MECHANOVA: AN AUTONOMOUS SCIENTIST FOR DISCOVERING FALSIFIABLE MECHANISMS OF IMMUNE-COLD-TO-HOT CONVERSION IN TUMORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous scientific discovery requires agents to revise causal explanations in response to evidence and translate those revisions into new experiments. We introduce Mechanova, an autonomous scientist for discovering falsifiable mechanisms of immune-cold-to-hot conversion in tumors. Mechanova couples competing causal graphs to experimental evidence and counterevidence, selects discriminating experiments, revises mechanistic structure in response to prediction–observation discrepancies, and compiles surviving explanations into testable intervention programs. Its Mechanism Evidence Chain (MEC) grounds causal claims and licenses structural edits through context-matched experimental evidence. Biological and operational world models predict intervention effects and research-action outcomes, respectively. Provenance-aware memory and observation-conditioned hierarchical control sustain adaptive, long-horizon investigation, while offline policy learning improves execution across investigations. In a comparison on 125 paper-context tasks, the rule-assisted system achieves 60.8% mechanism identification and 52.0% pre-observation intervention agreement, exceeding the strongest completed comparator for each endpoint by 19.6 and 12.4 percentage points. A separate matched-prompt study on 88 source families isolates structural editing under identical downstream probability and action rules: enabling edits reduces native Brier by 25.3%, from 0.1119 to 0.0835 (paired difference -0.0284; 95% CI $[-0.0439, -0.0146]$). Graph-replacement controls further demonstrate that downstream outputs depend on the revised mechanism. Supporting studies raise evidence recall from 25.2% to 36.8% with zero measured unsupported assertions and useful execution completion from 50% to 100%. Mechanova establishes a mechanism-centered approach to autonomous scientific discovery in which evidence governs causal claims, structural revision improves prediction, and surviving explanations generate explicit experimental commitments.

1 INTRODUCTION

Scientific discovery advances when an experiment changes an explanation and the revised explanation changes the next experiment. Immune-cold-to-hot conversion in tumors requires identifying barriers to antitumor immunity, their interactions, and effective perturbations. Stromal TGFβ signaling, linked to T-cell exclusion and checkpoint-blockade resistance, illustrates the importance of biological context (Mariathasan et al., 2018). The problem is causal and sequential: **which mechanism explains the response, what experiment distinguishes alternatives, and what intervention follows?** Failed interventions may challenge pathways or reveal unmet prerequisites or residual barriers; distinguishing these possibilities guides subsequent experiments.

AI Scientist-v2 integrates hypothesis formation, computational experimentation, and analysis through agentic exploration (Yamada et al., 2025); Co-Scientist refines hypotheses through critique, ranking, and competition (Gottweis et al., 2026). A central challenge is making causal revision explicit and establishing its consequences for prediction and action. Reweighting fixed hypotheses cannot introduce missing mediators or replace unsupported causal relations; revision requires evidence-justified structural changes that inform downstream decisions. We organize discovery accordingly:

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.

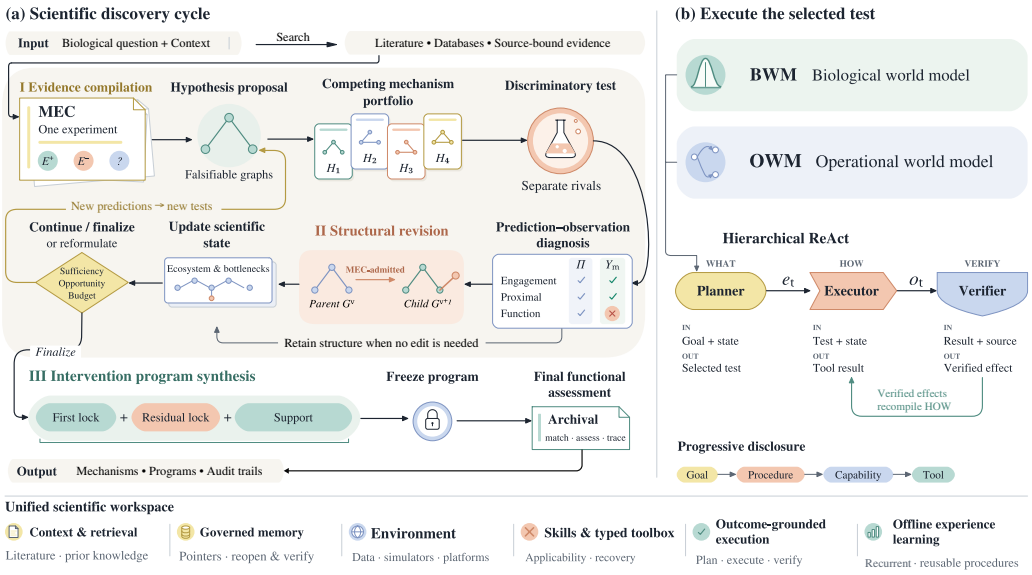

Figure 1: **Mechanova: a mechanism-centered scientific cycle.** (a) Experiment-grounded evidence supports competing explanations. A discriminating test opens an observation; evidence-licensed revision changes the mechanism when warranted, and a surviving mechanism generates an intervention program before archival assessment. (b) Biological and operational world models support WHAT–HOW control: verified effects update the scientific state and recompile execution. Memory reopens source evidence. The scientific interfaces connect separately evaluated realizations; executive weights remain frozen, and learned execution policies are assessed in their own study.

evidence constrains explanations, experiments distinguish alternatives, and revision creates experimental commitments.

We introduce **Mechanova**, an autonomous scientist for discovering falsifiable mechanisms of immune-cold-to-hot conversion in tumors. Mechanisms persist as revisable scientific states linking evidence, competing explanations, predictions, and intervention programs. **Pancreatic ductal adenocarcinoma (PDAC) serves as the flagship testbed for instantiating and evaluating Mechanova.** Targeting CXCL12-mediated stromal exclusion can expose checkpoint-blockade responsiveness in PDAC models (Feig et al., 2013), while focal adhesion kinase studies link tumor-cell signaling, fibrosis, suppressive immune populations, and responses to combined interventions (Jiang et al., 2016). These dependencies demand reasoning about residual barriers rather than isolated targets.

Mechanova couples competing causal graphs to experimental evidence and counterevidence to select discriminating experiments. The **Mechanism Evidence Chain (MEC)** binds causal claims to experimental systems, perturbations, comparators, contexts, and measured effects, distinguishing association, necessity, sufficiency, and rescue. Mechanisms carry potential falsifiers and predictions committed before observation. After localizing discrepancies, Mechanova revises relations, introduces missing pathways, or seeks further evidence without changing structure. Edits require relation- and context-matched evidence; missing observations remain unresolved. Surviving mechanisms yield intervention programs targeting primary and residual barriers and their combinations.

Biological and operational world models predict intervention effects and research-action outcomes, respectively, separating biological from operational uncertainty (Roohani et al., 2024; Adduri et al., 2026; Frazier et al., 2009; Toth et al., 2022). Provenance-aware memory preserves source access for reassessment. Observation-conditioned hierarchical control translates objectives into operations, adapts to verified effects, and allocates effort by unresolved questions and information value. Reusable skills support recovery; actionable research gaps govern continuation or stopping. Offline policy learning trains research-action selection from prior trajectories, with policies frozen during evaluation (Peng et al., 2019). Together, these support scientific revision within investigations and execution-policy improvement across investigations (Figure 1).

Computational studies connect system performance to controlled attribution. Evidence and task preparation draw on **469 human-reviewed papers and 872 adjudicated claims**; a separate **278-paper curation release yields 125 paper-context tasks**. In this comparison, rule-assisted Mechanova achieves **60.8% mechanism identification and 52.0% pre-observation intervention agreement**, exceeding the strongest completed comparator per endpoint by **19.6 and 12.4 percentage points**, respectively (Figure 2). These endpoints assess initial decisions; separate studies examine evidence use, acquisition, and execution.

A separate study on **88 source families** isolates edit permission under identical initial scientific information and downstream prediction and action rules. Enabling edits reduces native Brier from **0.1119 to 0.0835**, a **25.3% relative reduction** (paired difference $-\mathbf{0.0284}$; 95% CI $[-\mathbf{0.0439}, -\mathbf{0.0146}]$; Table 1). Graph-replacement controls show that downstream recommendations or probabilities depend on revised mechanisms. Supporting studies increase evidence recall from **25.2% to 36.8% with zero measured unsupported assertions** and useful execution completion from **50% to 100%**.

Contributions. **(1) Mechanism-centered autonomous discovery.** We connect competing causal graphs, experimental evidence, discriminating tests, and intervention programs through persistent scientific states with evidence-licensed revision. **(2) Adaptive, long-horizon investigation.** Dual world models, provenance-aware memory, observation-conditioned hierarchical control, and offline policy learning sustain investigation. **(3) System evaluation and structural-revision attribution.** Curated resources support comparisons; matched downstream rules and graph-replacement controls establish predictive gains from editing and downstream graph dependence, respectively.

2 Related work

Autonomous scientific discovery. AI Scientist-v2 uses agentic tree search (Yamada et al., 2025), while DeepScientist leverages accumulated findings and fidelity–cost trade-offs (Weng et al., 2026). Co-Scientist refines hypotheses (Gottweis et al., 2026); Robin integrates proposals with experimental analysis (Ghareeb et al., 2026). Biomni, SPARK, and OpenScientist support biomedical tool use, pathology-linked cohort analysis, and data-driven investigation, respectively (Huang et al., 2026; Trost et al., 2026; Roberts et al., 2026); SciAgents combines knowledge graphs with multi-agent hypothesis refinement (Ghafarollahi and Buehler, 2025). Mechanova centers discovery on competing causal graphs whose evidence-licensed revisions guide predictions, discriminating tests, and intervention programs.

Evidence and theory revision. SciFact grounds support and refutation in text (Wadden et al., 2020); INDRA assembles executable molecular models (Gyori et al., 2017). Theory refinement and self-revising discovery formalize changes to rules and scientific representations (Ourston and Mooney, 1994; Wang and Buehler, 2026). Mechanova's MEC grounds causal claims and structural edits in context-matched experiments, connecting revised graphs to prediction and intervention design. Controlled evaluations isolate editing effects under shared decision rules and test dependence on revised graphs.

World models and adaptive experimentation. GEARS and State predict cellular perturbation responses (Roohani et al., 2024; Adduri et al., 2026); knowledge-gradient policies value measurements under correlated beliefs (Frazier et al., 2009), and Active Bayesian Causal Inference selects query-specific interventions (Toth et al., 2022). BioDiscoveryAgent guides genetic experiments using language-model reasoning and observed outcomes (Roohani et al., 2025). Mechanova's dual world models predict biological effects and research-action outcomes; rival-aware acquisition distinguishes mechanisms and localizes disputed relations.

Long-horizon execution and experience-driven learning. Kosmos maintains research knowledge (Mitchener et al., 2025), ReAct interleaves reasoning and action (Yao et al., 2023), and Voyager accumulates executable skills (Wang et al., 2023); advantage-weighted regression supports off-policy learning (Peng et al., 2019). Mechanova combines provenance-aware memory with observation-conditioned hierarchical execution; evidence sufficiency and residual research value govern continuation or stopping. Offline policy learning adapts execution across investigations, complementing mechanism revision within them.

