

# CELLSCIENTIST: AUTONOMOUS BIOLOGICAL DISCOVERY VIA SCIENTIFIC ALIGNMENT AND MULTIFACETED BIOLOGICAL PRIORS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recent advances in large language models (LLMs) have catalyzed growing interest in autonomous AI-driven scientific discovery, yet existing frameworks face a fundamental limitation when applied to knowledge-intensive domains like cellular biology: they either produce scientifically invalid hypotheses that violate experimental constraints and physical laws, or generate methodologically sound but trivial findings that lack biological significance. To address this dual challenge, we observe that rigorous autonomous discovery requires both strict enforcement of validity constraints throughout the research process and systematic injection of domain knowledge to guide hypothesis search toward scientifically meaningful questions. We present CellScientist, an autonomous agent framework that implements this insight through a dual-grounding paradigm. First, Scientific Semantic Alignment enforces formal consistency among research questions, experimental protocols, statistical estimands, and biochemical plausibility, rejecting internally contradictory proposals before resource expenditure. Second, Comprehensive Biological Prior Injection integrates structured ontologies, automated literature retrieval, toolchain execution proficiencies, and empirical research heuristics across multiple decision points, steering the search toward biologically interpretable discoveries. Together, these mechanisms enable CellScientist to autonomously formulate research questions, design and execute multi-stage experiments, critically evaluate evidence through falsification tests, and synthesize findings into coherent scientific manuscripts. Evaluated across 20 single-cell and spatial omics research topics, CellScientist improves composite manuscript quality by 15–19% over the strongest general-purpose autonomous research agent and by 29–37% over prior AI Scientist systems, with the largest gains in reproducibility (+17–30%) and data-conclusion traceability (+24–35%). Running end to end without human intervention, CellScientist also recovered an interpretable composition-aware result in post-influenza lung: conditioning on cellular composition reduced unique stage-associated variance from **72.7% to 0.80%**, suggesting that much of the apparent transcriptional recovery reflects changes in cell-type composition rather than within-lineage recovery.

## 1 INTRODUCTION

The rapid advancement of large language models (LLMs) in recent years has inspired researchers worldwide to envision the transformative potential of AI-augmented scientific discovery (Zhang et al., 2024). From accelerating hypothesis generation to automating data analysis, AI systems are increasingly demonstrating their capacity to enhance human scientific inquiry. This progress has emboldened a growing community to pursue an even more ambitious goal: enabling AI to conduct research autonomously, mirroring the creativity and rigor of human scientists.

Early explorations into autonomous AI-driven research have yielded promising results across diverse domains. *AlphaEvolve* (Novikov et al., 2025) leveraged LLMs as operators within evolutionary algorithms, successfully discovering novel heuristic optimization strategies that outperform



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

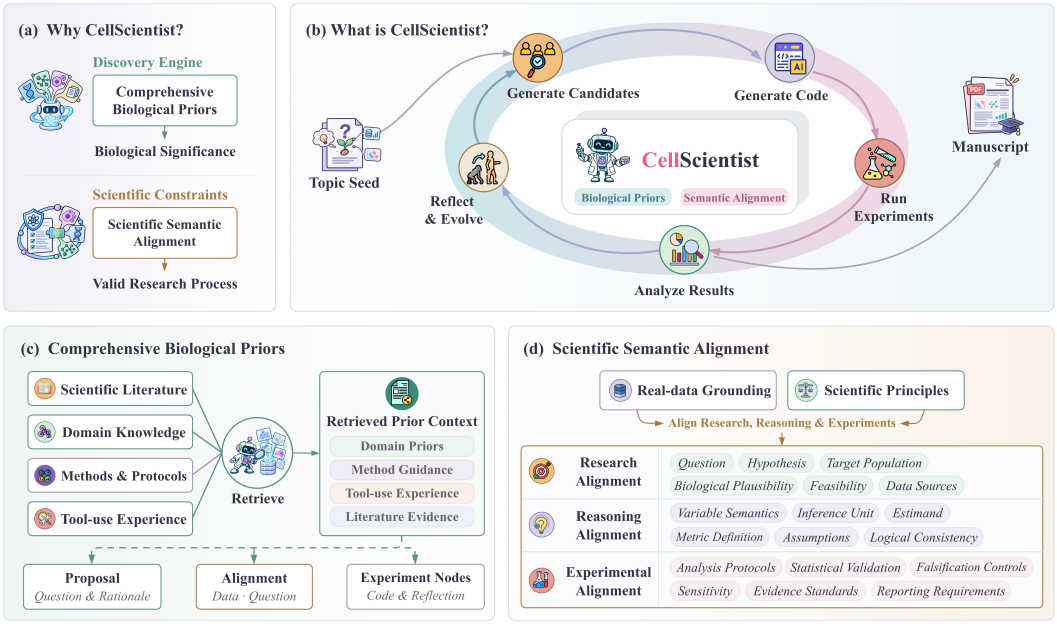


Figure 1: **CellScientist framework and core mechanisms.** The system integrates scientific semantic alignment (research constraint) with multifaceted biological priors (discovery engine) to enable autonomous, verifiable biological discovery. Alignment enforces consistency across research question formulation, experimental design, and evidence interpretation, while biological priors from domain knowledge, literature, tool proficiencies, and historical experience guide scientifically meaningful hypothesis generation and validation.

hand-designed methods. *AlphaFold* (Jumper et al., 2021) revolutionized structural biology by predicting protein structures with unprecedented accuracy, demonstrating AI's capacity to solve long-standing scientific challenges. Similarly, *FunSearch* (Romera-Paredes et al., 2024) employed LLMs to discover new mathematical constructs and combinatorial algorithms, expanding the frontiers of computational mathematics. These pioneering works collectively illustrate AI's potential to explore uncharted territories in science.

However, despite these successes, existing AI-driven discovery systems remain constrained in critical ways. Most approaches excel in narrowly defined, computation-intensive tasks—such as structure prediction or algorithm optimization—but lack the holistic autonomy required to emulate the full spectrum of scientific inquiry. They cannot independently formulate meaningful research questions, design and execute complex multi-stage experiments, critically evaluate evidence, and synthesize findings into coherent scientific narratives. In essence, they function as powerful tools rather than autonomous scientists.

The recent introduction of the *AI Scientist* framework (Lu et al., 2026) marks a paradigm shift in this landscape. AI Scientist pioneers a fully autonomous research pipeline capable of transforming a nascent idea into a complete scientific manuscript—encompassing hypothesis formulation, experimental design, execution, and manuscript generation. This comprehensive autonomy represents a significant leap forward, opening the door to scalable, end-to-end AI-driven research. Yet, this advancement also raises a natural and pressing question: **Can such an autonomous AI scientist conduct rigorous, impactful research in mature, knowledge-intensive scientific domains?**

This question is particularly salient in fields like cellular biology, a cornerstone of modern AI for science (AI4Science) research, where foundation models and LLM-derived representations are increasingly being developed for cellular data analysis (Theodoris et al., 2023; Cui et al., 2024; 2025; Baek et al., 2025; Chen & Zou, 2025). Cellular biology is characterized by intricate, multi-scale phenomena governed by complex regulatory networks, stochastic processes, and context-dependent mechanisms. Conducting meaningful research in this domain demands not only technical proficiency but also deep domain expertise, rigorous adherence to experimental constraints, and the abil-

ity to generate biologically interpretable insights—capabilities that remain challenging for current general-purpose LLMs (Zhang et al., 2026). As illustrated in Figure 1a, an autonomous AI scientist operating in such a domain must address two fundamental challenges:

**(1) Can an autonomous AI scientist formulate scientifically valuable research questions?** Research questions define the trajectory and impact of scientific inquiry. In mature fields, trivial or poorly motivated questions lead to incremental findings with limited novelty or relevance. An autonomous system must possess the contextual awareness and domain grounding necessary to identify open problems that are both feasible and scientifically significant.

**(2) Can such a system conduct methodologically sound research processes?** Invalid experimental designs, violations of physical or biochemical constraints, and improper controls can render findings spurious or irreproducible. An autonomous scientist must rigorously enforce scientific norms throughout the research lifecycle—from hypothesis generation to evidence synthesis—to ensure that discoveries are verifiable, meaningful, and aligned with domain principles.

These requirements are complementary: domain knowledge motivates important questions, whereas semantic rigor ensures that their estimands, controls, and evidence support the intended claims.

To address these challenges, we introduce CellScientist, an autonomous agent framework specifically designed for rigorous, verifiable scientific discovery in cellular biology. CellScientist is built upon a **dual-grounding paradigm** that synergistically integrates two core mechanisms:

- **Scientific Semantic Alignment**: A principled constraint system that formally governs the research process, enforcing alignment between research objectives, experimental designs, and biological plausibility. This mechanism ensures that every hypothesis, reasoning step, and experimental procedure adheres to domain-specific constraints—ranging from physical laws and biochemical feasibility to experimental protocols and statistical validity.
- **Comprehensive Biological Prior Injection**: A multifaceted prior system that infuses domain knowledge into the agent's reasoning at multiple stages. By integrating structured ontologies, automated literature retrieval, toolchain execution proficiencies, and empirical research heuristics, CellScientist navigates vast search spaces with domain-grounded intuition, prioritizing hypotheses and experiments that possess true scientific and biological significance.

Guided by this dual-grounding paradigm, CellScientist autonomously formulates research questions, designs and executes multi-stage experimental pipelines, critically evaluates evidence through falsification testing, and synthesizes findings into coherent scientific manuscripts. Extensive evaluations across diverse cellular analysis and discovery workflows demonstrate that CellScientist consistently produces actionable, verifiable, and novel biological discoveries, advancing the frontier of autonomous AI-driven science in knowledge-intensive domains.

