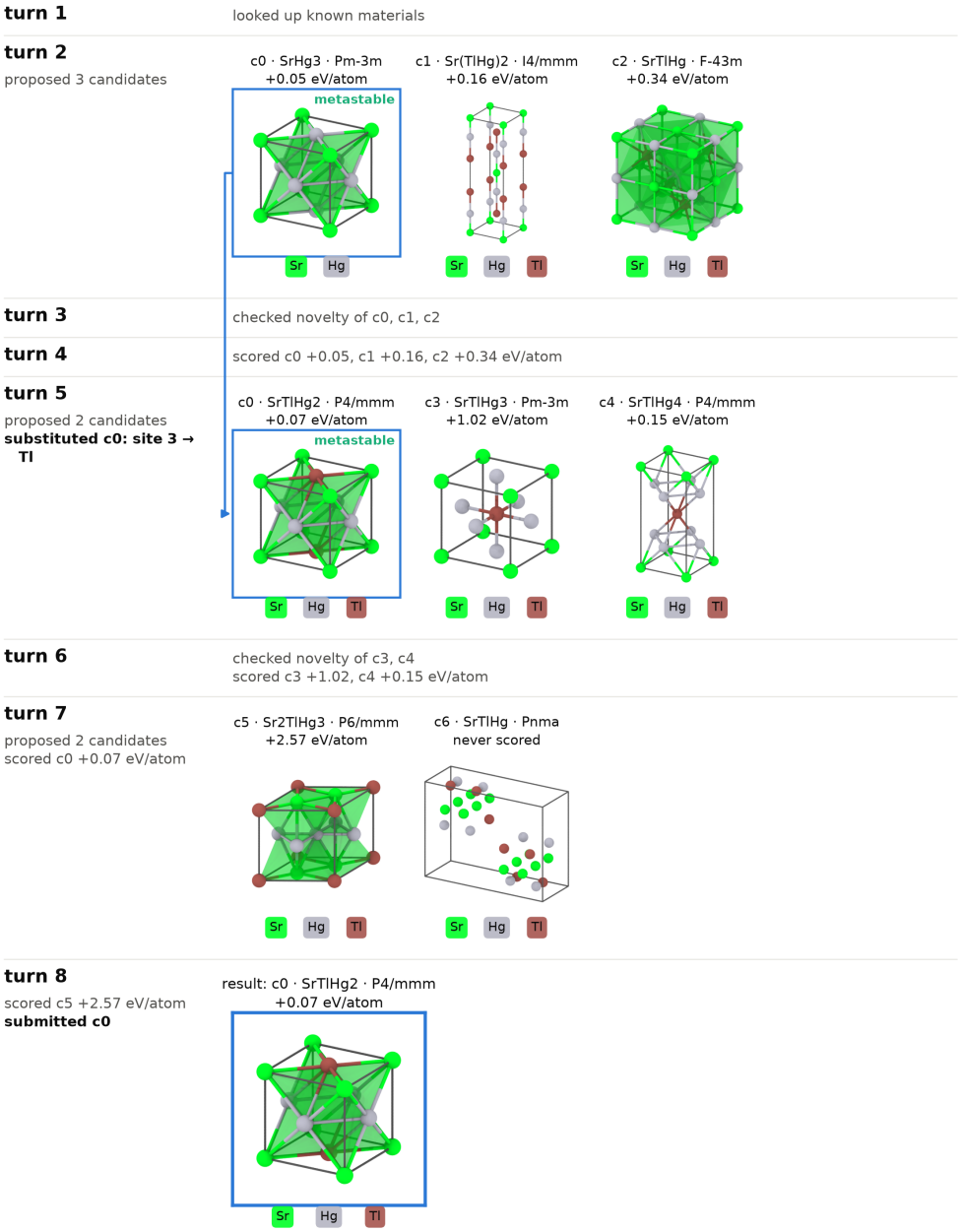


Figure 8: Multi-turn trajectory of the Qwen3.5-4B RL agent on an in-domain prompt with target composition $SrTlHg_2$. Each row is one turn, labelled with the tool calls made in that turn; the crystals in a row are the candidates created there, annotated with the energy above hull returned by the in-episode MLIP. The submitted structure is shown again as the result in the turn in which it was submitted. Blue frames and arrows mark the submitted candidate: proposed as $SrHg_3$ (Pm-3m) at turn 2, edited by substituting site 3 to Tl at turn 5, re-scored at turn 7 and submitted at turn 8. Renders: OVITO, with cation coordination polyhedra.

C DFT VALIDATION

Table 7 lists the structures from the final RL agent that we checked with DFT, from the in-domain and out-of-domain tasks and without f-electron elements. We relax each structure with Quantum ESPRESSO and compute its energy above the hull from its DFT energy, the DFT energies of the elements, and the Materials Project hull. Since this overestimates E_{hull} for halides and sulfides, for CdI_2 and $CrAgS_2$ we also compute the competing phases with DFT. On 16 known materials, 9 on the hull and 7 above it, this setup classifies all of them correctly. Of the 24 structures that both held-out MLIPs call stable, 23 lie within 0.1 eV/atom of the hull, while only 2 of the 11 that only the reward MLIP calls stable do.

Table 7: DFT check of structures from the final RL agent. MLIP columns give E_{hull} (eV/atom) after relaxation, with the held-out value being the larger of Orb-v3 and eSEN. $\leq$ marks an upper bound from a relaxation that did not finish, and failed means that DFT did not converge. †Oxide, for which our faster DFT setup overestimates E_{hull}. ‡From a self-consistent calculation that also computes the competing phases with DFT.

