PAPER PRESENTATION · NATURE 655, 497–505 (2026)

A multi-agent system for automating scientific discovery

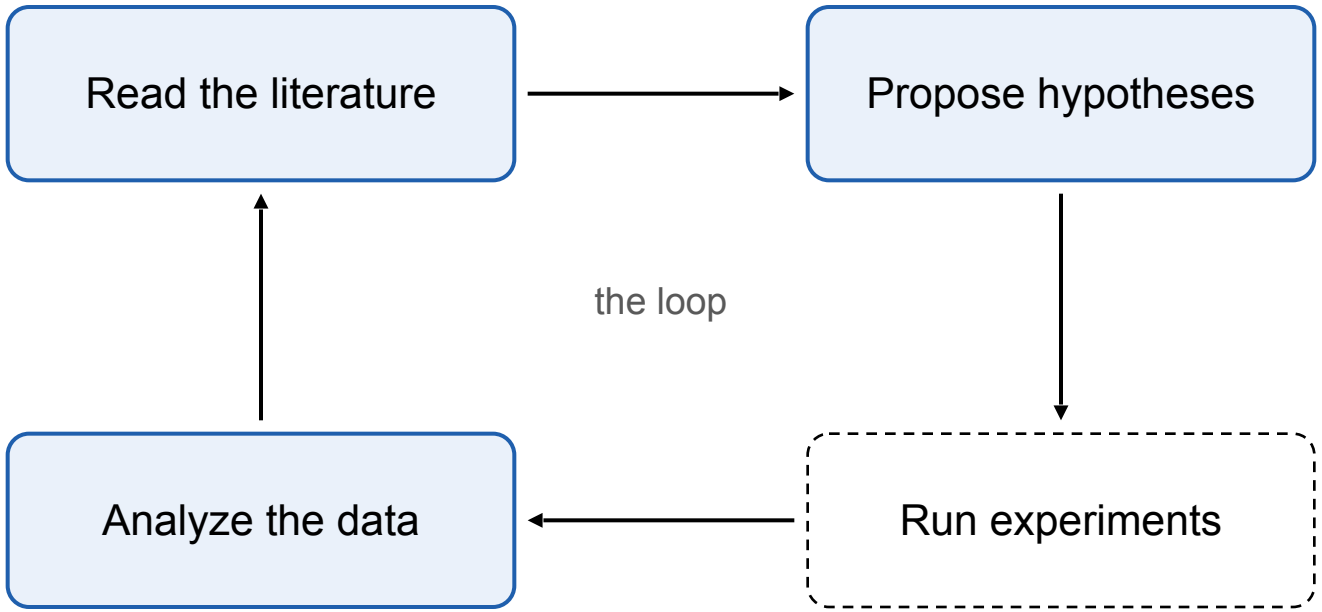
Robin: an AI that proposed, analyzed and refined drug candidates for dry macular degeneration

Ghareeb, Chang, Mitchener, … White, Finnemann, Hinks & Rodriques

FutureHouse · Fordham University · University of Oxford

Clemens and Linrong · COS 597Y

Can one AI run the whole research loop?



Earlier AI tools automate one step. Robin claims to run all the shaded ones.

The answer was already in published papers



The hearing protection follows directly from blocking BRAF. Nobody connected the facts.

Ten-year delays are not unusual



GLP-1: discovered in the Gila Monster

Could an AI that reads everything connect such facts sooner?

Roadmap: two parts, two presenters

PART 1 · HOW ROBIN WORKS

Clemens

1 Motivation
Why reading at scale could speed up discovery

2 The system
Three agents, a tournament, an analysis ensemble

3 The first experiment
Dry AMD, the lab test, and the first working drug

PART 2 · RESULTS AND RELIABILITY

Linrong

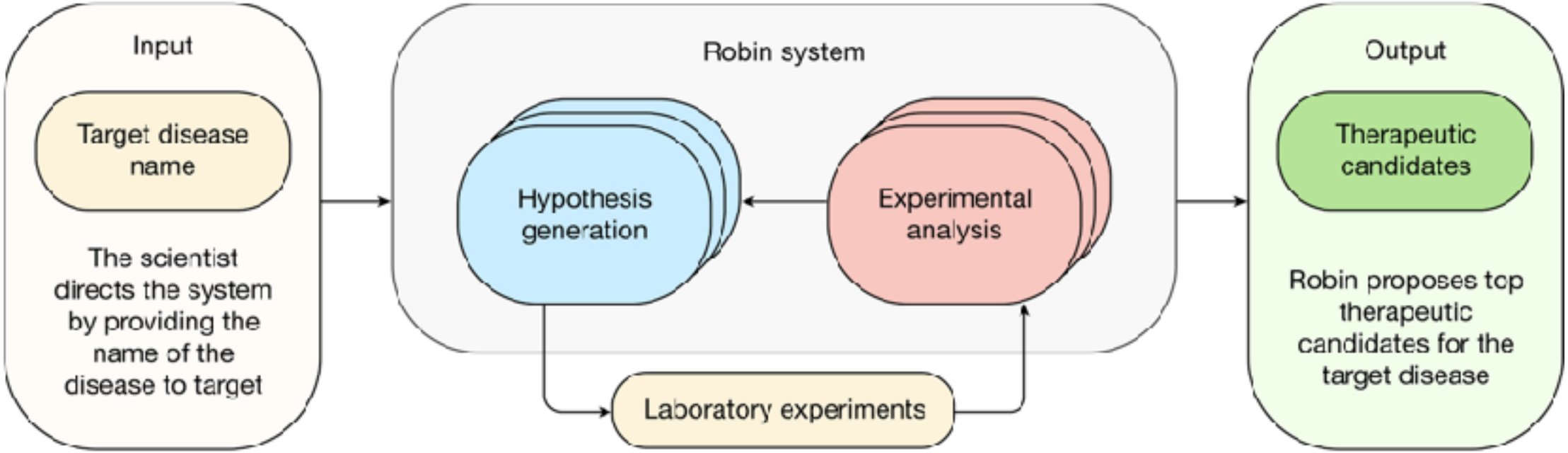
4 Robin's follow-up
Why it works, round 2, tests in human cells

5 Does Robin's design matter?
Removing parts, Finch's accuracy, a comparison, cost

6 Our assessment
What was automated, limits, takeaways

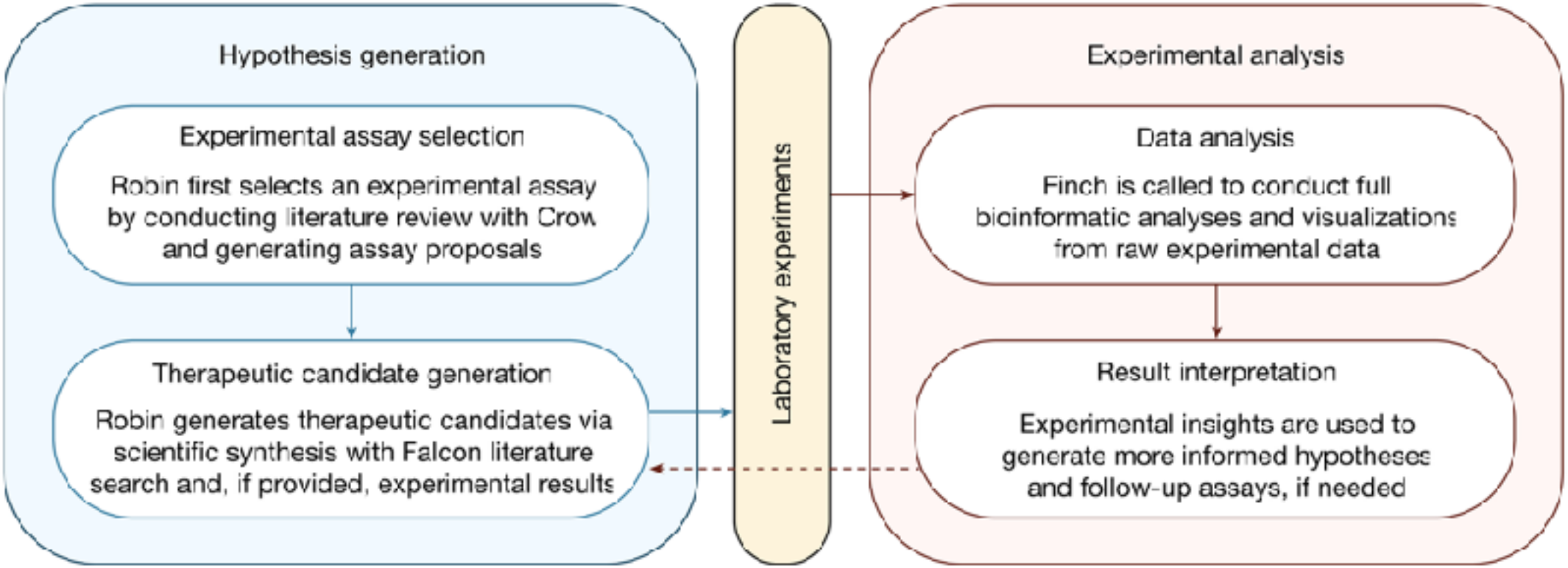
Then a discussion, led by both of us.

Input: a disease. Output: drug candidates



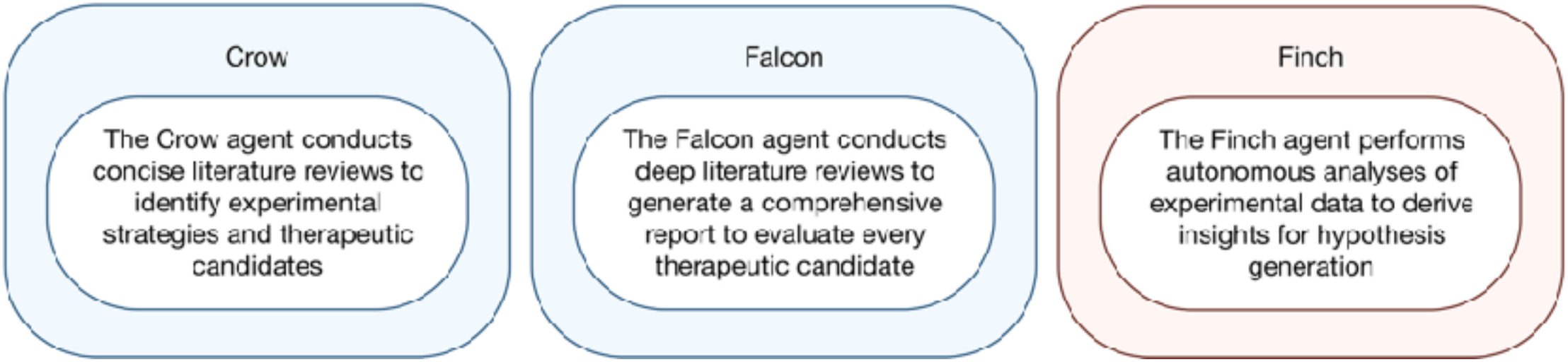
Scientists give the disease and run the lab work. Robin does the thinking in between.

Two halves, connected by lab experiments



Literature → hypotheses → lab → analysis → better hypotheses

Three specialist agents



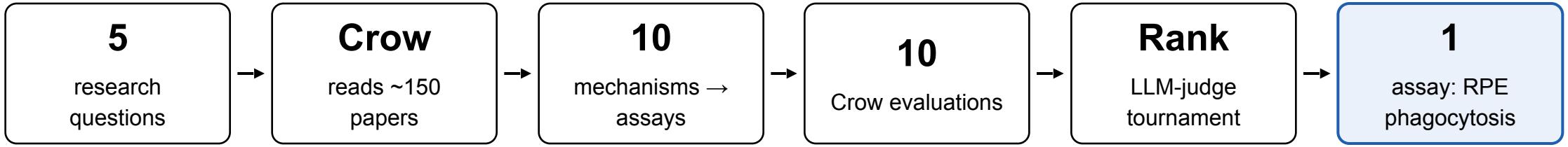
Crow and Falcon search the literature (built on PaperQA2), so claims come with real citations.

Finch writes and runs Python or R in a Jupyter notebook.

