SAGE: Evidence-First Biomarker Discovery through Multi-Agent Reasoning

Sahar Almahfouz Nasser[1,*,a,b,c,e]
Emory University, USA

Juan Francisco Pesantez Borja[2,*,a,b,c,e]
Georgia Institute of Technology, USA

Jincheng Liu[2,*,a,b,c,e]
Georgia Institute of Technology, USA

Sandeep Manandhar[1,a,c,d,e]
Emory University, USA

Shikhar Shiromani[3,b,c,e]
Georgia Institute of Technology, USA

Mohammad Tanvir Hasan[4,a,b,c,e]
University of Arkansas at Little Rock, USA

Zenghan Wang[2,a,b,c,e]
Georgia Institute of Technology, USA

Suman Ghosh[2,a,b,e]
Georgia Institute of Technology, USA

Jinchu Li[2,a,b,c,e]
Georgia Institute of Technology, USA

Xuejian Xu[2,b,c,e]
Georgia Institute of Technology, USA

Aniket Ramkrishnan Iyer[2,b,c,e]
Georgia Institute of Technology, USA

Naoto Tokuyama[1,d,e]
Emory University, USA

Twisha Shah[2,b,e]
Georgia Institute of Technology, USA

Tilak Pathak[1,d,e]
Emory University, USA

Soundharya Kumaresan[1,d,e]
Emory University, USA

Yohei Abe[1,d,e]
Emory University, USA

Himanshu Maurya[1,c,e]
Emory University, USA

Anant Madabhushi[1,a,d,e]
Emory University, USA

Author Contributions: [*]Indicates equal contributions; [a]Conceptualization and Methodology; [b]System Development and Implementation; [c]Experimental Design and Evaluation; [d]Biomedical and Clinical Expertise; [e]Writing and Revision.

Abstract

Engineered image-based biomarkers offer a clinically interpretable alternative to black-box AI in computational pathology, yet their discovery remains largely intuition-driven, guided by fragmented literature rather than rigorous biological validation. We introduce SAGE (Structured Agentic system for hypothesis Generation and Evaluation), a multi-agent framework that grounds biomarker discovery in biological evidence through three mechanisms: (i) knowledge-graph-anchored hypothesis generation via multi-path ontological reasoning, (ii) a debate-based multi-agent novelty assessment that stress-tests candidate biomarkers against existing literature, and (iii) an end-to-end automated validation pipeline that translates hypotheses directly into executable analyses on multimodal pathology datasets.

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

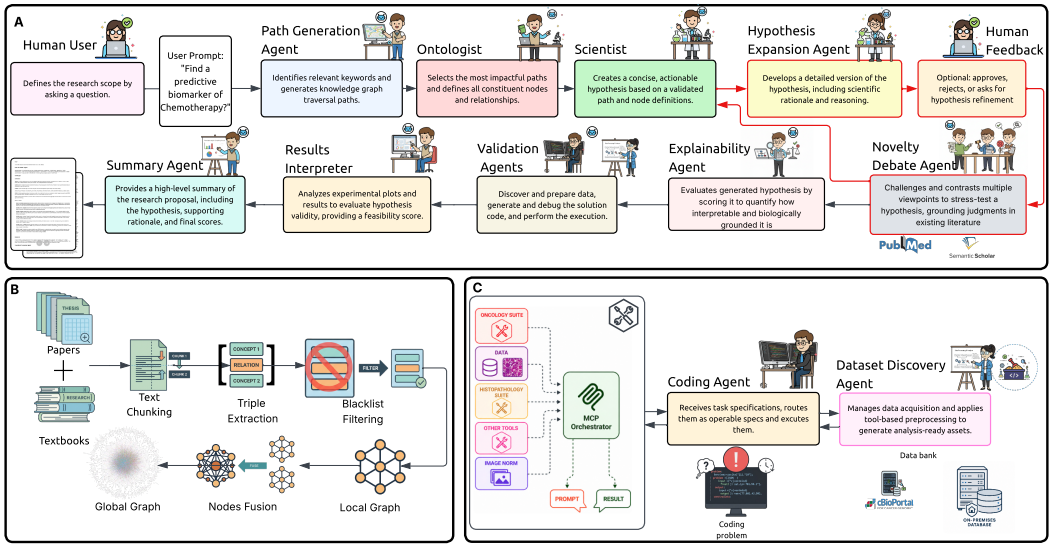

Figure 1: **Overview of the SAGE framework for computational pathology.** The pipeline comprises three stages: (A) *Hypothesis generation and refinement* (top), where coordinated agents produce interpretable hypotheses with optional human feedback at key decision points; (B) *Knowledge graph construction and reasoning* (bottom left), integrating pathology-aware biomedical entities to support structured biological reasoning; and (C) *Validation and summarization* (bottom right), which executes statistical analyses and generates clinically interpretable summaries.

> Together, these components shift biomarker discovery from an intuition-driven, literature-browsing exercise into a structured, traceable reasoning process that clinicians and researchers can inspect, trust, and build upon.

1 Introduction

Recent advances in large language models (LLMs) Luo et al. [2022], Singhal et al. [2023], Remy et al. [2022] have catalyzed agentic systems capable of synthesizing scientific literature and generating hypotheses. Prominent examples include the Virtual Lab Swanson et al. [2025] and AI Co-Scientist Gottweis et al. [2025], which demonstrate the power of multi-agent reasoning for biomedical discovery. However, both systems are designed around wet-lab validation cycles. Computational pathology operates under a fundamentally different constraint: discoveries must be grounded in measurable features spanning imaging, molecular, and clinical domains, and validated retrospectively on patient cohort data.

Within computational pathology, a *biomarker* is a measurable biological signal that correlates with disease state or clinical outcome Ballman [2015]. Discovering reliable biomarkers is central to precision medicine Collins and Varmus [2015], yet the field lacks systematic tools for doing so. ML-driven systems such as SciAgents Ghafarollahi and Buehler [2024] demonstrate that coordinated LLM agents can generate interesting hypotheses but stop short of biological grounding or empirical validation. Clinically motivated frameworks such as NOVA Vaidya et al. [2025] and WSI-Agents Lyu et al. [2025] target diagnostic objectives but rely on black-box models without explicit biological reasoning. The result is a field where biomarker discovery remains intuition-driven: hypotheses are generated without biological justification, and validation is treated as a separate manual step.

We introduce **SAGE** (Structured Agentic system for hypothesis Generation and Evaluation), an end-to-end agentic framework for biologically grounded biomarker discovery in computational pathology. SAGE makes four contributions: **(i)** an end-to-end discovery pipeline spanning hypothesis generation to computational validation on retrospective patient cohorts; **(ii)** multi-path ontological reasoning over a domain-specific knowledge graph linking imaging, molecular, and clinical concepts; **(iii)** a debate-based novelty assessment mechanism that stress-tests candidate biomarkers against existing literature; and **(iv)** a biologically grounded validation framework that produces interpretable and

clinically actionable results. We publicly release the SAGE hypothesis-generation framework at `https://github.com/Emory-Empathathetic-AI-for-Health-Inst/SAGE`.

2 Related Work

2.1 Biomarker Discovery in Computational Pathology

A biomarker is a measurable biological signal, derived from tissue morphology, molecular profiles, or their combination, that correlates with disease state or clinical outcome Ballman [2015]. Prognostic biomarkers stratify patients into high-risk groups (high probability of recurrence, relapse, or death) and low-risk groups. Predictive biomarkers identify patients likely to respond to a specific therapy. Both are central to precision oncology Collins and Varmus [2015], where treatment decisions depend on patient-specific biological evidence rather than population-level statistics.

Recent foundation models such as UNI Chen et al. [2024], CONCH Lu et al. [2024a], and Prov-GigaPath Xu et al. [2024] have demonstrated strong performance across diagnostic and prognostic tasks. Agentic systems such as PathChat Lu et al. [2024b], NOVA Vaidya et al. [2025], and WSI-Agents Lyu et al. [2025] enable interactive slide analysis and clinically motivated biomarker discovery. Despite this progress, most approaches treat biomarker discovery as a pattern recognition problem, producing associations that are statistically valid but biologically opaque, difficult to interpret, and rarely linked to mechanistic molecular hypotheses. SAGE complements these tools by providing a structured reasoning process that generates biologically grounded hypotheses and validates them computationally on multimodal patient data.

2.2 Agentic Systems for Scientific Discovery

The AI Co-Scientist Gottweis et al. [2025] is the closest prior work, employing a multi-agent tournament with inference-time scaling to generate and refine biomedical hypotheses. SAGE differs in two respects: it is purpose-built for computational pathology with domain-specific knowledge graphs and tools, and it validates hypotheses entirely in silico on retrospective patient cohorts rather than relying on wet-lab confirmation. The Virtual Lab Swanson et al. [2025] and Agent Laboratory Schmidgall et al. [2025] demonstrate agentic research automation but target wet-lab cycles and general ML workflows respectively, making them incomparable to SAGE on any shared task. SciAgents Ghafarollahi and Buehler [2024], which directly inspired SAGE, generates hypotheses through structured multi-agent reasoning but uses a planner-driven architecture with variability across runs and no empirical validation. AgentFlow Li et al. [2025] similarly uses a dynamic planner requiring on-policy reinforcement learning. SAGE instead adopts a fixed serial pipeline with well-defined agent roles, requiring no model training while ensuring reproducible, biologically grounded outputs.

2.3 Representation of Prior Knowledge and Validation

Text-based retrieval methods such as RAG Lewis et al. [2020] are effective at capturing surface-level evidence but cannot model relationships between concepts across domains. Knowledge graphs address this by encoding biomedical entities and relationships in structured form, enabling multi-hop reasoning. Foundational biomedical graphs such as Hetionet Himmelstein et al. [2017] and BioKG Zhang et al. [2025] support drug-repurposing and disease-mechanism discovery, while more recent work such as Tree-KG Niu et al. [2025] and GraphMERT Belova et al. [2025] focuses on scalable KG construction. SAGE adapts these principles to computational pathology, building the first domain-specific KG connecting imaging, molecular, and clinical concepts for hypothesis-driven biomarker discovery. We show empirically in Section 4.1 that graph-based reasoning produces more diverse hypotheses than text-based retrieval from a single prompt. On the validation side, ResearchCodeAgent Gandhi et al. [2025] integrates LLMs with predefined analysis scripts but its fixed execution logic limits adaptability. SAGE takes a complementary approach: its validation subsystem is equipped with a curated suite of computational pathology tools stress-tested on large multi-institutional cohorts. When a required analysis falls outside this toolset, SAGE's coding agent dynamically generates custom code on demand, ensuring that validation is neither constrained by predefined scripts nor sacrifices the reliability of battle-tested implementations where they exist.

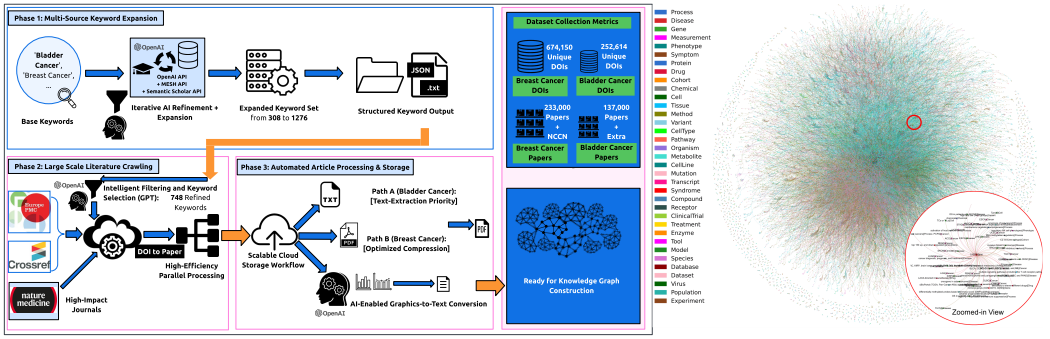

Figure 2: **Literature processing pipeline and biomedical knowledge graph.** *(Left)* Three-phase pipeline for corpus construction: keyword expansion via OpenAI, MeSH, and Semantic Scholar APIs; large-scale literature crawling from high-impact sources; and automated text extraction and storage. *(Right)* The resulting biomedical knowledge graph, where nodes represent biological entities and edges encode relationships mined from literature and medical ontologies. Colors indicate entity types (see legend); the zoomed-in subgraph highlights dense connectivity among disease-relevant entities.

3 Method

SAGE operationalizes biomarker discovery as a structured sequential pipeline comprising three stages (Figure 1): knowledge graph construction, which grounds all downstream reasoning in curated biological evidence; hypothesis generation and evaluation, where coordinated agents produce and refine candidate biomarkers; and empirical validation, where accepted hypotheses are automatically translated into executable analyses on multimodal patient cohort data.

3.1 Knowledge Graph Construction

Most agentic systems rely on text-based retrieval for prior knowledge, which is constrained to a single query context and cannot model relationships across biological domains. SAGE instead constructs a domain-specific biomedical knowledge graph (KG) encoding entities, relations, and supporting evidence in structured form, enabling multi-hop reasoning across imaging, molecular, and clinical domains.

Literature Processing and Triple Extraction. Candidate papers are screened using expert-curated keyword sets and scored on semantic relevance, SapBERT similarity Reimers and Gurevych [2019], and publication venue. Of all initially retrieved papers, approximately 40% are filtered as insufficiently relevant. Retained PDFs are segmented into overlapping 1,500-token chunks, each processed by `GPT-4o-mini` to extract biomedical triples $(h, t_h, r_{\text{text}}, r_{\text{norm}}, t, t_t, c, e)$, where h and t are head and tail entities, t_h and t_t their ontology types, r_{text} the surface relation, r_{norm} its normalized form, $c \in [0, 1]$ a confidence score, and e the supporting evidence span. Triples with $c \geq 0.5$ are retained and relations normalized via ontology-aware rule-based mapping Peng et al. [2017], Zhang et al. [2018]. See Figure 2

Graph Construction and Fusion. Local directed graphs $G_d = (V_d, E_d)$ are built per document with edges weighted by extraction confidence. Entity redundancy across documents is resolved using `BGE-Large` embeddings Xiao et al. [2023] at cosine similarity threshold $\tau = 0.9$, with redundant edges merged by weight summation. Disconnected components are pruned and the resulting global graph $\mathcal{G}_{\text{global}}$ serialized in GraphML format Brandes et al. [2001]. The final KG integrates 1,650 sources comprising 41,053 nodes and 56,338 edges. Quality assessment confirms 99.0% factual grounding and 100% relation accuracy, with entity type accuracy improving from 64.6% to 82.5% after ontology refinement (Appendix A.3, Figure 2).

