MOOSE-Chem: Large Language Models for Rediscovering Unseen Chemistry Scientific Hypotheses

Zonglin Yang [1] Wanhao Liu [2] Ben Gao [2] Tong Xie [3] Yuqiang Li [2] Wanli Ouyang [2] Soujanya Poria [4] Erik Cambria [1] Dongzhan Zhou [2]

[1]Nanyang Technological University [2]Shanghai AI Lab [3]University of New South Wales [4]Singapore University of Technology and Design

Center Question

- Can LLMs automatically discover novel and valid chemistry research hypotheses for any chemistry research question?

Yes, It Is Possible.

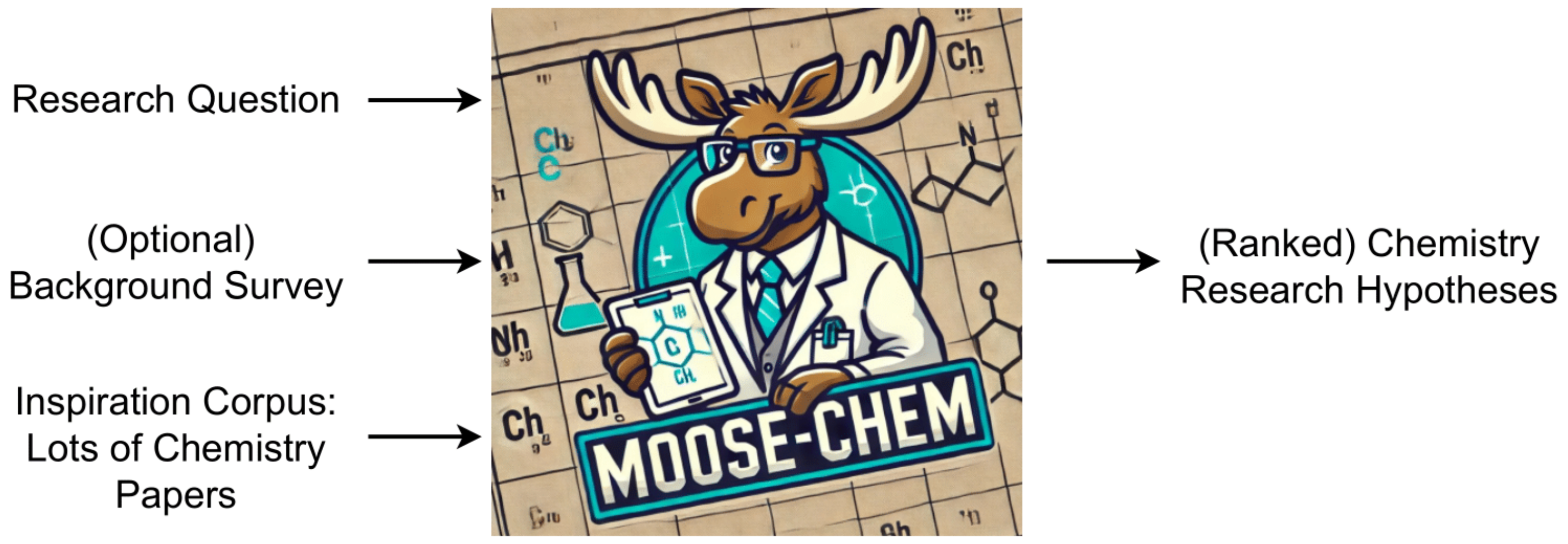
Research Question
(Optional) Background Survey
Inspiration Corpus: Lots of Chemistry Papers
MOOSE-CHEM
(Ranked) Chemistry Research Hypotheses

Figure 1. Input / Output of the MOOSE-Chem framework.

Fundamental Assumption

In chemistry, a majority of published hypotheses originate from one research background and several inspirations.

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

For example,

- Background Survey: The best performing methods are electrocatalytic methods.
- Inspiration 1: Ruthenium as catalyst
- Inspiration 2: Nitrogen-doped electrode
- Inspiration 3: D_2O as chemical solution
- Hypothesis: A nitrogen-doped ruthenium (Ru) electrode can effectively catalyze the reductive deuteration of (hetero)arenes in the presence of D_2O in an electrocatalytic method, leading to efficient D_2 gas production.

