MOOSE-Chem3: Closed-Loop Hypothesis Ranking with Simulated Feedback and Component Reasoning

Anonymous ACL submission

Abstract

Large language models have shown remarkable promise in generating scientific hypotheses. However, the high cost and slow turnaround of wet-lab validation often restrict hypothesis evaluation to static, pre-experiment assessments, impeding rapid iterative refinement through real experimental feedback. To address this bottleneck, we introduce MOOSE-Chem3, a closed-loop optimization framework for efficient scientific hypothesis screening. MOOSE-Chem3 (MC3) comprises two core components: (1) MC3-Sim, a mechanism-aware simulator that serves exclusively as a development-time surrogate to overcome the feedback scalability gap, not as a replacement for wet-lab validation. During deployment, the same ranking policy seamlessly transitions to real experimental feedback ($E_{\text{sim}} \rightarrow E_{\text{lab}}$). (2) MC3-Rank, an in-context reinforcement learning agent that decomposes hypotheses into functional components and attributes feedback to them, transferring evidence across hypotheses with shared mechanisms; it learns from simulated feedback during development and leverages wet-lab outcomes during deployment to refine hypothesis rankings. Evaluated on TOMATO-chem and a real wet-lab outcome sequential deployment setting, MC3-Rank outperforms static, adaptive, and human expert baselines, establishing an efficient and scalable paradigm for feedback-driven scientific discovery. [1]

1 Introduction

Scientific discovery is a cornerstone of societal and technological advancement (Coccia, 2019). The rise of large language models (LLMs) has ignited interest in automating scientific workflows, particularly the generation of novel, testable hypotheses (Luo et al., 2025). Yet hypothesis generation alone is insufficient; a critical, underexplored challenge lies in hypothesis ranking: *given a large set of AI-proposed candidates, which should be tested first under severely limited experimental budgets?*

In the natural sciences, wet-lab validation is expensive, slow, and low-throughput, making the choice of initial experiments pivotal: an effective ranking strategy must maximize early discovery while minimizing wasted trials. Current practice largely relies on pre-experiment ranking, scoring hypotheses statically via LLMs' internal reasoning without empirical feedback (Yang et al., 2025c; Si et al., 2024)—efficient, but inherently non-adaptive. Feedback-driven methods such as Bayesian optimization (Han et al., 2025) and active learning (Kozyrev et al., 2025) offer an alternative, but typically assume continuous or candidate-level vectorized search spaces, ill-suited for scientific hypotheses, which are discrete, combinatorial, and semantically rich. A separate line of work exploits feedback loops in domains with fast, low-cost verifiers such as mathematics and algorithm discovery (Romera-Paredes et al., 2024; Shojaee et al., 2024; Novikov et al., 2025; Weng et al., 2025)—a strategy unavailable in natural sciences. To leverage experimental outcomes for dynamic re-ranking, we need a new paradigm that treats each experiment not just as a binary validator but as a source of transferable insight: by discerning which functional components contribute to success, we can prune swathes of the combinatorial search space. Yet scaling such experiment-guided ranking with real experiments remains impractical—physical experiments cannot support the repeated interaction required for policy learning.

To overcome this bottleneck (Liao et al., 2025; Cheng et al., 2025), we propose a shift in perspective: instead of waiting for scarce real-world feedback, we develop a simulation-based proxy that captures the informational essence of wet-lab results, enabling scalable, closed-loop development of adaptive ranking policies while respecting the constraints of real experimentation. To realize

[1] All code and data can be found at https://anonymous.4open.science/r/MC3.

experiment-guided hypothesis ranking at scale, we introduce MOOSE-Chem3, a closed-loop framework comprising two synergistic components:

1. MC3-Sim, a surrogate environment that emulates the informational structure of wet-lab feedback by modeling hypothesis performance as a function of latent similarity to a hidden reference anchor.

2. MC3-Rank, a decomposition-based in-context reinforcement learning agent that re-ranks hypotheses by attributing experimental outcomes to reusable functional components.

Evaluated on a benchmark of real-world natural science hypotheses spanning chemistry, material science, physics, and biology, MOOSE-Chem3 identifies high-performing candidates within approximately 15 trials, reducing the experimental burden by about 50% over strong pre-experiment baselines, and further outperforms adaptive and human-expert baselines under real wet-lab outcome deployment. In summary, our contributions are:

- We formulate *experiment-guided hypothesis ranking* as a closed-loop learning problem, and address the feedback scalability bottleneck via a principled simulation paradigm.

- We propose MOOSE-Chem3, integrating a mechanism-aware simulator (MC3-Sim) with a decomposition-based ranking agent (MC3-Rank) that transfers feedback across hypotheses with shared functional components.

- We show that MOOSE-Chem3 surpasses static, adaptive, and human-expert baselines on both the TOMATO-chem benchmark and a real wet-lab sequential deployment.

2 Method

2.1 Overview

Identifying the optimal hypothesis h^* under stringent experimental budgets necessitates dynamic ranking strategies that adapt based on prior empirical outcomes. This stands in stark contrast to static pre-experiment ranking approaches that rely solely on a language model's internal priors without incorporating real-world feedback (Yang et al., 2025c). However, the development of such adaptive strategies is hindered by a fundamental bottleneck: real wet-lab experiments are severely throughput-limited, rendering conventional iterative policy learning impractical.

To overcome this limitation, we propose a simulator-mediated framework that explicitly decouples policy development, conducted in simulation, from policy deployment, executed in the wet lab. As illustrated in Figure 1, our approach bridges the feedback scalability gap through two synergistic components:

- **The Agent (MC3-Rank).** An in-context reinforcement learning agent that learns a ranking strategy π (§2.3). By iteratively incorporating experimental feedback, the agent narrows the hypothesis space and focuses on identifying high-performing candidates.

- **The Environment (MC3-Sim).** A domain-informed simulator grounded in three conceptual foundations (§2.4), which approximates experimental outcomes via functional decomposition and mechanism-aware similarity.

Together, these components form MOOSE-Chem3, which provides a practical and scalable solution for hypothesis screening under limited experimental resources.

2.2 Problem Formulation

Given a research question q, let $\mathcal{H}$ denote a candidate hypothesis set comprising both machine-generated hypotheses from scientific discovery systems (Yang et al., 2025c) and human-curated reference hypotheses from top-tier journals. Each hypothesis $\mathbf{h} \in \mathcal{H}$ is evaluated by an experiment executor $E : \mathcal{H} \to [0, 1]$, which maps $\mathbf{h}$ to a normalized performance score $s \in [0, 1]$. This score is calibrated against domain-specific state-of-the-art benchmarks to ensure cross-problem comparability. The objective is to identify the optimal hypothesis:

$$h^* = \arg\max_{h \in \mathcal{H}} E(h). \tag{1}$$

using as few experimental trials as possible, denoted N_{trials}. Achieving this requires a stateful strategy π that actively incorporates feedback from previously evaluated hypotheses to guide future selections, effectively balancing exploration and exploitation under extreme data scarcity.

The Scalability Bottleneck. Conventional reinforcement learning requires thousands of environment interactions, which is practical in digital domains with cheap feedback but untenable in

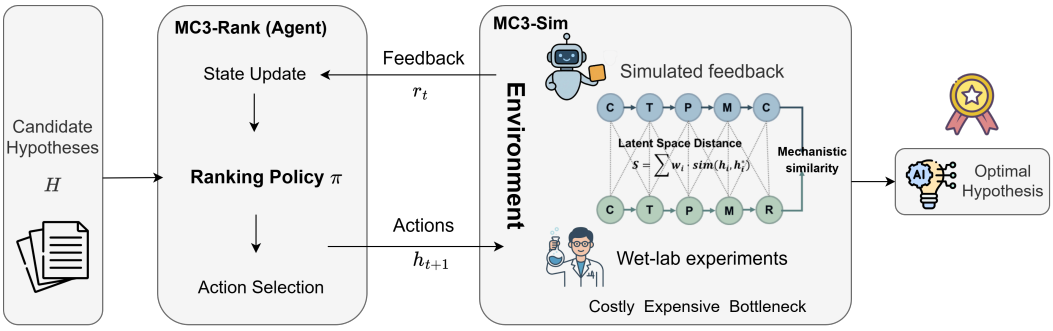


Figure 1: The overall MOOSE-Chem3 framework: the agent-environment interaction loop.

the natural sciences, where each wet-lab experiment is slow and costly. This mismatch creates a *feedback scalability gap*: real experiments provide high-fidelity signals yet cannot scale, while pre-experiment ranking avoids cost but forgoes empirical feedback. We therefore reformulate the problem as learning a strategy π that minimizes the number of experimental trials N^{π}_{trials} needed to identify the optimal hypothesis h^*.

2.3 The Agent (MC3-Rank)

Given costly wet-lab validation, **MC3-Rank** treats hypothesis ranking as a sequential decision-making problem: it selects hypotheses sequentially, learns from experimental feedback, and gradually improves its selection strategy. Figure 1 provides an overview of the interaction between the agent and the experimental environment.

Sequential Decision Formulation. We formulate experiment-guided ranking as a sequential decision process. At each timestep t, the agent observes a state s_t, which summarizes the results of all experiments conducted so far. Based on this state, the agent selects an action a_t by choosing a hypothesis $h_t \in \mathcal{H}$ to evaluate. The experiment executor then runs the experiment and returns an outcome, which is used as a reward r_t. The resulting interaction trajectory can be written as

$$\tau = [s_0, a_0, r_0, s_1, a_1, r_1, \ldots]. \tag{2}$$

Unlike standard reinforcement learning settings that aim to maximize cumulative reward, our objective reflects the high cost of scientific experiments. Specifically, the agent aims to identify the optimal hypothesis using as few experimental trials as possible. Formally, the goal is to learn a policy π^* that minimizes the expected number of trials:

$$\pi^* = \arg\min_{\pi} \ \mathbb{E}_{\pi}[N_{\text{trials}}]. \tag{3}$$

A good policy $\pi(s_t)$ therefore leverages past experimental outcomes to make increasingly informed hypothesis selections.

Iterative Ranking Procedure. MC3-Rank implements the policy π through a structured, multi-phase procedure, illustrated in Figure 2. Each iteration consists of four main phases.

Phase 1: State Construction. The agent decomposes each hypothesis $h \in \mathcal{H}$ into functional components, classifies them as effective, uncertain, or ineffective based on feedback, prunes ineffective ones, and clusters the rest by functional similarity. Each cluster represents a distinct mechanism that may contribute to solving the research question q. Together, these clusters form a structured and interpretable state representation s_t.

Phase 2: Action Selection. Given state s_t, the agent selects the next hypothesis through a two-stage process. It first identifies the most promising cluster based on accumulated evidence and domain knowledge, then chooses a hypothesis $h_t \in \mathcal{H}$ within that cluster as the action a_t. This connects mechanistic insights with experimental decisions.

Phase 3: Reward Observation. Hypothesis h_t is evaluated by the experiment executor E, either MC3-Sim or wet-lab execution; E returns a normalized reward $r_t = E(a_t) \in [0, 1]$. This provides quantitative feedback on the selected hypothesis.

Phase 4: State Update. After receiving the outcome, the agent updates its summary of past experiments. The updated summary forms the next state s_{t+1} for subsequent selection, enabling better decisions over time.

Interpretability and Transparency. As shown in Algorithm 1, MC3-Rank is interpretable by design. By decomposing hypotheses into functional components and clustering them, the agent yields a transparent decision process. Each selection can be traced to specific components and outcomes,

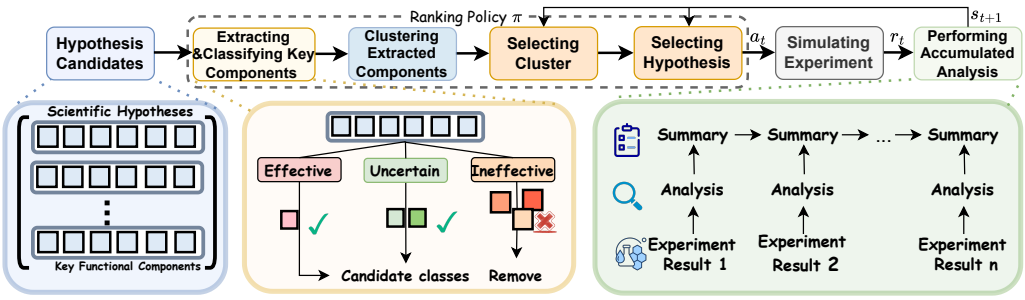


Figure 2: Experiment-guided ranking policy within an in-context reinforcement learning framework.

