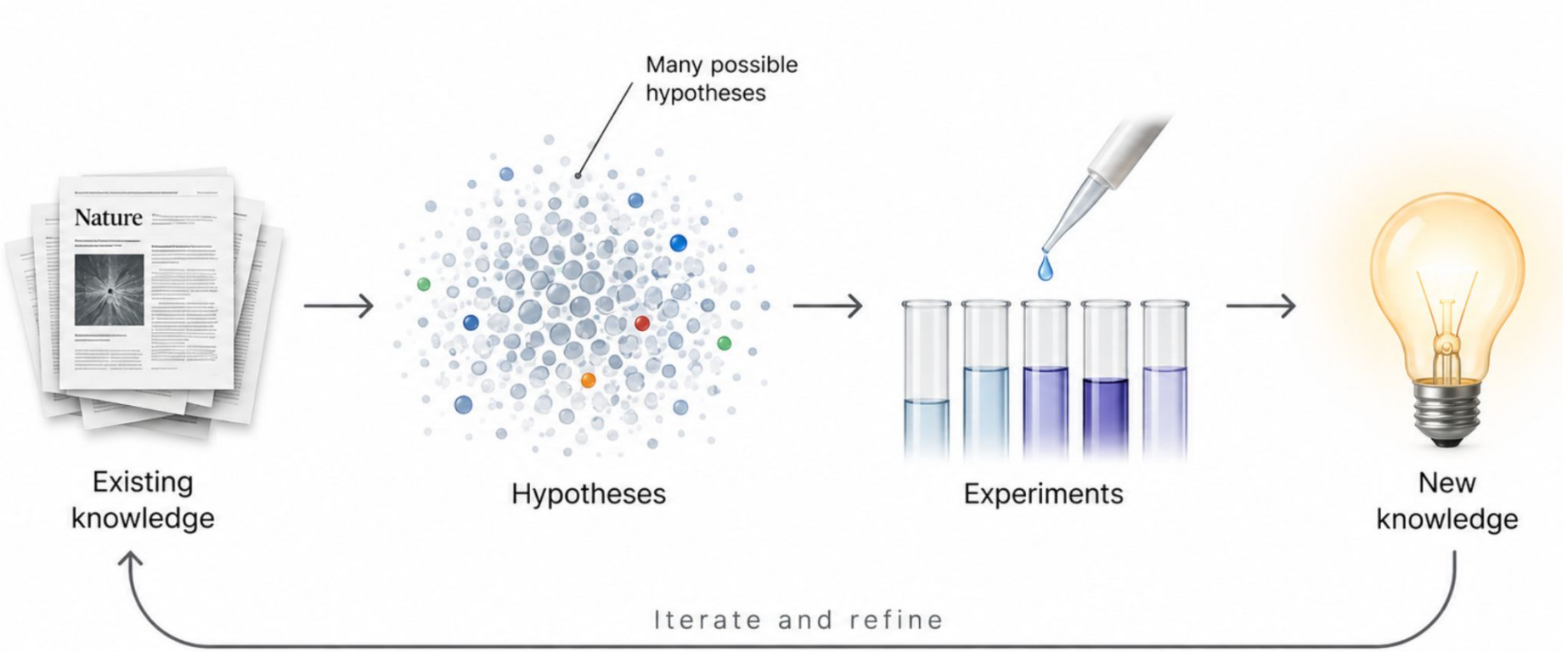
Accelerating scientific discovery with co-scientist

Gottweis et al., Nature 655 (2026)
Received: 20 March 2025; Published: 19 May 2026