## 2 METHOD

### 2.1 OVERVIEW

CellScientist addresses autonomous scientific discovery through a dual-grounding paradigm: (1) **Scientific Semantic Alignment**, which enforces consistency between research objectives, experimental designs, and biological plausibility, thereby filtering proposals that are scientifically inconsistent or experimentally infeasible, and (2) **Comprehensive Biological Prior Injection**, which integrates domain ontologies, literature evidence, and toolchain proficiencies to guide biologically meaningful discoveries within a substantially narrowed and better-informed search space.

Figure 1 illustrates these complementary mechanisms. Given a research topic, the system establishes feasibility constraints from available data, generates proposals grounded in retrieved priors, validates them against alignment criteria, and iteratively refines experiments through best-first search with scientific reflection. This interaction allows prior knowledge to guide exploration while alignment constraints continuously screen the resulting research decisions. Evidence packages are assembled with explicit traceability to experimental artifacts for manuscript generation (Figure 2).

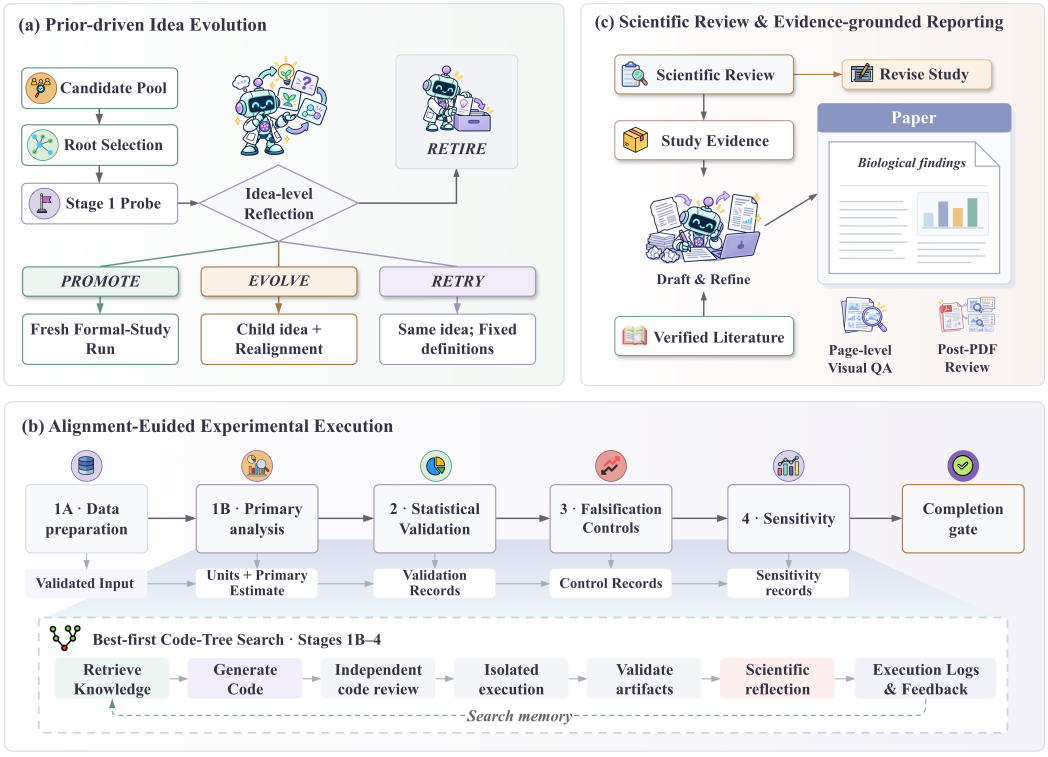


Figure 2: **Autonomous research workflow.** The system evolves research hypotheses through probe-based exploration, executes alignment-constrained formal experiments across four stages (data preparation, primary analysis, resampling validation, falsification tests), and generates evidence-grounded manuscripts with full traceability to experimental artifacts and real data.

### 2.2 Scientific Semantic Alignment

Scientific semantic alignment ensures generated research workflows remain scientifically valid and experimentally realizable. CellScientist enforces explicit alignment between the research question, experimental protocol, primary estimand, and decision criteria.

Before generating proposals, the system compiles a feasibility report $\mathcal{F}_0$ by inspecting data sources—recording schema information, missing patterns, and metadata constraints without loading full datasets. Each research proposal is then formalized as an alignment specification:

$$\mathcal{A} = (Q, H, U, P, G, \theta, \mu, \delta, V, F), \tag{1}$$

where $Q$ is the research question, $H$ the hypothesis, $U$ the inference unit, $P$ the target population, $G$ the comparison groups, $\theta$ the primary estimand, $\mu$ the primary metric with explicit formula, $\delta$ the decision rules, $V$ the validation strategy, and $F$ the mandatory falsification tests.

This specification enforces alignment across three levels: *research alignment* ensures question-data compatibility; *inference alignment* maintains consistency between variable definitions, inference units, and estimands; *experimental alignment* validates that protocols support prespecified evidence requirements. The validator enforces cross-field consistency through schema validation and symbolic reasoning, rejecting internally contradictory proposals before consuming experimental resources.

### 2.3 Comprehensive Biological Prior Injection

Biological priors guide the system toward discoveries of genuine scientific significance. CellScientist integrates four complementary knowledge sources: domain priors encoding common pitfalls, canonical methods covering standard workflows, literature evidence from high-impact publications,

and tool execution proficiencies documenting adapter interfaces and parameter constraints. Its literature grounding builds on retrieval-augmented scientific assistants that connect generated answers to primary sources (Lála et al., 2023).

All knowledge sources are normalized into a unified representation $\mathcal{K}_i = (\text{id}, \tau, \text{tier}, \text{title}, \text{summary}, \mathbf{k}, \mathbf{f}, s, m)$, where $\tau$ indicates the knowledge type and tier encodes priority. Following the retrieval-augmented generation paradigm (Lewis et al., 2020), relevant priors for a research proposal text $q$ are retrieved through semantic embedding:

$$\mathcal{K}_t = \text{TopK}\left(\text{Rank}_{\text{cosine}}\left(\phi(q), \{\phi(e) : e \in \mathcal{K}\}\right)\right), \tag{2}$$

where $\phi(\cdot)$ is a local sentence transformer (bge-m3) mapping text to normalized vectors. Retrieval computes cosine similarities between query and pre-computed entry embeddings, selecting top-K entries per knowledge type with tier and facet filtering.

Retrieved priors are injected at multiple decision points: during proposal generation to establish search context, during experiment planning to constrain parameter configurations, and during manuscript writing to provide citation context.

### 2.4 Best-First Experiment Search with Reflection

The discovery engine combines best-first tree search with scientific reflection to iteratively refine experimental implementations while maintaining semantic alignment. This design builds on language-agent paradigms that interleave reasoning and action, search over deliberate reasoning paths, and improve through verbal self-feedback (Yao et al., 2023b;a; Shinn et al., 2023; Madaan et al., 2023). Each experiment node $v_i = (p_i, c_i, e_i, \phi_i, F_i, r_i)$ contains the analysis plan, executable code, execution trace, implementation utility, output artifacts, and structured reflection.

Figure 2 illustrates the workflow. The search operates through four analysis stages: Stage 1 constructs the analysis cohort and computes the primary estimand; Stage 2 executes resampling validation; Stage 3 performs falsification tests; Stage 4 conducts sensitivity analyses. Each stage has an artifact contract, and downstream stages access frozen upstream results through a run-scoped catalog.

After execution, each node undergoes structured reflection $r_i = \text{Reflect}(p_i, c_i, e_i, F_i, \mathcal{A}, \mathcal{R}_{i-1})$ examining: (1) data-design concerns including label leakage and shortcuts; (2) metric semantics; (3) statistical validity; and (4) falsification adequacy. Reflections generate scientific interpretations (propagated as compressed context) and blocking anomalies (tracked via obligation ledger).

A two-tier memory system maintains context efficiency: short-term memory retains complete information from the most recent node, while long-term memory incrementally compresses earlier nodes into a fixed-budget summary. The reflection mechanism prevents redundant exploration by propagating insights about failed strategies, steering subsequent generations away from identified pitfalls through adaptive pruning.

### 2.5 Evidence Compilation and Manuscript Generation

Manuscript generation is enabled when all four analysis stages complete successfully or a prespecified terminal outcome is documented. The evidence package contains: the alignment specification $\mathcal{A}$, data preparation audit from $\mathcal{F}_0$, stage execution records with artifact inventories, the obligation ledger, deterministic figure scripts, and recomputation entry points. This package, together with retrieved literature citations, enables evidence-grounded authoring where claims are directly traceable to experimental outputs. The final archive records the complete idea lineage, tool invocations, code executions with artifacts, environment metadata, and the terminal outcome, providing full auditability and reproducibility.

## 3 Experiments

### 3.1 Experimental Setup

**Task and Data.** We evaluate end-to-end research generation on 20 single-cell and spatial omics topics spanning lung repair, cancer immunology, and immune-cell dynamics, each paired with real-

Table 1: **Comparative results.** Mean scores over 60 papers per system; Std denotes standard deviation. Bold and underline mark the best and second-best values in each column.

| System | Nov. | Snd. | Clar. | Sig. | Rep. | Des. | Tra. | Comp. | Std |
|---|---|---|---|---|---|---|---|---|---|
| AI Scientist-Gemini | 4.13 | 4.17 | 6.68 | 3.93 | 5.13 | 4.95 | 4.97 | 4.85 | 1.05 |
| AI Scientist-GPT | 3.58 | 5.13 | 6.12 | 3.10 | 5.15 | 5.77 | 4.23 | 4.73 | 1.18 |
| OpenScientist | 4.13 | 5.67 | 6.82 | <u>4.80</u> | 4.42 | 6.40 | 5.73 | 5.42 | 0.81 |
| CellScientist-GPT | <u>4.27</u> | **6.35** | **7.87** | 3.87 | <u>6.05</u> | **7.52** | **7.75** | <u>6.24</u> | 0.89 |
| CellScientist-Gemini | **5.27** | <u>6.13</u> | <u>7.70</u> | **5.28** | **6.72** | <u>7.03</u> | <u>7.13</u> | **6.47** | 1.20 |

world data. Appendix A reports the complete dataset registry, provenance, runtime objects, and inferential units. Given only a topic seed, each system formulates a hypothesis, designs and runs an analysis, produces figures, and writes a manuscript.

