MINIMAL EPISTEMIC CLOSED-LOOP AGENTS FOR SCIENTIFIC DISCOVERY: CONSTRAINTS, INFORMATION GAIN, AND DISCRIMINATIVE TESTING IN A TOY GENOMICS ENVIRONMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Closed-loop "AI scientist" systems can generate hypotheses, call tools, and iterate, but many implicitly optimize for plausibility or predicted success rather than explicitly reducing uncertainty about an underlying mechanism. We study a minimal design for *epistemic* closed-loop agents that (i) enforce feasibility via hard constraints during proposal, (ii) select experiments by Expected Information Gain (EIG), and (iii) accelerate refutation by generating discriminative "Achilles" tests that maximize disagreement among hypotheses. We evaluate only algorithmic behavior in a small, reproducible toy genomics simulator with discrete latent hypotheses and noisy outcomes (not biological validation). Across random seeds, EIG-based selection reduces posterior entropy faster than success-seeking baselines, and Achilles-style testing reduces experiments-to-refute incorrect leading hypotheses under an operational refutation criterion. Finally, we use a minimal taxonomy for negative outcomes (infeasible, inaccessible, null, execution-failure) to avoid conflating distinct failure semantics during belief updates.

1 INTRODUCTION

Closed-loop agents are increasingly capable of drafting hypotheses, invoking tools, and iterating on results (Lu et al., 2024; Bran et al., 2024; Boiko et al., 2023). However, many systems prioritize outputs that appear plausible or likely to succeed under learned heuristics, rather than actions that maximally reduce uncertainty about a hidden mechanism. Bayesian experimental design formalizes epistemic progress via Expected Information Gain (EIG) (Lindley, 1956; Chaloner and Verdinelli, 1995; Rainforth et al., 2024), and falsification-oriented workflows emphasize actively seeking discriminative tests that separate competing explanations (Liu et al., 2025; Popper, 1959).

Scope. We intentionally restrict claims to a toy genomics simulator to demonstrate algorithmic behavior, not wet-lab capability: (i) hard feasibility constraints, (ii) EIG-based experiment selection, (iii) discriminative "Achilles" tests for rapid refutation, plus a minimal negative-outcome taxonomy.

Related work. Recent "AI scientist" agents demonstrate end-to-end automation across ideation, tool use, and writing (Lu et al., 2024; Bran et al., 2024; Boiko et al., 2023) but do not primarily optimize epistemic uncertainty reduction. Separately, Bayesian experimental design studies EIG objectives and estimators (Lindley, 1956; Chaloner and Verdinelli, 1995; Rainforth et al., 2024), while falsification agents explicitly target refutation behaviors (Liu et al., 2025). Our contribution is a minimal, reproducible closed loop that integrates (constraints + EIG + discriminative testing) and reports simple metrics (entropy reduction and experiments-to-refute) in a controlled simulator.

2 MINIMAL EPISTEMIC CLOSED-LOOP

Let $h \in \mathcal{H}$ denote a discrete latent mechanism/hypothesis, $x \in \mathcal{X}$ an experiment, and y an observed outcome. The agent maintains $p(h \mid \mathcal{D})$ and iteratively selects x to reduce uncertainty.

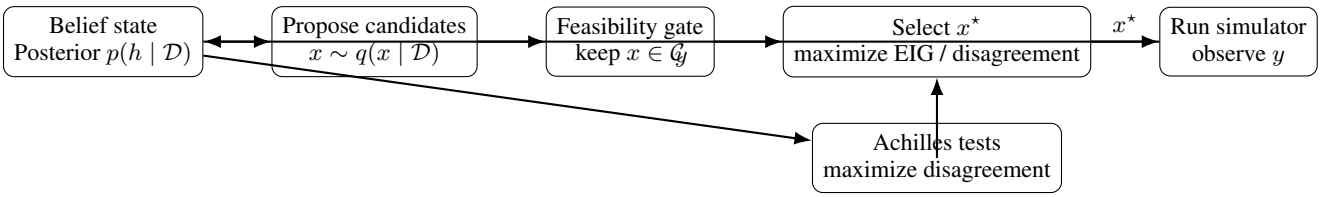


Figure 1: Minimal epistemic closed loop. "Operational refutation" means driving an incorrect leading hypothesis' posterior below a threshold (Sec. 3).

A: Hard feasibility constraints. Many proposals are invalid due to sequence/protocol/assay constraints. We model feasibility as a hard set $\mathcal{C}$ and do not execute $x \notin \mathcal{C}$. In our toy instantiation, $\mathcal{C}$ includes length bounds, GC bounds, and low-complexity filters. Hard checking can be deterministic or solver-based (De Moura and Bjørner, 2008); here we use deterministic filters.

B: Epistemic planning via EIG. The planner selects experiments maximizing mutual information:

$$x^\star = \arg\max_{x \in \mathcal{X}} I(h; y \mid x, \mathcal{D}). \tag{1}$$

In our discrete toy setting, EIG is computed by enumerating h and marginalizing outcomes.

