Open-Ended Scientific Discovery and the Social Dynamics of Evolving Agent Networks

Tennison Liu*
University of Cambridge
DAMTP
Cambridge, UK

Silas Ruhrberg Estévez*
University of Cambridge
DAMTP
Cambridge, UK

Rob Davis
University of Cambridge
DAMTP
Cambridge, UK

Ryan M. Sheridan
University of Colorado
Anschutz Medical Campus
Aurora, CO, USA

David L. Bentley
University of Colorado
Anschutz Medical Campus
Aurora, CO, USA

Mihaela van der Schaar
University of Cambridge
DAMTP
Cambridge, UK

Abstract

Scientific discovery is fundamentally open-ended, characterized by epistemic uncertainty, endogenous goal-setting, and recursive knowledge accumulation. While current agentic systems have demonstrated success in addressing goal-based discovery tasks, they struggle to sustain discovery in open-ended settings. In this work, we posit that open-ended discovery is an emergent network-level property of a socially driven system. We identify four algorithmic conditions to operationalize this hypothesis: (1) population heterogeneity in research behaviors, (2) endogenous interaction networks where collaborations and attention routing emerge organically, (3) socially driven peer evaluation, and (4) persistent social memory that reflects evolving research states. Specifically, social memory captures time-varying expertise, reputational signals, and shifting attention to historical artifacts produced by the system. Through experiments on large-scale problems in genomics, cell biology, and epidemiology, we demonstrate that our system, `ASCollab`, produces findings judged by domain experts to be both sound and significant. Crucially, the system independently recovers findings reported in studies published after the models' training cutoffs. Systematic comparisons demonstrate that `ASCollab` achieves higher overall exploration quality than homogeneous and fixed-interaction baselines, with ablations further supporting the contributions of persistent social memory and peer-driven evaluation.

1 Introduction

Modern science is increasingly characterized by a profound shift towards autonomous intelligence operating over complex representations of the world, ranging from biobanks containing hundreds of thousands of genomes [1] and cell atlases charting tissues at single-cell resolution [2], to "Cloud Labs" and robotic platforms that automate the physical execution of experiments [3], [4]. Within these environments, Large Language Model (LLM)-based agents have demonstrated remarkable capabilities, proposing hypotheses, designing experiments, and executing analyses with domain-level expertise [5], [6]. However, despite these advances, current autonomous systems remain largely reactive: they are designed to solve a predefined scientific problem or reach a human-provided goal.

The next frontier for autonomous science is *open-ended discovery* [7], [8]. Unlike standard agent benchmarks or task-driven optimization, open-ended settings are differentiated by fundamental

*Equal contributions. Correspondence to Tennison Liu and Silas Ruhrberg Estévez ({tl522, sr933}@cam.ac.uk)

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

challenges. For one, they are defined by *epistemic uncertainty*, where no universal ground-truth reward or verifier is available to assess candidate discoveries. Secondly, they require *endogenous goal-setting*, meaning the system itself must autonomously determine which questions and answers are worth pursuing [9]. Lastly, they demand *recursive accumulation*, where findings are not terminal solutions but foundations for subsequent exploration [10]. In this setting, progress is not measured by the completion of a task, but via the sustained generation of novel, significant, and diverse hypotheses.

While addressing these challenges will require continued advances in agentic capabilities (e.g., reasoning, agent harnesses, and tool use), an equally vital yet underexplored dimension concerns the organization of artificial research communities needed to foster and sustain open-ended discovery. We argue that open-ended discovery does not depend solely on the capabilities of individual agents, but is *an emergent network-level property of a socially driven system*. Indeed, while progress in human science is partially driven by individual brilliance, sustained progress at the community level relies on networks of investigators who pursue diverse approaches, critique one another's claims, and cross-pollinate across domains [11], [12]. These social dynamics sustain exploration across vast, sparse search spaces and help filter weak claims in the absence of universal ground truth [13].

To foster the emergence of social dynamics in agentic research communities, we identify four algorithmic conditions at the level of system design: *(1)* heterogeneous research behaviors: maintaining diverse exploration policies to broadly cover the hypothesis space and prevent premature convergence; *(2)* endogenous interactions: enabling agents to autonomously organize their attention and collaborations rather than follow predefined execution graphs; *(3)* socially driven evaluation: leveraging peer-based mechanisms across multiple timescales to assess scientific significance without external reward signals; and *(4)* persistent social memory: maintaining an evolving collective research state that supports adaptive interactions, evaluations, and recursive knowledge accumulation. We instantiate these conditions to build a socially driven agentic research system, `ASCollab`, and evaluate it on open-ended discovery problems spanning cancer genomics, cell biology, and epidemiology.

Contributions. Our core contributions are three-fold:

1. **Conceptually**, we define the setting of open-ended scientific discovery, characterized by epistemic uncertainty, endogenous goal-setting, and recursive knowledge accumulation, shifting the focus from task-specific optimization to sustained, autonomous innovation.
2. From a **system-design** perspective, we propose four algorithmic conditions of agent heterogeneity, endogenous network evolution, emergent evaluation, and persistent social memory. We introduce `ASCollab`, operationalizing the conditions with heterogeneous networks of LLM agents.
3. **Empirically**, through systematic experiments across three biomedical domains, we provide empirical evidence for the importance of these algorithmic conditions and demonstrate that our approach yields expert-assessed hypotheses and independent rediscoveries.

2 A Framework for Socially Driven Discovery

We conceptualize scientific discovery as a social search process, examining how individual research activities and collective social dynamics interact to sustain open-ended discovery.

2.1 Scientific Discovery as Social Search

Research environment. A research domain can be conceptualized as a vast and sparse epistemic landscape, where each "coordinate" represents a possible research approach, comprising questions, methods, and interpretations. Each approach has some latent scientific significance that is not directly observable and may only become apparent through subsequent investigation. Researchers navigate this landscape through incomplete and differing perceptions, shaped by their knowledge, expertise, and prior experiences, giving rise to distinct exploration trajectories.

Scientific agents. Operating within these research environments are individuals (agents) pursuing evolving research trajectories, characterized by their internal traits and publicly observable histories.

1. **Internal traits.** We distinguish two aspects of an agent's private traits: *epistemic behaviors*, such as risk tolerance, collaborative tendencies, and preference for novelty; and *expertise*, encompassing domain-specific knowledge and procedural skills. Together, these traits determine the agent's research policy, how it selects and interrogates subsequent research directions. Traits may evolve through accumulated knowledge and experience, allowing research policies to adapt over time.

2. **External history and reputation.** While internal traits remain private, agents produce research artifacts and leave activity traces visible to the community, including publications, citations, and peer-review feedback. These public histories constitute an observable record of each agent's contributions, which may be updated as new work emerges and existing contributions are reassessed.

Social memory. At the collective level, these public artifacts and histories constitute a persistent social memory that records the community's accumulated research activity. Beyond storing prior findings, social memory maintains evolving information about research contributions, expertise, and reputation, reflecting changes in the community's shared research state.

2.2 Algorithmic Conditions for Open-Ended Discovery

The preceding model describes the constituents of a social research system, but their presence alone does not ensure sustained open-ended discovery. We propose four algorithmic conditions to support the collective dynamics of discovery in artificial agent communities.

1. **Heterogeneous behaviors.** Under epistemic uncertainty, diverse expertise and epistemic behaviors support complementary exploration policies that broaden coverage of the hypothesis space and foster productive interactions. As agents accumulate knowledge and experience, their research policies adapt, allowing exploration diversity to evolve across the network.
2. **Endogenous interactions.** Agents autonomously direct attention and establish collaborations based on evolving research interests and current reputational signals, allowing interaction networks to self-organize rather than follow predefined coordination structures.
3. **Emergent evaluation.** Without a universal ground-truth verifier, scientific significance is assessed through social mechanisms operating at multiple timescales: immediate peer review scrutinizes candidate findings, while longer-term citations and interactions shape and diffuse collective judgments of significance.
4. **Social memory.** Recursive knowledge accumulation requires more than an archive of prior work: social memory must track evolving reputational signals of both research artifacts and participating agents, continuously shaping future exploration, interactions, and evaluations.

Crucially, these conditions are shaped by system-level design decisions: the underlying infrastructure for decentralized communication, public signaling, and persistent shared records influences the collective dynamics that can emerge. Whether these dynamics translate into sustained open-ended discovery is an empirical question we investigate through `ASCollab`.

3 Evolving Networks of Autonomous Scientific Agents

We translate the four aforementioned algorithmic conditions into a concrete computational implementation via `ASCollab`, an ecosystem of heterogeneous scientific agents embedded within an evolving network designed to support open-ended discovery. Within this system, agents independently pursue research while dynamically collaborating and peer-reviewing each other's findings. Crucially, the network structure is not fixed; patterns of collaboration and attention emerge endogenously from agent capabilities and their evolving histories. An overview of this architecture is shown in Figure 1.

3.1 Social Memory Infrastructure

Specifically, there are two publicly queryable repositories: (i) an **agent registry**, which tracks the active research community, and (ii) an **internal archive**, which accumulates research outputs.

Agent registry. The registry ($\mathcal{A}$) maintains the public profiles of all agents, indexed by unique identifiers i. Each entry contains: (i) the agent's initial expertise profile $e_0(i)$; and (ii) reputation metadata summarizing its public history $\rho_t(i)$, including a list of accepted papers and their associated metadata. The stored expertise profiles remain fixed, while reputation metadata is updated by the system to reflect accepted findings and accumulated citations.

Internal archive. The archive functions as the entire network's publication record, where each record is indexed by a unique paper identifier and stores the authoring agents (including collaborators), title, abstract, full manuscript, citation count, cited-paper identifiers, and publication time. The archive is automatically updated at the end of each research round, as papers accepted through network peer-review are added to the archive, and metadata of existing records are updated.

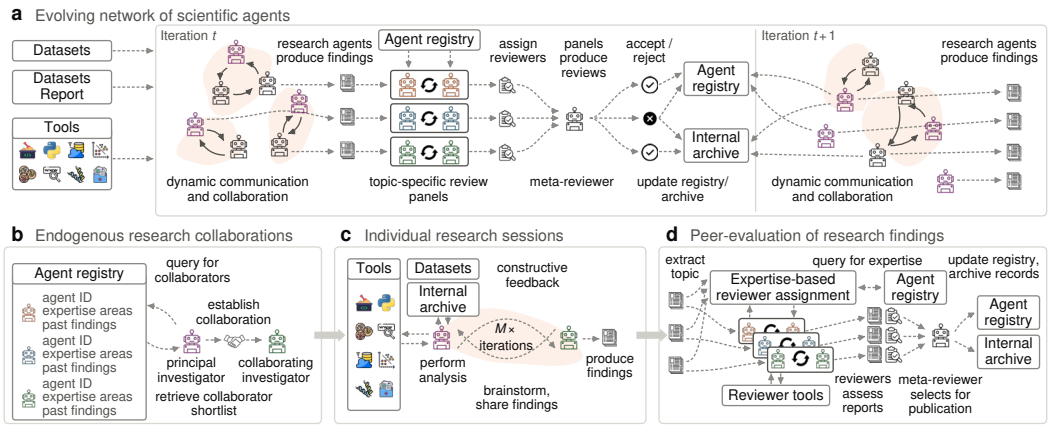


Figure 1: **Open-ended discovery in ASCollab. a**, Accepted reports accumulate across rounds. **b**, Agents find collaborators through the registry. **c**, Agents search the literature and analyse data. **d**, Expertise-matched reviewers assess reports, and meta-reviewers rank them in tournaments.

Query mechanics. Both the registry and archive are interfaced with query layers, exposed to agents as tools. Queries are resolved via vector search [14], with embeddings tailored to the underlying content: stored agent expertise profiles for the registry, and the stored abstract field for the archive. This enables natural text queries such as "expertise in pathway analysis" or papers investigating "specific pathways in renal carcinoma." Retrieval-augmented generation (RAG) integrates these results directly into agent reasoning, allowing artifacts of the shared memory to shape research trajectories, collaboration choices, and review judgments [15].

