MOOSE-Chem3: Toward Experiment-Guided Hypothesis Ranking via Simulated Experimental Feedback

Anonymous Author(s)
Affiliation
Address
email

Abstract

Hypothesis ranking is a crucial component of automated scientific discovery, particularly in natural sciences where wet-lab experiments are costly and throughput-limited. Existing approaches focus on *pre-experiment ranking*, relying solely on large language model's internal reasoning without incorporating empirical outcomes from experiments. We introduce the task of *experiment-guided ranking*, which aims to prioritize candidate hypotheses based on the results of previously tested ones. However, developing such strategies is challenging due to the impracticality of repeatedly conducting real experiments in natural science domains. To address this, we propose a simulator grounded in three domain-informed assumptions, modeling hypothesis performance as a function of similarity to a known ground truth hypothesis, perturbed by noise. We curate a dataset of 124 chemistry hypotheses with experimentally reported outcomes to validate the simulator. Building on this simulator, we develop a pseudo experiment-guided ranking method that clusters hypotheses by shared functional characteristics and prioritizes candidates based on insights derived from simulated experimental feedback. Experiments show that our method outperforms pre-experiment baselines and strong ablations.

1 Introduction

Scientific discovery plays a major role in advancing human society (Coccia, 2019). Recently, there have been promising advances in automating certain stages of the scientific process using large language models (LLMs) (Luo et al., 2025; Cambria et al., 2023).

One of the most critical stages is hypothesis ranking: given a large set of automatically generated hypotheses, which one should be tested in a real experiment first? This question is particularly important in natural science domains, where experiments are costly and resource-constrained, necessitating efficient prioritization strategies to minimize experimental effort.

Previous methods for hypothesis ranking (Yang et al., 2024b; Si et al., 2024) primarily rely on evaluations based solely on LLMs' internal reasoning, without incorporating any empirical feedback from experiments. We refer to this approach as *pre-experiment ranking*, as hypotheses are prioritized before any experimental evidence is gathered.

In contrast, we propose a new task: *experiment-guided ranking*, which focuses on dynamically prioritizing hypotheses by leveraging feedback from sequentially performed experiments. Rather than conducting all experiments upfront, this approach iteratively updates the ranking based on available experimental results, aiming to accelerate the discovery of promising hypotheses while minimizing the total number of experiments required. However, developing strategies for experiment-guided

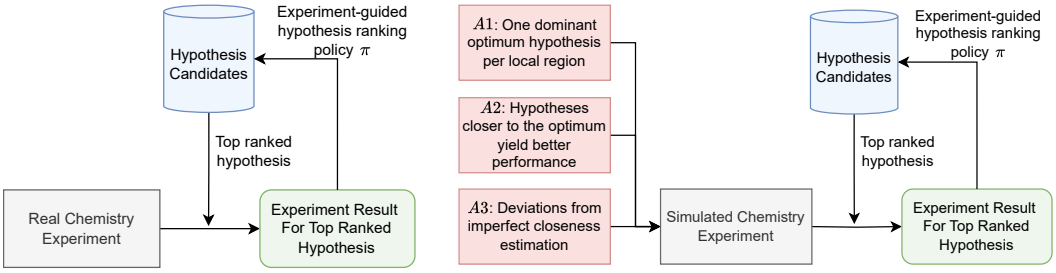
Submitted to the AI for Science workshop (NeurIPS 2025)

ranking in natural science domains such as chemistry is challenging, as it is impractical to rely on real laboratories to repeatedly conduct experiments. In other words, the lack of scalable access to meaningful experimental feedback remains a key barrier to researching experiment-guided ranking strategies.

Despite the challenges of obtaining real experimental feedback, we posit that simulating such feedback is feasible under three foundational assumptions. To illustrate these, consider a latent space where the x-axis (potentially multidimensional) parameterizes candidate hypotheses, such that each coordinate corresponds to a distinct hypothesis variant, and the y-axis denotes the associated experimental feedback (e.g., performance). Assumption 1 (*A1*) posits that within any sufficiently local neighborhood of the hypothesis space, there exists at most one dominant optimum, corresponding to a ground truth hypothesis (e.g., reported in the literature). Assumption 2 (*A2*) states that hypotheses closer to this dominant maximum are more likely to yield more competitive experimental feedback. Assumption 3 (*A3*) states that real experimental feedback can be viewed as idealized feedback—defined by *A1* and *A2*—perturbed by an unknown deviation term due to the imperfect representation of hypothesis closeness in the hypothesis space.

Specifically, the ground truth hypothesis is treated as the local optimum in the hypothesis space, and the performance of neighboring hypotheses is modeled as a function of their similarity to this optimum. Since real-world representations of hypothesis similarity are inherently imperfect, this relationship is subject to systematic deviations, whose effects on the simulator's fidelity are analyzed in this work.

The experiment-guided ranking task with real and simulated experiment feedback can be described by Figure 1. The primary goal of the simulator is to *enable systematic research on experiment-guided ranking strategies* by providing accessible and high-fidelity approximations of experimental feedback, which are otherwise prohibitively costly or unavailable. Ultimately, the aim is to deploy these strategies in real experimental settings to reduce the overall experimental costs.



(a) Real experiment feedback. (b) Simulated experiment feedback.

Figure 1: Experiment-guided hypothesis ranking using real and simulated feedback. $A1$, $A2$, and $A3$ illustrate our three foundational assumptions in a concise manner (introduced in § 2.1.1).

Guided by these assumptions and insights from chemistry experts, we construct a simulator that closely approximates real wet-lab experimental outcomes. To evaluate its fidelity, we curate a dataset of 30 groups of chemistry hypotheses, each consisting of 3–6 related hypotheses along with their experimentally reported performance, sourced from published literature (in total 124 <hypothesis, performance> pairs). Our simulator outperforms strong baselines, including widely used similarity-based evaluation metrics for hypothesis comparison (Yang et al., 2024b). Building on this foundation, we propose a new task: developing more accurate simulators for experimental feedback. Using the simulator to approximate experimental outcomes, we develop an pseudo experiment-guided ranking method leveraging functional clustering of hypotheses. Clustering enables effective transfer of insights from previously tested hypotheses to untested ones sharing similar functional elements, rather than evaluating each hypothesis in isolation.

Specifically, hypotheses containing elements with similar functional relevance—regardless of whether identical or distinct—are grouped together, allowing hypotheses to belong to multiple clusters. Our method prioritizes clusters based on accumulated experimental feedback and subsequently selects the most promising hypothesis within each. Experiments demonstrate that our approach outperforms

both pre-experiment ranking methods and strong ablation variants. Overall, the contributions of this paper are:

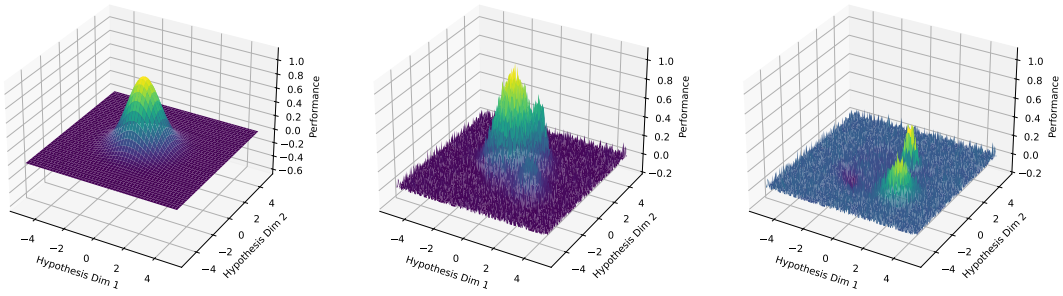
- We formalize the task of *experiment-guided ranking* and highlight a key challenge in the natural sciences: the lack of scalable access to wet-lab experimental feedback. To address this, we propose the use of simulators and release a curated dataset of 124 chemical hypotheses with annotated performance collected from the literature.
- We introduce three foundational assumptions for simulating experimental feedback, provide a mathematical formalization of the simulation process, and construct a high-fidelity simulator that approximates real wet-lab outcomes under these assumptions.
- We develop a clustering-based pseudo experiment-guided ranking method that leverages simulated feedback and structural similarities among hypotheses. Experimental results show that our method outperforms both pre-experiment baselines and strong ablation variants.

