EVOLAB: FEEDBACK-DRIVEN AGENTS FOR BUDGETED WET-LAB DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific agents increasingly use feedback to choose experiments, but a finite measurement budget makes each choice consequential twice: it yields discoveries now and determines which evidence can inform later decisions. We present EVOLAB, which couples budgeted measurement selection with the interpretation and use of selectively acquired evidence. After each batch, it reviews what the observations change about the remaining decision problem, foregrounds provenance-linked evidence while retaining a retrievable record, and investigates feasible next allocations with numerical and retrieval tools. Its decision context can evolve while language-model weights remain fixed; outcomes of unmeasured alternatives remain hidden. We evaluate EVOLAB by sequential replay in molecular hit discovery, protein library design, and two drug-combination screening settings. Across these four reported settings, it improves mean final discovery yield over task-adapted ReAct by approximately 5–17% and reduces token consumption by 12–33%. Trajectories show delayed gains after prerequisite measurements in Hit and GFP and more productive revisits in Drug C2; a full-information Hit oracle quantifies remaining allocation headroom. These results support designing what to measure and how acquired evidence shapes the next decision together.

1 INTRODUCTION

Scientific agents are beginning to participate in closed-loop experimentation. From early robot scientists to automated chemistry and materials platforms, laboratory systems have increasingly connected planning, execution, and measurement (King et al., 2004; Burger et al., 2020; Szymanski et al., 2023). Language-model agents can now search scientific resources, operate tools, design experiments using earlier results, and assemble biomedical workflows (Boiko et al., 2023; Huang et al., 2026; Roohani et al., 2025). These advances sharpen a practical question: when only a limited number of experiments can be performed, how should an agent use evidence created by its own earlier choices?

The scarce resource is the measurement budget (Häse et al., 2019). Unlike a static dataset, the experimental record is policy-dependent: the selected batch determines which outcomes become visible in the next round. Untested alternatives are missing evidence, not negative results. In protein engineering, measured properties guide later variant batches; in drug-combination screening, observed responses guide later dose choices (Yang et al., 2019; Tosh et al., 2025; Gao et al., 2026). Figure 1 illustrates how evidence accumulates selectively while the remaining budget contracts. A measurement can therefore matter both for what it discovers immediately and for which follow-ups it makes worthwhile or feasible. Budgeted experimentation requires choosing among future information states as well as candidates.

Existing approaches address parts of this loop. Numerical acquisition methods choose measurements under uncertainty and cost (Kirsch et al., 2019; Kandasamy et al., 2017; Tosh et al., 2025), and LMABO adapts the choice of acquisition function (Ngo et al., 2026). Coscientist and BioDiscoveryAgent incorporate observed outcomes into subsequent experimentation, while Biomni combines biomedical reasoning with tools (Boiko et al., 2023; Roohani et al., 2025; Huang et al., 2026). Reflexion and ExpeL retain interpretations of experience without changing model weights (Shinn et al., 2023; Zhao et al., 2024). Yet presenting previous outcomes does not establish how they influence the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

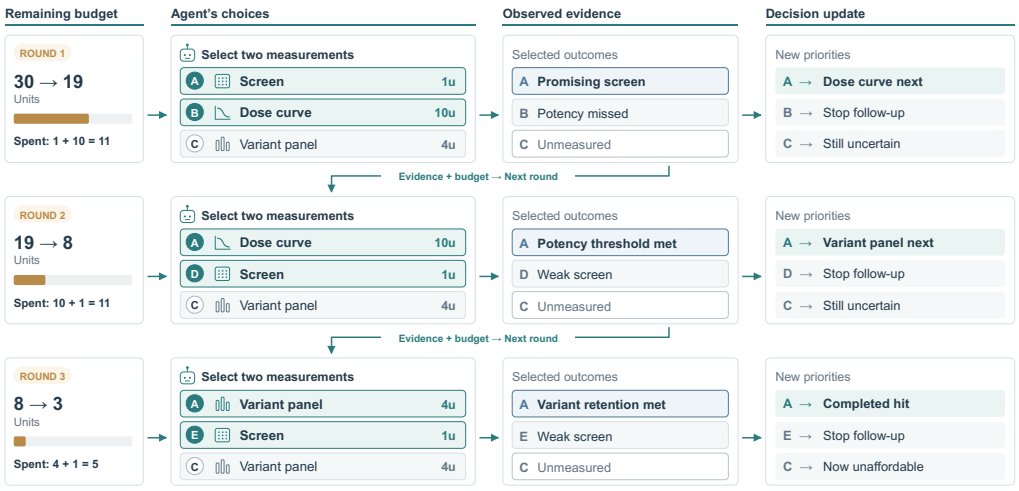


Figure 1: **Budgeted choices shape a sequence of experimental decisions.** In each round, the agent selects measurements with different costs and observes only their outcomes. Favorable evidence advances candidate A toward a completed hit, while weak evidence stops other follow-up. The shrinking budget eventually makes the deferred panel for C unaffordable. Filled badges mark selected measurements and hollow badges mark deferred alternatives. Evidence and remaining budget jointly inform the next batch. Candidate outcomes and budget values are schematic. B and C have prior screening records, and each illustrated panel contains four measurements.

next allocation. The evidence an agent can interpret was selectively created by its own earlier actions; *what should it retain, examine, and act on next?*

This problem has three practical features. Observations depend on earlier allocations, so an unmeasured outcome must remain distinct from a measured failure. The acquired record grows while deliberation remains bounded, requiring a focused view with access to the complete record. And experimental value can be delayed: a cheap prerequisite or an unresolved result may change which costly follow-up is worthwhile. We introduce EvoLab to couple two decisions: which measurements to acquire and how their outcomes shape the next allocation. After each batch, it reviews what newly acquired evidence changes about remaining opportunities, foregrounds relevant records and interpretations with links to their provenance, and carries the decision context into the next round. Retrieval and numerical tools support investigation of feasible next batches before commitment; only committed measurements reveal outcomes. Model weights remain fixed as the decision state evolves. Across three replay domains, we compare EvoLab with task-adapted ReAct (Yao et al., 2023) and BioDiscoveryAgent (Roohani et al., 2025).

