# Can AI Scientists Discover Neural Mechanisms? Evaluating Agentic Biological Discovery in a Digital Fly

Aarav Sinha [1]

## Abstract

Autonomous scientific agents are beginning to move beyond narrow assistance roles toward systems that can generate hypotheses, run experiments, and draft manuscripts. Yet current evaluation still says little about whether such systems can uncover mechanisms in biology rather than merely summarize correlations or orchestrate known pipelines. We propose a benchmark for agentic biological discovery centered on neural mechanism discovery in a digital fly, and we instantiate it as a pilot study on six tasks drawn from a whole-brain Drosophila model. The benchmark casts discovery as a budgeted hypothesis–experiment–update loop in which an agent observes compressed neural and behavioral readouts, selects interventions, and outputs a mechanistic explanation together with a held-out counterfactual prediction. On feeding and grooming tasks, a planning agent reaches mean Node F1 of 0.856/0.952/1.000/1.000 at budgets 2/4/6/8, outperforming a one-shot GPT-5.4 stand-in (0.532), a no-memory variant (0.578/0.667/0.667/0.667), and a memory-only variant (0.789/0.897/0.897/0.897). In a named-versus-anonymized test, the one-shot baseline drops from 1.000 to 0.532 Node F1 whereas planning remains at 1.000 in both settings. We interpret these results as a pilot benchmark-sensitivity study rather than a definitive leaderboard, but they suggest that the digital-fly setting can expose meaningful differences in iterative mechanistic reasoning, memory, and shortcut robustness.

[1]Independent Researcher. Correspondence to: Aarav Sinha <aaravsinha002@gmail.com>.

Accepted to the GenBio workshop at ICML 2026.

## 1. Introduction

The scientific capabilities of language-model-based agents are expanding rapidly. Recent systems such as The AI Scientist and The AI Scientist-v2 pursue increasingly autonomous loops over ideation, code generation, experimentation, analysis, and writing, while broader surveys describe a shift from conventional AI for Science toward more autonomous agentic science (Lu et al., 2024; Yamada et al., 2025; Gridach et al., 2025; Wei et al., 2025; Xin et al., 2025). Biology is now part of this shift: recent papers and surveys describe AI agents for nanobody design, gene-set analysis, computational biology workflows, and single-cell analysis, and GenBio 2026 explicitly highlights hypothesis generation, experimental planning, and evaluation frameworks for autonomous biological systems as central themes (Swanson et al., 2025; Wang et al., 2025; Xiao et al., 2024; Alber et al., 2026; Qi et al., 2026; Huang et al., 2026; GenBio 2026 Workshop Organizers, 2026).

What remains missing is a strong answer to a more specific question: can an AI scientist discover a biological mechanism? Many existing demonstrations show that agents can execute pipelines, retrieve domain knowledge, or produce persuasive scientific prose. Those are useful capabilities, but they do not directly test whether an agent can distinguish causation from correlation, choose informative interventions under budget, revise its beliefs after contradictory evidence, and produce an explanation that survives held-out perturbation. Science-oriented benchmark work increasingly argues that these workflow and process questions matter at least as much as the final artifact (Chen et al., 2024; Abram, 2026). Mechanism discovery therefore imposes a stricter contract than ordinary question answering: the system must identify which components matter, justify that claim from interventions, and make predictions about experiments it has not yet seen. That framing is closer to intervention-based causal reasoning than to one-shot answer generation (Peters et al., 2017).

Neuroscience offers a particularly demanding version

of this problem. Mechanism discovery requires partial observability, sequential experimentation, and counterfactual reasoning. At the same time, recent progress in fly connectomics and simulation has created a tractable middle ground between toy causal graphs and expensive wet-lab validation. A whole-brain fly wiring diagram is now available; connectome-constrained models can predict activity in the fly visual system; whole-brain computational models can generate experimentally testable sensorimotor circuit hypotheses; and neuromechanical digital twins support embodied, closed-loop evaluation (Dorkenwald et al., 2024; Lappalainen et al., 2024; Shiu et al., 2024; Lobato-Rios et al., 2022; Wang-Chen et al., 2024; Wang-Chen & Ramdya, 2026). Together, these advances make a digital fly a plausible substrate for evaluating agentic discovery. Figure 1 summarizes the resulting benchmark loop.

This paper proposes neural mechanism discovery in a digital fly as a benchmark for agentic biological discovery and reports a first pilot instantiation. Our contributions are fourfold. First, we formalize biological discovery as a budgeted hypothesis–experiment–update loop rather than one-shot question answering. Second, we instantiate this framework on six pilot tasks from feeding and grooming circuits in a whole-brain Drosophila model. Third, we show that the benchmark is sensitive to the capabilities we care about: memory and experiment planning improve performance sharply, and anonymization reveals shortcutting in a one-shot baseline.

