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

Zonglin Yang

Motivation

- Lack details
  - *fine-grained hypothesis discovery*
- Not yet best exploit LLM’s capacity and potential

Fine-grained Hypothesis Discovery

- $h_c \rightarrow Method \rightarrow h_f$
- $h_f = \{h_c, d_1, \ldots, d_m\}$
- d represents:
  - Adding an element/detail
  - Deleting an existing element/detail
- $d \in D$
- $P(h_f | b, h_c)$
  - $= P(\{d_1, \ldots, d_m\} | b, h_c, D)$
- $|D| = n$
- Complexity: $C_n^m = \frac{n!}{m!(n-m)!}$

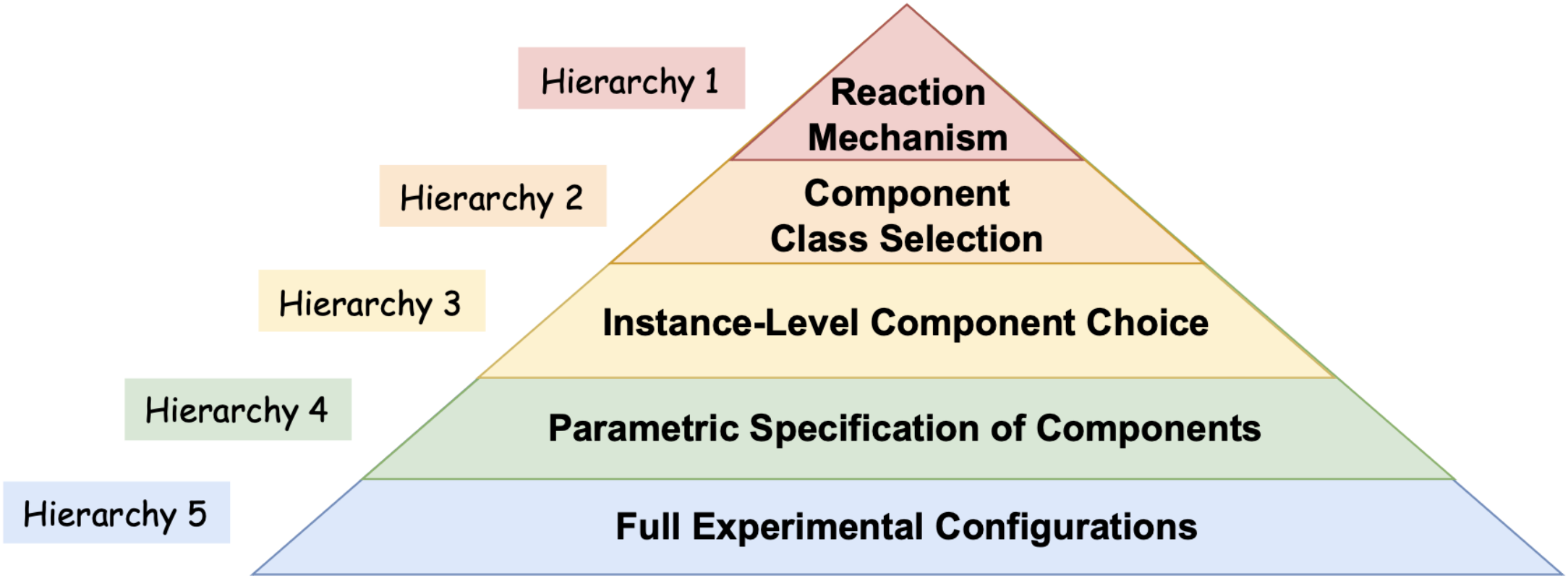
Challenges

- D is implicit
- $|D|$ can be very large
- Complexity (C_n^m) is intractable

Intrackable Complexity

- Classical algorithm for combinatorial complexity
  - Dynamic programming
- Problem structure
  - Optimal solution ← optimal solution of its subproblem
- Partition $\{d_1, \dots, d_m\}$ into p hierarchies
  - High-level concepts → low-level details
- Optimal solution of full $\{d_1, \dots, d_m\}$ ← optimal solution of $\{d\}$ in high-level concepts