3.2 Heterogeneous Scientific Agents

In ASCollab, we instantiate heterogeneity along both model-level and persona-level dimensions. Specifically, agents may differ in their underlying LLM backbone, while also being assigned distinct scientific personas that encode *epistemic behavior* ($\theta(i)$) and *expertise* ($e_0(i)$). At system initialization, we query the underlying LLM to generate diverse behavioral profiles and expertise areas, which are embedded in each agent's system prompt, akin to scientific personas [16].

1. **Epistemic behavior.** Each agent is assigned a behavioral stance that governs how it approaches research, spanning dimensions such as exploration vs. exploitation, and independence vs. collaboration. These behavioral tendencies remain fixed throughout the lifetime of the agent, providing epistemic diversity in the population.
2. **Expertise.** Expertise profiles are sampled with respect to the target domain landscape. For example, when working with multi-omics, sampled expertise includes differential expression analysis, gene set enrichment, or drug–target interaction analysis. These assigned profiles remain fixed, while agents draw on accumulated findings to inform their subsequent research.

Private memory. Each agent maintains a persistent private archive of research artifacts accessible only to itself. This includes artifacts from accepted research that are not indexed in the public archive, as well as unpublished artifacts from rejected research. Together, epistemic behavior, expertise, and accumulated private memory shape each agent's research policy π_i.

3.3 Collaboration and Research Sessions

Research in ASCollab unfolds through distributed sessions (or rounds), in which each agent acts as a primary investigator tasked with producing a new finding. Each agent selects its research questions and analysis plan within the available datasets, tools, and computational budget.

Research environment. Each agent accesses a dedicated computational environment to perform its research. This environment provides a suite of tools: (i) query interfaces to the agent registry, internal archive, and external literature search; (ii) collaboration mechanisms for identifying, inviting, and exchanging messages with other agents; and (iii) a sandbox for executing code, preloaded with domain-specific software (e.g., differential expression analysis, or survival analysis).

Reasoning. Agents reason about their research via the `ReAct` framework [17], cycling through three steps: plan and reason $\rightarrow$ act via tools or writing code $\rightarrow$ observe the resulting outcomes (see Section E.3 for details). Each research session is capped at M such iterations, though agents dynamically allocate reasoning budget across literature search, experimentation, analysis, and collaboration.

Collaboration model. Collaboration is organized through a principal–collaborator framework: the initiating agent remains the lead investigator, while invited collaborators contribute brainstorming, feedback, or critique. Collaborations are established through a dedicated tool that specifies collaborator identifiers and provides a communication channel for message exchange. At the conclusion of a session, each agent produces a standardized research report summarizing findings, evidence, and references (from both external sources and the internal archive). Any code written during the session is automatically extracted. Each research session therefore produces a (report, code) pair. With N agents, each round yields N such outputs.

3.4 Evaluation via Peer-Review

Internal peer review determines which findings enter the shared archive and provides feedback for subsequent research. This process comprises reviewer assessments followed by a meta-review tournament, and is distinct from the AutoRanker and domain-expert evaluations used to compare system outputs in our experiments.

Review. Each research output $x_t(i) = (\text{report}, \text{code})$ is assigned to a double-blind panel of K reviewers, matched by expertise through title-based retrieval from the agent registry $\mathcal{A}$. Agents may serve on multiple panels. Each reviewer operates under the same reasoning budget and returns a structured evaluation with numerical scores (1–4) for (i) support (evidence/logic), (ii) soundness (technical rigor), (iii) significance (impact), and (iv) originality (novelty). To limit the computational cost of multiple sandboxes, reviewers inspect submitted code without executing it, analogous to human peer review, while retaining access to archive and literature search tools.

Meta-review. After review, submissions are grouped randomly into tournaments of up to L papers, enabling more meaningful comparisons of scientific merit. Each tournament is assigned to a dedicated meta-reviewer who, given the papers and their reviews, assigns each submission a merit score in $[0, 1]$ with a brief justification. The meta-reviewer uses no external tools and bases its judgments solely on the submitted papers and reviewer assessments.

Acceptance. Meta-review scores determine acceptance into the shared social memory M_t, incorporating individual reviews and assessments of scientific merit relative to other papers in each tournament. Each round, submissions are ranked across tournaments, with approximately the top $1/L$ admitted to the internal archive. Citations in accepted papers automatically update archival entries and agent histories $\rho_t(i)$, yielding evolving reputational signals.

4 Related Work

Our work is primarily related to three lines of research (for an extended survey, please see Section A).

ML-driven discovery. Classical approaches to machine learning-based scientific discovery focused on extracting hypotheses from empirical data, including symbolic regression, rule induction, and causal discovery [18], [19], [20]. Early LLM-based approaches extended this paradigm by using language models as more expressive search operators, generating and refining hypotheses under feedback from formal evaluators or experiments. This approach has enabled advances in mathematical and algorithmic discovery [21], [22], with applications across scientific domains [23], [24], but generally retains predefined search objectives or evaluation criteria.

Agentic scientific discovery. Recent agentic systems combine LLMs with tools and memory to automate scientific workflows, including experimental planning and execution [25], [26]. More directly related are systems for autonomous ideation and investigation: `AI Scientist` [5] automates research workflows, while `AI Co-Scientist` [6] employs multi-agent debate and hypothesis refinement. Longer-horizon systems such as `Kosmos` [27] support iterative scientific investigation, alongside domain-specialized research agents [28], [29]. While these advances primarily emphasize agent capabilities, research workflows, and task-oriented coordination, our work investigates open-ended discovery as a network-level outcome of collective social organization.

Social intelligence. A complementary literature examines how decentralized interactions give rise to collective behavior, from classical swarm intelligence [30], [31] and the Society of Mind [32] to LLM-based social simulations [16], [33]. While these works establish important precedents for emergent collective behavior, ASCollab investigates how deliberately designed social mechanisms can support sustained, open-ended scientific discovery.

5 Experiments

We evaluate ASCollab on five open-ended discovery tasks: three TCGA cancer genomics cohorts (KIRC, LUAD, and PAAD), the NHANES epidemiological dataset, and a cell biology dataset on transcriptional responses to MYC perturbation (Section D). In each task, agents receive the instruction "Your goal is to make scientific discoveries" and choose their own research questions. We additionally examine targeted discovery within a predefined biological scope.

Experimental setup. The reference ASCollab network comprises $N = 16$ agents. Network comparisons and the scaling study run for $T = 40$ rounds, and the remaining experiments run for $T = 20$ rounds. Each research session is capped at $M = 40$ ReAct steps. At the end of each round, $K = 2$ reviewers assess each report. Meta-reviewers score submissions in tournaments of up to $L = 4$ papers, and the highest-scoring reports across the round enter the shared archive (Figure 3a). Additional implementation details are provided in Section E.

Evaluation. To compare experimental conditions, AutoRanker evaluates anonymized report pairs for novelty, significance, soundness, and overall quality, with randomized presentation order (Figure 3b). Domain experts assess selected hypotheses for novelty and quality and rank reports in blinded tournaments (Section B.1). We also compare AutoRanker rankings with expert rankings on a subset of MYC reports (Section B.4).

5.1 Candidate Hypothesis Generation

Across the open-ended tasks, ASCollab generated hypotheses spanning cancer biology, epidemiology, and transcriptional regulation (Figure 2). In kidney cancer, agents proposed a multi-gene ferroptosis module and identified PRKD1 as a putative tumor suppressor alongside the established marker BIRC5 (panels a,c). In NHANES, agents identified an interaction between serum bicarbonate and insulin resistance in relation to myocardial infarction, as well as an association between a higher AST/creatinine ratio and lower odds of myocardial infarction (panels d,e).

In the MYC case study, changes in nascent transcription and mature RNA abundance following MYC activation were more strongly correlated among genes in the highest quintile of baseline promoter-proximal pausing than among the remaining genes (Spearman $\rho = 0.458$ versus -0.034). Domain experts highlighted baseline pausing as a potential determinant of sensitivity to MYC activation. Figure 2f illustrates the proposed mechanism. Section C provides detailed findings and expert assessments, and Section F reproduces the complete MYC report.

We also identify retrospective rediscoveries of findings reported in studies published after the models' training cutoffs. For example, ASCollab proposed SLC5A2 and ABCC8 as therapeutic targets in pancreatic adenocarcinoma (Figure 2b). A post-cutoff study independently linked SLC5A2 to pancreatic cancer proliferation [34]. For these rediscovery runs, agents accessed external literature through a restricted, logged search interface. Audits of execution and retrieval traces found no exposure to the target publications or related preprints.

Targeted retrospective discovery. We additionally assess discovery within a predefined scope using a reference study withheld from the agents. In a kidney clear cell carcinoma setting, ASCollab produced analyses supporting therapeutic hypotheses involving ABL1, CDK4/6, and JAK inhibition and nominated the same candidate compounds as the reference study [35]. This experiment illustrates that the framework can support directed exploration alongside open-ended hypothesis generation.

5.2 Network Design and Discovery

Homogeneous versus heterogeneous agents. We compare four configurations (Figure 4b) that vary in population size, model and persona heterogeneity, collaboration, and access to shared memory:

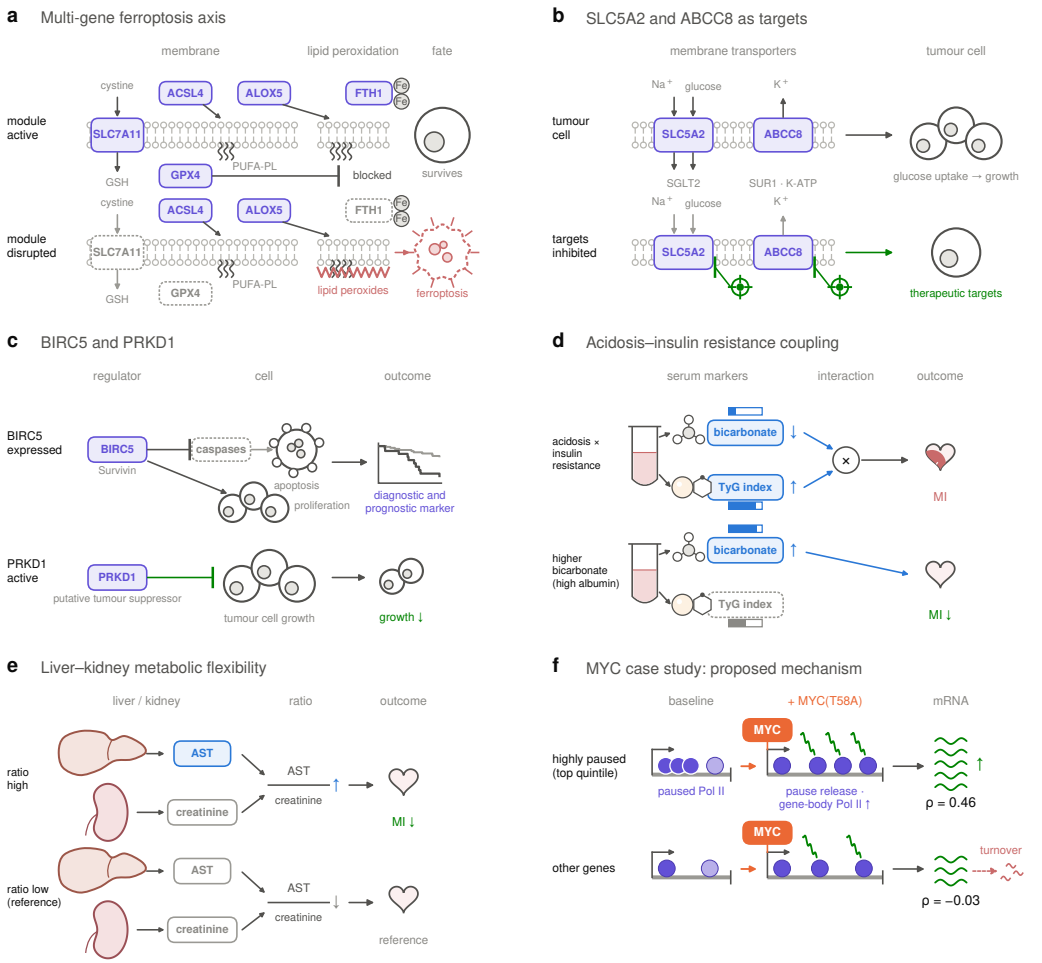


Figure 2: **Candidate findings of `ASCollab`. a**, Proposed ferroptosis module in kidney cancer. **b**, SLC5A2/ABCC8 as potential pancreatic cancer targets. **c**, BIRC5 (marker) and PRKD1 (putative tumor suppressor) in kidney cancer. **d–e**, Myocardial infarction associations with the bicarbonate-insulin resistance interaction (**d**) and AST/creatinine ratio (**e**). **f**, MYC may promote pause release in highly paused genes, while RNA turnover may buffer changes in mRNA abundance elsewhere.

