

MOOSE-Copilot: A Web-Based Interactive Assistant for Unified Exploratory and Fine-Grained Scientific Hypothesis Discovery

Hongran An

Central Conservatory of Music
hongran761@gmail.com

Zonglin Yang[†]

Nanyang Technological University
zonglin001@ntu.edu.sg

Abstract

Large language models (LLMs) show remarkable potential in scientific hypothesis discovery. However, existing approaches face two critical limitations: they treat divergent exploratory search and convergent fine-grained refinement as isolated tasks, and they operate autonomously with little to no human guidance. We present MOOSE-Copilot, the first unified framework to bridge this abstraction gap through a formalized human-AI interaction (HAIL) protocol. Our system empowers scientists to steer the generative process via three explicit signals: initial blueprints, inter-stage routing, and intra-stage feedback. Using an oracle-simulated evaluation in which an LLM provides idealized expert signals, we show that injecting these structured signals significantly outperforms purely autonomous baselines, characterizing the gains achievable under high-quality guidance. Furthermore, we build a web-based interface that turns the framework into a no-code workflow: researchers pose a question, watch the hypothesis search unfold as an interactive tree, and steer it by selecting hypotheses, routing between stages, and injecting feedback—no command-line agents required. This makes end-to-end hypothesis discovery directly accessible to interdisciplinary researchers. Website: <https://moosedemo.com> Video: www.youtube.com/watch?v=_GSa-42ArIA

1 Introduction

Recently, large language models (LLMs) have demonstrated remarkable potential in assisting various stages of scientific research, including hypothesis generation, experiment design and execution, manuscript writing, and peer review (Luo et al., 2025). Among these, scientific hypothesis discovery lies at the heart of the research process, as it plays a critical role in shaping research direction

and influencing the potential significance and impact of the resulting work.

We conceptualize the landscape of automated scientific hypothesis discovery as encompassing two distinct paradigms: exploratory and fine-grained discovery. Exploratory discovery focuses on broadly diverging to generate diverse, high-level research directions from a given background context. However, this often yields rough, underspecified ideas. Most existing agentic systems for scientific discovery operate within this exploratory regime (Yang et al., 2024; Zhou et al., 2024; Li et al., 2024; Shojaee et al., 2024; Hu et al., 2024; Pu et al., 2024). Among these, MOOSE-Chem (Yang et al., 2025b) is the first to formalize exploratory discovery as an explicit search problem.

In contrast, fine-grained discovery—recently introduced by MOOSE-Chem2 (Yang et al., 2025a)—operates convergently. It builds upon an initial conceptual blueprint, leveraging LLMs to optimize plausibility and coherence while systematically enriching the idea with methodological and experimental details. This process ultimately translates an abstract inspiration into a concrete, actionable research hypothesis.

We observe that these two tasks correspond to two complementary stages of the scientific discovery process: exploration and exploitation. Exploratory discovery enables the broad search for promising directions, while fine-grained discovery deepens and refines those directions into actionable research hypotheses.

Despite their complementarity, two key gaps remain. First, it is entirely unexplored how these two stages might be synergistically combined within a unified framework. Unifying them is inherently challenging due to the abstraction gap between divergent exploratory search and convergent fine-grained refinement, yet this integration is an essential step toward a unified framework of automated scientific discovery. Second, current approaches

[†]Corresponding author.

to both tasks typically rely on standalone LLM workflows or agentic methods, offering little to no support for human input or control. Consequently, a major bottleneck persists: we lack a cohesive system that successfully bridges the exploratory and fine-grained phases while simultaneously empowering scientists to actively steer, course-correct, and validate this end-to-end automated process.

To overcome this bottleneck, we introduce a principled human–AI interaction (HAI) paradigm that formally bridges exploratory and fine-grained hypothesis discovery. We instantiate this paradigm through MOOSE-Copilot, a unified framework that keeps the human scientist in control across the two stages of exploration and exploitation. Rather than treating human-in-the-loop as a mere feature, we define a structured interaction protocol where scientists inject initial blueprints, make inter-stage routing decisions (when to move between the two stages), and provide intra-stage feedback (how to steer within a stage). This synergy mitigates the search space explosion and resolves the ambiguity of when to diverge versus when to refine, transforming an intractable automated process into a manageable, human-guided workflow.