Example Hierarchy for Chemistry



Intrackable Complexity

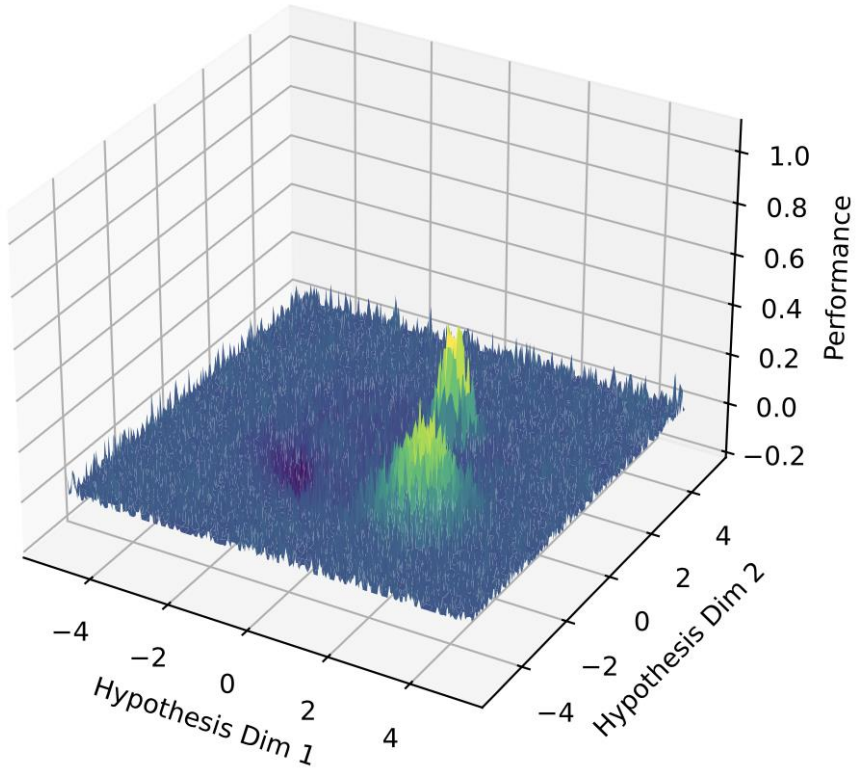
- Determining d in all p hierarchies ← iteration of determining d in each hierarchy sequentially
- Use approximate solution by heuristics rather than exhaustively search for the exact solution
  - LLM

Motivation

- Lack details
  - *fine-grained hypothesis discovery*
- Not yet best exploit LLM's capacity and potential

Exploit LLM's Upper Limit

- Definition of *upper limit*
  - *Of all the hypotheses an LLM can possibly generate, the one that the LLM itself consider the best*
- Consider a latent space
  - x-axis: hypothesis
  - y-axis: internal reward from LLM
- Upper limit:
  - The global maximum of the latent space
- In practice
  - Searching for a better local maximum