Agentic AI for science. Recent AI scientist systems demonstrate that autonomous research workflows are no longer purely speculative. Lu et al. (2024) and Yamada et al. (2025) show increasingly end-to-end pipelines for idea generation, experimentation, and writing, while surveys synthesize a broader literature on scientific agents, tool use, and autonomous discovery (Gridach et al., 2025; Wei et al., 2025). At the same time, benchmark papers argue that scientific agents should be evaluated on validated workflow tasks, contamination resistance, and multi-turn reasoning rather than on paper-shaped outputs alone (Chen et al., 2024; Abram, 2026). Our work follows this evaluation-first view, but targets mechanistic discovery in biology.

Agentic AI for biology. Biology-specific agent systems are emerging quickly. Recent work spans experimentally validated computational design, self-verifying bioinformatics agents, automated single-cell pipelines, agent-oriented bioinformatics benchmarks, and broader surveys of biological agents (Swanson et al., 2025; Wang et al., 2025; Xiao et al., 2024; Alber et al., 2026; Fa et al., 2026; Tang, 2025; Qi et al., 2026; Huang et al., 2026). These papers establish that agentic workflows can already be useful in biology. However, most focus on pipeline execution, artifact generation, or literature-grounded reasoning, not on iterative recovery of a causal circuit mechanism under interventions. That gap matters because a system can successfully complete a multi-step pipeline while still failing at deciding which biological variables are explanatory rather than merely predictive (Fa et al., 2026).

Digital twins and fly-scale mechanism modeling. The fly is a compelling testbed because it combines biological richness with unusually strong simulation support. Whole-brain connectomics provides structural coverage at the scale of an adult brain (Dorkenwald et al., 2024). Connectome-constrained models predict neural activity across the fly visual system (Lappalainen et al., 2024). Whole-brain computational models have already been used to identify compact feeding and grooming circuits and to generate experimentally testable sensorimotor hypotheses (Shiu et al., 2024). On the embodiment side, NeuroMechFly and related work argue that digital twins can act as hypothesis-generation platforms bridging brain, body, and behavior (Lobato-Rios et al., 2022; Wang-Chen et al., 2024; Wang-Chen & Ramdya, 2026). We build on this convergence and treat the digital fly as an evaluation substrate for agentic discovery.

## 2. Problem Formulation

### 2.1. What counts as discovery?

We model agentic biological discovery as interaction with an environment containing a latent mechanism $m^\star$. Given a scientific goal $g$ and an experiment budget $B$, the agent receives observations, chooses interventions, and returns a hypothesis together with supporting evidence and a held-out counterfactual prediction. Formally, one benchmark instance can be viewed as

$$\mathcal{T} = (\mathcal{E}, \mathcal{O}, \mathcal{A}, \mathcal{H}, g, B),$$

where $\mathcal{E}$ is the digital biological environment, $\mathcal{O}$ the observation space, $\mathcal{A}$ the allowed interventions, and $\mathcal{H}$ the space of mechanistic hypotheses. At round $t$, the agent observes a history $h_t = \{(c_i, a_i, o_i)\}_{i=1}^{t-1}$ and chooses $a_t \in \mathcal{A}$ until it stops or exhausts its budget.

This definition deliberately separates discovery from three easier objectives. First, a predictive model may forecast neural responses without identifying which neurons are necessary or sufficient. Second, a controller may exploit the environment to maximize re-

