MOOSE-Chem2: Exploring LLM Limits in Fine-Grained Scientific Hypothesis Discovery via Hierarchical Search

Zonglin Yang[1,2‡], **Wanhao Liu**[3,2], **Ben Gao**[4,2], **Yujie Liu**[2], **Wei Li**[5], **Tong Xie**[6], **Lidong Bing**[7], **Wanli Ouyang**[2,8], **Erik Cambria**[1†], **Dongzhan Zhou**[2†]
[1] Nanyang Technological University [2] Shanghai Artificial Intelligence Laboratory
[3] University of Science and Technology of China [4] Wuhan University [5] National University of Singapore
[6] University of New South Wales [7] MiroMind [8] The Chinese University of Hong Kong
{zonglin001,cambria}@ntu.edu.sg, zhoudongzhan@pjlab.org.cn

Abstract

Large language models (LLMs) have shown promise in automating scientific hypothesis generation, yet existing approaches primarily yield coarse-grained hypotheses lacking critical methodological and experimental details. We introduce and formally define the new task of *fine-grained scientific hypothesis discovery*, which entails generating detailed, experimentally actionable hypotheses from coarse initial research directions. We frame this as a combinatorial optimization problem and investigate the upper limits of LLMs' capacity to solve it when maximally leveraged. Specifically, we explore four foundational questions: (1) how to best harness an LLM's internal heuristics to formulate the fine-grained hypothesis it itself would judge as the most promising among all the possible hypotheses it might generate, based on its own internal scoring-thus defining a latent reward landscape over the hypothesis space; (2) whether such LLM-judged better hypotheses exhibit stronger alignment with ground-truth hypotheses; (3) whether shaping the reward landscape using an ensemble of diverse LLMs of similar capacity yields better outcomes than defining it with repeated instances of the strongest LLM among them; and (4) whether an ensemble of identical LLMs provides a more reliable reward landscape than a single LLM. To address these questions, we propose a hierarchical search method that incrementally proposes and integrates details into the hypothesis, progressing from general concepts to specific experimental configurations. We show that this hierarchical process smooths the reward landscape and enables more effective optimization. Empirical evaluations on a new benchmark of expert-annotated fine-grained hypotheses from recent literature show that our method consistently outperforms strong baselines.[1]

1 Introduction

Large language models (LLMs) have increasingly been applied to assist scientific research (Luo et al., 2025), with one of the most ambitious applications being the automated discovery of novel and valid scientific hypotheses. However, current methods produce hypotheses that are criticized for being overly coarse, lacking sufficient detail, offering simplistic suggestions, or omitting concrete implementation strategies (Wang et al., 2024; Hu et al., 2024; Si et al., 2025).

[1] All code and data can be found in https://github.com/ZonglinY/MOOSE-Chem2

[‡] Contribution during internship at Shanghai Artificial Intelligence Laboratory. [†] Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

https://doi.org/10.52202/085713-2979

We present the first systematic investigation into how LLMs can be leveraged to formulate fine-grained scientific hypotheses—those enriched not only with major concepts but also with precise methodological details and clearly specified experimental configurations. For example, a coarse-grained hypothesis in chemistry might state, "*synthesize hierarchical 3D copper*," while a fine-grained counterpart could elaborate, "*Copper foils are chemically oxidized by immersion in a solution of 0.5 M ammonium persulfate and 2 M sodium hydroxide for 15 minutes at room temperature, forming a pentagonal hierarchical* CuO *nanostructure.*" Such fine-grained hypotheses significantly enhance clarity, feasibility, and experimental implementability.

Formally, we define the task as generating a fine-grained hypothesis given a research background—comprising a research question and a survey of established methodologies—and a coarse-grained hypothesis direction. We show that fine-grained scientific hypothesis discovery is a combinatorial search problem, as it requires selecting and composing a coherent set of concrete details from a vast space of plausible options—making it particularly challenging in practice. The difficulty is compounded by the fact that scientific hypothesis discovery is an inherently out-of-domain (OOD) problem: the correctness of a hypothesis is fundamentally *unknown* at the time of formulation.

In this work, we focus on the pre-experimental stage of discovery, mirroring how human scientists—prior to empirical testing—iteratively search through the hypothesis space using heuristics and domain knowledge to identify the hypothesis they themselves would judge as the most promising among all plausible candidates they could think of during the hypothesis search process.

Our goal is to emulate this cognitive search process using LLMs, which increasingly rival human scientists in heuristic reasoning and scientific knowledge understanding. This motivates our central research question (*Q1*): *how to best harness an LLM's internal heuristics to formulate the fine-grained hypothesis it itself would judge as the most promising among all possible hypotheses it might generate?* We conceptualize a hypothesis space where each point along the input dimensions (the x-axis, potentially multidimensional) represents a candidate hypothesis, and each point is assigned a reward value (on the y-axis) by the LLM based on its internal heuristics. This defines a reward landscape over the hypothesis space, with the highest peak corresponding to the hypothesis the LLM internally judges as most promising. Framed this way, *Q1* becomes an optimization problem: *how can we navigate this landscape to find stronger local optima—or ideally the global optimum—thus eliciting the best fine-grained hypothesis the LLM can generate?*

A straightforward baseline is greedy search over the reward landscape. However, its non-convex and complex structure makes naive greedy strategies prone to poor local optima. To address this, we propose a hierarchical search framework that explicitly models how a finite-capacity reasoning agent—human or LLM—navigates the hypothesis space. Specifically, it first explores higher-level conceptual spaces and then incrementally refines into more specific detail spaces. This hierarchical approach smooths the reward landscape at each hierarchy level—especially at higher, more abstract levels—enabling convergence to superior local optima compared to greedy search and greedy search with self-consistency (Wang et al., 2023). The proposed framework naturally scales with the capability of the underlying LLM's heuristics, yielding better optima as those heuristics become stronger.

Having investigated how to identify stronger local optima in *Q1*, we now turn to our second question (*Q2*): *whether hypotheses judged better by LLMs exhibit stronger alignment with ground-truth hypotheses?* To rigorously address *Q2* while avoiding data contamination, we construct a benchmark of research backgrounds paired with expert-annotated fine-grained hypotheses from chemistry papers published after January 2024, ensuring these examples were unseen by our LLM (`GPT-4o-mini`, October 2023 cutoff). Using this benchmark, we indirectly evaluate *Q2* by comparing the recall of hypotheses discovered by our hierarchical approach—which locates better LLM-internal local optima—with hypotheses identified by baseline methods. Our results consistently show that hypotheses generated by our method achieve higher recall than those from baselines, providing empirical support for the reliability of the LLM's internal reward signal in guiding fine-grained hypothesis discovery.

Until now, the reward landscape guiding hypothesis search has been defined by a single LLM serving as the evaluator. We now address our third question (*Q3*): *whether defining this landscape with an ensemble of diverse LLMs of similar capacity yields better outcomes than using equally sized ensembles of the strongest LLM within that group.* Our experiments show that ensembles of repeated instances of the strongest LLM consistently outperform equally sized ensembles of diverse models, suggesting that peak model quality is more critical than model diversity in this setting.

Finally, we consider a fourth question (*Q4*): *whether an ensemble of identical LLMs provides better reward landscape than a single instance of the same LLM*. While *Q3* compares ensembles of different models, *Q4* isolates the effect of aggregation alone by controlling for model identity. We find that even identical LLMs, when sampled independently and aggregated via summarization, yield a reward signal that better captures novelty without sacrificing overall quality—highlighting a subtle but important dimension in optimizing hypothesis discovery.

Notably, while our experiments focus on chemistry, the task formulation, methodology, and analysis of *Q1*-*Q4* are discipline-agnostic. The only domain-specific component is the manually designed hierarchy (used by the methodology), which can be instantiated for each new discipline encountered.

Overall, the contributions of this work are:

1. We introduce and formalize the *fine-grained scientific hypothesis discovery* task as a combinatorial optimization problem, and release a post-2024 benchmark with expert-annotated fine-grained hypotheses, specifically designed to prevent data contamination.
2. We explore the limits of LLMs for fine-grained scientific hypothesis discovery by framing it as an optimization problem over a reward landscape defined by LLM heuristics, with pairwise comparisons serving as the gradient signal. We also propose a hierarchical heuristic search framework that theoretically smooths the reward landscape, reduces search complexity, and identifies superior local optimum by interpolating among discovered optima. Empirically, this framework consistently outperforms strong baselines in locating better local optima.
3. We analyze this optimization formalization through 4 foundational research questions.

2 Methodology

2.1 Background and Task Motivation

Yang et al. (2025) assume that many chemistry hypotheses can be constructed from a research background b—typically including the research question and/or background survey—and a set of inspirations $i_1, \ldots, i_k$, representing concepts or findings from the literature. It can be formulated as:

$$h = f(b, i_1, \ldots, i_k) \tag{1}$$

In practice, however, most hypotheses h generated from Equation 1 tend to be coarse-grained: while they form cohesive associations between b and the i, they often lack clear hypothesis specification and the detailed experimental configurations required for direct implementation in a laboratory setting. Additionally, many such hypotheses contain redundant elements—either due to the inclusion of unnecessary inspirations or from noise present in the literature that is unrelated to the core knowledge intended for hypothesis construction.

2.2 Problem Formulation: Fine-Grained Hypothesis Generation as Combinatorial Search

Let h_c be a coarse-grained hypothesis direction and h_f its fine-grained counterpart, defined as:

$$h_f = \{h_c, d_1, \ldots, d_m\} \tag{2}$$

Here, $\{h_c, d_1, \ldots, d_m\}$ denotes the meaningful integration of edits $d_1, \ldots, d_m$ into h_c, resulting in a coherent, fine-grained hypothesis. Each edit d corresponds to either (1) the addition of a fine-grained detail to an existing concept i or the introduction of a new concept i into h_c, or (2) the deletion of a redundant detail or concept from h_c. We define two edit candidate sets: D^+, containing all details and concepts that may be added to h_c; and D^-, containing all details and concepts within h_c that may be removed. The overall edit space is then given by $D = D^+ \cup D^-$.

