SMARTMAT EXPLORER: AUTOMATED SCIENTIFIC RESEARCH FOR MATERIALS DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in large language models (LLMs) have transformed artificial intelligence by positioning language as a universal interface for reasoning and interaction. Building on this foundation, scientific agents have emerged that can autonomously conduct literature analysis, generate hypotheses, guide experiments, and perform reflective evaluation. Despite this progress, enabling such agents with powerful inspiration-mining capabilities, rigorous scientific reasoning, and reliable feasibility verification remains challenging. We present SmartMat Explorer, a scientific agentic framework for functional molecular materials discovery centered on host–guest supramolecular complexes. It integrates a knowledge graph with a causally-grounded RAG to capture deeper causal structures for inspiration mining and employs a beam search–driven multi-level scoring mechanism informed by computational tools to enhance rational design decisions. To rigorously evaluate autonomous research performance, we introduce SmartMat-Bench, an extensive benchmark built from cutting-edge materials science publications and designed to cover both guided innovation and unconstrained exploration tasks. Empirical results show that SmartMat Explorer powered by Claude-3.7-Sonnet improves reasoning accuracy by 35% across five literature tasks and surpasses state-of-the-art deep research models by 6%. Coupling it with domain-specific tools further yields 33% gains in material discovery tasks. Overall, SmartMat Explorer offers a systematic and interpretable paradigm for robust and verifiable molecular materials discovery.

1 INTRODUCTION

Host–guest supramolecular complexes (Savyasachi et al., 2017; Mattia & Otto, 2015), as a class of functional materials, are molecular assemblies rationally designed to realize targeted functionalities and have demonstrated broad utility in sensing (Liu et al., 2024), drug delivery (Zhou et al., 2024), and catalysis (Li et al., 2021). Leveraging technological advances in structural characterization and thermodynamic measurements, it is now possible to achieve precise construction of host molecules and systematic regulation of interactions with guest molecules (Biedermann & Schneider, 2016). Although such advances demonstrate the importance of the traditional paradigm of hypothesis-driven synthesis and experimentation, current research remains highly dependent on empirical heuristics and trial-and-error strategies. This hinders research efficiency and scalability (Zhang et al., 2025). Bridging this gap requires automated methodologies capable of accelerating the discovery of host–guest complexes with tailored properties and applications, thereby expanding their impact across diverse scientific domains.

Recently, advances in automated scientific discovery have demonstrated remarkable progress, underscoring the transformative potential of large language models (LLMs) in accelerating scientific research (Guo et al., 2023; Tang et al., 2025b; Hong, 2023; Ghafarollahi & Buehler, 2024; Zhou et al., 2025; Inizan et al., 2025). In contrast to existing task-specific machine learning pipelines, which typically rely on large, high-quality datasets and often provide limited interpretability, LLM-based approaches can directly leverage the extensive physicochemical knowledge embedded in the scientific literature. This capability enables them to move beyond data-driven prediction toward the generation of novel hypotheses (Bazgir et al., 2025). But automated research systems are still in their infancy: they frequently hallucinate and lack the capability to tackle challenging scientific problems (Ren et al., 2025).

While promising recent advances have been reported in scientific domains such as gene editing (Roohani et al.; Wang et al., 2025) and protein design (Huang et al., 2025), the application of agentic system to materials research remains comparatively underexplored, particularly with respect to robust verification (Sprueill et al., 2024; Takahara et al., 2025). For example, MatPilot (Ni et al., 2024) integrates autonomous experimental validation to assess generated hypotheses, yet it still depends on repeated trials and demonstrates limited generality. Similarly, the OSDA agent (Hu et al., 2025) employs deep learning models to predict binding energies as an evaluation criterion, but its accuracy remains uncertain when extrapolated to novel molecules. An emerging frontier in accelerating research is scientific agentic systems, which are increasingly recognized as powerful tools. (Yue, 2025; Jiang et al., 2024; Liao et al., 2024). However, their application in materials research still faces several unresolved challenges.

Firstly, current approaches (Zhao et al., 2025; Zheng et al., 2025) are highly dependent on large models for idea generation, making them prone to chemically invalid molecules; in particular, complex molecular structures are often represented incorrectly, disrupting subsequent simulation and evaluation. Secondly, the scarcity and uneven distribution of high-quality materials data exacerbate the tendency of models to capture superficial correlations while overlooking the mechanistic and causal relationships that govern material behavior. Thirdly, the vastness of the scientific solution space further complicates exploration (Tang et al., 2025a): while specialized systems for materials design exist, prevailing agentic frameworks are only loosely grounded in physical and chemical principles, frequently producing candidate materials that violate fundamental constraints. Taken together, these limitations reveal a critical gap: the absence of a robust closed-loop validation framework capable of constraining idea generation, embedding mechanistic knowledge, and guiding exploration toward physically meaningful and experimentally feasible candidates.

To address these challenges, we propose an agentic system for material design that integrates generative idea exploration with knowledge graph–assisted reasoning to support diverse downstream objectives. The framework coordinates collaborative interactions among specialized agents, enabling efficient handling of complex tasks and domain-specific constraints. Our approach introduces three key innovations. First, to reduce invalid molecules generation, we combine coding exploration with database-assisted to construct molecular, followed by preliminary chemical validation, ensuring syntactically valid and chemically consistent structures. Second, to address data sparsity, we build a molecules–mechanisms–applications knowledge graph that captures multi-dimensional relationships, facilitating design principle extraction, knowledge transfer, and discovery of latent connections. Third, to navigate the vast solution space, we employ scientific tools to enforce physical constraints and perform comparative analyses, pruning infeasible candidates while yielding interpretable results aligned with physicochemical principles. Further, to rigorously evaluate SmartMat Explorer, we present SmartMat-Bench, a benchmark for systematic evaluation of literature comprehension and materials discovery. Together, these advances balance generative exploration with grounded validation, enhancing reliability, interpretability, and scalability in automated material discovery.

In summary, our contribution can be summarized as follows:

- We introduce *SmartMat Explorer*, an agentic framework that combines LLMs-driven idea exploration, knowledge reasoning, and physics-informed validation to improve reliability, interpretability, and scalability in automated materials discovery.
- We develop two complementary evaluation frameworks: *SMART* and *FIND-M*. The *SMART* evaluates the system's performance in literature comprehension and knowledge mining, while the *FIND-M* focuses on material screening.
- We introduce SmartMat-Bench, a comprehensive benchmark for evaluating literature synthesis, materials discovery accuracy, mechanistic validity, and experimental feasibility of host-guest supramaterial design, providing a unified platform for advancing smart materials design.

2 Related Work

2.1 Scientific Hypothesis Generation

Recent work has increasingly focused on automating scientific ideation to enhance creativity and reduce dependence on purely human-driven hypothesis generation. SciPIP(Wang et al., 2024b) advances systematic hypothesis generation by building a multi-granularity semantic literature database and introducing a dual-path creativity mechanism. However, its evaluation remains largely subjective, with limited domain coverage and high system complexity. SciMOM (Wang et al., 2024a) leverages contextual background modeling to explore open-ended problem spaces, optimizing for novelty within an unconstrained hypothesis landscape. In contrast, domain-specific efforts such as the MooseChem series target hypothesis discovery in chemistry. MooseChem(Yang et al., 2024) primarily generates coarse-grained hypotheses that lack experimental detail and are thus difficult to validate. MooseChem3 (Liu et al., 2025) incorporates simulated experimental feedback to improve ranking, yet such simulators inevitably diverge from real wet-lab outcomes, limiting reliability for novel or out-of-distribution hypotheses.

2.2 Agents for Materials Discovery

Recent advances in artificial intelligence have led to the emergence of increasingly sophisticated scientific agents designed to support complex research workflows (Huang et al.; Roohani et al.). Tool-assisted agents (M. Bran et al., 2024; Zhang et al., 2024) extend the capabilities of LLMs by integrating compute packages and calling tool (Chen et al., 2025), enabling them to perform diverse tasks ranging from data retrieval and statistical analysis to high-fidelity simulations. In parallel, Retrieval-Augmented Generation (RAG) has proven effective in grounding the generative power of LLMs with external knowledge bases, thereby facilitating nuanced tasks such as literature

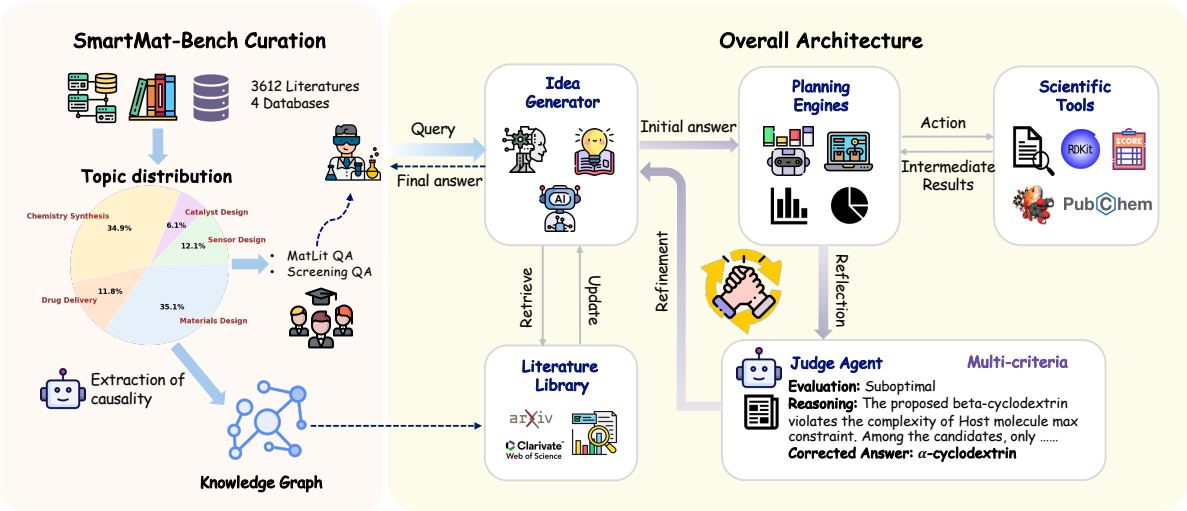

Figure 1: Overview of *SmartMat Explorer*. Our architecture consists of three main components: 1) Idea Generator, 2) Planning Engine, and 3) Judge Agent. These modules are further supported by the integration of literature library, domain-specific scientific tools, and web-based search utilities. The left panel illustrates the benchmark preparation process.

review, hypothesis generation, and domain-specific question answering through dynamic information retrieval and synthesis.