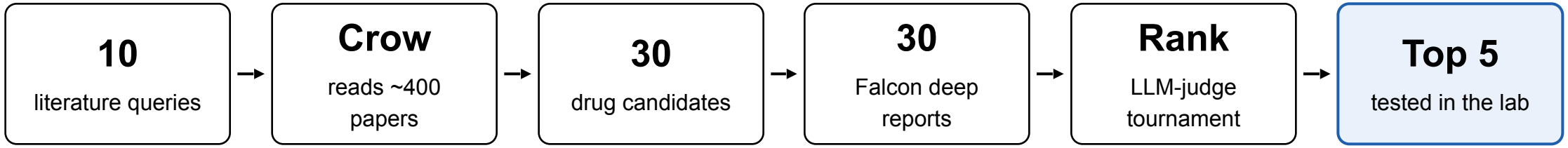
Ghareeb et al., Nature 2026 · Fig. 1c

From disease to drug list in two stages

Stage 1 · Choose a lab model of the disease

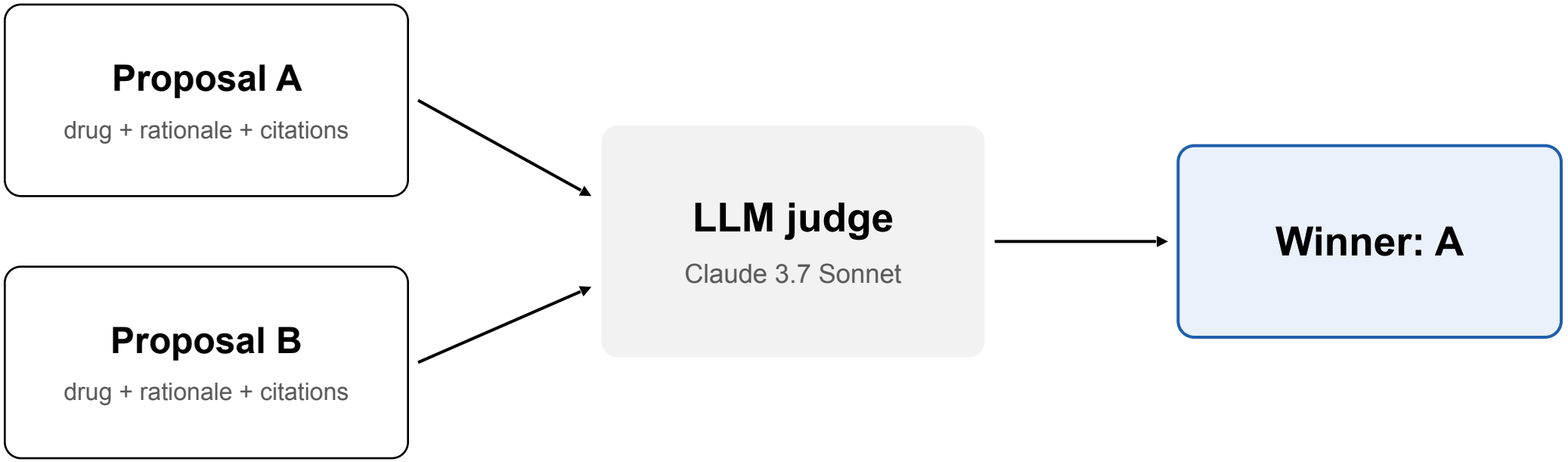


Stage 2 · Find drugs to test in it



Same recipe twice: search → propose many → evaluate each → rank → keep the best

A judge compares two ideas at a time



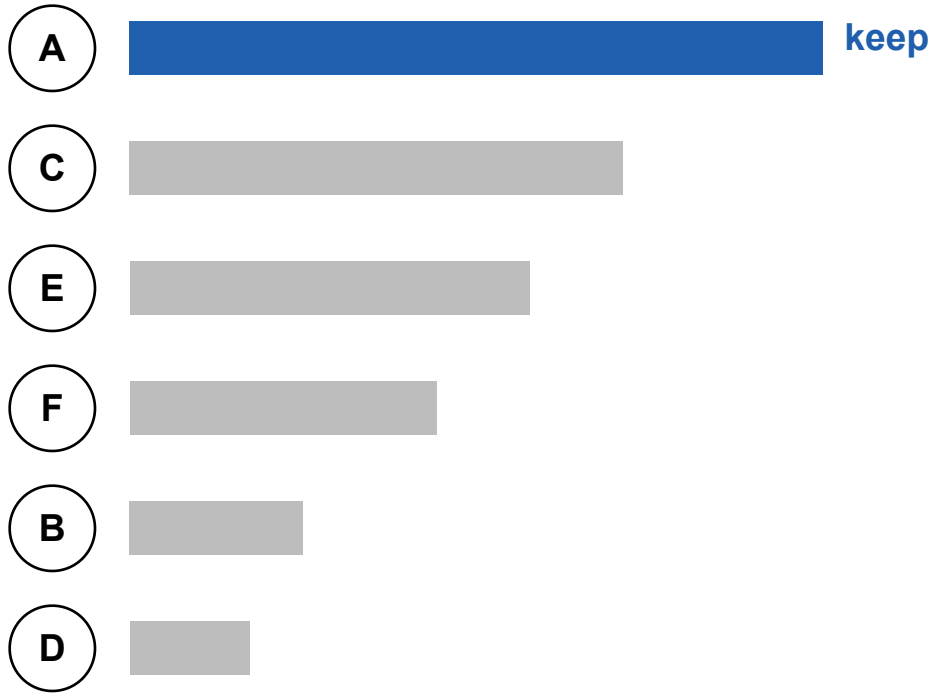
Up to 25 ideas: every pair is compared. More than 25: 300 random pairs.

Wins become a strength score

$$P(i \text{ beats } j) = s(i) / [\, s(i) + s(j) \,]$$

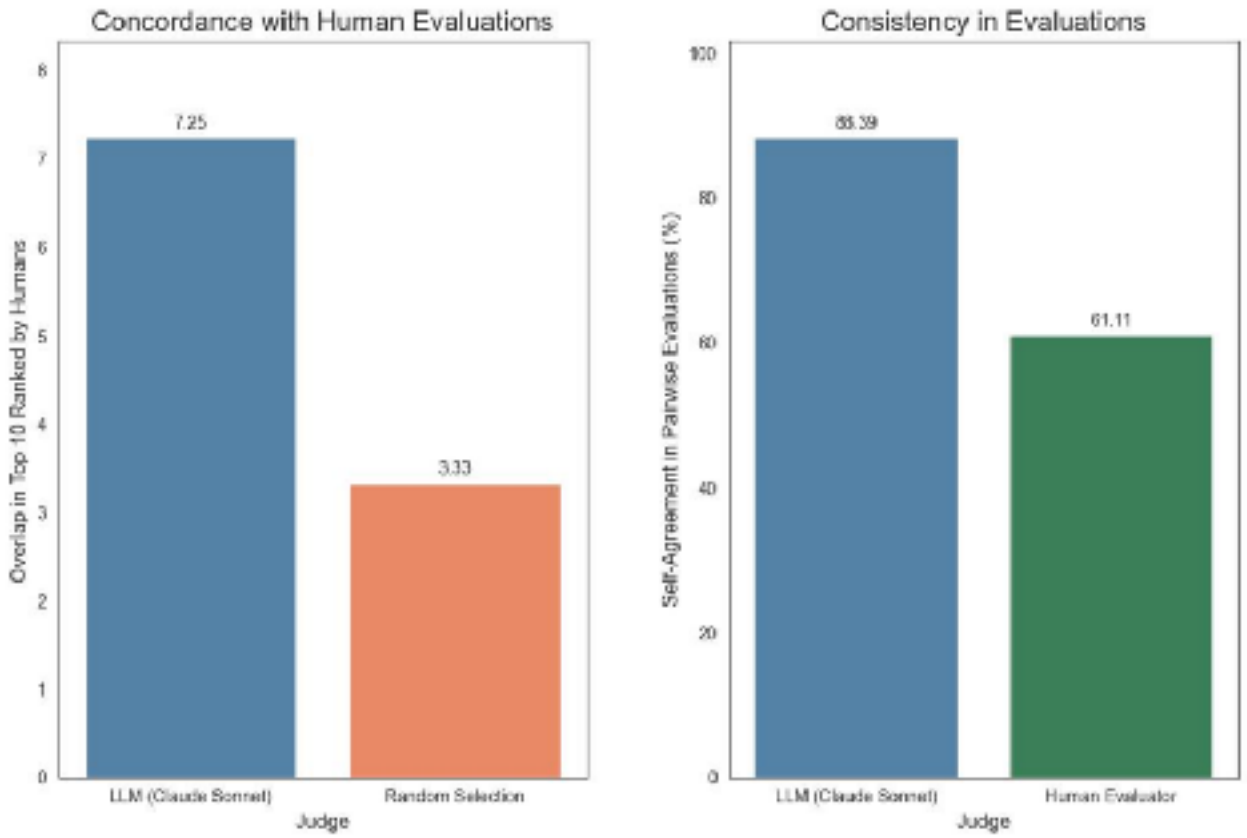
Bradley–Terry–Luce model

Fit one strength s per idea from all the match results, then rank by s.



(illustration)

The judge agrees with experts



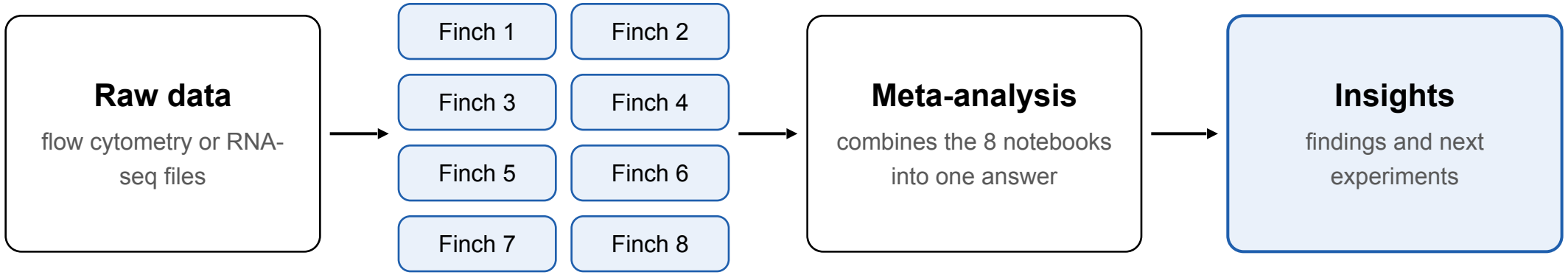
7.25 / 10

of the judge's top 10 match the experts' top 10 (random: 3.33)

88% vs 61%

same choice when shown a pair again: judge vs human experts

Eight separate analyses, one combined answer



Why eight? Analysis involves judgment calls, and LLM runs vary.

A consensus is more reliable than any single run.