3 MECHANOVA: EVIDENCE-GROUNDED THEORY REVISION

Mechanova maintains competing mechanisms, source-bound evidence, committed predictions, and unresolved obligations. Evidence licenses claims; experiments distinguish alternatives; observations update confidence or structure; surviving mechanisms generate new tests and interventions. Figure 1 summarizes the cycle; Appendix I specifies the study-specific implementations.

3.1 SCIENTIFIC STATE AND EXPERIMENT-GROUNDED EVIDENCE

A versioned mechanism is

$$h_i^{(v)} = \left(\mathrm{id}_i, v, G_i^{(v)}, \mathcal{B}_i, c_i, \mathcal{E}_i^{+,-,?}, \Pi_i^{(v)}, \mathcal{F}_i, \mathcal{U}_i\right). \tag{1}$$

Here, $G_i^{(v)}$ is a typed graph of proposed causal relations, $\mathcal{B}_i$ its bottlenecks, c_i its context, and $\mathcal{E}_i^{+,-,?}$ its supporting, opposing, and unresolved evidence. Predictions $\Pi_i^{(v)}$, falsifiers $\mathcal{F}_i$, and interventions $\mathcal{U}_i$ connect explanation to experiment; persistent identities preserve revision lineage.

The **Mechanism Evidence Chain (MEC)** grounds claims in experiments rather than citations alone (Wadden et al., 2020; Gyori et al., 2017), binding systems, perturbations, comparators, dose/time, assays, endpoints, and effects to recoverable sources. Normalization preserves source bindings; missing fields create acquisition obligations. For relation η, verified evidence dimensions D_η^+ license predicates with satisfied requirements $R(p)$:

$$\mathcal{V}_\eta = \{p : R(p) \subseteq D_\eta^+\}. \tag{2}$$

Predicate requirements distinguish association, perturbational dependence, necessity, sufficiency, rescue, and ordering; admission also matches directed endpoints, compartments, and target context. Counterevidence is claim- and context-relative; nulls and exceptions retain their qualifications. Repeated source use adds no independent support. Cross-paper compositions remain bounded hypotheses requiring tests (Appendix B).

PDAC's seven diagnostic axes, inspired by the cancer-immunity cycle (Chen and Mellman, 2013), distinguish upstream prerequisites, downstream deficits, and parallel suppression. Downstream deficits remain actionable after upstream relief (Appendix A).

3.2 COMPETING MECHANISMS AND INITIAL DECISIONS

The **rule-assisted authority policy** combines quote-supported competing hypotheses with public title/context signals and pre-cutoff evidence, excluding target-paper retrieval. Title signals provide an initial prior; supported challengers can revise the choice, with deterministic fallback when support is insufficient. MEC screens contradicted candidates, while relation-level admission governs subsequent edits. Mechanism distributions and experiment rankings are separately committed before observation. The 125-task evaluation reuses hypothesis packages for pre-observation decisions (Appendix F.3).

Separately, archival graph composition ranks candidates from supplied edges by bottleneck coverage, coherence, actionability, contradiction, and complexity, retaining distinct structural representatives for subsequent testing (Appendix C).

3.3 PREDICTIVE WORLD MODELS AND DISCRIMINATING EXPERIMENTS

The **Biological World Model (BWM)** forecasts context-specific intervention responses. Its response-decision realization fits source-separated joint Gaussian beliefs over transformed effects and measurement error, including shared controls. Permitted measurements update beliefs about correlated, unmeasured interventions; finite lookahead weighs expected decision value against measurement cost and stopping (Frazier et al., 2009). Recommendations use the original effect scale.

The **Operational World Model (OWM)** predicts evidence change, obligation closure, cost, and failure from pre-action state, action semantics, and compatible verified experience. Support and error checks govern forecast use; imagined transitions remain planning states, and only verified observations update the empirical record (Appendix D).

Archival rival-aware testing binds graph-to-readout effects and confidences to mechanism versions and experiments, separately from fitted BWM likelihoods. Let $D_{ij}(e)$ denote confidence- and causal-level-weighted disagreement between mechanisms i, j over shared predicted readouts. The core score is

$$A_{\mathrm{mech}}(e) = C_{\mathrm{pred}}(e)\Big[w_1\overline{D}(e) + w_2 D_{\min}(e) + w_3 R_{\mathrm{retire}}(e) + w_4 L_{\mathrm{edge}}(e) + w_5 D_{\mathrm{lock}}(e) + w_6 V_{\mathrm{readout}}(e)\Big] - \lambda C(e). \tag{3}$$

The terms reward mean and worst-pair separation, anticipated rival retirement, disputed-edge localization, residual-barrier diagnosis, and readout discrimination; $C(e)$ is cost. All precede observation access. Coverage is $C_{\mathrm{pred}}(e) = |\mathcal{H}_e|/|\mathcal{H}_t|$, where $\mathcal{H}_t$ is the current mechanism set and $\mathcal{H}_e$ its members with predictions for e. If any mechanism lacks a prediction, $D_{\min} = 0$, preventing missing rivals from inflating worst-pair separation. Feasibility adjustments, score bounds, and deterministic tie-breaking complete selection (Guo et al., 2026) (Appendix C.2).

3.4 Observation-licensed structural revision

The structural kernel matches readouts by experiment, normalized identity, and causal level, separating **agreement, coverage, and biological success**. Agreement with a predicted null is not prerequisite success; missing observations remain unknown without erasing observed contradictions. With confirmed engagement, proximal contradictions implicate corresponding relations; concurrent distal success motivates an alternative explanation. Proximal success followed by functional failure motivates a residual-barrier hypothesis. When observations are insufficient for a diagnosis, topology is retained and missing readouts requested.

Given diagnosis d_i, implicated relations $\mathcal{I}_i$, edge library L, updated bottlenecks $\mathcal{B}'_t$, evidence $\mathcal{E}_t$, and matched observation Y_t, eligible revisions follow

$$\begin{aligned} \Delta_i &= \mathrm{Propose}\left(G_i^{(v)}, d_i, \mathcal{I}_i, L, \mathcal{B}'_t\right), \\ G_i^{(v+1)} &= \mathrm{Apply}\left(G_i^{(v)}, \mathrm{Admit}_{\mathrm{MEC}}\left(\Delta_i; \mathcal{E}_t, c_i, Y_t\right)\right). \end{aligned} \tag{4}$$

At most three operations delete, replace, split, or insert relations. Admission binds the relation payload, parent, observation request, context, and evidence snapshot; application rechecks these bindings. Replacement and splitting are atomic: all additions must be admitted before the original relation is removed. Independent semantic checking reconstructs coverage, admissibility, and the child. Unresolved proposals remain in the audit lineage without licensing revised predictions or programs. Eligible children retain parent–observation–edit lineage and regenerate graph-conditioned predictions from the bound provider; posteriors additionally require child-specific likelihood bindings (Appendix C.3).

A separate **matched two-stage language-model realization** varies edit permission in stage one; stage two receives the working graph under identical probability and action rules, without arm labels. Parent or cross-family graph substitution changes only the working-graph input, testing output dependence separately from kernel admission guarantees (Section 4.2; Appendix C.5).

3.5 Mechanism-derived intervention programs

The compiler selects an eligible mechanism by diagnosis, primary-axis coverage, admitted additions, and support. It greedily covers primary and residual targets with atomic interventions, applies deterministic specificity, stage, and risk preferences, and removes redundancy. Probed actions require observed engagement and proximal response; dysfunction and rebound can motivate effector rescue and maintenance, respectively (Appendix C.4).

Primary-only, residual-only, joint, and full-program contrasts preserve experimental commitments. Graph-conditioned and action-composition forecasts remain distinct; different graphs can yield identical programs. Source mechanism, lineage, ordered actions, and predictions are committed before archival lookup, which assesses perturbation-class correspondence under the historical contract (Appendix H).

3.6 ADAPTIVE EXECUTION AND EXPERIENCE-DRIVEN LEARNING

Observation-conditioned execution. WHAT specifies the scientific goal and completion contract; HOW compiles the next permitted operation from verified state S_t: evidence, mechanisms, predictions, obligations, memory, trajectory, and resources. An intention hierarchy links goals, procedures, capabilities, and tools (Yao et al., 2023; Yang et al., 2024; Xiao et al., 2026); verified effects trigger reassessment or repair.

Governed memory reopens source pointers, checks identity and eligibility, routes identified conflicts to counterevidence, and withdraws stale or invalid pointers from active support. Evidence memory retains source provenance; operational memory retains verified action consequences. Progressive disclosure exposes relevant context; skills bind applicability, required operations, artifacts, and recovery conditions (Wang et al., 2023; Chen et al., 2026). Completion requires verified effects and artifacts.

Scientific continuation and stopping. Let s_t denote verified scientific and operational sufficiency, k_t reachable critical work, and r_t a reachable high-value residual question. With $\nu_t = k_t \vee (s_t \wedge r_t)$, after pending effects settle,

$$d_t = \begin{cases} \text{CONTINUE}, & \nu_t = 1 \text{ and budget permits}, \\ \text{COMPLETE}, & s_t = 1 \text{ and } \nu_t = 0, \\ \text{BOUNDED_INCOMPLETE}, & \text{otherwise}. \end{cases} \tag{5}$$

Thus, residual value can justify continuing beyond sufficiency; termination preserves unresolved obligations (Appendix E.3).

Learning across investigations. A masked recurrent policy selects search, verify, reject, or expand from pre-action history $x_{\le t}$. Evaluator-derived targets $a_t^\star$ use adjudicated support and retrieval conditions; targets and current outcomes never enter policy inputs. Over valid training steps $\mathcal{V}$,

$$\mathcal{L}(\theta) = -\frac{\sum_{t\in\mathcal{V}} q_t w_t \log \pi_\theta(a_t^\star \mid x_{\le t})}{\max\{1, \sum_{t\in\mathcal{V}} q_t w_t\}} + \frac{\lambda_{\mathrm{KL}}}{\max\{1, |\mathcal{V}|\}} \sum_{t\in\mathcal{V}} D_{\mathrm{KL}}(\mu_t \,\|\, \pi_\theta(\cdot \mid x_{\le t})) \,. \tag{6}$$

Here, μ_t is the logged behavior distribution and $\lambda_{\mathrm{KL}} = 0.35$. Legal-action masks precede normalization. Forced actions have $q_t = 0$; otherwise, q_t uses normalized inverse-square-root target-action frequencies from the fitting split. Supervision uses $w_t = 1$; the AWR-inspired variant (Peng et al., 2019) uses $w_t = \min\{\exp[(R(\tau) - b_t)/\beta], w_{\max}\}$, with registered return $R(\tau)$, fitting-set mean-return baseline b_t within task-family and target-action strata, $\beta = 1$, and $w_{\max} = 12$. Architecture, data, and regularization are matched. Policies are fitted offline and frozen for evaluation; foundation-model weights remain fixed (Appendix E.4).

Mechanisms adapt within investigations; execution preferences are learned across investigations.

