

CrystalSTAR: Structured Action Orchestration with Trio-Reflection for Constrained Novel Crystal Discovery

Chenyi Tong¹ Jingru Gan² Yanqiao Zhu² Kaiqiao Han² Haixin Wang²
Yizhou Sun² Wei Wang² Xiao Luo¹

¹UW-Madison ²UCLA

✉ Contact: ctong29@wisc.edu

Abstract

This paper studies the problem of constrained novel crystal discovery, which aims to identify a set of physical feasible and property-aligned material candidates within a complicated search space. Existing approaches typically integrate large language models (LLMs) with domain knowledge of material properties and constraints for material discovery. However, these approaches usually adopt shallow memory mechanisms or unstructured toolsets, leading to suboptimal performance. Towards this end, we propose a novel approach named Structured Action Orchestration with Trio-Reflection (**CrystalSTAR**) for constrained novel crystal discovery. The core of our **CrystalSTAR** is to incorporate hierarchical memories into a structured action orchestration from modification scopes and dominant components, facilitating a smooth and efficient crystal discovery procedure. In particular, our **CrystalSTAR** first formulates the toolset as a structured action space along two orthogonal dimensions, resulting in four complementary tools for material exploration. More importantly, we adopt a trio-reflection mechanism to guide the trial process in a hierarchical manner. Here, our bottom-up memory records local anomalies, stuck trajectories, and cross-run knowledge, enabling effective strategy planning with grounding evidence. Extensive experiments across multiple benchmark datasets validate the effectiveness of our proposed **CrystalSTAR** in comparison to extensive state-of-the-art approaches. Our codebase is open-sourced at <https://github.com/oldwizard7/CrystalSTAR>.

1 Introduction

Materials discovery underlies advances in electronics, energy storage, dielectrics, and structural materials (Jain et al., 2013; Butler et al., 2018; Merchant et al., 2023; Takahara et al., 2025; Glass et al., 2006; Zhang et al., 2024). Yet crystal design remains fundamentally difficult because the search space is vast, mixed discrete-continuous, and strongly coupled: changing composition can alter the stable lattice, while structural edits often require compatible stoichiometric changes. In realistic discovery settings, candidate materials must satisfy not only target property requirements but also physical constraints such as structural validity and thermodynamic metastability. Crystal discovery is therefore better viewed as a constrained, multi-objective sequential search problem than as unconstrained free-form generation (Jia et al., 2024; Abhyankar et al., 2025).

Recent deep generative models and LLM-based materials systems have enabled new approaches to crystal generation and materials design (Xie et al., 2021; Jiao et al., 2023; Miller et al., 2024; Antunes et al., 2024; Ding et al., 2024; Wang et al., 2026; Jia et al., 2024). Deep generative models, however, generally require large structural datasets. They also struggle when several property requirements must be met at once, and performance may even decline when the target differs from the training data. (Xie et al., 2021; Jiao et al., 2023; Miller et al., 2024). Although LLM-based methods are more flexible, their prompt-only or weakly grounded generation often produces candidates that are invalid, unstable, or repetitive (Wang et al., 2026; Jia et al., 2024; Abhyankar et al., 2025; Gan et al., 2025). More

broadly, many existing methods still treat materials discovery as one-shot generation or shallow refinement (Antunes et al., 2024; Jia et al., 2024; Wang et al., 2026), rather than as long-horizon search over physically grounded decisions.

In this work, we present a novel approach named **CrystalSTAR** for constrained crystal discovery built on *Structured Action* and *Trio-Reflection*. Instead of directly generating atomic structures, **CrystalSTAR** uses the language model as a high-level controller over a compact library of executable crystal tools. At each step, the agent selects a structured action, evaluates the resulting candidate with surrogate oracles, and feeds the outcome back into subsequent decisions. To support long-horizon exploration, we further introduce an M^3 memory mechanism, consisting of Micro, Meso, and Macro reflection, that converts trajectory-level successes and failures into reusable search guidance. We evaluate **CrystalSTAR** on 14 constrained materials-design tasks spanning electronic, dielectric, energy, and mechanical domains. Under a fixed search budget, **CrystalSTAR** consistently improves hit rate and metastability over strong deep generative and LLM-based baselines, while also achieving stronger novelty and more persistent discovery throughout the search process. Further analyses show that structured tool grounding is critical for avoiding invalid or thermodynamically collapsed proposals, and that Trio-Reflection improves discovery not by storing more raw context, but by organizing search experience into reusable guidance. These results support physically grounded action spaces and hierarchical reflection as an effective abstraction for agentic materials discovery.

Our contributions are summarized as follows: ① We formulate constrained crystal discovery as a closed-loop agentic search problem and propose **CrystalSTAR**, a framework that uses language models to guide search through executable crystal tools rather than unconstrained structure generation. ② We introduce Trio-Reflection through the M^3 hierarchy, which organizes search experience across Micro, Meso, and Macro levels to support long-horizon exploration. ③ Across 14 challenging materials-design tasks, we show that **CrystalSTAR** improves constraint satisfaction, thermodynamic plausibility, and novelty over prior deep generative and LLM-based approaches, while exhibiting more sustained discovery behavior in qualitative and ablation analyses.

2 Related Work

Deep Generative Models for Crystal Structures. Computational materials discovery has progressed from high-throughput simulation toward generative design and inverse search (Noh et al., 2020; Xie & Grossman, 2018; Chen et al., 2019; Choudhary et al., 2020). Deep generative approaches based on variational autoencoders, diffusion, and flow matching can model periodic structures and propose candidate materials (Xie et al., 2021; Jiao et al., 2023; Miller et al., 2024; Zeni et al., 2023), but typically depend on large training corpora and become less reliable under strict multi-property constraints.

LLM Agents and Hierarchical Reasoning. LLM-based agents use external tools and memory to support sequential decision-making (Yao et al., 2022; Shinn et al., 2023; Wang et al., 2023; Pan et al., 2026). Later work improved agent planning through tree-based search (Zhou et al., 2023) and self-refinement (Madaan et al., 2023). Other studies focused on tool use (Schick et al., 2023; Qin et al., 2023), or coordination among multiple agents (Wu et al., 2024; Hong et al., 2023). A persistent challenge is long-horizon memory: flat context pools quickly overflow, degrading planning quality. Recent coarse-to-fine grounded memory (Yang et al., 2025) addresses this via hierarchical abstraction. Our M^3 mechanism extends this direction to scientific discovery.

LLM-driven Materials Discovery. Recent work applies language models to crystal design through autoregressive generation and closed-loop agentic search. Fine-tuned approaches (Gruver et al., 2024; Antunes et al., 2024; Flam-Shepherd & Aspuru-Guzik, 2023; Cao et al., 2025) can generate structures but depend on domain-specific training data. CrystalAgent (Wang et al., 2026) decomposes generation into staged subtasks, LLMatDesign (Jia et al., 2024) adds self-reflection, MatLLMSearch (Gan et al., 2025) integrates LLMs with evolutionary algorithms, and LLEMA (Abhyankar et al., 2025) combines evolutionary refine-

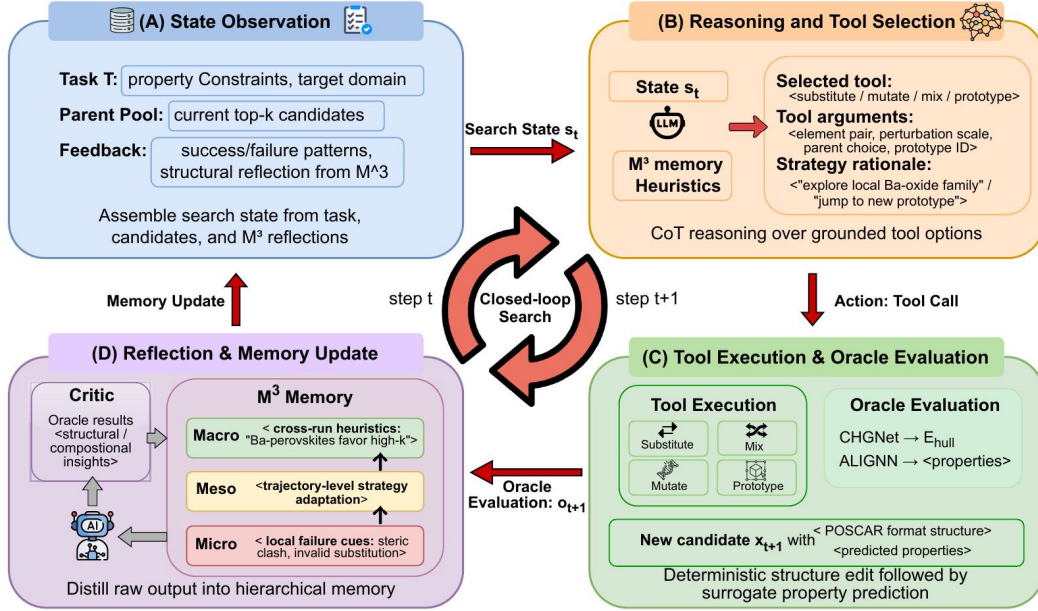


Figure 1: Overview of the **CrystalSTAR** framework: (A) *State Observation*, which assembles a grounded search state from the task, parent pool, and Trio-Reflection memory; (B) *Reasoning and Tool Selection*, where the agent selects a structured crystal operator and its parameters; (C) *Tool Execution and Oracle Evaluation*, which applies the selected operator and predicts task-relevant properties and thermodynamic stability; (D) *Reflection and Memory Update*, which distills oracle feedback into hierarchical search priors for subsequent decisions.

ment with memory-based feedback. However, most of these methods rely on flat memory and do not explicitly organize actions around compositional and structural changes. This can make long-horizon search difficult under tight constraints. **CrystalSTAR** addresses this with structured action orchestration and Trio-Reflection for long-horizon search.