Our contributions are:

- **Budgeted experimentation as information allocation.** We articulate the coupled decision in a finite campaign: each batch determines immediate discovery yield and the outcomes available for later choices. Costs and prerequisites make preparatory measurements valuable even when they complete no discovery in the current round. Unmeasured alternatives remain unknown rather than becoming presumed failures.
- **Acquisition coupled with evidence use.** EvoLab turns acquired feedback into a question about unfinished opportunities and their feasible completion paths. It connects the question to supporting records, optional provenance-linked notes, and tool-assisted comparison of next batches. The full acquired record remains accessible, while model weights stay fixed and forecasts remain distinct from observations.
- **Sequential evidence across scientific tasks.** Across Hit, GFP, and two Drug settings, EvoLab achieves higher mean final yield than task-adapted ReAct with lower token consumption. Trajectories reveal delayed returns; the Hit oracle and ablations expose remaining headroom and task-dependent note value.

2 Related Work

Numerical experimental allocation. Bayesian acquisition methods select informative or promising measurements under uncertainty, including cost-aware settings (Kirsch et al., 2019; Kandasamy et al., 2017; Tosh et al., 2025). LMABO chooses an acquisition function from a portfolio given the optimization state (Ngo et al., 2026). Our setting additionally asks how a measured result changes unfinished opportunities, including discoveries that require several assays with different costs and prerequisites. EvoLab can inspect numerical proposals, but relates them to acquired records, the task's completion criterion, and the remaining budget before selecting the next batch.

Biomedical agents. Biomedical agents combine scientific knowledge with tool use and experimental feedback. ReAct (Yao et al., 2023) interleaves reasoning with environment interaction. Coscientist (Boiko et al., 2023) integrates chemical planning and outcome-guided optimization; Biomni (Huang et al., 2026) supports biomedical workflows through retrieval and code execution; BioDiscoveryAgent (Roohani et al., 2025) uses prior results and biological knowledge to design genetic perturbations. These systems demonstrate how agents can act on scientific feedback. In our finite-budget campaigns, an agent also decides which outcomes to acquire, then reasons from only the evidence it actually purchased. We compare against task-adapted agents under shared catalogs, budgets, prerequisites, and observation rules.

Feedback-driven agent improvement. Agents can improve from experience by adapting context, tools, or decision strategies. Reflexion (Shinn et al., 2023) and ExpeL (Zhao et al., 2024) retain reflections and reusable insights; ACE (Zhang et al., 2025b) maintains contextual playbooks, and ALMA (Xiong et al., 2026) searches over memory designs. ADAS (Hu et al., 2025), AFlow (Zhang et al., 2025a), and AgentOptimizer (Zhang et al., 2024) adapt workflows or functions using evaluation feedback. EvoLab likewise uses external, revisable state with fixed model weights, but its state is tied to the measurement policy within the current campaign. Decision questions and supporting records evolve from selectively acquired outcomes. Optional notes retain interpretations with links to those records; the full archive stays accessible so later evidence can challenge an earlier interpretation.

3 EvoLab: Coupling Measurement and Evidence Use

At each round, EvoLab makes an evidence-use decision before a measurement decision. It first examines how newly acquired outcomes change the remaining opportunities, identifies a question for the next allocation, and selects the records needed to investigate it. It then compares feasible follow-ups and commits a batch. Figure 2 shows this loop. The two decisions are coupled: the selected batch determines which evidence can be acquired, while the interpretation of acquired evidence guides the next batch. The language model's weights remain fixed throughout the campaign.

3.1 Experimental choices shape future evidence

A campaign starts with public candidate information, initial measurements $\mathcal{D}_1$, and an additional measurement budget. Before round t, the agent has acquired records $\mathcal{D}_t$ and remaining budget b_t. A feasible batch a_t incurs domain-specific cost $c_d(a_t)$; only that batch reveals new outcomes:

$$\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(i, y_i, t) : i \in a_t\}, \qquad b_{t+1} = b_t - c_d(a_t). \tag{1}$$

Let $G_d(\mathcal{D})$ count discoveries certified by acquired observations under task d's endpoint. The immediate gain from a committed batch is

$$g_t = G_d(\mathcal{D}_{t+1}) - G_d(\mathcal{D}_t). \tag{2}$$

The objective is final discovery yield, although a useful batch may have $g_t = 0$. In Hit, a variant panel can justify a subsequent ten-unit dose curve without certifying a hit on its own. Thus, a purchase can change the evidence for later choices before it changes the discovery count. This connects the setting to cost-aware acquisition (Kandasamy et al., 2017; Tosh et al., 2025). We evaluate by sequential replay: public metadata and acquired outcomes are visible, but unpurchased outcomes remain hidden. Predictions from numerical tools are not new observations.

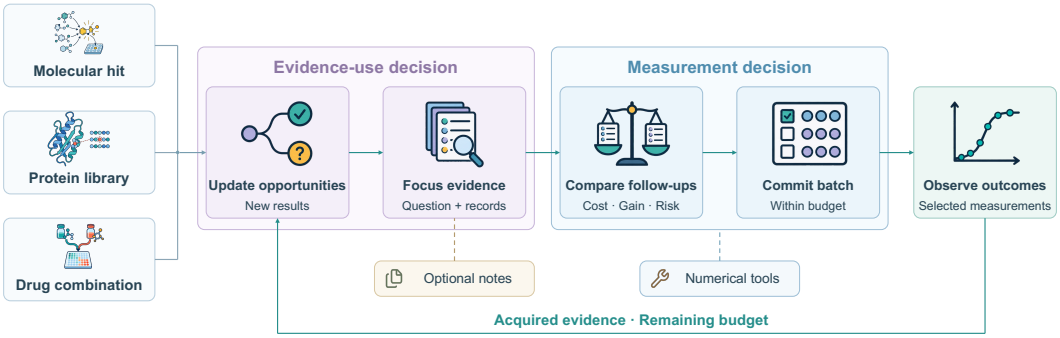


Figure 2: **Coupling measurement and evidence use in EvoLab.** Each committed batch reveals selected outcomes and reduces the remaining budget. The agent reviews what those outcomes change about unfinished opportunities, foregrounds relevant acquired records, and investigates feasible follow-ups before committing the next batch. The full acquired record remains retrievable; forecasts do not reveal unmeasured outcomes.