Multi-Path Ontological Reasoning. Rather than committing to a single KG traversal path, the **Path Generation Agent** identifies multiple candidate paths linking entity combinations such as *Gene–Pathway–Disease* or *Biomarker–Tissue–Outcome*, exposing the breadth of biologically plausible

connections from a single query. Each path P is scored on four metrics balancing plausibility against discovery potential, where $\mathbf{v}_Q$ and $\mathbf{v}_P$ are query and path embeddings; $\mathbf{e}_i$ is the embedding of entity e_i and $\mathbf{c}_{\text{target}}$ the target domain centroid; $\deg(e)$ is the degree of entity e in the global KG; and P_{actual}, P_{expected} are the entity category distributions of the path and full KG respectively:

$$S_{\text{logic}} = \frac{\mathbf{v}_Q \cdot \mathbf{v}_P}{\|\mathbf{v}_Q\|_2 \, \|\mathbf{v}_P\|_2} \tag{1}$$

$$S_{\text{rel}} = 1 - \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{e}_i - \mathbf{c}_{\text{target}}\|_2 \tag{2}$$

$$S_{\text{nov}} = \frac{1}{|P|} \sum_{e \in P} -\log P(e), \quad P(e) = \frac{\deg(e) + \epsilon}{\sum_v \deg(v) + \epsilon} \tag{3}$$

$$S_{\text{sur}} = \text{KL}(P_{\text{actual}} \parallel P_{\text{expected}}) \tag{4}$$

These are combined into a weighted aggregate:

$$S_{\text{total}} = w_{\text{logic}} S_{\text{logic}} + w_{\text{rel}} S_{\text{rel}} + w_{\text{nov}} S_{\text{nov}} + w_{\text{sur}} S_{\text{sur}}, \quad \sum_i w_i = 1 \tag{5}$$

The weights w_i are user-configurable: higher w_{logic} and w_{rel} favor plausible hypotheses, while higher w_{nov} and w_{sur} steer SAGE toward bolder discoveries. Here S_{sur} measures distributional surprise relative to background KG statistics, capturing paths whose entity-type composition deviates meaningfully from the global graph. Full derivations are in Appendix A.4.

3.2 Hypothesis Generation and Evaluation

The hypothesis generation stage coordinates six agents in a fixed sequential order, as illustrated in Figure 1-A: **Path Generation, Ontologist, Scientist, Hypothesis Expansion, Novelty Debate, and Explainability**. This fixed serial architecture produces consistent, reproducible outputs and reduces hallucination risk compared to planner-driven systems Ghafarollahi and Buehler [2024], while reducing inference cost by up to 75% through targeted context allocation, as shown in Appendix B (Table 7). A human-in-the-loop checkpoint between the Scientist and Hypothesis Expansion agents allows domain experts to approve, reject, or refine hypotheses before computational resources are committed. Agent roles and model assignments are detailed in the Appendix (Table 6).

The **Ontologist** selects the most impactful KG paths and grounds each in curated biomedical ontologies, mapping graph relations to biologically interpretable processes. The **Scientist Agent** then generates concise, dataset-aware hypotheses following a structured template that specifies the target population, biological and imaging variables, clinical endpoint, and expected directionality. The **Hypothesis Expansion Agent** subsequently enriches each approved hypothesis with mechanistic rationale and clinical interpretation, producing a fully reasoned scientific proposal ready for adversarial evaluation.

3.2.1 Debate-Based Novelty Assessment

A central contribution of SAGE is the **Novelty Debate Agent**, a multi-critic evaluation framework inspired by adversarial debate protocols Irving et al. [2018] that simulates the peer review process through structured argumentation. Three sub-agents with distinct epistemic stances participate: a **Prover** (optimistic) that advocates for the hypothesis by identifying unique mechanistic claims and gaps in existing literature; a **Verifier** (conservative) that challenges novelty claims through real-time retrieval across six scholarly databases (PubMed, Semantic Scholar, Europe PMC, bioRxiv, arXiv, and CrossRef); and a **Judge** (balanced) that synthesizes both sides and renders a final assessment.

Each critic assigns an initial score on a 1–10 scale. When significant disagreement is detected ($\sigma > 1.0$), a multi-round Bayesian debate Zhang et al. [2024] is initiated, updating scores iteratively until convergence ($\sigma < 0.5$) or a maximum of three rounds. The Verifier can flag arguments as *specious* when a novelty claim is directly contradicted by retrieved literature, triggering a score penalty when upheld by the Judge. The final novelty score is the unweighted mean across all three critics.

3.2.2 Explainability Agent

The **Explainability Agent** scores each hypothesis on biological and clinical interpretability using an *Explainability Index* (EI) composed of five criteria, each scored 0–2:

$$\mathrm{EI} = \mathrm{MD} + \mathrm{CP} + \mathrm{SBC} + \mathrm{CT} + \mathrm{MT} \tag{6}$$

The five components capture complementary dimensions of biomarker credibility: **Mechanistic Depth (MD)** evaluates grounding in established biological pathways; **Causal Plausibility (CP)** assesses whether biomarker variation could reasonably influence the disease state or outcome, rather than merely reflecting a downstream correlate; **Spatial and Biological Coherence (SBC)** checks localization within the relevant tissue architecture; **Clinical Traceability (CT)** evaluates measurability via pathology-compatible assays; and **Model Transparency (MT)** assesses whether the biomarker has a clear, reproducible quantitative definition. These dimensions are deliberately non-redundant: for instance, a biomarker mapping to an immune-exhaustion pathway may score high on MD yet low on CP if it appears only as a late consequence of tumor progression rather than a plausible driver of treatment resistance. Conversely, a tumor–immune spatial interaction pattern may score high on CP even when the precise molecular mechanism awaits further validation.

3.3 Empirical Hypothesis Validation

Accepted hypotheses are forwarded to the validation stage, coordinating the **Validation Agents** subsystem, the **Results Interpreter**, and the **Summary Agent** in sequence.

3.3.1 Dataset Discovery Agent

Rather than assuming data availability, the **Dataset Discovery Agent** determines which datasets can support a given hypothesis before any code is written. The hypothesis is first parsed into explicit constraints encoding the target cohort, predictor, outcome, exposure, and analysis objective. These constraints are organized into an *Artifact Requirement Graph* (ARG) $\mathcal{A} = (\mathcal{T}, \mathcal{R})$, where $\mathcal{T}$ is a set of validation targets (e.g., cohort definition, predictor measurement, outcome measurement) and $\mathcal{R}$ is the set of dependencies between them. Available files are indexed from public omics repositories, clinical registries, and institutional data banks. Each file f is evaluated against the ARG through a two-stage deterministic algorithm. In the first stage, each target $\tau \in \mathcal{T}$ is decomposed into fine-grained requirements $\{r_1, \ldots, r_k\}$ and each file receives a coverage status $\phi(f, \tau) \in \{\text{none, partial, full}\}$, reflecting how well the file satisfies the data needs of that target. We further define a binary satisfaction indicator:

$$\sigma(S, \tau) \;:=\; \mathbb{1}\left[\phi\left(\bigcup_{f \in S} f,\, \tau\right) = \text{full}\right] \tag{7}$$

In the second stage, a greedy selection identifies the smallest subset of files $\mathcal{F}^* \subseteq \mathcal{F}$ whose combined content fully satisfies every ARG target:

$$\mathcal{F}^* = \arg\min_{S \subseteq \mathcal{F}} |S| \quad \text{s.t.} \quad \forall\, \tau \in \mathcal{T},\ \sigma(S, \tau) = 1 \tag{8}$$

where $\bigcup_{f \in S} f$ denotes the union of information contributed by all selected files, and $\sigma(S, \tau) = 1$ indicates that target τ is completely satisfied. At each greedy step, the file contributing the highest weighted marginal coverage gain over still-unmet targets is added to $\mathcal{F}^*$, until all targets reach full coverage or no further gain is achievable. Hypotheses for which no suitable $\mathcal{F}^*$ exists are flagged for expert review rather than proceeding to code execution.

3.3.2 Coding Agent and Tool Orchestration

The **Coding Agent** operates in three stages: file inspection, hypothesis-conditioned code generation, and sandboxed execution within containerized environments. Failed scripts enter a repair loop using execution feedback, retrying up to a fixed budget before escalating to expert review (Appendix C.1).

Tool selection is mediated by a **Tool Orchestration Subsystem** in which each tool is registered through a three-layer specification enabling token-efficient routing, planning, and on-demand execution. The orchestrator dynamically selects from validated tools spanning survival analysis, histopathology, transcriptomics, proteomics, and genomic correlation; when no suitable tool exists, the Coding Agent synthesizes new Python code within sandbox constraints. Full architecture and illustrative examples are provided in Appendix C.2.

Table 1: Hypothesis quality evaluation across six configurations ($N = 10$ hypotheses per prompt, 46 prompts). Debate-agent scores: novelty, confidence, and consensus (1–10). Vendi Score: per-prompt diversity averaged across 46 prompts (max = 10). Human-expert scores: novelty and feasibility from four blinded domain specialists. **Bold** = best per metric. * = significantly different from SAGE (Full) based on non-overlapping 95% confidence intervals.

System	Debate Agent Scores				Human Expert Scores			
	Novelty	Confidence	Consensus	Vendi	Novelty	Feasibility	Confidence	Consensus
ChatGPT 5.2	6.91* [6.75, 7.07]	**6.70*** [6.49, 6.90]	7.06	5.00	6.74 [5.95, 7.53]	**10.00*** [10.00, 10.00]	7.07 [6.25, 7.88]	6.83 [5.91, 7.76]
Gemini 3.1 Pro	6.90* [6.72, 7.08]	6.30 [6.08, 6.52]	7.23	**6.10**	4.53* [4.08, 4.98]	8.56* [8.18, 8.94]	7.00 [6.09, 7.91]	6.92 [6.23, 7.61]
AgentFlow	6.82* [6.63, 7.01]	**6.76*** [6.52, 6.99]	6.76	5.22	4.83* [4.24, 5.42]	8.27* [7.73, 8.81]	7.67 [6.82, 8.52]	6.67 [5.82, 7.51]
SAGE (Text)	7.00* [6.87, 7.13]	6.18 [6.04, 6.32]	7.21	4.51	5.41* [4.89, 5.93]	7.26 [6.79, 7.73]	**7.93** [7.52, 8.35]	6.50 [5.77, 7.23]
SAGE (Single-Path)	7.33 [7.16, 7.50]	6.42* [6.25, 6.60]	7.18	4.11	6.36 [5.64, 7.08]	7.86 [7.41, 8.31]	7.13 [6.58, 7.69]	**7.25** [6.62, 7.88]
SAGE (Full)	**7.67** [7.50, 7.84]	6.00 [5.91, 6.09]	**7.85**	5.07	**7.10** [6.49, 7.71]	7.03 [6.54, 7.52]	6.40 [5.25, 7.55]	6.83 [6.10, 7.57]

3.3.3 Results Interpreter and Summary Agent

The **Results Interpreter** evaluates hypothesis validity from Coding Agent outputs based on statistical significance, effect size, and biological interpretability. The **Summary Agent** consolidates the hypothesis, novelty score, explainability index, and validation results into a structured research report.

4 Experimental Results

We evaluate SAGE across four dimensions corresponding to its core contributions: hypothesis quality and diversity, debate-based novelty assessment, automated validation performance, and clinical relevance of discovered biomarkers. All experiments use bladder cancer as the target disease domain, as its biomarker landscape remains comparatively underexplored relative to other cancers, making systematic discovery particularly valuable.

4.1 Hypothesis Quality and Diversity

We compare six configurations to evaluate knowledge graph reasoning and multi-path traversal: ChatGPT 5.2 and Gemini 3.1 Pro as direct prompting baselines; AgentFlow Li et al. [2025] as a multi-agent baseline (planer-based); SAGE (Text) replacing the KG with text retrieval; SAGE (Single-Path) using a single KG path; and SAGE (Full) using multi-path ontological reasoning.

Each system generated $N = 10$ hypotheses per prompt across 46 prompts (questions from four domain experts). For all baselines, SAGE (Text), and SAGE (Single-Path), diversity arises from running the same prompt 10 times independently. For SAGE (Full), diversity arises structurally from selecting the top-10 ranked KG paths from a single prompt. Diversity is measured using the Vendi Score Friedman and Dieng [2023] computed per prompt ($N = 10$, max = 10) and averaged across 46 prompts (Appendix B.7).

Results. Table 1 shows that SAGE (Full) achieves the highest automated novelty (7.67) and reviewer consensus (7.85). On diversity, SAGE (Full) outperforms both SAGE (Single-Path) (5.07 vs 4.11) and SAGE (Text) (5.07 vs 4.51), confirming that structured KG traversal produces more semantically distinct hypotheses than repeated prompting. Although Gemini 3.1 Pro achieves the highest Vendi score (6.10), its lower automated novelty (6.90) and expert novelty (4.53) suggest this reflects stochastic generation variability rather than biologically structured exploration, see the limitations of Vendi score in Appendix B.7.

Expert evaluation by four blinded domain specialists confirms that SAGE (Full) achieves the highest novelty score (7.10). A tradeoff emerged: novel hypotheses scored lower on feasibility because experts were less familiar with the proposed associations, while ChatGPT and Gemini generated hypotheses resembling reformulations of the original questions, appearing more immediately feasible. This tradeoff is consistent with SAGE's goal of surfacing underexplored rather than established associations, please refer to Appendix C.4 for more details on the expert evaluation protocol.

Cross-Domain Computational Scalability. A key challenge in extending agentic biomarker discovery to new cancer types is whether larger knowledge graphs increase inference cost and latency. To evaluate this, we benchmarked SAGE across bladder, breast, and prostate cancer knowledge

graphs containing 41,053, 28,088, and 21,814 nodes, respectively (56 research questions per domain; $n = 168$ runs). Despite the differences in graph size, mean execution times remained between 142.8 and 169.8 seconds, with inference costs of only \$0.042–\$0.045 per hypothesis. Graph size showed negligible correlation with runtime ($r = -0.02$) and weak correlation with inference cost ($r = 0.20$). These findings suggest that SAGE's targeted path retrieval and selective context allocation decouple per-hypothesis inference cost from knowledge-graph size, supporting computationally efficient expansion to additional cancer domains. Detailed results are provided in Appendix B.3.

Takeaway 1

SAGE (Full) achieved the highest novelty and improved diversity compared with the single-path and text-based SAGE variants, highlighting the value of multi-path knowledge graph reasoning.

4.2 Ablation Study: Knowledge-Graph Path Scoring

We assess the contribution of the four path-scoring components (logic, relevance, novelty, and surprise) through leave-one-out ablations, setting each component's weight to zero while retaining the remaining weights. Across 56 bladder cancer research questions, we evaluate changes in downstream hypothesis novelty using SAGE's novelty critic. Removing relevance produced the largest reduction in mean novelty score ($\Delta = -0.27$, $n = 55$), followed by surprise ($\Delta = -0.25$, $n = 56$), novelty ($\Delta = -0.14$, $n = 56$), and logic ($\Delta = -0.06$, $n = 54$).

We additionally assess path-ranking sensitivity across 54 questions. Removing surprise changed the top-ranked path in all cases (100%), whereas removing relevance changed it in only 3.7% of cases. Despite its limited effect on path ranking, relevance yielded the largest reduction in downstream novelty when removed. These findings suggest that path-ranking sensitivity alone does not capture a scoring component's contribution to the quality of generated hypotheses. Detailed results are provided in Appendix A.5.

