

Experiment selection and information gain

Expected information gain over competing hypotheses.

Reasoning agents, feedback, and scientific evidence

An unpredictable test can teach nothing

Two tests each return positive or negative with equal probability. Both look maximally unpredictable before we run them.

For the first test, rival mechanisms predict different outcomes. For the second, both mechanisms predict the same fair coin flip.

Only the first can change our relative beliefs. This exposes why raw outcome uncertainty is the wrong target when the goal is to learn a mechanism.

Comparing what outcomes would change

List the possible outcomes and ask how each would change the hypotheses. We need the likelihood of an outcome under each candidate mechanism.

Because the result is still unknown, average the possible learning before choosing the test. This constructs an expected information objective.

If the goal is a final action, ask instead whether the posterior changes that action enough to justify the measurement cost. Learning and decision value can diverge.

Choosing the acquisition objective

Candidate	Reason to choose it	Cost or limitation
Outcome uncertainty	Find unpredictable or poorly modeled observations	Can select noise unrelated to the hypothesis
Expected information gain	Distinguish candidate mechanisms	Depends on the hypothesis and outcome models
Expected decision value	Improve a specified final action	Needs utilities, costs and available actions

Use information gain for the stated learning goal. Use decision value when a consequential action and its losses define success.

Pedagogical reconstruction

[Chaloner and Verdinelli \(1995\)](#)

Expected information gain averages possible futures

$$\text{EIG}(e) = \mathbb{E}_y[\text{KL}(b(h \mid y, e) \parallel b(h))]$$

Expected information gain (EIG) averages the Kullback–Leibler (KL) divergence $\text{KL}(\text{posterior} \parallel \text{prior})$ over possible outcomes y of experiment e . Let $b(h)$ denote the current belief.

Equivalently, subtract expected posterior entropy from current entropy. The expectation is taken before the experiment.

An information-gain estimate is only as useful as its likelihood and hypothesis class; misspecification can make apparently decisive experiments misleading.

Realised surprise and ElG are not interchangeable

After observing y , Bayesian surprise is the KL change from the prior belief to the posterior. Predictive surprisal is instead minus the log probability of y .

ElG averages posterior changes over outcomes that might occur. A surprising observation can reveal an instrument fault or model misspecification rather than a useful discovery.

Identify which quantity a paper actually optimises before transferring its result to experiment selection.

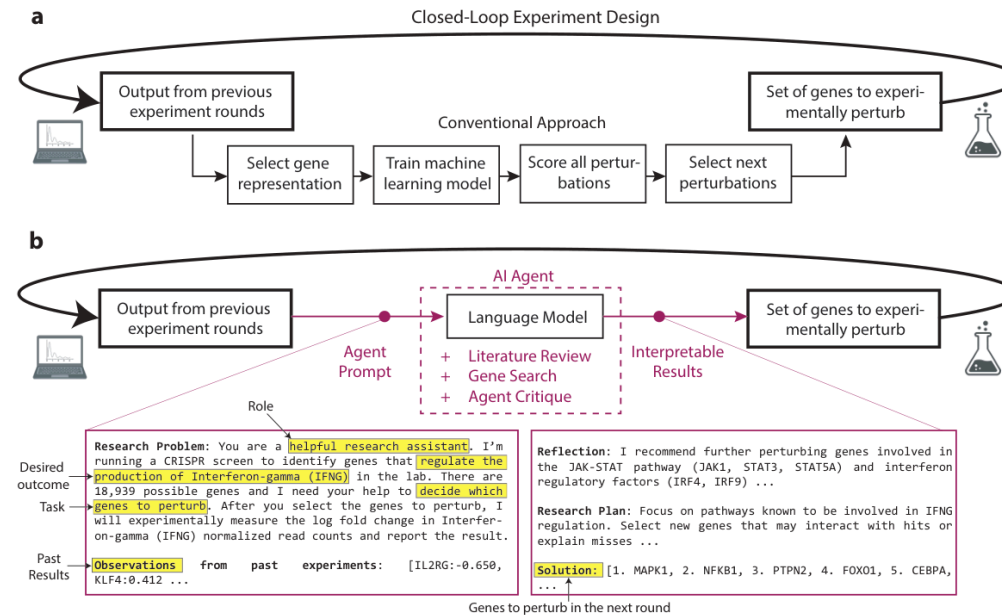
BioDiscoveryAgent: the method

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.

