

Biomedical discovery systems

Computational prioritization, retrospective recovery and prospective validation.

Reasoning agents, feedback, and scientific evidence

A biological association leaves a mechanism open

Imagine a gene is more active in diseased cells. It could help cause the disease, compensate for it, or mark a change in cell composition.

The same association therefore supports different intervention choices. A larger literature summary alone may not resolve which explanation is right.

This creates a reason to combine mechanistic hypotheses with measurements or perturbations that distinguish their consequences.

Constructing perturbations that separate explanations

Construct predictions before proposing a target. A causal driver, a compensatory response and a cell-mixture explanation need different controls.

A targeted perturbation, an independent assay or a cell-resolved measurement can be useful depending on the ambiguity. Feasibility limits which questions can be tested.

An agent may help retrieve evidence, prioritize candidates or organize tools. Its role follows from the bottleneck and intended endpoint, rather than from the availability of an agent framework.

Choosing a biomedical discovery strategy

Candidate	Reason to choose it	Cost or limitation
Literature synthesis	The relevant evidence already exists	Cannot create a missing intervention result
Predictive prioritization	Many candidates compete for a small assay budget	A high score need not identify a causal mechanism
Adaptive perturbation design	Outcomes can change the next useful experiment	Assay validity, feasibility and cost constrain the loop

Choose the system around the biological uncertainty and the attainable endpoint. A prioritized target still needs independent validation.

Pedagogical reconstruction

[Roohani et al. \(2024\)](#)

[Huang et al. \(2025\)](#)

A biological context is part of the hypothesis

A gene–phenotype relationship can depend on cell type, perturbation strength, time, environment, and measurement procedure.

Record these variables rather than reducing the claim to a context-free association between names. A result in one setting may motivate testing elsewhere but does not establish generalisation.

Evaluate whether the system preserves context when retrieving literature, comparing candidates, and summarising experimental outcomes.

Retrospective recovery has a useful but limited role

A public screen can serve as a hidden oracle for evaluating which candidates an agent prioritises under a fixed budget.

Recovery of known hits measures prioritisation in that dataset. It does not establish prospective novelty or rule out information leakage from training and retrieval.

Specify the knowledge cutoff, available literature, hit definition, and candidate universe. Compare random, established statistical, and domain-informed baselines.

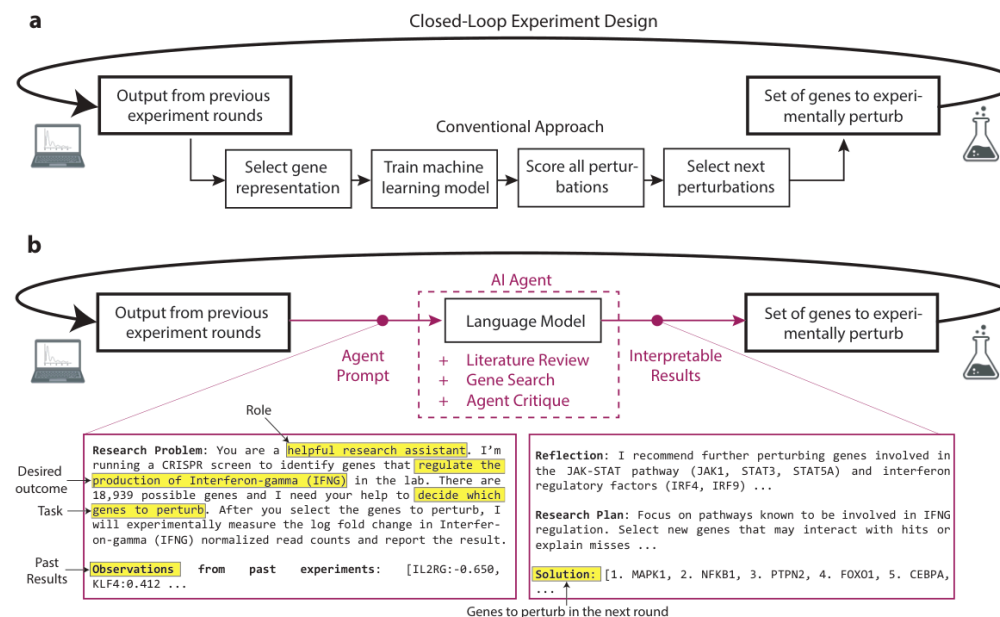
BioDiscoveryAgent: the method

The agent proposes genetic perturbations using previous experimental observations, domain knowledge, literature retrieval, and critique.

It is evaluated as a policy for selecting candidates in iterative campaigns. The language model's proposal process is compared with established selection methods under a defined action set.