Task	Formula	Reward MLIP	Held-out MLIP	DFT	Verdict
Stable under both held-out MLIPs (23 of 24 within 0.1 eV/atom)					
ID	CdI_2	-0.022	-0.018	$\leq 0^{\ddagger}$	within 0.1
ID	$CrAgS_2$	-0.004	-0.004	$\leq 0^{\ddagger}$	within 0.1
ID	Rb_3AgF_6	-0.043	-0.047	≤ 0.256	not finished
ID	$ScIr_2$	-0.026	-0.019	≤ -0.021	within 0.1
ID	SrTlHg	0.001	-0.000	≤ 0	within 0.1
ID	TaGeIr	0.076	-0.083	≤ 0	within 0.1
ID	Ti_2Ru	-0.013	-0.009	≤ 0	within 0.1
ID	ZrNiPb	-0.064	-0.071	≤ 0	within 0.1
OOD	$Ba(SnRh)_2$	-0.095	-0.092	≤ -0.108	within 0.1
OOD	BaSnRh	-0.082	-0.089	≤ -0.103	within 0.1
OOD	$CaWN_2$	-0.282	-0.285	≤ 0	within 0.1
OOD	$CaWN_2$	-0.099	-0.103	0.012	within 0.1
OOD	$CaWN_2$	-0.091	-0.091	0.037	within 0.1
OOD	CsInBi	-0.015	-0.013	≤ 0	within 0.1
OOD	NbVGe	-0.034	-0.032	≤ 0	within 0.1
OOD	Sc_2Al_3Ir	-0.076	-0.072	≤ 0	within 0.1
OOD	ScAlIr	-0.008	-0.016	≤ 0	within 0.1
OOD	$ScIr_2$	-0.026	-0.019	≤ 0	within 0.1
OOD	$Sr(SiNi)_2$	-0.040	-0.036	≤ 0	within 0.1
OOD	$Y(PIr)_2$	-0.064	-0.048	≤ 0	within 0.1
OOD	YPIr	-0.017	-0.026	≤ 0	within 0.1
OOD	YPIr	0.003	-0.002	≤ 0	within 0.1
OOD	ZrAsRh	-0.064	-0.098	≤ 0	within 0.1
OOD	ZrSbIr	-0.145	-0.200	≤ 0	within 0.1
Stable under the reward MLIP only (2 of 11 within 0.1 eV/atom)					
ID	$CaMn(Ir_4O_5)_2$	-526	14	–	failed
ID	Li_2GaPd_2	-11	0.991	–	failed
ID	Li_2GaPd_3	-8.600	12	0.385	above
ID	Li_2TiZnO_4	-0.164	1.636	–	failed
ID	Li_2TiZnO_5	-4.205	0.406	–	failed
ID	Li_6Ga_6Pd	-329	13	–	failed
ID	$LiSb_2O_7$	-0.305	0.812	1.453	above†
ID	$LiTiO_3$	-3.576	11	0.938	above†
ID	$LiZn_2$	-0.003	0.003	≤ 0	within 0.1
OOD	ScGeIr	-0.001	0.008	0.011	within 0.1
OOD	Ti_2Rh_2Se	-700	14	–	failed

