AN INFORMATION-THEORETIC APPROACH TO AGENTIC SCIENTIFIC EXPLORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Answering open scientific questions requires gathering evidence scattered across literature, databases, and search engines whose contents are too vast to model in advance. Retrieval-augmented generation selects actions by semantic similarity, treating each step independently and ignoring what the agent has already learned or still needs to resolve. We propose an information-theoretic framework that computes expected information gain over a set of competing hypotheses rather than over the unbounded space of possible observations, enabling principled action selection in open systems where observation-space information gain is intractable. The agent maintains an evolving memory graph, generates hypotheses, and selects actions by a multi-objective utility combining hypothesis-space information gain, novelty, coherence, and challenger-driven falsification pressure. Beliefs are updated by Bayesian scoring over accumulated evidence. On cancer-type prediction from gene coexpression networks, our agent is the only method to achieve balanced sensitivity and specificity and positive Matthews correlation, whereas strong agentic baselines (ReAct, Tree-of-Thoughts, MCTSr, and AlphaApollo) each collapse to a class bias, defaulting to affirmation or to rejection and achieving no better than chance-level discrimination. Ablations show that each utility term contributes to this balanced discrimination.

1 INTRODUCTION

Large language model (LLM)-based agents are beginning to move beyond information retrieval toward genuine scientific exploration (Gridach et al., 2025; Yamada et al., 2025; Zou & Topol, 2025; Ghafarollahi & Buehler, 2025; Boiko et al., 2023; Nguyen et al., 2026). Early systems, formalized under retrieval-augmented generation (RAG), largely operate in a single-pass regime: retrieve semantically related documents and synthesize an answer (Lewis et al., 2020). It works when the answer already exists in the literature, but many scientific questions are uncertainty-driven rather than retrieval problems. Standard RAG is not well suited for this setting, selecting retrieval actions independently by similarity without an explicit representation of what the agent has already learned or still needs to know. Recent adaptive approaches (Asai et al., 2024; Trivedi et al., 2023; Jiang et al., 2023) introduce conditional or iterative retrieval, but lack a formal framework for quantifying the information value of exploration actions. The problem is compounded by the open-ended nature of scientific knowledge sources, whose contents cannot be enumerated or modeled exhaustively.

Recent progress in LLM reasoning reflects an increasing awareness of these limitations. Chain-of-thought prompting (Wei et al., 2022) and self-consistency (Wang et al., 2023) showed that explicit intermediate reasoning and trajectory diversity improve performance. Tree-of-Thoughts (Yao et al., 2023a) and Monte Carlo Tree Search–based agents (Zhou et al., 2024; Li et al., 2025) introduced structured search over reasoning paths, while ReAct (Yao et al., 2023b) and Reflexion (Shinn et al., 2023) integrated reasoning with tool use and self-feedback. Graph-based approaches, e.g., GraphRAG (Edge et al., 2024) and Think-on-Graph (Sun et al., 2024) leverage entity-relation structure but rely on static summarization or greedy traversal rather than information-theoretic action selection. These methods improve *how* agents search while offering little guidance on *what* is a high-value action in open domains.

Balancing exploration and exploitation under uncertainty is well studied in reinforcement learning, through curiosity-driven rewards for novel states (Pathak et al., 2017; Burda et al., 2019) and 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.

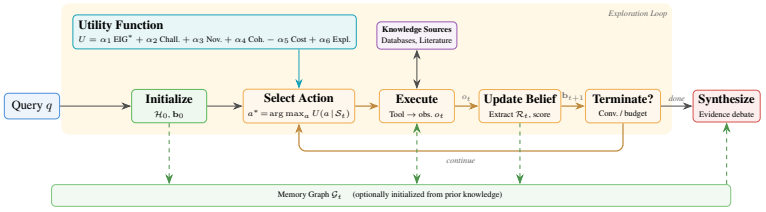


Figure 1: Architecture of the framework. The agent maintains a memory graph, generates competing hypotheses with uniform belief $\mathbf{b}_t$, and iteratively selects actions maximizing $U(a \mid \mathcal{S}_t)$; extracted evidence rules update beliefs via Bayesian scoring with a decaying graph-topology prior.

Bayes-Adaptive MDP (Duff, 2002), but they optimize policies over many episodes with a transition model, whereas scientific exploration must reason within a single encounter. Bayesian optimal experimental design (Lindley, 1956; Chaloner & Verdinelli, 1995; Rainforth et al., 2024) selects actions to maximally reduce posterior uncertainty (formalized as curiosity in Still & Precup, 2012), but assumes a known generative model over observations, breaking down in knowledge systems with an unbounded observation space.

A promising theoretical perspective comes from Schapiro et al. (2025), who formalize scientific conceptual spaces as directed acyclic graphs of generative rules and axioms. Within this framework, transformational scientific creativity arises from perturbations to foundational assumptions rather than incremental refinements, echoing Kuhn's account of scientific revolutions (Kuhn, 1962). The value of an action then lies in how it reshapes the space of plausible explanations. The framework is theoretical, however, and offers no computational implementation for agents operating over real knowledge sources.

In this paper, we propose an information-theoretic framework for agentic scientific exploration that addresses these challenges (Figure 1). Our key insight is that expected information gain should be computed over hypotheses rather than observations or documents in knowledge sources (MacKay, 1992), allowing agents to reason and explore effectively even when the underlying knowledge source cannot be modeled. The agent maintains a structured, evolving memory graph initialized from prior knowledge and progressively enriched through tool-augmented retrieval (Chater et al., 2010). Given a query, it generates competing hypotheses and iteratively selects actions maximizing a multi-objective utility that balances hypothesis-space information gain, novelty, coherence, and explicit falsification pressure (Popper, 2005). A challenger mechanism prevents premature convergence by sustaining exploration among non-leading hypotheses as confidence grows. Beliefs are updated via Bayesian scoring over accumulated evidence, and exploration proceeds until convergence or budget exhaustion. Experiments on disease prediction from gene coexpression networks demonstrate consistent improvements over ReAct, Tree-of-Thoughts, and MCTSr baselines. The graph-of-axioms perspective (Schapiro et al., 2025; Kuhn, 1962) motivates the framework conceptually, but the contribution we evaluate here is narrower: active, budget-constrained discrimination and falsification over a finite set of competing hypotheses grounded in a noisy knowledge graph.

2 Methodology

Following Schapiro et al. (2025), we organize scientific knowledge as a heterogeneous acyclic property graph in which low-level *research artifacts* connect to grounded *rules* that support or refute *hypotheses* (promoted to confirmed *axioms* once sufficiently supported). Our framework operationalizes this hierarchy: the agent populates the graph through tool-augmented retrieval and selects actions to maximally reshape its belief distribution over competing hypotheses, as depicted in Figure 1.

2.1 Problem Formulation

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a knowledge graph representing the agent's memory, with entities as nodes and relationships as edges. The framework fixes only the tier hierarchy: artifacts support rules, rules support hypotheses, and hypotheses become axioms. The entities filling these tiers are domain-

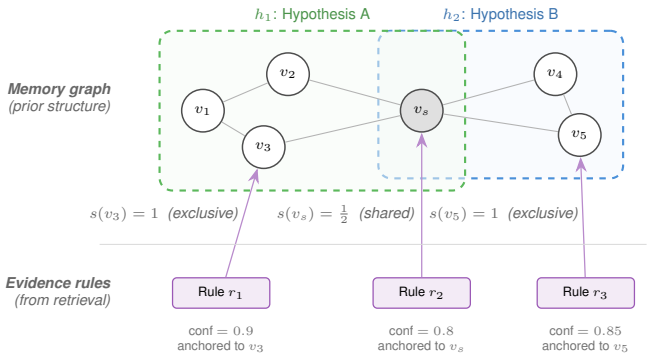


Figure 2: Belief state of the agent. Entity nodes (circles) form the memory graph; white nodes are exclusive to one hypothesis while shaded nodes are shared. The node specificity $s(v)$ weights evidence by discriminative power: rules anchored to exclusive nodes (e.g., r_1 at v_3) contribute fully to their hypothesis, while rules at shared nodes (e.g., r_2 at v_s) are down-weighted. Evidence rules (purple rectangles) are added to memory as the agent retrieves new information, progressively enriching the graph beyond its initial prior structure.

defined, so an artifact may be a gene, protein, study, or any object relevant to the query. The agent accumulates this content during exploration. The memory may be initialized from query-relevant structured data (seeding hypothesis generation via community detection and centrality) or start empty, and each action enriches it with evidence nodes. Given a natural-language query q, the agent produces an answer using a finite set of action types $\mathcal{A}$, each parameterized by target entities. At each step t, it maintains a belief state $\mathcal{S}_t = (\mathcal{G}_t, \mathcal{H}, \mathbf{b}_t, \mathbf{a}_{<t})$ over the memory graph, hypotheses $\mathcal{H} = \{h_1, \ldots, h_n\}$, their belief distribution $\mathbf{b}_t$, and action history. It selects $a_t \in \mathcal{A}$, obtains observation o_t, extracts evidence rules $\mathcal{R}_t$ ($\mathcal{R} = \bigcup_t \mathcal{R}_t$), updates beliefs, and repeats until termination, aiming to identify the best-supported hypothesis while minimizing total action cost.

2.2 Belief Representation

Each hypothesis is a tuple $h_i = (\ell_i, \mathcal{N}_i, f_g(h_i))$, where ℓ_i is a natural-language explanatory claim, $\mathcal{N}_i \subseteq \mathcal{V}$ is its grounding set of memory nodes, and $f_g(h_i) \in [0, 1]$ is an initial graph-support score. The claim ℓ_i is what the scoring terms embed and compare (e.g. query relevance f_q in Appendix B); $\mathcal{N}_i$ is what evidence rules and actions anchor to. For the cancer-type task, ℓ_i is a candidate answer such as "this network characterizes lung adenocarcinoma" and $\mathcal{N}_i$ is the community of coexpressed genes that motivated it. The $|\mathcal{H}|$ hypotheses are treated as mutually exclusive and exhaustive candidate answers to the query (Section 2.4). Hypotheses may overlap in their grounding sets, with a shared node $v_s \in \mathcal{N}_i \cap \mathcal{N}_j$ belonging to multiple hypotheses (as illustrated in Figure 2).

Node specificity. To weight evidence by discriminative power, we define $s(v) = 1/|\{h_i \in \mathcal{H} : v \in \mathcal{N}_i\}|$, so exclusive nodes have $s(v) = 1$ and nodes shared by k hypotheses have $s(v) = 1/k$. This adapts inverse document frequency (Sparck Jones, 1988) from terms across documents to entities across hypotheses: evidence at exclusive nodes contributes fully, while evidence at shared nodes is down-weighted.