Algorithm 1: MC3-Rank

Input : question q; hypotheses $\mathcal{H}$; executor E; budget B
Output : best hypothesis $\mathbf{h}^*$

$s_0 \leftarrow \emptyset$; $\mathcal{H}_{\text{eval}} \leftarrow \emptyset$
for $t \leftarrow 0$ **to** $B-1$ **do**
  State
  $\mathcal{Z}_t \leftarrow$ PRUNE(CLASSIFY(DECOMPOSE($\mathcal{H}, q$), s_t))
  $\mathcal{G}_t \leftarrow$ CLUSTER($\mathcal{Z}_t$); $s_t \leftarrow$ SUMMARIZE($\mathcal{G}_t, s_t$)
  Action
  $g_t \leftarrow$ SELECTCLUSTER($\mathcal{G}_t, s_t$); $\mathbf{h}_t \leftarrow$ SELECTHYPOTHESIS($g_t, \mathcal{H}$)
  Reward & update
  $r_t \leftarrow E(\mathbf{h}_t)$; $\mathcal{H}_{\text{eval}} \leftarrow \mathcal{H}_{\text{eval}} \cup \{(\mathbf{h}_t, r_t)\}$; $s_{t+1} \leftarrow$ UPDATE($s_t, \mathbf{h}_t, r_t$)
  if STOP($s_{t+1}, \mathcal{H}_{\text{eval}}, B$) **then**
    break
  end
end
$\mathbf{h}^* \leftarrow \arg\max_{(\mathbf{h},r)\in\mathcal{H}_{\text{eval}}} r$
return $\mathbf{h}^*$

supporting expert analysis and trust.

2.4 The Environment (MC3-Sim)

Wet-lab feedback is often expensive, slow, and data-scarce, which makes it difficult to develop and stress-test hypothesis screening methods purely through real experiments. We therefore introduce **MC3-Sim**, a simulator that instantiates the experiment executor in our framework. It follows the same interface as wet-lab evaluation:

$$E : \mathcal{H} \to [0,1], \qquad r_t = E(a_t), \tag{4}$$

but uses a low-cost simulation function E_{sim} during development. In each trial, the agent selects an action $a_t \triangleq h_t \in \mathcal{H}$, and the simulator returns a normalized performance score

$$r_t = E_{\text{sim}}(h_t \mid q, h^*) \in [0,1], \tag{5}$$

which is used directly as the reward in MC3-Rank. Here h^* denotes a reference anchor hypothesis for the research question q. In simulator-based development, h^* is the benchmark target used only to instantiate E_{sim}; MC3-Rank observes only executor feedback. With real experiments, E_{sim} is replaced by E_{lab}, and no anchor is required. We formulate **three conceptual foundations** that serve as the theoretical basis for our simulator (see Appendix A for the complete formalization):

1. (*A1: Local Optimum Assumption*) There exists a neighborhood $\mathcal{N}(h^*) \subseteq \mathcal{H}$ in which h^* is a dominant solution. Intuitively, hypotheses sufficiently close to h^* in mechanism tend to perform well, and hypotheses farther away tend to perform worse.

2. (*P1: Similarity Principle*) Hypotheses that are more similar in their underlying structure or function tend to yield more similar outcomes. Equivalently, performance varies smoothly with a suitable notion of mechanism-level closeness.

3. (*D1: Representation Distortion*) Any practical representation of hypothesis similarity is imperfect. This imperfection distorts the closeness between hypotheses, which can introduce noise, secondary peaks, and unexpected valleys in the performance landscape.

Collectively, these foundations theoretically justify the construction of a low-cost proxy for experimental validity. Guided by these principles, we propose the MC3-Sim pipeline, which evaluates a hypothesis by quantifying its alignment with the reference anchor at the mechanism level.

MC3-Sim Pipeline. MC3-Sim produces r_t by comparing the candidate hypothesis h_t with the reference h^* at the level of functional mechanisms. It also outputs an explanation trace (components, alignments, and per-component similarities), which

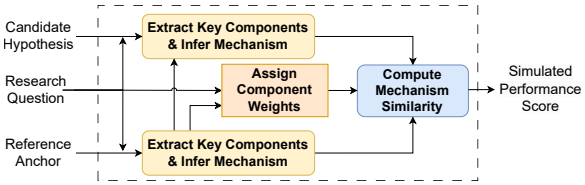


Figure 3: The internal structure of the simulator.

supports analysis and debugging. We summarize and depict the simulator structure in Figure 3.

Phase 1: Reference Construction. Given (q, h^*), MC3-Sim decomposes the reference anchor into K functional components $\{c_i^*\}_{i=1}^K$ and infers their mechanisms under q. It assigns each component a non-negative weight w_i, with $\sum_{i=1}^K w_i = 1$, and identifies a critical subset $\mathcal{C} \subseteq \{1, \ldots, K\}$ treated as necessary for success.

Phase 2: Candidate Parsing and Alignment. Given (q, h_t), MC3-Sim decomposes the candidate into components $\{c_i\}_{i=1}^K$ and aligns them with the reference components $\{c_i^*\}_{i=1}^K$. This ensures consistent component granularity between h_t and h^*.

Phase 3: Component Similarity Scoring. For each aligned pair (c_i^*, c_i), MC3-Sim computes a similarity score $s_i \in [0, 1]$ measuring the candidate's functional and mechanistic match to the reference component under q. Scores near 1 indicate strong agreement, while scores near 0 indicate missing or mismatched components.

Phase 4: Aggregation to Executor Output. MC3-Sim aggregates per-component similarities into a simulated performance score:

$$S(h_t \mid q; h^*) = \left(\prod_{i \in \mathcal{C}} \mathbf{1}_{\{s_i > 0\}} \right) \cdot \left(\sum_{i=1}^{K} w_i s_i \right). \quad (6)$$

where $\mathbf{1}_{\{s_i>0\}}$ enforces all critical components. By construction, $S(h_t \mid q; h^*) \in [0, 1]$ and $S(h^* \mid q; h^*) = 1$ under perfect alignment. We set $E_{\text{sim}}(h_t \mid q, h^*) \triangleq S(h_t \mid q; h^*)$ and $r_t = E_{\text{sim}}(h_t \mid q, h^*)$. Along with r_t, MC3-Sim returns a trace containing $\{c_i^*\}$, $\{c_i\}$, $\{w_i\}$, $\mathcal{C}$, and $\{s_i\}$, enabling component-level inspection.

Connection to MC3-Rank. As shown in Algorithm 2, MC3-Sim provides development-time feedback: MC3-Rank proposes h_t, MC3-Sim returns r_t, and the agent updates s_{t+1} using (h_t, r_t). When real experiments are available, E_{sim} can be replaced by a wet-lab executor E_{lab} without changing the agent pipeline.

Algorithm 2: MC3-Sim

Input : question q; anchor h^*; candidate h
Output : reward $r \in [0, 1]$; trace $\mathcal{T}$

Reference construction
$\mathbf{c}^* \leftarrow \textsc{Extract}(h^*, q)$; $\mathbf{m}^* \leftarrow \textsc{Infer}(\mathbf{c}^*, q)$
$\mathbf{w} \leftarrow \textsc{Weight}(\mathbf{c}^*, \mathbf{m}^*, q)$;
$\mathcal{C} \leftarrow \textsc{Critical}(\mathbf{c}^*, \mathbf{m}^*, q)$
Candidate parsing and alignment
$\mathbf{c} \leftarrow \textsc{Extract}(h, q)$; $\mathbf{m} \leftarrow \textsc{Infer}(\mathbf{c}, q)$;
$\mathcal{A} \leftarrow \textsc{Align}(\mathbf{c}^*, \mathbf{m}^*, \mathbf{c}, \mathbf{m}, q)$
Component similarity scoring
for $i \leftarrow 1$ **to** K **do**
  $s_i \leftarrow \begin{cases} \textsc{Sim}(i, \pi(i), q), & \pi(i) \neq \texttt{NULL} \\ 0, & \text{otherwise} \end{cases}$
end
Aggregation
$g \leftarrow \prod_{i \in \mathcal{C}} \mathbf{1}[s_i > 0]$; $r \leftarrow g \sum_{i=1}^K w_i s_i$
$\mathcal{T} \leftarrow \{\mathbf{c}^*, \mathbf{m}^*, \mathbf{w}, \mathcal{C}, \mathcal{A}, \mathbf{s}\}$
return $(r, \mathcal{T})$

2.5 The Overall Pipeline

MOOSE-Chem3 couples MC3-Rank with an executor $E \in \{E_{\text{sim}}, E_{\text{lab}}\}$ to form a closed loop for data-efficient hypothesis screening. Given a research question q, an LLM-based scientific discovery system generates candidate hypotheses $\mathcal{H}$. At each timestep t, MC3-Rank selects a hypothesis $h_t \in \mathcal{H}$ based on the current state s_t, receives feedback $r_t = E(h_t) \in [0, 1]$, and updates s_{t+1} using the new evidence. This loop refines component clusters, pruning decisions, and hypothesis rankings over time.

During *development*, MC3-Sim instantiates E_{sim} as a scalable feedback environment by comparing candidates with the reference anchor h^*. This enables MC3-Rank, and potentially other ranking policies, to be developed and stress-tested under repeated feedback that would be infeasible with wet-lab experiments alone, while preserving component-level traces for interpretability. During *deployment*, E_{sim} is replaced by E_{lab}, allowing the same agent-side policy to operate on real wet-lab outcomes without requiring a reference anchor.

3 Results

3.1 Evaluation Setting

We evaluate MOOSE-Chem3 from two perspectives: (i) simulator fidelity, that is, whether MC3-Sim produces feedback consistent with real experimental outcomes; and (ii) ranking efficiency, that is, whether MC3-Rank reduces the number of trials required to identify the reference anchor hypothesis. Unless otherwise specified, all experiments are

Table 1: Performance comparison of MC3-Sim against baselines and ablation studies on wet-lab experiments.

Method	**Trend Alignment**		**Predictive Accuracy**		
	Spearman ρ (↑)	PCI (↑)	RMSE (↓)	MAE (↓)	RMSLE (↓)
Main Comparison					
Matched Score	0.667	47/178	0.267	0.205	0.201
MC3-Sim (Ours)	**0.759**	**121/178**	**0.245**	**0.184**	**0.178**
Ablation Studies					
w/o Critical Points	0.751	89/178	0.263	0.199	0.193
w/o Comp. Extraction	0.683	47/178	0.313	0.246	0.232

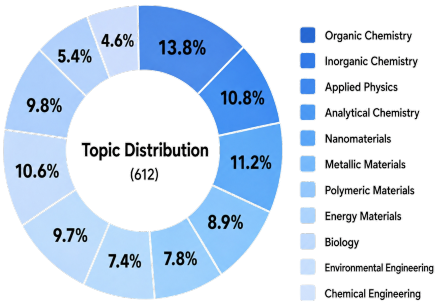


Figure 4: Distribution of real-world experimental benchmarks by discipline.

implemented with `GPT-4o-mini` (OpenAI, 2024).

Datasets. For simulator fidelity evaluation, we curate a benchmark of 178 research questions and 612 hypotheses with experimentally validated outcomes spanning domains, as shown in Figure 4.

For ranking evaluation, we use TOMATO-chem (Yang et al., 2025c), which contains 51 scientific problems, each with 64 hypotheses: one reference anchor and 63 generated candidates. The disciplinary distribution is provided in §D.3. For real-outcome sequential deployment, we construct a post-cutoff benchmark with three scientific questions and 300 experimentally validated hypotheses.

Executors. MC3-Rank interacts with an experiment executor $E : \mathcal{H} \to [0, 1]$ and receives feedback $r_t = E(a_t)$ for each tested hypothesis $a_t \triangleq h_t \in \mathcal{H}$. During development, we instantiate E with the simulator E_{sim} (MC3-Sim); when real experiments are available, the same interface allows switching to a wet-lab executor E_{lab}.

Metrics. For simulator fidelity, we compare MC3-Sim predictions against annotated outcomes using (i) trend alignment by Spearman correlation, (ii) Perfect Consistency Indicator (PCI), the number of questions with perfect trend alignment, and (iii) predictive accuracy by metrics such as RMSE (§E). For ranking efficiency, we report N_{trials}, the number of evaluations required to identify the target hypothesis for each problem; lower is better.