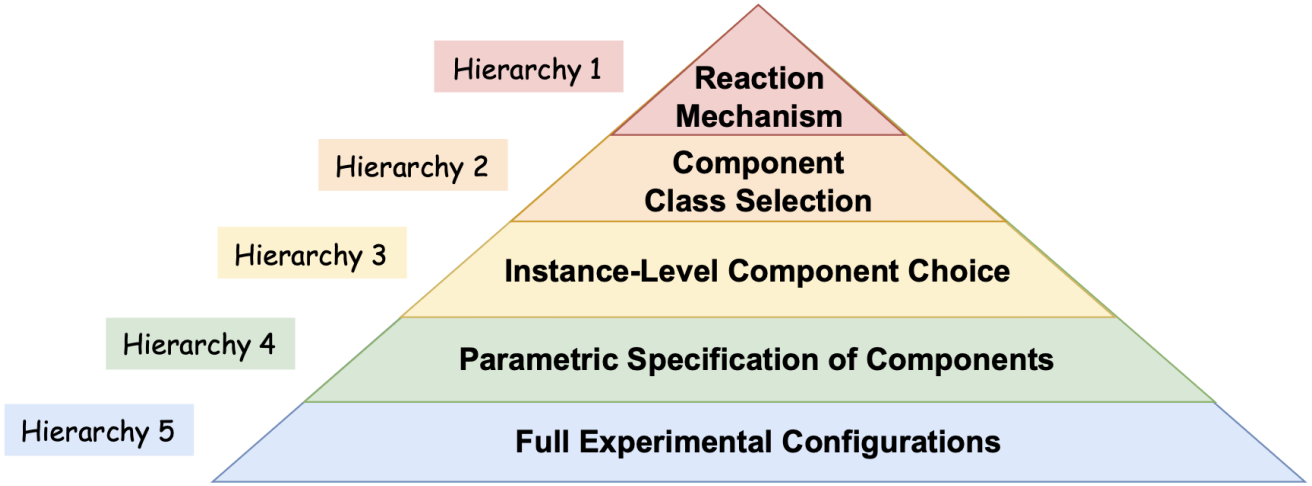


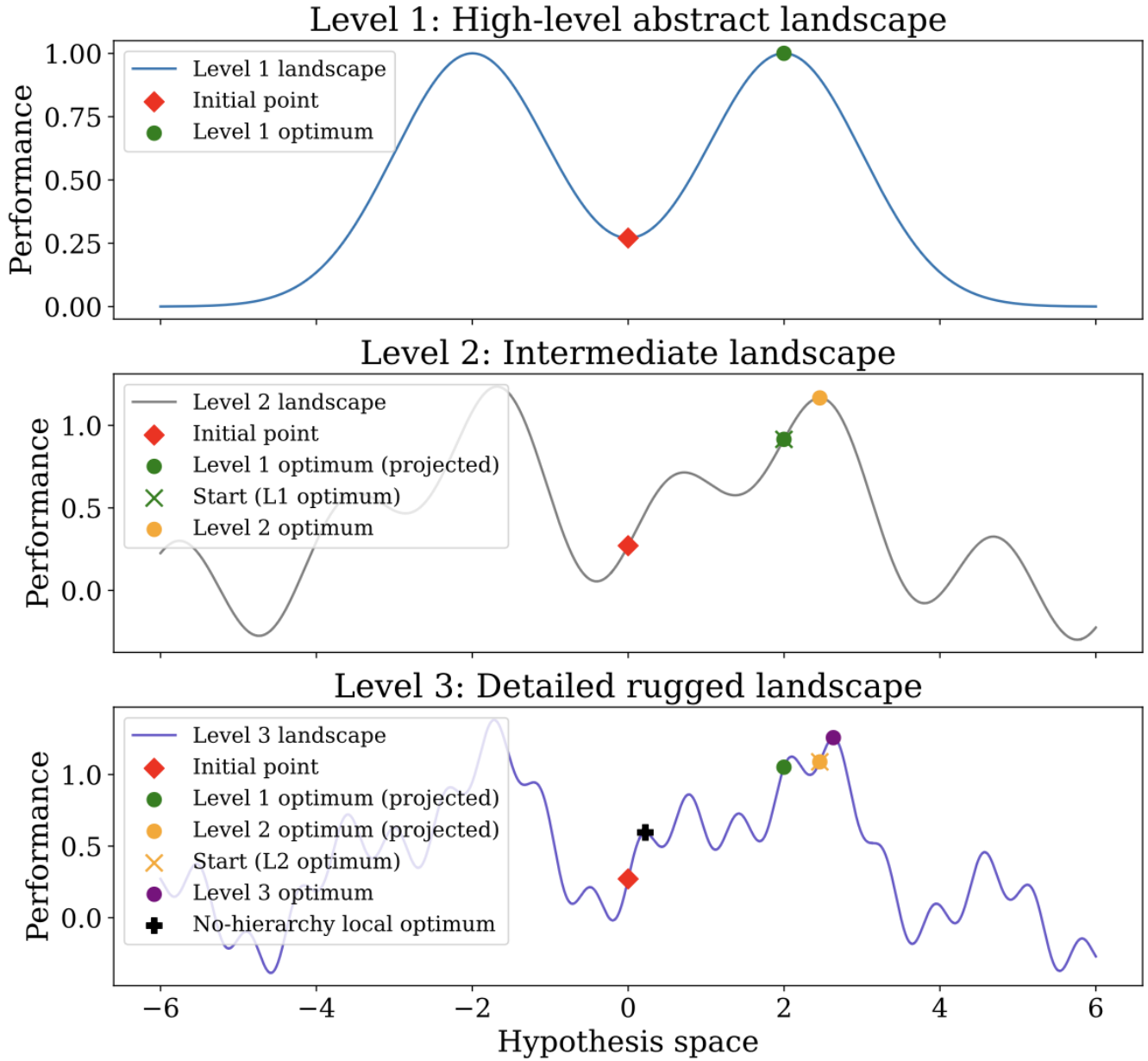
Exploit LLM’s Upper Limit

- How human scientist do
  - To think of a hypothesis that themselves would think the best for experiment