Evidence rules. A *rule* is a single factual statement extracted from a tool observation. Each rule $r \in \mathcal{R}$ carries a natural-language statement, a confidence $c(r) \in [0, 1]$, a source type and identifier (e.g. a PubMed ID), and anchor nodes $\mathcal{V}_r \subseteq \mathcal{V}$ linking it to the memory graph. For example, a rule stating *"entity A is affected by mechanism M and interacts with entity B"* carries its extraction confidence, source, and anchor nodes $\{A, B\}$. Each rule is added to memory as a node connected to its anchors (as shown in Figure 2). Rule specificity is the maximum over its anchors, $s(r) = \max_{v \in \mathcal{V}_r} s(v)$, so a rule anchored to an exclusive node counts as specific even if it also references shared nodes.

Contradictions. When a rule explicitly supports an alternative to hypothesis h_i, it is recorded as a contradiction, and the contradiction penalty aggregates the confidences of contradicting rules analogously to how evidence strength aggregates supporting rules.

2.3 HYPOTHESIS GENERATION

Given the query q and current memory $\mathcal{G}_0$, the agent generates competing hypotheses in three stages: (i) *proposal*, detecting communities and hub nodes and prompting a language model for up to N_h hypotheses, each with candidate nodes $\mathcal{N}_i$ and an initial graph-support score $f_g(h_i) \in [0, 1]$; (ii) *merging* equivalent hypotheses to prevent belief fragmentation; and (iii) *filtering*, discarding hypotheses without memory entities, recomputing graph support with exclusivity weighting, and initializing beliefs uniformly ($b_0(h_i) = 1/|\mathcal{H}|$). Generation continues during exploration, and hypotheses exceeding a promotion threshold τ_{axiom} become confirmed *axioms*.

2.4 EXPECTED INFORMATION GAIN (EIG) IN HYPOTHESIS SPACE

Traditional EIG formulations marginalize over possible observations o, $\text{EIG}(a) = \mathbb{E}_{o\sim P(o|a,\mathcal{S}_t)}[D_{\text{KL}}(P(\mathcal{H} \mid o, a, \mathcal{S}_t) \,\|\, P(\mathcal{H} \mid \mathcal{S}_t))]$, which requires a generative model $P(o \mid a, \mathcal{S}_t)$. For open knowledge sources this distribution is intractable, since the set of possible observations is unbounded and domain-dependent.

Partition-based EIG. Instead of modeling the raw observation space, we group the hypotheses by their predicted outcomes and measure information over those groups. For each candidate action a, we partition $\mathcal{H}$ into M groups $\{P_1, \ldots, P_M\}$ by which hypotheses would be *differentially affected* by the action's outcome: hypotheses predicting the same result share a group, those making competing claims are separated, and a leftover group captures hypotheses with no stake in the outcome. This defines a group index $Z \in \{1, \ldots, M\}$, the group whose prediction the outcome confirms, and we compute EIG as information about Z. The partition requires only a judgment of which hypotheses have competing predictions about the action's targets, a well-posed reasoning task over a small finite set that needs no model of what a knowledge source might return. In our implementation this judgment is made by a language model, but any partitioning that reflects genuine competing predictions serves the same role.

The EIG is then the entropy of the belief distribution over these groups:

$$\text{EIG}(a) = H_{\text{partition}}(a) = -\sum_{m=1}^{M} p(P_m) \log p(P_m), \tag{1}$$

where $p(P_m) = \sum_{h_i \in P_m} b_t(h_i)$ is the total belief mass assigned to group P_m. Intuitively, belief concentrated in one group means the action cannot discriminate, while belief spread across disagreeing groups means high potential to shift beliefs. The leftover group participates in the entropy computation since discovering that the target entity is uninvolved is itself informative against hypotheses that depend on it. Because the hypotheses are treated as mutually exclusive and exhaustive (exactly one is the true generator), confirming one group's prediction redistributes mass away from the others, including the leftover group. The grouping coarsens a single categorical truth rather than a set of independent claims. This partition entropy equals the observation-marginalized EIG under an idealized observer, and upper-bounds it in general.

Proposition 1. *Let $Z = g(H)$ be the partition-group index and assume the observation is informative about H only through its group, i.e. the Markov chain $H \to Z \to O$ holds (within a group, the outcome carries no discriminative signal between hypotheses). Then* $\text{EIG}(a) = I(H; O \mid a) \leq H_{\text{partition}}(a) = H(Z)$*, with equality when the outcome perfectly identifies the group.*

We use $H_{\text{partition}}$ as the acquisition criterion for action selection. The assumption is exact for a deterministic forward model and approximate otherwise; its failure mode (an action that separates hypotheses pooled in the same group) and the proof are analyzed in Appendix A.

To reward actions that may reveal new hypotheses, we add a hypothesis-expansion bonus, $\text{EIG}^*(a) = \text{EIG}(a) + \gamma \cdot \hat{n}_{\text{new}}(a)$, where $\hat{n}_{\text{new}}(a)$ estimates how many of the action's anchor nodes are not yet covered by existing hypotheses.

2.5 UTILITY FUNCTION

At each step t, the agent generates candidate actions $\mathcal{A}_t$ (literature searches, database queries, structural analyses) and selects the one maximizing an objective utility over the current belief $\mathcal{S}_t$:

$$U(a \mid \mathcal{S}_t) = \alpha_1 \,\mathrm{EIG}^*(a) + \alpha_2 \,\mathrm{EIG}_{\mathrm{chal}}(a) + \alpha_3 \,\mathrm{Nov}(a) + \alpha_4 \,\mathrm{Coh}(a) - \alpha_5 \,C(a) + \alpha_6 \,\mathrm{Expl}(a). \quad (2)$$

The weights $\alpha_1, \ldots, \alpha_6$ (denoted $\alpha_{\mathrm{eig}}, \alpha_{\mathrm{chal}}, \alpha_{\mathrm{nov}}, \alpha_{\mathrm{coh}}, \alpha_{\mathrm{cost}}, \alpha_{\mathrm{expl}}$ in Table 10) are fixed on a small development set rather than learned. Unlike an RL policy, they require no transition model or training episodes, and an expert can adjust them directly (e.g. raising α_2 for stronger falsification pressure). The terms are: **EIG*** (Section 2.4): prefers actions that discriminate between competing hypotheses; **challenger EIG**: recomputes IG over the non-leader beliefs to counter the premature convergence that EIG-optimal actions cause once a hypothesis leads; **novelty**: penalizes actions whose description embedding is close to past observations (Bellemare et al., 2016; Pathak et al., 2017); **coherence**: favors actions consistent with established evidence; **cost**: fixed per action type; and **exploration bonuses** for uninvestigated hypotheses, under-explored communities, and evidence gaps. An adaptive schedule dampens EIG and boosts exploration once the leading belief exceeds $\theta_c = 1.5/|\mathcal{H}|$, so a more confident agent shifts toward testing alternatives. Closed forms, the challenger-weight ramp, and the schedule are given in Appendix B.

2.6 Belief Update

After each action, the agent extracts evidence rules and recomputes every hypothesis score from *all* accumulated evidence rather than applying a single-step Bayesian update, combining specificity-weighted literature evidence with a decaying graph-topology prior and a contradiction penalty. Scores become beliefs via a temperature softmax with *prior anchoring*, whereby uninvestigated hypotheses retain their initial prior mass while investigated ones share the remainder. Consequently, early evidence cannot collapse the distribution before alternatives are examined (inspired by optimistic initialization in bandits Auer et al., 2002). A memory-decay multiplier shifts beliefs from topology-driven to evidence-driven as rules accumulate, analogous to empirical Bayes (Robbins, 1956). The scoring equation (Eq. 8), its eight factors, and the anchored update (Eq. 11) are given in Appendix B.

2.7 Termination

Exploration terminates on budget exhaustion, belief convergence (stable normalized entropy), or exploration stall (no new confirmed findings), after which an evidence-quality audit evaluates all hypotheses before the final synthesis. For the binary task, the hypotheses are themselves framed as competing *yes*/*no* claims about the queried type (e.g. "this is a STAD network" vs. "this is not a STAD network"), so the terminal belief distribution over these claims yields the answer: the final synthesis returns the verdict of the leading claim, and a verdict of "insufficient evidence" is recorded as *no*. The cost scales linearly in the number of steps T, with per-step complexity $O(|\mathcal{H}| \cdot |\mathcal{R}|)$. Full termination conditions and the audit are detailed in Appendix B.

3 Experiments

Task. Since there is no available standard benchmark for agentic scientific exploration, we evaluate on a protocol that requires hypothesis formation over noisy structured data, multi-source evidence integration, and crucially the ability to *reject* when evidence is insufficient.[1] We form the cancer-type prediction task from gene coexpression networks. Given a network of correlated genes from tumor samples, the agent determines the cancer type. This demands reasoning rather than lookup: the same genes recur across cancer types with different roles, and discriminative evidence is scattered across literature and databases rather than present in any single source.

Dataset. We extract gene coexpression networks from The Cancer Genome Atlas (TCGA), filtering to OncoKB cancer genes (Chakravarty et al., 2017) and thresholding correlations at $r \geq 0.7$ (full pipeline in Appendix D). The network initializes the memory graph and seeds hypothesis generation without revealing the cancer type. Following evidence that coexpression networks are unstable below 200 samples (Aguirre-Plans et al., 2025; Nulsen et al., 2021; Clayton et al., 2020; Silvey et al., 2026), we use the 18 of 33 TCGA types with at least 200 primary tumor samples (excluding the remaining 15), capped at 200 each. We pose a binary task, "Is this a [cancer type X] network?", with **hard** (same organ system), **medium** (adjacent), and **easy** (different organ) negatives per type, manually selected (Appendix C).

[1] In our study, "reject" means answering *no* to the binary query (declining to affirm the queried cancer type), not rejecting a null hypothesis in the statistical-testing sense.

Settings. We compare against four baselines: **ReAct** (Yao et al., 2023b), **Tree-of-Thoughts** (Yao et al., 2023a), **MCTSr** (Zhang et al., 2024), and **AlphaApollo** (Zhou et al., 2026), a deep agentic-reasoning system that refines answers through a propose–judge–update loop with tool-assisted verification. All agents receive the same graph, query, LLM, budget (10 steps), and tool suite (PubMed, UniProt, Reactome enrichment, graph analysis), with no rejection instructions during exploration (baseline configurations and prompt constructions in Appendix D.3). We report mean $\pm$ std over 3 runs of Accuracy, Sensitivity, Specificity, F1, and the Matthews correlation coefficient (MCC). We also evaluate a **ReAct+Reject** variant whose synthesis prompt adds explicit falsification instructions while exploration is unchanged, testing whether prompt-level rejection closes the specificity gap.