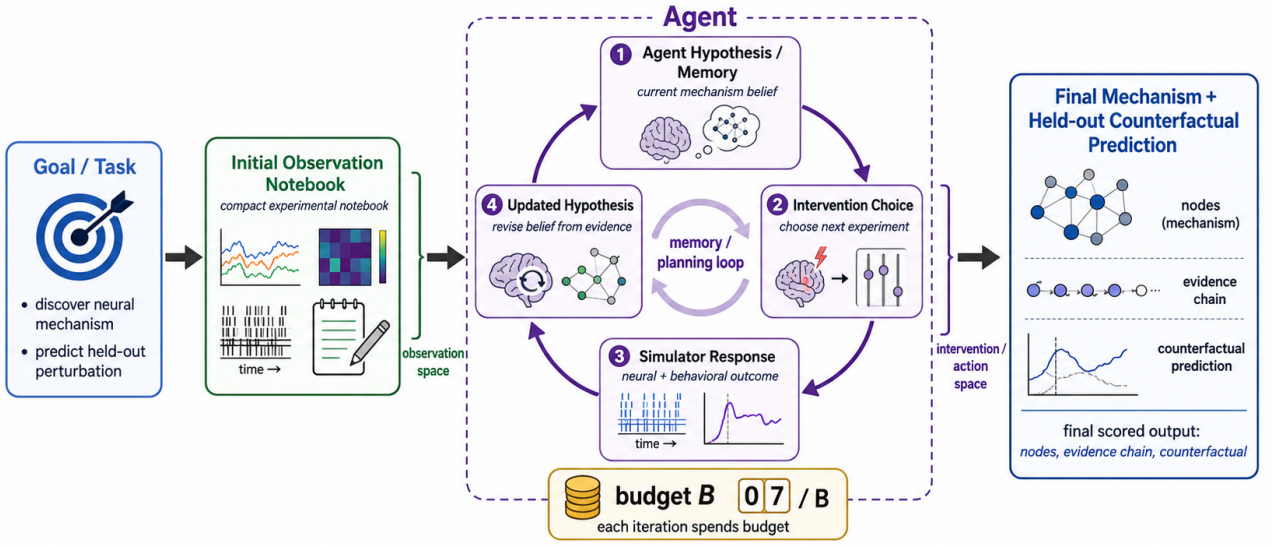


Figure 1. Digital-fly discovery loop. The agent observes compressed neural and behavioral readouts, chooses budgeted interventions, updates its mechanism hypothesis, and returns an evidence-supported explanation plus a held-out counterfactual prediction.

ward without learning an interpretable explanation. Third, a retrieval-heavy system may state the names of published neurons without grounding that answer in the evidence available during the benchmark episode. Our benchmark instead asks for a compact mechanism, an explicit evidence chain, and a counterfactual claim that can be checked under a held-out perturbation. That standard is closer to the logic of intervention-based causal inference than to ordinary supervised prediction (Peters et al., 2017).

### 2.2. Task families and outputs

The broader benchmark is intended to cover at least five task families. Mechanism localization asks the agent to identify the smallest node set that explains a behavior or intermediate computation. Causal attribution asks which components contribute positively or negatively and, in future versions, which edges or paths transmit those effects. Intervention design asks the agent to spend a limited budget on the experiments that best discriminate among competing explanations. Counterfactual prediction evaluates whether the learned explanation generalizes to held-out perturbations. Cross-context transfer asks whether a mechanism inferred in one sensory or behavioral condition still predicts behavior in another.

The pilot instantiation in this paper touches all five families, although unevenly. Node-set recovery is the primary scored target. Held-out interventions provide a first counterfactual test. Budgeted interaction supports intervention-design evaluation. The named-versus-anonymized condition addresses shortcut robustness, and the JO-CE versus JO-F contrast probes whether the same candidate circuit is interpreted differently across contexts. We do not yet claim full edge-level causal attribution; that is precisely one of the main gaps exposed by the pilot.

The required final output is therefore more structured than a free-form paragraph. For each task, the agent must return a set of hypothesized mechanism nodes, optional edges, a confidence score, a brief natural-language explanation, and a held-out experimental prediction. In future releases, this structure should make it possible to score both symbolic overlap with ground truth and the quality of the scientific narrative used to justify it.

### 2.3. Evaluation principles

A useful benchmark for scientific agents should reflect not only whether the final answer is correct, but also how that answer was obtained. We therefore treat evaluation as a vector rather than a single scalar. The pilot reports Node F1, held-out counterfactual accuracy, budget AUC, and Brier score, and logs the full experiment trace for qualitative inspection. Conceptually, the benchmark score can be written as

$$\text{Score} = (F1_{\text{node}},\ Acc_{\text{cf}},\ AUC_B,\ \text{Brier},\ Q_{\text{trace}}),$$

where $Q_{\text{trace}}$ denotes process quality, such as whether the agent chooses discriminative interventions and revises hypotheses after contradictory evidence.

Three design principles follow from this view. First,

the benchmark should be budgeted: exhaustive search is not the behavior we want to reward. Second, it should be multi-turn: mechanism discovery is an iterative process, not a single answer. Third, it should be shortcut-aware: biological names, cached summaries, or superficial activity salience should not be enough to win. These principles closely match the concerns raised in recent evaluations of scientific and bioinformatics agents (Chen et al., 2024; Fa et al., 2026; Abram, 2026).

## 3. Benchmark Instantiation in a Digital Fly

### 3.1. Why this substrate?

The digital fly is attractive because it sits in a rare part of the design space. It is much richer than a toy causal graph, but still far more controllable than wet-lab biology. The whole-brain fly connectome provides a global structural scaffold (Dorkenwald et al., 2024); connectome-constrained models show that anatomy can support predictive dynamical modeling (Lappalainen et al., 2024); and the Shiu et al. model demonstrates that a brain-wide simulator can recover compact sensorimotor hypotheses for feeding and grooming (Shiu et al., 2024). As a result, the substrate combines known anatomy, executable interventions, and published biological targets.