4 EXPERIMENTS

Study-specific protocols use recorded Qwen3.8-27B backends and task- or source-grouped seed averages (Appendices F and I). Useful scientific outcomes remain distinct from trajectory completion (Liu et al., 2024; Ruan et al., 2024).

4.1 INITIAL MECHANISM AND INTERVENTION DECISIONS

Seven configurations, 125 contexts, and two seeds yield **1,750 scored records**. Titles, pre-cutoff evidence, and menus are supplied; target-paper retrieval and private summaries, labels, and outcomes are excluded. Endpoints are reference-bottleneck and exact experiment-ID agreement with deterministic projections of curated summaries. External outputs use a shared projector; Mechanova scores rule-assisted pre-observation decisions from cached hypotheses. Figure 2(a,b) compares bounded configurations with configuration-specific compute budgets; panel (c) reports matched editing under shared downstream rules.

Mechanova achieves **60.8% mechanism and 52.0% intervention agreement**, exceeding the strongest completed comparators, Biomni and OpenScientist respectively, by **19.6 points** [9.6, 29.2]

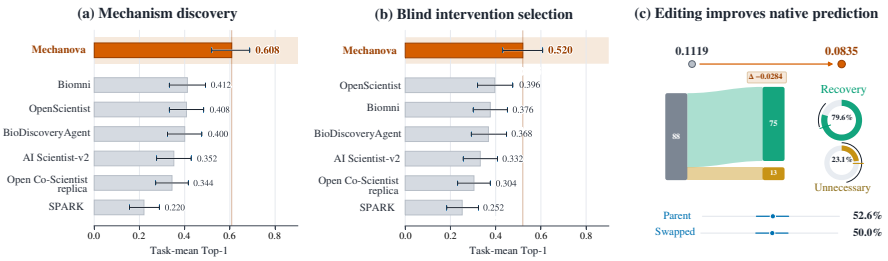

Figure 2: **Initial scientific decisions and matched editing.** (a,b) Top-1 on 125 tasks, two seeds; whiskers are marginal task-bootstrap 95% intervals. Open Co-Scientist is the public replica. (c) Matched editing on 88 families (Table 1).

Table 1: **Editing improves native prediction under matched decision rules.** 88 families; 264 records/arm. Effects: edit-on minus edit-off. Intervals: 10,000 paired family-bootstrap draws. Graph changes count recommendation *or* rounded probability changes in eligible families.

Endpoint / diagnostic	Estimate	95% CI
Native Brier: edit-off → edit-on	$0.1119 \rightarrow 0.0835$	—
Paired native Brier	$\mathbf{-0.0284}$	$[\mathbf{-0.0439}, \mathbf{-0.0146}]$
Parent replacement: output change (64 families)	52.6%	[40.6, 64.1]
Swapped-child replacement: output change (64)	50.0%	[38.0, 61.5]
Edit occurrence when needed (75 families)	79.6%	[70.2, 88.0]
Unnecessary editing (13 no-edit families)	23.1%	[0.0, 46.2]
Secondary native regret difference	−0.0290	$[-0.0886, 0.0268]$
Secondary frozen graph-Brier difference	+0.0186	$[-0.0111, 0.0471]$

and **12.4 points** [2.4, 22.4]. Gains over frozen `title_v1` are **5.6 and 1.6 points**, with fourteen and four net additional correct decisions (Table 11). External-comparator intervals use 10,000 paired task-bootstrap draws (Appendix F.3).

4.2 Controlled attribution of structural revision

The two-stage comparison uses 88 source families, two arms, and three seeds: **528 factual records**. Evidence, initial graphs, trigger observations, event descriptions, and panels are shared. Stage one varies edit permission; edit-off preserves the original triples. Stage two receives the working graph without arm labels under identical probability rules and a two-action cap. Native Brier scores withheld binary events; unknown outcomes are excluded and missing forecasts receive probability 0.5.

Editing reduces native Brier by **25.3%** under shared downstream rules (Figure 2c and Table 1). **376 graph-substitution calls** (188 trials / 64 families) establish downstream graph dependence. Edit occurrence quantifies structural activity; relation correctness is a separate endpoint. Secondary-regret and frozen graph-readout Brier intervals include zero (Appendix G.8).

4.3 Evidence-grounded reasoning and selective memory

Forty-five questions (five axes, nine lenses) use matched twelve-record packets (Lewis et al., 2020; DeYoung et al., 2020; Gao et al., 2023; Asai et al., 2024): eight configurations and three seeds produce **1,080 parent outputs plus 135 compiled outputs**. Against full-reference, compilation raises Recall@10 from **0.252 to 0.368** and supported yield from 0.108 to 0.190; unsupported assertions fall from 0.516 to **0.000**. Outputs average 2.52 cards, none empty; parent-inclusive tokens fall from 24,307 to 8,860. Source support and target recovery remain distinct; negative-target nonpromotion weakens (Appendix G.2).

Separate MEC ablations (144 outputs) distinguish typed grounding from counterevidence effects. Memory evaluation (432 executions) crosses three policies, four pointer conditions, twelve groups,

Table 2: **Capabilities supporting sustained investigation.** Study-specific paired effects; positive favors the component. MEC is developmental; full endpoints remain in Appendix G.3.

Component − reference (units)	Endpoint	Difference [95% CI]
Compiler − full-reference (45 tasks)	Recall@10	$+0.1167$ $[0.0391, 0.1959]$
MEC − no MEC (12 groups)	Typed grounding	$+0.6022$ $[0.5129, 0.6842]$
Full − no counterevidence (12 groups)	Contradiction recall	$+0.3053$ $[0.2332, 0.3813]$
Governed − episodic memory (12 groups)	Utility	$+0.0524$ $[0.0496, 0.0550]$
OWM memory/reflection − dynamics (24 groups)	Decision utility	$+0.1400$ $[0.0963, 0.1750]$
EIG − random queries (30 models)	Log score	$+0.0635$ $[0.0262, 0.1011]$
Hierarchy − flat loop (12 groups)	Completion (pp)	$+50.0$ $[25.0, 75.0]$
Supervised GRU − linear (106 groups)	Execution score	$+0.2942$ $[0.2129, 0.3759]$

Table 3: **Information-guided acquisition across budgets and response panels.** A: 30 models / 36 interventions; Top-1 (%), primary budget: five queries. B: developmental source-macro means, three groups / 31 overlapping views / eight panels; targets unmeasured.

A. Mechanism identification at 1 / 3 / 5 / 8 queries

Policy	1	3	5	8	Log score (5)↑
EIG	**43.3**	**63.3**	**56.7**	**60.0**	-1.0153
Random	37.5	40.9	42.9	45.6	-1.0788
Uncertainty only	36.7	43.3	33.3	46.7	-1.0968
Metadata-family cycle	30.0	36.7	40.0	53.3	-1.0840

B. Protected-response decisions and measurement costs

Policy	Regret↓	Queries	Cost	Top-1 (%)↑	Regret + cost↓
No queries	0.04675	0.000	0.00000	74.37	0.04675
Random queries	0.01722	1.556	0.02542	83.07	0.04264
Full active	**0.01652**	0.809	0.00886	**83.41**	**0.02538**

and three seeds. Its gains concentrate under stale, conflicting, and poisoned pointers (Table 2; Appendix G.3).

4.4 World models and informative acquisition

Hidden-dynamics OWM uses 24 locked groups and **216 executions**, excluding evaluation episodes from memory. Evidence-prefix control separately saves **20% of panels** on 80 DepMap models per release while meeting 0.01 Brier-utility noninferiority in 26Q1 and exposure-disjoint 24Q2 (Tsherniak et al., 2017).

Cross-screen, lineage-disjoint PRISM/DepMap comparisons (Corsello et al., 2020; Tsherniak et al., 2017) compare expected-information-gain (EIG) acquisition with random, uncertainty, and metadata policies (Walker and Ravisankar, 2019; Foster et al., 2020) (Table 3). Protected-target development achieves similar Top-1 to random querying with fewer measurements; one-step and full lookahead coincide. Four Replogle backgrounds contribute **4,058 paired perturbation–context records**, with Pearson gains of 0.0209–0.0844 in separate representation comparisons (Cui et al., 2024; Theodoris et al., 2023; Bunne et al., 2023; Lotfollahi et al., 2023; Replogle et al., 2022) (Appendix G.5).

4.5 Adaptive execution and experience-driven learning

Both controllers complete 18/18 clean trials; hierarchy recovers 18/18 disrupted trials after flat-loop context exhaustion. Disclosure and skills test separate interventions (Schick et al., 2023; Shinn et al., 2023) (Table 4A). On eight development tasks, adaptive horizons preserve essential slots using 177,485 versus 258,068 tokens (**31.2% saving**).

Offline-fitted policies execute source-disjoint tasks (Kumar et al., 2020; Kostrikov et al., 2022; Chen et al., 2021) (Table 4B). Supervised GRU improves research-action selection; AWR adds no reliable

Table 4: **Execution, experience-driven learning, and archival program assessment.** A: separate 12-group, three-seed studies; group-paired intervals. B: 450 episodes / 150 tasks / 106 groups / 573 requests; task means, with terminal decisions included in calls. C: 30 archival assessments / ten contexts / seven families.

A. Confirmation completion: reference → component

Comparison	Reference	Component	Gain in pp [95% CI]
Hierarchy vs. flat	18/36	**36/36**	$+50.0$ $[25.0, 75.0]$
Disclosure vs. full inventory	18/36	**32/36**	$+38.9$ $[25.0, 55.6]$
Outcome vs. protocol-only skills	8/36	**34/36**	$+72.2$ $[52.8, 88.9]$

B. Frozen policy execution after fitting 337 trajectories / 74 groups

Policy	Score↑	Recovery (%)↑	Nonnecessary (%)↓	Calls	Errors
Linear	0.4453	32.67	8.00	2.35	0
Supervised GRU	**0.7540**	**71.33**	**5.00**	**2.23**	0
GRU + AWR	0.7537	**71.33**	5.33	2.24	0

C. Proxy-assisted program assessment; predictions frozen before lookup

Policy	Evaluable / all	Success / all	Success / evaluable
Mechanova Full	10/10	**10/10**	**10/10**
Static bottleneck	9/10	0/10	0/9
Checkpoint only	10/10	0/10	0/10

score gain over matched supervision, $+0.0055$ $[-0.0154, 0.0300]$. These actual executions remain separate from historical log reweighting (Appendix G.7).

4.6 Mechanism-derived programs and scientific traces

Ten archived PDAC contexts evaluate programs committed before functional-outcome lookup (Jiang et al., 2016; Feig et al., 2013; Shakiba and Tuveson, 2025; Zhu et al., 2014; Beatty et al., 2011). The proxy-assisted predicate combines five benefit floors, immune dependence, and risk ceilings (Table 4C). Static policies select one bottleneck action or checkpoint rescue without observation-conditioned combinations.

FAK–KPC engagement without downstream rescue eliminates two of four rivals. The unchanged graph yields stromal remodeling plus effector rescue, matching FAK-plus-PD-1 classes (4/5 forecasts agree). This evaluates rival elimination and intervention-class program synthesis (Appendix H).