To validate this unified framework, we systematically evaluate the impact of the core guiding signals—initial blueprints, inter-stage routing, and intra-stage feedback—on the quality of the final hypotheses. Using an oracle-simulated evaluation in which an LLM provides idealized expert signals, our quantitative analyses show that structured intervention significantly outperforms purely autonomous baselines, offering insights into how scientists can most effectively guide the system through these signals.

Furthermore, a theoretical interaction protocol is only impactful if it is accessible to its target audience. Recognizing that many domain experts—particularly in fields like chemistry and biology—are hindered by the barrier of command-line agentic tools, we operationalize our HAI paradigm through a comprehensive web-based graphical user interface (GUI).

Designed for usability and transparency, *MOOSE-Copilot* visualizes the hypothesis search trajectory as an interactive tree. Through this tree, researchers trace the evolution of generated ideas and steer the search in either direction—drilling a promising branch down into fine-grained exploitation, or pivoting a refined hypothesis back to explore new directions—while injecting domain

knowledge wherever needed, grounding the automated discovery process in human expertise.

In conclusion, this paper makes the following four contributions:

- We propose the first work that unifies exploratory search and fine-grained exploitation in scientific hypothesis discovery.
- We formalize a principled human–AI interaction protocol that lets scientists steer the automated process through initial blueprints, inter-stage routing, and intra-stage feedback.
- We systematically evaluate the impact of these guiding signals under an oracle-simulated setting (idealized interventions standing in for expert input), showing that structured guidance significantly outperforms purely autonomous baselines.
- We operationalize this paradigm into an accessible web-based system featuring interactive tree visualization to ensure transparency and seamless human-in-the-loop control.

2 Methodology

2.1 Preliminaries

2.1.1 Exploratory Hypothesis Discovery

Generating a scientific hypothesis h from a research background b is often intractable. MOOSE-Chem (Yang et al., 2025b) addresses this through a fundamental assumption: that a majority of hypotheses can be composed from a research background b and a set of inspirations i_1, \dots, i_k retrieved from the literature,

$$h = f(b, i_1, \dots, i_k), \quad (1)$$

where each inspiration $i \in I$ is a piece of knowledge not previously known to be related to b , and I denotes the literature corpus serving as the inspiration space. Under this assumption, the intractable distribution $P(h | b)$ can be decomposed into a structured sequence of tractable sub-operations, transforming exploratory discovery into a search problem (with $h_0 = \emptyset$):

$$P(h | b) \approx \prod_{j=1}^k P(i_j | b, h_{j-1}, I) \cdot P(h_j | b, h_{j-1}, i_j), \quad (2)$$

This decomposition naturally defines a Markov Decision Process, in which the current hypothesis

h_{j-1} and the selected inspiration i_j form a state-action pair: $P(i_j | b, h_{j-1}, I)$ is the policy that retrieves the next inspiration, and $P(h_j | b, h_{j-1}, i_j)$ is the state transition that composes it into an updated hypothesis,

$$b \xrightarrow{i_1} h_1 \xrightarrow{i_2} h_2 \xrightarrow{\dots} h_{k-1} \xrightarrow{i_k} h_k = h. \quad (3)$$

In practice, an LLM-based agent realizes this process iteratively: at each step it retrieves several candidate inspirations $i_j \in I$ from the current hypothesis, and each retrieved inspiration is composed into a new hypothesis. A single step therefore branches into multiple successors, and repeating it across steps expands the hypotheses into a tree, where each node is an intermediate hypothesis and each root-to-node path is a distinct sequence of inspiration-driven updates. To keep this growth manageable, the agent applies beam search, using a ranking function $R(h)$ to retain only the top-scoring hypotheses at each step. We adopt MOOSE-Chem as the exploratory discovery stage of our system; accordingly, the MOOSE-Copilot interface exposes the resulting hypothesis tree and ranking for navigation (Section 3).