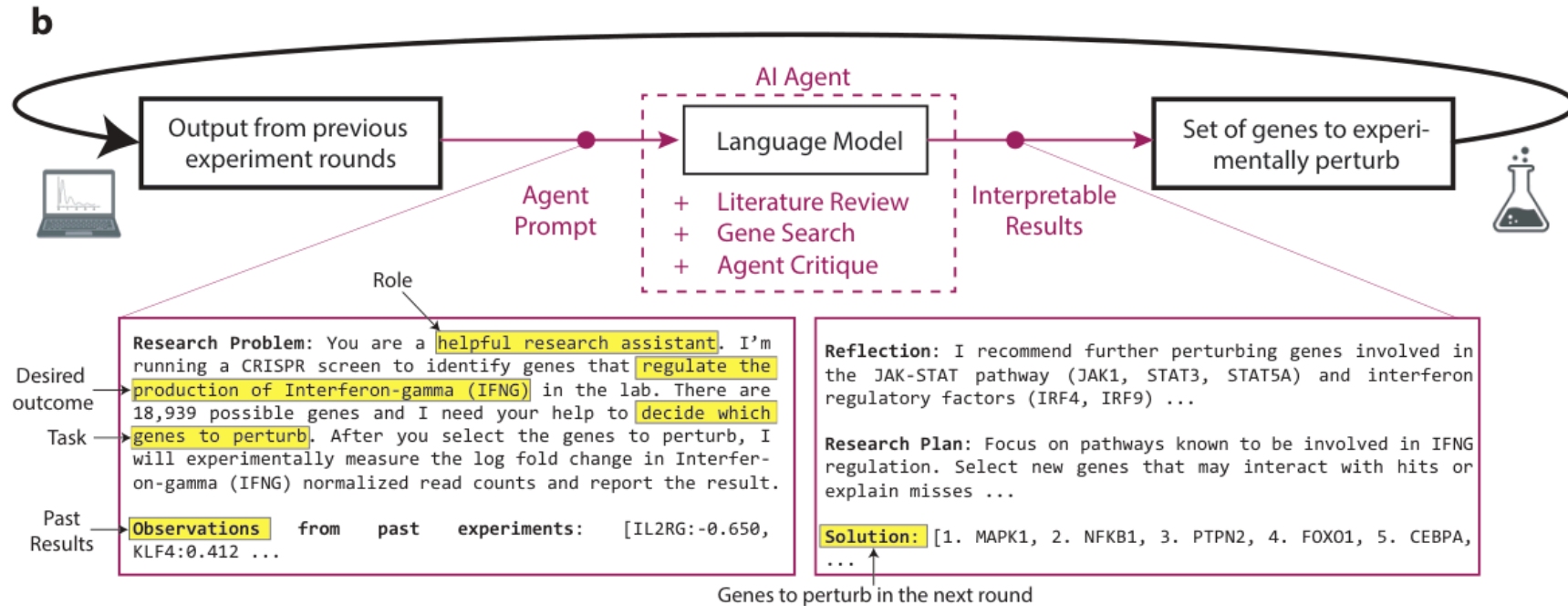
The method demonstrates a practical experiment-selection interface without automatically supplying a calibrated posterior over biological mechanisms or an explicit expected-information-gain calculation.

BioDiscoveryAgent: experiment-selection loop



The conventional and agent loops organize previous observations and candidate selection differently. The example shows how context and observed results enter the proposal process. This architecture alone does not establish causal identification.

BioDiscoveryAgent: mechanism detail



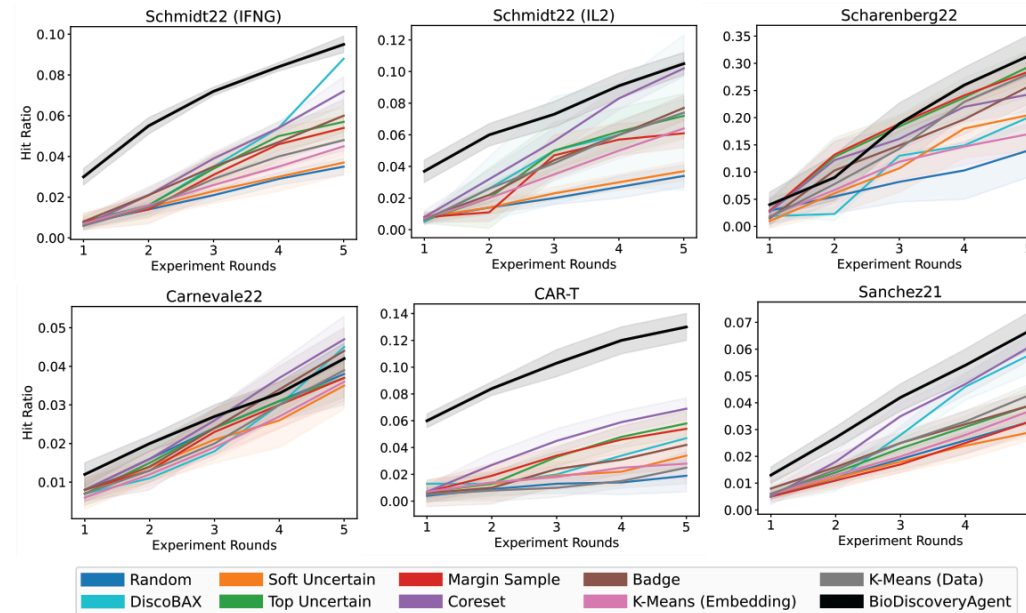
The selection component combines retained observations with new proposals. A valid negative result changes the evidence available for the next decision, even when it produces no positive discovery.