Inspired by coarse-to-fine strategies in computer vision—where a coarse image is first generated and then refined with fine-grained details (Tian et al., 2024)—we formulate the transition from h_c to h_f as an additional step building on Equation 1, which provides the initial h_c.

$$P(h_f|b, h_c) = P(\{d_1, \ldots, d_m\}|b, h_c, D) \tag{3}$$

https://doi.org/10.52202/085713-2979

This formulation turns $P(h_f \mid b, h_c)$ into a combinatorial optimization problem, where the objective is to select a subset of edits $d_1, \ldots, d_m \subseteq D$. Let $|D| = n$ and $|d_1, \ldots, d_m| = m$. The search space has at least combinatorial complexity $C_n^m = \frac{n!}{m!(n-m)!}$. This makes the problem particularly challenging due to three factors: (1) both m and n are unknown; (2) the candidate set D is itself implicit and potentially very large; and (3) the edits d_i are not independent—errors early in the reasoning chain can propagate and impair later decisions.

2.3 Algorithmic Motivation for Hierarchical Heuristic Search (*HHS*)

Fine-grained hypothesis generation is generally intractable due to the exponential growth of the search space, where the candidate set $|D|$ is often large or prohibitively so.

A notable exception occurs when the problem exhibits an optimal substructure—i.e., an optimal solution can be composed from the optimal solutions to its subproblems. This principle underlies dynamic programming, where solutions are built incrementally from smaller subproblems (first try to obtain the optimal solution for a smaller subproblem and then iteratively find the optimal solution for larger subproblems).

We observe that fine-grained hypothesis generation exhibits an optimal substructure. Specifically, the edits $d_1, \ldots, d_m$ can be organized hierarchically: some address high-level concepts (e.g., functional groups, catalyst classes), while others specify low-level details (e.g., reagents, catalysts, temperature, concentration). We assume these edits can be partitioned into p hierarchical levels ($p > 1$), with higher levels corresponding to finer details. Then the overall problem can be seen as to determine d in $\{1, \ldots, p\}$ hierarchies. The subproblem of it can be seen as the determination of d in $\{1, \ldots, p-1\}$ hierarchies, etc. Then it is obvious that the optimal solution of a problem can be derived from the optimal solution of its subproblem, etc.

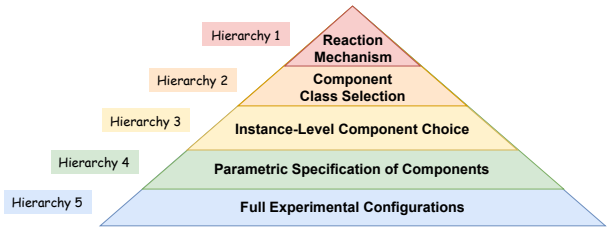

Figure 1: Hierarchies designed for chemistry and material science by PhD-level domain expert.

Figure 1 illustrates an example hierarchical decomposition for chemistry, developed in collaboration with domain experts (PhD-level chemists). The hierarchy spans from high-level mechanistic intent to low-level experimental configurations, reflecting the granularity typically considered when translating a conceptual hypothesis into a testable laboratory procedure in chemistry.

Now we have simplified the problem of determining d in all p hierarchies into the iteration of determining d in each hierarchy sequentially. Nonetheless, even within a single hierarchy, the number of candidates remains combinatorially large.

A practical approach to this combinatorial complexity is to use heuristics that approximate solutions rather than exhaustively searching for exact ones. This aligns with how chemists refine hypotheses: given that h_c often represents an unexplored direction, while individual details may be retrievable from existing databases, the complete set of details is rarely available. Instead, chemists often rely on domain knowledge and intuition to heuristically identify and progressively integrate plausible details.

Analogously, we propose to leverage LLMs' internal heuristics to guide the search for d at each hierarchical level. As LLMs advance, their heuristics—emerging from pretraining over extensive scientific corpora—will increasingly approximate, and may surpass, those of human experts. The proposed framework naturally scales with the strength of these heuristics, yielding increasingly better optima for fine-grained hypothesis discovery as LLM capabilities continue to grow.

In this setting, the candidate space D is not explicitly enumerated but is implicitly embedded within the LLM's internal knowledge and reasoning capabilities. The LLM does not select d from a predefined list, but rather proposes candidates by navigating this latent, heuristic-driven space.

2.4 Hierarchical Factorization of the Search Problem

For formalization, we partition the implicit candidate space D into p hierarchical levels, where $D^{(i)} \subseteq D$ represents all potential edits at level i, and $D^{*(i)} \subseteq D^{(i)}$ denotes the (unknown) ground-truth edits. The j-th ground-truth edit at level i is denoted as $d_j^{*(i)} \in D^{*(i)}$. Since D is implicitly determined by h_c, we have $P(D \mid h_c) = 1$, and explicitly condition on D for clarity in the subsequent factorization. Applying the chain rule hierarchically, Equation 3 can be reformulated as:

$$P(h_f \mid b, h_c) = P\left(\{D^{*(1)}, \ldots, D^{*(p)}\} \mid b, h_c, D\right) \tag{4}$$

$$= \prod_{i=1}^{p} P\left(D^{*(i)} \mid b, h_c, D^{*(<i)}, D^{(i)}\right) \tag{5}$$

$$= \prod_{i=1}^{p} \prod_{j=1}^{|D^{*(i)}|} P\left(d_j^{*(i)} \mid b, h_c, D^{*(<i)}, d_{<j}^{*(i)}, D^{(i)}\right), \tag{6}$$

where $D^{*(<i)} = \{D^{*(1)}, \ldots, D^{*(i-1)}\}$ and $d_{<j}^{*(i)} = \{d_1^{*(i)}, \ldots, d_{j-1}^{*(i)}\}$.

The key advantage of this hierarchical factorization is that at each level i, the search is restricted to the reduced candidate set $D^{(i)}$ rather than the full space D, significantly narrowing the search space. Moreover, as we will show in § 2.6, this hierarchical decomposition smooths the reward landscape at each hierarchy level, facilitating more stable optimization and enabling the discovery of stronger local optima in the hypothesis space.

2.5 LLM-Based Implementation of Hierarchical Heuristic Search

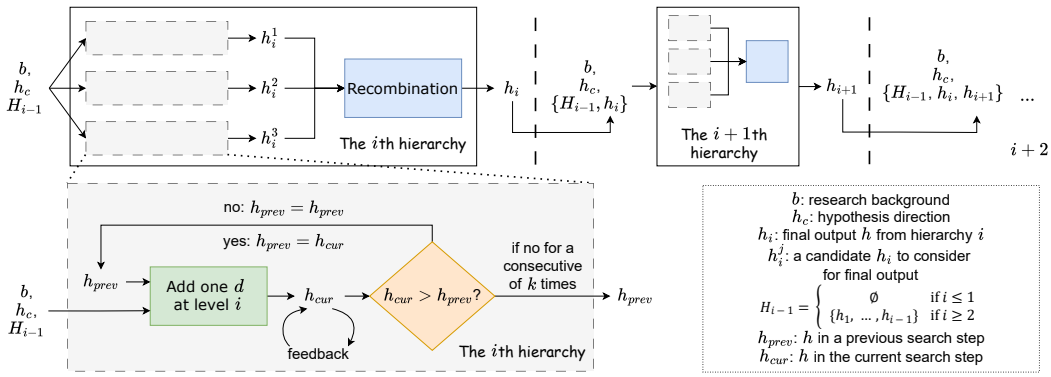

Figure 2: Overview of the proposed Hierarchical Heuristic Search (HHS) framework.

We implement *HHS* as an LLM-driven agentic process that directly follows the hierarchical factorization formalized in Equation 6. As shown in Figure 2, at each hierarchy level i, H_{i-1} represents the accumulated edits from all previous levels, corresponding to $D^{*(<i)}$. Within the current level, h_{prev} denotes the partial hypothesis incorporating the edits selected up to step $j-1$, i.e., $d_{<j}^{*(i)}$. The candidate set $D^{(i)}$ is not explicitly enumerated but emerges implicitly from the LLM's internal heuristics, conditioned on the background b, the hypothesis direction h_c, and the edits selected so far.

Specifically, the search for a local optimum h_i^j begins from the initial point h_{i-1}, using contextual information from b, h_c, and H_{i-2}. For hierarchy level $i = 1$, we set $h_0 = h_c$ and $H_0 = \emptyset$, making the hypothesis direction h_c the starting point.

At each iteration, the "*Add one d at level i*" module prompts an LLM to propose an edit d to h_{prev}, producing a candidate h_{cur}, which is then refined once for validity, novelty, and specificity. The "$h_{cur} > h_{prev}$" module evaluates whether the new hypothesis improves upon the previous one via LLM-based pairwise comparison, serving as an *internal gradient signal* for hypothesis optimization.

https://doi.org/10.52202/085713-2979

This search process continues until no further improvement is observed over k consecutive steps (default $k = 3$), at which point the current hypothesis is accepted as a local optimum. Each edit d may involve either an addition or a deletion, allowing the search path to include retrospection and self-correction as needed.

Within each hierarchy level, we adapt the design of an evolutionary unit (Yang et al., 2025) to our task, where the search for the local optimum h_i is independently repeated multiple times (set to three in our implementation), yielding several local optima h_i^1, h_i^2, and h_i^3. These candidates are then passed to a *recombination* module, which integrates their complementary strengths to interpolate a potentially superior local optimum h_i within the subspace spanned by h_i^1, h_i^2, h_i^3.