2 Methodology of Simulator Construction

2.1 Foundational Assumptions and Formalization

Our simulator construction is guided by three foundational assumptions derived from expert consultations in the chemistry domain. These assumptions provide a principled basis for modeling simulated experimental outcomes of untested chemical hypotheses, enabling systematic investigation of experiment-guided ranking strategies.

2.1.1 Foundational Assumptions



(a) Idealized performance landscape ($A1$ + $A2$)).

(b) Realistic performance landscape ($A1$ + $A2$ + $A3$).

(c) Deviations from imperfect closeness estimation ($A3$).

Figure 2: Illustration of the three fundamental assumptions for simulator construction.

We posit that real experimental feedback within a hypothesis space can be simulated under the following assumptions:

1. (*A1*) Within any sufficiently local neighborhood of the hypothesis space, there exists at most one dominant optimum, corresponding to a ground truth hypothesis.
2. (*A2*) Among hypotheses in the vicinity of a dominant optimum, those that are closer to it are more likely to yield better experimental feedback.
3. (*A3*) Real experimental outcomes deviate from the idealized structure described in *A1* and *A2* due to the imperfect representation of hypothesis closeness in the hypothesis space.

Figure 2 visually illustrates these assumptions. In the ideal scenario (Figure 2a), hypotheses are embedded in a latent hypothesis space such that the Euclidean distance ("closeness") between a hypothesis and the dominant optimum hypothesis accurately reflects similarity in terms of how they perform on a research question, creating a smooth, unimodal performance landscape.

However, practical scenarios differ substantially, since the distance ("closeness") between hypotheses—whether assessed by scientists or LLMs—may not accurately reflect functional similarity. For example, a chemical hypothesis may include a useful functional component whose contribution is not

fully recognized, causing it to be placed farther from the dominant peak than it should be—resulting in a spurious secondary peak. Conversely, a suboptimal hypothesis may appear closer to the dominant peak than warranted, forming a local valley. These distortions result in a performance landscape such as that in Figure 2b, with unexpected secondary peaks and valleys. Figure 2c further isolates this deviation, representing the discrepancy between oracle and practical understandings of closeness.

We now formalize these assumptions by defining a mathematical model that makes the relationship between hypothesis embeddings, similarity, and performance explicit.

2.1.2 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 for any hypothesis h in the vicinity of h^* as:

$$f(h, h^*; q, \phi^*(\cdot)) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\phi^*(h \mid q) - \phi^*(h^* \mid q)\|^2}{2\sigma^2}\right), \tag{1}$$

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 $\|\phi^*(h \mid q) - \phi^*(h^* \mid q)\|$.

We model the idealized performance surface as a Gaussian-like function centered at $\phi^*(h^* \mid q)$, yielding a strictly unimodal landscape that decays smoothly with increasing distance from the optimum h^* (Figure 2a). While the true performance landscape in chemical space may not be strictly Gaussian, the isotropic Gaussian form serves as a tractable and interpretable approximation in the latent space. This modeling choice directly reflects Assumptions *A1* and *A2*.

However, practical simulations rely on imperfect embeddings of hypotheses into the latent space, stemming from limitations in domain understanding—no matter whether the embedding is performed (internally) by human experts or LLMs. Consequently, this leads to distortions in perceived "closeness", effectively warping the positions of hypotheses in latent space. Such a distorted hypothesis embedding $\tilde{\mathcal{H}}$ yields a different observed structure:

$$\tilde{f}(h, h^*; q, \phi(\cdot)) = f(h, h^*; q, \phi^*(\cdot)) + \epsilon(h \mid q) \tag{2}$$

where $\phi(\cdot \mid q)$ is a practical embedding function that maps each hypothesis h into (somewhat distorted) positions in the latent hypothesis space for a research question q, and $\epsilon(h \mid q)$ represents a systematic correction term that accounts for the discrepancy between oracle embedding $\phi^*(\cdot \mid q)$ and the practical embedding $\phi(\cdot \mid q)$ under the context of q. As a result, the practical embedding $\tilde{\mathcal{H}}$ introduces systematic distortions in the latent space, leading to spurious local optima or unexpected valleys—effectively transforming the unimodal ideal surface into a noisier, multimodal one (Figure 2b).

Crucially, Figures 2a and 2b illustrate the same underlying performance-closeness relationship $f(h, h^*)$, differing only by $\phi(h)$, which is how hypotheses are embedded in the latent space. Figure 2c illustrates $\epsilon(h)$, the correction term that accounts for the discrepancy between the oracle embedding $\phi^*(\cdot)$ and the practical embedding $\phi(\cdot)$.

2.2 A Practical Implementation of $\phi(\cdot)$ with Chemistry Prior Knowledge

As discussed in § 2.1, the core objective of the simulator is to construct an embedding function $\phi(\cdot)$ that maps each hypothesis h into a latent space such that distances in this space reflect meaningful functional differences. Through extensive discussions with chemistry PhD students, we observe that a chemical hypothesis succeeds in addressing a research question primarily due to its underlying reaction mechanisms.

Specifically, an effective hypothesis typically comprises a set of chemically meaningful components—each contributing to distinct yet complementary sub-mechanisms—which together enable the

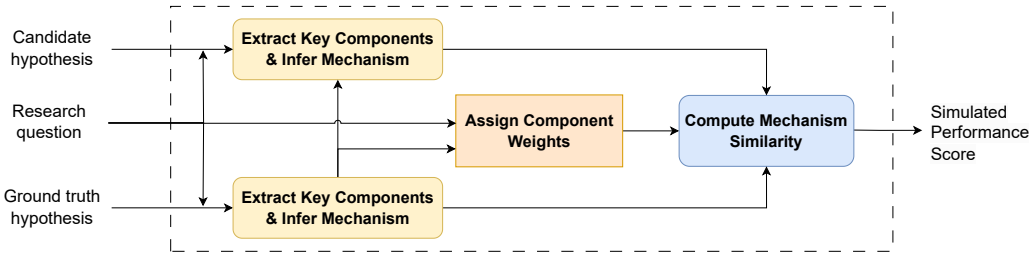


Figure 3: The internal structure of the simulator.

overall reaction to fulfill its intended function. The specific prompts and examples for extracting key chemical components and inferring mechanisms are provided in § C.