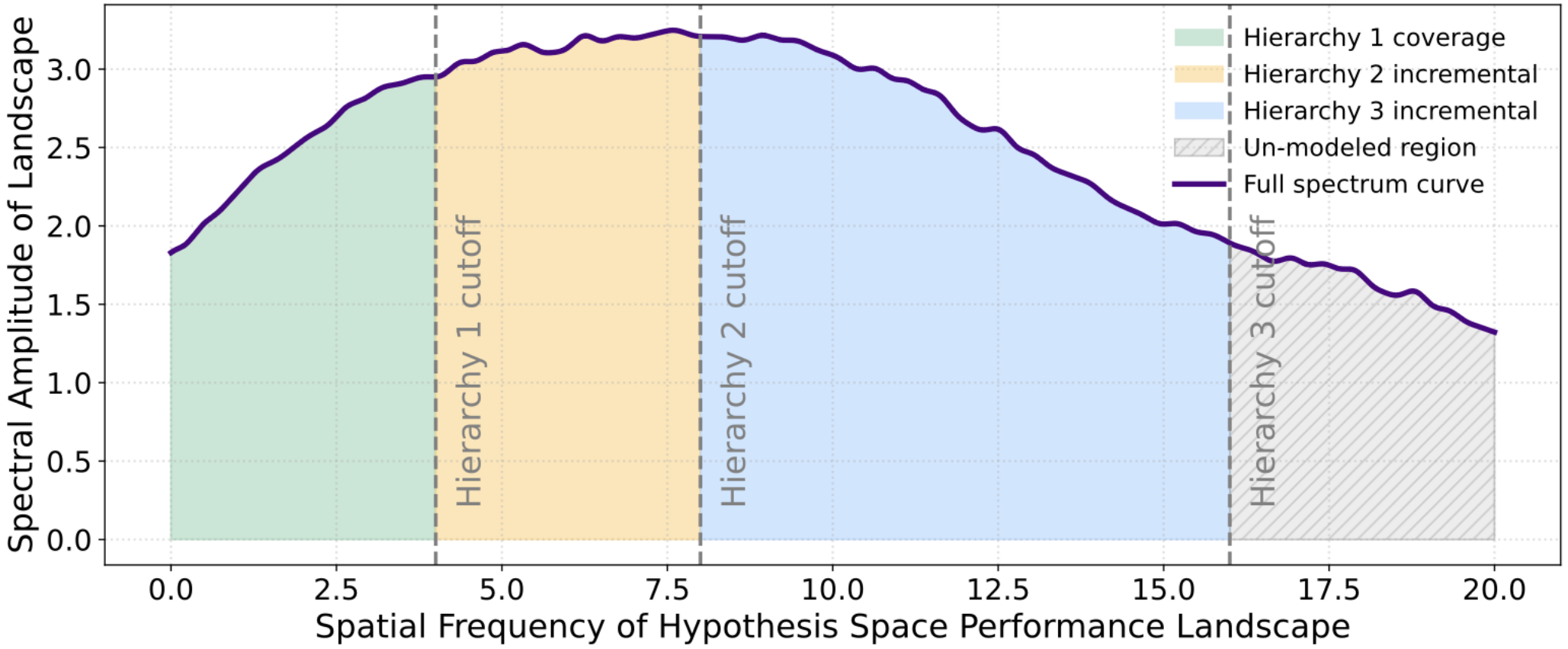
Challenge: Exploit LLM's Upper Limit

- Very rugged reward landscape
  - Difficult for optimization
- Method: the hierarchical design
  - Smoothen the reward landscape by averaging