3 The Proposed CrystalSTAR

Problem Definition. We formulate crystal discovery for a target task $\mathcal{T}$ as a constrained optimization problem over the candidate space $\mathcal{X}$. Each candidate crystal $x \in \mathcal{X}$ is specified by its composition and structure, including atomic species, stoichiometric ratios, lattice parameters, and fractional coordinates. For task $\mathcal{T}$, let $\rho_1, \dots, \rho_K : \mathcal{X} \rightarrow \mathbb{R}$ denote the objective-specific scalar scores, where larger values indicate better task performance, and let $\mathcal{C}_{\text{task}} \subseteq \mathcal{X}$ denote the subset of candidates satisfying the task-specific hard constraints. We further impose thermodynamic plausibility through the energy above hull and define the feasible region as:

$$\mathcal{X}_{\text{feas}}(\mathcal{T}) = \{x \in \mathcal{X} \mid x \in \mathcal{C}_{\text{task}}, E_{\text{hull}}(x) \leq \epsilon\}, \quad (1)$$

where $E_{\text{hull}}(x)$ denotes the energy above the thermodynamic convex hull of candidate x , and ϵ is a metastability threshold, set to 0.1 eV/atom in our experiments. To aggregate multiple objectives, we use a linear scalarization $S_{\mathcal{T}}(x) = \sum_{k=1}^K \beta_k \rho_k(x)$ where β_k denotes the relative importance of the k -th objective. In our benchmark, the coefficients β_k are fixed a priori for each task. Under these fixed task preferences, **CrystalSTAR** instantiates constrained multi-objective discovery through a task-specific linear scalarization, which provides a concrete objective for sequential search. The corresponding ideal optimization problem is $x_{\mathcal{T}}^* \in \arg \max_{x \in \mathcal{X}_{\text{feas}}(\mathcal{T})} S_{\mathcal{T}}(x)$. In practice, exhaustive optimization over $\mathcal{X}$ is intractable. Under a finite oracle evaluation budget B , the goal is therefore to sequentially discover feasible crystals with high task score $S_{\mathcal{T}}(x)$.

3.1 Framework Overview

CrystalSTAR is a closed-loop agentic framework for constrained crystal discovery. It does not ask the language model to generate atomic coordinates directly. Instead, the model selects from a set of executable crystal tools. The selected tool modifies some parent crystals, and the oracles evaluate the resulting candidate. This separation makes each search step easy to track and control. As shown in Figure 1, each iteration consists of four stages: state observation, reasoning and tool selection, tool execution and oracle evaluation, and reflection with memory update. At iteration t , the agent first assembles a compact search state from the task specification $\mathcal{T}$, a parent pool of currently promising candidates, and summaries retrieved from the Trio-Reflection memory. Using this search state, the agent selects a structured tool call, including both a tool τ from the structured action space $\mathcal{A}$ and its tool-specific arguments Θ (full interface definitions are provided in Appendix A.1). The selected tool then edits or recombines parent crystals to produce a new candidate, which is subsequently evaluated by surrogate oracles for the task properties and thermodynamic stability. Finally, the resulting feedback is distilled into hierarchical memory by the Trio-Reflection mechanism, and injected into the next iteration, allowing future decisions to benefit from both recent outcomes and longer-horizon search patterns. Section 3.2 details the structured action space, and Section 3.3 describes the Trio-Reflection mechanism.

3.2 Structured Action Orchestration Along Orthogonal Dimensions

Direct crystal generation provides limited control over how a candidate is modified and makes it difficult to relate search outcomes to specific design choices (Abhyankar et al., 2025). **CrystalSTAR** instead operates over a compact library of executable crystal tools. Formally, each tool $\tau \in \mathcal{A}$ defines a structured interface $\tau : \mathcal{X} \times \Theta \rightarrow \mathcal{X}$, where $\mathcal{X}$ denotes the space of candidate crystals and Θ denotes tool-specific parameters. Each tool takes one or more parent crystals and tool-specific arguments as input. It then returns a new candidate crystal. To give action selection an explicit structure, we organize the tool space along two orthogonal dimensions: *scope of modification* (local refinement versus global transformation) and *dominant control variable* (composition versus structure), yielding four canonical tool types (Figure 2).

On the composition side, **SUBSTITUTE** performs local compositional refinement by replacing selected elements while preserving the parent site framework, enabling chemically meaningful edits such as isovalent or same-group substitutions. **MIX** performs a global compositional transformation by recombining two parent crystals, allowing the search to move more aggressively across chemical systems. On the structure side, **MUTATE** applies local perturbations to lattice parameters and fractional coordinates while keeping composition fixed, supporting nearby structural exploration. **PROTOTYPE** performs a global structural transformation by instantiating a target composition within a different crystal template. Together, these tools define a structured action space over both what to modify and how far to move in the search space. To maintain executability, tool outputs are lightly sanitized before downstream evaluation; in particular, structurally aggressive transformations such as **PROTOTYPE** are followed by heuristic lattice rescaling based on atomic radii to reduce obviously invalid structures.

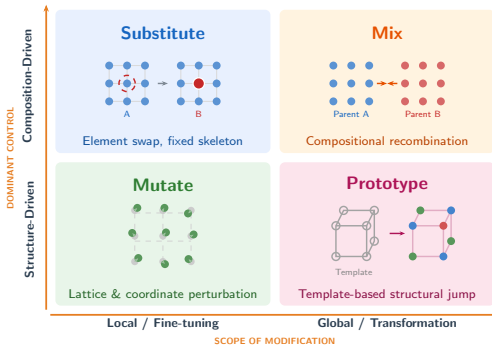


Figure 2: Tool taxonomy by scope and dominant control variable.

3.3 Trio-Reflection Mechanism with Bottom-up Memory

For crystal discovery, a single reflection buffer is often insufficient since the signals that guide planning arise at different scales. Some are immediate and candidate-specific, such

as an invalid structure, a failed constraint check, or a high energy above hull produced by a particular tool action. Others emerge only across multiple iterations, for example when a family of local edits repeatedly stalls. **CrystalSTAR** addresses this mismatch through *Trio-Reflection*, instantiated as the M^3 (*Micro, Meso, Macro*) framework, which organizes search experience into an external hierarchical memory state at iteration t as:

$$\mathcal{M}_t = (\mathcal{M}_{t,\mu}, \mathcal{M}_{t,m}, \mathcal{M}_{t,M}). \quad (2)$$

Inspired by coarse-to-fine grounded memory for LLM planning (Yang et al., 2025; Ge et al., 2025), but adapted to oracle-guided materials search, M^3 is designed to build reusable search guidance without overloading the planner with raw execution traces. Beyond avoiding repeated failures, it also encourages the agent to build reusable scientific heuristics from search experience, much as how human researchers refine intuition from experimental feedback.

The three levels differ in how they are written and what they retain. *Micro* memory $\mathcal{M}_\mu$ logs the oracle output for every evaluated child: evaluation status $\sigma(x) \in \{\text{hit}, \text{no_hit}, \text{invalid}\}$, decomposition energy, source tool, and per-constraint outcomes, serving as the raw evidential ground from which higher levels are derived. The two higher levels progressively summarize this evidence into increasingly abstract textual guidance: *Meso* stores within-run tactical tips, while *Macro* stores cross-run reusable rules. Upon the conclusion of iteration t , the agent consolidates a bounded recent window of *Micro* records together with tool statistics s_t into a *Meso* tip z_t , which is a within-run tactical recommendation about which tools, parent families, or search directions should be promoted or avoided next:

$$z_t = \text{LLM}_{\text{meso}}(\mathcal{M}_{t,\mu}, s_t), \quad \mathcal{M}_{t,m} \leftarrow \mathcal{M}_{t-1,m} \cup \{z_t\}. \quad (3)$$

Every T_{Macro} iterations, accumulated *Meso* tips are further distilled into a *Macro* rule r_t that encodes longer-horizon regularities transferable across runs:

$$r_t = \text{LLM}_{\text{macro}}(\mathcal{M}_{t,m}, \mathcal{M}_{t-1,M}), \quad \mathcal{M}_{t,M} \leftarrow \mathcal{M}_{t-1,M} \cup \{r_t\} \text{ iff } c(r_t) \geq c_{\min}. \quad (4)$$

$c(x)$ is the confidence score, and the confidence gate $c_{\min}$ ensures that only rules with sufficient empirical backing are promoted; weakly supported patterns are discarded rather than allowed to bias subsequent search. At planning time, the planner reads the three banks in reverse, *Macro* first, then *Meso*, then *Micro*, and concatenates them into the tool selector:

$$\text{prompt}_t = \text{TopK}_M(\mathcal{M}_{t,M}) \oplus \text{TopK}_m(\mathcal{M}_{t,m}) \oplus \text{TopK}_\mu(\mathcal{M}_{t,\mu}), \quad (5)$$

so that strategic priors frame the context before tactical and local signals refine it. Values of T_M , $c_{\min}$, and per-level TopK budgets are given in Appendix A.

At the next iteration, the planner reuses this memory in a top-down manner, drawing on *Macro* guidance, *Meso* strategy cues, and *Micro* guardrails when constructing the next search state. To keep high-level heuristics grounded in current evidence, the planner also retains a bounded evidence zone containing recent outcomes, tool statistics, and compact structural feedback. This allows **CrystalSTAR** to benefit from hierarchical memory without losing contact with the empirical signals generated during the ongoing search.

3.4 Surrogate Oracle and Feedback

After tool execution, each candidate must be evaluated within the iterative search loop. Since direct DFT evaluation is not practical for in-loop screening, **CrystalSTAR** employs ML-based surrogate models that provide accurate and efficient estimates of both task-relevant properties and thermodynamic stability. This enables reliable candidate scoring, filtering, and comparison while preserving the throughput needed for closed-loop discovery.