Evidence: a standard deviation is not a standard error

Model	Round 1	Round 2	Round 3	Round 4	Round 5
Random	2.6 ± 1.43	5.7 ± 2.83	8.9 ± 3.67	12.8 ± 3.74	16.4 ± 3.8
BioDiscoveryAgent (No-Tools)					
Claude 3 Haiku	5.1 ± 2.2	14.8 ± 2.9	23.6 ± 4.1	34.1 ± 5.8	45.1 ± 5.3
Claude 3.5 Sonnet	8.4 ± 1.36	15.7 ± 1.1	23.1 ± 2.59	31.7 ± 2.9	40.0 ± 3.77

At round five, Claude 3 Haiku reports 45.1 ± 5.3 cumulative hits and Claude 3.5 Sonnet reports 40.0 ± 3.77 across ten runs; $\pm$ denotes one standard deviation. Means and run variability describe this retrospective screen, rather than a prospective hit rate in another candidate population.

BioDiscoveryAgent: interpretation and limits



Published screen information may be available through training or retrieval. Biological context, measurement quality, and candidate selection affect the result. A successful retrospective campaign does not by itself establish a new therapeutic claim.

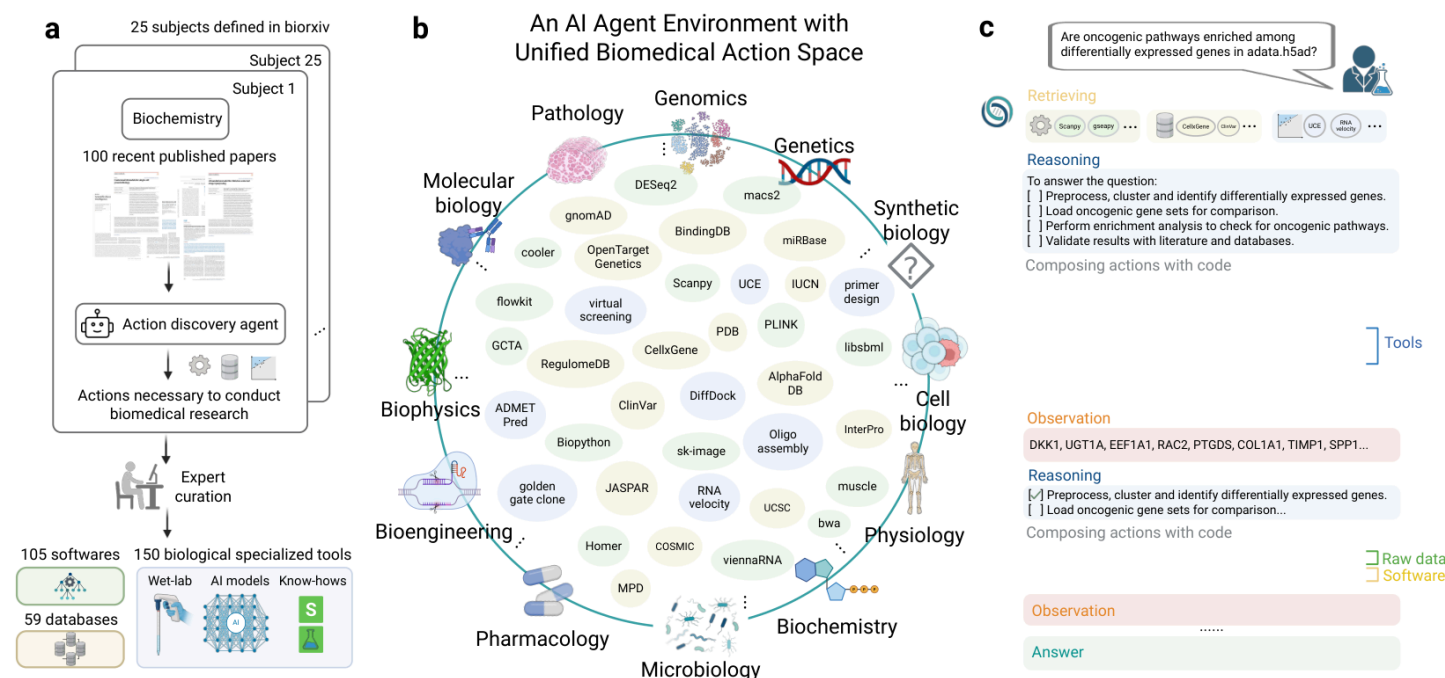
Biomni: the method

Biomni integrates a biomedical tool and data environment with retrieval-assisted planning and code execution. The agent can compose workflows across heterogeneous tasks.

Its action space is assembled from domain tools, databases, and software rather than a single fixed analysis template.