**Systems.** We compare CellScientist with three baseline configurations: OpenScientist (Roberts et al., 2026) with GPT-5.6 Sol, and AI Scientist v1 (Lu et al., 2026) with Gemini 3.8 Flash or GPT-5.6 Sol. We evaluate CellScientist with both backbones. Each configuration generates three papers per topic, yielding 60 papers per system and 300 in total.

**Evaluation.** Claude Opus 5 scores all manuscripts on seven 1–10 dimensions: Novelty, Soundness, Clarity, Significance, Reproducibility, Research Design Coherence, and Data-Conclusion Traceability. Their unweighted mean gives the composite score. Evaluation is automated, blind to system identity, and includes every generated paper. Appendix B provides the full protocol, and Appendix C gives the topic descriptions.

### 3.2 Comparative Results

**Overall Performance.** Both CellScientist variants lead in composite score (6.47 with Gemini and 6.24 with GPT), exceeding OpenScientist by 0.81–1.05 points and AI Scientist by 1.39–1.74 points. The largest margins occur in Reproducibility (+1.57 to +2.30) and Traceability (+1.40 to +3.52), followed by Design Coherence (7.03–7.52 vs. 4.95–6.40). The GPT variant trails OpenScientist in Significance (3.87 vs. 4.80), whereas the Gemini variant leads in both Novelty (5.27) and Significance (5.28).

**Where the Gains Arise.** Compared with OpenScientist, Clarity rises by 0.88–1.05 points, but larger gains in Reproducibility (+1.63 to +2.30) and Traceability (+1.40 to +2.02) show that the main advantage is stronger links between evidence and claims. GPT favors Soundness and Design Coherence, whereas Gemini favors Novelty, Significance, and Reproducibility, indicating that dual grounding transfers across backbones while preserving distinct strengths.

### 3.3 Visual Analysis

**Backbone Performance Is Topic Dependent.** Figure 3a compares GPT-minus-Gemini score gaps across all topic pairs. On the matched-topic diagonal, Gemini leads on 12 topics and GPT on 8, showing that neither backbone is uniformly superior. The off-diagonal cells form coherent blocks rather than unstructured variation, suggesting that the relative advantage may reflect backbone preferences for particular topics. This observation raises a more specific question: do semantically related topics favor the same backbone?

**Semantic Analysis Reveals Local Preference Structure.** To test this hypothesis, Figure 3b projects the matched-topic gaps onto a principal component analysis of topic-description embeddings constructed without evaluation scores. T01–T03 favor GPT, whereas topics in T05–T08 and T11–T15 often favor Gemini, revealing local regions of shared preference. However, neither same-preference neighborhood agreement ($p = 0.275$) nor Moran's $I$ on the continuous gap ($p = 0.291$) detects global clustering in the original 1024-dimensional space. Backbone preference is therefore locally structured but does not divide the topics into broad semantic partitions.

**Pareto Analysis Balances Quality and Reliability.** Given this topic sensitivity, a single average quality score is insufficient for comparing methods: a system may generate strong papers when

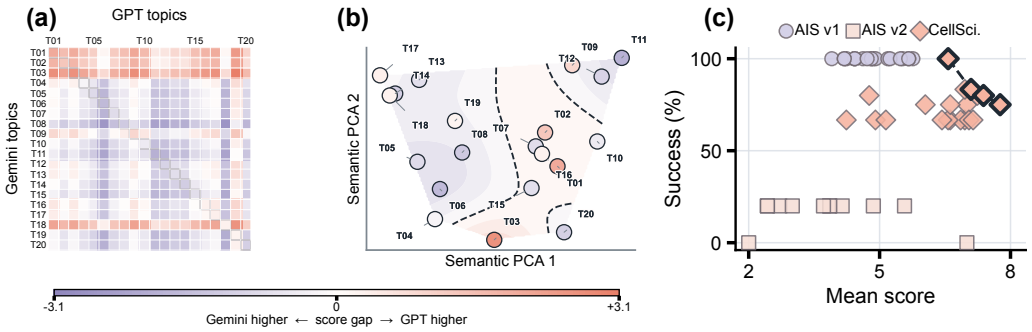


Figure 3: **Backbone variation and the quality-reliability trade-off.** (a) GPT-minus-Gemini Top-3 mean score gaps across topics; diagonal cells compare matched topics. (b) Matched-topic gaps projected by principal component analysis (PCA) of topic embeddings; the dashed curve marks zero interpolated gap. (c) Mean score among successful papers versus run success rate; the dashed line and bold markers indicate Pareto-optimal points.

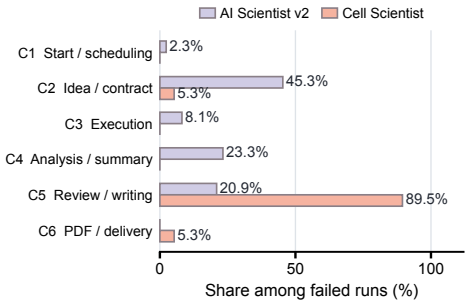


Figure 4: **Primary failure stages.** Conditional distribution of the first blocking failure among failed Gemini 3.8 Flash runs.

**Failures Shift Downstream.** Figure 4 shows a clear redistribution of the first blocking failure. Among failed Gemini runs, AI Scientist v2 fails most often during idea and research-contract formation (C2; 45.3%), followed by analysis and summary (C4; 23.3%) and review and writing (C5; 20.9%). Its failures therefore span early formulation, execution, and analysis stages rather than accumulating at a single terminal step. In contrast, CellScientist records no first blocking failures in C1, C3, or C4: 89.5% occur in C5, and 5.3% each in C2 and C6. Most failed CellScientist runs thus pass idea formation, execution, and analysis before encountering a delivery-stage obstacle. The remaining C5 failures arise from figure rendering and LaTeX compilation, localizing the main reliability bottleneck to manuscript production rather than upstream scientific planning.

successful yet fail frequently end to end. Figure 3c therefore evaluates conditional paper quality together with run success rate. All four Pareto-optimal points belong to CellScientist: T17 scores 6.57 at 100% success, while T20, T11, and T08 score 7.10, 7.38, and 7.76 at success rates of 83.3%, 80.0%, and 75.0%. T17 dominates every AI Scientist v1 topic at the same 100% success rate, whereas AI Scientist v2 averages 10.4% success and contributes no Pareto-optimal point. Quality is computed over successful runs, and AI Scientist v2 produced scorable papers for 12 of 20 topics. Together, the three panels progress from identifying topic-dependent backbone performance, to testing its semantic structure, to comparing methods under the resulting quality–reliability trade-off.

### 3.4 Case Study

**Post-influenza Lung.** In 123,154 cells from 25 donors through day 366, temporal stage explains 72.7% of expression variance. Conditioning on myeloid composition in isometric log-ratio coordinates reduces unique stage variance to 0.80%, a 98.9% attenuation ($F(3, 18) = 29.66$, $p = 3.5 \times 10^{-7}$) reproduced in all 1,000 donor bootstraps. The homeostatic program correlates with alveolar macrophage fraction ($r = 0.919$), whereas housekeeping controls show no stage association ($p = 0.98$).

**Renal Cortex and Medulla.** Counterfactual tissue reconstitution over 46,482 spots from eight donors attributes 92.09% of regional variance to lineage-intrinsic remodeling and 7.91% to com-

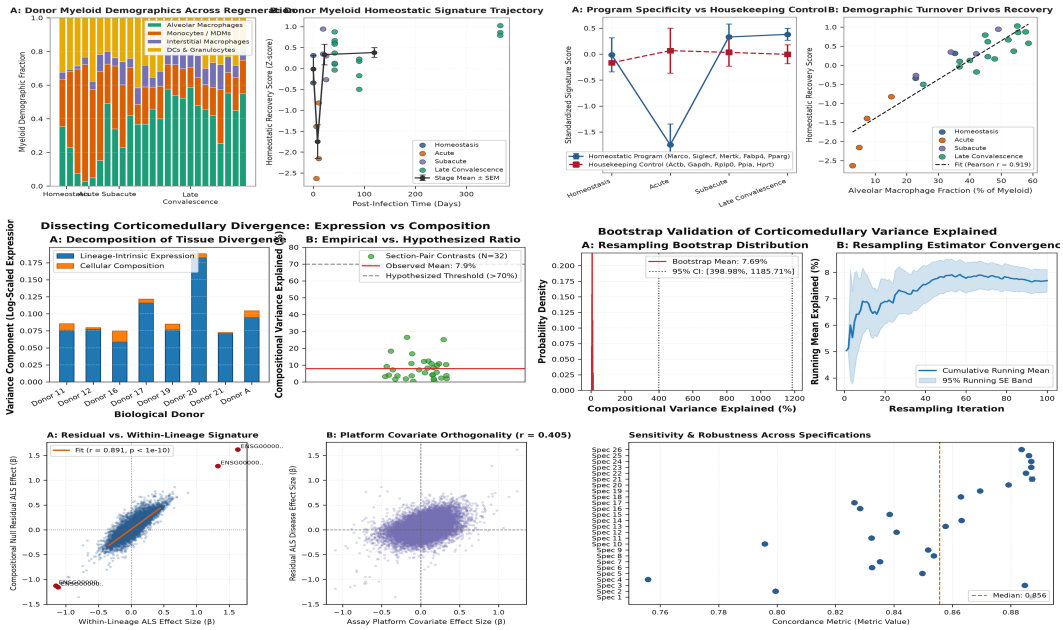


Figure 5: **Case studies.** (a) Composition-conditioned variance analysis in post-influenza lung. (b) Counterfactual decomposition of regional variance in human kidney. (c) Concordance between tissue-level residuals and within-population changes in ALS spinal cord.

