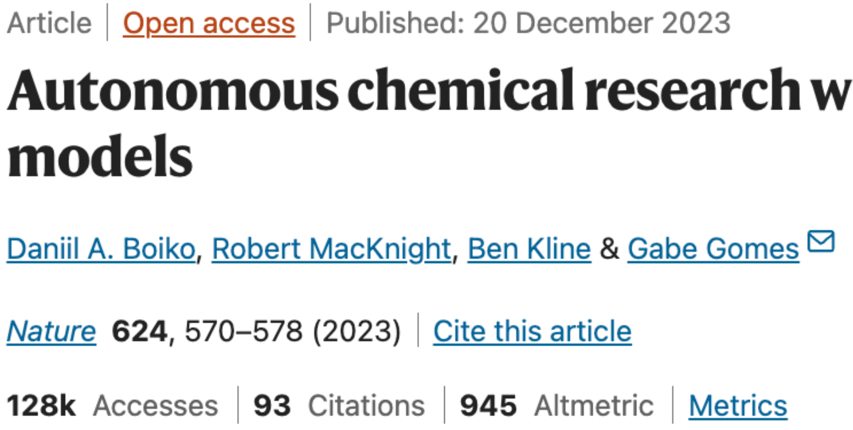


Article | Open access | Published: 20 December 2023

Autonomous chemical research with large language models

Daniil A. Boiko, Robert MacKnight, Ben Kline & Gabe Gomes

Nature **624**, 570–578 (2023) | Cite this article

128k Accesses | **93** Citations | **945** Altmetric | Metrics

Student : 林子軒

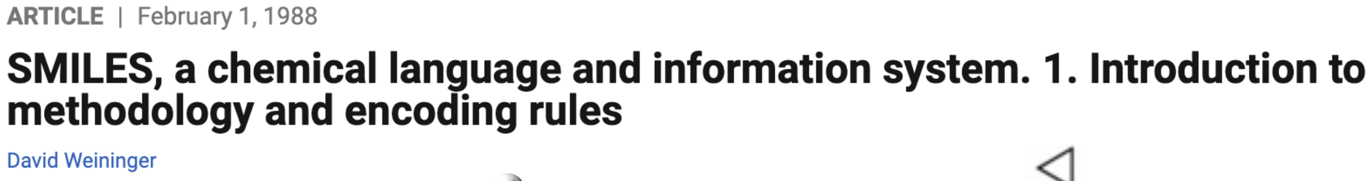
ID：113023503

Date : 2024/09/16

Lab :楊自雄老師 智動化分子研發實驗室

Introduction

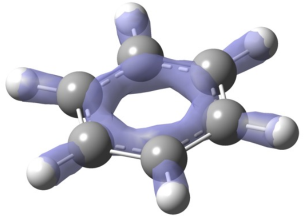
SMILES format

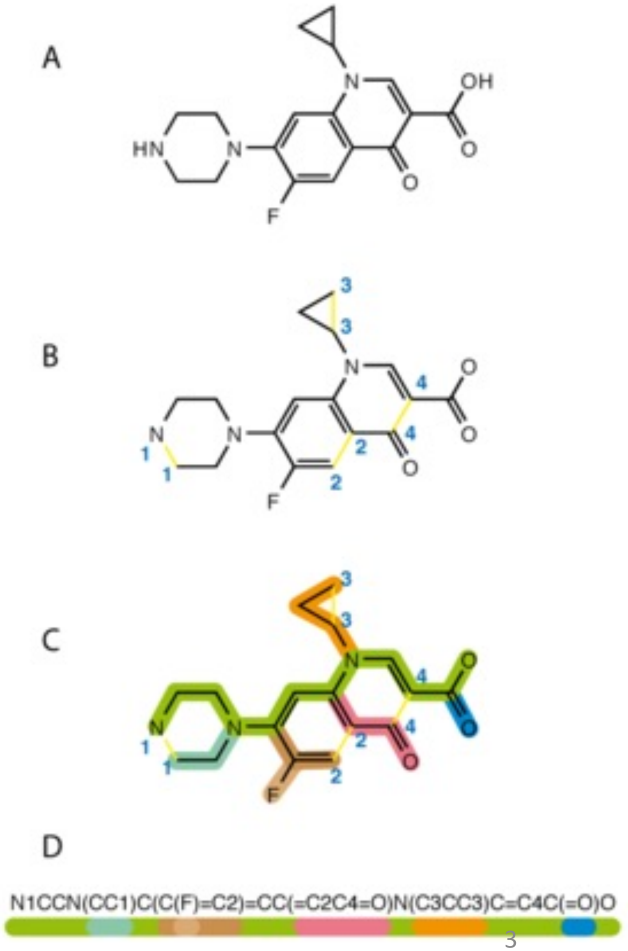


ARTICLE | February 1, 1988

SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules

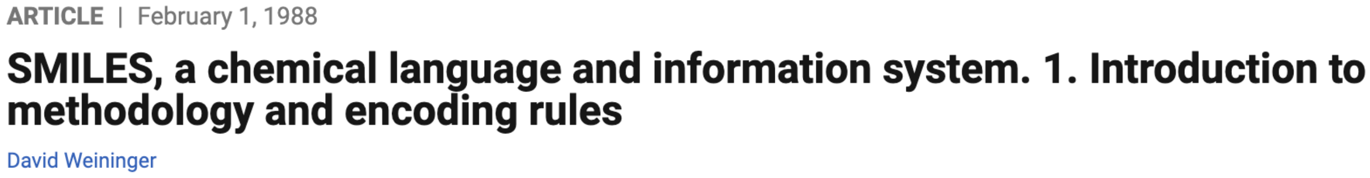
David Weininger





- **3D, 2D -> 1D, how?**
- **Simplicity**: SMILES uses **ASCII characters** to represent molecules, making it easy to input and process by computers.
- **Readability**: For simple molecules, **SMILES strings can be relatively easy** for humans to read and interpret.
- **Compactness**: SMILES can represent complex molecular structures **in a compact string format**.

SMILES format

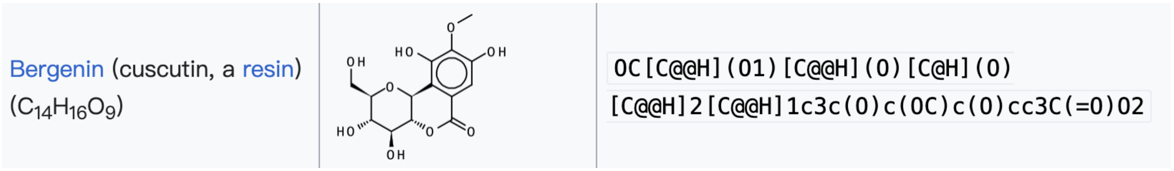


ARTICLE | February 1, 1988

SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules

David Weininger

- Atoms are represented by their **atomic symbols** (e.g., C for carbon, O for oxygen).
- **Single bonds** are implied and not explicitly written.
- **Double bonds** are represented by '=', triple bonds by '#’.
- **Branching** is shown using parentheses.
- **Rings** are indicated using numbers to show connection points.
  - Ethanol: CCO
  - Benzene: c1ccccc1
  - CC1=CN(C2CC(O)C(CO[P](O)(=O)O[P](O)(=O)O[P](O)(O)=O)O2)C(=O)NC1=O



Bergenin (cuscutin, a resin) ($C_{14}H_{16}O_9$)		`OC[C@@H](O1)[C@@H](O)[C@H](O)[C@@H]2[C@@H]1c3c(O)c(OC)c(O)cc3C(=O)O2`

Large language models (LLMs)

arXiv > cs > arXiv:2303.08774

Computer Science > Computation and Language

[Submitted on 15 Mar 2023 (v1), last revised 4 Mar 2024 (this version, v6)]

