



香港中文大學
The Chinese University of Hong Kong



Journal Club

Article

<https://doi.org/10.1038/s41467-024-54457-x>

An automatic end-to-end chemical synthesis development platform powered by large language models

Zihan Li

11th Apr. 2025

Nat. Commun. **2024**, 15, 10160.

1. Background for Research

2. LLM-RDF

3. Conclusion and Methods

1. Background for Research

Importance

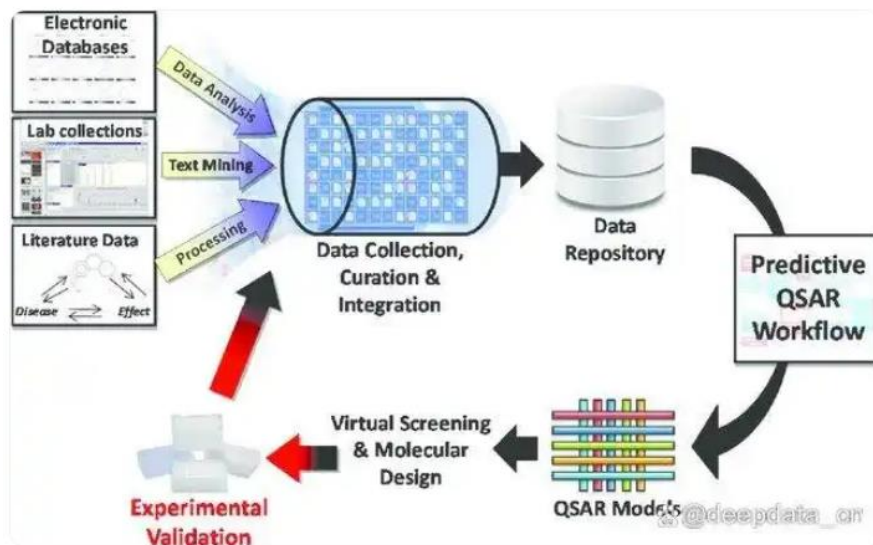
synthesis reaction design for drug discovery and chemical process

relies on expert chemists, requires time and cost



machine learning accelerate the development

QSAR



translation of multistep synthesis procedures from literature to experimental execution via natural language processing (NLP) models



Research Progress

OpenAI released the large language model (LLM) based ChatGPT

2022

Other LLM

Claude, Gemini, Llama3.1, Mistral, et al.

Applications

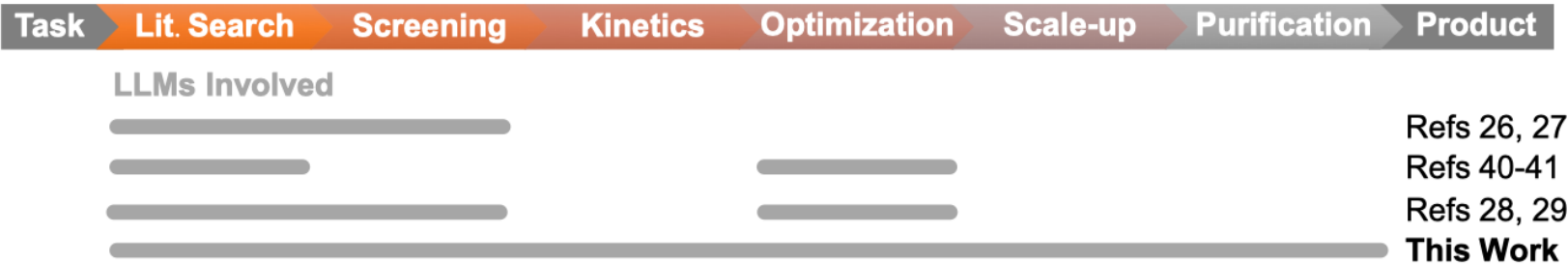
literature mining, molecule and material discovery, reaction condition recommendation, optimization and lab apparatus automation.