In our implementation, ALIGNN (Choudhary & DeCost, 2021) is used for task-specific property prediction, while CHGNet (Deng et al., 2023) is used for stability-oriented assessment. These predictions are checked against the task-specific feasibility conditions and the metastability threshold defined in Section 3; feasible candidates are then ranked according to the scalarized task score $S_{\mathcal{T}}(x)$, while infeasible ones still provide informative signals for downstream reflection.

4 Experiment

4.1 Experimental Setup

Benchmark and Metrics. We evaluate `CrystalSTAR` on the 14-task materials discovery benchmark introduced by (Abhyankar et al., 2025), which spans electronic and optical materials (e.g., wide-bandgap semiconductors, high- κ dielectrics, and transparent conductors), energy materials (e.g., solid-state electrolytes, photovoltaic absorbers, and perovskite oxides), and structural or mechanical materials (e.g., SAW/BAW acoustic substrates, hard ceramics, and low-density structures). Each task defines a constrained multi-objective discovery problem with strict property requirements; full task specifications are provided in Appendix B. We report three primary metrics. *Hit Rate* (*H.R.*, %) measures the fraction of evaluated candidates that satisfy all task-specific property constraints. *Metastability Rate* (*Stab.*, %) measures the fraction of evaluated candidates with surrogate-predicted $E_{\text{hull}} \leq 0.1$ eV/atom. *Novel Hit Rate* (*N.H.R.*, %) measures the fraction of evaluated candidates that are both task-valid and absent from existing materials databases; details of the novelty check are provided in Appendix B.3.

Baselines. We compare against seven baselines spanning two paradigms. The deep generative baselines, CDVAE (Xie et al., 2021), G-SchNet (Gebauer et al., 2019), and DiffCSP (Jiao et al., 2023), model crystal structure distributions directly. The LLM-guided baselines include End2end, which directly prompts an LLM to generate candidate materials, LL-MatDesign (Jia et al., 2024), and LLEMA (Abhyankar et al., 2025) with Mistral and GPT backends. Since our benchmark matches that of Abhyankar et al. (2025), Table 1 reports their published *Hit Rate* and *Metastability Rate* results under the original evaluation protocol. Under the protocol-aligned evaluation, `CrystalSTAR` still achieves a higher *Metastability Rate* than LLEMA (GPT) on all 14 tasks (Appendix C.2). For *Novel Hit Rate* in Table 2, which is not reported in their work, we rerun the strongest baseline LLEMA (GPT) under the same search budget as `CrystalSTAR`: 50 iterations with five evaluated candidates per iteration.

Implementation. `CrystalSTAR` uses gpt-4o-mini (Achiam et al., 2023) as the reasoning engine, together with ALIGNN (Choudhary & DeCost, 2021) for task-specific property prediction and CHGNet (Deng et al., 2023) for stability-oriented assessment. Detailed implementation settings are provided in Appendix C.

4.2 Performance Comparison

Tables 1 and 2 summarize `CrystalSTAR`’s quantitative performance across the 14 benchmark tasks. Overall, `CrystalSTAR` achieves the highest *Hit Rate* on 13 of 14 tasks and the highest *Metastability Rate* on all 14, indicating that its gains are broad rather than task-specific. These improvements are especially visible on tightly constrained problems such as Wide-Bandgap Semiconductors and High- κ Dielectrics, where satisfying multiple target requirements must be balanced against thermodynamic plausibility. More importantly, `CrystalSTAR` achieves a higher *Novel Hit Rate* than LLEMA (GPT) on 12 of the 14 tasks. The largest gains occur on Aerospace Materials (+46.8%), Low-Density Structural Materials (+35.0%), and Hard Coating Materials (+24.6%). This pattern suggests that the method is not simply producing more feasible candidates, but is exploring broader and less charted regions of chemical space. A detailed breakdown of these novel hits is provided in Appendix B.4. On Acousto-Optic Hybrids, `CrystalSTAR` achieves a lower *Hit Rate* than LLEMA (GPT): 14.4% versus 26.3%. However, its *Metastability Rate* reaches 75.2%, compared with 0.8% for LLEMA (GPT). A discovery efficiency analysis is provided in Appendix D.4.

4.3 Search Dynamics and Discovery Trajectories

Beyond aggregate metrics, we further examine whether `CrystalSTAR`’s discoveries remain chemically grounded. On the Wide-Bandgap Semiconductors task, the left panel of Figure 3 shows that `CrystalSTAR` concentrates candidates more effectively inside the feasible region than LLEMA, yielding a denser set of high quality solutions rather than broader coverage of lower quality regions. This pattern is reflected not only in the feasible region occupancy,

Table 1: Hit Rate (H.R., %) and Metastability Rate (St., %) across 14 materials-design tasks. HR = Hit Rate, St. = Metastability Rate. Best results in each column are **bold**.

Method	Wide-Bandgap		SAW/BAW		High-k Dielect.		Solid-State Elect.		Piezo Energy		Transp. Cond.		Insul. Dielect.	
	HR	St.	HR	St.	HR	St.	HR	St.	HR	St.	HR	St.	HR	St.
CDVAE	0.0	0.0	0.3	0.0	0.8	0.0	0.0	0.0	42.2	0.0	0.0	0.0	1.1	0.1
G-SchNet	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0
DiffCSP	0.0	0.0	0.4	0.0	0.8	0.0	0.0	0.0	41.2	0.0	0.0	0.0	1.1	0.0
End2end	1.0	0.8	10.3	0.7	0.0	0.0	0.5	0.3	10.3	0.3	0.0	0.0	0.0	0.0
LLMatDesign	4.2	1.1	47.6	0.1	1.4	0.3	2.5	2.4	32.2	1.4	0.0	0.0	0.2	0.1
LLEMA (Mistral)	17.1	10.7	31.6	6.8	7.5	3.6	31.8	20.8	67.1	4.8	43.9	18.5	21.5	9.4
LLEMA (GPT)	33.6	22.4	59.9	10.7	20.0	12.7	46.2	25.4	63.5	3.2	39.1	14.9	17.6	4.6
CrystalSTAR	71.2	81.2	69.2	68.4	65.9	75.9	65.6	62.4	68.4	76.0	57.2	62.0	55.0	71.5

Method	Photo-voltaics		Hard Coatings		Hard Ceramics		Aerospace Mat.		Acousto-optic		Low Density		Perovskites	
	HR	St.	HR	St.	HR	St.	HR	St.	HR	St.	HR	St.	HR	St.
CDVAE	1.1	0.0	0.0	0.0	15.3	0.1	1.2	0.0	21.9	0.0	0.0	0.0	0.0	0.0
G-SchNet	0.0	0.0	0.0	0.0	0.2	0.2	0.1	0.0	0.0	0.0	0.2	0.0	0.0	0.0
DiffCSP	1.1	0.0	0.0	0.0	14.8	0.0	0.1	0.0	21.5	0.0	0.0	0.0	0.0	0.0
End2end	24.6	10.7	0.0	0.0	14.3	5.1	0.0	0.0	8.6	0.6	2.0	0.4	0.0	0.0
LLMatDesign	3.9	0.0	0.0	0.0	19.0	0.4	0.0	0.0	15.5	0.6	0.1	0.0	1.1	0.8
LLEMA (Mistral)	20.5	3.7	10.8	1.4	27.9	2.7	1.5	0.5	14.0	0.5	1.5	0.1	22.9	2.8
LLEMA (GPT)	22.9	4.8	17.8	4.6	61.0	5.7	1.0	0.3	26.3	0.8	0.5	0.1	19.4	2.8
CrystalSTAR	53.6	71.4	74.8	67.2	66.7	57.8	91.6	70.7	14.4	75.2	53.6	40.3	46.0	63.2

Table 2: Novel Hit Rate (N.H.R., %) across 14 tasks. N.H.R. measures the proportion of valid candidates absent from existing materials databases.

Method	Wide-Bandgap	SAW/BAW	High-k Dielect.	Solid-State Elect.	Piezo Energy	Transp. Cond.	Insul. Dielect.
LLEMA (GPT)	31.0	32.3	18.6	28.4	59.7	26.2	16.7
CrystalSTAR	44.4	32.4	32.1	48.8	56.0	28.8	36.6

Method	Photo-voltaics	Hard Coatings	Hard Ceramics	Aerospace Mat.	Acousto-optic	Low Density	Perovskites
LLEMA (GPT)	15.2	17.4	39.1	8.6	24.9	0.1	1.7
CrystalSTAR	34.7	42.0	39.4	55.4	11.2	35.1	16.8

but also in how candidate density shifts toward favorable trade-off regions in the left panel of Figure 3. Many of these hits also remain connected to experimentally established alkali-fluoride families. For example, K_2NaGaF_6 is a known elpasolite phosphor host (Qiu et al., 2017), and **CrystalSTAR** proposes nearby family consistent variants rather than chemically isolated outliers. This suggests that tool-grounded search preserves meaningful structural motifs while still enabling novelty.