4.3 Ablation Study: Debate-Based Novelty Assessment

We evaluate whether adversarial deliberation, defined as iterative evidence-driven score revision triggered by critic disagreement, improves novelty assessment calibration beyond multi-critic averaging. All experiments use time-travel backtesting on 150 historical proposals spanning five novelty tiers to prevent temporal leakage Ye et al. [2026]. Full dataset composition and deliberation statistics are provided in Appendix B.8.

Experimental design. We compare four configurations isolating the effect of deliberation: (i) *Single-Critic*; (ii) *Two-Critic* ensemble (Prover and Verifier) without interaction; (iii) *Multi-Critic (No Debate)*, where three critics score independently and are averaged; and (iv) *Multi-Critic (Full)*, our proposed design, where critics enter iterative Bayesian debate when initial disagreement exceeds $\sigma > 1.0$. The sole difference between (iii) and (iv) is disagreement-triggered deliberation, isolating the contribution of our debate mechanism.

Results. Table 2 reports overall accuracy and A–E separability, defined as the mean score difference between breakthrough (A) and specious (E) proposals. Multi-Critic (Full) achieves the largest marginal gain: a 1.76-point increase in A–E separability over the no-debate condition ($^*p < 0.05$), and a 35.3% absolute accuracy improvement over single-critic scoring. Deliberation is triggered most frequently for extreme categories (A: 76.7%, E: 83.3%), where it resolves high initial disagreement through structured evidence exchange. Error analysis and computational cost breakdown are provided in Appendix B.8.

Takeaway 2

Debate across three critics improves novelty assessment accuracy by 35.3% over a single agent.

4.4 Ablation Study: Validation Pipeline

We evaluate the importance of dedicated dataset discovery by benchmarking the Coding Agent on KramaBench Lai et al. [2026], a public data-science benchmark spanning diverse tabular reasoning

Table 2: Effect of deliberation on novelty separability across 150 proposals. A–E Gap denotes the mean score difference between breakthrough (A) and specious (E) proposals. *Statistically significant improvement over the preceding configuration ($p < 0.05$).

Configuration	Accuracy	A–E Gap
Single-Critic	56.0%	1.12
Two-Critic (Verifier + Prover)	75.3%*	4.18*
Multi-Critic (No Debate)	81.3%*	5.02*
Multi-Critic (Full)	**91.3%***	**6.78***

Table 3: Coding Agent performance on KramaBench under oracle file selection (upper bound) and unguided autonomous search (lower bound) across four model backends. *Exec.%*: fraction of tasks where the generated script ran successfully. *Exact%*: fraction of tasks with a correct answer. The gap between conditions quantifies the performance headroom unlocked by dedicated dataset discovery.

Oracle Discovery ($N = 94$)					**Unguided Discovery** ($N = 106$)				
Model	**Score**	**Exact%**	**Exec.%**	**Failure**	**Model**	**Score**	**Exact%**	**Exec.%**	**Failure**
codex	42.18	31.9%	78.7%	Script	codex	18.28	12.3%	37.7%	No files
5.4-03-05	40.79	28.7%	73.4%	Script	5.4-03-05	15.76	9.4%	29.2%	No files
5.4-mini	32.04	25.5%	63.8%	Script	5.4-mini	12.89	9.4%	39.6%	Exec.
5.4-nano	29.90	20.2%	61.7%	Script	5.4-nano	12.96	9.4%	37.7%	No files

tasks. We compare two conditions that bound the impact of data retrieval quality: **(i) Oracle Discovery** ($N = 94$ tasks), where the task-relevant files are pre-attached to the Coding Agent, simulating a perfect file-selection oracle and establishing an upper bound on achievable performance; and **(ii) Unguided Discovery** ($N = 106$ tasks), where the Coding Agent autonomously searches a raw dataset root without any structured retrieval, representing the lower bound when no dedicated discovery mechanism is available.

Table 3 summarizes performance across both conditions and four model backends. Under oracle conditions, execution success rates range from 61.7% to 78.7%, reflecting the Coding Agent's capability when supplied with correct data. Under unguided discovery, execution rates drop sharply to 29.2%–39.6%, with the dominant failure mode being failure to locate a valid file set rather than failures in code generation or execution. This gap directly demonstrates that data retrieval, not code generation, is the primary bottleneck in end-to-end hypothesis validation. These findings motivated the design of the Dataset Discovery Agent in SAGE, which replaces unguided file search with a deterministic ARG-based retrieval algorithm that formally matches hypothesis requirements to available patient data. Note that Table 3 does not directly measure the performance of SAGE's Dataset Discovery Agent, but rather quantifies the performance headroom that a structured discovery mechanism can unlock. Full attrition analysis and per-model breakdown are provided in Appendix C.1.

Takeaway 3

Poor data retrieval, not code generation, is the primary bottleneck in automated hypothesis validation.

4.5 Case Study: End-to-End Biomarker Discovery

Starting from a single research question on prognostic signatures in bladder cancer, SAGE autonomously traversed its knowledge graph, generated and stress-tested a hypothesis, and executed its validation on 412 TCGA-BLCA patients without manual intervention.

Discovery. SAGE identified a joint prognostic interaction between FABP5 expression and the abundance of tertiary lymphoid structures (TLS), immune aggregates quantified from whole-slide images, a multimodal biomarker spanning molecular and imaging domains absent from prior literature.

Validation. Patients with high FABP5 and scarce TLS exhibited significantly worse overall survival compared to those with low FABP5 and abundant TLS (Figure 3), confirmed by Cox proportional hazards modeling. Additional validated hypotheses generated by SAGE are provided in Appendix D.

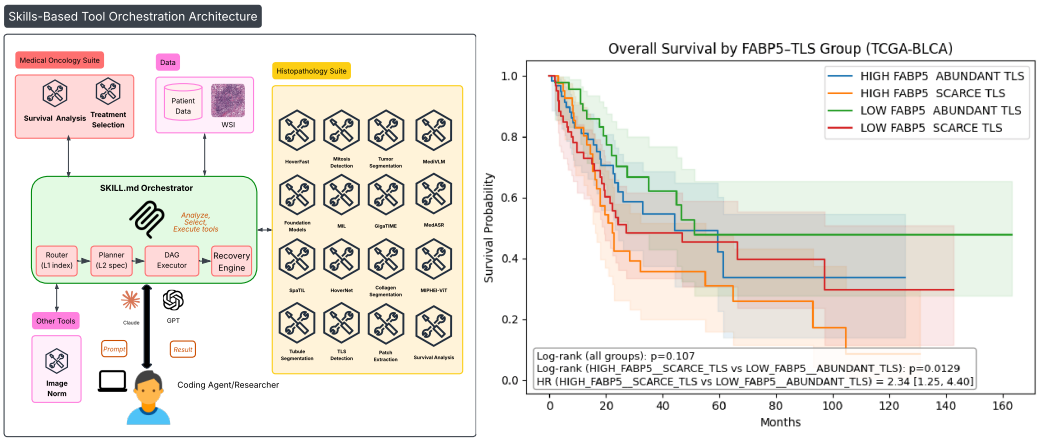

Figure 3: **Tool orchestration and end-to-end validation.** *(Left)* The SKILL.md orchestrator routes, plans, and executes tools across oncology and histopathology suites. *(Right)* Overall survival by joint FABP5 and TLS stratification in TCGA-BLCA (HR = 2.34 [1.25, 4.40], $p = 0.013$).

Takeaway 4

SAGE autonomously discovered and validated a novel multimodal biomarker, absent from prior literature.

5 Conclusion

SAGE treats biomarker discovery as a structured reasoning problem, unifying knowledge-graph construction, multi-path hypothesis generation, adversarial novelty debate, and automated in silico validation in a single pipeline. Multi-path graph-based reasoning produced significantly more diverse hypotheses than text-based retrieval, three-critic debate improved novelty assessment by 35.3% over single-agent scoring, and a dedicated dataset discovery agent improved validation success by 40%. Applied to bladder cancer, SAGE autonomously discovered and validated novel multimodal biomarkers absent from prior literature.

Current limitations include the restriction of the knowledge graph to text-based sources, the absence of direct benchmarking against systems such as AI Co-Scientist Gottweis et al. [2025] due to unavailable code, and the use of a general-purpose benchmark as a proxy for validation pipeline evaluation. Future work will extend the knowledge graph to genomic, proteomic, and spatial transcriptomic data across multiple cancer types, and quantify validation failure modes on real computational pathology tasks. Furthermore, when adapting SAGE to new scientific domains, future work will introduce an automated model discovery module that searches online repositories such as GitHub for domain-validated computational models, dynamically constructing a domain-specific tool registry for the coding agent rather than relying on a manually curated set of tools.

The design principles of SAGE are domain-agnostic and transferable to any scientific domain requiring interpretable, empirically grounded discovery, with the automated tool registry construction serving as the key bridge for seamless domain adaptation.

Acknowledgments

Research reported in this publication was supported by the National Cancer Institute under award numbers R01CA249992-01A1, R01CA257612-01A1, R01CA264017-01, R01CA268287-01A1, U01CA113913-16A1, U01CA269181-01, U24CA274494-01, U54CA302465-01, and R01CA268207-01; the National Heart, Lung and Blood Institute under award number R01HL158071-01A1; the National Institute of Allergy and Infectious Diseases under award numbers R01AI175555-01A1 and R01HL151277-01A1; the National Institute of Biomedical Imaging and Bioengineering under award number 75N92022D00015; the National Library of Medicine under award number R01LM013864-01A1; the National Institute on Aging under award number R01AG089759; the Na-

tional Institute of Diabetes and Digestive and Kidney Diseases under award number R01DK118431; the Office of the Director, National Institutes of Health under award number 1OT2OD038065-01; the Kidney Mapping and Atlas Project (KMAP) under award number U01DK133090-01;the United States Department of Veterans Affairs VA Merit Review award under award number IBX004121; the VA Biomedical Laboratory Research and Development Service under award number IK6BX006185; the VA Research and Development Office through the Lung Precision Oncology Program (LPOP-L0021); and sponsored research agreements from Astrazeneca, Bristol Myers Squibb, the Prevent Cancer Foundation, Innovation in Cancer Informatics, the Institute for Technology in Health Care, the Breast Cancer Research Foundation and the Scott R. Mackenzie Foundation; and the Advanced Research Projects Agency for Health (ARPA-H) under award number D25CA00140-00.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the U.S. Department of Veterans Affairs, or the United States Government.

Conflict of Interest Disclosure. Dr. Madabhushi is a Research Career Scientist at the Atlanta Veterans Affairs Medical Center and an equity holder in Picture Health, Elucid Bioimaging, and Inspirata Inc. He serves on the advisory board of Picture Health, consults for Takeda Inc., and has sponsored research agreements with AstraZeneca and Bristol Myers-Squibb. His technology has been licensed to Picture Health and Elucid Bioimaging.

References

Karla V Ballman. Biomarker: predictive or prognostic? *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 33(33):3968–3971, 2015.

Margarita Belova, Jiaxin Xiao, Shikhar Tuli, and Niraj K. Jha. Graphmert: Efficient and scalable distillation of reliable knowledge graphs from unstructured data, 2025. URL `https://arxiv.org/abs/2510.09580`.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32:D267–D270, 2004.

Ulrik Brandes, Markus Eiglsperger, Ivan Herman, Michael Himsolt, and M. Scott Marshall. GraphML progress report: Structural layer proposal. In *Graph Drawing*, pages 501–512, 2001.

Nuo Chen, Yicheng Tong, Yuzhe Yang, Yufei He, Xueyi Zhang, Zou Qingyun, Qian Wang, and Bingsheng He. Diversity collapse in multi-agent llm systems: Structural coupling and collective failure in open-ended idea generation. *arXiv preprint arXiv:2604.18005*, 2026.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning, 2023. *URL https://arxiv. org/abs/2210.02410*, 2023.

Shubham Gandhi, Dhruv Shah, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. Research-codeagent: An llm multi-agent system for automated codification of research methodologies, 2025. URL `https://arxiv.org/abs/2504.20117`.

Alireza Ghafarollahi and Markus J. Buehler. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning, 2024. URL `https://arxiv.org/abs/2409.05556`.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Aric Hagberg, Pieter Swart, and Daniel A. Schult. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11–15, 2008.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726, 2017.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate, 2018. URL https://arxiv.org/abs/1805.00899.

Eugenie Lai, Gerardo Vitagliano, Ziyu Zhang, Om Chabra, Sivaprasad Sudhir, Anna Zeng, Anton A. Zabreyko, Chenning Li, Ferdi Kossmann, Jialin Ding, Jun Chen, Markos Markakis, Matthew Russo, Weiyang Wang, Ziniu Wu, Mike Cafarella, Lei Cao, Samuel Madden, and Tim Kraska. KRAMABENCH: A benchmark for AI systems on data-to-insight pipelines over data lakes. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fZfUdeCC5X.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Zhuofeng Li, Haoxiang Zhang, Seungju Han, Sheng Liu, Jianwen Xie, Yu Zhang, Yejin Choi, James Zou, and Pan Lu. In-the-flow agentic system optimization for effective planning and tool use, 2025. URL https://arxiv.org/abs/2510.05592.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024a.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, 2024b.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), September 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac409. URL http://dx.doi.org/10.1093/bib/bbac409.

Xinheng Lyu, Yuci Liang, Wenting Chen, Meidan Ding, Jiaqi Yang, Guolin Huang, Daokun Zhang, Xiangjian He, and Linlin Shen. Wsi-agents: A collaborative multi-agent system for multi-modal whole slide image analysis. *arXiv preprint arXiv:2507.14680*, 2025.

Songjie Niu, Kaisen Yang, Rui Zhao, Yichao Liu, Zonglin Li, Hongning Wang, and Wenguang Chen. Tree-KG: An expandable knowledge graph construction framework for knowledge-intensive domains. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18516–18529, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.907. URL https://aclanthology.org/2025.acl-long.907/.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5:101–115, 2017.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

François Remy, Kris Demuynck, and Thomas Demeester. Biolord: Learning ontological representations from definitions (for biomedical concepts and their textual descriptions), 2022. URL https://arxiv.org/abs/2210.11892.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5977–6043, 2025.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023. URL https://arxiv.org/abs/2305.09617.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723, 2025.

Anurag J. Vaidya, Felix Meissen, Daniel C. Castro, Shruthi Bannur, Tristan Lazard, Drew F. K. Williamson, Faisal Mahmood, Javier Alvarez-Valle, Stephanie L. Hyland, and Kenza Bouzid. Nova: An agentic framework for automated histopathology analysis and discovery, 2025. URL https://arxiv.org/abs/2511.11324.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. Bge: Baai general embedding. GitHub repository, 2023. URL https://github.com/FlagOpen/FlagEmbedding.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

Bingyang Ye, Shan Chen, Jingxuan Tu, Chen Liu, Zidi Xiong, Samuel Schmidgall, and Danielle S. Bitterman. Proof of time: A benchmark for evaluating scientific idea judgments, 2026. URL https://arxiv.org/abs/2601.07606.

Weitong Zhang, Chengqi Zang, and Bernhard Kainz. Truth or deceit? a Bayesian decoding game enhances consistency and reliability, 2024. URL https://arxiv.org/abs/2410.01064.