2.1.2 Fine-Grained Hypothesis Discovery

MOOSE-Chem2 (Yang et al., 2025a) introduced the task of fine-grained hypothesis discovery, which refines a coarse-grained hypothesis into a fully specified one to maximize its scientific plausibility and quality. The exploratory stage (Section 2.1.1) produces a hypothesis h_j that is, in principle, agnostic to granularity; in practice, without a dedicated specification step it typically remains coarse-grained, and we therefore treat it as the skeleton h_c passed to the fine-grained stage. Concretely, given such a skeleton h_c , a fine-grained hypothesis is defined as

$$h_f = \{h_c, d_1, \dots, d_m\}, \quad (4)$$

where each edit d_i adds, instantiates, or removes a detail relative to h_c , drawn from a specification space $D = \mathcal{D}(b, h_c)$ that is implicitly induced by the background b and the skeleton h_c . Refinement is thus framed as selecting a coherent subset of edits from D :

$$P(h_f | b, h_c) = P(\{d_1, \dots, d_m\} | b, h_c, D). \quad (5)$$

Because D is implicit and combinatorially large, this is cast as an optimization problem in which

an LLM, acting as both the proposer of candidate edits and the evaluator that judges whether each edit improves the hypothesis, searches for a locally optimal h_f whose internal reward approaches the global optimum.

To make this search tractable over a noisy reward landscape, MOOSE-Chem2 organizes the edits in D into a hierarchy of abstraction levels and searches them *level by level*, fixing higher-level choices before descending to finer ones. The hierarchy is domain-agnostic, ranging from high-level conceptual choices down to low-level operational specifics; in chemistry, for example, it spans from the reaction mechanism and the class of reagent required, down to the exact compounds, concentrations, and temperatures. This hierarchical ordering yields two benefits. First, each level searches only its own reduced candidate subset rather than the full space D , shrinking the per-step branching factor. Second, it smooths the reward landscape: the reward of a high-level choice can be viewed as an aggregate over the many fine-grained variants beneath it, so searching at higher levels first averages out the high-frequency irregularities of the detail space—a low-pass filtering effect that reduces premature convergence to poor local optima. As the only existing work on fine-grained hypothesis discovery, we adopt MOOSE-Chem2 for the fine-grained refinement stage of our system.

2.2 Unified Hypothesis Discovery Framework

Existing research on hypothesis discovery falls into two complementary tasks: exploratory and fine-grained discovery. These correspond to the classical notions of *exploration* and *exploitation*—broadly searching for promising research directions, and deepening a chosen direction into a fully specified hypothesis. We unify them into a single framework, illustrated in Figure 1.

The *exploratory stage* (left), implemented with MOOSE-Chem (Section 2.1.1), searches the inspiration space I : at each step it selects an inspiration i_j (the title and abstract of a paper) and updates the working hypothesis, expanding a tree of diverse coarse-grained research directions (skeletons h_c). The *fine-grained stage* (right), implemented with MOOSE-Chem2 (Section 2.1.2), takes a skeleton h_c as input and refines it into a fully specified hypothesis h_f through hierarchical search over its induced specification space. Together, the two stages balance exploration and exploitation, transforming an initial research question into a fully specified,

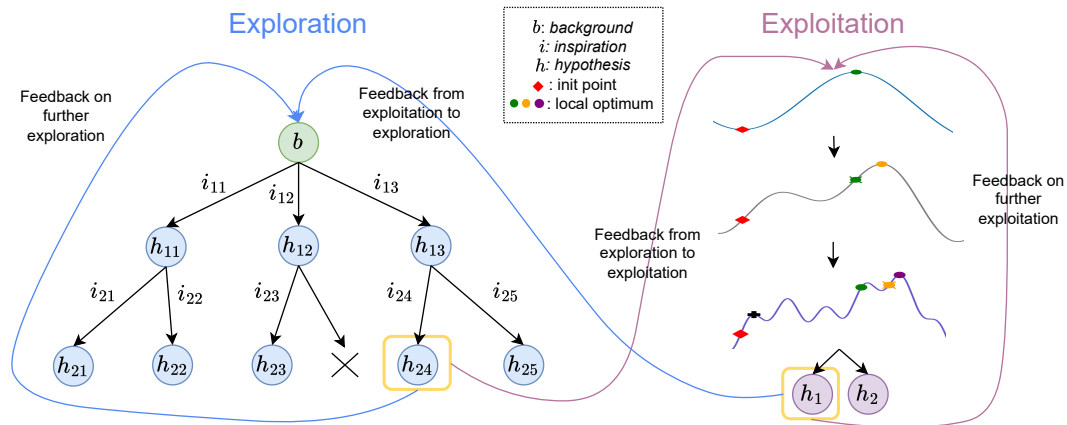


Figure 1: Overview of MOOSE-Copilot. Yellow rectangles indicate hypotheses actively selected by the user for further exploration or exploitation. Blue and purple paths represent user-directed transitions to additional exploration or fine-grained exploitation stage, respectively. Feedback can be provided along with each additional stage selection.

actionable hypothesis.