3.2 Results on Real-world Experiments

Fidelity of MC3-Sim. We compare simulated results from *MC3-Sim* against experimentally validated outcomes on the curated benchmark of 178 research questions and 612 hypotheses. As shown in Table 1, MC3-Sim achieves better trend alignment (higher Spearman correlation and PCI) and lower predictive error than the Matched Score baseline (Yang et al., 2025c). Ablations suggest that component extraction and critical-component modeling contribute to simulator fidelity: removing Critical Points causes a modest degradation, while omitting Component Extraction leads to larger drops in both alignment and predictive accuracy. These results support the use of MC3-Sim as an informative development-time proxy for developing experiment-guided ranking policies.

Table 2: Search efficiency comparison. Lower N_{trials} is better.

Method	N_{trials}	
	$N = 64$	$N = 128$
Baselines		
Uninformed Search	32.50	64.50
Pre-Exp. Ranking	28.61	51.30
Experiment-Guided Ranking		
MC3-Rank (GPT-4o-mini)	15.20	30.70
MC3-Rank (Claude 3.5)	**14.12**	**27.35**

Efficiency of MC3-Rank. We evaluate experiment guided ranking on TOMATO-chem (Yang et al., 2025c) using N_{trials}, the number of evaluations needed to identify the reference anchor hypothesis. As shown in Table 2, Uninformed Search and Pre-Experiment Ranking require around 30 trials on average in the standard setting ($N = 64$), while MC3-Rank requires 15.20. This advantage grows in the larger search space ($N = 128$), where MC3-Rank remains robust as baselines de-

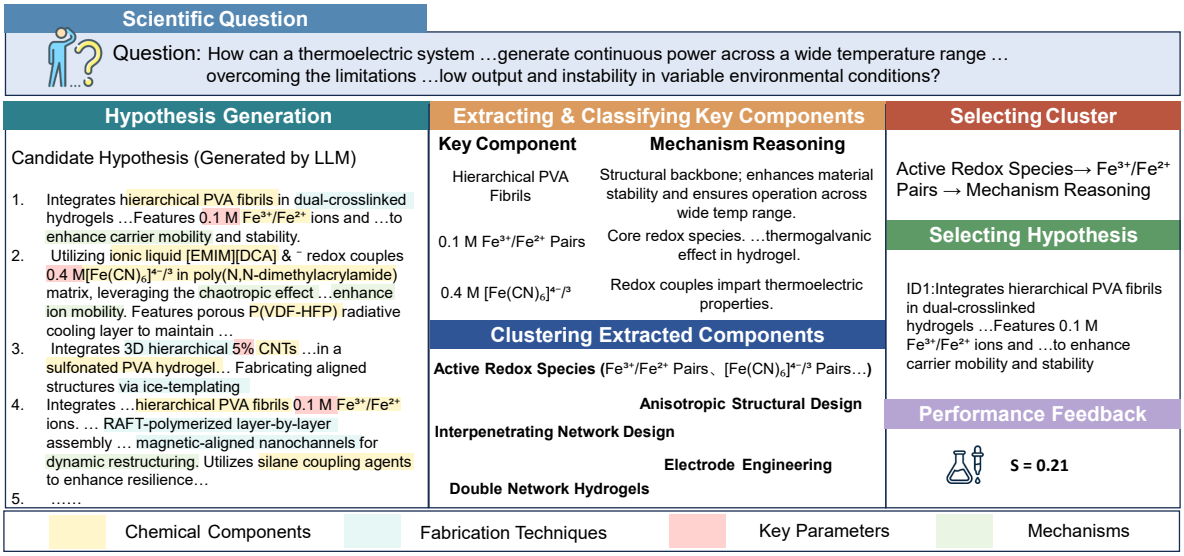


Figure 5: Case study of the iterative hypothesis ranking workflow.

grade. Figure 5 provides a case study showing how component-level reasoning helps narrow the search space. These results highlight the scalability benefit of iterative feedback. Detailed analysis is provided in §J.

Comparison with adaptive baselines. To further test whether the advantage of MC3-Rank comes merely from generic adaptivity, we compare it with adaptive baselines from Bayesian optimization (Yu et al., 2026), bandit optimization (Shi et al., 2024), and active learning (Herde et al., 2025). As shown in Table 3, the best non-MC3 adaptive baseline, GreedyTargetRF, requires 23.06 trials, whereas MC3-Rank requires only 15.20 trials. This suggests that MC3-Rank's advantage comes from component-level mechanistic attribution and cross-hypothesis feedback transfer, rather than from adaptivity alone.

Table 3: Adaptive baselines on TOMATO-chem ($N = 64$). Lower N_{trial} is better.

Family	Method	$N_{\text{trial}} \downarrow$
Bayesian Optimization	text-embedding-3-small	30.62
	Qwen3-Embedding-0.6B	32.59
	SciBERT	30.37
Bandit Optimization	Seq. Halving	30.73
	Emb. k-means	34.62
Active Learning	UncertaintyRF	25.45
	GreedyTargetRF	23.06
MC3	MC3-Rank	**15.20**

Real-outcome sequential deployment. We construct a post-cutoff real-outcome benchmark with three scientific questions and 300 experimentally validated hypotheses. At each step, MC3-Rank or human experts select one untested hypothesis, observe its wet-lab outcome only after selection, analyze the feedback, and then choose the next candidate. Thus, the evaluation follows the same sequential protocol as deployment with E_{lab}, rather than simulated feedback. To prevent leakage, public records are curated from 2025–2026 literature, supplemented with unpublished collaborating-lab results, and outcome labels are hidden before selection. As shown in Table 4, MC3-Rank reaches the best-performing hypothesis in 32.3 trials on average, compared with 45.7 trials for experts, reducing screening cost by 29.3%. This demonstrates more efficient use of wet-lab feedback under the same sequential protocol. See Appendix F for details.

Table 4: Real wet-lab outcome sequential deployment.

Method	Mean $N_{\text{trial}} \downarrow$
Human experts	45.7
MC3-Rank	**32.3**

3.3 Ablation Study

MC3-Rank Ablations. To isolate the contributions of key components in *MC3-Rank*, we evaluate ablated variants: (1) removing functional clustering (*MC3-Rank w/o Clustering*); (2) further disabling feedback analysis (*MC3-Rank w/o Clustering & Feedback Analysis*); and (3) restricting the comparison to the idealized noise-free setting. As shown in Table 6, each removal degrades performance, suggesting that clustering, analytical summarization,

Table 5: Composition of different noise levels.

Noise Level	Extrema	Plateaus	Cliffs
Simple	0–10	0–2	0–2
Medium	0–30	0–6	0–6
Complex	≥ 30	≥ 3	≥ 3

Table 6: Ablation study and robustness analysis. Lower N_{trials} indicates better efficiency.

Method	Clean	Simulated Noise		
		Simple	Medium	Complex
MC3-Rank	**15.20**	**21.8**	**26.6**	**32.7**
Ablation Studies				
w/o Clustering	27.98	32.7	35.8	36.5
w/o Clus. & Anal.	35.63	37.2	38.4	40.5

and sufficient feedback all contribute to efficient hypothesis screening.

Robustness to Noisy Feedback. To assess robustness to imperfect feedback from simulator approximation and LLM-based judgments, we introduce domain-grounded perturbations to $\phi(\cdot)$ (details in §G). Table 5 defines Simple, Medium, and Complex settings with increasing local extrema, plateaus, and cliffs.

We evaluate *MC3-Rank* and its ablations under these settings. As shown in Table 6, noise increases N_{trials}, but MC3-Rank retains an efficiency margin under Complex Noise. These results suggest that clustering and feedback analysis help mitigate misleading feedback and preserve search efficiency.

Complexity Reduction. We provide a theoretical formalization of why component-level feedback can reduce search complexity. Let $\mathcal{K}$ denote the set of functional clusters and define

$$\mathcal{H} = \{h \subseteq \mathcal{K} : |h| \leq m\}. \tag{7}$$

where m bounds the structural complexity of a hypothesis. Holistic search scales as $\mathcal{O}(|\mathcal{K}|^m)$. By attributing feedback to marginal utilities of clusters $k \in \mathcal{K}$ and updating $P(k \mid y)$, MC3-Rank effectively prunes large subsets $\{h' \in \mathcal{H} \mid k \in h'\}$ without individual testing, reducing the effective dimensionality toward $\mathcal{O}(|\mathcal{K}|)$. A formal derivation is provided in §N.

4 Related Work

Pre-experiment Ranking. Prior methods mainly rely on *pre-experiment ranking*, either by assigning individual scores (Yang et al., 2024, 2025c; Zhou et al., 2024) or performing pairwise comparisons (Si et al., 2024; Liu et al., 2025; Yang et al., 2025b). However, they depend on static LLM priors and do not adapt to empirical outcomes, overlooking the iterative nature of scientific discovery (see §O for a detailed comparison with MOOSE-Chem and MOOSE-Chem2 (Yang et al., 2025c,b)).

Active Learning and Bayesian Optimization. Experimental feedback is central to active learning (Zhang et al., 2025) and Bayesian optimization (Cheng et al., 2026; Yang et al., 2025a) (detailed in §K). Although adaptable to discrete hypothesis pools via text embeddings, these methods typically depend on encoder quality at the candidate level. In contrast, MC3-Rank attributes feedback to functional components and transfers evidence across hypotheses with shared mechanisms, sidestepping the encoder bottleneck.

Discovery with Digital Verifiers. LLM-based methods in mathematics (Romera-Paredes et al., 2024; Shojaee et al., 2024) and programming (Novikov et al., 2025; Weng et al., 2025) also use feedback loops for discovery. However, these domains are fully automatable: with sufficient compute, verification runs end-to-end without human intervention, enabling massive-scale trial-and-error. Wet-lab experiments cannot: they demand human labor, instrumentation, and long turnaround cycles, making such feedback loops infeasible.

5 Conclusion

In this work, we tackle the critical bottleneck of scarce experimental feedback that hinders adaptive hypothesis ranking in AI-driven scientific discovery. We present MOOSE-Chem3, a closed-loop framework that couples MC3-Sim, a mechanism-aware development-time simulator, with MC3-Rank, a component-reasoning agent for feedback-guided hypothesis selection. Experiments show that MC3-Sim captures useful trend-level alignment with wet-lab outcomes, while MC3-Rank identifies high-performing hypotheses more efficiently than static, adaptive, ablated, and human-expert baselines. Theoretically, we show that component-level feedback can transfer across hypotheses sharing functional mechanisms, reducing the effective search space in decomposable discovery tasks. Together, these results establish an efficient, interpretable, and scalable framework for advancing feedback-driven hypothesis discovery, with potential impact across the natural sciences.

Limitations

A primary limitation of this work is that the constructed simulator does not provide perfectly accurate experimental feedback. Specifically, the simulator is based on foundational assumptions developed through extensive consultations with domain experts. While it represents the first attempt to build such a simulator for experiment-guided hypothesis ranking, its outputs remain an approximation rather than exact experimental results.

The rationale for developing this simulator stems from the absence of any prior tools with comparable functionality. Its purpose is to *enable* research on experiment-guided ranking methods, which can later be applied and validated with real experimental feedback in practical settings.

Perfect simulator fidelity is not required, but useful trend-level alignment remains important By developing and testing ranking methods within a simulated environment characterized by noisy rewards, we ensure their robustness. This prepares the policy to leverage real experimental feedback, which is inherently stochastic due to environmental factors like humidity and reproducibility challenges, to effectively identify optimal hypotheses in real-world scenarios.

MC3-Sim is a development-time proxy rather than a replacement for wet-lab validation. Its reference anchor h^* is used only to instantiate simulated feedback during development; MC3-Rank observes only executor feedback and does not access the anchor as an oracle answer. During real deployment, E_{sim} is replaced by E_{lab}, and no reference anchor is required.

Nevertheless, the quality of policies developed in simulation depends on how well MC3-Sim captures useful trends in the underlying experimental landscape.

Our real-outcome sequential deployment provides evidence that MC3-Rank can use wet-lab feedback efficiently under a sequential protocol, but it remains limited in scale. The deployment benchmark contains three scientific questions and 300 experimentally validated hypotheses. Although the data are post-cutoff and outcome labels are hidden before selection, broader validation across more domains, laboratories, protocols, and expert groups is needed to establish generality. Because real-outcome deployments require substantial wet-lab resources, expert time, and reproducibility checks, scaling this evaluation is an important direction for future work.