Denoting:

- research background: b
- inspiration: i
- hypothesis: h

The center challenge: $P(h \mid b)$ seems impossible to solve.

Markov Property in Scientific Discovery

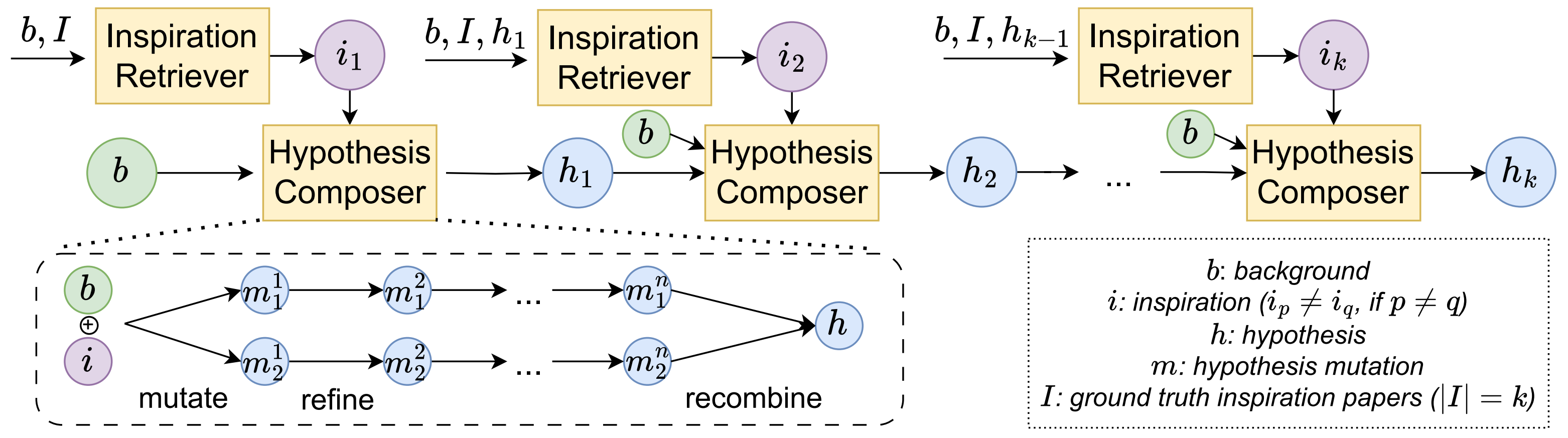

Figure 2. The scientific discovery process can be represented as a Markov process: h_j is a state, and i_j is an action.

Mathematical Derivation

$$\begin{aligned}
P(h \mid b) &= P(i_1, \dots, i_k, h_1, \dots, h_k \mid b) && (2)\\
&= P(i_1, h_1 \mid b) \cdot P(i_2, h_2 \mid b, i_1, h_1) \cdot \ldots \cdot P(i_k, h_k \mid b, i_1, \dots, i_{k-1}, h_1, \dots, h_{k-1}) && (3)\\
&\approx P(i_1, h_1 \mid b) \cdot P(i_2, h_2 \mid b, h_1) \cdot \ldots \cdot P(i_k, h_k \mid b, h_{k-1}) && (4)\\
&= \prod_{j=1}^{k} P(i_j \mid b, h_{j-1}, I) \cdot P(h_j \mid b, i_j, h_{j-1}), \text{ where } h_0 = \emptyset && (5)
\end{aligned}$$

$$P(H_{\text{ranked}}) = P(H, R), \text{ where } H_{\text{ranked}} = \{h_1, h_2, \dots, h_n \mid R(h_i) \geq R(h_{i+1}) \text{ for all } i\} \tag{6}$$

To solve $P(h \mid b)$, a sufficient set of sub-tasks are

1. $P(i_j \mid b, h_{j-1}, I)$ Inspiration Retrieval
2. $P(h_j \mid b, i_j, h_{j-1})$ Hypothesis Composition
3. $R(h_i)$ Hypothesis Ranking

