Accelerated Inorganic Materials Design with Generative AI Agents

Izumi Takahara
The University of Tokyo
kougen@iis.u-tokyo.ac.jp

Teruyasu Mizoguchi
The University of Tokyo
teru@iis.u-tokyo.ac.jp

Bang Liu
University of Montreal & Mila
bang.liu@umontreal.ca

Abstract

Designing inorganic crystalline materials with tailored properties is critical to technological innovation, yet current generative computational methods often struggle to efficiently explore desired targets with sufficient interpretability. Here, we present MatAgent, a generative approach for inorganic materials discovery that harnesses the powerful reasoning capabilities of large language models (LLMs). By combining a diffusion-based generative model for crystal structure estimation with a predictive model for property evaluation, MatAgent uses iterative, feedback-driven guidance to steer material exploration precisely toward user-defined targets. Integrated with external cognitive tools—including short-term memory, long-term memory, the periodic table, and a comprehensive knowledge base—MatAgent emulates human expert reasoning to vastly expand the accessible compositional space. Our results demonstrate that MatAgent robustly directs exploration toward desired properties while consistently achieving high compositional validity, uniqueness, and material novelty. This framework thus provides a highly interpretable, practical, and versatile AI-driven solution to accelerate the discovery and design of next-generation inorganic materials.

1 Introduction

Developing materials with desired properties is a critical challenge, driving innovations across essential industrial fields such as catalysis, energy storage, and beyond[1–3]. Computational approaches, including density functional theory[4, 5] simulations and high-throughput virtual screenings, have been widely employed to identify candidate materials[6]. However, these traditional methods typically require extensive manual modeling by human experts or exhaustive enumeration of candidate materials, limiting their efficiency and scalability in exploring broader materials spaces. To overcome these limitations and facilitate more autonomous, scalable, and efficient materials discovery, there is an increasing demand for workflows capable of effectively exploring broader materials space to identify novel inorganic materials while reducing the reliance on human-expert intervention.

Recently, generative artificial intelligence has made significant progress, particularly in the fields of natural language processing and computer vision. These advances have also influenced computational materials design[7], where generative models such as variational autoencoders[8], generative adversarial networks[9], generative flow networks[10], diffusion models[11–13], and autoregressive models[14] are increasingly being applied, individually[15–26] or in combination[27–30], to generate and explore novel crystalline materials. This opens new avenues for effectively navigating and

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Mat: AI for Accelerated Materials Design.

exploring the vast materials design space utilizing generative artificial intelligence techniques[31, 32]. For example, MatterGen[26], a diffusion-based generative model, has successfully generated crystal structures by simultaneously generating lattice vectors, atomic coordinates, and elemental species, enabling the exploration and discovery of novel materials beyond conventional approaches. Similarly, large language models (LLMs), fine-tuned by Gruver *et al.*[19] or trained from scratch by Antunes *et al.*[20], have directly generated stable materials, highlighting the broad potential of generative AI approaches in accelerating materials discovery.

However, these generative approaches typically produce materials through a single-step generation process, potentially limiting their ability to precisely meet specified target properties. Furthermore, in many target-aware generation methods, materials are optimized within latent spaces, making it challenging to interpret the underlying reasoning behind the generated results. Recent advances in the foundational knowledge and reasoning capabilities of LLMs have enabled these models to perform complex and sophisticated tasks[33–35]. Leveraging these enhanced capabilities, multi-step refinement approaches have emerged as a promising direction. For instance, MatExpert[36] decomposed the material generation process into three distinct steps and demonstrated improved performance in generating target-oriented materials. Jia *et al.* proposed an LLM-driven iterative framework that refines materials via modifications, accompanied by human-interpretable reasoning provided in natural language[37]. Nevertheless, existing LLM-based methods primarily rely on the inherent knowledge embedded within the LLMs, thereby constraining their exploration to relatively limited compositional spaces or frequently requiring fine-tuning tailored specifically to each target-specific property or constraint, thus reducing their practical scalability. To achieve more effective materials discovery, it is essential to develop scalable frameworks capable of autonomously exploring broader materials spaces, while incorporating external knowledge and enabling interpretable, iterative refinement toward target properties.

In this study, we introduce MatAgent, a framework using a general-purpose LLM as its central generative engine for materials discovery. Enhanced with external tools that mimic human-expert reasoning processes and integrated with models for crystal structure estimation and property evaluation, MatAgent enables feedback-driven autonomous exploration of broad materials space. Our experiments targeting specific formation energies demonstrate MatAgent's effectiveness in proposing materials with desired properties while maintaining high compositional validity, uniqueness, and novelty. The framework's natural language integration enables intuitive constraint implementation for practical workflows. By providing explicit reasoning for each composition proposal, our framework ensures interpretability in materials design process, serving as an intelligent collaborator that enhances expert knowledge and accelerates inorganic materials discovery.

2 Results

2.1 MatAgent framework

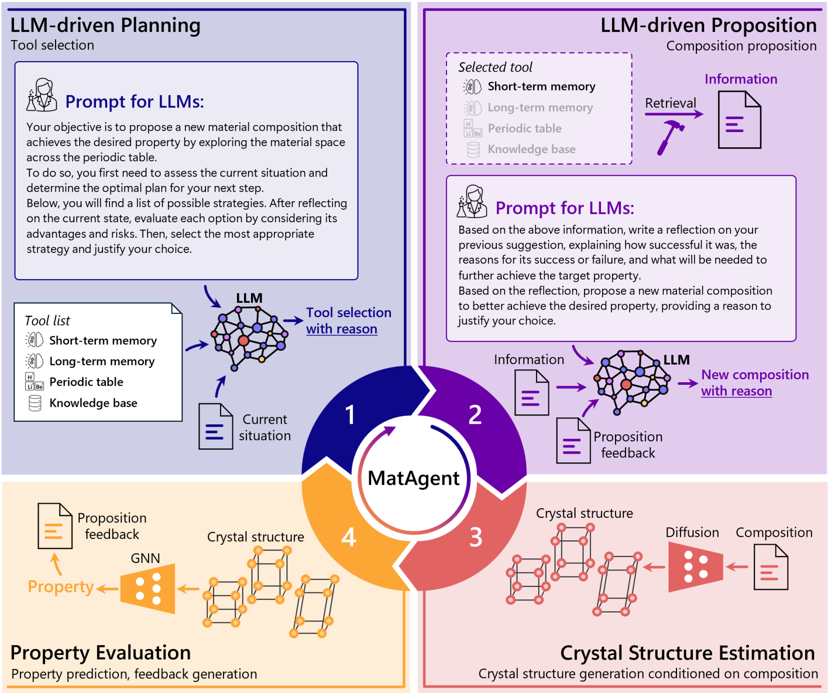
An overview of the proposed MatAgent framework is shown in Fig. 1. MatAgent is a feedback-driven iterative framework that involves four distinct stages: (1) LLM-driven Planning stage, (2) LLM-driven Proposition stage, (3) generation of three-dimensional crystal structures in Crystal Structure Estimation stage, and (4) evaluation of materials properties and the generation of feedback in the Property Evaluation stage.

LLM as agent and tool-use inspired by human-experts. Recent advancements in LLMs have driven remarkable improvements in their reasoning capabilities, enabling their use in complex problem-solving tasks[38–40]. Our framework leverages these capabilities for materials design through a two-stage process: (1) Planning and (2) Proposition, as shown in Fig. 1.

During the Planning stage, the LLM strategically selects one of four tools with explicit justification based on current situation. While previous methods have relied on reasoning using knowledge inherent to LLMs and its self-reflection[37], our approach aims to enhance materials exploration by LLMs and enable effective AI-driven materials design through the multifaceted integration of tools that mimic the reasoning processes of human experts. Specifically, we integrated four distinct tools—short-term memory, long-term memory, the periodic table, and the knowledge base—reflecting the human approach of leveraging short-term experience, long-term insights, fundamental knowledge such as the periodic table, and previously accumulated knowledge to facilitate iterative exploration and informed decision-making in materials-design.

The short-term memory recalls the compositions proposed in recent iterations and the corresponding feedback received. The intent of this is to help the LLM understand recent performance trends and revise potentially redundant composition proposals. The long-term memory retrieves not only previously successful compositions but also the associated reasoning processes used by the LLM, thus providing insights into why certain compositions led to favorable outcomes. The periodic table provides elements related to those used in the previous composition, specifically elements within the same group, based on the expectation that such substitutions can fine-tune material properties while preserving compositional plausibility. Lastly, the knowledge base is a compiled database that records how material properties change when transitioning from one composition to another.

In the Proposition stage, relevant information is retrieved based on the tool selected during the Planning stage. Combining this retrieved information with the composition proposed in the previous iteration and the corresponding feedback, the LLM then generates a new composition proposal accompanied by explicit reasoning, providing interpretable insights into the underlying design choices. By strategically utilizing these tools, the framework moves beyond the intrinsic knowledge of the LLM, enhancing its reasoning capabilities and broadening exploration within the materials design space. The complete prompt templates used for the Planning and Proposition stage corresponding to each tool are shown in Appendix A.



Your objective is to propose a new material composition that achieves the desired property by exploring the material space across the periodic table.
To do so, you first need to assess the current situation and determine the optimal plan for your next step.
Below, you will find a list of possible strategies. After reflecting on the current state, evaluate each option by considering its advantages and risks. Then, select the most appropriate strategy and justify your choice.

Based on the above information, write a reflection on your previous suggestion, explaining how successful it was, the reasons for its success or failure, and what will be needed to further achieve the target property.
Based on the reflection, propose a new material composition to better achieve the desired property, providing a reason to justify your choice.

Figure 1: **The overview of the proposed materials design framework.** MatAgent is a feedback-driven iterative framework that involves four distinct stages: (1) LLM-driven Planning: where the LLM analyzes current situation and selects an appropriate tool with explicit reasoning; (2) LLM-driven Proposition: where the LLM proposes new material compositions with explicit reasoning based on retrieved information and previous feedback; (3) Crystal Structure Estimation: where a diffusion model transforms proposed compositions into three-dimensional crystal structures; and (4) Property Evaluation: where a GNN-based model evaluates physical properties and generates feedback for further refinement. During the Planning stage, the LLM selects the appropriate tools (short-term memory, long-term memory, periodic table, and knowledge base), providing the natural-language justifications. In the Proposition stage, the LLM proposes revised compositions by reasoning based on previous feedback and insights obtained from these tools. The proposed compositions are then processed in the Crystal Structure Estimation stage, where three-dimensional crystal structures are generated from the proposed composition. These structures are evaluated and the corresponding feedback is generated in the Property Evaluation stage, and the feedback is iteratively provided to the LLM for further refinement.

Crystal Structure Estimation. Although the LLM proposes candidate materials as discrete compositions, accurately evaluating the corresponding material properties typically requires knowledge of three-dimensional crystal structures, including unit cell geometries and atomic coordinates. To address this requirement, Crystal Structure Estimation stage plays a crucial role in the proposed framework by estimating three-dimensional crystal structures corresponding to the compositions proposed by the LLM.