Finally, MOOSE-Chem3 is intended to assist, not replace, scientific judgment. The system prioritizes hypotheses and provides component-level traces, but experimental feasibility, safety, reproducibility, and domain validity should remain under human expert review. This is especially important when AI-suggested hypotheses involve hazardous materials, biological systems, high-cost experiments, or safety-critical applications.

Ethical Considerations

The datasets used in this study are curated from publicly available peer-reviewed scientific journals and contain no personally identifying information or offensive content, ensuring no violation of privacy or copyright. The human-expert baseline and distortion-design consultations involve domain researchers participating as collaborators rather than crowdworkers; see Appendix Q for the full protocol. For public literature, we use the data only for scientific evaluation and do not include private personal information. For collaborating-lab results, we report only aggregated benchmark statistics and do not disclose sensitive experimental details beyond what is permitted by the collaborators. No crowdworkers were employed. The human expert comparison involved domain experts acting in their professional capacity. No personal or sensitive information was collected, and all reported expert-comparison results are aggregated. The purpose of the expert comparison is to evaluate hypothesis-prioritization efficiency under the same sequential feedback protocol, not to assess individual expert performance.

While our framework facilitates automated hypothesis ranking for materials science, we emphasize that all AI-suggested experimental plans should be validated by human experts to ensure safety and ethical compliance. We believe this work presents no significant risk of harm or negative societal impact.

AI Assistance Statement

In this work, LLMs, specifically GPT-4o-mini, Claude 3.5 Sonnet, and Gemini, were employed in four main ways: (i) to aid and polish the writing for clarity and style; (ii) to assist in the retrieval and discovery of related work; (iii) to support research ideation by suggesting possible directions and refining problem formulations; and (iv) to pro-

vide coding assistance, including code generation, debugging, and optimization suggestions.

All outputs were critically reviewed, verified, and refined by the authors. The core scientific ideas, methodology, and contributions remain entirely the authors' own.

References

Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander D Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2015. *Essential cell biology*. Garland Science.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 995–1005.

William D Callister and David G Rethwisch. 1999. *Materials science and engineering: an introduction*. Wiley New York.

Liang Cheng, Zhenzhi Tan, Zongbin Jia, Qifeng Lin, Qi Yang, and Sanzhong Luo. 2025. Heuristic data-driven approach for synergistic cobalt (iv)–enamine catalysis. *Nature Synthesis*, pages 1–11.

Mouyang Cheng, Chu-Liang Fu, Ryotaro Okabe, Abhijatmedhi Chotrattanapituk, Artittaya Boonkird, Nguyen Tuan Hung, and Mingda Li. 2026. Artificial intelligence-driven approaches for materials design and discovery. *Nature Materials*, pages 1–17.

Mario Coccia. 2019. Why do nations produce science advances and new technology? *Technology in society*, 59:101124.

Yongcheng Ding, José D Martín-Guerrero, Yolanda Vives-Gilabert, and Xi Chen. 2023. Active learning in physics: From 101, to progress, and perspective. *Advanced Quantum Technologies*, page 2300208.

Dong Han, Zhehong Ai, Pengxiang Cai, Shanya Lu, Jianpeng Chen, Zihao Ye, Shuzhou Sun, Ben Gao, Lingli Ge, Weida Wang, and 1 others. 2025. Chembomas: Accelerated bo in chemistry with llm-enhanced multi-agent system. *arXiv preprint arXiv:2509.08736*.

Corwin Hansch, A Leo, and 1 others. 1995. Fundamentals and applications in chemistry and biology.

Marek Herde, Minh Tuan Pham, Daniel Kottke, Alexander Benz, Lukas Lührs, Pascal Mergard, Christoph Sandrock, Jiaying Cheng, Atal Roghman, Mehmet Müjde, and 1 others. 2025. scikit-activeml: A comprehensive and user-friendly active learning library.

Yuriy Khalak, Gary Tresadern, David F Hahn, Bert L De Groot, and Vytautas Gapsys. 2022. Chemical space exploration with active learning and alchemical free energies. *Journal of Chemical Theory and Computation*, 18(10):6259–6270.

Vladimir Kozyrev, François Sindt, and Didier Rognan. 2025. Active learning to select the most suitable reagents and one-step organic chemistry reactions for prioritizing target-specific hits from ultralarge chemical spaces. *Journal of Chemical Information and Modeling*, 65(2):693–704.

Hongguang Liao, Sheng Hu, Hu Yang, Lei Wang, Shinya Tanaka, Ichigaku Takigawa, Wei Li, Hailong Fan, and Jian Ping Gong. 2025. Data-driven de novo design of super-adhesive hydrogels. *Nature*, 644(8075):89–95.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. 2025. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. In *Findings of the Association for Computational Linguistics ACL 2026*.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. LLM4SR: A survey on large language models for scientific research. *CoRR*, abs/2501.04306.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, and 1 others. 2025. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2025-05-16.

D Christopher Radford, Matthew Tamasi, Elena Di Mare, and Adam Gormley. 2025. Automation and active learning for the multi-objective optimization of antibody formulations.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, and 1 others. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.

Burr Settles. 2009. Active learning literature survey. Technical Report 1648, University of Wisconsin-Madison Department of Computer Sciences.

Chengshuai Shi, Kun Yang, Zihan Chen, Jundong Li, Jing Yang, and Cong Shen. 2024. Efficient prompt optimization through the lens of best arm identification. *Advances in Neural Information Processing Systems*, 37:99646–99685.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. 2024.

Llm-sr: Scientific equation discovery via programming with large language models. *arXiv preprint arXiv:2404.18400*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Justin S Smith, Benjamin Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E Roitberg. 2018. Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics*, 148(24):241733.

Andrew Wagenmaker, Kevin Huang, Liyiming Ke, Kevin Jamieson, and Abhishek Gupta. 2024. Overcoming the sim-to-real gap: Leveraging simulation to learn to explore for real-world rl. *Advances in Neural Information Processing Systems*, 37:78715–78765.

Yixuan Weng, Minjun Zhu, Qiujie Xie, Qiyao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. 2025. Deepscientist: Advancing frontier-pushing scientific findings progressively. *arXiv preprint arXiv:2509.26603*.

John Wiley. 1986. Introduction to solid state physics. *New York*, 185.

Zhuo Yang, Jiaqing Xie, Ben Gao, Shuaike Shen, Wanhao Liu, Liujia Yang, Beilun Wang, Tianfan Fu, Yuqiang Li, and 1 others. 2025a. Molact: An agentic rl framework for molecular editing and property optimization. *arXiv preprint arXiv:2512.20135*.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13545–13565.

Zonglin Yang, Wanhao Liu, Ben Gao, Yujie Liu, Wei Li, Tong Xie, Lidong Bing, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. 2025b. Moose-chem2: Exploring llm limits in fine-grained scientific hypothesis discovery via hierarchical search. *Advances in Neural Information Processing Systems*, 38:89045–89076.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2025c. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ye Yu, Zikai Xie, Man Luo, Adam Redfearn, Arianna Brandolese, Chenxi Sheng, Andrew P Dove, Xiaolong Zhang, Linjiang Chen, and Jun Jiang. 2026. Optimizing toward discovery: Ai-driven exploration of lewis acid–base catalysts for pet glycolysis. *Journal of the American Chemical Society*, 148(4):4635–4644.

Yu Zhang, Yang Han, Shuai Chen, Ruijie Yu, Xin Zhao, Xianbin Liu, Kaipeng Zeng, Mengdi Yu, Jidong Tian, Feng Zhu, and 1 others. 2025. Large language models to accelerate organic chemistry synthesis. *Nature Machine Intelligence*, pages 1–13.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 117–139.

A Theoretical Formulation of MC3-Sim

MC3-Sim is designed as a low-cost surrogate feedback environment for developing and analyzing experiment-guided ranking policies. Its goal is to produce a normalized performance signal in $[0, 1]$ that is informative for screening hypotheses, while remaining controllable and interpretable. This appendix provides an explanatory formulation that motivates the simulator design and clarifies why the feedback landscape can exhibit noise and spurious optima.

Conceptual Foundations (A1–P1–D1). For a fixed research question q, let $\mathcal{H}$ denote the candidate hypothesis set, and let $h^* \in \mathcal{H}$ denote a reference (ground-truth) hypothesis validated by experts or prior evidence. We posit three conceptual foundations:

1. (*A1: Local Optimum Assumption*) There exists a neighborhood $\mathcal{N}(h^*) \subseteq \mathcal{H}$ in which h^* is a dominant solution. Intuitively, hypotheses sufficiently close to h^* in mechanism tend to perform well, and hypotheses farther away tend to perform worse.

2. (*P1: Similarity Principle*) Hypotheses that are more similar in their underlying structure or function tend to yield more similar outcomes. Equivalently, performance varies smoothly with a suitable notion of mechanism-level closeness.

3. (*D1: Representation Distortion*) Any practical representation of hypothesis similarity is imperfect. This imperfection distorts the perceived closeness between hypotheses, which can introduce noise, secondary peaks, and unexpected valleys in the observed performance landscape.

Oracle Closeness and an Idealized Performance Surface. To make *A1* and *P1* concrete, consider

an oracle embedding function

$$\phi^*(\cdot \mid q) : \mathcal{H} \to \mathbb{R}^d,$$

which maps each hypothesis to a latent point that reflects its true mechanism-level relation to q. Define the oracle distance to the reference hypothesis as

$$d^*(h \mid q, h^*) \triangleq \|\phi^*(h \mid q) - \phi^*(h^* \mid q)\|_2 .$$

An idealized normalized performance function can be modeled as a monotone decay of this distance:

$$r^*(h \mid q, h^*) \triangleq g(d^*(h \mid q, h^*)),$$
$$g(\rho) \in [0,1], \quad g(0) = 1, \quad g'(\rho) \leq 0.$$

A common choice is an isotropic Gaussian-like form,

$$g(\rho) = \exp\left(-\frac{\rho^2}{2\sigma^2}\right), \tag{8}$$

which yields a smooth unimodal landscape centered at $\phi^*(h^* \mid q)$. This serves as an interpretable approximation in the latent space, capturing the idea that hypotheses closer in mechanism to h^* tend to achieve higher normalized performance.

Practical Closeness and Distorted Landscapes. In realistic settings, the oracle embedding ϕ^* is unavailable. Similarity must be estimated using a practical proxy embedding

$$\phi(\cdot \mid q) : \mathcal{H} \to \mathbb{R}^d,$$

which introduces distortions relative to ϕ^*. One convenient view models this as a perturbation in the latent representation:

$$\phi(h \mid q) = \phi^*(h \mid q) + \Delta(h \mid q), \tag{9}$$

where $\Delta(h \mid q)$ captures systematic representation error. The induced practical distance is

$$d(h \mid q, h^*) \triangleq \|\phi(h \mid q) - \phi(h^* \mid q)\|_2 ,$$

and the corresponding observed performance proxy can be written as

$$\tilde{r}(h \mid q, h^*) \triangleq g(d(h \mid q, h^*)) + \epsilon(h \mid q). \tag{10}$$

Here $\epsilon(h \mid q)$ summarizes additional deviations that are convenient to model explicitly, for example, measurement noise or systematic underestimation of certain mechanisms. Even when $g(\cdot)$ is smooth, distortions in ϕ and nonzero ϵ can generate irregular landscapes with secondary peaks and valleys, aligning with *D1*. This explanatory view clarifies why a simulator can expose ranking policies to realistic non-ideal feedback, even when grounded in a smooth oracle principle.

Connection to MC3-Sim. MC3-Sim instantiates a practical executor E_{sim} that returns a normalized score in $[0,1]$ for any tested hypothesis h:

$$r = E_{\text{sim}}(h \mid q, h^*) \in [0,1].$$

Instead of explicitly constructing $\phi(\cdot \mid q)$ in $\mathbb{R}^d$, MC3-Sim estimates mechanism-level closeness through a structured, component-based comparison against the reference h^*. Concretely, it decomposes hypotheses into functional components, scores per-component similarity, and aggregates them into the simulator output (Eq. 6 in the main text). From the viewpoint of Eq. 10, the component-based scoring defines a proxy closeness measure and an induced performance surface, while imperfections in extraction, alignment, and similarity judgment act as structured sources of distortion. This design keeps the simulator interpretable and controllable, while producing feedback landscapes that are sufficiently rich to develop and evaluate experiment-guided ranking strategies.