2.6 Theoretical Analysis: Smoothing Effects of Hierarchical Heuristic Search

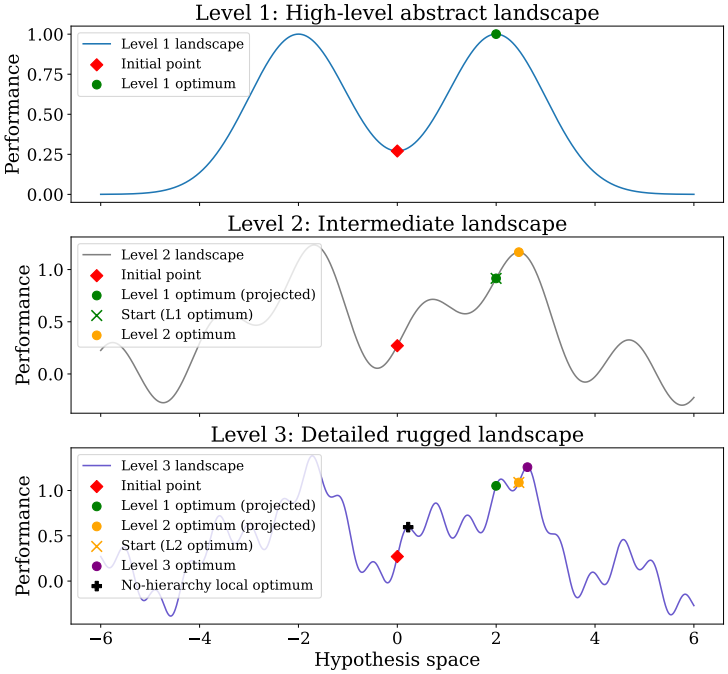

Figure 3: The smoothing effect of hierarchy on the reward landscape of hypothesis space.

A key observation is that a hypothesis candidate's performance at lower (more abstract) hierarchy level can be viewed as an aggregated estimate—approximating an average or soft maximum—of its higher-level subspace (more concrete). For instance, when evaluating a coarse-grained (more abstract) concept like "hierarchical 3D copper," the LLM may implicitly account for its diverse fine-grained (more concrete) structural variants, some highly relevant, others ineffective. We hypothesize that the LLM's assessment to the coarse-grained concept is an aggregation of its fine-grained outcomes, weighting promising variants within the broader distribution to produce an overall estimate of the coarse-grained concept's expected potential.

Building on this observation, the hierarchical abstraction smooths the reward landscape at lower levels by attenuating local irregularities in the fine-grained space, as the performance of a point at a lower level can be interpreted as an approximate aggregation or average of the performance across its corresponding higher-level subspace. This effect is illustrated in Figure 3 (a simplified schematic projection into a 1D space). Consequently, direct search over the flat, non-hierarchical space tends to be highly rugged and non-convex, often leading to premature convergence to suboptimal local optima. In contrast, introducing hierarchical structure progressively smooths the landscape, enabling more stable and efficient optimization, particularly at lowest levels.

This smoothing effect can also be interpreted in the frequency domain as a form of low-pass filtering, where high-frequency components of the landscape are attenuated, resulting in a (roughly) spectral cutoff in the spatial frequency domain, as illustrated in Figure 4.

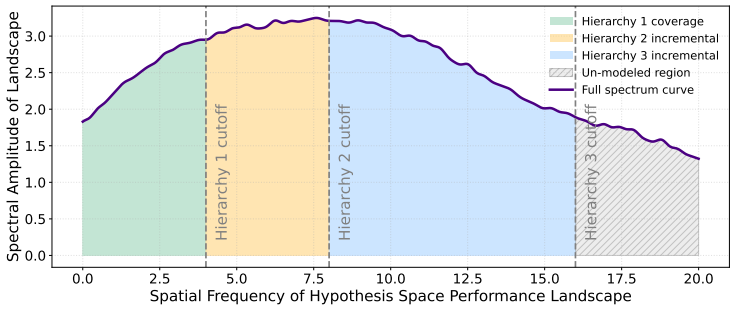

Figure 4: Hierarchical design as a low-pass filtering over the spectrum of the reward landscape.

3 Experiment: Investigating the Four Fundamental Questions

3.1 Benchmark Construction, LLM Selection, and Baselines

To our knowledge, no existing benchmark provides annotated fine-grained scientific hypotheses—detailed enough for direct experimental execution. We extend the TOMATO-Chem dataset (Yang et al., 2025), which includes 51 chemistry papers published after January 2024 in leading journals such as *Nature* and *Science*. Each entry contains a research background b and a coarse-grained hypothesis h_c. We further annotated it with fine-grained hypotheses h_f, serving as ground-truth references created by two PhD-level chemists. To prevent data contamination, all experiments are conducted using `GPT-4o-mini`, whose pretraining data cutoff is October 2023.

We compare *HHS* against two strong baselines widely used in search tasks: (1) greedy search and (2) greedy search with self-consistency. The latter serves as an ablation of *HHS* where the hierarchical decomposition is removed, performing the search in a single stage with each d sampled directly from the full candidate set D rather than hierarchy-specific subsets $D^{(i)}$. The self-consistency mechanism is similar to the *Recombination* module in Figure 2, which interpolate multiple local optima trying to find a better one. Greedy search represents a further ablation, disabling the *Recombination* module entirely and following a single search trace where the first found local optimum (h_i^1) is directly adopted as the output ($h_i = h_i^1$ in Figure 2).

3.2 *Q1*: How to Best Harness an LLM's Internal Heuristics to Formulate the Fine-Grained Hypothesis It Would Judge Most Promising Among All It Might Generate?

We frame this question as an optimization problem: Given only a coarse-grained hypothesis h_c as the starting point, and relying entirely on a single LLM, how can we navigate the hypothesis space to approach the global optimum of the reward landscape, as defined by this same LLM's internal heuristics, where each optimization step consists of adding an edit d to h_c? In this setting, the LLM plays a dual role: it serves both as the *proposal generator*, proposing candidate edits d to formulate new hypotheses within the hypothesis space, and as the *gradient provider*, judging whether the new hypothesis improves upon the current one via its own internal heuristics (e.g., pairwise comparison).

While it is inherently infeasible to determine whether a found local optimum represents the global optimum, we can empirically compare local optima obtained by different methods with the *gradient provider* (pairwise comparison), and therefore check which one is a better local optimum.

As detailed in § 2, the hierarchical design of *HHS* offers two key advantages over flat search strategies: (1) less search space to propose each d (from $D^{(i)}$, instead of D), and (2) smoothing the reward landscape progressively in the hypothesis space. Among these, the smoothing effect is particularly critical, as it reduces the risk of early convergence to suboptimal local optima and facilitates progress toward higher peaks in the LLM's internal reward landscape.

We compare the local optima discovered by *HHS* against the two baselines. For each pair of local optima, we conduct both overall evaluations and dimension-specific assessments across four key criteria: **effectiveness**, **novelty**, **detailedness**, and **feasibility**. In this context, **feasibility** reflects the practical ease of implementing the proposed hypothesis, encompassing factors such as implementation

https://doi.org/10.52202/085713-2979

complexity and the minimization of redundant steps. Hypotheses that are easy to implement and free of redundant components are preferred.

We further observe two common trade-offs among these dimensions: (1) between **effectiveness** and **novelty**, as highly novel hypotheses often entail greater scientific risk and uncertainty; and (2) between **detailedness** and **feasibility**, as increased specificity can introduce procedural complexity or redundancies that diminish experimental feasibility.

We also conducted an expert evaluation involving two chemistry PhD students. For each benchmark item, one hypothesis was randomly sampled from each method, and the experts were tasked to rank the three hypotheses. The results from both the LLM-based and expert evaluations on the quality of local optima discovered by each method are presented in Table 1. To mitigate known position bias in LLM-based pairwise comparisons—where models tend to favor the first option (Li et al., 2024)—each pair of local optima was compared six times, with the order of presentation alternated every three times. A hypothesis was considered to win if it received more than three votes; a tie was recorded if both received exactly three votes.

Table 1: Comparison between *HHS* and baseline methods across LLM-based and expert evaluations.

	Effectiveness (LLM)	Novelty (LLM)	Detailedness (LLM)	Feasibility (LLM)	Overall (LLM)	Overall (Expert)
***HHS* v.s. Greedy Search**						
Win	**74.51%**	**41.18%**	**71.57%**	**67.65%**	**73.53%**	**76.47%**
Tie	18.63%	18.63%	28.43%	10.78%	18.63%	15.69%
Lose	6.86%	40.20%	0.00%	21.57%	7.84%	7.84%
***HHS* v.s. Greedy Search + Self-consistency**						
Win	**59.31%**	42.16%	**56.37%**	**48.53%**	**53.43%**	**74.51%**
Tie	24.02%	8.33%	43.14%	18.63%	33.82%	17.65%
Lose	16.67%	**49.51%**	0.49%	32.84%	12.75%	7.84%
Greedy Search + Self-consistency v.s. Greedy Search						
Win	**57.84%**	**48.04%**	**29.41%**	**51.96%**	**54.90%**	**62.75%**
Tie	22.55%	11.76%	65.69%	18.63%	34.31%	21.57%
Lose	19.61%	40.20%	4.90%	29.41%	10.78%	15.69%

3.3 *Q2*: Whether Hypotheses Judged Better by LLMs Exhibit Stronger Alignment With Ground-Truth Hypotheses?

§ 3.2 shows that HHS consistently discovers superior local optima compared to baseline methods. We further investigate whether these optima exhibit stronger alignment with the ground-truth hypotheses.

Table 2: Recall of ground-truth components by discovered hypotheses. #Steps represents the number of reasoning steps used. *HHS* represents *HHS-3* referred in § 3.5.

	Soft Recall	Hard Recall	#Steps
ChemCrow (M. Bran et al., 2024)	12.28%	7.20%	-
Qi et al. (2024)	19.57%	11.15%	-
SciMON (Wang et al., 2024)	18.57%	10.09%	-
MOOSE (Yang et al., 2024)	20.04%	11.76%	-
MOOSE-Chem (Yang et al., 2025)	19.99%	11.98%	-
Greedy Search	16.60%	9.92%	9.69
w/ In-context RL	16.76%	10.28%	16.80
w/ Self-consistency	31.53%	17.73%	67.55
HHS (*HHS-3*)	**40.35%**	**23.04%**	282.04
w/ In-context RL	33.63%	21.29%	531.08
w/ Single LLM Gradient (*HHS-1*)	32.40%	19.95%	747.92

Given the lack of established metrics for this task, we introduce an LLM-based evaluation that measures how well the discovered hypotheses recover the methodological and experimental details of the ground-truth hypotheses. The detailed formulations of the two metrics—*Soft Recall* and *Hard Recall*—are provided in Appendix C.