In this study, we employed a diffusion model[12, 13, 41] to estimate crystal structures from the proposed composition. Diffusion models have successfully been applied to generative tasks within image and language generation fields, and their effectiveness has also been demonstrated in crystal structure generation[26, 41, 18]. We constructed the MP-60 dataset by collecting stable crystal structure data with unit cells containing up to 60 atoms from the Materials Project[42] database. Using this dataset, we trained a conditional crystal structure generation model, which generates multiple candidate structures for each given reduced composition with varying numbers of formula units per unit cell (Z). This approach is based on the idea that simultaneously generating and evaluating structures with different formula units for the same composition helps identify the most stable unit cell configuration. The structure estimator is not limited to diffusion models but can also utilize other crystal structure prediction methods, enabling it to directly leverage advancements in both generative models and crystal structure prediction technologies.

Property Evaluation. The Property Evaluation stage plays a key role by quantitatively assessing and providing feedback on the material compositions proposed by the LLM. Specifically, it evaluates the physical properties of three-dimensional crystal structures generated in the Crystal Structure Estimation stage. In this study, we constructed a property predictor based on graph neural networks (GNNs), trained on the MP-60 dataset, specifically for predicting the formation energy per atom. The formation energy of all candidate structures is evaluated, and the structure with the lowest formation energy per atom is considered the most stable and selected. Feedback derived from the formation energy predictions is subsequently returned to the LLM, enabling further refinement and improvement of the proposed compositions. A template for generating the feedback is provided in Appendix A.3.

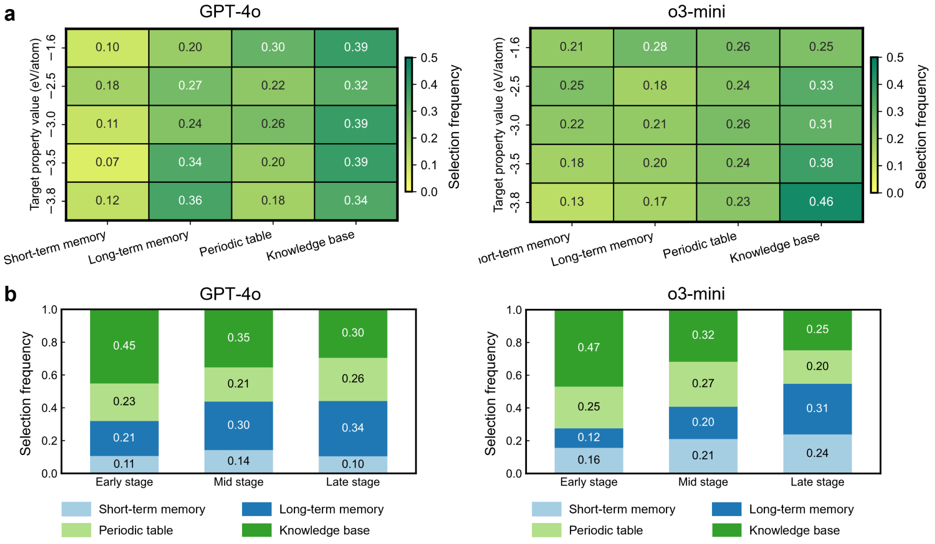


Figure 2: **Comparison of tool usage patterns during the Planning stage by the LLM. a** Tool selection frequency with which each tool (short-term memory, long-term memory, periodic table, and knowledge base) was selected by the LLM for each target value in the Planning stage to guide subsequent material composition proposals. **b** Tool selection frequency evolved over iterations, categorized into Early (iteration 1–5), Mid (iteration 6–10), and Late (iterations 11-15) stages.

2.2 Iterative materials generation guided by target property

Our framework iteratively refines materials through feedback cycles toward specific target properties. We selected *formation energy per atom* as the target property, defining target values based on quantiles from the MP-60 dataset: -3.8 eV/atom (1.0% quantile), -3.5 eV/atom (2.5% quantile), -3.0 eV/atom (10% quantile), -2.5 eV/atom (20% quantile), and -1.6 eV/atom (40% quantile). The distribution of formation energy per atom for the MP-60 dataset is provided in Appendix B. We employed OpenAI's GPT-4o and o3-mini as LLM backbones, conducting 20 independent runs of 15 iterations each, starting from randomly sampled compositions. GPT-4o is a general-purpose LLM developed by OpenAI, applicable to various types of tasks, while o3-mini is a specialized model optimized for focused reasoning applications with strength in scientific contexts.

Tool selection Patterns. We first analyzed how each LLM selected tools during the Planning stage. Fig. 2 summarizes tool selection frequency, with Fig. 2 **a** showing the tool selection frequency for each target value and Fig. 2 **b** illustrating the time evolution of tool selection. From Fig. 2 **a**, GPT-4o and o3-mini showed distinct patterns depending on target formation energy. GPT-4o frequently selected the knowledge base and long-term memory when targeting lower formation energy values (e.g., -3.8 eV/atom), whereas at higher values (e.g., -1.6 eV/atom), it preferred using the periodic table alongside the knowledge base. In contrast, o3-mini heavily relied on the knowledge base when targeting -3.8 eV/atom but exhibited more uniform tool selection for -1.6 eV/atom.

Examining tool-selection evolution over iterations, GPT-4o initially showed a strong preference for the knowledge base during early stages (iterations 1–5), but gradually reduced its reliance in later stages (iterations 11–15), increasingly utilizing long-term memory to leverage previous experiences. o3-mini similarly shifted toward increased use of long-term memory in later stages, accompanied by greater reliance on short-term memory. This suggests that both GPT-4o and o3-mini exhibit contextually adaptive behavior in their tool selection strategies, dynamically modifying their choices in response to the evolving task requirements and previously observed outcomes.

Evaluation of generated materials—success rate comparison. To explore the capability of the proposed framework in proposing materials, success rates were calculated based on these proposals as shown in Fig. 3, where a "successful" run is defined as achieving at least one composition with a predicted property value within ±0.25 eV/atom of the target. We conducted comparative experiments without tool assistance (GPT-4o w/o tools and o3-mini w/o tools) and also included results from MatterGen, a diffusion-based generative model fine-tuned on the MP-20 dataset using classifier-free (CF) guidance[43] for guided generation. For consistent evaluation, both MatAgent and MatterGen used the same property predictor without post-processing steps like structural relaxation. It should be noted that while MatterGen generates materials in a single step, MatAgent's iterative approach incurs increasing computational costs as iterations progress. All MatAgent experiments were initialized with compositions randomly sampled from the MP-20 dataset.

From Fig. 3, success rates were initially low for the first proposed compositions but steadily increased with iterations, demonstrating the effectiveness of our iterative refinement approach in navigating materials generation towards target property. MatterGen typically achieved higher success rates at early iterations compared to both LLMs, indicating effectiveness in rapidly generating materials close to target formation energies. GPT-4o w/ tools achieved higher success rates compared to GPT-4o w/o tools, especially when targeting formation energies of -3.8 eV/atom and -2.5 eV/atom. However, for the most challenging target (-3.8 eV/atom), both GPT-4o configurations exhibited relatively low success rates. Even for such a case, the success rate can be further enhanced by integrating knowledge-informed initialization, as detailed in Appendix C. The analysis from Fig. 3 revealed that o3-mini consistently demonstrated remarkably high success rates approaching 1.0, even when targeting the most challenging formation energy value of -3.8 eV/atom.

Evaluation of generated materials—compositional metrics. Although the high success rates achieved by GPT-4o and o3-mini indicate their capability to generate material compositions close to target properties, these success rates alone may not fully reflect the practical usefulness of the framework in materials exploration. In practice, compositions that are chemically invalid or already well-known typically have limited usefulness. Therefore, the compositional validity, uniqueness, novelty, and overall V.U.N. score of the generated materials compositions when targeting -3.8 eV/atom were evaluated (Fig. 4 **a**). Here, validity refers to the proportion of compositions assessed valid according to SMACT[44]; uniqueness denotes the proportion of unique compositions among candidate compositions; novelty indicates the proportion of compositions not present in the MP-20

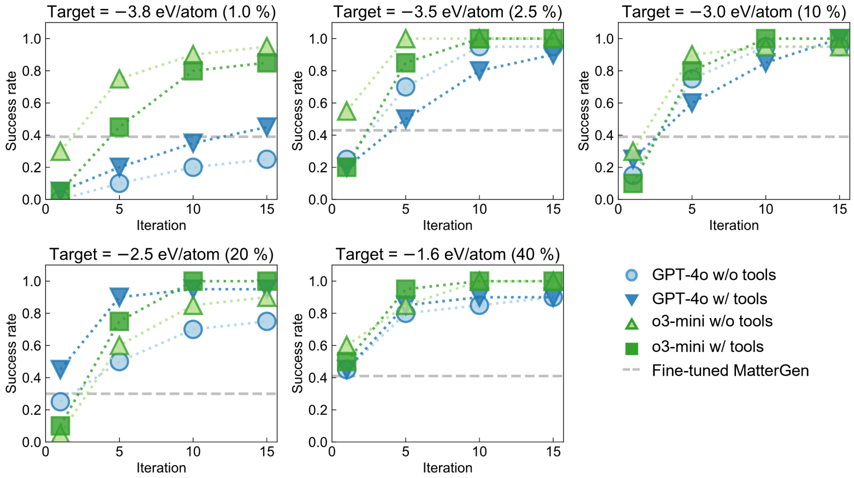


Figure 3: **Performance evaluation of the proposed framework on the task of targeted materials design.** Success rate progression over iterations for different settings of the MatAgent framework using GPT-4o and o3-mini as LLM backbones. Results are shown under conditions with and without tool-assisted Planning and Proposition, referred to as GPT-4o w/ tools, GPT-4o w/o tools, o3-mini w/ tools, and o3-mini w/o tools, respectively. Reference success rates from MatterGen[26] fine-tuned on formation energy per atom using the MP-20 dataset are also included.

training set; and the V.U.N. score represents the proportion of compositions that are simultaneously valid, unique, and novel. Here, while MatterGen was fine-tuned on the MP-20 dataset, the LLMs in MatAgent were not exclusively trained on this dataset and their broader pretraining may influence the novelty evaluation, thus a direct comparison of novelty scores between MatAgent and MatterGen may not be entirely fair. For MatAgent, all metrics were evaluated based on the structures identified as closest to the target property value within each independent run, while all 256 compositions generated through CF-guided generation were included in the assessment for MatterGen.

Clear differences emerged between w/ tools and w/o tools scenarios. With GPT-4o, despite high compositional validity without tools, uniqueness and novelty scores remained low. Adding tool-assisted Planning and Proposition significantly enhanced these metrics, raising the V.U.N. score from 0.20 to 0.60. Similarly, o3-mini w/o tools maintained high compositional validity but showed zero novelty, resulting in a V.U.N. score of 0.0. In contrast, o3-mini w/ tools significantly boosted uniqueness and novelty, achieving a V.U.N. score of 0.55, comparable to MatterGen.

To quantitatively evaluate the tendency of generated compositions, the elemental distributions in the proposed compositions were analyzed, as presented in Fig. 4 **b**. Elemental distribution analysis revealed that incorporating tool-assisted Planning and Proposition significantly expanded compositional space diversity. GPT-4o w/o tools primarily proposed materials dominated by Ti and O, while with tools, it produced more diverse compositions involving F, Ca, and Al. Similarly, o3-mini w/o tools suggested compositions with limited elements like Hf and O, where approximately 60% of suggested compositions were HfO_2, indicating heavy reliance on prior knowledge rather than effective exploration. In contrast, with tool-assisted Planning and Proposition, it introduced a broader range of elements including F, Th, and Y, enhancing compositional diversity and novelty.