A.0.1 Conceptual Foundations

A1, *P1*, and *D1* are all reasonable and sufficiently grounded. *P1* reflects the fundamental axiom that "structure determines properties, and properties determine outcome," which underpins multiple disciplines: molecular structure and material function in chemistry and materials science (Callister and Rethwisch, 1999; Hansch et al., 1995), crystal structure and physical properties in physics (Wiley, 1986), and protein structure and biological function in biology (Alberts et al., 2015). *D1* follows logically from *A1* and *P1*, since any practical representation of hypothesis similarity must introduce distortions. *A1* is mostly valid but not guaranteed: even within a sufficiently small neighborhood, the labeled reference anchor hypothesis may not be the strict local optimum, as there could exist another hypothesis in that region with higher performance. This limitation, however, does not affect the simulator's role in developing ranking policies, whose goal is to recover the labeled ground truth. When deployed in real experiments, any superior hypotheses beyond the labeled ground truth would be revealed,

ensuring that policies developed under a simulator supported by *A1* remain effective in practice.

Figure 6 visually illustrates these conceptual foundations. In the ideal scenario (Figure 6a), *A1* ensures the presence of a dominant local optimum, while *P1* enforces that hypotheses closer in structure or function to this optimum yield more similar outcomes. Together, these yield a smooth, unimodal performance landscape, where Euclidean distance in hypothesis space faithfully reflects structural and functional similarity.

However, practical scenarios differ substantially, as the measured distance ("closeness") between hypotheses, whether estimated by scientists or LLMs, may not faithfully capture structural and functional similarity. For instance, a chemical hypothesis might contain a useful functional component whose contribution is underrepresented, placing it farther from the dominant peak than warranted and creating a spurious secondary maximum. Conversely, a weaker hypothesis may appear deceptively close to the optimum, forming a local valley. These distortions yield a more irregular performance landscape, as illustrated in Figure 6b, with unexpected secondary peaks and valleys. Figure 6c further isolates these deviations, highlighting the gap between the idealized oracle landscape and practical estimates of closeness.

We formalize these foundations by defining a mathematical model relating hypothesis embeddings, similarity, and performance.

A.0.2 Mathematical Formulation

Let $\mathcal{H} \subset \mathbb{R}^d$ denote the hypothesis space, where each hypothesis $h \in \mathcal{H}$ is represented as a point in a d-dimensional latent space, conditioned on a specific research question q. Let $h^* \in \mathcal{H}$ denote the ground truth hypothesis for q, representing an experimentally validated optimum. We define the idealized performance function $f(h, h^*; q, \phi^*(\cdot))$ for any hypothesis h in the vicinity of h^* as:

$$\frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\phi^*(h \mid q) - \phi^*(h^* \mid q)\|^2}{2\sigma^2}\right) \quad (11)$$

where $\phi^*(\cdot \mid q)$ is an oracle embedding function that maps each hypothesis h to a point in the latent hypothesis space under the context of research question q. The embedded positions capture the oracle's understanding of closeness, measured by the Euclidean distance $\|\phi^*(h \mid q) - \phi^*(h^* \mid q)\|$.

A.1 Comparison of Hypothesis Ranking Paradigms

Building on this foundation, we develop an in-context reinforcement learning (ICRL) framework for experiment-guided hypothesis ranking. Within this framework, we instantiate a clustering-based agentic policy that decomposes hypotheses into functional components and groups them by shared mechanistic roles. After each experimental trial, the agent analyzes the tested hypothesis to infer which components contributed to its performance, then prioritizes untested hypotheses that incorporate the most promising functional elements. This enables efficient transfer of insights across structurally related candidates and helps navigate the hypothesis space more effectively. Experiments show that this policy significantly outperforms pre-experiment baselines and strong ablations. Combined with the simulator, our ICRL framework forms a general-purpose toolkit for studying experiment-guided ranking, with our policy serving as a strong proof of concept.

Figure 7 provides an overview of the three paradigms: pre-experiment ranking, experiment-guided ranking with real experiments, and our simulator-driven approach for developing ranking policies.

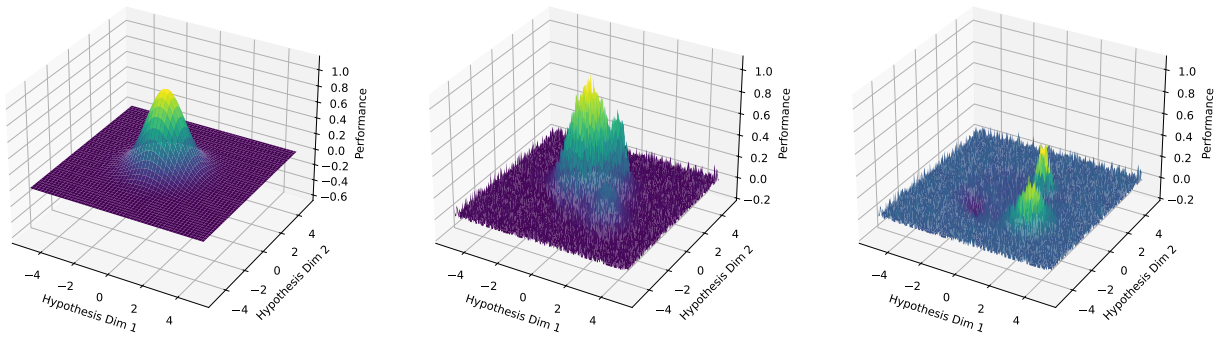
B Extracting Key Functional Components in the Simulator

B.1 A Framework for Extracting Critical Functional Components in the Simulator

The specific framework of *MC3-Sim* for extracting key functional components is illustrated in Figure 8. To demonstrate this process, we categorize the critical components and conclusions embedded within scientific hypotheses using a representative example from the field of chemistry. We then analyze the specific role and underlying mechanism of each component in the context of the research problem and the derived conclusions. Finally, the system synthesizes and extracts the key components, their corresponding mechanisms, and the validated conclusions from the hypothesis.

B.2 Prompt for Extracting Key Chemical Components in the Simulator

The prompt for extracting key chemical components in the simulator, along with examples, is as follows:



(a) Idealized performance landscape (*A1* + *P1*).

(b) Realistic performance landscape (*A1* + *P1* + *D1*).

(c) Deviations from imperfect closeness estimation (*D1*).

Figure 6: Illustration of the three conceptual foundations (*A1*–*P1*–*D1*) for simulator construction.

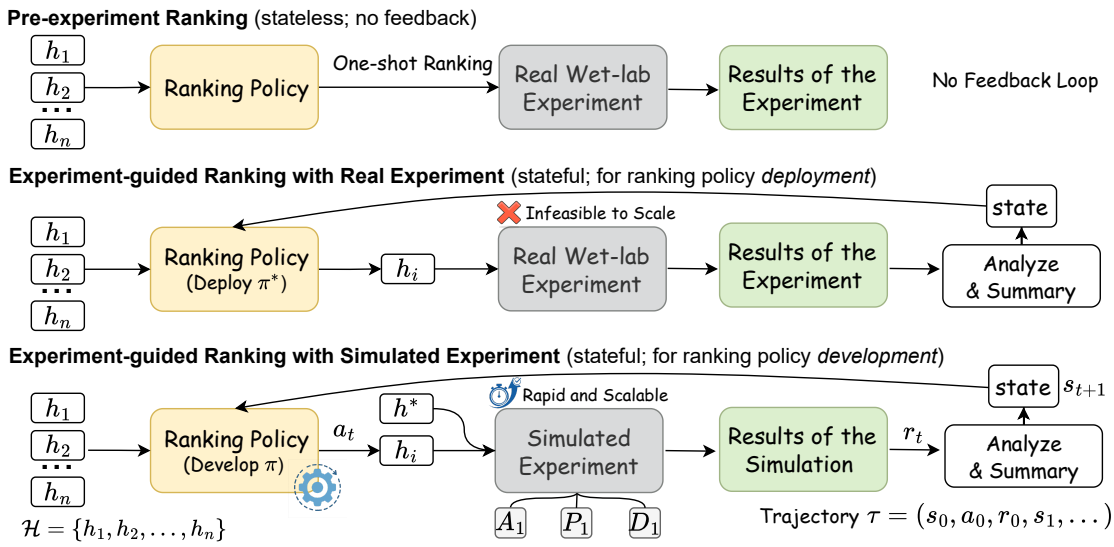


Figure 7: Overview of ranking strategies. Pre-experiment ranking is stateless and ignores feedback. Experiment-guided ranking with real experiments is stateful but infeasible to scale. Our simulator enables efficient testing of ranking policies through simulated feedback before real deployment.

You are an experienced expert. I will provide you with a scientific question and a scientific hypothesis. Your task is to identify the chemical key points within the hypothesis that are essential for addressing the scientific question. Chemical key points are the core elements—such as basic chemical components, reactions, or mechanistic methods—critical to solving the problem effectively. Analyze these key points by linking them to the scientific question, determining how they contribute to resolving it.

When identifying chemical key points, consider the following:

Each substance may be a key point. If it includes specific parameters like concentration or mass fraction (e.g., 0.3M NaCl, 10wt% PVA), ensure these details are retained in the division process without losing specificity. If multiple substances are related and function together (e.g., potassium ferricyanide and potassium ferrocyanide as an oxidizing-reducing pair), group them as a single chemical key point based on their shared role or interdependence. Exclude elements from the scientific question that reappear in the hypothesis as prerequisites. Distinguish key points from validation methods. For each identified chemical key point, conduct a detailed and rigorous analysis of its role and function in relation to the scientific question. Use your chemical knowledge to explain the specific mechanism by which it addresses the problem, focusing on how it enhances the relevant properties or performance outlined in the question. Provide a clear, mechanistic explanation of its contribution and, if multiple key points exist, describe their interconnections.

Additionally, identify the results—effects or phenomena caused by these key points—representing

Framework for Analyzing Scientific Hypotheses in Chemical Problems

1. Identify Key Chemical Components and Conclusions

$$H_i = \{K_i, C_i\}$$

$$(H_i\text{: Hypothesis } i,\ K_i\text{: Key Chemical Components},\ C_i\text{: Conclusion})$$

2. Explain Mechanism of Key Chemical Components

$$M_i = f(K_i, S, C_i)$$

$$(M_i\text{: Mechanism},\ S\text{: Scientific Problem},\ K_i,\ C_i \text{ as above})$$

3. Verify and Output Key Points, Mechanisms, and Conclusions

$$O_i = \{P_i, M_i, C_i\}$$

$$(O_i\text{: Output},\ P_i\text{: Key Points incl. } K_i,\ M_i,\ C_i \text{ as above})$$

Figure 8: A framework for extracting chemical components in the simulator.

the experiment's outcomes. In your output, focus on listing and explaining the chemical key points, followed by the results, ensuring no prerequisites from the scientific question are included.

Output format:

Chemical Key Points
Chemical substance/component/method 1
Role and Function: Describe the role and function of the substance or method, including a detailed mechanistic explanation of how it addresses the scientific question and enhances relevant properties.
Chemical substance/component/method 2
Role and Function: Describe the role and function of the substance or method, including a detailed mechanistic explanation of how it addresses the scientific question and enhances relevant properties.
End Chemical Key Points
Results
Result 1: Describe the effects caused by the aforementioned reasons.
Result 2: Further describe other effects related to the experimental objectives.
End Results

Example:

Chemical Key Points
1. 10wt% PVA (Polyvinyl Alcohol)
Role and Function: Polyvinyl alcohol hydrogel acts as the base material, providing structural support and mechanical performance for thermoelectric gels.
2. Gdm_2SO_4 (Guanidine Sulfate)
Role and Function: Guanidine sulfate is integrated into the $K_3[Fe(CN)_6]$ / $K_4[Fe(CN)_6]$ system to improve thermoelectric performance by increasing solvent entropy and enhancing thermopower.
3. Directional Freezing Method
Role and Function: By employing directional freezing technology, aligned channels are created, enhancing the electrical conductivity and mechanical strength of the material.
4. Potassium Ferricyanide and Potassium Ferrocyanide ($K_3[Fe(CN)_6]$ / $K_4[Fe(CN)_6]$)
Role and Function: These compounds are crucial electrolytes that facilitate redox reactions within the polymer gel.
End Chemical Key Points
Results
Carnot-relative Efficiency: The Carnot-relative efficiency of the FTGA exceeds 8%.
Thermopower and Mechanical Robustness: Thermopower and mechanical robustness are enhanced.
End Results

To better illustrate the effectiveness of extracting key components, we provide a detailed example in materials science where we compare the performance of our simulator against human experts on a real-world problem.