This is also why the fly is a better fit for the present paper than a generic sandbox simulator. The benchmark is intended to measure whether an agent can perform a biologically meaningful form of reasoning. The richer the substrate, the more likely it is that the benchmark rewards hypothesis testing instead of brittle prompt engineering. At the same time, neuromechanical-digital-twin work suggests a path outward from the current pilot: once the discovery loop is working on brain-scale circuit tasks, future versions can add embodiment, environment interaction, and brain–body coupling without abandoning the same evaluation logic (Wang-Chen et al., 2024; Wang-Chen & Ramdya, 2026).

### 3.2. Pilot tasks and ground truth

Our pilot uses the whole-brain Drosophila model of Shiu et al. (2024). We ran the benchmark on an archived 630-neuron completeness release because it preserved the paper-specific grooming identifiers needed for the selected tasks. After initial sanity checks, we used a 100 ms observation window, which was more stable than a shorter 40 ms window on the feeding_water task.

We instantiated six tasks spanning feeding and grooming: feeding with sugar input, feeding with water input, feeding with combined sugar-plus-water input, grooming under full JON stimulation, grooming under JO-CE stimulation, and grooming under JO-F stimulation. Feeding tasks were built around locally verified groups including G2N_1, rattle, usnea, FMIn, bract, Fdg, roundup, clavicle, and MN9. Grooming tasks used JON_CE, JON_F, JON_full, aBN1, aDN1_bilateral, aDN2_left, and the published JO-F inhibitory bundle. To prevent simple lookup, we froze an anonymized manifest in which the agent saw opaque tokens such as N004; human-readable names were stored separately in an ID map. We also augmented candidate lists with high-response non-core decoys drawn from the initial context response.

Each task included a held-out causal check. For example, the feeding-water task held out G2N_1, the combined feeding task held out FMIn, and the JO-F grooming task held out the JO-F inhibitory bundle. The goal was not only to recover a node set but also to propose a counterfactual intervention consistent with the recovered mechanism. This choice makes the benchmark stricter than a post hoc explanation task: a mechanism that sounds plausible but fails on the held-out perturbation is incomplete by construction.

### 3.3. Observation interface, baselines, and protocol

The simulator wrapper compressed raw model outputs into a compact experimental notebook containing target firing rates, candidate-node deltas, top changed nodes, and a short text summary. This design deliberately restricts the agent to a scientist-like observation channel rather than a full simulator dump. In practice, that matters for benchmark validity. Unrestricted simulator access would make it easier to key off implementation artifacts or exhaustive search rather than to reason from the same level of evidence a scientist would plausibly inspect.

We used three seeds and four experiment budgets $B \in \{2, 4, 6, 8\}$. The main comparison included random and greedy automated baselines, a one-shot direct_llm baseline using only the initial observation, and a full planning agent that produced a hypothesis–experiment–update trajectory. We also ran an agentic ablation with four LLM-style variants: direct_llm, no_memory, memory, and planning.

### 3.4. Metrics and contamination controls

Node F1 is the primary metric in the present paper because it is the most stable quantity across all six tasks and all baseline variants. We also track held-out

counterfactual accuracy, budget AUC, and Brier score, but edge-level mechanism scoring is not yet available because the current hypothesis schema leaves hypothesis_edges empty. This is a meaningful limitation, but it is also informative: the pilot identifies exactly where symbolic mechanism evaluation needs to improve next.

We additionally built the pilot around two shortcut-resistance devices. First, anonymization tests whether a method relies on meaningful experimental evidence or on biologically suggestive names. Second, decoy nodes test whether a method equates strong first-pass activation with mechanism membership. Both devices are motivated by recent concerns that scientific-agent benchmarks can be gamed through contamination, retrieval, or other superficial cues rather than genuine reasoning (Chen et al., 2024; Fa et al., 2026; Abram, 2026).

## 4. Pilot Experiments

### 4.1. Main benchmark comparison

Table 1 reports mean Node F1 across the six pilot tasks. The main result is that sequential planning substantially outperforms the reference baselines and reaches near-perfect recovery with modest budgets. At $B = 2$, planning already attains 0.856 mean Node F1; by $B = 4$ it rises to 0.952; and by $B = 6$ it reaches 1.000 and remains there at $B = 8$. In contrast, the one-shot GPT-5.4 stand-in remains at 0.532, only modestly above the random reference (0.494) and well above the greedy heuristic (0.412). This gap is exactly what the benchmark is supposed to reveal: plausible one-shot explanation is not the same as budgeted causal discovery.