C: Achilles tests (discriminative testing). Beyond informativeness in expectation, Achilles tests target *disagreement*: propose x that maximally separates predictions under the current leading hypothesis $\hat{h}$ from plausible alternatives, accelerating refutation when $\hat{h}$ is wrong (Liu et al., 2025; Popper, 1959).

3 Toy Genomics Environment and Evaluation

Environment. We use a synthetic sequence-to-outcome simulator inspired by promoter-like design. Each hypothesis h defines a conditional outcome model $p(y \mid x, h)$ with noise. The agent begins with a prior over h and updates $p(h \mid \mathcal{D})$ by Bayes' rule after observing outcomes.

Baselines. We compare: `random`; `greedy_success` (maximize predicted mean under MAP hypothesis); `uncertainty` (maximize predictive variance); `eig` (Eq. 1); and an `achilles` variant that prioritizes disagreement-based tests.

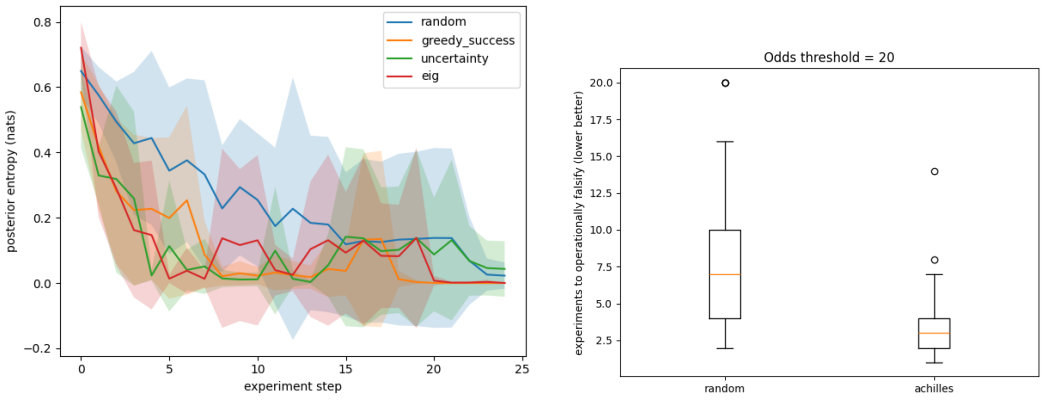
Metrics. (1) Posterior entropy $H[p(h \mid \mathcal{D})]$ vs. step. (2) Experiments-to-refute: steps until an incorrect leading hypothesis satisfies $p(\hat{h} \mid \mathcal{D}) < \epsilon$).

3.1 Results

EIG-based selection reduces posterior entropy faster than `random` and `greedy_success`, consistent with EIG optimizing expected posterior change (Rainforth et al., 2024). Achilles-style selection reduces experiments-to-refute by prioritizing queries that maximize predictive separation across hypotheses.

4 Negative Outcomes and Failure Semantics

Negative outcomes are not monolithic: conflating them can induce incorrect belief updates. We use a minimal four-way taxonomy: **(i) infeasible** (violates $\mathcal{C}$), **(ii) inaccessible** (assay limitation / unobservable), **(iii) null** (no signal under noise), **(iv) execution failure** (the experiment did not execute correctly. Execution failures are treated as epistemically null events and do not contribute evidence for or against any hypothesis.) This supports principled updates and motivates structured logging of negative outcomes (Fanelli, 2012; Ioannidis, 2005).



(a) Posterior entropy vs. step.

(b) Experiments-to-refute.

Figure 2: Toy environment results.

5 Limitations and Conclusion

This paper demonstrates algorithmic behavior in a small simulator and does not claim biological validation or autonomous wet-lab capability. Within this scope, combining feasibility constraints, EIG-based selection, and discriminative testing yields more epistemic progress per experiment. Next steps include scaling to established closed-loop benchmarks (Gandhi et al., 2025; Jansen et al., 2024) and testing whether failure semantics improve calibration and decision quality.

References

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Bran, A.-M., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. ChemCrow: Augmenting large-language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Lindley, D. V. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Chaloner, K. and Verdinelli, I. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Rainforth, T., Foster, A., Ivanova, D. R., and Bickford Smith, F. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–127, 2024.

Liu, Z., Liu, K., Zhu, Y., Lei, X., Yang, Z., Zhang, Z., Li, P., and Liu, Y. BABY-AIGS: An LLM-agent framework with explicit falsification for AI-generated science. *arXiv preprint*, 2025.

Popper, K. *The Logic of Scientific Discovery*. Hutchinson, 1959.

De Moura, L. and Bjørner, N. Z3: An efficient SMT solver. In *TACAS*, pp. 337–340, 2008.

Fanelli, D. Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3):891–904, 2012.

Ioannidis, J. P. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.

Gandhi, K., Li, M. Y., Goodyear, L., Li, L., Bhaskar, A., Zaman, M., and Goodman, N. D. BoxingGym: Benchmarking progress in automated experimental design and model discovery. *arXiv preprint*, 2025.

Jansen, P. A., Côté, M.-A., Khot, T., Bransom, E., Dalvi, B., Majumder, B. P., Tafjord, O., and Clark, P. DiscoveryWorld: A virtual environment for developing and evaluating automated scientific discovery agents. *arXiv preprint*, 2024.