D LEAD-FREE PEROVSKITES

Setup. We use three lead compounds as references, $CsPbI_3$ (1.50 eV), $CsPbBr_3$ (2.16 eV), and $CsPbCl_3$ (2.28 eV). We predict their band gaps from their Materials Project structures with the same model the agent uses as a tool, so any systematic error of the model affects the target and the candidates in the same way. An episode counts as a success if the submitted structure contains no lead, has ABX_3 stoichiometry, has a band gap within 0.2 eV of the reference, is stable ($E_{\text{hull}} < 0.1$ eV/atom under the reward MLIP), and is novel (similarity below 0.9 to the closest known structure). We run the Qwen3.5-4B RL agent for 512 episodes per reference, with 10 turns and 6 stability evaluations per episode.

Results. Out of 1,536 episodes, 99.7% end with a submitted structure, 98.6% with one that contains no lead, and 87.8% with one that is also ABX_3. About 30% are also stable and novel, but only 24 (1.6%) also have the right band gap, so the band gap is by far the hardest condition. The success rate stays about the same as we run more episodes, but the agent keeps finding new materials, with 2 distinct ones after 128 episodes per reference, 7 after 256, and 8 after 512. Table 8 lists them, Figure 10 shows how the band gaps of all lead-free submissions spread around the targets, and Figure 9 compares one of them with its lead reference. For the tolerance and octahedral factors in Figure 3 we use Shannon radii, with 12-fold coordination for the A site and 6-fold for the B site and the halide. Sn^{2+} has no Shannon radius, so we use 1.10 Å. With 1.18 Å instead, $RbSnI_3$ stays in the perovskite region and $KSnBr_3$ moves to its edge ($t = 0.81$). All of them are cubic perovskites ($Pm\bar{3}m$). Only $KGeI_3$ and $RbMgI_3$ are new compositions, and a self-consistent DFT calculation places both within 0.1 eV/atom of the hull (0.080 and 0.075 eV/atom). The other five are already in Materials Project with lower-energy structures that are not cubic, so for these the agent found new cubic polymorphs rather than new compounds. $TlCdCl_3$ contains no lead, but it contains thallium, which is also toxic. Figure 11 shows one of the episodes that found $KSnBr_3$. The agent proposes several Sn and Ge variants at once, drops the three that the novelty check reports as known, checks the band gap of the one left, and tries a few backups before submitting it.

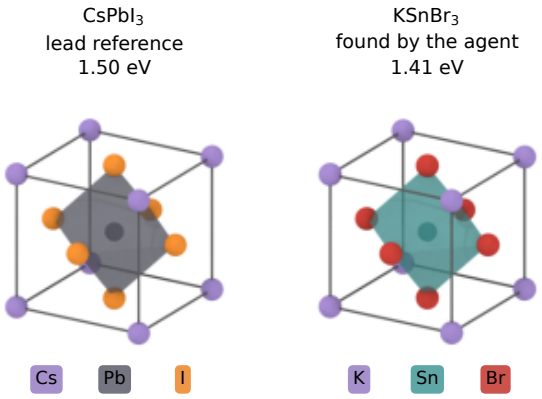


Figure 9: $CsPbI_3$ and $KSnBr_3$, one of the materials found by the agent, have the same cubic perovskite structure, with the metal (Pb or Sn) at the center of an octahedron of halide atoms. Band gaps are predicted by the band-gap model.