A second takeaway is that the digital-fly tasks are solvable but not trivial. If the benchmark were too easy, one-shot prediction would saturate immediately. If it were too hard, even planning would fail under modest budgets. Instead, the pilot yields a useful regime in which richer discovery loops provide a clear advantage. The budget curve is especially important here: the gain from 0.856 to 1.000 shows that additional experiments are useful, but only when the agent can exploit them intelligently.

### 4.2. Interpreting the task family mix

The six pilot tasks are intentionally not interchangeable. The feeding tasks emphasize convergence from gustatory input toward a compact feeding-initiation pathway and motor output, while the grooming tasks include subtype-specific routing and inhibition. That distinction matters because the second family is harder to solve from static activity summaries alone. In the published simulator study, JO-CE and JO-F stimulation do not behave identically; the mechanistic explanation involves the way different mechanosensory subtypes interact with downstream relays and inhibition (Shiu et al., 2024). A benchmark that only rewarded first-pass salience would miss that distinction.

This makes the pilot more than a generic node-ranking exercise. Some tasks are close to "which nodes propagate a sensory drive to an output?" Others are closer to "why does one context fail to recruit the same downstream pathway despite related anatomy?" The latter is where sequential experimentation is most valuable, because it gives the agent an opportunity to test absence-of-effect explanations rather than only follow the largest positive responses.

### 4.3. Ablating memory and planning

The right half of Table 1 decomposes which parts of the agentic loop matter. A sequential policy without notebook memory improves only modestly over one-shot inference, moving from 0.532 to 0.578 at $B = 2$ and plateauing at 0.667 by $B = 4$. Adding memory produces a much larger gain: 0.789 at $B = 2$ and 0.897 by $B = 4$. Full planning improves further, reaching 1.000 by $B = 6$. This pattern supports the paper's core claim that mechanism discovery is not well captured by a single free-form answer. Persistent state matters, and actively selecting discriminative interventions matters even more.

The plateau of the no-memory variant is especially informative. Additional experiments are available, but their marginal value collapses when the agent cannot maintain an explicit record of which hypotheses have already been supported or weakened. The memory-only variant does better because it can accumulate evidence, yet it still leaves gains on the table relative to planning. In other words, the benchmark is sensitive to the very ingredients that should define an AI scientist: not just access to a model, but a loop that remembers prior evidence, chooses what to test next, and revises its hypothesis over time.

### 4.4. Robustness to naming shortcuts

A strong mechanistic benchmark should penalize shortcutting. To test this, we compared a named condition, in which the agent saw labels such as aBN1 or MN9, with an anonymized condition, in which all groups were replaced by opaque tokens. Table 2 shows a sharp contrast. The one-shot baseline scores 1.000 Node F1 in the named condition but falls to 0.532 after anonymization. Planning, by contrast, remains at

Table 1. Mean Node F1 on the six pilot tasks. Random and greedy are budget-insensitive reference baselines in this pilot because their implementation reused one final heuristic summary at each budget rather than recomputing a prefix-specific answer.

| Main comparison | | | | | Agentic ablation | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | $B$=2 | $B$=4 | $B$=6 | $B$=8 | Method | $B$=2 | $B$=4 | $B$=6 | $B$=8 |
| Random | 0.494 | 0.494 | 0.494 | 0.494 | Direct LLM | 0.532 | 0.532 | 0.532 | 0.532 |
| Greedy | 0.412 | 0.412 | 0.412 | 0.412 | No memory | 0.578 | 0.667 | 0.667 | 0.667 |
| Direct LLM | 0.532 | 0.532 | 0.532 | 0.532 | Memory | 0.789 | 0.897 | 0.897 | 0.897 |
| Planning | 0.856 | 0.952 | 1.000 | 1.000 | Planning | 0.856 | 0.952 | 1.000 | 1.000 |

Table 2. Named versus anonymized mean Node F1.

| Method | Named | Anonymized |
|---|---|---|
| Direct LLM | 1.000 | 0.532 |
| Planning | 1.000 | 1.000 |

1.000 in both settings.

This is a desirable property of the benchmark. Biological nomenclature can leak function, especially in small pilot tasks derived from published circuits. An evaluation that never checks for this leakage risks rewarding systems for recognizing names rather than interpreting evidence. The anonymization result therefore does double duty: it is both a robustness check on the methods and a validity check on the benchmark itself.