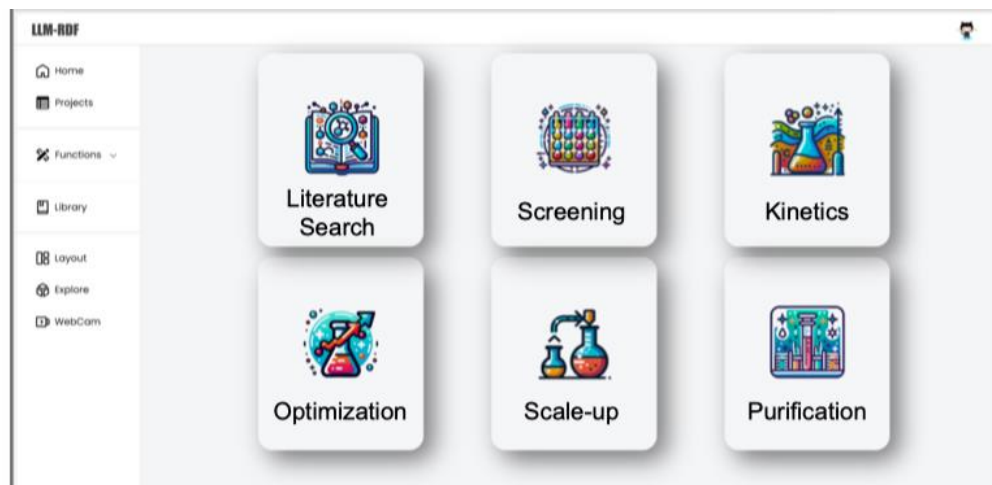
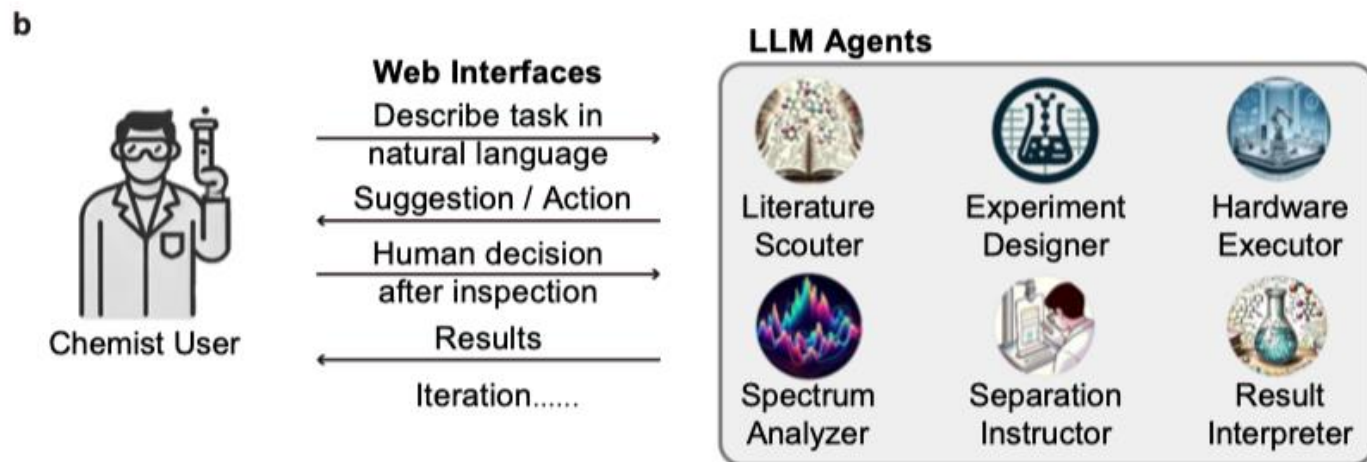
Deficiencies and Challenges

Fragmented coverage of the synthesis development stages

Fail to achieve fully autonomous end-to-end design



This Work



Cu/TEMPO catalyzed aerobic alcohol oxidation

LLM-based reaction development framework

- GPT-4 based
- Pre-training
- No coding
- External tools available
- Human-machine cooperation

2. LLM-RDF



Literature Search and Information Extraction!

literature-oriented

Literature Search

User: Search for methods that can use air oxidation of primary alcohols into aldehydes and provide references.

Answer: Here are some notable methods (literatures):

- Sustainable catalytic oxidation of alcohols over the interface between air and water⁴⁹.
- Highly practical copper(I)/TEMPO catalyst system for chemoselective aerobic oxidation of primary alcohols⁴⁸.
- Pd(II)-hydrotalcite-catalyzed oxidation of alcohols to aldehydes and ketones using atmospheric pressure of air⁵⁰.
- Photochemical oxidation of benzylic primary and secondary alcohols utilizing air as the oxidant⁵¹.

Results Analysis

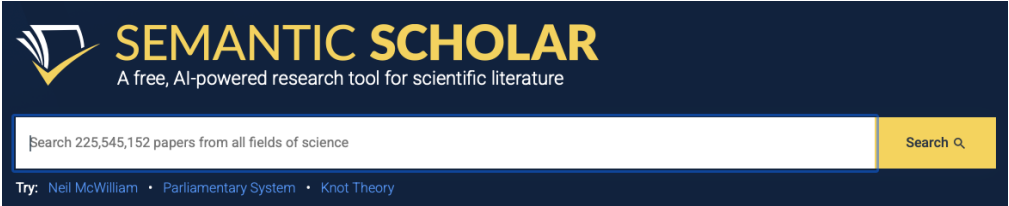
User: Which method has the best potential for practical applications?

Answer: Among the methods discussed, method 2 stands out for ease of catalyst preparation, use of ambient air as oxidant, compatibility with various alcohols, operational simplicity, and chemoselectivity.

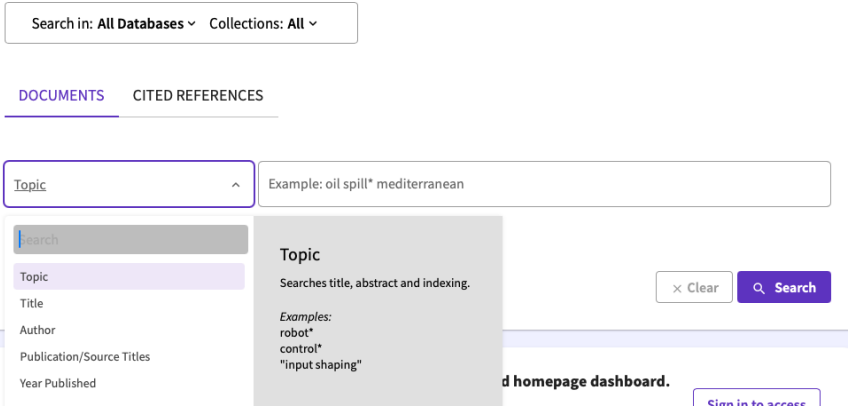
Information Extraction

User: Extract the procedure and design space from the uploaded literature document.