Test case: dry macular degeneration

#1

cause of irreversible sight loss in developed countries

1.5 M

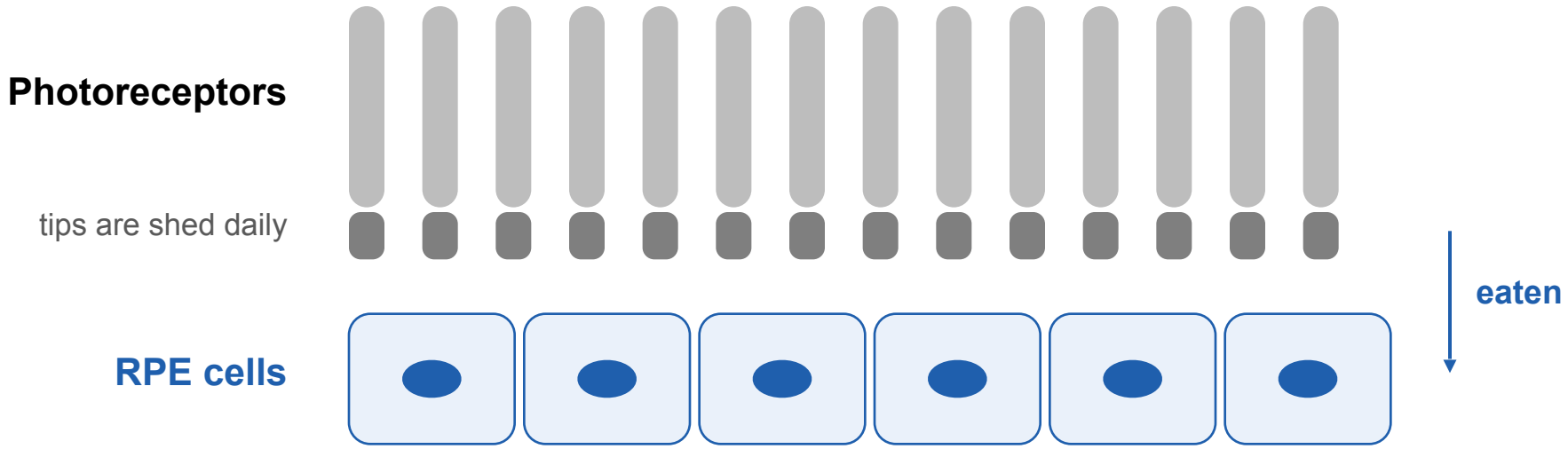
people in the US with vision-threatening dry AMD

~3×

projected rise in AMD blindness by 2050

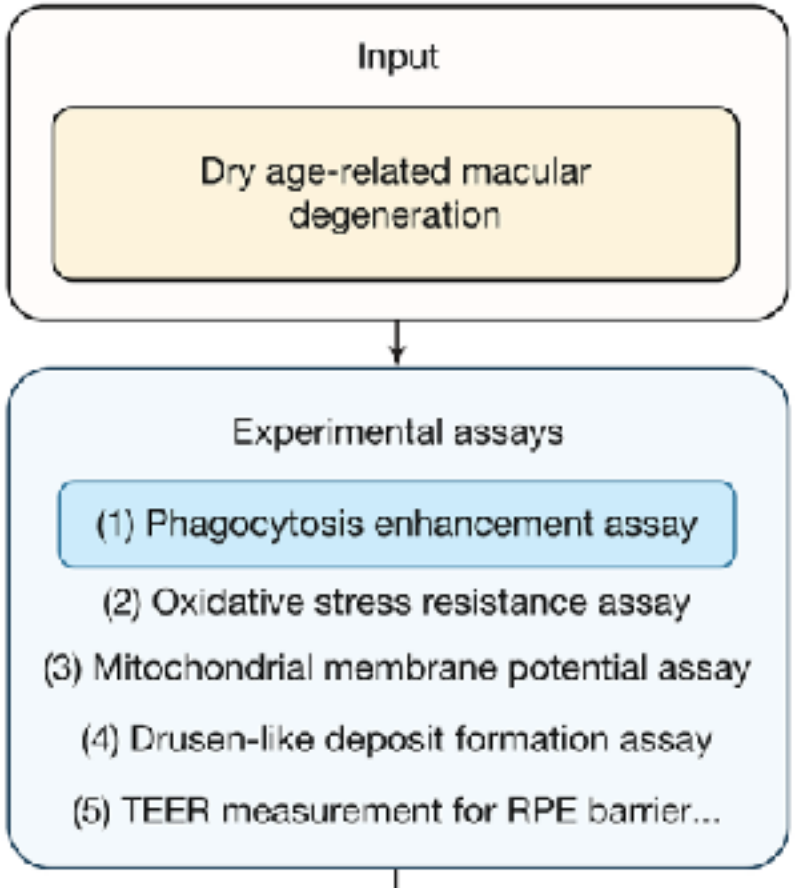
And few treatments exist for the dry form.

RPE cells clean up after photoreceptors



This daily clean-up (phagocytosis) declines with age and is impaired in AMD.

Robin's choice: help RPE cells eat more



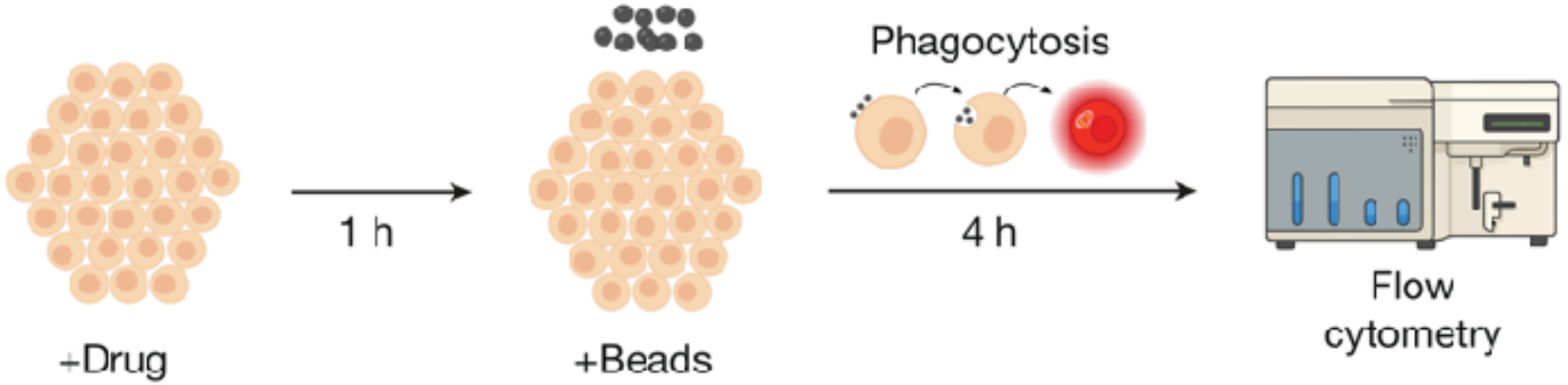
Robin proposed 10 assays, each testing a different disease mechanism.

A tournament picked #1: the phagocytosis assay.

Excerpt of Robin's own output

Ghareeb et al., Nature 2026 · Fig. 2a

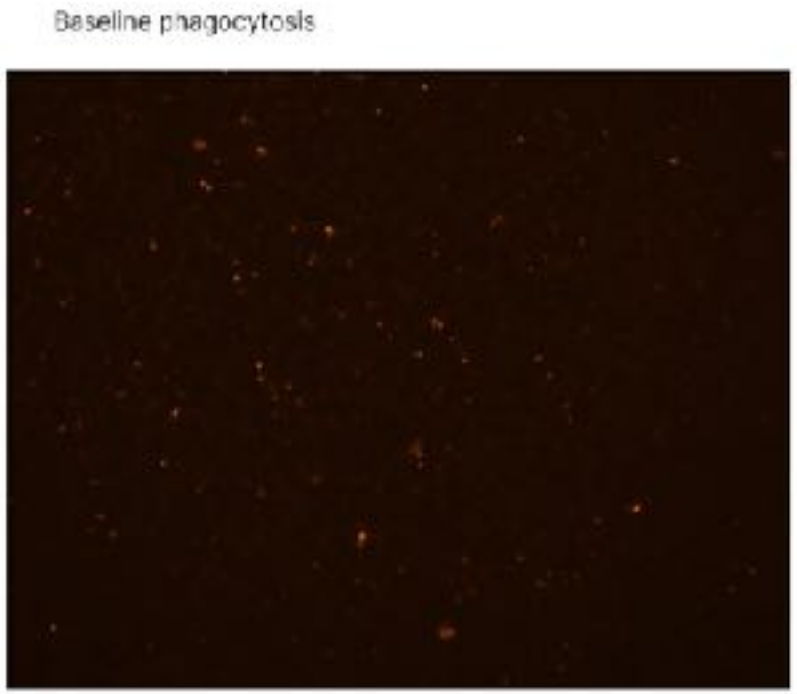
The assay: count the beads cells eat



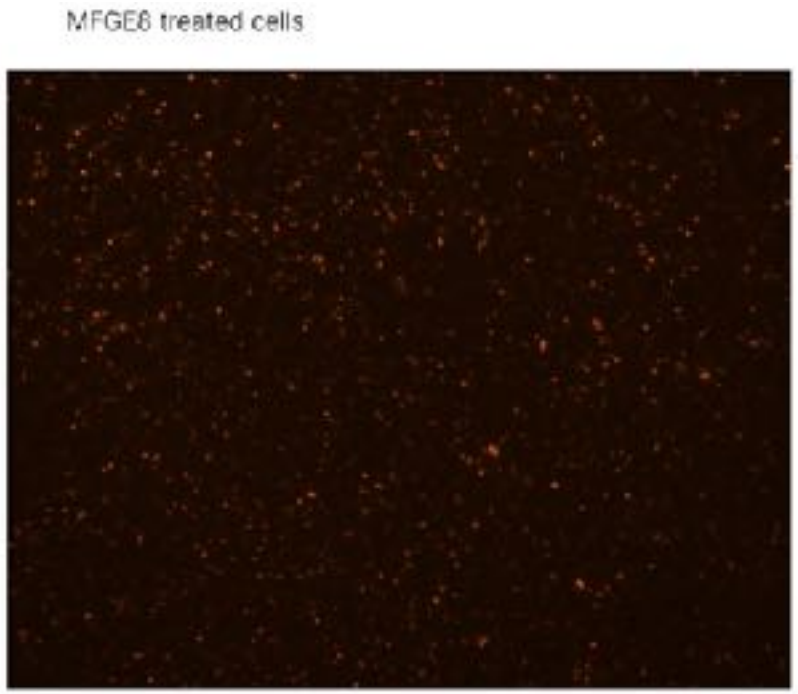
Beads glow only inside the cell's acidic digestion compartment.

So more glow means more eating. Flow cytometry measures every cell.

Check: a known enhancer makes cells eat more



Untreated

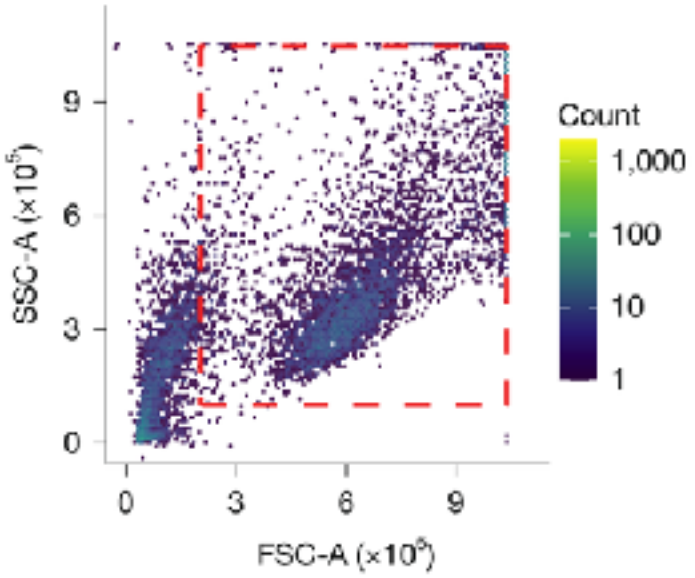


MFGE8, a known booster

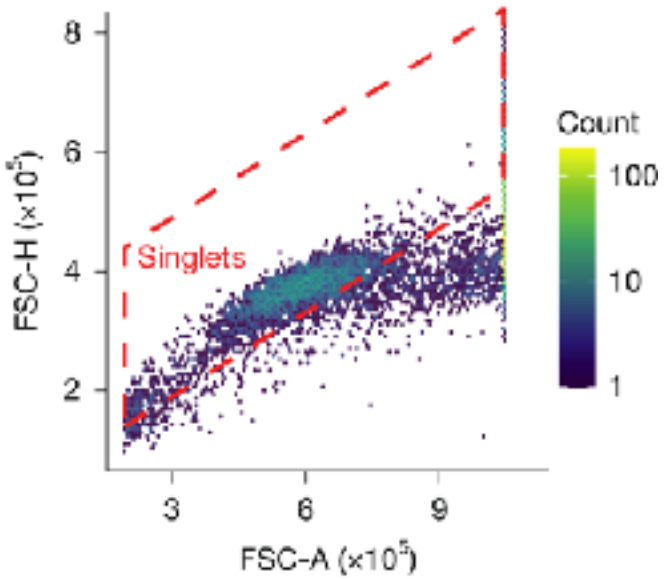
Ghareeb et al., Nature 2026 · Supp. Fig. 12a