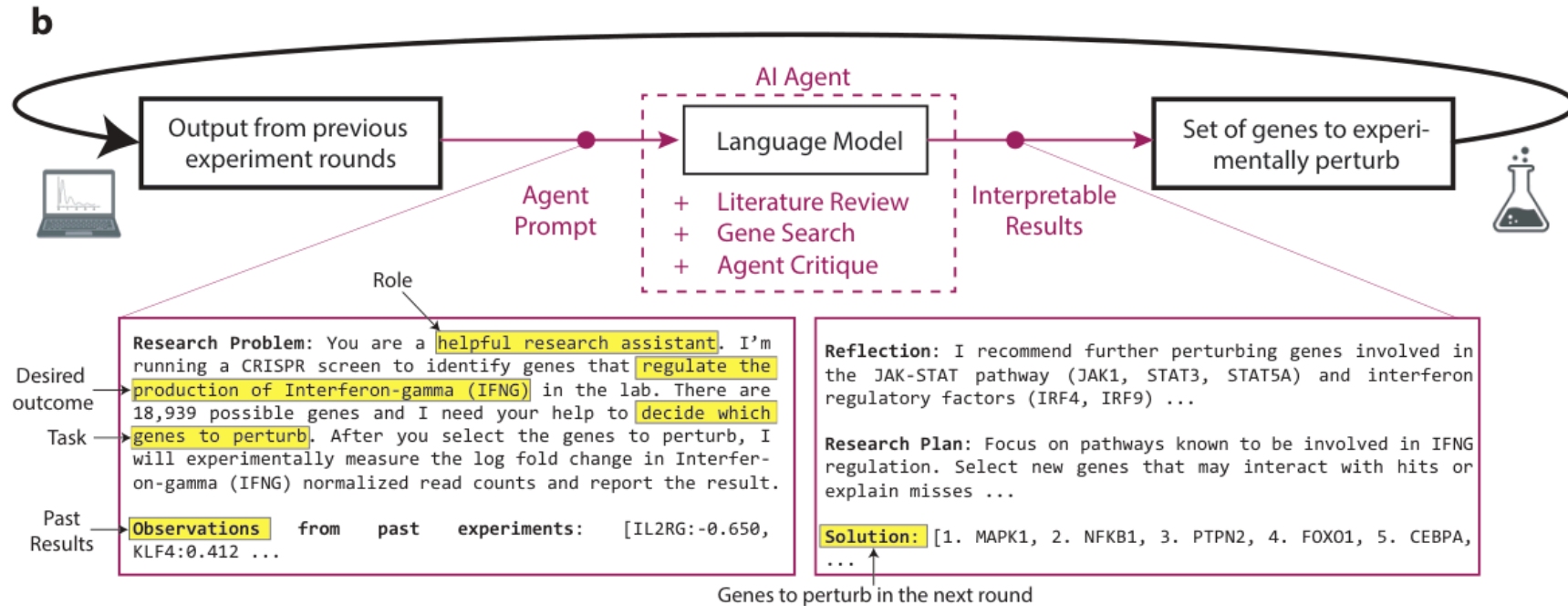
The method demonstrates a practical experiment-selection interface without automatically supplying a calibrated posterior over biological mechanisms or an explicit expected-information-gain calculation.

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



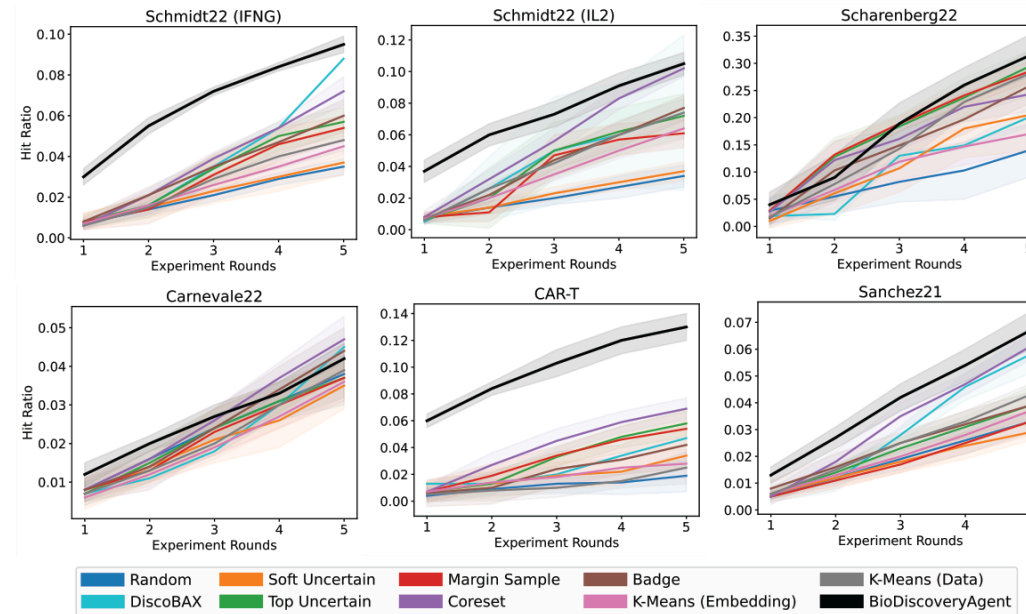
The selection component combines retained observations with new proposals. A valid negative result changes the evidence available for the next decision, even when it produces no positive discovery.

Evidence: read cumulative hits and uncertainty

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

In this retrospective two-gene screen, 32 pairs are selected per round from 100,576 possible pairs. Round-five cumulative hits average 40.0 for Claude 3.5 Sonnet and 16.4 for random selection over ten runs. The displayed error intervals are one standard deviation. A hit count is not an information-gain measurement.