3.2 From acquired outcomes to a decision state

The first review uses initialization measurements; later reviews use the preceding batch's feedback together with the existing record and remaining budget. Rather than summarizing the batch alone, the review identifies a decision-relevant question about what can still be completed or ruled out. In molecular hit discovery, for example, a passing variant panel can motivate an expensive dose curve, whereas a failed panel can redirect spending elsewhere. In protein library design, measured constituents can make an unfinished multi-mutant project feasible. These examples illustrate why a zero-yield observation can still change the next allocation.

The record and remaining budget delimit the batches that can be purchased:

$$\mathcal{A}_t = \{a : c_d(a) \le b_t,\ \mathrm{legal}_d(a \mid \mathcal{D}_t, t) = 1\}. \tag{3}$$

Here legal_d checks the task's prerequisites and round rules for the batch as a whole: a Hit panel needs a screen from an earlier round, whereas GFP constituents and their multi-mutant can be measured together. Membership in $\mathcal{A}_t$ says nothing about predicted success. A failed dose can leave a Drug grid unresolved, whereas a discovered grid cannot contribute another count. The review therefore asks which acquired observations support an unfinished opportunity and which affordable follow-up could complete it. This interpretation uses the task's discovery criterion rather than a generic summary of feedback.

Together, g_t and $\mathcal{A}_t$ separate immediate certification from what the campaign can still pursue. A prerequisite can leave $g_t = 0$ yet provide evidence for an affordable completion path in a later round. Conversely, a promising candidate may be impossible to finish within the remaining budget, even if the next assay is legal. This is why the next decision considers both the acquired evidence and the cost of completing an opportunity. EvoLab does not assume that this tradeoff can be computed exactly from unknown outcomes.

The review also determines which acquired evidence to foreground for that question. The complete acquired record remains accessible. Task-specific views connect measurements to unfinished opportunities, while evidence-linked notes can retain a question and its supporting records alongside the remaining budget. In the single-loop configuration, all acquired raw records are presented; foregrounding evidence therefore organizes attention rather than withholding other records. This is an evidence-use decision: it changes attention to the acquired record without changing the record or treating unmeasured alternatives as negative results. Unlike carrying forward a full reasoning and tool transcript (Yao et al., 2023; Roohani et al., 2025), the retained state is organized around the current question and its supporting observations. Reflection and persistent insights have precedents in agent systems (Shinn et al., 2023; Zhao et al., 2024); here they remain linked to measurements selected during the same finite-budget campaign.

The evidence-use decision is a reasoning step within the actor loop, not necessarily a separate model request. Optional notes can be updated while investigating or committing a batch. Execution limits are specified in Appendix A.2, and the evaluated interaction configurations are described in Appendix C.

3.3 From the decision state to a measurement

The agent investigates follow-ups for the question identified during review. It can retrieve supporting records, inspect numerical proposals, and compare candidate batches by cost, prerequisites, and predicted value, using the reasoning–tool interaction pattern of ReAct (Yao et al., 2023). In the Hit example, this means comparing confirmation of a panel-supported candidate with other affordable opportunities rather than treating its panel result as a completed discovery. LMABO adapts acquisition functions from optimization state (Ngo et al., 2026); here, numerical proposals can also be examined alongside individual acquired records before the agent selects an explicit batch. Forecasts remain estimates of unmeasured outcomes.

The domain interface checks the final batch against costs, prerequisites, and the remaining allowance. Commitment releases its outcomes, which return to the next review with a smaller budget. This closes the coupling between evidence acquisition and evidence use without assuming that the review is optimal or that any stored interpretation is correct. The task measurements and discovery endpoints appear in Section 4.1; measurement budgets and request limits appear in Appendix A.

4 Experiments

4.1 Task Settings

Molecular hit discovery. The KIRHub replay (Saifudeen et al., 2026) contains 13 compounds and 175 wild-type assays. Actions purchase a screen, ten-point dose curve, or reported variant panel, costing 1, 10, or the number of panel measurements. A panel requires a screen from an earlier round. The endpoint, $\mathrm{VRH}(900)$, counts non-initial pairs whose purchased curve reaches 50% inhibition at or below $1\ \mu\mathrm{M}$ and whose purchased panel has residual activity at most 50% for every reported finite measurement. Curves use monotone fitting and log-concentration interpolation without extrapolation. This measures retention across the reported variants, not off-target selectivity.

Protein library design. Our protein-library evaluation uses the GFP substitution pool (Sarkisyan et al., 2016) from ProteinGym (Notin et al., 2023). Each sequence costs one measurement. A successful project requires a multi-mutant j and all constituent single mutants $\mathcal{C}(j)$ to be measured, with

$$\{j\} \cup \mathcal{C}(j) \subseteq \mathcal{O}_t \quad \text{and} \quad y_j > \max_{i \in \mathcal{C}(j)} y_i. \tag{4}$$

Here $\mathcal{O}_t$ is the purchased set. Shared constituents are charged once. The endpoint H_{96} counts newly successful projects after 96 total measurements; it is an observed score comparison, not a statistical test of epistasis.

Drug-combination screening. We use GDSC combination screens (Jaaks et al., 2022) for SIDM00874 (C1) and SIDM01062 (C2), both under S1. A grid is discovered when a purchased dose satisfies

$$v \leq 0.50 \quad \text{and} \quad s_A s_B - v \geq 0.20, \tag{5}$$

where v is combination response and s_A, s_B are public dose-matched single-agent references. S1 counts each discovered grid once and requires neither paid single-agent prerequisites nor a second confirming dose. Other Drug protocols remain development variants.