Building on these paradigms, agentic frameworks have been proposed to enhance reasoning and collaboration (Zhao et al., 2025). SciAgents (Ghafarollahi & Buehler, 2025), for example, integrate ontology-driven knowledge graphs with cooperative agent reasoning to uncover latent relationships and refine conceptual frameworks, but their reliance on static ontologies limits responsiveness to emerging literature and evolving knowledge. HoneyComb (Zhang et al., 2024) demonstrates the potential of modular, tool-integrated LLM agents for automating materials tasks, though its validation remains largely conceptual and underexplored in complex design scenarios. To enhance predictive and optimization capabilities, some approaches embed pre-trained models within the agent loop (Hu et al., 2025), yet their limited generalization constrains exploration of uncharted chemical spaces and, consequently, their contribution to genuine material discovery.

3 SmartMat-Bench Construction

To enable quantitative evaluation of the scientific discovery pipeline, we introduce **SmartMat-Bench**, a comprehensive benchmark for evaluating idea generation through literature synthesis and supramolecular screening. SmartMat-Bench supports two primary research objectives: (i) literature synthesis, which focuses on extracting knowledge related to supramolecular from scientific publications; and (ii) functional material screening, aimed at identifying promising candidate supramolecular based on targeted criteria.

MatLitQA benchmark. To evaluate the scientific reasoning of the system in the scientific literature, inspired by LitQA2 (Laurent et al., 2024), we selected the literature related to supramolecular complexes that has been published in top journals since 2020 for the construction of multiple-choice questions, named MatLitQA. The detailed literature collection is provided in the Appendix C.1.We randomly selected 84 multi-choice questions from 3620 scientific papers, extracted by GPT-4.1-nano, and divided into 5 subscenarios, reviewed by doctoral students in chemistry and materials majoring. After construction, the expert checks again (i) whether the material and molecular name are correct; (ii) whether the question does not contain any information in the abstract; (iii) whether the question meets the requirements of material design.

ScreeningQA benchmark. We evaluate material design capabilities in two sub-scenarios: cyclodextrin screening (CD) and liquid surface tension property prediction (ST), both aimed at designing supramolecular complexes that satisfy specific application requirements. For the CD task, we utilize the open cyclodextrin host–guest molecular pair database, which comprises 3,459 complexes, including 549 unique cyclodextrin molecules and 209 distinct guest molecules. These data were compiled from publications spanning 1963 to 2021. For the ST task, we adopt experimental surface tension data from real-world measurements reported in (Goussard et al., 2017), covering 262 molecules. The objective is to design materials optimized for surface tension. A complete description of the screening benchmark construction is provided in Appendix C.4.

4 METHODOLOGY

In this section, we introduce the agentic architecture for supramolecular complexes design, *SmartMat Explorer*. The architecture is organized around three LLM-powered components: (i) **Idea generator**, (ii) **Planning engine**, and (iii) **Judgment agent**. Among these, the planning engine interface with an external tool library to execute computational tasks (The overall pipeline is shown in Fig. 1). Further, *SmartMat Explorer* provides two evaluation frameworks: **S**tructured **MA**terials-inspiration **R**easoning and **T**ext-mining(**SMART**) and a **F**lexible approach for **IN**tegrated **M**aterials **D**iscovery(**FIND-M**). SMART assesses the capabilities of the system in literature review and idea generation, while FIND-M focuses on evaluating material screening and candidate selection. Detailed descriptions of the SMART and FIND-M components are presented in Subsections 4.1 and 4.2, respectively.

4.1 SMART: STRUCTURED MATERIALS-INSPIRATION REASONING AND TEXT-MINING

4.1.1 THE CONSTRUCTION OF KNOWLEDGE GRAPH

To conduct a literature review and capture causality of supramolecular complexes studied in the scientific literature, we extracted causal triples from 3620 documents to construct a knowledge graph. The triple is expressed as: $\mathcal{T} = \{(m, f, a) \mid m \in \mathcal{M}, f \in \mathcal{F}, a \in \mathcal{A}\}$, where $\mathcal{M}$, $\mathcal{F}$, and $\mathcal{A}$ denote *molecules*, *functions*, and *applications*, respectively. We first employ LLMs for named entity recognition and triple extraction. To identify domain-specific entities, we employ a hybrid strategy: combining rule-based pattern processing for explicit relationships while utilizing model-based parsing for implicit or context-dependent relationships. Extracted entities are organized into triples (m, f, a), which can be further connected into knowledge chains that encode domain-specific causal semantics. To reduce noise, we assign a confidence score to each extracted triple by integrating both cross-document co-occurrence frequency and semantic similarity.

We capture molecules–functions–applications in a structured form. A knowledge graph is then constructed by aggregating these triples, with entities represented as nodes and causal relations(e.g., required, driven, as shown in Fig. 2) as edges. Specifically, let a candidate triple be denoted as $\theta = (m, f, a)$. Its confidence score is defined as $w(\theta) = \alpha \cdot \varphi(\theta) + (1-\alpha) \cdot \cos(\theta)$, where $\varphi(\theta)$ denotes the normalized co-occurrence frequency of θ across the corpus, and $\cos(\theta)$ represents the semantic similarity between entities, computed in an embedding space (e.g., as the cosine similarity of their contextual embeddings). The trade-off parameter $\alpha \in [0, 1]$ balances statistical evidence against semantic coherence. During knowledge graph construction, triples with $w(\theta)$ below a threshold τ=0.1 are discarded, while higher weights strengthen the corresponding edges. This mechanism effectively emphasizes relations that are both frequently observed and semantically consistent, thereby reinforcing high-confidence links and suppressing spurious ones.

4.1.2 CAUSALLY-GROUNDED IDEA GENERATION

Direct prompting of LLMs often leads to hallucinations, while RAG mitigates this issue by grounding responses in external documents. However, standard RAG typically yields only shallow lexical evidence and lacks support for deeper reasoning. To address this limitation, we propose *Causally-grounded RAG*, which leverages causal chains as structured retrieval contexts. Rather than returning raw documents, our approach extracts and expands causal paths, transforming knowledge graph retrieval into structured multi-hop reasoning, as show in Fig. 2.

We formulate idea generation as a multi-stage search over a structured **concept space** ($\mathcal{C}$), which extends beyond embedding spaces by jointly modeling entities, relations, and graph. This representation enables the discovery of novel and query-aligned causal paths that serve as candidate hypotheses. Formally, the conceptual space $\mathcal{C}$ consists of an entity space $\mathcal{E}$, where each domain-specific entity i is encoded by an encoder $\mathbf{v}(\cdot)$ into a continuous representation $e_i \in \mathbb{R}^{n \times d}$. In addition, a query embedding $\mathbf{e}_q \in \mathbb{R}^{n \times d}$ is introduced to constrain the search process to query-relevant regions of $\mathcal{C}$, thereby ensuring that subsequent reasoning and retrieval operate on concepts that are semantically aligned with the query context.

To balance relevance and diversity during retrieval, we employ Maximal Marginal Relevance (MMR) reranking (Carbonell & Goldstein, 2017), which prioritizes entities that are both highly related to the query and mutually dissimilar. The MMR score for a candidate entity d_i with embedding $\mathbf{e}_{d_i}$ is defined as:

$$\text{MMR}(d_i; S_e, \mathbf{e}_q, \lambda) = \lambda \cdot sim(\mathbf{e}_q, \mathbf{e}_{d_i}) - (1 - \lambda) \cdot \max_{d_j \in S_e} sim(\mathbf{e}_{d_i}, \mathbf{e}_{d_j}), \tag{1}$$

where S_e denotes the set of already selected entities, $sim(\cdot)$ represents cosine similarity, and $\lambda \in [0, 1]$ controls the trade-off between relevance and novelty. We retain the top-k ranked entities. Subsequently, rather than relying directly on unstructured documents, we convert the retrieved candidates into a structured causal space by extracting explicit causal chains:

$$\mathcal{G}_c = (e_1 \xrightarrow{r_1} e_2 \xrightarrow{r_2} \ldots \xrightarrow{r_m} e_{m+1}), \tag{2}$$

where $e_{m+1} \in \mathcal{E}$ and $r_m \in \mathcal{R}$. These structured triples encode directional dependencies between entities, transforming unstructured textual evidence into interpretable reasoning paths suitable for downstream causal inference.