The right panel of Figure 3 provides a more explicit example on the High- κ Dielectrics task. **CrystalSTAR** revisits Ba-based oxide and double-perovskite candidates before arriving at the novel candidate $\text{Ba}_2\text{HfScO}_6$. Figure 3 shows the related candidates: Ba_2HfO_4 , Ba_2ZrO_4 and $\text{Ba}_2\text{ScTaO}_6$, with the corresponding substitution steps observed during the search. Related compounds in this family have been experimentally studied. $\text{Ba}_2\text{ScNbO}_6$ single crystals have been grown for use as lattice matched substrates for BaSnO_3 films (Gugushev et al., 2019), and $\text{Ba}_2\text{ScTaO}_6$ has been synthesized and characterized as a large-band-gap double perovskite (Himanshu et al., 2022). **CrystalSTAR** records these local hit patterns at the Micro level and consolidates them into tool selection guidance at the Meso level. At the

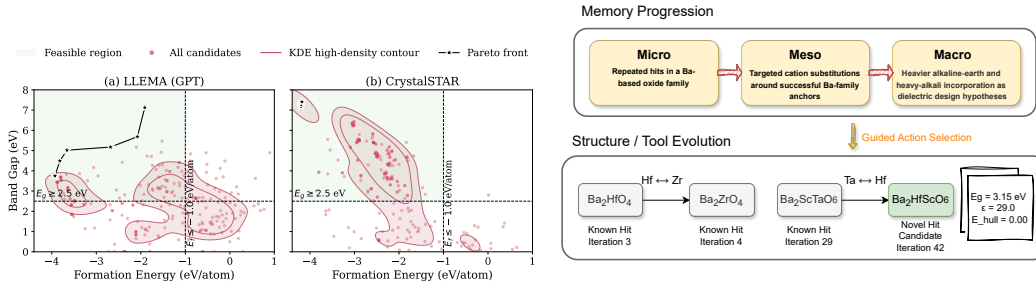


Figure 3: **Left:** Candidate distribution on the Wide-Bandgap Semiconductors task. Red contours denote high-density regions estimated by Gaussian kernel density estimation. Pareto frontiers indicate the best observed trade-offs between formation energy and band gap. **Right:** Discovery lineage on the High- κ Dielectrics task. The lineage connects known hits to the novel candidate Ba₂HfScO₆ and shows how the M^3 memory derives higher-level design guidance from local substitution patterns.

Macro level, it abstracts them into broader family level and scientific heuristics that guide subsequent tool use and candidate generation. **CrystalSTAR** follows the ionic-substitution principle used in inorganic materials discovery (Hautier et al., 2011), but learns the relevant substitution patterns online from search feedback.

4.4 Ablation on Structured Actions

To test the effect of the structured action space, we compare **CrystalSTAR** against a *Direct Generation* baseline on two representative tasks spanning electronic and mechanical materials design. This baseline keeps the same task input, search loop, surrogate evaluation, and feedback pipeline as **CrystalSTAR**, but replaces the tool-execution step with a direct LLM proposal of the next crystal candidate. We also report *Syntactic Validity*, the percentage of generated candidates that can be successfully parsed. As shown in Figure 4, this direct-generation variant can often produce parsable crystal structures, but many of these candidates fail to remain physically plausible or satisfy the target property constraints. This gap is especially pronounced on the electronic task, where the constraints are more tightly coupled, whereas **CrystalSTAR** remains robust across both settings. These results suggest that **CrystalSTAR**’s advantage comes not from direct LLM-based generation, but from using the LLM to guide search through structured, executable crystal actions.

We next ask whether **CrystalSTAR**’s gains come from any single class of actions, or from the interplay between global exploration and local refinement. To answer this, we compare the full method against restricted variants that retain only global-transformation tools (MIX and PROTOTYPE) or only local-refinement tools (SUBSTITUTE and MUTATE). As shown in Table 3, neither restricted variant recovers the performance of the full system. The global-only variant attains moderate metastability but very low hit rate, suggesting large jumps are useful for reaching physically plausible regions of the search space, but are too coarse to reliably satisfy task constraints. The local-only variant achieves a substantially higher hit rate, showing that local edits are effective for refinement once the search is near a promising region. However, it still falls short of the full method, indicating local refinement alone cannot explore the search space broadly enough. Collectively, these results show the structured action space serves as a coarse-to-fine search mechanism: global moves identify promising basins, while local moves exploit them to satisfy tightly coupled design objectives.

4.5 Ablation on Trio-Reflection (M^3)

Effective memory in materials discovery should do more than expose the LLM to longer search histories; it should distill past search outcomes into guidance that improves subsequent exploration. We study this question on the High- κ Dielectrics task, where recurring family-level search patterns make the effect of memory particularly visible. In this setting, Hit Rate alone can be misleading, since an agent may repeatedly recover known structures

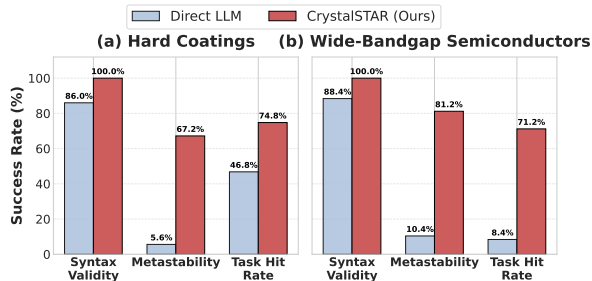


Figure 4: Direct generation vs. tool-grounded search. Direct LLM proposal preserves less syntactic validity and leads to substantially worse thermodynamic stability and task hit rate than CrystalSTAR.

Variant	H.R.	St.
Global Transformation	9.2	50.2
Local Fine-tuning	39.2	42.8
CrystalSTAR Toolset	65.9	75.9

Table 3: Full method vs. restricted action variants on the High- κ Dielectrics task. Global only uses MIX+PROTOTYPE; local only uses SUBSTITUTE+MUTATE.

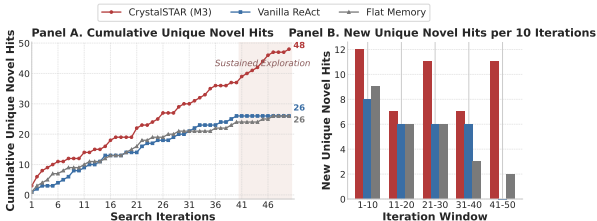


Figure 5: Sustained discovery on the High- κ Dielectrics task. (A) Cumulative novel valid discoveries. (B) Yield per 10-iteration window.

Config.	Div.	Nov.	H.R.	N.H.R.
ReAct	36.8	46.0	77.2	24.4
Raw Context	36.4	52.8	73.2	28.0
CrystalSTAR	53.8	60.3	65.9	32.1

Table 4: Ablation on the High- κ Dielectrics task. Compared with ReAct and Raw Context, M^3 yields the strongest Diversity Rate, Novelty Rate, and Novel Hit Rate.

without sustaining genuinely novel search. We therefore also report Diversity Rate (Div.), which measures the fraction of evaluated candidates that remain structurally unique under a StructureMatcher-based deduplication rule; Novelty Rate (Nov.), which measures the fraction of evaluated candidates identified as novel relative to the reference database under our novelty checker, regardless of task validity; and Novel Hit Rate (N.H.R.), which further restricts novelty to task-valid discoveries. Detailed definitions are provided in Appendix B. We compare three configurations: *ReAct* (Yao et al., 2022), which uses short-horizon reflection over a sliding window of recent steps; *Raw Context*, which prepends unsummarized execution histories from previous runs to the context window; and *CrystalSTAR* (M^3), which organizes experience into Micro, Meso, and Macro memory.

Table 4 and Figure 5 together show that the benefit of M^3 lies in sustaining novel discovery rather than maximizing short-term Hit Rate. ReAct attains the highest H.R., but its lower Diversity Rate, Novelty Rate, and N.H.R. suggest a search pattern biased toward repeatedly recovering familiar structures. Raw Context modestly improves Novelty Rate over ReAct, indicating that exposing the LLM to more historical traces is somewhat helpful, but its Diversity Rate remains nearly unchanged and its discovery yield still decays quickly, so additional raw context alone does not produce broader exploration. In contrast, *CrystalSTAR* with M^3 achieves the strongest Diversity Rate, Novelty Rate, and N.H.R., while Figure 5(A) shows more cumulative novel valid discoveries and Figure 5(B) shows that its per-window yield remains active after the baselines begin to flatten. Although this comes with a lower H.R., the overall pattern suggests that Trio-Reflection is not merely storing more search history, but transforming past outcomes into reusable guidance that supports more sustained and genuinely novel exploration. The same sustained-discovery pattern also holds on Hard Coatings and Wide-Bandgap Semiconductors, where *CrystalSTAR* achieves the highest Diversity Rate and cumulative number of unique novel hits. Component-level ablations further show that removing inherited Macro rules reduces N.H.R. from 32.1% to 24.4%, while removing Meso reduces cumulative unique novel hits from 48 to 35 (Appendix D.2).

5 Conclusion

We presented a closed-loop agentic framework named **CrystalSTAR** for constrained crystal discovery that couples *Structured Action* with *Trio-Reflection*. Rather than using the language model as a free-form generator of crystal structures, **CrystalSTAR** uses it to guide search over executable crystal tools and to update future decisions through hierarchical memory distilled from oracle feedback. Across 14 benchmark tasks spanning electronic, dielectric, energy, and mechanical materials design, **CrystalSTAR** consistently improves discovery quality, achieving broad gains in constraint satisfaction, thermodynamic plausibility, and especially genuine novelty. Qualitative analyses further show that **CrystalSTAR** steers search toward chemically grounded material families. Future work could extend the tool library to cover a broader range of materials operations.

Acknowledgment

This work was partially supported by NSF 2616633, NSF 2312501, NSF 2202693, NSF 2106859, NSF 2609582, NSF 2303037, NSF 2531008, DARPA HR001126CE054, Amazon, NEC, SRC JUMP 2.0 Center, UCLA CDSC Center, Amazon Research Awards, Snapchat, and Google Gifts. We also acknowledge the Center for High Throughput Computing (CHTC) at the University of Wisconsin–Madison for providing computational resources and support (Center for High Throughput Computing, 2006).