Main Framework

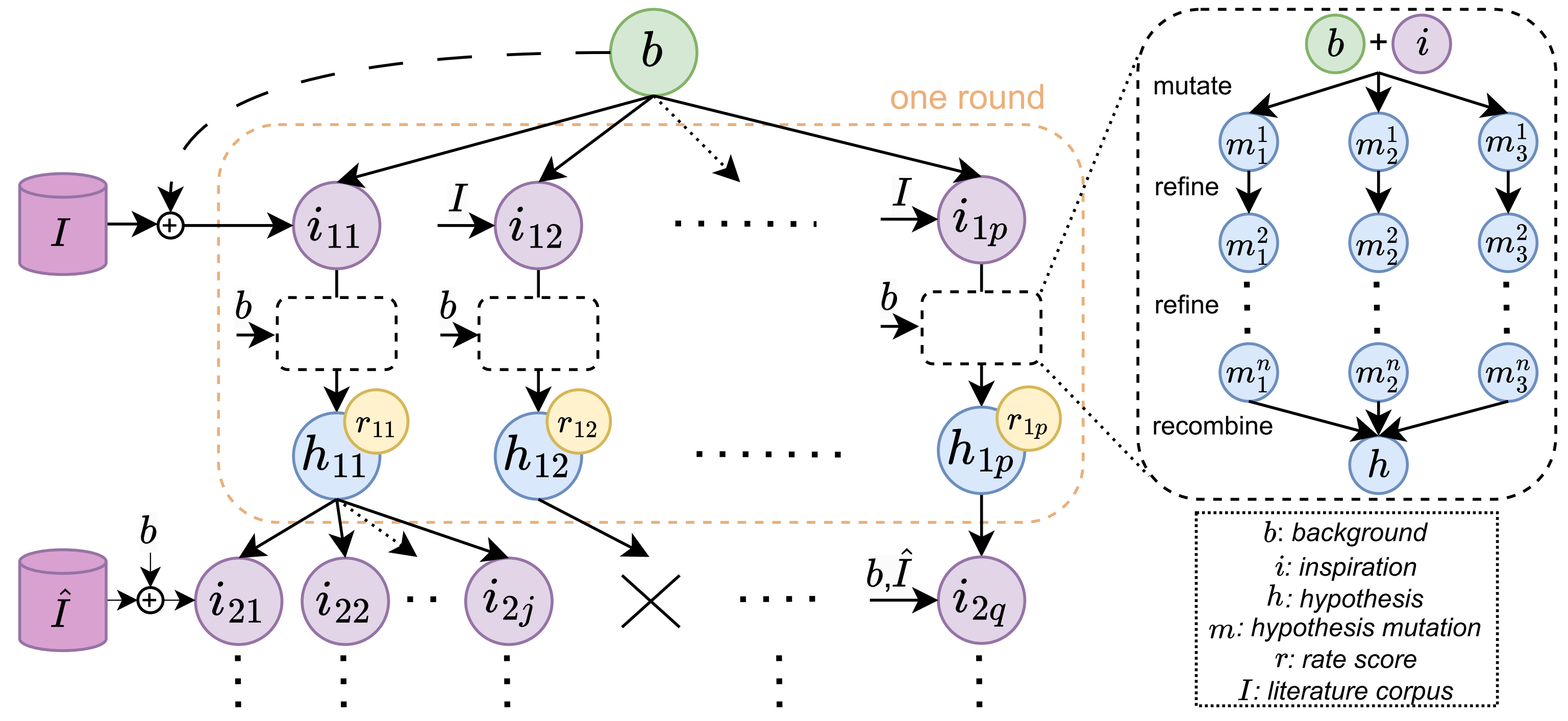

Figure 3. The MOOSE-Chem framework. It receives b and I as input, and outputs a list of ranked h. The bottom-right legend describes the symbols in the figure.

Analysis: Inspiration Retrieval

For each hypothesis, normally it needs 2 or 3 (groundtruth) inspirations. We collect the most cited chemistry papers published in *Nature*, mix the groundtruth inspirations and the collected papers (inspirations) together to create a corpus $|I| = 300$.