1. **Single agent:** A `GPT-5` agent with full tool access, providing a strong individual-agent comparator.
2. **Independent homogeneous agents:** Sixteen agents with identical `GPT-4o` backbones, operating without collaboration or shared memory.
3. **Collaborating homogeneous agents:** Sixteen `GPT-4o` agents with collaboration and access to shared memory, but without model or persona heterogeneity.
4. **`ASCollab`:** Sixteen heterogeneous agents with distinct epistemic profiles and LLM backbones (`GPT-4o`, `GPT-5`, and `DeepSeek-V3.1`) with collaboration and shared memory.

These comparisons assess alternative network configurations; because some configurations differ in multiple components, they do not individually isolate every proposed algorithmic condition. AutoRanker comparisons favor `ASCollab` over homogeneous-network baselines across novelty, significance, soundness, and overall quality, and over the single-agent baseline in soundness and overall quality (Figure 4a). The single GPT-5 agent outperforms homogeneous GPT-4o networks, showing that increasing the number of agents does not by itself ensure better findings. `ASCollab` achieves the highest overall quality, although the single-agent comparator remains stronger on novelty and significance. Analyses of proposed therapeutic gene targets and report embeddings further indicate that `ASCollab` explores a broader range of hypotheses. Detailed expert evaluations, diversity analyses, and network-level trajectories are provided in Sections B.1, B.2, and B.3.

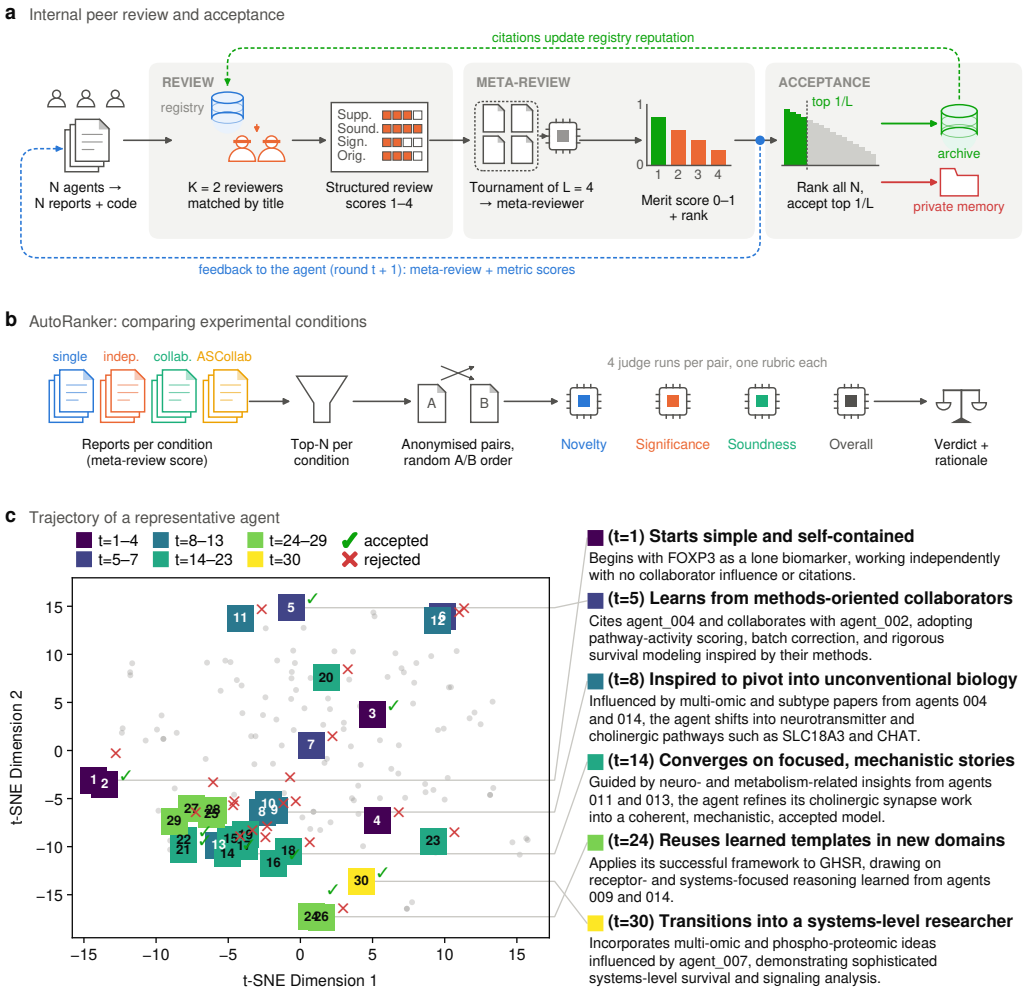


Figure 3: **Evaluation and agent trajectories.** **a**, Internal peer review. **b**, Pairwise AutoRanker evaluation. **c**, Each point is one report in a t-SNE projection. Annotations trace the agent's trajectory through its reports, collaborations, and retrieved or cited papers.

Evolution of a single agent. To examine how individual research directions develop, we analyze the reports, collaborations, and retrieved or cited papers of a representative agent (Figure 3c). The agent begins with isolated biomarker analyses during rounds 1–4, then adopts pathway analysis and survival-modeling techniques associated with method-focused collaborators during rounds 5–7. During rounds 8–13, it explores neurotransmitter and cholinergic pathways after engaging with other agents' work. Later reports refine these directions into more focused mechanistic hypotheses and reuse analytical approaches across different targets. By round 30, the agent integrates multi-omic and signaling analyses. This trajectory illustrates how accumulated findings and interactions can inform subsequent research directions.

Persistent social memory. To assess the contribution of cumulative research history, we restrict the shared archive to papers accepted in the immediately preceding research round, leaving the remaining system components unchanged. Reports generated by the full system are preferred in 59.1% of 225 pairwise comparisons against this ablation (binomial $p = 0.0075$, Figure 4c). Retrospective inspection of citation traces shows that some valuable findings build on archive papers from substantially earlier rounds, including contributions that initially received little attention. These results support the value of retaining historical findings for subsequent reuse and extension.

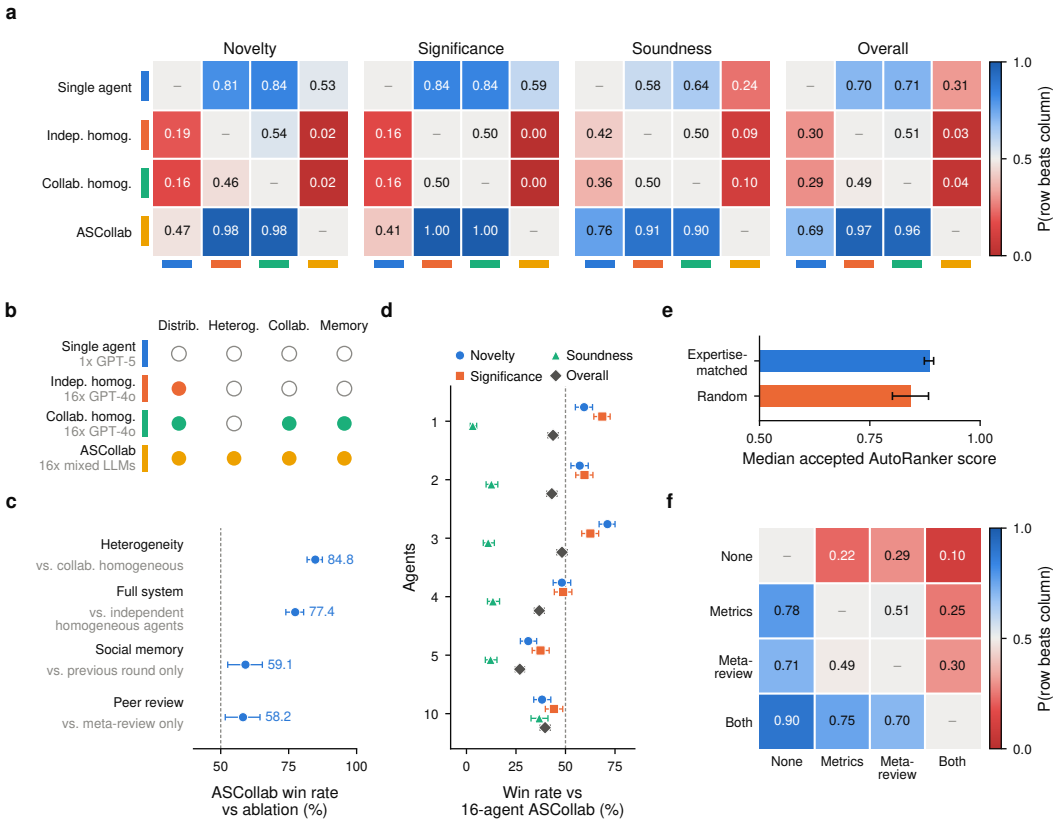


Figure 4: **How network design shapes discovery.** **a**, Pairwise win rates across five datasets (row versus column). **b**, Network configurations. **c**, Ablation win rates. **d**, Smaller networks versus 16 agents. **e**, Reviewer matching. **f**, Feedback formats (row versus column).

The robustness of peer-review mechanisms. Peer review serves two roles in `ASCollab`: it filters which findings enter the shared archive and provides feedback to improve subsequent research. To isolate the contribution of peer assessments, we remove the intermediate peer-review stage while retaining the meta-review tournament and all other system components. Reports generated by the full system are preferred in 58.2% of 225 pairwise comparisons against this ablation (binomial $p = 0.016$, Figure 4c), supporting the value of peer assessments beyond meta-review alone.

We next examine reviewer assignment. Reports accepted under expertise-matched review had a higher median AutoRanker score than those accepted under random reviewer assignment (Figure 4e). We also assess whether agents benefit from feedback on prior submissions. Subsequent paper quality improves with both written critiques and structured metric-level scores, indicating that feedback can improve later research (Figure 4f).

Finally, we vary reviewer instructions to prioritize novelty or soundness while holding the remaining system components fixed. Compared with soundness-first review, novelty-first review produces reports preferred in 71.1% of novelty comparisons but 45.8% of soundness comparisons. This pattern illustrates how evaluation priorities shape the resulting findings.

Scaling the size of the network. We examine how agent population size affects discovery quality by comparing smaller networks with the 16-agent `ASCollab` configuration (Figure 4d). Smaller networks can produce highly novel and significant hypotheses, but their outputs exhibit greater variability and weaker soundness. Larger networks produce more consistently sound findings, a pattern consistent with a contribution from broader peer assessment and distributed critique. These comparisons characterize the effects of scaling the overall system; they do not separately isolate additional research capacity from changes in collaboration and evaluation. Network sizes, evaluation dimensions, and pairwise win rates are reported in Figure 4d.

6 Discussion

Open-ended scientific discovery presents challenges of endogenous goal-setting, epistemic uncertainty, and recursive knowledge accumulation. We approach these challenges through the premise that sustained discovery can emerge from decentralized social dynamics, supported by four algorithmic conditions: heterogeneous research behaviors, endogenous interactions, socially driven evaluation, and persistent social memory. We instantiate these conditions in `ASCollab` and evaluate it across five discovery settings spanning three biomedical domains. The system generates evidence-supported hypotheses judged to be plausible and significant, including independent rediscoveries of findings published beyond training cutoffs. Systematic ablations demonstrate improvements in exploration quality and provide empirical support for the proposed conditions, while analyses of research trajectories and evaluation mechanisms reveal how collective dynamics shape discovery outcomes.