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.

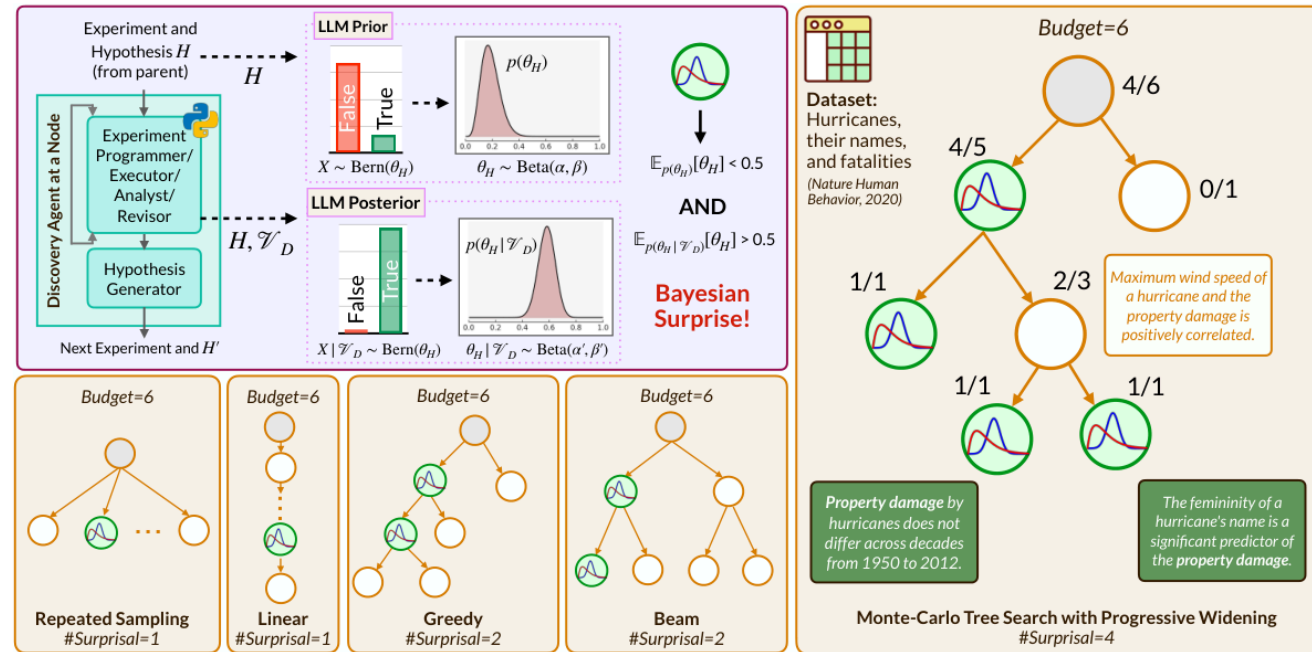
AutoDiscovery: the method

AutoDiscovery explores nested hypotheses using a search procedure guided by changes between elicited prior and posterior beliefs.

Data analysis and hypothesis evaluation produce feedback used in an MCTS-style exploration process. Bayesian surprise supplies an objective related to belief change.

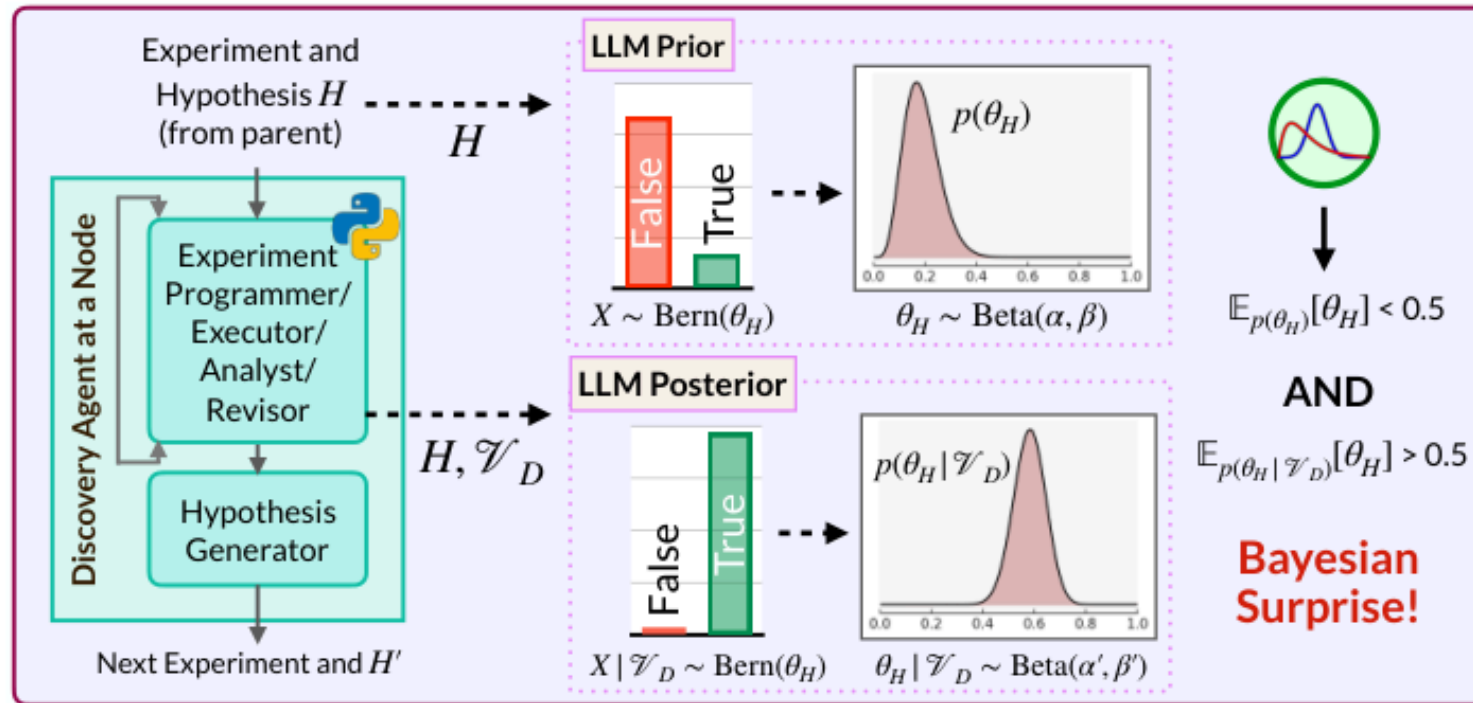
The method's operational belief estimates and search reward should be distinguished from a fully calibrated generative model and from expected information gain evaluated before an experiment.