To enable cross-document inference while preventing unbounded expansion, entities in causal triples are extended via a graph search on $\mathcal{G}$. For a given entity embedding e, we retrieve all neighboring relations up to depth L as follows:

$$\mathcal{N}_L(e; \mathcal{G}) = \{(e \xrightarrow{r} e') \mid \text{dist}_{\mathcal{G}}(e, e'') \leq L\}, \tag{3}$$

where e'' denotes an entity reachable within L hops from e, and e' represents its directly connected neighbor. This procedure links otherwise isolated causal triples into extended multi-hop paths, thereby bridging disjoint evidence sources. Hypotheses are then generated by iteratively expanding causal paths in $\mathcal{G}$, guided by semantic similarity in the entity space $\mathcal{E}$, relational coherence in the relation space $\mathcal{R}$, and alignment with the query embedding $\mathbf{e}_q$. This integration ensures that the resulting candidate causal paths are both structurally valid and contextually relevant to the research question.

4.1.3 Path Scoring and Ranking

To identify the most plausible causal chains, we reformulate path scoring and ranking as a beam search process over the causal chain. Starting from the query node, the search iteratively expands candidate paths by exploring neighboring entities in $\mathcal{G}$. At each expansion step t, each partial path $\mathcal{G}_t$ is assigned a cumulative confidence score $\mathcal{C}(\mathcal{G}_t)$ defined as:

$$\mathcal{C}(\mathcal{G}_t) = \sum_{i=1}^{t} \alpha \cdot \mathcal{S}(e_i, r_i, e_{i+1}) + \beta \cdot \mathcal{R}(e_i, e_{i+1} \mid q), \tag{4}$$

where $\mathcal{S}(e_i, r_i, e_{i+1})$ measures the semantic coherence and factual support of a causal edge $(e_i \xrightarrow{r_i} e_{i+1})$, and $\mathcal{R}(e_i, e_{i+1} \mid q)$ quantifies its relevance to the query q. The weights α and β balance local edge reliability and global query alignment. At each iteration, only the top-k paths with the highest $\mathcal{C}(\mathcal{G}_t)$ are retained in the beam, pruning low-confidence branches and focusing search on promising causal chains. The process terminates when either a maximum depth $L_{\max}$ is reached or no higher-scoring expansions remain. The ultimate goal is to identify the most plausible causal chain conditioned on the query:

$$f(q, \mathcal{G}_c) = \arg\max \Phi(\mathcal{G}_c \mid q). \tag{5}$$

This retrieval-augmented causal expansion provides three benefits: (i) it goes beyond *document-level retrieval*, yielding structured causal graphs; (ii) it supports *multi-hop reasoning*, enabling connections across disjoint evidence sources; and (iii) it ensures *robust ranking*, prioritizing paths that are both semantically aligned with the query and structurally coherent. Together, these properties allow the system to act not only as a retriever but as a reasoning engine capable of generating novel scientific ideas by effectively navigating and exploiting the defined concept space.

4.2 FIND-M: A Flexible approach for Integrated Materials Discovery

Evaluating newly discovered materials presents substantial challenges. Traditional binding energy calculations rely on density functional theory, which is accurate but computationally prohibitive, whereas deep learning approaches often lack robustness Hu et al. (2025) when extrapolating to novel molecular structures. To address this gap, we propose a heuristic evaluation framework tightly integrated with tool-assisted interaction. This framework systematically evaluates molecular descriptors, including size, structural complexity, and aromaticity, to heuristically assess the suitability of host molecules for guest inclusion. We denote it as FIND-M.

4.2.1 Planning Engine

The planning engine serves as the central coordinator of the scientific agent, orchestrating the entire discovery process. It converts advanced scientific ideas into a series of executable tasks, such as idea validation, tool invocation, code execution, and database querying (Wei et al., 2025). In our task, given that molecular names generated from the idea generator are generally not directly usable for structure optimization, and directly generated SMILES are prone to hallucinations, the planning engine can employ a collaborative approach of database queries and code refinement to generate new molecular structures. Specifically, when molecular design is required, the agent first queries the PubChem Library via the planning engine to obtain raw molecular representations, including both the SMILES string and the corresponding structural file. Guided by the task specification, the agent then performs structural modifications (e.g., inserting functional groups at primary hydroxyl sites) to generate initial candidate molecules. This process is automated through coding that is executed within the environment to ensure reproducibility and minimize manual intervention.

Once the initial SMILES is obtained, the system proceeds to the property evaluation pipeline. First, lightweight cheminformatics descriptors, such as molecular complexity and aromaticity, are computed to provide rapid heuristic assessments of molecular feasibility. Next, selectively evaluate more accurate physicochemical properties by interfacing with a chemistry software package, such as xtb (Bannwarth et al., 2019)), to approximate electronic properties and more. For further details on the use of chemical tools, refer to Appendix A.

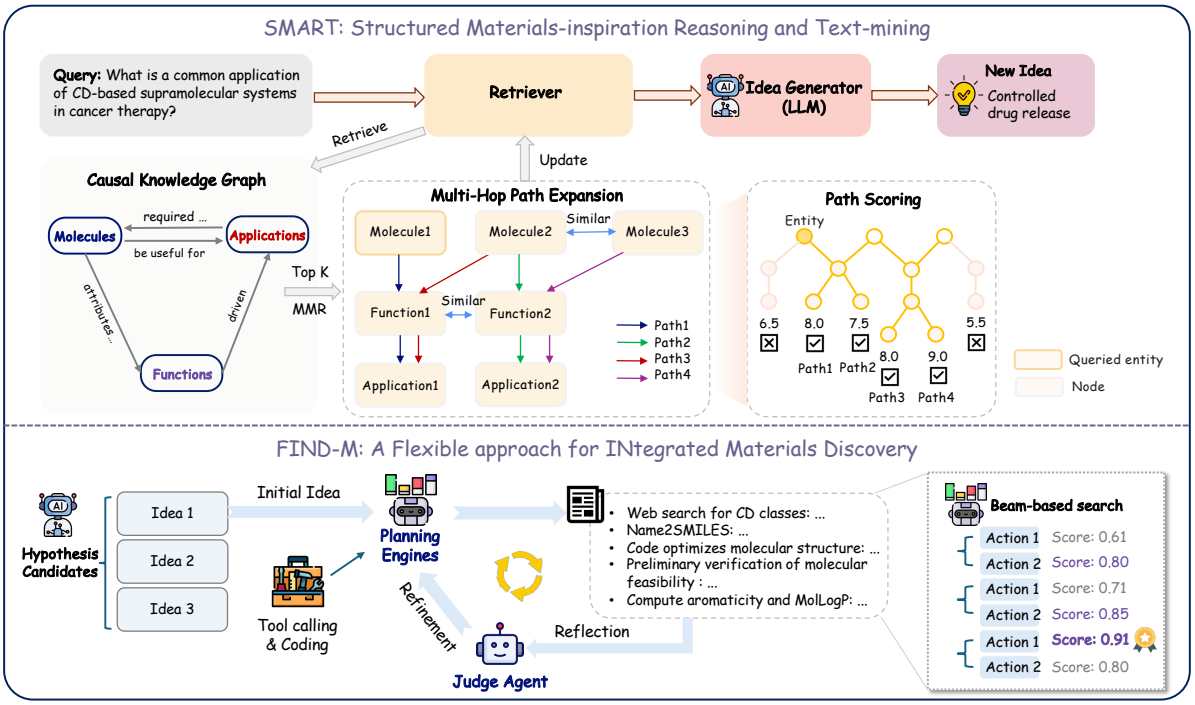

Figure 2: Overview of Two Complementary Evaluation Frameworks: SMART(a Structured MAterials design Reasoning and Text-mining scheme) and FIND-M(a Flexible approach for INtergated Materials Discovery).

Formally, the property set for a candidate molecule m can be expressed as $\mathcal{P}(m) = \mathcal{P}_{\text{heuristic}}(m) \cup \mathcal{P}_{\text{semi}}(m)$, where $\mathcal{P}_{\text{heuristic}}$ denotes descriptor-level features and $\mathcal{P}_{\text{semi}}$ corresponds to physics-grounded quantities derived from electronic structure calculations. This hierarchical assessment ensures that the reasoning process is simultaneously informed by fast, heuristic descriptors and rigorous, physics-based estimations, thereby balancing computational efficiency with physical fidelity.