### 4.5. Qualitative case study

The grooming_JOF trace illustrates the intended discovery loop. The task asks the agent to explain the discrepancy between JO-CE and JO-F stimulation and to reason about the role of the JO-F inhibitory bundle documented in the simulator substrate (Shiu et al., 2024). The saved trace shows a planning-style trajectory rather than a static answer: the agent starts with a broad antennal-grooming hypothesis, uses targeted interventions to test competing explanations, and revises its mechanism before making the held-out counterfactual prediction.

This example is scientifically revealing because it is not only about finding "the most active nodes." A plausible but weak strategy would simply report whichever grooming-related units respond most strongly to the first observation. The stronger strategy is to use the discrepancy between contexts to ask what kind of hidden causal story could explain it: missing excitation, active suppression, or convergent downstream compensation. That is the kind of distinction a benchmark for mechanism discovery should surface.

### 4.6. What these pilot experiments do and do not show

The positive interpretation is that the benchmark appears to be measuring something nontrivial. The tasks are grounded in a real digital-fly substrate; the planning agent benefits from extra budget; memory and experiment selection matter; and anonymization reveals a concrete shortcut. The results therefore support the paper's benchmark claim: this environment can separate qualitatively different reasoning styles rather than merely rewarding verbosity or prior biological name knowledge.

The negative interpretation is equally important: these numbers are not yet a final leaderboard. Edge-level outputs are currently absent, and some reference baselines need prefix-specific budget recomputation. For those reasons, we present the results as a pilot benchmark instantiation and focus on the benchmark's discriminative behavior rather than on absolute method ranking.

## 5. Discussion and Limitations

The pilot results support the central premise of the paper: a digital-fly benchmark can expose meaningful differences in agentic biological discovery. In particular, three observations are encouraging. First, there is a large gap between one-shot explanation and sequential experimentation. Second, memory is not enough on its own; the experiment policy also matters. Third, name anonymization is an informative stress test rather than a nuisance, because it detects exactly the sort of biological shorthand that could otherwise inflate performance.

More broadly, the benchmark helps clarify what should count as progress in biological AI agents. Pipeline completion and manuscript drafting are useful, but they are not yet evidence of mechanistic reasoning. A system that recovers a compact circuit under a budget, supports it with interventions, and predicts a held-out perturbation is much closer to the standard scientists care about. In that sense, our re-

sults complement recent biology-agent demonstrations rather than compete with them: work such as Virtual Lab, GeneAgent, and CellVoyager shows that agentic workflows can already be useful, whereas our benchmark asks whether they are useful for a harder epistemic goal (Swanson et al., 2025; Wang et al., 2025; Alber et al., 2026).

At the same time, several limitations are important for any honest reading. Construct validity. The current output schema leaves hypothesis_edges empty, so we report node-level mechanism recovery instead of edge/path F1. Baseline validity. The current random and greedy implementations reused a final heuristic summary at each budget, so their flat curves should be treated as reference orderings rather than definitive budget traces. Biological coverage. Not all originally planned named groups were mapped cleanly in the local repo snapshot; for example, several feeding nodes and aBN2 were omitted. These are exactly the kinds of benchmark-design details that recent scientific-agent evaluation work urges the community to expose rather than hide (Chen et al., 2024; Abram, 2026).

There is also an important question of external validity. A digital fly is not a living fly. The substrate inherits the simplifications of the underlying simulator, including incomplete treatment of neuromodulation, gap junctions, morphology, and other factors that matter for real circuit function (Shiu et al., 2024). More generally, digital-twin work argues that simulation is most useful when treated as a hypothesis-generation partner rather than as a final arbiter of biological truth (Wang-Chen & Ramdya, 2026). For that reason, we view benchmark success as evidence about agentic reasoning under biologically structured constraints, not as evidence that the same agent is ready for unsupervised wet-lab autonomy.

These limitations point directly to the next version of the benchmark. A stronger release should execute live agent loops, recompute all baselines at each budget prefix, add edge-level scoring, and extend the task family to a second substrate such as a connectome-constrained visual-system model or an embodied NeuroMechFly-style environment (Lappalainen et al., 2024; Wang-Chen et al., 2024). It should also include stricter robustness sweeps over decoys, corrupted observations, and prompt bloat, following the spirit of recent scientific-agent and bioinformatics-agent benchmarks (Fa et al., 2026). Even in its current form, however, the benchmark already serves a useful purpose: it makes the question "can an AI scientist discover a neural mechanism?" operational enough to test, critique, and improve.

Finally, the paper suggests a narrower and more defensible narrative for AI scientists in biology. The right near-term goal is not a fully autonomous biologist, but a system whose reasoning can be inspected, stressed, and falsified. Mechanism-centered benchmarks can contribute to that goal because they reward auditable experiment traces and punish shallow shortcutting. In a period when agentic-science claims are expanding quickly, that kind of evidentiary discipline is itself a useful contribution (Xin et al., 2025; Nature Machine Intelligence, 2026).