AutoDiscovery: surprise-guided search



The figure links a hypothesis, analysis program, prior and posterior beliefs, and search expansion. Realized surprise concerns an observed outcome. Prospective information gain averages over outcomes that have not yet been observed.

AutoDiscovery: mechanism detail



The diagram links prior-to-posterior change to branch expansion. Whether a search reward is available before or after data analysis determines its decision-theoretic interpretation.

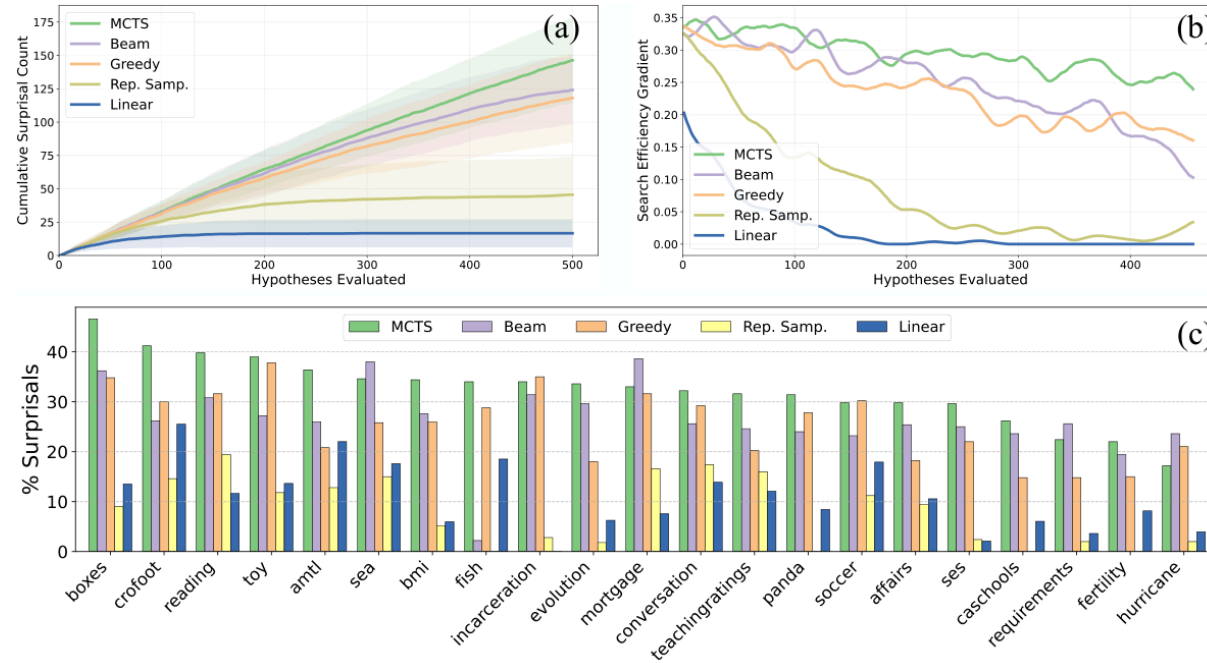
Evaluation scope

The study compares search procedures using discovery-related rewards and human assessments of selected outputs.

The original search-performance figure separates cumulative discoveries, yield, and performance across datasets. Interpret each axis with its evaluation definition.