Informed by this domain knowledge, we design a simulator architecture illustrated in Figure 3. Each module corresponds to a subroutine implemented using an LLM with task-specific prompting. The simulator's goal is to estimate the latent-space distance $\|\phi(h \mid q) - \phi(h^* \mid q)\|$ between a candidate hypothesis h and a ground truth hypothesis h^*, conditioned on a research question q.

The simulation begins by decomposing both the candidate and ground truth hypotheses into a set of key chemical components, and identifying the reaction mechanism associated with each component in the context of the research question. The decomposition of h^* is performed first, serving as a reference. These reference components and mechanisms guide the decomposition of h, ensuring alignment in both granularity and mechanistic interpretation.

Concurrently, the *Assign Component Weights* module estimates the relative importance w_i of each component in the ground truth hypothesis, given the research question. A subset of these components—denoted $\mathcal{C}$—are labeled as critical, meaning they are considered necessary for the reaction to succeed. To elaborate on the role of $\mathcal{C}$, we provide illustrative examples in § D.

Next, the *Compute Mechanism Similarity* module compares each key component in h^* with its corresponding component in h, assigning a similarity score $s_i \in [0, 1]$ to each pair. These scores are then aggregated using a weighted sum, combined with a multiplicative penalty that enforces the presence of all critical components:

$$S(h \mid q; h^*) = \left(\prod_{i \in \mathcal{C}} \mathbf{1}_{s_i > 0} \right) \cdot \left(\sum_{i=1}^{K} w_i \cdot s_i \right), \quad \text{where} \quad \sum_{i=1}^{K} w_i = 1 \tag{3}$$

This formulation guarantees that $S(h^* \mid q; h^*) = 1$, since all components are present with maximal similarity ($s_i = 1$ for all i), resulting in zero distance from the ground truth. Similarity score S are thereby bounded in $[0, 1]$, and lower distances correspond to stronger functional alignment with the ground truth hypothesis. The resulting value is used as the simulated performance score.

The final distance between the candidate and ground truth hypotheses is then calculated as:

$$|\phi(h \mid q) - \phi(h^* \mid q)| = |S(h \mid q; h^*) - 1| \tag{4}$$

3 Methodology of Experiment-Guided Ranking

3.1 Problem Formulation

Given a research question q, a set of candidate hypotheses $\mathcal{H}$ is formed by selecting hypotheses generated by existing scientific discovery systems (Yang et al., 2024b) and ground-truth hypotheses from top-tier chemistry journals reporting high-quality lab experiments. The goal of experiment-guided ranking is to identify the optimal hypothesis $h^* \in \mathcal{H}$ with the highest experimentally measured performance using an experiment executor E.

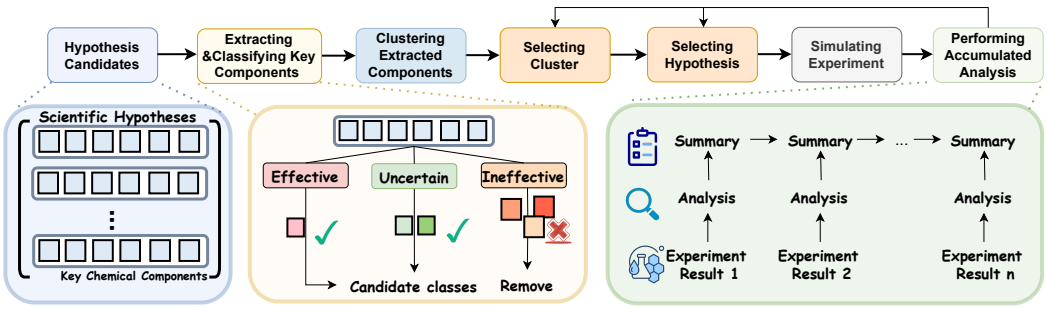


Figure 4: Experiment-guided ranking method.

Formally, we define the experiment executor as a function:

$$E : \mathcal{H} \to [0, 1] \tag{5}$$

that maps each hypothesis $h \in \mathcal{H}$ to a normalized performance score $s \in [0, 1]$. The normalization provides a unified performance metric across heterogeneous research hypotheses and varying problem settings q, and can be defined relative to a domain-specific state-of-the-art benchmark established by experts.

The primary goal is to find h^*. However, since each evaluation of $E(h)$ corresponds to a real or simulated experiment—which may be costly or time-consuming—a critical requirement is to identify h^* using as few experimental trials as possible. Accordingly, an effective experiment-guided ranking strategy must actively incorporate feedback from prior evaluations to guide subsequent selections, balancing exploration and exploitation under a limited experimental budget.

Thus, the problem can be reframed as finding a selection strategy that minimizes the number of trials required to identify the optimal hypothesis:

$$\arg\min_{\pi} \; N_{\text{trials}}^{\pi} \quad \text{subject to} \quad h^* = \arg\max_{h \in \mathcal{H}} E(h), \tag{6}$$

where π denotes the hypothesis selection strategy, and N_{trials}^{π} is the number of experiments required under strategy π to successfully discover h^*.

3.2 Methodology

We propose an experiment-guided ranking framework leveraging LLM agents, as illustrated in Figure 4. This design is informed by extensive consultations with chemistry domain experts, capturing key insights into hypothesis effectiveness.

These discussions identified a key criterion for hypothesis efficacy: effective hypotheses typically contain a sufficient number of key chemical components that collectively fulfill complementary mechanistic roles relevant to the research question q. Building upon this insight, our framework employs a structured, iterative approach comprising several distinct stages.

Step 1: Extraction, Classification, and Clustering of Functional Components. Each candidate hypothesis $h \in \mathcal{H}$ is decomposed into its functional chemical components—distinct substructures or motifs potentially contributing to the target reaction mechanism. These components are then classified into three categories: effective, uncertain, and ineffective. Components deemed ineffective are excluded from further consideration to reduce computational overhead, as the initial hypothesis set $\mathcal{H}$ may yield a large number of components. The remaining components are clustered based on functional similarity, with each cluster representing a distinct mechanistic contribution to solving q. Individual elements within a cluster correspond to specific functional components, each traceable to its originating hypothesis h.

Step 2: Cluster and Hypothesis Selection. Guided by the LLM's prior chemical knowledge, the framework identifies the cluster most likely to contain components highly relevant to q. Within this selected cluster, the LLM agent further selects a hypothesis h deemed most promising based on component relevance and prior understanding.

Step 3: Experiment Execution and Result Analysis. The selected hypothesis h is evaluated using the experimental executor (or simulator) E, yielding a normalized performance score s. The outcome of this experiment is then analyzed to evaluate the effectiveness of the chosen cluster and to validate or update the mechanistic assumptions made.

Step 4: Iterative Summarization and Refinement. Following each experimental evaluation, a detailed analysis is conducted, and the insights gained are integrated into a cumulative summary. This continually updated summary synthesizes insights from all prior analyses, highlighting effective clusters and guiding future hypothesis and cluster selections.

By iteratively leveraging prior chemical knowledge and empirical feedback, this framework systematically refines hypothesis prioritization. The overall objective is to efficiently identify the optimal hypothesis while minimizing the total number of experimental trials.

4 Experiment

We name our simulator as *CSX-Sim*, and the experiment-guided ranking method as *CSX-Rank*. All experiments are implemented with `GPT-4o-mini` (OpenAI, 2024).