Without tool assistance, LLMs frequently proposed well-known compositions, achieving high success rates but limited novelty. Incorporating tool assistance significantly improved uniqueness and novelty while maintaining high success rates, highlighting the advantages of integrating external tools into LLM-driven reasoning. The contribution of each tool was further analyzed in Appendix E.

Interpretability in materials design processes. In previous sections, we have demonstrated that the proposed MatAgent exhibits high success rates in proposing materials that meet target property values, while also showing the capability to navigate diverse compositional spaces. In addition to these strengths, a key advantage of our approach compared to many previous generative methods

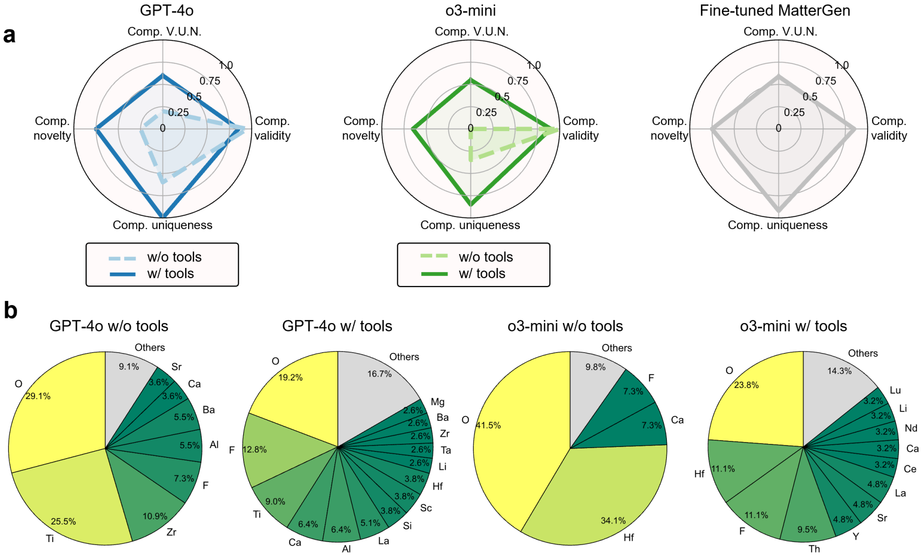


Figure 4: **Compositional metrics on proposed compositions. a** Comparison of compositional validity, uniqueness, novelty, and the combined V.U.N. score across five different generation approaches. **b** Pie charts showing elemental proportions of compositions closest to the target, sorted by frequency for each approach. Elements comprising less than 2.5% are grouped together in the gray region for GPT-4o w/o tools, GPT-4o w/toos, o3-mini w/o tools, and o3-mini w/ tools, respectively. The complete distribution including all elements is available in Appendix D.

is the incorporation of LLM-driven reasoning processes into the materials design process, enabling interpretation through natural language.

In the Planning stage, the LLM strategically selects tools with accompanying reasoning based on the current situation. For example, o3-mini most frequently selected the knowledge base, providing justifications such as: "Leveraging an external database can provide broader insight and more data points on compounds with low formation energies, guiding us towards better substitutions." Additionally, reasoning such as "Continuing with the external database appears most promising as it provided the largest improvement, suggesting further fine-tuning may reach -3.8 eV/atom." demonstrates adaptive behavior in response to previous outcomes. These explanations reveal how the LLM dynamically adjusts it strategy based on the performance, enabling the interpretation of the decision-making process by the LLM.

During the Proposition stage, composition proposals are made with accompanying reasoning. For example, when targeting a formation energy of -3.8 eV/atom with o3-mini, if the previous attempt resulted in a value lower than the target, the LLM might reason: "Replacing F with O in $CeThO_3F$ to create $CeThO_4$ should weaken the ionic bonds slightly, increasing the formation energy per atom closer to the -3.8 eV/atom target". Conversely, when the previous attempt yielded a value higher than the target, it might suggest: "Replacing Ba with Sr and Fe with Hf should achieve stronger ionic bonds, as seen in Sr_2HfO_4, further lowering formation energy toward -3.8 eV/atom." These natural language explanations make MatAgent's decision-making transparent and interpretable. Importantly, the reasoning it provides is often insightful even for domain experts, helping them uncover new materials design principles. Therefore, MatAgent functions not merely as a generator of candidate materials, but as an intelligent collaborator that enhances expert knowledge by expressing implicit design logic in human-readable form. Examples of such reasoning in the Planning and Proposition stages are provided in Appendix F.

2.3 Constrained generation with natural language: Application to industrial scenarios

While the proposed framework is promising for fundamental research, industrial applications of materials design often require navigating practical constraints imposed by environmental regulations, resource availability, and manufacturing requirements. Conventional generative approaches have typically struggled with incorporating such qualitative constraints, being limited to numeric or categorical conditions. MatAgent, however, leverages the natural language understanding capabilities of LLMs to seamlessly integrate diverse constraints expressed in plain language. To evaluate this industrial applicability, we examined the performance under several practical constraints:

1. Exclusion of environmentally damaging elements: Materials generation was performed while explicitly excluding Pb, Hg, Cd, and Cr. The constraint was communicated via the prompt: “Here, do not include Pb, Hg, Cd, Cr in any proposed compositions.”
2. Exclusion of actinide elements: Generation was carried out excluding actinide elements, with the prompt: “Here, do not include the actinides in any proposed compositions.”
3. Restriction to non-metal elements: Generation was constrained to compositions containing only non-metal elements, with the prompt: “Here, only propose compositions consisting of non-metal elements.”

These constraints effectively altered elemental distributions in generated compositions. Constraint 1 completely eliminated targeted elements (96 % → 100 % exclusion), Constraint 2 reduced actinide presence (3% → 1 %), and Constraint 3 dramatically increased non-metal-only compositions (0 % → 83 %). The target formation energy was set to -3.8 eV/atom for Constraints 1–2 and -2.5 eV/atom for Constraint 3. The detailed elemental distributions corresponding to each constraint condition are provided in Appendix G. By enabling natural language constraints, MatAgent bridges the gap between theoretical AI frameworks and real-world industrial applications.

3 Discussion

In this work, we present MatAgent, an LLM-driven generative framework for inorganic materials design, developed to iteratively propose and refine materials toward specific target properties. MatAgent leverages explicit, interpretable reasoning capabilities of the LLM to propose promising material compositions. Inspired by the reasoning process of human experts, MatAgent integrates external tools to extend beyond the inherent knowledge of the LLM and facilitate exploration across a broader materials space. Compositions proposed by the LLM are subsequently evaluated by two complementary stages: the Crystal Structure Estimation stage employing a diffusion model, and the Property Evaluation stage employing a GNN-based prediction model. These stages enable the provision of precise feedback, guiding the LLM to iteratively refine and improve the generated compositions, effectively steering the exploration toward materials with user-specified target properties.

Our framework demonstrated high effectiveness in guiding inorganic materials generation toward specific target properties with controlled formation energies. Its key strength is interpretability through natural language reasoning, making the design process transparent. While LLM-only compositions showed limited diversity, integrating external tools that mimic expert reasoning significantly enhanced exploration capabilities while maintaining validity. The transparent reasoning process not only produces better material candidates but also accelerates human understanding by providing insights into fundamental materials design principles and property-composition relationships. Our experiments further demonstrated the framework’s ability to incorporate qualitative constraints in natural language, addressing practical industrial requirements.

One limitation of the current study is that the proposed framework focuses exclusively on guiding materials generation toward a single target property–formation energy. In practical materials discovery tasks, it is often necessary to simultaneously optimize multiple materials properties, necessitating more sophisticated property evaluators capable of accurately predicting diverse properties and more complex feedback mechanisms. Therefore, future work should focus on extending the framework to effectively guide the LLM toward practical multi-objective materials design. Additionally, the current framework does not evaluate the correctness of reasoning provided by the LLM. Extending the framework to incorporate a human-in-the-loop approach, wherein human experts provide feedback to ensure accurate reasoning, would be important for developing a more effective framework.

A promising direction for future research is to integrate synthesizability considerations into the MatAgent framework. By leveraging the text-based reasoning capabilities of LLMs, extensive knowledge derived from textbooks, scientific literature, and other comprehensive sources can be effectively incorporated. This integration would allow the LLM-based framework to autonomously assess and reason about the practical synthesizability of proposed materials. Developing such a framework that seamlessly combines extensive synthesis-related knowledge with efficient exploration of materials spaces represents a significant advancement toward practical autonomous materials discovery systems.

4 Methods

4.1 Tools

Short-term memory retrieval. In short-term memory retrieval, information regarding recent composition proposals, the reasoning behind these proposals, and their corresponding feedback are retrieved. In this study, we retrieve information about the three most recent proposals.

Long-term memory retrieval. In long-term memory retrieval, information is retrieved regarding past proposals that successfully generated compositions close to the target property, including the reasoning behind these proposals. In this study, information about the top three such proposals is retrieved.

Periodic table retrieval. In periodic table retrieval, elements belonging to the same group as each element included in the previously proposed composition are retrieved and listed.

Knowledge base retrieval. In knowledge base retrieval, information is retrieved from a pre-constructed knowledge base. To construct this knowledge base, 10,000 random composition pairs were generated from the MP-20 dataset, and GPT-4o was used to analyze and explain why the material properties change when one composition (source composition) transitions into another (target composition). The prompt template used for generating the explanation is provided in supplementary material A.5. The knowledge base consists of 10,000 data entries, each containing a pair of compositions, their corresponding properties, and GPT-4o-generated explanations. During retrieval, the previously proposed composition is first used as a query to identify the five most similar source compositions from the knowledge base based on compositional similarity. The compositional similarity is evaluated by vectorizing compositions using the CBFV[45] package. These five retrieved candidate entries are then ranked by the LLM, and the top three entries judged by the LLM to be most relevant and useful for the current task are ultimately selected.

4.2 Dataset

To curate the MP-60 dataset for training models used in Crystal Structure Estimation stage and Property Evaluation stage, we collected crystal structure data from the Materials Project[42] database. Structures were restricted to those containing up to 60 atoms per unit cell, with an energy above hull of 0.02 eV/atom or less. Additionally, we excluded structures consisting exclusively of gaseous elements and those with lattice vectors exceeding 20 Å. After applying these criteria, we obtained a final dataset comprising 48,323 structures, which was subsequently divided into training, validation, and test sets using a 6:2:2 ratio.

4.3 Crystal Structure Estimation

The Crystal Structure Estimation stage is responsible for generating three-dimensional crystal structures given a specific material composition proposed by LLMs. As the backbone model for this component, we employ DiffCSP[41], a diffusion-based crystal structure prediction model, trained on the MP-60 dataset to conditionally generate structures based on the provided compositions. For each given reduced composition, we simultaneously generate five crystal structures for formula units ranging from 1 to 4. Additionally, to ensure that the total number of atoms in each crystal structure did not exceed 34, we adjusted the maximum formula units accordingly.

4.4 Property Evaluation

The Property Evaluation stage assesses the compositions and crystal structures proposed by the LLM, providing feedback that guides subsequent composition proposals. Specifically, it employs iComFormer[46], a GNN-based transformer model trained on formation energy per atom, as its backbone for predicting material properties. All crystal structures generated in the Crystal Structure Estimation stage are evaluated, and the structure with the lowest predicted formation energy per atom is considered the most stable and selected accordingly. When the LLM submits a composition with formatting problems, the system returns specific feedback that instructs the LLM to correct these issues in its next attempts. Detailed descriptions of the feedback format are provided in supplementary material A.3.