Presented by: Kelsey Fu & Tao Zhong

Scientific Discovery Is a Search Problem



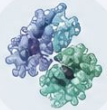
The Breadth–Depth Problem

DEPTH

Increasing specialization



More complex techniques



Deeper domain knowledge



More sophisticated tools and methods



Go deeper in one area

Look across many areas



A single scientist

BREADTH

Exploding knowledge across fields



Rapidly growing literature



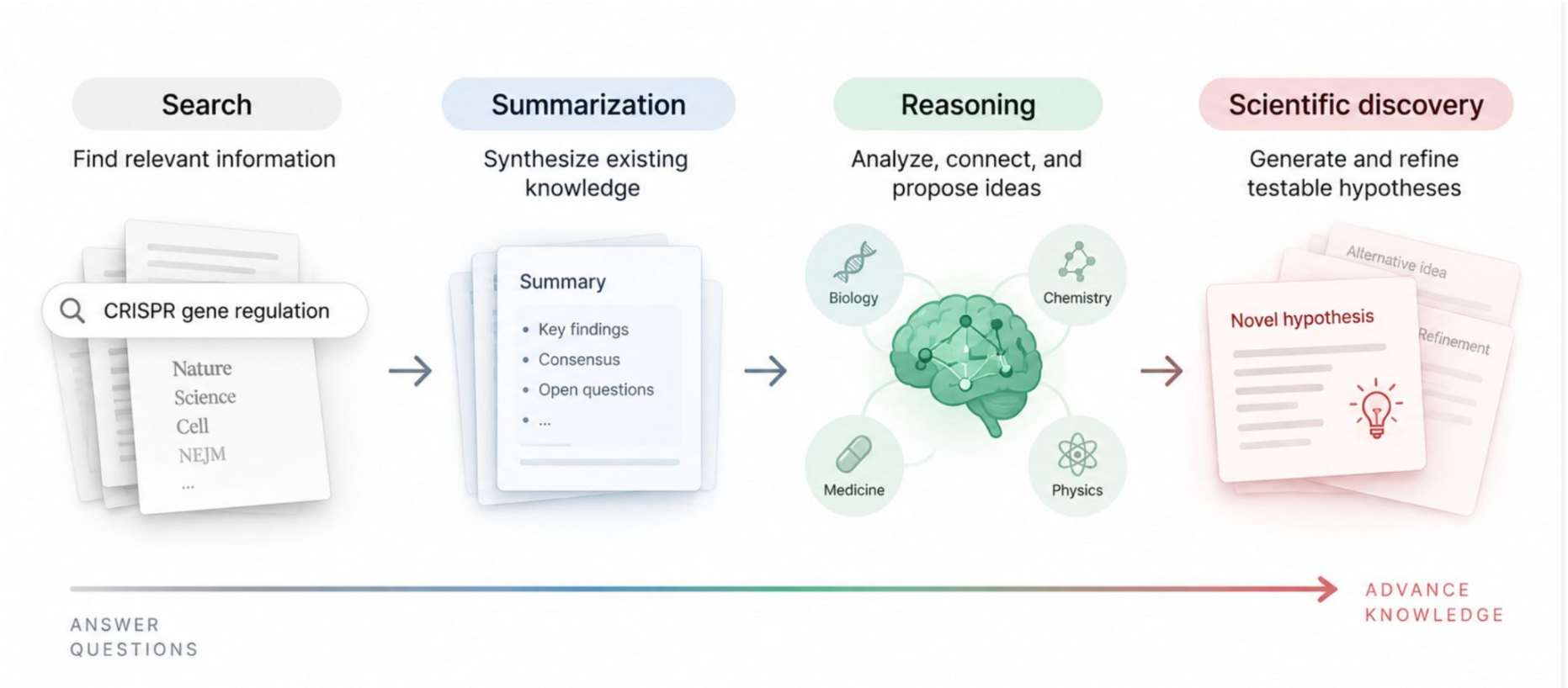
Cross-disciplinary connections



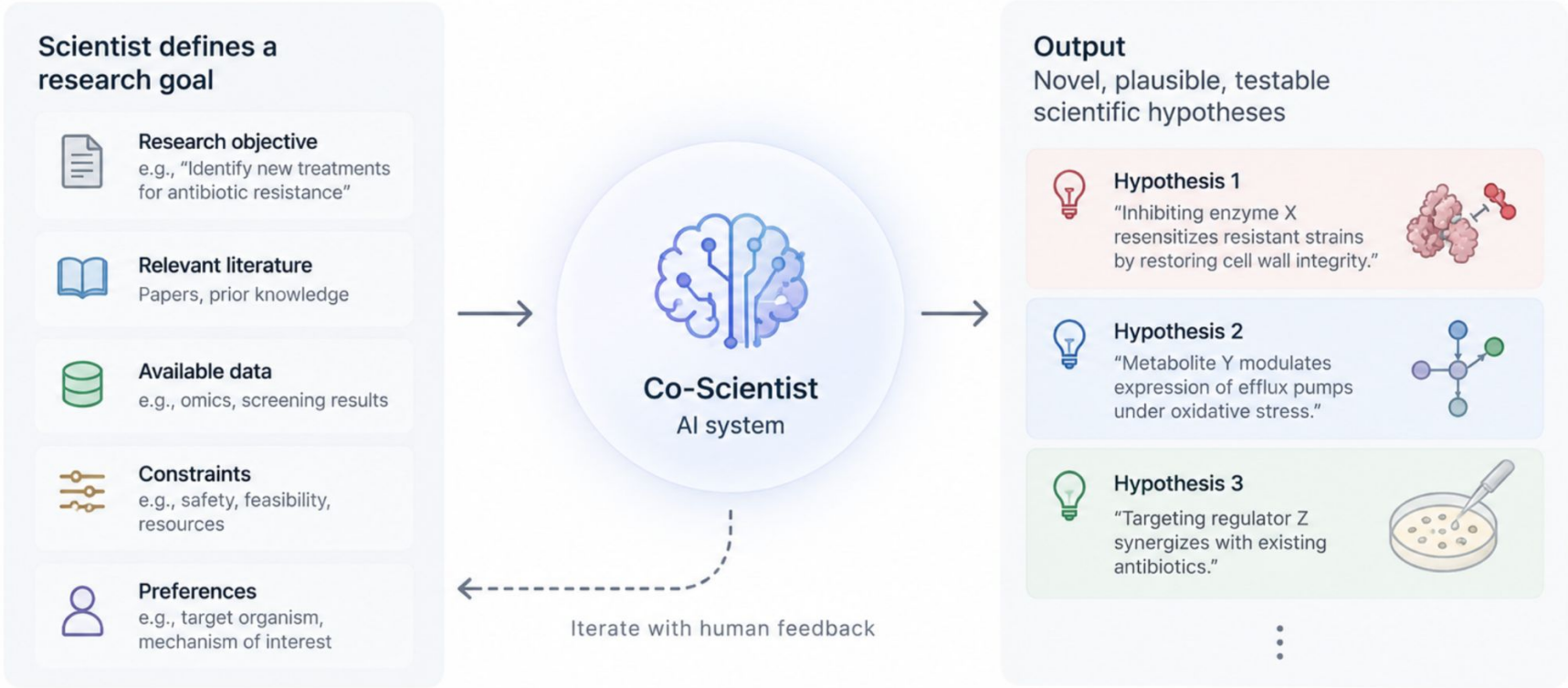
Knowledge across biology, chemistry, physics, medicine, and beyond

No individual scientist can comprehensively navigate both the depth and breadth of modern science.

From AI Tools to AI Scientific Collaborators

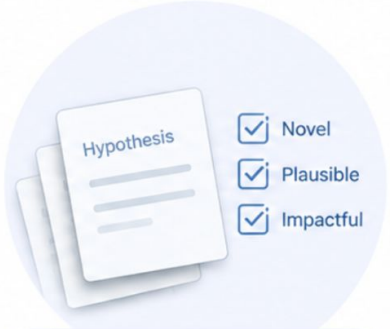


Problem Formulation



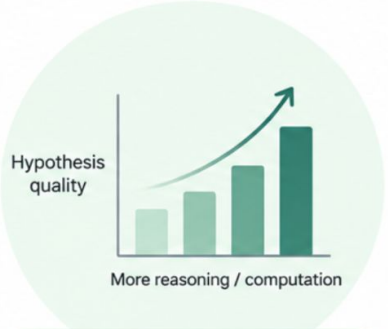
The Core Research Question

Can an AI system iteratively generate,