4.1 Simulator: Evaluating the Simulator with Real Experiment Results

To rigorously evaluate the performance of our simulator on advanced chemical problems, we curated a benchmark of 30 cutting-edge research questions, each associated with 3–6 mutually related candidate hypotheses, totaling 124 hypotheses. Ground-truth experimental outcomes were sourced from published literature, covering major subfields of chemistry, as detailed in § E.1

For each hypothesis, simulated results were generated using the proposed *CSX-Sim* and compared against the annotated experimental outcomes. This trend comparison is illustrated in § E.2. Evaluation focused on two key criteria: (1) trend alignment, measured by Spearman rank correlation, which assesses whether the predicted performances correctly reflect the relative ranking of ground-truth annotations within each research question. This criterion is critical, as hypothesis ranking primarily depends on relative performance differences; absolute offsets (e.g., uniform biases of ± 0.2) have limited impact on ranking outcomes. Here we use "Perfect Consistency Indicator" (PCI) as metric, which quantifies the number of research questions for which the simulator achieves perfect trend alignment with experimental outcomes (2) predictive accuracy, quantified by root mean square error (RMSE), measuring absolute deviations between predicted and annotated performances. Detailed explanations of this metric and other predictive accuracy indicators are available in the § F. The comparative results are summarized in Table 1.

Simulator	Spearman Correlation (↑)	Perfect Consistency Indicator (↑)	RMSE (↓)
Matched Score	0.843	12/30	0.232
CSX-Sim	**0.960**	**26/30**	**0.213**
w/o CriticalPoints	0.950	23/30	0.229
w/o ComponentExtraction	0.864	12/30	0.272

Table 1: Validating the simulator with collected chemistry experiment results from literature.

Baseline and Ablation We adopt the "Matched Score" (Yang et al., 2024b) as our primary baseline, which evaluates hypotheses by measuring their similarity to ground-truth references through a reference-based comparison. Additionally, we conduct two ablation studies on *CSX-Sim* to assess the contribution of its key components: (1) The first ablation (w/o CriticalPoints) disables the labeling of critical components $\mathcal{C}$, as defined in Equation 3, allowing hypotheses that lack essential components to still receive positive feedback from the simulator; (2) The second ablation (w/o ComponentExtraction)

skips the extraction and weighting of critical components, directly computing mechanism similarity using prompts analogous to the final module in Figure 3.

Results Interpretation As shown in Table 1, *CSX-Sim* achieves superior performance across all metrics, with a Spearman correlation of 0.960, perfect consistency in 26 out of 30 questions, and the lowest RMSE of 0.213. Compared to the Matched Score baseline, *CSX-Sim* demonstrates substantial improvements in both trend alignment (+0.117 in Spearman) and robustness (+14 in PCI), while also reducing predictive error. Ablation studies further highlight the importance of critical component identification: removing CriticalPoints slightly degrades performance (Spearman 0.950, PCI 23/30), whereas omitting component extraction leads to significant drops in both alignment (Spearman 0.864) and accuracy (RMSE 0.272). These results underscore the necessity of fine-grained component analysis in achieving high-fidelity simulation feedback.

4.2 Experiment-guided Ranking

Due to the page limit, experiments in terms of the experiment-guided rankning method *CSX-Rank* can be found in Appendix § A. Ablation on the simulator in terms of the robustness can be found in Appendix § B.

5 Related Work

Most prior work on hypothesis ranking has focused on pre-experiment ranking. Some approaches assign a score to each hypothesis and rank them accordingly, providing a simple and efficient solution (Yang et al., 2024a,b). Others adopt a pairwise ranking strategy, evaluating hypothesis pairs one at a time (Si et al., 2024; Liu et al., 2025). However, these methods rely solely on the internal reasoning of LLMs and do not incorporate feedback from experimental outcomes.

To our knowledge, few existing works leverage experimental feedback in hypothesis-driven tasks. Notably, recent methods in mathematics (Romera-Paredes et al., 2024; Shojaee et al., 2024) and programming (Qiu et al., 2024) incorporate feedback loops by refining hypotheses based on verification outcomes. These approaches rely on domains where extremely efficient verifiers are available, allowing for rapid hypothesis testing and direct refinement rather than explicit ranking. In contrast, our work focuses on natural science domains, where real experiments are significantly more costly, making such exhaustive trial-and-error strategies impractical. This difference motivates the need for a more deliberate experiment-guided ranking process, where each experiment must inform the prioritization of future hypotheses due to limited experimental bandwidth.

Roohani et al. (2024) address hypothesis generation in a genetic perturbation setting, where task-specific feedback can be directly computed (e.g., via gene overlap). In contrast, our work focuses on constructing general-purpose simulators for natural science domains, with an emphasis on chemistry due to the availability of annotated novel hypotheses from the literature (Yang et al., 2024b).

6 Conclusion

We present a systematic framework for experiment-guided hypothesis ranking in chemistry, addressing the critical challenge of limited access to real experimental feedback. By formalizing three foundational assumptions, we develop a high-fidelity simulator that approximates experimental outcomes based on hypothesis similarity, validated against a curated dataset of 124 hypotheses with reported wet-lab results. Building on this simulator, our proposed *CSX-Rank* method leverages functional clustering and iterative feedback analysis to efficiently prioritize hypotheses during the discovery process.

Empirical evaluations demonstrate that *CSX-Rank* significantly outperforms pre-experiment baselines, reducing the number of trials required to identify ground-truth hypotheses by more than 50% on the TOMATO-chem dataset. Ablation studies and controlled noise experiments further highlight the importance of analytical components and feedback integration for robust performance under increasingly challenging conditions.

References

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 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*, pp. 995–1005, 2012.

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

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

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

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. LLM4SR: A survey on large language models for scientific research. *CoRR*, abs/2501.04306, 2025. doi: 10.48550/ARXIV.2501.04306. URL https://doi.org/10.48550/arXiv.2501.04306.

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

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, et al. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. Biodiscoveryagent: An ai agent for designing genetic perturbation experiments. *arXiv preprint arXiv:2405.17631*, 2024.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. Llm-sr: Scientific equation discovery via programming with large language models. *arXiv preprint arXiv:2404.18400*, 2024.

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

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

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

A Experiment-Guided Ranking: Baselines and Ablation Study

Data and Evaluation Metrics We evaluate experiment-guided ranking on the TOMATO-chem dataset (Yang et al., 2024b), which includes 51 chemical problems, each annotated with a ground-truth (gdth) hypothesis. For each problem, we use the MOOSE-Chem framework (Yang et al., 2024b) to generate 63 additional candidate hypotheses that are distinct from the ground truth, resulting in 64 hypotheses per research question (1 gdth and 63 negatives).

To measure performance, we define the metric N_{trials}, representing the number of simulation-based evaluations required to identify the ground-truth hypothesis for each of the 51 problems. Lower values of N_{trials} indicate more efficient hypothesis prioritization. Results are summarized in Table 2.

Method	N_{trials} ($\downarrow$)
Random Sampling	32.000
Pre-Experiment Ranking	33.280
CSX-Rank	**15.196**
w/o Clustering	27.980
w/o Clustering & Analysis	35.627
w/o Clustering & Analysis & Full Feedback	37.667

Table 2: Number of experiments required to identify the ground truth hypothesis across methods.