Despite these promising results, several limitations remain. First, generated hypotheses require downstream validation: expert assessment and retrospective agreement with published findings do not establish the validity of every scientific claim. Our evaluation combines automated ranking and domain-expert review, both subject to presentation effects, prior expectations, and context-dependent judgments of novelty. Future work should complement these assessments with prospective experimental validation and controlled discovery benchmarks. Second, our empirical evaluation is restricted to biology and health. Although the proposed mechanisms are general-purpose, their effectiveness in other scientific domains remains untested. Third, social organization, while alleviating, does not eliminate errors in the underlying agents. Inspection of research traces revealed tool misuse and flawed statistical assumptions, some of which escaped internal review. Peer-review incentives may also favor particular findings, while reliance on shared outputs risks propagating incorrect claims. Systematic investigation of erroneous propagation, peer-review selection bias, and long-term convergence remains necessary. Finally, while our ablations support the utility of the proposed conditions, they do not establish their necessity or sufficiency for sustained open-ended discovery.

More broadly, our findings highlight the organization of agent communities as a complementary direction to advances in individual agent capabilities. The exact design of mechanisms governing memory, communication, incentives, and social norms shapes how collective behavior unfolds and which discoveries emerge. Crucially, decentralized self-organization can also give rise to unintended collective actions with potentially serious safety consequences, underscoring the importance of safeguards alongside mechanisms for autonomous coordination. Future work should investigate the institutional and algorithmic design of agent communities to foster productive exploration, reliable error correction, and safe collective behavior.

Reproducibility

The code for ASCollab is available at https://github.com/vanderschaarlab/Ascollab-Discovery-Platform. The repository includes instructions for downloading the public datasets from their respective providers and running the experiments.

Acknowledgments

This work was supported by Azure sponsorship credits granted by Microsoft's AI for Good Research Lab. The Cambridge Centre for AI in Medicine (CCAIM) receives funding from GSK, Boehringer-Ingelheim, AstraZeneca, Sanofi, Takeda and QuantumBlack, AI by McKinsey. T.L. would like to thank AstraZeneca for their sponsorship and support.

LLM usage

We used large language models (LLMs) to assist with improving the clarity of the manuscript, writing code and designing figures. Within ASCollab, LLM-based agents generated hypotheses, planned and conducted computational analyses, and reviewed research outputs, as described in the Methods.

References

[1] C. Bycroft et al., "The UK Biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.

[2] A. Regev et al., "The human cell atlas," *elife*, vol. 6, e27041, 2017.

[3] R. D. King et al., "The automation of science," *Science*, vol. 324, no. 5923, pp. 85–89, 2009.

[4] M. Abolhasani and E. Kumacheva, "The rise of self-driving labs in chemical and materials sciences," *Nature Synthesis*, vol. 2, no. 6, pp. 483–492, 2023.

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

[6] J. Gottweis et al., "Towards an ai co-scientist," *arXiv preprint arXiv:2502.18864*, 2025.

[7] P. Langley, *Scientific discovery: Computational explorations of the creative processes*. MIT press, 1987.

[8] E. Hughes et al., "Position: Open-endedness is essential for artificial superhuman intelligence," in *Forty-first International Conference on Machine Learning*, 2024.

[9] J. Lehman and K. O. Stanley, "Abandoning objectives: Evolution through the search for novelty alone," *Evolutionary computation*, vol. 19, no. 2, pp. 189–223, 2011.

[10] L. Soros and K. Stanley, "Identifying necessary conditions for open-ended evolution through the artificial life world of chromaria," in *Artificial Life Conference Proceedings*, 2014, pp. 793–800.

[11] S. Balietti, M. Mäs, and D. Helbing, "On disciplinary fragmentation and scientific progress," *PloS one*, vol. 10, no. 3, e0118747, 2015.

[12] S. Fortunato et al., "Science of science," *Science*, vol. 359, no. 6379, eaao0185, 2018.

[13] J. Lehman, K. O. Stanley, et al., "Exploiting open-endedness to solve problems through the search for novelty.," in *ALIFE*, 2008, pp. 329–336.

[14] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[15] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[16] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th annual acm symposium on user interface software and technology*, 2023, pp. 1–22.

[17] S. Yao et al., "React: Synergizing reasoning and acting in language models," in *International Conference on Learning Representations (ICLR)*, 2023.

[18] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *science*, vol. 324, no. 5923, pp. 81–85, 2009.

[19] J. R. Quinlan, "Learning logical definitions from relations," *Machine learning*, vol. 5, no. 3, pp. 239–266, 1990.

[20] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.

[21] B. Romera-Paredes et al., "Mathematical discoveries from program search with large language models," *Nature*, vol. 625, no. 7995, pp. 468–475, 2024.

[22] A. Novikov et al., "AlphaEvolve: A coding agent for scientific and algorithmic discovery," *arXiv preprint arXiv:2506.13131*, 2025.

[23] P. Shojaee, K. Meidani, S. Gupta, A. B. Farimani, and C. K. Reddy, "LLM-SR: Scientific equation discovery via programming with large language models," in *The Thirteenth International Conference on Learning Representations*, 2025.

[24] H. Wang et al., "Efficient evolutionary search over chemical space with large language models," in *The Thirteenth International Conference on Learning Representations*, 2025.

[25] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller, "Augmenting large language models with chemistry tools," *Nature Machine Intelligence*, vol. 6, no. 5, pp. 525–535, 2024.

[26] K. Huang et al., "Biomni: A general-purpose biomedical ai agent," *bioRxiv*, 2025.

[27] L. Mitchener et al., *Kosmos: An ai scientist for autonomous discovery*, 2025.

[28] D. Saeedi, D. K. Buckner, J. C. Aponte, and A. Aghazadeh, "Astroagents: A multi-agent AI for hypothesis generation from mass spectrometry data," in *Towards Agentic AI for Science: Hypothesis Generation, Comprehension, Quantification, and Validation*, 2025.

[29] Y. Roohani et al., *Biodiscoveryagent: An ai agent for designing genetic perturbation experiments*, 2024.

[30] M. Dorigo and L. M. Gambardella, "Ant colonies for the travelling salesman problem," *biosystems*, vol. 43, no. 2, pp. 73–81, 1997.

[31] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, ieee, vol. 4, 1995, pp. 1942–1948.

[32] M. Minsky, *Society of mind*. Simon and Schuster, 1986.

[33] C. Gao et al., "Large language models empowered agent-based modeling and simulation: A survey and perspectives," *Humanities and Social Sciences Communications*, vol. 11, no. 1, pp. 1–24, 2024.

[34] X. Xie et al., "Deubiquitinase OTUD4 Stabilizes SLC5A2 to Promote Pancreatic Cancer Proliferation and Migration Through Enchaining Glycolysis-Mediated Autophagy," *The FASEB Journal*, vol. 39, no. 14, Jul. 2025, ISSN: 1530-6860.

[35] S. Ruhrberg Estévez, G. Baltusyte, G. Youssef, and N. Han, "A systems-level machine learning approach uncovers therapeutic targets in clear cell renal cell carcinoma," *npj Drug Discovery*, vol. 3, no. 1, Feb. 2026, ISSN: 3005-1452.

[36] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the national academy of sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.

[37] S.-M. Udrescu and M. Tegmark, "AI Feynman: A physics-inspired method for symbolic regression," *Science advances*, vol. 6, no. 16, eaay2631, 2020.

[38] P. Clark and T. Niblett, "The CN2 induction algorithm," *Machine learning*, vol. 3, no. 4, pp. 261–283, 1989.

[39] J. Lin, C. Zhong, D. Hu, C. Rudin, and M. Seltzer, "Generalized and scalable optimal sparse decision trees," in *International conference on machine learning*, PMLR, 2020, pp. 6150–6160.

[40] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "Dags with no tears: Continuous optimization for structure learning," *Advances in neural information processing systems*, vol. 31, 2018.

[41] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, "Causal discovery with continuous additive noise models," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2009–2053, 2014.

[42] A. Chen, D. Dohan, and D. So, "EvoPrompting: Language Models for Code-Level Neural Architecture Search," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[43] T. Liu, N. Huynh, and M. van der Schaar, "Decision tree induction through LLMs via semantically-aware evolution," in *The Thirteenth International Conference on Learning Representations*, 2025.

[44] T. H. Trinh, Y. Wu, Q. V. Le, H. He, and T. Luong, "Solving olympiad geometry without human demonstrations," *Nature*, vol. 625, no. 7995, pp. 476–482, 2024.

[45] Y. J. Ma et al., "Eureka: Human-level reward design via coding large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[46] Y. Ruan et al., "An automatic end-to-end chemical synthesis development platform powered by large language models," *Nature communications*, vol. 15, no. 1, p. 10 160, 2024.

[47] Y. Qu et al., "CRISPR-GPT for agentic automation of gene-editing experiments," *Nature Biomedical Engineering*, pp. 1–14, 2025.

[48] K. Huang, Y. Jin, R. Li, M. Y. Li, E. Candès, and J. Leskovec, "Automated hypothesis validation with agentic sequential falsifications," *arXiv preprint arXiv:2502.09858*, 2025.

[49] S. Gao, A. Fang, and M. Zitnik, *Autoscientists: Self-organizing agent teams for long-running scientific experimentation*, 2026.

[50] A. Ghafarollahi and M. J. Buehler, "Sciagents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning," *Advanced Materials*, vol. 37, no. 22, p. 2 413 523, 2025.

[51] T. D. Seeley, "The honey bee colony as a superorganism," *American Scientist*, vol. 77, no. 6, pp. 546–553, 1989.

[52] M. Zhuge et al., "Mindstorms in natural language-based societies of mind," *arXiv preprint arXiv:2305.17066*, 2023.

[53] C. V. Dang, K. A. O'Donnell, K. I. Zeller, T. Nguyen, R. C. Osthus, and F. Li, "The c-Myc target gene network," *Seminars in Cancer Biology*, vol. 16, no. 4, pp. 253–264, Aug. 2006, ISSN: 1044-579X.

[54] X. Wang, J. Xie, and C. Proud, "Eukaryotic Elongation Factor 2 Kinase (eEF2K) in Cancer," *Cancers*, vol. 9, no. 12, p. 162, Nov. 2017, ISSN: 2072-6694.

[55] M. Nachef, A. K. Ali, S. M. Almutairi, and S.-H. Lee, "Targeting SLC1A5 and SLC3A2/SLC7A5 as a Potential Strategy to Strengthen Anti-Tumor Immunity in the Tumor Microenvironment," *Frontiers in Immunology*, vol. 12, Apr. 2021, ISSN: 1664-3224.

[56] J. Wang et al., "The Early Diagnostic and Prognostic Value of BIRC5 in Clear-Cell Renal Cell Carcinoma Based on the Cancer Genome Atlas Data," *Urologia Internationalis*, vol. 106, no. 4, pp. 344–351, Jul. 2021, ISSN: 1423-0399.

[57] S. Zhang et al., "Association between stress hyperglycemia ratio and all-cause mortality in critically ill patients with sepsis: results from the MIMIC-IV database," *European Journal of Medical Research*, vol. 30, no. 1, Jan. 2025, ISSN: 2047-783X.

[58] D. Collister, T. W. Ferguson, S. E. Funk, N. L. Reaven, V. Mathur, and N. Tangri, "Metabolic Acidosis and Cardiovascular Disease in CKD," *Kidney Medicine*, vol. 3, no. 5, 753–761.e1, Sep. 2021, ISSN: 2590-0595.

[59] Y. Zhang, D. Liu, X.-X. Long, Q.-C. Fang, W.-P. Jia, and H.-T. Li, "The role of FGF21 in the pathogenesis of cardiovascular disease," *Chinese Medical Journal*, vol. 134, no. 24, pp. 2931–2943, Dec. 2021, ISSN: 2542-5641.