Yijia Zhang, Wei Zheng, Hongfei Lin, Jian Wang, Zhihao Yang, and Michel Dumontier. Drug-drug interaction extraction via hierarchical rnns on sequence and shortest dependency paths. *Bioinformatics*, 34(5):828–835, 2018. doi: 10.1093/bioinformatics/btx546.

Yuan Zhang, Xin Sui, Feng Pan, Kaixian Yu, Keqiao Li, Shubo Tian, Arslan Erdengasileng, Qing Han, Wanjing Wang, Jianan Wang, et al. A comprehensive large-scale biomedical knowledge graph for ai-powered data-driven biomedical research. *Nature Machine Intelligence*, pages 1–13, 2025.

A Knowledge Graph Construction

A.1 Literature Processing Details

Candidate papers are retrieved using domain-specific keyword queries covering computational pathology, disease biomarkers, spatial analysis, and whole-slide imaging. Each paper is scored using a composite metric combining keyword frequency, semantic proximity via SapBERT similarity, and journal impact factor. Thresholds are calibrated empirically to balance recall and precision, resulting in approximately 40% of retrieved papers being excluded as insufficiently relevant.

PDF text extraction uses a three-tier fallback strategy employing `pdfplumber`, `PyMuPDF`, and `pypdf` to accommodate heterogeneous document formats. Extracted text is normalized and segmented into overlapping chunks of approximately 1,500 tokens to preserve local semantic context for relation extraction.

Triple extraction uses `GPT-4o-mini` with structured prompts enforcing entity typing, relation normalization, and evidence attribution. The ontology schema includes fine-grained entity types such as *Gene*, *Pathway*, *Disease*, *ClinicalEndpoint*, *TissueRegion*, *StainingMethod*, and *Algorithm*. Disambiguation rules resolve patient cohorts, clinical outcomes, and gene identifiers.

A.2 Graph Construction Details

For each document d, a local graph $G_d = (V_d, E_d)$ is constructed where nodes correspond to unique entities and edges aggregate all triples linking entity pairs, with weights defined as the maximum confidence among supporting triples. Local graphs are fused into a global KG using `BGE-Large` embeddings Xiao et al. [2023] with cosine similarity threshold $\tau = 0.9$. Redundant edges are merged by summing weights and disconnected components with fewer than three nodes are pruned. Graphs are serialized in GraphML format Brandes et al. [2001], Hagberg et al. [2008].

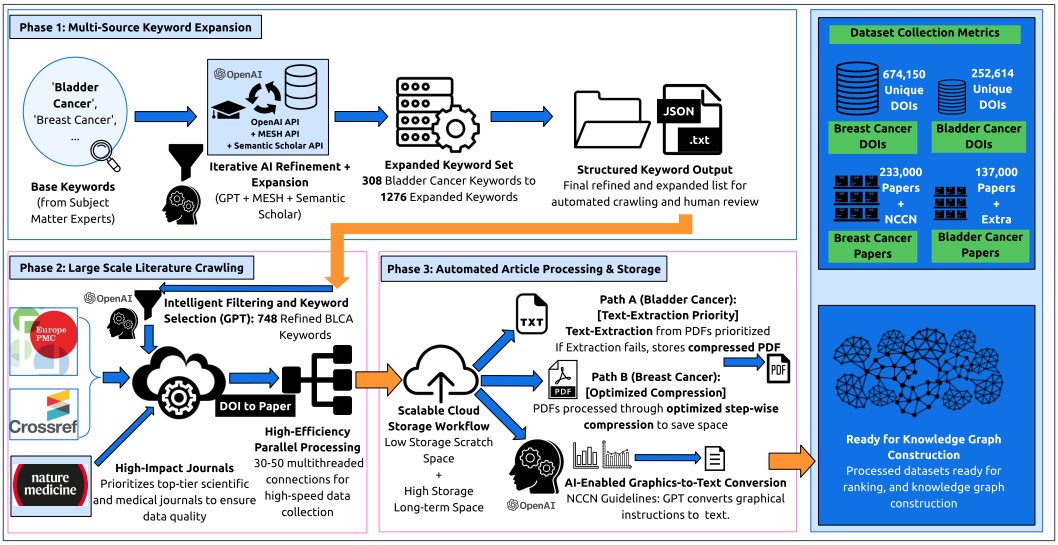

Figure 4: Three-phase literature processing pipeline for KG construction: keyword expansion via OpenAI, MeSH, and Semantic Scholar APIs; large-scale crawling from high-impact sources; and automated text extraction and storage.

A.3 Knowledge Graph Quality Assessment

Quality is assessed via stratified random sampling of 200 triples across relation types, evaluated on three dimensions: factual grounding with respect to the cited evidence span, correctness of entity names and types, and accuracy of the extracted relation. Results are summarized in Table 4.

Table 4: Knowledge graph quality assessment on a stratified random sample of 200 extracted triples.

Metric	Initial	Refined
Factual grounding	99.0%	99.0%
Relation accuracy	100.0%	100.0%
Entity type accuracy (strict)	64.6%	82.5%
Entity type accuracy (lenient)	99.0%	92.5%
Hallucination rate	0.5%	<1.0%

Remaining challenges include entity boundary delineation for complex noun phrases and disambiguation of clinical endpoints, motivating future integration of ontology-constrained classifiers and external resources such as UMLS Bodenreider [2004].

A.4 Path Scoring Metric Derivations

This section provides full derivations of the four path-level metrics introduced in Section 3.1.

A.4.1 Logic Score

Each path P is converted into a natural-language sentence by concatenating its entities and relations in traversal order. Both the query Q and path description are embedded into a shared semantic space. The Logic score is the cosine similarity between query embedding $\mathbf{v}_Q$ and path embedding $\mathbf{v}_P$:

$$S_{\text{logic}} = \text{CosSim}(\mathbf{v}_Q, \mathbf{v}_P) = \frac{\mathbf{v}_Q \cdot \mathbf{v}_P}{\|\mathbf{v}_Q\|\|\mathbf{v}_P\|} \tag{9}$$

Higher values indicate stronger semantic alignment; lower values indicate conceptual gaps or implausible transitions.

A.4.2 Relevance Score

Each entity e_i in path $P = \{e_1, \ldots, e_N\}$ is represented by normalized embedding $\mathbf{e}_i$, and the target domain by centroid $\mathbf{c}_{\text{target}}$. Relevance is the average proximity of path entities to the domain centroid:

$$S_{\text{rel}} = 1 - \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{e}_i - \mathbf{c}_{\text{target}}\|_2 \tag{10}$$

The score is linearly rescaled and clamped to $[0, 1]$ for comparability across paths.

A.4.3 Novelty Score

The Novelty score favors paths containing rare, underexplored entities. An empirical occurrence probability is defined for each entity based on its degree in the global KG:

$$P(e) = \frac{\deg(e) + \epsilon}{\sum_{v \in V} \deg(v) + \epsilon} \tag{11}$$

where ϵ is a small smoothing constant. The Novelty score is the average information content of path entities:

$$S_{\text{nov}} = \frac{1}{|P|}\sum_{e \in P} -\log P(e) \tag{12}$$

Scores are normalized using min-max scaling across candidate paths.

A.4.4 Surprise Score

The Surprise score captures atypical cross-domain entity combinations. Let $P_{\text{actual}}(c)$ denote the entity category distribution within path P and $P_{\text{expected}}(c)$ the global KG category distribution. Surprise is the KL divergence between these distributions:

$$S_{\text{sur}} = \text{KL}(P_{\text{actual}} \,\|\, P_{\text{expected}}) = \sum_c P_{\text{actual}}(c) \log \frac{P_{\text{actual}}(c)}{P_{\text{expected}}(c)} \tag{13}$$

Divergence values are normalized to $[0, 1]$ across all candidate paths. Higher values indicate paths that combine entity categories in unexpected ways.

A.4.5 Weighted Aggregation

The four normalized metrics are combined as:

$$S_{\text{total}} = w_{\text{logic}} S_{\text{logic}} + w_{\text{rel}} S_{\text{rel}} + w_{\text{nov}} S_{\text{nov}} + w_{\text{sur}} S_{\text{sur}}, \quad \sum_i w_i = 1 \tag{14}$$

In our experiments, higher weights are assigned to Novelty and Surprise to emphasize discovery-oriented criteria. Candidate paths are ranked in descending order and the top-ranked paths forwarded to the Ontologist agent.

A.5 Path-Scoring Ablation Study

Experimental setup. To quantify the contribution of each component of the knowledge-graph path-scoring function, we conducted a leave-one-out ablation study of four scoring terms: logic, relevance, novelty, and surprise. The default weights were 0.2, 0.2, 0.3, and 0.3, respectively. For each ablation, we set the corresponding weight to zero while keeping the remaining three weights fixed. We evaluated two complementary outcomes: (i) changes in path selection and ranking, and (ii) changes in the novelty scores of the downstream generated hypotheses.

Path-ranking sensitivity. We evaluated path-ranking changes across 54 bladder cancer research questions; two of the original 56 questions were excluded because source-entity inference failed. For each question and ablation, we compared the selected paths against those obtained using the complete scoring function. We measured the percentage of questions for which the highest-ranked path changed (Top-1 change) and the mean number of shared paths among the top five ranked candidates (Top-5 overlap).

As shown in Table 5, removing surprise changed the highest-ranked path in all 54 questions (100%) and reduced the mean Top-5 overlap to 1.11 out of 5. Removing novelty changed the highest-ranked path in 14.8% of questions, followed by logic (11.1%) and relevance (3.7%). These findings indicate that surprise has the greatest influence on path ranking under the default weighting configuration.

Downstream hypothesis novelty. To determine whether changes in path scoring also affect the generated hypotheses, we repeated the hypothesis-generation pipeline for all 56 bladder cancer research questions under each ablation and the complete scoring function. We compared the resulting hypotheses using SAGE's novelty critic overall novelty score (`overall_novelty_score`, 1–10 scale). For each question, we calculated the score difference between the ablated configuration and the complete model.

All four ablations reduced the mean downstream novelty score. Removing relevance produced the largest mean reduction ($\Delta = -0.27$, $n = 55$), followed by surprise ($\Delta = -0.25$, $n = 56$), novelty ($\Delta = -0.14$, $n = 56$), and logic ($\Delta = -0.06$, $n = 54$).

Notably, path-ranking sensitivity did not directly correspond to downstream novelty-score changes. Although removing relevance changed the top-ranked path in only 3.7% of questions, it produced the largest reduction in downstream novelty score. Conversely, removing surprise changed the top-ranked path in every question but resulted in a smaller novelty-score reduction. These findings suggest that path-ranking sensitivity alone is insufficient to characterize the contribution of individual scoring terms to downstream hypothesis generation.

B Hypothesis Generation Pipeline

B.1 Agent Descriptions and Model Assignments

Table 6 details the role, context source, model assignment, and model-selection rationale for each agent in the hypothesis generation pipeline. The fixed serial architecture assigns each agent only task-relevant context from its immediate predecessor, avoiding the token overhead and contextual noise of shared-memory approaches.

Table 5: Leave-one-out ablation of the four knowledge-graph path-scoring components. Top-1 change and Top-5 overlap are evaluated on 54 research questions. The downstream novelty-score change (Δ) is measured relative to the complete scoring function, with sample sizes indicated in parentheses. Negative values indicate a lower novelty score after removing the corresponding component.

Removed component	**Top-1 change**	**Top-5 overlap**	**Novelty Δ**
Logic	11.1%	4.57/5	-0.06 (54)
Relevance	3.7%	4.70/5	-0.27 (55)
Novelty	14.8%	3.24/5	-0.14 (56)
Surprise	100%	1.11/5	-0.25 (56)

Table 6: Complete agent workflow. Each agent is defined by its context source in the graph-based pipeline (GP), model assignment, role, and model-selection rationale.

Agent	**Context Source**	**Model**	**Role**	**Rationale**
Path Generation	Query only	GPT-5 Nano	Generate paths between KG nodes via graph traversal.	Cheapest model; requires only tool calling with minimal reasoning.
Ontologist	Path Generation	GPT-5 Mini	Define biological relationships and technical terms in generated paths.	Balanced model; domain knowledge required but not maximal reasoning.
Scientist	Ontologist	GPT-5 (high reasoning)	Formulate concise, dataset-aware, empirically testable hypotheses.	Best available model; critical reasoning task determining overall hypothesis quality.
Hypothesis Expansion	Scientist, Ontologist	GPT-5 Mini	Enrich hypothesis with scientific rationale, background, and clinical implications.	Balanced model; competence needed but conciseness valued over maximal reasoning.
Novelty Debate	Hypothesis Expansion	GPT-5 Mini	Stress-test hypothesis novelty through adversarial three-critic debate grounded in real-time literature retrieval.	Balanced model; requires literature search and comparative analysis.
Explainability	Hypothesis Expansion	GPT-5 Mini	Score hypothesis on five interpretability criteria to compute the Explainability Index (EI).	Balanced model; structured scoring task requiring domain awareness.
Dataset Discovery	Hypothesis Expansion	Deterministic	Parse hypothesis into explicit constraints, construct the Artifact Requirement Graph (ARG), and identify the minimal file set satisfying all validation requirements via greedy selection.	Rule-based deterministic algorithm; no LLM required as retrieval is governed by explicit constraint matching and coverage scoring.
Coding	Dataset Discovery, Hypothesis Expansion	GPT-5 (high reasoning)	Synthesize and execute hypothesis-conditioned analysis code on selected patient cohort data within a sandboxed environment.	Best available model; code generation and debugging require strong reasoning and adaptability.
Results Interpreter	Coding, Hypothesis Expansion, Dataset Discovery	GPT-5 (high reasoning + visual)	Interpret statistical outputs, assess validation outcomes, and assign a feasibility score.	Multimodal high-reasoning model; requires quantitative and visual analysis.
Summary	Results Interpreter, Hypothesis Expansion, Novelty Debate	GPT-5 Mini	Synthesize hypothesis, rationale, novelty score, EI, and feasibility score into a structured report.	Balanced model; synthesis task prioritizing clarity and structured reporting.

B.2 Pipeline Architecture and Context Allocation

SAGE uses a specific-memory graph-based pipeline (GP) in which each agent receives only task-relevant context from its immediate predecessor. This contrasts with a shared-memory chat-based pipeline (CP) in which every agent receives the full conversation history. Table 7 shows that GP reduces prompt tokens by 57.5% and completion tokens by 14.9%, cutting inference cost by 25–75% while maintaining comparable hypothesis novelty (7/10 vs. 8/10). These results confirm that targeted context allocation preserves discovery quality while enabling predictable, scalable execution.

Table 7: Comparison of graph-based (GP) and chat-based (CP) pipelines. GP reduces token usage and cost while maintaining comparable hypothesis quality.

Metric	GP	CP	Change
Prompt tokens (SD)	$9{,}002 \pm 2{,}367$	$21{,}210 \pm 3{,}146$	-57.5%
Completion tokens (SD)	$10{,}011 \pm 2{,}909$	$11{,}765 \pm 817$	-14.9%
Runtime sec. (SD)	222 ± 73	227 ± 29	-2.2%
Novelty score	7/10	8/10	-12.5%
Relative cost	$1\times$	1.3–$4\times$	25–75%↓

The rationale for excluding information at each stage is that upstream content is already distilled into the predecessor agent's output, making re-exposure redundant and increasing the risk of context dilution, where long prompts degrade LLM focus and increase hallucination. Table 8 details the context provided and excluded for each agent.