4.2.2 Reasoning and Reflection

Given a user query q, the planning engine's initial response, an associated confidence score c_0, and the complete interaction history $\mathcal{T}$. Building on 4.2.1, we improve the idea agent with causal triples to stimulate novel hypotheses and to promote knowledge generalization. Details of the prompt templates and examples are provided in Appendix C.3. To ensure both *validity* and *optimality* in material discovery, we incorporate a structured reflection mechanism into the reasoning pipeline. This reflection process conducts a post-hoc evaluation of the model's initial proposal and applies corrections when necessary, proceeding through two phases:

Reasoning Phase. The agent is initialized with domain-specific instructions and a tool library $\mathcal{K} = \{k_i\}$ (e.g., aromaticity checkers, MolLogP calculators, complexity estimators). At each iteration t the planning engine issues tool calls which are validated against function for type-consistency. Molecule identifiers are canonicalized to SMILES to guarantee compatibility with cheminformatics tools. Tool outputs are serialized and appended to T, producing a verifiable evidence trace. The loop terminates when the model emits a candidate molecule m_0 or a maximum iteration budget $T_{\max}$ is reached.

Reflection Phase. The reflectionx phase occurs in the judgment agent. Let $\mathcal{M}_{\text{cand}}$ denote the pool of candidates collected during reasoning and let $\mathcal{P} = \{p(m) \mid m \in \mathcal{M}_{\text{cand}}\}$ be the summary of consolidated properties, where $p(m)$ is a property for the molecule m. The reflection process ϕ is denoted as

$$\hat{m} = \phi(m_0, \mathcal{P}) = \phi\Big(\arg\max_{m \in \mathcal{O}(m_0, \mathcal{P})} s(m; \mathcal{P}),\ \mathcal{P} \Big), \tag{6}$$

where ϕ is the update operator, $\mathcal{O}$ is the validity operator, and s is the selection operator. In practice we implement a deterministic acceptance rule to decide whether the reflection result replaces the original proposal. Let $s_0 = s(m_0; \mathcal{P})$ and $m^\star = \arg\max_{m \in \{(m_0, \mathcal{P})} s(m; \mathcal{P})$. Then

$$\hat{m} \leftarrow \begin{cases} \phi(m^\star, \mathcal{P}, s(m^\star; \mathcal{P})), & \text{if } s(m^\star; \mathcal{P}) > s_0 + \tau, \\ m_0, & \text{otherwise}, \end{cases} \tag{7}$$

Table 1: Performance comparison of agentic systems on literature comprehension tasks. Results report accuracy and per-1M-token cost averaged over five runs. The best model is **bolded**, the second best is underlined, and relative improvements are given in bracket. We provide standard variance in the appendix 7.

Model	Catalysis	Drug	Sensor	Chemistry	Materials	Average	Cost ($)
GPT-4.1-nano	84.62	84.00	80.00	77.03	**93.33**	83.80 (+18%)	0.10
GPT-5-nano	84.62	84.00	80.00	75.68	84.44	81.75 (+11%)	0.05
GPT-5-mini	78.85	86.00	82.00	79.73	82.22	81.76 (+0.5%)	0.25
GPT-5	**92.31**	86.00	80.00	**87.84**	86.67	86.56(+8.4%)	1.25
GPT-4o-mini	80.77	74.00	82.00	72.97	73.33	76.61 (+0.6%)	0.15
GPT-4o	84.62	80.00	82.00	74.32	84.44	81.08 (+4.5%)	2.50
o3-mini	84.62	88.00	82.00	82.43	84.44	84.30 (+13%)	1.10
o3	88.46	86.00	82.00	86.49	88.89	86.37 (+18%)	2.00
o4-mini	88.46	88.00	**88.00**	83.78	86.67	**86.98**(+18%)	1.10
Claude 3.5 Sonnet	82.69	74.00	72.00	75.68	84.44	77.76 (+76%)	1.50
Claude 3.7 Sonnet	86.54	**90.00**	86.00	85.14	86.67	86.87(+35%)	1.50
o3 Deep Research	80.77	86.00	82.00	74.32	91.11	82.84	10.00

where $\tau \geq 0$ is acceptance threshold that prevents oscillation due to minor score fluctuations. Otherwise, the reflection may be iterated until improvement falls below threshold or no further feasible candidates are generated. This reflection mechanism yields a interpretable pipeline that mitigates premature convergence, balances exploration of external knowledge with exploitation of confident reasoning, and frames inference as a constrained search robust to noisy retrieval.

5 Experiments

5.1 Experimental Setup

Implementation Details. During evaluation, we carefully control the configurations to ensure stable and comparable performance. Specifically, for most models, we set the sampling temperature to 0.5, whereas for certain reasoning-oriented models such as OpenAI o3 and OpenAI o3-mini, we follow their default settings without temperature adjustment. For Claude 3.7 Sonnet, we enable the "think" mode to explicitly elicit structured reasoning traces. In reasoning settings, final answers are extracted using a regular-expression-based parsing strategy, ensuring reliable identification of the selected option for subsequent accuracy analysis.

Evaluated Models. We tested 11 LLMs of varying complexity for *SmartMat Explorer* (GPT-4o, GPT-4o-mini, GPT-4.1-nano, GPT-5-mini, GPT-5, GPT-5-nano, GPT-o3, o3-mini, o4-mini, Claude 3.5 Sonnet, and Claude 3.7 Sonnet). Since most organizations lack transparent parameter counts, we use the price per token as a proxy for model size. Additionally, we evaluate Deep Research model: OpenAI o3 Deep Research.

Evaluation Protocols. To ensure a rigorous and multi-faceted evaluation, we adopt four complementary metrics: Accuracy (Laurent et al., 2024), Success Rate (Huang et al., 2023; Sun et al., 2025), Semantic Similarity (Sprueill et al., 2024), and BLEU (Papineni et al., 2002). Together, these metrics capture complementary aspects of evaluation, offering a holistic view of system performance across literature comprehension and material discovery tasks. Further details of the evaluation scheme are provided in Appendix B.

5.2 Main results

To assess the effectiveness of our *SmartMat Explorer*, we benchmarked it against 11 LLMs, encompassing both reasoning models and deep research models, on literature comprehension tasks. To ensure fair comparison, all models were evaluated under identical settings. We adopted two widely recognized evaluation metrics: accuracy and computational cost.

SmartMat Explorer improve all baselines for literature comprehension experiments. We evaluated performance on literature comprehension task across a diverse set of LLMs (Table 1). We observe that reasoning models such as GPT-o4-mini and Claude 3.7 Sonnet deliver strong average accuracy (86.98% and 86.87% respectively), with notable performances on specific domains. For example, Claude 3.7 Sonnet reaching 90.0% on drug delivery task, and GPT-4.1-nano achieving 93.33% on materials science task. In contrast, chat models such as GPT-4.1-nano and GPT-5-nano achieve slightly lower overall accuracy (83.80% and 81.70%) but exhibit exceptional cost-efficiency (<$0.12 per 1M tokens). Meanwhile, larger models like GPT-4o show no clear advantage in accuracy (81.08%) despite incurring substantially higher costs. Considering that the performance improved

substantially after incorporating our system into a broader range of LLMs, our *SmartMat Explorer* adopts Claude 3.7 Sonnet (improve +35%) as backbone model and further integrates it with knowledge graphs, domain-specific scientific tools, and reflection reasoning mechanisms. This design enables the generation of accurate and interpretable reasoning trajectories, positioning *SmartMat Explorer* as a competitive solution for challenging literature comprehension tasks.

SmartMat Explorer provides an efficient and reliable aid for accelerating material screening. Overall performance s are present in Table 2. As shows in Table 2, *SmartMat Explorer* consistently surpasses strong conversational and reasoning LLM baselines on both literature comprehension (MatLitQA) and materials screening (ScreeningQA) benchmarks. Averaged over five runs, *SmartMat Explorer* achieves 86.87% on MatLitQA, representing a 4.6% improvement absolute gain over the best baseline (82.24%). The improvement is even more pronounced on ScreeningQA: *SmartMat Explorer* attains 83.82% on Surface Tension (ST) task and 77.50% on Cyclodextrin (CD) task, whereas the strongest baseline scores 79.41% on ST but only 27.00% on CD. The large improvement on CD experiment highlights *SmartMat Explorer*'s ability to perform iterative grounding, targeted retrieval, and idea refinement, which substantially improve concept-level discovery beyond factoid-style discovery. These improvements are consistent across all domains, suggesting broad generalization rather than domain-specific overfitting.

Table 2: Comparison of agent performance between vanilla LLMs and the proposed *SmartMat Explorer* on literature comprehension and materials screening tasks. Reported values correspond to average accuracy (%) over five independent experimental rounds. The best model is **bolded** and the second best is underlined.

Model	MatLitQA						ScreeningQA	
	Catalysis	Drug	Sensor	Chemistry	Materials	Avg.	ST	CD
Chat LLMs								
GPT-4.1-nano	73.08	76.00	66.00	60.81	77.78	70.73	52.94	24.50
GPT-5-nano	75.00	82.00	70.00	67.57	73.33	73.58	54.41	9.50
GPT-5-mini	80.77	86.00	82.00	75.68	82.22	81.33	72.06	10.00
GPT-5	86.54	84.00	70.00	81.08	77.78	79.88	79.41	27.00
Claude-3.5-Sonnet	42.31	44.00	40.00	45.95	48.89	44.23	44.12	18.50
Reasoning LLMs								
GPT-4o-mini	76.92	72.00	76.00	75.68	80.00	76.12	45.59	5.00
GPT-4o	75.00	84.00	76.00	72.97	80.00	77.59	57.35	23.50
o3-mini	75.00	82.00	66.00	68.92	82.22	74.83	69.12	10.50
o3	75.00	80.00	66.00	60.81	84.44	73.25	79.41	20.00
o4-mini	74.00	82.00	70.00	60.81	80.00	73.36	72.06	7.00
Claude-3.7-Sonnet	78.85	64.00	54.00	56.76	68.89	64.50	47.06	23.00
Deep Research Models								
o3 DeepResearch	80.77	86.00	82.00	74.32	**91.11**	82.24	60.29	20.00
SmartMat Explorer	**86.54**	**90.00**	**86.00**	**85.14**	86.67	**86.87**	**80.88**	**77.50**

Overall, *SmartMat Explorer* establishes a state of the art performance across a benchmarks, demonstrating superiority in literature-grounded reasoning, mechanistic understanding, and cross-domain scientific discovery. These findings confirm that integrating causal-aware retrieval with iterative reasoning significantly enhances performance beyond both general-purpose LLMs and domain-agnostic deep-research models.