Bridging the two stages, however, requires deciding where to begin, when to switch between them, and how to steer each stage—decisions that are difficult to automate but natural for a domain expert. We therefore identify three guiding signals through which a human can steer the unified process, formalized in Section 2.3:

1. **Initial blueprints** that constrain the starting region of the exploration–exploitation process.
2. **Inter-stage routing**: when to transition between exploration and exploitation, and on which intermediate hypothesis.
3. **Intra-stage feedback**: in which direction to continue refining within the current stage.

2.3 Formalizing the HAI Protocol

A complete hypothesis requires both exploratory search for a promising skeleton and fine-grained exploitation of its details. Carrying out both without human guidance, however, is computationally prohibitive. A fully specified hypothesis $h_f = \{h_c, d_1, \dots, d_m\}$ is determined by two coupled searches: a search over the inspiration space I that composes a skeleton $h_c = f(b, i_1, \dots, i_k)$ (Section 2.1.1), and a search over the specification space $D = \mathcal{D}(b, h_c)$ that selects the edits $\{d_1, \dots, d_m\}$ refining h_c into h_f (Section 2.1.2). The difficulty is twofold. First, jointly searching I and D induces a combinatorial explosion of inspiration-driven branches and microscopic edits. Second, and more

fundamentally, the specification space is not fixed but induced by the skeleton, $D = \mathcal{D}(b, h_c)$: any change to h_c re-induces a different D and invalidates the edits accumulated under the previous skeleton. The two searches are therefore entangled rather than separable, which is what makes an autonomous joint search intractable.

To overcome this, we formalize MOOSE-Copilot as a human-intervened search process in which expert interventions act as routing operators that prune and redirect the search trajectory. We collect these interventions into a set of three guiding signals $\mathcal{F} = \{f_{init}, f_{route}, f_{dir}\}$:

- f_{init} (**Initial Blueprint**): root-node constraints injected to prune the initial region of the inspiration search, biasing which skeletons h_c are reachable from b .
- f_{route} (**Inter-stage Routing**): a transition operator that selects a skeleton h_c^* from the exploratory search and commits it as the entry point of fine-grained exploitation. The operator is symmetric: a refined hypothesis may also be elevated back to seed renewed exploration, in which case its skeleton is re-opened and the previously accumulated edits, being tied to the old skeleton, no longer directly apply once $D = \mathcal{D}(b, h_c)$ is re-induced under the new skeleton, though they may still inform the renewed search.
- f_{dir} (**Intra-stage Feedback**): a feedback-conditioned regeneration step available within either stage. Given the current hypothesis

h_{prev} and a user critique, it augments the context, $b' = b \cup \{h_{prev}, \text{feedback}\}$, and re-generates the next hypothesis conditioned on b' , steering the output toward the direction indicated by the critique.

With these three signals, the entangled joint search no longer needs to be carried out at once. Instead, the expert first explores I to obtain a skeleton, commits one via f_{route} , and only then refines its details over D —turning a single intractable search into two tractable single-stage searches, each conditioned on the guiding signals:

$$P(h_f | b, \mathcal{F}) \approx \underbrace{P(h_c | b, I, f_{init}, f_{dir})}_{\text{exploration over } I} \cdot \underbrace{P(\{d_1, \dots, d_m\} | b, h_c^*, D, f_{dir})}_{\text{exploitation over } D}. \quad (6)$$

The three signals enter the factorization at different points. f_{init} and f_{dir} condition the two generative terms, shaping how each stage generates. f_{route} , by contrast, acts between the two terms. In the forward direction it selects one skeleton h_c^* from those produced by the exploratory search and commits it as the entry point of exploitation, thereby inducing the specification space $D = \mathcal{D}(b, h_c^*)$ over which refinement proceeds. It can also operate in reverse, elevating a refined hypothesis back into the exploratory stage to seek a new skeleton; we return to the role of this reverse transition below.

The two terms correspond directly to the two stages defined earlier. The first term instantiates the exploratory decomposition of Equation 2: a skeleton h_c is composed from inspirations retrieved over I , bounded by f_{init} and redirected by f_{dir} . The second term instantiates the fine-grained refinement of Equation 5, now conditioned on f_{dir} : edits selected from the induced space D refine h_c^* into h_f .