As shown in Table 2, the hypotheses discovered by HHS—corresponding to better local optima than those produced by greedy search baselines—achieve consistently higher recall scores than both greedy search baselines and other comparative methods. Here, *in-context RL* refers to a mechanism in which, if the current hypothesis h_{cur} does not outperform the previous one h_{prev}, h_{cur} is inserted into the LLM's context to generate a new candidate. We also report the total number of reasoning steps used by each method. A general trend emerges: increasing the number of reasoning steps tends to improve recall up to a point, beyond which excessive steps lead to diminishing or negative returns.

3.4 *Q3*: Whether Defining the Reward Landscape With an Ensemble of Diverse LLMs Yields Better Outcomes Than Using the Same Number of the Strongest LLMs Among Them?

Table 3: "EF": Effectiveness, "NV": Novelty, "DT": Detailedness, "FS": Feasibility, "OV": Overall. "(GT)" and "(GM)" indicate that the pairwise comparisons were conducted by `GPT-4o-mini` and `Gemini-1.5-flash`, respectively.

	EF (GT)	NV (GT)	DT (GT)	FS (GT)	OV (GT)	EF (GM)	NV (GM)	DT (GM)	FS (GM)	OV (GM)
	Mixed committee v.s. `GPT-4o-mini` committee					`GPT-4o-mini` committee v.s. `Gemini-1.5-flash` committee				
Win	20.83%	33.33%	**14.58%**	33.33%	29.17%	**27.08%**	**31.25%**	**14.58%**	0.00%	**18.75%**
Tie	41.67%	20.83%	72.92%	18.75%	33.33%	58.33%	52.08%	77.08%	95.83%	68.75%
Lose	**37.50%**	**45.83%**	12.50%	**47.92%**	**37.50%**	14.58%	16.67%	8.33%	**4.17%**	12.50%
	`Gemini-1.5-flash` committee v.s. `GPT-4o-mini` committee					Mixed committee v.s. `Gemini-1.5-flash` committee				
Win	16.67%	25.00%	6.25%	37.50%	16.67%	**16.67%**	**33.33%**	12.50%	**6.25%**	**18.75%**
Tie	41.67%	27.08%	79.17%	25.00%	52.08%	68.75%	35.42%	75.00%	93.75%	64.58%
Lose	**41.67%**	**47.92%**	**14.58%**	37.50%	**31.25%**	14.58%	31.25%	12.50%	0.00%	16.67%
	Mixed committee v.s. `Gemini-1.5-flash` committee					Mixed committee v.s. `GPT-4o-mini` committee				
Win	**29.17%**	**45.83%**	**10.42%**	**47.92%**	**27.08%**	8.33%	29.17%	14.58%	**6.25%**	8.33%
Tie	56.25%	16.67%	85.42%	10.42%	50.00%	77.08%	39.58%	70.83%	93.75%	64.58%
Lose	14.58%	37.50%	4.17%	41.67%	22.92%	**14.58%**	**31.25%**	14.58%	0.00%	**27.08%**

The optimization in *HHS* relies on the "$h_{cur} > h_{prev}$?" module (Figure 2), which acts as the gradient signal driving the search. This raises a key question: does a diverse ensemble of comparably capable LLMs improve search performance, or is it more effective to use the same number of parallel instances of the single strongest LLM among the ensemble?

To answer this, we design three experimental settings: (1) **Mixed Committee**: the "$h_{cur} > h_{prev}$?" module is implemented by an ensemble of three different LLMs—`GPT-4o-mini` (OpenAI, 2024), `Gemini-1.5-flash` (Georgiev et al., 2024), and `Claude-3-haiku` (Anthropic, 2024); (2) **GPT-4o-mini Committee**: the module is implemented by three instances of `GPT-4o-mini`; (3) **Gemini-1.5-flash Committee**: the module is implemented by three instances of `Gemini-1.5-flash`. Each committee's three judgments are then aggregated by a `GPT-4o-mini`, which produces the final decision for "$h_{cur} > h_{prev}$?" representing that committee. All three settings use `GPT-4o-mini` as the proposer module for generating edits d at each hierarchy level i.

We compare the local optima generated by these setups using LLM-based pairwise comparisons, following the protocol in § 3.2, where each pair is evaluated six times to mitigate position bias. However, since the evaluator is itself an LLM, an additional bias may occur—favoring optima discovered using gradients from the same model. To control for this, we conduct two sets of evaluations: one using `GPT-4o-mini` as the evaluator, and the other using `Gemini-1.5-flash`.

As shown in Table 3, across both evaluators, the `GPT-4o-mini` committee consistently outperforms the mixed committee, which in turn outperforms the `Gemini-1.5-flash` committee. These results suggest that leveraging repeated instances of the strongest single model provides a more effective gradient signal for hypothesis optimization than combining different models of similar capacity.

3.5 *Q4*: Do Multiple Identical LLMs Yield a Better Reward Landscape Than One?

In experiments of Tables 1 and 2 (except for the *HHS-1* line in table 2), the reward landscape was defined using an ensemble of three identical LLMs, followed by a fourth instance of the same LLM that aggregated the three judgments into a final, reasoned decision. However, it is unclear on whether

https://doi.org/10.52202/085713-2979

one LLM would be already enough. To evaluate this, we compare two variants: *HHS-3*, which uses an ensemble of three identical instances of `GPT-4o-mini` (and a fourth instance of the same LLM for aggregation) to provide the reward signal, and *HHS-1*, which relies on a single instance of the same model. Table 4 reports LLM-based pairwise evaluations between the two setups across four criteria. While overall quality, effectiveness, and detailedness are largely comparable, *HHS-3* outperforms in novelty, whereas *HHS-1* shows an advantage in feasibility.

Table 4: Pairwise comparison between *HHS-1* and *HHS-3* with LLM-based evaluation.

	Effectiveness (LLM)	Novelty (LLM)	Detailedness (LLM)	Feasibility (LLM)	Overall (LLM)
			HHS-1 v.s. *HHS-3*		
Win	21.08%	25.49%	4.41%	**41.67%**	8.82%
Tie	57.35%	28.92%	94.12%	28.92%	82.35%
Lose	21.57%	**45.59%**	1.47%	29.41%	8.82%

This result is somewhat counterintuitive. Notably, the summarization step is not a simple majority vote: the aggregating LLM assesses the relative strength of reasoning across the three perspectives and selects the most compelling argument. This allows it to surface minority-supported but well-reasoned views, promoting exploration of more novel or unconventional hypotheses. Consequently, the aggregated reward signal in *HHS-3* becomes more sensitive to creative or atypical ideas that a single-shot evaluation might overlook, whereas *HHS-1* relies on a single comparative judgment at each step—favoring conventional and well-established hypotheses, at the expense of novelty.

Table 2 presents the recall of ground-truth components for hypotheses generated by *HHS-1* and *HHS-3*. Across both soft and hard recall metrics, *HHS-3* outperforms *HHS-1*, indicating stronger alignment with expert-annotated reference hypotheses.

4 Related Work and Discussion

LLM-driven scientific discovery methods typically fall into two categories: (1) direct generation of hypotheses from a research background—comprising a research question and background survey (Qi et al., 2024); or (2) retrieval of seemingly unrelated yet potentially useful knowledge fragments, or *inspirations*, which are then combined with the background to construct a hypothesis (Yang et al., 2024, 2025; Wang et al., 2024; Liu et al., 2025b). While these methods show promise in generating novel ideas, they are often criticized for producing hypotheses that are overly coarse, lacking detail, or omitting actionable experimental steps (Wang et al., 2024; Hu et al., 2024; Si et al., 2025). In contrast, our goal is to investigate how LLMs can generate *fine-grained scientific hypotheses*—those sufficiently detailed to be directly implemented in laboratory settings. Although this work centers on the *pre-experimental* stage of discovery, the framework can in principle extend to the *experiment-guided* stage (Liu et al., 2025a; Romera-Paredes et al., 2024; Shojaee et al., 2025; Novikov et al., 2025) by incorporating experimental feedback into the background survey. Likewise, details retrieved from papers relevant to the proposed hypothesis can also be integrated into the background survey.

5 Conclusion

We introduce and formalize the *fine-grained scientific hypothesis discovery* task as a combinatorial optimization problem. To explore the upper limit of LLMs' capacity for this task, we frame it as an optimization problem, and propose hierarchical heuristic search (*HHS*), which theoretically smooths the reward landscape, reduces combinatorial complexity through optimal substructure exploitation, and identifies superior local optimum by interpolating among discovered optima. Empirical results show that (1) *HHS* reliably discovers better local optima than flat greedy search baselines, (2) hypotheses preferred by LLMs often align more closely with expert annotations, (3) repeated use of the strongest model defines a more effective reward landscape than diverse ensembles, and (4) aggregating identical LLMs yields a reward signal more sensitive to novelty and higher in recall than single-instance evaluation. These findings illustrate how hierarchical search can better harness LLMs' internal heuristics for scientific discovery. Although evaluated on a chemistry dataset, the proposed task formulation and methodology are expected to generalize across disciplines, with only the manually designed hierarchy (Figure 1) requiring domain-specific adaptation.

Acknowledgments

This work was supported by a locally commissioned task from the Shanghai Municipal Government. This work is supported by Shanghai Artificial Intelligence Laboratory. This research/project is supported by the Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (STEM RIE2025 Award MOE-T2EP20123-0005).

References

Anthropic. The claude 3 model family: Opus, sonnet, haiku. `https://www.anthropic.com/news/claude-3-family`, March 2024. Accessed: 2025-05-16.

Erik Cambria, Rui Mao, Melvin Chen, Zhaoxia Wang, and Seng-Beng Ho. Seven pillars for the future of artificial intelligence. *IEEE Intelligent Systems*, 38(6):62–69, 2023.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar

 https://doi.org/10.52202/085713-2979

Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeynep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, and Gu. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. URL `https://arxiv.org/abs/2403.05530`.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. Nova: An iterative planning and search approach to enhance novelty and diversity of LLM generated ideas. *CoRR*, abs/2410.14255, 2024. doi: 10.48550/ARXIV.2410.14255. URL `https://doi.org/10.48550/arXiv.2410.14255`.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. Split and merge: Aligning position biases in llm-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11084–11108, 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*, 2025a.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. *arXiv preprint arXiv:2503.21248*, 2025b.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*, 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.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`, 2024. Accessed: 2025-05-16.

Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. Large language models as biomedical hypothesis generators: A comprehensive evaluation. In *First Conference on Language Modeling*, 2024.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. Llm-sr: Scientific equation discovery via programming with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? A large-scale human study with 100+ NLP researchers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 279–299. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.18. URL https://doi.org/10.18653/v1/2024.acl-long.18.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. Large language models for automated open-domain scientific hypotheses discovery. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 13545–13565. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.804. URL https://doi.org/10.18653/v1/2024.findings-acl.804.

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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

 https://doi.org/10.52202/085713-2979

A MOOSE-Chem2 Overview I/O Figure

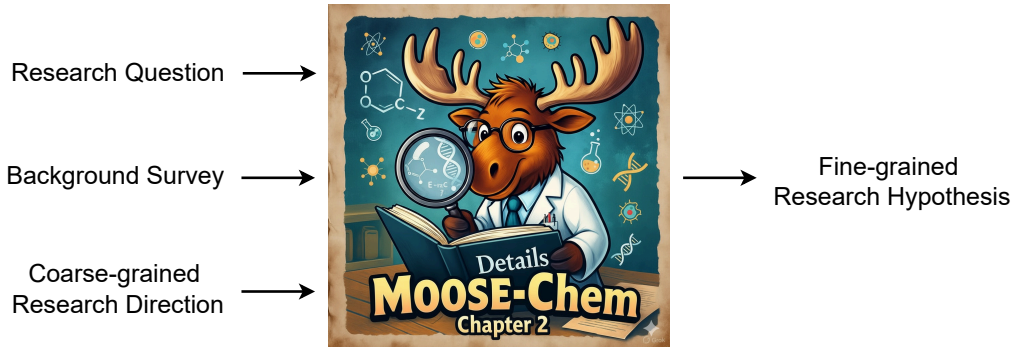

Figure 5: Overview of the input and output of the Hierarchical Heuristic Search (HHS) framework, also known as MOOSE-Chem2.

Figure 5 presents an overview of the input and output of the Hierarchical Heuristic Search (HHS) framework, also referred to as MOOSE-Chem Chapter 2 (MOOSE-Chem2). The input to the framework consists of three components: a *research question*, a *background survey*, and a *coarse-grained research direction*.

The research question defines the scientific goal to be addressed. The background survey summarizes existing methodologies relevant to the research question. It may optionally include methodological or experimental details from prior studies that investigate similar questions or employ similar methodologies to the given research direction, enabling the framework to leverage richer contextual knowledge. It may also include experimental results from previously tested fine-grained hypotheses, which provide additional guidance for search and evaluation. The coarse-grained research direction—potentially derived from MOOSE-Chem (Yang et al., 2025)—serves as a high-level starting point that can vary in granularity, ranging from a brief sentence describing a general research direction to a preliminary hypothesis with partial methodological details.

B Expert Evaluation Instructions

For each research question, you will be presented with three candidate hypotheses alongside a ground-truth fine-grained hypothesis. The order of the hypotheses is randomized. Your task is to rank the three candidate hypotheses based on their quality, using the ground-truth hypothesis as a reference.

Please evaluate the hypotheses based on the following four criteria:

- Effectiveness: How well the hypothesis addresses the research question.
- Novelty: The degree of originality relative to existing knowledge.
- Detailedness: The specificity and clarity of the hypothesis.
- Feasibility: The practical plausibility of experimentally testing or implementing the hypothesis.

Note that two tradeoffs may arise:

- Between effectiveness and novelty
- Between detailedness and feasibility

Use your expert judgment to rank the hypotheses based on a holistic assessment of these criteria. In rare cases where two hypotheses appear to be of similar quality, assigning them the same rank is acceptable.

C Recall Metric Calculation

Each evaluation compares a pair of hypotheses—a *ground-truth hypothesis* (from literature) and a *candidate hypothesis* (generated by the model). The goal is to quantify how much methodological or experimental content from the ground-truth hypothesis is covered by the candidate.

Step 1. Decomposition. Both hypotheses are decomposed into methodological and experimental components (details) using structured prompts co-designed with chemistry PhD students. The prompts guide an LLM to extract key components such as procedures, specific materials, or reaction parameters. Each component is annotated with its role, function, and context.

Step 2. Scoring via LLM-Judge. For each ground-truth component, the LLM judge searches for functionally corresponding components in the candidate hypothesis and assigns a *coverage score* from 0–3:

- **0**: no match;
- **1**: vague or partial match;
- **2**: close but not exact match;
- **3**: exact or specific match.

The scoring prompts and examples were jointly developed and audited by PhD students to ensure calibration and consistency.

Step 3. Aggregation. Two recall metrics are computed:

$$\text{Soft Recall} = \frac{\#\{\text{components with score} > 0\}}{\#\{\text{ground-truth components}\}}, \tag{7}$$

$$\text{Hard Recall} = \frac{\sum_i \text{score}_i}{3 \times \#\{\text{ground-truth components}\}}. \tag{8}$$

Soft Recall captures whether relevant components are recovered at all, while Hard Recall further reflects how precisely they are reproduced.

D Case Study

Example 1:

- Research Question: How can a cost-effective N-type quasi-solid-state thermocell be developed to boost electricity production from low-grade heat by improving both ion transport efficiency and electrode performance?
- Hypothesis Candidate (from HHS): The development of a cost-effective N-type quasi-solid-state thermocell will be achieved through the strategic integration of three core components to enhance electricity production from low-grade heat ($\leq$100°C):
  1. **Hierarchical Metal Electrodes**: Constructed from a copper-nickel alloy (70% Cu, 30% Ni) optimized for conductivity and corrosion resistance, these electrodes will feature a dual-layer design incorporating a titanium dioxide (TiO_2; CAS number: 13463-67-7) coating, precisely controlled at a thickness of 100 nm to enhance charge transport by providing a stable interface that reduces charge recombination losses. An aluminum oxide (Al_2O_3; CAS number: 1344-28-1) layer will be included to improve corrosion resistance and reinforce mechanical stability, operating synergistically to enhance the overall electrochemical performance. The fabrication process will utilize an eco-friendly dual-step electrochemical deposition in a 0.5 M potassium sulfate electrolyte at a controlled temperature of 25°C, ensuring micro- and nanoscale porosity targeting 50-100 nm to maximize surface area as supported by literature demonstrating that this range optimally enhances charge transfer and ion migration efficiency. This degree of porosity is expected to lower charge transfer resistance significantly, fostering improved electrochemical kinetics, which will be verified using **scanning electron microscopy (SEM)** for monitoring thickness and porosity.

https://doi.org/10.52202/085713-2979

2. **Metal-Based Redox Couples**: The thermocell will utilize **copper/copper(I)** and nickel/nickel(II) redox couples, selected for their favorable redox potentials to minimize side reactions. An integrated cobalt co-catalyst (0.1 M) will serve as an effective stabilizing agent, enhancing electron transfer kinetics and maintaining the oxidation states of Cu^{2+} and Ni^{2+} during thermal cycling, as demonstrated by prior studies indicating its role in fostering electron transfer efficiency. Real-time monitoring will maintain pH levels between 4-7, with adjustable concentrations of redox couples systematically optimized between 0.5 to 1.5 M based on insights from the literature regarding their stability and reactivity under varying operational conditions, with specific methodologies for pH adjustments clearly defined to ensure minimal disruption during testing.
3. **Anisotropic Polymer Materials**: The polymer matrix will feature aligned functional groups (-COOH and $-SO_3H$), which will be developed through controlled mechanical stretching (5 mm/min at 70°C), a technique known to enhance ionic transport pathways by promoting favorable interactions between the polymer backbone and ionic species. High-conductivity ionic species, specifically Na^+ and Li^+, will be incorporated at optimized concentrations (0.5 to 1.0 M), with clear justifications based on their advantageous solvation properties and diffusion characteristics. The impact of variations in thermal conditions on ionic conductivity will be quantitatively assessed through impedance spectroscopy, with diffusion coefficients measured to elucidate the mechanisms of ionic mobility, providing a detailed connection between thermal gradients and polymer performance.

This approach clarifies the synergistic interactions and functional enhancements among hierarchical metal electrodes, optimized redox couples, and engineered polymer materials, establishing a cohesive mechanism that uses low-grade heat to drive efficient ion migration and electron transfer within the quasi-solid-state thermocell.

- Expert's Analysis: The study addresses the question of how to develop a cost-effective N-type quasi-solid-state thermocell to enhance electricity production from low-grade heat by improving both ion transport efficiency and electrode performance. The response provides a detailed and logically structured explanation, covering key aspects related to electrodes, redox pairs, and polymer gel media. Specifically, the preparation of Hierarchical Metal Electrodes is highlighted, noting the primary use of copper-containing electrodes, with a clear principle of providing a stable interface, which aligns well with the ground truth (gdth) description of 3D hierarchical copper electrodes. Additionally, the study includes specific characterization of the electrode surface using SEM. However, while the polymer gel mentions an anisotropic alignment structure, it omits the specific operational methods for forming this alignment, which is inconsistent with the gdth. Nevertheless, the overall response encompasses detailed mechanistic explanations and core points related to the specific operations.

Example 2:

- Research Question: How can a Supramolecular amorphous glass material be engineered to simultaneously exhibit strong mechanical properties, optical transparency, self-healing capabilities, and adhesive strength at room temperature?
- Hypothesis Candidate (from HHS): The engineering of multifunctional **peptide-based supramolecular amorphous glass** materials will concentrate on creating a robust dynamic network facilitated by cooperative reversible hydrogen bonds derived from interactions among four key amino acids: lysine, glutamic acid, aspartic acid, and tyrosine, along with proline to enhance rigidity. The hypothesized molar ratios will be set at 1:1 for lysine and glutamic acid, complemented by approximately 0.5:0.5 ratios for aspartic acid and tyrosine. These ratios are supported by empirical studies that have shown that such compositions can optimize hydrophilic and hydrophobic interactions, which are essential for improving mechanical **strength, adhesion**, and**self-healing capabilities** .