5 Discussion and conclusion

Mechanova connects evidence, explanation, and experimental commitment through persistent, testable mechanisms. The Mechanism Evidence Chain grounds causal claims in evidence and counterevidence; rival-aware tests distinguish explanations; evidence-licensed revision turns discrepancies into structural changes and new prediction obligations. Surviving mechanisms generate intervention programs addressing primary and residual barriers.

Mechanova leads the completed 125-task comparisons and improves both initial-decision endpoints over frozen title-only routing (Section 4.1). In the separate 88-family study, editing reduces native Brier by **25.3% under identical downstream probability and action rules**, while graph-substitution controls establish downstream dependence on the working mechanism (Section 4.2).

Evidence compilation, governed memory, informative acquisition, and hierarchical recovery sustain investigation (Sections 4.3–4.5). Scientific sufficiency and residual value guide continuation; offline policy learning improves execution across investigations, complementing mechanism revision within them. With PDAC as its flagship testbed, Mechanova makes this cycle **explicit, executable, and measurable**.

AI USE STATEMENT

Generative AI tools assisted literature discovery and synthesis, refinement of methodological ideas and hypotheses, implementation and debugging of research code, interpretation of recorded results, and manuscript drafting, translation, editing, and figure preparation. Model-assisted evidence and annotation proposals retain their provenance and are distinguished from human-reviewed records (Appendix F.1). Language models also serve as experimental components of Mechanova; these roles and their study-specific configurations are described in Sections 3–4 and Appendix I. The authors retain full responsibility for verifying AI-assisted contributions and for the final text, citations, methods, reported results, and accompanying artifacts.

REPRODUCIBILITY STATEMENT

Section 3 and Appendices A–E describe the algorithms and study-specific implementations. Section 4 and Appendix F specify task construction, information-access restrictions, comparison settings, and statistical procedures. Appendix I provides the implementation map, model and decoding settings, random seeds, source and result hashes, and commands for reconstructing reported statistics from saved records. It distinguishes statistical reconstruction, deterministic replay, and fresh model execution, together with their respective dependency requirements. These references connect each reported study to its recorded configuration, inputs, outputs, and scoring procedure.

REFERENCES

Abhinav K. Adduri et al. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 189(19):5914–5931.e20, 2026. doi: 10.1016/j.cell.2026.07.052. URL https://doi.org/10.1016/j.cell.2026.07.052. Available online August 31, 2026.

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

Gregory L. Beatty, Elena G. Chiorean, Matthew P. Fishman, et al. CD40 agonists alter tumor stroma and show efficacy against pancreatic carcinoma in mice and humans. *Science*, 331(6024):1612–1616, 2011. doi: 10.1126/science.1198443. URL https://doi.org/10.1126/science.1198443.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Raetsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20:1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

Daniel S. Chen and Ira Mellman. Oncology meets immunology: The cancer-immunity cycle. *Immunity*, 39(1):1–10, 2013. doi: 10.1016/j.immuni.2013.07.012. URL https://pubmed.ncbi.nlm.nih.gov/23890059/.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/7f489f642a0ddb10272b5c31057f0663-Abstract.html.

Guoxin Chen, Zile Qiao, Xuanzhong Chen, Donglei Yu, Haotian Xu, Wayne Xin Zhao, Ruihua Song, Wenbiao Yin, Huifeng Yin, Liwen Zhang, Kuan Li, Minpeng Liao, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. IterResearch: Rethinking long-horizon agents with interaction scaling. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2511.07327.

Steven M. Corsello, Rohith T. Nagari, Ryan D. Spangler, et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, 1:235–248, 2020. doi: 10.1038/s43018-019-0018-6. URL https://doi.org/10.1038/s43018-019-0018-6.

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.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.bionlp-1.13. URL https://aclanthology.org/2020.bionlp-1.13/.

Christine Feig, James O. Jones, Matthew Kraman, Richard J. B. Wells, Andrew Deonarine, Derek S. Chan, Claire M. Connell, Edward W. Roberts, Qi Zhao, Otavia L. Caballero, Sarah A. Teichmann, Tobias Janowitz, Duncan I. Jodrell, David A. Tuveson, and Douglas T. Fearon. Targeting CXCL12 from FAP-expressing carcinoma-associated fibroblasts synergizes with anti-PD-L1 immunotherapy in pancreatic cancer. *Proceedings of the National Academy of Sciences*, 110(50):20212–20217, 2013. doi: 10.1073/pnas.1320318110.

Adam Foster, Martin Jankowiak, Matthew O'Meara, Yee Whye Teh, and Tom Rainforth. A unified stochastic gradient approach to designing bayesian-optimal experiments. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 2020. URL https://proceedings.mlr.press/v108/foster20a.html.

Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing*, 21(4):599–613, 2009. doi: 10.1287/ijoc.1080.0314. URL https://doi.org/10.1287/ijoc.1080.0314.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://aclanthology.org/2023.acl-long.910/.

Alireza Ghafarollahi and Markus J. Buehler. SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):e2413523, 2025. doi: 10.1002/adma.202413523. URL https://doi.org/10.1002/adma.202413523.

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.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655:487–496, 2026. doi: 10.1038/s41586-026-10644-y. URL https://www.nature.com/articles/s41586-026-10644-y.

Yujia Guo, Daolang Huang, Xinyu Zhang, Sammie Katt, Samuel Kaski, and Ayush Bharti. Constrained bayesian experimental design via online planning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.26990.

Benjamin M. Gyori, John A. Bachman, Kartik Subramanian, Jeremy L. Muhlich, Lucian Galescu, and Peter K. Sorger. From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology*, 13(11):954, 2017. doi: 10.15252/msb.20177651. URL https://doi.org/10.15252/msb.20177651.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. doi: 10.48550/arXiv.1803.10122. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. URL https://arxiv.org/abs/2301.04104.

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

Hong Jiang, Samarth Hegde, Brett L. Knolhoff, Yu Zhu, John M. Herndon, Melissa A. Meyer, Timothy M. Nywening, William G. Hawkins, Irina M. Shapiro, David T. Weaver, Jonathan A. Pachter, Andrea Wang-Gillam, and David G. DeNardo. Targeting focal adhesion kinase renders pancreatic cancers responsive to checkpoint immunotherapy. *Nature Medicine*, 22:851–860, 2016. doi: 10.1038/nm.4123.

Jiajie Jin, Yuyang Hu, Kai Qiu, Qi Dai, Chong Luo, Guanting Dong, Xiaoxi Li, Tong Zhao, Xiaolong Ma, Gongrui Zhang, Zhirong Wu, Bei Liu, Zhengyuan Yang, Linjie Li, Lijuan Wang, Hongjin Qian, Yutao Zhu, and Zhicheng Dou. Toward generalist autonomous research via hypothesis-tree refinement. *arXiv preprint arXiv:2606.11926*, 2026. URL https://arxiv.org/abs/2606.11926.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html.

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, pages 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. doi: 10.1214/aoms/1177728069. URL https://doi.org/10.1214/aoms/1177728069.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e9df36b21ff4ee211a8b71ee8b7e9f57-Abstract-Conference.html.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, Mohsen Naghipourfar, Riza M. Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19:e11517, 2023. doi: 10.15252/msb.202211517.

A. Rupam Mahmood, Hado van Hasselt, and Richard S. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/1c64ee92596e8ea5050fc435a1d57459-Abstract.html.

Sanjeev Mariathasan, Shannon J. Turley, Dorothee Nickles, Alessandra Castiglioni, Kobe Yuen, Yulei Wang, Edward E. Kadel III, Hartmut Koeppen, Jillian L. Astarita, Rafael Cubas, Suchit Jhunjhunwala, Romain Banchereau, Yagai Yang, Yinghui Guan, Cecile Chalouni, James Ziai, Yasin Senbabaoglu, Stephen Santoro, Daniel Sheinson, Jeffrey Hung, Jennifer M. Giltnane, Andrew A. Pierce, Kathryn Mesh, Steve Lianoglou, Johannes Riegler, Richard A. D. Carano, Pontus Eriksson, Mattias Hoglund, Loan Somarriba, Daniel L. Halligan, Michiel S. van der Heijden, Yohann Loriot, Jonathan E. Rosenberg, Lawrence Fong, Ira Mellman, Daniel S. Chen, Marjorie Green, Christina Derleth, Gregg D. Fine, Priti S. Hegde, Richard Bourgon, and Thomas Powles. TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature*, 554:544–548, 2018. doi: 10.1038/nature25501.

B. McClellan, Q. Wang, K. Aung, and W. Matsui. Quiescent tumor cells shape the immunosuppressive microenvironment in pancreatic cancer. *Nature Communications*, 17:8889, 2026. doi: 10.1038/s41467-026-75883-z. URL https://pubmed.ncbi.nlm.nih.gov/42481500/.

Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C. Landsness, Daniel L. Barabasi, Siddharth Narayanan, Nicky Evans, et al. Kosmos: An AI scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*, 2025. URL https://arxiv.org/abs/2511.02824.

Dirk Ourston and Raymond J. Mooney. Theory refinement combining analytical and empirical methods. *Artificial Intelligence*, 66(2):273–309, 1994. doi: 10.1016/0004-3702(94)90028-0. URL https://doi.org/10.1016/0004-3702(94)90028-0.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019. URL https://arxiv.org/abs/1910.00177.

John R. Platt. Strong inference. *Science*, 146(3642):347–353, 1964. doi: 10.1126/science.146.3642.347. URL https://pubmed.ncbi.nlm.nih.gov/17739513/.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, et al. Mapping information-rich genotype–phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013. URL https://doi.org/10.1016/j.cell.2022.05.013.

Kaleigh F. Roberts, Zachary B. Abrams, Luca Cappelletti, Mahdi Moqri, Nicholas Heugel, J. Harry Caufield, Mathieu Bourdenx, Yan Li, Jineta Banerjee, Luca Foschini, Diego Galeano, Nomi L. Harris, Melody Li, Kejun Ying, Justin A. Melendez, Nicolas R. Barthélemy, James G. Bollinger, Yingxin He, Vitaliy Ovod, Tammie L. S. Benzinger, Shaney Flores, Brian A. Gordon, Adegoke A. Ojewole, Mukta Phatak, Donald L. Elbert, Sarah Biber, Eric C. Landsness, Christopher J. Mungall, Randall J. Bateman, and Justin T. Reese. OpenScientist: Evaluating an open agentic AI co-scientist to accelerate biomedical discovery. *medRxiv*, 2026. doi: 10.64898/2026.03.15.26348338. URL https://www.medrxiv.org/content/10.64898/2026.03.15.26348338v1.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42:927–935, 2024. doi: 10.1038/s41587-023-01905-6. URL https://www.nature.com/articles/s41587-023-01905-6.

Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. BioDiscoveryAgent: An AI agent for designing genetic perturbation experiments. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4252dc94531833029000f85dc5fac792-Abstract-Conference.html.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7274ed909a312d4d869cc328ad1c5f04-Abstract-Conference.html.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

Mojdeh Shakiba and David A. Tuveson. Macrophages and fibroblasts as regulators of the immune response in pancreatic cancer. *Nature Immunology*, 26:678–691, 2025. doi: 10.1038/s41590-025-02134-6.