4.2 Experimental Setup

We evaluate two classes of decision-making methods: numerical baselines and agentic methods. Random supplies a randomized acquisition reference. RF-Greedy and RF-UCB refit random-forest predictors on acquired measurements, using estimated discovery value and an uncertainty-augmented score, respectively, under task-specific costs and feasibility constraints. The agentic comparison includes EvoLab, task-adapted ReAct (Yao et al., 2023), and BioDiscoveryAgent (BDA) (Roohani

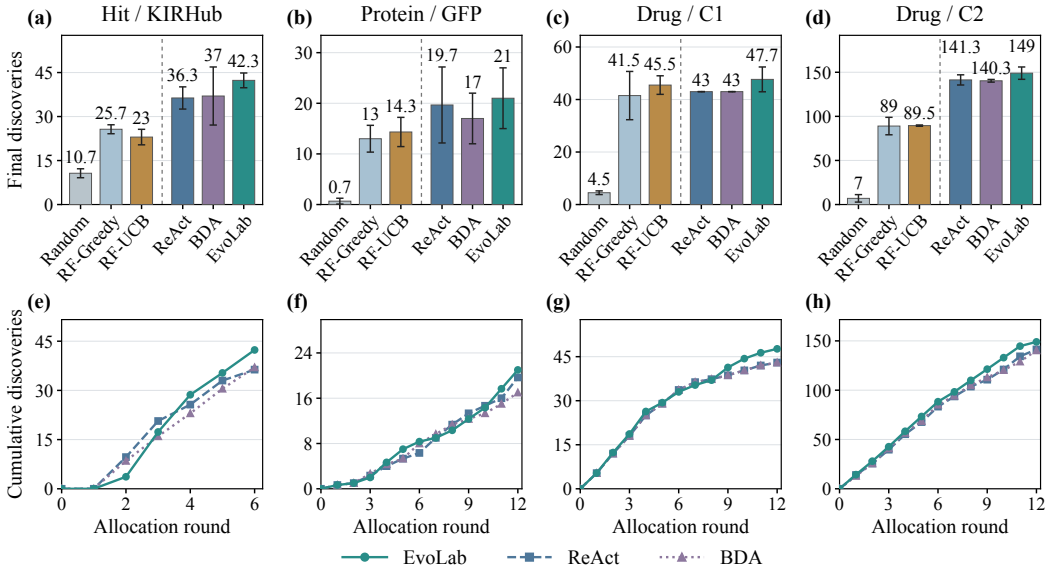


Figure 3: **Discovery yield and allocation trajectories.** Upper panels compare average final discovery counts for six methods; lower panels show average cumulative discoveries over allocation rounds. Error bars represent the standard deviation across runs. Dashed dividers separate numerical acquisition methods from agents. Each task uses a fixed measurement budget. Numerical baselines are historical, unpaired references; Hit agent comparisons also use different cohorts. Execution parameters are detailed in Appendix A.2.

et al., 2025). We use DeepSeek-V4.1-Flash (DeepSeek-AI, 2026) as the backbone for the agentic frameworks. Within each task, the agents face the same measurement budget and can observe outcomes only for selected measurements. The task configurations and execution parameters are in Appendix A; Appendix B traces one Hit campaign from acquired evidence to subsequent allocations.

4.3 Main Results

As shown in Figure 3, EvoLab achieves the highest average discovery yield among the six methods across molecular hit discovery, protein library design, and both drug-combination settings. The comparison highlights the importance of connecting current measurements to future discovery opportunities. On C1, the strongest numerical baseline outperforms both agentic baselines, yet EvoLab achieves a higher mean yield. This comparison motivates examining how acquired evidence informs subsequent allocation. The trajectories reveal how this advantage emerges under different task structures: prerequisite measurements delay visible gains, whereas productive follow-up can sustain an earlier lead.

Molecular hit discovery: using prerequisites to guide costly follow-up. A successful hit requires both a qualifying dose curve and a passing variant panel, while a curve costs ten times as much as a screen. Immediate discovery counts therefore give an incomplete view of allocation quality: inexpensive measurements can establish whether a later, costly follow-up is worthwhile. EvoLab initially trails ReAct, overtakes it in round four, and finishes with 16.5% more discoveries. Its curve expenditure on pairs whose purchased panels fail is less than half that of either agentic baseline. All three methods avoid purchasing curves after an observed panel failure, so the distinction concerns when prerequisite evidence becomes available relative to expensive follow-up. By bringing candidate completion costs and acquired panel support into the decision context, EvoLab provides a basis for prioritizing follow-up on better-supported candidates. The observed trajectory is consistent with the value of this task-aware ordering of evidence acquisition.

Protein library design: converting prepared projects into discoveries. The protein library task rewards a multi-mutant only after its constituent single mutants have been measured and its score exceeds theirs. Measuring a constituent can thus prepare a promising project without immediately increasing discovery yield. EVOLAB turns this delayed value into stronger late-round gains: it overtakes ReAct in the final two rounds and improves final yield by 6.8% over ReAct and 23.5% over BDA. The completion analysis links the gain over ReAct to successful projects whose constituents were already available before the round, partially offset by fewer successes requiring newly measured constituents. This makes the state of unfinished projects central to allocation: the agent must identify which prepared projects merit completion while accounting for the cost of missing prerequisites. EVOLAB explicitly presents these dependencies and follow-up costs, connecting acquired evidence to feasible next steps. The late divergence suggests that retaining an interaction history or adding planning stages does not by itself ensure effective completion choices. The advantage holds against each agentic baseline on two of the three seeds; the data support better selection of prepared projects, without establishing uniformly greater constituent reuse.