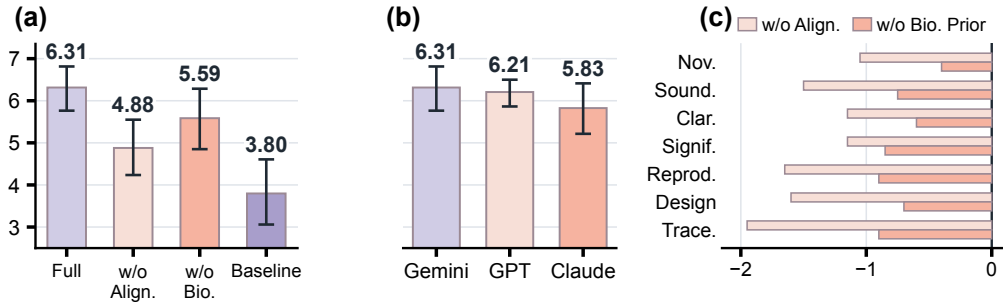


Figure 6: **Ablation study.** (a) Component ablations and the AI Scientist v2 baseline. (b) Backbone comparison. (c) Per-dimension effects of removing semantic alignment or biological priors.

position (95% CI [3.99%, 11.86%]). Negative-control scrambling and within-donor permutations preserve this result.

**ALS Spinal Cord.** Generative compositional null modeling over 62,711 nuclei from 12 donors separates cell-intrinsic dysregulation from motor-neuron loss. Tissue-level residuals agree with within-population alterations (Spearman $\rho = 0.8869$, $p < 10^{-300}$) across 26 sensitivity analyses, including leave-one-donor-out validation and lineage ablations.

### 3.5 Ablation Study

**Both Components Improve Quality.** Removing semantic alignment lowers the composite score from 6.31 to 4.88, while removing biological priors lowers it to 5.59 (Figure 6a). AI Scientist v2 (Yamada et al., 2025) scores 3.80 and completes 12 of 100 attempts.

**The Components Affect Different Dimensions.** Removing alignment produces the largest drops in Traceability ($-1.95$), Reproducibility ($-1.65$), and Design Coherence ($-1.60$). Removing bi-

ological priors has smaller but broader effects on Reproducibility and Traceability (both $-0.90$), Significance ($-0.85$), and Soundness ($-0.75$). Clarity remains above 6.50 in both ablations.

**Backbone Sensitivity.** Gemini and GPT obtain similar composite scores (6.31 and 6.21), while Claude scores 5.83 (Figure 6b). Gemini is strongest in Novelty and Significance; all three backbones exceed the 3.80 baseline.

## 4 Related Work

Early AI-driven scientific discovery focused on solving specific computational problems within well-defined domains. *AlphaFold* (Jumper et al., 2021) revolutionized protein structure prediction through deep learning on structural databases, with accessible implementations such as ColabFold further streamlining structure-prediction workflows (Kim et al., 2025). *FunSearch* (Romera-Paredes et al., 2024) discovered novel mathematical constructs by leveraging LLMs for combinatorial search, while a growing line of work has used LLM-guided evolutionary search to design and refine heuristic algorithms (Liu et al., 2024; Yao et al., 2025; Ye et al., 2024). *AlphaEvolve* (Novikov et al., 2025) scaled this paradigm to scientific and algorithmic discovery, and *GNoME* (Merchant et al., 2023) accelerated materials discovery by predicting stable crystal structures. In chemistry, Coscientist automated experimental planning and execution, while ChemCrow augmented LLMs with expert chemical tools (Boiko et al., 2023; Bran et al., 2024). Together, these systems demonstrate AI's capacity to advance scientific frontiers, but their scope remains centered on particular domains or task families rather than the complete, general scientific lifecycle.

A fundamentally different paradigm emerged with the *AI Scientist* framework (Lu et al., 2026), which pioneered fully autonomous, end-to-end research workflows—from idea generation through experimental execution to manuscript drafting. Its successor, AI Scientist v2, introduced agentic tree search to extend automated discovery to broader computational settings (Yamada et al., 2025), while OpenScientist explored an open agentic co-scientist for biomedical discovery (Roberts et al., 2026). More recently, Robin demonstrated multi-agent automation of scientific discovery across hypothesis generation, experimentation, and validation (Ghareeb et al., 2026). However, general-purpose autonomous scientists face critical limitations when applied to knowledge-intensive natural science domains: they lack the deep domain expertise necessary to navigate complex biological contexts and exhibit high failure rates when experimental designs require strict adherence to biochemical constraints, standardized protocols, and domain-specific plausibility checks. This is precisely the challenge CellScientist addresses through its dual-grounding paradigm of scientific semantic alignment and comprehensive biological prior injection.

Complementary work has strengthened scientific language processing and evidence grounding. Domain-pretrained models such as SciBERT improve representations of scientific text (Beltagy et al., 2019), while retrieval-augmented agents such as PaperQA ground scientific question answering in the literature (Lála et al., 2023). These advances provide important components for literature-aware agents, but do not by themselves enforce consistency among biological hypotheses, executable analyses, statistical estimands, and final claims.

## 5 Conclusion

We present CellScientist, an autonomous agent framework that combines Scientific Semantic Alignment with Comprehensive Biological Prior Injection to keep research questions, analyses, and claims consistent while guiding the search toward meaningful biological problems. Across 20 single-cell RNA-sequencing topics, CellScientist improves composite manuscript quality by 15–19% over the strongest general-purpose baseline and by 29–37% over prior AI Scientist systems, with the largest gains in reproducibility and data-conclusion traceability. The case studies further show that the framework can recover interpretable, falsifiable biological findings rather than only producing complete manuscripts.

The current evaluation is limited to computational analyses of existing single-cell data. Extending the system to additional modalities and prospectively validating its discoveries with human experts and laboratory experiments are important next steps.

## AI USE STATEMENT

Generative AI systems are central to the research evaluated in this paper: CellScientist and the baseline systems use large language models to formulate research questions, generate and execute analysis code, interpret results, and draft manuscripts. During preparation of this paper, generative AI tools were also used for language editing, LaTeX formatting, and code assistance. The authors reviewed the AI-assisted text, citations, code, analyses, and figures included in the paper and take responsibility for the final content and conclusions.

## REFERENCES

William E Allen, Timothy R Blosser, Zuri A Sullivan, Catherine Dulac, and Xiaowei Zhuang. Molecular and spatial signatures of mouse brain aging at single-cell resolution. *Cell*, 186(1): 194–208, 2023.

Seungbyn Baek, Kyungwoo Song, and Insuk Lee. Single-cell foundation models: bringing artificial intelligence into cell biology. *Experimental & Molecular Medicine*, 57(10):2169–2181, 2025.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3615–3620. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1371.

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

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from chatgpt. *Nature biomedical engineering*, 9(4):483–493, 2025.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: Toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Haotian Cui, Alejandro Tejada-Lapuerta, Maria Brbić, Julio Saez-Rodriguez, Simona Cristea, Hani Goodarzi, Mohammad Lotfollahi, Fabian J Theis, and Bo Wang. Towards multimodal foundation models in molecular cell biology. *Nature*, 640(8059):623–633, 2025.

Ali E. Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Dmytro Shved, Gavin J. Gyimesi, Jon M. Laurent, Samantha M. Wright, Muhammed T. Razzak, Andrew D. White, Silvia C. Finnemann, Michaela M. Hinks, and Samuel G. Rodriques. A multi-agent system for automating scientific discovery. *Nature*, 655:497–505, 2026. doi: 10.1038/s41586-026-10652-y.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Gyuri Kim, Sewon Lee, Eli Levy Karin, Hyunbin Kim, Yoshitaka Moriwaki, Sergey Ovchinnikov, Martin Steinegger, and Milot Mirdita. Easy and accurate protein structure prediction using colabfold. *Nature Protocols*, 20(3):620–642, 2025.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

Tammy CT Lan, Alison Kochersberger, Ruth Raichur, Sophia Szady, Hien Tran, Radiana Simeonova, Ashley Helmuth, Andrew Minagar, Maria José Orozco Fuentes, Vipin Kumar, et al. Aging disrupts spatiotemporal coordination in the cycling murine ovary. *Nature Aging*, 6(6):1244–1266, 2026.

Michael Lattke, Wee Leng Tan, Salil Kalarikkal Sukumaran, Kagistia Hana Utami, Marcos Sintes, Srinivasan Sakthivel, Jonathan Tan, Auriel Lim, Vibhavari Aysha Bansal, Katerina Rekopoulou, et al. Single-cell atlas of the developing down syndrome brain cortex. *Nature medicine*, 32(3): 1061–1072, 2026.

Jusung Lee, Ji Yun Jeong, Mi Jeong Hong, Yoon Ha Choi, Ju Young Kim, Jang Hyuck Lee, Jin Eun Choi, Moonsik Kim, Young Woo Do, Eung Bae Lee, et al. Defining the cellular and molecular identities of histologic subtypes in lung adenocarcinoma. *Experimental Hematology & Oncology*, 15(1):12, 2026.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Daniel Y Li, Soumya Kundu, Paul Cheng, Wenduo Gu, Matthew D Worssam, William R Jackson, Quanyi Zhao, Trieu Nguyen, Amelia M Yu, João P Monteiro, et al. Vascular smooth muscle cell state trajectories mediate molecular mechanisms of coronary disease risk. *Nature Communications*, 17(1):4059, 2026.