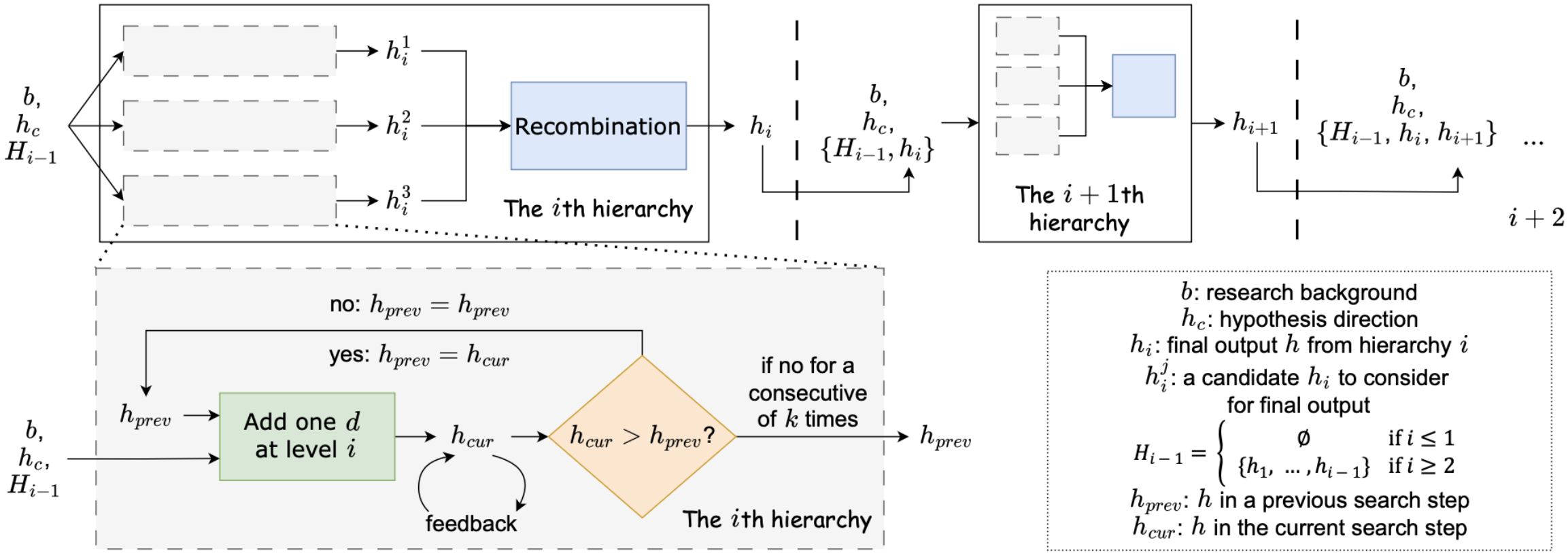



Level 1: High-level abstract landscape
Performance
1.00
0.75
0.50
0.25
0.00
Level 1 landscape
Initial point
Level 1 optimum
Level 2: Intermediate landscape
Performance
1.0
0.5
0.0
Level 2 landscape
Initial point
Level 1 optimum (projected)
Start (L1 optimum)
Level 2 optimum
Level 3: Detailed rugged landscape
Performance
1.0
0.5
0.0
Level 3 landscape
Initial point
Level 1 optimum (projected)
Level 2 optimum (projected)
Start (L2 optimum)
Level 3 optimum
No-hierarchy local optimum
−6
−4
−2
0
2
4
6
Hypothesis space