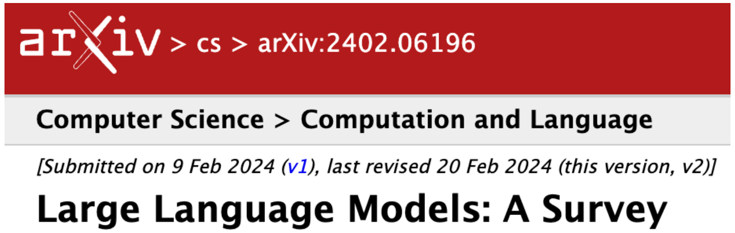
GPT-4 Technical Report

arXiv > math > arXiv:2404.03647

Mathematics > Optimization and Control

[Submitted on 4 Apr 2024]

Capabilities of Large Language Models in Control Engineering: A Benchmark Study on GPT-4, Claude 3 Opus, and Gemini 1.0 Ultra


arXiv > cs > arXiv:2402.06196

Computer Science > Computation and Language

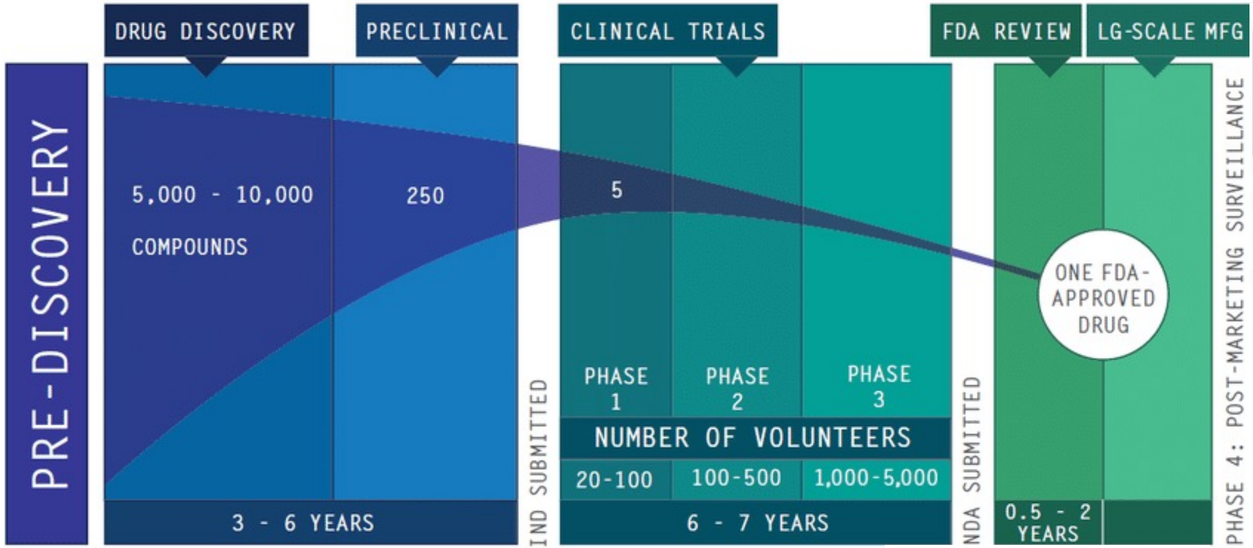
[Submitted on 9 Feb 2024 (v1), last revised 20 Feb 2024 (this version, v2)]

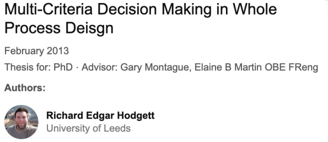
Large Language Models: A Survey

- Large Language Models (LLMs) are AI systems that process and generate **human language**. They are "large" due to their billions to trillions of parameters.
- **Massive Training Data**: LLMs are trained on **vast amounts** of text data from diverse sources, including books, websites, and articles.

Large language models (LLMs)

- But how is LLMs important in chemistry?
- What if it can predict a new drug or a new structure within only a few seconds and dollars?




Multi-Criteria Decision Making in Whole Process Deisgn

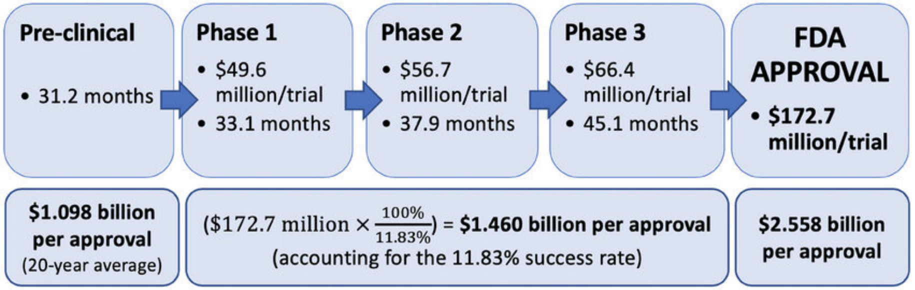
February 2013

Thesis for: PhD · Advisor: Gary Montague, Elaine B Martin OBE FReng

Authors:

Richard Edgar Hodgett
University of Leeds

Article: Multi-Criteria Decision Making in Whole Proce
Author: Gary Montague, Elaine B Martin OBE FReng


Pre-clinical
- 31.2 months

Phase 1
- $49.6 million/trial
- 33.1 months

Phase 2
- $56.7 million/trial
- 37.9 months

Phase 3
- $66.4 million/trial
- 45.1 months

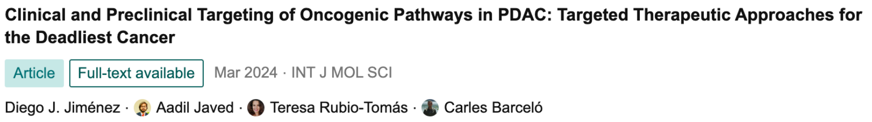
FDA APPROVAL
- **$172.7 million/trial**

$1.098 billion per approval (20-year average)

($172.7 million $\times \frac{100\%}{11.83\%}$) = **$1.460 billion per approval** (accounting for the 11.83% success rate)

$2.558 billion per approval

Data from DiMasi et al. (2016)

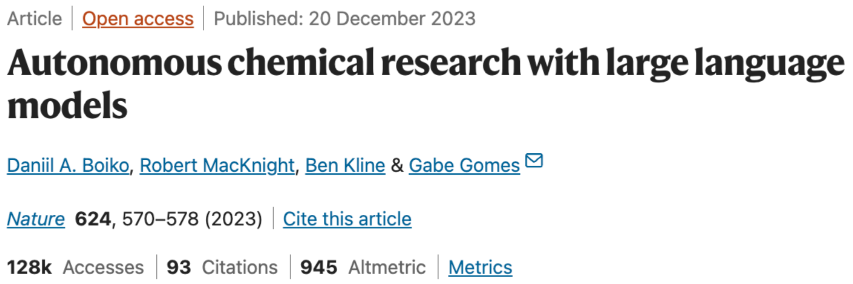

Clinical and Preclinical Targeting of Oncogenic Pathways in PDAC: Targeted Therapeutic Approaches for the Deadliest Cancer

Article | Full-text available | Mar 2024 · INT J MOL SCI

Diego J. Jiménez · Aadil Javed · Teresa Rubio-Tomás · Carles Barceló

Coscientist

Coscientist 6 tasks



Article | Open access | Published: 20 December 2023

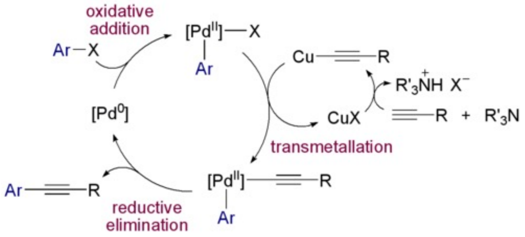
Autonomous chemical research with large language models