Equation 6 describes a single explore-then-exploit pass, assuming f_{route} commits a suitable skeleton on the first attempt. In practice this is rarely optimal at once, as refinement over D may reveal that the committed skeleton is itself limiting. This is why f_{route} is bidirectional: a refined hypothesis can be elevated back to reopen exploration for a better skeleton, after which exploitation resumes over the newly induced D . The overall process therefore alternates between the two stages, progressively improving both the skeleton and its details rather than resolving them in a single pass.

Figure 2: Input interface of MOOSE-Copilot.

3 MOOSE-Copilot Web Interface

The interface operationalizes the three guiding signals of Section 2.3. Figure 2 shows the input interface: users enter their LLM API credentials, specify a research question, optionally provide a literature survey, and upload a custom inspiration corpus to guide the exploratory search—through which the initial blueprint f_{init} is injected. Figure 3 illustrates the tree view of the generated hypotheses, where each node is a hypothesis generated at a particular step, forming an interpretable trace of how ideas evolve through iterative inspiration-driven updates. Figure 4 presents the rank view, listing hypotheses by their averaged LLM self-evaluation scores ($R(h)$) so that users can compare hypothesis quality. Figure 5 displays the feedback interface. After selecting a hypothesis, the user can provide intra-stage feedback (f_{dir}) and make the inter-stage routing decision (f_{route}): continuing exploration with MOOSE-Chem or transitioning to fine-grained exploitation with MOOSE-Chem2. The routing is bidirectional—a refined hypothesis can likewise be taken back into exploration to seek a new direction.

4 Experiments

We evaluate our system on the TOMATO-Chem2 dataset (Yang et al., 2025a), leveraging its detailed annotations (research questions, literature surveys, and fine-grained hypotheses) across 51 top-tier papers. Following Yang et al. (2025a), we compute the recall of ground-truth elements recovered by the generated hypotheses (Table 1). Our experi-

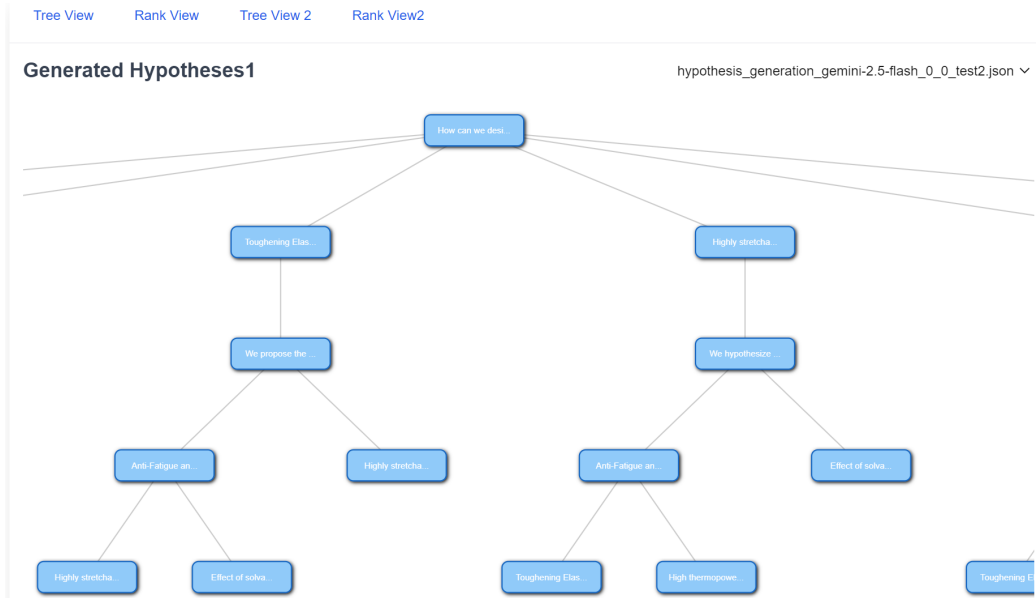


Figure 3: Tree view of the hypothesis generation process in MOOSE-Copilot.