Code availability

The source code and the dataset used in this study are publicly available on GitHub (https://github.com/izumitkhr/matagent).

Acknowledgement

This work was supported by JST ACT-X (Grant Number JPMJAX24DB) and JST BOOST (Grant Number JPMJBS2418), Japan. I. T. was supported by the MERIT-WINGS, The University of Tokyo.

References

[1] Chi Chen, Yunxing Zuo, Weike Ye, Xiangguo Li, Zhi Deng, and Shyue Ping Ong. A critical review of machine learning of energy materials. *Advanced Energy Materials*, 10(8):1903242, 2020.

[2] Takashi Toyao, Zen Maeno, Satoru Takakusagi, Takashi Kamachi, Ichigaku Takigawa, and Ken-ichi Shimizu. Machine learning for catalysis informatics: Recent applications and prospects. *ACS Catalysis*, 10(3):2260–2297, 2 2020.

[3] Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 12 2023.

[4] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[5] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, 1965.

[6] Juhwan Noh, Geun Ho Gu, Sungwon Kim, and Yousung Jung. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chem. Sci.*, 11:4871–4881, 2020.

[7] Hyunsoo Park, Zhenzhu Li, and Aron Walsh. Has generative artificial intelligence solved inverse materials design? *Matter*, 7(7):2355–2367, 2024.

[8] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, 2014.

[10] Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. In *Advances in neural information processing systems*, 2021.

[11] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[13] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[15] Zekun Ren, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G. Aberle, Shijing Sun, Xiaonan Wang, Yi Liu, Qianxiao Li, Senthilnath Jayavelu, Kedar Hippalgaonkar, Yousung Jung, and Tonio Buonassisi. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 5(1):314–335, 1 2022.

[16] Sungwon Kim, Juhwan Noh, Geun Ho Gu, Alan Aspuru-Guzik, and Yousung Jung. Generative adversarial networks for crystal structure prediction. *ACS Central Science*, 6(8):1412–1420, 8 2020.

[17] Mistal, Alex Hernández-García, Alexandra Volokhova, Alexandre AGM Duval, Yoshua Bengio, Divya Sharma, Pierre Luc Carrier, Michał Koziarski, and Victor Schmidt. Crystal-gfn: sampling materials with desirable properties and constraints. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023.

[18] Sherry Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable diffusion for materials generation. In *International Conference on Learning Representations*, 2024.

[19] Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C. Lawrence Zitnick, and Zachary Ward Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *International Conference on Learning Representations*, 2024.

[20] Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):10570, 12 2024.

[21] Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M Wood. FlowMM: Generating materials with riemannian flow matching. In *Internatioinal conference on machine learning*, 2024.

[22] Junkil Park, Youhan Lee, and Jihan Kim. Multi-modal conditional diffusion model using signed distance functions for metal-organic frameworks generation. *Nature Communications*, 16(1):34, 1 2025.

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

[24] Yong Zhao, Edirisuriya M. Dilanga Siriwardane, Zhenyao Wu, Nihang Fu, Mohammed Al-Fahdi, Ming Hu, and Jianjun Hu. Physics guided deep learning for generative design of crystal materials with symmetry constraints. *npj Computational Materials*, 9(1):38, 3 2023.

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

[26] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, mar 2025.

[27] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations*, 2022.

[28] Youzhi Luo, Chengkai Liu, and Shuiwang Ji. Towards symmetry-aware generation of periodic materials. In *Advances in Neural Information Processing Systems*, 2023.

[29] Anuroop Sriram, Benjamin Kurt Miller, Ricky T. Q. Chen, and Brandon M. Wood. FlowLLM: Flow matching for material generation with large language models as base distributions. In *Advances in Neural Information Processing Systems*, 2024.

[30] Xiaoshan Luo, Zhenyu Wang, Pengyue Gao, Jian Lv, Yanchao Wang, Changfeng Chen, and Yanming Ma. Deep learning generative model for crystal structure prediction. *npj Computational Materials*, 10(1):254, 11 2024.

[31] Hyunsoo Park, Anthony Onwuli, and Aron Walsh. Exploration of crystal chemical space using text-guided generative artificial intelligence. *ChemRxiv*, 11 2024.

[32] Sherry Yang, Simon Batzner, Ruiqi Gao, Muratahan Aykol, Alexander L. Gaunt, Brendan McMorrow, Danilo J. Rezende, Dale Schuurmans, Igor Mordatch, and Ekin D. Cubuk. Generative hierarchical materials search. In *Advances in Neural Information Processing Systems*, 2024.

[33] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

[34] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

[35] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

[36] Qianggang Ding, Santiago Miret, and Bang Liu. Matexpert: Decomposing materials discovery by mimicking human experts. In *International Conference on Learning Representations*, 2025.

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

[38] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 12 2023.

[39] 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*, 2024.

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

[41] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In *Advances in Neural Information Processing Systems*, 2023.

[42] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 07 2013.

[43] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

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

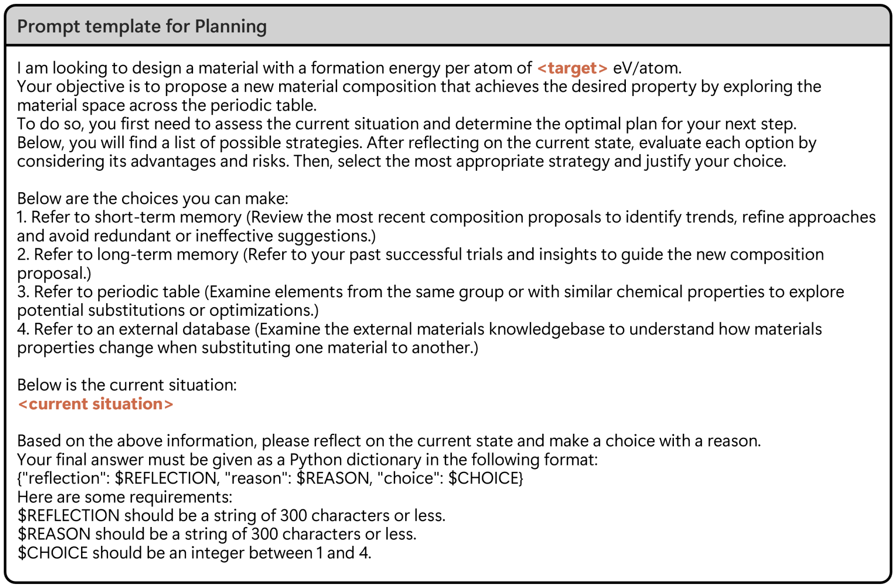
[45] Cbfv. `https://github.com/Kaaiian/CBFV`.

[46] Keqiang Yan, Cong Fu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Complete and efficient graph transformers for crystal material property prediction. In *International conference on Learning Representations*, 2024.

A Prompt templates

A.1 Prompt template for the Planning stage

Prompt template used during the Planning stage is shown in Fig. 5.



Prompt template for Planning

I am looking to design a material with a formation energy per atom of <target> eV/atom.
Your objective is to propose a new material composition that achieves the desired property by exploring the material space across the periodic table.
To do so, you first need to assess the current situation and determine the optimal plan for your next step.
Below, you will find a list of possible strategies. After reflecting on the current state, evaluate each option by considering its advantages and risks. Then, select the most appropriate strategy and justify your choice.

Below are the choices you can make:
1. Refer to short-term memory (Review the most recent composition proposals to identify trends, refine approaches and avoid redundant or ineffective suggestions.)
2. Refer to long-term memory (Refer to your past successful trials and insights to guide the new composition proposal.)
3. Refer to periodic table (Examine elements from the same group or with similar chemical properties to explore potential substitutions or optimizations.)
4. Refer to an external database (Examine the external materials knowledgebase to understand how materials properties change when substituting one material to another.)

Below is the current situation:
<current situation>

Based on the above information, please reflect on the current state and make a choice with a reason.
Your final answer must be given as a Python dictionary in the following format:
{"reflection": $REFLECTION, "reason": $REASON, "choice": $CHOICE}
Here are some requirements:
$REFLECTION should be a string of 300 characters or less.
$REASON should be a string of 300 characters or less.
$CHOICE should be an integer between 1 and 4.

Figure 5: **Prompt template for Planning stage. <target>** and **<current situation>** are the variables. Here, **<current situation>** includes historical information regarding previously selected tools and how the material properties changed as a result of those selections.

A.2 Prompt templates for the Proposition stage

Different prompt templates are used in the Proposition stage depending on which tool was selected during the Planning stage. The prompt template used when short-term memory is selected is shown in Fig. 6. Similarly, the prompt templates used when long-term memory, periodic table, and materials knowledge base are selected are shown in Fig. 7, 8, and 9, respectively.

A.3 Prompt template for feedback generation

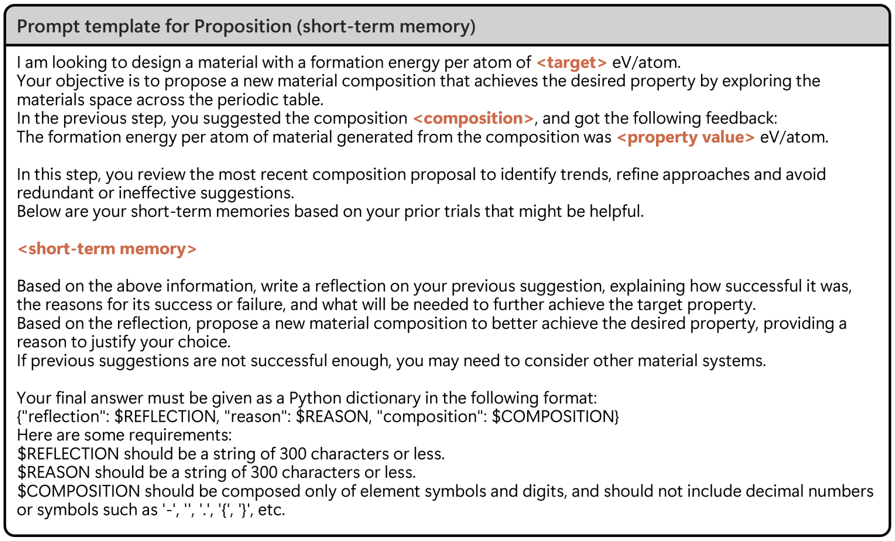
A prompt template used in the Property Evaluation stage to generate feedback for the LLM is shown in Fig. 10. This template is applied when the proposed composition is provided in a valid format. If the composition is submitted in an unreadable or incorrect format—for example, if it includes invalid element symbols or fractional subscripts—feedback instructing the LLM to resubmit the proposal using the correct format is provided.

A.4 Prompt template for comparative experiments

Fig. 11 shows the prompt template used for having the LLM propose new compositions without tool-assisted Planning and Proposition. In this case, the LLM receives feedback on the previously proposed composition and, based on that feedback, returns a revised composition proposal.

A.5 Prompt template for observation generation

Fig. 12 shows the prompt template used for generating observation when one composition transitions to another during the construction of the knowledge base.



Prompt template for Proposition (short-term memory)

I am looking to design a material with a formation energy per atom of <target> eV/atom.
Your objective is to propose a new material composition that achieves the desired property by exploring the materials space across the periodic table.
In the previous step, you suggested the composition <composition>, and got the following feedback:
The formation energy per atom of material generated from the composition was <property value> eV/atom.