critique and improve novel scientific hypotheses
—and produce ideas valuable enough
to **test experimentally**?



1. Quality

Are the hypotheses
scientifically strong?



2. Scaling

Does more reasoning
improve them?



3. Discovery

Can they lead to
real experimental findings?

Related Work: What Already Existed?

1. Information tools

Retrieve and summarize scientific knowledge

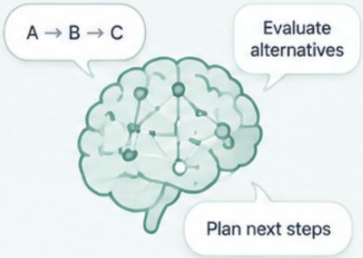


Strength
Access and synthesize vast literature

Limitation
Primarily organize existing knowledge

2. Reasoning models

Perform complex, multi-step reasoning



Strength
Tackle complex scientific problems

Limitation
Often single-model reasoning, not tailored to scientific ideation

3. AI agents

Use tools and take multi-step actions

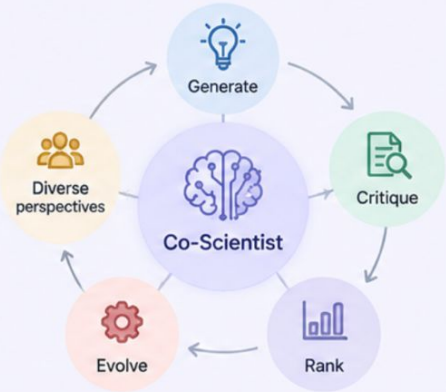


Strength
Iterative reasoning, tool use, multi-step workflows

Limitation
Not specifically designed for scientific hypothesis generation and evolution