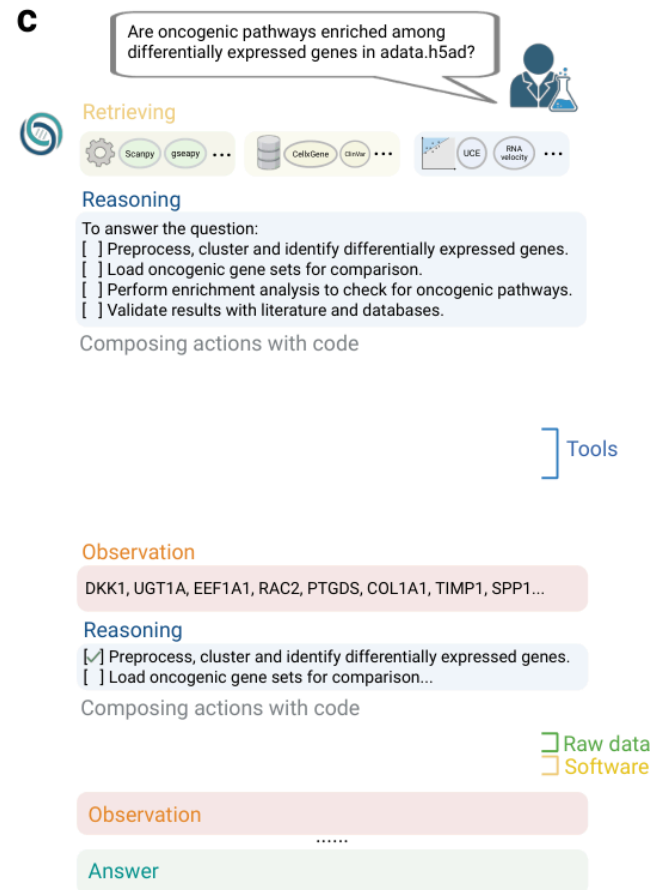
The useful architectural question is how retrieval and execution expose the right capabilities while preserving the meaning and provenance of their outputs.

Biomni: tools, retrieval and execution



Environment, retrieval, planning and execution components have distinct roles. Tool availability, appropriate selection and suitable input data are separate conditions for a successful workflow.

Biomni: mechanism detail



The handoff from retrieval to executable code depends on database versions, identifier mappings and data types. Interface errors can change the interpretation of a downstream analysis.

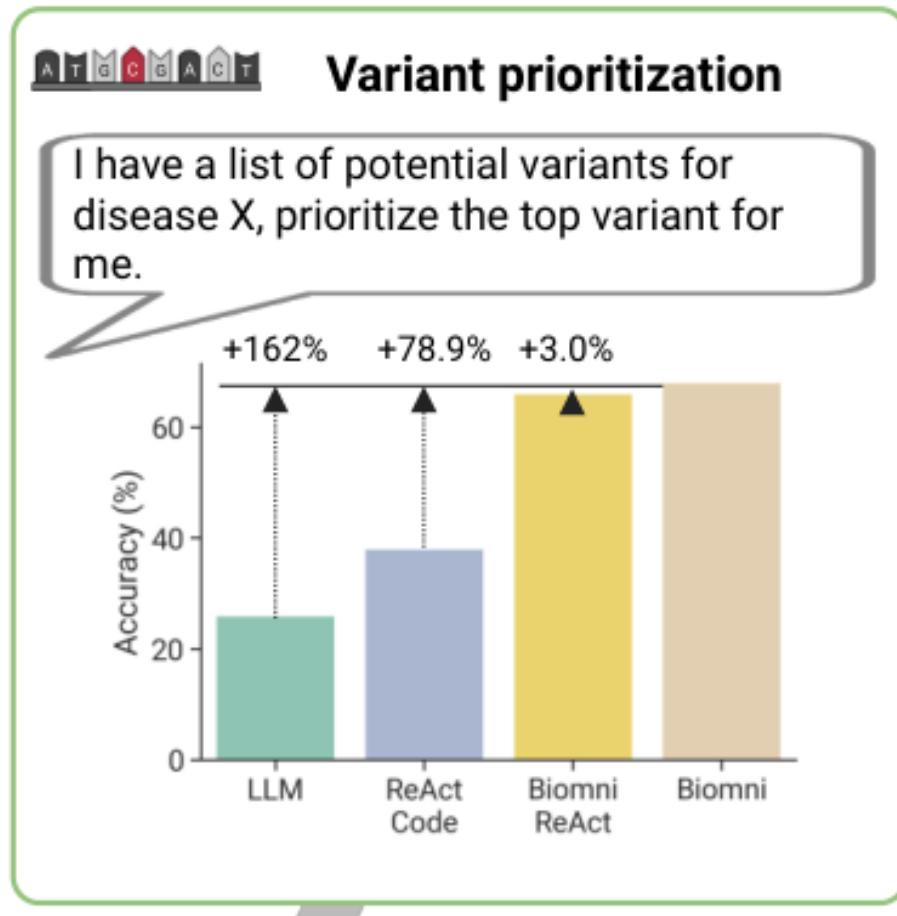
Evaluation scope

The paper evaluates multiple biomedical task families and presents detailed analysis case studies.

Each panel has its own task, scoring rule and baseline. Metrics from different panels are not interchangeable.

A successful computational analysis supports the evaluated workflow. Biological interpretation and claims about interventions require the additional evidence appropriate to the specific task.

Biomni: interpretation and limits



Task heterogeneity makes a single aggregate difficult to interpret. Model knowledge, tool access, data selection, and evaluation design can all affect the reported comparison.

Simulation and measurement answer different questions

A simulator can prioritize a biomedical candidate cheaply. An assay asks about a specified biological system; a later outcome may concern another system entirely.