Answer: Procedure: To a solution of alcohol in CH₃CN in a tube was added (1) CuOTf , (2) bpy, (3) TEMPO, and (4) NMI. Design Space: Cu Salts: Cu(OTf), CuBr, Cu(OTf)₂, CuBr₂, Bases: NMI, DBU, Time:1, 6, 24 hours.



<https://www.semanticscholar.org>





Substrate Scope and Condition Screening

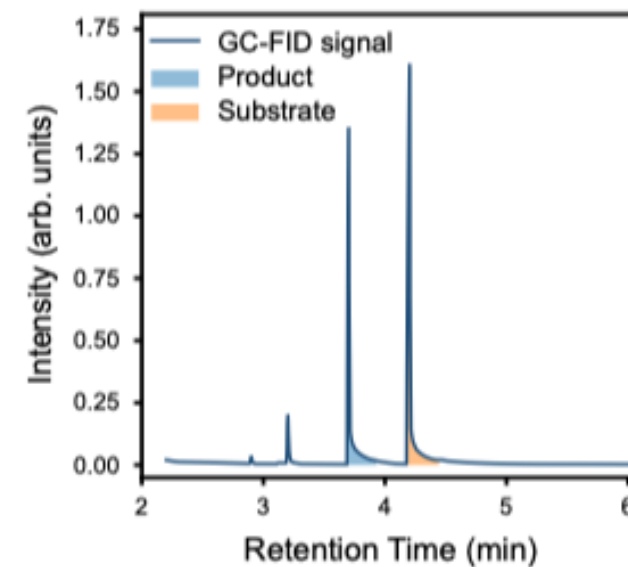
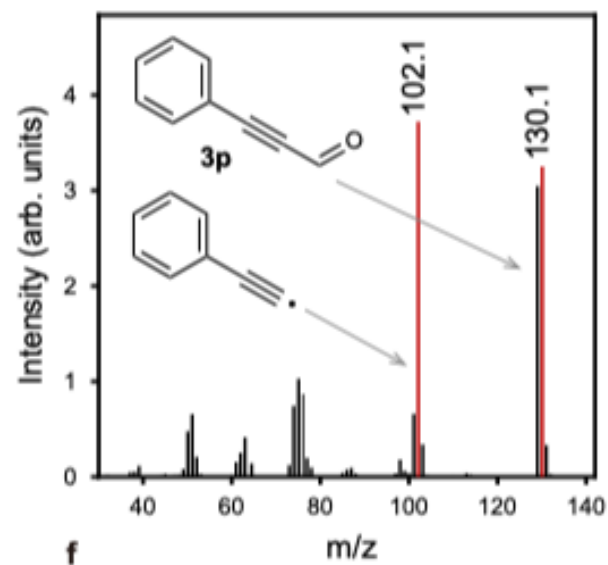
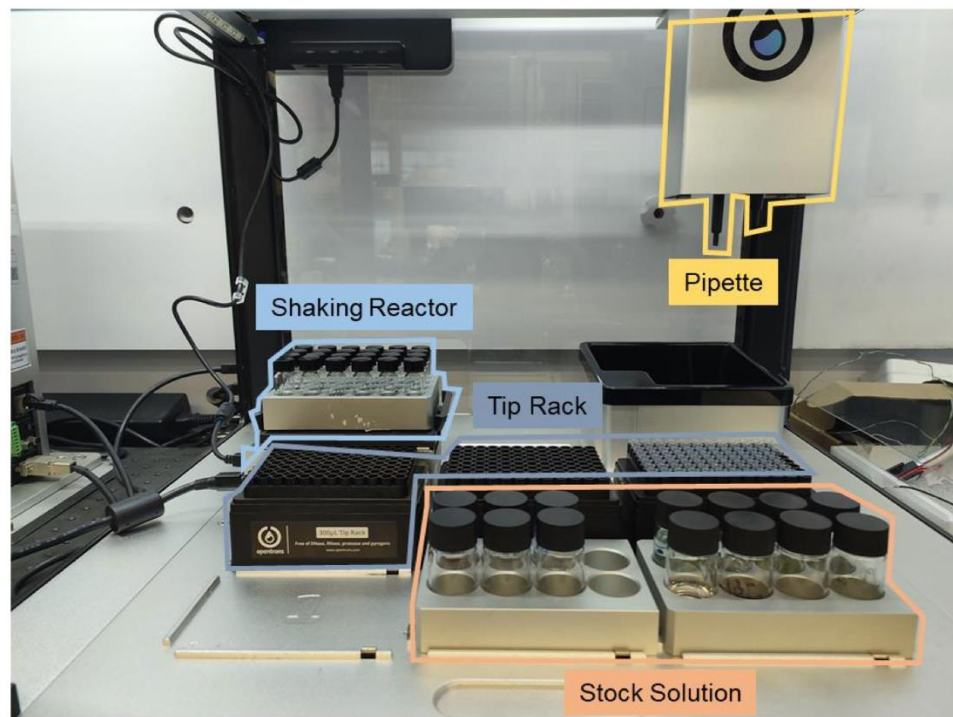
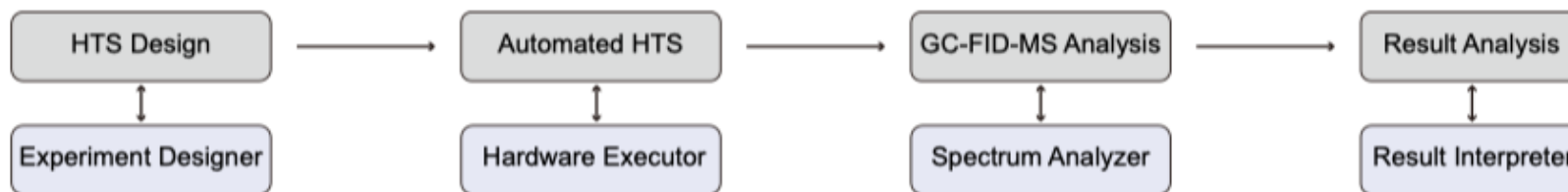
First-principle theories hard to predict yield