## 6. Conclusion

We argued that the right question for agentic AI in biology is not whether a system can produce biology-sounding text, but whether it can iteratively uncover a mechanism in a structured biological environment. We proposed a benchmark formulation for this problem and instantiated it on six pilot tasks in a digital fly. The pilot results suggest that the benchmark is sensitive to memory, planning, and shortcut robustness, which are precisely the capabilities that should matter for biological discovery. We hope this benchmark can help move the field from broad claims about scientific autonomy toward more rigorous, mechanism-centered evaluation.

## Impact Statement

This paper proposes a benchmark for evaluating agentic biological discovery rather than a deployment-ready autonomous scientist. Its positive potential impact is to raise the evidentiary bar for claims about biological reasoning by emphasizing interventions, counterfactuals, calibration, and transparent experiment traces. At the same time, benchmark success could be over-interpreted if simulator performance is treated as evidence that a system is ready for unsupervised wet-lab science. We therefore stress that this work is diagnostic, not certifying. More broadly, stronger agentic-science benchmarks may accelerate useful scientific tooling, but they also raise familiar concerns about automation overreach, misplaced trust, and premature delegation of high-stakes reasoning (Xin et al., 2025; Nature Machine Intelligence, 2026). Our aim is to support more responsible evaluation by making biological mechanism discovery harder to fake and easier to audit.

## References

Abram, M. Toward evaluation frameworks for multi-agent scientific AI systems. arXiv preprint

arXiv:2603.26718, 2026. doi: 10.48550/arXiv.2603.26718. URL https://arxiv.org/abs/2603.26718.

Alber, S., Chen, B., Sun, E., Isakova, A., Wilk, A. J., and Zou, J. Cellvoyager: AI compbio agent generates new insights by autonomously analyzing biological data. Nature Methods, 23:749–759, 2026. doi: 10.1038/s41592-026-03029-6. URL https://www.nature.com/articles/s41592-026-03029-6.

Chen, Z., Chen, S., Ning, Y., Zhang, Q., Wang, B., Yu, B., Li, Y., Liao, Z., Wei, C., Lu, Z., Dey, V., Xue, M., Baker, F. N., Burns, B., Adu-Ampratwum, D., Huang, X., Ning, X., Gao, S., Su, Y., and Sun, H. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. arXiv preprint arXiv:2410.05080, 2024. doi: 10.48550/arXiv.2410.05080. URL https://arxiv.org/abs/2410.05080.

Dorkenwald, S., Matsliah, A., Sterling, A. R., et al. Neuronal wiring diagram of an adult brain. Nature, 634:124–138, 2024. doi: 10.1038/s41586-024-07558-y. URL https://www.nature.com/articles/s41586-024-07558-y.

Fa, D.,
vCuljak, M., Pand
vza, B., and
vCupi
'c, M. Bioagent bench: An AI agent evaluation suite for bioinformatics. arXiv preprint arXiv:2601.21800, 2026. doi: 10.48550/arXiv.2601.21800. URL https://arxiv.org/abs/2601.21800.

GenBio 2026 Workshop Organizers. The 2026 workshop on generative and agentic AI for biology. https://genbio-workshop.github.io/2026/, 2026. Accessed 2026-04-13.

Gridach, M., Nanavati, J., Zine El Abidine, K., Mendes, L., and Mack, C. Agentic AI for scientific discovery: A survey of progress, challenges, and future directions. arXiv preprint arXiv:2503.08979, 2025. doi: 10.48550/arXiv.2503.08979. URL https://arxiv.org/abs/2503.08979.

Huang, S., Lang, M., Chen, Z., Yang, C., Huang, X., Mohtashaminia, Z., and Peng, Y. From foundation models to autonomous agents in biology. Genomics Communications, 3:e006, 2026. doi: 10.48130/gcomm-0026-0005. URL https://www.maxapress.com/article/doi/10.48130/gcomm-0026-0005.

Lappalainen, J. K., Tschopp, F. D., Prakhya, S., McGill, M., Nern, A., Shinomiya, K., Takemura, S.-y., Gruntman, E., Macke, J. H., and Turaga, S. C. Connectome-constrained networks predict neural activity across the fly visual system. Nature, 634:1132–1140, 2024. doi: 10.1038/s41586-024-07939-3. URL https://www.nature.com/articles/s41586-024-07939-3.

Lobato-Rios, V., Ramalingasetty, S. T., Ozdil, P. G., Arreguit, J., Ijspeert, A. J., and Ramdya, P. Neuromechfly, a neuromechanical model of adult Drosophila melanogaster. Nature Methods, 19:620–627, 2022. doi: 10.1038/s41592-022-01466-7. URL https://www.nature.com/articles/s41592-022-01466-7.