Drug-combination screening: allocating follow-up to unresolved opportunities. Each drug-pair grid contributes at most one discovery, so a repeat measurement is useful only if it uncovers a qualifying dose in an unresolved grid. EVOLAB recovers from a mid-campaign deficit on C1, while maintaining an earlier lead on C2; its final yield exceeds both agentic baselines on every matched seed in both cell lines. The C2 decomposition explains the distinction: first-visit discoveries are nearly unchanged relative to ReAct, while revisits contribute eight additional discoveries on average. Similar total revisit expenditure indicates more productive follow-up choices within that budget. This behavior is consistent with EVOLAB's evidence-centered context, which connects prior grid outcomes, unresolved opportunities, and the cost of another measurement. The comparison illustrates why additional reasoning or planning must remain grounded in the current experimental state: further investigation is valuable when it targets a remaining discovery opportunity. These trajectory decompositions describe the observed allocation behavior; they do not isolate the effect of an individual component.

4.4 BIOLOGICAL EVIDENCE FOR DECISION RELIABILITY

Assay-grounded evaluation of sequential decisions. We examine decision reliability in molecular hit discovery by linking allocation choices to measured wild-type and variant responses. To assess how each allocation changes future discovery opportunities, we construct a retrospective oracle from the complete wild-type dose-response and reported-variant assay tables. It obeys the original Hit rules: six rounds of 150 units, screen/curve/panel costs of one/ten/the reported variant count, a screen acquired in an earlier round before a panel, and exclusion of initialization-curve pairs. Let s_r denote the state before round r, including purchase status, acquired outcomes, and prerequisite timing, $\mathcal{F}_r(s_r)$ the feasible remaining allocation sequences, T the assay-replay transition, and $D(s_7)$ the final discovery count, with $s_{k+1} = T(s_k, A_k)$. Define

$$V_r^*(s_r) = \max_{(A_r,\ldots,A_6)\in\mathcal{F}_r(s_r)} D(s_7), \qquad V_7^*(s_7) = D(s_7), \tag{6}$$

$$\Delta_r = V_r^*(s_r) - V_{r+1}^*\big(T(s_r, A_r^{\mathrm{obs}})\big)\,. \tag{7}$$

Thus Δ_r measures terminal discoveries forgone by the observed batch, while crediting the future contribution of its prerequisite purchases. Integer optimization and legal witness schedules certify every optimum. Table 1 reports the oracle value before each round and the corresponding loss; the oracle has access to unpurchased outcomes solely for evaluation.

From variant feedback to more selective follow-up. The first two rounds expose the cost of acquiring evidence under uncertainty. All three EVOLAB campaigns spend round one on screens and emphasize variant panels in round two, deferring most costly curve measurements. No method can complete a new hit in round one because a panel requires an earlier screen. This prerequisite explains zero immediate yield, but does not itself explain positive oracle loss: the oracle already credits future completions and can select productive measurements with foreknowledge. The larger early loss therefore does not by itself establish a failure of the feedback design, nor does delayed payoff make every preparatory purchase optimal. It quantifies headroom relative to an outcome-informed allocation; the present analysis cannot separate the value of exploration from avoidable inefficiency.

Quantity	Method	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Oracle V_r^*	ReAct	61.00	57.67	55.33	53.33	48.67	37.33
	BDA	61.00	58.50	55.50	52.50	48.00	38.00
	EvoLab	61.00	57.00	53.00	52.00	50.67	42.67
Loss $\Delta_r \downarrow$	ReAct	3.33	**2.33**	2.00	4.67	11.33	1.00
	BDA	**2.50**	3.00	3.00	4.50	10.00	1.00
	EvoLab	4.00	4.00	**1.00**	**1.33**	**8.00**	**0.33**

Table 1: **Attainable terminal yield and round-specific loss in Hit.** Upper rows give the full-information oracle's maximum final discovery count from each method's state before the indicated round. Lower rows give the terminal discoveries forgone by its observed batch, with optimal continuation thereafter. Entries are campaign means; bold denotes the lowest loss in each round. The oracle retains the original costs and prerequisites. Selected cohorts differ in seeds and recorded model identifiers.

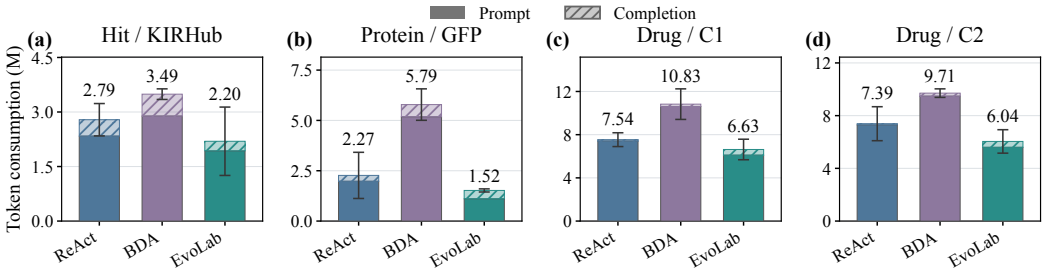


Figure 4: **Token consumption across the four evaluation settings.** Solid and hatched segments show average prompt and completion tokens per campaign, respectively. Error bars represent the standard deviation of total consumption across runs. Consumption is measured from archived successful model responses.

Variant profiling has a biological rationale: wild-type inhibition need not persist across kinase variants, as the KIRHub study demonstrates through mutation-specific inhibitor responses (Saifudeen et al., 2026). In one campaign, the round-two decision explicitly prioritizes panel results to guide later curve purchases despite zero immediate discoveries. Across all three campaigns, round three then purchases 43 curves for previously panel-qualified pairs, of which 41 satisfy the wild-type criterion. This links an acquired biological observation to a subsequent confirmation outcome. Correspondingly, EvoLab's mean loss falls from 4.00 in round two to 1.00 in round three, before its cumulative yield overtakes ReAct. It remains below both agent baselines in every subsequent round, although the round-five increase makes this loss trajectory nonmonotonic. The pattern is consistent with EvoLab's task-aware presentation of acquired panel support and completion costs: intermediate evidence informs which costly follow-ups to pursue. It supports more effective later allocation in these cohorts, without identifying a causal effect of the feedback presentation or establishing conformity to a biological law.