4 RESULTS

Main results. Table 1 summarizes performance across 18 cancer types. Every baseline collapses to a class bias, spiking one class-wise metric while cratering the other. The affirmative agents inflate sensitivity: Tree-of-Thoughts answers *yes* to 96.3% of positive queries but also to 90.7% of negatives (9.3% specificity), and ReAct behaves similarly. AlphaApollo shows the opposite, rejection bias, with the highest specificity (75.9%) but only 14.8% sensitivity. Each of these biases buys a headline number: high sensitivity, high specificity, or (for AlphaApollo) accuracy on par with ours, since under the 1:3 imbalance a near-always-*no* agent scores well on both accuracy and specificity. None of it reflects genuine discrimination. This is why we report MCC as the primary metric: a trivial always-*no* classifier reaches 75% accuracy at MCC=0, and every baseline lands at or below zero MCC. No agent, including ours, exceeds that 75% accuracy, precisely because always-*no* is the degenerate optimum accuracy rewards under this imbalance while attaining zero sensitivity and MCC=0; raw accuracy is therefore uninformative here. Our agent is the only method with positive MCC (0.141, positive in every repeat) and the best balanced accuracy (58.0%), and its 60.2% accuracy comes with genuine positive detection (53.7% sensitivity), staying competitive on *both* classes rather than sacrificing one for the other.

Agent	Accuracy	Sensitivity	Specificity	F1	Bal. Acc.	MCC
Ours	$\mathbf{60.2 \pm 0.8}$	53.7 ± 8.5	62.3 ± 2.1	0.401 ± 0.043	**58.0**	$\mathbf{0.141 \pm 0.056}$
ReAct	41.2 ± 4.5	81.5 ± 17.9	27.8 ± 0.0	0.407 ± 0.073	54.7	0.099 ± 0.179
ToT	31.0 ± 3.5	$\mathbf{96.3 \pm 3.2}$	9.3 ± 3.7	$\mathbf{0.411 \pm 0.020}$	52.8	0.082 ± 0.097
MCTSr	39.4 ± 4.9	63.0 ± 12.8	31.5 ± 3.2	0.341 ± 0.062	47.3	-0.050 ± 0.137
AlphaApollo	$\mathbf{60.6 \pm 2.1}$	14.8 ± 8.5	$\mathbf{75.9 \pm 0.0}$	0.156 ± 0.085	45.4	-0.100 ± 0.094

Table 1: Main results on 18 TCGA cancer types ($\geq$200 samples), mean $\pm$ std over 3 runs. Balanced accuracy is (sensitivity+specificity)/2. Best per column in bold; accuracy is a statistical tie between Ours and AlphaApollo.

Difficulty. Table 2 breaks accuracy down by query type. Each of the three negative tiers contains only negative queries, so accuracy there is simply the rejection rate on that tier: an agent that answers *no* indiscriminately tops all three tiers without discriminating at all, at the cost of the positive column. The table shows exactly this pattern of class bias. The affirmative baselines score high on the positive query and low on every negative tier, while AlphaApollo does the reverse, topping all three negative tiers (66.7–85.2%) purely by rejecting almost everything (14.8% on positives). These tier scores therefore reward bias, not skill, and should be read together with the positive column. Our agent is the only one whose accuracy stays in a comparable band across both positive queries (53.7%) and negative tiers (59.3–66.7%), rather than excelling at one by abandoning the other. A separate effect appears within the negatives: every agent scores *lower* on the nominally easier (different-organ) negatives than on the hard (same-organ) ones. Because our tiers vary by anatomical distance rather than by belief difficulty, this ordering likely reflects a literature-frequency confound. Breast invasive carcinoma (BRCA) serves as the easy negative for 10 of the 18 cohorts (Appendix C, Table 7) and is among the most heavily studied malignancies; because the networks are filtered to general OncoKB drivers (e.g. *TP53*, *PIK3CA*, *MYC*) that co-occur densely with breast cancer in PubMed, a query "is this a BRCA network?" on an unrelated tumor retrieves abundant confirmatory-looking literature, inflating false positives for every agent. The easy-tier numbers should be read with this artifact in mind; grounding negative assignments in a formal anatomical ontology is left to future work. Our agent leads on most cancer types, though not all (cf. Appendix E).

Agent	Positive	Hard	Medium	Easy
Ours	53.7 ± 8.5	66.7 ± 0.0	61.1 ± 5.6	59.3 ± 8.5
ReAct	81.5 ± 17.9	33.3 ± 14.7	35.2 ± 17.9	14.8 ± 6.4
ToT	**96.3 ± 3.2**	14.8 ± 11.6	7.4 ± 3.2	5.6 ± 5.6
MCTSr	63.0 ± 12.8	42.6 ± 11.6	33.3 ± 5.6	18.5 ± 8.5
AlphaApollo	14.8 ± 8.5	**85.2 ± 3.2**	**75.9 ± 3.2**	**66.7 ± 5.6**

Table 2: Accuracy by query type (mean ± std over 3 runs). "Positive" is accuracy on the positive query (equivalently sensitivity, cf. Table 1); Hard/Medium/Easy are accuracy on same-organ, adjacent-organ, and different-organ negatives. Best per column in bold. Some cells show ±0.0 when all three runs yield identical counts (Appendix D.2).

Exploration dynamics. Figure 3 shows how the agent behaves within a run. Belief concentrates over time (rising Gini) while the top-two belief gap first widens then narrows, as the challenger term reopens competition once a leader emerges; cumulative KL grows steadily, confirming that beliefs are revised by incoming evidence rather than fixed early. The utility decomposition shifts accordingly: exploration and coverage bonuses dominate the first few steps to secure broad hypothesis coverage, then give way to EIG and novelty as the agent moves from surveying hypotheses to discriminating among them.

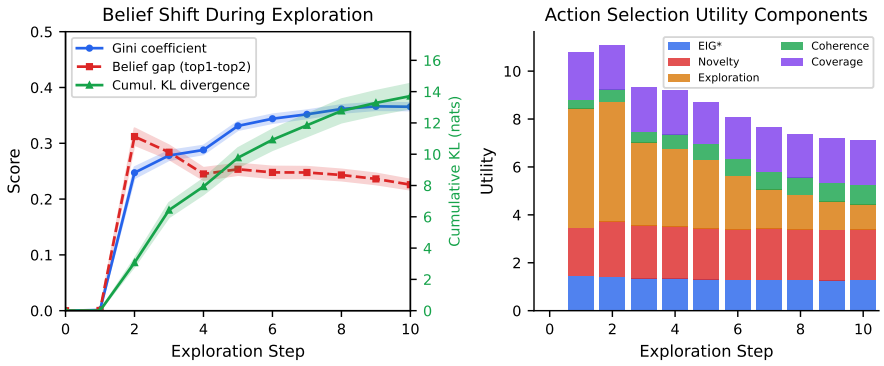


Figure 3: Exploration dynamics averaged over all runs. **Left:** the Gini coefficient (blue) rises as beliefs concentrate, the belief gap (red) narrows after an initial spike as the challenger sustains competition, and cumulative KL (green) tracks evidence-driven revision. Shaded regions show ±1 standard error of the mean. **Right:** the main utility components (Equation 2) shift from exploration-dominated early steps to EIG- and novelty-driven later steps, with coverage shown as a separate exploration sub-bonus. The challenger EIG shares the form of EIG*; its effect appears in the narrowing belief gap on the left.

Ablations. We test the two design choices the method rests on: the multi-objective utility and the meaningfulness of the partition. Each variant shares the exploration machinery, tools, budget, and belief update of the full agent, differing only in the single change. Removing any single utility term (Table 3) drops MCC from 0.141 to a mean at or below zero, and the full agent is the only configuration whose MCC stays positive within one standard deviation. The removals fail in two ways: dropping expansion or coherence restores the affirmative bias of the baselines (high sensitivity, low specificity), while dropping the challenger or EIG instead depresses sensitivity, under-detecting true positives. Removing the challenger also raises the specificity standard deviation 4.5 times (from ±2.1 to ±9.6), since EIG alone favors confirmatory actions once a leader emerges. Replacing the meaningful partition with a random two-group assignment (Table 4) collapses MCC to 0.017 and specificity to 35.2%, isolating the partition as the source of the gains: EIG over an uninformative grouping has no discriminative signal to exploit.

Agent	Accuracy	Sensitivity	Specificity	F1	MCC
Ours (full)	**60.2 ± 0.8**	53.7 ± 8.5	**62.3 ± 2.1**	**0.401 ± 0.043**	**0.141 ± 0.056**
− Challenger (α_2=0)	50.9 ± 5.3	37.0 ± 8.5	55.6 ± 9.6	0.272 ± 0.029	−0.065 ± 0.027
− EIG (α_1=0)	47.7 ± 6.9	46.3 ± 12.8	48.1 ± 4.9	0.307 ± 0.087	−0.048 ± 0.153
− Novelty (α_3=0)	44.4 ± 0.0	53.7 ± 12.8	41.4 ± 4.3	0.323 ± 0.055	−0.043 ± 0.074
− Coherence (α_4=0)	46.8 ± 13.2	**66.7 ± 5.6**	40.1 ± 15.7	0.391 ± 0.079	0.054 ± 0.192
− Expansion (γ=0)	37.5 ± 2.8	59.3 ± 12.8	30.2 ± 2.1	0.320 ± 0.057	−0.095 ± 0.106

Table 3: Utility-term ablations, each removing one term of Equation 2 (α_i=0, or γ=0 for expansion). Mean ± std over 3 runs on all 18 cancer types; best in bold. Some cells show ±0.0 when all three runs yield identical counts (Appendix D.2).

Agent	Accuracy	Sensitivity	Specificity	F1	MCC
Ours (meaningful partition)	**60.2 ± 0.8**	53.7 ± 8.5	**62.3 ± 2.1**	**0.401 ± 0.043**	**0.141 ± 0.056**
Random partition	43.1 ± 3.7	**66.7 ± 5.6**	35.2 ± 6.7	0.369 ± 0.007	0.017 ± 0.017

Table 4: Partition-quality ablation (mean ± std over 3 runs on all 18 cancer types; best in bold). Replacing the meaningful hypothesis partition with a random two-group partition collapses MCC and specificity while sensitivity rises.