Daniil A. Boiko, Robert MacKnight, Ben Kline & Gabe Gomes

Nature **624**, 570–578 (2023) | Cite this article

128k Accesses | **93** Citations | **945** Altmetric | Metrics

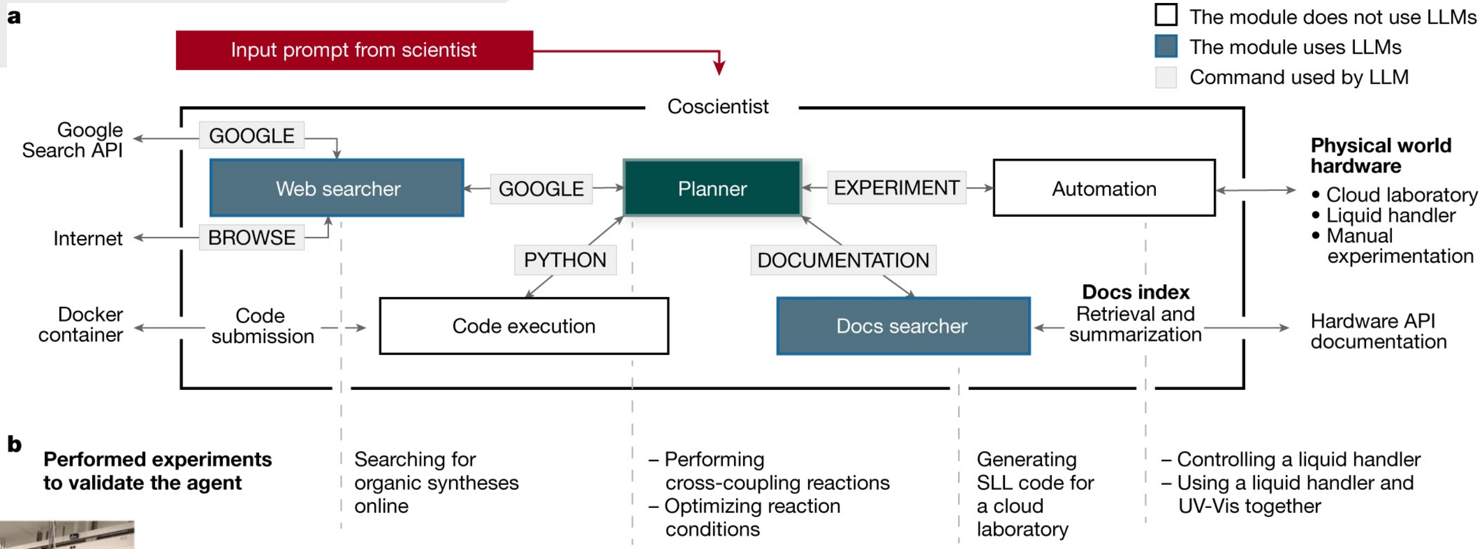
- ✓planning chemical syntheses of known compounds using **publicly available data**
- ✓**efficiently searching** and navigating through extensive hardware documentation
- ✓using documentation to execute high-level commands in a **cloud laboratory**
- ✓precisely controlling **liquid handling instruments** with low-level instructions
- ✓tackling **complex scientific tasks** that demand simultaneous use of multiple hardware modules and integration of diverse data sources
- ✓**solving optimization problems** requiring analyses of previously collected **experimental data**



Palladium-catalyzed cross-coupling reactions
Nobel prize 2010: Richard Fred Heck

Coscientist system architecture

Coscientist system architecture



a

- ☐ The module does not use LLMs
- 🟦 The module uses LLMs
- ▢ Command used by LLM

Input prompt from scientist

Coscientist

Google Search API ← GOOGLE → Web searcher ← GOOGLE ← Planner

Internet ← BROWSE → Web searcher

Planner ← EXPERIMENT → Automation

PYTHON

DOCUMENTATION

Docker container ← Code submission → Code execution

Docs searcher ← **Docs index** Retrieval and summarization → Hardware API documentation

Automation ↔ **Physical world hardware**

- Cloud laboratory
- Liquid handler
- Manual experimentation

b

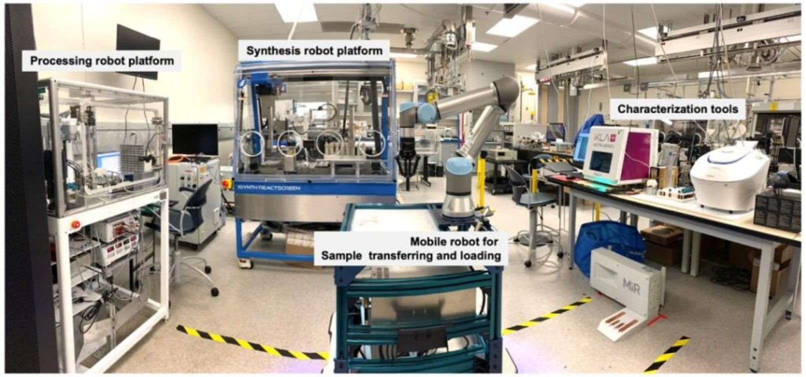
Performed experiments to validate the agent