Baselines We consider two baselines: Pre-Experiment Ranking and Random Sampling. Pre-Experiment Ranking follows the strategy used in MOOSE-Chem (Yang et al., 2024b), where hypotheses are scored based on the model's prior knowledge and ranked accordingly, without incorporating any experimental feedback. Random Sampling selects hypotheses uniformly at random, serving as a simple yet unbiased baseline.

As shown in Table 2, both baselines require over 32 trials on average to identify the ground-truth hypothesis, with Pre-Experiment Ranking (33.28 trials) slightly underperforming Random Sampling (32.00 trials). This counterintuitive result indicates that relying solely on prior model knowledge—without feedback—can lead to suboptimal prioritization, as initial estimation errors may mislead the ranking more than random choice.

Ablation Study To assess the contribution of key components in *CSX-Rank*, we conducted ablation studies under three conditions: (1) removing functional clustering (*CSX-Rank w/o Clustering*); (2) further disabling feedback analysis (*CSX-Rank w/o Clustering & Feedback Analysis*); and (3) additionally limiting feedback to the 10 most recent simulation results (*CSX-Rank w/o Clustering & Feedback Analysis & Full Feedback*). As shown in Table 2, progressively removing these components leads to marked performance degradation, confirming the importance of clustering, analytical summarization, and sufficient feedback quantity for efficient hypothesis ranking.

B Simulator: Ablation on Different $\phi(\cdot)$ with Different Levels of Distortion

To assess how simulator quality affects ranking performance, we leverage the observation that experiment-guided ranking is fundamentally an optimization process: navigating the hypothesis space to identify candidates with superior experimental performance. A high-fidelity simulator facilitates this search by providing informative feedback, while a degraded simulator misleads the process, making it harder to reach the optimum. Based on this perspective, we systematically introduce controlled distortions that worsen simulator fidelity from an optimization standpoint.

Specifically, we collaborated with chemistry PhD students to design three types of distortions commonly encountered in chemical research: local maxima/minima, plateaus, and cliffs. These noise patterns capture typical challenges in hypothesis evaluation, informed by domain expertise and heuristics. We defined three distortion levels—Simple Noise, Moderate Noise, and Complex Noise—and incorporated them into the hypothesis embedding function $\phi(\cdot)$ to simulate increasingly challenging feedback conditions. The composition and classification of constructed noise are detailed in § G

We evaluated *CSX-Rank*, *CSX-Rank w/o Clustering*, and *CSX-Rank w/o Clustering & Analysis* across three noise scenarios of increasing complexity: Simple Noise (3 maxima, 3 minima, 1 cliff, 2 plateaus), Moderate Noise (8 maxima, 12 minima, 3 cliffs, 4 plateaus), and Complex Noise (38 maxima, 22 minima, 4 cliffs, 4 plateaus).

As shown in Table 3, increasing noise complexity progressively degraded performance across all methods, as reflected by higher N_{trials}. *CSX-Rank* consistently outperformed its ablated variants, maintaining a substantial efficiency margin even under Complex Noise (32.7 vs. 36.5 and 40.5 trials). These results highlight the robustness of functional clustering and feedback analysis in mitigating misleading signals and preserving search efficiency. The findings align with Section A, underscoring the critical role of each component in navigating noisy hypothesis spaces.

Method	N_{trials} (Simple Noise)	N_{trials} (Medium Noise)	N_{trials} (Complex Noise)
CSX-Rank	21.804	26.608	32.706
w/o Clustering	32.706	35.843	36.471
w/o Clustering & Analysis	37.235	38.373	40.451

Table 3: Simulator with different noise conditions

C Extracting Key Chemical Components in the Simulator

C.1 A Framework for Extracting Critical Chemical Components in the Simulator

To better illustrate the specific framework of *CSX-Sim* for extracting key chemical components, as shown in Figure 5. For scientific hypotheses addressing specific problems, we categorize key chemical components and conclusions within the hypothesis. We then analyze the role and mechanism of each key chemical component based on the chemical problem and the conclusions drawn from the hypothesis. Finally, we review and output the key chemical components, their corresponding mechanisms, and the conclusions from the hypothesis.

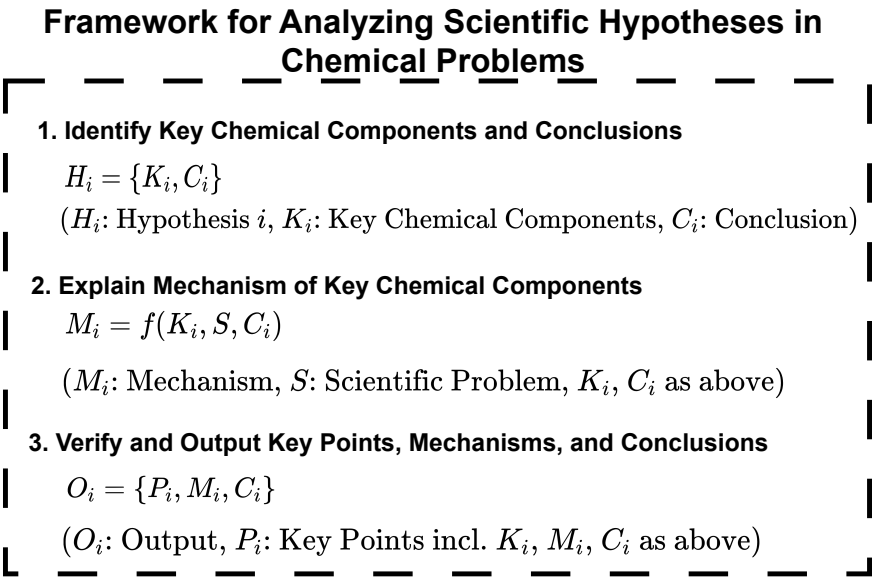


Figure 5: A Framework for Extracting Chemical Components in the Simulator.

C.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:

You are an experienced chemistry 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 (e.g., if the question involves improving MXene nanosheets and the hypothesis enhances them with liquid metal, MXene nanosheets are a prerequisite, not a key point; liquid metal is the key point). Prerequisites should not be output or analyzed as key points. Distinguish key points from validation methods (e.g., elemental analysis to verify properties). Validation methods support the hypothesis but are not chemical key points. 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 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 (e.g., performance improvement, efficiency changes).
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 (PVA) hydrogel acts as the base material, providing structural support and mechanical performance for thermoelectric gels. PVA with a mass fraction of 10% can provide mechanical support through hydrogen bonds in its structure and interact with potassium ferricyanide and potassium ferrocyanide to offer electrical changes.
2. Gdm_2SO_4 (Guanidine Sulfate)
Role and Function: Guanidine sulfate (Gdm_2SO_4) is integrated into the $K_3[Fe(CN)_6]$ / $K_4[Fe(CN)_6]$ to improve thermoelectric performance. The introduction of guanidine salt increases solvent entropy and effectively enhances 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. The presence of these ions enhances ion mobility and conductivity due to their ability to undergo reversible redox processes, thereby boosting the thermoelectric properties of the 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, outperforming traditional quasi-solid-state thermoelectric cells.
End Results

Here's a detailed example in chemistry: To better illustrate the effectiveness of extracting key chemical components, we compare the performance of our simulator against human chemistry experts by analyzing a real-world chemical 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 (DFT) process, which involves applying a temperature gradient during freezing to guide ice crystal growth for polymer chain alignment. Repeated cycles further enhance the alignment and crosslinking, creating anisotropic pores that reduce ion transport resistance. 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. This synergistic design achieves a remarkable 1500% increase in power density, reaching 0.51 mW m^{-2} at a ΔT of 5°C, with a thermopower of 0.7 mV K^{-1} and ionic conductivity improved by 20%. Mechanical tests reveal significant strength with a tensile strain at break of 350% and the system maintains stability under various mechanical deformations. This approach offers a cost-effective, adaptable solution for low-grade heat harvesting.