Falsification prompting. A ReAct+Reject variant (identical exploration, rejection-augmented synthesis; Table 5) inverts the affirmative bias rather than calibrating it, collapsing sensitivity to 25.9% while specificity rises to 73.5%, yielding MCC≈0 (-0.006 ± 0.210) at nearly 4 times our standard deviation. The shift in specificity relative to plain ReAct is significant (Fisher's exact test on the 2×2 table of correct-vs-incorrect negative-query verdicts, ReAct vs. ReAct+Reject: $p<0.001$). The prompt genuinely changes behavior, but it produces a different failure mode rather than balanced discrimination. This isolates *post-hoc synthesis* prompting only: the 10-step exploration trace is unchanged, so the result shows that adding a rejection instruction at answer time cannot repair a confirmatory search trajectory. It does not rule out exploration-time interventions, e.g. prompting the agent to seek counter-evidence or reflect mid-trajectory as in Reflexion (Shinn et al., 2023), which remain an open comparison.

Agent	Sensitivity	Specificity	MCC
ReAct	81.5 ± 17.9	27.8 ± 0.0	0.099 ± 0.179
ReAct+Reject	25.9 ± 16.0	73.5 ± 5.3	−0.006 ± 0.210
Ours	53.7 ± 8.5	62.3 ± 2.1	0.141 ± 0.056

Table 5: Falsification prompt ablation (3 repeats). Adding rejection instructions inverts the bias but MCC collapses to zero with higher variance, confirming balanced discrimination requires architectural support.

Significance and runtime. We assess significance with McNemar's test at two levels and report both (Table 6): pooling all 216 predictions, and a query-level test that majority-votes each of the 72 queries across its three repeats into 72 independent paired units. Under the query-level test our agent significantly outperforms the three affirmative baselines (ReAct p=0.007, ToT $p<0.001$, MCTSr p=0.003), with query-level accuracy 63.9% versus 38.9% (ReAct), 30.6% (ToT), and 37.5% (MCTSr). The comparison with AlphaApollo is not significant under McNemar; as noted in Table 6, this reflects a rejection-biased agent matching our raw per-case correctness on a negative-heavy set, while our discriminative advantage shows in MCC ($+0.141$ vs -0.100). On runtime, our agent averages 1465s per query, 3.2 times ReAct (454s) and comparable to MCTSr (1535s), bounded by the 10-step budget.

Comparison	Pooled (n=216)		Query-level (n=72)	
	χ^2	p	χ^2	p
Ours vs. ReAct	13.45	<0.001	7.22	0.007
Ours vs. Tree-of-Thoughts	32.85	<0.001	13.22	<0.001
Ours vs. MCTSr	16.83	<0.001	8.76	0.003
Ours vs. AlphaApollo	0.00	1.00	0.04	0.85

Table 6: McNemar's test (continuity-corrected) comparing our agent to each baseline. Pooled treats all 216 predictions as independent; query-level majority-votes each query's three repeats into 72 independent units. Our agent is significant against the three affirmative baselines under both. The comparison with AlphaApollo is not significant because McNemar tests raw per-case correctness: on a negative-heavy set a rejection-biased agent gets a comparable *number* of cases right, so the two agents disagree on roughly as many cases in each direction, even though they are correct for opposite reasons. The discriminative difference is captured by MCC (Table 1): $+0.141$ for our agent versus -0.100 for AlphaApollo.

5 Conclusions

We presented an information-theoretic framework for agentic scientific exploration that selects actions by expected information gain over competing hypotheses rather than over unbounded knowledge-source content. On cancer-type prediction across 18 TCGA cancer types, it is the only method to achieve balanced sensitivity and specificity and consistently positive MCC, whereas every baseline collapses to a class bias, defaulting to affirmative answers (ReAct, Tree-of-Thoughts, MCTSr) or to blanket rejection (AlphaApollo), a limitation that post-hoc rejection prompting inverts rather than corrects. This points to the bias being rooted in the confirmatory exploration trajectory rather than the final answer step, though whether exploration-time falsification prompting could repair it without a belief distribution over competing hypotheses remains an open question.

Limitations and future work. Action selection depends on meaningful hypothesis partitions, since an uninformative partition provides no discriminative signal for EIG, though low-belief misassignments have bounded entropy impact (Proposition 1). Re-evaluating the partition for every candidate action at each step is the dominant cost: the agent averages 1,465s per query, 3.2× ReAct (Section 4), so larger action sets or longer horizons would benefit from amortized or batched partitioning. The utility also exposes over twenty fixed weights (Table 10). Although the adaptive schedule and prior anchoring make it robust to moderate perturbations, this manual configuration is a real dependency for transfer, and automated weight scheduling or Bayesian optimization is a natural next step. Finally, the evaluation covers a single domain, and the benchmark's difficulty tiers carry a literature-frequency confound (Section 4). While the core mechanisms are domain-agnostic, transferability and evaluation on standardized multi-relational knowledge bases remain to be demonstrated.

AI use statement

Because our method is itself an LLM-driven agent, language models appear in this paper in two distinct roles. As objects of study, the agent and its baselines run on a fixed LLM whose configuration, tool interface, and full prompts we document in Section 3 and Appendix D.3. As authoring aids, we additionally used generative AI to help draft and edit prose and to help write the evaluation and analysis code. Every AI-assisted passage and code path was checked and validated by the authors, who are solely accountable for the claims and artifacts presented here.

Reproducibility statement

The paper is written to be reproducible from its description. How we build the benchmark, i.e., extracting coexpression networks from TCGA, restricting to OncoKB genes, thresholding correlations, and choosing same-organ, adjacent, and distant negatives, is laid out in Section 3 with full detail in Appendices C and D. The exact agent and baseline setups, the common tool suite, and the construction of each prompt appear in Appendix D.3, while Appendix D.4 tabulates every weight and threshold and Appendix B states the algorithm and each scoring term. Reported numbers are averaged over three independent runs with standard deviations.

REFERENCES

Joaquim Aguirre-Plans, Bingsheng Chen, Susan Dina Ghiassian, Alex Jones, Viatcheslav R. Akmaev, Alif Saleh, Deisy Morselli Gysi, and Albert-László Barabási. Enhancing the accuracy of network medicine through understanding the impact of sample size in gene co-expression networks. *Network and Systems Medicine*, 25 (1):0002, May 2025. doi: 10.14293/NSM.25.1.0002.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Mach. Learn.*, 47(2–3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 1479–1487, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(8073):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1lJJnR5Ym.

Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E. Rudolph, Rona Yaeger, Tara Soumerai, Moriah H. Nissan, Matthew T. Chang, Sarat Chandarlapaty, Tiffany A. Traina, Paul K. Paik, Alan L. Ho, Feras M. Hantash, Andrew Grupe, Shrujal S. Baxi, Margaret K. Callahan, Alexandra Snyder, Ping Chi, Daniel C. Danila, Mrinal Gounder, James J. Harding, Matthew D. Hellmann, Gopa Iyer, Yelena Y. Janjigian, Thomas Kaley, Douglas A. Levine, Maeve Lowery, Antonio Omuro, Michael A. Postow, Dana Rathkopf, Alexander N. Shoushtari, Neerav Shukla, Martin H. Voss, Ederlinda Paraiso, Ahmet Zehir, Michael F. Berger, Barry S. Taylor, Leonard B. Saltz, Gregory J. Riely, Marc Ladanyi, David M. Hyman, José Baselga, Paul Sabbatini, David B. Solit, and Nikolaus Schultz. OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, 1(1):1–16, November 2017. ISSN 2473-4284. doi: 10.1200/po.17.00011.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pp. 273–304, 1995.

Nick Chater, Mike Oaksford, Ulrike Hahn, and Evan Heit. Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):811–823, 2010.

Evan A. Clayton, Toyya A. Pujol, John F. McDonald, and Peng Qiu. Leveraging TCGA gene expression data to build predictive models for cancer drug response. *BMC Bioinformatics*, 21(S14), September 2020. ISSN 1471-2105. doi: 10.1186/s12859-020-03690-4.

M.O.G. Duff. *Optimal Learning: Computational Procedures for Bayes-adaptive Markov Decision Processes*. University of Massachusetts at Amherst, 2002.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. April 2024. URL https://www.microsoft.com/en-us/research/publication/from-local-to-global-a-graph-rag-approach-to-query-focused-summarization/.

Alireza Ghafarollahi and Markus J. Buehler. SciAgents: Automating Scientific Discovery Through Bioinspired Multi-Agent Intelligent Graph Reasoning. *Advanced Materials*, 37(22):2413523, 2025. doi: https://doi.org/10.1002/adma.202413523.

Mourad Gridach, Jay Nanavati, Christina Mack, Khaldoun Zine El Abidine, and Lenon Mendes. Agentic AI for scientific discovery: A survey of progress, challenges, and future directions. In *Towards Agentic AI for Science: Hypothesis Generation, Comprehension, Quantification, and Validation*, 2025.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495.

Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 1962.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Qingyao Li, Wei Xia, Xinyi Dai, Kounianhua Du, Weiwen Liu, Yasheng Wang, Ruiming Tang, Yong Yu, and Weinan Zhang. RethinkMCTS: Refining erroneous thoughts in Monte Carlo tree search for code generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8092–8110, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.410.

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

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4): 590–604, 07 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.4.590.

Nhung TH Nguyen, Dmytro S Lituiev, Zhimin Liu, Aditya Kashyap, Garrett Jenkinson, Kevin Kuhl, Christopher Corrado, Naisargi Manishkumar Patel, Kirti Snigdha, Sirwe Saeedi, et al. Med. AI ASK: an agentic system for biomedical question answering. *Journal of the American Medical Informatics Association*, pp. ocag038, 2026.

Joel Nulsen, Hrvoje Misetic, Christopher Yau, and Francesca D. Ciccarelli. Pan-cancer detection of driver genes at the single-patient resolution. *Genome Medicine*, 13(1), February 2021. ISSN 1756-994X. doi: 10.1186/s13073-021-00830-0.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 2778–2787. JMLR.org, 2017.

Karl Popper. *The logic of scientific discovery*. Routledge, 2005.

Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Herbert Robbins. An Empirical Bayes Approach to Statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 157–163. University of California Press, 1956.

Samuel Schapiro, Jonah Black, and Lav R Varshney. Transformational Creativity in Science: A Graphical Theory. In *Proceedings of the 16th International Conference on Computational Creativity (ICCC)*, 2025. URL https://computationalcreativity.net/iccc25/papers/iccc25-schapiro2025transformational.pdf.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Scott Silvey, Amy Olex, Shaojun Tang, and Jinze Liu. Sample size requirements for machine learning classification of binary outcomes in bulk rna-seq data. *BMC Bioinformatics*, 27(1):06372, January 2026. doi: 10.1186/s12859-026-06372-9.