  Lysine's positively charged ammonium group is hypothesized to establish strong ionic interactions with the negatively charged carboxylate groups of glutamic acid and aspartic acid, enhancing the stability of the hydrogen-bonding framework critical for effective energy

dissipation during mechanical stress. Tyrosine will contribute to the network through π-stacking interactions, which are expected to maintain the structural integrity and optical transparency of the material under load. Proline's unique cyclic structure is anticipated to provide localized rigidity, supporting favorable peptide conformations and facilitating effective stress distribution throughout the dynamic network.

A pivotal component of this dynamic system will be the incorporation of **structured water** , maintained at an optimal concentration of 10–15% by weight. Structured water is theorized to engage in specific hydrogen bonding interactions with the peptide backbone, promoting molecular mobility and enabling rapid bond reformation necessary for self-healing capabilities at room temperature. The investigation will differentiate between structured and unstructured water forms, examining the specific interactions that influence bond lifetimes and recovery dynamics under mechanical stress. Previous research indicates that the interactions between structured water and the peptide backbone are critical for maintaining bond lifetimes during stress, and these dynamics will be quantitatively assessed through methods such as differential scanning calorimetry (DSC) and rheological measurements.

The synthesis of these peptide assemblies will be executed under rigorously controlled environmental conditions. The reaction temperature will be maintained at 25°C to optimize amino acid interactions and prevent degradation, while systematic variations in pH (5.5 to 8.5) will be conducted to explore their effects on the ionization states of the amino acids and corresponding **hydrogen bonding** dynamics. Additionally, ionic strength will be regulated at approximately 0.15 M using sodium chloride, which is expected to enhance electrostatic interactions and stabilize the hydrogen bonding network.

To thoroughly investigate these interactions and material properties, a combination of experimental methodologies will be utilized. Dynamic Mechanical Analysis (DMA) will assess mechanical properties such as tensile strength and elasticity, while rheological assessments will evaluate the material's response under stress. Spectroscopic techniques, including NMR spectroscopy and Fourier Transform Infrared (FTIR) spectroscopy, will be employed to elucidate hydrogen bonding dynamics and monitor molecular interactions. This comprehensive approach aims to clarify the intricate relationships among amino acids and structured water dynamics, as well as the influences of environmental conditions on the multifunctional properties of the engineered peptide-based supramolecular materials.

- Expert's Analysis: In addressing the research question—"How can a supramolecular amorphous glass material be engineered to simultaneously exhibit strong mechanical properties, optical transparency, self-healing capabilities, and adhesive strength at room temperature?"—the generated scientific hypothesis proposes a notably complex system. This system is envisioned to comprise five distinct amino acids: lysine, glutamic acid, aspartic acid, tyrosine, and proline.

  Despite the increased complexity of this multi-component approach compared to the simpler system underlying the real hypothesis (the scientific finding concerning YYY peptide glass), several key conceptual parallels are evident: Shared Foundation in Peptide-Based Materials: At their core, both the generated hypothesis and the real scientific finding are centered on peptide-based materials as the fundamental building blocks for the desired amorphous glass.Convergent Aim for Dynamic Networks and Functional Properties: Both frameworks leverage their respective peptide systems with the goal of establishing a dynamic network. This network is considered crucial for imbuing the material with critical functionalities, particularly self-healing capabilities and effective adhesive strength.Emphasis on the Role of Structural Water: In their mechanistic considerations, both hypotheses significantly highlight the indispensable role of structural water. The real hypothesis (the scientific finding on YYY glass) successfully demonstrated that a dense, random hydrogen-bonding network, mediated by water molecules, is fundamental to the YYY glass's unique structure and its observed properties. The generated hypothesis also underscores the centrality of cooperative and reversible hydrogen bonds in the construction and operational dynamics of its proposed network.

Example 3:

https://doi.org/10.52202/085713-2979

- Research Question: How can computational methods be used to accurately predict and improve the reactivity and selectivity of modular diazo transfer (MoDAT) reactions, especially reactions with primary amines? And to design new reagents based on computational models.
- Hypothesis Candidate 1 (from HHS): We propose to systematically investigate the reactivity and selectivity of modular diazo transfer (MoDAT) reactions utilizing azide-based reagents, with a specific focus on para-substituted benzyl azide derivatives modified with strong electron-withdrawing groups (EWGs) such as nitro (–NO_2) and cyano (–CN), as well as weaker electron-withdrawing groups (e.g., fluoro (–F) and chloro (–Cl)), and electron-donating groups (EDGs) like methoxy (–OCH_3). Our central hypothesis posits that the electronic nature and precise positioning of these substituents will significantly modulate the electrophilicity of the azide moiety, which will in turn influence the stability and geometrical configurations of intermediates and transition states during nucleophilic attacks by primary amines.

  The experimental work will be executed under controlled laboratory conditions using a Schlenk line to maintain an inert nitrogen atmosphere for at least 30 minutes prior to reaction initiation, minimizing moisture exposure. Reactions will be conducted at a temperature of 95–105°C, chosen based on literature findings indicating optimal kinetic performance while preserving the stability of diazo intermediates. We will employ polar aprotic solvents such as dimethylformamide (DMF) and dimethyl sulfoxide (DMSO), which are anticipated to enhance solvation of the azide and improve nucleophilicity of the primary amines. A stoichiometric ratio of 1:1.5 (benzyl halide to sodium azide) will be applied, and reactant concentrations will be maintained at approximately 10–20 mM, a range supported by preliminary studies demonstrating optimal reactivity and solubility.

  To deepen our mechanistic understanding, we will utilize advanced computational techniques, primarily Density Functional Theory (DFT) with specific emphasis on the B3LYP functional and a 6-31G(d) basis set. This will allow us to thoroughly assess the impacts of substituent variations on charge distributions and transition state energies. In particular, we will quantify how the introduction of strong EWGs influences the electrophilicity of the azide, predicted to lower activation energy for nucleophilic attacks, while the EDGs may provide stabilization during transition state formation. A rigorous electronic structure analysis will be employed to visualize the overlap between the nucleophilic orbital of primary amines and the electrophilic centers of the diazo derivatives.

  To ensure a comprehensive evaluation of solvent effects, we will systematically conduct reactions in both polar protic solvents, such as ethanol, and polar aprotic solvents, like DMSO, maintaining a concentration of 50% v/v relative to the reaction volume. This will elucidate how solvent polarity and hydrogen-bonding capacity impact the stability of zwitterionic intermediates and transition states. Furthermore, we will establish a robust iterative feedback mechanism that directly correlates our computational predictions to empirical data from high-resolution techniques, including NMR spectroscopy and mass spectrometry. Discrepancies exceeding a predefined threshold—such as a 10% difference between predicted and observed yields—will prompt specific adjustments to our computational models, thus refining our understanding of the factors influencing MoDAT reactions. This structured approach aims to enhance the design of azide-based reagents tailored for improved reactivity and selectivity in synthetic applications involving primary amines.
- Hypothesis Candidate 2 (Greedy + self-consistency): This study aims to investigate the reactivity and selectivity of modular diazo transfer (MoDAT) reactions by designing tailored azide-based reagents that incorporate strategically positioned electron-withdrawing groups (EWGs), specifically nitro (–NO_2) and cyano (–CN) moieties. These groups will be systematically evaluated at distances of 2 Å, 4 Å, and 6 Å from the azide functionality. This selection is grounded in existing literature, which suggests that such variations in EWG positioning can significantly influence the electrophilicity of the azide, thereby stabilizing transition states during nucleophilic interactions with primary amines. The specific rationale for these distances is based on prior studies indicating that alterations beyond these ranges may lead to diminished effects on reactivity or unfavorable sterics.

  The primary amines selected for this investigation—n-butylamine, phenethylamine, benzylamine, and ethylamine—exhibit varying degrees of steric hindrance and electronic characteristics, allowing for a comprehensive analysis of their reactivity profiles when paired with engineered azide reagents. This choice effectively captures a spectrum of nucleophilic

behavior and enhances our understanding of how both electronic and steric factors operate within MoDAT reactions.

An integrated computational strategy will be employed utilizing density functional theory (DFT) calculations focused on the B3LYP functional, supplemented by other functionals (e.g., M06-2X and PBE0) to ensure accuracy and validation of outcomes. In addition to DFT, we will incorporate transition state theory and potentially machine learning approaches to analyze the reactivity patterns further. We will conduct extensive mapping of activation energy profiles and transition state geometries to uncover the relationships between EWG positioning, electron density distributions, and transition state stability. Molecular dynamics (MD) simulations will also be utilized to explore solvation effects across differing solvent environments, including acetonitrile, methanol, and DMSO, examining how these solvents influence reaction pathways and stabilize charged intermediates.

Empirical validation of computational models will incorporate a structured approach to varying critical parameters such as azide reagent concentrations and molar ratios of primary amines to azides, along with solvent compositions, to derive quantitative metrics, including reaction yields, rate constants, and activation energies. Statistical analyses will employ techniques such as ANOVA and regression models to extract significant trends from the experimental data. This iterative feedback mechanism will facilitate a dynamic refinement process, whereby experimental outcomes directly inform adjustments to computational predictions. Through this comprehensive methodological framework, we aim to elucidate the interplay between EWG distances and steric factors, ultimately leading to the design of innovative azide-based reagents optimized for selective transformations of primary amines.

- Hypothesis Candidate 3 (Greedy): This research aims to systematically investigate the reactivity and selectivity of modular diazo transfer (MoDAT) reactions utilizing azide-based reagents, focusing on a set of primary amines: benzylamine, 2-aminopropane, and cyclohexylamine. This selection combines varying steric and electronic profiles, enabling comprehensive evaluation of how solvent and reaction conditions influence reactivity and selectivity across different nucleophilicity ranges. Initial studies will determine baseline reactivities through systematic kinetic measurements, assessing critical parameters such as rate constants and product ratios under controlled conditions.