Fei Liu, Xialiang Tong, Mingxuan Yuan, Xi Lin, Fu Luo, Zhenkun Wang, Zhichao Lu, and Qingfu Zhang. Evolution of heuristics: Towards efficient automatic algorithm design using large language model. *arXiv preprint arXiv:2401.02051*, 2024.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jamie L Marshall, Teia Noel, Qingbo S Wang, Haiqi Chen, Evan Murray, Ayshwarya Subramanian, Katherine A Vernon, Silvana Bazua-Valenti, Katie Liguori, Keith Keller, et al. High-resolution slide-seqv2 spatial transcriptomics enables discovery of disease-specific cell neighborhoods and pathways. *Iscience*, 25(4), 2022.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

Yoshiharu Muto, Parker C Wilson, Nicolas Ledru, Haojia Wu, Henrik Dimke, Sushrut S Waikar, and Benjamin D Humphreys. Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nature communications*, 12(1):2190, 2021.

Terren K Niethamer, Joseph D Planer, Michael P Morley, Apoorva Babu, Gan Zhao, Maria C Basil, Edward Cantu, David B Frank, Joshua M Diamond, Ana N Nottingham, et al. Longitudinal single-cell profiles of lung regeneration after viral infection reveal persistent injury-associated cell states. *Cell Stem Cell*, 32(2):302–321, 2025.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Richard K Perez, M Grace Gordon, Meena Subramaniam, Min Cheol Kim, George C Hartoularos, Sasha Targ, Yang Sun, Anton Ogorodnikov, Raymund Bueno, Andrew Lu, et al. Single-cell rna-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, 376(6589): eabf1970, 2022.

Chengxiang Qiu, Beth K Martin, Ian C Welsh, Riza M Daza, Truc-Mai Le, Xingfan Huang, Eva K Nichols, Megan L Taylor, Olivia Fulton, Diana R O'Day, et al. A single-cell time-lapse of mouse prenatal development from gastrula to birth. *Nature*, 626(8001):1084–1093, 2024.

Kaleigh F Roberts, Zachary B Abrams, Luca Cappelletti, Mahdi Moqri, Nicholas Heugel, J Harry Caufield, Mathieu Bourdenx, Yan Li, Jineta Banerjee, Luca Foschini, et al. Openscientist: evaluating an open agentic ai co-scientist to accelerate biomedical discovery. *medRxiv*, 2026.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Eriko Takeuchi, Yoshiaki Yasumizu, Junko Morita, Masakazu Ishikawa, Kotaro Ogawa, Daisuke Motooka, Daisuke Okuzaki, Miho Nagata, Yasuki Ishihara, Yohei Miyashita, et al. Single-nucleus multiome shows motor neuron glutamate overactivation in amyotrophic lateral sclerosis. *Brain*, 149(7):2480–2494, 2026.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, 2023. doi: 10.1038/s41586-023-06139-9.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Yun Yan, Yiyun Lin, Tapsi Kumar, Shanshan Bai, Aatish Thennavan, Jianzhuo Li, Emi Sei, Tuan Tran, Min Hu, Mitchell Rao, et al. Ecotypes of triple-negative breast cancer in response to chemotherapy. *Nature*, pp. 1–10, 2026.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822, 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023b.

Shunyu Yao, Fei Liu, Xi Lin, Zhichao Lu, Zhenkun Wang, and Qingfu Zhang. Multi-objective evolution of heuristic using large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27144–27152, 2025.

Haoran Ye, Jiarui Wang, Zhiguang Cao, Federico Berto, Chuanbo Hua, Haeyeon Kim, Jinkyoo Park, and Guojie Song. Reevo: Large language models as hyper-heuristics with reflective evolution. *Advances in neural information processing systems*, 37:43571–43608, 2024.

Fan Zhang, Tianyu Liu, Zhihong Zhu, Hao Wu, Haixin Wang, Donghao Zhou, Yefeng Zheng, Kun Wang, Xian Wu, and Pheng-Ann Heng. Cellverse: Do large language models really understand cell biology? *Advances in Neural Information Processing Systems*, 38, 2026.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8783–8817. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.498.

Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049, 2017.

# A Dataset Registry and Provenance

Table 2–4 distinguish the full source object from the object actually supplied at runtime. Fixed subsets contain observed records. All inferential analyses retain the donor, sample, mouse, or animal as the independent unit, and population-level estimates from unequal-probability subsets use the recorded sampling weights.

Table 2: Dataset registry (D01–D04). Dimensions are observations × features. "Unit" denotes the highest-level unit available for inference.

| ID | Dataset and modality | Source size | Runtime object | Unit and design | Topics |
|---|---|---|---|---|---|
| D01 | Mouse post-influenza lung, 10x 3' v3 scRNA-seq | $123{,}154 \times 52{,}445$ | Full object, same size | 25 sample/donors; eight sacrifice-day labels; two homeostasis donors | 01–02 |
| D02 | Human LUAD histological subtypes, 10x 3' v3 scRNA-seq | $117{,}266 \times 57{,}398$ | Full object, same size | 18 donors; four recorded histological subtypes | 03–04 |
| D03 | Untreated TNBC, 10x 3' v2/v3 scRNA-seq | $427{,}823 \times 32{,}354$ | Fixed 50,000-cell subset | 101 donors; pCR, RD, and Excluded labels; two assays | 05–06 |
| D04 | Healthy human kidney, Slide-seqV2 | $677{,}256 \times 32{,}014$ | Fixed 50,000-spot subset | 9 donors, 17 samples, 34 donor-section units; cortex/medulla | 07–08 |

Table 3: Dataset registry (D05–D09).

| ID | Dataset and modality | Source size | Runtime object | Unit and design | Topics |
|---|---|---|---|---|---|
| D05 | Mouse ovary, Slide-seqV2 | $608{,}096 \times 29{,}286$ | Fixed 1/10 subset, $60{,}809 \times 29{,}286$ | 22 mice; Young/Middle/Old $= 9/6/7$; Estrus/Metestrus $= 15/7$ | 09–10 |
| D06 | Mouse brain aging atlas, MERFISH | $378{,}918 \times 374$ | Fixed 1/10 subset, $37{,}891 \times 374$ | 12 donors; three ages; eight regions; targeted 374-gene panel | 11–12 |
| D07 | Human fetal cortex in Down syndrome, 10x Multiome-derived RNA object | $248{,}998 \times 35{,}477$ | Fixed 1/10 subset, $24{,}899 \times 35{,}477$ | 30 donors, 15 control/15 DS; three developmental-stage bins | 13–14 |
| D08 | Adult human kidney, processed scATAC-seq | $27{,}034 \times 19{,}515$ | Full processed object | 5 donors; archived dataset | retired |
| D09 | Mouse vascular smooth muscle, processed scATAC-seq | $87{,}718 \times 18{,}922$ | Full object, same size | 23 animals; control/Tcf21-KO $= 14/9$; partially crossed genotype, HFD, and time | 16 |

Table 4: Dataset registry (D10–D13).

| ID | Dataset and modality | Source size | Runtime object | Unit and design | Topics |
|---|---|---|---|---|---|
| D10 | Human ALS lumbar spinal cord, mixed 10x 3' v3/multiome technologies, analyzed as RNA | $62{,}711 \times 35{,}467$ | Full object, same size | 12 donors, 6 ALS/6 control; assay may be confounded with disease | 17–18 |
| D11 | Mouse neural crest/PNS neurons, sci-RNA-seq3 | $103{,}999 \times 45{,}525$ | Fixed 1/10 subset, $10{,}399 \times 45{,}525$ | 65 donors; 43 day labels; 16 experimental batches; 6 cell classes | 19 |
| D12 | Fresh 68k PBMC Donor A, filtered scRNA-seq matrix | $68{,}579 \times 32{,}738$ | Converted full CSR object, same size | One sample; no donor, disease, cell-type, or batch metadata | 20 |
| D13 | Human SLE PBMC, 10x 3' v2 scRNA-seq | source size unavailable in the audit record | Local $50{,}000 \times 30{,}172$ subset | 261 donors; SLE/normal; four processing cohorts; 11 cell types | 15 |

## A.1 Dataset sources and scientific scope

**D01** Niethamer et al. (2025), *Cell Stem Cell*; CELLxGENE collection `606dc65c-ebd8-4f93-ac86-d0789c04863f`. Sacrifice day defines a cross-sectional time point. Analyses compare sample composition and within-compartment transcriptional state at each sacrifice time.

**D02** Lee et al. (2026), *Experimental Hematology & Oncology*; CELLxGENE collection `0bebef1a-4607-4584-9070-dacf89a0d635`. Subtype sample sizes are approximately 6, 5, 4, and 3 donors, and all inferential comparisons use the donor as the independent unit.

**D03** Yan et al. (2026); CELLxGENE collection `ceef2841-5333-46ac-92ef-ccbe0c20fe55`, "Single-cell atlas of untreated human triple negative breast cancer." The subset was sampled without replacement with seed `20260913`, preserving donor coverage and recording source row, sampling probability, and weight. The Excluded category is omitted from the primary pCR-versus-RD comparison.

**D04** Marshall et al. (2022), *iScience*; CELLxGENE collection `8e880741-bf9a-4c8e-9227-934204631d2a`. Sections are modeled as measurements nested within donors.

**D05** Lan et al. (2026), *Nature Aging*; CELLxGENE collection `ed04d975-b0f9-4527-ba1d-7c249ba7b8a0`. The fixed subset uses seed `20260914` and retains all mice, stages, pucks, segments, and cell-type strata. Analyses use mice as independent units, treat spots as nested observations, and account for the observed imbalance between age and estrous stage.

**D06** Allen et al. (2023), *Cell*; CELLxGENE collection `31937775-0602-4e52-a799-b6acdd2bac2e`. Analyses are restricted to the 374 measured genes, and the two stored coordinate representations are verified separately before use.

**D07** Lattke et al. (2026), *Nature Medicine*; CELLxGENE collection `0e9fd1d3-ef4c-47c6-a2e4-ef4bfadf7c79`. The runtime analysis uses the RNA matrix and interprets its scores as transcriptional associations. Paired accessibility analysis requires an ATAC matrix or fragment file.