L05 modeled cheap outputs as θ + shared bias + noise. Four correlated runs left posterior variance 17/21; one direct noisy measure reduced it to $\frac{1}{2}$.

That arithmetic is illustrative. For a real candidate, estimate discrepancy for the endpoint and context at issue; repeating the same simulator does not supply new biological validation.

Screening performance depends on prevalence

Sensitivity and specificity describe conditional test behaviour. The fraction of positive calls that are true also depends on how common true hits are in the screened population.

In a low-prevalence search, even a seemingly accurate screen can produce many false positives.

Report the candidate population, thresholds, and uncertainty. A precision claim from an enriched benchmark may not hold in a broader discovery campaign.

An assay effect suggests competing biological routes

Hypothetical assay: applying perturbation X changes readout Y . A mediator M changes too. Does M carry the effect, or merely track the perturbation?

Direct candidate: $M=X$, $Y=X$. Mediated candidate: $M=X$, $Y=M$. Both reproduce the initial observation and the response to changing X .

This motivates selective perturbation of M , rather than another fluent explanation of the same result. The construction follows L12's two causal routes.

Evidence that distinguishes the proposed routes

Comparison	What the models predict	What must be checked
Set $X=1$, hold $M=0$	Direct route: $Y=1$. Mediated route: $Y=0$.	M manipulation is selective; X and Y remain comparable.
Observe $Y=0$	Supports the mediated route relative to this direct rival.	Toxicity or off-target action could also alter the readout.
Change the system	Neither model yet predicts a different biological context.	Identify which mechanisms and measurement links persist.

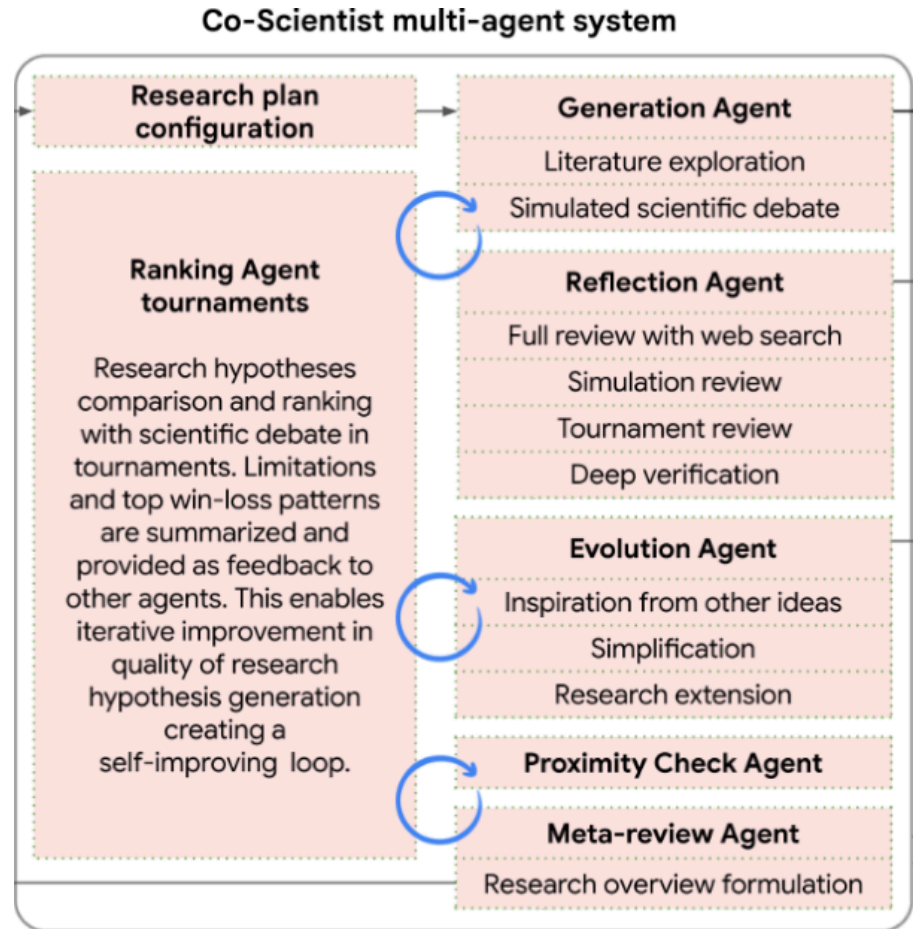
Support for an assay mechanism is not evidence, by itself, of efficacy in another system.

Robin's loop includes essential human work

Contribution	Role in the reported system	Interpretation
Literature agents	Crow and Falcon support literature research	Generate and evaluate candidate directions
Analysis agent	Finch analyzes experimental data	Feed measurements back into the loop
Researchers	Review candidates and write protocols	Decide and operationalize experiments
Laboratory work	Physical experiments produce observations	Computational proposals meet biological evidence

Robin: May 2025 preprint; Nature publication in 2026. Credit the full human–agent process rather than treating every step as autonomous.

Co-Scientist: mechanism detail



Generation, reflection and ranking form a feedback loop: candidates enter comparisons whose results influence later proposals. A high rank can reflect the evidence available to the evaluator or the persuasiveness of an explanation; the information supplied to each stage determines the interpretation.