Tree View Rank View Tree View 2 Rank View 2

Ranking of Hypotheses

Rank	Averaged Score	Hypothesis	Rounds	More
1	4	<p>We hypothesize that a flexible thermogalvanic device can be engineered through the integration of a polycaprolactone (PCL) and polyvinylidene fluoride (PVDF) hybrid matrix, employing a dual-directional freezing approach combined with strategic ion selection informed by the Hofmeister series. This design aims to simultaneously maximize Carnot-relative efficiency and mechanical robustness, addressing the critical challenges of energy harvesting from human body heat. 1. Material Composition - The proposed hybrid matrix will consist of PCL and PVDF in a 70:30 weight ratio, exploiting PCL's mechanical flexibility coupled with PVDF's piezoelectric properties. Selecting thermogalvanic ions—potassium (K+), sodium (Na+), and barium (Ba2+)—utilizing their positions within the Hofmeister series will be key to enhancing ionic conductivity. Initial molar ratios will be set at 0.1 M (K+), 0.05 M (Na+), and 0.01 M (Ba2+) to optimize ion transport and method efficacy. 2. Directional Freezing Methodology - The composite polymer solution will undergo a two-step directional freezing process, leveraging a custom-built apparatus featuring aluminum molds equipped with thermoelectric cooling elements. The first phase involves cooling the polymer solution from ambient temperature to -15°C at a rate of 1°C/min to induce initial crystallization. Once the desired crystal structure is established, a second phase will introduce a temperature gradient between -15°C and -5°C, promoting lamellar alignment while further controlling ice crystallization. Following this process, sublimation will occur at a controlled vacuum pressure of 150 mTorr at -3°C to stabilize the nanostructured matrix. 3. Structural Integration - To enhance mechanical integrity, cellulose nanocrystals (CNCs) will be incorporated into the hybrid matrix at a concentration of 5%, following template-assisted alignment methods during the casting process. This integration will achieve enhanced tensile strength and fatigue resistance while facilitating better ion transport pathways. Mechanical performance of the composite will be evaluated through tensile tests, targeting an increase of at least 50% in mechanical toughness compared to traditional thermogalvanic devices. 4. Thermogalvanic Mechanisms - The device will operate via redox reactions between the Hofmeister series-optimized ions and a redox-active conducting polymer such as polyaniline (PANI). Thermal gradients induced through body heat (approximately 37°C) will drive the ionic movement and subsequent electrochemical reactions within the device. The ionic dynamics, ionic conductivity, and thermal gradient effects will be characterized using electrochemical impedance spectroscopy (EIS) to elucidate conduction pathways. 5. Performance Evaluation - Comprehensive evaluations will involve quantifying Carnot-relative efficiency under physiological temperature conditions, aiming for at least 30% of theoretical maxima. Mechanical fatigue tests will simulate realistic stress conditions over 20,000 cycles, assessing structural integrity through cyclic tensile and environmental testing. Advanced imaging techniques, including scanning electron microscopy (SEM) and atomic force microscopy (AFM), will be employed to visualize microstructural alignment and integrity before and after performance assessments. This hypothesis seeks to innovate the field of thermogalvanics by developing a device that effectively harvests energy from body heat, ensuring both high efficiency and durability essential for sustainable wearable energy solutions.</p>	1	More

Figure 4: Hypothesis ranking interface in MOOSE-Copilot.

Hypothesis Details

Question:

How can we design a flexible thermogalvanic device that simultaneously maximizes both Carnot-relative efficiency and mechanical robustness for sustainable energy harvesting from body heat?

Hypothesis:

We hypothesize that a flexible thermogalvanic device can be engineered through the integration of a polycaprolactone (PCL) and polyvinylidene fluoride (PVDF) hybrid matrix, employing a dual-directional freezing approach combined with strategic ion selection informed by the Hofmeister series. This design aims to simultaneously maximize Carnot-relative efficiency and mechanical robustness, addressing the critical challenges of energy harvesting from human body heat. 1. Material

Feedback (Optional):

Enter your feedback here (optional)

MOOSE1

MOOSE2

Figure 5: Feedback interface in MOOSE-Copilot.

