STRUCT-R1: MATERIAL STRUCTURE SEARCH WITH MULTI-FIDELITY MEASUREMENT IN AGENTIC LOOPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Materials discovery requires a research loop in large structure space with multi-tier measurements with different fidelity and costs, from machine-learned surrogates to DFT calculations and experiments. Manual exploration can be time-consuming, while statistical search methods and existing agent designs struggle to achieve high search efficiency with trade-off between fidelity and cost.

We present Struct-R1, a self-evolving agentic framework with two coupled loops. The inner *search loop* performs physics-grounded reasoning and perturbation proposal, and manages cross-tier relationships to improve search efficiency. The outer *evolution loop* reflects on individual search steps and converts their feedback into reusable knowledge. Through the two loops, Struct-R1 progressively refines both where to explore and how to allocate search budget on the measurements. We evaluate Struct-R1 on representative material-property search tasks, such as Raman and phonon properties, and further validate it using real research data for well-studied materials to assess its effectiveness in realistic scientific settings.

1 INTRODUCTION

The study of atomic structures and their perturbations is a central theme in materials science (Oganov et al., 2019). A material structure's composition and atomic arrangement govern its important properties such as band gap, optical absorption, and mechanical strength. To tune these properties toward specific targets, researchers perturb a stable parent structure through strain, vacancies, or substitution (Li et al., 2014; Queisser & Haller, 1998). For example, strain tunes the band gap and carrier dynamics of halide perovskites such as $MAPbI_3$ (Zhu et al., 2019), while Te vacancies can drive $HfTe_5$ across a topological phase boundary (Jo et al., 2024). Such research on structural perturbations drives the discovery of new materials and guides their design and synthesis (Zunger, 2018).

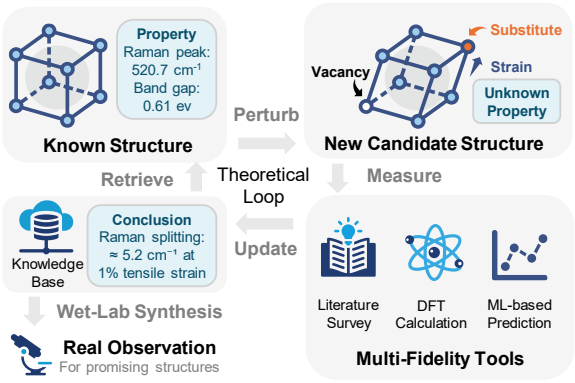

Figure 1: Illustration of the theoretical loop in computational materials science.

Materials researchers commonly follow a **theoretical loop** (Curtarolo et al., 2013). As illustrated in Fig. 1, researchers start from a known structure and propose perturbations based on physical knowledge and the desired property changes. The resulting candidates can be evaluated through multiple approaches (Fare et al., 2022): retrieving and summarizing findings from existing literature, performing high-fidelity DFT simulations, or using ML-based models to predict properties and screen candidates. The results are analyzed to determine how each perturbation affects the material's properties and can be integrated with the existing knowledge base to provide theoretical guidance. Theoretical promising candidates are further selected for real synthesis and validation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

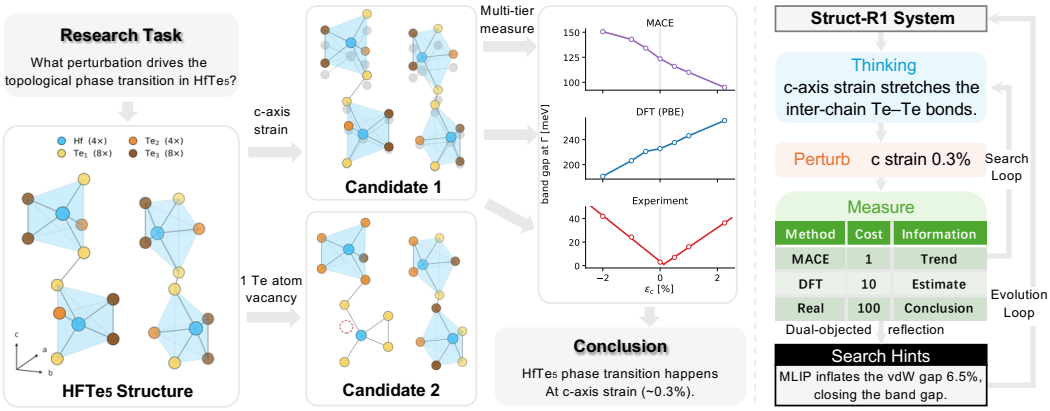

Figure 2: Overview of Struct-R1. The inner search loop manages multi-fidelity measurements during structure exploration, while the outer evolution loop reflects on search trajectories and updates reusable knowledge to improve search efficiency.

Such theoretical design loop for material structures remains labor-intensive. Researchers have devoted extensive time and effort to studying structure–property relationships, testing perturbation hypotheses, and screening large numbers of candidate structures through repeated cycles of proposal and verification to achieve target properties. For example, $HfTe_5$ was predicted to be a topological insulator in 2014 (Weng et al., 2014), yet conflicting measurements left its ground state unresolved for nearly a decade, until controlled strain (Liu et al., 2024) and defect (Jo et al., 2024) studies separated the competing explanations.