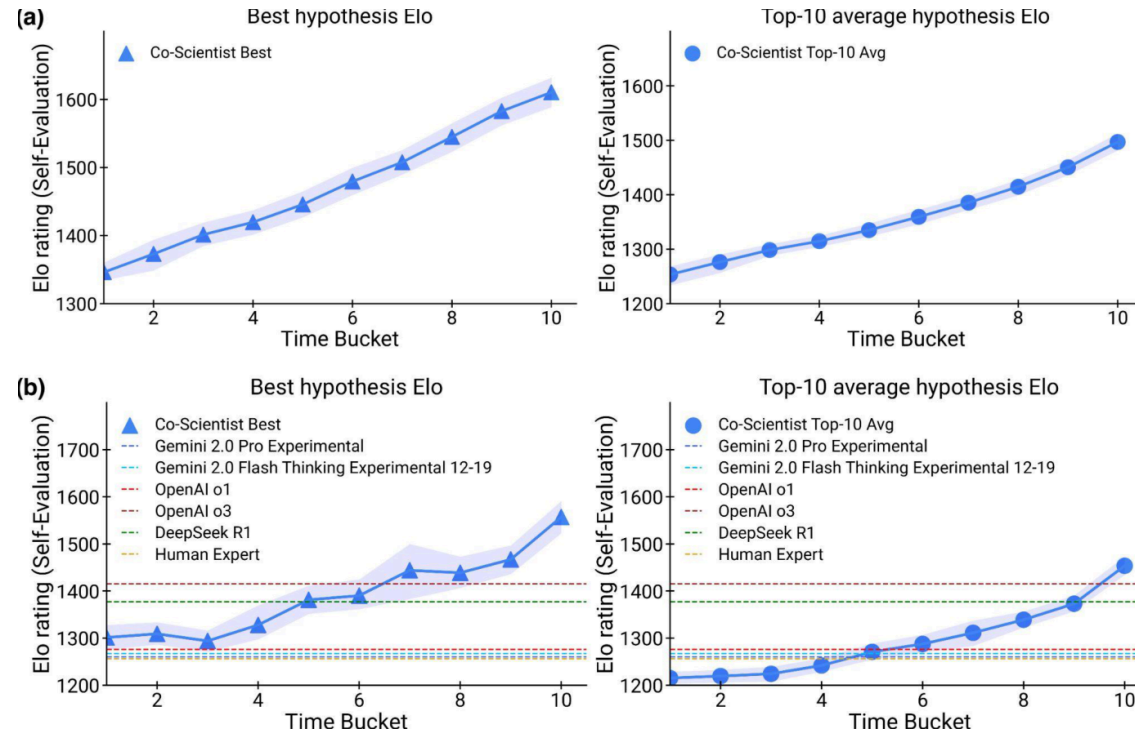
Co-Scientist: ranking and independent support

Internal Elo measures relative hypothesis rankings under the tournament evaluator. It is distinct from expert assessment and from an independently measured experimental outcome.

Extended Data Table 2 reports seven selected combinations examined in two contexts. The outcomes include support in both contexts, mixed support and unsupported cases.

Selection and measurement repetitions answer different questions. These selected cases establish context-specific evidence, not the success probability of all generated hypotheses.

Co-Scientist: interpretation and limits



These panels show internal Elo and auto-evaluation comparisons. They measure ranking under selected evaluators. The paper also reports experimental validations elsewhere; neither provides a calibrated success rate for every generated hypothesis or general autonomous performance across domains.

TxAgent as a biomedical reasoning case

TxAgent combines a fine-tuned language model, tool access and iterative reasoning for therapeutic question answering. Its reported evaluation uses biomedical reasoning benchmarks.

Study the task definition, information sources and component ablations. Keep question-answering performance distinct from prospective scientific or clinical outcomes.

The next experiment depends on the missing link

Uncertainty	Candidate next action	Why choose it?
Simulation discrepancy	Pair simulated predictions with matched measurements.	Estimate systematic mismatch; repeats alone cannot do this.
Causal route	Manipulate the mediator with suitable controls.	Rival mechanisms disagree under this intervention.
Transfer of effect	Test the claim in the relevant new model system.	Check the mechanism and endpoint in the new context.

Choose the observation that can change the current decision, not simply the cheapest available readout.

Pedagogical reconstruction

[Kennedy and O'Hagan \(2000\)](#)

[Pearl \(2009\)](#)

State exactly which claim the evidence supports

Effect: changing X changes assay Y in this setting. Mechanism: M carries that effect rather than a specified competing route.

Transfer: the relevant causal and measurement relationships persist in a new context. This requires additional assumptions and evidence about that context.

An agent can connect these questions and prioritize tests. Its success is the justified change in what is known, not the length of its generated explanation.