5.3 Ablation Study

SmartMat Explorer uses retrieved knowledge and scientific tools to interpretively guide material screening. We conduct an ablation study to analysis the contribution of each component by respectively removing RAG, tools, and reflection mechanism. Table 3 reports results on the ScreeningQA benchmark across four baseline models. The vanilla agent exhibits relatively low accuracy across all settings, underscoring the inherent difficulty of the task. Removal of RAG results in a decrease in performance, but is still above the baseline, indicating that the external knowledge base provides a useful background, but is still insufficient for complex reasoning. Incorporating specific tools produces more substantial improvements, highlighting the importance of structured computational support beyond text-based retrieval. Reflection further enhances robustness by enabling iterative self-correction, though the degree of improvement varies across models. When all modules are integrated into *SmartMat Explorer*, performance improves both significantly and consistently (e.g., 77.50% on Claude 3.7 Sonnet and 74.00% on o3-mini). On the other side, although the additional modules introduce moderate inference overhead, the computational cost remains far lower than that of wet-lab experiments or deep research models. Overall, the ablation confirms that the three modules are complementary and jointly essential for effective materials discovery.

Table 3: Ablation study of *SmartMat Explorer* performance on materials screening with varying components. Here, the abbreviation for causally-grounded RAG is labeled RAG. Reported values represent the average accuracy (%) and time consumption (s) over 5 independent runs.

Method	o3		o3-mini		GPT-4o		Claude-3.5-Sonnet		Claude-3.7-Sonnet	
	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time
Ours	**77.00**	57.93	**74.00**	31.60	**70.00**	75.71	**42.50**	26.08	**77.50**	32.16
w/o CausalRAG	65.00	45.95	65.00	40.05	25.00	61.95	20.00	23.71	25.50	20.24
w/o Tools	55.00	50.43	70.00	31.55	45.00	52.58	15.00	24.76	50.00	30.81
w/o Reflection	31.00	40.01	51.50	16.12	24.50	15.18	19.00	11.05	39.50	16.02
Vanilla	20.00	9.88	10.50	10.98	23.50	1.48	14.00	9.01	24.50	14.56

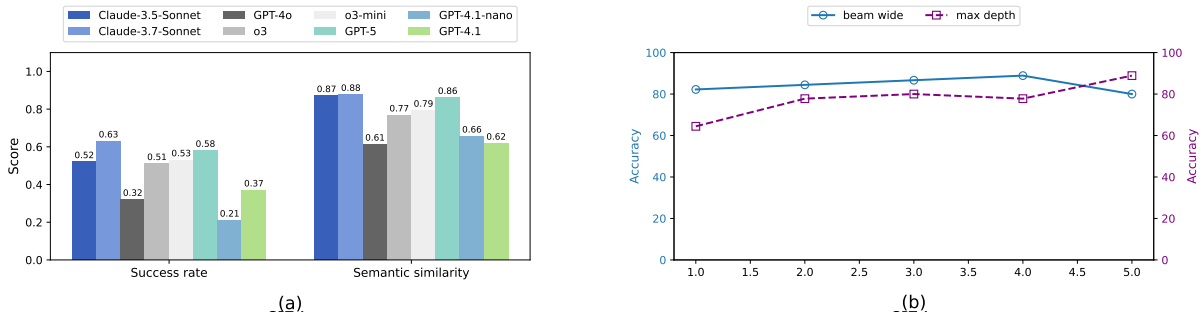

Figure 3: Overview of additional experiments. (a) Evaluation of screening quality in open-ended scenarios, measured by semantic similarity and success rate, with comparative results for cyclodextrin-based material discovery. (b) Comparative analysis of beam search performance under different hyperparameter configurations.

5.4 Additional Experiments

In Fig. 3(a), we observe substantial variation in material screening performance across different models. In particular, both success rate and semantic similarity indicate that GPT-5 and Claude-3.7-Sonnet consistently achieve higher scores, suggesting that these models are more reliable in generating physically consistent candidate structures. By contrast, GPT-4.1 and GPT-4o exhibit lower performance, especially in semantic similarity, reflecting their limited ability to capture chemically meaningful structures in an open search space. Fig. 4 presents the BLEU scores of various LLM backbones in a radar chart. The results show that o3-mini achieves the highest overall BLEU score, indicating a stronger capability to produce outputs that align with chemically plausible structures, whereas o3 and GPT-4.1-nano perform notably weaker. Taken together, these findings underscore the complementary role of multiple evaluation metrics in assessing model performance and highlight their importance in ensuring robust and physically grounded material generation. On the other hand, Fig. 3(b) analyzes the impact of beam search methods. Increasing beam width generally leads to steady improvements in accuracy, demonstrating the benefit of exploring broader candidate sets. Conversely, deeper search settings improve accuracy up to a point but show diminishing returns, with performance plateauing beyond a certain depth.

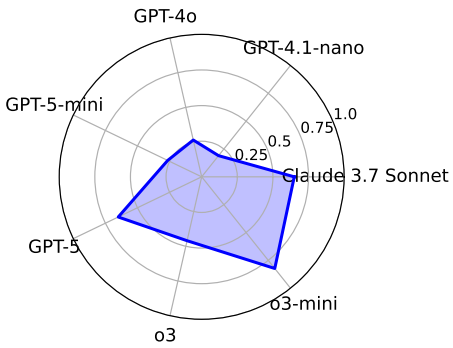

Figure 4: Radar charts summarizing BLEU scores for different LLMs backbone on material screening-CD task.

6 Conclusion

In this work, we present SmartMat Explorer, an agentic framework that integrates literature mining, hypothesis generation, and material screening. Applied to the design of host–guest functional materials, it autonomously identifies chemically efficient and structurally compatible supramolecular systems. SmartMat Explorer demonstrates the ability to design promising materials, execute complex reasoning, and validate outcomes through rigorous evaluation, marking a step toward fully automated scientific discovery. These results highlight its potential for material innovation and inspiration mining. Future work will extend its applicability across broader scientific domains and incorporate diverse objectives into the optimization process.

REFERENCES

Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation*, 15(3):1652–1671, 2019.

Adib Bazgir, Yuwen Zhang, et al. Matagent: A human-in-the-loop multi-agent llm framework for accelerating the material science discovery cycle. In *AI for Accelerated Materials Design-ICLR 2025*, 2025.

Frank Biedermann and Hans-Jörg Schneider. Experimental binding energies in supramolecular complexes. *Chemical reviews*, 116(9):5216–5300, 2016.

Jaime G. Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. *SIGIR Forum*, 51(2):209–210, 2017. doi: 10.1145/3130348.3130369. URL `https://doi.org/10.1145/3130348.3130369`.

Huajun Chen, Keyan Ding, Jing Yu, Junjie Huang, Yuchen Yang, and Qiang Zhang. Scitoolagent: A knowledge graph-driven scientific agent for multi-tool integration. 2025.

Alireza Ghafarollahi and Markus J Buehler. Atomagents: Alloy design and discovery through physics-aware multi-modal multi-agent artificial intelligence. *arXiv preprint arXiv:2407.10022*, 2024.

Alireza Ghafarollahi and Markus J Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025.

Valentin Goussard, François Duprat, Vincent Gerbaud, Jean-Luc Ploix, Gérard Dreyfus, Véronique Nardello-Rataj, and Jean-Marie Aubry. Predicting the surface tension of liquids: Comparison of four modeling approaches and application to cosmetic oils. *Journal of chemical information and modeling*, 57(12):2986–2995, 2017.

Taicheng Guo, Kehan Guo, Zhengwen Liang, Zhichun Guo, Nitesh V Chawla, Olaf Wiest, Xiangliang Zhang, et al. What indeed can gpt models do in chemistry? a comprehensive benchmark on eight tasks. *arXiv preprint arXiv:2305.18365*, 3, 2023.

Zijian Hong. Chatgpt for computational materials science: a perspective. *Energy Material Advances*, 4:0026, 2023.

Zhaolin Hu, Yixiao Zhou, Zhongan Wang, Xin Li, Weimin Yang, Hehe Fan, and Yi Yang. Osda agent: Leveraging large language models for de novo design of organic structure directing agents. In *The Thirteenth International Conference on Learning Representations*, 2025.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, et al. Biomni: A general-purpose biomedical ai agent. *biorxiv*, pp. 2025–05, 2025.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Mlagentbench: Evaluating language agents on machine learning experimentation, 2024. *URL https://arxiv. org/abs/2310*, 3302.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Mlagentbench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*, 2023.