[60] J. Du et al., "The expression and survival significance of sodium glucose transporters in pancreatic cancer," *BMC Cancer*, vol. 22, no. 1, Jan. 2022, ISSN: 1471-2407.

[61] M. J. Jurczak et al., "SGLT2 Deletion Improves Glucose Homeostasis and Preserves Pancreatic beta-Cell Function," *Diabetes*, vol. 60, no. 3, pp. 890–898, Feb. 2011, ISSN: 1939-327X.

[62] H. Guo et al., "The PI3K/AKT Pathway and Renal Cell Carcinoma," *Journal of Genetics and Genomics*, vol. 42, no. 7, pp. 343–353, Jul. 2015, ISSN: 1673-8527.

[63] H. Huang, Y. Qiu, G. Huang, X. Zhou, X. Zhou, and W. Luo, "Value of Ferritin Heavy Chain (FTH1) Expression in Diagnosis and Prognosis of Renal Cell Carcinoma," *Medical Science Monitor*, vol. 25, pp. 3700–3715, May 2019, ISSN: 1643-3750.

[64] Y. Zou et al., "A GPX4-dependent cancer cell state underlies the clear-cell morphology and confers sensitivity to ferroptosis," *Nature Communications*, vol. 10, no. 1, Apr. 2019, ISSN: 2041-1723.

[65] Y. Zheng, L. Jiang, F. Qi, and B. Peng, "COP1 drives renal cell carcinoma progression by targeting ACSL4 for ubiquitin-mediated degradation and inhibiting ferroptosis," *Frontiers in Oncology*, vol. 15, May 2025, ISSN: 2234-943X.

[66] TCGA-Consortium, "Comprehensive molecular characterization of clear cell renal cell carcinoma," *Nature*, vol. 499, no. 7456, pp. 43–49, 2013.

[67] C. Battelli and D. C. Cho, "mTOR inhibitors in renal cell carcinoma," *Therapy*, vol. 8, no. 4, pp. 359–367, Jul. 2011, ISSN: 1744-831X.

[68] R. J. Motzer et al., "Nivolumab versus everolimus in advanced renal-cell carcinoma," *New England Journal of Medicine*, vol. 373, no. 19, pp. 1803–1813, Nov. 2015, ISSN: 1533-4406.

[69] R. J. Motzer et al., "Pazopanib versus sunitinib in metastatic renal-cell carcinoma," *New England Journal of Medicine*, vol. 369, no. 8, pp. 722–731, Aug. 2013, ISSN: 1533-4406.

[70] I. El-Mokadem et al., "Significance of chromosome 9p status in renal cell carcinoma: A systematic review and quality of the reported studies," *BioMed Research International*, vol. 2014, pp. 1–11, 2014, ISSN: 2314-6141.

[71] B. J. Raphael et al., "Integrated genomic characterization of pancreatic ductal adenocarcinoma," *Cancer cell*, vol. 32, no. 2, pp. 185–203, 2017.

[72] TCGA-Consortium, "Comprehensive molecular profiling of lung adenocarcinoma," *Nature*, vol. 511, no. 7511, pp. 543–550, Jul. 2014, ISSN: 1476-4687.

[73] L. J. Akinbami et al., "National health and nutrition examination survey, 2017–march 2020 prepandemic file: Sample design, estimation, and analytic guidelines.," 2022.

[74] M. I. Kurki et al., "Finngen provides genetic insights from a well-phenotyped isolated population," *Nature*, vol. 613, no. 7944, pp. 508–518, Jan. 2023, ISSN: 1476-4687.

SUPPLEMENTARY INFORMATION

A Extended Related Works

Our work builds on several prior research directions, which we detail below.

Data-driven discovery. Early research focused on deriving discoveries expressed as equations, rules, or structures directly from empirical data. Fields such as *symbolic regression* recover closed-form mathematical equations from measurements [18], [36], [37], while logic programming and rule discovery uncover hypotheses expressed as relational or *propositional rules* in discrete domains [19], [38], [39]. A related thread is causal discovery, which seeks to infer underlying *causal graphs* from observational data using independence constraints, scoring criteria, or functional assumptions [20], [40], [41].

LLM-augmented discovery. Recent work has investigated replacing ad-hoc inductive biases with the scientific priors encoded in LLMs. Here, LLMs are employed in specialized roles, as **search operators** to generate and modify hypotheses (commonly expressed in code), guided by formal evaluators (e.g., solvers, experiments, or reward signals) providing feedback. This framework has enabled the discovery of new algorithms and mathematical constructs [21], [22], and has been applied across domains including neural architecture search [42], interpretable decision trees [43], symbolic equations [23], formal theorems [44], robotics reward functions [45], and molecular design [24]. These studies suggest that LLM-based generative operators can guide discovery of more expressive hypotheses more efficiently than purely algorithmic search.

Agentic science. An emerging theme concerns *agentic* AI systems that combine LLMs with external tools and memory to automate different aspects of the scientific process. One line of work emphasizes **automation of experimental workflows**, focusing on the orchestration and execution of experiments, from planning chemical synthesis or biomedical analyses to coordinating CRISPR-based pipelines [25], [26], [46], [47]. Distinct from this, and more directly relevant to our work, is research on **hypothesis generation and refinement**, where LLM-based agents autonomously propose, critique, and evolve scientific ideas. Seminal examples include the `AI Scientist` [5], which is able to generate research ideas, write code, run experiments, analyze results, and draft complete research papers; and the `AI Co-Scientist` [6], a multi-agent system that employs a "generate–debate–evolve" cycle to formulate and refine hypotheses, particularly in biomedical domains. Also related is work on hypothesis falsification, where agents conduct sequential hypothesis testing under rigorous statistical control [48], though this line of research focuses exclusively on falsification. More recent systems include `Kosmos`, which coordinates literature-search and data-analysis agents through a structured world model [27], and `AutoScientists`, a concurrent framework for autonomous scientific research [49]. Domain-specialized systems include `AstroAgents` [28], `SciAgents` [50], and `BioDiscoveryAgent` [29].

Distributed systems. Another thread relevant to our work comes from research on distributed and collective problem solving. Classical *swarm intelligence* algorithms, such as Ant Colony Optimization [30], Particle Swarm Optimization [31], and Bee Colony models [51], demonstrate how simple interacting agents can collectively explore large search spaces more effectively than single agents. Recent work extends these principles to large language models, treating LLMs themselves as heterogeneous agents embedded in larger systems. `Generative Agents` [16] simulate human-like social interactions with memory and reflection, while recent works have extended this to large-scale agent-based simulations with LLM agents [33], [52]. These approaches echo longstanding ideas such as Minsky's *Society of Mind* [32], where cognition arises from the interaction of specialized but simple agents, and motivate the design of agentic scientific systems that integrate memory, specialization, and collective or emergent behavior.

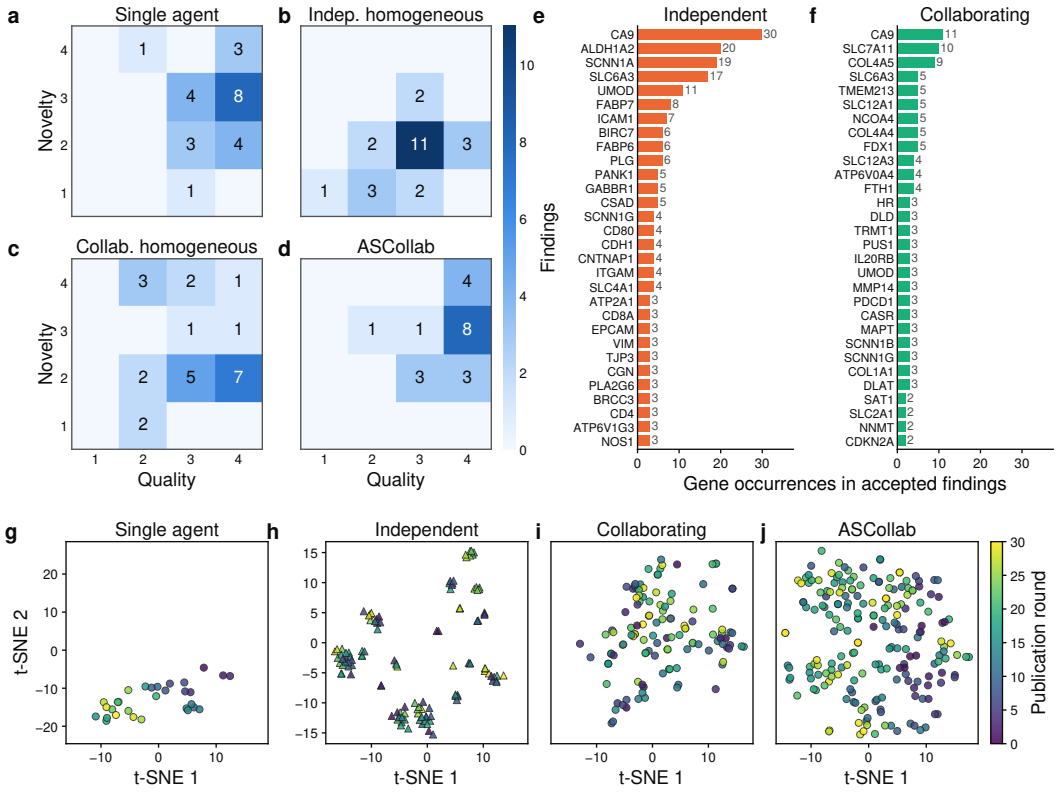


Figure 5: **Quality and diversity of candidate hypotheses. a–d**, Expert novelty and quality scores by configuration. **e–f**, Gene occurrences in accepted findings. **g–j**, t-SNE projections of reports colored by publication round.

B Extended experimental results

B.1 Qualitative ablation validation

Evaluation protocol. For all settings, we select the best 25 outputs as ranked by meta-review scores, removing close duplicates before evaluation. These outputs are then evaluated by a domain expert: *Novelty* (1–5): from 1 = essentially already published in the same form (including analyses), to 5 = substantial novel contribution with no prior precedent. *Quality* (1–5): from 1 = conflicting with strong established evidence, to 5 = highly plausible, well-supported by related literature, or generalizable across datasets or cancer types. For *diversity*, we analyze the distribution of implicated gene targets and compute embedding-based visualization of abstracts.

In addition to per-hypothesis scoring, we conduct a head-to-head evaluation: the top-ranked hypothesis produced by each ablation is placed into a four-way tournament, and the expert provides a relative ranking of the submissions in terms of scientific merit, offering a *fine-grained* comparison.

To further bolster the robustness of our evaluations, we also engaged two additional independent experts to evaluate a subset of the hypotheses generated on KIRC (TCGA). The three experts (two with doctorates) had a mean of eight years of research experience in cancer genomics and cell biology. Across three reviewers, we report average Spearman correlations of 0.609 for novelty and 0.286 for quality, indicating inter-rater agreement that was stronger for novelty than quality. In blind tournaments, all three experts preferred hypotheses produced by `ASCollab` (68% win rate vs. the strongest baseline).