Table 8: Context allocation per agent in the specific-memory pipeline. Each agent receives only the information essential to its function.

Agent	Context Provided	Context Excluded	Rationale
Path Generation	User query	None (pipeline start)	Needs only the query to traverse the KG.
Ontologist	All paths from Path Generation	User query	Paths already encode query intent.
Scientist	Ontologist output	Path Generation output	Ontologist already refines path information.
Hypothesis Expansion	Scientist hypothesis	Ontologist, Path Generation outputs	Scientist hypothesis synthesizes all upstream content.
Novelty Debate	Expanded hypothesis	All upstream outputs	Needs only final hypothesis for literature comparison.
Explainability	Expanded hypothesis	All upstream outputs	Scores hypothesis independently of generation history.
Dataset Discovery	Expanded hypothesis	All upstream outputs	Deterministic algorithm; requires only hypothesis constraints to construct ARG and retrieve matching files.
Coding	Expanded hypothesis, selected file set from Dataset Discovery	All upstream conceptual outputs	Requires hypothesis variables and identified data files to synthesize and execute analysis code.
Results Interpreter	Expanded hypothesis, coding outputs, and selected file set	Upstream conceptual outputs	Requires hypothesis, execution results, and data context for statistical interpretation.
Summary	Results Interpreter, Hypothesis Expansion, Novelty Debate	Coding outputs, upstream conceptual outputs	Needs validation results, final hypothesis, and novelty score only.

B.3 Cross-Domain Computational Scalability

Motivation. Expanding SAGE to additional cancer domains requires constructing disease-specific biomedical knowledge graphs (KGs), which may vary substantially in size depending on the available literature and extracted biological relationships. A potential limitation of knowledge-graph-driven agentic systems is that increasing graph size could lead to higher retrieval overhead, larger language-model contexts, and greater computational cost. We therefore evaluated whether SAGE's per-hypothesis computational requirements increase with KG size across different cancer domains.

Experimental setup. We evaluated three cancer-specific KGs: bladder cancer (41,053 nodes), breast cancer (28,088 nodes), and prostate cancer (21,814 nodes). For each domain, SAGE processed 56 research questions, yielding 168 runs in total. We recorded four computational metrics: execution time, token consumption, inference cost, and the number of LLM calls.

Table 9: Computational scalability of SAGE across three cancer-specific knowledge graphs. Each domain includes 56 research questions. Costs are estimated using GPT-5 mini API pricing.

Cancer	KG nodes	Time (s)	Tokens	Cost ($)
Bladder	41,053	159.3	44,734	0.045
Breast	28,088	142.8	43,044	0.043
Prostate	21,814	169.8	42,747	0.042

Inference costs were calculated using GPT-5 mini API pricing of $0.25 per million input tokens and $2.00 per million output tokens, as used in the rebuttal experiments. We additionally computed Pearson correlation coefficients between KG size and each computational metric to assess whether larger graphs were associated with increased per-hypothesis computational demands.

Results. Table 9 summarizes the results. Despite differences in KG size, mean execution times remained within 142.8–169.8 seconds across the three domains. Mean token consumption was similarly comparable, ranging from 42,747 to 44,734 tokens per hypothesis. Estimated inference costs varied only slightly, from $0.042 to $0.045 per hypothesis, with a total cost of $7.26 across all 168 runs.

Graph size exhibited negligible correlation with execution time (Pearson $r = -0.02$) and the number of LLM calls ($r = -0.05$), and weak positive correlations with token consumption ($r = 0.10$) and inference cost ($r = 0.20$). Notably, prostate cancer, which had the smallest KG, exhibited the highest mean execution time but the lowest mean inference cost. Thus, larger graphs did not consistently result in greater computational demands within the evaluated range.

Interpretation and implications. These findings are consistent with SAGE's targeted retrieval and selective context allocation strategy. Rather than passing the complete knowledge graph to every agent, SAGE retrieves and ranks candidate paths relevant to the research question and provides downstream agents with task-specific information. Consequently, the amount of context processed during hypothesis generation is not directly proportional to the total number of nodes in the underlying graph.

This property is important for extending SAGE to additional cancer domains, where the availability of biomedical literature and the resulting KG size may vary. The results suggest that disease-specific KG expansion need not produce a corresponding increase in per-hypothesis inference cost.

Limitations. The reported costs characterize per-hypothesis inference and do not include the one-time computational expense of literature processing and KG construction, which depends on corpus size. Furthermore, these experiments evaluate computational efficiency rather than cross-domain hypothesis quality, biomarker validity, or clinical generalizability. The results are restricted to the three evaluated cancer domains and KG sizes; additional benchmarking is required to establish scalability at substantially larger graph sizes.

B.4 Architectural Ablations

Motivation. SAGE adopts an evidence-first architecture in which knowledge-graph retrieval precedes hypothesis generation, and specialized agents perform successive reasoning tasks. We investigated whether these architectural decisions improve hypothesis generation compared with alternative workflows. Specifically, we evaluated the contributions of evidence-first reasoning, agent specialization, and pipeline ordering.

Experimental setup. We compared four pipeline configurations:

- **Full SAGE:** The proposed evidence-first architecture, consisting of knowledge-graph retrieval, the Ontologist, Scientist, Hypothesis Expansion, and Novelty Assessment agents.
- **Hypothesis-first:** The Scientist first generates a hypothesis without retrieved evidence. Knowledge-graph retrieval is then conditioned on the initial hypothesis, followed by ontology analysis, hypothesis revision, and novelty assessment.

Table 10: Architectural ablation results across five bladder cancer research questions and five repetitions per configuration. Values are reported as mean ± standard deviation, except for the Vendi Score, which follows the same reporting format.

Configuration	Novelty	Confidence	Consensus	Vendi	Time (s)	Completion tokens
Full SAGE	7.75 ± 0.48	0.729 ± 0.040	8.96 ± 0.68	2.34 ± 0.43	229.06 ± 107.40	$10{,}732 \pm 3{,}151$
Hypothesis-first	7.64 ± 0.74	0.745 ± 0.054	8.76 ± 1.01	1.75 ± 0.25	140.41 ± 15.41	$10{,}461 \pm 896$
Merged-agent	7.85 ± 0.51	0.770 ± 0.076	8.80 ± 0.96	2.04 ± 0.48	96.83 ± 76.56	$8{,}265 \pm 6{,}632$
Reordered	7.65 ± 0.59	0.765 ± 0.069	8.56 ± 1.04	2.36 ± 0.40	156.21 ± 18.83	$10{,}310 \pm 1{,}088$

- **Merged-agent:** The Ontologist, Scientist, and Hypothesis Expansion agents are replaced with a single agent performing all three functions, while preserving the surrounding pipeline.
- **Reordered:** Novelty assessment is performed before Hypothesis Expansion, allowing expansion to use the novelty score and revision guidance.

Rather than simply removing stages, we preserved the expected interfaces and output formats wherever possible to avoid introducing failures caused by incompatible agent inputs. All configurations used the same knowledge graph, embeddings, and research questions.

We evaluated five clinically motivated bladder cancer research questions, with five independent repetitions per configuration, yielding 100 runs. We measured automated novelty, confidence, consensus, hypothesis diversity using the Vendi Score, execution time, and completion tokens. The Vendi Score was computed from concise hypotheses to ensure consistent comparisons across configurations.

Results and interpretation. Full SAGE achieved the highest consensus score (8.96), compared with 8.76 for hypothesis-first, 8.80 for merged-agent, and 8.56 for reordered. Hypothesis-first generation produced the lowest diversity (Vendi Score 1.75 versus 2.34 for Full SAGE), suggesting that generating a hypothesis before retrieving evidence may constrain the range of hypotheses subsequently explored.

Agent specialization introduced a computational trade-off. The merged-agent configuration achieved slightly higher novelty (7.85 versus 7.75) and confidence (0.770 versus 0.729), while reducing mean execution time from 229.06 to 96.83 seconds and completion tokens from 10,732 to 8,265. Thus, the full architecture did not maximize every metric and required additional computation.

Reordering novelty assessment before hypothesis expansion produced diversity comparable to Full SAGE (2.36 versus 2.34), but lower consensus (8.56 versus 8.96).

Overall, these results suggest that evidence-first reasoning and specialized agent interactions provide benefits in consensus and hypothesis diversity, while simpler configurations can offer comparable novelty at substantially lower computational cost. The appropriate architecture therefore depends on the relative importance of hypothesis diversity, assessment consensus, and computational efficiency.

B.5 Cross-Domain Biomarker-Type Adherence

Motivation. A biomarker discovery system should generate hypotheses that match the user's requested scientific objective. In particular, prognostic biomarkers, which are associated with clinical outcomes, must be distinguished from predictive biomarkers, which concern differential response or benefit from a specific treatment. We therefore examined whether SAGE maintains the requested biomarker type across cancer domains.

Experimental setup. We evaluated SAGE across three cancer domains: bladder, breast, and prostate cancer. For each domain, we requested either a prognostic or predictive biomarker across three cancer subtypes and three data modalities. Each configuration was repeated twice, yielding 108 evaluations in total.

For every generated hypothesis, we compared the biomarker type indicated in SAGE's output with the requested type. We report the proportion of hypotheses whose assigned type did not match the request.

Table 11: Biomarker-type mismatch rates across three cancer domains. Each domain and requested-type combination contains 18 evaluations.

Cancer domain	Prognostic mismatch	Predictive mismatch
Bladder	72.2%	16.7%
Breast	66.7%	11.1%
Prostate	22.2%	44.4%

Results. Table 11 shows substantial variation in task adherence across cancer domains. For prognostic requests, mismatch rates were 72.2% in bladder cancer, 66.7% in breast cancer, and 22.2% in prostate cancer. For predictive requests, mismatch rates were 16.7%, 11.1%, and 44.4%, respectively.

These findings indicate that SAGE's ability to return the requested biomarker type depends on the disease context and the requested endpoint.

Error analysis. Mismatch cases frequently involved biomarkers with both prognostic and predictive associations reported in the literature, including tertiary lymphoid structure (TLS)-like aggregates, tumor-infiltrating lymphocyte density, and stromal-to-tumor ratio.

Because SAGE assigns a single biomarker type to each generated hypothesis, such biomarkers may be labeled according to only one of their potential clinical roles. Consequently, some mismatches may reflect incomplete endpoint characterization rather than an inherently incorrect biomarker hypothesis.

Differences in the representation of prognostic and predictive biomarkers across disease-specific literature may also contribute to the observed patterns. However, this explanation remains hypothetical and was not directly tested.

Implications. These results reveal an important limitation in task adherence that is not captured by computational scalability experiments. Before extending SAGE to a new disease domain, its ability to follow biomarker-type and endpoint constraints should be evaluated independently.

Future improvements should include explicit verification of the requested biomarker type and support for hypotheses with both prognostic and predictive interpretations when supported by the underlying evidence.

B.6 Failure Analysis and Robustness to Flawed Inputs

Motivation. Although SAGE uses structured knowledge-graph retrieval and multi-agent novelty assessment, these mechanisms do not guarantee that every generated hypothesis is adequately supported by evidence. We therefore evaluated failure modes related to weak retrieval, insufficient grounding, and incorrect assumptions in user-provided inputs.

Weak retrieval and insufficient grounding. We first examined 108 non-adversarial runs across bladder, breast, and prostate cancer. Weak retrieval was defined as a case in which the Path Generation Agent returned only one knowledge-graph path.

This occurred in 9 of 108 runs (8.3%). Some retrieved paths also contained generic entity descriptions that provided limited biological information. Nevertheless, SAGE could continue generating a detailed hypothesis without explicitly indicating that the underlying evidence was limited.

We additionally examined cases in which a novelty score of at least 7/10 was assigned despite single-path grounding. This occurred in 3 of 108 runs (2.8%).

This operational definition captures a potential mismatch between novelty assessment and evidentiary support; it does not establish that the resulting hypotheses were necessarily biologically incorrect.

Explicitly conflicting literature evidence was not observed in these runs. Therefore, the experiment does not establish SAGE's ability to resolve genuine conflicts between published findings.

Table 12: Responses to 60 deliberately flawed input hypotheses. "Approved" indicates that the pipeline returned an approved hypothesis, whereas "Revised" indicates that the generated hypothesis was changed relative to the input. No input was explicitly rejected as invalid.

Input error	Approved	Revised
Directional inversion	3	9
Fabricated relationship	6	6
Wrong-context transfer	4	8
Confounded attribution	3	9
Self-contradictory premise	5	7
Total	21	39

Robustness to flawed input hypotheses. We next constructed 60 deliberately flawed input hypotheses spanning five error categories: directional inversion, fabricated relationships, wrong-context transfer, confounded attribution, and self-contradictory premises.

Each category contained 12 evaluations, covering three cancer domains, two examples per domain, and two repetitions.

Results and interpretation. As shown in Table 12, 39 of 60 inputs were revised and 21 were approved, while none was explicitly rejected.

In several cases, SAGE modified the original premise, such as changing an incorrect biological direction or replacing a relationship transferred from an inappropriate cancer context. However, the pipeline did not explicitly notify users that the initial premise was flawed or explain why it had been changed.

This behavior presents an important limitation for scientific traceability. Although revision may prevent some incorrect premises from being carried forward, silent changes make it difficult for researchers to distinguish accepted evidence from assumptions that were corrected or replaced.

Furthermore, a revised output should not automatically be interpreted as a successful correction: the system may also redirect a question toward an unrelated hypothesis.

Implications and future improvements. These experiments identify two priorities for improving SAGE's reliability.

First, the system should explicitly report the strength and completeness of retrieved evidence, including when candidate hypotheses are supported by only a single path.

Second, the pipeline should distinguish between accepting, revising, and rejecting input premises, and provide an explanation whenever a premise is modified.

Such mechanisms would improve transparency and allow researchers to review potential errors before committing resources to downstream validation.

B.7 Hypothesis Diversity Evaluation: Graph-Based vs. Text-Based Retrieval

Diversity Metric: Vendi Score. We adopt the Vendi Score Friedman and Dieng [2023] as our primary diversity metric. Given N hypotheses embedded as vectors $\{x_1, \ldots, x_N\}$, the Vendi Score is defined as:

$$\mathrm{VS}(X) = \exp\left(-\sum_{i=1}^{N} \lambda_i \log \lambda_i\right) \tag{15}$$

where $\{\lambda_i\}$ are the eigenvalues of the normalized cosine kernel matrix $\tilde{K} = K/N$, with $K_{ij} = k(x_i, x_j)$ the pairwise cosine similarity. The Vendi Score measures the *effective number of independent semantic modes* in the hypothesis set: a value of 1 indicates all hypotheses are identical, while a value of N indicates they are pairwise orthogonal. Hypotheses forming k tight semantic clusters yield VS $\approx k$.