Theo Jaffrelot Inizan, Sherry Yang, Aaron Kaplan, Yen-hsu Lin, Jian Yin, Saber Mirzaei, Mona Abdelgaid, Ali H Alawadhi, KwangHwan Cho, Zhiling Zheng, et al. System of agentic ai for the discovery of metal-organic frameworks. *arXiv preprint arXiv:2504.14110*, 2025.

Jacob N Israelachvili. *Intermolecular and surface forces*. Academic press, 2011.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*, 2024.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Guotai Li, Fanny Trausel, Michelle P van Der Helm, Benjamin Klemm, Tobias G Brevé, Susan AP van Rossum, Muhamad Hartono, Harm HPJ Gerlings, Matija Lovrak, Jan H van Esch, et al. Tuneable control of organocatalytic activity through host–guest chemistry. *Angewandte Chemie International Edition*, 60(25):14022–14029, 2021.

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X Zhang. Llms as research tools: A large scale survey of researchers' usage and perceptions. *arXiv preprint arXiv:2411.05025*, 2024.

Jia-ning Liu, Kui Du, Jia-hao Guo, Dan Wang, Cheng-bin Gong, and Qian Tang. Visual sensor with host–guest specific recognition and light–electrical co-controlled switch. *Small*, 20(31):2311823, 2024.

Wanhao Liu, Zonglin Yang, Jue Wang, Lidong Bing, Di Zhang, Dongzhan Zhou, Yuqiang Li, Houqiang Li, Erik Cambria, and Wanli Ouyang. Moose-chem3: Toward experiment-guided hypothesis ranking via simulated experimental feedback. *arXiv preprint arXiv:2505.17873*, 2025.

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

Elio Mattia and Sijbren Otto. Supramolecular systems chemistry. *Nature nanotechnology*, 10(2):111–119, 2015.

Ziqi Ni, Yahao Li, Kaijia Hu, Kunyuan Han, Ming Xu, Xingyu Chen, Fengqi Liu, Yicong Ye, and Shuxin Bai. Matpilot: an llm-enabled ai materials scientist under the framework of human-machine collaboration. 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*, 2025.

Yusuf Roohani, Jian Vora, Qian Huang, Zachary Steinhart, Alexander Marson, Percy Liang, and Jure Leskovec. Biodiscoveryagent: An ai agent for designing genetic perturbation experiments, 2024. *URL https://arxiv.org/abs/2405.17631*.

Aramballi J Savyasachi, Oxana Kotova, Sankarasekaran Shanmugaraju, Samuel J Bradberry, Gearóid M Ó'Máille, and Thorfinnur Gunnlaugsson. Supramolecular chemistry: a toolkit for soft functional materials and organic particles. *Chem*, 3(5):764–811, 2017.

Henry W Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. Chemreasoner: Heuristic search over a large language model's knowledge space using quantum-chemical feedback. *arXiv preprint arXiv:2402.10980*, 2024.

Qiushi Sun, Zhoumianze Liu, Chang Ma, Zichen Ding, Fangzhi Xu, Zhangyue Yin, Haiteng Zhao, Zhenyu Wu, Kanzhi Cheng, Zhaoyang Liu, et al. Scienceboard: Evaluating multimodal autonomous agents in realistic scientific workflows. *arXiv preprint arXiv:2505.19897*, 2025.

Izumi Takahara, Teruyasu Mizoguchi, and Bang Liu. Accelerated inorganic materials design with generative ai agents. *arXiv preprint arXiv:2504.00741*, 2025.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. Ai-researcher: Autonomous scientific innovation. *arXiv preprint arXiv:2505.18705*, 2025a.

Xiangru Tang, Tianyu Hu, Muyang Ye, Yanjun Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, et al. Chemagent: Self-updating library in large language models improves chemical reasoning. *arXiv preprint arXiv:2501.06590*, 2025b.

Hanchen Wang, Yichun He, Paula P Coelho, Matthew Bucci, Abbas Nazir, Bob Chen, Linh Trinh, Serena Zhang, Kexin Huang, Vineethkrishna Chandrasekar, et al. Spatialagent: An autonomous ai agent for spatial biology. biorxiv, 2025.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 279–299, 2024a.

Wenxiao Wang, Lihui Gu, Liye Zhang, Yunxiang Luo, Yi Dai, Chen Shen, Liang Xie, Binbin Lin, Xiaofei He, and Jieping Ye. Scipip: An llm-based scientific paper idea proposer. *arXiv preprint arXiv:2410.23166*, 2024b.

Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, et al. From ai for science to agentic science: A survey on autonomous scientific discovery. *arXiv preprint arXiv:2508.14111*, 2025.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. *arXiv preprint arXiv:2410.07076*, 2024.

Murong Yue. A survey of large language model agents for question answering. *arXiv preprint arXiv:2503.19213*, 2025.

Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. Honeycomb: A flexible llm-based agent system for materials science. *arXiv preprint arXiv:2409.00135*, 2024.

Jinglan Zhang, Xinyi Chen, Xu Ye, Yulin Yang, and Bin Ai. Large language model in materials science: Roles, challenges, and strategic outlook. *Advanced Intelligent Discovery*, pp. 202500085, 2025.

Bingxi Zhao, Lin Geng Foo, Ping Hu, Christian Theobalt, Hossein Rahmani, and Jun Liu. Llm-based agentic reasoning frameworks: A survey from methods to scenarios. *arXiv preprint arXiv:2508.17692*, 2025.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, pp. 1–11, 2025.

Jiawei Zhou, Yamin Li, Lutao Wang, Pin Lv, Miao Chen, Feijian Xiao, Tian Si, Jun Tao, and Bo Yang. Bifunctional drug delivery system with carbonic anhydrase ix targeting and glutathione-responsivity driven by host-guest amphiphiles for effective tumor therapy. *Carbohydrate Polymers*, 326:121577, 2024.

Lianhao Zhou, Hongyi Ling, Keqiang Yan, Kaiji Zhao, Xiaoning Qian, Raymundo Arróyave, Xiaofeng Qian, and Shuiwang Ji. Toward greater autonomy in materials discovery agents: Unifying planning, physics, and scientists. *arXiv preprint arXiv:2506.05616*, 2025.

A DETAILS OF SCIENTIFIC TOOLS

We integrate a set of domain-specific tools to support molecule design, property estimation, and knowledge retrieval. These tools ensure that our system produces scientifically reliable outputs rather than relying solely on language generation. Below, we provide details of the key categories of tools employed in our framework.

- **RDKit tools.** We use RDKit[1] for molecular graph construction, SMILES parsing, conformer generation, and property calculation (e.g., molecular weight, LogP, Complexity, Aromaticity). RDKit provides a reliable cheminformatics backbone for structure-based tasks.
- **Chemical tools.** A collection of specialized computational chemistry utilities are integrated, including molecular docking software, Gaussian input file generators, and property prediction modules. These tools facilitate preliminary validation of designed molecules before downstream evaluation. **SCScore**,
- **Web search.** General-purpose web APIs (i.e., SerpAPI, Google Search) are employed to retrieve up-to-date information, material databases, and scientific resources. They enable dynamic augmentation of the knowledge base, ensuring access to real-world, non-static data.
- **Literature search.** We integrate multiple literature retrieval services (i.e., CrossRef, Semantic Scholar, PubMed, Web of Science) to gather abstracts and metadata based on task-specific queries. These tools are essential for grounding model responses in verified scientific evidence and for enabling Retrieval-Augmented Generation pipelines.

B DETAILS OF EVALUATION METRICS

Here, we provides the detailed definitions, formulations, and implementation details of the evaluation metrics used in this work. Specifically, we describe the computation of *Accuracy*, *Success Rate*, *Semantic Similarity*, and *BLEU*, ensuring reproducibility and clarity of the evaluation protocol.

Accuracy. We adopt exact match accuracy, defined as the proportion of predictions that exactly match the ground-truth labels. Given N samples,

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i), \tag{8}$$

where $\mathbf{1}(\cdot)$ is an indicator function, $\hat{y}_i$ is the model prediction answer, and y_i is the standard answer. Accuracy is used for closed-ended, uniquely correct answer multi-choice question.

Success Rate. Success Rate measures the proportion of tasks that are completed successfully according to predefined correctness criteria. Formally, it is defined as:

$$\text{Success Rate} = \frac{1}{M}\sum_{i=1}^{M} \big(\mathcal{C}(\hat{y}_i, y_i) = 1\big), \tag{9}$$

where $\mathcal{C}(\hat{y}_i, y_i)$ is a task-specific boolean function that returns 1 if the predicted output $\hat{Y}_i$ satisfies the correctness requirements with respect to the reference Y_i, and 0 otherwise. Success rate is used for open short answer question evaluations.