What the biomedical evaluations measure

Study / selected object	Observed endpoint	Limit of the inference
BioDiscoveryAgent / retrospective pool	Cumulative hits over 10 runs	SD across runs; not prospective hit rate
Biomni / variant prioritisation	Accuracy on defined cases	Tool and data access differ
Co-Scientist / 7 selected combinations	Expert-run tests in two contexts	Mixed support; no general success rate
Robin / human-reviewed candidates	In-vitro tests performed by humans	Selected-case results; not clinical efficacy

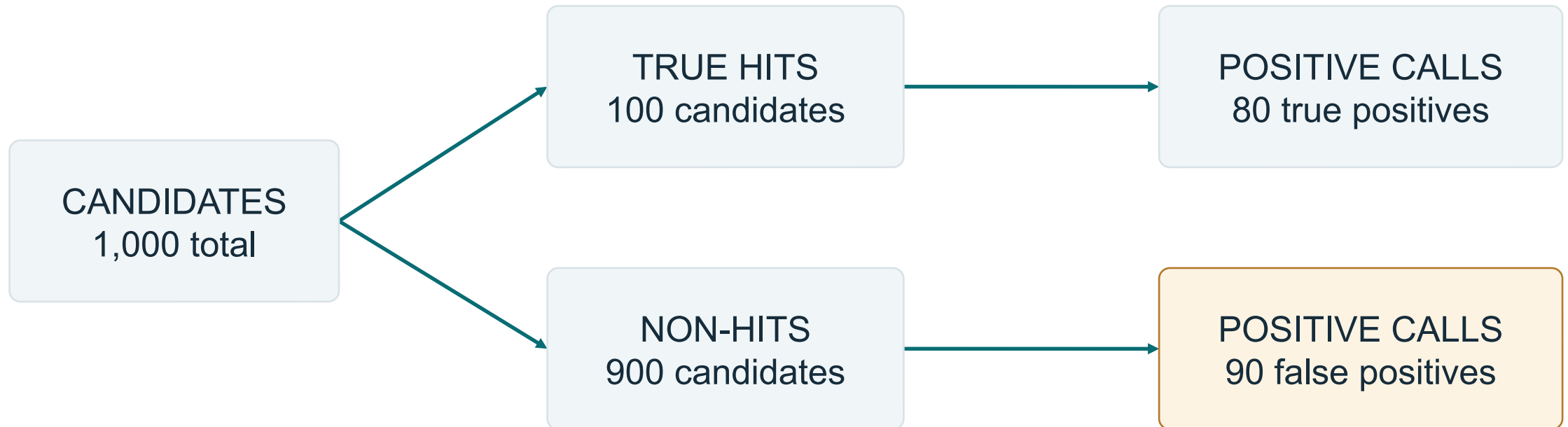
Different endpoints and selection denominators prevent a single league table of scientific agents.

A research follow-up threshold depends on utility

Operation	Calculation or state	Consequence
Abstract research action	Valid lead yields +9 units	Invalid lead yields -1 unit
Calibrated probability p	Expected value $9p - (1-p)$	Simplifies to $10p - 1$
Threshold	$10p - 1 > 0$	Follow up only when $p > 0.1$
Two hypothetical leads	$p=0.08$: -0.2; $p=0.20$: +1	Different decisions

Synthetic decision analysis for an abstract research choice. The numbers are utility units rather than a therapeutic or experimental prescription.

Worked screening: a low-prevalence candidate set



Sensitivity 0.8 selects 80 of the 100 true hits. Specificity 0.9 leaves a false-positive rate of 0.1 among 900 non-hits. The selected set combines both positive branches. These are synthetic expectations under fixed test characteristics; they are not measurements of a named biomedical model.

Calculate the precision of the selected set

Expected counts	Actual hit	Actual non-hit
Screen positive	80	90
Screen negative	20	810

Of 170 expected positive calls, 80 are true hits: precision = $80/170 \approx 0.471$. Sensitivity is $80/100 = 0.8$ and specificity is $810/900 = 0.9$. This table makes the different denominators explicit.

Change the population and observe the base-rate effect

With sensitivity 0.8 and specificity 0.9, a prevalence of 1% among 1,000 candidates yields eight expected true positives and 99 expected false positives.

Precision is now eight divided by 107, approximately 7.5%.

The test's conditional characteristics have not changed; the population has. This is why results from an enriched benchmark need careful interpretation before use in a broad search.

Retrospective biomedical prioritisation

Retrospective prioritisation benchmarks can compare agent policies with random, uncertainty-based and literature-informed baselines under the same hidden split and resource budget.

Hit recovery, calibration, context transfer and information available at selection time define different evaluation questions. Failed and unmapped candidates affect the denominator.

The result evaluates a prioritisation policy on the specified dataset. It leaves prospective performance and external scientific validation unresolved.

Precision follows from Bayes' rule

$$\text{PPV} = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity})(1 - \text{prevalence})}$$

Positive predictive value (PPV) is precision among positive screens. Its numerator is the probability of a true hit and positive screen; its denominator is the probability of any positive screen.

The worked example changes prevalence while holding sensitivity and specificity fixed; this isolates the base-rate effect.

Real deployment can also change conditional test behaviour. Re-estimate these quantities when the candidate population or measurement process changes.

Robin: an experimental feedback loop

Robin connects literature agents Crow and Falcon with the data-analysis agent Finch. Analysed results inform a new round of hypotheses, making feedback part of the discovery process.

Human scientists review ranked candidates, write protocols, conduct the physical experiments and provide data. The reported workflow is semi-autonomous.