We prefer the Vendi Score over alternatives such as pairwise cosine distance (PCD), Self-BLEU, or LLM judges for three reasons: (i) it measures the number of distinct semantic centers rather

than average spread, capturing cluster structure that PCD cannot; (ii) it is insensitive to surface lexical variation that inflates Self-BLEU-based metrics; and (iii) it is fully deterministic given fixed hypotheses and embedding model, introducing no LLM sampling variance. Chen et al. Chen et al. [2026] report that the Vendi Score achieves 87% agreement with expert human diversity judgments across 500 pairwise comparisons, the highest among four evaluated metrics. In our experiments, the Vendi Score is computed per prompt on $N = 10$ hypotheses using `text-embedding-3-small` embeddings with L2 normalization, and averaged across all 46 prompts. Eigenvalues are computed via `numpy.linalg.eigvalsh` with a numerical floor of 10^{-12}.

Limitations. The Vendi Score operates on full hypothesis text rather than extracted biological entities, meaning mechanistically similar hypotheses phrased differently may appear more diverse than they are. Both conditions share the same embedding model, ensuring the comparison is controlled. Absolute Vendi values are not directly comparable across embedding models, though relative orderings are robust to embedding choice Chen et al. [2026].

B.8 Novelty Debate: Extended Evaluation

B.8.1 Evaluation Dataset Composition

We constructed a ground-truth dataset of 150 scientific proposals spanning five novelty tiers using time-travel backtesting, evaluating each proposal only against literature available prior to its publication or retraction date. Category A (*Breakthrough*) includes Nobel-winning discoveries and paradigm-shifting methods such as AlphaFold2, CRISPR–Cas9, Transformer, and GAN. Category B (*High Novelty*) covers significant advances extending prior work such as BERT, ViT, and SAM. Category C (*Moderate*) captures meaningful non-paradigmatic contributions. Category D (*Incremental*) includes routine extensions and standard applications. Category E (*Specious*) comprises fabricated or invalid claims including STAP Cells, Arsenic Life, and ICEP2. Dataset composition is summarized in Table 13.

Table 13: Evaluation dataset of 150 proposals across five novelty tiers with ground-truth score ranges and temporal evaluation windows.

Category	n	Ground Truth	Evaluation Window
A: Breakthrough	30	8.5–10.0	2 years pre-publication
B: High Novelty	30	7.0–8.5	1 year pre-publication
C: Moderate	30	5.0–7.0	Publication year
D: Incremental	30	3.5–5.0	Publication year
E: Specious	30	1.0–3.5	Post-retraction
Total	150		

B.8.2 Per-Category Deliberation Statistics

Table 14 reports deliberation frequency, debate depth, and literature retrieval behavior across novelty categories. Deliberation is triggered most frequently for extreme categories (A: 76.7%, E: 83.3%), where initial critic disagreement is highest. Breakthrough proposals retrieve fewer references with minimal contradiction, consistent with genuine gaps in prior literature. Specious claims exhibit high contradiction rates (8.7 ± 2.9 per proposal), validating the system's ability to distinguish absence of evidence from evidence of contradiction.

Table 14: Debate frequency and literature retrieval statistics across five novelty categories.

Category	Debate Triggered	Rounds (Avg.)	Refs Retrieved	Contradictions Found
A: Breakthrough	76.7%	2.43	4.4 ± 1.8	0.3 ± 0.6
B: High Novelty	63.3%	2.11	5.9 ± 2.2	1.2 ± 1.1
C: Moderate	46.7%	1.64	8.0 ± 2.8	3.8 ± 1.8
D: Incremental	40.0%	1.50	9.4 ± 3.1	6.2 ± 2.3
E: Specious	83.3%	2.68	8.5 ± 3.2	8.7 ± 2.9

Table 15: Sensitivity of novelty assessment to the convergence threshold and maximum round budget. Convergence rates and average rounds are calculated over 93 triggered debates. Token consumption is normalized to the default configuration.

Setting	Accuracy	A–E gap	Converged	Avg. rounds	Rel. tokens
Default: $\sigma_{\text{conv}} = 0.50$, $R_{\max} = 3$	91.3%	6.78	91.4%	2.19	1.00×
$\sigma_{\text{conv}} = 0.25$, $R_{\max} = 3$	90.7%	6.83	80.6%	2.61	1.17×
$\sigma_{\text{conv}} = 1.00$, $R_{\max} = 3$	89.3%	6.25	97.8%	1.48	0.76×
$\sigma_{\text{conv}} = 0.50$, $R_{\max} = 1$	86.7%	5.72	21.5%	1.00	0.60×
$\sigma_{\text{conv}} = 0.50$, $R_{\max} = 5$	90.7%	6.86	97.8%	2.32	1.08×

B.8.3 Cross-Domain Validation and Error Analysis

Bias analysis revealed a slight conservative tendency for breakthrough hypotheses (mean signed error $= -0.17$) and a small positive bias for moderately novel hypotheses ($+0.12$). Of 13 misclassifications (8.7% error rate), 10 arose from boundary ambiguity between creative application and incremental extension, with errors concentrated near category boundaries rather than reflecting gross misjudgments.

B.9 Novelty Debate Hyperparameter Sensitivity

Motivation. SAGE's novelty assessment uses an iterative debate mechanism to resolve disagreements between critics. The stopping criteria determine how long the debate continues and may influence both assessment accuracy and computational cost.

We therefore examined the sensitivity of novelty assessment to the convergence threshold and maximum number of debate rounds.

Experimental setup. We used the same 150 historical research proposals evaluated in the main debate ablation experiment. The disagreement trigger threshold was fixed at $\sigma_{\text{trigger}} = 1.0$.

We varied the convergence threshold σ_{conv} and the maximum round budget $R_{\max}$ independently, using the default configuration $\sigma_{\text{conv}} = 0.5$ and $R_{\max} = 3$ as the reference.

Across the 150 proposals, 93 triggered debates. We measured overall classification accuracy, the score difference between breakthrough (A) and specious (E) proposals, the proportion of triggered debates that converged, the average number of rounds, and token consumption relative to default.

Results. The default configuration achieved 91.3% accuracy, an A–E gap of 6.78, and convergence in 91.4% of triggered debates, with an average of 2.19 rounds.

Reducing the convergence threshold to 0.25 increased token consumption by 17% without improving accuracy. Increasing the threshold to 1.0 reduced token consumption by 24%, but also reduced accuracy to 89.3%.

Limiting debates to one round reduced accuracy to 86.7%, with only 21.5% of triggered debates meeting the convergence criterion. Increasing the maximum number of rounds to five did not improve accuracy over the default setting and increased token consumption by 8%.

Interpretation. These results indicate that the default configuration provides a favorable accuracy–cost trade-off among the evaluated settings. Allowing three rounds improves performance compared with a single-round budget, while additional rounds provide no further accuracy improvement in this experiment.

The results also show that debate performance depends on its stopping criteria, emphasizing the importance of reporting these parameters when reproducing the novelty assessment.

C Empirical Hypothesis Validation

C.1 Coding Agent: Extended Evaluation on KramaBench

To assess generalizability beyond pathology-specific hypotheses, we benchmarked the Coding Agent on KramaBench Lai et al. [2026], a public data-science benchmark spanning diverse tabular reasoning tasks. This setting is strictly harder than our internal evaluation because the agent must autonomously locate relevant files before writing and executing code. We compare two conditions across four model backends: a baseline where task-relevant files are pre-attached ($N = 94$ tasks), and a full dataset-discovery workflow where the agent searches a raw dataset root autonomously ($N = 106$ tasks).

Tables 16 and 17 report performance under each condition. Under the oracle-file baseline, execution success rates range from 61.7% to 78.7%, establishing an upper bound for the full pipeline. Adding autonomous discovery reduces execution rates to 29.2%–39.6%, with the dominant failure mode being failure to finalize a valid file set rather than failures in code generation or execution.

Table 16: Coding Agent performance on KramaBench with oracle file selection ($N = 94$ tasks). *Score%*: composite benchmark score. *Exact% (Reports)*: exact-match rate over all tasks. *Exact% (Executed)*: exact-match rate conditioned on executed tasks. *Executed%*: fraction of tasks where the script ran successfully.

Model	Score%	Exact	Exact% (Reports)	Exact% (Executed)	Executed	Exec.%
gpt-5.3-codex	42.18	30/94	31.9%	40.5%	74/94	78.7%
gpt-5.4-2026-03-05	40.79	27/94	28.7%	39.1%	69/94	73.4%
gpt-5.4-mini	32.04	24/94	25.5%	40.0%	60/94	63.8%
gpt-5.4-nano	29.90	19/94	20.2%	32.8%	58/94	61.7%

Table 17: Coding Agent performance on KramaBench with autonomous dataset discovery ($N = 106$ tasks, discovery step budget $= 6$). The drop in *Executed%* relative to Table 16 reflects tasks that fail at the discovery stage and never reach execution.

Model	Score%	Exact	Exact% (Reports)	Exact% (Executed)	Executed	Exec.%
gpt-5.3-codex	18.28	13/106	12.3%	32.5%	40/106	37.7%
gpt-5.4-2026-03-05	15.76	10/106	9.4%	32.3%	31/106	29.2%
gpt-5.4-mini	12.89	10/106	9.4%	23.8%	42/106	39.6%
gpt-5.4-nano	12.96	10/106	9.4%	25.0%	40/106	37.7%

Table 18 decomposes performance across the three pipeline stages. Stronger models (codex, gpt-5.4-2026-03-05) fail primarily at the discovery stage, while lighter models (gpt-5.4-mini) advance further through discovery but incur higher execution-retry costs. This confirms that improving dataset discovery is the primary lever for advancing end-to-end validation performance.

Table 18: Three-stage task attrition under autonomous dataset discovery ($N = 106$ tasks). Each row traces task throughput across file selection, execution, and exact match stages.

Model	Selected	Sel.%	Exec.	Exec.%	Exact	Ex.%	Failure
gpt-5.3-codex	42/106	39.6%	40/106	37.7%	13/106	12.3%	No files
gpt-5.4-2026-03-05	33/106	31.1%	31/106	29.2%	10/106	9.4%	No files
gpt-5.4-mini	67/106	63.2%	42/106	39.6%	10/106	9.4%	Script fail
gpt-5.4-nano	52/106	49.1%	40/106	37.7%	10/106	9.4%	No files

C.2 Tool Registry and Orchestrator

The tool orchestration subsystem provides a unified control plane for interacting with heterogeneous clinical and computational tools through a plugin architecture. Each tool is registered through a

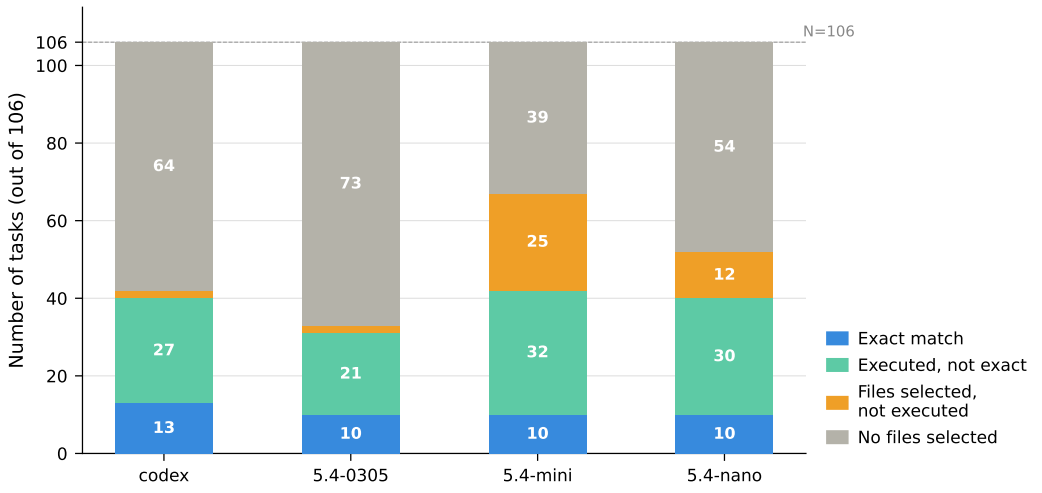

Figure 5: **Task attrition across three pipeline stages on KramaBench ($N = 106$ tasks per model).** Each bar shows task distribution across four outcomes: exact match, executed but not exact, files selected but not executed, and no files selected. The dominant bottleneck for stronger models is dataset discovery rather than code generation.

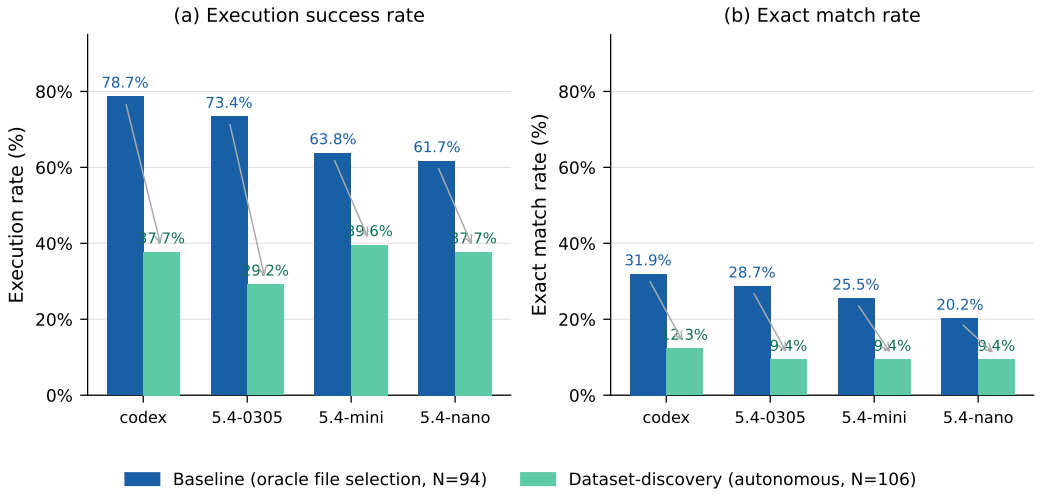

Figure 6: **Impact of dataset discovery on execution and exact-match rates across four model backends.** Dark blue: oracle file selection baseline ($N = 94$). Teal: autonomous discovery ($N = 106$). Execution rates fall by 24–44 percentage points under discovery, confirming that code generation is performant when supplied with correct inputs.

structured specification file parsed into three progressively richer layers: a lightweight index for fast routing, an operational specification for planning, and a full description loaded on demand for execution. This layered design keeps the system token-efficient while scaling across a large tool catalog, and separates tool discovery, selection, planning, and execution into distinct concerns. New tools can be integrated without any backend code changes, making the subsystem fully modular and extensible.