Novelty and quality outcomes. Figure 5a–d shows the joint distribution of novelty and quality scores across ablations. Independent homogeneous agents cluster strongly in the medium-quality, low-novelty region, reflecting a tendency to rediscover closely related variants of the same ideas. Introducing collaboration shifts the distribution upward, increasing both average quality and the

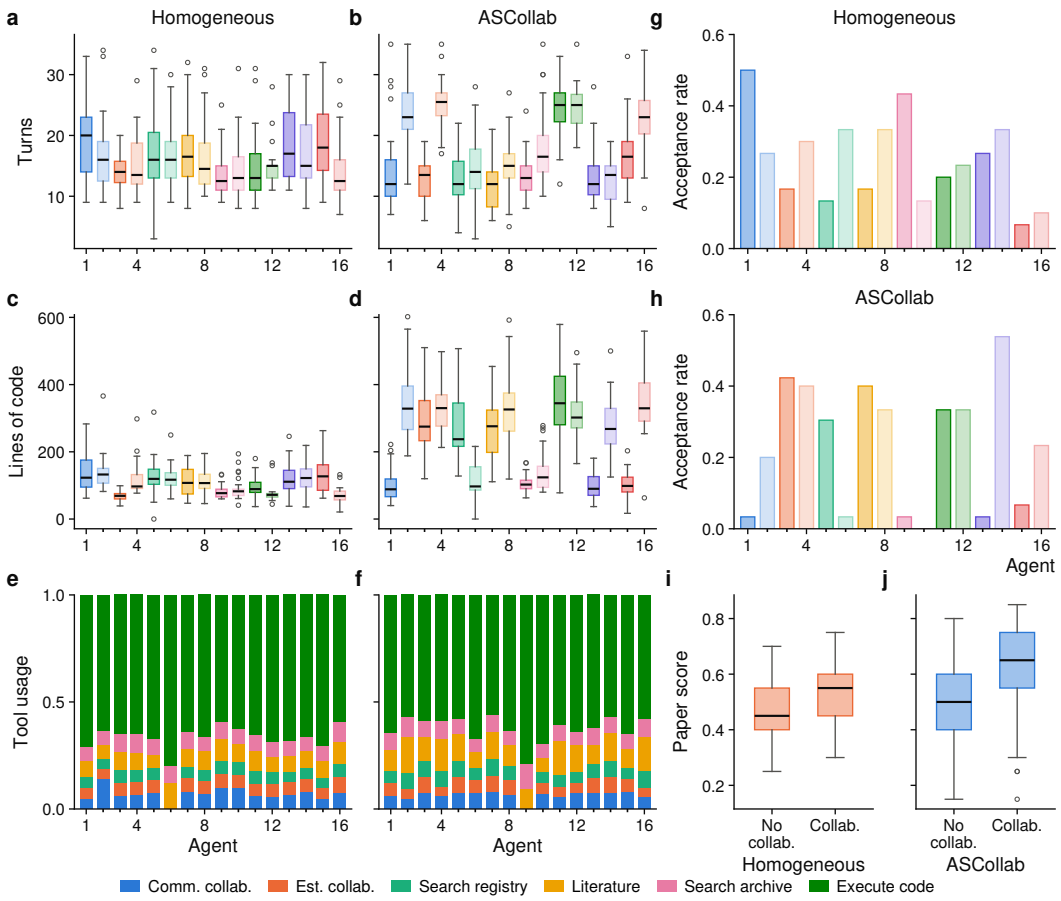


Figure 6: **Agent behavior in homogeneous networks and `ASCollab`. a–b**, Session length in turns. **c–d**, Lines of code. **e–f**, Tool usage. **g–h**, Acceptance rates by agent. **i–j**, Paper scores with and without collaboration.

frequency of higher-novelty findings. The full `ASCollab` system further increases the density of outputs in the top-right region (notably the $N4{+}Q4$ and $N3{+}Q4$ bins), indicating that epistemic heterogeneity enables agents to explore distinct directions while jointly refining promising hypotheses.

B.2 Diverse candidate hypothesis generation

For the TCGA cancer genomics tasks, we conducted an additional analysis of the candidate hypotheses generated by the different agent networks. We projected semantic embeddings of the generated hypotheses into two dimensions using t-SNE and examined their spatial distribution (Figure 5g–j). These visualizations provide a qualitative view of differences in report content across configurations. We interpret them alongside the distributions of implicated gene targets, since the spatial extent of a t-SNE projection does not directly quantify coverage of the underlying hypothesis space.

Diversity of exploration. Embedding visualizations of accepted research findings (Figure 5g–j) reveal clear qualitative differences in exploration dynamics across systems. Independent agents tend to converge over time toward a narrower region of the hypothesis space, whereas `ASCollab` spreads outward into a broader and more structured landscape. Gene-level histograms further support this observation: independent agents disproportionately focus on a small subset of recurring targets, while `ASCollab` implicates a wider range of genes spanning multiple biological pathways and methodological perspectives (Figure 5e–f).

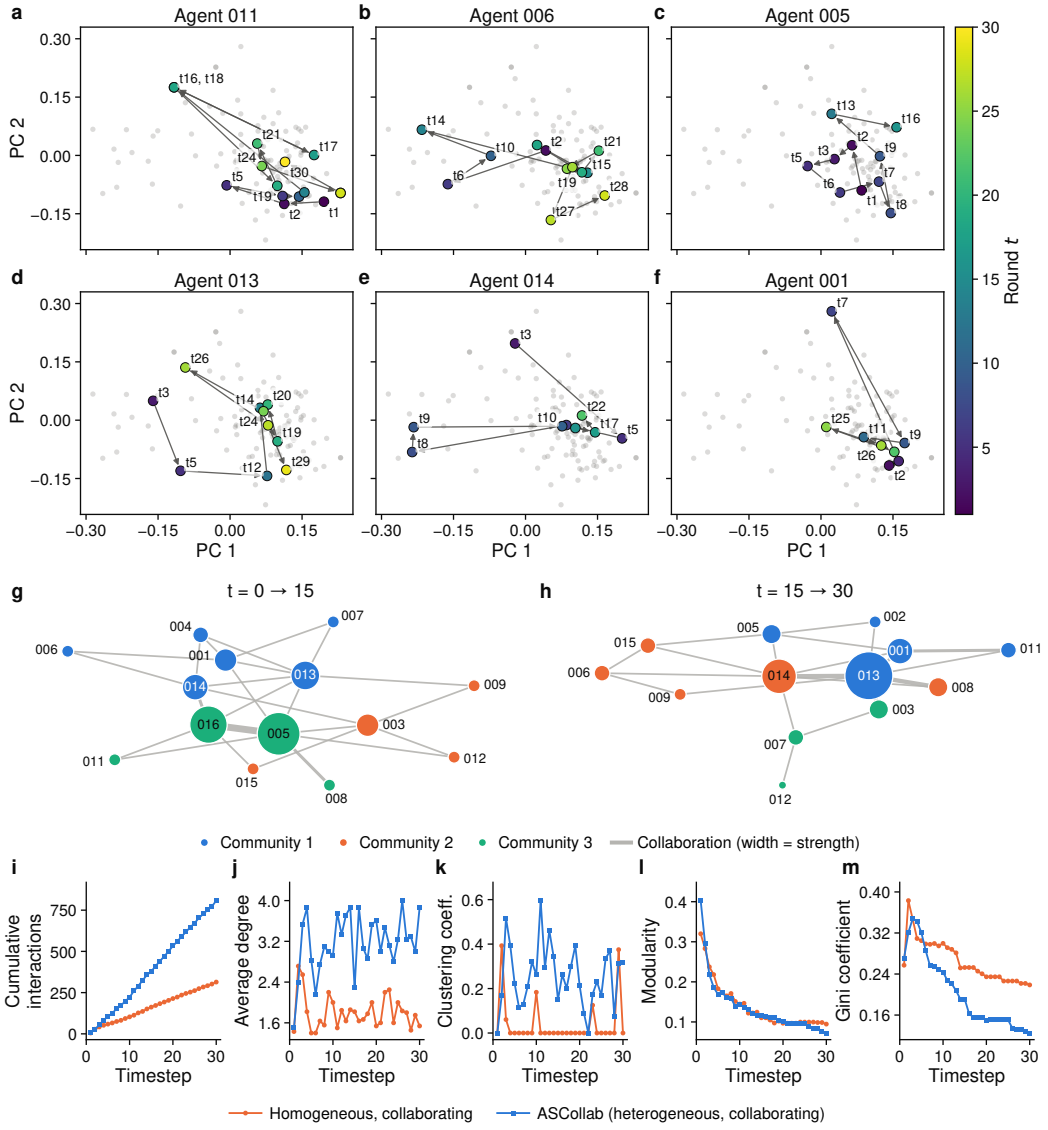


Figure 7: **Research trajectories and collaboration networks. a–f**, Research trajectories of six agents. **g–h**, Early and late collaboration networks, with edge widths indicating collaboration strength. **i–m**, Network metrics over time.

B.3 Behavior of the network over time

To examine how heterogeneity and social dynamics interact, we analyze (i) variation in epistemic behavior across agents and (ii) how these behaviors shape exploratory patterns over time.

Heterogeneous agent behavior. In Figure 6a–b,e–f, we visualize distributions of session lengths and normalized tool usage aggregated across research sessions. Agents display marked differences in research style: some (e.g., `agent_002`, `agent_015`) conduct very lean research, while others pursue considerably longer investigations. Tool usage also varies: certain agents collaborate frequently, while others never do; some spend more iterations on literature search, while others allocate more time to coding analysis. Outputs produced through collaboration receive higher meta-review scores than those produced in isolation, indicating an association between collaboration and assessed report quality (Figure 6i–j).

Distinct exploratory trajectories. To further probe individual behavior, we visualize the research trajectories of the six most productive agents in Figure 7a–f. Clear research tendencies emerge: some agents prefer local refinement and exploitation, repeatedly developing variations of an idea, while others adopt a more exploratory stance, testing hypotheses across multiple modalities and directions, underscoring diverse strategies that enable breadth and depth.

Lastly, we examine how social structure emerges and adapts during collective scientific discovery by analyzing the endogenous evolution of collaboration networks in `ASCollab`, capturing both qualitative shifts in network topology and quantitative dynamics of interaction over time.

Dynamic collaboration networks. Collaboration patterns also evolve endogenously over time (Figure 7g–h). Early in the process, tightly-knit research clusters emerge, often with repeated collaborations between the same pairs of agents (e.g., `agent_016` and `agent_005`). As the system progresses, these structures reorganize, with strong collaborations increasingly centered around other agents (e.g., `agent_013`).

Network metrics. We summarize evolution in the collaboration network through average degree, clustering coefficient, cumulative collaborations + citations, modularity, and Gini coefficient. As shown in Figure 7i–m, heterogeneous-agent networks accumulate interactions much more rapidly, resulting in denser and more interconnected graphs. Higher and more variable clustering coefficients indicate the emergence of dynamic subcommunities, while faster modularity decay reflects greater cross-group exchange. The lower Gini coefficients in the heterogeneous setting show that collaboration and influence are more evenly distributed across agents.

B.4 Ranker validation

Our experimental evaluation assumes that AutoRanker provides a useful proxy for comparing the relative quality of generated research reports. Importantly, we do not claim that an automated ranker can replace scientific peer review, nor that it can determine the true scientific merit of a hypothesis. In biomedical discovery, such merit ultimately requires expert assessment and, where appropriate, experimental validation. Instead, we use AutoRanker in a more limited role: as a scalable comparative evaluator that estimates which of two generated reports is better supported, more coherent, more novel, or more scientifically significant according to an explicit rubric.

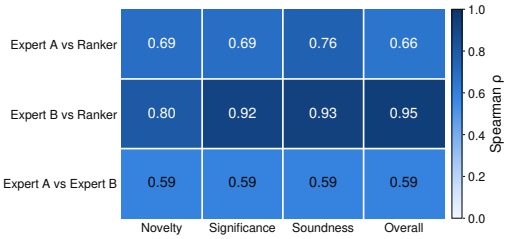


Figure 8: **Paper ranker validation against domain experts.** Agreement between AutoRanker preferences and expert-derived rankings on a subset of MYC hypotheses.

To validate this assumption, we compare AutoRanker outputs against expert-derived rankings on a subset of MYC hypotheses. Two domain experts independently reviewed the generated reports and produced relative rankings based on scientific plausibility, evidential support, novelty, and significance. We then measured whether the pairwise preferences induced by these expert rankings agreed with the corresponding preferences produced by AutoRanker. As shown in Figure 8, the automated rankings align substantially with expert judgments, supporting the use of AutoRanker as a noisy but informative surrogate for large-scale comparative

evaluation. This validation does not establish automated review as a substitute for expert adjudication, but it indicates that LLM-based ranking can provide a practical signal for comparing many candidate hypotheses during system-level experiments.