Nature published the study in May 2026, following a May 2025 preprint. Selected in-vitro outcomes support the case study; Table 1 estimates human time from surveys rather than a matched timed trial.

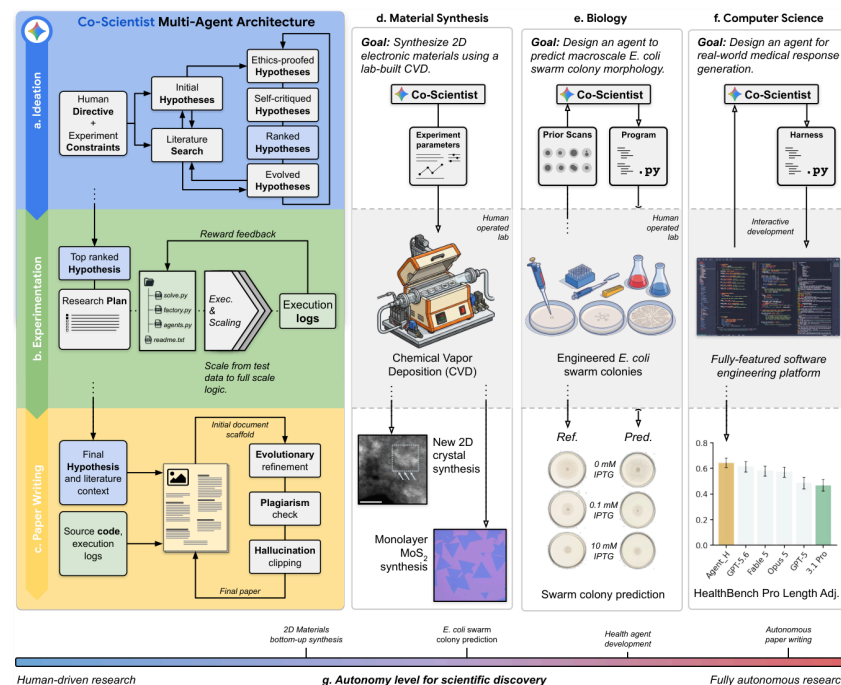
Co-Scientist in 2026: validation across domains

The August 2026 extension studies materials science, biology and computer science. Its level of autonomy varies by case, from human-executed physical work to automated computational evaluation.

Expert assessment, benchmark performance and experimental measurements are separate endpoints. Their denominators and human contributions must remain attached to each reported result.

Compared with Robin, the central question is how external evidence changes the next hypothesis. A completed workflow alone cannot establish reliable discovery across new domains.

Co-Scientist: hypothesis-to-execution workflow



The workflow connects hypothesis generation, execution and reporting. External measurements can ground the loop, while the individual experiments determine the strength of the validation.

Recent methods in context

Co-Scientist (Nature 2026). Co-Scientist organizes hypothesis generation, critique and tournament-based refinement. The Nature publication adds peer-reviewed evidence for selected applications. Internal rankings and external scientific validation support different claims.

FrontierScience (OpenAI report, 2025/26). Separates Olympiad questions from expert-written scientific research sub-tasks with rubric-based grading. Bounded scientific tasks cover only part of discovery, and scores depend on task selection and the evaluator.

EXP-Bench (ICLR 2026). Evaluates agents on experimental procedures reconstructed from AI research papers and code. Faithful execution of a prescribed experiment is narrower than open-ended creativity or a new scientific discovery.

References (1/2)

[Yusuf Roohani, Andrew Lee, Qian Huang et al. \(2024\). BioDiscoveryAgent: An AI Agent for Designing Genetic Perturbation Experiments](#)

[Kexin Huang, Serena Zhang, Hanchen Wang et al. \(2025\). Biomni: A General-Purpose Biomedical AI Agent](#)

[Juraj Gottweis, Wei-Hung Weng, Alexander Daryin et al. \(2025\). Accelerating scientific discovery with Co-Scientist](#)

[Kyle Swanson, Wesley Wu, Nash L. Bulaong et al. \(2025\). The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation \(preprint used for figures; journal version linked\)](#)

[Marc C. Kennedy, Anthony O'Hagan \(2000\). Predicting the output from a complex computer code when fast approximations are available](#)

[Kathryn Chaloner, Isabella Verdinelli \(1995\). Bayesian Experimental Design: A Review](#)

References (2/2)

[Samuel Schmidgall, Xiaokai Zhu, Marian Shaw et al. \(2026\). Accelerating Scientific Research with Gemini in the Real-World](#)

[Shanghua Gao, Richard Zhu, Zhenglun Kong et al. \(2025\). TxAgent: An AI Agent for Therapeutic Reasoning Across a Universe of Tools](#)

[Miles Wang, Robi Lin, Kat Hu et al. \(2025\). FrontierScience: Evaluating AI's Ability to Perform Expert-Level Scientific Tasks](#)

[Patrick Tser Jern Kon, Qiuyi Ding, Jiachen Liu et al. \(2025\). EXP-Bench: Can AI Conduct AI Research Experiments?](#)

[Ali E. Ghareeb, Benjamin Chang, Ludovico Mitchener et al. \(2026\). A multi-agent system for automating scientific discovery](#)

[Judea Pearl \(2009\). Causal inference in statistics: An overview](#)