Machine learning based methods need amount of experimental data

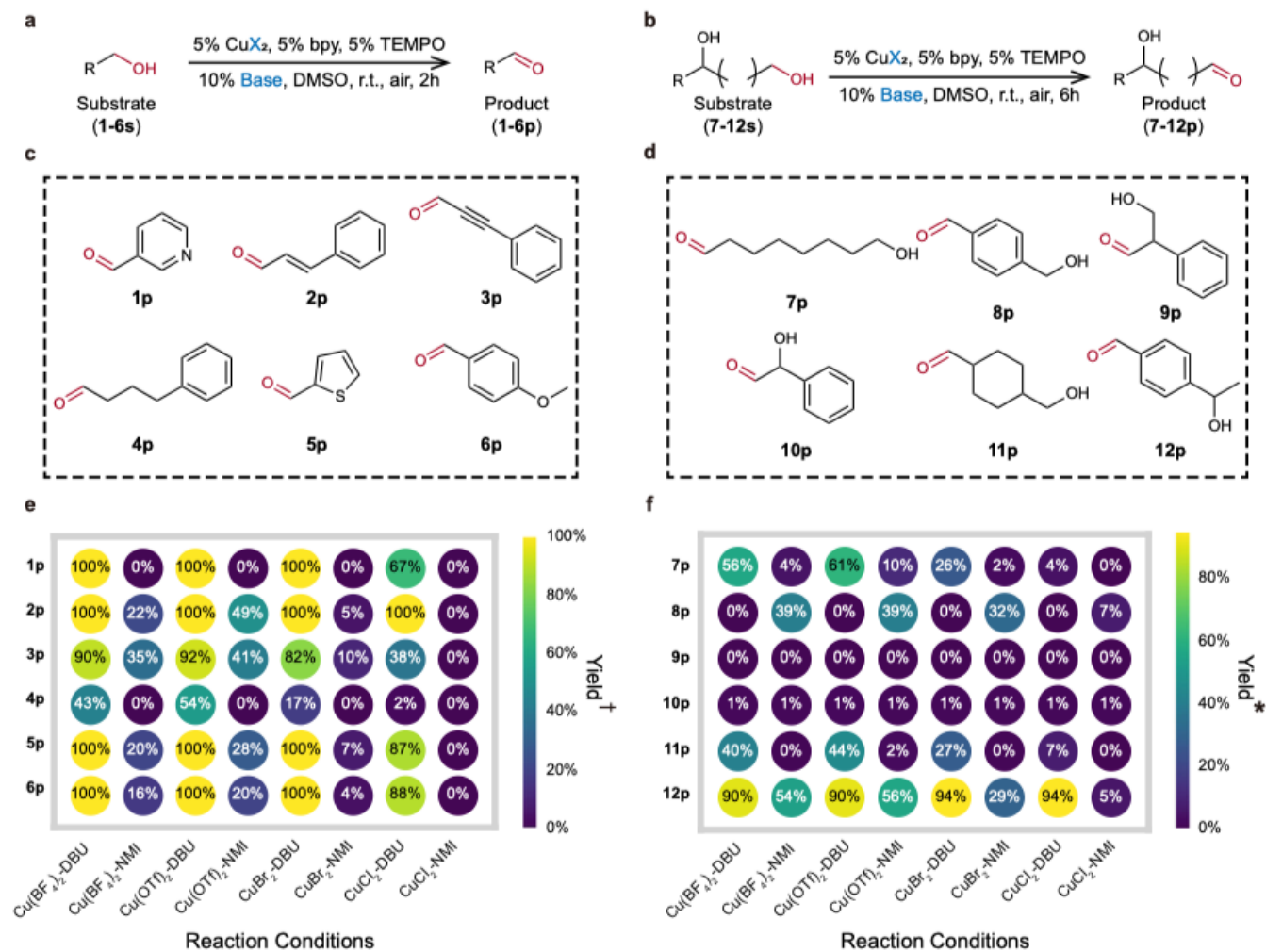
High-throughput screening need programming, manual analysis and expensive

It uses **microplates** as the carrier for experimental tools, an **automated operating system** to execute the experimental process, a sensitive and rapid **detection instrument** to collect experimental result data, and a **computer to analyze** and process the experimental data.