- Scientific 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?

- Scientific Hypothesis: Develop a flexible N-type quasi-solid-state thermocell by integrating **anisotropic** polymer networks and **hierarchical 3D copper electrodes** to enhance ion transport, mechanical robustness, and thermoelectric performance. Utilizing Polyvinyl Alcohol (PVA) as the hydrogel matrix, the anisotropic structure is achieved through a directional freeze-thawing process. Ionic crosslinking with a 0.7 M $CuSO_4$ electrolyte and 0.1 M H_2SO_4 strengthens the hydrogel while retaining flexibility. Meanwhile, hierarchical 3D copper electrodes, fabricated via oxidation, etching, and thermal reduction, provide a high surface area, enhancing redox kinetics of the Cu^{2+}/Cu^0 couple and obviating platinum electrode reliance.

Here is expert extraction of key components:

- Key Points:
  1. Anisotropic Polymer Networks
  2. Hierarchical 3D Copper Electrodes
  3. Polyvinyl Alcohol (PVA) as the Hydrogel Matrix
  4. Directional Freeze-Thawing Process
  5. 0.7 M $CuSO_4$ Electrolyte
  6. 0.1 M H_2SO_4

- Results: Increased power density, improved thermopower, enhanced ionic conductivity, and improved mechanical properties.

Here is *MC3-Sim* extraction of key components:

- Key Points:
  1. Anisotropic Polymer Networks
  2. Hierarchical 3D Copper Electrodes
  3. Copper Sulfate ($CuSO_4$) Electrolyte
  4. Sulfuric Acid (H_2SO_4)
  5. Directional Freeze-Thawing Process

- Results: Power density increase, enhanced thermopower, improved ionic conductivity, and mechanical strength under deformation.

By comparing the approaches of an expert and *MC3-Sim* in extracting key scientific components for the specific scientific issues of ion transport efficiency and electrode performance, *MC3-Sim* successfully identifies solutions in its scientific hypotheses, including anisotropic polymer networks and hierarchical 3D copper electrodes. Compared to the human expert, *MC3-Sim* captures five out of six key points, missing only "Polyvinyl Alcohol (PVA) as the Hydrogel Matrix." The points it does identify align accurately with those proposed by the human expert based on the hypothesis, demonstrating the high accuracy of *MC3-Sim* in extracting key scientific components.

C The Role of Critical Points in *MC3-Sim*

$$S(h \mid q; h^*) = \left(\prod_{i \in \mathcal{C}} \mathbf{1}_{\{s_i > 0\}}\right) \cdot \left(\sum_{i=1}^{K} w_i s_i\right) \tag{12}$$

To better illustrate the role of labeling critical components $\mathcal{C}$ in *MC3-Sim*, as defined in Equation 12, we provide an example for clarity. For simplicity, we define the term $\left(\prod_{i \in \mathcal{C}} \mathbf{1}_{\{s_i > 0\}}\right)$ from Equation 12, related to Critical Points, as the *Correction Factor*. This factor takes values of either 0 or 1.

The scientific problem under study is: *How can a polymer gel material be designed to enhance the Seebeck coefficient by optimizing the matrix material and redox pair, thereby improving the energy conversion efficiency of a thermoelectric device utilizing the temperature difference between body heat and the environment?*

This scientific problem corresponds to four real experimental hypotheses, outlined as follows:

1. **Hypothesis 1**: By combining gelatin with KCl, prepare a gel with high ionic conductivity to investigate its Seebeck coefficient performance with the $[Fe(CN)_6]^{3-}/[Fe(CN)_6]^{4-}$ redox pair.

2. **Hypothesis 2**: By combining a PVA matrix with HCl, prepare a gel with high ionic conductivity and investigate its Seebeck coefficient performance under the influence of the Fe^{3+}/Fe^{2+} redox pair.

3. **Hypothesis 3**: By preparing a pure PVA gel, investigate its Seebeck coefficient performance under the influence of the Fe^{3+}/Fe^{2+} redox pair.

4. **Hypothesis 4**: By polymerizing acrylamide to prepare a hydrogel and investigate its thermoelectric performance.

The results of the actual wet-lab experiments are summarized in Table 7. For the first three hypotheses, the Critical Points are included in the assumptions, resulting in a Correction Factor of 1 and non-zero final results. In contrast, for the fourth hypothesis, the Critical Points are not included in the assumption, leading to a Correction Factor of 0 and a final result of 0.

Our simulator analyzes Hypothesis 4, where the Correction Factor is 0, as follows:

Analysis: *The ground truth scientific hypothesis identifies three core key points: Gelatin, KCl, and the redox pair, each playing significant roles in enhancing the Seebeck coefficient for thermoelectric efficiency. The generated hypothesis uses Polyacrylamide in place of Gelatin, which can provide a similar structural function but lacks some specific characteristics of Gelatin important for thermoelectric stability. It discusses ionic conductivity, paralleling that of KCl without directly employing the salt. Notably, it does not mention the redox pair, which is critical to the thermoelectric function as outlined in the ground truth hypothesis.*

However, while the generated hypothesis incorporates functional aspects related to ionic conductivity and chemical structure, it omits the crucial redox couple and does not fully replicate the essential components necessary for observable thermoelectric phenomena defined by the ground truth. The lack of adherence to all core key points leads to a conclusion that observable phenomena are limited. Given these evaluations, we assign $R = 0$ *because there is a substantial gap in essential components for the reaction proposed in the ground truth hypothesis.*

Our simulator analyzes Hypothesis 4 by noting that enhancing the Seebeck coefficient of a polymer gel material requires a redox pair. The first three hypotheses incorporate redox pairs, either $[Fe(CN)_6]^{3-}/[Fe(CN)_6]^{4-}$ or Fe^{3+}/Fe^{2+}, which facilitate efficient conversion of thermal energy to electrical energy. In contrast, Hypothesis 4 only involves polymerizing acrylamide and lacks a redox pair, rendering it unable to effectively convert thermal energy into electrical energy. Consequently, the thermoelectric potential is zero.

D Evaluating the Simulator with Real Experiment Results

In this section, we present the validation of our simulator's accuracy using a dataset of 612 experimental hypotheses, detailing their classification and composition. We further compare the trends of the simulated results with the corresponding real experimental outcomes to assess the simulator's predictive performance and reliability in capturing real-world experimental behaviors.

D.1 Dataset Composition and Analysis

To evaluate the performance of the simulator, we conducted a thorough analysis using real-world experimental data. We curated a set of 178 cutting-edge research questions, each designed to probe significant aspects of scientific research. These questions were carefully selected to encompass multiple areas within the scientific domain, ensuring a diverse and representative evaluation framework. Each question was associated with 3 to 6 hypotheses, resulting in a total of 612 authentic wet-lab experiment results. This extensive dataset forms a robust foundation for assessing the simulator's predictive accuracy and reliability.

The 612 experiment results were sourced from key subfields of natural science to provide broad coverage of the discipline. The distribution of these results across subfields is presented in Table 8.

0.96The use of authentic wet-lab results bolsters the credibility of our findings. By grounding the evaluation in real experimental data, we ensured that the simulator's predictions were tested against the intricacies and variability of actual laboratory conditions. This approach not only validates the simulator's performance but also underscores its potential to guide subsequent research by delivering reliable and actionable predictions.

D.2 Trend Comparison with Real Experiment Results

To further assess the capabilities of *MC3-Sim*, we utilized it to simulate 612 wet-lab experiments. These experiments corresponded to 178 cutting-edge science questions, and their simulated outcomes were subsequently aggregated for a comprehensive analysis. For each of the 612 experiments,

Table 7: The Seebeck coefficient of polymer thermoelectric materials.

Matrix	The Seebeck coefficient Se / mV K^{-1}
$[Fe(CN)_6]^{3-}/[Fe(CN)_6]^{4-}$ KCl/Gelatin	17
Fe^{3+}/Fe^{2+} HCl/PVA	1.02
Fe^{3+}/Fe^{3+} PVA	0.85
Polymerizing Acrylamide (PAM)	0

Table 8: Disciplinary distribution of the expanded real-outcome validation set.

Subfield	# Samples
Organic Chemistry	84
Inorganic Chemistry	66
Applied Physics	69
Analytical Chemistry	54
Nanomaterials	48
Metallic Materials	45
Polymeric Materials	59
Energy Materials	65
Biology	60
Environmental Engineering	33
Chemical Engineering	29
Total	612

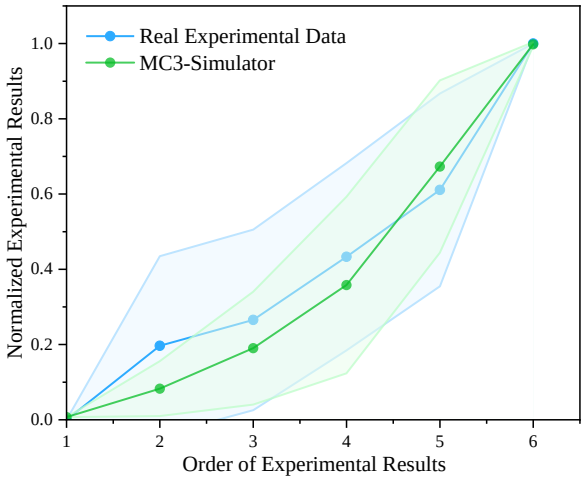


Figure 9: Trend comparison with real experiment results.

the simulated result was derived from the average of three trials conducted by *MC3-Sim*. These results were systematically arranged in ascending order along the "Order of Experimental Results" axis, enabling a unified comparison between the simulated and actual experimental outcomes.

Figure 9 compares the trends observed in *MC3-Sim* predictions (green line) with those from real experimental data (blue line). Error bars, representing the population standard deviation, illustrate the variability of the data points. Statistical significance was further established using the bootstrap method, with results indicating $p < 0.01$ (Berg-Kirkpatrick et al., 2012). The aggregated analysis reveals that the simulator effectively predicts the mean trends for all 178 sets of results, demonstrating strong consistency with the mean of the actual experimental outcomes.

D.3 Disciplines of the TOMATO-chem Benchmark Dataset

The TOMATO-chem benchmark dataset, spanning 12 distinct categories and totaling 3264 data records (Table 9), is a powerful resource due to its **inherent interdisciplinary nature** and wide scientific coverage. Far exceeding a focus on traditional chemical branches such as organic, inorganic, and analytical chemistry, the dataset features a strong emphasis on **advanced materials science** by including significant contributions from Energy Materials, Polymeric Materials, Metallic Materials, and Nanomaterials. Furthermore, it effectively bridges fundamental research with practical applications through the inclusion of data from both chemical engineering and environmental engineering. Critically, the dataset extends its reach into the **chemical-biological interface**, incorporating samples from molecular biology and systems biology. This comprehensive disciplinary matrix underscores the dataset's utility as a robust benchmark for evaluating AI models across complex, intersecting scientific challenges, rather than isolated domain-specific problems.

E Evaluation of Trend Alignment and Accuracy

E.1 Evaluation of Trend Alignment

To quantitatively assess trend alignment between simulated and experimental results, we employed the *Spearman Rank Correlation Coefficient* (denoted as ρ). This non-parametric measure evaluates the monotonic relationship between the rankings of simulated and experimental outcomes, making

Table 9: Disciplinary classification of the TOMATO-chem benchmark dataset.

Category	Count
Energy Materials	363
Polymeric Materials	359
Metallic Materials	268
Nanomaterials	316
Applied Physics	127
Analytical Chemistry	317
Inorganic Chemistry	392
Organic Chemistry	482
Chemical Engineering	196
Environmental Engineering	298
Molecular Biology	84
Systems Biology	62
Total	3264

it suitable for capturing trend consistency across diverse scientific problems.

The Spearman Correlation Coefficient is calculated as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{13}$$

where d_i is the difference between the ranks of the i-th simulated and experimental result, and n is the number of hypotheses in a given group. A Spearman Correlation Coefficient near 1 indicates strong trend alignment, meaning the simulated results closely mirror the relative ordering of experimental outcomes. Our *MC3-Sim* achieved a mean Spearman Correlation Coefficient of $\rho = 0.759$, significantly outperforming the baseline, as shown in Table 1.