Karen Sparck Jones. *A statistical interpretation of term specificity and its application in retrieval*, pp. 132–142. Taylor Graham Publishing, GBR, 1988. ISBN 0947568212.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012. ISSN 1611-7530. doi: 10.1007/s12064-011-0142-z.

Aleksandar Stojmirović and Yi-Kuo Yu. Robust and accurate data enrichment statistics via distribution function of sum of weights. *Bioinformatics*, 26(21):2752–2759, 11 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq511.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nnVO1PvbTv.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.557.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *NeurIPS*, 2022. ISBN 9781713871088.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. 2025. URL https://arxiv.org/abs/2504.08066.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822. Curran Associates, Inc., 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*, 2023b.

Di Zhang, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. Accessing GPT-4 level Mathematical Olympiad Solutions via Monte Carlo Tree Self-refine with LLaMa-3 8B. *arXiv preprint arXiv:2406.07394*, 2024. URL https://arxiv.org/abs/2406.07394.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language Agent Tree Search Unifies Reasoning, Acting, and Planning in Language Models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62138–62160. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhou24r.html.

Zhanke Zhou, Chentao Cao, Xiao Feng, Xuan Li, Zongze Li, Xiangyu Lu, Jiangchao Yao, Weikai Huang, Tian Cheng, Jianghangfan Zhang, Tangyu Jiang, Linrui Xu, Yiming Zheng, Brando Miranda, Tongliang Liu, Sanmi Koyejo, Masashi Sugiyama, and Bo Han. Alphaapollo: A system for deep agentic reasoning. *Transactions on Machine Learning Research (TMLR)*, 2026. arXiv:2510.06261.

James Zou and Eric J Topol. The rise of agentic AI teammates in medicine. *The Lancet*, 405(10477):457, 2025.

A PARTITION ENTROPY AND EIG

We treat the hypotheses $\mathcal{H} = \{h_1, \ldots, h_n\}$ as mutually exclusive and exhaustive, so that $\mathbf{b}_t$ is a genuine distribution over a single latent truth H (in our benchmark, exactly one cancer type generates the network). The traditional expected information gain of an action marginalizes over possible observations:

$$\mathrm{EIG}(a) = \mathbb{E}_{o \sim P(o|a,\mathcal{S}_t)} \left[D_{\mathrm{KL}}(P(\mathcal{H} \mid o, a, \mathcal{S}_t) \parallel P(\mathcal{H} \mid \mathcal{S}_t)) \right] = I(H; O \mid a), \tag{3}$$

the mutual information between the latent truth H and the observation O. Let $Z \in \{1, \ldots, M\}$ be the partition-group index, a deterministic coarsening $Z = g(H)$ that maps each hypothesis to its group, with prior mass $p(P_m) = \sum_{h_i \in P_m} b_t(h_i) = P(Z = m)$. We restate Proposition 1 (Section 2.4): under the Markov chain $H \to Z \to O$ (equivalently $O \perp H \mid Z$), $\mathrm{EIG}(a) = I(H; O \mid a) \leq H_{\text{partition}}(a) = H(Z)$, with equality when the outcome perfectly identifies the group.

Proof. Under $H \to Z \to O$, the data-processing inequality gives $I(H; O) \leq I(Z; O)$. Since $Z = g(H)$ is a deterministic function of H, and mutual information is non-negative,

$$I(Z; O) = H(Z) - H(Z \mid O) \leq H(Z) = H_{\text{partition}}(a), \tag{4}$$

so $\mathrm{EIG}(a) = I(H; O) \leq H_{\text{partition}}(a)$. For the equality case, suppose the outcome perfectly identifies the group: observing o_m (confirming group P_m) yields posterior $b'(h_i \mid o_m) = b_t(h_i)/p(P_m)$ for $h_i \in P_m$ and zero otherwise, so

$$D_{\mathrm{KL}}(b' \| b_t) = \sum_{h_i \in P_m} \frac{b_t(h_i)}{p(P_m)} \log \frac{1}{p(P_m)} = -\log p(P_m). \tag{5}$$

The outcome o_m occurs with probability $p(P_m)$, hence

$$\mathrm{EIG}(a) = \sum_{m=1}^{M} p(P_m) \left[-\log p(P_m) \right] = H(Z) = H_{\text{partition}}(a). \tag{6}$$

Here $H(Z \mid O) = 0$ because the outcome resolves group membership; more generally $H(Z \mid O) > 0$ and the inequality is strict. □

Remark. The Markov condition $H \to Z \to O$ decomposes into two parts. The map $Z = g(H)$ is deterministic, so $H \to Z$ holds exactly. The substantive assumption is $Z \to O$, i.e. $I(H; O \mid Z) = 0$: once the correct group is known, the observation carries no further signal about which hypothesis *within* that group is true. This is not automatic. By the chain rule $I(H; O) = I(Z; O) + I(H; O \mid Z)$, so whenever the intra-group term is positive the bound can be violated and $\mathrm{EIG}(a)$ may exceed $H_{\text{partition}}(a)$: an action that resolves two hypotheses placed in the *same* group contributes information the partition entropy never counts. The condition therefore holds only to the extent that the partition groups hypotheses by genuinely distinct predicted outcomes, and its weakest point is the leftover group, where unrelated hypotheses are pooled and an incidental observation is most likely to separate them. A good partition drives $I(H; O \mid Z) \to 0$ and the criterion is informative. A poor one leaves large intra-group information uncounted, so EIG no longer tracks the action's true value. This is exactly the degradation measured in the partition-quality ablation (Table 4): replacing meaningful groupings with random ones inflates $I(H; O \mid Z)$, and the discriminative signal collapses.

The bound is tight when actions target entities about which hypotheses make sharply distinct, testable predictions, and loosest when multiple groups would produce similar observational signatures. By preferring actions with high partition entropy, the agent selects actions with high *potential* information gain, which is realized to the extent that retrieved evidence cleanly distinguishes between groups.

The bound is optimistic, selecting actions as if observations will perfectly resolve group membership, in the spirit of UCB strategies. Misassigning a hypothesis of belief mass ϵ perturbs $H_{\text{partition}}$ by $O(\epsilon \log(1/\epsilon))$, so the action ranking is preserved when the utility margin between top actions exceeds the cumulative error. This guarantee is weakest early, when beliefs are near-uniform and no high-belief hypothesis dominates the entropy; there the exploration and novelty bonuses (Equation 2) and prior anchoring (Equation 11) buffer partitioning noise until beliefs sharpen.

Acquisition proxy. Proposition 1 characterizes $H_{\text{partition}}$ under exact Bayesian conditioning on an idealized observer. The deployed agent does not perform this exact update: beliefs are recomputed from all accumulated evidence via the composite score (Equation 8), a temperature softmax, and prior anchoring (Equation 11). We therefore use $H_{\text{partition}}$ strictly as an *acquisition function* for action selection, motivated by the idealized observer, not as a prediction of the realized change in the agent's operational belief $\mathbf{b}_{t+1}$. The empirical value of the criterion is established directly by the partition-quality ablation (Table 4), where replacing meaningful partitions with random ones removes the signal.

B ALGORITHM AND SCORING-TERM DEFINITIONS

This appendix gives the full exploration procedure (Algorithm 1), the belief-update sub-procedure (Algorithm 2), and precise definitions of every scoring term in Equation 8.

Algorithm 1 Information-theoretic exploration loop

Input: query q; initial memory graph $\mathcal{G}_0$ (possibly empty); action space $\mathcal{A}$; budget $B_{\max}$, $T_{\max}$
Output: best-supported hypothesis and evidence-based synthesis

$\mathcal{H} \leftarrow$ GENERATEHYPOTHESES$(q, \mathcal{G}_0)$ ▷ propose, merge, filter (Sec. 2)
$\mathbf{b}_0(h_i) \leftarrow 1/|\mathcal{H}|$ for all $h_i \in \mathcal{H}$ ▷ uniform prior
$t \leftarrow 0$
while not TERMINATE$(\mathcal{S}_t, B_{\max}, T_{\max})$ **do** ▷ budget / convergence / stall
  $\mathcal{A}_t \leftarrow$ GENERATECANDIDATEACTIONS$(\mathcal{S}_t)$
  $\{P_1, \ldots, P_M\}_a \leftarrow$ PARTITION$(a, \mathcal{H})$ for each $a \in \mathcal{A}_t$ ▷ group hypotheses by how a differentiates them
  $a_t \leftarrow \arg\max_{a \in \mathcal{A}_t} U(a \mid \mathcal{S}_t)$ ▷ Eq. 2; EIG from Eq. 1
  $o_t \leftarrow$ EXECUTE(a_t) ▷ PubMed / UniProt / Reactome / graph analysis
  $\mathcal{R}_t \leftarrow$ EXTRACTRULES(o_t); $\mathcal{G}_{t+1} \leftarrow \mathcal{G}_t \cup \mathcal{R}_t$
  $\mathbf{b}_{t+1} \leftarrow$ UPDATEBELIEF$(\mathcal{H}, \mathcal{R}_{\le t}, \mathcal{G}_{t+1})$ ▷ Algorithm 2
  promote h_i to axiom if $b_{t+1}(h_i) > \tau_{\text{axiom}}$; spawn downstream hypotheses
  $t \leftarrow t + 1$
end while
return EVIDENCEDEBATESYNTHESIS$(\mathcal{S}_t)$ ▷ quality audit before final answer

Termination conditions (Section 2). The loop terminates, checked after each belief update, when any of the following holds: (1) *budget exhaustion*, the total action cost exceeds $B_{\max}$ or the step count reaches $T_{\max}$; (2) *belief convergence*, the normalized entropy $H(\mathbf{b}_t)/\log|\mathcal{H}|$ stays below a stability threshold ϵ for W_e consecutive steps; or (3) *exploration stall*, no new confirmed findings for W_s consecutive steps. On termination, an evidence-quality audit evaluates all hypotheses holistically, assessing whether evidence is specific or generic, whether early-promoted axioms are contradicted by later findings, and whether belief magnitudes reflect evidence quality, before the final synthesis. Graph-structural features (centrality, community detection) are computed once at initialization.