**D08** Muto et al. (2021), *Nature Communications*; CELLxGENE collection `9b02383a-9358-4f0f-9795-a891ec523bcc`.

**D09** Li et al. (2026), *Nature Communications*; CELLxGENE collection `7a3044e4-6b16-4693-9504-212d9a573f80`. The matrix may represent peaks, gene scores, or another processed feature space; this semantics is resolved before analysis. The animal is the inferential unit.

**D10** Takeuchi et al. (2026), *Brain*; CELLxGENE collection `0986e4cd-7a58-405d-9b91-4b199bb4124e`. The dataset contains 12 donors, and disease–assay overlap determines which comparisons are estimable.

**D11** Qiu et al. (2024); *Major cell cluster: Neural crest (PNS neurons)*; CELLxGENE collection `45d5d2c3-bc28-4814-aed6-0bb6f0e11c82`. The subset uses seed `20260914`, preserves all 326 observed joint strata, and records sampling weights. Developmental interpretation is based on donor-level concordance after accounting for the structured relationship between day and experimental batch.

**D12** Zheng et al. (2017); 10x Genomics, *Fresh 68k PBMCs Donor A*. The filtered Matrix Market object was transposed to cells $\times$ genes and stored as float32 CSR. The single-sample study evaluates representation stability across preprocessing and modeling choices.

**D13** Perez et al. (2022); CELLxGENE collection `436154da-bcf1-4130-9c8b-120ff9a888f2`.

## B Experimental Protocol and Reproducibility

This appendix documents the experimental protocol, dataset provenance, topic pool, alignment contract, execution stages, and fixed prompt templates used by CellScientist. Topic text, inspected schemas, retrieved literature, contracts, executable code, numerical results, and figures are run-specific inputs; the control instructions reproduced in Appendix E are fixed templates. All primary, auxiliary, validation, and falsification experiments use observed data; resampling operates on observed experimental units.

### B.1 END-TO-END TASK

Each run starts from a topic seed and a declared local dataset. The system (i) inspects the schema and prepares a read-only data card, (ii) retrieves relevant scientific knowledge, (iii) proposes candidate studies, (iv) freezes a Research Contract before examining experimental results, (v) selects one root after bounded probes, (vi) executes four artifact-separated analysis stages, and (vii) compiles citations, figures, and a manuscript from recorded evidence. Cells or spots are treated as nested observations; the independent inferential unit is the donor, sample, mouse, animal, or other experimental unit specified by the topic and contract.

### B.2 SYSTEMS, GENERATIONS, AND SCORING

The benchmark evaluation covers 20 topics with five configurations: CellScientist with Gemini and GPT backbones, OpenScientist with GPT, and AI Scientist v1 with Gemini and GPT. The target design requests three manuscripts per topic and configuration. Manuscripts are scored blindly by an automated reviewer on seven 1–10 dimensions: Novelty, Soundness, Clarity, Significance, Reproducibility, Research Design Coherence, and Data–Conclusion Traceability. The composite score is the unweighted arithmetic mean of these seven values. Quality summaries are computed over scorable manuscripts, and run success is reported separately.

### B.3 TOPIC-LEVEL UNCERTAINTY FOR THE AGGREGATE BENCHMARK

The 60 manuscript scores within a system are nested within 20 topics and are therefore not treated as 60 independent observations. For system $s$ and topic $t$, let $\bar{y}_{s,t}$ be the mean composite score of the three benchmark manuscripts. Each paired comparison uses

$$\Delta_t = \bar{y}_{\text{CellScientist},t} - \bar{y}_{\text{baseline},t}, \tag{3}$$

and reports the mean of the 20 topic-level differences. The 95% confidence interval is the percentile interval from 100,000 nonparametric bootstrap resamples of topics, keeping each topic's three manuscripts grouped. We also report an exact two-sided paired randomization test obtained from all $2^{20}$ sign flips of the topic-level differences. This directly targets the paired topic estimand without treating within-topic generations as independent.

Table 5: **Topic-level uncertainty for composite-score comparisons.** Confidence intervals resample the 20 topics; $p_{\text{perm}}$ is the exact two-sided paired randomization $p$-value. The $\Delta_t > 0$ column counts topics favoring the first system; the first comparison also has one tie.

| **Comparison** | **Mean $\Delta_t$** | **95% CI** | $\Delta_t > 0$ | $p_{\text{perm}}$ |
|---|---|---|---|---|
| CellScientist-Gemini − OpenScientist | 1.04 | $[0.50, 1.53]$ | 15/20 | 0.0017 |
| CellScientist-GPT − OpenScientist | 0.81 | $[0.45, 1.11]$ | 19/20 | 0.0004 |
| CellScientist-Gemini − AI Scientist-Gemini | 1.61 | $[1.24, 1.94]$ | 19/20 | $< 0.0001$ |
| CellScientist-GPT − AI Scientist-GPT | 1.51 | $[1.03, 1.97]$ | 19/20 | $< 0.0001$ |
| CellScientist-Gemini − CellScientist-GPT | 0.23 | $[-0.32, 0.74]$ | 12/20 | 0.418 |

The four system–baseline comparisons have confidence intervals above zero. In contrast, the interval for the 6.47 versus 6.24 comparison includes zero, so the aggregate benchmark does not distinguish the two CellScientist backbones. Because a single automated judge scored each manuscript, these intervals quantify uncertainty across topics conditional on that judge; the available benchmark does not identify judge-to-judge variance.

### B.4 RUN RECORDS AND AUDIT TRAIL

For every run, the minimum reproducibility record contains the topic identifier, source seed path and SHA-256, frozen runtime seed and SHA-256, batch directory, actual H5AD path and SHA-256, H5AD dimensions, biological unit, exposure and outcome, primary comparison, exclusions, Research Contract, software/model configuration, random seeds, artifact catalog, terminal status, and manuscript.

# C TOPIC SEED POOL

The topic seed pool is summarized below at the level of the research question. Detailed seed text and run-specific analysis specifications are maintained in the supplementary materials; empirical outcomes are recorded separately for each run.

Table 6: **Topic seed pool.** High-level research questions used to initialize the benchmark runs.

| ID | High-level topic seed |
|---|---|
| T01 | Lung recovery—recruitment versus within-compartment repair |
| T02 | Lung recovery coordination beyond shared time and markers |
| T03 | LUAD histology—immune recruitment versus immune state |
| T04 | LUAD histological signatures—transportability versus patient idiosyncrasy |
| T05 | TNBC response signals beyond assay and composition |
| T06 | TNBC immune heterogeneity beyond sampling depth |
| T07 | Kidney regional biology versus section sampling |
| T08 | Kidney regional signatures across sections and donors |
| T09 | Ovarian aging and cycle—composition versus state |
| T10 | Ovarian physiological signal versus sampling variation |
| T11 | Brain aging—shared versus region-specific biology |
| T12 | Microglial aging—cross-region transfer versus regional identity |
| T13 | Down syndrome cortex—developmental timing versus disease-associated RNA |
| T14 | Down syndrome RNA programs across developmental stages |
| T15 | Protocol-invariant SLE representations |
| T16 | SMC perturbation—redistribution versus chromatin-state change |
| T17 | ALS RNA signals beyond cell loss and assay differences |
| T18 | ALS assay correction—benefit versus signal removal |
| T19 | Neural crest development—progression versus batch alignment |
| T20 | PBMC structure—preprocessing stability versus algorithmic imposition |

# D RESEARCH CONTRACT AND STAGE-SEPARATED EXECUTION

Before full execution, every proposal is converted to a frozen Research Contract. The contract records the pre-experiment specification before full result inspection. It contains the research question, hypothesis, unit of inference, target population, included and excluded comparison groups, primary estimand, and a primary metric with an explicit numerator, denominator, aggregation, direction, analysis unit, and interpretation. It further records full-data and resampling decision rules (statistical threshold, effect threshold, direction requirement, and multiplicity scope), covariates/exposures, feasibility checks, no more than two required execution-critical falsification tests, success and failure criteria, required raw records, assumptions, and the evidence basis for consequential design choices.

## D.1 SEMANTIC AND DETERMINISTIC VALIDATION

The contract is generated in four repairable blocks: *core*, *inference*, *controls*, and *bindings*. The language model supplies all semantic choices, while the compiler adds measured execution facts. Provider-native schemas enforce types and required fields. A separate semantic review compares the complete contract with the topic, proposal, measured feasibility, and retrieved evidence. It requests revision only for a concrete supported inconsistency; genuinely undecidable cases return `insufficient_information` and may proceed.

## D.2 FOUR ANALYSIS STAGES

1. **Stage 1: cohort and primary analysis.** Freeze the cohort, hypothesis-specific exposure/features, primary estimand, and minimum model diagnostic. Schema discovery and identifier normalization are performed once.
2. **Stage 2: resampling validation.** Apply the prespecified resampling rule at the true inferential unit and retain every valid resample, including unit assignments, seeds, and recomputable records.

3. **Stage 3: required falsification.** Run only the one or two contracted execution-critical null, permutation, or negative-control tests on frozen upstream artifacts.
4. **Stage 4: sensitivity and ablation.** Evaluate prespecified or scientifically motivated robustness analyses without replacing the primary estimand. Plotting and writing occur only after this stage and consume saved artifacts rather than recomputing statistics.

Each downstream stage starts from a run-scoped artifact catalog and receives already loaded immutable objects. Downstream computation is restricted to the assigned stage and its registered upstream artifacts. Permutations are kept distinct from donor/patient resampling, and all required raw records are indexed.

### D.3 Reflection, mutation, and terminal outcomes

Bounded root probes provide advisory evidence to a separate selector. Promotion uses importance, novelty, feasibility, diversity, and the likelihood of interpretable execution. Preliminary effect size and $p$-value are outside the promotion criterion. After full execution, positive, negative, null, borderline, underpowered, and sensitivity-dependent results proceed to reporting. Mutation is reserved for a structural failure such as an unmeasurable estimand, an unidentifiable required comparison, a principled falsification, or execution of the wrong data, unit, outcome, or definition. A mutation preserves the observed data source and updates the contract leaves justified by that failure.