References

- Nikhil Abhyankar, Sanchit Kabra, Saaketh Desai, and Chandan K Reddy. Accelerating materials design via llm-guided evolutionary search. *arXiv preprint arXiv:2510.22503*, 2025.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):10570, 2024.
- Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.
- Zhendong Cao, Xiaoshan Luo, Jian Lv, and Lei Wang. Space group informed transformer for crystalline materials generation. *Science Bulletin*, 2025.
- Center for High Throughput Computing. Center for high throughput computing, 2006. URL <https://chtc.cs.wisc.edu/>.
- 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.
- Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021.
- Kamal Choudhary, Kevin F Garrity, Andrew CE Reid, Brian DeCost, Adam J Biacchi, Angela R Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A Gilad Kusne, Andrea Centrone, et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj computational materials*, 6(1):173, 2020.
- Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023.

- Qianggang Ding, Santiago Miret, and Bang Liu. Matexpert: Decomposing materials discovery by mimicking human experts. *arXiv preprint arXiv:2410.21317*, 2024.
- Daniel Flam-Shepherd and Alán Aspuru-Guzik. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as xyz, cif, and pdb files. *arXiv preprint arXiv:2305.05708*, 2023.
- Jingru Gan, Peichen Zhong, Yuanqi Du, Yanqiao Zhu, Chenru Duan, Haorui Wang, Daniel Schwalbe-Koda, Carla P Gomes, Kristin A Persson, and Wei Wang. Matllmsearch: Crystal structure discovery with evolution-guided large language models. *arXiv preprint arXiv:2502.20933*, 2025.
- Yubin Ge, Salvatore Romeo, Jason Cai, Monica Sunkara, and Yi Zhang. Samule: Self-learning agents enhanced by multi-level reflection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16602–16621, 2025.
- Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Advances in neural information processing systems*, 32, 2019.
- Colin W Glass, Artem R Oganov, and Nikolaus Hansen. Uspex—evolutionary crystal structure prediction. *Computer physics communications*, 175(11-12):713–720, 2006.
- Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379*, 2024.
- C Gugushev, D Klimm, M Brützm, TM Gesing, M Gogolin, H Paik, A Dittmar, VJ Fratello, and DG Schlom. Single crystal growth and characterization of ba2scnbo6—a novel substrate for basno3 films. *Journal of Crystal Growth*, 528:125263, 2019.
- Geoffroy Hautier, Chris Fischer, Virginie Ehrlicher, Anubhav Jain, and Gerbrand Ceder. Data mined ionic substitutions for the discovery of new compounds. *Inorganic chemistry*, 50(2):656–663, 2011.
- AK Himanshu, Sujit Kumar, Urmimala Dey, and Rajyavardhan Ray. Multiple low-energy excitons and optical response of d0 double perovskite ba2scta06. *Physica B: Condensed Matter*, 637:413856, 2022.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiwu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*, 2023.
- 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.
- Shuyi Jia, Chao Zhang, and Victor Fung. Llmatform: Autonomous materials discovery with large language models. *arXiv preprint arXiv:2406.13163*, 2024.
- 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.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.
- Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gwoon 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.
- Juhwan Noh, Geun Ho Gu, Sungwon Kim, and Yousung Jung. Uncertainty-quantified hybrid machine learning/density functional theory high throughput screening method for crystals. *Journal of Chemical Information and Modeling*, 60(4):1996–2003, 2020.
- Zhiming Pan, Junyu Luo, Zhiping Xiao, Kaize Ding, Xiao Luo, and Ming Zhang. Zipact: Zipping interaction history into a compact state for efficient llm agents. *Transactions on Machine Learning Research*, 2026.
- Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023.
- Shijie Qiu, Hengwei Wei, Mengmeng Wang, Shuai Zhang, Yang Zhou, Ling Xu, Xiaoming Wang, and Huan Jiao. Synthesis and photoluminescence of mn 4+ activated ternary-alkaline fluoride k₂ nagaf₆ red phosphor for warm-white led application. *RSC advances*, 7(79):50396–50402, 2017.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.
- Izumi Takahara, Teruyasu Mizoguchi, and Bang Liu. Accelerated inorganic materials design with generative ai agents. *Cell Reports Physical Science*, 6(12), 2025.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.
- Ruobing Wang, Qiaoyu Tan, Yili Wang, and Xin Wang. Crystalagent: Towards autonomous crystal generation via agentic reasoning, 2026. URL <https://openreview.net/forum?id=SOAfumn8bA>.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*, 2024.
- Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.
- Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.
- Wei Yang, Jinwei Xiao, Hongming Zhang, Qingyang Zhang, Yanna Wang, and Bo Xu. Coarse-to-fine grounded memory for llm agent planning. *arXiv preprint arXiv:2508.15305*, 2025.
- 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*, 2022.
- Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. Honeycomb: A flexible llm-based agent system for materials science. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3369–3382, 2024.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models. *arXiv preprint arXiv:2310.04406*, 2023.

A Additional Method Details

A.1 Structured Action Space: Tool Specifications

Substitute. Replaces one or more elements while preserving the parent’s lattice and site framework. Supports multi-element substitution (e.g., Ba, Ti $\rightarrow$ Sr, Zr). Parameters: `old_element` (comma-separated), `new_element` (comma-separated).

Mutate. Applies random perturbations to fractional coordinates and optionally to the lattice. Let $\mathbf{r}_i \in [0, 1)^3$ denote the fractional coordinate of the i -th atom and $\mathbf{L} \in \mathbb{R}^{3 \times 3}$ the parent lattice matrix:

$$\mathbf{r}'_i = (\mathbf{r}_i + \delta_i) \bmod 1, \quad \delta_i \sim \text{Uniform}(-s, s)^3, \quad (6)$$

$$\mathbf{L}' = \begin{cases} \alpha \cdot \mathbf{L}, & \text{with probability } 0.5, \\ \mathbf{L}, & \text{otherwise,} \end{cases} \quad \alpha \sim \text{Uniform}(1 - s/2, 1 + s/2), \quad (7)$$

where $s \in [0.01, 0.1]$ is the mutation strength. The modulo operation wraps perturbed fractional coordinates back into the unit cell, ensuring $\mathbf{r}'_i \in [0, 1)^3$ under periodic boundary conditions. The composition is preserved.

Mix. Performs a global compositional jump by recombining two parent structures. Let $\mathbf{L}_1, \mathbf{L}_2 \in \mathbb{R}^{3 \times 3}$ denote the parent lattice matrices. We interpolate the child lattice as

$$\mathbf{L}' = r \cdot \mathbf{L}_1 + (1 - r) \cdot \mathbf{L}_2,$$

with mixing ratio $r \in (0, 1)$, and populate it by randomly combining atoms from the two parents in proportion to r and $1 - r$. The resulting child is then geometrically sanitized by removing atoms involved in pairwise distances smaller than $d_{\min} = 0.75 (R_i + R_j)$, where R_i and R_j are covalent radii of the involved atoms, for up to three cleanup passes.

Prototype. Performs a global structural jump by instantiating a target composition in a crystal template that defines the lattice geometry and site topology. We use both fixed templates from a prior database and dynamic templates derived from the current parent pool; at each iteration, up to 10 high-scoring parent structures are added as reusable templates. To adapt a source template to new target elements, we rescale the lattice by a covalent-radius-based factor, which helps preserve reasonable interatomic distances after substitution.

A.2 Search Algorithm

Algorithm 1 presents the full `CrystalSTAR` search loop. We denote by C the number of children generated per iteration, by K the number of population groups, and by P the number of parents retained per group, so that the parent pool contains $K \times P$ structures in total.

A.3 Trio-Reflection (M^3) Configuration

Each M^3 level operates on a different temporal scale and serves a distinct role during search. At each iteration, all three levels are assembled and injected into the tool selector agent prompt in top-down order: macro prior, then meso tactical policy, then micro signals, so that the LLM planner consults cross-run strategic knowledge before finer-grained, within-run feedback. Table 5 lists the key hyperparameters; we describe each level below.

Micro events. Each evaluated structure produces one micro event recording its formula, evaluation status (`invalid/no_hit/hit`), decomposition energy E_d , source tool, and constraint satisfaction. Events are ranked by importance, hits first, then by recency, and the top 3 are injected into the next prompt as immediate feedback signals.

Algorithm 1 CrystalSTAR: Structured Action Orchestration with Trio-Reflection

Require: Task $\mathcal{T}$, parent pool $\mathcal{P}_0$ ($K \times P$ structures), M³ memory $\mathcal{M} = (\mathcal{M}_\mu, \mathcal{M}_m, \mathcal{M}_M)$

- 1: **for** iteration $t = 1, 2, \dots, T$ **do**
- 2: // — *Generate* —
- 3: Update dynamic templates from top-scoring structures in $\mathcal{P}_{t-1}$
- 4: Assemble search state s_t from $\mathcal{T}$, $\mathcal{P}_{t-1}$, and M³ summaries
- 5: Select C tool actions $\{(\tau_j, \Theta_j)\}_{j=1}^C$ via tool selector agent conditioned on s_t
- 6: **for** each action (τ_j, Θ_j) **do**
- 7: Execute tool: $x_j \leftarrow \tau_j(\text{parent}, \Theta_j)$
- 8: Relax structure via CHGNet (FIRE, $f_{\max}=0.05$ eV/Å, 500 steps)
- 9: Compute $E_{\text{hull}}(x_j)$ against per-chemsys convex hull
- 10: Predict task properties via ALIGNNN
- 11: Evaluate constraints: $\text{hit}_j \leftarrow \mathbb{1}[x_j \in \mathcal{C}_{\text{task}}]$
- 12: **end for**
- 13: // — *Reflect* —
- 14: Compute iteration statistics: valid rate, hit rate, diversity, tool performance
- 15: Generate LLM reflection analyzing generation quality
- 16: **Update M³:**
- 17: $\mathcal{M}_\mu \leftarrow$ record evaluation outcomes as micro events
- 18: **if** meso trigger **then** $\mathcal{M}_m \leftarrow$ LLM-derived tactical tip from recent micro events
- 19: **if** $t \bmod 10 = 0$ **then** $\mathcal{M}_M \leftarrow$ LLM-derived strategic rule from recent meso tips
- 20: // — *Curate* —
- 21: Rank $\mathcal{P}_{t-1} \cup \{x_1, \dots, x_C\}$ by weighted score $S_{\mathcal{T}}(x)$
- 22: Deduplicate via StructureMatcher
- 23: $\mathcal{P}_t \leftarrow$ top- $(K \times P)$ structures
- 24: **end for**
- 25: **return** All evaluated candidates and final parent pool $\mathcal{P}_T$

Table 5: M³ memory hierarchy configuration.