To further assess the robustness of the simulator across diverse problems, we introduced the Perfect Consistency Indicator (PCI), a stringent metric that counts the number of question groups where the simulated results achieved perfect trend alignment with the experimental results ($\rho = 1$). Notably, our *MC3-Sim* achieved perfect trend alignment in 121 out of 178 question groups, highlighting its robustness and predictive fidelity.

E.2 Evaluation of Simulator Accuracy

For evaluating prediction accuracy, we used the *Root Mean Square Error* (RMSE) to quantify the deviation between simulated and experimental values:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{14}$$

where y_i is the experimental result for the i-th hypothesis and $\hat{y}_i$ is the simulated result. The *MC3-Sim* exhibited a lower RMSE than the Matched Score baseline (Yang et al., 2025c), signifying improved predictive accuracy.

To further evaluate the predictive accuracy of MC3-Sim against real-world experimental outcomes, we test it on an expanded dataset of 612 authentic scientific hypotheses. In addition to trend-level metrics such as Spearman correlation and the Perfect Consistency Indicator (PCI), we report predictive-error metrics including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Root Mean Squared Logarithmic Error (RMSLE). These complementary metrics evaluate whether MC3-Sim not only preserves the relative ordering of hypotheses but also produces calibrated normalized performance scores close to real experimental outcomes.

Below, we define each metric used in the evaluation:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{15}$$

$$\text{RMSLE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\log(\hat{y}_i + 1) - \log(y_i + 1))^2} \tag{16}$$

As reported in Table 1, MC3-Sim achieves RMSE $= 0.245$, MAE $= 0.184$, and RMSLE $= 0.178$, improving over the Matched Score baseline with RMSE $= 0.267$, MAE $= 0.205$, and RMSLE $= 0.201$. *MC3-Sim* consistently outperforms the Matched Score baseline (Yang et al., 2025c) across all metrics. Ablation studies further reveal the contributions of individual components: the removal of Critical Points results in a slight performance decline, while the exclusion of Component Extraction leads to more significant degradation. These results underscore the importance of both critical point identification and component extraction in achieving high predictive accuracy and robustness in simulation outcomes.

F Real Wet-lab Outcome Sequential Deployment

Benchmark construction. We construct a real-wet-lab outcome deployment benchmark from

Table 10: Per-question results for real-outcome sequential deployment. Lower N_{trial} is better.

Question	# Hypotheses	Human Experts	MC3-Rank
Q1	100	63	40
Q2	100	26	21
Q3	100	48	36
Average	100	45.7	**32.3**

three scientific questions, each associated with 100 candidate hypotheses and experimentally validated outcomes. The benchmark contains 300 hypotheses in total. To prevent benchmark leakage, we enforce temporal and source separation: all public records are curated from 2025–2026 literature, and additional outcomes are drawn from unpublished results from collaborating laboratories. During ranking, neither MC3-Rank nor human experts have access to the outcome labels of untested hypotheses.

Sequential feedback protocol. The evaluation instantiates the deployment setting with a real experimental executor E_{lab}. At timestep t, the selector, either MC3-Rank or a human expert, chooses an untested hypothesis h_t. Only after this choice is made, the corresponding real wet-lab outcome $r_t = E_{\text{lab}}(h_t)$ is revealed. The selector may then analyze the observed feedback and update the next choice. This process continues until the best-performing hypothesis in the candidate pool is selected. We report N_{trial}, the number of selected hypotheses required to reach the best-performing one.

Human expert baseline. The human baseline is produced by three domain experts following the same sequential protocol as MC3-Rank. Experts observe only outcomes of previously selected hypotheses and do not access hidden labels for untested candidates.

Results. Across the three scientific questions, MC3-Rank reaches the best-performing hypothesis in 32.3 trials on average, whereas human experts require 45.7 trials on average. This corresponds to a 29.3% reduction in screening cost. The result suggests that MC3-Rank can use real experimental feedback more efficiently than human expert sequential selection in this post-cutoff deployment setting.

Table 11: Composition of different noise levels.

Noise Level	Extrema	Plateaus	Cliffs
Simple	0–10	0–2	0–2
Medium	0–30	0–6	0–6
Complex	≥ 30	≥ 3	≥ 3

G Different Levels of Distortion

We collaborated with scientific PhD students to identify and design three common types of distortions encountered in scientific research: local maxima/minima, plateaus, and cliffs. These distortion patterns reflect typical challenges in hypothesis evaluation, drawing on domain expertise and established heuristics to ensure relevance. We defined three distinct distortion levels, Simple Noise, Medium Noise, and Complex Noise, and incorporated them into the hypothesis embedding function $\phi(\cdot)$ to simulate increasingly challenging feedback conditions.

In scientific hypotheses, biases in understanding key factors can result in specific distortion patterns. For instance, when adding guanidine sulfate to polymer thermoelectric materials, recognizing it solely as a salt providing hydrogen bonds for the reaction, while overlooking its influence on the entropy of redox pairs, can lead to a local maximum. Similarly, misjudging irrelevant factors can create a plateau effect. Conversely, misjudging critical factors can produce a cliff if the factor is incorrectly assumed to inhibit the reaction entirely. These elements present significant challenges in optimization problems within scientific research.

Through extensive discussions with scientific experts, we conducted a statistical analysis to evaluate the discrepancies between wet-lab results and empirical expected outcomes across diverse experimental scenarios. Based on the observed distribution, we categorized the top 35% of observed gaps as Simple Noise, the middle 40% as Medium Noise, and the bottom 25% as Complex Noise.

These distortions, along with their detailed quantities, are outlined in Table 5. We constructed three distinct noise levels to evaluate the robustness of *MC3-Rank* under complex scientific feedback conditions. By comparing Table 6, we observed that with the introduction of noise, the experiment-guided ranking method requires a significantly higher number of simulation feedback iterations to identify the ground truth scientific hypothesis as the complexity of the noise increases.

H Performance Comparison of Different Functions in the Simulator

This supplementary study was conducted to validate the robustness of our core mathematical modeling. While the Gaussian function was selected for its well-behaved mathematical properties and intuitive alignment with our core assumptions, the framework's success is not tied to any single function form. The choice of function can be viewed as a tunable hyperparameter. Table 12 shows that, on the evaluated subset, all tested functions provide effective ranking guidance, underscoring the framework's overall robustness.

I Evaluation of Experiment-Guided Ranking and Its Societal Benefits

The intricate web of scientific knowledge, combined with the multitude of factors influencing hypothesis analysis, often leads to the gradual accumulation of small cognitive biases. These biases can significantly distort the final experimental outcomes, creating substantial disparities between expected and observed results. To address this challenge, we conducted a comparative analysis between experiment-guided ranking and pre-experiment ranking. Our findings reveal that the experiment-guided ranking method demonstrates a marked improvement because it allows reflective use of previous simulation and experimental results. This iterative reflection provides more contextually relevant information, enabling the selection of the next hypothesis with greater precision.

Beyond academic and scientific advancements, this approach holds promising societal impacts. By reducing the need for extensive wet-lab experiments, it can lower research costs and accelerate the development of new materials and drugs, potentially improving healthcare access and environmental sustainability. Additionally, enhanced efficiency in hypothesis testing could foster innovation in industrial applications, such as cleaner energy solutions.

J Scalability Analysis

To evaluate the scalability of our proposed method, we expanded the pool of candidate hypotheses from $N = 64$ to $N = 128$. The results, presented in Table 13, show that *MC3-Rank* required 15.2 trials for 64 candidates and 30.7 trials for 128 candidates. In contrast, Uninformed Search and Pre-experiment Ranking required 64.5 and 51.3 trials, respectively, for the larger candidate pool. The observed slope of approximately $0.24 \times N$ confirms the near-linear scalability of *MC3-Rank* and shows that it retains a substantial cost advantage even as the candidate pool expands.

K Discussion on the Relationship with Active Learning and Sim2Real

While our proposed experiment-guided ranking framework shares the high-level conceptual reliance on iterative feedback loops with active learning (AL) (Settles, 2009) and sim-to-real transfer (Wagenmaker et al., 2024), it is fundamentally distinct in its operational paradigm, optimization objective, and interpretability.

K.1 Operational Paradigm: Training-Free Reasoning vs. Model Retraining

Standard active learning strategies typically rely on a *retraining-based* workflow. Similarly, traditional sim-to-real methods often require the collection of real-world data to fine-tune policies and bridge the domain gap via gradient updates (Wagenmaker et al., 2024). In contrast, our approach operates under a training-free paradigm powered by in-context reinforcement learning. By keeping the underlying LLM frozen, our agent performs reasoning and hypothesis prioritization from the very first trial without the computational overhead or data requirements of gradient-based updates.

K.2 Objective: Optimization-Centric vs. Generalization-Centric

The primary objective of classical active learning in physics (Ding et al., 2023) and chemical space exploration (Radford et al., 2025; Smith et al., 2018; Khalak et al., 2022) is often rooted in generalization. Conversely, our framework is strictly optimization-centric:

$$\arg\min_{\pi} \; N^{\pi}_{\text{trials}} \quad \text{subject to} \quad h^* = \arg\max_{h \in \mathcal{H}} E(h). \tag{17}$$

We are not concerned with reducing uncertainty in irrelevant regions of the space. Instead, our agent focuses solely on identifying the optimal hypothesis h^* with the minimum number of trials.

K.3 Interpretability: Structure-Aware Reasoning vs. Scalar Acquisition

A significant limitation of traditional AL is the opacity of its decision-making process. Candidates are typically selected based on scalar acquisition

Table 12: Performance comparison of different simulator variants on a 30-question validation subset.

Function	**Spearman Corr.** (↑)	**RMSE** (↓)	**Perfect Consistency** (↑)
Linear Function	0.9708	0.1959	24/30
Gaussian Function	0.9600	0.2147	26/30
Absolute Value Function	0.9626	0.2595	23/30
Quadratic Function	0.9682	0.3996	22/30

Table 13: Number of experiments required to identify the ground truth hypothesis across methods.

Method	**Trials** ($N = 64$)	**Trials** ($N = 128$)
Uninformed Search	32.5	64.5
Pre-experiment Ranking	28.6	51.3
MC3-Rank (ours)	**15.2**	**30.7**

functions, which offer little insight into why a specific candidate is promising. Our framework addresses this by employing structure-aware reasoning. As detailed in Section 2.3, our agent prioritizes candidates through explicit component clustering and reasoning steps. This allows domain experts to audit the decision logic and understand the rationale behind each selection.

L Sensitivity Analysis on Model Architectures and Environmental Noise

To demonstrate the robustness of our framework, we conducted a comprehensive sensitivity analysis focusing on two critical dimensions: the choice of LLM backbones and the impact of simulator noise levels.

L.1 Robustness Across LLM Architectures

We evaluated the generalizability of our component-based reasoning framework by varying the underlying models for both the simulator environment and the agent's policy. The results indicate model agnosticism: our method consistently outperforms the baseline regardless of the model combination used. Replacing the policy model with a stronger reasoning engine (Claude 3 Sonnet) further reduces the average trials to 14.12, confirming that the core advantage of our approach lies in the structural reasoning framework rather than the parametric knowledge of a specific LLM.

L.2 Resilience to Simulator Noise Assumptions

Beyond model architecture, we stress-tested the policy's resilience to environmental stochasticity. While the absolute efficiency naturally correlates with signal fidelity, MC3-Rank demonstrates effective mitigation of misleading signals. Even under the Complex noise regime, our approach significantly outperforms ablated variants, proving that the structural analysis module effectively filters noise and prevents the agent from being misled by individual erroneous data points.

M Analysis of Failure Modes and Handling of Emergent Properties

M.1 Case Study on Failure Modes: Complex Physicochemical Mechanisms

While MC3-Rank demonstrates high accuracy in general retrieval, failures can occur when the underlying scientific mechanism relies on subtle, high-order physical interactions rather than direct compositional effects. A concrete instance of divergence was observed in the domain of thermoelectric materials. The ground truth hypothesis introduced **(Gdm)$_2$SO$_4$** to interact with the **K$_3$Fe(CN)$_6$/K$_4$Fe(CN)$_6$** redox pair. Theoretically, this addition boosts the Seebeck coefficient by significantly increasing the reaction entropy difference:

$$S_e = \frac{\Delta E}{\Delta T} = \frac{\Delta S_{rc}}{nF} \tag{18}$$

The system also involves complex transport dynamics governed by the Eastman entropy of transfer:

$$S_{td} = \frac{\sum_i q_i n_i^0 \hat{S}_i D_i}{\sum_i q_i^2 n_i^0 D_i} \tag{19}$$

In this case, the ranking policy correctly identified the key component but misinterpreted its functional

Table 14: Comparison between standard active learning / sim-to-real and our proposed framework.