Finch analyzes the raw data like an expert

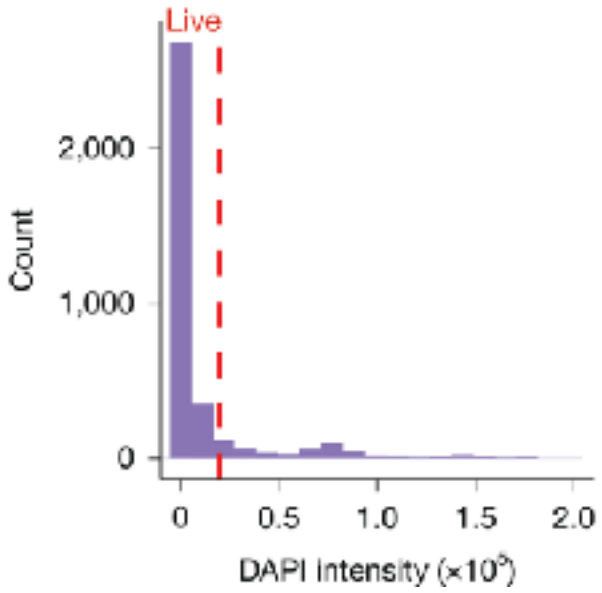
1 · Remove debris



2 · Keep single cells

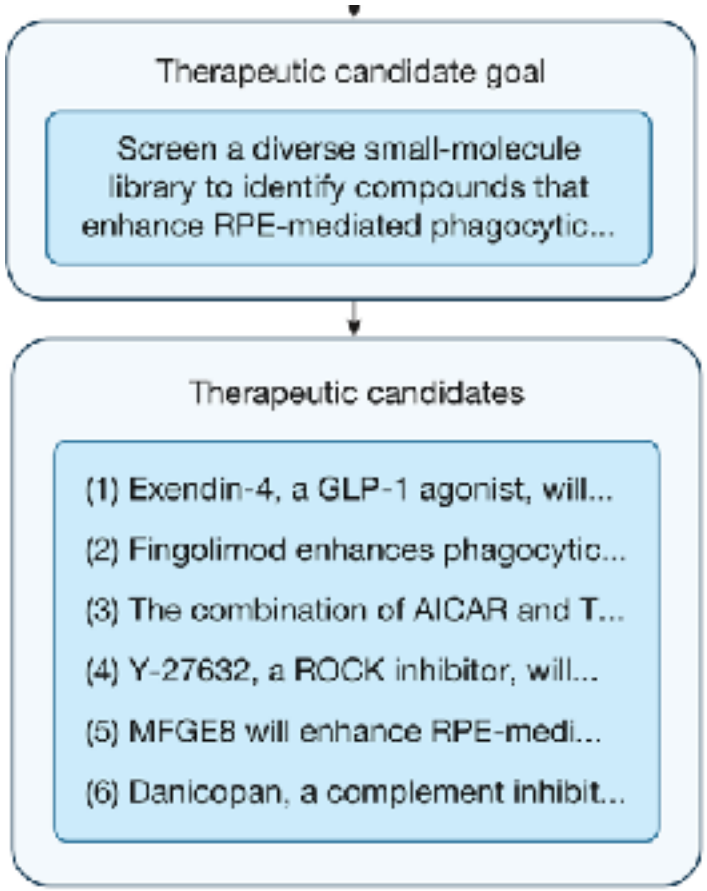


3 · Keep live cells



Then it compares each drug's average glow with the DMSO control.

Round 1: five drugs are tested in the lab



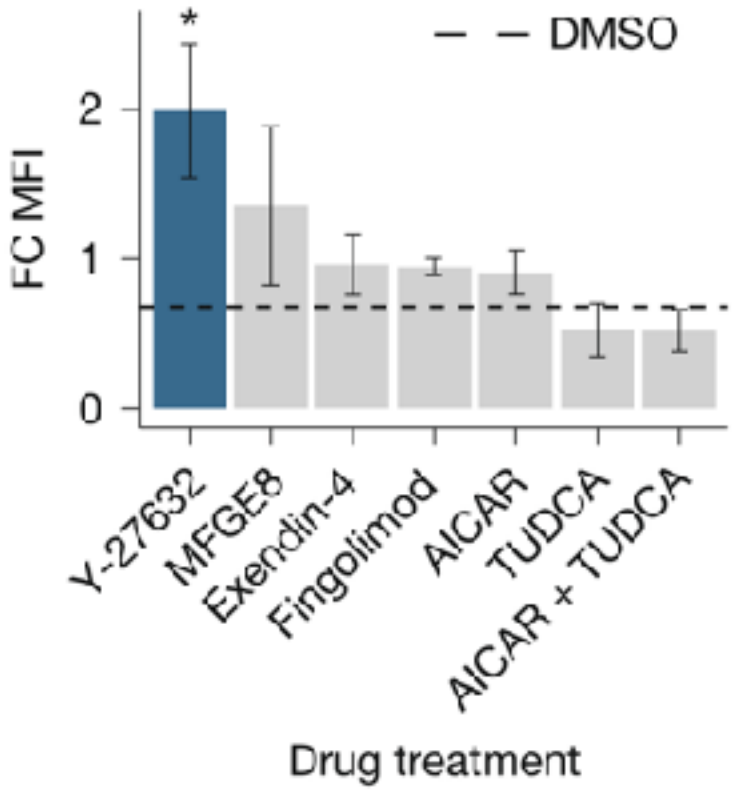
Robin ranked 30 drugs; the scientists tested the top five.

One of them, MFGE8, is a known booster, used as a positive control.

Excerpt of Robin's own output

Ghareeb et al., Nature 2026 · Fig. 2a

Round 1 result: Y-27632 works best



About 2× more eating than control

the only significant candidate (P < 0.05)

Y-27632 blocks ROCK, a kinase that controls the actin skeleton cells use to engulf particles.

Ghareeb et al., Nature 2026 · Fig. 2f

PART 2 OF 2

What Robin found, and is it reliable?

Presented by Linrong

4 Robin's follow-up

5 Does Robin's design matter?

6 Our assessment

Robin asks why, and plans a gene test

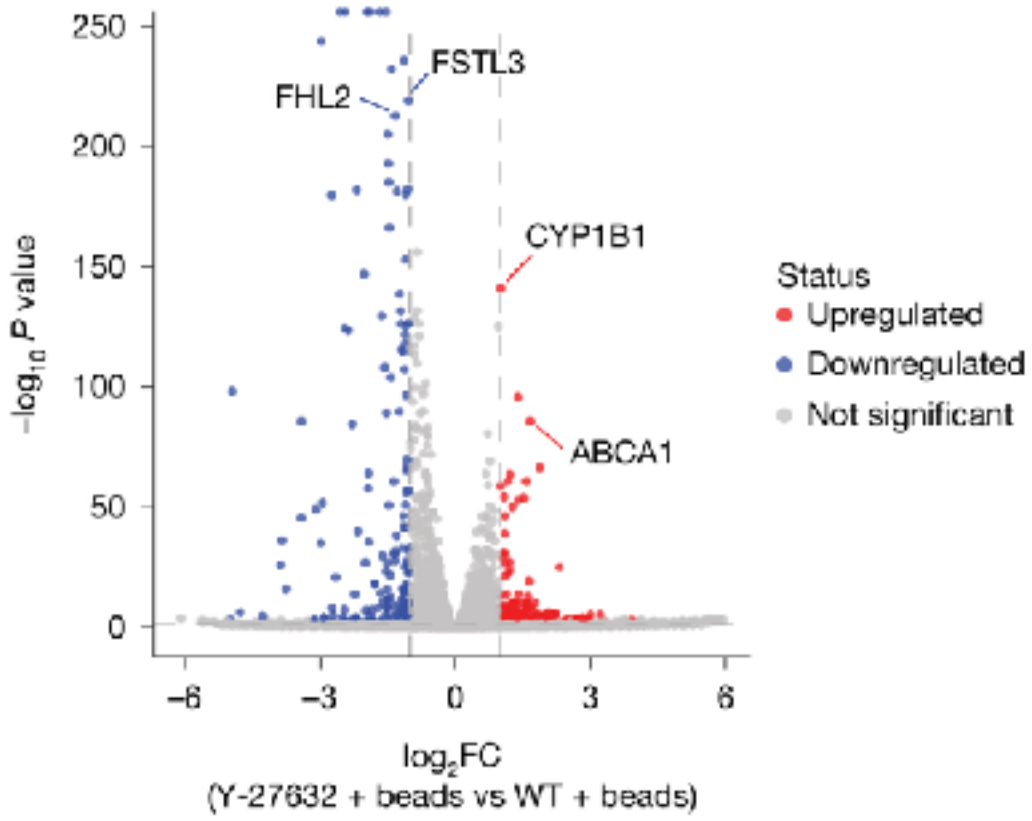


Robin read its own round-1 results and wrote this interpretation.

Its follow-up: measure which genes Y-27632 switches on (RNA-seq).

Ghareeb et al., Nature 2026 · Fig. 3a

A lipid pump, ABCA1, rises 3-fold

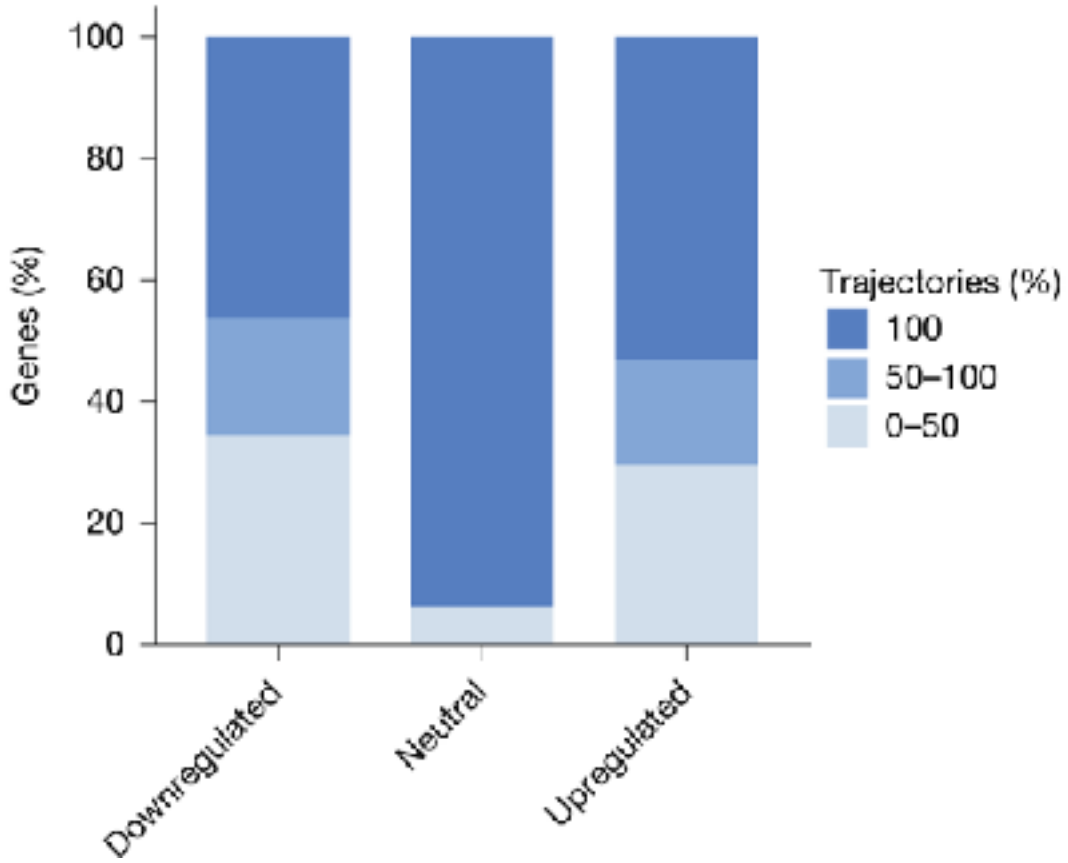


ABCA1 up about 3×

adjusted P = 2.1 × 10^-83

Each dot is a gene: red = up, blue = down after Y-27632.