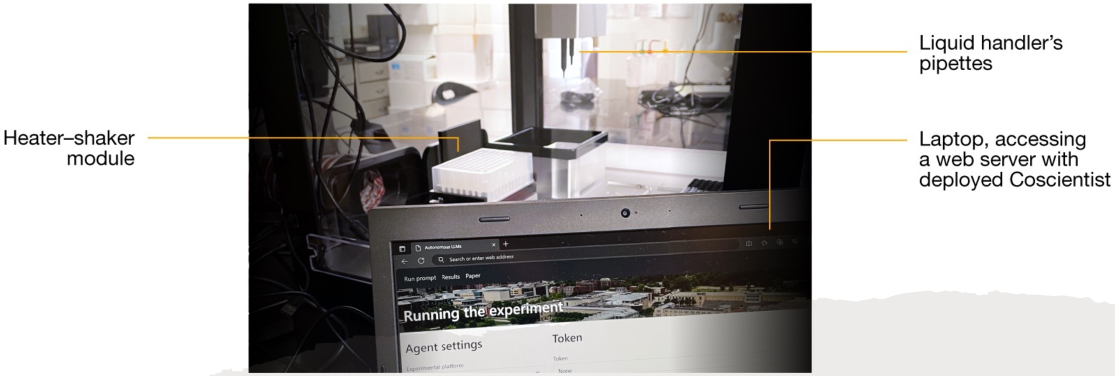
Searching for organic syntheses online

- Performing cross-coupling reactions
- Optimizing reaction conditions

Generating SLL code for a cloud laboratory

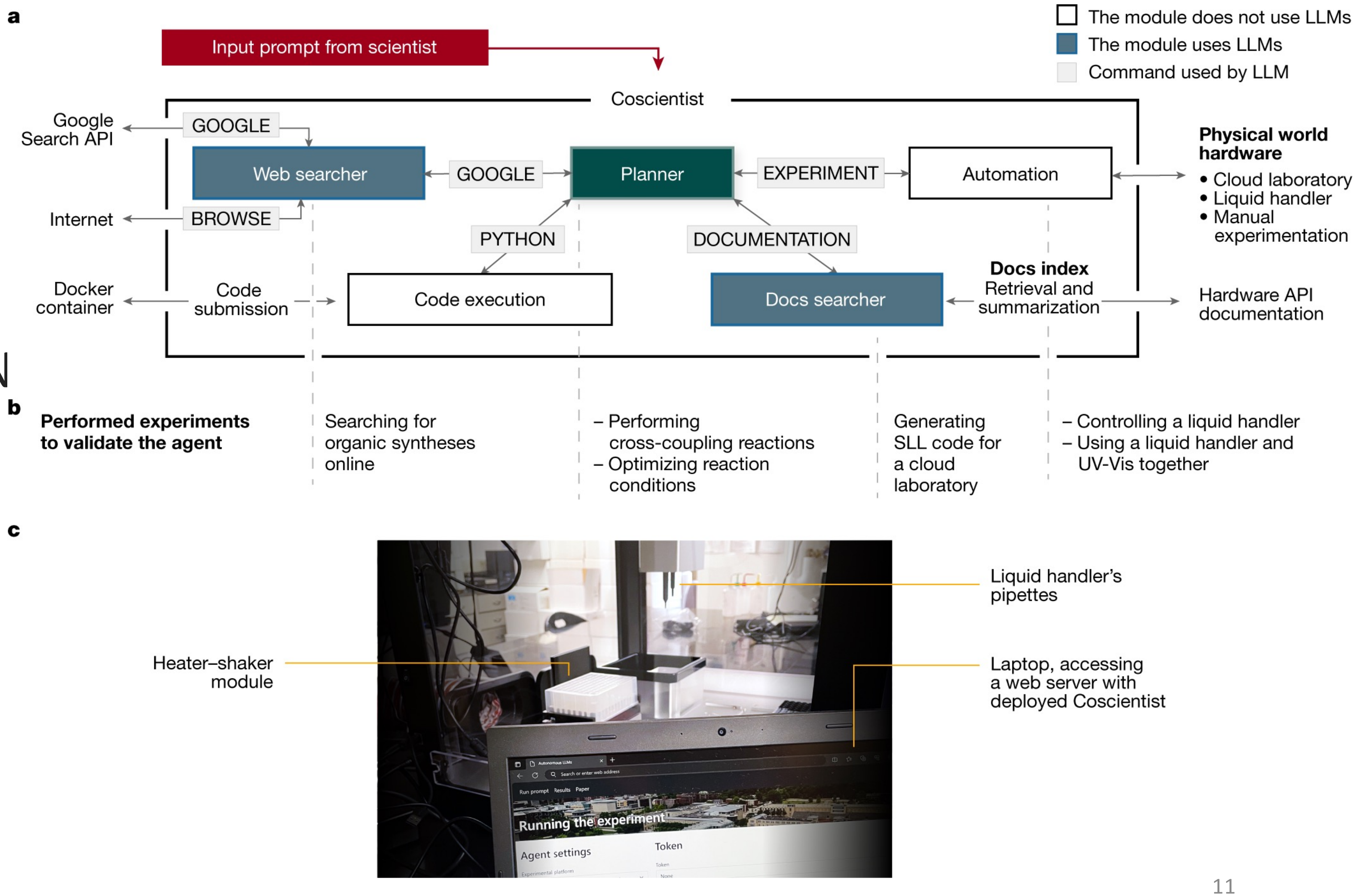
- Controlling a liquid handler
- Using a liquid handler and UV-Vis together


Processing robot platform
Synthesis robot platform
Characterization tools
Mobile robot for Sample transferring and loading


Heater–shaker module
Liquid handler's pipettes
Laptop, accessing a web server with deployed Coscientist
Running the experiment
Agent settings
Token

Command

- ➢GOOGLE
- ➢PYTHON
- ➢DOCUMENTATION
- ➢EXPERIMENT



a

Input prompt from scientist

☐ The module does not use LLMs
■ The module uses LLMs
▢ Command used by LLM

Coscientist

Google Search API

GOOGLE

Web searcher

GOOGLE

Planner

EXPERIMENT

Automation

Internet

BROWSE

PYTHON

DOCUMENTATION

Docker container

Code submission

Code execution

Docs searcher

Docs index
Retrieval and summarization

Physical world hardware
- Cloud laboratory
- Liquid handler
- Manual experimentation