With the rapid progress of LLM-driven autonomous research (Zheng et al., 2025), a natural question arises: *can LLM agents be applied to this theoretical loop and improve the efficiency of materials structure discovery?* However, direcly applying autonomous agents to the structure-discovery loop poses distinct challenges: **First**, the perturbation space is combinatorially large, while a high-fidelity DFT evaluation of a complex system can require hundreds of CPU-hours. This makes exhaustive verification impractical. ML-based surrogates provide fast estimates but may lack the accuracy needed for reliable materials conclusions. Efficient discovery therefore requires tools and data across multiple fidelity tiers. **Second**, the relationship among these tiers is itself material-, property-, and region-dependent. A low-fidelity method may accurately capture a trend in one region, shift a transition boundary in another, or become uninformative altogether. An effective search strategy must therefore determine not only *where* to explore, but also *which fidelity* to use, *when* to switch fidelity, and *how* to interpret information across tasks. These decisions still rely heavily on expert experience.

To address these challenges, we present Struct-R1, a **close-looped agentic framework for cost-aware materials structure search**. As shown in Figure 2, Struct-R1 consists of two nested loops. In the inner *search loop*, Struct-R1 combines physics-grounded exploration with multi-fidelity measurements to efficiently navigate the perturbation space. To reason about heterogeneous measurements, it maintains structured tier records that capture the cost, cross-tier relationship, and useful information provided by each fidelity tier, allowing the agent to adaptively select and interpret measurements during search. In the outer *evolution loop*, Struct-R1 continuously reflects on completed search trajectories to improve subsequent searches. We design a dual-objective reflection mechanism that considers both search effectiveness and measurement cost, and analyzes individual search steps together with complete trajectories to update the harness system.

2 Background

2.1 Electronic structure and perturbation

Problem formulation. We represent a crystalline material as $M = (L, B)$, where $L \in \mathbb{R}^{3\times3}$ denotes the lattice matrix and $B = \{(e_i, x_i)\}_{i=1}^{N}$ denotes the atomic basis, with element type e_i and

fractional coordinate x_i for each atom. A structural perturbation is an operator π that transforms a parent structure into a new configuration, $M' = \pi(M) = (L', B')$.

We consider three common classes of perturbations. Let $\widetilde{M} = (L', \widetilde{B})$ denote the structure immediately after a perturbation and $M' = (L', B')$ the structure after ionic relaxation. *Strain* modifies the lattice as $L' = (I + \varepsilon)L$, where ε is the strain tensor; the atomic coordinates are initially inherited from the parent structure and relaxed within the strained cell. *Vacancy* perturbations remove a selected subset of atoms $S \subseteq 1, \ldots, N$ from the atomic basis, while *substitution* perturbations replace selected atoms of type e_a with e_b. The resulting perturbed structures are then relaxed to obtain B'.

Key challenges: expensive exploration and complex relationships across measurements. Automating this search loop poses two key challenges. **First**, the perturbation space is large, while reliable verification is expensive. High-fidelity DFT or experiments cannot be applied exhaustively, whereas cheaper machine-learned models may be unreliable across materials and search regions. Efficient exploration must therefore balance *where* to search with *which fidelity* to use, reserving expensive verification for cases where it is necessary. **Second**, the relationships among perturbations, properties, and measurement tiers are complex and vary across materials and tasks. Determining which trends are reliable, which measurements are informative, and how to adapt the search strategy often requires domain knowledge accumulated through years of theoretical and experimental practice. Capturing and reusing such expertise is therefore essential for efficient autonomous search.

2.2 Exploring the Perturbation Space: Existing Approaches and Limitations

Automated DFT frameworks. Significant progress has been made in automating materials DFT computation. Workflow engines such as AiiDA (Pizzi et al., 2016) and atomate (Mathew et al., 2017) can orchestrate large numbers of first-principles calculations, while recent agentic frameworks (Wang et al., 2025d; Hu et al., 2026; Xia et al., 2025) further automate computational parameter configuration, workflow construction, and execution. However, these systems primarily automate the computational process for a given material candidate. Search directions and plans in the broader structure search loop still rely heavily on human expertise. Moreover, relying solely on DFT for search would incur prohibitive computational cost.

Statistical and surrogate-based approaches. Bayesian optimization and active learning (Lookman et al., 2019; Xue et al., 2016) model the target property from evaluated candidates and select subsequent measurements through an acquisition function. Multi-fidelity methods further relate evaluation tiers through linear (Kennedy & O'Hagan, 2000) or nonlinear (Perdikaris et al., 2017) mappings, allowing cheap measurements to guide expensive ones. Learned surrogates, including descriptor-based models (Ouyang et al., 2018), graph neural networks (Chen et al., 2019; Xie & Grossman, 2018), and universal interatomic potentials (Chen & Ong, 2022; Batatia et al., 2022), instead predict properties directly at much lower cost. Both approaches learn from sampled data rather than physical reasoning, and can remain tied to their training distribution. As a result, both can become unreliable in poorly sampled regions, leading to search misguidance.

LLM-agent-based approaches. LLM agents have demonstrated goal-directed search across different scientific domains, including protein structures (Wu et al., 2026; Wang et al., 2025c), molecular structures (Chen et al., 2026; Li et al.) and chemical design (Wang et al., 2025b;a). In the field of material design, LLMatDesign (Jia et al., 2024) prompts an LLM to propose structural modifications to a known material and reflects on past decisions, using ML-based models for relaxation and property prediction. MatEvolve (Zhong et al.) focuses on knowledge-guided material editing, leveraging literature-derived modification rules and symbolic operators to optimize existing materials. LLEMA (Abhyankar et al., 2026) performs LLM-driven evolutionary search over candidate materials using surrogate-based multi-objective feedback. However, these agents primarily rely on surrogate models for feedback, without the verification mechanism between high-fidelity DFT calculations and low-fidelity surrogate data. To our knowledge, in materials structure search, agentic systems that coordinate measurement tiers with different fidelities, learn their cross-tier relationships, and adaptively determine the fidelity for effieicnt search remain largely unexplored.