Here's Chemistry Expert Extracting Key Chemical Component:

- Chemical Key Points:
  1. Anisotropic Polymer Networks
     Role and Function: The layered polymer network structure enhances ion transport rates, thereby improving thermoelectric conversion efficiency.
  2. Hierarchical 3D Copper Electrodes
     Role and Function: The use of 3D copper electrodes effectively increases the reaction contact area, enhancing the thermoelectric potential.
  3. Polyvinyl Alcohol (PVA) as the Hydrogel Matrix
     Role and Function: PVA hydrogel is used as the matrix to provide good mechanical properties and structural support.
  4. Directional Freeze-Thawing (DFT) Process
     Role and Function: The DFT process forms ion channels along the freezing direction, thereby increasing ion transport rates.
  5. 0.7 M $CuSO_4$ Electrolyte
     Role and Function: This electrolyte is used for ionic crosslinking to enhance the mechanical properties of the thermoelectric material.
  6. 0.1 M H_2SO_4
     Role and Function: This acid is used together with a 0.7 M $CuSO_4$ electrolyte for ionic crosslinking to enhance the mechanical properties of the thermoelectric material.
- Results:
  - Enhanced Electrical Properties of the Thermoelectric Material
    - Increased power density
    - Improved thermopower
    - Enhanced ionic conductivity

- Improved Mechanical Properties of the Thermoelectric Material
  - Increased strength with a tensile strain
  - Enhanced stability under various mechanical deformations

Here's *CSX-Sim* Extracting Key Chemical Component:

- Chemical Key Points:
  1. Anisotropic Polymer Networks
     Role and Function: These networks are engineered to enhance ion transport efficiency while providing mechanical robustness through alignment of polymer chains, reducing ion transport resistance in thermocells.
  2. Hierarchical 3D Copper Electrodes
     Role and Function: The use of hierarchical 3D copper electrodes increases the surface area, enhancing redox kinetics of the Cu^{2+}/Cu^{0} couple, thus improving the thermoelectric performance without relying on expensive platinum electrodes.
  3. Copper Sulfate ($CuSO_4$) Electrolyte
     Role and Function: A 0.7 M $CuSO_4$ solution serves as an ionic crosslinker in the hydrogel, enhancing mechanical strength while facilitating efficient ion transport in the system.
  4. Sulfuric Acid (H_2SO_4)
     Role and Function: The addition of 0.1 M H_2SO_4 in the hydrogel contributes to ionic crosslinking and enhancing the overall conductivity of the electrochemical system.
  5. Directional Freeze-Thawing (DFT) Process
     Role and Function: This method is used to create an ordered structure in the polymer hydrogel, promoting alignment of polymer chains and ensuring anisotropic pores that further reduce ion transport resistance.
- Results:
  - Power Density Increase
  - Enhanced Thermopower
  - Improved Ionic Conductivity
  - Mechanical Strength under Deformation

Here's a comparison of the analysis results between our simulator and human experts:

By comparing the approaches of a chemistry expert and *CSX-Sim* in extracting key chemical components for the specific chemical issues of ion transport efficiency and electrode performance, *CSX-Sim* successfully identifies solutions in its scientific hypotheses, including anisotropic polymer networks and hierarchical 3D copper electrodes. Compared to the human chemistry expert, *CSX-Sim* captures five out of six key points, missing only one: "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 *CSX-Sim* in extracting key chemical components.

D The Role of CriticalPoints in *CSX-Sim*

To better illustrate the role of labeling critical components $\mathcal{C}$ in *CSX-Sim*, as defined in Equation 3, 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 3, related to CriticalPoints, 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 (Se) 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 (Se) performance with the $[Fe(CN)_6]^{3-}/[Fe(CN)_6]^{4-}$

redox pair. KCl, as an electrolyte, significantly enhances the gel's ionic conductivity, while the $[Fe(CN)_6]^{3-}$/$[Fe(CN)_6]^{4-}$ redox pair boosts the Seebeck coefficient through temperature-gradient-driven ion diffusion. Gelatin provides biocompatibility and mechanical strength, making it suitable for efficient thermoelectric energy conversion.

2. **Hypothesis 2**: By combining a PVA matrix with HCl, prepare a gel with high ionic conductivity and investigate its Seebeck coefficient (Se) performance under the influence of the Fe^{3+}/Fe^{2+} redox pair. HCl, as a strong electrolyte, significantly enhances the gel's ionic conductivity, while the Fe^{3+}/Fe^{2+} redox pair boosts the Seebeck coefficient through temperature-difference-driven ion diffusion. PVA provides flexibility and transparency, and by optimizing the HCl concentration and PVA crosslinking degree, ion migration efficiency can be further improved, enhancing the Seebeck coefficient and making it suitable for efficient energy conversion in body-heat thermoelectric devices.
3. **Hypothesis 3**: By preparing a pure PVA gel, investigate its Seebeck coefficient (Se) performance under the influence of the Fe^{3+}/Fe^{2+} redox pair. PVA, as a hydrophilic polymer, possesses a certain level of ionic conductivity, and the Fe^{3+}/Fe^{2+} redox pair generates a Seebeck coefficient through temperature-difference-driven ion diffusion.
4. **Hypothesis 4**: By polymerizing acrylamide (PAM) to prepare a hydrogel and investigate its thermoelectric performance. The porous network structure of the polyacrylamide hydrogel enhances the gel's ionic conductivity.

The results of the actual wet-lab experiments are summarized in Table 4. For the first three hypotheses, the CriticalPoints are included in the assumptions, resulting in a Correction Factor of 1 and non-zero final results. In contrast, for the fourth hypothesis, the CriticalPoints are not included in the assumption, leading to a Correction Factor of 0 and a final result of 0.

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 4: The Seebeck coefficient (Se) of polymer thermoelectric materials

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 (Potassium Chloride), and the [Fe(CN)$_6$]$^{3-}$/[Fe(CN)$_6$]$^{4-}$ Redox Pair, each playing significant roles in enhancing the Seebeck coefficient for thermoelectric efficiency. The Generated Hypothesis uses Polyacrylamide (PAM) 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 presence of PAM and evolving ionic conductivity offers a partial basis for plausible thermoelectric behavior under hypothetical conditions, but the complete omission of the redox pair is significant. The lack of adherence to all core key points and reliance on acceptable substitutions leads to a conclusion that while observable phenomena might still occur, they are limited. Given these evaluations, I assign $R = 0$ as there is a substantial gap in essential components for the reaction proposed in the Ground Truth Hypothesis.

Final Correction Factor 0 End

The our simulator analyzes Hypothesis 4, 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 (PAM) and lacks a redox pair, rendering it unable to effectively convert thermal energy into electrical energy. Consequently, the thermoelectric potential (Seebeck coefficient, Se) is zero.