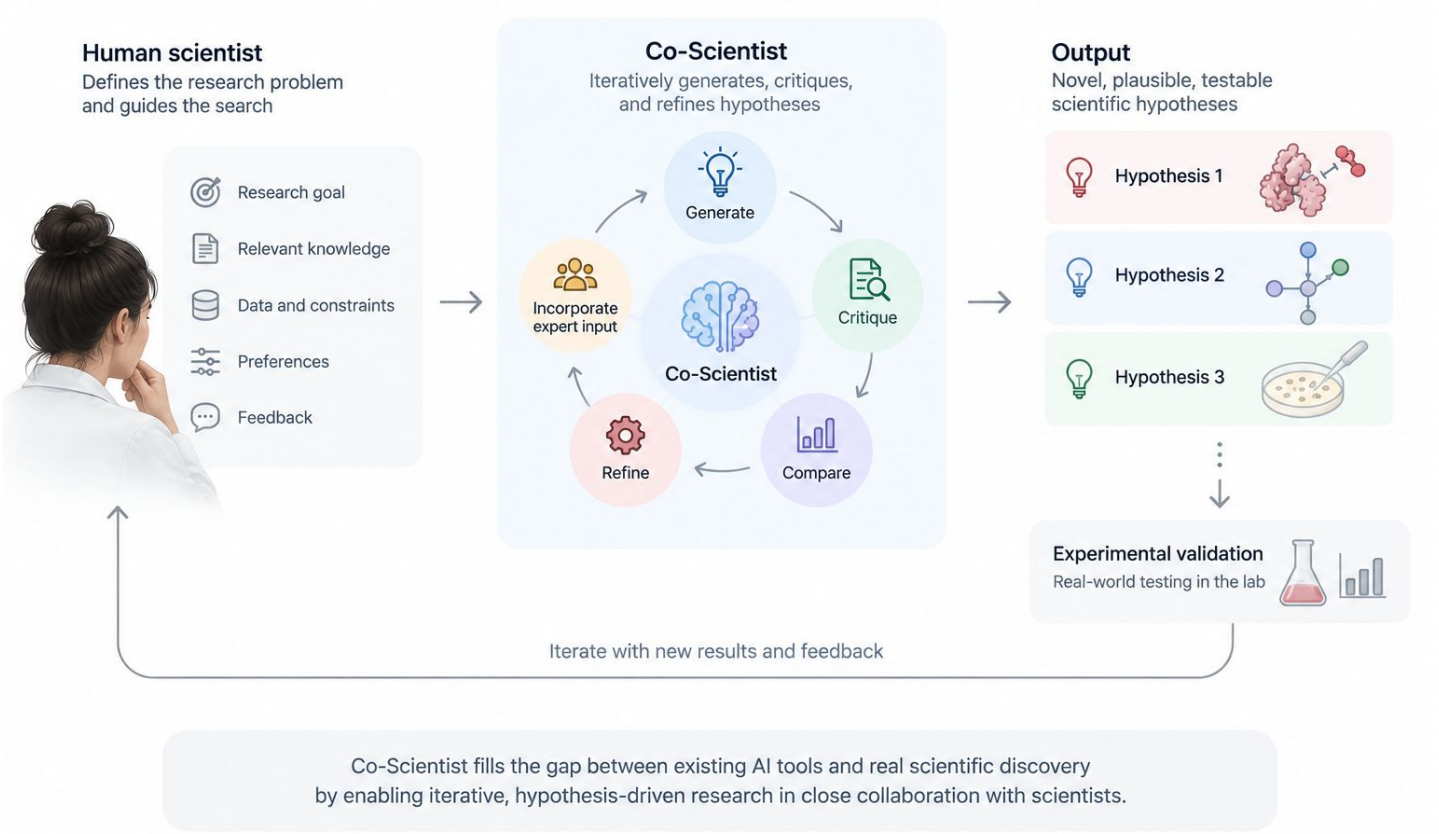
Co-Scientist (this work)

A structured, multi-agent system for scientific discovery



Key difference
Iteratively generate, evaluate, and improve novel, testable scientific hypotheses

The Gap Co-Scientist Tries to Fill



Co-Scientist fills the gap between existing AI tools and real scientific discovery by enabling iterative, hypothesis-driven research in close collaboration with scientists.