Level	Trigger	Stored / Injected [†]	Derivation	Cross-Run
Micro	Every evaluation	100 / 3	Direct from oracle	No
Meso	Every iteration	30 / 3	LLM synthesis	No
Macro	Every 10 iters	20 / 2	LLM synthesis, $\text{conf} \geq 0.55$	Yes

[†]Max stored entries / top- k injected into the tool selector agent per iteration.

Meso tips. At the end of each iteration, the LLM is prompted with recent micro events and historical reflections to produce a single tactical recommendation (e.g., “switch from SUBSTITUTE to PROTOTYPE for the Ba-family”), accompanied by a confidence score $c \in (0, 1]$ and supporting evidence. The confidence score reflects the strength of empirical support behind the recommendation and is used for downstream ranking and filtering.

Macro rules. Every 10 iterations, the LLM distills recent meso tips and historical entries into a single strategic rule intended to be transferable across runs for the same task (e.g., “Ba-based double perovskites consistently achieve high κ with moderate band gaps”). Rules whose confidence falls below a minimum threshold (Table 5) are discarded to prevent weakly supported heuristics from propagating. Accepted rules are retained in task-level persistent memory, enabling cross-run knowledge transfer: when a new run begins on the same task, previously stored macro rules are loaded as priors.

A.4 Tool Selection Prompt Structure

Below is a condensed illustration of the tool-selection prompt at a representative iteration on the High- κ Dielectrics task, highlighting the M³ injection points:

Tool Selection Prompt (Condensed)

[TASK CONTEXT]

Task: High- κ Dielectrics. Constraints: $10 \leq \kappa \leq 90$; $2.5 \leq E_g \leq 6.5$ eV.Soft targets: $\kappa \rightarrow 50$ ($\beta=0.35$), $E_g \rightarrow 4.5$ ($\beta=0.35$), $E_d \rightarrow 0$ ($\beta=0.30$).[M³ DECISION BACKBONE]

=== Macro Prior ===

1. Ba-based double perovskites favor high κ with moderate E_g . (conf=0.82)
2. Heavier alkaline-earth cations improve dielectric response. (conf=0.71)

=== Meso Tactical Policy ===

1. Targeted cation substitutions around Ba-family anchors show highest hit rate.
2. Tool signal: prioritize SUBSTITUTE (hit=42.0%, valid=88.0%).

=== Micro Immediate Signals ===

1. Ba₂HfO₄ $\rightarrow$ hit ($E_d=0.02$, $\kappa=31.2$, $E_g=3.8$)
2. SrTiO₃ $\rightarrow$ no_hit ($\kappa=7.8$, below threshold)
3. Ba₂ScTaO₆ $\rightarrow$ hit ($E_d=0.00$, $\kappa=28.7$, $E_g=4.1$)

[PARENT POOL]

Top candidates with formulas, properties, and weighted scores.

[INSTRUCTION]

Select 5 tool actions. For each, specify: tool name, parameters, parent choice, and strategy rationale. Output as JSON array.

B Benchmark, Metrics, and Evaluation Protocol**B.1 Task Specifications**

Table 6 lists the hard constraints for all 14 tasks. A candidate is counted as a *hit* only if it satisfies every constraint in its task. Decomposition energy ($E_{\text{hull}} \leq 0.1$ eV/atom) serves as a shared metastability filter and is evaluated separately from the task-specific hit rate.

Scoring function. As introduced in the main text, the task score $S_{\mathcal{T}}(x) = \sum_k \beta_k \rho_k(x)$ aggregates multiple objectives into a single scalar. Each per-property score ρ_k is normalized to $[0, 1]$ via *target-saturation*: for a property to be maximized (e.g., E_g , K , κ), $\rho_k = \min(v/v^*, 1)$ where v is the predicted value and v^* the target; for a property to be minimized toward zero (e.g., E_d), $\rho_k = \exp(-v/\epsilon)$ with scale ϵ set to the metastability threshold; for a property with a negative target (e.g., $E_f \rightarrow -1.5$), $\rho_k = \min(v/v^*, 1)$ when $v < 0$. Candidates violating any hard constraint receive $S_{\mathcal{T}} = 0$ regardless of their property values, ensuring that constraint satisfaction is a prerequisite for ranking.

B.2 Evaluation Metrics

Given N evaluated candidates $\{x_1, \dots, x_N\}$ produced during a search run, we define the following metrics:

Hit Rate (H.R.) The fraction of candidates satisfying all task-specific hard constraints:

$$\text{H.R.} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[x_i \in \mathcal{C}_{\text{task}}]. \quad (8)$$

Metastability Rate (Stab.) The fraction with surrogate-predicted energy above hull at most $\epsilon = 0.1$ eV/atom:

$$\text{Stab.} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[E_{\text{hull}}(x_i) \leq \epsilon]. \quad (9)$$

Table 6: Full task specifications for the 14-task benchmark. Hard constraints define the feasible region $\mathcal{C}_{\text{task}}$. Beyond constraint satisfaction, each task jointly optimizes the constrained properties and thermodynamic stability (E_d) via weighted scalarization $S_{\mathcal{T}}(x)$.

#	Task Name	Hard Constraints
1	Wide-Bandgap Semicond.	$E_g \geq 2.5 \text{ eV}; E_f \leq -1.0 \text{ eV/atom}$
2	SAW/BAW Acoustic Substr.	$25 \leq G \leq 150 \text{ GPa}; 3.7 \leq \kappa \leq 95$
3	High- κ Dielectrics	$10 \leq \kappa \leq 90; 2.5 \leq E_g \leq 6.5 \text{ eV}$
4	Solid-State Electrolytes ^a	$E_g \geq 2.0 \text{ eV}; E_f \leq -1.0 \text{ eV/atom}$
5	Piezoelectric Harvesters	$d_{ij} \geq 8 \text{ pC/N}; 10 \leq \kappa \leq 8000$
6	Photovoltaic Absorbers	$0.7 \leq E_g \leq 2.0 \text{ eV}; E_f \leq 0 \text{ eV/atom}$
7	Hard Coatings	$200 \leq K \leq 500 \text{ GPa}; 100 \leq G \leq 300 \text{ GPa}$
8	Hard, Stiff Ceramics	$100 \leq K \leq 300 \text{ GPa}; 60 \leq G \leq 200 \text{ GPa}$
9	Aerospace Materials ^b	$K \geq 100 \text{ GPa}; G \geq 40 \text{ GPa}; \rho \leq 5.0 \text{ g/cm}^3$
10	Acousto-Optic Hybrids	$2 \leq d_{ij} \leq 9 \text{ pC/N}; 8 \leq \kappa \leq 95$
11	Low-Density Structural	$\rho \leq 3.5 \text{ g/cm}^3; 65 \leq G \leq 195 \text{ GPa}$
12	Toxic-Free Perovsk. Oxides ^c	$E_g \geq 2.0 \text{ eV}; 90 \leq K \leq 135 \text{ GPa}$
13	Insulating Dielectrics	$\kappa \geq 8.0; E_g \geq 2.5 \text{ eV}$
14	Transparent Conductors	$E_g \geq 3.0 \text{ eV}$

Notation: E_g = band gap, E_f = formation energy, E_d = decomposition energy (E_{hull}), K = bulk modulus, G = shear modulus, κ = dielectric constant, d_{ij} = piezoelectric coefficient, ρ = density.

^aMust contain at least one of {Li, Na, K, Mg, Ca, Al}. ^bPrefer {Al, Ti, Mg, Li, Si, B, C, N, V, Zr, Nb}.

^cExclude {Pb, Cd, Hg, Tl, Be, As, Sb, Se, U, Th}.

Novel Hit Rate (N.H.R.) The fraction of candidates that are both task-valid and structurally absent from the reference database $\mathcal{D}_{\text{ref}}$ (e.g. Materials Project):

$$\text{N.H.R.} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[x_i \in \mathcal{C}_{\text{task}} \wedge x_i \notin \mathcal{D}_{\text{ref}}]. \quad (10)$$

Diversity Rate (Div.) The fraction of structurally unique candidates after deduplication via pymatgen’s StructureMatcher (ltol=0.2 Å, stol=0.3, angle_tol=5°):

$$\text{Div.} = \frac{|\{x_i : x_i \text{ unique under StructureMatcher}\}|}{N}. \quad (11)$$

Novelty Rate (Nov.) The fraction identified as novel relative to $\mathcal{D}_{\text{ref}}$, regardless of task validity:

$$\text{Nov.} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[x_i \notin \mathcal{D}_{\text{ref}}]. \quad (12)$$

B.3 Novelty Checking Protocol

Our novelty checker implements a three-tier cascade, proceeding from fast compositional lookups to detailed structural comparison:

- Tier 1 — Composition Check.** We query the Materials Project (MP) API for all entries in the candidate’s chemical system (chemsys). If the chemsys is entirely absent from MP, or if the candidate’s reduced composition does not appear among MP entries within that chemsys, the candidate is declared *novel* at Tier 1.
- Tier 2 — Quick Structural Filters.** For candidates whose composition matches known MP entries, we apply three rapid structural filters against each composition-matched entry:
 - Space group number (via SpacegroupAnalyzer, symprec = 0.1)

- Atom count (± 2 tolerance for primitive vs. conventional cell differences)
- Density ($\pm 10\%$ relative tolerance)

If all three filters differ from every composition-matched MP entry, the candidate is declared *novel* at Tier 2.