The terminal review asks whether the declared question and estimand remain identifiable and reportable from the observed evidence. Its scope is structural validity; novelty, effect direction, interval width, venue judgment, and optional analyses remain reporting considerations. Runs with support, non-support, or insufficient evidence pass to writing with the corresponding limitations.

## E Fixed Prompt Templates

The following templates reproduce the fixed instructions used by the default idea-evolution path. Braced, upper-case names denote values injected at runtime. Provider-native function schemas validate structured outputs; the instantiated schema and payload are stored in the corresponding run archive. The inventory covers the default end-to-end path; legacy entry points remain in the source archive.

Table 7: Prompt inventory for the default end-to-end workflow.

| ID | Stage | Purpose |
|---|---|---|
| P0 | Feasibility | Local schema scan and semantic data-card construction |
| P1 | Retrieval | Scientific query expansion for the local knowledge index |
| P2 | Ideation | Diverse candidate proposal generation |
| P3 | Alignment | Research Contract blocks, compilation, and semantic review |
| P4 | Evolution | Root selection, probe reflection, terminal review, and mutation |
| P5 | Experiment | Stage 1–4 code generation, debugging, and evidence review |
| P6 | Reporting | Memory, node selection, summaries, and evidence-only plots |
| P7 | Writing | Citations, LaTeX drafting, polishing, and figure consistency |
| P8 | Review | LLM manuscript review, meta-review, and VLM figure review |

**Runtime placeholders.** Braced names inject the topic seed, inspected schema and data card, retrieved knowledge, frozen Research Contract, code and execution artifacts, recorded results and figure descriptions, Semantic Scholar results, and few-shot review examples. These run-specific values provide evidence-bearing context and are not part of the fixed prompt text reproduced below.

### E.1 P0: global real-data boundary and data card

The topic is augmented with the following real-data boundary.

**P0-A Global real-data boundary**

```
SYSTEM-WIDE REAL-DATA CONSTRAINT
Use only the real observed datasets declared in this Topic Seed. Synthetic or
simulated patients, cells, clonotypes, outcomes, labels, counts, neighborhoods,
and injected effects are prohibited in primary, auxiliary, validation, and
falsification experiments. Resampling observed experimental units is allowed.

RUN OUTPUT CONTRACT
An unmutated root idea that is promoted must run the complete Stage 1-4 search.
A result-guided scientific mutation must use the controller's deterministic
Research Contract invalidation boundary, reuse only unaffected parent artifacts,
and complete every remaining stage. Every successful path produces a four-page
ICBINB/workshop paper by default. The paper must still report the full experiment
and its limitations; page length does not reduce the required scientific modules.
```

Inspected facts are injected as read-only planning context:

**P0-B Injected pre-contract data facts**

```
PRE-CONTRACT LOCAL DATA FACTS (system-generated, read-only):
{{PRECONTRACT_FACTS_JSON}}
These are observations, not scientific results. Do not optimize thresholds or
features against them or claim a biological finding.
```

The data-card system instruction is:

**P0-C Data-card system instruction**

```
You interpret dataset schemas for scientific planning. Read every provided
column profile and map scientifically meaningful columns to roles using names,
dtypes, observed low-cardinality values, and matrix metadata together. Do not
infer roles from keyword rules. Do not invent columns, values, relationships,
or biological facts. High-cardinality values were intentionally omitted.
Produce one descriptive Data Card for all sources.
```

The user supplies `INSPECTED DATASET FACTS:` followed by {{`SCHEMA_VERSION_AND_READY_SOURCES_JSON`}}. If validation fails, the next call also contains the previous error, treats the inspected namespace and column lists as authoritative, and accepts fields defined by the `BuildDatasetCard` schema.

### E.2 P1: QUERY EXPANSION

**P1 Query-expansion prompt**

```
System: You are a scientific retrieval component. Return only valid JSON. Do
not answer the research question. Return concise, established terminology that
can be passed to a local embedding retriever.

User: Expand this scientific question for local embedding retrieval.

Query:
{{QUERY}}

Return this JSON shape:
{
  "search_terms": ["specific English term or method name"],
  "research_tasks": ["analysis task"],
  "constraints": ["confounder, validity risk, or execution constraint"]
}

Use established English scientific terminology and tool names. Include useful
synonyms, but keep the combined output to at most {{MAX_QUERY_TERMS}} short
phrases.
```

The three arrays are deduplicated and truncated to the configured maximum before local embedding retrieval.

### E.3 P2: CANDIDATE PROPOSAL

**P2 Candidate-proposal prompt**

```
System: Propose a creative and feasible scientific study that answers the Topic
Seed and differs from earlier candidates. Use retrieved literature and
scientific knowledge as optional sources of inspiration, prior-work awareness,
and methodological guidance; the proposal need not be derived from an existing
publication. It may instead originate from the biological question, observed
data structure, or your own scientific reasoning. Generate only the seven
narrative proposal fields. Specify the question, hypothesis, target population,
independent analysis unit, comparisons and proposed experiments clearly in that
narrative. Respect all Topic Seed constraints. Call SearchScientificKnowledge,
then FinalizeScientificProposal. The Research Contract and execution bindings
are generated in a separate pass. Do not generate those objects now or claim
unrun results.

User:
{{TOPIC_SEED}}

{{PRECONTRACT_FEASIBILITY_CONTEXT}}

Here are the proposals that you have already generated:
'''
{{PREVIOUS_PROPOSALS}}
'''

{{DOMAIN_KNOWLEDGE_CONTEXT}}

Begin by generating an interestingly new high-level research proposal that
differs from what you have previously proposed.
```

The seven narrative fields are Name, Title, Short Hypothesis, Related Work, Abstract, Experiments, and Risk Factors and Limitations.

### E.4 P3: RESEARCH CONTRACT

All four semantic blocks share the following scientific instruction.

**P3-A Shared Research Contract instruction**

```
Before finalizing, convert the research question into a Research Contract using
the retrieved literature/knowledge and your own scientific reasoning. This is a
pre-experiment commitment: do not tune it to results that have not been run.

The final idea must include a Research Contract object with exactly these
scientific components:
- research_question, hypothesis, unit_of_inference, target_population
- comparison_groups: explicit included and excluded groups
- primary_estimand
- primary_metric: name, formula, numerator, denominator, aggregation,
  analysis_unit, direction, and the exact interpretation supported by the value
- full_data_decision_rule: statistical_threshold, effect_threshold,
  direction_requirement, multiplicity_scope
- resampling_validation_rule: the same four fields plus
  valid_resample_definition and pass_fraction. Do not silently make this rule
  weaker than the full-data rule; explain any necessary difference.
- multiplicity_control and covariates_and_exposures
- data_feasibility_checks: objects with check, minimum_requirement, and
  failure_response
- falsification_tests: objects with test, rationale, acceptance_criterion, and
  required (boolean). Choose at most two required, execution-critical tests that
  can run from frozen upstream artifacts. Put other robustness analyses in
  Stage 4 sensitivity analysis or future work.
- success_criteria, failure_criteria, required_raw_records, assumptions
- evidence_basis: objects with source, finding, and contract_implication.

The contract must define what a negative or underpowered result means. Required
raw records must be sufficient to recompute every headline metric at its true
unit of inference, including each valid resampling unit rather than summaries
alone. Record measured values, unit assignments, seeds, parameters, and native
tool receipts, but never request SHA-256, checksums, file/content hashes,
code-tree hashes, or other file-integrity fingerprints.

Keep the implementation boundary explicit: Stage 1 prepares data and computes
the single primary analysis; Stage 2 performs resampling validation; Stage 3
```

```
runs only the required falsification controls; Stage 4 runs sensitivity/ablation
analyses. Plotting and manuscript writing happen after those modules and must
consume their saved artifacts rather than recomputing the study.
```

Each block is requested using this payload:

**P3-B Contract-block payload**

```
TOPIC SEED:
{{TOPIC_SEED}}

{{PRECONTRACT_FEASIBILITY_CONTEXT}}

KNOWLEDGE EVIDENCE:
{{KNOWLEDGE_CONTEXT}}

SCIENTIFIC PROPOSAL:
{{PROPOSAL_JSON}}

ALREADY GENERATED CONTRACT BLOCKS:
{{PREVIOUS_BLOCKS_JSON}}

EXECUTION REFERENCES (facts and available choices):
{{COMPILER_CONTEXT}}

CURRENT BLOCK: {{core|inference|controls|bindings}}
Generate every field in this block. Scientific choices, including whether a
falsification test is required, must be made explicitly by you. The compiler
will only add fixed execution facts and will never fill a missing semantic value.

EXACT ARGUMENT SHAPE:
{{BLOCK_JSON_SCHEMA}}

CORRECTION FEEDBACK:
{{OPTIONAL_VALIDATION_OR_REVIEW_FEEDBACK}}
```

The system then says: "Generate only the requested semantic block. Supply every required field yourself; the compiler cannot choose a semantic default." The completed contract is reviewed with the following instructions:

**P3-C Contract semantic review**

```
System: You are a scientific contract reviewer. Judge semantic consistency from
evidence, never from keyword presence. Do not redesign a study just because its
result might be negative.

User: Review the complete pre-experiment contract below. Interpret scientific
meaning; do not apply keyword rules. Return `revise` only for a concrete error
supported by the supplied evidence. Return `insufficient_information` when the
evidence cannot decide; that result permits continuation.

TOPIC: {{TOPIC_SEED}}
PROPOSAL: {{PROPOSAL_JSON}}
CONTRACT: {{CONTRACT_JSON}}
MEASURED FEASIBILITY: {{FEASIBILITY_CONTEXT}}
RETRIEVED EVIDENCE: {{KNOWLEDGE_CONTEXT}}
```