4.5 Token Consumption

Figure 4 shows that EvoLab combines higher average discovery yield with substantially lower token consumption across all four settings. It reduces average consumption by 12.1–32.9% relative to ReAct and 37.1–73.7% relative to BDA. The largest reduction occurs in protein library design, where EvoLab uses nearly three-quarters fewer tokens than BDA while achieving a higher discovery yield. The prompt–completion breakdown links this efficiency advantage to reduced input volume: in protein library design and both drug-combination settings, prompt consumption falls even as completion consumption increases.

The design addresses a recurring cost of sequential experimentation: every round adds observations, but the next allocation depends on their implications for the remaining opportunities. ReAct carries

prior reasoning and tool interactions into subsequent prompts, and BDA additionally retains research plans and uses an advisor stage. These workflows repeatedly present the history of how evidence was obtained. EVOLAB carries forward acquired evidence and structured persistent state, together with the preceding execution receipt, while using current-round tools to investigate the next decision. All acquired observations remain accessible. This separation allows the agent to preserve scientific continuity without repeatedly including the entire prior reasoning and tool transcript. It is particularly useful when successful discovery requires combining evidence across rounds, as in completing a protein project or revisiting an unresolved drug grid.

4.6 ABLATION STUDY

The ablations evaluate EVOLAB's explicit-review configuration, described in Appendix C. Table 2 reports measured contrasts for feedback review, persistent notes, and record presentation. Each intervention changes the full configuration independently. These historical campaigns test that configuration rather than isolating components of the single-loop Hit/protein implementation used in the main comparison.

Table 2: **Measured component comparisons.** Entries are average final discovery counts over matched campaigns. Each intervention changes the full explicit-review configuration independently, with the same task budget and discovery criterion.

Configuration	Hit	Drug C1	Drug C2
Full explicit-review agent	42.0	47.7	149.0
Review optional	38.0	42.7	142.7
Notes not carried	43.3	48.7	141.3
Full-record context	41.0	44.3	144.3

Review and evidence selection. The measured Hit and Drug contrasts favor a required review checkpoint over optional review. The proportional gap is largest in Hit and C1, where the optional-review means fall to 38.0 and 42.7 from the full agent's 42.0 and 47.7. Review gives the latest measurements a defined place in the next decision, although this comparison changes the broader interaction path as well as the checkpoint itself. Placing every acquired record in the initial context also yields less than selected-record presentation in the measured settings. The full archive remains retrievable, and per-run effects are mixed. The distinction is therefore between preserving evidence and deciding which evidence to foreground when attention and requests are limited.

Memory and action formation. Persistent notes show a task-dependent pattern: disabling note carry-over improves the measured Hit and C1 means but reduces the C2 mean. Retained notes coincide with higher C2 yield, whereas the Hit and C1 contrasts show that carrying interpretations forward does not always help. These contrasts suggest that retained interpretations require task-specific scrutiny; memory is not uniformly beneficial in these cohorts. The historical Hit trajectories contain no nonempty procedure installation, so their endpoint gain cannot be explained by procedure reuse. The measured contrasts concern review, note retention, and record presentation; they do not isolate adaptive tool use.

5 CONCLUSION

EVOLAB connects acquired evidence to subsequent budgeted decisions through task-aware context, persistent state, and tool-assisted investigation. In Hit, GFP, and two Drug settings, its mean yield exceeds the five compared baselines. The trajectories distinguish delayed prerequisite returns from productive follow-up, while token accounting shows lower mean input-token consumption than ReAct and BDA. These results highlight the importance of deciding both what to measure and how to carry the resulting evidence forward.

AI USE STATEMENT

We used generative AI tools for drafting assistance (text polishing and LaTeX preparation). We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Benjamin Burger, Phillip M Maffettone, Vladimir V Gusev, Catherine M Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M Alston, Buyi Li, Rob Clowes, et al. A mobile robotic chemist. *Nature*, 583(7815):237–241, 2020.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. *CoRR*, abs/2606.19348, 2026. doi: 10.48550/ARXIV.2606.19348. URL https://doi.org/10.48550/arXiv.2606.19348.

Jin Gao, Juntu Zhao, Zirui Zeng, Jiaqi Shen, Junhao Shi, Dukun Zhao, Yuming Lu, and Dequan Wang. TadA-Bench: A million-variant benchmark for future-round discovery toward agentic protein engineering. In *The 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.02624.

Florian Häse, Loïc M. Roch, and Alán Aspuru-Guzik. Next-generation experimentation with self-driving laboratories. *Trends in Chemistry*, 1(3):282–291, 2019. doi: 10.1016/j.trechm.2019.02.007.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=t9U3LW7JVX.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Ryan Li, Yusuf Roohani, Lin Qiu, Shiyi Cao, Gavin Li, Junze Zhang, Di Yin, Rick Wierenga, Deniz Kavi, Sherry Liu, Tianwei She, Shruti Marwaha, Jennefer N. Carter, Xin Zhou, Matthew T. Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael P. Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Autonomous biomedical research with an artificial intelligence agent. *Science*, 393 (6813):eadz4351, 2026. doi: 10.1126/science.adz4351. URL https://www.science.org/doi/abs/10.1126/science.adz4351.

Patricia Jaaks, Elizabeth A Coker, Daniel J Vis, Olivia Edwards, Emma F Carpenter, Simonetta M Leto, Lisa Dwane, Francesco Sassi, Howard Lightfoot, Syd Barthorpe, et al. Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature*, 603(7899):166–173, 2022.

Kirthevasan Kandasamy, Gautam Dasarathy, Jeff G. Schneider, and Barnabás Póczos. Multi-fidelity bayesian optimisation with continuous approximations. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1799–1808. PMLR, 2017. URL http://proceedings.mlr.press/v70/kandasamy17a.html.

Ross D King, Kenneth E Whelan, Ffion M Jones, Philip GK Reiser, Christopher H Bryant, Stephen H Muggleton, Douglas B Kell, and Stephen G Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247–252, 2004.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 7024–7035, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/95323660ed2124450caaac2c46b5ed90-Abstract.html.