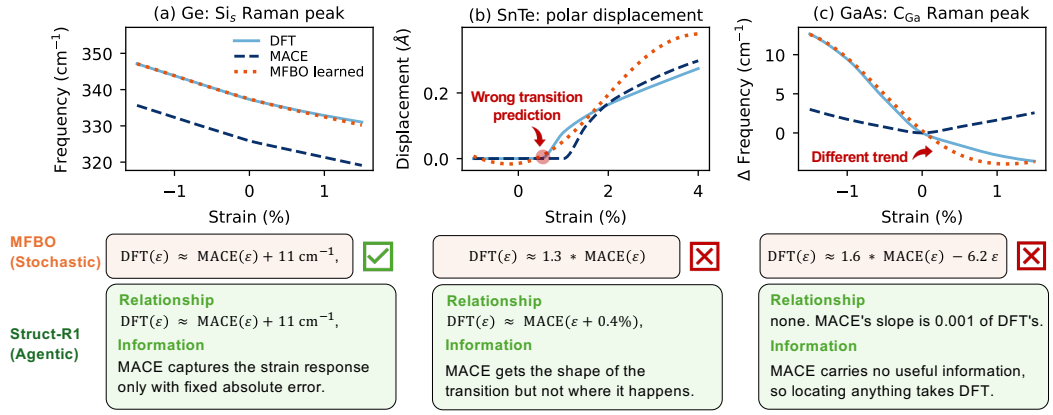

Figure 3: Structured cross-tier relationships for multi-fidelity measurements.

3 Struct-R1: Design and Implementation

3.1 Structured Reasoning and Search Loop

Each search step follows a reasoning–action loop (Yao et al., 2022) to keep scientific information in the context: `<think>` → `<perturb>` → `<measure>`. The agent first explains in `<think>` why a perturbation may improve the target property. It then proposes a concrete perturbation in `<perturb>` using the operator π defined in Section 2.1. The operator supports strain, vacancy, substitution, and doping, as well as their compositions. Each operation specifies concrete operands, such as the crystallographic axis, atomic species, site, concentration, and perturbation magnitude. A structure-construction backend built on `pymatgen` (Ong et al., 2013) applies the proposed operation by constructing the required supercell and updating the lattice and atomic basis. Then in `<measure>`, the agent gets feedback from a series of external measurements, ranging from low-cost surrogate predictions to DFT calculations, as described in Section 3.2. This loop continues throughout the search, allowing the agent to iteratively explore the perturbation space.

3.2 Multi-Fidelity and Cost-Aware Measurement

Multi-fidelity theoretical measurements. To reduce the cost of computational search, Struct-R1 mainly provides two fidelity tiers. ❶ At the low-fidelity tier, Struct-R1 evaluates each candidate using a machine-learned interatomic potential (MLIP) (Batatia et al., 2022), typically within seconds. This tier is used to identify trends and prune candidates rather than to produce final results. ❷ At the high-fidelity tier, Struct-R1 performs the complete DFT workflow with well-converged numerical settings, providing reliable property evaluation at much higher computational cost.

For some target properties, however, the MLIP cannot directly produce the required measurement, such as phonon or Raman properties. In these cases, Struct-R1 uses the MLIP to pre-relax the structure and then performs the corresponding property calculation with DFT. Since structural relaxation often accounts for a large fraction of the DFT workflow, this pre-relaxation can reduce the number of DFT relaxation steps and lower the overall cost. However, an inaccurate MLIP relaxation may lead the structure toward an incorrect local minimum, so this path provides a cheaper but less reliable approximation to the full DFT calculation results.

Structured cross-tier records. The key to exploiting multi-tier data is understanding what information transfers across multi-tier data, how it transfers, and where that relationship and fidelity remains valid. To make this knowledge explicit, Struct-R1 maintains a structured record for each measurement tier with four fields: `cost` records the esitimated execution time of the method, `relationship` describes the current belief about how measurements at this tier relate to the expected ground truth, and `information` describes the role of this tier in the search process.

The `relationship` field is the core for selection between multi-tier data. Instead of relying on statistical models used in Bayesian optimization (Poloczek et al., 2017; Kandasamy et al., 2016), it

captures more flexible and comprehensive information. Figure 3 shows three representative cases. For Ge, MACE follows the DFT response with a nearly fixed offset, which can also be captured by MFBO. For SnTe, however, MACE captures the transition shape but shifts its location, causing a statistical fit to predict an incorrect transition point. For GaAs, MACE and DFT exhibit different trends, where fitting a numerical cross-tier relationship provides little useful information. In contrast, Struct-R1 explicitly records these relationships and their limitations. The `information` field translates each relationship into its practical role. The lower-fidelity tier can be used to track a trend for Ge, identify a transition region for SnTe, or be excluded from the search for GaAs.

Tier selection with lookahead cost-fidelity tradeoff. Instead of selecting each measurement independently , Struct-R1 plans a short sequence of measurements and specifies the role of each measurement, such as identifying a trend, narrowing the candidate region, or validating a tier record. For each planned measurement, the agent queries the tier records and selects the lowest-cost tier that can provide the required information. The agent does not plan measurements only from statistical black-box data analysis. Instead, it enables the agent to analyze each measurement with physical knowledge together with the data analysis. For example, the agent can use physical constraints such as symmetry and known scaling relations to eliminate candidate regions or determine the direction of subsequent exploration without additional measurements. Each measurement is therefore planned based on what it can resolve physically, rather than only its expected numerical gain.