The orchestrator interprets prompts from a user or autonomous agent, dynamically selects appropriate tools, and coordinates their execution over shared patient-level data and whole-slide images. Supported tool suites cover medical oncology workflows (survival analysis and treatment selection), histopathology pipelines (tumor, mitosis, tubule, collagen, and TIL segmentation, spatial transcriptomics, multiple instance learning, and foundation model inference), and emerging multimodal

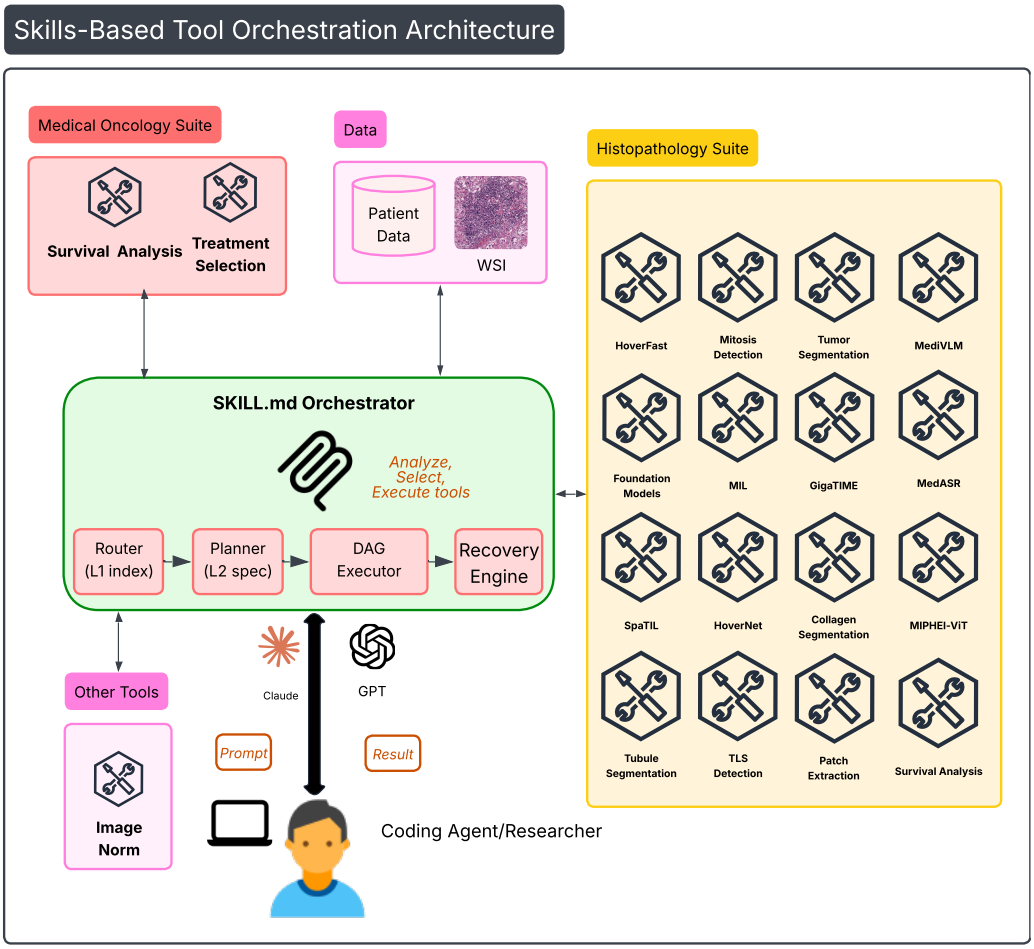

Figure 7: **Tool orchestration subsystem architecture.** The Orchestrator mediates between a coding agent or researcher and distributed domain-specific tool suites by analyzing prompts, selecting appropriate capabilities, and coordinating execution over shared patient data and whole-slide images.

capabilities. Session-based state management persists tool outputs across analysis steps, enabling multi-step investigations without requiring users to re-specify intermediate results. A built-in tile server supports in-browser visualization of extracted masks and multi-channel overlays.

Figure 7 provides an overview of the subsystem architecture. Figures 8 and 9 illustrate an example workflow in which a whole-slide image is uploaded, a tumor detection request is issued, and the subsystem invokes the appropriate segmentation tool to return a binary tumor mask followed by region-aware patch extraction across tumor, peritumoral, and remaining tissue regions.

C.3 Comparison with Scientific Discovery Agents

Motivation. To complement comparisons with general-purpose language models and multi-agent baselines, we evaluated SAGE against two scientific-discovery frameworks: AI Co-Scientist and SciAgents. These comparisons examine whether SAGE's evidence-first, multi-path reasoning pipeline provides advantages over alternative scientific hypothesis-generation approaches.

Baseline implementations. AI Co-Scientist was evaluated through its available API. Each question required approximately six hours to generate 10 hypotheses.

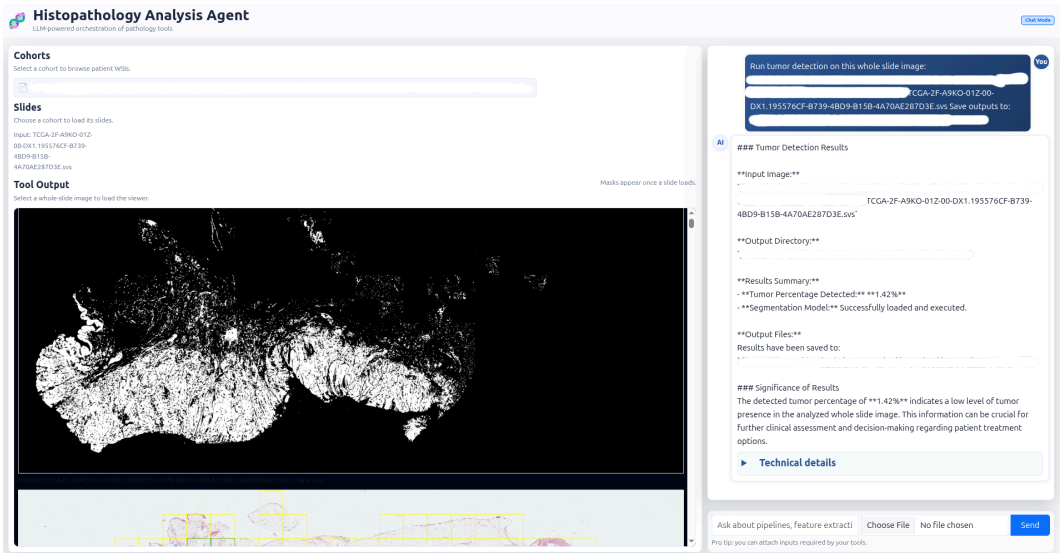

Figure 8: **Tumor detection via tool orchestration.** The subsystem responds to a tumor detection request by invoking the tumor segmentation tool and returning a binary mask highlighting detected tumor regions.

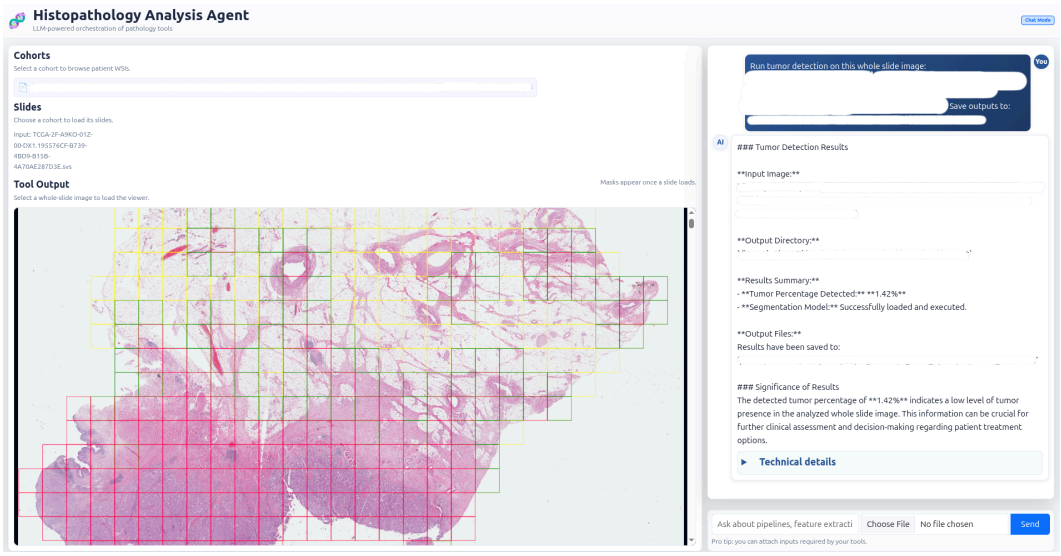

Figure 9: **Region-aware patch extraction via tool orchestration.** Following tumor segmentation, the subsystem composes segmentation and spatial reasoning tools to overlay color-coded patches corresponding to tumor (red), peritumoral (green), and remaining tissue (yellow) regions.

For SciAgents, we constructed a domain-adapted version of the original framework. The original implementation was developed for materials science and relied on human participation, dynamic agent selection, and outputs not directly aligned with the present evaluation.

We retained its core scientific reasoning mechanism, including stochastic single-path knowledge-graph retrieval, ontology-mediated reasoning, and multi-component hypothesis generation. We adapted the domain instructions to bladder cancer and introduced automated agent selection and termination to enable consistent evaluation.

The resulting comparison therefore uses a domain-adapted SciAgents implementation rather than the unmodified original system.

Table 19: Comparison of SAGE with scientific discovery agents. Values are mean ± standard deviation. AI Co-Scientist was evaluated on two questions, while Legacy SciAgents and SAGE were evaluated on five questions. Execution times and token statistics are reported as recorded for each system and are not necessarily based on identical execution units.

System	Novelty	Confidence	Consensus	Vendi	Execution time	Output cost measure
AI Co-Scientist	7.47 ± 0.83	0.743 ± 0.053	8.60 ± 1.27	5.90 ± 0.20	$\sim$6 h / 10 hypotheses	2,500 AI credits / 10 hypotheses
Legacy SciAgents	7.75 ± 0.52	0.754 ± 0.071	8.72 ± 1.06	2.13 ± 0.12	163.24 ± 19.30 s	$13{,}292 \pm 1{,}054$ tokens
SAGE (Full)	7.75 ± 0.48	0.729 ± 0.040	8.96 ± 0.68	2.34 ± 0.43	229.06 ± 107.40 s	$10{,}732 \pm 3{,}151$ tokens

Evaluation protocol. Full SAGE and Legacy SciAgents were evaluated on five clinically motivated bladder cancer questions, with five repetitions per question, yielding 25 runs per configuration.

Due to AI Co-Scientist's longer execution time, its evaluation covered two of the five questions, with ten repetitions per question, yielding 20 runs.

We compared automated novelty, confidence, consensus, hypothesis diversity, execution time, and token consumption where available.

Results and interpretation. As shown in Table 19, SAGE achieved a mean novelty score of 7.75, comparable to Legacy SciAgents (7.75) and slightly higher than AI Co-Scientist (7.47). SAGE also achieved the highest consensus score (8.96), compared with 8.72 for Legacy SciAgents and 8.60 for AI Co-Scientist.

However, the methods exhibited different trade-offs. AI Co-Scientist achieved substantially higher hypothesis diversity, with a Vendi Score of 5.90 compared with 2.34 for SAGE. Legacy SciAgents required less execution time than SAGE (163.24 versus 229.06 seconds on average), although SAGE used fewer completion tokens (10,732 versus 13,292).

These results suggest that SAGE provides competitive automated novelty and consensus relative to the evaluated scientific-discovery baselines, while differences in diversity and computational requirements remain.

Limitations of the comparison. The evaluations are not perfectly matched. AI Co-Scientist was tested on fewer unique questions, and its computational resource measure is expressed in platform-specific credits rather than directly comparable API tokens or monetary cost.

Moreover, the SciAgents baseline required domain adaptation and orchestration changes to support automated evaluation. These differences limit direct conclusions about relative efficiency and general superiority across systems.

Accordingly, this experiment should be interpreted as a complementary comparison, rather than a definitive ranking of scientific-discovery frameworks.

C.4 Expert Evaluation Protocol

Expert panel composition. Four domain specialists independently evaluated hypotheses generated by each system: two board-certified urologists with experience in bladder cancer management and translational research, one computational pathologist specializing in whole-slide image analysis, and one molecular biologist with expertise in tumor immunology. Each expert was presented with hypotheses from their primary domain of expertise only, ensuring that scoring reflected genuine domain knowledge rather than generalist judgment.

Scoring rubric.

Scoring rubric. Each hypothesis was evaluated using two criteria, *Novelty* and *Feasibility*, on a 1–10 ordinal scale, where scores of 1–4 indicate weak performance, 5–7 indicate moderate performance, and 8–10 indicate strong performance.

Novelty measures the originality of the hypothesis. Weak scores correspond to hypotheses based on well-known mechanisms or simple reformulations of existing ideas. Moderate scores indicate partial

novelty or limited new insight. Strong scores indicate clearly new biological or clinical insights with meaningful conceptual innovation.

Feasibility measures how practical the hypothesis is to validate experimentally or computationally. Weak scores correspond to hypotheses with no clear validation pathway or unrealistic requirements. Moderate scores indicate hypotheses that are testable but may require substantial effort, resources, or additional tool development. Strong scores indicate hypotheses that can be readily tested using currently available datasets, models, or experimental protocols.

Evaluation procedure. Experts were blinded to the system identity that generated each hypothesis.

D Experimental Results

D.1 CXCL13 as a Prognostic Biomarker in Muscle-Invasive Bladder Cancer

Generated Hypothesis. In muscle-invasive bladder cancer (MIBC), higher CXCL13 expression is associated with improved progression-free survival (PFS) and overall survival (OS). This hypothesis is grounded in the biological role of CXCL13 as a key mediator of tertiary lymphoid structure (TLS) formation: CXCL13 recruits B cells and T follicular helper cells, promoting immune niches that enhance antigen presentation and $CD8^+$ T cell activation, contributing to favorable clinical outcomes and immune checkpoint blockade response.

Validation Dataset and Design. The hypothesis was validated on the TCGA-BLCA cohort ($N = 404$ patients, stage II–IV). Patients were stratified into CXCL13-high and CXCL13-low groups using the 75th percentile expression threshold, selected as the cutoff yielding statistically significant separation for both OS and PFS. Kaplan–Meier survival analysis with log-rank tests and multivariable Cox proportional hazards models adjusting for age, sex, and stage were used for validation.

Results. Kaplan–Meier analysis demonstrated that CXCL13-high patients had significantly improved OS (log-rank $p = 0.0059$) and PFS (log-rank $p = 0.035$) compared to CXCL13-low patients (Figure 10). In multivariable Cox models, CXCL13 remained independently associated with reduced hazard after adjustment for clinical covariates: OS hazard ratio 0.60 (95% CI 0.42–0.85, $p = 0.005$, C-index = 0.66) and PFS hazard ratio 0.58 (95% CI 0.40–0.82, $p = 0.002$, C-index = 0.64) (Figure 11). Stage-stratified analysis revealed that the prognostic effect was strongest in Stage III patients (OS: $p = 0.015$, HR = 0.45; PFS: $p = 0.035$, HR = 0.48), while Stage II and Stage IV patients showed no significant associations (Figure 12).