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. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

C. W. Steele, S. A. Karim, J. D. G. Leach, et al. CXCR2 inhibition profoundly suppresses metastases and augments immunotherapy in pancreatic ductal adenocarcinoma. *Cancer Cell*, 29(6):832–845, 2016. doi: 10.1016/j.ccell.2016.04.014. URL https://pubmed.ncbi.nlm.nih.gov/27265504/.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181–211,

1999. doi: 10.1016/S0004-3702(99)00052-1. URL https://www.sciencedirect.com/science/article/pii/S0004370299000521.

Yingying Tang, Shijie Tang, Wenjuan Yang, et al. MED12 loss activates endogenous retroelements to sensitise immunotherapy in pancreatic cancer. *Gut*, 73(12):1999–2011, 2024. doi: 10.1136/gutjnl-2024-332350. URL https://pubmed.ncbi.nlm.nih.gov/39216984/.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. 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.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2139–2148, 2016. URL https://proceedings.mlr.press/v48/thomasa16.html.

Christian Toth, Lars Lorch, Christian Knoll, Andreas Krause, Franz Pernkopf, Robert Peharz, and Julius von Kügelgen. Active Bayesian causal inference. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2206.02063.

Florian Trost, Bide Zhang, Ines Aring, Marcus Bauer, Lennert Glamann, Michael Wessolly, Kyra Johnson, Heike Göbel, Tristan Lerbs, Taban Sangenne, Peter Herrmann, et al. An agentic framework for autonomous scientific discovery in cancer pathology. *Nature Medicine*, 32:2254–2266, 2026. doi: 10.1038/s41591-026-04357-y.

Aviad Tsherniak, Francisca Vazquez, Phil G. Montgomery, et al. Defining a cancer dependency map. *Cell*, 170(3):564–576.e16, 2017. doi: 10.1016/j.cell.2017.06.010. URL https://doi.org/10.1016/j.cell.2017.06.010.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Eric A. Walker and Kishore Ravisankar. Bayesian design of experiments: Implementation, validation and application to chemical kinetics. *arXiv preprint arXiv:1909.03861*, 2019. doi: 10.48550/arXiv.1909.03861. URL https://arxiv.org/abs/1909.03861.

Fiona Y. Wang and Markus J. Buehler. Self-revising discovery systems for science: A categorical framework for agentic artificial intelligence. *arXiv preprint arXiv:2606.01444*, 2026. URL https://arxiv.org/abs/2606.01444.

Haorui Wang, Parshin Shojaee, Kazem Meidani, Kunyang Sun, Jose Miguel Hernandez-Lobato, Teresa Head-Gordon, Jiajun He, Chandan K. Reddy, Chao Zhang, and Yuanqi Du. Towards diverse scientific hypothesis search with large language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.10587.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023. URL https://arxiv.org/abs/2305.16291.

Yixuan Weng, Minjun Zhu, Qiujie Xie, Qiyao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. DeepScientist: Advancing frontier-pushing scientific findings progressively. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4f64494ecc3442f1c9261baa036378bc-Abstract-Conference.html.

Rafael Winograd et al. Induction of T-cell immunity overcomes complete resistance to PD-1 and CTLA-4 blockade and improves survival in pancreatic carcinoma. *Cancer Immunology Research*, 3(4):399–411, 2015. doi: 10.1158/2326-6066.CIR-14-0215. URL https://pubmed.ncbi.nlm.nih.gov/25678581/.

Qiao Xiao, Haochen Shi, Yisen Gao, Wenbin Hu, Huihao Jing, Tianshi Zheng, Baixuan Xu, Ziheng Zhang, Weiqi Wang, Haoran Li, Jiaxin Bai, and Yangqiu Song. SING: Synthetic intention graph for scalable active tool discovery in LLM agents. *arXiv preprint arXiv:2606.16591*, 2026. URL https://arxiv.org/abs/2606.16591.

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. doi: 10.48550/arXiv.2504.08066. URL https://arxiv.org/abs/2504.08066.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. *arXiv preprint arXiv:2405.15793*, 2024. URL https://arxiv.org/abs/2405.15793.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Yu Zhu, Brett L. Knolhoff, Melissa A. Meyer, Timothy M. Nywening, Brian L. West, Jingqin Luo, Andrea Wang-Gillam, S. Peter Goedegebuure, David C. Linehan, and David G. DeNardo. CSF1/CSF1R blockade reprograms tumor-infiltrating macrophages and improves response to T-cell checkpoint immunotherapy in pancreatic cancer models. *Cancer Research*, 74:5057–5069, 2014. doi: 10.1158/0008-5472.CAN-13-3723. URL https://doi.org/10.1158/0008-5472.CAN-13-3723.

ADDITIONAL MATERIAL

This supplement expands the mechanism-centered scientific cycle of **Mechanova**. Appendices A–E specify scientific state, evidence compilation, structural revision, biological and operational world models, and adaptive execution and learning. Appendices F–I provide benchmark construction, extended results, archival trajectories, and reproducibility details. The supplement follows the same evidential structure as the main text: initial scientific decisions, controlled attribution of structural revision, and the capabilities that sustain investigation. Each study retains its own inputs, comparator, scoring rule, and analysis unit. Figure 2 and Table 1 summarize the two central evaluations; the appendices below specify their mechanisms and reproducible consequences.

A EXECUTABLE SCIENTIFIC STATE AND UPDATE SEMANTICS

Mechanova maintains an explanation together with the evidence that licenses it, the observations that could challenge it, and the intervention commitments it produces. This appendix specifies how those objects remain connected as an investigation changes; the transformations themselves are detailed in Appendices B–E.

A.1 PERSISTENT SCIENTIFIC OBJECTS AND THEIR BINDINGS

The episode interface is

$$S_t = (\mathcal{E}_t, \mathcal{H}_t, \mathcal{P}_t, \mathcal{O}_t, \mathcal{M}_t, \mathcal{A}_t, \mathcal{T}_t, B_t^{\mathrm{rem}}), \tag{7}$$

where $\mathcal{E}_t$ contains source-bound evidence, $\mathcal{H}_t$ versioned mechanisms, $\mathcal{P}_t$ predictions, and $\mathcal{O}_t$ unresolved scientific obligations. Memory $\mathcal{M}_t$, available actions $\mathcal{A}_t$, the verified trajectory $\mathcal{T}_t$, and remaining budget B_t^{rem} support execution. This notation links typed scientific records across study-specific implementations; each experiment retains its own state representation and serialization.

The mechanism-local evidence and predictions in Equation 1 are linked through evidence, experiment, context, and mechanism-version identities. Reusing a source pointer does not create independent support. An atomic-action catalogue specifies what can be composed; a compiled intervention program is a separate artifact retaining its source mechanism, revision lineage, selected actions, and prediction commitment. The trajectory connects these artifacts to the observations and operations that produced them.

Evidence admission changes what supports a claim; a graph edit changes its explanation; prediction regeneration changes its testable consequences; verified execution changes the operational record. These updates have separate justifications. A tool's successful completion alone neither licenses a biological claim nor converts a forecast into an observation. Experiment-specific providers and assessment bindings are given in Appendix I.1.

A.2 ECOSYSTEM STATE AND BOTTLENECK TOPOLOGY

In the evaluated archival realization, numerical diagnosis uses a task-specific ecosystem vector $x_t \in [0,1]^{13}$, supplied by the baseline packet and subsequently updated from permitted readouts. Higher values indicate competence or barrier relief, except for adverse-direction rebound pressure. These are declared state encodings, not response fractions. Seven coarse-grained diagnostic axes summarize the vector: antigenicity, antigen presentation, priming, recruitment, infiltration, suppression, and dysfunction/escape. They provide an initial scaffold for organizing mechanisms, rather than an exhaustive biological taxonomy of immune-cold-to-hot conversion.

Antigenicity, presentation, priming, and dysfunction competence use their corresponding tumor, dendritic-cell, and T-cell fields directly. Recruitment takes the minimum of chemokine recruitment and vascular access; infiltration takes the minimum of stromal permissiveness and T-cell infiltration; suppression competence takes the minimum of myeloid permissiveness and metabolic fitness. Driver identities remain available beneath these aggregates, distinguishing, for example, myeloid from metabolic suppression.

At the research-design level, the agent should use collected evidence to assess whether a newly identified factor refines an existing mechanism, calls for finer within-axis distinctions or cross-axis

dependencies, or motivates a candidate additional dimension. A proposed dimension should specify the explanatory gap, context-matched supporting and opposing evidence, observable readouts, and distinct predictive or intervention consequences. Adding a new mechanism does not by itself require a new axis. Incorporating additional dimensions into numerical diagnosis would require versioned state and readout definitions, topology, and scoring rules; the reported archival evaluation retains the fixed seven-axis projection below.

Inspired by the cancer-immunity cycle (Chen and Mellman, 2013), this archival diagnostic projection orders six axes serially: antigenicity, presentation, priming, recruitment, infiltration, then dysfunction/escape. Suppression is parallel to immune-cell arrival but can explain downstream dysfunction. For a serial axis b, let q_b be competence, ρ_b the minimum competence of its serial ancestors (one when none exist), and ℓ_b its number of serial descendants divided by five. The root-attribution heuristic is

$$r_b = (1 - q_b)(0.55 + 0.25\ell_b + 0.20\rho_b)[1 - 0.48(1 - \rho_b)]. \tag{8}$$

Suppression instead uses $r_b = 0.83(1 - q_b)$. These coefficients are algorithmic choices, not fitted causal probabilities (Appendix I.2).

An axis is deficient below 0.55 and severely deficient below 0.45. A deficient ancestor a shadows downstream root attribution when $r_a \geq 0.82 r_b$; suppression can shadow dysfunction by the same rule. The ranking retains up to three unshadowed deficient roots, with the highest-ranked axis as a fallback when none qualify. Actionable axes retain these roots and, when the dominant root is serial, append its deficient descendants in serial order, up to four axes. Unshadowed roots outside the same serial chain identify parallel barriers.

Explaining a deficit does not discharge its repair obligation. A downstream axis discounted as a root can remain actionable and become a residual barrier after upstream relief. Conversely, the fallback ranking does not force a severe-deficiency finding in an otherwise adequate ecosystem. Optional axis uncertainty reduces recorded confidence without changing competence or root ranking.

A.3 Observation-driven updates and scientific completion

An opened observation updates the fields licensed by its experiment/readout binding. In the archival adapter, a mapped signed readout y_r becomes $x'_j = \mathrm{clip}_{[0,1]}(0.5 + 0.5y_r)$; other fields retain their prior values except for two explicit coupled rules specified in Appendix I.2. The revised vector is diagnosed again before structural revision and program synthesis. This state update is distinct from Bayesian reweighting over a fixed mechanism inventory.

Admitted graph edits retain parent–observation–child lineage and regenerate child-specific predictions; a new graph requires its own likelihood binding before posterior assignment. Program synthesis then binds its actions and predictions to the retained source graph. Graph, prediction, and program changes remain separately attributable: rival elimination need not edit the surviving graph, and action compression can make distinct graphs yield the same program (Appendix H). Final archival outcomes are opened only after program and prediction commitment.