A high surprise score can help prioritise investigations, but its scientific value depends on the reliability of the analysis and the relevance of the target belief.

AutoDiscovery: interpretation and limits



Elicited beliefs can be miscalibrated, and an unexpected result can reflect a model or data error. Human assessments and benchmark-specific rewards do not imply a universal stopping or experiment-selection rule.

A high-entropy observation may be uninformative

An outcome can be unpredictable under every hypothesis in exactly the same way. It has high predictive entropy but does not help distinguish the hypotheses.

EIG measures dependence between the outcome and the target hypothesis, conditional on the chosen experiment.

This is why choosing the noisiest measurement is not equivalent to choosing the best measurement. Separate epistemic uncertainty from irreducible observation noise.

A cost ratio can choose the wrong one-test allocation

Test	Expected information	Cost	Information / cost
e_1 : noisy discriminator	0.278 bits	1	0.278
e_3 : perfect discriminator	1 bit	10	0.100

Budget ten; only one test is allowed. The perfect discriminator yields more information, but the ratio selects e_1 . This myopic rule need not optimise the allocation. Specify the allowed actions, total budget, and whether measurements can repeat or adapt.

Surrogates require a calibrated connection

Simulations, predictive models, and retrieved observations can inform beliefs about physical systems when their relationship to the target is modelled.

A surrogate observation should carry fidelity, uncertainty, and known bias. It need not be excluded simply because it is computational.

In multifidelity design, a cheap result is useful when it reliably changes the ranking of expensive actions. Domain shift can weaken that reliability.

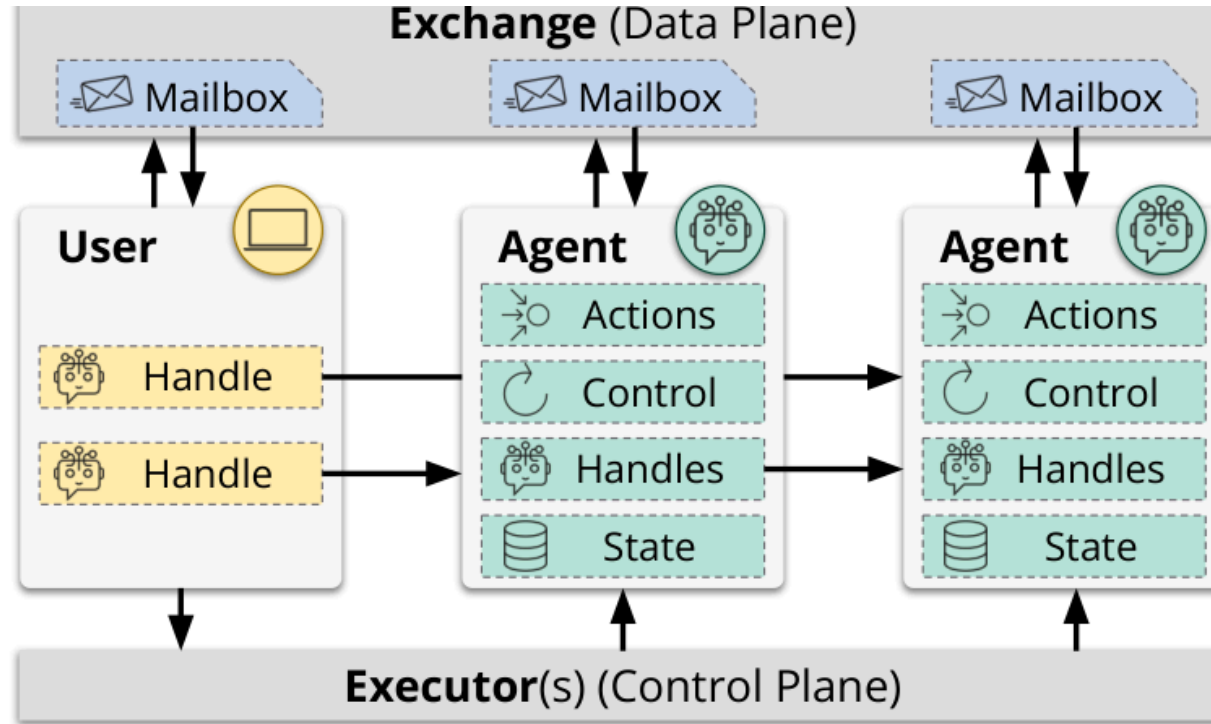
ACADEMY federated agents: the method

ACADEMY provides infrastructure for agents that interact asynchronously across scientific computing resources. Handles expose actions, while executors and an exchange separate control from communication.

The framework addresses deployment, messaging, and integration with scientific workflows.