Algorithm 2 UPDATEBELIEF: recompute beliefs from all accumulated evidence

Input: hypotheses $\mathcal{H}$; accumulated rules $\mathcal{R}_{\le t}$; memory graph $\mathcal{G}_t$
Output: updated belief distribution $\mathbf{b}_t$

route each rule in $\mathcal{R}_t$ to hypotheses whose nodes it anchors to; record contradictions
$s(v) \leftarrow 1/|\{h_i : v \in \mathcal{N}_i\}|$ for each node v ▷ specificity
for each $h_i \in \mathcal{H}$ **do**
  compute factors $f_e, f_c, f_d, f_g, f_b, f_q, f_n, f_p$ ▷ defined below
  $\delta_t \leftarrow \max(\delta_{\min}, 1 - E_{\text{total}}/\tau)$ ▷ memory decay
  $\text{Score}(h_i) \leftarrow w_e f_e + \delta_t(w_c f_c + w_d f_d + w_g f_g) + w_b f_b + w_q f_q + w_n f_n - w_p f_p$
end for
$b_t^{\text{post}} \leftarrow$ SOFTMAX$_{h_i \in \mathcal{I}_t}(\text{Score}/\tau_s)$ ▷ normalized over investigated set $\mathcal{I}_t$
blend with renormalized prior via anchoring (Eq. 11); uninvestigated h_i keep $b_0(h_i)$
return $\mathbf{b}_t$

Scoring-term definitions (Equation 8). Let $\mathcal{R}_i$ be the rules supporting h_i, $\mathcal{N}_i$ its candidate nodes, and $s(r) = \max_{v \in \mathcal{V}_r} s(v)$ the rule specificity. All node-topology factors are normalized to $[0, 1]$ by the maximum over graph nodes.

- $f_e(h_i) = \log(1 + |\mathcal{R}_i|) \cdot \frac{1}{|\mathcal{R}_i|} \sum_{r \in \mathcal{R}_i} c(r)\, s(r)$: specificity-weighted *evidence strength* (Eq. 9).
- $f_c(h_i) = \text{mean}_{v \in \mathcal{N}_i} \widehat{\text{btw}}(v)$: normalized *betweenness centrality* of the hypothesis nodes, averaged over $\mathcal{N}_i$ (reducing to the single node's value when $|\mathcal{N}_i| = 1$).
- $f_d(h_i) = \text{mean}_{v \in \mathcal{N}_i} \deg(v)/\max_u \deg(u)$: normalized *degree connectivity*.
- $f_g(h_i)$: *graph support*, the precomputed hub score assigned at hypothesis proposal (max normalized hub score over $\mathcal{N}_i$, with exclusivity weighting).

- $f_b(h_i)$: *biomedical support* accumulated from structured-database evidence (UniProt/Reactome), incremented by supporting and decremented by contradicting findings; not subject to δ_t as it is an independent evidence source.
- $f_q(h_i) \in [0,1]$: *query relevance*, the embedding cosine similarity between the hypothesis description and the query q.
- $f_n(h_i) = \log(1 + |\mathcal{R}_i|)$: *evidence count* (log rule count).
- $f_p(h_i) = \log(1 + |\mathcal{C}_i|) \cdot \frac{1}{|\mathcal{C}_i|} \sum_{r \in \mathcal{C}_i} c(r)$: *contradiction penalty*, aggregating the confidences of the contradicting rules $\mathcal{C}_i$ analogously to f_e.

The evidence factors f_e and f_p are defined to be 0 for a hypothesis with no supporting rules ($\mathcal{R}_i = \emptyset$) or no contradicting rules ($\mathcal{C}_i = \emptyset$), respectively. The $\frac{1}{|\mathcal{R}_i|}$ and $\frac{1}{|\mathcal{C}_i|}$ factors are only applied when the corresponding rule set is non-empty, consistent with the limiting value $\log(1+k) \cdot \bar{c} \to 0$ as $k \to 0$.

Challenger EIG and novelty (Equation 2). The challenger term sustains falsification pressure on the leading hypothesis $h^* = \arg\max_i b_t(h_i)$. We renormalize beliefs over the non-leader hypotheses into a challenger distribution $b_t^{\text{chal}}(h_i) \propto b_t(h_i)$ for $h_i \neq h^*$ and compute $\text{EIG}_{\text{chal}}(a)$ by the same partition method (Section 2.4) applied to b_t^{chal}. Its weight ramps as the leader grows dominant, $\alpha_2^{(t)} = \alpha_2 \cdot \max(1,\, b_t(h^*)/\tau_{\text{chal}})$, so the agent gathers evidence against the leader before committing. Novelty penalizes actions whose description embedding e_a lies close to the embeddings of past observations, $\text{Nov}(a) = \min_j \|e_a - e_j\|^2 / 2\sigma^2$, adapting count-based exploration bonuses (Bellemare et al., 2016; Pathak et al., 2017). Exact repeats and previously failed actions receive near-zero novelty.

Coherence and exploration (Equation 2). Coherence is $\text{Coh}(a) = w_{\text{rule}}\text{Coh}_{\text{rule}}(a) + w_{\text{ev}}\text{Coh}_{\text{ev}}(a) + w_{\text{cross}}\text{Coh}_{\text{cross}}(a)$, where Coh_{rule} is one minus the fraction of contradicting rule pairs among rules sharing anchors with a, Coh_{ev} is the mean evidence support of hypotheses overlapping a's targets, and $\text{Coh}_{\text{cross}} = |\{h_i : \mathcal{N}_i \cap \mathcal{V}_a \neq \emptyset\}|/|\mathcal{H}|$. When a targets new nodes overlapping no existing hypothesis, Coh_{ev} and Coh_{rule} default to a neutral 0.5 (no evidence for or against consistency), so coherence neither rewards nor penalizes actions into unexplored regions. The exploration term $\text{Expl}(a)$ sums a fixed bonus for targeting an uninvestigated hypothesis, a coverage bonus for under-explored communities, and an evidence-gap bonus for literature searches on confirmed findings lacking references (weights in Table 10). The adaptive weight schedule dampens the EIG weight and boosts the novelty and exploration weights once the leading belief exceeds the convergence threshold $\theta_c = 1.5/|\mathcal{H}|$:

$$\alpha_k^{(t)} = \begin{cases} \alpha_k / \rho & k = 1 \text{ (EIG)} \\ \alpha_k \cdot \rho & k \in \{3, 6\} \text{ (novelty, exploration)} \end{cases}, \quad \rho = \max\left(1,\ \frac{\max_i b_t(h_i)}{\theta_c}\right). \tag{7}$$

The clamp $\rho \geq 1$ ensures the schedule is inactive while beliefs are near-uniform ($\max_i b_t(h_i) \leq \theta_c$): early steps then run at the base weights α_k, keeping exploration dominant, and the EIG-damping/exploration-boosting only engages once a leader emerges. Without it, $\rho < 1$ in the early regime would invert the intended trajectory, inflating EIG and suppressing exploration exactly when broad coverage is wanted.

Belief-update scoring (Section 2). Each hypothesis h_i receives the composite score

$$\begin{aligned} \text{Score}(h_i) = w_e\, f_e(h_i) + \delta_t\, (w_c\, f_c(h_i) + w_d\, f_d(h_i) + w_g\, f_g(h_i)) \\ + w_b\, f_b(h_i) + w_q\, f_q(h_i) + w_n\, f_n(h_i) - w_p\, f_p(h_i), \end{aligned} \tag{8}$$

with the eight factors defined above and the specificity-weighted evidence strength given explicitly by

$$f_e(h_i) = \log(1 + |\mathcal{R}_i|) \cdot \frac{1}{|\mathcal{R}_i|} \sum_{r \in \mathcal{R}_i} c(r) \cdot s(r). \tag{9}$$

Evidence-derived weights (w_e, w_b, w_n) are set higher than structural priors (w_c, w_d, w_g), so retrieved evidence dominates as exploration progresses. This is reinforced by a memory-decay multiplier $\delta_t = \max(\delta_{\min},\ 1 - E_{\text{total}}/\tau)$, where $E_{\text{total}} = \sum_i |\mathcal{R}_i|$ is the total rule count and τ the decay horizon: at step 0 topology is the only signal, and as evidence accumulates $\delta_t \to \delta_{\min}$ and beliefs become evidence-driven, analogous to prior-to-posterior transitions in empirical Bayes methods (Robbins, 1956). The structured-database factor f_b is not decayed, being an independent evidence source.

Softmax with prior anchoring. A hypothesis is *investigated* once an executed action has targeted at least one of its candidate nodes ($\mathcal{N}_i \cap \mathcal{V}_{a_\tau} \neq \emptyset$ for some $\tau \leq t$). We write $\mathcal{I}_t$ for this set and $m_{\text{un}} = \sum_{h_i \notin \mathcal{I}_t} b_0(h_i)$ for the prior mass held by the rest. Scores are converted to beliefs via a temperature softmax taken *over the investigated set* so that the blended distribution normalizes exactly:

$$b_t^{\text{post}}(h_i) = \frac{\exp(\text{Score}(h_i)/\tau_s)}{\sum_{h_j \in \mathcal{I}_t} \exp(\text{Score}(h_j)/\tau_s)}, \quad h_i \in \mathcal{I}_t. \tag{10}$$

To prevent premature belief collapse, uninvestigated hypotheses retain their initial prior $b_0(h_i)$, while investigated hypotheses share the remaining mass $1 - m_{\text{un}}$ via a prior-posterior blend:

$$b_t(h_i) = \left[\lambda_t \cdot \bar{b}_0(h_i) + (1 - \lambda_t) \cdot b_t^{\text{post}}(h_i)\right] \cdot (1 - m_{\text{un}}), \quad h_i \in \mathcal{I}_t, \tag{11}$$

where $\bar{b}_0(h_i) = b_0(h_i) / \sum_{h_j \in \mathcal{I}_t} b_0(h_j)$ is the prior renormalized over investigated hypotheses, and $\lambda_t = \lambda_0(1 - |\mathcal{I}_t|/|\mathcal{H}|)$ decays as more hypotheses are investigated. Since $\sum_{h_i \in \mathcal{I}_t} \bar{b}_0(h_i) = \sum_{h_i \in \mathcal{I}_t} b_t^{\text{post}}(h_i) = 1$, the investigated block sums to $1 - m_{\text{un}}$ and the full distribution over $\mathcal{H}$ sums to 1. This mechanism, inspired by optimistic initialization in bandit algorithms (Auer et al., 2002), ensures under-explored hypotheses retain sufficient belief mass to attract future investigation while allowing beliefs to converge to the pure evidence-based posterior as exploration progresses.

Relation to Bayesian updating. The belief update is not the exact Bayesian conditioning assumed by Proposition 1, but it is not an arbitrary heuristic either. Since $b_t^{\text{post}}(h_i) \propto \exp\big(\sum_k (w_k/\tau_s)\, f_k(h_i)\big)$, each term $w_k f_k$ acts as a log-evidence contribution and the softmax normalizes them into a posterior, with τ_s controlling how sharply evidence is weighed. Prior anchoring (Eq. 11) blends this toward the prior b_0 with a weight λ_t that fades as evidence accumulates. Two things keep it approximate rather than exact: the factors are added as if independent when they are not, so correlated evidence is partly double-counted, and the weights w_k are tuned on a development set rather than estimated as calibrated likelihoods. The update therefore trades exact conditioning for robustness to noisy, correlated evidence extracted by an LLM.