Method Name	Description	Recall	# Search Steps
baseline_MC (Yang et al., 2025b)	MC	11.44%	-
baseline_MC2 (Yang et al., 2025a)	MC2	10.33%	478.6
MC_with_hint	MC + initial blueprint	15.37%	-
MC_with_soft_feedback_with_hint	MC + initial blueprint + (oracle ranking + soft feedback) + MC	16.78%	-
MC_with_feedback_with_hint	MC + initial blueprint + (oracle ranking + feedback) + MC	16.93%	-
MC2_with_MC_input_self_rank	MC + initial blueprint + (self-ranking) + MC2	12.74%	321.2
MC2_with_MC_input_oracle_rank	MC + initial blueprint + (oracle-ranking) + MC2	18.26%	336.6
MC2_with_feedback_oracle_rank	MC + initial blueprint + (oracle-ranking) + MC2 + [(oracle ranking + feedback) + MC2]x1	21.98%	166.1
MC2_with_feedback_x2_oracle_rank	MC + initial blueprint + (oracle-ranking) + MC2 + [(oracle ranking + feedback) + MC2]x2	21.91%	149.9
MC2_with_feedback_x3_oracle_rank	MC + initial blueprint + (oracle-ranking) + MC2 + [(oracle ranking + feedback) + MC2]x3	22.35%	123.1
MC2_with_feedback_x4_oracle_rank	MC + initial blueprint + (oracle-ranking) + MC2 + [(oracle ranking + feedback) + MC2]x4	20.99%	107.4
MC2_with_strong_feedback_oracle_rank	MC + initial blueprint + (oracle-ranking) + MC2 + [(oracle ranking + strong feedback) + MC2]x1	23.10%	129.4
MC2_with_strong_feedback_x2_oracle_rank	MC + initial blueprint + (oracle-ranking) + MC2 + [(oracle ranking + strong feedback) + MC2]x2	25.54%	102.1
MC2_with_strong_feedback_x3_oracle_rank	MC + initial blueprint + (oracle-ranking) + MC2 + [(oracle ranking + strong feedback) + MC2]x3	25.70%	123.1
MC2_with_strong_feedback_x4_oracle_rank	MC + initial blueprint + (oracle-ranking) + MC2 + [(oracle ranking + strong feedback) + MC2]x4	26.96%	90.1

Table 1: Experimental results. “MC” represents MOOSE-Chem, “MC2” represents MOOSE-Chem2.

ments isolate and assess the impact of the protocol’s three guiding signals: (1) initial blueprints, (2) inter-stage routing, and (3) intra-stage feedback.

Justification for Oracle-Simulated Evaluation.

To measure the system’s intrinsic responsiveness to these signals—free from the confounding variance of human expertise—we adopt an *oracle-simulated evaluation*. Feedback is provided by an oracle LLM with access to the ground-truth hypothesis but instructed to give guiding critiques without disclosing the answer, and node selection is simulated by oracle ranking against the ground truth. This characterizes the recall achievable under idealized guidance, serving as an upper bound rather than an estimate of real-world human gains.

Results. Table 1 supports three observations. First, **initial blueprints** improve effectiveness, raising recall from 11.44% to 15.37% over the MC baseline. Second, **inter-stage routing** largely determines final performance: oracle ranking instead of self-ranking at the routing step raises recall from 12.74% to 18.26%. Third, **intra-stage feedback** drives stronger refinements, and crucially its *quality* matters more than its quantity. At matched rounds, strong feedback consistently reaches higher recall than ordinary feedback while using fewer optimization steps (e.g., at four rounds, 26.96%

at 90.1 steps vs. 20.99% at 107.4 steps). Moreover, repeating refinement under strong feedback steadily improves recall (23.10% \rightarrow 26.96% over one to four rounds), whereas under ordinary feedback additional rounds reduce the optimization steps but no longer improve recall (\sim 21% throughout). This indicates that high-quality critiques, rather than simply more refinement iterations, are what convert extra search into better hypotheses.

5 Conclusion

We presented MOOSE-Copilot, the first framework to bridge the abstraction gap between exploratory search and fine-grained exploitation in scientific hypothesis discovery. By formalizing a principled human–AI interaction protocol, we let scientists steer the automated process via initial blueprints, inter-stage routing, and intra-stage feedback. Under an oracle-simulated evaluation, structured guidance significantly outperforms purely autonomous baselines, characterizing the gains achievable under high-quality guidance. Furthermore, our interactive web interface lowers the technical barrier of command-line agentic tools, making AI-assisted hypothesis discovery accessible to interdisciplinary researchers beyond machine-learning experts.