The eight Finch runs mostly agree

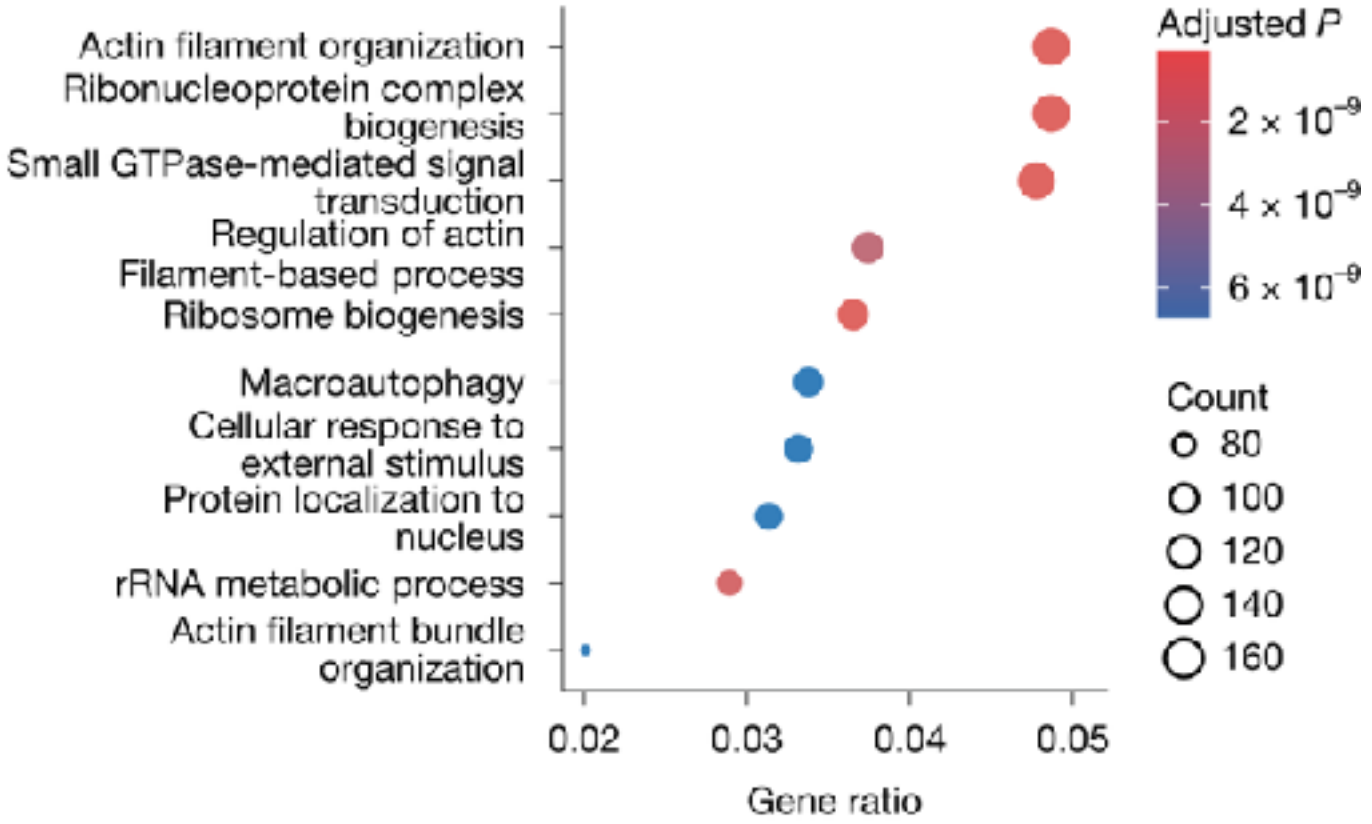


Most significant genes were called by more than half of the 8 Finch runs.

A human analysis gave similar results.

Ghareeb et al., Nature 2026 · Fig. 3c

Other gene groups that changed

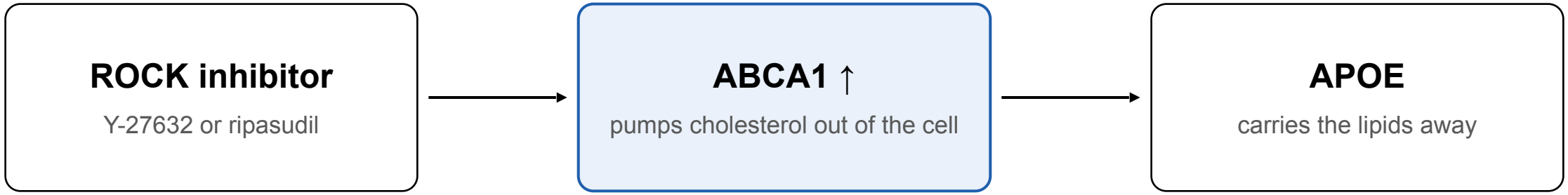


Fits the story: ROCK reshapes the actin skeleton.

Autophagy genes hint at better digestion, not yet tested.

Ghareeb et al., Nature 2026 · Fig. 3d

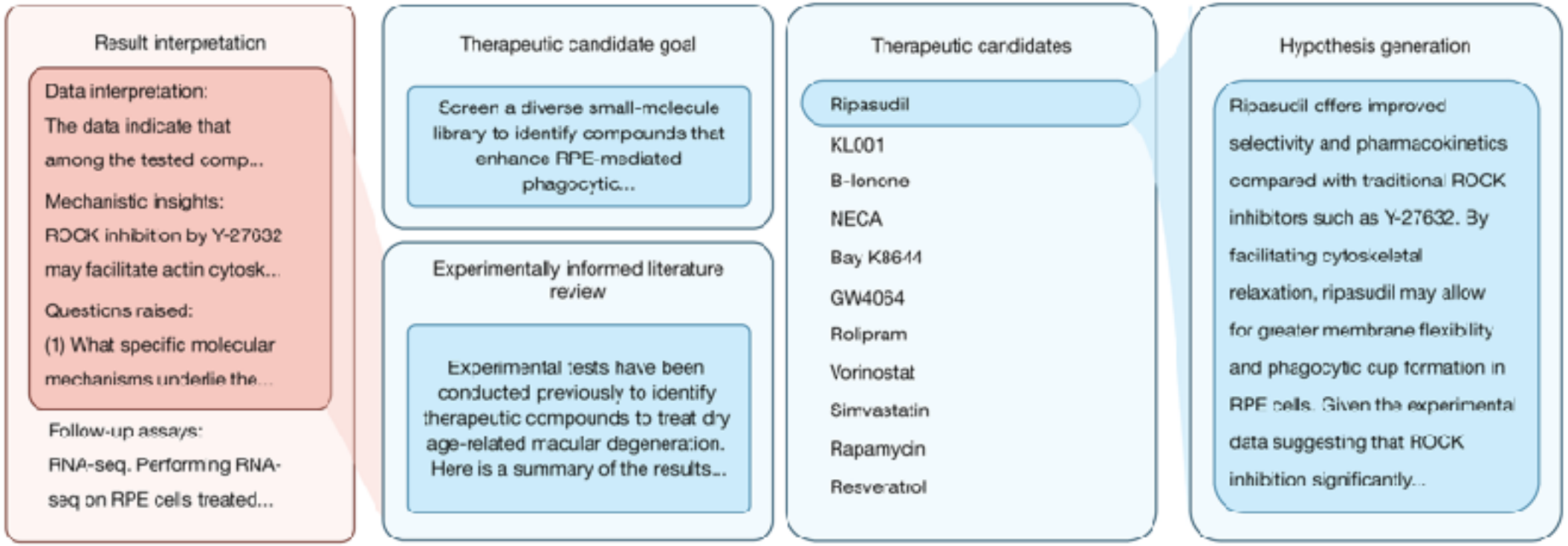
Why ABCA1 matters for AMD



APOE variants are a genetic risk factor for macular degeneration

A possible new drug target, found because Robin asked a follow-up question.

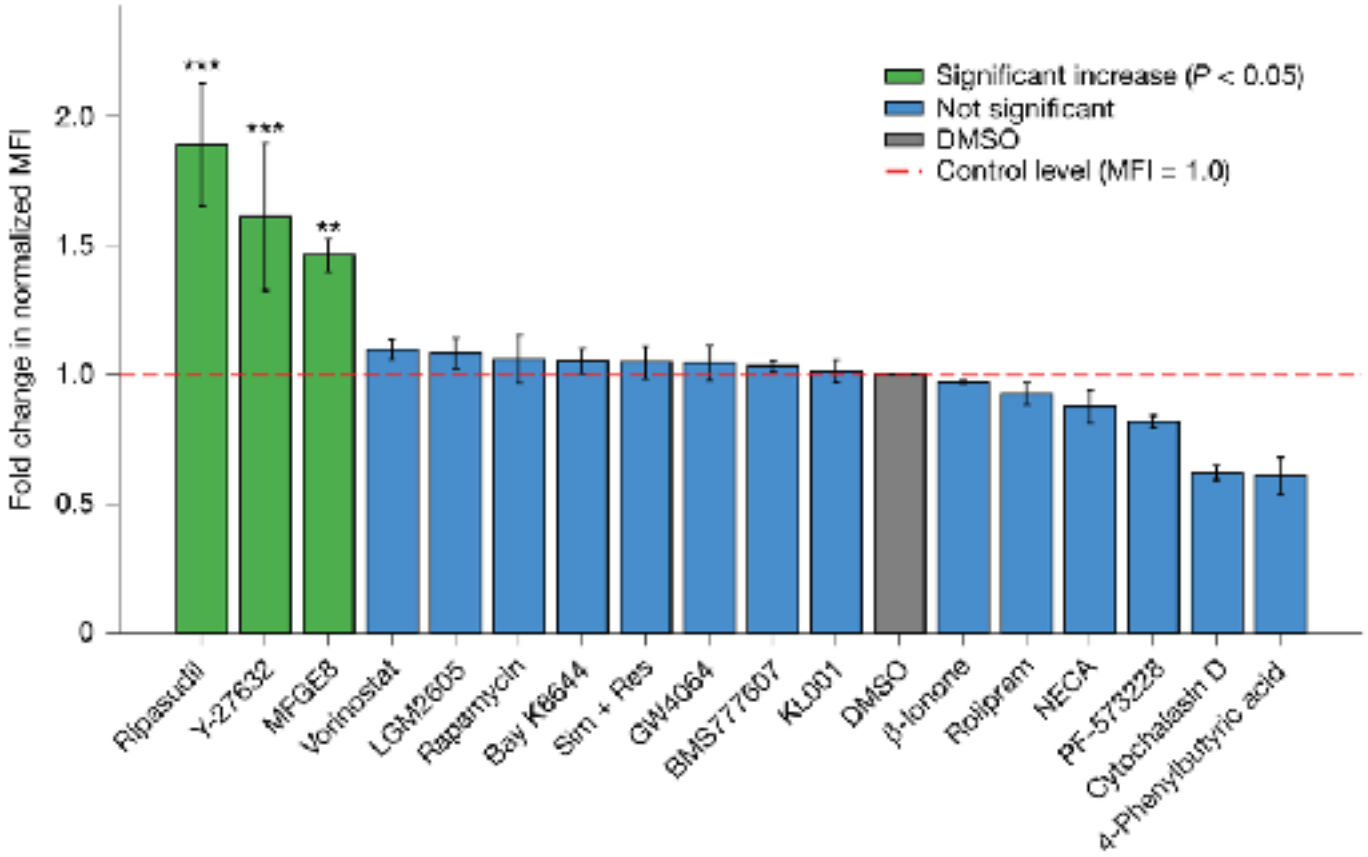
Round 2: Robin uses what it learned



Robin's round-1 results feed a new literature review and a new drug list.

Top pick: ripasudil, a ROCK inhibitor already sold as a glaucoma eye drop in Japan.