C Ground Truth of the Classification Task

For each cancer type, we define three categories of negative classes based on similarity, while the positive class is the cancer type itself. The negative classes are manually selected to represent varying degrees of difficulty for the agent, ensuring a comprehensive evaluation of its discriminative capabilities. The criteria for selecting negative classes are as follows:

- **Hard negative**: a cancer type from the same organ system.
- **Medium negative**: a cancer type with adjacent organ system.
- **Easy negative**: a cancer type from a different organ system.

Table 7 presents the resulting negative class assignments.

No.	Cancer full name	Acronym	N	Hard Neg.	Medium Neg.	Easy Neg.
1	Breast invasive carcinoma	BRCA	200	OV	LUAD	KIRC
2	Lung adenocarcinoma	LUAD	200	LUSC	BRCA	KIRC
3	Lung squamous cell carcinoma	LUSC	200	LUAD	HNSC	BRCA
4	Colon adenocarcinoma	COAD	200	STAD	LIHC	BRCA
5	Prostate adenocarcinoma	PRAD	200	KIRC	BLCA	COAD
6	Stomach adenocarcinoma	STAD	200	COAD	LIHC	BRCA
7	Liver hepatocellular carcinoma	LIHC	200	COAD	STAD	BRCA
8	Kidney renal clear cell carcinoma	KIRC	200	KIRP	PRAD	BRCA
9	Kidney renal papillary cell carcinoma	KIRP	200	KIRC	KICH	BRCA
10	Kidney chromophobe	KICH	91	KIRC	KIRP	BRCA
11	Bladder urothelial carcinoma	BLCA	200	PRAD	KIRC	BRCA
12	Uterine corpus endometrial carcinoma	UCEC	200	KIRC	BRCA	SKCM
13	Uterine carcinosarcoma	UCS	57	SARC	BRCA	PAAD
14	Head and neck squamous cell carcinoma	HNSC	200	LUSC	STAD	BRCA
15	Thyroid carcinoma	THCA	200	HNSC	LUAD	MESO
16	Ovarian serous cystadenocarcinoma	OV	100	UCEC	BRCA	KIRC
17	Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	200	KIRC	BRCA	SKCM
18	Skin cutaneous melanoma	SKCM	200	UVM	LUAD	BRCA
19	Uveal melanoma	UVM	80	SKCM	GBM	MESO
20	Glioblastoma multiforme	GBM	200	LGG	SKCM	BRCA
21	Brain lower grade glioma	LGG	200	GBM	SKCM	TGCT
22	Pheochromocytoma and paraganglioma	PCPG	187	ACC	BRCA	THCA
23	Esophageal carcinoma	ESCA	198	PAAD	PCPG	BRCA
24	Rectum adenocarcinoma	READ	177	PAAD	PCPG	BRCA
25	Cholangiocarcinoma	CHOL	44	PAAD	PCPG	BRCA
26	Pancreatic adenocarcinoma	PAAD	183	CHOL	PCPG	BRCA
27	Sarcoma	SARC	200	UCS	BRCA	PAAD
28	Mesothelioma	MESO	87	UCS	BRCA	PAAD
29	Adrenocortical carcinoma	ACC	79	PCPG	BRCA	THCA
30	Thymoma	THYM	122	PCPG	BRCA	THCA
31	Testicular germ cell tumors	TGCT	156	KIRC	BRCA	SKCM
32	Acute myeloid leukemia	LAML	150	DLBC	BRCA	CHOL
33	Lymphoid neoplasm diffuse large B-cell lymphoma	DLBC	48	LAML	BRCA	CHOL

Table 7: Cancer types and negative class assignments for the binary classification task. N denotes the number of TCGA samples used for network construction, capped at 200 for comparability across cancer types. Shaded rows indicate cancer types with fewer than 200 total samples available in TCGA, which are excluded from the primary evaluation due to insufficient samples for reliable coexpression network construction (Aguirre-Plans et al., 2025; Nulsen et al., 2021).

D EVALUATION SETTINGS

D.1 NETWORK CONSTRUCTION PIPELINE

We construct one gene coexpression network per TCGA cohort as follows. (i) *Expression matrix.* We use TCGA primary-tumor RNA-seq expression quantified as TPM (log-transformed), restricted to protein-coding genes. (ii) *Gene filtering.* We retain only OncoKB cancer genes (Chakravarty et al., 2017), focusing the network on established oncogenic drivers rather than the full transcriptome. (iii) *Sampling.* For each cohort we draw up to 200 primary-tumor samples. Cohorts with fewer than 200 available samples are excluded from the primary evaluation (Table 7) because coexpression estimates are unstable below this size (Aguirre-Plans et al., 2025; Nulsen et al., 2021; Clayton et al., 2020; Silvey et al., 2026). (iv) *Coexpression edges.* We compute pairwise gene–gene correlations across samples and place an undirected edge between two genes when their correlation satisfies $r \geq 0.7$. The resulting network is stored as a property graph and initializes the agent's memory without revealing the cohort label. Table 8 reports the size and density of each of the 18 evaluated networks.

Cohort	Nodes	Edges	Density	Cohort	Nodes	Edges	Density
BRCA	631	3,946	0.020	BLCA	525	3,071	0.022
LUAD	554	3,270	0.021	UCEC	662	15,306	0.070
LUSC	492	2,282	0.019	HNSC	740	19,462	0.071
COAD	726	10,012	0.038	THCA	965	67,177	0.144
PRAD	884	37,025	0.095	CESC	469	3,585	0.033
STAD	782	17,774	0.058	SKCM	655	19,061	0.089
LIHC	789	32,614	0.105	GBM	955	44,546	0.098
KIRC	913	32,026	0.077	LGG	776	10,728	0.036
KIRP	878	46,104	0.120	SARC	471	4,750	0.043

Table 8: Size and density of the 18 evaluated coexpression networks (OncoKB genes, TPM, $r \geq 0.7$). Density is $2|E|/(|V|(|V|-1))$. Node counts reflect OncoKB genes surviving the correlation threshold with at least one edge.

D.2 AGENT SETTINGS

All four agents operate under identical constraints:

- Same LLM: Claude Sonnet 4.5 via AWS Bedrock.
- Same tools: PubMed search (via Weaviate), UniProt queries, Reactome pathway enrichment (via SaddleSum, Stojmirović & Yu, 2010), and graph analysis (LLM-generated NetworkX code).
- Same budget: 10 exploration steps.
- Same input: Identical graph and query.

The LLM runs at its default sampling temperature, so all agents are stochastic across the three repeated runs (hypothesis generation, partition grouping, retrieval queries, and evidence interpretation vary between runs). A few table cells nonetheless report a standard deviation of exactly zero. These arise when all three runs yield identical confusion-matrix counts for that metric and tier rather than from any determinism or caching. It happens most often for the strongly biased baselines: ReAct, for example, answers *yes* to nearly every query, so the fraction of negatives it correctly rejects lands on the same value in each run, giving a specificity standard deviation of zero. It also occurs occasionally for our own agent and the ablations when a metric resolves to the same count each run, such as our agent correctly rejecting the same subset of hard negatives across all three runs.

D.3 BASELINE CONFIGURATIONS

All baselines share the LLM, tool suite, 10-step budget, and input graph of our agent, and use a common tool interface, evidence requirement, and citation format. Their search hyper-parameters are given in Table 9.

Baseline	Parameter	Value
ReAct	max steps	10
Tree-of-Thoughts	beam width	2
	max depth	5
MCTSr	max rollouts	10
	exploration constant (UCB)	1.0

Table 9: Baseline search hyper-parameters. ReAct interleaves thought/action/observation; Tree-of-Thoughts performs beam search over reasoning paths that make real tool calls; MCTSr runs Monte Carlo tree search with self-refinement (Zhang et al., 2024).

ReAct prompt. The shared exploration prompt instructs the agent to interleave `THOUGHT` and `ACTION` steps over the tool suite (`pubmed_search`, `uniprot_query`, `pathway_enrichment`, `graph_analysis`), build evidence from structural to functional to mechanistic to clinical, and emit `ACTION: FINISH` with a final analysis. Tree-of-Thoughts and MCTSr reuse the same tool interface and evidence/citation instructions, differing only in their search procedure. No baseline receives rejection instructions during exploration.

ReAct+Reject. The ReAct+Reject variant is identical to ReAct during exploration and differs only in the synthesis prompt, which adds explicit falsification instructions: answer *no* unless multiple lines of specific, discriminative evidence uniquely characterize the queried cancer type; treat generic biological facts or evidence that could apply to several cancer types as insufficient; and default to *no* when in doubt. This isolates whether prompt-level rejection guidance, rather than a change to exploration, can close the specificity gap.

D.4 Hyper-parameters for our agent

All hyperparameters are fixed across runs and cancer types, tuned on a small held-out subset not included in the main evaluation. Values are presented in Table 10.