3.3 Evolution Loop with Dual-Objective Reflection

Self-evolving harness. Struct-R1 uses self-evolution with reflection on historical search trajectories to improve search efficiency. The system continuously evolves three knowledge fields:

- *Measurement configuration* records how each measurement tier should be configured, such as numerical parameters and pesudopotential choices for DFT calculation. These entries help avoid repeated failures caused by incorrect or inefficient measurement configurations.
- *Search hints* record reusable strategies for exploring the perturbation space, such as how to probe a trend, when to verify a candidate, and when to switch measurement strategies. These hints allow the agent to reuse successful strategies across search loops.
- *Tier records* capture the relationship among measurement tiers, as described in Section 3.2.

Dual-objective self-reflection. Existing harness evolution systems typically reflect on an entire trajectory (Agrawal et al., 2026; Zhang et al., 2026), where task-level success or failure provides a clear signal for evolution. However, structure search introduces the fidelity–cost tradeoff: a measurement can be scientifically accurate but unnecessarily expensive, while an inexpensive measurement may provide useful search feedback without being sufficiently accurate. Struct-R1 therefore adopts a two-stage reflection strategy to gain improvement from sparse and expensive search trajectories.

In the first stage, Struct-R1 reflects on individual search steps from two aspects. ❶ *Measurement validity.* The agent examines execution logs and returned results for possible issues in measurement execution, such as numerical convergence or incomplete property calculations. ❷ *Fidelity and cost.* The agent compares available measurements across fidelity tiers to assess how informative the lower-fidelity result is and whether additional high-fidelity verification may be useful. These observations are used to refine the measurement configurations and tier records.

In the second stage, Struct-R1 reflects on the complete search trajectory to improve the search strategy. It reviews whether parts of the trajectory appear redundant, overly expensive, or insufficiently explored, and summarizes useful patterns for future searches. For example, the reflection may suggest probing additional regions, introducing higher-fidelity verification, or avoiding measurements that provided little useful information. The resulting suggestions are merged into the search hints to guide subsequent search loops.

3.4 System implementation

Struct-R1 builds its computational backend on components of TritonDFT (Hu et al., 2026), including automatic parameter configuration and convergence testing. The same converged DFT configuration is used throughout each search to ensure consistent measurements across evaluation. We further extend the backend with machine-learned interatomic potentials (Batatia et al., 2022) through

Table 1: Details of materials, perturbation space, and target properties in the Struct-R1 benchmark. Each configuration is combined with uniaxial strain along x, y, or z. Subscripts denote the replaced site; H_i is an interstitial hydrogen and V is a vacancy.

Host	Configurations	Uniaxial strain	Target properties
Si	pristine, $\mathrm{C_{Si}}$, $\mathrm{B_{Si}}$, V_{Si}	x, y, z; $\pm 2.5\%$	Raman, energy, charge level, decay threshold
Ge	pristine, $\mathrm{Si_{Ge}}$, $\mathrm{P_{Ge}}$, V_{Ge}	x, y, z; $\pm 1.5\%$	Raman, energy, charge level
GaAs	pristine, substitutional defects	x, y, z; $\pm 1.5\%$	Raman, energy, Born charge
MgO	pristine, $\mathrm{Li_{Mg}}$, H_i, V_{O}	x, y, z; $\pm 2.5\%$	Raman, energy, charge level

ASE (Hjorth Larsen et al., 2017), supporting both direct MLIP prediction and MLIP pre-relaxation followed by DFT. All measurement tiers expose the same interface to the agent, which specifies a structure and target property and receives the measurement result and execution log.

For efficient benchmark, Struct-R1 caches previously visited DFT points, including their relaxed structures and computed properties, and reuses them when the same configuration is queried again. We additionally pre-compute dense DFT and MLIP scans over the perturbation spaces used in our benchmark, allowing different agents and search methods to directly query the same measurement results. The resulting benchmark contains over 1,000 cached DFT evaluations across the materials in Section 3.5, constructed using more than 8,000 CPU-core-hours of computation.

3.5 Benchmark Design

Material property search with hidden perturbations. We design the benchmark as an property-search problem with a verifiable ground truth. Each task starts from a host crystal and secretly applies a perturbation consisting of a material configuration (pristine or a point defect) and uniaxial strain. The agent observes only a target material property produced by this hidden structure and must recover a perturbation that reproduces the target. During search, it can propose arbitrary candidates and query the available measurement tiers under a fixed budget. This hidden-perturbation design prevents the agent from recalling from literature knowledge or known structure data. The target is generated from a randomly selected perturbation and is never exposed to the agent.

Materials and properties. Our benchmark covers four representative semiconductor and insulating material. As summarized in Table 1, we construct search tasks over five properties spanning different physical mechanisms: Raman peak, elastic energy, charge-transition level, three-phonon decay threshold, and Born effective charge. Together, these tasks require the agent to distinguish different structural configurations, reason about continuous perturbations, and adapt its search strategy.

Multi-fidelity search and evaluation. We evaluate whether the final structure reproduces the hidden target property under DFT within its numerical tolerance. We report both the success rate and the cost required to reach the solution. Each tier's costs are normalized by computation time.

4 Evaluation

4.1 Experimental Setup

Expensive DFT tier. We use Quantum ESPRESSO (Giannozzi et al., 2009) 7.5 with PBE and PseudoDojo (Van Setten et al., 2018) norm-conserving pseudopotentials, with plane-wave cutoffs converged per element and Fermi-Dirac smearing. Defects are modeled in 64-atom supercells. Neutral defect calculations are used for structural and vibrational properties, while charge-transition levels are obtained from the corresponding charged defect states. Born charges are computed with DFPT in 8-atom cells. $\mathrm{HfTe_5}$ is computed on the experimental lattice with PBE+SOC.

Cheap proxy tier. The low-fidelity tier uses MACE-MPA-0 (Batatia et al., 2022) through ASE (Hjorth Larsen et al., 2017) with fixed-cell ionic relaxation to 0.02 eV/Å. When MACE cannot directly provide the target property, it is used for pre-relaxation followed by the corresponding DFT property calculation. For the WBM experiments, we additionally use MACE-MPA-0, MACE-MP-0, and MEGNet predictions from Matbench Discovery.