E Evaluating the Simulator with Real Experiment Results

In this section, we present the validation of our simulator's accuracy using a dataset of 124 chemical 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 chemical behaviors.

E.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 30 cutting-edge chemical questions, each designed to probe significant aspects of chemical research. These questions were carefully selected to encompass multiple areas within the chemistry domain, ensuring a diverse and representative evaluation framework. Each question was associated with 3 to 6 hypotheses, resulting in a total of 124 authentic wet lab chemical experiment results. This extensive dataset forms a robust foundation for assessing the simulator's predictive accuracy and reliability.

The 124 experiment results were sourced from key subfields of chemistry to provide broad coverage of the discipline. The distribution of these results across subfields is presented in Table 5. Specifically, Polymer Chemistry contributed 16 results, Organic Chemistry provided 36, Inorganic Chemistry accounted for 33, and Analytical Chemistry comprised 39, totaling 124 results. This distribution across multiple subfields ensures that the test set reflects the diversity and complexity of real-world chemical experiments, enhancing the robustness of our evaluation.

A statistical analysis of the 124 authentic wet lab results was conducted to rigorously evaluate the simulator's performance. By including a substantial number of experiments from various subfields, we ensured that the dataset captures a wide range of challenges encountered in chemical research. This approach minimizes potential biases from over-representing any single subfield, thereby strengthening the reliability of our evaluation. The dataset's diversity and scale provide a solid basis for assessing the simulator's ability to predict experimental outcomes accurately, offering valuable insights for future research and applications.

Category	Count
Polymer Chemistry	16
Organic Chemistry	36
Inorganic Chemistry	33
Analytical Chemistry	39
Total	124

Table 5: Distribution of categories.

The 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. The diverse dataset and representation of multiple subfields collectively contribute to a comprehensive and effective evaluation, paving the way for advancements in chemical simulation and experimentation.

E.2 Trend Comparison with Real Experiment Results

To further assess the capabilities of our *CSX-Sim*, we utilized it to simulate 124 wet lab experiments. These experiments corresponded to 30 cutting-edge chemical science questions, and their simulated

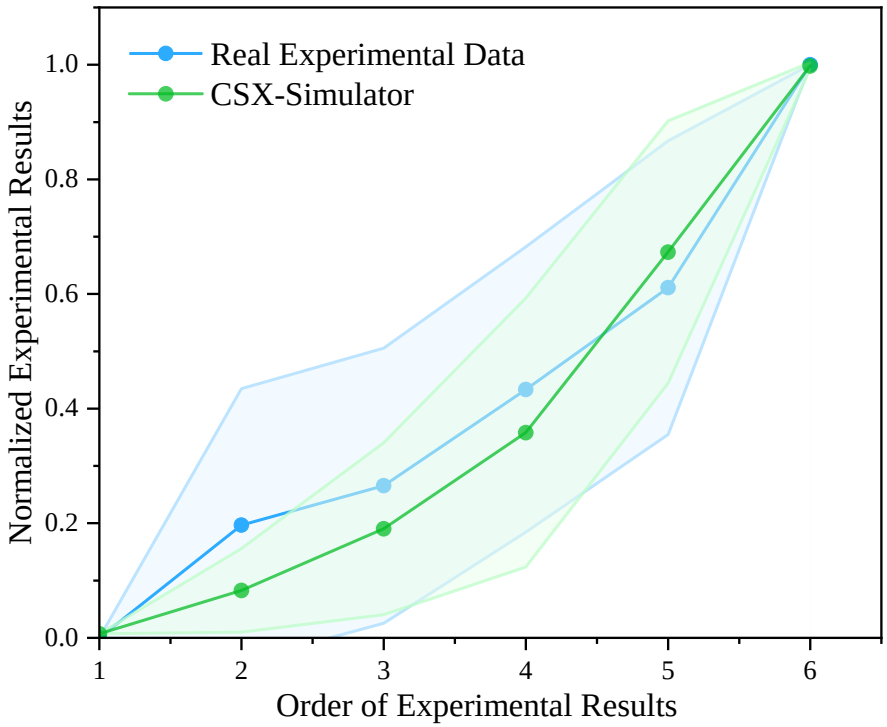


Figure 6: Comparison of simulated real experimental results with CSX-Simulator.

outcomes were subsequently aggregated for a comprehensive analysis. For each of the 124 experiments, the simulated result was derived from the average of three trials conducted by *CSX-Sim*. These results, each corresponding to one of the curated chemical questions, were systematically arranged in ascending order along the "Order of Experimental Results" axis, as depicted in Figure 6. This organization enabled a unified comparison between the simulated and actual experimental outcomes, with the vertical axis representing normalized experimental results to standardize the evaluation across the dataset.

Figure 6 compares the trends observed in *CSX-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 30 sets of results, demonstrating a strong consistency with the mean of the actual experimental outcomes. This alignment of mean trends across the diverse questions underscores the simulator's ability to model chemical processes accurately, capturing the overall behavior of the experimental data, regardless of the specific subfield.

The use of normalized results ensures that differences in scale do not affect the comparison, allowing a fair assessment of the simulator's trend-matching capability. The close correspondence between the simulated and real mean data, as visualized in the figure, highlights the *CSX-Sim* broad applicability across the chemistry domain. By successfully replicating the mean trends of the 124 results, the simulator proves to be a versatile tool, offering reliable predictions that can support a wide range of chemical research and applications.

F Evaluation of Trend Alignment and Accuracy

F.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 it suitable for capturing trend consistency across diverse chemical problems.

The Spearman Correlation Coefficient is calculated as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{7}$$

Where: d_i: The difference between the ranks of the i-th simulated and experimental result. n: The number of hypotheses in a given group (ranging from 3 to 6 per scientific question). ρ: The correlation coefficient, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no monotonic relationship. A Spearman Correlation Coefficient (ρ) near 1 indicates strong trend alignment, meaning the simulated results closely mirror the relative ordering of experimental outcomes. Our CSX simulator achieved a mean Spearman Correlation Coefficient of $\rho = 0.960$, significantly outperforming the baseline, as shown in Table 1, and demonstrating superior trend alignment.

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 (out of the 30 scientific questions) where the simulated results achieved perfect trend alignment with the experimental results ($\rho = 1$). Perfect trend alignment requires an exact match in the ranking of simulated and experimental outcomes, making PCI a robust measure of the simulator's ability to consistently replicate experimental trends across all problems. Notably, our CSX simulator achieved perfect trend alignment ($\rho = 1$) in 26 out of 30 question groups, significantly surpassing the baseline methods and highlighting its exceptional robustness and predictive fidelity.

F.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. The RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{8}$$

Where: y_i: The experimental result for the i-th hypothesis. $\hat{y}_i$: The simulated result for the i-th hypothesis. The CSX simulator exhibited a lower RMSE than the "Matched Score" baseline (Yang et al., 2024b), signifying improved predictive accuracy, as substantiated by the results in Table 1.

To thoroughly evaluate the predictive accuracy of our CSX simulator compared to real-world experimental outcomes, we tested its performance on a dataset of 124 authentic scientific hypotheses. For a comprehensive comparison, we calculated several performance indicators, as presented in Table 6. Building on the previously discussed metrics, we introduced three additional measures: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Logarithmic Error (RMSLE). These metrics, defined below, enhance the robustness of our analysis by capturing different aspects of prediction error.