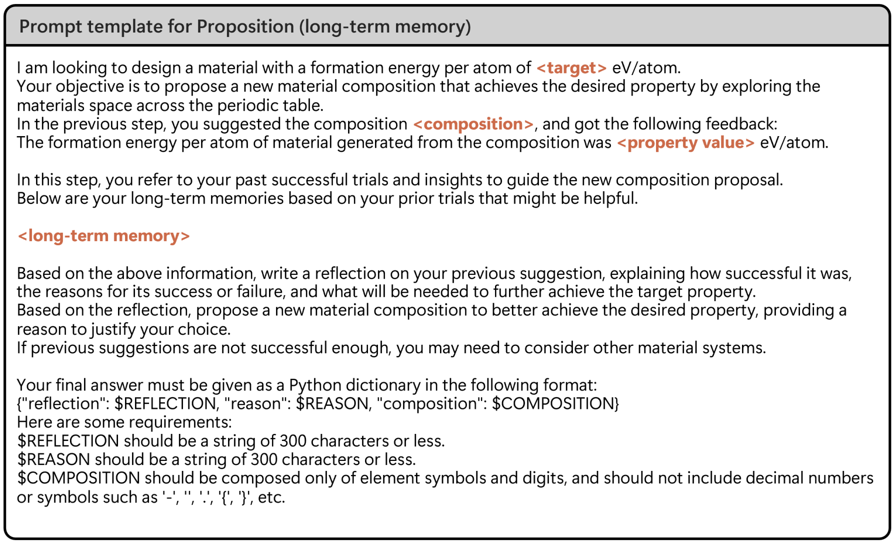
In this step, you review the most recent composition proposal to identify trends, refine approaches and avoid redundant or ineffective suggestions.
Below are your short-term memories based on your prior trials that might be helpful.

<short-term memory>

Based on the above information, write a reflection on your previous suggestion, explaining how successful it was, the reasons for its success or failure, and what will be needed to further achieve the target property.
Based on the reflection, propose a new material composition to better achieve the desired property, providing a reason to justify your choice.
If previous suggestions are not successful enough, you may need to consider other material systems.

Your final answer must be given as a Python dictionary in the following format:
{"reflection": $REFLECTION, "reason": $REASON, "composition": $COMPOSITION}
Here are some requirements:
$REFLECTION should be a string of 300 characters or less.
$REASON should be a string of 300 characters or less.
$COMPOSITION should be composed only of element symbols and digits, and should not include decimal numbers or symbols such as '-', '', '.', '{', '}', etc.

Figure 6: Prompt template for the Proposition stage when short-term memory has been selected. The template contains the following variables: **<target>**, **<composition>**, **<property value>**, and **<short-term memory>**. The **<short-term memory>** contains information regarding recent composition proposals, the reasoning behind those proposals, and the feedback received from the Property Evaluator.



Prompt template for Proposition (long-term memory)

I am looking to design a material with a formation energy per atom of <target> eV/atom.
Your objective is to propose a new material composition that achieves the desired property by exploring the materials space across the periodic table.
In the previous step, you suggested the composition <composition>, and got the following feedback:
The formation energy per atom of material generated from the composition was <property value> eV/atom.

In this step, you refer to your past successful trials and insights to guide the new composition proposal.
Below are your long-term memories based on your prior trials that might be helpful.

<long-term memory>

Based on the above information, write a reflection on your previous suggestion, explaining how successful it was, the reasons for its success or failure, and what will be needed to further achieve the target property.
Based on the reflection, propose a new material composition to better achieve the desired property, providing a reason to justify your choice.
If previous suggestions are not successful enough, you may need to consider other material systems.

Your final answer must be given as a Python dictionary in the following format:
{"reflection": $REFLECTION, "reason": $REASON, "composition": $COMPOSITION}
Here are some requirements:
$REFLECTION should be a string of 300 characters or less.
$REASON should be a string of 300 characters or less.
$COMPOSITION should be composed only of element symbols and digits, and should not include decimal numbers or symbols such as '-', '', '.', '{', '}', etc.

Figure 7: Prompt template for the Proposition stage when long-term memory has been selected. The template contains the following variables: **<target>**, **<composition>**, **<property value>**, and **<long-term memory>**. The **<long-term memory>** includes information on previous cases in which material compositions leading to structures with properties close to the target were successfully proposed, detailing the reasoning behind those proposals and the corresponding feedback received from the Property Evaluator.

Prompt template for Proposition (periodic table retrieval)

I am looking to design a material with a formation energy per atom of <target> eV/atom.
Your objective is to propose a new material composition that achieves the desired property by exploring the materials space across the periodic table.
In the previous step, you suggested the composition <composition>, and got the following feedback:
The formation energy per atom of material generated from the composition was <property value> eV/atom.

In this step, you can examine elements from the same group or with similar chemical properties to explore potential substitutions or optimizations.
Below are the elements related to previously suggested composition.
<elements>

Based on the above information, write a reflection on your previous suggestion, explaining how successful it was, the reasons for its success or failure, and what will be needed to further achieve the target property.
Based on the reflection, propose a new material composition to better achieve the desired property, providing a reason to justify your choice.
If previous suggestions are not successful enough, you may need to consider other material systems.

Your final answer must be given as a Python dictionary in the following format:
{"reflection": $REFLECTION, "reason": $REASON, "composition": $COMPOSITION}
Here are some requirements:
$REFLECTION should be a string of 300 characters or less.
$REASON should be a string of 300 characters or less.
$COMPOSITION should be composed only of element symbols and digits, and should not include decimal numbers or symbols such as '-', '', '.', '{', '}', etc.

Figure 8: Prompt template for the Proposition stage when the periodic table has been selected. The template contains the following variables: **<target>**, **<composition>**, **<property value>**, and **<elements>**. The **<elements>** contains information about elements belonging to the same group as those included in the previous proposal. For example, if a previous proposal contains Li and O, **<elements>** will include Li: H, Na, K, Rb, Cs and O: S, Se, Te.

Prompt template for Proposition (knowledgebase retrieval)

I am looking to design a material with a formation energy per atom of <target> eV/atom.
Your objective is to propose a new material composition that achieves the desired property by exploring the materials space across the periodic table.
In the previous step, you suggested the composition <composition> and got the following feedback:
The formation energy per atom of material generated from the composition was <property value> eV/atom.

In this step, you refer to an external database to understand how materials properties change when substituting one material to another.
Below is the information on how property values change with composition for several materials, which might be useful.

<Information>

Based on the above information, write a reflection on your previous suggestion, explaining how successful it was, the reasons for its success or failure, and what will be needed to further achieve the target property.
Based on the reflection, propose a new material composition to better achieve the desired property, providing a reason to justify your choice.
If previous suggestions are not successful enough, you may need to consider other material systems.

Your final answer must be given as a Python dictionary in the following format:
{"reflection": $REFLECTION, "reason": $REASON, "composition": $COMPOSITION}
Here are some requirements:
$REFLECTION should be a string of 300 characters or less.
$REASON should be a string of 300 characters or less.
$COMPOSITION should be composed only of element symbols and digits, and should not include decimal numbers or symbols such as '-', '', '.', '{', '}', etc.

Figure 9: Prompt template for the Proposition stage when the materials knowledge base has been selected. The template contains the following variables: **<target>**, **<composition>**, **<property value>**, and **<Information>**. The **<Information>** includes insights into how material properties change when compositions related to the previously proposed composition are modified, along with observations explaining why these changes occur.

Feedback template

The formation energy per atom of material generated from the composition was <property value> eV/atom.

Figure 10: Template for feedback generation. The **<property value>** is a variable.

Baseline prompt template

```
I am looking to design a material with a formation energy per atom of <target> eV/atom.
Your objective is to propose a new material composition that achieves the desired property by exploring the
materials space across the periodic table.
In the previous step, you suggested the composition <composition>, and got the following feedback:
The formation energy per atom of material generated from the composition was <property value> eV/atom.

Based on the feedback, propose a new material composition to better achieve the desired property.
If previous suggestions are not successful enough, you may need to consider other material systems.

Your final answer must be given as a Python dictionary in the following format:
{"composition": $COMPOSITION}
Here are some requirements:
$COMPOSITION should not include decimal numbers and symbols such as '-', '', '.', etc.
```

Figure 11: Prompt template for material composition proposal by the LLM without tool-assisted Planning and Proposition. **<target>**, **<composition>**, and **<property value>** are variables.

Prompt template for generating transition path

```
You are given compositions and formula units of two materials, <source composition> with formula units <source
formula units> and <target composition> with formula units <target formula units>.
Here, "composition" describes the types and proportions of elements in a material, whereas "formula units"
indicates how many times the normalized composition is repeated within a single unit cell.
The formation energy per atom of <source composition> is <source property value> eV/atom, while that of
<target composition> is <target property value> eV/atom.
Observe the change and predict why the formation energy differs when <source composition> transforms into
<target composition>, considering elemental properties, crystal structure, and chemical bonding.

Your output must be a Python dictionary in the following format:
{"observation": "$OBSERVATION"}
Here, $OBSERVATION is a concise explanation (≤500 characters) of the key factors influencing the formation energy
change.
```

Figure 12: Prompt template used to generate explanations of observations when transitioning from one composition to another during the construction of the knowledge base. The variables are follows: **<source composition>**, **<source formula units>**, **<source property value>**, **<target composition>**, **<target formula units>**, and **<target property value>**.

B Data distribution

Fig. 13 shows the distribution of formation energies in the MP-60 dataset. The dashed lines indicate the values corresponding to the 1.0% (-3.8 eV/atom), 2.5% (-3.5 eV/atom), 10% (-3.0 eV/atom), 20% (-2.5 eV/atom), and 40% (-1.6 eV/atom) quantiles, calculated by sorting the formation energy values in ascending order.

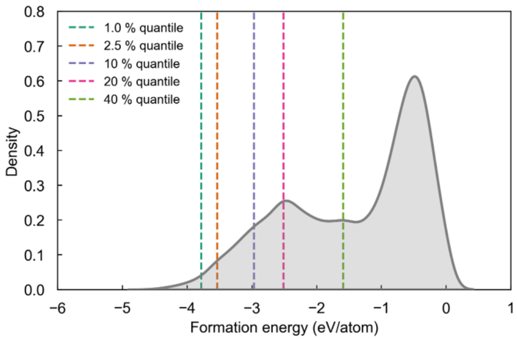


Figure 13: Distribution of formation energy per atom in MP-60 dataset.

C Natural language integration to initialization: Towards knowledge-informed materials design

In real-world materials exploration by domain experts, it is often preferable to begin with promising initial compositions based on prior knowledge or intuition. To advance toward such a knowledge-informed framework, we implemented two distinct methods for initializing compositions through natural language: one where the LLM proposes an initial composition based on a provided prompt (LLM-based sampling), leveraging its inherent chemical knowledge, and another that employs a retriever trained to identify relevant compositions from a database using natural-language queries (Retriever-based sampling).