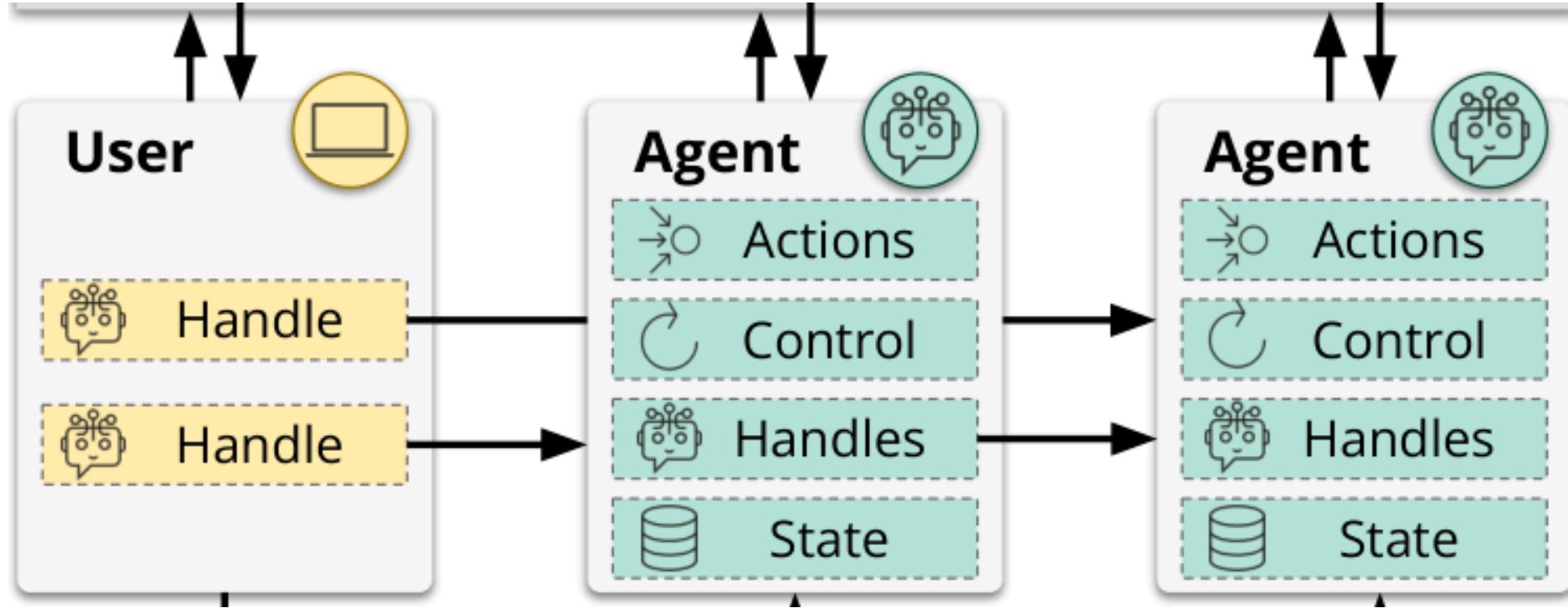
These system capabilities can make a research policy executable, but they do not define the scientific objective or prove that the agent selects informative experiments.

ACADEMY: communication and execution architecture



User or agent handles connect through the exchange to actions and execution contexts. The data plane and control plane serve different roles. These arrows represent infrastructure communication and execution relationships, rather than causal biological effects.

ACADEMY federated agents: mechanism detail



Action, control and state interfaces organize calls across agents and executors. An immediate handle and an eventual result represent different states of execution. Pending work supplies no completed observation.