Low-Pass Filtering on Reward Landscape



Method



b: research background
h_c: hypothesis direction
h_i: final output h from hierarchy i
h_i^j: a candidate h_i to consider for final output

$$H_{i-1} = \begin{cases} \emptyset & \text{if } i \leq 1 \\ \{h_1, \ldots, h_{i-1}\} & \text{if } i \geq 2 \end{cases}$$

h_{prev}: h in a previous search step
h_{cur}: h in the current search step

Research Questions

- How to reach to better local optimum?
- Does it really better?
- Reward landscape ← Multiple diverse LLMs
- Reward landscape ← Multiple same LLMs

Benchmark

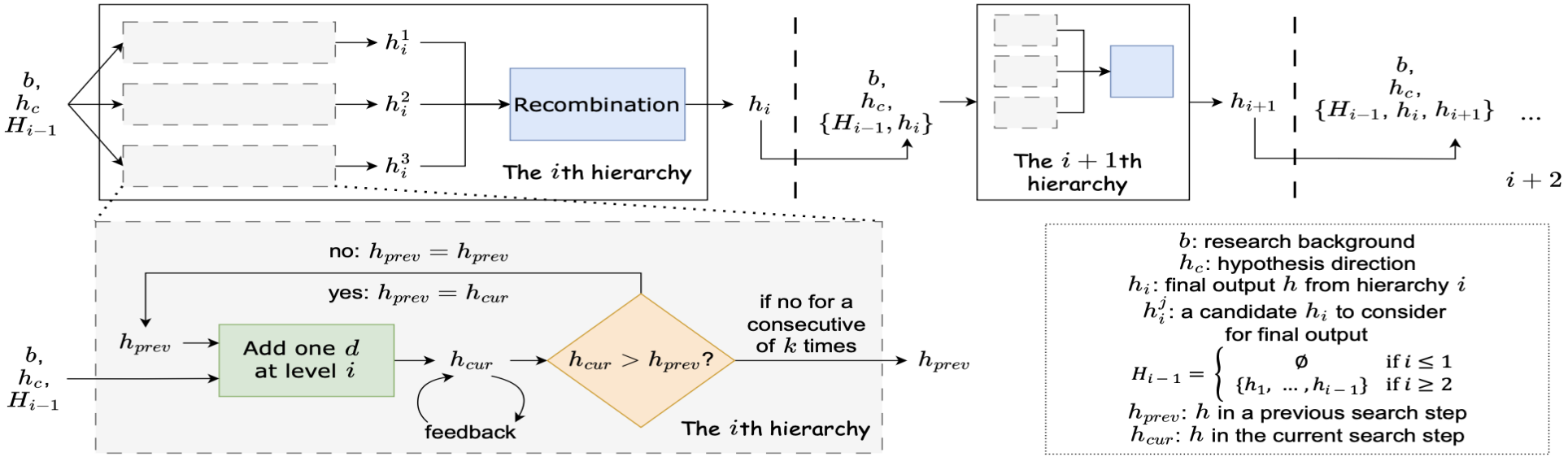
- Extend TOMATO-Chem
- 51 papers published in good venues
  - Research background
  - Fine-grained hypothesis

Evaluation

- LLM pairwise evaluation
  - Which one is better optimum?
- Reference-based evaluation
  - How many details in annotated fine-grained hypothesis are recalled?

Baselines

- Greedy Search
- Greedy Search + Self-consistency



Question 1: How to reach to better local optimum?

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

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

Question 2: Does it really better?

	Soft Recall	Hard Recall
Greedy Search	16.60%	9.90%
w/ Self-consistency	31.50%	17.70%
HHS	**40.40%**	**23.00%**

Table 2: Recall of ground-truth components by discovered hypotheses.

Question 3: Reward landscape ← Multiple Diverse LLMs

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

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

Question 4: Reward landscape ← Multiple Same LLMs

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

Table 4: Pairwise comparison between *HHS-1* and *HHS-3*.

Thank you!

- Q & A

My Twitter

Code