Table 2: Search performance on the benchmark. Each property cell reports the success rate (%) / mean number of DFT calls. We evaluate 80 Raman-peak, 80 elastic-energy, 60 charge-level, 20 decay-threshold, and 20 Born-charge tasks, generated from materials in Section 3.5.

	Success rate (%) / DFT calls					Overall	
Method	Raman	Elastic	Charge	Decay	Born	Solved	DFT calls
Deterministic Algorithms							
Grid	9 / 9.8	45 / 11.8	87 / 5.8	40 / 8.2	60 / 10.1	44%	9.4
MLIP only	1 / 0.0	11 / 0.0	0 / 0.0	0 / 0.0	0 / 0.0	4%	0.0
MFBO (AR1)	70 / 6.8	52 / 8.0	58 / 8.5	55 / 8.2	75 / 7.6	61%	7.7
NARGP	26 / 10.6	66 / 7.3	58 / 8.5	40 / 9.1	75 / 7.6	51%	8.8
LLM-based (GPT-5.6 Terra)							
Single shot	0 / 0.0	3 / 0.0	13 / 0.0	3 / 0.0	1 / 0.0	4%	0.0
Plan & execute	55 / 9.7	58 / 8.1	45 / 9.3	0 / 12.0	9 / 6.1	46%	9.0
Struct-R1							
GPT-5.6 Terra	100 / 5.1	92 / 5.2	100 / 4.5	100 / 5.3	100 / 3.8	**98%**	4.9
GPT-5.2	79 / 9.3	71 / 8.8	97 / 12.4	85 / 8.7	100 / 12.8	83%	10.1
GPT-4o	15 / 6.5	34 / 6.8	92 / 10.9	35 / 3.7	95 / 11.3	46%	7.8
Claude Sonnet-5	92 / 5.1	81 / 3.9	100 / 5.1	80 / 4.8	100 / 5.0	90%	4.7
Claude Opus-5	88 / 3.8	91 / 4.0	97 / 3.8	90 / 4.2	100 / 3.9	92%	**3.9**

Table 3: Crystal-structure discovery across cheap tiers of different fidelity.

	Cheap-tier	Search algorithms			Struct-R1 (GPT-5.6 Terra)	
Proxy choice	r / MAE	SFBO	MLIP-Ranking	MFBO	No record	Evolution
MACE-MPA-0	0.93 / 0.025		**191** (**10.3**)	178 (11.5)	181 (10.8)	189 (10.5)
MACE-MP-0	0.85 / 0.053	61 (>20)	**164** (**11.4**)	156 (12.7)	158 (12.8)	163 (11.7)
MEGNet	0.69 / 0.133		108 (16.3)	113 (16.8)	122 (16.4)	**123** (16.3)

4.2 Overall Performance

Search efficiency across search tasks. Table 2 compares different search strategies for property inversion. Grid search uniformly scans the perturbation space, but its limited measurement budget provides insufficient coverage. MLIP-only search relies entirely on low-fidelity predictions, of which performance degrades when the MLIP is inaccurate near the target. MFBO and NARGP instead fit multi-fidelity surrogate models and adaptively select new measurements, achieving higher success rates, but they still require more DFTs and can be misled by an inaccurate cross-tier relation.

We also compare against LLM baselines without iterative feedback. Single-shot prediction solves only 4% of the tasks, while plan-and-execute reaches 46%. This shows that neither recalling solutions from the model's internal knowledge nor planning the search in advance is sufficient; the search must adapt to feedback obtained during exploration. In contrast, Struct-R1 closes the feedback loop and achieves the best performance: GPT-5.6 Terra solves 98% of the tasks with 4.9 DFT calls, while Claude Opus-5 reaches 92% with the lowest cost of 3.9 calls.

Evaluation with stable-structure discovery benchmark. We further evaluate Struct-R1 on crystal-structure discovery using the WBM test set of Matbench Discovery (Riebesell et al., 2025). We consider ten prototype families with 200–400 candidates each and repeat the experiment with three published models of different quality as the cheap tier. We compare Struct-R1 with MLIP ranking, which follows the Matbench Discovery protocol by sorting all candidates using cheap-tier predicted energy above the hull and spending DFT evaluations in that fixed order until the budget is exhausted. As shown in Table 3, direct ranking works well when the cheap tier is reliable but degrades as its predictions become less accurate. In contrast, Struct-R1 adapts its search from reference feedback and remains close to the strongest fixed strategy across all three tiers. Evolution further helps when the cross-tier relationship is transferable, allowing the agent to summarize recurring errors from pre-

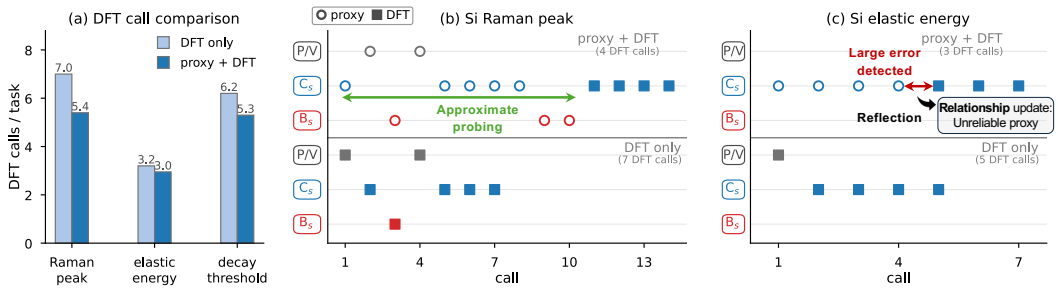

Figure 4: Search-loop analysis. (a) DFT cost across tasks on Silicon. (b) Proxy-guided probing with selective DFT verification. (c) Reflection detects an unreliable proxy and switches to DFT.