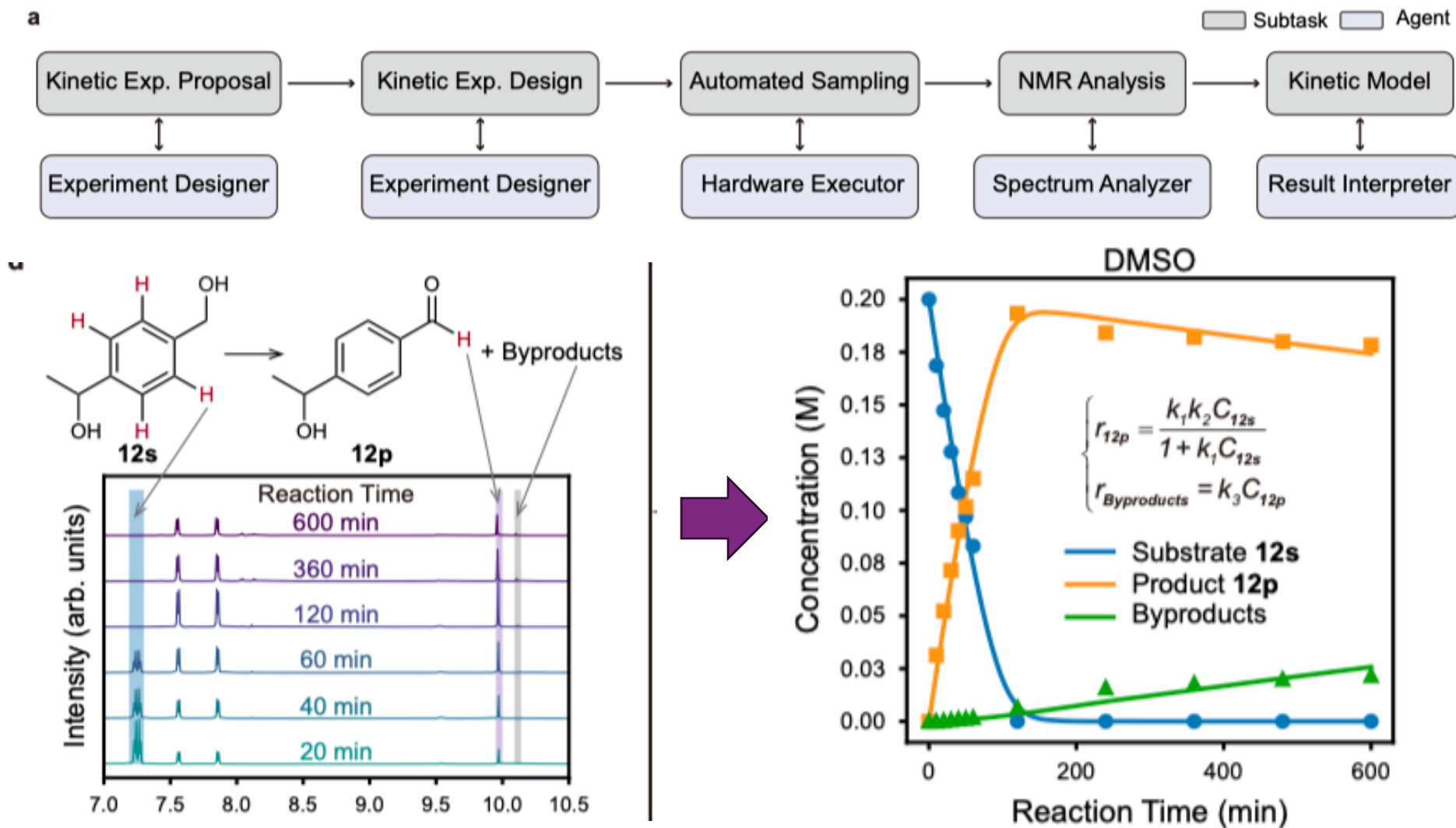
Substrate Scope and Condition Screening



Substrate Scope and Condition Screening



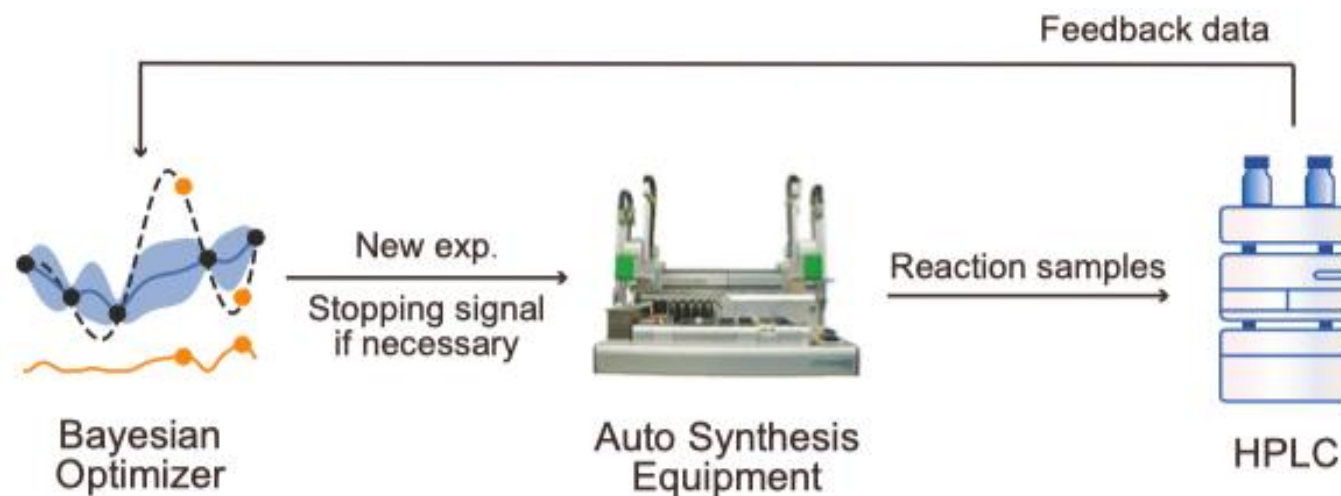
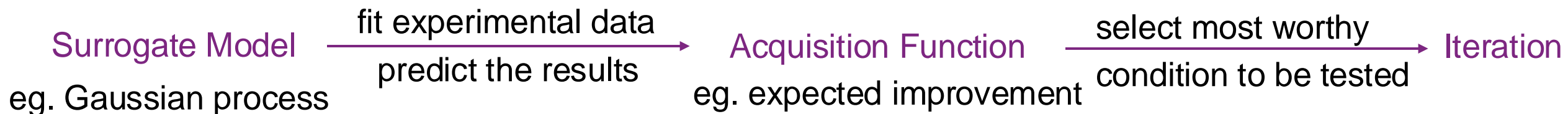
Reaction Kinetics Study