The internal pseudo-hot flag requires recruitment and infiltration at least 0.65, immune-dependent control below 0.40, and suppression or dysfunction competence below 0.45. Internal functional readiness requires all seven competences and immune-dependent control at least 0.60, persistence at least 0.50, and rebound pressure at most 0.40. Neither flag is the final archival success predicate, which uses the separate assessment in Appendix H.

Missing source fields create acquisition obligations; new explanatory edges create follow-up prediction and test requirements; residual functional deficits retain repair obligations. Each is resolved against its corresponding evidence or completion condition, not by confidence, tool success, or budget exhaustion. Termination therefore records sufficient completion or bounded incompleteness, while feasible residual opportunity can justify continuation (Appendix E).

Table 5: **Experimental requirements and executable causal licenses.** Each row requires hop-bound experimental evidence. Predicate names are shown in readable form; the frozen policy supplies their exact identifiers.

Verified requirement	Licensed relation	No automatic promotion to
Observational association	Associated / correlated / co-localized	Perturbational dependence
Single perturbation	Modulates / promotes / suppresses	Direct regulation or mediation
Perturbation ∧ dose/time response	Dose/time-dependent modulation	A stronger causal mechanism
Perturbation ∧ rescue/reversal	Contributes to / partially mediates	Complete mediation
Necessity; sufficiency (separate)	Necessary for; sufficient for	Each other
Mediation ∧ epistasis/ordering	Mediates	Mediation from connectivity alone
Physical association; direct binding (separate)	Physically associates; directly binds	Direct binding from association
Regulatory occupancy	Occupies regulatory region	Expression-effect direction

B Experiment-grounded evidence compilation and selective memory

B.1 Experiment-level compilation and admission

MEC makes the experiment, rather than the citation, the unit of scientific binding, connecting claim verification (Wadden et al., 2020) with executable molecular-model assembly (Gyori et al., 2017). Results, methods, figure panels, and table regions supply fields of one experiment anchor: biological system, perturbation, comparator, dose/time, assay, endpoint, and effect. For field j, the compiler retains

$$a_{e,j} = (v_{e,j}^{\text{source}}, v_{e,j}^{\text{norm}}, \mathcal{R}_{e,j}, \text{role}_{e,j}, \xi_{e,j}, \sigma_{e,j}), \tag{9}$$

where ξ records explicitness and σ binding status. Exact pointers bind source bytes, document representation, node, and text span; source values must replay from those pointers. Normalization preserves the original value rather than supplying missing experimental details.

Binding status is compiler-derived, not extractor-declared. Explicit or cross-region-linked values can become bound; inferred candidates require a versioned inference rule and remain partial. A unique bound value supplies the resolved view; competing bound values remain conflicts. Missing critical groups remain unresolved, with dose/time optionally not required for observational designs. These states constrain downstream admission and expose acquisition needs: a missing comparator cannot be repaired by borrowing another experiment's control. Field binding establishes provenance; stronger causal language requires the additional hop-level checks below.

B.2 Causal licensing and bounded cross-paper synthesis

A proposed hop η binds subject, predicate, object, direction, context, and supporting experiment/evidence identities. Let D_η^+ contain dimensions verified for that hop and $R(p)$ the frozen requirements of predicate p:

$$\mathcal{V}_\eta = \{p : R(p) \subseteq D_\eta^+\}. \tag{10}$$

The 14-dimensional grammar distinguishes present, absent, unavailable, and not-applicable evidence. A failed check establishes absence only when representation coverage permits that conclusion. Missing coverage remains unavailable. Finalization requires endpoint-binding and causal-profile checks for the same hop and question context, together with resolved endpoint identities. Passing evidence for one relation cannot license another. Table 5 summarizes the experiment-specific requirements and the causal relations they license.

Ordering evidence separately licenses upstream/downstream relations. Site-specific regulatory and catalytic dependencies have their own dimensions; orthogonal perturbation adds no standalone predicate. Context compatibility is a further admission constraint, not another causal predicate.

Cross-paper composition defines testable mechanism hypotheses. Cross-paper synthesis joins two to four finalized hops from at least two source bindings. Canonical endpoints must form a contiguous, noncyclic path, without unresolved hop assumptions. At least one configured biological-context field must be present in two supporting experiments and agree; any compared conflict blocks composition. Disease, model, species, and cell/tissue compatibility are checked separately from each hop's local experimental license. The output is a *bounded compositional hypothesis* whose cross-experiment causal and mediation implications motivate the rival-discriminating tests in Appendix C.

B.3 Claim-relative counterevidence and acquisition obligations

Evidence needs specify a question, target context, required experimental classes, and support, counterevidence, and null/exception search routes. These minimum routes cannot be removed by a caller's search hints. Retrieval provenance does not determine polarity: a result returned by a contradiction query is not automatically counterevidence.

A separate verifier binds each evidence item to a specific need and assigns support, contradiction, null/exception, irrelevant, or unresolved status. Null results and context exceptions remain qualified rather than being promoted into direct contradictions. Unresolved fields and relations retain their source and need identities for further acquisition. Shared provenance identifies dependent observations instead of counting repeated extraction as replication. Source corrections, retractions, and invalidations are interpreted through the applicable lifecycle snapshot, preserving the information boundary of the investigation.

B.4 Public-evidence-constrained repair

The 45-question study instantiates a public-span compiler over generated parent cards, distinct from the structured hop compiler above. For claim c, let s_c be effective support, d_c effective contradiction, and $r_c = \max(d_c, 1 - s_c)$ the bounded risk proxy. The source license is

$$L(c) = \mathbf{1}\{\text{citations resolved} \wedge \text{support verdict} \\ \wedge\, s_c \geq 0.60 \wedge r_c \leq 0.15\}. \tag{11}$$

The stricter $r_c \leq 0.03$ flag is recorded separately for risk-safe coverage; r_c is a rule-defined score, not a calibrated error probability. Source-support checking remains separate from evaluator-owned target matching (DeYoung et al., 2020; Gao et al., 2023; Asai et al., 2024).

Algorithm B.1: frozen public-span compilation

Input: public task q, evidence packet P, parent cards C, verifier V, frozen rules θ, capacity $K \leq 10$.

1. $A \leftarrow \mathrm{Audit}_V(C, P)$; $C \leftarrow \mathrm{StripSecondary}(C)$.
2. $\Omega \leftarrow \mathrm{BindSameSourceSpans}(C, P)$, with identity-overlap checks.
3. $S \leftarrow \mathrm{SelectLicensed}_\theta(\{\Omega_c : c \in C\}, A, q, V)$; record unsalvaged parents.
4. $S \leftarrow \mathrm{SourceCapAndDeduplicate}_\theta(S, q, V)$.
5. Apply frozen coverage upgrades, complementary spans, and bounded unused-source additions to S; recheck $L(c)$ after each change.
6. $S \leftarrow \mathrm{FinalProject}_V(S, P, K)$; preserve repair, expansion, and rejection records.
7. **Return** S, source licenses, and measured generation cost; evaluator targets are never inputs.

Coverage rules use public focus terms, source identity, support, and deterministic tie-breaking. A same-source complementary span requires additional focus coverage; bounded replacement rules address specific coverage patterns. These coverage rules were selected during development and frozen before evaluation. Final projection rechecks retained primaries and removes secondary assertions without padding the answer. All 135 evaluated compiler outputs used zero additional generative-editor requests; parent generation and source verification remain distinct costs. Target recovery, nonempty output, assertion support, and cost are reported separately in Appendix G.

B.5 Governed pointer reuse and revalidation

The controlled memory implementation stores claim-free pointers from supported, relevant training outputs: source trial/card identities, study family, mechanism axis, retrieval terms, and citation

identifiers. Snapshots exclude held-out task/family identities; the runtime additionally rejects selected pointers citing the target paper. Selection uses same-axis entries when available, otherwise the full eligible snapshot, ranked by query-term overlap and stable identity, with default top-5 retrieval.

No-memory, episodic, and governed arms share the task, model, retrieval snapshot, verifier, and tool inventory. Under registered perturbation labels, episodic memory exposes the injected pointer as ordinary context. Governed memory routes a conflicting pointer to an explicit counterevidence channel; stale or poisoned pointers leave active support, enter quarantine, and trigger rollback. The controlled intervention measures routing and recovery conditional on the registered hazard labels. Protocols and paired utility are in Appendices F and G.

Before a stored pointer supports a new premise, the system revalidates source identity, verification bindings, and lifecycle eligibility under the current or benchmark-frozen information scope. Revalidation governs eligibility restoration, and refresh remains within the information cutoff of a frozen evaluation. Procedural skills preserve different applicability and recovery contracts and are treated in Appendix E. Evidence therefore earns its inferential role at compilation and must retain that role at reuse.

C Rival-aware testing, structural revision, and program synthesis

C.1 Compose mechanisms that can disagree experimentally

A mechanism is a typed causal subgraph whose edges bind compartments, affected axes, readouts, and intervention targets (Appendices A–B). The composer ranks bounded subsets by

$$J(G) = \mathrm{clip}_{[-1,1]}\left[\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}(G) + 0.06 r_{\mathrm{esc}}\mathbf{1}_{E_{\mathrm{esc}}\neq\varnothing} - 0.28\bar{\kappa} - \Omega(|E(G)|)\right]. \tag{12}$$

The nine features are weighted-axis, root, and cascade coverage; coherence; parallel-lock coverage; compartment span; readout diversity; target coverage; and mean context-weighted edge strength. Their weights are $(.23, .19, .16, .14, .08, .05, .06, .04, .05)$; $\bar{\kappa}$ is mean contradiction burden, r_{esc} escape pressure, and $\Omega(m) = .035(m-3)_+ + .006(m-1)_+$. Connectivity contributes to coherence rather than imposing a universal hard filter.

Enumeration visits $\sum_{j=1}^{k}\binom{n}{j}$ subsets, with $n \le 32$ and $k \le 5$, retaining the highest-scoring representative per structural signature. The archival realization requests four candidates with at most three initial edges. It uses ranked structural representatives; niche-based portfolio retention is a separate extension of diverse hypothesis search (Jin et al., 2026; Wang et al., 2026). Structural difference becomes experimentally useful only through different predictions: two explanations may agree that a first barrier was relieved while disagreeing about whether an additional barrier prevents functional conversion.

C.2 Acquire evidence that separates the hardest remaining rivals

Following discriminating-test reasoning (Platt, 1964), predictions are constructed before experiment selection. The archival provider maps target/disputed-edge overlap and causal level to fixed signed effects $\mu_{ier} \in [-1, 1]$ and confidences c_{ier}. These are declared graph-to-readout rules, not fitted BWM likelihoods. For shared predicted readouts $\mathcal{R}_{ije}$ within experiment e, define

$$D_{ij}(e) = \frac{\sum_{r\in\mathcal{R}_{ije}} w_{ijr}|\mu_{ier} - \mu_{jer}|/2}{\sum_{r\in\mathcal{R}_{ije}} w_{ijr}}, \qquad w_{ijr} = \max(\omega_{\ell_i(r)}, \omega_{\ell_j(r)})\sqrt{c_{ier}c_{jer}}. \tag{13}$$

Zero denominator yields zero separation. The level weights for engagement, immediate, intermediate, functional, and persistence readouts are $(.55, 1, .90, 1, .85)$. Let $\mathcal{H}_e$ contain mechanisms with a prediction object for e. Then $C_{\mathrm{pred}} = |\mathcal{H}_e|/|\mathcal{H}|$ measures mechanism coverage, not readout-cell completeness. Mean separation uses available pairs; $D_{\min}$ is their minimum only when every declared mechanism is covered, and zero otherwise. Missing a difficult rival therefore cannot improve the worst-pair term.