Inside the Co-Scientist Loop



Scientist

Goal, constraints and feedback



Supervisor

Research plan and task scheduling



Worker pool

Specialized agents run asynchronously



Shared memory: ideas and feedback

Ideas, feedback and comparison scores carry forward to the next task

Six Agents Divide the Scientific Work



Generation

Propose hypotheses



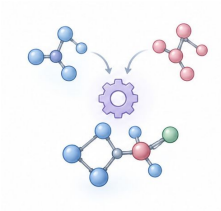
Reflection

Check evidence and originality



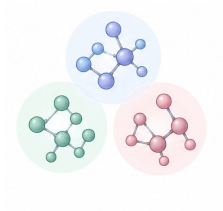
Ranking

Compare ideas and keep score



Evolution

Revise or combine ideas



Proximity

Group related proposals

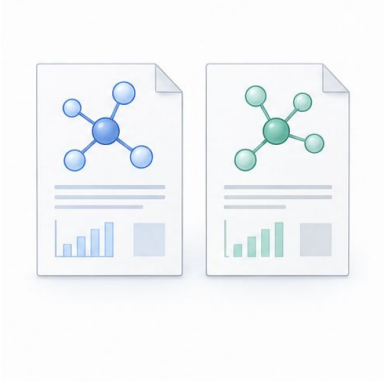


Meta-review

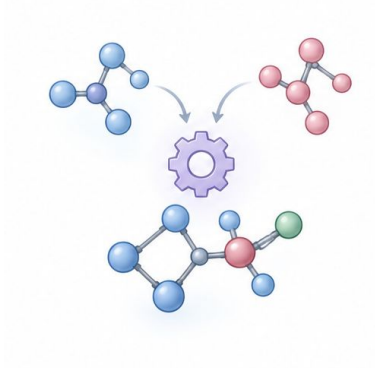
Feed back recurring lessons

Specialized prompts organize work around a shared hypothesis pool

Debate and Evolution Drive the Search







Compare

Original, plausible
and testable in the lab

Rank by debate

Winner, critique
and preference score

Evolve

Revise or combine ideas
Keep the originals

Re-enter the tournament

Elo tracks wins in AI-judged comparisons. Experiments check whether ideas work

More Compute Extends the Search



More model calls for the same problem

Explore	More alternatives
Review	More evidence checks
Compare	More debate matches
Refine	More candidate revisions

Recurring lessons guide later model calls

Ideas improve and feedback accumulates. The model weights stay fixed

What Do the Component Tests Measure?

Baseline vs component-assisted versions

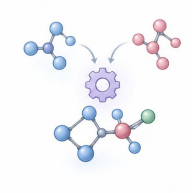


Reflection + search

Recognize already-known ideas
AI novelty rating, 0–10

6.14 → 2.38

Search avoids false novelty
Lower is better for known ideas



Evolution

GPQA: hard science questions
% matching the answer key

70.9% → 75.4%

Refined vs original answers
+4.5 points in a selected subset



Meta-review

Spot correct vs wrong answers
Review AUC (0.5 = chance)

0.521 → 0.597

Higher = better separation
1.0 would be perfect

These tests probe the design. They do not measure real-world discovery success

Evaluation Tests Three Different Claims



Scaling

203 goals

Better ideas with more search?
AI judges compare ideas



Quality

15 goals / 11 rated

Original and useful ideas?
Experts see no model names



Discovery

3 case studies

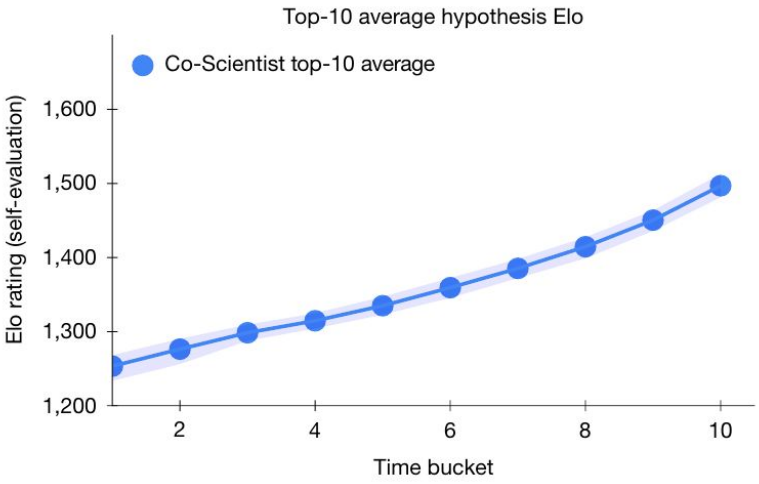
Do proposed ideas work?
Cell tests and gene transfer