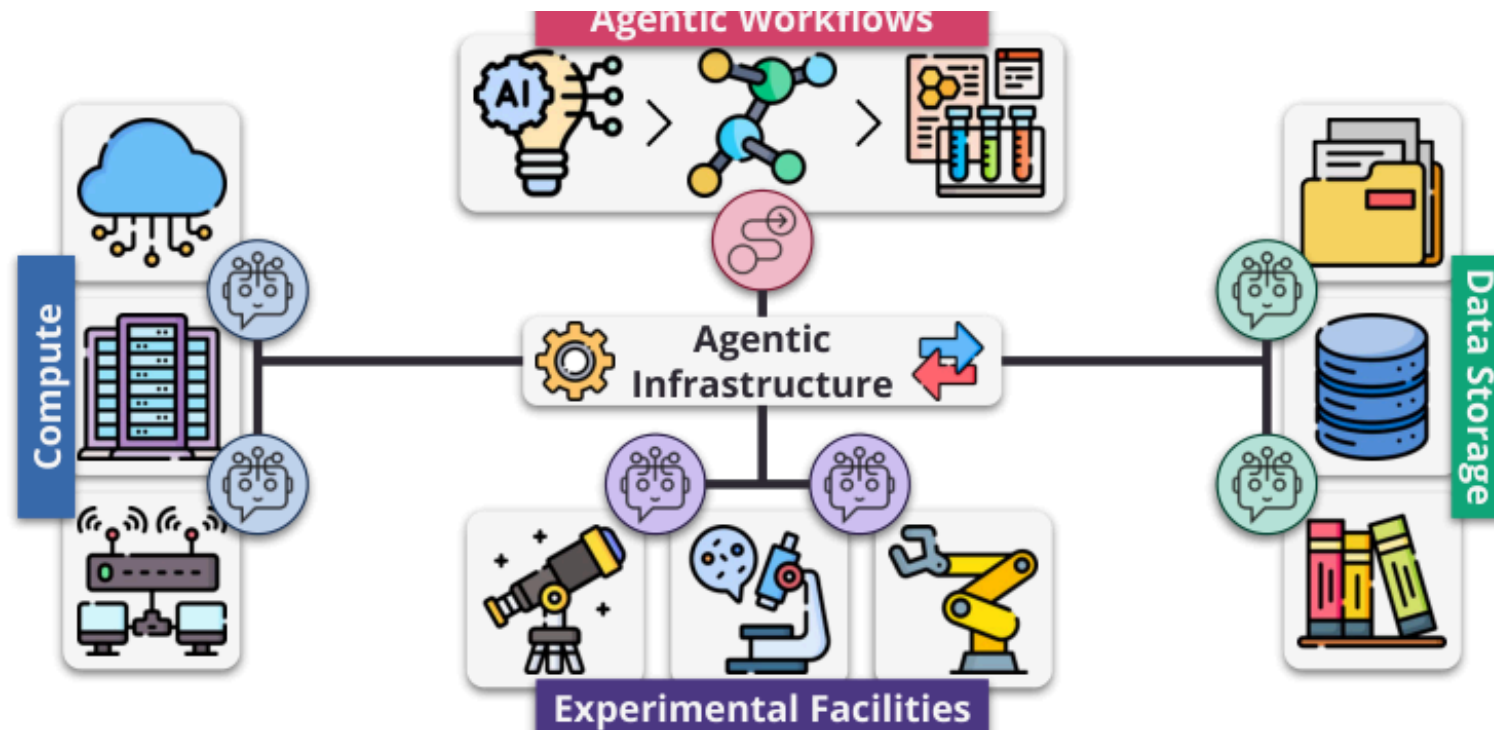
Evaluation scope

The source describes a framework and scientific-workflow use cases. The extracted figures explain how agents connect to computational and experimental resources.

They are evidence of the proposed system organisation, not measurements of discovery quality.

Evaluate a deployment using task completion, latency, resilience, and resource accounting, then separately assess whether the scientific policy using that infrastructure produces useful evidence.

ACADEMY federated agents: interpretation and limits



Infrastructure generality does not establish a better acquisition policy. Queueing delays, permissions, resource failures, and data movement can alter the set of feasible scientific actions and their effective costs.

Cheap simulations suggest a second evidence source

An expensive measurement observes a target quantity θ . A cheap simulator approximates it. The opportunity is more information per unit of budget.

But repeated simulations can share the same discrepancy. Treating their agreement as independent evidence about reality is the questionable assumption.

Candidates: use measurements alone, pool all outputs, or model how simulation differs from measurement. The third choice can exploit cheap data without hiding bias.

Represent the shared discrepancy explicitly

Prior: $\theta \sim N(0,1)$. Cheap run i : $y_i = \theta + b + \varepsilon_i$. Shared bias $b \sim N(0,4)$; independent noise $\varepsilon_i \sim N(0,1)$.

Direct measurement: $z = \theta + \eta$, with $\eta \sim N(0,1)$. All prior variables and noise terms are independent, except that every cheap run shares b .

Average n cheap runs: effective noise variance $v = 4 + 1/n$. Posterior variance $V = v/(1+v)$; information gain = $\frac{1}{2} \ln(1/V)$ nats.

A finite budget changes which source to use

Budget = 4 units	Posterior variance V	Information about θ
Four cheap runs (cost 1 each)	$(4 + \frac{1}{4}) / (5 + \frac{1}{4})$ $= 17/21 \approx 0.810$	$\frac{1}{2} \ln(21/17)$ ≈ 0.106 nats
One direct measure (cost 4)	$1 / (1+1) = \frac{1}{2}$	$\frac{1}{2} \ln 2 \approx 0.347$ nats
Unlimited cheap runs	V approaches $4/5$	Shared bias leaves an uncertainty floor.

More repeats reduce random noise. They do not identify an unknown shared discrepancy.

Pedagogical reconstruction

[Chaloner and Verdinelli \(1995\)](#)

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

Calibration can reverse the preferred choice

If b were known exactly, subtract it. Four cheap runs then give $V = 1/5$ and information gain $\frac{1}{2} \ln 5 \approx 0.805$ nats: now cheap runs win.

Paired simulation–measurement data can estimate discrepancy, but calibration costs budget and may not transfer to a new region or model system.

The design question becomes which uncertainty to reduce next: the target, simulator noise, or discrepancy. L16 returns to the same choice in biomedical evidence.

Information value follows the terminal decision

Operation	Calculation or state	Consequence
Prior	$P(A)=0.4, P(B)=0.6$	Stopping value 0.60
Test model	$P(+ A)=0.9; P(+ B)=0.3$	Positive mass 0.54
Best choices	A after +; B after –	Correct mass $0.36+0.42$
Value of information	$0.78-0.60=0.18$	Buy only below this cost

Synthetic correctness utility. Expected information value depends on the best action after each outcome, not simply on how much the posterior moves.