Semantic Similarity. To capture cases where predicted molecules are semantically correct but not lexically identical to molecule in the references, we compute semantic similarity using a pre-trained sentence encoder $\phi(\cdot)$ (specifically, `all-MiniLM-L6-v2`). Further, both the predicted answer $\hat{y}$ and the reference answer y are mapped into a shared embedding space, and their similarity is measured by cosine similarity:

$$\text{Sim}(\hat{y}, y) = \frac{\langle \phi(\hat{y}), \phi(y) \rangle}{\|\phi(\hat{y})\| \cdot \|\phi(y)\|}, \tag{10}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, and $\| \cdot \|$ is the Euclidean norm. The final semantic similarity score is obtained by averaging $\text{Sim}(\hat{y}_i, y_i)$ over all evaluation instances:

$$\text{Semantic Similarity} = \frac{1}{N}\sum_{i=1}^{N} \text{Sim}(\hat{y}_i, y_i). \tag{11}$$

[1] `https://www.rdkit.org/`

BLEU. To evaluate surface-level fluency and lexical overlap between predicted and reference answers, we adopt the BLEU metric (Papineni et al., 2002). In our setting, predictions are typically short text (i.e., molecular names), for which a simplified BLEU formulation suffices. Given a reference string y and a candidate string $\hat{y}$, both are tokenized into word sequences. The BLEU score is then computed as the product of a brevity penalty (BP) and the modified n-gram precision:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{12}$$

where p_n denotes the n-gram precision, w_n is a weighting factor, and BP penalizes overly short predictions. In the special case of single-token answers, BLEU reduces to unigram precision with BP, and reaches 1.0 if and only if the predicted token exactly matches the reference.

C BENCHMARK CONSTRUCTION

C.1 THE COLLECTION OF MATLITQA DATASETS

We searched academic database Web of Science (WOS) for the keywords supramolecular materials and functional materials, and found 3,620 papers published in top journals since 2020. We have collected relevant information such as titles, abstracts, keywords, etc. on WOS. Subsequently, we employ GPT-4.1-nano to automatically classify the application areas of the paper to ensure appropriate difficulty. We select 75 papers from them and downloaded the relevant pdf documents for knowledge extraction, which were reviewed by 2 Ph.D. majoring in chemical materials. Inspired by LitQA2 (Laurent et al., 2024), our multiple-choice questions cannot be retrieved directly from the abstract, but relevant information can be obtained from the main text. This filtering process yielded a final dataset containing 84 multiple-choice questions in 5 application scenarios.

Table 4: Distribution of categories.

Category	Count
Chemical Synthesis	26
Materials Science	11
Drug Delivery	12
Sensor Design	13
Catalyst Design	13
Total	75

C.2 PROMPT FOR THE EXTRACTION OF CATEGORY CLASSIFICATION.

Prompt for the classification

<system_role>
You are a helpful assistant that categorizes research papers into multiple application domains.

<user_prompt>
You are an expert in categorizing academic research papers by their application domain. Your task is to analyze the provided title, abstract, and keywords of a research paper and identify ALL relevant application domains from the following list: $\{',' .join(sampledomains)\}$
If a paper fits into multiple domains, list all that apply. If the paper does not fit clearly into any of these domains, or if the provided information is insufficient, please classify it as "Unclassified".
Provide your answer as a comma-separated list of domain names. For example: "Sensor, Chemistry" or "Unclassified". Do not include any explanations or additional text.
Here is the paper information: {input text}
Application Domains:

C.3 PROMPTING FOR MULTI-TURN JUDGMENT AND EVALUATION.

To ensure that the Judgment Agent can remember previous interactions and task-specific requirements across multiple dialogue turns, we incorporate both the interaction history and the relevant task context into the prompt. Specifically, the interaction history records all prior queries, tool calls, and intermediate responses, providing a chronological trace of the agent's reasoning process. The task context includes the original user request, reference

information on candidate hosts, and the defined evaluation criteria. By combining these components, the Judgment Agent can:

- Accurately verify whether the proposed host satisfies all specified conditions, including physicochemical properties such as pH compatibility, solvent environment, target binding, molecular complexity, aromaticity, and MolLogP.
- Compare the proposed host with alternative candidates in the list, identifying potential improvements or suboptimal selections based on computed properties and task objectives.
- Recommend corrective actions, providing a new optimal host if the original selection is insufficient, or confirming the current choice when it meets all criteria.

This structured prompt design enables the Judgment Agent to perform rigorous, context-aware evaluation, maintain consistency across multiple dialogue turns, and provide interpretable, actionable recommendations for host selection. Examples are as follows:

Prompt for the judgment

<system_role>
You are an expert AI assistant specializing in materials chemistry. You can use tools to gather information and critically evaluate choices.

<user_prompt>
Critically evaluate the following task and the AI's proposed solution. Your goal is to ensure the selected host molecule rigorously meets all specified conditions.

Original User Request Context:{prompt}
Full Interaction History:{interaction history}
Optimal Host Selection:{agent final answer}
Reference Information: {host summary}

Evaluation Criteria and Reflection Tasks:
1. Verification of Conditions:
- Does the proposed host {final answer} satisfy all stated conditions in the original user request and the provided RAG context (e.g., pH, Solvent, Target, δ G, Complexity Host max, MolLogP target)?
- Specifically, check the computed Aromaticity, MolLogP, and Complexity of the proposed host against the requirements. (The tool calls in the history provide these values for the candidate hosts).
2. Comparison with Alternatives:
- Were there other hosts in the provided candidate list that might have been a better fit based on the computed properties and the overall objective (e.g., hydrophilicity enhancement, size matching)?
- If the current selection is suboptimal, identify the specific reasons why.
3. Corrective Action:
- If the proposed host {final answer} is NOT the optimal choice, please provide a NEW, improved optimal host from the candidate list.
- Clearly explain why this new choice is superior, referencing the conditions and computed properties.
- If the proposed host is optimal, confirm this decision and state that no changes are needed.

Output Format:
Please structure your response as follows:
- Evaluation: [Your assessment of the proposed host's suitability, e.g., "Satisfactory", "Suboptimal", "Incorrect"]
- Reasoning: [Detailed explanation for your evaluation, referencing conditions and properties]
- Corrected Optimal Host: $\langle answer \rangle (.*?) \langle /answer \rangle$ [Only the new and best host name from the *candidate list* should be here, or if satisfied, only repeat the original host name]

Table 5: Examples of benchmark metadata for literature comprehension.

Sub-Category	Metadata
Chemical Synthesis	**Question**: What determines the type of actuation mode shown by UC-1@PVDF films? **Choice**: A) Cage content distribution B) Hydrogen bonding strength C) Thickness of the film D) Polarity and steric hindrance of solvents
Materials Science	**Question**: What is the primary limitation preventing stimuli-responsive materials from being interactive? **Choice**: A) Lack of orthogonality under physiological conditions B) They can only switch between states A and B C) Insufficient self-regulating properties D) Low thermal stability
Drug Delivery	**Question**: What mechanism is primarily responsible for the drug release in the nanogels under UV light? **Choice**: A) Thermal degradation of nanogels B) Reduction by enzymes C) pH-sensitive swelling of the core D) cis-trans isomerization of azobenzene molecules
Sensor Design	**Question**: What is the role of FcDMA in the LECCVS system? **Choice**: A) Optical switching B) Electrochemical intermediary C) Structural support D) Photosensitization
Catalyst Design	**Question**: Which of the following are the advantages that the host-guest coordination structure may bring? **Choice**: A) Improving the selectivity of catalytic sites B) Reduce the reaction energy barrier C) Enhance anti-toxicity D) improve the conductivity of metals

C.3.1 THE EXTRACTION OF MULTIPLE-CHOICE QUESTION.

Prompt for the extraction of QA

<system_role>
You are a helpful assistant that categorizes research papers into multiple application domains.

<user_prompt>
You are a professional research assistant, specializing in extracting structured knowledge from biochemical ring materials papers. Please extract key scientific research information based on the following paper content and return it to JSON format:
Thesis content:

`{paper text}`

Please output the following:

1. **Research Questions and Motivation** (`research_problem`)
2. **Methods and technical routes** (`methods`)
3. **Experiment settings and data situations** (`experiments`)
4. **Key Findings and Conclusions** (`key_findings`)
5. **Limitations and future work directions** (`future_work`)
6. **Named entity recognition**: material, method, model noun
7. **Extract 3-5 Q&A pairs that require inference, and cannot get the answer directly from the summary or title** (`qa_pairs` needs to give options, answer can only be one of A/B/C/D)

The output format is as follows:

```
{
  "title": "",
  "authors": [],
  "abstract": "This is the complete summary of the paper, copied and pasted
      from the original text, without any rewriting or summary.",
  "research_problem": "",
  "methods": [],
  "experiments": "",
  "key_findings": "",
  "limitations": "",
  "future_work": "",
  "entities": {
    "materials": [],
    "methods": [],
    "applications": []
  },
  "qa_pairs": [
    {"question": "", "answer": ""},
    ...
  ],
  "class": "chemistry"
}
```

If the information is missing, please fill in ”N/A”.

C.4 THE COLLECTION OF SCREENINGQA DATASETS

Database resources. The databases used in this work include: PubChem[2], OpenCyclodb[3], surface tension dataset Goussard et al. (2017) and data from book *Intermolecular and surface forces* (Israelachvili, 2011).

The cyclodextrin dataset is derived from OpenCycloDB, with each host–guest complex annotated with binding energy, TPSA, MolLogP, and molecular complexity. To assess the molecular screening and inference capabilities of our system, we construct a ScreeningQA-CD benchmark where each query includes the correct complex along

[2]https://pubchem.ncbi.nlm.nih.gov
[3]http://opencyclodb.org

with three structurally similar distractor molecules. Furthermore, to evaluate the system's performance in open-ended scenarios, we design an additional set of short answer questions that test its ability to identify suitable candidates beyond the provided dataset. Demo example can be find in Table 5.