Equation 3 combines these terms with the fraction of pairs satisfying $D_{ij} \ge .25$, Jaccard overlap between disputed and predicted causal edges, readout-level availability for secondary-lock diagnosis, and the fraction of declared readouts whose predicted range is at least .30. Secondary-lock



Figure 3: **Evidence licenses inference at compilation and reuse.** (a) Source spans bind fields to one experiment (cross-experiment splice blocked); the grammar licenses a predicate only when its required dimensions are present; a licensed hop retains experiment, span, and rule identities; pointer memory reopens sources rather than trusting stored prose. The lower lane is the public-span repair realization evaluated in (b–e). (b) Licensed compiler versus Core, atomic repair, and full-reference on Recall@10, supported yield, and Precision@10. (c) Per-question Recall@10 difference versus full-reference (45 questions). (d) The same difference by immune axis and lens. (e) Unsupported assertion rate, licensed coverage, and retained card counts. Precision@10 versus full-reference has a task-paired interval that includes 0. Panels (b–e) use the public-span repair study (45 questions × 3 seeds), not the structured hop compiler alone.

diagnosticity is 1 for engagement/immediate/functional coverage, .55 for immediate/functional coverage, .15 for immediate without functional coverage, and zero otherwise. Acquisition weights are $(.25, .32, .20, .08, .09, .06)$ with cost coefficient .07; feasibility adds .02, infeasibility subtracts .40, and the score is clipped to $[-1, 1]$. Hard restrictions belong to the caller's inventory. Ties use worst-pair separation, retirement potential, mean separation, lower cost, then experiment identity.

C.3 Observation discrepancies become typed graph edits

The structural kernel implements `structural_revision_contract_v2`: observation coverage determines which diagnoses are justified, evidence determines which relations are admissible, and atomic application binds each accepted operation to its child graph. The language-model interface is specified separately in Appendix C.5. Study-specific execution bindings and the scope of semantic verification are consolidated in Appendices G.9 and I.1.

Agreement is conditional on an observation. Parents and probe predictions are frozen before the mechanistic observation opens. A readout matches only when its normalized name and causal level agree; duplicate normalized names and level mismatches are rejected. For matched readouts $\mathcal{M}$,

$$s_r = c_r \max\{0, 1 - |\mu_r - y_r|\} + (1 - c_r)/2, \qquad \bar{S} = \frac{\sum_{r \in \mathcal{M}} \alpha_{\ell(r)} s_r}{\sum_{r \in \mathcal{M}} \alpha_{\ell(r)}}. \tag{14}$$

The weights for engagement, immediate, intermediate, functional, and persistence levels are $\alpha = (.50, 1.20, 1, 1.20, .90)$. Both $\bar{S}$ and the level mean S_ℓ are unknown when their denominator is zero. Missing readouts remain explicit; a measured zero effect is an observation, not missingness.

Let $\mathcal{P}_\ell$ be the predicted readouts at level ℓ and $\mathcal{M}_\ell$ their matched subset. Complete coverage is $C_\ell = \mathbf{1}\{0 < |\mathcal{P}_\ell| = |\mathcal{M}_\ell|\}$. The positive-success predicate P_ℓ additionally requires $S_\ell \geq .75$ and $\mu_r, y_r \geq .45$ for every predicted readout in that level. This separates *agreement with a prediction* from *success of a biological prerequisite*. Functional failure N_F requires complete functional coverage, $S_F < .50$, and $\mu_r \geq .45$, $y_r < .45$ for every functional readout. A strong local contradiction is

$$\begin{aligned} F_\ell = \mathbf{1}\big[\exists r \in \mathcal{M}_\ell : |\mu_r| \geq .45 \wedge s_r < .45 \\ \wedge \big(\mu_r y_r < -.08 \ \vee \ (|\mu_r| \geq .60 \wedge |y_r| \leq .15)\big)\big]. \end{aligned} \tag{15}$$

These thresholds define an executable diagnostic rule, not biological significance tests. Engagement g requires both the confirmation flag and matching observed target effects at least .45; a Boolean assertion alone cannot enable an engagement-dependent branch. When target readouts are predicted, all must be observed. The interface also recognizes the declared target/engagement aliases when that level was not predicted.

Failed matched readouts ($s_r < .50$) implicate the parent edges that declare them. Weakening increases their contradiction burden by .25, clipped at one. Replacement searches alternatives with deficient-axis coverage, compatible source nodes, and graph attachment. Residual-barrier proposals use eligible axes, attachment, failed-readout overlap, and strength. The proposer emits at most three operations. A configured deterministic fallback handles eligible malformed proposals, but cannot conceal identity conflicts or create the observations required by a diagnostic branch.

A receipt licenses one complete relation in one scope. Admission checks directed endpoints, predicate, axis, source/target compartments, and readout/intervention-target scope against the trusted evidence catalogue. Sharing an axis or a single endpoint is insufficient. Allowed contexts come from the target request and explicit compatibility rules, not a union of scopes asserted by candidate evidence. A declared independent-admission flag and self-consistent hashes are insufficient unless the receipt identity and catalogue digest resolve to the trusted record. Supporting and opposing receipts retain their required pairing.

The archival catalogue supplies frozen, curated computational assertions. Its receipts identify the asserted relation and permitted context; source-paper compilation provides a separate level of evidence verification (Appendix B). Catalogue membership alone does not establish a verbatim source span. Missing support remains unresolved, and incomplete provenance remains candidate-only.

Table 6: **Coverage-aware observation-to-edit rule.** Rows are evaluated in order. I, M, F denote immediate, intermediate, and functional levels; h_ℓ denotes a predicted level. Earlier observed contradictions are not erased by missing readouts elsewhere. Structural changes additionally require admissible proposals.

Ordered condition	Diagnosis	Scientific consequence
$g \wedge (F_I \vee F_M) \wedge P_F$	Requires revision	Replace the contradicted explanation while accounting for observed distal success.
$g \wedge (F_I \vee F_M)$	Structurally falsified	Remove implicated relations or propose an admitted replacement; an empty child is retired.
$g \wedge P_I \wedge (\neg h_M \vee P_M) \wedge N_F$	Secondary lock revealed	Preserve successful proximal structure and propose a residual barrier.
All predicted readouts observed; at least one is non-target; $\bar{S} \geq .82$; predicted immediate/functional levels each have $S_\ell \geq .75$	Supported	Preserve topology; test persistence or transfer. Agreement alone does not imply a positive biological effect.
Missing readouts, no matched readout, no non-target prediction, or a confirmation flag with unobserved engagement and $(F_I \vee F_M \vee N_F)$	Insufficient observation	Preserve topology and acquire missing discriminating readouts; no revised prediction/program is licensed.
Otherwise	Weakened	Preserve topology, retain local discrepancies, and acquire discriminating evidence.

Bind validation to application. Before dictionary construction, the kernel checks normalized identifiers across the library, parent, and proposal batch. Conflicting payloads cannot share an identifier, and ambiguous operation fields are rejected. Each admission decision binds the full edge payload, edit, parent graph, observation request, target context, and evidence snapshot. Application independently reconstructs diagnosis and re-admits the actual edge, then compares the complete decision binding before mutation.

Replacement and path splitting are atomic at the operation level: every required addition must be admitted before the corresponding old relation is removed. A split must form a directed, endpoint- and compartment-connected path. Rejected or unresolved proposals remain visible in the lineage. The retained parent can remain an audit object, but an incompletely admitted revision cannot authorize follow-up predictions or a revision-derived program.

Verify semantics as well as replay identity. The independent semantic checker does not import the primary diagnosis, admission, or editing functions. From the bound raw inputs it reconstructs coverage, permitted operations, evidence identity, atomic effects, child identity, and downstream eligibility. Rehashing an invalid trajectory therefore does not repair its scientific prerequisites. The focused regression evidence and reconstruction scope appear in Appendix G.9.

New structure creates new obligations. Eligible children receive graph-conditioned first-lock, residual-lock, and joint-test predictions, retaining the parent and motivating observation. The archival provider's declared functional contrast values .10, .34, and .84 illustrate these three configurations; they are rule-based forecasts, not fitted treatment effects. A child requires a separately bound likelihood before receiving a new mechanism posterior. Structural, predictive, and program identities remain separately inspectable.

C.4 Compile the revised explanation into an intervention program

The compiler excludes insufficient-observation and unresolved or non-applied revisions before selecting a surviving source. Selection is governed by scientific eligibility and target coverage, rather than composition score alone. Each evaluated execution retains its recorded compiler binding (Appendix I.1). Let $b_\star$ be the promoted primary axis, $\mathcal{B}_i$ the child's axes, and b_{eff} dysfunction/escape.

Selection maximizes the lexicographic key

$$K_i = \big(\mathbf{1}_{d_i=\text{secondary}},\ \mathbf{1}_{b_\star \in \mathcal{B}_i},\ |\Delta_i^+|,\ \bar{S}_i,\ -|\mathcal{B}_i \setminus \{b_\star, b_{\text{eff}}\}|\big). \tag{16}$$

Here d_i is the diagnosis and Δ_i^+ contains admitted additions; ties preserve record order. Primary-axis promotion selects the most deficient eligible non-effector axis, with root-score/identity tie breaks and a dominant-root fallback. A shadowed descendant is eligible only when its competence is more than $.15$ below the worst eligible ancestor.

For secondary-lock records, first-lock edges are retained relations with successful immediate readouts, falling back to primary-axis or retained edges; other records use primary-axis edges with a retained-edge fallback. Newly added edges supply residual locks. For uncovered targets T and action target sets T_a, greedy coverage selects

$$\begin{aligned} a^* &= \operatorname*{argmax}_{\text{lex}\ a:T_a \cap T \neq \varnothing} \big(|T_a \cap T|, -|T_a|, -\text{stage}(a), -\text{risk}(a), \text{id}(a)\big), \\ T &\leftarrow T \setminus T_{a^*}. \end{aligned} \tag{17}$$

Maintenance is excluded from coverage; uncovered non-effector axes receive an available axis-specific action. Axis-level compression favors specificity, then residual coverage or stage/risk/identity precedence. A probed action is committed only with engagement and at least one immediate response at $.45$ or above; dedicated effector rescue is handled separately. Rescue is requested when dysfunction competence is below $.60$ or matched functional responses exist and all are below $.45$; maintenance is added at rebound pressure $.55$. Final actions follow stage, maintenance flag, and identity.