Worked design: two equally likely mechanisms

Experiment	$P(+ A)$	$P(+ B)$	$P(+)$
e_1 : discriminating test	0.8	0.2	0.5
e_2 : pure noise for this question	0.5	0.5	0.5

The example uses equal priors over A and B and base-two logarithms. Both experiments have balanced outcomes; only e_1 predicts different outcomes across mechanisms. The calculation assumes valid measurement and correctly specified likelihoods.

Posterior uncertainty for e_1

$$\text{EIG}(e_1) = 1 - H_2(0.8) \approx 0.278 \text{ bits}$$

The prior entropy is one bit. A positive outcome gives posterior probabilities 0.8 and 0.2; a negative outcome reverses them.

Both posteriors have entropy approximately 0.722 bits. Since the outcomes are equally likely, expected posterior entropy is also 0.722 bits.

EIG is therefore about 0.278 bits. Symmetry simplifies this example; generally every possible outcome needs its own posterior and predictive weight.

Explain why e_2 teaches nothing about the mechanism

$$p(A \mid y, e_2) = 0.5, \quad \text{EIG}(e_2) = 0$$

Under e_2 , both hypotheses assign the same likelihood to either outcome. Bayes' rule therefore leaves the prior unchanged after a positive or negative result.

The observation still has one bit of entropy, but its mutual information with the hypothesis is zero.

This example separates noisy measurements from informative measurements. The important feature is how predictions differ across hypotheses, not how uncertain the raw outcome looks.

Robustness to model and measurement assumptions

Sensitivity to likelihoods, priors and cost estimates affects experiment selection. Measurement failure and omitted model classes can alter the preferred action.

Decision changes and detectable prediction–observation mismatches reveal the sensitivity of the procedure to its assumptions.

An acquisition decision is interpretable alongside its assumptions and sensitivity analysis. Public-data or simulation results establish computational behavior within their specified setting.

Information gain across possible outcomes

$$I(H; Y | e) = \sum_{h, y} b(h) p(y | h, e) \log_2 \frac{p(y | h, e)}{p(y | e)}$$

The predictive denominator is $p(y|e) = \sum_h b(h)p(y|h,e)$. Terms with zero joint mass contribute zero.

If every hypothesis predicts the same outcome distribution, numerator and denominator agree and every nonzero term vanishes.

In the binary worked example the prior is uniform and the symmetric test is correct with probability 0.8, giving $1 - H_2(0.8) \approx 0.278$ bits.

Information gain need not improve a decision

Suppose the final action is unchanged for every posterior reachable after a test. The test can reduce entropy while having zero expected value for that particular decision, before costs.

Conversely, a small information gain near a decision boundary can be valuable.

Specify the loss of each final action and compare posterior expected losses. This converts a preference for learning into a decision-theoretic objective when the application requires one.

A confident posterior can favor a bad model

Setup or calculation	Model M1	Model M2
Success probability; equal prior weights	0.9	0.8
Observed sequence: five independent failures	Likelihood = 0.1^5	Likelihood = 0.2^5
Posterior within this two-model family	$1/33 \approx 0.0303$	$32/33 \approx 0.9697$

Yet the prior-predictive probability of that exact sequence is only 0.000165. Relative preference is not absolute adequacy.

Model expansion requires a new comparison

Stage	Scientific action	Bookkeeping
Unexpected failures	Check instruments, dependence and sampling	Do not assume model failure is the only cause
A third model is proposed	M3 predicts success probability 0.2	Record that M3 was proposed after the failures
A fresh trial is selected	Compare M2: 0.8 with M3: 0.2	Declare new priors and the prospective rule
Fresh evidence arrives	Update and inspect absolute predictive fit	Keep exploration separate from confirmation

Experiment design chooses within a model space; scientific revision can change that space. Those are different operations.

Recent methods in context

Model Discovery Agent (preprint; revised 2026). Combines Bayesian model/parameter inference, LLM proposals of model structures and value-of-information experiment choice. Version 4 reports fixing prompt leakage and rerunning benchmarks; comparisons must use the corrected version.

AutoDiscovery (NeurIPS 2025). AutoDiscovery combines hypothesis exploration with Bayesian surprise. It links an observed belief change to further search, while surprise and the prospective value of a future experiment remain different quantities.

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.

References (1/2)

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

Dhruv Agarwal, Bodhisattwa Prasad Majumder, Reece Adamson et al. (2025). AutoDiscovery: Open-ended Scientific Discovery via Bayesian Surprise

Alok Kamatar, J. Gregory Pauloski, Yadu Babuji et al. (2025). Empowering Scientific Workflows with Federated Agents

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

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

References (2/2)

Kathryn Chaloner, Isabella Verdinelli (1995). Bayesian Experimental Design: A Review

Chenglei Si, Zitong Yang, Yejin Choi et al. (2026). Towards Execution-Grounded Automated AI Research

Kevin Murphy (2026). Model Discovery Agent: LLM-assisted Bayesian experiment design for data-efficient discovery of mechanistic world models

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