In LLM-based sampling, the initial composition is proposed directly by the LLM. Specifically, the LLM is prompted using a request in natural language, such as: "I am looking to design a material with a formation energy per atom of **<target>** eV/atom. Could you suggest one possible material composition?", where **<target>** denotes the target property value, specifically the desired formation energy value. For the Retriever based sampling, we trained a Retriever using contrastive learning[1], employing the T5[2] as a text encoder, following the approach introduced in MatExpert[3]. Specifically, we performed contrastive learning by encoding both the property descriptions and their corresponding compositional descriptions (chemical formulas) of materials from the MP-20 dataset using the T5 encoder. This enables the retriever to effectively select relevant initial compositions based on natural-language queries describing desired material properties.

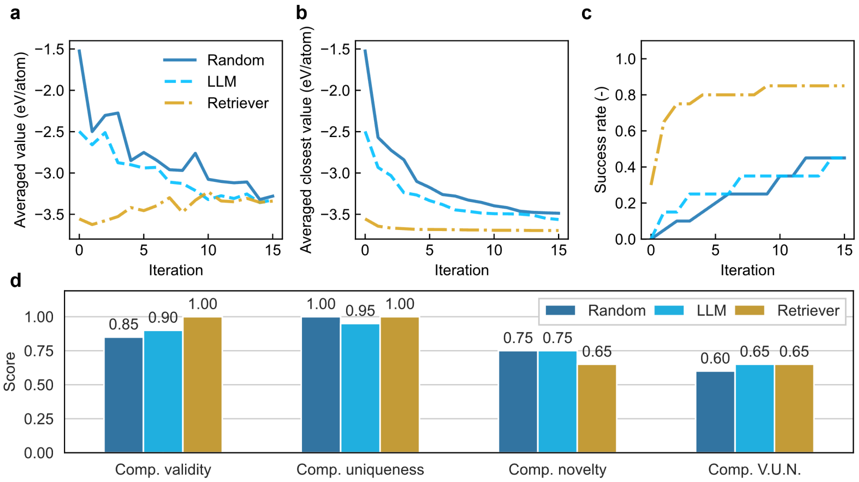


Figure 14: **Comparison of three initialization methods (Random, LLM, Retriever).** Target formation energy was set to −3.8 eV/atom. **a** Average property values at each iteration across 20 independent runs. **b** Average of the property values closest to the target achieved up to each iteration, calculated across 20 independent runs. **c** Success rates calculated at each iteration, averaged over 20 independent runs. **d** Compositional metrics (validity, uniqueness, novelty, and V.U.N. score) for compositions identified as closest to the target.

Fig. 14 compares three different initialization methods used to reach a target formation energy of −3.8 eV/atom. Fig. 14 **a**, **b**, and **c** show how the average property values evolve across iterations averaged over 20 independent runs, the average of the closest property values to the target observed in all preceding iterations, and the success rates over iterations, respectively. Finally, panel **d** summarizes compositional metrics (validity, uniqueness, novelty, and the V.U.N. score) for the compositions successfully identified as being close to the target value.

As shown in Fig. 14 **a**, both natural language-based initialization methods (LLM and Retriever) demonstrated superior starting points compared to Random initialization, with initial property values much closer to the target. The Retriever-based method, in particular, consistently provided initial compositions with formation energies nearest to the target value. This advantageous starting point enabled more focused exploration of the compositional space, leading to a significantly higher success rate throughout the iterative process, as evidenced in Fig. 14 **c**. Furthermore, Fig. 14 **d** reveals that compositions generated using the Retriever-based initialization achieve comparable V.U.N. score compared to Random initialization, indicating superior overall quality of the generated materials. These compelling results suggest that knowledge-informed initialization through natural language queries can substantially improve both the efficiency and effectiveness of targeted materials design while maintaining high compositional quality.

1. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning (2020)
2. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21 (2020).
3. Ding, Q., Miret, S. & Liu, B. Matexpert: Decomposing materials discovery by mimicking human experts. In International Conference on Learning Representations (2025).

D Element distribution

In Fig. 15, the left-column panels show elemental frequencies calculated from all proposed compositions, while the right-column panels show frequencies calculated only from compositions identified as closest to the target property. Rows one through four correspond to GPT-4o w/o tools, GPT-4o w/ tools, o3-mini w/o tools, and o3-mini w/ tools, respectively.

E Contribution of each tool

To better understand the contribution of each individual tool, we conducted additional experiments using GPT-4o, evaluating each tool separately, as shown in Table 1. Results suggest the knowledge base contributes to proposing unique compositions with high success rates. Long-term memory enhances compositional uniqueness and novelty, while the periodic table aids in exploring novel elemental spaces.

Method	Comp. validity	Comp. uniqueness	Comp. novelty	Num. of elements	Success rate
w/o tools	0.95	0.60	0.25	13	0.25
w/ Short-term memory	0.95	0.85	0.60	19	0.30
w/ Long-term memory	0.90	1.00	0.85	25	0.25
w/ Periodic table	0.80	1.00	0.60	36	0.10
w/ Knowledge base	0.95	1.00	0.70	24	0.65
w/ tools	0.85	1.00	0.75	27	0.45

Table 1: **Evaluation results for materials generated using each tool independently.** The table includes Performance metrics for different configurations: without tool-assisted Planning and Proposition (w/o tools), with each tool used independently, and with tool-assisted Planning and Proposition (w/ tools), where GPT-4o was used as the LLM backbone.

F Reasoning examples

One of the key advantages of the proposed framework lies in its interpretability, which stems from employing an LLM as the central reasoning engine. To investigate how the LLM reasons during the materials design process, we analyzed concrete examples of the reasoning outputs generated by the LLMs. Specifically, in the Planning stage, we collected the reasoning texts associated with each selected tool, converted them into embedding representations using SciBERT[1], and applied t-SNE[2] to project them into a two-dimensional space. The resulting visualization is shown in Fig. 16. As can be confirmed from Fig. 16, the embedding representations form distinct clusters, and the k-means clustering was applied to the embedded reasoning representations. The color of each point in the figure corresponds to its assigned cluster. To further interpret the characteristics of each cluster, raw reasoning texts associated with points in each cluster are shown in Fig. 18 for GPT-4o and in Fig. 19 for o3-mini.

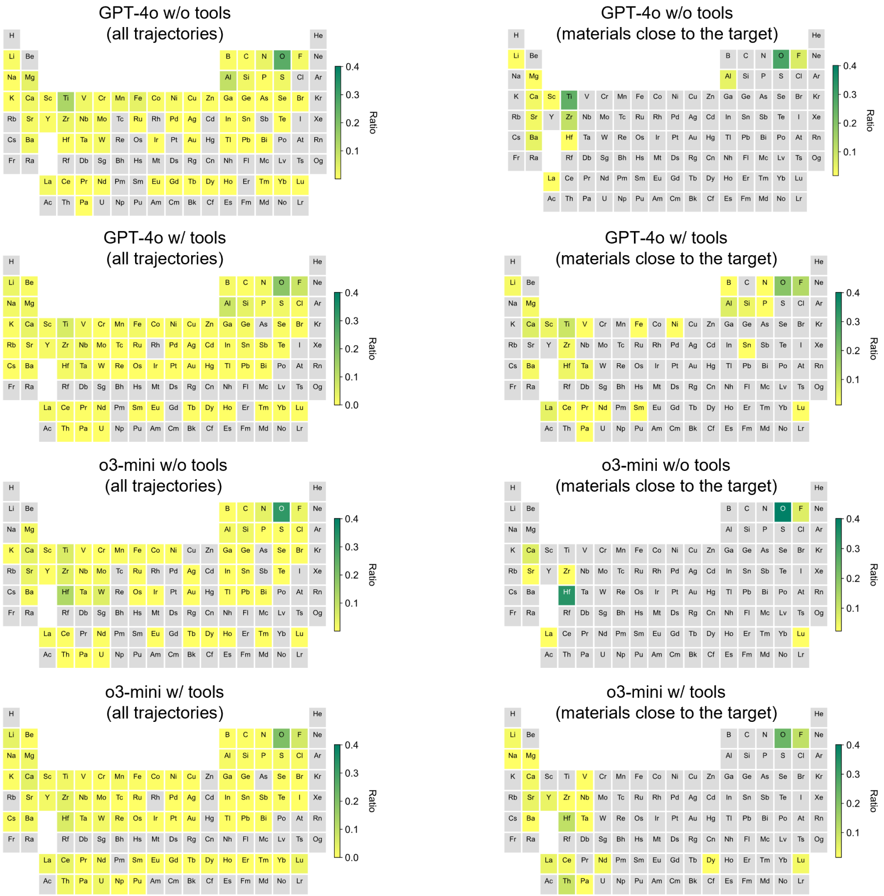


Figure 15: **Elemental distributions of proposed compositions.** Elemental distributions were calculated from compositions proposed across 15 iterations in 20 independent runs targeting a formation energy of -3.8 eV/atom. The left panels show frequencies from all proposed compositions; the right panels show frequencies from compositions closest to the target. Rows one to four correspond to GPT-4o w/o tools, GPT-4o w/ tools, o3-mini w/o tools, and o3-mini w/ tools, respectively.

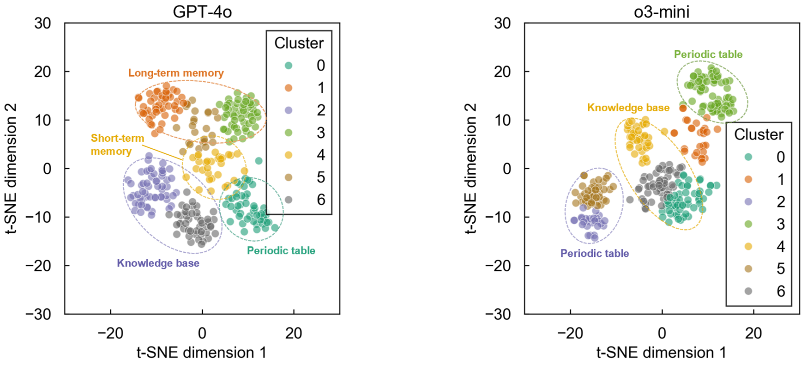


Figure 16: t-SNE visualization of the embedding representations of reasoning texts generated by the LLM during the Planning stage for GPT-4o and o3-mini. Each point represents a reason associated with a selected tool, and the colors indicate clusters obtained via k-means clustering.