Parameter	**Value**	**Description**
Utility function weights in Equation 2		
α_{eig} (α_1)	1.0	Weight on EIG* term
α_{chal} (α_2)	0.5	Base weight for challenger EIG
α_{nov} (α_3)	0.3	Weight on novelty Nov(a)
α_{coh} (α_4)	0.5	Weight on coherence Coh(a)
α_{cost} (α_5)	0.1	Cost penalty weight (subtracted)
α_{expl} (α_6)	1.0	Weight on exploration term Expl(a); fixed to unity, with the term's magnitude set by the internal bonuses below
γ	0.3	Hypothesis-expansion bonus coefficient (Eq. 1)
τ_{chal}	0.25	$\max_i b_i$ above which challenger weight ramps up
Novelty estimator constants (Section 2.4, Appendix B)		
σ	0.5	Gaussian kernel bandwidth: Nov $= \min_j \|e_a - e_j\|^2/(2\sigma^2)$
embedding	`all-MiniLM-L6-v2` (384-dim) for e_a, e_j	
Nov_{cap}	10.0	Maximum novelty value
$\text{Nov}_{\text{repeat2+}}$	0.05	Novelty for actions executed ≥ 2 times
$\text{Nov}_{\text{repeat1}}$	0.20	Novelty for an already-executed action
$\text{Nov}_{\text{failed}}$	0.02	Novelty for retrying a previously failed action
β_{action}	0.6	Blend weight on action-based novelty
β_{obs}	0.4	Blend weight on observation-based novelty
τ_{overlap}	0.7	Jaccard word-overlap threshold for "very similar" actions
Coherence estimator weights		

Continued on next page

Parameter	Value	Description
w_{rule}	0.35	Rule-contradiction consistency
w_{ev}	0.40	Evidence trajectory of matched hypotheses
w_{cross}	0.25	Cross-hypothesis coverage
Action costs		
PUBMED SEARCH	1.5	
PATHWAY ENRICHMENT	1.5	
GRAPH ANALYSIS	0.8	
UNIPROT QUERY	0.5	
TERMINATE	0.0	
Adaptive bonuses and penalties (exploration term Expl(a), Eq. 2)		
θ_c	$1.5/\lvert\mathcal{H}\rvert$	Belief above which EIG is dampened and exploration is boosted (Eq. 7)
B_{hyp}	5.0	Bonus for targeting an uninvestigated hypothesis
$\Delta_{\text{axiom-node}}$	$+2.0$	Evidence-gap bonus: PubMed on axiom node awaiting literature
$\Delta_{\text{comm-member}}$	$+1.5$	Evidence-gap bonus: PubMed on community member of pending axiom
Δ_{pubmed}	$+0.5$	Evidence-gap bonus: general PubMed when any axiom awaits literature
Δ_{enrich}	-1.0	Evidence-gap penalty: pathway enrichment when literature still needed
$B^{\varnothing}_{\text{cov}}$	$+2.0$	Coverage bonus for a community with no axioms and no rules
$B^{\partial}_{\text{cov}}$	$+1.5$	Coverage bonus for a community with rules but no confirmed axioms
B^{1}_{cov}	$+0.5$	Coverage bonus for a community with exactly one axiom
U_{term}	$0.5 \cdot \max_i b_i$	Utility assigned to the TERMINATE action
Belief scoring weights (Equation 8)		
w_e	1.5	Evidence strength $f_e = \log(1+n_r) \cdot \bar{c} \cdot s$
w_c	0.8	Betweenness centrality f_c
w_d	0.6	Degree connectivity f_d
w_g	0.8	Graph support score f_g
w_b	0.8	Biomedical support f_b
w_q	0.6	Query relevance f_q
w_n	0.2	Evidence rule count f_n
w_p	1.0	Contradiction penalty f_p
τ_s	1.0	Softmax temperature
λ_0	0.4	Prior anchor strength (Equation 11)
δ_{min}	0.2	Floor for memory decay
τ	50	Evidence count for decay horizon
Hypothesis manager parameters (Section 2)		
N_h	15	Hypotheses proposed at initialization (Sec. 2)
$\lvert\mathcal{H}\rvert^{\max}$	20	Maximum number of active hypotheses
$\lvert\mathcal{H}\rvert^{\min}$	5	Minimum hypotheses retained during pruning
τ_{young}	3	Steps since creation below which pruning is blocked
$N_{\text{ev}}^{\text{protect}}$	3	Evidence rules required to gain pruning protection
b_{min}	0.01	Minimum belief to retain a hypothesis
$\delta s^{+}_{\text{bio}}$	$0.2 \cdot c \cdot \rho$	Bio-support increment per supporting rule
$\delta s^{-}_{\text{bio}}$	$0.15 \cdot c \cdot \rho$	Bio-support decrement per contradicting rule
c_0	0.3	Initial confidence for mid-run hypotheses

Continued on next page

Parameter	Value	Description
K_{deep}	3	Max new hypotheses generated per promoted axiom
Termination thresholds (Appendix B)		
B_{max}	250.0	Maximum total accumulated action cost
T_{max}	10	Hard step-count limit
τ_{axiom}	0.85	Belief threshold to promote a hypothesis to axiom
$N_{\text{axiom}}^{\min}$	10	Minimum axiom count before soft termination
W_s	10	Steps without a new confirmed finding before stall-termination
W_e	5	Rolling window size for entropy stability check
ϵ	0.03	Max normalized-entropy change to declare convergence
Retrieval and rule extraction parameters		
α_{hybrid}	0.25	BM25 vs. vector blend in Weaviate hybrid search
k_{docs}	4	Documents fetched per PubMed search call
$R_{\text{doc}}^{\max}$	3	Max rules extracted per document
$R_{\text{chunk}}^{\max}$	8	Max rules extracted per text chunk
L_{chunk}	5000	Character limit per text chunk
W_{dedup}	10	Number of prior rules compared for deduplication

Table 10: Hyperparameters for our agent. All values are fixed across runs and cancer types, determined through preliminary tuning on a small development subset. These weights are interpretable and steerable: a domain expert can adjust individual values to prioritize, e.g., stronger falsification (α_{chal}) or broader exploration (α_{nov}, γ), requiring no learned policy or training data.

The framework exhibits robustness to moderate weight perturbations for two reasons. First, the adaptive schedule (Equation 7) automatically compensates for suboptimal fixed weights by modulating EIG and exploration pressure based on the current belief state. Second, prior anchoring (Equation 11) prevents early-step decisions from causing irreversible belief collapse, providing a self-correcting mechanism that absorbs weight imprecision.

E PER-CANCER RESULTS

This appendix breaks the aggregate results down by individual cancer type: a per-cancer accuracy heatmap across agents (Figure 4), the overall metric comparison with error bars (Figure 5), and the accuracy-by-difficulty-tier breakdown (Figure 6). Our agent leads on the majority of cancer types rather than winning on a few outliers, but the advantage is not uniform: on PRAD it reaches only 17%, trailing the other agents, and on LUSC and LUAD it is matched or edged out by ReAct and MCTSr respectively (Figure 4). These cases concentrate in cohorts whose discriminative signal overlaps a same-organ neighbor (prostate/kidney, the two lung subtypes), where the agent's willingness to reject costs it true positives that the affirmative baselines happen to capture.

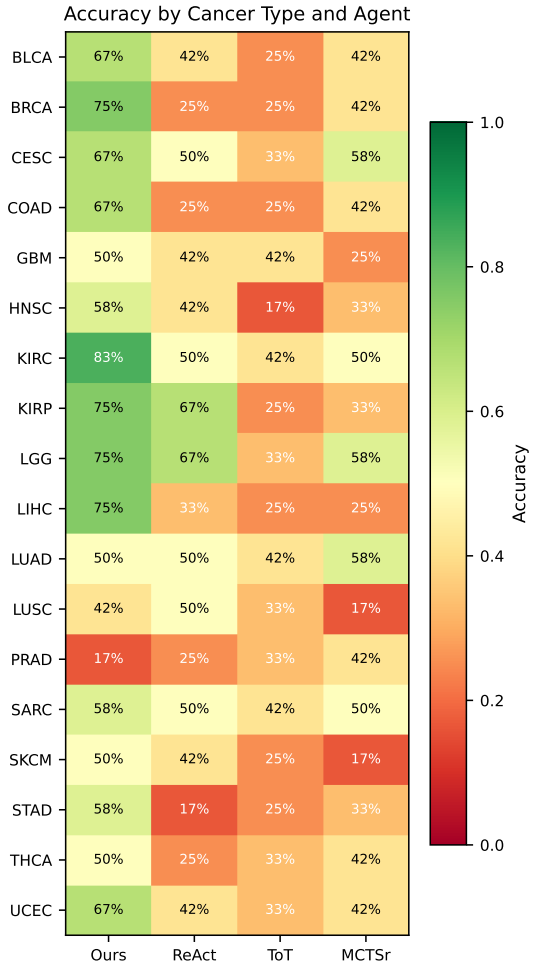


Figure 4: Per-cancer accuracy across agents. Each cell shows the fraction of correct predictions for a given cancer type across all test cases (positive + three negatives).

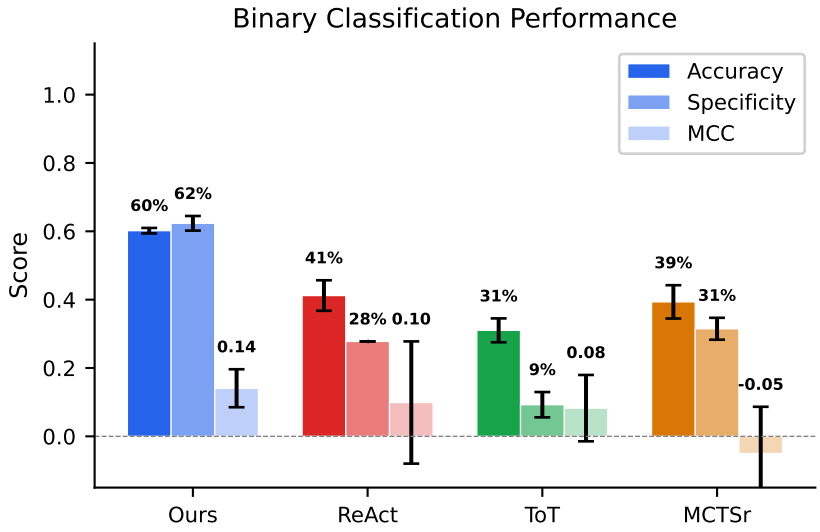


Figure 5: Overall performance comparison across agents. Accuracy, specificity, and MCC shown with error bars (±1 std over 3 repeated runs).

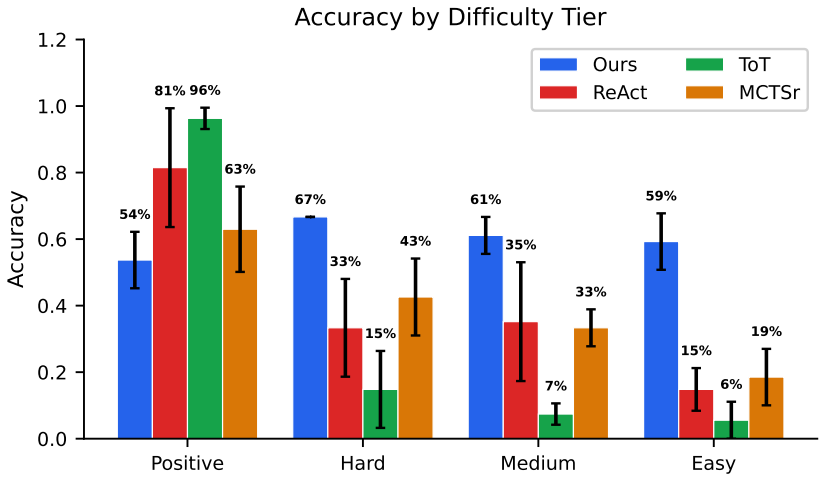


Figure 6: Accuracy by difficulty tier across agents. Our agent maintains consistent performance across all tiers while baseline agents degrade sharply on negative cases.