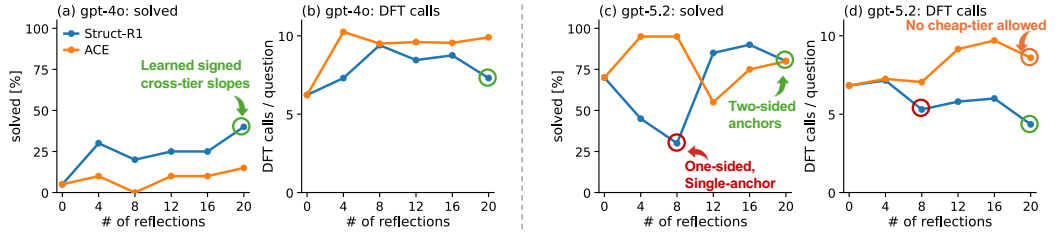

Figure 5: Evolution curve on success rate and DFT calls per task the Ge elastic-energy task.

vious families and reuse them in subsequent search; when such errors are family-specific, however, the learned record provides little benefit.

Moreover, ranking requires a full cheap-tier scan, whereas Struct-R1 queries it selectively. In our evaluation, it uses only 9.8 cheap calls per question, compared with 38.4 for MLIP-ranked. WBM does not report the computational cost of the cheap tier, so we cannot include it in the comparison, although this cost is non-negligible in practical materials search.

Search Loop Analysis. Figure 4(a) summarizes the DFT cost across search tasks with Silicon as the material. Compared with DFT-only search, combining low-fidelity probing with selective DFT verification reduces the number of DFT calls across Raman peak, elastic energy, and decay-threshold tasks. The reduction is larger when the proxy provides useful guidance such as in Raman peak search, but also remains possible even when the proxy is imperfect because Struct-R1 can adapt its measurement strategy during search. Figure 4(b) shows one of the representative behavior pattern, where the low-fidelity tier provides a useful approximation of the search trend. The agent therefore performs several inexpensive proxy measurements to probe the candidate region, and invokes DFT only after the promising range has been narrowed. For the Si Raman task, this approximate-probing pattern reduces the required DFT evaluations from 7 to 4. Figure 4(c) shows another behavior pattern, where the proxy becomes unreliable. After observing a large disagreement between the low- and high-fidelity measurements, the agent reflects on the mismatch and updates the corresponding tier relationship as unreliable. Subsequent exploration therefore only relies on DFT for verification. This avoids the search loop to repeatedly rely on the incorrect prediction.

Evolution Loop Analysis. We evaluate evolution on the challenging Ge elastic-energy task with GPT-4o and GPT-5.2, measuring performance every four reflections, as shown in Figure 5. We implement an ACE (Zhang et al., 2026)-like reflection on the complete trajectory, which can conflate cross-tier mismatch with errors in the search strategy. For example, with GPT-5.2, it eventually attributes failures to the cheap tier and evolves a rule that disables cheap measurements, increasing DFT usage while limiting search performance. In contrast, Struct-R1 shows model-specific improvements. For GPT-4o, where inaccurate cross-tier estimation is the main bottleneck, evolution gradually learns the correct signed slopes between cheap and DFT measurements, improving the success rate. For GPT-5.2, evolution first produces an overly aggressive one-sided, single-anchor strategy, but subsequent reflections identify insufficient exploration and repair it with two-sided

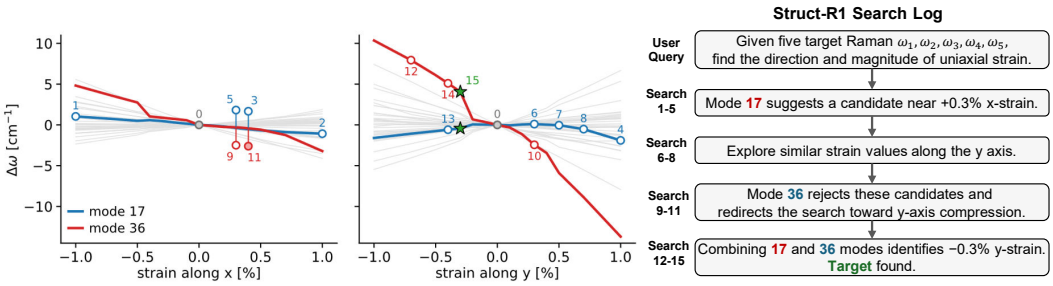

Figure 6: An example of searching residual strain in $MAPbI_3$ from Raman measurements.

anchors. These behaviors arise from the dual-objective reflection in Struct-R1, which separates per-step fidelity mismatch from trajectory-level search strategy and improves them independently.

4.3 Case Studies

Residual strain search in $MAPbI_3$. $MAPbI_3$ is a widely studied photovoltaic material whose residual strain can affect material stability and device performance (Zhao et al., 2017), making strain identification important for practical characterization. We construct this case study from a prior study of its mechanical properties and strain-dependent Raman response (Talit & Strubbe, 2024), using the complete DFPT sweep over three strain axes as the measurement oracle. From the 36 computed Raman modes, we expose a minimal set of five that are each non-linear along at least one axis and jointly distinguish candidate strains by at least 0.32 cm^{-1}. We further exclude cases solvable by linear extrapolation from the $\pm 0.4\%$ calibration window. Over 30 episodes on four resulting hidden strains, Struct-R1 solves 26 (87%) with 6.9 measurements on average, touching a median of only four strained structures compared with 43 in the published sweep.