Simulator	MSE (↓)	MAE (↓)	RMSLE (↓)
Matched Score	0.068	0.179	0.166
CSX-Sim	**0.058**	**0.161**	**0.147**
w/o CriticalPoints	0.064	0.174	0.159
w/o ComponentExtraction	0.087	0.215	0.192

Table 6: Validating the simulator with collected chemistry experiment results from literature.

Below, we define each metric used in the evaluation, along with their respective formulas, to ensure scientific rigor:

Mean Squared Error (MSE): MSE measures the average squared difference between predicted values $\hat{y}_i$ and actual values y_i across n samples. It is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{9}$$

A lower MSE indicates higher predictive accuracy, with larger errors penalized more heavily due to squaring.

Mean Absolute Error (MAE): MAE quantifies the average absolute difference between predicted and actual values, calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{10}$$

This metric is less sensitive to outliers than MSE, providing a more balanced measure of error.

Root Mean Squared Logarithmic Error (RMSLE): RMSLE focuses on relative errors by evaluating the logarithmic difference between predicted and actual values:

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log(\hat{y}_i + 1) - \log(y_i + 1)\right)^2} \tag{11}$$

This metric is particularly useful for datasets with exponential trends or varying error scales.

As shown in Table 6, *CSX-Sim* consistently outperforms the "Matched Score" baseline (Yang et al., 2024b) across all metrics, achieving an MSE of 0.058, an MAE of 0.161, and an RMSLE of 0.147. Ablation studies further reveal the contributions of individual components: the removal of CriticalPoints results in a slight performance decline (MSE of 0.064, MAE of 0.174, RMSLE of 0.159), while the exclusion of ComponentExtraction leads to more significant degradation (MSE of 0.087, MAE of 0.215, RMSLE of 0.192). These results underscore the importance of both critical point identification and component extraction in achieving high predictive accuracy and robustness in simulation outcomes.

G Different Levels of Distortion

We collaborated with chemistry PhD students to identify and design three common types of distortions encountered in chemical 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, Moderate Noise, and Complex Noise—and incorporated them into the hypothesis embedding function $\phi(\cdot)$ to simulate increasingly challenging feedback conditions.

In chemical 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, as this oversight may enhance thermoelectric performance unexpectedly. Similarly, misjudging irrelevant factors, such as additives in organic reactions with no actual impact, can create a plateau effect. Conversely, misjudging critical factors, like the temperature's role in enzyme activity during enzyme studies, can produce a cliff if the temperature is incorrectly assumed to inhibit the reaction entirely. These elements—local maxima/minima, plateaus, and cliffs—present significant challenges in optimization problems within chemical research.

Through extensive discussions with chemistry experts, we conducted a statistical analysis to evaluate the discrepancies between wet lab results and empirical expected outcomes across diverse experimental scenarios. This process enabled us to statistically analyze the frequency of the three types of distortions—local maxima/minima, plateaus, and cliffs—across various chemical scenarios. We then quantified the occurrence of these distortions in different scenarios and sorted them by frequency, from low to high. Based on this distribution, we categorized the discrepancies: the top 35% of observed gaps were classified as Simple Noise, the middle 40% as Moderate Noise, and the bottom 25% as Complex Noise. Furthermore, we integrated the three distortion levels—Simple Noise, Moderate

Noise, and Complex Noise—into the hypothesis embedding function $\phi(\cdot)$ to simulate increasingly challenging feedback conditions. This structured stratification provided a clear framework to evaluate the varying impacts of different scenarios on our simulator, facilitating a deeper understanding of the simulator's performance under diverse conditions.

Noise Conditions	Local Maxima/Minima	Plateaus	Cliffs
Simple	0-10	0-2	0-2
Medium	0-30	0-6	0-6
Complex	≥ 30	≥ 3	≥ 3

Table 7: The composition of different types of noise.

These distortions, along with their detailed quantities, are outlined in the accompanying Table 7, which illustrates the composition of different types of noise across various conditions. For instance, simple noise conditions are associated with 0-10 local maxima/minima, 0-2 plateaus, and 0-2 cliffs. Medium noise conditions escalate these figures to 0-30 local maxima/minima, 0-6 plateaus, and 0-6 cliffs. In complex noise scenarios, the challenges intensify, with ≥ 30 local maxima/minima, ≥ 3 plateaus, and ≥ 3 cliffs, reflecting the increased difficulty in achieving optimal solutions. We constructed three distinct noise levels to evaluate the robustness of our *CSX-Rank* under complex chemical feedback conditions.

By comparing Table 3, 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. This is primarily due to the growing discrepancy between highly complex noise and real experimental feedback, where simulation feedback contains substantial erroneous information, thereby degrading the performance of screening the ground truth scientific hypothesis from the generated scientific hypotheses.

H Evaluation of Experiment-Guided Ranking and Its Societal Benefits

The intricate knowledge system of chemistry, 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 two distinct approaches: the experiment-guided ranking method, which leverages simulation feedback or real experimental results to refine hypothesis selection, and the pre-experiment method, which relies solely on the model's prior knowledge for screening the ground truth hypothesis. Our findings reveal that the experiment-guided ranking method demonstrates a marked improvement over its counterpart. By integrating simulation feedback, this method allows for a reflective process that considers previous simulation (and experimental) results. This iterative reflection provides more contextually relevant information, enabling the selection of the next hypothesis with greater precision. Consequently, this approach effectively mitigates the accumulation of biases, thereby enhancing the efficiency and accuracy of experimental screening processes.

The ranking of hypotheses emerges as a pivotal element in automated scientific discovery, particularly in natural sciences, where wet-lab experiments are costly and are constrained by low throughput. Traditional approaches, such as pre-experiment ranking, depend exclusively on the internal reasoning of large language models, lacking integration with empirical experimental outcomes. In contrast, we introduce the novel task of experiment-guided ranking, designed to prioritize candidate hypotheses by leveraging insights from previously tested results. However, the development of such strategies is hindered by the impracticality of repeatedly conducting real experiments in natural science domains due to time, cost, and resource limitations. To overcome this obstacle, we propose a simulator grounded in three domain-informed assumptions, modeling hypothesis performance as a function of its similarity to a known ground truth hypothesis, with performance perturbed by noise to reflect real-world variability. To validate this simulator, we curated a dataset comprising 124 chemistry hypotheses, each accompanied by experimentally reported outcomes, providing a robust foundation for evaluation.

Building on this simulator, we developed a pseudo experiment-guided ranking method that clusters hypotheses based on shared functional characteristics and prioritizes candidates using insights derived from simulated experimental feedback. Our experimental results demonstrate that this method outperforms both pre-experiment baselines and strong ablations, highlighting its potential to revolutionize hypothesis selection in chemical research. 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, the enhanced efficiency in hypothesis testing could foster innovation in industrial applications, such as cleaner energy solutions, contributing to global efforts to address climate change and promote sustainable development.

I 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 three 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.

Importantly, the simulator's absolute accuracy is not critical for this line of research. As long as the experiment-guided ranking methods are robustly developed and tested within this simulated environment, they can subsequently leverage real experimental feedback to identify optimal hypotheses when deployed in real-world scenarios.