AI preference, human judgment and lab evidence answer different questions

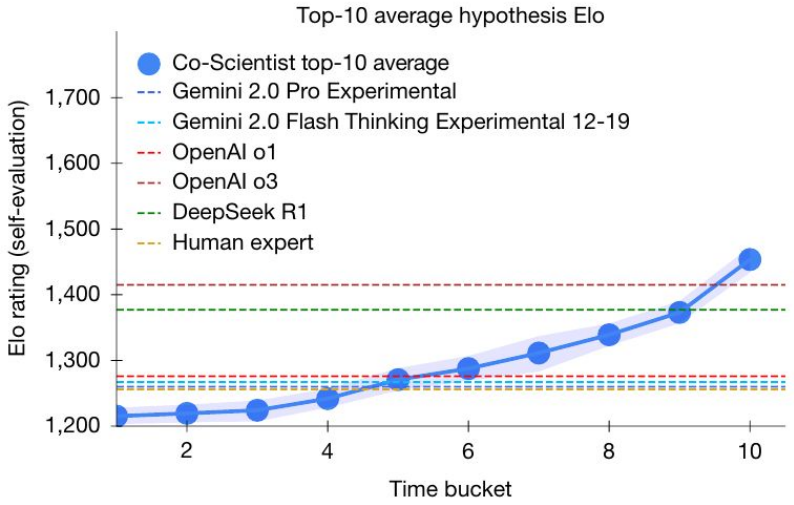
Elo Improves with More Search

Later in a run = more search. Higher Elo = stronger preference from an AI judge

203 research goals



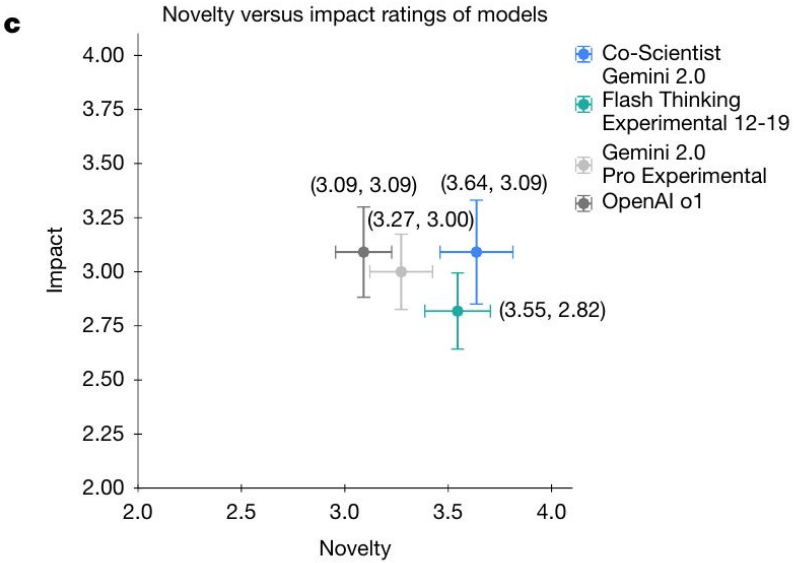
15 expert-curated biomedical goals



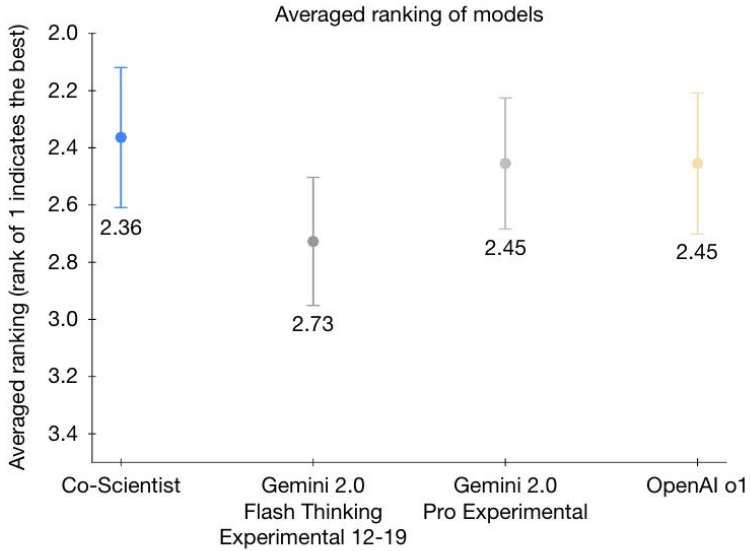
More search earns the AI judge’s approval. Experimental success still needs testing