Round 2 result: ripasudil works best

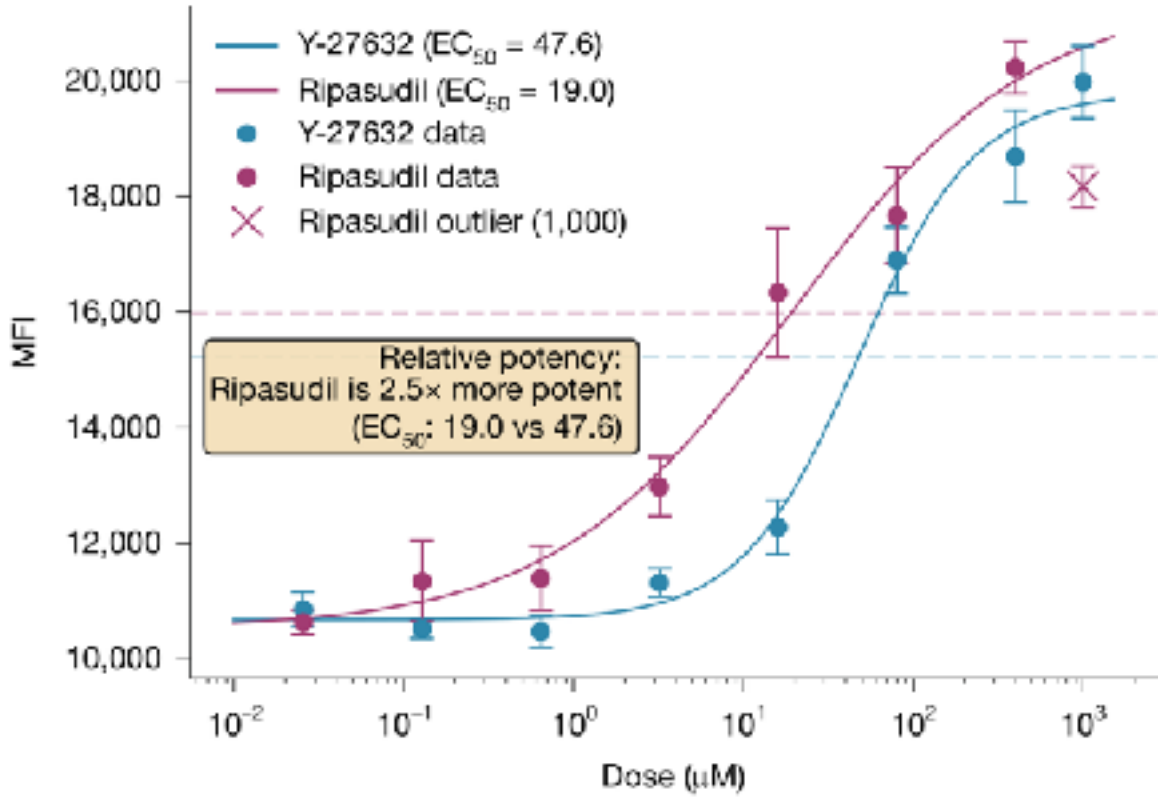


1.89×

more eating than control, the best of all candidates

An approved drug, so its safety in the eye is known.

It also works at lower doses



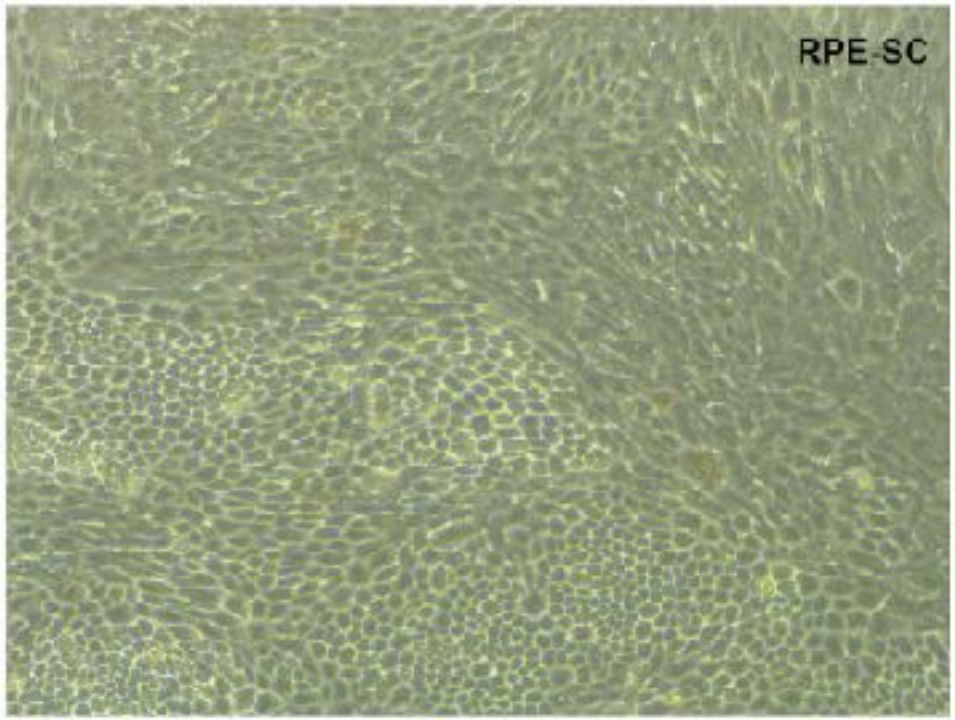
2.5× more potent

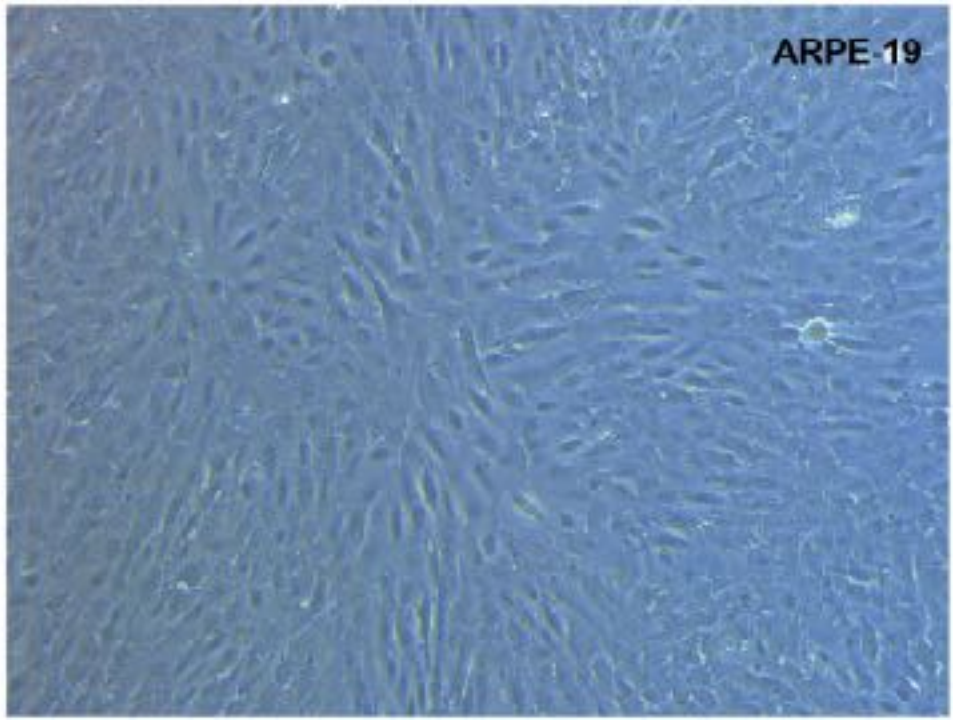
than Y-27632 (EC50: 19.0 vs 47.6 μM)

EC50 = the dose giving half the maximal effect

Ghareeb et al., Nature 2026 · Fig. 4c

Next test: eye cells from an older donor

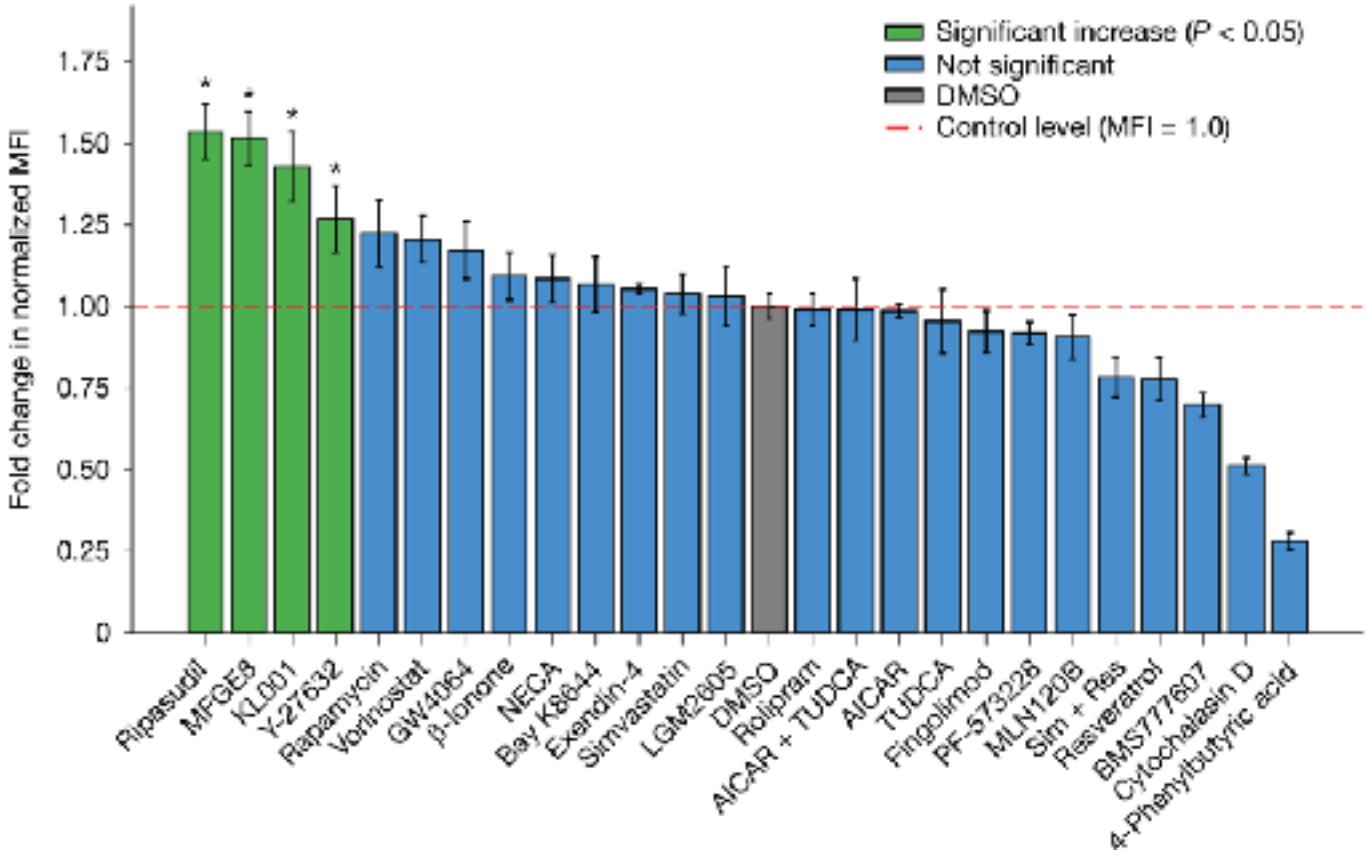




Cells from a donor over 60 (left) look like real RPE; the ARPE-19 line (right) less so.

They are fed real photoreceptor fragments instead of beads.

Ripasudil and KL001 still work



Ripasudil 1.54×

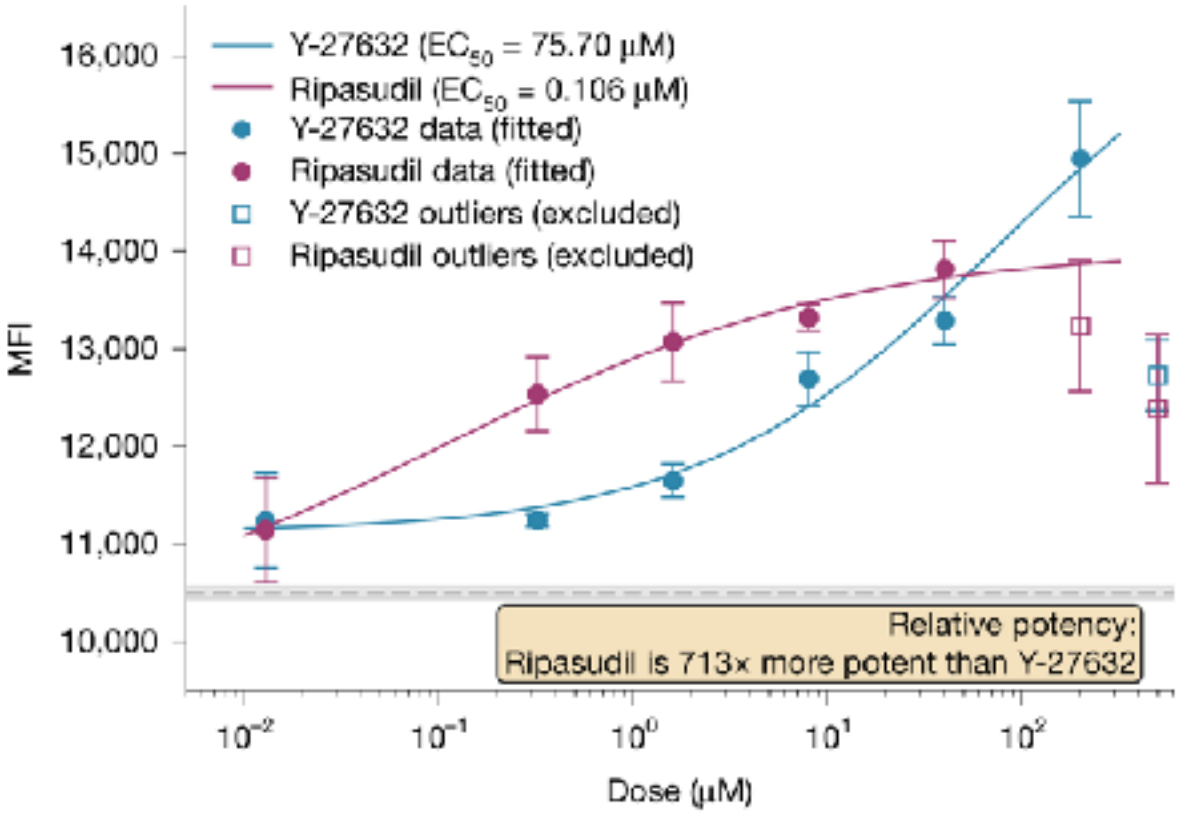
KL001 1.43×

KL001 tunes the circadian clock; it had never been linked to phagocytosis.