Limitations

MOOSE-Copilot currently has two main limitations. First, it does not integrate automated experiment execution (Liu et al., 2025; Romera-Paredes et al., 2024; Novikov et al., 2025), which would close the loop between hypothesis generation and empirical validation. The system does, however, expose a hook for incorporating external experimental results through the intra-stage feedback signal f_{dir} . Second, it does not yet leverage post-training methods specifically developed for scientific hypothesis discovery (Yang and Bing, 2026), which could improve the generation quality of both the exploratory and fine-grained stages. We leave both as directions for future work.

Ethics Statement

This work aims to support scientific research by developing a system that assists in hypothesis generation through human–AI collaboration. All datasets used in this study are derived from publicly available scientific literature and do not contain personally identifiable or sensitive information. The system does not produce or promote misinformation intentionally, and user interactions remain local unless explicitly shared.

While MOOSE-Copilot has the potential to accelerate scientific discovery, we acknowledge possible misuse—for example, generating misleading or unfounded hypotheses if used without proper domain knowledge or oversight. We emphasize that the system is intended to augment, not replace, expert judgment, and we encourage responsible use by qualified researchers. Future work should further investigate the societal and epistemic implications of AI-augmented scientific workflows, including issues of bias, overreliance, and reproducibility.

We have made efforts to promote transparency, reproducibility, and accessibility by releasing our code and interface under a permissive open-source license.

Broader Impact Statement

MOOSE-Copilot aims to democratize and accelerate scientific discovery by providing researchers with a powerful, interactive assistant for hypothesis generation. By integrating both exploratory and fine-grained reasoning processes into a single human–AI framework, the system has the potential to support more systematic, creative, and efficient

research workflows across a wide range of scientific disciplines, including chemistry, biology, and materials science.

This broader accessibility may empower smaller research teams, early-career scientists, and institutions with fewer computational resources to participate more actively in hypothesis-driven research. It also opens new opportunities for interdisciplinary collaboration, as researchers can more easily explore ideas outside their immediate domain expertise with LLM-assisted support.

However, there are also potential risks. Improper use of the system—such as uncritical acceptance of AI-generated hypotheses, misuse in ideologically driven or pseudoscientific contexts, or reinforcement of biases present in training data—could negatively affect the scientific process. To mitigate this, MOOSE-Copilot is designed to keep humans meaningfully in the loop, encouraging critical reflection, transparency, and control.

Overall, we view this work as a step toward more inclusive, efficient, and assistive scientific research, while recognizing the importance of responsible use, continuous evaluation, and community oversight in its deployment and development.

References

- Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. 2024. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255*.
- Ruo Chen Li, Liqiang Jing, Chi Han, Jiawei Zhou, and Xinya Du. 2024. Learning to generate research idea with dynamic control. *arXiv preprint arXiv:2412.14626*.
- Wanhao Liu, Zonglin Yang, Jue Wang, Lidong Bing, Di Zhang, Dongzhan Zhou, Yuqiang Li, Houqiang Li, Erik Cambria, and Wanli Ouyang. 2025. Moosechem3: Toward experiment-guided hypothesis ranking via simulated experimental feedback. *arXiv preprint arXiv:2505.17873*.
- Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*.
- Alexander Novikov, Ng n V , Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. 2025. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*.

- Kevin Pu, KJ Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. Ideasynt: Iterative research idea development through evolving and composing idea facets with literature-grounded feedback. *arXiv preprint arXiv:2410.04025*.
- Bernardino Romera-Paredes, Mohammadamin Barekatin, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.
- Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. 2024. [LLM-SR: scientific equation discovery via programming with large language models](#). *CoRR*, abs/2404.18400.
- Zonglin Yang and Lidong Bing. 2026. Moose-star: Unlocking tractable training for scientific discovery by breaking the complexity barrier. In *Proceedings of the 43rd International Conference on Machine Learning*.
- Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. [Large language models for automated open-domain scientific hypotheses discovery](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13545–13565, Bangkok, Thailand. Association for Computational Linguistics.
- Zonglin Yang, Wanhao Liu, Ben Gao, Yujie Liu, Wei Li, Tong Xie, Lidong Bing, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. 2025a. Moose-chem2: Exploring llm limits in fine-grained scientific hypothesis discovery via hierarchical search. In *Advances in Neural Information Processing Systems*.
- Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2025b. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. [Hypothesis generation with large language models](#). *CoRR*, abs/2404.04326.