Reaction Condition Optimization!

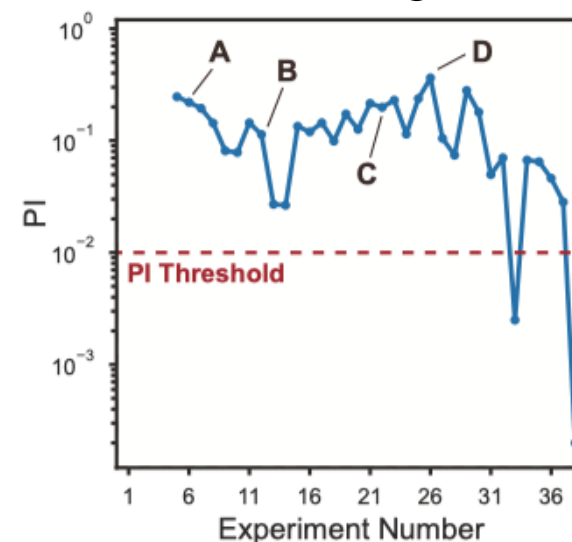
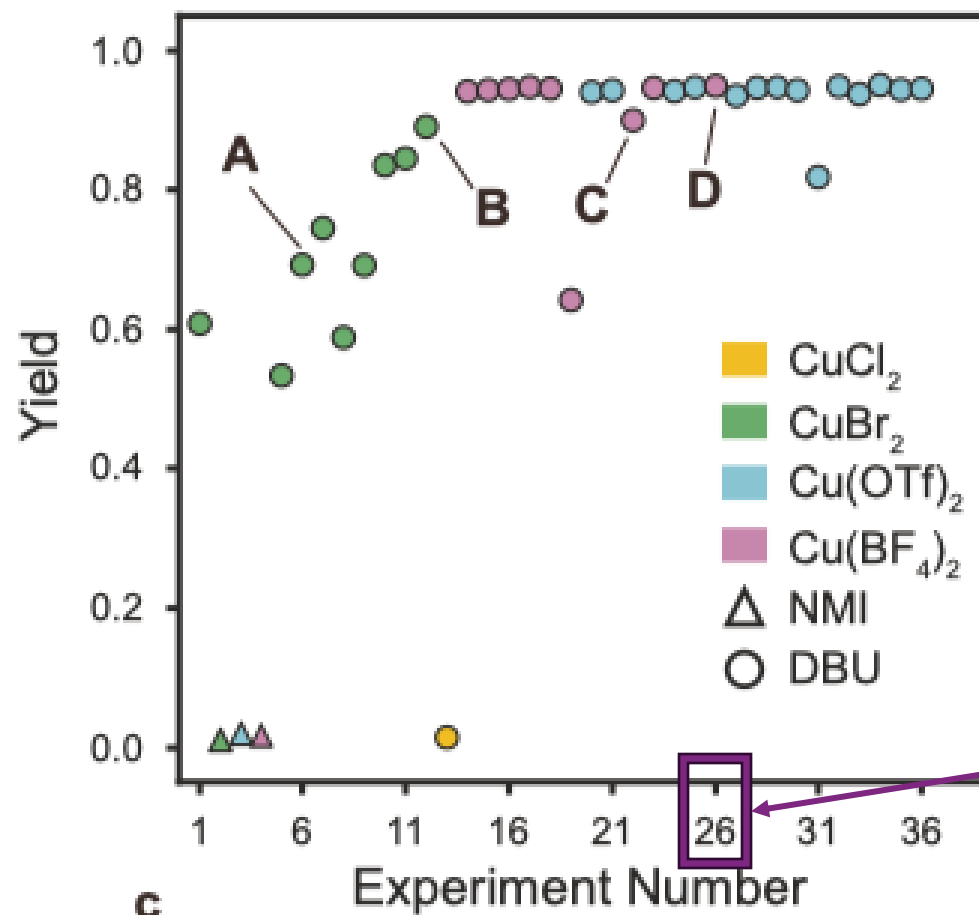
Bayesian Optimization

- An efficient global optimization algorithm
- Suitable for the "black-box function"



Reaction Condition Optimization

Use probability of improvement and LLM to terminate the BO algorithm.