Figure 6 illustrates how this efficiency arises. Mode 17 first suggests a candidate near $+0.3\%$ x-strain, but similar responses along y leave the direction ambiguous. Rather than continuing the same search, Struct-R1 introduces the more strain-sensitive mode 36, rejects the x branch, and redirects the search toward y compression, eventually identifying the target -0.3% strain. This adaptive use of observables is difficult for fixed search rules: a bisection baseline assumes a monotonic strain response and remains at 25% accuracy even as its budget increases from 6 to 30 calls, while exhaustive enumeration reaches 68% at the full budget. The agent instead uses measurements to first explore the search space and then resolve ambiguity only when needed.

$HfTe_5$ case study. We further evaluate Struct-R1 on the strain-driven topological transition of $HfTe_5$ using the reference data from Liu et al. (Liu et al., 2024). Given only a few reference points along the c axis, the agent updates the `relationship` record from a simple offset to a folded affine mapping between its PBE+SOC measurements and the V-shaped reference response. It then uses the `information` record to recognize that two reference points cannot determine the transition vertex and requests an additional point on the other branch. When testing whether this calibration transfers to the untested a axis, the agent proposes a targeted measurement and updates the relationship after the returned value disagrees with its prediction. For missing quantities, the agent first attempts in `<think>` to retrieve relevant evidence from the isostructural $ZrTe_5$ literature before requesting additional measurements. This case illustrates how the structured records guide what relation to maintain, what information is still missing, and when high-fidelity measurements are necessary, although broader use in real research settings remains to be further studied.

5 Conclusion

We present Struct-R1, an agentic framework for cost-aware materials structure search with multi-fidelity measurements. Struct-R1 combines physics-grounded exploration, adaptive tier selection, and reflection over search trajectories to improve search efficiency. Experiments show that Struct-R1 achieves higher success with fewer high-fidelity evaluations across diverse tasks. Case studies on $MAPbI_3$ and $HfTe_5$ further demonstrate its ability to work with real research data.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation. We have not used generative AI tools to propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results, and generate synthetic data sets, help develop theoretical models or conceptual frameworks. To formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to create or edit software code. We have reviewed all AI-assisted work. AI-edited code was inspected, tested, and used by the authors in the actual experimental workflow, while AI-edited manuscript text was manually checked and further refined by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nikhil Abhyankar, Sanchit Kabra, Saaketh Desai, and Chandan K Reddy. Llema: Evolutionary search with llms for multiobjective materials discovery. In *The Fourteenth International Conference on Learning Representations*, 2026.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, et al. Gepa: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 8479–8565, 2026.

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in neural information processing systems*, 35:11423–11436, 2022.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

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.

Xiangsen Chen, Ruilong Wu, Yanyan Lan, Ting Ma, and Yang Liu. Molevolve: Llm-guided evolutionary search for interpretable molecular optimization. *arXiv preprint arXiv:2603.24382*, 2026.

Stefano Curtarolo, Gus LW Hart, Marco Buongiorno Nardelli, Natalio Mingo, Stefano Sanvito, and Ohad Levy. The high-throughput highway to computational materials design. *Nature materials*, 12(3):191–201, 2013.

Clyde Fare, Peter Fenner, Matthew Benatan, Alessandro Varsi, and Edward O Pyzer-Knapp. A multi-fidelity machine learning approach to high throughput materials screening. *npj Computational Materials*, 8(1):257, 2022.

Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, et al. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *Journal of physics: Condensed matter*, 21(39):395502, 2009.

Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, et al. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017.

Zhengding Hu, Kuntal Talit, Zhen Wang, Haseeb Ahmad, Yichen Lin, Prabhleen Kaur, Christopher Lane, Elizabeth A Peterson, Zhiting Hu, Elizabeth A Nowadnick, et al. Tritondft: Automating dft with a multi-agent framework. *arXiv preprint arXiv:2603.03372*, 2026.

Shuyi Jia, Chao Zhang, and Victor Fung. Llmatdesign: Autonomous materials discovery with large language models. *arXiv preprint arXiv:2406.13163*, 2024.

Na Hyun Jo, Omar A Ashour, Zhixue Shu, Chris Jozwiak, Aaron Bostwick, Yang Wang, Eoghan Downey, Sae Hee Ryu, Kai Sun, Tai Kong, et al. Effects of strain and defects on the topological properties of hfte 5. *Physical Review B*, 109(23):235122, 2024.

Kirthevasan Kandasamy, Gautam Dasarathy, Junier B Oliva, Jeff Schneider, and Barnabás Póczos. Gaussian process bandit optimisation with multi-fidelity evaluations. *Advances in neural information processing systems*, 29, 2016.

Marc C Kennedy and Anthony O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

Chengzhu Li, Haichuan Tan, Wenyu Zhu, Bowen Gao, Jiqing Zheng, Ya-Qin Zhang, Wei-Ying Ma, and Yanyan Lan. Molopt-eval: Can frontier llms perform structure-based hit-to-lead optimization? In *The 2026 Workshop on Generative and Agentic AI for Biology*.

Ju Li, Zhiwei Shan, and Evan Ma. Elastic strain engineering for unprecedented materials properties. *Mrs Bulletin*, 39(2):108–114, 2014.

Jinyu Liu, Yinong Zhou, Sebastian Yepez Rodriguez, Matthew A Delmont, Robert A Welser, Triet Ho, Nicholas Sirica, Kaleb McClure, Paolo Vilmercati, Joseph W Ziller, et al. Controllable strain-driven topological phase transition and dominant surface-state transport in hfte5. *Nature Communications*, 15(1):332, 2024.