Feature	**Standard AL / Sim2Real**	**Ours (Experiment-Guided Ranking)**
Paradigm	Training-dependent; requires iterative retraining or fine-tuning.	Training-free; uses in-context RL with a frozen model.
Data Requirement	High data demand to update model weights effectively.	Effective in extreme data scarcity.
Objective	Generalization; reduces global uncertainty.	Optimization; directly targets the optimal hypothesis h^*.
Interpretability	Opaque; based on scalar acquisition values.	Transparent; based on component reasoning and clustering.

Table 15: Performance comparison across different LLM backbones.

Configuration (Simulator / Policy)	**Avg. Trials**	**Improvement**
Baseline (Pre-Experiment Ranking)	28.60	-
Ours (Gemini 2.5 Flash-Lite / GPT-4o-mini)	19.56	+31.6%
Ours (GPT-4o-mini / GPT-4o-mini)	15.20	+46.8%
Ours (GPT-4o-mini / Claude 3 Sonnet)	**14.12**	**+50.6%**

role, underestimating the hypothesis's rank. This suggests that while the framework excels at structural and functional matching, it may struggle with mechanisms involving implicit higher-order thermodynamic derivatives.

M.2 Handling Emergent and Non-Decomposable Properties

M.2.1 Mathematical Handling via Multiplicative Gating

Our scoring model (Equation 12) is designed to move beyond simple additive similarity. Crucially, it incorporates a multiplicative gating term:

$$\text{Score}(h) \propto \prod_{i \in \mathcal{C}} \mathbf{1}(s_i > 0) \tag{20}$$

This term ensures that if a critical point is missing, the total score drops to zero. This effectively models all-or-nothing emergent behaviors often seen in physics and biology.

M.2.2 Functional Abstraction vs. Atomic Decomposition

Our decomposition strategy (Figure 3) operates at the functional mechanism level rather than the atomic level. For example, in high-temperature superconductors, a hypothesis is not decomposed into raw atoms, but into functional units such as cuprate planes, charge reservoirs, and interlayer spacing. By abstracting these emergent behaviors into distinct functional units, the emergence is effectively encapsulated within the component, allowing the linear ranking framework to remain valid.

M.2.3 Robustness to Non-Linear Landscapes

To empirically validate this, we introduced landscape cliffs in our distortion experiments. These cliffs simulate scenarios where a small change in similarity results in a drastic drop in performance. Our results demonstrate that MC3-Rank maintains high ranking efficiency even in these non-smooth regimes, confirming that the combination of functional abstraction and critical gating provides robustness against non-linear complexities.

N Theoretical Analysis of Search Complexity Reduction via Functional Decomposition

In this section, we provide a formal analysis of how experiment-guided ranking, driven by functional decomposition, fundamentally reduces the search complexity compared to traditional pre-experiment ranking baselines. We model the hypothesis discovery process as an optimization problem over a combinatorial space.

N.1 Problem Formulation

Definition 1 (Hypothesis Space and Component Space). Let $\mathcal{K} = \{k_1, k_2, \ldots, k_m\}$ be the univer-

sal discrete set of functional component clusters, where $K = |\mathcal{K}|$ is the total number of unique functional modules available in the domain. We define a hypothesis h as a composition of functional modules selected from $\mathcal{K}$. The hypothesis space $\mathcal{H}$ is the set of all valid combinations bounded by a maximum structural complexity m:

$$\mathcal{H} = \{h = \{k_1, \ldots, k_n\} \subseteq \mathcal{K} \mid n \leq m\} \quad (21)$$

Assuming a standard combinatorial setting, the cardinality of the hypothesis space scales exponentially with the complexity m:

$$N = |\mathcal{H}| \approx \sum_{n=1}^{m} \binom{K}{n} \approx \mathcal{O}(K^m) \quad (22)$$

where N represents the total number of candidate hypotheses.

N.2 Baseline Complexity: Black-box Search

In traditional pre-experiment ranking or naive search, hypotheses are treated as atomic black boxes. The internal structure is opaque to the selection policy.

Proposition 1 (Information Isolation). Without decomposition, the feedback from an experiment on hypothesis h_i, denoted as $y_i = f(h_i)$, provides information strictly limited to the joint probability $P(h_i)$. The expected number of trials needed to identify the optimal hypothesis scales linearly with the size of the hypothesis space:

$$\mathbb{E}[T_{\text{atomic}}] \propto N \approx \mathcal{O}(K^m) \quad (23)$$

Conclusion: The baseline suffers from the curse of dimensionality, yielding exponential complexity with respect to the hypothesis structural complexity m.

N.3 MC3-Rank Complexity: Modular Optimization

Our approach leverages functional decomposition, transforming the problem from identifying the optimal combination h^* to identifying the set of optimal functional modules $\mathcal{K}^* \subset \mathcal{K}$.

Proposition 2 (Feedback Propagation and Pruning). MC3-Rank attributes the experimental outcome y to the marginal utility of individual modules $k \in \mathcal{K}$. A negative feedback on a module k_{bad} allows the agent to update the module-level posterior $P(k_{\text{bad}} \mid y)$ and effectively prune not just the tested hypothesis, but the entire subset of hypotheses $\mathcal{H}_{\text{sub}} \subset \mathcal{H}$ containing that module:

$$\mathcal{H}_{\text{sub}} = \{h \in \mathcal{H} \mid k_{\text{bad}} \in h\} \quad (24)$$

Complexity Analysis. The efficiency of this reduction relies on the degree to which the target property can be decomposed. We analyze the asymptotic complexity across three regimes:

- **Best-case ($\mathcal{O}(K)$):** In ideal scenarios where hypotheses are perfectly functionally modular, the agent essentially performs a parallel screening of the component space $\mathcal{K}$.
- **Average-case:** In practical settings, while component interactions introduce noise, the search complexity is fundamentally governed by the effective number of functional modules per hypothesis, not the raw count of candidates. The observed slope in Table 13 reflects this advantage.
- **Worst-case ($\mathcal{O}(K^m)$):** In scenarios with strong entanglement or pure emergence, feedback on individual components provides zero information gain. In this limit, the strategy degenerates to baseline black-box search, recovering exponential combinatorial complexity.

Applicability and Limitations. This theoretical boundary highlights that our decomposition-based framework is particularly well-suited for experimental sciences such as chemistry, biology, and materials science, where specific functional motifs often carry intrinsic, transferable utility. Conversely, we acknowledge limitations in domains like theoretical physics or highly abstract mathematical derivation tasks, where properties are often strictly emergent or governed by high-order nonlinear equations.

O Detailed Comparison with MOOSE-Chem and MOOSE-Chem2

This section clarifies the task-level and methodological distinctions between MOOSE-Chem3 (MC3) and other works in the MOOSE-Chem series, namely MOOSE-Chem (Yang et al., 2025c) (MC1) and MOOSE-Chem2 (Yang et al., 2025b) (MC2). While the three works share the broader goal of LLM-driven scientific hypothesis discovery, they address fundamentally different problems and operate under different feedback regimes.

Task-level distinction. MC1 introduces *exploratory hypothesis discovery*: given a research background and an inspiration corpus, it generates candidate hypotheses via inspiration retrieval and structured search over functional components, and ranks them statically using LLM-internal scoring. MC2 targets *fine-grained hypothesis discovery* via hierarchical search, refining hypotheses at progressively finer granularity to recover detailed mechanistic content. MC3 addresses a third, orthogonal problem: *experiment-guided hypothesis ranking*, in which a fixed candidate pool is sequentially prioritized under iterative experimental feedback, with the explicit goal of minimizing the number of trials required to identify the optimal hypothesis.

Feedback regime. MC1 and MC2 operate entirely in the *pre-experiment phase*: their ranking and refinement decisions rely on static LLM priors and have no access to empirical outcomes. In contrast, MC3 is the first in the series to introduce a *closed-loop* setting, where each selection is informed by feedback from previously tested hypotheses. This shift from static scoring to feedback-driven sequential decision-making is the central methodological departure of MC3 from MC1 and MC2.

Shared methodological thread. Despite these differences, all three works share a common foundation: hypothesis decomposition into functional components. MC1 establishes this decomposition view to structure hypothesis generation and retrieval; MC2 leverages it for hierarchical refinement; and MC3 builds on it to attribute experimental feedback to reusable components, enabling evidence transfer across structurally related hypotheses. The decomposition primitive thus serves as a connective thread, but is operationalized for distinct purposes across the series.

P Reproducibility and Computational Setup

P.1 Models and API Configuration

All main experiments are conducted using GPT-4o-mini (OpenAI, 2024) via the OpenAI API as the default backbone. For sensitivity analysis, we additionally use Claude 3 Sonnet (Anthropic API) and Gemini 2.5 Flash-Lite (Google API). As these are proprietary closed-source API models, parameter counts and underlying compute infrastructure are not publicly disclosed by the providers.

Decoding parameters use temperature $= 0$ across all phases (component extraction, clustering, cluster selection, hypothesis selection, scoring, and feedback analysis) for deterministic behavior. No hyperparameter search was performed over LLM decoding parameters. The prompt templates used in each phase are provided in the released code repository.

P.2 Computational Budget

The dominant computational cost is API inference. For each scientific question in the TOMATO-chem ranking evaluation, MC3-Rank requires on average $\sim$15 ranking iterations, with each iteration invoking the LLM for component extraction, clustering, cluster selection, hypothesis selection, and feedback analysis. Across the 51 TOMATO-chem problems and ablation/sensitivity configurations, the total experiments consume on the order of 10^5 API calls. We do not use local GPUs for inference; all computation is performed through API endpoints.

P.3 Dataset and Code Licensing

Existing datasets. TOMATO-chem (Yang et al., 2025c) is released by its authors for research use on hypothesis ranking, which matches our use. We use it under its original release terms.

Newly curated benchmarks. Our 612-hypothesis simulator-fidelity benchmark and 300-hypothesis real-outcome sequential deployment benchmark are curated from publicly available peer-reviewed scientific literature, supplemented in the latter case with previously unreleased experimental results contributed by collaborating laboratories. Both benchmarks are intended for research use on experiment-guided hypothesis ranking and will be released under the **CC BY 4.0** license, consistent with the research purpose of this work.

Code. The released code will be made available under the **Apache 2.0** license.

Data content. All curated data consists of research questions, scientific hypotheses, and experimental outcomes extracted from scientific publications. The data contains no personally identifying information and no offensive content.

Q Human Participation Protocol

Several parts of this work involve human participation. We summarize each below.

Q.1 Real-outcome Sequential Deployment Comparison

The human-expert baseline in the real-outcome sequential deployment benchmark involves domain researchers performing sequential hypothesis selection under the same protocol as MC3-Rank.

Participants. Three domain researchers in chemistry, materials science, and biology participated across the three scientific questions, recruited as research collaborators. No external crowdsourcing platform was used.

Protocol and instructions. Each participant was given the research question and the candidate hypothesis pool with outcome labels hidden. The instruction was as follows: at each step, select one untested hypothesis based on prior observations; after submission, the wet-lab outcome of the selected hypothesis is revealed; then proceed to the next selection until either the best-performing hypothesis is identified or all hypotheses are tested. The number of trials required to identify the best-performing hypothesis is recorded.

Compensation. Participants contributed as part of standard research collaboration; no separate per-task payment was made.

Consent. All participants were informed in advance that their selection trajectories would be aggregated and reported in this paper as part of the human-expert baseline, and consented to such use.

Q.2 Distortion Design Consultations

For the noise robustness experiments (Appendix G), we consulted scientific PhD students to identify common failure modes in hypothesis evaluation (local maxima/minima, plateaus, and cliffs). These consultations were conducted as informal research discussions with collaborators; no structured data collection or evaluation was performed with them.

Q.3 Ethics Review

The human involvement in this work consists of research collaborators participating in protocol design and benchmark evaluation as part of standard scientific collaboration, and does not constitute human-subjects research requiring institutional ethics review under our institutional guidelines. No personal data, sensitive attributes, or behavioral data were collected.