Hardware API documentation

b

Performed experiments to validate the agent

Searching for organic syntheses online

– Performing cross-coupling reactions
– Optimizing reaction conditions

Generating SLL code for a cloud laboratory

– Controlling a liquid handler
– Using a liquid handler and UV-Vis together

c

Liquid handler's pipettes

Heater–shaker module

Laptop, accessing a web server with deployed Coscientist

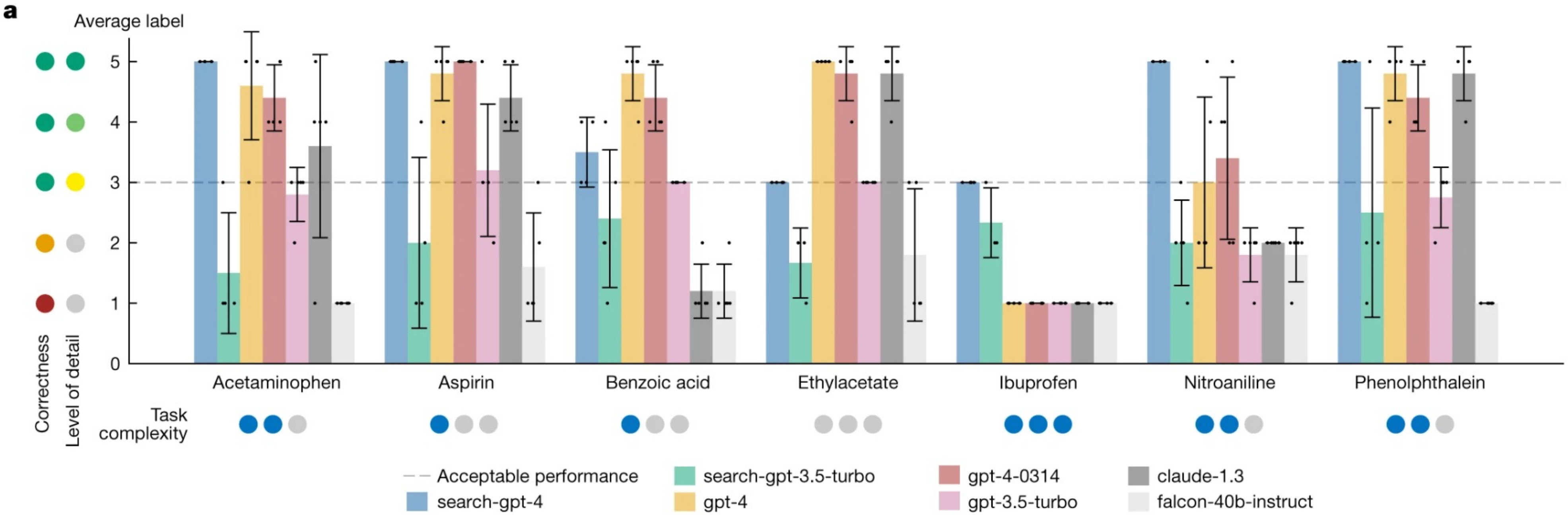
Web search module

Ranking

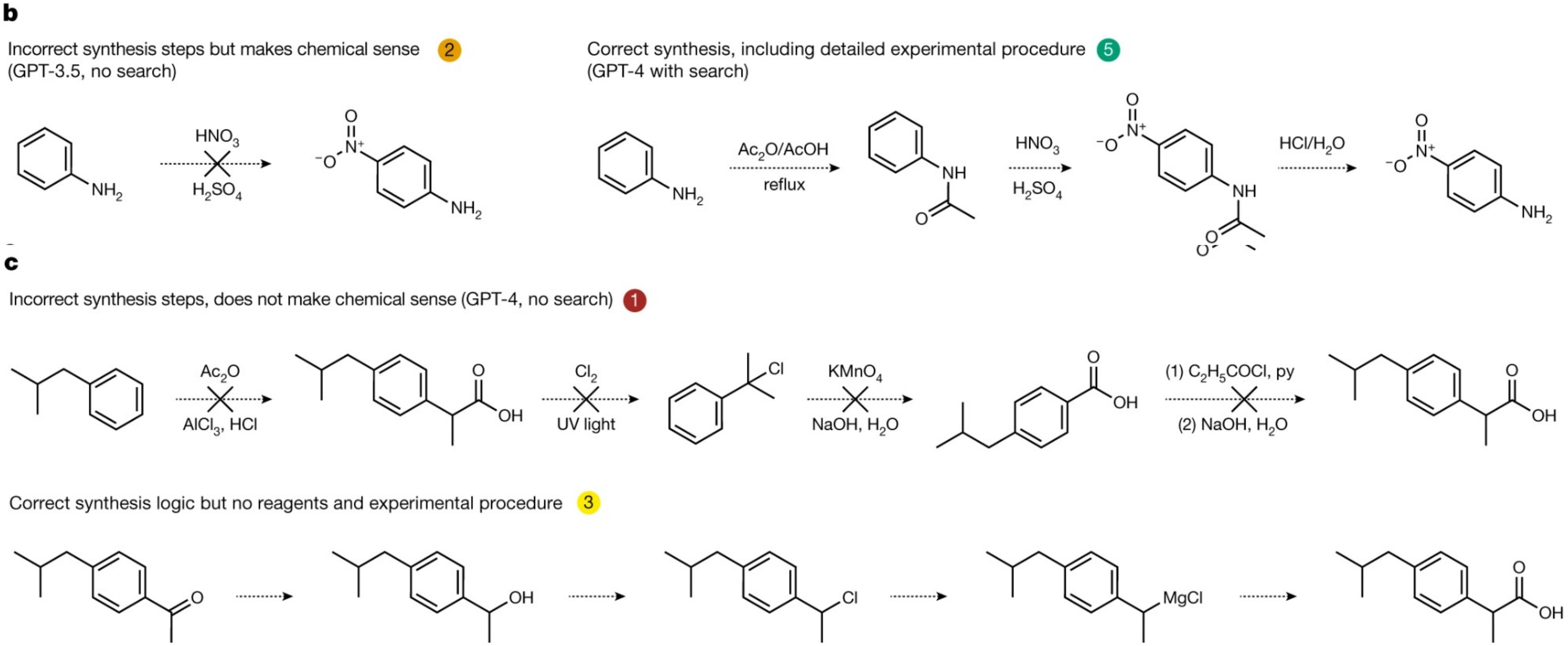
- 5 for a **very detailed** and **chemically accurate** procedure description
- 4 for a **detailed and chemically accurate** description but without reagent quantities
- 3 for a correct chemistry description that **does not include step-by-step procedure**
- 2 for **extremely vague or unfeasible** descriptions
- 1 for **incorrect responses or failure** to follow instructions

All scores below 3 indicate task failure

Web search module



Evaluation



b

Incorrect synthesis steps but makes chemical sense ②
(GPT-3.5, no search)

Correct synthesis, including detailed experimental procedure ⑤
(GPT-4 with search)

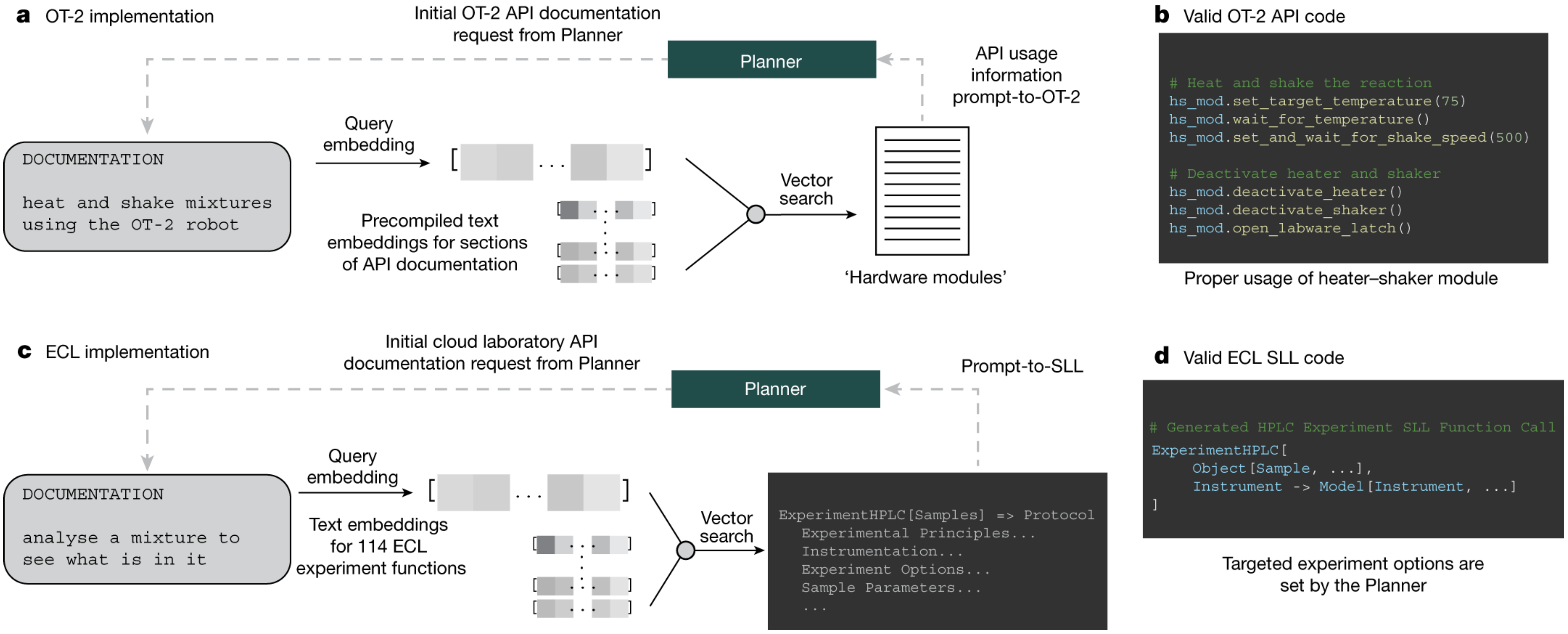
c

Incorrect synthesis steps, does not make chemical sense (GPT-4, no search) ①

Correct synthesis logic but no reagents and experimental procedure ③

Documentation search module

Documentation search module


a
OT-2 implementation
Initial OT-2 API documentation request from Planner
Planner
API usage information prompt-to-OT-2
DOCUMENTATION
heat and shake mixtures using the OT-2 robot
Query embedding
Precompiled text embeddings for sections of API documentation
Vector search
'Hardware modules'
b
Valid OT-2 API code

```
# Heat and shake the reaction
hs_mod.set_target_temperature(75)
hs_mod.wait_for_temperature()
hs_mod.set_and_wait_for_shake_speed(500)

# Deactivate heater and shaker
hs_mod.deactivate_heater()
hs_mod.deactivate_shaker()
hs_mod.open_labware_latch()
```

Proper usage of heater–shaker module

c
ECL implementation
Initial cloud laboratory API documentation request from Planner
Planner
Prompt-to-SLL
DOCUMENTATION
analyse a mixture to see what is in it
Query embedding
Text embeddings for 114 ECL experiment functions
Vector search