3. **Tier 3 — Detailed Structure Matching.** Remaining candidates are compared against all MP entries in the chemsys using pymatgen’s StructureMatcher with mode-dependent tolerances. In our experiments, we use the *standard* configuration:

- Lattice tolerance: 0.2 Å
- Site tolerance: 0.3
- Angle tolerance: 5°

If no MP entry produces a match, the candidate is declared *novel* at Tier 3.

This cascaded design minimizes expensive structure-matching calls: most novel candidates are caught at Tier 1 (new composition) or Tier 2 (new polymorph with distinct symmetry), while Tier 3 resolves the hardest cases involving subtle structural differences within the same composition and symmetry class.

B.4 Novelty Analysis by Composition and Structure

Novel Hit Rate combines two types of novelty. A candidate has *composition-level* novelty if its reduced composition is absent from the reference database. This corresponds to Tier 1 of the novelty check in Appendix B.3. A candidate has *structure-level* novelty if its composition is already known but its generated structure does not match any entry in the reference database after the quick structural filters and StructureMatcher. This includes Tiers 2 and 3.

We also examine whether the search reaches different structures for the same reduced composition. Table 7 shows that structure-level cases account for 12.61% to 73.81% of novel hits across tasks and exceed 50% in four tasks. In every task, the search also reaches multiple distinct structures for at least one reduced formula; 1.32% to 16.52% of unique compositions have multiple structures, with a maximum of two to seven structures per formula. This shows that novelty is not merely formula-level: many novel hits are structure-level cases, and the multi-structure statistics confirm that CrystalSTAR can generate distinct realizations at fixed composition.

B.5 Constraint Boundary Analysis

Since the evaluation relies on property predictions from CHGNet and ALIGNN, the reported results should be interpreted as surrogate-guided evidence rather than direct first principle validation. We therefore consistently refer to successful candidates as surrogate-predicted metastable hits. To examine whether these hits are concentrated near the task constraint boundaries, we conduct a boundary proximity analysis.

For each discovered hit x , we define a conservative Boundary Distance (BD) that measures its margin from the closest task constraint boundary. For a bounded constraint $l_j \leq v_j(x) \leq u_j$, we define

$$d_j(x) = \frac{\min\{v_j(x) - l_j, u_j - v_j(x)\}}{(u_j - l_j)/2}.$$

For one-sided constraints, we normalize the feasible-side margin by the empirical interquartile range $\text{IQR}_j = Q_{3,j} - Q_{1,j}$, where $Q_{1,j}$ and $Q_{3,j}$ denote the first and third quartiles of the predicted values for property j :

$$d_j(x) = \frac{v_j(x) - l_j}{\text{IQR}_j} \quad \text{for } v_j(x) \geq l_j,$$

and

$$d_j(x) = \frac{u_j - v_j(x)}{\text{IQR}_j} \quad \text{for } v_j(x) \leq u_j.$$

Table 7: Novelty analysis by composition and structure across the 14 materials design tasks. Composition-level novelty and Structure-level novelty are reported as percentages of novel hits. *Unique compositions* report the number of unique reduced compositions; *Composition with ≥ 2 Structures* denotes the fraction of unique novel hit compositions for which CrystalSTAR generated at least two structurally distinct candidates; *Max Struct. per Formula* denotes the maximum number of structurally distinct candidates generated for any single reduced formula.

Metric	Wide-Bandgap	SAW/BAW	High-k Dielect.	Solid-State Elect.	Piezo Energy	Transp. Cond.	Insul. Dielect.
N.H.R. (%)	44.4	32.4	32.1	48.8	56.0	28.8	36.6
Composition-level (%)	87.4	53.1	47.5	83.6	51.4	40.3	39.6
Structure-level (%)	12.6	46.9	52.5	16.4	48.6	59.7	60.4
Unique compositions	137	129	122	147	151	151	136
Comp. w/ ≥ 2 Struct. (%)	11.7	6.2	5.7	3.4	1.3	9.9	6.6
Max Struct. per Formula	5	3	6	3	2	3	4

Metric	Photo-voltaics	Hard Coatings	Hard Ceramics	Aerospace Mat.	Acousto-optic	Low Density	Perovskites
N.H.R. (%)	34.7	42.0	39.4	55.4	11.2	35.1	16.8
Composition-level (%)	81.4	60.0	62.2	68.8	75.0	77.0	26.2
Structure-level (%)	18.6	40.0	37.8	31.2	25.0	23.0	73.8
Unique compositions	139	132	115	103	149	130	143
Comp. w/ ≥ 2 Struct. (%)	6.5	9.9	16.5	16.5	4.0	13.9	7.7
Max Struct. per Formula	4	5	7	5	2	4	3

Table 8: Median Boundary Distance over discovered hits for seven materials design tasks. Larger values indicate a greater normalized margin from the closest constraint boundary.

Method	Wide-Bandgap	SAW/BAW	High-k Dielect.	Solid-State Elect.	Piezo Energy	Transp. Cond.	Insul. Dielect.
LLEMA	0.337	0.138	0.074	1.462	0.008	0.235	0.292
CrystalSTAR	0.944	0.622	0.291	1.794	0.003	1.285	1.060

For tasks with multiple constraints, we define

$$\text{BD}(x) = \min_j d_j(x),$$

so that the metric captures the closest constraint boundary. Larger BD values indicate a larger margin from the closest constraint boundary, making feasibility labels less sensitive to small surrogate-prediction errors.

As an immediate robustness check, Table 8 reports the median BD over discovered hits for the seven tasks included in this analysis. **CrystalSTAR** generally exhibits larger constraint margins than LLEMA, with Piezoelectric Energy Harvesters as the only exception.

B.6 Parent Pool Construction

The parent pool $\mathcal{P}_0$ seeds both the evolutionary search and the dynamic template library for PROTOTYPE. For each task, we curate a prior database from the Materials Project and related repositories, retaining structures that fall within or near the task’s constraint region, and annotate each entry with precomputed surrogate predictions so that downstream sampling requires no additional oracle calls. From this database we draw $K \times P$ structures via stratified sampling over crystal system, element count, and site count to ensure broad structural coverage rather than clustering around a single prototype family; element-level inclusion or exclusion constraints specified per task are enforced during selection. The top-scoring structures in the resulting pool are then registered as dynamic templates at each iteration, providing the agent with task-relevant lattice scaffolds from the outset.

C Implementation Details

C.1 Surrogate Oracle Details

Structural relaxation and stability. Every candidate generated by the agent undergoes a two-stage surrogate evaluation before entering the parent pool: structural relaxation followed by thermodynamic stability assessment. We adopt CHGNet (Deng et al., 2023), a pretrained universal machine-learning interatomic potential (MLIP) that jointly models energies, forces, and stresses across the periodic table, enabling relaxation of arbitrary compositions without task-specific retraining. Structures are relaxed using the FIRE optimizer with $f_{\max} = 0.05$ eV/Å and a maximum of 500 steps. Thermodynamic stability is then quantified by the decomposition energy E_{hull} , computed against a per-chemsys convex hull constructed from up to 500 Materials Project reference entries whose energies are recomputed with CHGNet for consistency with the candidate energies.

Property prediction. Task-specific properties are predicted by ALIGNN (Choudhary & DeCost, 2021) using pretrained checkpoints from the JARVIS repository. Table 9 lists the property-to-model mapping.

Table 9: ALIGNN model checkpoints used for property prediction.

Property	ALIGNN Checkpoint
Band gap (E_g)	jv_optb88vdw_bandgap_alignn
Bulk modulus (K)	jv_bulk_modulus_kv_alignn
Shear modulus (G)	jv_shear_modulus_gv_alignn
Formation energy (E_f)	jv_formation_energy_peratom_alignn
Piezoelectric coeff. (d_{ij})	jv_dfpt_piezo_max_dij_alignn
Dielectric constant (κ)	jv_dfpt_piezo_max_dielectric_alignn

ALIGNN inference is executed in a separate conda environment (matagent-alignn) via subprocess call with a 300-second timeout. Density (ρ) is computed directly from the relaxed structure’s lattice and composition without a surrogate model.

C.2 Protocol-Aligned Stability Evaluation

Table 1 reports the published baseline results under the original LLEMA evaluation protocol (Abhyankar et al., 2025). To examine whether the Metastability Rate comparison depends on the stability oracle, we re-evaluate the same candidates generated by CrystalSTAR using a protocol aligned with LLEMA, querying the Materials Project when values are available and using ALIGNN predictions otherwise. As shown in Table 10, CrystalSTAR retains higher Metastability Rate than LLEMA (GPT) across the benchmark.

C.3 Hyperparameter Summary

Table 11 summarizes the key hyperparameters used across all experiments.

C.4 Ablation Configurations

Generation mechanism. The four crystal tools impose a strong structural prior on the search: every candidate is derived from an existing parent through a well-defined geometric transformation. **Direct Generation** removes this prior entirely, the LLM is instead prompted to output complete CIF-format crystal structures from scratch, given the same task description, parent pool, and reflection feedback. The search loop, surrogate evaluation, and curation pipeline remain identical. This ablation tests whether the tool abstraction is merely a convenience or a critical inductive bias: can an LLM produce physically plausible crystal structures without being grounded through parent-based transformations?

Table 10: Metastability Rate (St., %) across 14 tasks under the original and protocol-aligned evaluation settings. MP+ALIGNN and CHGNet evaluate the same candidates generated by CrystalSTAR.