D: 26th Experiment

The recommendation is to terminate optimization because high yields have been achieved, and a wide range of conditional spaces have been explored.

Reaction Scale-up and Product Purification

Scale-up reaction

Choose Condition

User: Select the best condition, considering cost, efficiency for large-scale production.

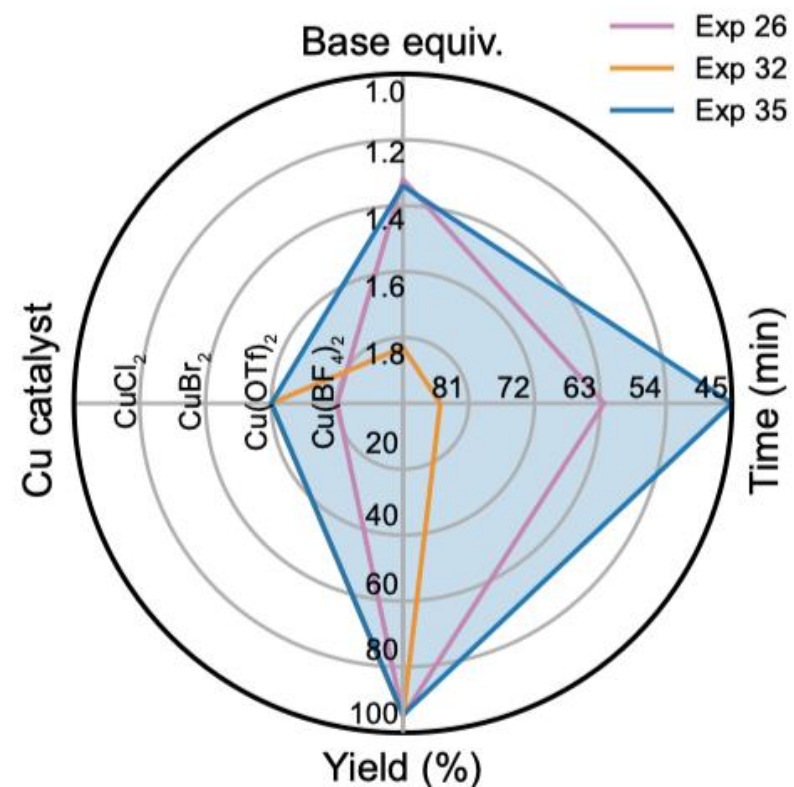
Answer: To select the optimal experimental condition, numerical scoring was applied to reaction time, base equivalence, and catalyst cost ($\text{CuCl}_2 < \text{CuBr}_2 < \text{Cu}(\text{OTf})_2 < \text{Cu}(\text{BF}_4)_2$), with lower values preferred. Each parameter was normalized to a 0-1 scale. The 35th experiment was chosen from among those with high yields, as it had the highest average score across the evaluated parameters. The right radar chart highlighted that the experiment 35 had the largest area, reflecting superior performance across the evaluated criteria.

Scale-up Design*

User: Design a scale-up strategy for the gas-liquid biphasic aerobic alcohol oxidation reaction, currently at 7.5 mg in 1.25 mL reaction mixture.

Answer: The core of scaling up this reaction is ensuring efficient gas-liquid contact for effective oxygen transfer. Propose a two-stage scale-up:

- 1 g scale to validate reproducibility: Reactor: 250 mL round-bottom flask; Mass transfer: Mechanical stirring; Oxygen source: Ambient air.
- 100 g scale to assess industrial production feasibility: Reactor: > 20 L glass or stainless steel reactor; Mass transfer: bubbling device or gas distributor; Oxygen source: Air or oxygen introduced through gas distributor or bubbling device.



Reaction Scale-up and Product Purification

Purification

Eluent Optimization*

User: With 1:1 hexane : ethyl acetate, the product's R_f is 0.63, and an impurity's R_f is 0.83. Adjust eluent composition to make product R_f value range of 0.2 to 0.3 in TLC.

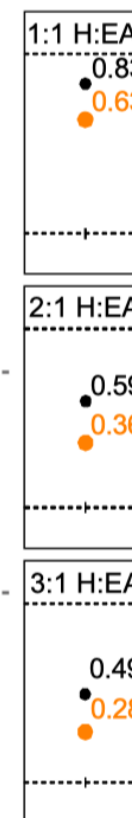
Answer: To lower R_f values and improve separation, increase the proportion of the less polar solvent hexane. Adjust to hexane:ethyl acetate = 2:1.

User: With 2:1 hexane : ethyl acetate, the product's R_f is 0.36, and an impurity's R_f is 0.59.

Answer: To further optimize, consider increasing the proportion of hexane by adjusting the ratio to hexane : ethyl acetate = 3:1.

User: With 3:1 hexane : ethyl acetate, the product's R_f is 0.28, and an impurity's R_f is 0.49.

Answer: This ratio is effective, bringing the product's R_f into the ideal range for column chromatographic separation.