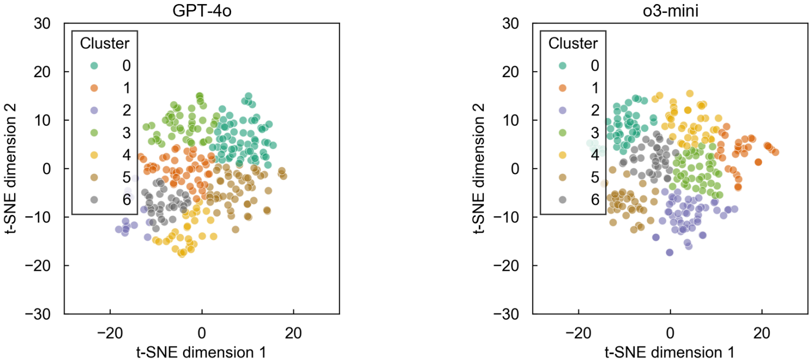
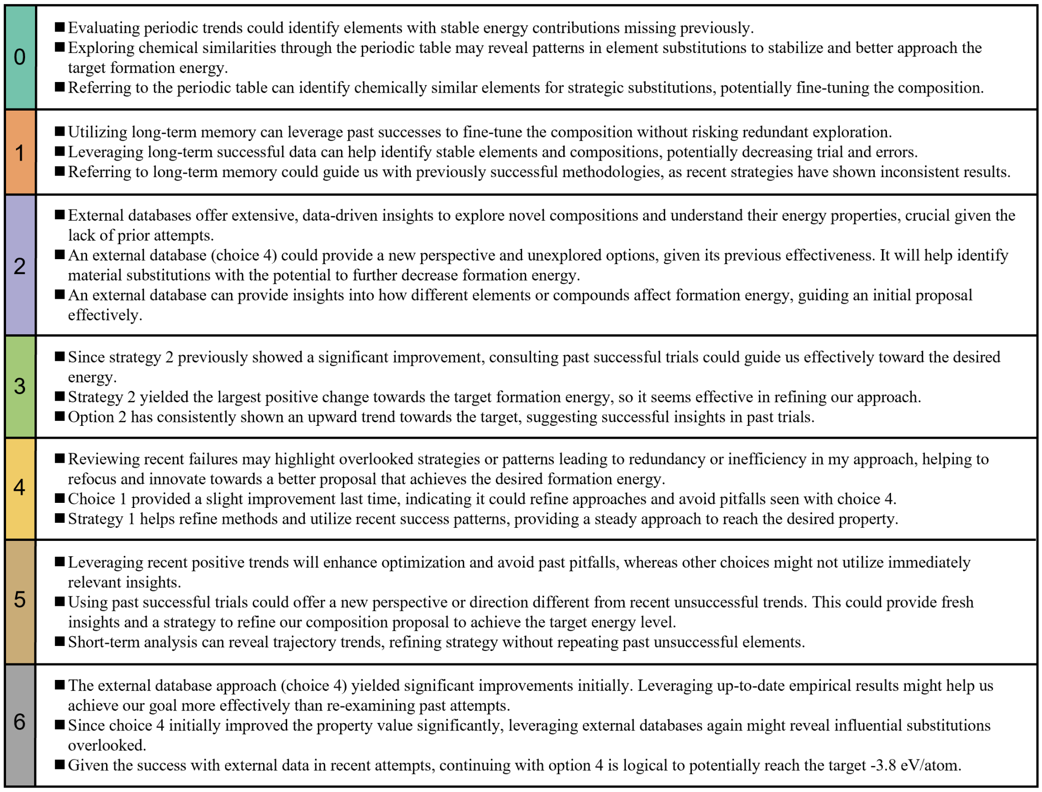


Figure 17: t-SNE visualization of the embedding representations of reasoning texts generated by the LLM during the Proposition stage for GPT-4o and o3-mini. Each point represents a reason associated with a selected tool, and the colors indicate clusters obtained via k-means clustering

By examining the reasoning texts within each cluster in Fig. 18 and Fig. 19, it becomes evident that each cluster corresponds to a specific tool selected by the LLM, indicating that the reasoning patterns are correlated with tool choice. Additionally, in Fig. 18, when the knowledge base is selected, the reasoning texts are further divided into two distinct clusters (purple Cluster 2 and gray Cluster 6). The texts in the gray cluster suggest that the LLM selected the knowledge base again based on prior experience, where using it previously led to successful outcomes. A similar trend was also observed for o3-mini as well. These results suggest that the LLM selects tools by taking past outcomes into account, demonstrating experience-informed decision-making.

A similar analysis conducted for the Proposition stage is shown in Fig. 17. Unlike in the Planning stage, the reasoning texts associated with composition proposals in the Proposition stage did not form clearly separatable clusters in the t-SNE analysis. This is likely because tool selection is a task involving a choice among four fixed options, whereas composition proposal is a more flexible task. For both GPT-4o and o3-mini, Fig. 20 and Fig. 21 show randomly selected reasoning texts from each region obtained by k-means clustering in the t-SNE space. Upon examining the reasoning texts, it was observed that texts containing words such as "substitution" or "replace" tend to be embedded in nearby regions. Although the physical and chemical correctness of the reasoning is not guaranteed in this study, it was observed that the LLM provides explanations using relevant keywords such as types of chemical bonding (e.g., ionic or covalent bonds), electronegativity, and atomic affinity. By further refining these explanations and improving their reliability, LLMs have the potential to provide valuable insights that human experts can interpret and effectively leverage in the materials design process.

1. Beltagy, I., Lo, K. & Cohan, A. Scibert: Pretrained language model for scientific text. In Conference on Empirical Methods in Natural Language Processing (2019).
2. van der Maaten, L. & Hinton, G. Visualizing data using t-sne. J. Mach. Learn. Res. 9, 2579–2605 (2008).

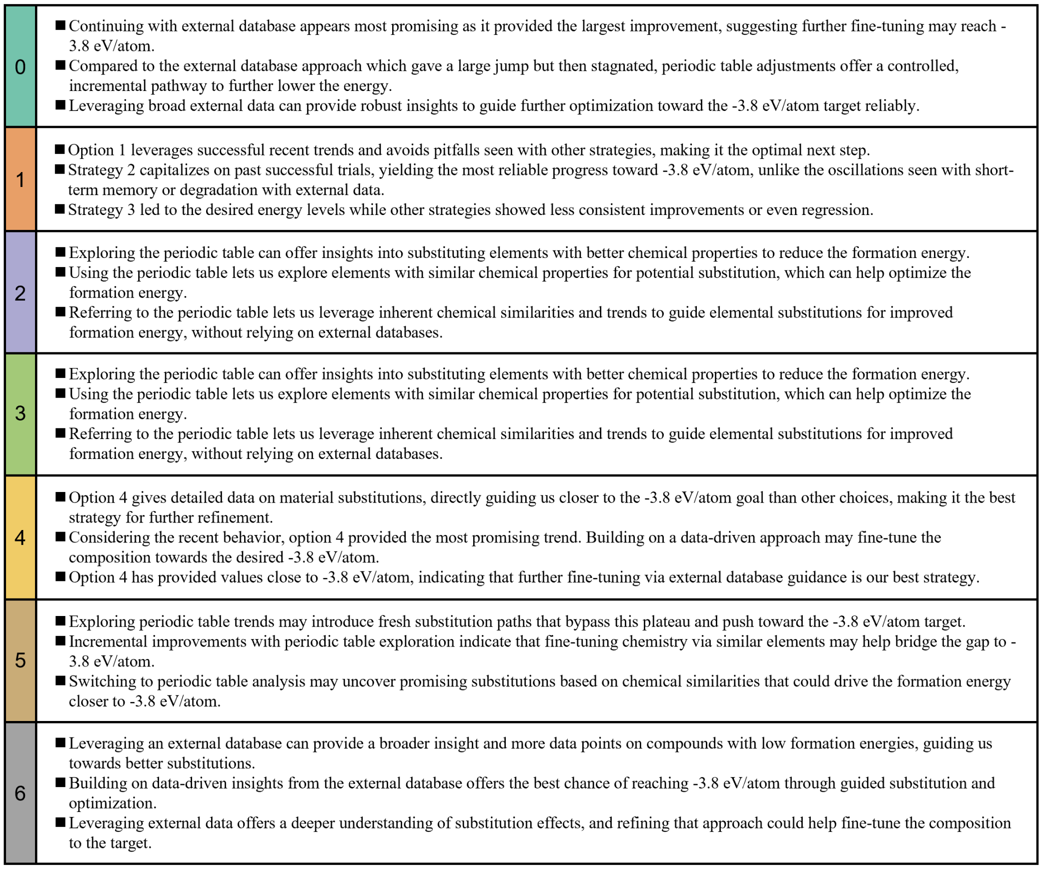


0	■ Evaluating periodic trends could identify elements with stable energy contributions missing previously. ■ Exploring chemical similarities through the periodic table may reveal patterns in element substitutions to stabilize and better approach the target formation energy. ■ Referring to the periodic table can identify chemically similar elements for strategic substitutions, potentially fine-tuning the composition.
1	■ Utilizing long-term memory can leverage past successes to fine-tune the composition without risking redundant exploration. ■ Leveraging long-term successful data can help identify stable elements and compositions, potentially decreasing trial and errors. ■ Referring to long-term memory could guide us with previously successful methodologies, as recent strategies have shown inconsistent results.
2	■ External databases offer extensive, data-driven insights to explore novel compositions and understand their energy properties, crucial given the lack of prior attempts. ■ An external database (choice 4) could provide a new perspective and unexplored options, given its previous effectiveness. It will help identify material substitutions with the potential to further decrease formation energy. ■ An external database can provide insights into how different elements or compounds affect formation energy, guiding an initial proposal effectively.
3	■ Since strategy 2 previously showed a significant improvement, consulting past successful trials could guide us effectively toward the desired energy. ■ Strategy 2 yielded the largest positive change towards the target formation energy, so it seems effective in refining our approach. ■ Option 2 has consistently shown an upward trend towards the target, suggesting successful insights in past trials.
4	■ Reviewing recent failures may highlight overlooked strategies or patterns leading to redundancy or inefficiency in my approach, helping to refocus and innovate towards a better proposal that achieves the desired formation energy. ■ Choice 1 provided a slight improvement last time, indicating it could refine approaches and avoid pitfalls seen with choice 4. ■ Strategy 1 helps refine methods and utilize recent success patterns, providing a steady approach to reach the desired property.
5	■ Leveraging recent positive trends will enhance optimization and avoid past pitfalls, whereas other choices might not utilize immediately relevant insights. ■ Using past successful trials could offer a new perspective or direction different from recent unsuccessful trends. This could provide fresh insights and a strategy to refine our composition proposal to achieve the target energy level. ■ Short-term analysis can reveal trajectory trends, refining strategy without repeating past unsuccessful elements.
6	■ The external database approach (choice 4) yielded significant improvements initially. Leveraging up-to-date empirical results might help us achieve our goal more effectively than re-examining past attempts. ■ Since choice 4 initially improved the property value significantly, leveraging external databases again might reveal influential substitutions overlooked. ■ Given the success with external data in recent attempts, continuing with option 4 is logical to potentially reach the target -3.8 eV/atom.

Figure 18: Examples of reasoning texts generated by GPT-4o during the Planning stage. For each cluster shown in Fig. 16, three reasoning examples were randomly selected from the points belonging to that cluster.

G Element distribution under different constraints

Figure 22 **a** illustrates the elemental distributions without any additional constraints, while panels **b–d** present the results obtained under specific constraints. Each distribution represents the ratio of elements contained in the proposed compositions generated by the LLM, aggregated over 10 independent runs—each consisting of 15 iterations, using GPT-4o as the LLM backbone. For the generation experiments, the target property values were set to -3.8 eV/atom for unconstrained generation and constrained generation under Constraints 1 and 2, while Constraint 3 had a target property value of -2.5 eV/atom.