Method	Wide-Bandgap	SAW/BAW	High-k Dielect.	Solid-State Elect.	Piezo Energy	Transp. Conduct.	Insul. Dielect.
LLEMA (GPT)	22.4	10.7	12.7	25.4	3.2	14.9	4.6
CrystalSTAR (MP+ALIGNN)	44.0	58.0	66.8	54.2	52.8	60.8	65.0
CrystalSTAR (CHGNet)	81.2	68.4	75.9	62.4	76.0	62.0	71.5

Method	Photovoltaics	Hard Coatings	Hard Ceramics	Aerospace Mat.	Acousto-optic	Low Density	Perovskites
LLEMA (GPT)	4.8	4.6	5.7	0.3	0.8	0.1	2.8
CrystalSTAR (MP+ALIGNN)	48.8	56.5	49.4	47.6	59.6	38.9	63.7
CrystalSTAR (CHGNet)	71.4	67.2	57.8	70.7	75.2	40.3	63.2

Table 11: CrystalSTAR hyperparameters.

Category	Parameter	Value
LLM	Model	gpt-4o-mini
	Temperature	0.8
	Max tokens	8192
	Provider	OpenAI
Search	Population groups (K)	5
	Parents per group (P)	4
	Total parent pool	$K \times P = 20$
	Children per iteration (C)	5
Evaluation	MLIP backend	CHGNet
	Relaxation optimizer / $f_{\max}$ / steps	FIRE / 0.05 eV/Å / 500
	Metastability threshold (ϵ)	0.1 eV/atom
	MP entries per chemsys	≤ 500
Pre-filtering	Elemental skip filter	Single-element O, F
	Min distance factor	0.75

Memory architecture. The full CrystalSTAR system uses M^3 to compress search history into a hierarchical prompt backbone (Section A.3). We compare two degraded variants that remove different aspects of this design. **ReAct** retains the tool-calling pipeline but strips the M^3 backbone entirely. The planner receives only the current iteration’s evaluation statistics—hit rate, validity rate, diversity—and the top-performing structures as short-horizon context. No information accumulates across iterations: tactical patterns observed at iteration 5 are invisible at iteration 15. This ablation isolates the value of cross-iteration memory accumulation. **Raw Context** takes the opposite approach: instead of discarding history, it preserves as much as possible. Raw evaluation logs from recent iterations are prepended verbatim into the prompt without summarization or hierarchical organization. This tests whether the LLM’s in-context learning can substitute for explicit memory structuring—in principle, all the information that M^3 distills is present in the raw logs, but the planner must extract patterns from unstructured text under a finite context budget.

Tool portfolio. Our four tools span a spectrum from local perturbation (SUBSTITUTE, MUTATE) to global structural recombination (MIX, PROTOTYPE). **Global Jump only** disables SUBSTITUTE and MUTATE, restricting the agent to global operators; **Local Fine-tuning only** disables MIX and PROTOTYPE, restricting the agent to local perturbations. These two ablations quantify the complementarity between exploration and exploitation: global tools can escape local basins but produce less stable structures, while local tools refine promising candidates but risk premature convergence.

D Additional Results

D.1 Tool Usage Distribution

Figure 6 shows the distribution of tool calls across the 14 tasks. Several patterns emerge: (1) SUBSTITUTE is the most frequently selected tool overall, consistent with its role as the primary local refinement operator; (2) mechanical-property tasks (Hard Coatings, Aerospace Materials) show higher usage of PROTOTYPE, likely because structural topology plays a stronger role in determining mechanical properties; (3) MIX usage increases on tasks with broader element constraints, where compositional diversity is more beneficial.

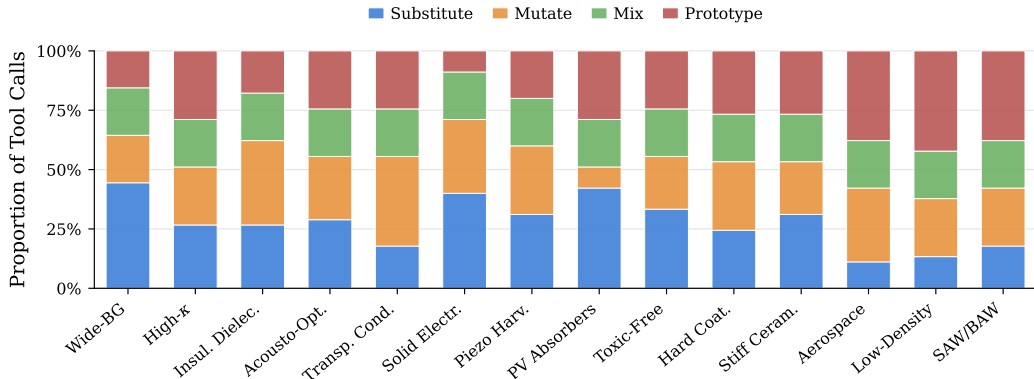


Figure 6: Tool usage distribution across 14 benchmark tasks.

D.2 Additional Ablations on Trio-Reflection (M^3)

We first repeat the ReAct, Raw Context, and **CrystalSTAR** comparison on two additional tasks from different domains: Hard Coatings, representing mechanical materials design, and Wide-Bandgap Semiconductors, representing electronic materials design. As shown in Table 12, **CrystalSTAR** achieves the highest Diversity Rate and cumulative number of unique novel hits on both tasks. On Hard Coatings, ReAct attains a slightly higher N.H.R. and Raw Context attains a higher H.R.; nevertheless, **CrystalSTAR** discovers substantially more unique novel candidates. On Wide-Bandgap Semiconductors, **CrystalSTAR** achieves the highest Diversity Rate, Novelty Rate, N.H.R., and cumulative number of unique novel hits. These results support the conclusion that M^3 primarily improves the breadth and persistence of discovery rather than simply maximizing short-term H.R.

To examine the role of each memory level, we conduct a component-level ablation on High- κ Dielectrics. The results suggest that the three levels contribute in different ways. Removing Macro causes the largest drop in N.H.R., from 32.1% to 24.4%, highlighting the importance of inherited cross-run priors for valid novel discovery. Without Meso, the Diversity Rate drops to 41.8% and the cumulative number of unique novel hits decreases from 48 to 35, suggesting that within-run tactical adaptation helps sustain exploration. Removing Micro has a smaller and more mixed effect: several rate-based metrics improve slightly, while H.R. decreases from 65.9% to 60.4% and cumulative unique novel hits decrease from 48 to 45. This suggests that Micro mainly stabilizes local feedback rather than directly expanding the search space.

D.3 Learned Macro Rule Examples

Table 13 shows representative macro rules learned by the M^3 mechanism during search. These rules illustrate how **CrystalSTAR** progressively distills search-level experience into reusable scientific heuristics. These examples show that macro rules capture domain-relevant chemical intuition, such as the suitability of Ba-perovskites for dielectric applica-

Table 12: Additional cross-task and component-level ablations of Trio-Reflection. Cum. Hits denotes cumulative unique novel hits. Best results within each task block are **bold**.

Task	Configuration	Div.	Nov.	H.R.	N.H.R.	Cum. Hits
Hard Coatings	ReAct	34.1	52.2	89.1	44.2	45
	Raw Context	31.2	40.4	91.2	34.4	47
	CrystalSTAR	60.4	61.6	74.8	42.0	74
Wide-Bandgap Semiconductors	ReAct	46.4	69.2	72.0	42.8	37
	Raw Context	49.2	60.4	70.4	37.2	50
	CrystalSTAR	64.0	70.8	71.2	44.4	74
High- κ Dielectrics	w/o Micro	57.2	64.0	60.4	32.8	45
	w/o Meso	41.8	56.2	69.9	26.5	35
	w/o Macro	54.8	61.2	56.8	24.4	42
	CrystalSTAR	53.8	60.3	65.9	32.1	48

tions or boride/nitride systems for lightweight structural materials that would typically require expert knowledge to encode manually.

Table 13: Examples of macro rules learned during search. Confidence scores reflect the LLM’s self-assessed reliability of each rule based on accumulated evidence.

Task	Conf.	Macro Rule
High- κ Dielectrics	0.82	Ba-based double perovskites consistently achieve high dielectric constants with moderate band gaps. Prioritize A-site substitutions within Ba-perovskite family.
Wide-Bandgap Semiconductors	0.78	Alkali fluoride families (K–Na–Ga–F systems) provide a robust basin for wide-gap hits. Explore compositional variants within elpasolite-like structures.
Aerospace Materials	0.71	Light-element borides and nitrides (B–C–N based) with ≤ 4 elements yield the best stiffness-to-density ratio. Avoid heavy transition metals beyond Ti and Zr.

D.4 Efficiency Analysis

To evaluate discovery efficiency, we compare **CrystalSTAR** with LLEMA using per-discovery resource consumption: total tokens, output tokens, API calls, wall-clock time, and the number of candidates evaluated before the first hit. Table 14 summarizes these results. **CrystalSTAR** reaches its first hit roughly $3\times$ faster and requires approximately half the total tokens per hit compared to LLEMA. The strongest gap appears on the output side ($\sim 20\times$ fewer output tokens per hit), because **CrystalSTAR** emits compact JSON tool actions whereas LLEMA generates longer free-form structure proposals. Overall, **CrystalSTAR** achieves higher discovery quality while requiring substantially less compute per successful hit, suggesting that tool-grounded search offers a favorable efficiency performance for closed-loop materials discovery.

Table 14: Per-discovery efficiency on the 14 benchmark tasks.

Method	Cand. to 1st Hit	Tokens/Hit	Out Tokens/Hit	API Calls/Hit	Time/Hit (s)
CrystalSTAR	5.4	$\sim 2.9\text{k}$	19.3	1.01	51.3
LLEMA	18.2	$\sim 5.4\text{k}$	395.7	8.38	93.9