3. Conclusion and Methods



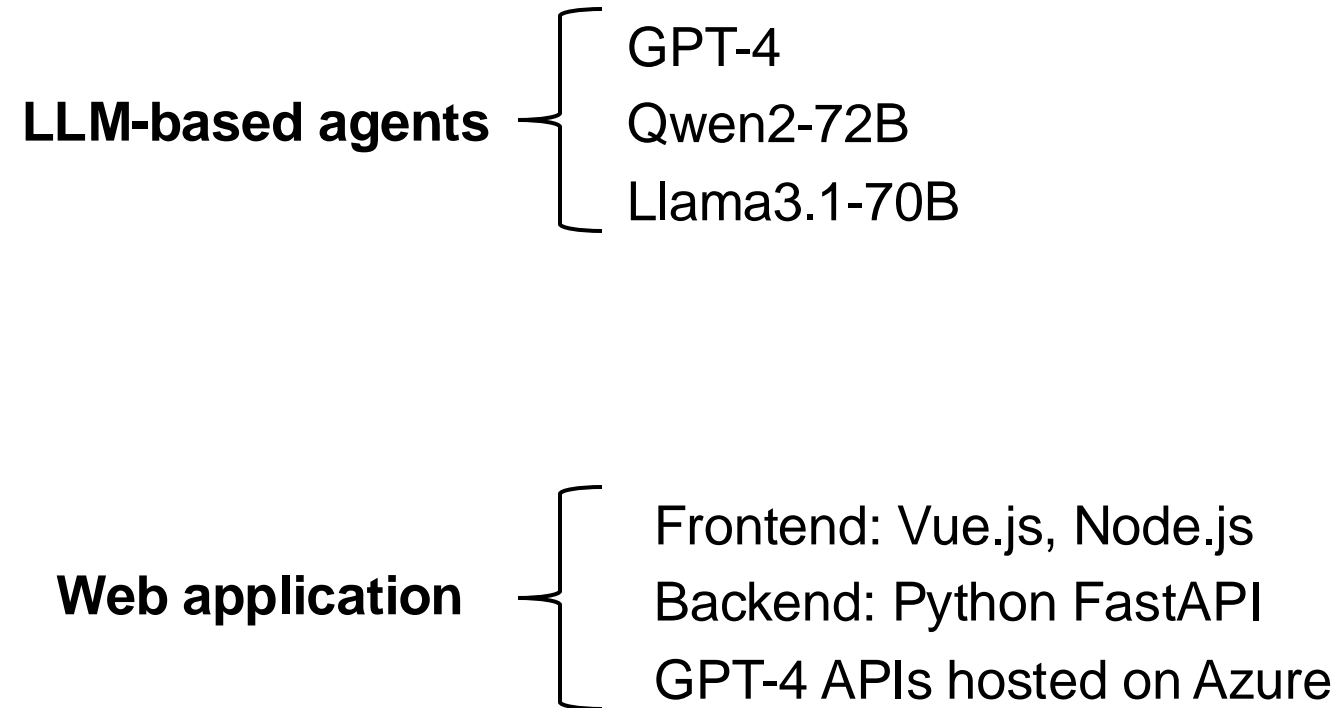
Conclusion

LLM-RDF demonstrates a transformative approach to chemical synthesis that integrates chemist users, LLM-based agents, and automated experimental platforms, significantly streamlining the traditional expert-driven and labor-intensive workflow of reaction development.

Limitations

- LLM-based agents may provide incorrect responses ← Introducing another LLM to double-check
- Lack of domain knowledge ← Retrieval-augmented generation
- Mathematical operations (recognized limitations) ← Equip with integrated tools
- Reproducibility and transparency ← More training
- Communication among LLM-based agents ← People can only do critical decisions

Methods



Methods

JSON

A lightweight data-interchange format

Easy for humans to read and write

Convenient for machines to parse and generate

Adopts an independent text format

Array (in square brackets)

Each object is enclosed in curly braces

Including key - value pairs

```
1. [
2.   {
3.     "reagent": "alcohol",
4.     "parameter": 0.25,
5.     "unit": "ml"
6.   },
7.   {
8.     "reagent": "Cu catalyst",
9.     "parameter": 0.25,
10.    "unit": "ml"
11.  },
12.  {
13.    "reagent": "bpy",
14.    "parameter": 0.25,
15.    "unit": "ml"
16.  },
```

Methods

Bayesian Algorithm

$$\begin{aligned} \left\{ \mathbf{x}_{new}^{(k)} \right\}_{k=1}^q &= \operatorname{argmax}_q \operatorname{EI} \left(\left\{ \mathbf{x}^{(k)} \right\}_{k=1}^q \right) \\ &= \operatorname{argmax} \mathbb{E}_n \left(\operatorname{ReLu} \left(\max_{i=1, \dots, q} f(\mathbf{x}_i) - f_n(\mathbf{x}^+) \right) \right) \end{aligned} \quad (3)$$

where $\{\mathbf{x}_{new}^{(k)}\}_{k=1}^q$ is the q newly proposed reaction conditions, \mathbf{x}^+ is the current optimal condition, and \mathbb{E}_n indicates that the expectation is taken under the posterior distribution at time n .

PI Algorithm

$$\operatorname{PI}(\mathbf{x}) = \mathbb{P}(f(\mathbf{x}) \geq f(\mathbf{x}^+) + \xi) = \Phi \left(\frac{\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi}{\sigma(\mathbf{x})} \right) \quad (4)$$

where $\mu(\cdot)$ is GP's mean, $\sigma(\cdot)$ is GP's standard deviation, $\Phi(\cdot)$ is the normal cumulative distribution function, and ξ is the trade-off parameter of exploitation and exploration.

Thank You