  Reactant concentrations will be evaluated at specific increments of 0.1 M (0.1 M, 0.5 M, and 1.0 M), and the temperature will be optimized through a systematic approach involving stepwise assessments from 25°C to 60°C, analyzing how these variations affect reaction progress. A comprehensive assessment of solvent effects will be performed, including the examination of solvent mixtures (e.g., varying concentrations of water, DMSO, and possible co-solvents) to quantify their influence on nucleophilicity and overall reactivity.

  Advanced computational methods, including density functional theory (DFT) calculations with the M06-2X functional and a 6-31G basis set, will be employed to simulate the MoDAT reaction environment accurately. We will analyze key molecular descriptors such as nucleophilicity, electrophilicity, and steric hindrance to construct predictive models of reactivity. These analyses will guide experimental design, with a feedback mechanism where discrepancies between computational predictions and experimental observations will result in specific adjustments to molecular descriptors or computational parameters, refining the predictive capabilities of the models.

  Following these investigations, the design of innovative azide-based reagents will be undertaken to optimize MoDAT reactions. This design process will emphasize the incorporation of electron-withdrawing groups like trifluoromethyl and cyano, aimed at enhancing both stability and selectivity by stabilizing the transition state. Rigorous standardized experimental protocols will ensure reproducibility, including specific techniques for measuring yields and selectivity ratios over controlled reaction durations. By integrating mechanistic insights from computational and empirical findings, this research will elucidate the key factors influencing reactivity and selectivity in diazo transfer reactions, enhancing our understanding of these critical processes.

- Expert's Analysis: 1 conducted a relatively comprehensive analysis, for instance, suggesting that modifying the azide reagent with functional groups could improve it, which aligns with the original text. However, 2 and 3 did not. This time, 1 has an obvious error: the speculated temperature is incorrect, and the proposed temperature is experimentally unfeasible, as azide reagents are prone to explosion at high temperatures. Of course, temperature is a minor

 https://doi.org/10.52202/085713-2979

point, and overall, 1 is still acceptable. 2 deviates significantly from the original text in terms of the research design approach. Compared to 2, 3 lacks consideration of the group effect in the research design, making 3 the weakest. Finally, all three mentioned using DFT calculations, and although there are deviations in details from the original text, the approach is correct.

Example 4:

- Research Question: How can photoredox catalysis be used to exploit the latent reactivity of phosphorus ylides, allowing them to participate in a formal three-component cycloaddition that converts inert C–H and C=P bonds into C–C and C=C bonds, creating versatile synthetic building blocks in an efficient, controlled manner?
- Hypothesis Candidate 1 (from HHS): The mechanism for activating phosphorus ylides in a formal three-component cycloaddition via photoredox catalysis can be articulated in four key steps, each supported by optimized experimental conditions:

  1. Initiation of Single-Electron Transfer (SET): Irradiation of phosphorus ylides with specific wavelengths of visible light (400–450 nm) from a high-intensity LED source (approximately 20 mW/cm^2), validated by studies demonstrating effective radical generation at this intensity (Smith et al., 2020), promotes SET using suitable photoredox catalysts (e.g., [Ru(bpy)$_3$]$^{2+}$ or [Ir(dF(CF$_3$)ppy)$_2$(bpy)]). The resulting radical cation exhibits enhanced electrophilicity due to significant charge localization, which is further assisted by strong electron-withdrawing substituents such as carbonyl or nitro groups. Empirical evidence indicates an increase in reactivity by up to 2.5-fold as supported by Hammett parameters.

  2. Stabilization via Zwitterionic Intermediate: The radical cation transitions to a zwitterionic intermediate, characterized by resonance stabilization through delocalized π-electrons and non-covalent interactions, such as hydrogen bonding in polar aprotic solvents like acetonitrile (dielectric constant ≈ 37) and DMSO (dielectric constant ≈ 47). To optimize stabilization, a 1:1 (v/v) mixture of these solvents will be used, taking advantage of their combined dielectric properties (≈ 38) to enhance charge separation and stabilize reactive intermediates. Literature supports this approach, showing improved reaction kinetics (Miller et al., 2021).

  3. Selective Nucleophilic Attack: The zwitterionic intermediate selectively engages in nucleophilic attacks on activated C–H and C=P bonds, particularly those adjacent to strong electron-withdrawing groups. Maintaining phosphorus ylide concentrations at 0.1-0.5 M and controlling reaction temperatures precisely within an optimized range of 10-25 °C, as indicated by previous studies on radical stability, will minimize side reactions. An inert atmosphere (nitrogen or argon) will be established by purging the reaction vessel for 30 minutes before use, effectively mitigating oxidation. Real-time NMR (utilizing 1D and 2D techniques) and GC-MS metrics will be employed to monitor yield and product distribution effectively, specifying analytical conditions (e.g., temperature settings and flow rates) to ensure accurate assessment of outcomes.

  4. Concerted Formation of Products: The reaction culminates in the concerted formation of new C–C and C=C bonds, facilitating the synthesis of valuable carbocycles and synthetic building blocks. The influence of substituent identity and positioning (ortho, meta, para) on reactivity will be quantitatively analyzed using NMR and HPLC techniques. This systematic approach will provide insight into the efficiency and selectivity of the cycloaddition process, explaining how each factor contributes to overall reactivity.

  By integrating these components clearly and methodically, this hypothesis presents a comprehensive exploration of how photoredox catalysis can unveil new reactivity pathways for phosphorus ylides, fully addressing the research question with explicitly defined roles of each mechanistic step and comprehensive definitions for specialized terms provided for clarity.
- Hypothesis Candidate 2 (Greedy + self-consistency): This study aims to investigate how photoredox catalysis can elucidate specific reactivity mechanisms in diphenylphosphinyl ylides, focusing on their participation as intermediates in formal three-component cycloaddition reactions that convert inert alkyl C–H bonds and coordinated C=P bonds into C–C and C=C bonds. We will concentrate on diphenylphosphinyl ylides due to their capacity to facilitate charge inversion processes, which are crucial for generating stabilized radical

intermediates. Specifically, we will examine how the electronic structure of diphenylphosphinyl ylides transforms upon excitation under visible light (e.g., at wavelengths of 400-450 nm) into polar intermediates, leading to the generation of stable benzylic radicals from C–H bond cleavage and allylic radicals from C=P bond activation—both essential for effective cycloaddition.

1. Charge Inversion Mechanisms: We will clarify the mechanisms of charge inversion by detailing the electronic transitions involved, emphasizing the role of excited state lifetimes and how solvent polarity affects the stabilization of generated radicals. Our discussion will include specific references to known photoredox systems and their operational parameters in literature, enhancing our findings with context.

2. Role of π-Stacking Interactions: We will investigate the influence of π-stacking interactions, specifying the optimal inter-ring distances and energy considerations between the stabilized radicals and electrophilic carbon centers from substrates like 2,6-dibromoaryl and 2,4-difluoroaryl. This section will include quantitative assessments of radical coupling reactivity, aiming to determine coupling efficiencies and product selectivity through methods such as ^{1}H NMR and UV-Vis spectroscopy.

3. Solvent and Temperature Effects: We will analyze solvent effects by conducting reactions in a range of solvents with varied dielectric constants, including a thorough comparison of polar solvents like acetonitrile and non-polar solvents like dichloromethane. We will detail how these solvents are expected to influence radical stabilization and the kinetics of cycloaddition, providing a theoretical framework based on established solvent interaction models. Additionally, we will justify the controlled temperature range of room temperature to 50°C by linking it to the expected stability of radical intermediates and the kinetics of the reactions, ensuring optimal conditions for product yield and selectivity.

4. Experimental Conditions: Reactions will be conducted under carefully controlled conditions, with temperature justification focusing on maintaining the balance between radical stability and reaction kinetics. We will outline how these conditions directly relate to the completed cycloaddition mechanism and the anticipated outcomes of the study.

By systematically elucidating these mechanisms—specifically charge inversion, π-stacking interactions, and solvent effects—we aim to develop robust methodologies for the efficient generation of versatile synthetic building blocks from simple molecular precursors.

- Hypothesis Candidate 3 (Greedy): Investigate how photoredox catalysis enhances the reactivity of phosphorus ylides through selective nucleophilic attack on α, β-unsaturated carbonyl compounds, such as crotonaldehyde, which are activated to form stable radical cation intermediates via photoredox-driven single-electron transfer (SET) processes. These radical cations, characterized by their electrophilicity, promote effective nucleophilic attacks by phosphorus ylides, generating stabilized carbon radical intermediates that significantly enhance their reactivity in subsequent bond-forming transformations. Conduct a formal three-component cycloaddition by introducing a nucleophilic amine, such as ammonia or an aniline derivative, selected based on its electronic properties which influence the stabilization of the radical intermediates and affect product selectivity. Detail specific optimized reaction conditions, including the use of polar aprotic solvents like acetonitrile, which facilitate radical stability, and employ a specific light wavelength of 400 nm to ensure efficient excitation of the photocatalyst. These conditions will be designed to minimize potential side reactions and maximize the conversion of inert C–H and C=P bonds into desired C–C and C=C bonds through well-defined mechanistic pathways, addressing the nuanced interplay between reaction parameters and final product outcomes.

- Expert's Analysis: 1 accurately predicted the light source wavelength range, metal catalyst system, and solvent system, such as the use of Ir catalyst and acetonitrile as the solvent, all of which align with the original text. In contrast, 2 only correctly predicted the wavelength range and solvent system but failed to specify the metal catalyst system, which is crucial in organic chemical reactions. Therefore, 2 is inferior to 1. Finally, 3 did not predict the light source wavelength range or the metal catalyst system, missing several key pieces of information, making it the weakest.

https://doi.org/10.52202/085713-2979

E Expert Analysis of Hypothesis Quality

E.1 Convergence to Ground-Truth Local Optima

To complement our quantitative evaluations, we asked domain experts to qualitatively assess how well the hypotheses generated by *HHS* aligned with the expert-annotated fine-grained hypotheses in our benchmark.