0	■ Continuing with external database appears most promising as it provided the largest improvement, suggesting further fine-tuning may reach -3.8 eV/atom. ■ Compared to the external database approach which gave a large jump but then stagnated, periodic table adjustments offer a controlled, incremental pathway to further lower the energy. ■ Leveraging broad external data can provide robust insights to guide further optimization toward the -3.8 eV/atom target reliably.
1	■ Option 1 leverages successful recent trends and avoids pitfalls seen with other strategies, making it the optimal next step. ■ Strategy 2 capitalizes on past successful trials, yielding the most reliable progress toward -3.8 eV/atom, unlike the oscillations seen with short-term memory or degradation with external data. ■ Strategy 3 led to the desired energy levels while other strategies showed less consistent improvements or even regression.
2	■ Exploring the periodic table can offer insights into substituting elements with better chemical properties to reduce the formation energy. ■ Using the periodic table lets us explore elements with similar chemical properties for potential substitution, which can help optimize the formation energy. ■ Referring to the periodic table lets us leverage inherent chemical similarities and trends to guide elemental substitutions for improved formation energy, without relying on external databases.
3	■ Exploring the periodic table can offer insights into substituting elements with better chemical properties to reduce the formation energy. ■ Using the periodic table lets us explore elements with similar chemical properties for potential substitution, which can help optimize the formation energy. ■ Referring to the periodic table lets us leverage inherent chemical similarities and trends to guide elemental substitutions for improved formation energy, without relying on external databases.
4	■ Option 4 gives detailed data on material substitutions, directly guiding us closer to the -3.8 eV/atom goal than other choices, making it the best strategy for further refinement. ■ Considering the recent behavior, option 4 provided the most promising trend. Building on a data-driven approach may fine-tune the composition towards the desired -3.8 eV/atom. ■ Option 4 has provided values close to -3.8 eV/atom, indicating that further fine-tuning via external database guidance is our best strategy.
5	■ Exploring periodic table trends may introduce fresh substitution paths that bypass this plateau and push toward the -3.8 eV/atom target. ■ Incremental improvements with periodic table exploration indicate that fine-tuning chemistry via similar elements may help bridge the gap to -3.8 eV/atom. ■ Switching to periodic table analysis may uncover promising substitutions based on chemical similarities that could drive the formation energy closer to -3.8 eV/atom.
6	■ Leveraging an external database can provide a broader insight and more data points on compounds with low formation energies, guiding us towards better substitutions. ■ Building on data-driven insights from the external database offers the best chance of reaching -3.8 eV/atom through guided substitution and optimization. ■ Leveraging external data offers a deeper understanding of substitution effects, and refining that approach could help fine-tune the composition to the target.

Figure 19: Examples of reasoning texts generated by o3-mini during the Planning stage. For each cluster shown in Fig. 16, three reasoning examples were randomly selected from the points belonging to that cluster.

0	■ By substituting Lu with Sm, Hf with Zr, and O with S, we expect stronger metal-chalcogen bonds and increased exothermicity to better achieve the desired formation energy. ■ Substituting the larger Sr with the smaller Ca can increase the lattice energy due to stronger Ca-O interactions, potentially lowering the formation energy toward -3.8 eV/atom. ■ Replacing Ca with Ba and Hf with Ti is expected to enhance ionic interactions and lattice energy. Choosing S maintains a similar anionic framework while potentially deepening formation energy.
1	■ I propose TiN because titanium nitride forms robust metal-nitrogen bonds, has a stable crystal lattice, and is known to have formation energies near -3.8 eV/atom. ■ I propose TiC because titanium carbide exhibits strong Ti-C bonds and robust covalent interactions, with formation energies reported near -3.8 eV/atom, which may fulfill the target property. ■ Replacing Al with B, which has strong bonding potential, and Ho with Nd can enhance the cohesive energy, helping to reach the -3.8 eV/atom target.
2	■ I propose HfO2 because hafnium forms very strong bonds with oxygen. Binary oxides like HfO2 often exhibit highly negative formation energies due to strong ionic/covalent interactions, making it a promising candidate near -3.8 eV/atom. ■ Incorporating an element like Zr can enhance the metal-oxide network. HfZrO4, with an energy around -3.87 eV/atom, is a promising candidate near the -3.8 eV/atom target. ■ Increasing the Zr content can shift the average formation energy closer to -3.8 eV. The composition HfZr3O8, preserving the MO2 stoichiometry for each metal, is predicted to achieve this balance.
3	■ Increasing the oxygen content can enhance ionic interactions and lower formation energy. I propose EuPaO5 to potentially drive the energy closer to the desired -3.8 eV/atom. ■ Replacing Ho and Np with Er and Ti, as seen in the shift toward ErTiO3 in similar systems, should enhance ionic bonding and lower the formation energy closer to -3.8 eV/atom. ■ Replacing Hf with Zr in a fluoride framework (EuZrF7) should slightly weaken the ionic bonds compared to EuHfF7, potentially raising formation energy to near -3.8 eV/atom.
4	■ To approach -3.8 eV/atom, a robust oxide structure with strong ionic bonds is beneficial. BaLa2FeO6 is expected to achieve lower energy via stable Fe-O bonds and enhanced lattice ordering. ■ Substituting sulfur with nitrogen to form BaTiN3 should enhance ionic/covalent bonding, driving the formation energy closer to -3.8 eV/atom. ■ Switching to a spinel oxide like NiFe2O4, known for strong metal-oxygen ionic bonds, can provide the enhanced stability needed to approach a formation energy near -3.8 eV/atom.
5	■ I propose CeZr2O6 as oxygen enhances bonding stability and substituting Th with Zr may tune the formation energy closer to -3.8 eV/atom. ■ By increasing the oxygen and metal ratio, we enhance ionic bond strength. CeZr3O8 exhibits formation energy closer to -3.8 eV/atom, making it a promising candidate. ■ Substituting S with the more electronegative O and Eu with Ce can enhance bond strength. CeO2 is expected to demonstrate the needed stabilization.
6	■ Substituting Ge with C, which forms stronger bonds, and Ru with Os to further enhance stability, while using La for its chemical similarity, may drive the formation energy closer to -3.8 eV/atom. ■ In a double perovskite-type system, substituting Br with F boosts lattice energy. Cs2KBiF6 should exhibit stronger ionic bonds and a formation energy closer to -3.8 eV/atom. ■ By replacing Ce with La and Zr with Hf, the stronger bonds from Hf and La's similar chemistry may enhance exothermicity, moving closer to -3.8 eV/atom while retaining oxygen for its strong bonding.

Figure 20: Examples of reasoning texts generated by GPT-4o during the Proposition stage. For each cluster shown in Fig. 17, three reasoning examples were randomly selected from the points belonging to that cluster.

0	■ Using Al and F might form a highly stable fluoroaluminate. Introducing Al in NaAlF4 leverages strong Al-F bonds to enhance formation energy towards -3.8 eV/atom. ■ Substituting Na with Li increases ionic character and may lower formation energy due to Li's higher electronegativity and stability in fluoride lattices. Al is kept due to its stable bonding with fluorides, while F remains the same for its high electronegativity. ■ Replacing Al with Sc in LiScF4 introduces stronger bonding due to Sc's higher affinity for F, potentially increasing stability and formation energy per atom closer to -3.8 eV/atom.
1	■ Switching to a composition like NaThO3 may increase formation energy due to Th's oxophilicity and strong ionic bonds, potentially achieving the target -3.8 eV/atom. ■ Substitute Ge with Sn in Li2SnO3, as Sn's larger size than Ge and enhancing covalency with Sn–O bonds may better stabilize structure and lower formation energy. ■ Replacing Ni with Pt, and O with S may improve stability due to similar electronic properties and increased bond strengths, potentially achieving a more negative formation energy.
2	■ To achieve a more negative formation energy, we increase fluorination and include stronger bonds by considering elements like Ti that can enhance the stability of fluorine-rich compounds. ■ Choosing V2O5, with V-O bonds known for strong ionic character and suitable formation energy trends, could help achieve the desired energy. ■ Introducing Sc, which forms stable oxides, in place of some Hf might lower the formation energy as Sc-O bonds are strong. Aim for balance in valency and stability.
3	■ Replacing Si with Ge in Li2GeO3 might strengthen bonds due to Ge's larger atomic size, potentially achieving formation energy closer to -3.8 eV/atom. ■ Introducing a ternary oxide like Al2TiO5 might enhance stability due to Al's bonding, potentially reducing formation energy. ■ Incorporating a semi-metal like Ge and a transition metal like Fe in FeGeO3 could balance ionic and covalent bonds to tune the formation energy closer to -3.8 eV/atom.
4	■ Consider using Zr instead of Ti in combination with Na and F to explore enhanced stability through stronger ionic and covalent characteristics. Zr's presence in diverse stable fluoride forms justifies its potential to decrease formation energy. ■ Nd and Sm can stabilize structures with lower energy. Nb as a Ta substitute might also decrease formation energy. These changes aim to reach the desired formation energy more closely. ■ Incorporating Zr could enhance stability due to its strong ionic bonding like in successful LaTaO2F2. Thus, propose ZrTaO2F4 for improved formation energy.
5	■ Inspired by the stability in NaUF6 due to strong ionic bonds, we propose a composition with similar features. Using U and F with Na as a base could help approach the target. ■ Introducing Si could offer additional covalent character to strengthen interaction, further reducing energy. Si's role in forming stable silica analogs should enhance overall stability. ■ Incorporating an additional element like Ni to the YTaO2F2 structure might improve its stability by enhancing bonding strength.
6	■ By moving to a ternary oxide with mixed ionic-covalent bonds, we aim to achieve a more optimized formation energy. Combining elements like Nb, Sn, and O can yield promising results. ■ Exploring more stable, highly ionic compounds like TiO2 should help enhance formation energy due to their well-known stability and strong ionic bonds. ■ Exploring ZnO, with strong ionic bonds, is promising. Known for thermal and chemical stability, ZnO could help bridge the gap to reach the desired formation energy as it successfully combines light and transition metals.

Figure 21: Examples of reasoning texts generated by o3-mini during the Proposition stage. For each cluster shown in Fig. 17, three reasoning examples were randomly selected from the points belonging to that cluster.

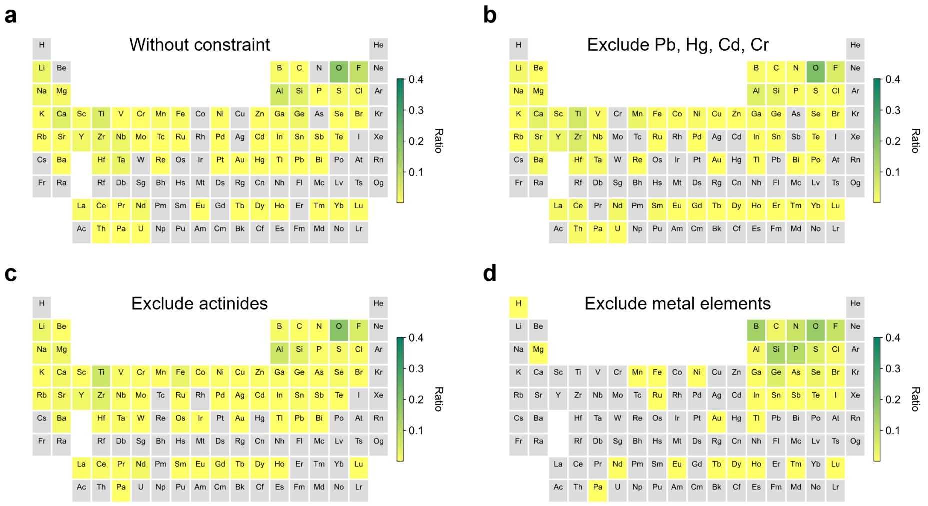


Figure 22: **Comparison of elemental distributions explored by the framework under different additional prompts.** **a** Elemental distribution without any additional constraints. **b** Elemental distribution under the constraint excluding Pb, Hg, Cd, and Cr. **c** Elemental distribution under the constraint excluding actinides elements. **d** Elemental distribution under the constraint restricting compositions to non-metallic elements only.