Giang Ngo, Dat Phan Trong, Dang Nguyen, Sunil Gupta, and Svetha Venkatesh. Adaptive acquisition selection for bayesian optimization with large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EPKmSgXvRe.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan J. Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora S. Marks. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/cac723e5ff29f65e3fcbb0739ae91bee-Abstract-Datasets_and_Benchmarks.html.

Yusuf H. Roohani, Andrew H. Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. Biodiscoveryagent: An AI agent for designing genetic perturbation experiments. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=HAwZGLcye3.

Mehlam Saifudeen, Songli Zhu, Shuguang Liang, Mia Eason, Alison Goupil, Deanna F Mische, Christian M Loch, Haiching Ma, Marina Chan, and Taranjit S Gujral. Comprehensive profiling of clinically approved kinase inhibitors reveals mutation-specific inhibitors and opportunities for drug repurposing. *Nature Biotechnology*, pp. 1–15, 2026.

Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624(7990):86–91, 2023.

Christopher Tosh, Mauricio Tec, Jessica B White, Jeffrey F Quinn, Glorymar Ibanez Sanchez, Paul Calder, Andrew L Kung, Filemon S Dela Cruz, and Wesley Tansey. A bayesian active learning platform for scalable combination drug screens. *Nature communications*, 16(1):156, 2025.

Yiming Xiong, Shengran Hu, and Jeff Clune. Learning to continually learn via meta-learning agentic memory designs. *CoRR*, abs/2602.07755, 2026. doi: 10.48550/ARXIV.2602.07755. URL https://doi.org/10.48550/arXiv.2602.07755.

Kevin K. Yang, Zachary Wu, and Frances H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8):687–694, 2019. doi: 10.1038/s41592-019-0496-6.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. Aflow: Automating agentic workflow generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a. URL https://openreview.net/forum?id=z5uVAKwmjf.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. *CoRR*, abs/2510.04618, 2025b. doi: 10.48550/ARXIV.2510.04618. URL https://doi.org/10.48550/arXiv.2510.04618.

Shaokun Zhang, Jieyu Zhang, Jiale Liu, Linxin Song, Chi Wang, Ranjay Krishna, and Qingyun Wu. Offline training of language model agents with functions as learnable weights. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60315–60335. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/zhang24cd.html.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: LLM agents are experiential learners. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 19632–19642. AAAI Press, 2024. doi: 10.1609/AAAI.V38I17.29936. URL https://doi.org/10.1609/aaai.v38i17.29936.

A Experimental Setup and Parameters

A.1 Experimental domains

Each round consists of investigating feasible follow-ups and committing one measurement batch. Only the committed batch reveals new outcomes, which inform the next allocation. All three domains use sequential replay of previously measured biological data.

Molecular hit discovery. The KIRHub task contains 2,275 compound–target pairs. A discovery requires both a qualifying wild-type dose curve and a passing reported-variant panel, as defined in Section 4.1. Screens cost one unit, curves ten units, and panels one unit per reported variant; a panel requires a screen from an earlier round. Campaigns start with 143 units of acquired measurements and receive 150 units per round for six rounds, giving an additional budget of 900 units.

Protein library design. We use the GFP substitution dataset, containing 51,714 variants. Each sequence costs one measurement. A multi-mutant qualifies only after all constituent single mutants have been measured and its score exceeds theirs, following Equation 4. Constituents may be acquired earlier or in the same batch, and shared constituents are charged once. Campaigns start with 12 measured variants and acquire seven per round for twelve rounds, totaling 96 measurements.

Drug-combination screening. The GDSC replay contains 915 drug-pair grids for C1 and 965 for C2. Each purchased dose condition costs one unit, with public dose-matched single-agent references. A grid contributes at most one discovery under the response and Bliss criteria in Equation 5; initialization discoveries are excluded. Campaigns start with six measured grids and receive 63 units per round for twelve rounds. Unspent allowance carries forward within the additional 756-unit budget.

A.2 Agent execution parameters

Table 3 summarizes the backbone, decoding configuration, and computational allowances. These limits are distinct from the measurement budgets above. Prompt configurations are provided in Appendix C.

Table 3: **Agent execution parameters.** Model-request and analysis-operation limits apply to each allocation round.

Parameter	Setting
Backbone	DeepSeek-V4.1-Flash
Decoding	Temperature 0; thinking enabled; high reasoning effort
Maximum generated tokens per response	65,536
Model requests / analysis operations	At most 8 / 12

Numerical baselines. Random selects measurements through a seeded project ordering. RF-Greedy uses random-forest predictions to rank feasible projects by estimated discovery gain relative to remaining cost; RF-UCB adds an uncertainty bonus of one ensemble standard deviation. Both forest baselines use 200 trees with a minimum leaf size of two, update from acquired measurements, and obey the same task costs and prerequisites.

B CASE STUDY: VARIANT EVIDENCE GUIDES HIT COMPLETION

We examine one EVOLAB Hit campaign trial, selected retrospectively because its 42 discoveries are the median endpoint of the three plotted campaigns. Table 4 reports the allocation sequence across all six rounds.

Table 4: **Allocation and discovery in the representative Hit campaign.** The first three rows report measurement expenditure, which sums to 150 in every round. Panel expenditure counts variant measurements rather than the number of panels. The last two rows report discoveries satisfying both the curve and panel criteria.

	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Screen expenditure	150	1	0	4	14	1
Panel expenditure	0	129	0	16	6	29
Curve expenditure	0	20	150	130	130	120
New discoveries	0	2	14	11	10	5
Cumulative discoveries	0	2	16	27	37	42

Evidence acquisition before confirmation. Round one purchases 150 screens. Since a panel requires an earlier screen, those purchases create future panel opportunities without completing hits. Round two acquires 116 panels at a cost of 129 units, two curves, and one screen. Its recorded rationale emphasizes using panel results to guide later curves while four rounds remain. The two immediate discoveries therefore capture only part of the allocation's consequence: the panels also distinguish supported and unsuccessful follow-up opportunities.