Expert Preferences Offer Modest Support

Originality and usefulness
Scores 1–5; higher = better



Which proposal would you choose?
Rank 1 = best, 4 = worst



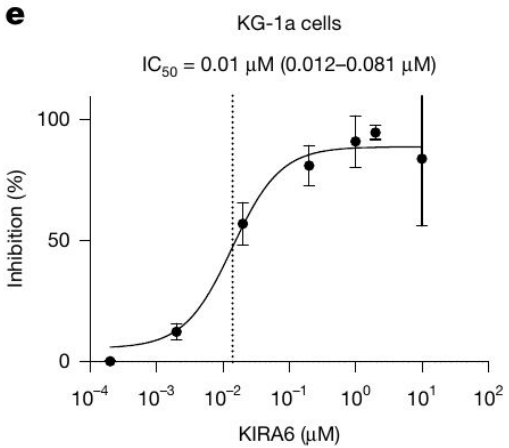
Experts slightly preferred Co-Scientist; usefulness tied o1. Only 11 goals

Testing a Drug Idea for Blood Cancer

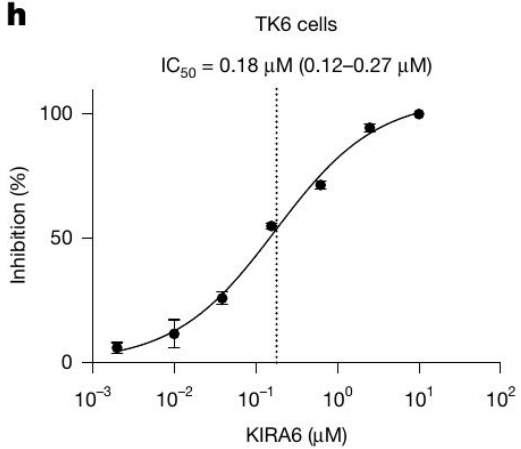
Ranked proposals → **Expert selection** → **Grow cells + add drug**

Readout: living-cell activity after treatment, compared with untreated cells

AML blood-cancer cell model



Non-AML comparison cells



KIRA6

IC50: dose that halves living-cell activity

10 nM

AML model (KG-1a)

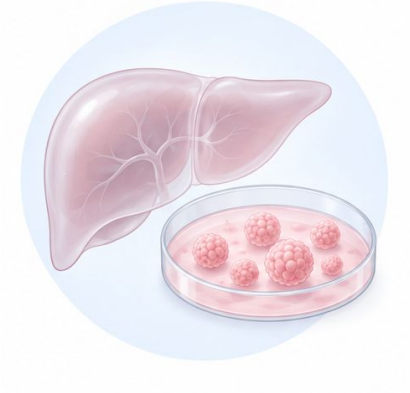
180 nM

Non-AML cells (TK6)

One AML model responded at a lower dose. Patient benefit remains untested

Two More Tests of Scientific Usefulness

Liver fibrosis = excess scarring

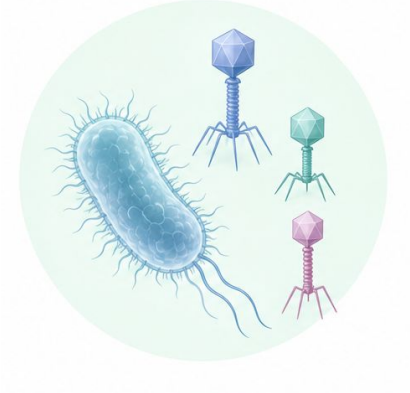


2 of 3

Selected drugs reduced scarring signals

Organoids: lab-grown liver tissue
Less scarring without observed toxicity

How can resistance genes spread?



2 days

To propose an explanation

DNA packages use different virus tails to reach different bacterial species

Lab tissue activity and an independently supported mechanism offer reality checks

What the Paper Establishes



Strengths

Concrete search architecture
Component tests show what helps

Evidence beyond text judgments
Some chosen ideas worked in lab models

Scientists steer the process
Experts frame goals and select experiments



Limits on the claims

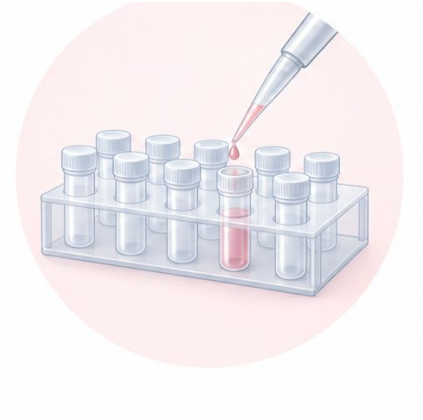
Highly rated ideas can still be wrong
The AI judge may reward its own preferences

Small, selected validation
11 expert-rated goals, 3 biomedical cases

Extra compute may explain gains
Unequal compute, full source unavailable

The strongest evidence supports useful collaboration on selected scientific tasks

Open Questions and Better Tests



Better teamwork, or just more attempts?