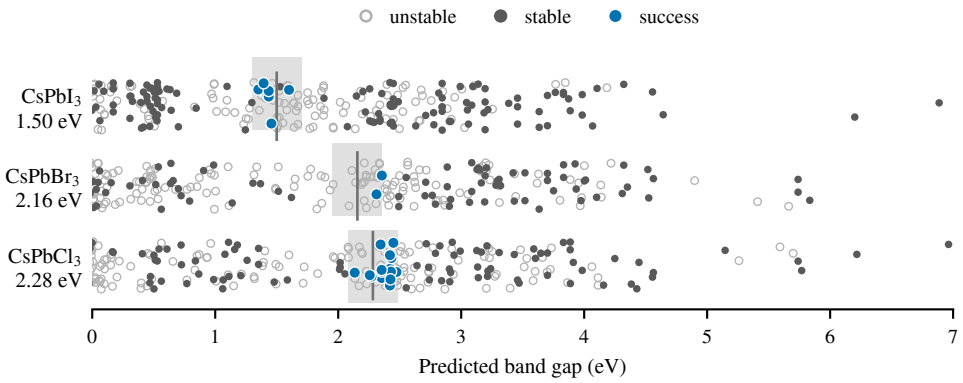


Figure 10: Predicted band gaps of the lead-free ABX_3 submissions, one row per lead compound, with its band gap (line) and the ± 0.2 eV target (shaded). Each dot is one episode, colored as unstable ($E_{\text{hull}} \geq 0.1$ eV/atom), stable, or a full success (stable, novel, and in the target). Only the 798 episodes where the agent predicted the band gap of its submission are shown.

Table 8: Lead-free perovskites found by the agent. Band gap and E_{hull} are medians over the successful episodes, and all structures are cubic ($Pm\bar{3}m$). The last column gives the lowest-energy Materials Project phase of the same composition, if any.

Reference	Material	Episodes	Band gap (eV)	E_{hull} (eV/atom)	MP phase
$CsPbI_3$ (1.50 eV)	$KSnBr_3$	4	1.41	0.071	$Pnma$
	$KGeI_3$	2	1.44	0.092	–
	$RbSnI_3$	1	1.60	0.077	$Pnma$
$CsPbBr_3$ (2.16 eV)	$CsMgI_3$	2	2.33	0.076	$P6_3/mmc$
$CsPbCl_3$ (2.28 eV)	$CsMgI_3$	8	2.39	0.066	$P6_3/mmc$
	$RbMgI_3$	3	2.26	0.073	–
	$CsBaI_3$	3	2.43	0.074	$Pnma$
	$TlCdCl_3$	1	2.47	0.053	$Pnma$

Episode for $CsPbI_3$ (target 1.50 eV)

```
1 retrieve_materials:  Rb-Sn-I, Rb-Sn-Br, Cs-Ge-I, Cs-Sn-Br
2 propose_structure (Pm-3m):  CsSnCl3, RbSnCl3, RbGeI3, KSnBr3
3 check_novelty:  only KSnBr3 is novel
4 evaluate_structure(KSnBr3):  E_hull = 0.067 eV/atom
  predict_band_gap(KSnBr3):  1.42 eV
  propose_structure:  CsCaI3, KCaI3, RbCaI3, CsSrI3 (backups)
5--6 screen backups:  known, E_hull 0.12--0.20 eV/atom, or gap 3.06 eV
7 submit(KSnBr3)
```

Figure 11: One successful episode for the $CsPbI_3$ target, with one line per turn and shortened tool outputs.