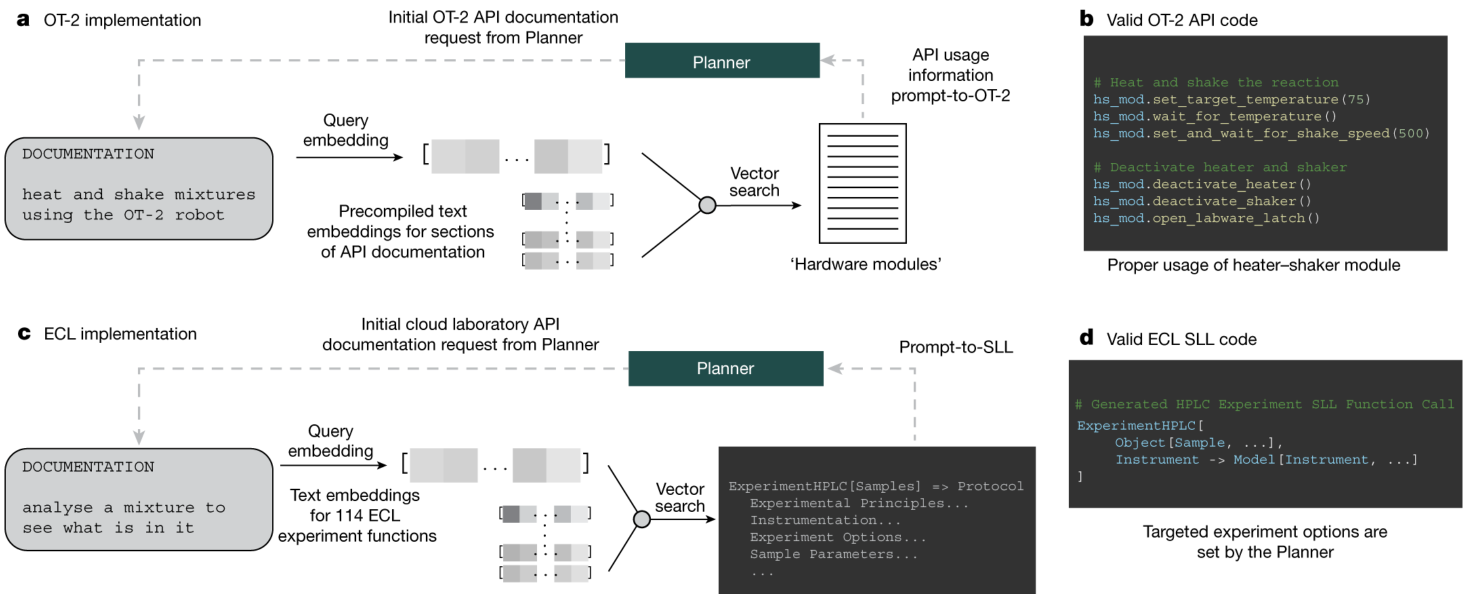
```
ExperimentHPLC[Samples] => Protocol
   Experimental Principles...
   Instrumentation...
   Experiment Options...
   Sample Parameters...
   ...
```

```
# Generated HPLC Experiment SLL Function Call
ExperimentHPLC[
     Object[Sample, ...],
     Instrument -> Model[Instrument, ...]
]
```

Targeted experiment options are set by the Planner

Why is it important?

Documentation search module


a OT-2 implementation
Initial OT-2 API documentation request from Planner
Planner
API usage information prompt-to-OT-2
DOCUMENTATION
heat and shake mixtures using the OT-2 robot
Query embedding
Precompiled text embeddings for sections of API documentation
Vector search
'Hardware modules'
b Valid OT-2 API code
Proper usage of heater–shaker module
c ECL implementation
Initial cloud laboratory API documentation request from Planner
Planner
Prompt-to-SLL
DOCUMENTATION
analyse a mixture to see what is in it
Query embedding
Text embeddings for 114 ECL experiment functions
Vector search
d Valid ECL SLL code
Targeted experiment options are set by the Planner

```
# Heat and shake the reaction
hs_mod.set_target_temperature(75)
hs_mod.wait_for_temperature()
hs_mod.set_and_wait_for_shake_speed(500)

# Deactivate heater and shaker
hs_mod.deactivate_heater()
hs_mod.deactivate_shaker()
hs_mod.open_labware_latch()
```

```
ExperimentHPLC[Samples] => Protocol
  Experimental Principles...
  Instrumentation...
  Experiment Options...
  Sample Parameters...
  ...
```

```
# Generated HPLC Experiment SLL Function Call
ExperimentHPLC[
    Object[Sample, ...],
    Instrument -> Model[Instrument, ...]
]
```

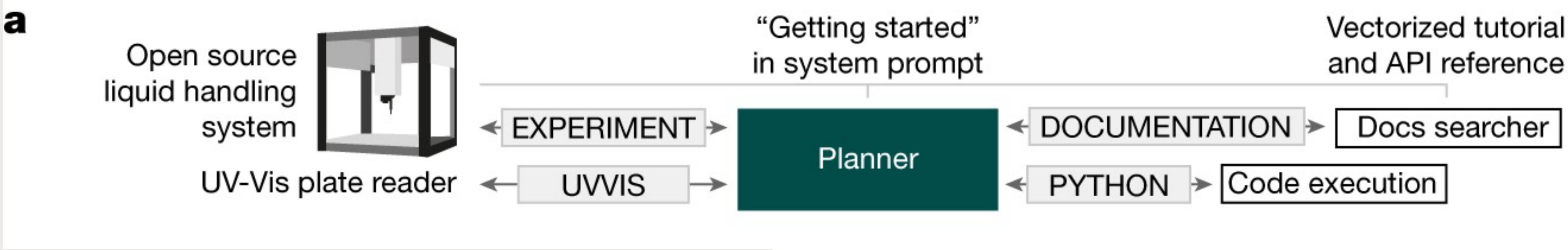
Why is it important?

- ✓**Technical Integration**
  - ✓Capable of addressing the complexity of **software components** and their interactions
  - ✓Crucial for integrating LLMs with **laboratory automation**
- ✓**Effective Utilization of Technical Documentation**
  - ✓Enables Coscientist to **understand** and use technical documentation
- ✓**Learning New Languages and Systems**
  - ✓Demonstrates GPT-4's ability to **learn new programming languages** (like ECL SLL)

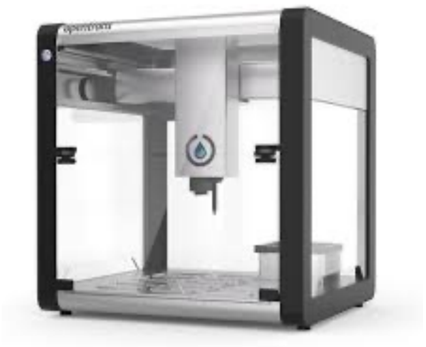
Controlling laboratory hardware

Controlling laboratory hardware

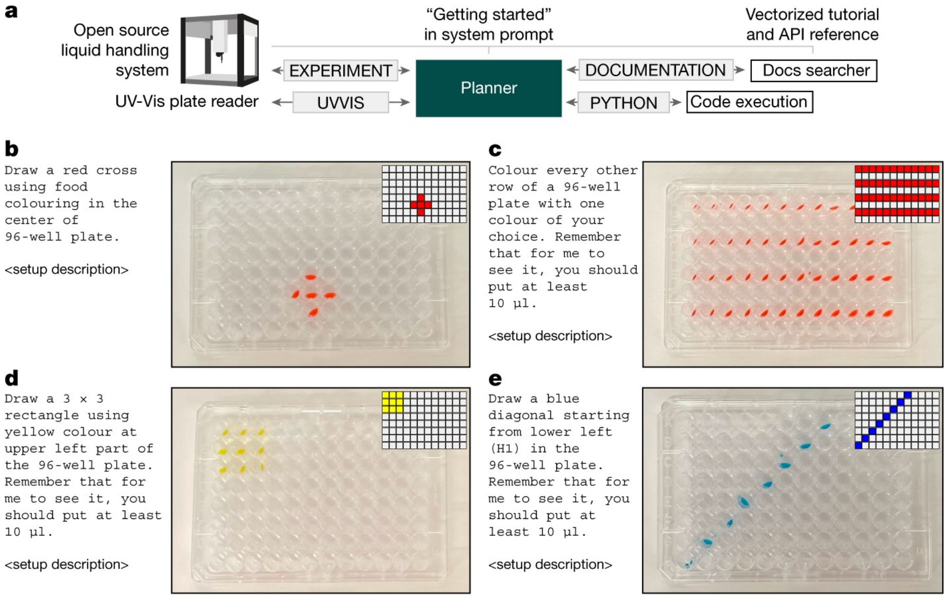
Robotic liquid handler control capabilities and integration with analytical tools.