Same model-call budget and sources for each method

Does Elo predict scientific value?

Test high- and low-ranked ideas without model labels

Can claims be traced to evidence?

Citations to exact evidence, including failed tests

Does the loop learn from the lab?

Test other fields and use lab results in the next search

Measure useful discoveries per total compute, lab effort and scientist time

What Did This Paper Teach Us?



Quality

Critique gives search direction

Check sources, then revise



Scaling

Later attempts build on earlier feedback

AI preference rises with search



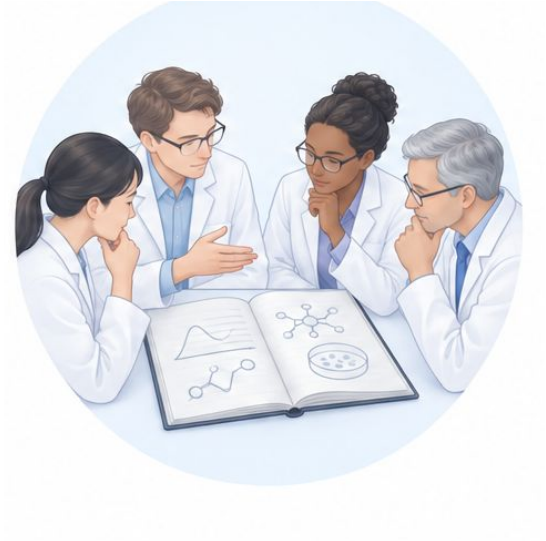
Discovery

Experiments anchor the claim

Some chosen ideas worked in lab

Scientific value depends on the whole workflow, including the experiment

Discussion: What Should We Optimize?



You can run only 20 experiments

Top AI scores, a range of ideas,
or tests that teach us the most?

What would convince you it accelerates discovery?

Choose a baseline, an outcome and a fair budget

Backup: What Does Elo Measure?

Like chess ratings: winning comparisons raises a proposal's relative score



$$E(A) = \frac{1}{1 + 10^{(R(B) - R(A)) / 400}}$$

$$R'(A) = R(A) + K \times [S(A) - E(A)]$$

R: proposal rating E: expected win chance
All new proposals start at 1200

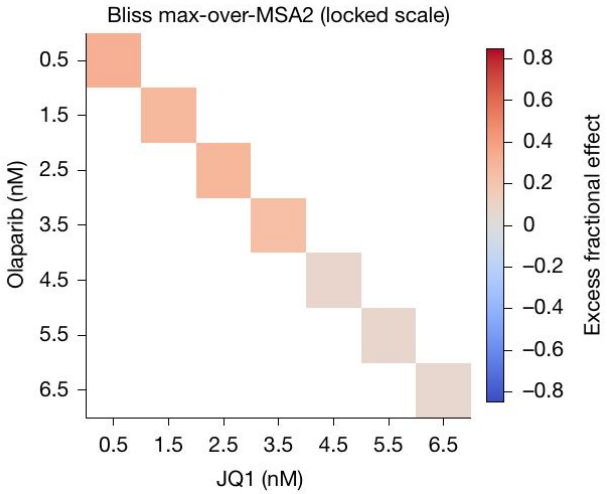
S: observed outcome K: update step size

Winning comparisons does not mean a hypothesis is likely to be scientifically true

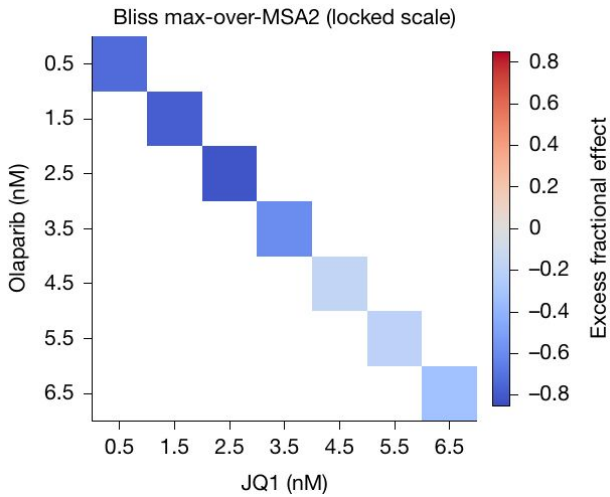
Backup: Testing Drug Combinations

Same three drugs in two AML cell models

MOLM-13 cell model



KG-1a cell model



Red: more inhibition than expected. Blue: less.

Same mix: extra inhibition in one model, less than expected in the other