For each round of inspiration retrieval, 3 (output) out of 15 (incorporated in prompt) is selected, so the remaining ratio of inspiration candidates after one round is 20%, two rounds is 4%, and three rounds is 0.8%.

$$\text{Hit Ratio} = \frac{\text{number of groundtruth inspirations remained in that round}}{\text{number of all groundtruth inspirations}} \tag{7}$$

Model	Hit Ratio (top 20%)	Hit Ratio (top 4%)	Hit Ratio (top 0.8%)
`Llama-3.1-8B`	71.6%	43.5%	26.8%
`Llama-3.1-70B`	95.1%	83.0%	59.5%
`Llama-3.1-405B`	95.7%	78.7%	52.7%
`GPT-4o`	96.7%	83.7%	60.8%

Table 1. Comparison of `Llama` series and `GPT-4o` on inspiration retrieval. The corpus size is 300. For each screen window of 15 papers, 3 papers are selected.

Analysis: Hypothesis Composition

Evaluation metric: Matched Score. It measures the number of main innovations covered (with very similar content) in the groundtruth hypothesis.

	5	4	3	2	1	0	Total
	w/ background survey						
Average MS (`GPT-4o`)	2	9	18	17	5	0	51
Top MS (`GPT-4o`)	28	1	19	3	0	0	51
Top MS (Experts)	9	12	22	6	2	0	51
	w/o background survey						
Average MS (`GPT-4o`)	1	7	17	19	7	0	51
Top MS (`GPT-4o`)	25	2	19	5	0	0	51

Table 2. Average/Top MS means the average/highest Matched Score of all generated h from one b.

Analysis: Hypothesis Ranking

#Matched i	3	2	1	0
Average Rank Ratio	NA	0.411	0.474	0.521
Size	0	302	2458	4899

Table 3. Relation between the number of matched ground truth i and the average ranking ratio (↓).

Overall Performance

Method	Top MS	Average MS
SciMON	2.549	2.281
MOOSE	2.882	2.464
Qi et. al.	2.686	2.356
MOOSE-Chem	**4.020**	2.564
w/o multi-step	3.765	2.730
w/o multi-step & EU	2.863	2.578

Table 4. Experiments and ablation study. The Matched Score (MS) is evaluated by `GPT-4o`.

Case Study

Ground truth h: *The main hypothesis is that a **nitrogen-doped ruthenium (Ru)** electrode can effectively catalyze the reductive deuteration of (hetero)arenes in the presence of $\mathbf{D_2O}$, leading to high deuterium incorporation into the resulting saturated cyclic compounds. The findings validate this hypothesis by demonstrating that this electrocatalytic method is highly efficient, scalable, and versatile, suitable for a wide range of substrates.*

Generated h: *A pioneering integrated electrocatalytic system leveraging **ruthenium** nanoparticles embedded in **nitrogen-doped** graphene, combined with a dual palladium-coated ion-exchange membrane reactor, will catalyze efficient, scalable, and site-selective reductive deuteration of aromatic hydrocarbons and heteroarenes. Utilizing deuterium sources from both D_2 gas and $\mathbf{D_2O}$, this system will optimize parameters through real-time machine learning-driven dynamic adjustments. Specific configurations include ruthenium nanoparticle sizes (2-4 nm), nitrogen doping levels (12-14%), precisely engineered palladium membranes (5 micrometers, ensuring 98% deuterium-selective permeability), and advanced cyclic voltammetry protocols (1-5 Hz, -0.5V to -1.5V).*

Expert's analysis: *The proposed hypothesis effectively covers two key points from the ground truth hypothesis: **the incorporation of ruthenium (Ru) and the use of $\mathbf{D_2O}$ as a deuterium source** within the electrocatalytic system. However, the current content does not detail the mechanism by which Ru-D is produced, which is essential for explaining the process of reductive deuteration. Nevertheless, the results are still insightful. The specific level of nitrogen doping, for example, is highly suggestive and warrants further investigation. Overall, the match remains strong in its alignment with the original hypothesis while also presenting opportunities for deeper exploration.*

https://github.com/ZonglinY/MOOSE-Chem ICLR 2025 zonglin.yang@ntu.edu.sg