a
Open source liquid handling system
UV-Vis plate reader
EXPERIMENT
UVVIS
“Getting started” in system prompt
Planner
Vectorized tutorial and API reference
DOCUMENTATION
Docs searcher
PYTHON
Code execution

Controlling laboratory hardware


OT-2

- Coscientist conducts experiments using an open-source **liquid handler OT-2** and its **API**, obtaining information from **documentation**.
- Coscientist is capable of integrating multiple modules, like using **UV-Vis spectrometry**, and solving complex tasks through data analysis, such as **determining sample colors and positions without prior information**.


a
Open source liquid handling system
UV-Vis plate reader
EXPERIMENT
UVVIS
"Getting started" in system prompt
Planner
DOCUMENTATION
PYTHON
Vectorized tutorial and API reference
Docs searcher
Code execution
b
c
d
e

`Draw a red cross using food colouring in the center of 96-well plate.`

<setup description>

`Colour every other row of a 96-well plate with one colour of your choice. Remember that for me to see it, you should put at least 10 µl.`

<setup description>

`Draw a 3 × 3 rectangle using yellow colour at upper left part of the 96-well plate. Remember that for me to see it, you should put at least 10 µl.`

<setup description>

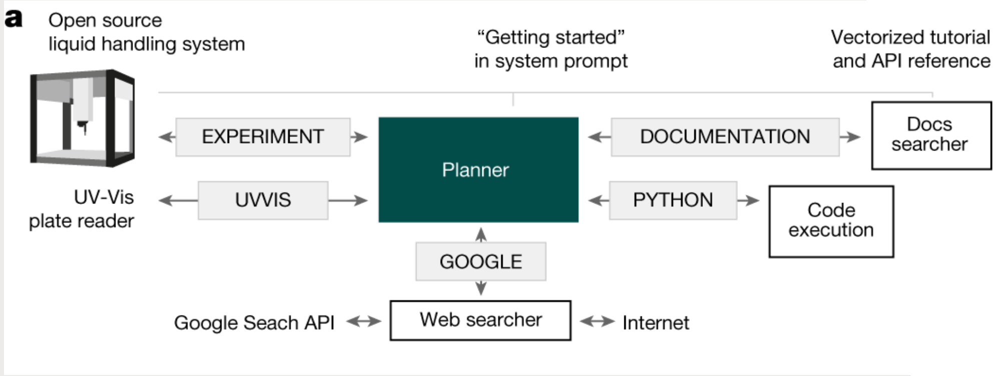
`Draw a blue diagonal starting from lower left (H1) in the 96-well plate. Remember that for me to see it, you should put at least 10 µl.`

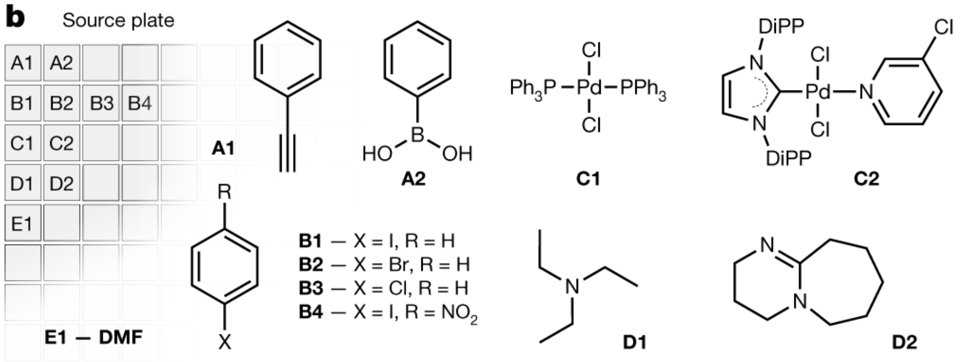
<setup description>

Integrated chemical experiment design

Controlling laboratory hardware

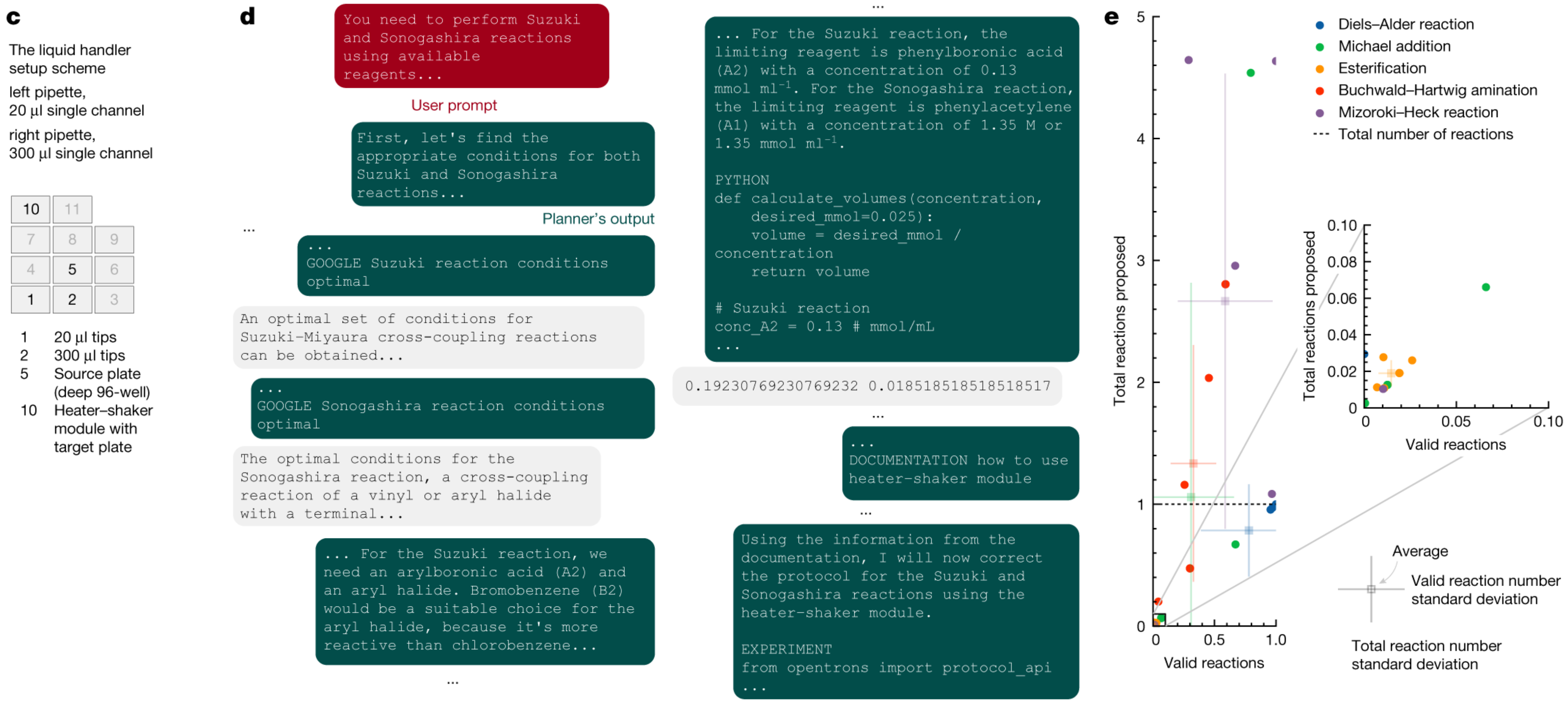
Cross-coupling Suzuki and Sonogashira reaction experiments designed and performed by Coscientist