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. doi: 10.48550/arXiv.2408.06292. URL https://arxiv.org/abs/2408.06292.

Nature Machine Intelligence. Multi-agent AI systems need transparency. Nature Machine Intelligence, 8:1, 2026. doi: 10.1038/s42256-026-01183-2. URL https://www.nature.com/articles/s42256-026-01183-2.

Peters, J., Janzing, D., and Schölkopf, B. Elements of Causal Inference: Foundations and Learning Algorithms. The MIT Press, Cambridge, MA, 2017. ISBN 9780262037310. URL https://mitpress.mit.edu/9780262037310/elements-of-causal-inference/.

Qi, C., Wang, W., Jiang, S., Liu, Q., Song, X., Fang, H., and Wei, Z. Artificial intelligence agents for biological research: A survey. Briefings in Bioinformatics, 27(1):bbag075, 2026. doi: 10.1093/bib/bbag075. URL https://academic.oup.com/bib/article/27/1/bbag075/8499367.

Shiu, P. K., Sterne, G. R., Spiller, N., Franconville, R., Sandoval, A., Zhou, J., Simha, N., et al. A Drosophila computational brain model reveals sensorimotor processing. Nature, 634:210–219, 2024. doi: 10.1038/s41586-024-07763-9. URL https://www.nature.com/articles/s41586-024-07763-9.

Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E., and Zou, J. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. Nature, 646:716–723, 2025. doi: 10.1038/s41586-025-09442-9. URL https://www.nature.com/articles/s41586-025-09442-9.

Tang, L. Artificial intelligence agents for biology. Nature Methods, 22:2496–2497, 2025. doi: 10.1038/s41592-025-02958-y. URL https://www.nature.com/articles/s41592-025-02958-y.

Wang, Z., Jin, Q., Wei, C.-H., Tian, S., Lai, P.-T., Zhu, Q., Day, C.-P., Ross, C., Leaman, R., and Lu, Z. Geneagent: Self-verification language agent for gene-set analysis using domain databases. Nature Methods, 22:1677–1685, 2025. doi: 10.1038/s41592-025-02748-6. URL https://www.nature.com/articles/s41592-025-02748-6.

Wang-Chen, S. and Ramdya, P. The embodied brain: Bridging the brain, body, and behavior with neuromechanical digital twins. arXiv preprint arXiv:2601.08056, 2026. doi: 10.48550/arXiv.2601.08056. URL https://arxiv.org/abs/2601.08056.

Wang-Chen, S., Stimpfling, V. A., Lam, T. K. C., Ozdil, P. G., Genoud, L., Hurtak, F., and Ramdya, P. Neuromechfly v2: Simulating embodied sensorimotor control in adult Drosophila. Nature Methods, 21:2353–2362, 2024. doi: 10.1038/s41592-024-02497-y. URL https://www.nature.com/articles/s41592-024-02497-y.

Wei, J., Yang, Y., Zhang, X., Chen, Y., Zhuang, X., Gao, Z., Zhou, D., Wang, G., Gao, Z., Cao, J., Qiu, Z., He, X., Zhang, Q., You, C., Zheng, S., Ding, N., Ouyang, W., Dong, N., Cheng, Y., Sun, S., Bai, L., and Zhou, B. From AI for science to agentic science: A survey on autonomous scientific discovery. arXiv preprint arXiv:2508.14111, 2025. doi: 10.48550/arXiv.2508.14111. URL https://arxiv.org/abs/2508.14111.

Xiao, Y., Liu, J., Zheng, Y., Xie, X., Hao, J., Li, M., Wang, R., Ni, F., Li, Y., Luo, J., Jiao, S., and Peng, J. Cellagent: An LLM-driven multi-agent framework for automated single-cell data analysis. arXiv preprint arXiv:2407.09811, 2024. doi: 10.48550/arXiv.2407.09811. URL https://arxiv.org/abs/2407.09811.

Xin, H., Kitchin, J. R., and Kulik, H. J. Towards agentic science for advancing scientific discovery. Nature Machine Intelligence, 7:1373–1375, 2025. doi: 10.1038/s42256-025-01110-x. URL https://www.nature.com/articles/s42256-025-01110-x.

Yamada, Y., Lange, R. T., Lu, C., Hu, S., Lu, C., Foerster, J., Clune, J., and Ha, D. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. arXiv preprint arXiv:2504.08066, 2025. doi: 10.48550/arXiv.2504.08066. URL https://arxiv.org/abs/2504.08066.