### E.5 P4: IDEA EVOLUTION AND TERMINAL REVIEW

Candidate roots are selected before results exist:

**P4-A Candidate-root selection**

```
You select research directions before any experimental result exists. Rank by
scientific importance, novelty, real-data feasibility, and pairwise diversity.
Never invent or assume experimental outcomes.

Select exactly {{ROOT_COUNT}} distinct roots from this complete candidate pool
using the same validated dataset interpretation supplied during candidate
generation.

DATASET CARD: {{DATASET_CARD_JSON}}
CANDIDATE POOL: {{COMPACT_CANDIDATE_POOL_JSON}}
```

After bounded probes, a second selector chooses the root most likely to execute and yield interpretable evidence. Selection criteria exclude preliminary effect size and $p$-value; null and inconclusive probes remain informative. The terminal review uses the following fixed instruction:

**P4-B Terminal scientific review**

```
You are the structured scientific terminal reviewer. This review runs after
Stage 4 sensitivity/validation and before citations or manuscript generation.
Decide only whether the declared primary question and prespecified estimand
remain identifiable and honestly reportable from the observed data. Do not
judge novelty, venue fit, writing quality, number of figures, citation count,
or whether the result is exciting. A result is allowed to be positive, negative,
null, inconclusive, underpowered, small-effect, borderline, or to disagree
across sensitivity analyses. These cases MUST return disposition=pass and can
be written as a null/inconclusive paper with limitations. A reviewer Reject, a
p-value above 0.05, a wide CI, low novelty, missing optional analysis, or an
unexpected result is never by itself a mutation reason.

Return disposition=mutate only for a structural case that prevents the original
primary question from being answered: (1) estimand mismatch; (2) unidentifiable
question because a required group/outcome/unit/key variable is absent or labels
do not match; (3) principled falsification with adequate information; or (4)
implementation_contract, where the executed analysis used the wrong data, unit,
outcome, or prespecified definition. Mutation repairs that structural failure,
not significance. Preserve the observed data source and set
preserve_data_source=true. If the question can be answered as support,
non-support, or inability to determine, pass it.

IDEA AND FROZEN CONTRACT:
{{COMPACT_IDEA_AND_CONTRACT_JSON}}

COMPLETED STAGE 1-4 ANALYSIS RECORD:
{{BOUNDED_FULL_RUN_EVIDENCE_JSON}}
```

A justified mutation returns a sparse patch. The parent remains authoritative for omitted fields, nested contract changes are recursively merged, the dataset and real-data policy remain fixed, and tool-plan changes require support from the reflection and registered tool catalog.

### E.6 P5: ANALYSIS CODE AND SCIENTIFIC REFLECTION

All code-generation responses contain a short natural-language plan followed by exactly one complete executable Python block. Stage 1 receives:

**P5-A Stage 1 core-analysis prompt**

```
You are implementing the small core-analysis module for a publishable real-data
study. Stage 1A has already frozen schema, provenance, and identifier
normalization. Implement only cohort construction, the hypothesis-specific
primary feature/exposure, the primary estimand, and its minimum model diagnostic.
Later modules own resampling, falsification, sensitivity analysis, independent
validation, and plots.

Research idea: {{RESEARCH_CONTRACT_AND_PROPOSAL}}
Memory: {{LONG_AND_SHORT_TERM_MEMORY}}
Instructions:
- This core module must be narrow and executable.
- Do not redo Stage 1A schema discovery or identifier normalization.
- {{REAL_DATA_LOADING_INSTRUCTION}}
- Use stable analysis-runtime helpers and write the experiment index.
Evaluation Metric(s): {{METRICS}}
Implementation guideline: {{COMMON_IMPLEMENTATION_GUIDELINE}}
Installed Packages: {{SHUFFLED_PACKAGE_LIST}}
Data Overview: {{OPTIONAL_DATA_PREVIEW}}
```

Stages 2–4 receive the following artifact boundary:

**P5-B Stages 2–4 artifact boundary**

```
Implement Stage {{STAGE_NUMBER}} ({{STAGE_NAME}}) as one small executable module.
The upstream experiment is frozen. Read its saved artifacts through
load_analysis_inputs([...]) and add only this stage's evidence; never copy or
rerun the complete upstream program.

Implement only: {{MODULE_RESPONSIBILITY}}
Use upstream analysis-unit records as immutable inputs; do not read canonical
```

```
source tables in this module. Use only resampling and permutation counts in the
binding contract. Keep permutation controls separate from patient/donor
resampling and preserve the true unit of inference. Save every raw record
assigned to this module and write the module's artifact index. Do not silently
replace an infeasible contracted analysis with a weaker proxy.
```

Generated code is pre-reviewed for execution readiness, registered artifact access, mandatory outputs, stage-scope compliance, observed-data use, and plain alignment to the contracted question, groups, and unit. The frozen contract supplies the estimand, formula, statistical test, threshold, normalization, covariate strategy, and biological interpretation.

Execution-result review marks a bug only for a runtime exception, missing mandatory output, or unparsable result. Null, negative, underpowered, unstable, or hypothesis-falsifying findings are valid outputs. The subsequent in-loop reflection follows this audit protocol:

**P5-C In-loop scientific audit**

```
DATA DESIGN: reconstruct included/excluded labels and actual group counts; check
imbalance, unit of inference, independence/pseudoreplication, zero-contribution
resamples, data provenance, exposure/offsets, and required covariates.
METRIC SEMANTICS: write down the implemented numerator, denominator, aggregation
levels, and threshold sequence. Distinguish an individual resampling score from
the fraction of selected items passing a threshold.
STATISTICAL VALIDITY: compare full-data and resampling decision rules, including
multiple-testing correction, effect threshold and direction; check assumptions,
convergence, uncertainty, imbalance, and whether valid resamples change data.
FALSIFICATION: verify from code and artifacts which required permutation/null
tests, negative controls, baselines, sensitivity analyses and distinct seeds ran.
CONTRACT: compare implementation and saved records with the Research Contract.
SCIENTIFIC REFLECTION IS ADVISORY: biological interpretation, assumptions,
limitations and future work are advisory.
BLOCKING ANOMALIES ARE NARROW: only runtime failure, missing mandatory output,
unparsable/non-finite result, synthetic-data substitution, or mutation of the
binding primary estimand/cohort/endpoint can block execution.
```

For real-data nodes, best-implementation selection prioritizes contract compliance, complete recomputable artifacts, numerical stability, and a focused implementation. Effect magnitude, significance, direction, and anticipated publishability are outside the selection criterion.

### E.7 P6: MEMORY, SUMMARIES, AND PLOTTING

Long-term memory preserves concrete evidence, failed approaches, validity concerns, and required next actions while merging repetitions within a fixed character budget. Final plotting uses this fixed boundary:

**P6 Final plotting boundary**

```
FINAL REAL-DATA PLOTTING STEP:
All analysis modules have finished. Do not fit models, resample, or recompute
statistics. Read working/artifact_catalog.json first. It is the canonical
dict-shaped catalog. Prefer load_analysis_inputs([...]) and use exact saved
CSV/JSON/NPY/NPZ values. Create only 3-5 paper-relevant figures. Prefer
patient-level estimates, confidence/stability summaries, falsification results,
and sensitivity comparisons actually supported by artifacts. Do not invent
values and do not add training curves unless genuine training records exist.
Save PNG files directly in working/ and close every figure.
```

Plotting guidelines require a white background, high-contrast text, colorblind- safe colors, explicit axes and units, sample size and uncertainty where supported, catalogued artifacts, unconnected unordered categories, 300 dpi, unclipped legends, and unique descriptive filenames. The final aggregator is limited to 12 figures and three panels per figure, with every plotted array and value traced to an experiment summary.

### E.8 P7: CITATIONS AND MANUSCRIPT WRITING

Citation gathering uses only Semantic Scholar search results, rejects duplicates, and records why and where each selected paper is needed. The writing stages share this system skeleton:

**P7 Manuscript-writing system skeleton**

```
{{STAGE_ROLE}}
{{VENUE_FORMATTING_RULE}}
The main text, including figures and tables, is limited to {{PAGE_LIMIT}} pages;
references, the impact statement, and appendices are excluded. The venue style
is a formatting requirement, not a preference for any kind of experimental
result. Preserve the document class, venue package, graphicspath, and embedded
references.bib. Use only supplied verified citations and recorded results.
Return the complete LaTeX manuscript in a single fenced latex block, without
placeholders, translations, or editorial checklists in the paper.

{{WRITING_SKILL_TEXT}}
```

The draft role writes a scientifically accurate paper from completed experiments. The polish role improves structure and prose while preserving all numerical values, methods, uncertainty, limitations, and conclusion strength. The figure role aligns every panel, table, caption, label, and textual claim to recorded evidence. Missing evidence is reported as uncertainty. Mechanical repair covers LaTeX, template, and filename errors while preserving scientific meaning.

### E.9 P8: MANUSCRIPT AND FIGURE REVIEW

The default manuscript reviewer is told to be critical and cautious and returns a structured review with Summary, Strengths, Weaknesses, Originality, Quality, Clarity, Significance, Questions, Limitations, Ethical Concerns, Soundness, Presentation, Contribution, Overall, Confidence, and an Accept/Reject Decision. A reflection pass checks review accuracy and JSON validity while retaining the substance of the original review unless it contains a clear error. Multiple reviews are aggregated by an area-chair prompt that seeks consensus and respects all reviewers.

Figure VLM review describes only visible content, checks axes, legends, trends, captions, and text references, and flags missing labels, unsupported conclusions, or caption mismatch. Under a page limit it additionally judges whether a figure is informative, contains subfigures, or belongs in the appendix. The common output protocol is:

**P8 Structured review output**

````
THOUGHT:
<THOUGHT>

REVIEW JSON:
```json
<JSON>
```
````