Y-27632: 1.27×

Ghareeb et al., Nature 2026 · Fig. 4d

About 700× more potent in aged cells



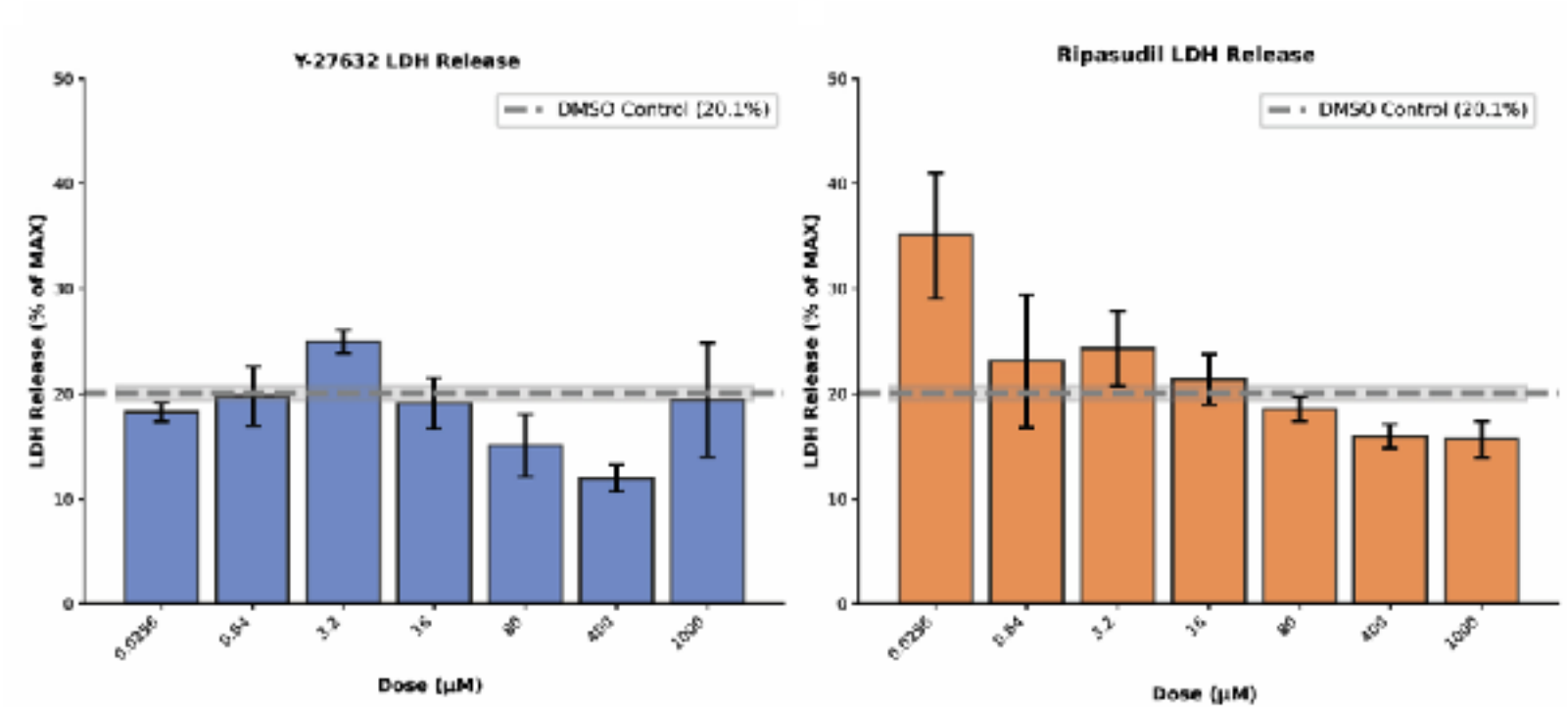
EC50: 0.106 vs 75.7 µM

ripasudil vs Y-27632 in aged human cells

Caveat: Y-27632 reaches a higher maximum at the top dose.

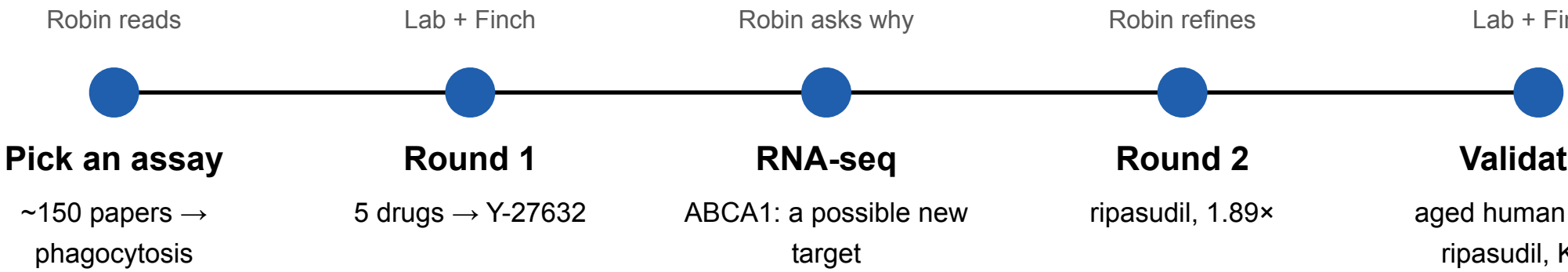
Ghareeb et al., Nature 2026 · Fig. 4e

No sign that the drugs damage cells



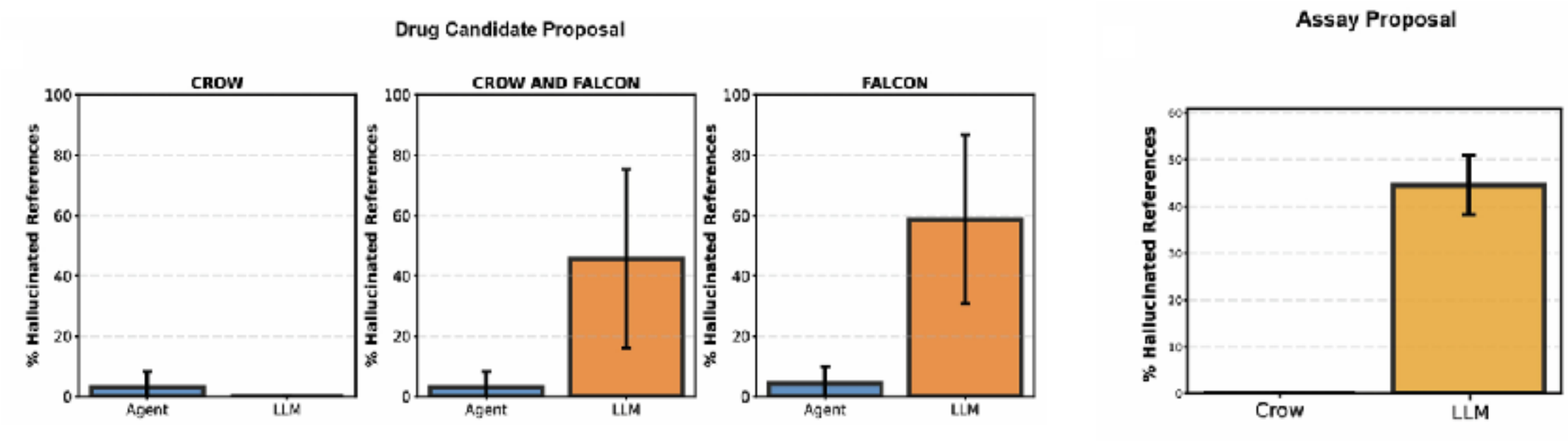
LDH, a cell-damage marker, does not rise with either drug, and falls as the ripasudil dose increases.

Recap: two rounds of the loop



Each round's data changed the next round's ideas. This feedback is the key idea.

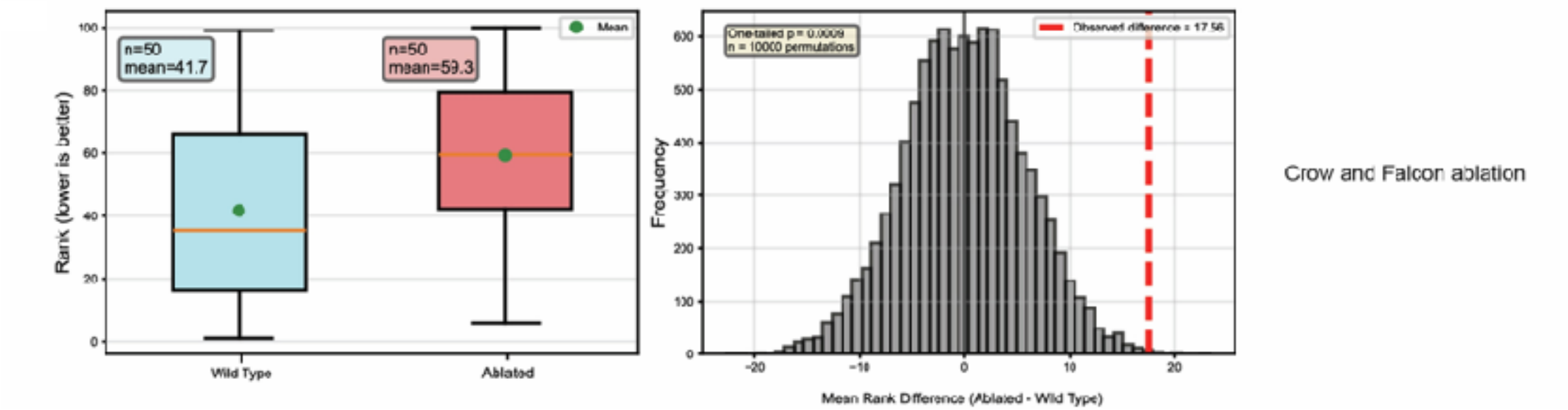
Without the agents, citations are made up