A traceable biological observation. One example is the archived Apatinib–RAF1 pair. Its round-one screen reports zero residual activity. Round two measures the reported RAF1(R391W) variant at 0.225% residual activity, satisfying the panel criterion. At this point, wild-type dose qualification remains unmeasured; the passing panel is supporting evidence for a follow-up, not a completed hit. Round three selects its ten-unit curve and cites both earlier records. The curve reports 40.93% inhibition at $0.04115\,\mu$M and 68.39% at $0.1235\,\mu$M, bracketing the absolute 50% crossing below $1\,\mu$M. The pair now satisfies both measured criteria. This source-reported assay sequence illustrates how variant evidence can motivate a distinct wild-type confirmation, without equating inhibition of one with inhibition of the other.

From the question to the batch. The round-three rationale ranks affordable curves by predicted wild-type qualification among pairs with previously passing panels. All 15 panels precede the decision; 14 curves subsequently qualify. The forecast of 13.14 discoveries remains distinct from the 14 observed successes, and the unsuccessful confirmation is retained. This trace connects acquired evidence, cited support, allocation, and measured consequences. It neither establishes panel-first optimality nor isolates the effect of memory or review.

C PROMPT CONFIGURATION

The boxes below present evaluated instructions, organized into inputs, decision procedure, and expected output. Headings and prose are edited for readability; these are not verbatim transcripts. Task metadata, acquired records, candidate handles, and current tool responses are populated at execution time. The single-loop and explicit-review configurations are presented separately below.

C.1 SHARED MEASUREMENT CONTRACT

Budgeted Biological Experiment Selection

Inputs

The task objective and discovery criterion; public candidate metadata; acquired measurements; legal actions and prerequisites; the current round and remaining measurement budget.

Decision procedure

1. Select measurements using the evidence available in this campaign. Unmeasured outcomes are unknown. Numerical forecasts are predictions, not acquired observations.
2. Retrieve records or request analysis when its result can change the choice. Compare complete, legal batches under the remaining budget and task prerequisites.
3. Return a legal final batch within eight model requests and twelve analysis operations. The last request must finalize. Only execution of that batch reveals new measurements.

Expected output

One finite JSON object: either a tool name, its arguments, and a rationale; or a final action, rationale, immediate-discovery forecast, and supporting acquired-record identifiers. A final action identifies explicit measurements or a candidate returned in the current round. Supporting identifiers may be empty; supplied identifiers must refer to acquired records.

C.2 EVOLAB DECISION AND EVIDENCE USE

Single-Loop Evidence Use and Allocation

Inputs

All acquired raw observations, public task relationships, the previous execution receipt, current tool responses, optional persistent notes, and the remaining allowances.

Decision procedure

1. Interpret what acquired outcomes imply for unfinished opportunities. Use requested numerical alternatives and task-specific evidence to decide which follow-up deserves the next allocation.
2. Distinguish structural feasibility from predicted success. Historical success among selected measurements is not a success probability for the remaining pool. Keep provisional interpretations revisable.
3. Commit a legal action when sufficient evidence is available. No separate critic, campaign plan, mandatory review response, or executable procedure is required. Carry forward the actual executed purchase, including any fallback, rather than an unexecuted intention.

Optional memory update

Attach a selective note update to an ordinary tool or action response. Add or revise a stable note with acquired-record references, or retire an obsolete note with a reason. Unmentioned notes persist. A valid update commits with its associated valid response and consumes one analysis operation. Notes do not restrict access to the full acquired record.

Expected output

The shared tool or action response, optionally accompanied by a memory update. The evidence-use decision occurs within this interaction; it is not an additional required model call.

Task-Specific Interpretation

Molecular hit discovery

Connect screens, variant panels, and dose curves to unfinished hit paths. Compare stage costs and missing prerequisites. Distinguish panel retention, wild-type qualification, and completed discoveries. Acquired panel cohorts can support interpretation but do not reveal unchosen outcomes. A final-round screen cannot enable a later panel.

Protein library design

Compare requested portfolios by measured constituents, missing records, shared requirements, and follow-up cost. A constituent and its combination may complete a project in the same batch, including the final round. A point forecast below a measured constituent is uncertain evidence, not an observed failure. Structural closure is not sufficient for success.

Drug-combination screening

Allocate to dose conditions under the distinct-grid endpoint. Distinguish a qualifying condition from a newly discovered grid; repeated measurements of an already discovered grid do not increase yield. Treat candidate forecasts and any forecast–outcome comparisons as estimates and evidence from selected measurements, respectively.

Explicit-Review Configuration

Inputs

Initialization evidence in the first round; subsequently, latest measurements, retained older records, notes, the previous execution receipt, and the remaining budget. Other acquired records remain retrievable.

Review and investigation

First state what the feedback supports or leaves unresolved and identify the next decision question. Replace the retained-record and note state, including previous entries that should survive. Retain or revise the optional analysis procedure. Then investigate and select a batch through the shared tool interface. A stored procedure contains permitted analysis calls and cannot purchase measurements.

Expected output

A review response followed by tool or action responses. Each review consumes one model request and one analysis operation; each executed procedure call consumes an analysis operation. This configuration applies to the selected Drug campaigns and the existing ablations, rather than to the single-loop Hit/GFP campaigns.

C.3 AGENTIC COMPARISON PROMPTS

ReAct and BioDiscoveryAgent

ReAct

Reason from the available observations, request a tool, inspect its response, and choose the next action. Prior interaction and action history accompany the acquired observations. Adapt the investigation within the shared request and operation allowances; finalize through the shared measurement contract.

BioDiscoveryAgent: proposal

Reflect on the observed experiments, formulate a research plan, investigate public candidates and acquired evidence, and propose a batch. Include the reflection and research plan in the proposal. Submit by the sixth model request to leave room for the advisor within the same round allowance.

BioDiscoveryAgent: advisor

Critique the proposed batch and research plan using acquired evidence and available tool results. Approve the proposal or return a corrected legal batch with a critique. No purchase has occurred at the proposal stage; approval or a valid replacement precedes execution.

Information boundary

Both baselines observe only selected measurements and permitted public metadata. They use the same task costs, prerequisites, and discovery criteria as EVOLAB.