The distribution of expert assessments across all evaluated examples is as follows:

- **Reached a completely different region—likely a distinct local optimum with scientific plausibility:** 60%
- **Reached the vicinity of the ground-truth local optimum, but with differing details:** 24%
- **Reached the vicinity of the ground-truth local optimum, but failed to fully elaborate or specify the key details:** 16%

Here, the "ground-truth local optimum" refers to the expert-extracted fine-grained hypothesis from a publication, which serves as the reference target. "Reaching the vicinity of a local optimum" indicates that the generated hypothesis converges to a coherent and internally consistent formulation that is conceptually close to the ground-truth hypothesis, though not necessarily identical in detail.

The relatively high divergence rate (60%) reflects an inherent tradeoff in our experimental setup. For many research questions, multiple hypotheses can be plausible yet structurally distinct. Guiding the model toward the exact ground-truth hypothesis requires:

1. initializing the search process from a starting point sufficiently close to the ground-truth optimum;
2. but avoiding initialization that is too close or too specific, as this would risk leaking the ground-truth answer.

To strike this balance, we derive the initial search point from the annotated coarse-grained hypothesis h_c by applying an *ambiguation* procedure. This involves removing or abstracting key details to produce a generalized version of h_c—for example, replacing "a specific protein" with "a protein" or "a catalyst"—thus preserving the overall research direction while preventing answer leakage.

Consequently, even when the search begins in the correct conceptual region, the model may naturally diverge toward a nearby but distinct local optimum, especially given the openness of the hypothesis space and the heuristic-driven nature of the optimization process.

E.2 Coverage of Experimentally Critical Details

In addition to alignment with the reference hypotheses, we evaluated the extent to which the generated hypotheses captured the critical experimental details required for practical implementation.

Among all the details mentioned in the generated hypotheses, approximately **40% are experimentally important**—regardless of their accuracy (which is not the focus of this analysis). The remaining **60% are peripheral or have minimal impact** on the actual experiment.

Among all the important details that should be included, about **50% are mentioned** in the generated hypotheses.

Peripheral details refer to contextual or environmental factors with limited relevance to the core experiment—for instance, ambient humidity or weather conditions, which may only affect specific reactions.

This highlights a key challenge: while LLMs can generate rich and context-aware hypotheses, they often fail to prioritize the most essential components for experimental planning. Future work may explore techniques to guide LLM attention toward experimentally salient information.

Table 5: Error analysis of hypotheses generated by HHS. Two PhD-level chemistry experts conducted the evaluation: one analyzed the top 30 samples and the other the remaining 21.

Missing key chemical substances	14/30
Excessive details in characterization methods	28/30
Feasibility issues	18/30
Limitations of characterization methods	08/30
Insufficient basis for material selection	22/30
Lack of design comparison experiments	12/30
Ignoring data validation and reproducibility	10/30
Severe deviation from feasibility	8/21
Missing or incorrect key chemicals or reaction systems	9/21
Incorrect explanation of chemical principles	12/21
Incorrect prediction of experimental system	10/21

Table 6: Error analysis on why HHS's hypotheses are better than the greedy search baselines'. Two PhD-level chemistry experts conducted the evaluation: one analyzed the top 30 sample pairs and the other the remaining 21.

Insufficient performance metrics	25/30
Complexity of experimental conditions	16/30
Insufficient explanation details	29/30
Inadequate description of preparation plan	22/30
Vague research objectives	28/30
Cost and scalability issues	13/30
Poor feasibility	12/21
Errors in research plan details	21/21
Insufficient explanation details	19/21
Clear experimental system	21/21

F Error Analysis

Two PhD-level chemistry experts conducted the error analyses. Table 5 summarizes the main error types observed in hypotheses generated by HHS, while Table 6 analyzes the reasons why HHS outperforms the greedy search baseline.

G Hypothesis Search Prompt

The following prompt is used to guide both the baseline methods and our proposed method, *HHS*, during the hypothesis refinement process. To ensure fair comparison, the prompt is designed in a controlled way: we use a shared core prompt across all methods, with minimal differences. Specifically, the portion highlighted in orange is unique to *HHS* and introduces the hierarchical structure used in its search process.

This design isolates the effect of hierarchical search. As illustrated below, the only difference between *HHS* and the baseline (*Greedy Search + Self-Consistency*) lies in the hierarchical prompting. The core content—including the role of the assistant, editing instructions, and structural expectations—is kept identical.

This enables a controlled ablation-style comparison, attributing observed improvements specifically to the hierarchical design.

https://doi.org/10.52202/085713-2979

The complete prompt is as follows:

You are assisting with scientist's research. Given their research question, a survey on the past methods for the research question, and a preliminary coarse-grained research hypothesis for the research question, please help to make modifications into the coarse-grained hypothesis, to make it one step closer to a more effective and more complete fine-grained hypothesis.

The modification can be two-folds: (1): delete or change an existing improper detail or information in the existing hypothesis; (2) add and integrate one detail to the existing hypothesis. If you choose to add a detail, do not simply append new information to the existing hypothesis. Instead, think thoroughly how the new detail relates to the existing components and integrate it seamlessly into the hypothesis to create a new coherent and unified hypothesis. In addition if you choose to add a detail to a general information, if the corresponding general information is correct, you should try to keep the corresponding general information in the updated hypothesis and also mention the details, instead of replacing the general information with the details. In this way, it would be much easier for scientists to understand both the general infomration/structure and the details from your generated hypothesis. It would be also easier for scientists to propose better details, inspired by your suggested details, following the general information.

Please remind that this is about research: research is about discover a new solution to the problem that ideally is more effective and can bring new insights. Usually we don't need the hypothesis to contain lots of known tricks to make it work better: we want to explore the unknown, which ideally is more effective than the known methods and can also bring in new insights. Therefore, a research hypothesis is usually about a small set (usually less than eight) of major components (and lots of details on how to implement these major components), which overall composes a novel and complete solution to the research question, which potentially can bring in new insights. Hypotheses that include an excessive number of irrelevant or unnecessary major components, which do not contribute to addressing the research question, are less favorable, as we only want to know exactly what are the key components that fundamentally make the hypothesis work. If you think any ancillary components that can truly assist with the research question, you may mention what are the key components and what are the ancillary components to avoid the ambiguity of which components are the key component. The reaction mechanism, however, is not classified as a major component or detail (and therefore not limited by the number of major components). Instead, a novel and valid reaction mechanism can be a good source of insights. If previous hypothesis already contains too many major components, you should consider to replace some of the major components with more effective ones (but not to add more major components), or to give more details to the existing major components for clarity and ease of implementation (instead of adding or replacing major components).

Here we are searching for the fine-grained hypothesis in a hierarchical way. The rationale is that, we can classify any complete set of modifications into several hierarchy, with different levels of details. If we do not search in a hierarchical way, we need to consider all the available details in all hierarchy levels for each search step, which (1) has a very high complexity, and (2) first search a low-level detail might largely influence the following search of a high-level detail: it might stuck in one high-level detail corresponding to the already searched low-level detail without considering the other low-level details corresponding to other high-level details, making the search process stuck in a local minumum at the beginning.

Here we roughly classify all possible modifications into five hierarchies: (1) Mechanism of the Reaction: Describes how the reaction proceeds at a conceptual level, focusing on electron flow, bond formation and breaking, and any intermediates or transition states involved. This is the theoretical "blueprint" that explains why the reaction works; (2) General Concept or General Component Needed: Identifies the type of reagent or functional group required (e.g., "a strong acid," "a Lewis base," "an activated aromatic ring") without committing to a specific chemical. It outlines the broader roles that are necessary for the mechanism to proceed; (3) Specific Components for the General Concept: Narrows down from the general category to a particular substance (e.g., "concentrated HCl" for a strong acid, "benzene" for an aromatic ring). This makes the reaction hypothesis testable by specifying which chemicals fulfill the roles; (4) Full Details of the Specific Components: Provides exact structural or molecular information—such as SMILES strings, IUPAC names, purity, or CAS numbers. These details ensure clarity and reproducibility so researchers know precisely which substances to use; (5) Experimental Conditions: Specifies the practical setup—temperature, pressure, solvent system, reaction time, atmosphere, and any work-up procedures. This final layer describes

how to carry out the reaction in a laboratory setting. And we are searching for modifications hierarchy by hierarchy: hierarchy (1) first, and then hierarchy (2), and so on. Hypothesis from a higher hierarchy is an expansion of the hypothesis from its previous hierarchy, with additional information described above.

The research question is:

The survey is:

Now please help to make modifications into the coarse-grained hypothesis, to make it one step closer to a more effective and more complete fine-grained hypothesis. Please do not include the expected performance or the significance of the hypothesis in your generation. Please answer the question in the following response format. (response format: 'Reasoning Process: Revised Hypothesis: ')

H Experiment Compute Resources

We implement our proposed framework as an agentic LLM system using `GPT-4o-mini` using OpenAI's official API. Generating the final hypothesis via the *HHS* optimization process—converging to the final local optimum at hierarchy level 5—typically involves several hundred or even to a thousand iterative search steps.

I Limitation

While *HHS* consistently discovers higher-quality local optima compared to baseline methods, it does not guarantee convergence to the global optimum. Addressing this limitation remains an open direction for future research.

https://doi.org/10.52202/085713-2979

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, we have summarized the contributions in the end of the introduction section. They accurately reflect the scope of this paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mention it in the conclusion section: "The limitation is that *HHS* is not guaranteed to reach to a global optimum, and we leave it as a future research direction."

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: In Section 2, we provide detailed information about the assumptions and corresponding proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed information about the experimental configurations and evaluation protocols. More details can be found in the supplement.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/085713-2979

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code, data, and instructions can be found in https://github.com/ZonglinY/MOOSE-Chem2

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All details about hyperparameters are provided in the supplement.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All details about statistical significance are provided in the supplement.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We mention that we use "GPT-4o-mini". More details are mentioned in the supplement.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We check the code of Ethics, and we don't think this research has violated it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Discussions about social impacts are provided in the supplement.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/085713-2979

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets used in this paper are properly credited and respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code, data, and instructions can be found in https://github.com/ZonglinY/MOOSE-Chem2

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

https://doi.org/10.52202/085713-2979

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development does not involve LLMs as any important components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.