- No human decision

Cross-coupling Suzukiand Sonogashira reaction



c

The liquid handler setup scheme

left pipette, 20 µl single channel

right pipette, 300 µl single channel

1	20 µl tips
2	300 µl tips
5	Source plate (deep 96-well)
10	Heater–shaker module with target plate

d

```
You need to perform Suzuki
and Sonogashira reactions
using available
reagents...
```

User prompt

```
First, let's find the
appropriate conditions for both
Suzuki and Sonogashira
reactions...
```

Planner's output

...

```
...
GOOGLE Suzuki reaction conditions
optimal
```

```
An optimal set of conditions for
Suzuki-Miyaura cross-coupling reactions
can be obtained...
```

```
...
GOOGLE Sonogashira reaction conditions
optimal
```

```
The optimal conditions for the
Sonogashira reaction, a cross-coupling
reaction of a vinyl or aryl halide
with a terminal...
```

```
... For the Suzuki reaction, we
need an arylboronic acid (A2) and
an aryl halide. Bromobenzene (B2)
would be a suitable choice for the
aryl halide, because it's more
reactive than chlorobenzene...
```

...

...

... For the Suzuki reaction, the limiting reagent is phenylboronic acid (A2) with a concentration of 0.13 mmol ml^{-1}. For the Sonogashira reaction, the limiting reagent is phenylacetylene (A1) with a concentration of 1.35 M or 1.35 mmol ml^{-1}.

```
PYTHON
def calculate_volumes(concentration,
    desired_mmol=0.025):
    volume = desired_mmol /
concentration
    return volume

# Suzuki reaction
conc_A2 = 0.13 # mmol/mL
...
```

```
0.19230769230769232 0.018518518518518517
```

...

```
...
DOCUMENTATION how to use
heater-shaker module
```

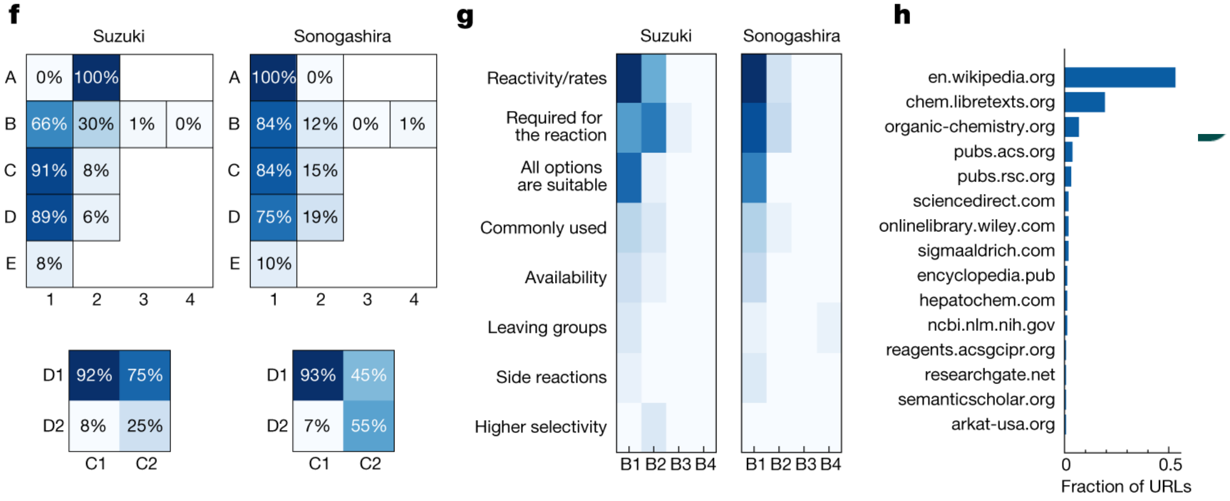
...

```
Using the information from the
documentation, I will now correct
the protocol for the Suzuki and
Sonogashira reactions using the
heater-shaker module.

EXPERIMENT
from opentrons import protocol_api
...
```

e

Cross-coupling Suzuki and Sonogashira reaction



Suzuki

	1	2	3	4
A	0%	100%		
B	66%	30%	1%	0%
C	91%	8%		
D	89%	6%		
E	8%			

Sonogashira

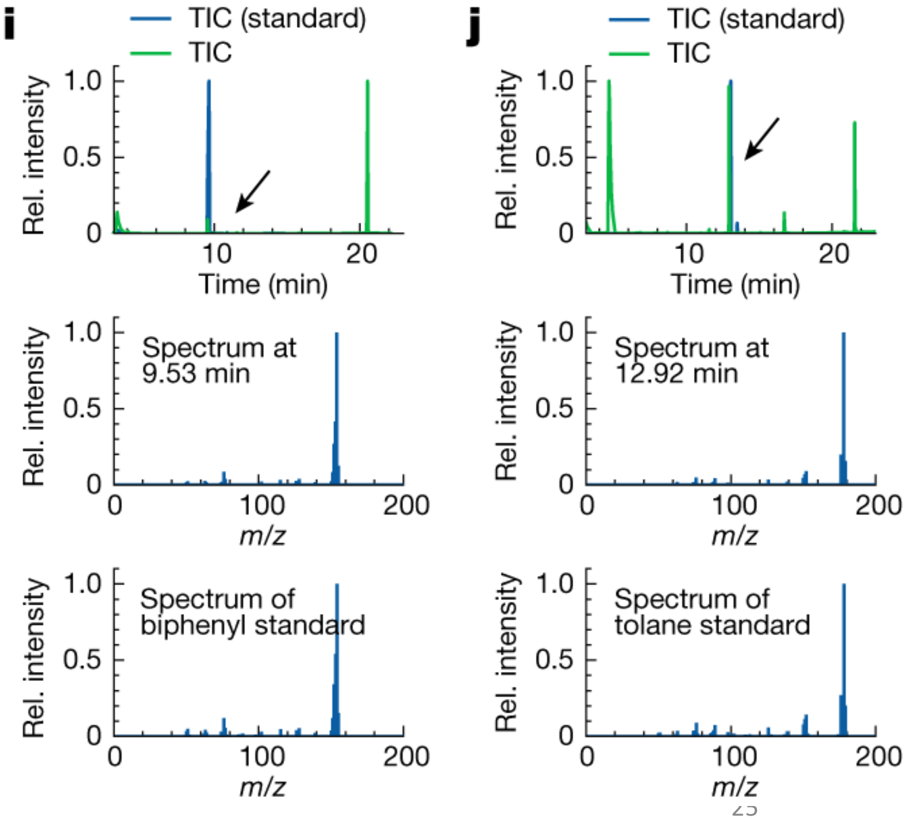
	1	2	3	4
A	100%	0%		
B	84%	12%	0%	1%
C	84%	15%		
D	75%	19%		
E	10%			

Suzuki

	C1	C2
D1	92%	75%
D2	8%	25%

Sonogashira

	C1	C2
D1	93%	45%
D2	7%	55%



- ➢The model is able to **provide explanations** for specific choices
- ➢ demonstrating its **capability to handle concepts such as reactivity and selectivity**.

Discussion

Discussion

- This study demonstrates an AI system capable of **autonomously designing** and **executing complex scientific experiments**.
- The system combines **large language models** with research tools, showcasing **advanced reasoning and experimental design capabilities**.

Thanks for listening!