Turab Lookman, Prasanna V Balachandran, Dezhen Xue, and Ruihao Yuan. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials*, 5(1):21, 2019.

Kiran Mathew, Joseph H Montoya, Alireza Faghaninia, Shyam Dwarakanath, Muratahan Aykol, Hanmei Tang, Iek-heng Chu, Tess Smidt, Brandon Bocklund, Matthew Horton, et al. Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. *Computational Materials Science*, 139:140–152, 2017.

Artem R Oganov, Chris J Pickard, Qiang Zhu, and Richard J Needs. Structure prediction drives materials discovery. *Nature Reviews Materials*, 4(5):331–348, 2019.

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

Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, and Luca M Ghiringhelli. Sisso: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials*, 2(8):083802, 2018.

Paris Perdikaris, Maziar Raissi, Andreas Damianou, Neil D Lawrence, and George Em Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings. Mathematical, physical, and engineering sciences*, 473(2198):20160751, 2017.

Giovanni Pizzi, Andrea Cepellotti, Riccardo Sabatini, Nicola Marzari, and Boris Kozinsky. Aiida: automated interactive infrastructure and database for computational science. *Computational Materials Science*, 111:218–230, 2016.

Matthias Poloczek, Jialei Wang, and Peter Frazier. Multi-information source optimization. *Advances in neural information processing systems*, 30, 2017.

Hans J Queisser and Eugene E Haller. Defects in semiconductors: some fatal, some vital. *Science*, 281(5379):945–950, 1998.

Janosh Riebesell, Rhys EA Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A Lee, Anubhav Jain, and Kristin A Persson. A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence*, 7(6):836–847, 2025.

Kuntal Talit and David A Strubbe. Resolving discrepancies in calculations of mechanical properties of ch\textsubscript {3} nh\textsubscript {3} pbi\textsubscript {3} perovskites. *arXiv preprint arXiv:2412.17771*, 2024.

Michiel J Van Setten, Matteo Giantomassi, Eric Bousquet, Matthieu J Verstraete, Don R Hamann, Xavier Gonze, and G-M Rignanese. The pseudodojo: Training and grading a 85 element optimized norm-conserving pseudopotential table. *Computer Physics Communications*, 226:39–54, 2018.

Haorui Wang, Jeff Guo, Lingkai Kong, Rampi Ramprasad, Philippe Schwaller, Yuanqi Du, and Chao Zhang. Llm-augmented chemical synthesis and design decision programs. *arXiv preprint arXiv:2505.07027*, 2025a.

Haorui Wang, Marta Skreta, Cher-Tian Ser, Wenhao Gao, Lingkai Kong, Felix Strieth-Kalthoff, Chenru Duan, Yuchen Zhuang, Yue Yu, Yanqiao Zhu, et al. Efficient evolutionary search over chemical space with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 51694–51727, 2025b.

Ruheng Wang, Hang Zhang, Trieu Nguyen, Shasha Feng, Hao-Wei Pang, Xiang Yu, Li Xiao, and Peter Zhiping Zhang. Pepthink-r1: Llm for interpretable cyclic peptide optimization with cot sft and reinforcement learning. *arXiv preprint arXiv:2508.14765*, 2025c.

Ziqi Wang, Hongshuo Huang, Hancheng Zhao, Changwen Xu, Shang Zhu, Jan Janssen, and Venkatasubramanian Viswanathan. Dreams: Density functional theory based research engine for agentic materials simulation. *arXiv preprint arXiv:2507.14267*, 2025d.

Hongming Weng, Xi Dai, and Zhong Fang. Transition-metal pentatelluride zrte 5 and hfte 5: A paradigm for large-gap quantum spin hall insulators. *Physical review X*, 4(1):011002, 2014.

Fang Wu, Weihao Xuan, Heli Qi, Hanqun Cao, Heng-Jui Chang, Zeqi Zhou, Haokai Zhao, Ma Jian, Carl Ma, Yu-Chi Cheng, et al. Proteo-r1: Reasoning foundation models for de novo protein design. *arXiv preprint arXiv:2605.02937*, 2026.

Zeyu Xia, Jinzhe Ma, Congjie Zheng, Shufei Zhang, Yuqiang Li, Hang Su, P Hu, Changshui Zhang, Xingao Gong, Wanli Ouyang, et al. An agentic framework for autonomous materials computation. *arXiv preprint arXiv:2512.19458*, 2025.

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.

Dezhen Xue, Prasanna V Balachandran, John Hogden, James Theiler, Deqing Xue, and Turab Lookman. Accelerated search for materials with targeted properties by adaptive design. *Nature communications*, 7(1):11241, 2016.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, et al. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, volume 2026, pp. 86069–86100, 2026.

Jingjing Zhao, Yehao Deng, Haotong Wei, Xiaopeng Zheng, Zhenhua Yu, Yuchuan Shao, Jeffrey E Shield, and Jinsong Huang. Strained hybrid perovskite thin films and their impact on the intrinsic stability of perovskite solar cells. *Science advances*, 3(11):eaao5616, 2017.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17744–17761, 2025.

Qianyou Zhong, Dingyang Lv, Bowen Fang, Baole Wei, Xiangran Cheng, Hongwei Du, and Feng Yu. Matevolve: A synergistic symbolic–llm agent for multi-objective materials design.

Cheng Zhu, Xiuxiu Niu, Yuhao Fu, Nengxu Li, Chen Hu, Yihua Chen, Xin He, Guangren Na, Pengfei Liu, Huachao Zai, et al. Strain engineering in perovskite solar cells and its impacts on carrier dynamics. *Nature communications*, 10(1):815, 2019.

Alex Zunger. Inverse design in search of materials with target functionalities. *Nature Reviews Chemistry*, 2(4):0121, 2018.