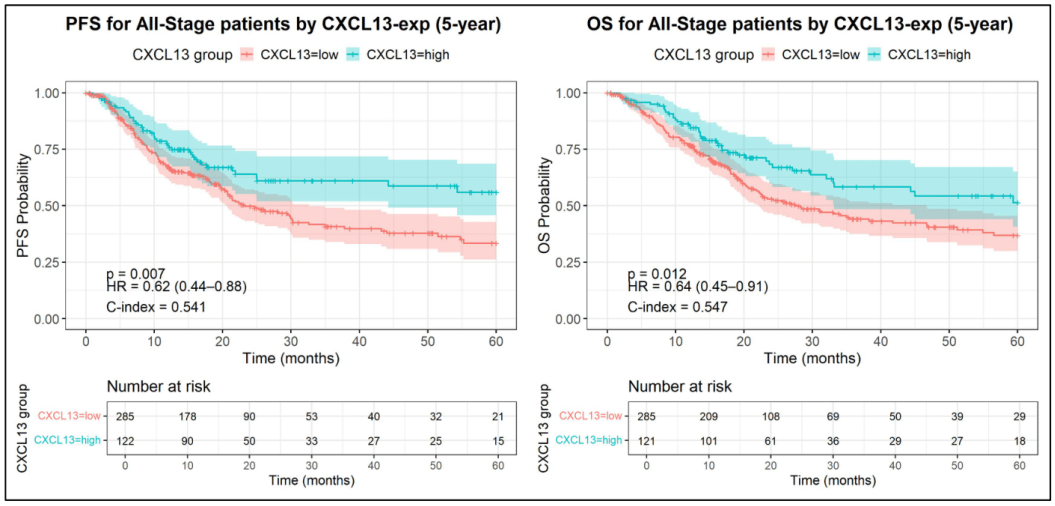

Figure 10: Kaplan–Meier survival curves for PFS (left) and OS (right) in TCGA-BLCA stratified by CXCL13 expression. CXCL13-high patients exhibit significantly longer survival under both endpoints.

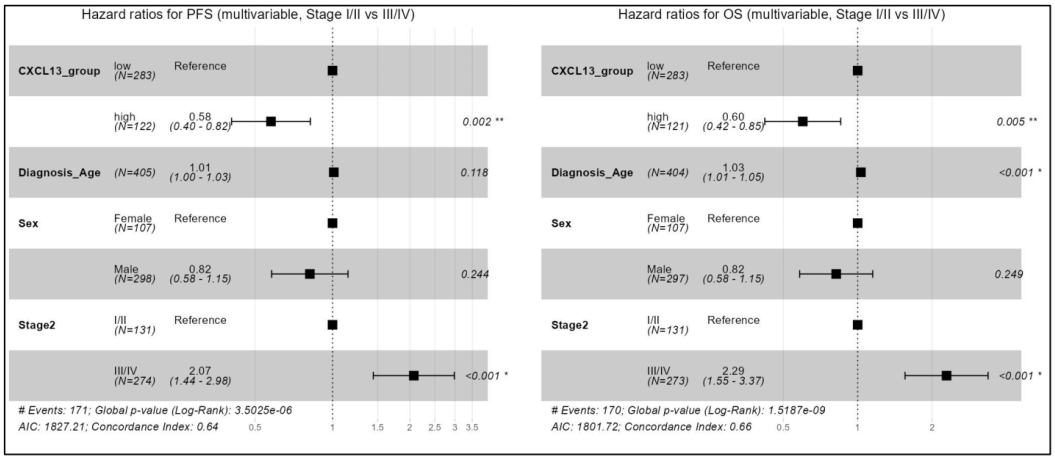

Figure 11: Forest plots from multivariable Cox models for PFS and OS, adjusting for age, sex, and stage. CXCL13 remains independently associated with reduced hazard of progression and death.

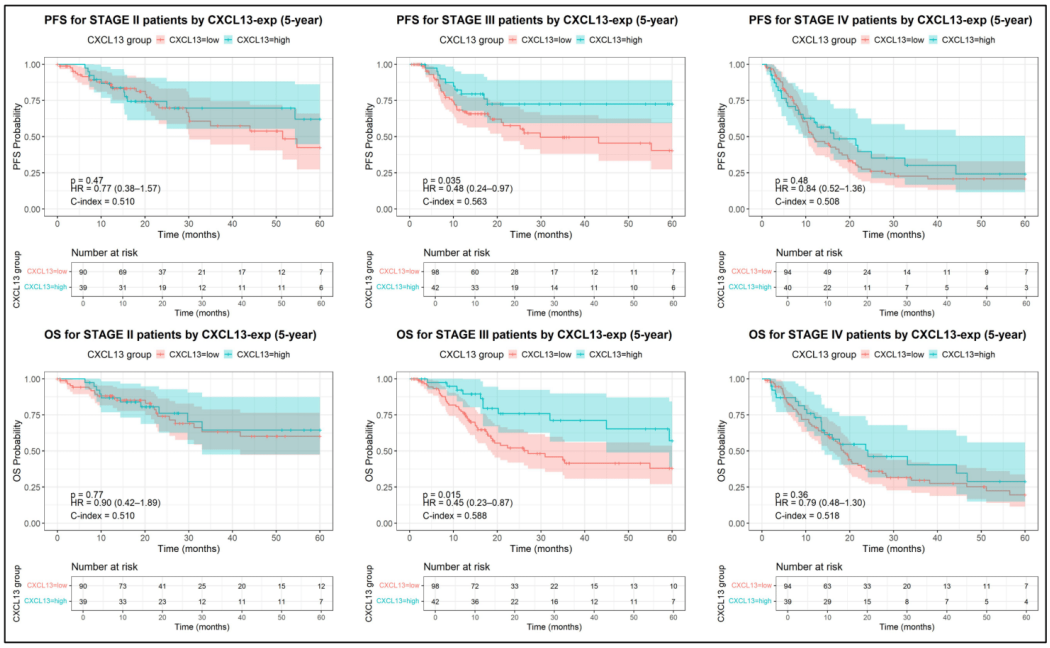

Figure 12: Stage-stratified Kaplan–Meier curves for PFS (top) and OS (bottom) across Stage II, III, and IV patients. The prognostic effect of CXCL13 is most pronounced in Stage III disease.

D.2 LAG3 Tumor Density as a Predictive Biomarker of Treatment Response in Bladder Cancer

Generated Hypothesis. LAG3 tumor density, quantified from whole-slide images as an imaging-derived feature, is associated with response to immune checkpoint blockade in bladder cancer. This hypothesis is grounded in the established role of LAG3 as an inhibitory immune checkpoint receptor expressed on exhausted T cells: lower LAG3 tumor density in the tumor microenvironment is expected to reflect reduced immune exhaustion and greater susceptibility to checkpoint inhibitor therapy.

Validation Dataset and Design. The hypothesis was validated on an institutional bladder cancer cohort ($N = 57$ patients) with paired LAG3 tumor density measurements and treatment response

labels (response_label: 0 = non-responder, 1 = responder). Two endpoints were evaluated: (i) a predictive endpoint assessing association between LAG3 tumor density and treatment response using a Mann–Whitney U test, and (ii) a prognostic endpoint evaluating overall survival (OS) stratified by LAG3 tumor density via Kaplan–Meier analysis on training ($N = 46$) and holdout ($N = 11$) sets.

Results. **Predictive analysis.** LAG3 tumor density was significantly lower in responders (mean $= 0.00215$, median $= 0.00142$) compared to non-responders (mean $= 0.00567$, median $= 0.00432$), Mann–Whitney U $p = 0.0008$ (Figure 13), providing strong evidence supporting LAG3 tumor density as a predictive biomarker of treatment response.

Prognostic analysis. Survival analysis did not reveal statistically significant separation of OS by LAG3 tumor density in either the training set (log-rank $p = 0.462$) or the holdout set (log-rank $p = 0.111$). The holdout set is small ($N = 11$, 10 events), making estimates unstable and the null result difficult to interpret definitively.

Interpretation. This example illustrates an important distinction that SAGE explicitly handles: predictive and prognostic biomarker claims are evaluated separately. LAG3 tumor density shows a strong association with treatment response but does not demonstrate independent prognostic value for overall survival in this cohort. This nuanced outcome reflects the transparent validation design of SAGE, which reports both supported and unsupported claims rather than selectively presenting positive results.

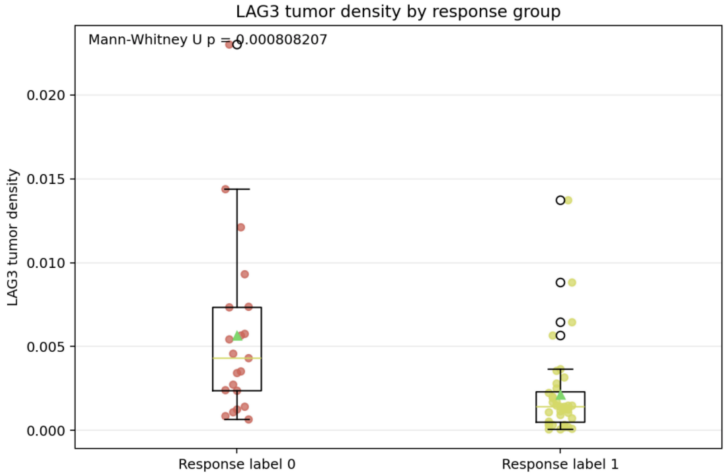

Figure 13: **LAG3 tumor density by treatment response group.** LAG3 tumor density is significantly lower in responders (response label 1) compared to non-responders (response label 0), Mann–Whitney U $p = 0.0008$. Each point represents a patient; triangles indicate group means.

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 clearly state all four contributions: end-to-end biomarker discovery pipeline, multi-path ontological reasoning, debate-based novelty assessment, and biologically grounded validation. Each claim is directly supported by experimental results in the Results section.

   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 the Conclusion

   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: The paper does not include theoretical results or formal proofs. All contributions are algorithmic and empirical. The path scoring metrics and ARG-based retrieval algorithm are defined formally but do not require proofs.

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: The full pipeline architecture, agent roles, model assignments, and context allocation strategy are described in the Methods section and Appendix. Path scoring metrics are formally defined in Section on path generation and the corresponding Appendix. The novelty debate configuration, including disagreement threshold ($\sigma > 1.0$), convergence criterion ($\sigma < 0.5$), and maximum debate rounds (3), is specified in the Section on novelty agent, and the 150-proposal evaluation dataset is documented with temporal evaluation windows in Appendix. Survival analyses are conducted on TCGA-BLCA, a publicly available dataset accessible via cBioPortal, with all preprocessing steps and statistical methods specified. The KramaBench coding agent evaluation uses a public benchmark with model backends, discovery step budget (6), and execution retry budget (5) fully reported. The expert evaluation rubric and scoring procedure are provided in Appendix. Code is provided in the supplementary material.

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: An anonymized code repository is provided in the supplementary material, containing the full SAGE pipeline implementation. The main validation experiments use TCGA-BLCA, publicly accessible via cBioPortal with no registration required. The KramaBench evaluation uses a public benchmark available at its original repository. The institutional cohort used in the LAG3 validation is available upon reasonable request to the corresponding institution. Instructions for data access, environment setup, and experiment reproduction are included in the supplementary material alongside the anonymized repository link.

   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: SAGE requires no model training, or optimizers as all agents use pre-trained LLMs via API. The computational pathology tools used in the validation subsystem are locked pre-trained models requiring no hyperparameter tuning, as described in Appendix. Agent model assignments and context sources are specified in Appendix. Path scoring weights and confidence thresholds are defined in the main text. Novelty debate hyperparameters, including disagreement threshold ($\sigma > 1.0$), convergence criterion ($\sigma < 0.5$), and maximum debate rounds (3), are stated in the main text. For the main case study in the results Section, all 412 available TCGA-BLCA patients were used in evaluation mode.

For the additional validated hypotheses in Appendix, dataset split details are specified per experiment. Coding agent execution budgets are reported, and expert evaluation criteria and scoring procedure are detailed in 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: All tables report means with 95% confidence intervals and statistical significance where applicable. Survival analyses report hazard ratios with 95% CIs and log-rank p-values. The novelty debate reports accuracy with significance markers ($p < 0.05$). Diversity results include bootstrap CIs and permutation test p-values.

   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: All LLM-based experiments are conducted via commercial API calls (OpenAI, Google, and Claude), requiring no local GPU resources. Per-hypothesis running time and token costs are reported in the results Section (Tables) and Appendix. Computational pathology validation tools and survival analyses run on a single NVIDIA A100 GPU. No additional compute beyond what is reported here and in the paper was required.

   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: All experiments use de-identified public datasets or data collected under existing institutional oversight. Expert evaluators participated voluntarily and are co-authors of this work. The paper preserves anonymity in submission. No aspects of the research deviate from 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: Positive impacts include accelerating biologically interpretable biomarker discovery in underexplored cancers, reducing reliance on expert intuition, and providing a domain-agnostic reasoning pipeline that can be extended to other scientific domains by supplying the appropriate domain-specific tool suite, as discussed in Section 5 (Conclusion). SAGE is designed as a research tool; any biomarker it discovers requires independent validation across multiple cohorts before clinical use, which is the responsibility of the end user. A potential negative impact is that biomarker discovery may not generalize equitably across demographic groups if the underlying knowledge graph reflects historical disparities in research coverage.

    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: SAGE poses no significant misuse risk. It does not release pre-trained models or scraped datasets. The knowledge graph is constructed from publicly available peer-reviewed scientific literature, and all LLM interactions use commercial APIs governed by their respective providers' usage policies.

    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 datasets, models, and tools are properly cited with their original publications. TCGA-BLCA is accessed via cBioPortal under its open-access data use policy. KramaBench is used under its original license. The SAGE codebase and knowledge graph are released under the CC BY-ND 4.0 license, which permits sharing with attribution while prohibiting modifications without permission.

    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: Two new assets are introduced: the SAGE knowledge graph (41,053 nodes, 56,338 edges) documented in the main text and appendix, and the 150-proposal novelty evaluation dataset documented in Appendix. Both are released under CC BY-ND 4.0 alongside an anonymized code repository provided in the supplementary material. The expert evaluation questions per domain, scoring rubric, and model-generated hypothesis outputs across all system configurations are also included in the supplementary material as additional assets.

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: [Yes].

Justification: Human evaluation was conducted by four domain expert co-authors (two urologists, one pathologist, one molecular biologist). The full scoring rubric, evaluation instructions, and domain-specific questions are provided in the Appendix and the supplementary material. No crowdsourcing platforms were used. Experts participated as co-authors and were not separately compensated.

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: [Yes]

Justification: All patient data used in validation experiments are either de-identified public datasets (TCGA-BLCA) or institutional datasets held in the institutional repository, all of which are covered by existing IRB approvals at the contributing institutions. Expert co-authors evaluated computational hypotheses only and were not exposed to identifiable patient data, posing no risk to human subjects. No new patient data collection was performed for this study.

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: LLMs are a core and original component of SAGE, used across all pipeline stages including hypothesis generation, novelty debate, explainability scoring, code generation, and results interpretation. Model assignments per agent are fully described in Appendix. LLMs were also used for writing assistance in preparing this manuscript, but all scientific content, results, and conclusions were verified by the authors.

    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.