Agent = Robin's literature agent · LLM = plain o4-mini with the same prompt

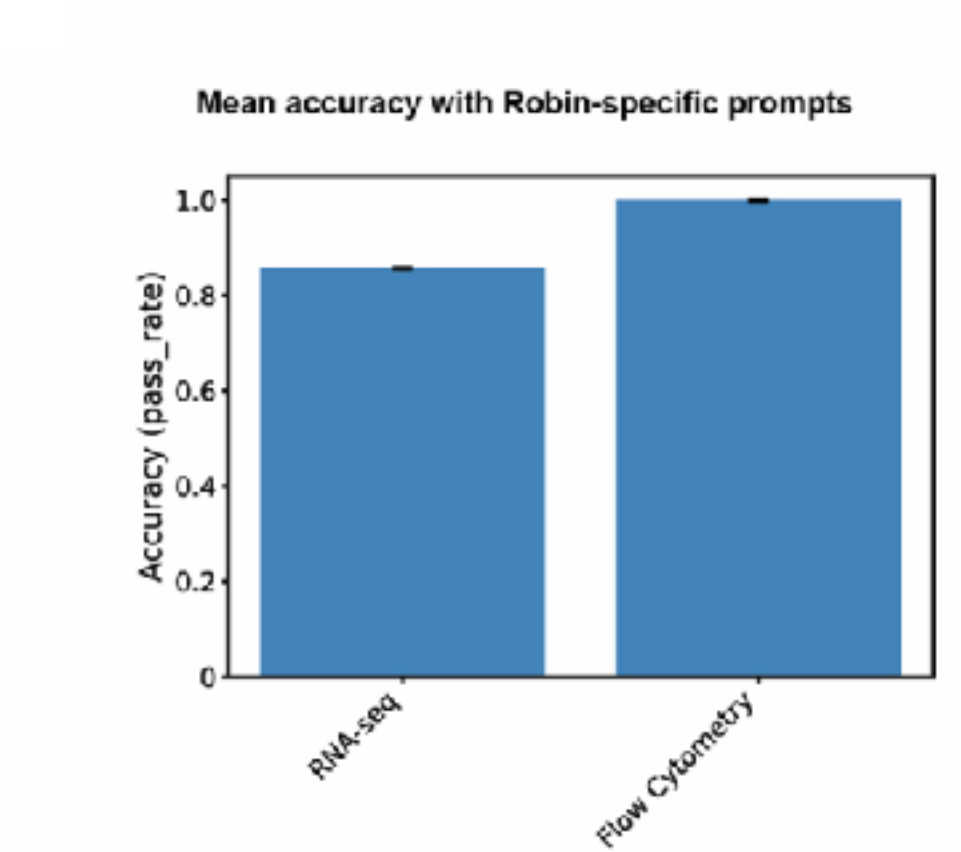
A blinded scientist tried to find every cited paper.

Without them, proposals get worse



Both agents swapped: 50 vs 50 proposals in one LLM-judged tournament; lower rank is better ($P < 0.001$).

Finch is reliable on Robin's tasks

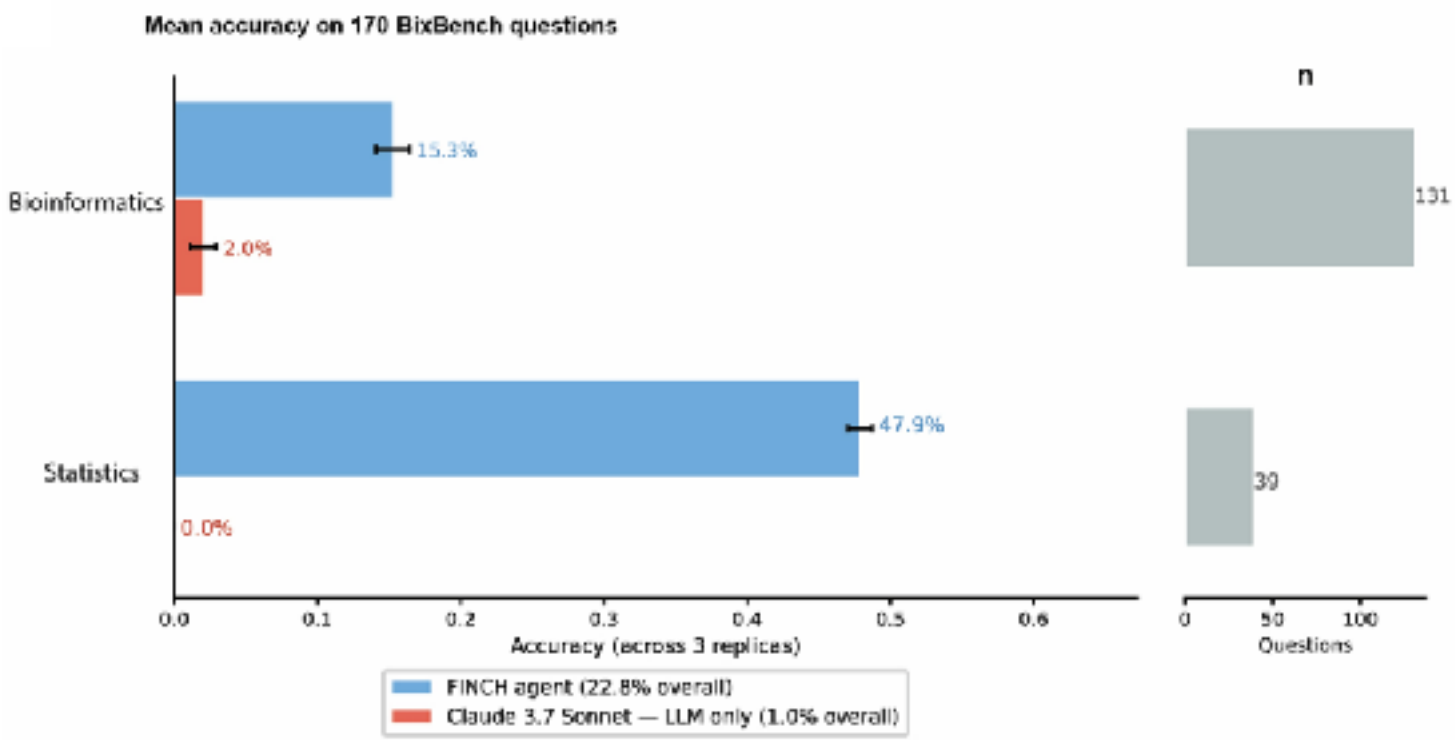


100% flow cytometry

86% RNA-seq

share of an expert-written rubric met, 3 runs each

On harder questions, Finch is often wrong

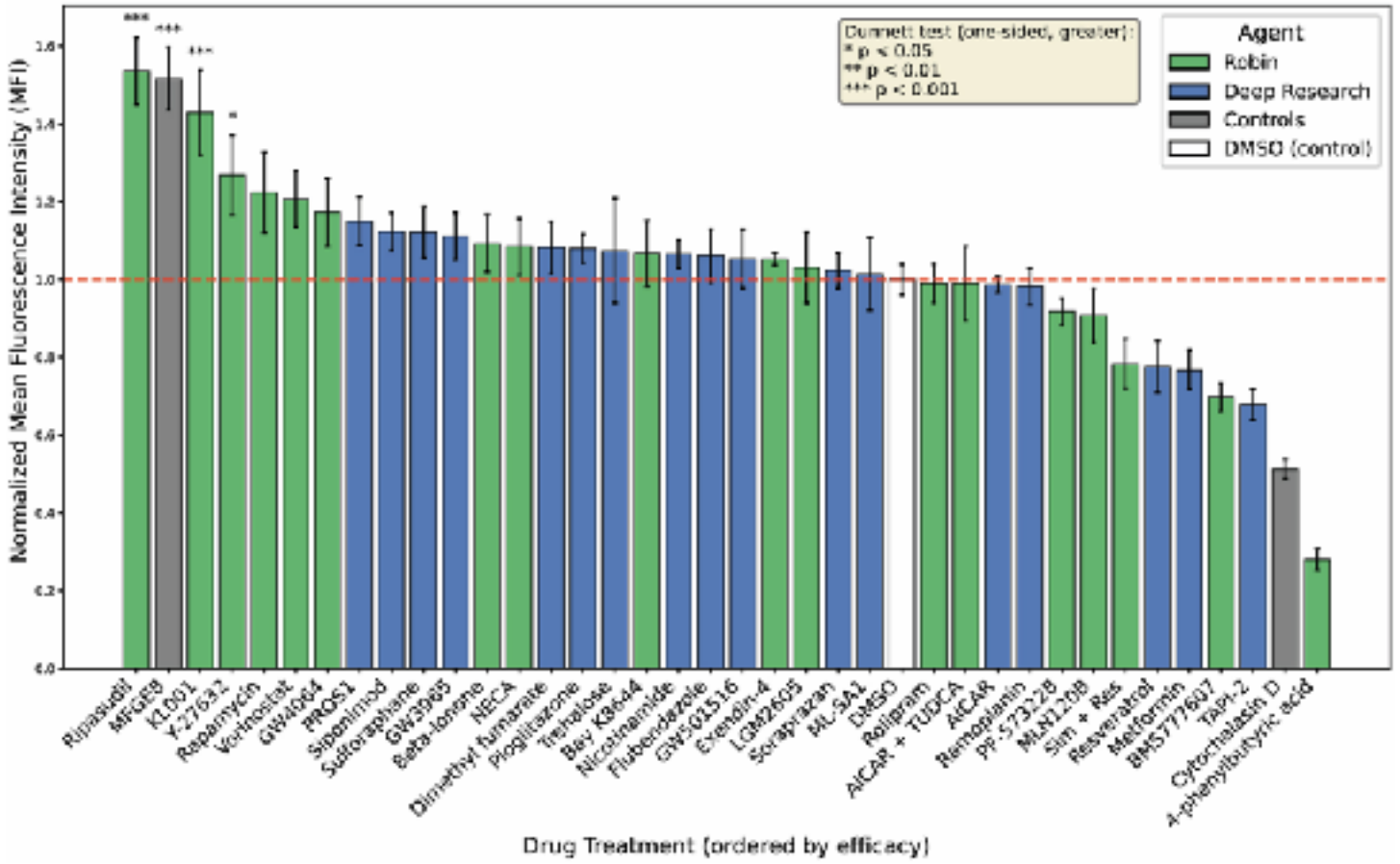


22.8% vs 1.6%

Finch vs the LLM alone, on 170 BixBench questions

Only 15% on multi-step bioinformatics

OpenAI's Deep Research found no working drugs



Robin: 3 working drugs

green bars

Deep Research: 0

blue bars, same prompt and assay

Cost and time of one round

$10.76

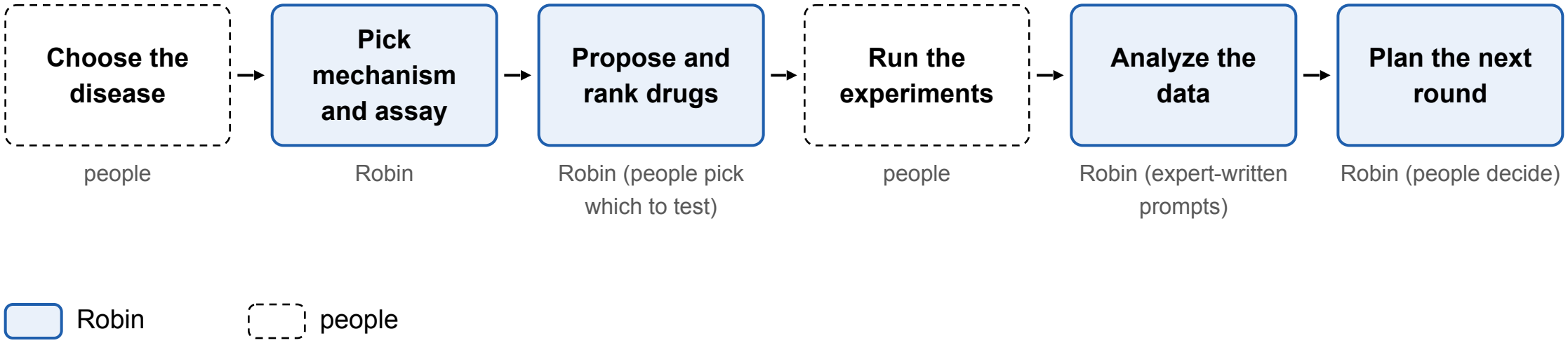
LLM cost of one Robin run

~400 h → <2 h

estimated human vs Robin thinking time per cycle

Human hours are survey-based estimates (Table 1); bench time is not included.

Robin plans; people run the experiments



The thinking steps are automated; the lab work and the decisions are not.

Takeaways

1

Robin runs the full research loop

Literature → hypothesis → experiment → analysis → better hypothesis

2

Searching real papers matters

Literature-search agents cut made-up citations from about half to 4% or less

3

A promising but early result

Ripasudil and KL001 boost phagocytosis in aged human RPE; animal tests are next

Questions for discussion

Q1 Is connecting already-published facts real discovery? Who gets the credit?

Q2 How should we benchmark AI scientists: many diseases, or head-to-head with experts?

Q3 Which steps should stay human, and what safeguards does automated biology need?

Thank you