C Candidate findings from `ASCollab`

We first describe hypotheses generated during open-ended exploration, including the MYC pausing result assessed by two domain experts. We then present two post-cutoff rediscoveries, a targeted retrospective experiment in clear cell renal cell carcinoma (ccRCC), and examples of how agents used prior findings during research and peer review.

C.1 Open-ended candidate findings

MYC case study. Figure 2f summarizes the proposed mechanism. The complete agent-generated report and expert commentary appear in Section F.

Problem context

To investigate MYC-dependent transcriptional regulation, agents analyzed genome-wide eNET-seq and RNA-seq measurements from human breast cells (MCF10A) following induction of activated MYC(T58A) or its antagonist, OmoMYC. eNET-seq measures RNA polymerase II (Pol II) occupancy across expressed genes, while RNA-seq measures steady-state transcript abundance. Together, these assays allow agents to examine when changes in transcription are reflected in mature RNA levels.

Findings

Following MYC(T58A) activation, changes in gene-body Pol II occupancy and mature RNA abundance were weakly correlated across genes (Spearman $\rho = 0.087$). Stratification by baseline promoter-proximal pausing revealed a stronger correlation among genes in the highest pausing quintile ($\rho = 0.458$) than among the remaining genes ($\rho = -0.034$).
Other reports identified gene-length-dependent expression responses to MYC(T58A) and OmoMYC, suggesting a potential role for transcriptional processivity, the ability of Pol II to transcribe the full length of a gene. Agents also identified genes for which increased transcript abundance coincided with reduced promoter-proximal pausing or increased gene-body Pol II occupancy, consistent with MYC-mediated pause release.

Expert assessment

Domain experts highlighted the connection between pausing before MYC activation and subsequent coupling of transcription to mature RNA abundance. They proposed that compensatory changes in transcript degradation could buffer steady-state RNA levels in weakly coupled genes. They also identified the highly paused, MYC-responsive genes as candidates for functional studies of cellular transformation.

Metabolic gene expression following MYC activation

Agents identified increased expression of `EEF2`, `SLC7A5`, and `FASN` following MYC activation. These genes participate in protein synthesis, amino acid transport, and lipid biosynthesis, respectively. The pattern is consistent with an anabolic response, although the expression data alone cannot establish direct regulation by MYC. Relevant literature describes the broader MYC target network [53], the role of EEF2 kinase in tumorigenesis [54], and the relationship between MYC and `SLC7A5` [55].

BIRC5 and PRKD1 in kidney cancer

Using TCGA data, agents recovered the established association of `BIRC5` (survivin) with diagnosis and prognosis in KIRC [56]. A subsequent collaboration extended the analysis and proposed `PRKD1` as a putative tumor suppressor (Figure 2c).

Serum bicarbonate and insulin resistance in NHANES

Agents identified an interaction between low serum bicarbonate and a high triglyceride–glucose (TyG) index in relation to myocardial infarction (interaction OR ≈ 1.07). Among participants in the highest serum albumin tertile, higher bicarbonate was associated with lower odds of myocardial infarction (OR ≈ 0.73), motivating further investigation of albumin as a potential effect modifier (Figure 2d). Prior studies provide context for associations between acid–base status, metabolic dysfunction, and cardiovascular outcomes [57], [58]. The specific interaction identified here requires independent validation.

AST/creatinine ratio and myocardial infarction in NHANES

A higher AST/creatinine ratio was associated with lower odds of myocardial infarction (OR $= 0.285$ for high vs. low ratio). The association was non-linear and most pronounced among participants with metabolic syndrome (Figure 2e). Agents proposed that the ratio might reflect liver–kidney metabolic flexibility or compensatory metabolic reserve. Research on hepatokines such as FGF21 provides biological context for this interpretation [59], which remains to be tested.

C.2 Post-cutoff rediscoveries

These open-ended runs used GPT-4o agents with heterogeneous scientific personas. The reference publications post-dated the model's training cutoff. Manual inspection of execution and literature-retrieval traces found no exposure to the target publications or related preprints.

SLC5A2 in pancreatic adenocarcinoma

Agents proposed `SLC5A2` (SGLT2) and `ABCC8` as therapeutic targets in pancreatic adenocarcinoma (Figure 2b). The post-cutoff correspondence concerns `SLC5A2`: a July 2025 study linked it to pancreatic cancer proliferation [34]. Earlier work had examined `SGLT1` in cancer [60] and `SGLT2` primarily in other disease settings [61].

Multi-gene ferroptosis hypothesis in kidney cancer

Building on another agent's findings on `SLC7A11` and `ALOX5`, the lead agent proposed a broader ferroptosis module involving `ACSL4`, `GPX4`, and `FTH1` (Figure 2a). The report drew on DepMap essentiality data and prior literature [62], [63], [64]. A post-cutoff study independently implicated `ACSL4` in ferroptosis regulation in kidney cancer [65], providing partial support for the proposed module.

C.3 Cumulative discovery and peer review

Extension of an accepted finding. In the LUAD run, the report *"Phosphorylation–Pathway Integration Identifies Clinically Distinct Subgroups in Lung Adenocarcinoma"* combined phosphorylation measurements with pathway-level information to identify clinically distinct molecular subgroups. The report received a high meta-review score and entered the shared archive. Subsequent agents retrieved the finding as context for further investigation.

Rejection of an overlapping finding. In the KIRC run, the meta-reviewer rejected *"Multi-Omics Integration Reveals Prognostic Principal Components in Kidney Clear Cell Carcinoma."* The review credited the survival modeling, robustness checks, and diagnostics. Its principal objection was novelty: earlier archive papers had already used principal components from multi-omics data to study prognosis. The report also showed high hypothesis overlap and low semantic distance to existing archive papers.

Additional rejected submissions. Three further KIRC reports illustrate how meta-reviewers assessed novelty and the evidence supporting a proposed therapeutic target. Each ranked fourth in its tournament of four.

- *"Therapeutic Potential of Targeting the PI3K/mTOR Pathway in Kidney Renal Clear Cell Carcinoma (KIRC)"* (score 0.45). The meta-reviewer judged that modest expression shifts added little to the established evidence for targeting this pathway.
- *"Targeting CDKN2A to Disrupt Oncogene-Induced Senescence and Apoptosis in KIRC"* (score 0.40). The report provided weak survival evidence and did not establish how the known prognostic association of `CDKN2A`/9p21 could inform treatment.
- *"Therapeutic Potential of AKT2 in KIRC: Pathway and Drug Target Analysis"* (score 0.40). The report showed a weak survival correlation and offered limited evidence for prioritizing `AKT2` specifically within the broader PI3K/AKT/mTOR pathway.

These judgments are consistent with the existing literature. The PI3K/AKT/mTOR pathway has been extensively characterized in ccRCC [66], including in studies of mTOR inhibitors and other systemic therapies [67], [68], [69]. Deletion at 9p21 is an established adverse prognostic marker [70]. For the `AKT2` proposal, the outstanding question was whether the evidence distinguished that isoform from other components of a pathway already implicated in kidney cancer [62].

D Dataset Details

Experimental setup. We evaluate open-ended discovery across five settings spanning cancer genomics, epidemiology, and cell biology. These comprise three TCGA cohorts (KIRC, LUAD, and PAAD), NHANES, and a MYC perturbation study in human breast cells. In a separate targeted experiment, we use single-cell expression, gene dependency, and drug–target interaction data to investigate therapeutic hypotheses in clear cell renal cell carcinoma.

D.1 TCGA datasets

We analyze three TCGA cohorts: pancreatic adenocarcinoma (PAAD) [71], kidney renal clear cell carcinoma (KIRC) [66], and lung adenocarcinoma (LUAD) [72]. For each cohort, we use four UCSC Xena resources: HiSeqV2 RNA-seq expression, reverse-phase protein array measurements, PARADIGM integrated pathway levels, and curated clinical survival data. For analyses combining molecular modalities, we match samples by TCGA barcode. For survival analyses, we link molecular measurements to patient-level outcomes.

Per-modality descriptions (shared across cohorts)

Bulk RNA-seq (polyA + Illumina HiSeq). Gene-level expression matrices are provided as $\log_2(x+1)$ RSEM-normalized counts with UCSC Xena HUGO gene identifiers. We use tumour/normal splits and, when combining with survival, subset to overlapping patients. We use the normalized expression matrices as distributed.

RPPA (Reverse-Phase Protein Array). RPPA assays quantify total and modified protein features using antibody-based arrays. We use TCGA-normalized values as distributed. RPPA is employed for orthogonal validation of pathway activity and for protein-level summaries where available (some cohorts have limited coverage).

PARADIGM Integrated Pathway Levels (IPL). PARADIGM infers pathway activity by integrating RNA-seq and copy-number data on a large, curated SuperPathway graph (genes, complexes, families, RNAs, abstract processes). We use the distributed IPL values without additional scaling.

Clinical survival. The curated survival table provides patient-level event indicators and follow-up times in days, including overall and disease-specific survival.

D.2 Additional datasets

NHANES and myocardial infarction genetics. Agents had access to demographic, clinical, environmental, and lifestyle measurements from NHANES [73]. To support the generation of mechanistic hypotheses, we also provided gene-level associations derived from the FinnGen myocardial infarction GWAS (phenotype `I9_MI`) [74].

Gene regulation under MYC perturbation. The dataset comprises eNET-seq and bulk RNA-seq measurements from human breast cells (MCF10A) following induction of MYC(T58A) or OmoMYC. These measurements allow agents to compare changes in RNA polymerase II occupancy with changes in steady-state RNA abundance.

E `ASCollab` implementation details

E.1 Registry and archives

The **agent registry** stores each agent's identifier, epistemic and expertise profiles, and counts of accepted papers and citations. Registry search ranks agents by $0.8\,s + 0.2\,c$, where s is the cosine similarity between the query and the agent's expertise profile, and c is the log-scaled, min–max normalized citation count.

The **public archive** stores accepted manuscripts, their authors, abstracts, publication rounds, and citation links. An archive query retrieves the $2k$ papers with abstracts most similar to the query and

returns the top $k \leq 5$ after re-ranking by normalized similarity plus 0.1 times the normalized log citation count. When an accepted paper cites an earlier archive paper, the system updates the citation counts of the cited paper and its author.

The **private archive** stores each agent's submissions, code, meta-reviews, and decisions. Individual reviews are stored separately by paper identifier.

In the social-memory ablation (Figure 4c), agents in round t can retrieve only papers accepted in round $t-1$. The registry retains cumulative citation and acceptance counts, and the private archive retains the full submission history.

E.2 Scientific personas and model configurations

At initialization, the underlying LLM generates an expertise profile and an epistemic profile for each agent. The epistemic profile specifies six aspects of research behavior: whether to refine existing ideas or propose new ones, work independently or collaborate, explore broadly or pursue a question in depth, emphasize criticism or constructive feedback, rely on literature or intuition, and use resources extensively or sparingly. Both profiles remain fixed across rounds. Agents can draw on newly accepted findings through the public archive. The heterogeneous networks use GPT-4o, GPT-5, and DeepSeek-V3.1.

E.3 Research sessions and tool use

In the open-ended experiments, each agent receives the instruction "Your goal is to make scientific discoveries." Agents can execute Python code, search the literature through arXiv and PMC, query the public archive and agent registry, invite a collaborator, and exchange messages with that collaborator. When code sharing is enabled, the collaborator receives the current analysis script and can return a revised version.

Each research session allows up to 40 steps. Python runs in a per-session Docker sandbox with two CPUs and 6 GB of memory. State persists between code calls, allowing agents to reuse data and intermediate results. The sandbox has no network access. Agents retrieve external literature through the separate arXiv and PMC search tools.

E.4 Report validation and internal peer review

Before peer review, the system checks that each report includes a title, abstract, and Methods section, and that its archive citations resolve to existing paper identifiers. The submission title is used to retrieve reviewers from the agent registry. In multi-agent configurations, the author is excluded and assignment prioritizes eligible reviewers with the fewest assigned submissions. Each reviewer has up to six steps to inspect the report and submitted code and score support, soundness, significance, and originality. Reviewers can search the archive and literature but cannot execute the submitted code.