The surface tension(ST) dataset used in our benchmark is derived from molecular dynamics simulation calculation reported in the literature (Goussard et al., 2017). The original dataset[4] compiles surface tension values for a wide range of organic cosmetic oils measured by molecular dynamic simulations.

To build the ScreeningQA-ST benchmark, we collated and reformatted the original data into a structured QA dataset, an example of which is shown in Table 6. Each sample is represented as a query that links molecular identity to its corresponding surface tension value. The resulting dataset contains a total of 262 compounds composed of carbon, oxygen, hydrogen, or silicon atoms and spanning multiple functional groups, thereby covering a diverse chemical space.

This QA-style transformation enables systematic evaluation of a model's capability to retrieve property-relevant knowledge, reason over structure–property relationships, and predict quantitative physicochemical values. By grounding the benchmark in experimentally measured data, ScreeningQA provides a reliable and interpretable testbed for assessing the performance of language agents in property prediction tasks.

Table 6: Examples of benchmark for materials screening.

Sub-Category	Metadata
CD	**Instruction**: Select the optimal host molecule for trans-anethole drug delivery based on given conditions **Condition:** pH: 7.4, Solvent: H_2O **Answer**: β-Cyclodextrin
ST	**Instruction:** Given the molecule, write its surface tension at normal measurement conditions. **Input:** Pentane **Output:** 15.5

Multi-Choice Prompt

<system_role>
You are an expert AI assistant specializing in material chemistry. You can use tools to gather information and critically evaluate choices.

<user_prompt>
Given the list of application conditions:
Guest molecule: sample['IsomericSMILES']
pH: sample['pH']
Solvent: H_2O
Target: In order to enhance the targeted delivery ability of the guest molecule and increase the hydrophilicity by using the inclusion, please select the most suitable host molecule according to the size of the molecular cavity, hydrophilicity, etc
Host list for selection: {", ".join(host_list)}
formatted in tags as follows: <answer >Your answer here </answer >

D Additional Experiments

Case Study. Fig. 5 illustrates several cyclodextrin host–guest complexes predicted by our SmartMat Explorer through a tool-augmented workflow. Compared with previous approaches, our framework integrates code generation to assist SMILES-based molecular design, employs molecular descriptors for heuristic reasoning and rapid validation, and leverages a continuously updated web search engine. This enables the agentic system to achieve multi-level matching between candidate and known molecules while exhibiting high structural compatibility and synthetic feasibility.

As illustrated in Fig. 5, more complex molecules do not necessarily exhibit higher binding energies. For example, the third candidate in the first row forms more stable complexes with the guest molecule despite its higher structural complexity and smaller pore size, which ultimately reduces the stability of certain host–guest interactions. In

[4] https://pubs.acs.org/doi/suppl/10.1021/acs.jcim.7b00512/suppl_file/ci7b00512_si_001.pdf

Table 7: Agent performance improvement in literature comprehension tasks. Results show the accuracy and standard variance (in bracket) averaged over 5 runs(due to the computational cost, the deep research model only run the experiment only once). The best model is **bolded**, the second best is underlined.

Model	Catalysis	Drug	Sensor	Chemistry	Materials	Average
GPT-4.1-nano	84.62 (± 0.6)	84.00 (± 1.4)	80.00 (± 0.7)	77.03 (± 0.1)	**93.33**(± 1.1)	83.80
GPT-5-nano	84.62 (± 0.0)	84.00 (± 0.2)	80.00 (± 0.6)	75.68 (± 1.0)	84.44 (± 1.2)	81.75
GPT-5-mini	78.85 (± 0.9)	86.00 (± 0.0)	82.00 (± 0.0)	79.73 (± 0.8)	82.22 (± 1.2)	81.76
GPT-5	**92.31**(± 1.2)	86.00 (± 0.2)	80.00 (± 0.5)	**87.84**(± 0.2)	86.67 (± 2.5)	86.56
GPT-4o-mini	80.77 (± 0.2)	74.00 (± 0.8)	82.00 (± 1.0)	72.97 (± 1.4)	73.33 (± 2.0)	76.61
GPT-4o	84.62 (± 2.2)	80.00 (± 2.0)	82.00 (± 0.0)	74.32 (± 1.0)	84.44 (± 0.7)	81.08
o3-mini	84.62 (± 0.6)	88.00(± 1.4)	82.00 (± 0.2)	82.43 (± 1.3)	84.44 (± 0.8)	84.30
o3	88.46(± 1.2)	86.00 (± 1.6)	82.00 (± 0.0)	86.49(± 1.0)	88.89(± 1.2)	86.37
o4-mini	88.46(± 1.0)	88.00(± 2.0)	**88.00**(± 0.8)	83.78 (± 1.4)	86.67 (± 0.6)	**86.98**
Claude 3.5 Sonnet	82.69 (± 0.5)	74.00 (± 1.0)	72.00 (± 0.2)	75.68 (± 1.3)	84.44 (± 1.1)	77.76
Claude 3.7 Sonnet	86.54 (± 1.9)	**90.00**(± 1.5)	86.00(± 2.0)	85.14 (± 0.6)	86.67 (± 1.3)	86.87
o3 Deep Research	80.77	86.00	82.00	74.32	91.11	82.84

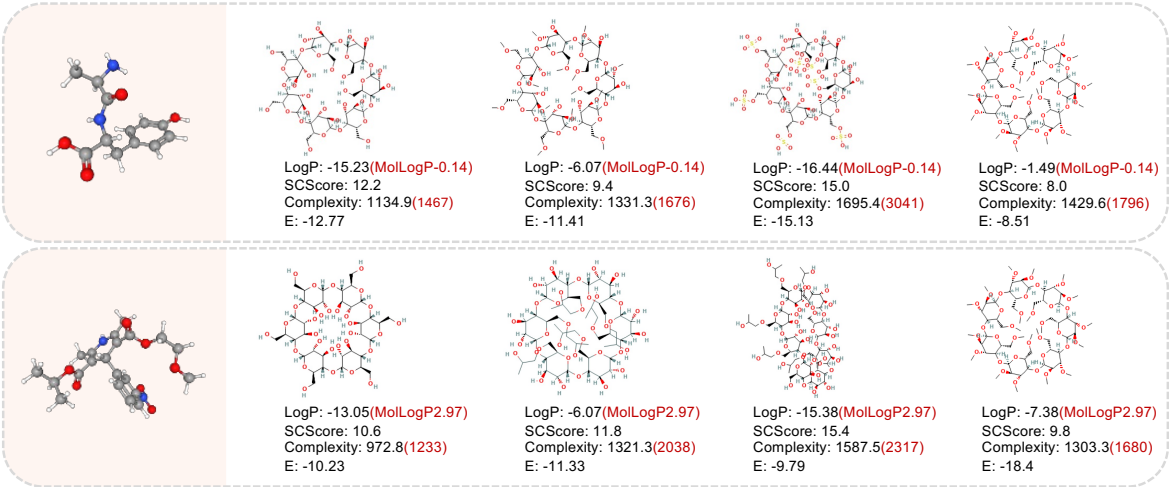

Figure 5: Case study: Illustrated are two examples of a material screening task to select the appropriate host molecule for the guest molecule for inclusion design. The first in each row is the hose molecule, and the four on the right are the candidate guest molecules. The red value is the data for the source dataset, and the black value is the result of the calculation of the scientific tool.

contrast, the fourth host in the second row is less complex and exhibits lower binding energy, making it easier to synthesize. Consequently, judging molecules based on a single attribute is insufficient, and computing combined energies is often computationally expensive. Instead, heuristic reasoning using multifaceted molecular descriptors allows rapid screening, enabling the agent system to identify optimal candidates and perform more sophisticated scientific reasoning. In practice, different calculation methods will yield calculation results with different precisions, but the differences between molecules do not change with the calculation accuracy, so simple descriptors are helpful for heuristic screening.

It is worth noting that our agentic framework does not incorporate scientific computing tools merely for molecular property prediction. Instead, it utilizes comparative analysis of fundamental molecular properties for rapid material screening and feasibility assessment. This demonstrates how large language models can empower inspiration-driven discovery, bridging critical information gaps in scientific research and enabling more feasible pathways toward automated scientific discovery.

E LIMITATIONS

Although this work achieves state-of-the-art performance in material screening and literature comprehension by integrating scientific computing tools with causally-grounded RAG, several limitations remain. First, the selection and orchestration of external tools have not been systematically optimized, potentially constraining the

framework's overall efficiency and robustness. Second, the LLM's interpretation of computational outputs may still suffer from hallucinations, posing risks to reliability in downstream decision-making. Finally, our study primarily targets materials representable by the SMILES format, while many real-world systems require richer structural descriptors to capture their intrinsic properties. Extending the framework to support multiple representation modalities is therefore an important direction for future work.

F USE OF LARGE LANGUAGE MODELS

We disclose using LLMs solely to aid and polish writing enhancing academic expression accuracy, argument coherence, and text logic. All core ideas, research designs experiments, and conclusions are independently developed by the authors.