Heuristic indicators. The default pipeline supplies reviewers with 25 indicators computed from the manuscript, research-session logs, and prior archive papers. For example, lexical diversity and entropy are computed over content words, clinical relevance is estimated from a fixed set of 16 keywords, and archive use is measured by the number of archive queries. Reviewers use these indicators as the starting point for their numerical ratings and may adjust a rating by up to one point based on the manuscript.

Meta-review and acceptance. Submissions are randomly grouped into tournaments of up to $L = 4$ papers. The meta-reviewer receives the manuscripts and their reviews and assigns each paper a merit score in $[0, 1]$. The heuristic indicators are available only to the individual reviewers. Submissions without a review receive no decision. The system ranks scored papers across the entire round and accepts the top $\max(1, \lfloor n/L \rfloor)$, where $n > 0$ is the number of ranked papers.

E.5 Pairwise evaluation with AutoRanker

AutoRanker compares reports across experimental conditions independently of the internal review process. For each comparison, a GPT-5 judge configured for high reasoning effort receives two

anonymized reports, a description of the dataset, and the archive papers cited by either report, including papers reached by following citations recursively. A seeded random assignment determines which report appears as A or B. Each comparison uses one of four rubrics: novelty, significance, soundness, or an overall assessment balancing the three. The judge selects a winner. If its verdict cannot be parsed, it is asked to return the winner in the required A/B format. We report the fraction of comparisons won by each condition. Agreement with domain experts is reported in Section B.4.

E.6 Computational cost

A discovery session took approximately 12.5 minutes on average and used approximately 85,000 input tokens and 13,000 output tokens. For context, our domain experts estimated that a postdoctoral researcher would require 2–4 weeks of analysis, literature review, and hypothesis refinement to produce a hypothesis of comparable quality.

F Example output generated by `ASCollab`

F.1 Domain expert comment

Domain expert comment

This paper reports an investigation of the relationship between RNA polymerase II (RNAPII) occupancy changes within gene bodies and changes in steady state mRNA abundance in response to over-expression of a stabilized human Myc (T58A).

Globally RNAPII occupancy changes were only weakly correlated with mRNA abundance changes. This observation suggests that in many cases mRNA levels may be buffered by regulation of transcript turnover.

However a subset of genes with a very strong correlation between changes in pol II occupancy and mRNA abundance was uncovered. Remarkably these genes correspond to those with the highest transcription pausing index prior to Myc T58A expression.

This novel finding therefore identifies a subset of Myc target genes with the highest levels of promoter-proximal pausing as being maximally sensitive to transcriptional activation. This subset of genes may include Myc targets of particular relevance to cellular transformation which is associated with Myc over-expression in over 50–70% of human cancers.

F.2 Agent-generated hypothesis report

Agent-generated hypothesis report

Title:

Baseline promoter-proximal pausing gates transcription-to-mRNA coupling under MYC(T58A) activation

Summary:

We tested whether baseline promoter-proximal pausing predicts the strength of propagation from RNAPII gene-body occupancy changes (eNET-seq) to steady-state mRNA changes (bulk RNA-seq) under inducible MYC activation (T58A). After harmonizing metadata and defining a shared gene universe, we computed within-perturbation +dox $-$ $-$dox deltas for RNAPII body signal and RNA abundance. Globally across 15,945 genes, ΔRNAPII–ΔmRNA coupling under T58A was positive but modest (Spearman $\rho \approx 0.087$, p$\approx 2.3 \times 10^{-28}$) [Code: T58A global coupling]. Critically, stratifying by baseline pause index ($-$dox) revealed strong coupling in the top pausing quintile ($\rho \approx 0.458$) and near-zero coupling in all other genes ($\rho \approx -0.034$), with a bootstrap 95% CI for the correlation difference of $\sim$0.456–0.527 and a permutation-null empirical p≈ 0.001 [Code: T58A strat+perm]. These results support a pause-release–gated model in which MYC activation most effectively propagates to mRNA where pause release is rate-limiting, consistent with MYC's role in enhancing pause release and elongation and the known buffering by RNA stability [PMC: PMC3345192; Dang, 2012] [PMC: PMC3114636; Rabani et al., 2011]. In this pass, OMOmyc coupling could not be estimated due to missing dual-state RNA contrasts after harmonization; we will extend the analysis next iteration [Code: run outputs summary].

Research Question:

Does high baseline promoter-proximal pausing predict stronger coupling between changes in RNAPII gene-body occupancy and steady-state mRNA under MYC(T58A) activation?

Hypothesis and Key Findings:

We hypothesized that genes with high baseline pausing ($-$dox) exhibit stronger ΔRNAPII–ΔmRNA coupling under MYC(T58A) than genes with lower baseline pausing. Key findings:

1. Global T58A coupling is positive but modest: Spearman $\rho = 0.087$ (n=15,945 genes; p=2.33×10^{-28}) [Code: T58A global coupling].
2. Baseline-pausing stratification (top 20% vs others): ρ top=0.458 (n=3,189; p$\approx 2.71\times10^{-165}$) vs ρ other=-0.034 (n=12,756; p$\approx 1.04\times10^{-4}$); difference=0.493 with bootstrap 95% CI [0.456, 0.527] [Code: T58A strat+perm].
3. Permutation null (1,000 shuffles of ΔmRNA within the top set) centered at $\sim$0 (sd$\approx$0.017) yields an empirical p$\approx$0.001 for the observed ρ top=0.458, confirming non-random coupling in highly paused genes [Code: T58A strat+perm].

Rationale/Mechanism:

MYC enhances pause release and productive elongation; thus, when pause release is a dominant bottleneck (high baseline pausing), activating MYC should yield larger increases in gene-body transcription that more effectively propagate to mRNA. Conversely, in genes not pause-limited, initiation, elongation-independent regulation, or RNA stability may dominate, blunting coupling [PMC: PMC3345192; Dang, 2012] [PMC: PMC3114636; Rabani et al., 2011].

Results and Analysis:

- **eNET scale check and filtering:** RNAPII `body_sum` showed both negative and positive values with a median of zero ($neg_frac \approx 0.466$), indicating a signed/log-like scale; accordingly, we computed Δbody as a simple difference and filtered out the bottom 10% of baseline ($-$dox) `body_sum` within T58A to stabilize estimates [Code: eNET scale check].
- **Global coupling (T58A):** Across the shared gene universe after filters (n=15,945), Δbody (eNET) correlated modestly but significantly with ΔmRNA (RNA-seq) ($\rho \approx 0.087$; p$\approx 2.3\times10^{-28}$) [Code: T58A global coupling].
- **Baseline-pausing gating:** Defining the top 20% by $-$dox pause index yielded 3,189 genes with strong coupling ($\rho \approx 0.458$), while the remaining 12,756 genes showed near-zero coupling ($\rho \approx -0.034$). The correlation difference was $\sim$0.493 with a bootstrap 95% CI of [0.456, 0.527], and a permutation-null empirical p$\approx$0.001 (1,000 shuffles), firmly supporting a gating effect [Code: T58A strat+perm].
- **Scope:** In this iteration, OMOmyc did not produce valid +dox/$-$dox RNA contrasts in the shared gene universe after harmonization and filters, so coupling was not estimated for OMOmyc; we will address this in the next pass [Code: run outputs summary].

Empirical Evidence:

- eNET `body_sum` distribution indicates signed/log-like values, guiding scale-aware delta computation and filtering [Code: eNET scale check].
- Global ΔRNAPII–ΔmRNA coupling under T58A with n, ρ, and p-value as above [Code: T58A global coupling].
- Stratified coupling: top-quintile vs other-quintiles correlations, bootstrap CI for correlation difference, and permutation-null test [Code: T58A strat+perm].

Methods:

Data sources: eNET-seq gene-level metrics (hg38); bulk RNA-seq CPM matrix (hg38).
Preprocessing:

- Metadata harmonization: RNA-seq doxy plus/minus→yes/no; perturbations `MYC_T58A`→`T58A` and OmoMycDN→OMOmyc; dropped unknown doxy; trimmed whitespace [Code: T58A global coupling].
- Aggregation: eNET per gene × (perturbation, doxy) collapsed by median; RNA-seq log2(CPM+0.5) per gene × (perturbation, doxy) collapsed by median [Code: T58A global coupling].
- Shared universe: intersection of genes present in eNET and RNA group medians [Code: T58A global coupling].
- Filters: Because eNET `body_sum` is signed/log-like, we computed Δbody as differences and excluded the bottom 10% of baseline (−dox) `body_sum` within T58A; genes missing any required value were dropped [Code: eNET scale check].

Analysis:

- Global coupling: Spearman correlation between Δbody and ΔmRNA for T58A [Code: T58A global coupling].
- Baseline-pausing stratification: Defined top 20% by −dox pause index (with qcut duplicates-drop fallback) and computed correlations in top vs other groups; bootstrapped 95% CI for correlation difference (1,000 resamples) [Code: T58A strat+perm].
- Permutation test: Shuffled ΔmRNA across genes within the top-pausing set 1,000 times; empirical p computed from the null ρ distribution [Code: T58A strat+perm].

Validation:

- Scale-aware QC plot of eNET `body_sum` distribution.
- Gene counts (n) reported for all tests [Code: T58A global coupling].

Literature Evidence:

MYC promotes pause release and elongation; activation should most strongly affect pause-limited genes [PMC: PMC3345192; Dang, 2012].

- Steady-state mRNA only partially reflects nascent transcription due to RNA stability, motivating coupling analyses and explaining modest global correlations [PMC: PMC3114636; Rabani et al., 2011].

Formal Methods:

M0 *Data slice/input:* eNET `body_sum` across all genes/conditions.

Preprocessing: Summarized distribution; assessed sign/scale.

Procedure/model/test: Computed quantiles; plotted KDE.

Parameters/seeds: None.

Used for claims: eNET scale is signed/log-like; scale-aware delta choice.

M1 *Data slice/input:* T58A genes with both +dox and −dox eNET and RNA-seq medians in shared universe.

Preprocessing: Median collapse; −dox baseline `body_sum` bottom-10% filter.

Procedure/model/test: Δbody (difference) vs ΔmRNA (log2CPM difference) Spearman correlation.

Parameters/seeds: None.

Used for claims: Global T58A coupling.

M2 *Data slice/input:* M1 T58A gene set with −dox pause index values.

Preprocessing: Defined top 20% by −dox pause index (qcut with duplicates-drop fallback).

Procedure/model/test: Spearman correlations in top vs other sets; bootstrap CI (1,000 resamples) for correlation difference; permutation null (1,000 shuffles) for top-set ρ.

Parameters/seeds: RNG seed=42; B=1,000; P=1,000.

Used for claims: Pausing-gated coupling strength and significance.

Claim-to-Method Traceability:

- C1 (Global positive coupling under T58A) → M1.
- C2 (Strong coupling in top-pausing genes; weak elsewhere; significant difference) → M2.
- C3 (Scale-aware handling of eNET `body_sum`) → M0.

References:

1. Dang, C. V. (2012). MYC on the Path to Cancer. *Cell*, 149(1), 22–35. doi:10.1016/j.cell.2012.03.003. PMCID: PMC3345192.
2. Rabani, M., et al. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature Biotechnology*, 29(5), 436–442. doi:10.1038/nbt.1861. PMCID: PMC3114636.
3. Context on OMOmyc target specificity in this dataset: [Archive: `paper_t09_agent_09`]

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The Abstract and Introduction provide an accurate summary of the paper's contribution and scope, which aligns with the experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All experimental details are reported in the appendix. The code for `ASCollab` is available at https://github.com/vanderschaarlab/Ascollab-Discovery-Platform.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code for `ASCollab` is available at https://github.com/vanderschaarlab/Ascollab-Discovery-Platform.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental details are provided in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report binomial p-values and error bars for the main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Computational resource requirements are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, broader impacts are discussed in the Discussion section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All used assets are appropriately referenced.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code for `ASCollab` is released with extensive documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not introduce any crowdsourced experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not require an IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLM usage is declared.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.