Keep structural and action consequences separately attributable. Four contrast records retain first-lock, residual-lock, joint, and full-program predictions. Compression can leave empty or identical action contrasts; these remain explicit. Graph readouts and program ordinals are stored separately: for the full program, a *marked* action-composition prediction can override a weaker graph-derived ordinal. Neither is substituted for the archival observation. Algorithm 1 formalizes this transformation; comparative results remain in Appendix G, and mapping, functional assessment, and worked trajectories in Appendix H.

Figure 4 separates measured acquisition performance from the mechanism-portfolio and program changes recorded after an opened probe; full numerical comparisons remain in Appendix G, and final archival assessment in Appendix H.

C.5 Matched mechanism revision and graph-conditioned decisions

The two-stage interface separates **revising an explanation** from **using that explanation to predict and act**. Both arms receive the same context, evidence, trigger observation, initial graph, follow-up events, and intervention panel. The controlled treatment is permission to change the working graph. The deterministic structural kernel and this language-model realization share scientific objects but retain distinct execution and assessment bindings.

Stage one: vary structural-edit permission. Both arms share the scientific preamble and diagnosis checklist; their system prompts differ only in the edit-permission paragraph. Edit-on may add, delete, or replace at most three evidence-grounded edges. Edit-off must preserve the exact $(\text{src}, \text{rel}, \text{dst})$ inventory, enforced by fixed-structure projection. Confidence and status updates are available to both arms. This preserves the ability to reconsider an explanation while isolating permission to change its structure.

Stage two: hold prediction and action rules fixed. A byte-identical system prompt consumes the public task and `working_graph`, with shared probability bands, action-arity rules, and a two-action cap. No arm label is supplied. The resulting event probabilities and recommendation, rather than the corresponding stage-one fields, define the native scored output. Thus, an edit can influence the endpoint only through the representation supplied to the common decision interface; edit-on receives neither an extra probability-sharpening rule nor a larger action menu. An empty working-graph payload uses the declared parent-edge fallback, separately from the deterministic kernel's retirement semantics.



Figure 4: **Informative experiment selection and recorded mechanism-to-program changes.** (a) Top-1 mechanism identification at 1, 3, 5, and 8 queries on the same 30 models. (b) Paired five-query log-score gains versus random, uncertainty-only, and metadata-family, with 30 model-level deltas and 95% CIs; the companion panel splits the random contrast by mechanism class. (c) Four candidate mechanisms per archival context after the opened probe, split by whether they remain supported, weakened, structurally edited, or retired. (d) Ordered intervention actions emitted from the selected graph; a gold outline marks the one context whose selected graph was edited. Panels (a,b) use the lineage-disjoint confirmation screen (30 models, 36-intervention menu); at five queries, EIG reaches 56.7% top-1 versus 42.9% for random. Panels (c,d) use the separate ten-context archival execution. Final program assessment is in Figure 7.

Algorithm 1: coverage-aware, evidence-licensed revision and synthesis

Input: frozen parents $\mathcal{H}$ and predictions $\Pi_i(e)$; observed probe Y_m; target context c; ecosystem x; trusted evidence $\mathcal{E}$, edge library L, and action library $\mathcal{A}$.
Output: complete lineage $\mathcal{R}$ and, when eligible, a committed program.

Validate normalized identities and prediction–observation bindings; snapshot $(\mathcal{H}, \Pi, Y_m, c, \mathcal{E}, L)$.
$x' \leftarrow \mathrm{Update}(x, Y_m)$; $B' \leftarrow \mathrm{DiagnoseEcosystem}(x')$; initialize $\mathcal{R}$ and eligible sources $\mathcal{S}$.
for $h_i \in \mathcal{H}$ **do**
  $(d_i, \mathcal{I}_i, \mathrm{coverage}) \leftarrow \mathrm{Diagnose}(\Pi_i(e), Y_m)$ [Table 6].
  if $d_i = \mathrm{insufficient}$, retain parent and acquisition obligations in $\mathcal{R}$; **continue**.
  $\Delta_i \leftarrow \mathrm{Propose}(h_i, d_i, \mathcal{I}_i, L, B')$, $|\Delta_i| \leq 3$; reject batch ID conflicts.
  Bind each admission decision to its full edge, operation, parent, request, context, and evidence snapshot.
  for operation $\delta \in \Delta_i$, recheck diagnosis and re-admit the actual payload.
    if its required additions are all admitted, atomically apply δ; otherwise record its non-applied disposition.
  Reconstruct child identity and run independent semantic verification against the bound inputs.
  Record parent, proposals, decisions, observation, child, coverage, and verification in $\mathcal{R}$.
  if child is empty, retire it; **if** revision is unresolved or ineligible, retain audit state only.
  otherwise, version/rescore the child, regenerate its predictions, and add it to $\mathcal{S}$.
end for; **if** $\mathcal{S} = \varnothing$, **return** lineage and no program.
Select $h'_j \in \mathcal{S}$ using K_j [Equation 16]; identify operative first and residual locks.
Cover their targets with actions [Equation 17]; compress redundant coverage and retain only observationally committed probed actions.
Add indicated effector rescue and maintenance; order by registered causal stage and action identity.
if no valid action program remains, **return** lineage and no program.
Bind graph-conditioned contrasts and program forecasts to the selected eligible source.
return committed lineage, graph, predictions, and program **before** final functional lookup Y_f.

Test whether the working graph is used. For edit-on trials with a changed structural signature, repeat stage two with either the initial parent graph or the next task's edit-on child at the same seed. The factual comparison uses the actual child. Only `working_graph` changes: evidence, trigger observation, event descriptions, action menu, and decision instructions remain fixed. Output change is a changed recommendation or a changed event prediction, with probabilities rounded to three decimals and abstention flags retained. Parent substitution tests dependence on the revision; cross-family substitution tests sensitivity to the supplied mechanism. These are output-dependence diagnostics, not recommendation-only switch rates or certificates of biological edge correctness.

The protocol prioritizes native Brier, child-versus-parent output change, and edit occurrence when needed (registered as *needed-edit recovery*). Native regret, selection accuracy, and frozen graph-readout metrics remain secondary. Appendix F.6 defines family pairing and eligibility; Appendix G.8 reports the complete predictive and structural diagnostics.

D BIOLOGICAL AND OPERATIONAL WORLD MODELS

D.1 TWO PREDICTIVE QUESTIONS, TWO DECISION INTERFACES

Mechanova separates *biological consequence* from *research-process consequence*. BWM predicts intervention responses to guide measurements and recommendations; OWM predicts research-action outcomes to guide execution. A biologically informative query may be operationally infeasible, while an inexpensive operation may leave the decisive biological uncertainty unchanged. Table 7 separates their decision and feedback channels. The experiment-specific bindings in Appendix I.1 identify which provider is executed; mechanism-graph admission remains with MEC (Appendix C).

Below we specify the response-decision BWM, the operational transition forecaster, and the separately evaluated evidence-horizon controller. Perturbation-response representation studies use their own context-bound checkpoints and held-out perturbations (Roohani et al., 2024; Adduri et al., 2026; Replogle et al., 2022); they do not replace the decision model described here.

Table 7: **Two predictive interfaces, distinct decisions and feedback.** Predictions guide action; only verified observations update the empirical scientific record.

Model	Conditioning information	Forecast and decision	Admitted feedback
BWM	Intervention, biological context, and source-separated response data.	Joint response beliefs; permitted measurement or STOP; unmeasured-target recommendation.	Selected, context-bound response conditions the biological belief.
OWM	Pre-action operational state, action semantics, and verified experience.	Outcome distribution, information gain, cost, and state change for planning.	Actual verified transition supplies an outcome and prediction-error record.

D.2 Biologically structured, source-separated response beliefs

Representation and response scale. Inputs encode intervention target, direction, modality, functional role, compartment, and experimental context. The Full head augments typed linear features with target/role pairs, immune-context interactions, and a radial-basis similarity term; source/panel nuisance covariance remains explicit. Pretreatment measurements require context-bound evidence, and missing slots remain masked. For therapeutic burden readouts, the control-normalized effect is $y_i = 1-\bar{t}_i/\bar{c}$, where $\bar{t}_i$ and $\bar{c}$ are treatment and control means. The provider models $z_i = -\log(1-y_i)$ jointly as $z \sim \mathcal{N}(\mu, K)$; effects requiring censoring are not clipped into this likelihood.

Fitting and support-sensitive transfer. Training excludes the entire test source family. Source/panel-balanced weights and nested training-source folds select ridge regularization and mean shrinkage using marginal and within-panel contrast Gaussian negative log-likelihoods. Outer held-source residuals, with measurement contributions subtracted, calibrate latent covariance. For intervention i, let n_i be the minimum training-source count across its target–direction–modality signatures and signature pairs. Then

$$\begin{aligned} s_i &= \frac{1+n_i}{1+n_i+\lambda_s}, \qquad \mu_i^{\mathrm{ref}} = z_{\mathrm{global}} + K_{ii}/2, \\ \mu_i &= \mu_i^{\mathrm{ref}} + \alpha s_i(\mu_i^{\mathrm{raw}} - \mu_i^{\mathrm{ref}}). \end{aligned} \tag{18}$$

Here $z_{\mathrm{global}} = -\log \sum_j \omega_j e^{-z_j}$ for normalized training weights ω_j. Support limits the transferred *mean*; covariance calibration is separate. The protected-target configuration fixes $\lambda_s = 2$ and selects ridge from $\{0.05, 0.2, 1\}$ and α from $\{0, 0.5, 1\}$. Recommendations use the original-scale expectation

$$\mathbb{E}[y_i] = 1 - \exp(-\mu_i + K_{ii}/2), \tag{19}$$

not simply the largest transformed mean.

D.3 Protected-target information transfer and finite decision value

Condition biology and measurement error jointly. Training measurement covariance uses delta-method log-ratio variances with shared-control terms. Query covariance R uses training-only median squared coefficients of variation and the declared arm sample counts: its shared-control contribution is $\widehat{\mathrm{CV}}_c^2/N_c$, plus $\widehat{\mathrm{CV}}_t^2/N_i$ on each treatment diagonal. Unrelated panels receive no shared-control term. For potential measurements $\tilde{z} = z + \epsilon$, initialize

$$m = \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} K & K \\ K & K+R \end{bmatrix}. \tag{20}$$

A permitted experiment has column operator $h_e = (0, w_e)$, source binding, and cost c_e. On receiving its selected observation o_e, update

$$\begin{aligned} v_e &= h_e^{\mathsf{T}}\Sigma h_e, \qquad k_e = \Sigma h_e, \\ m' &= m + k_e(o_e - h_e^{\mathsf{T}} m)/v_e, \qquad \Sigma' = \Sigma - k_e k_e^{\mathsf{T}}/v_e. \end{aligned} \tag{21}$$

Retaining the full augmented posterior preserves shared-control dependence across sequential measurements. For example, observing an allowed intervention can reorder two correlated protected candidates without either candidate being measured.

Every operator has zero weight on protected targets, including composite contrasts. With consumed measurement coordinates U and remaining query budget b, the decision state is $B = (m, \Sigma, U, b)$. The available set $\mathcal{A}(B)$ excludes protected, overlapping, or zero-variance measurements. An observation must match the pending experiment and context before another proposal is issued.

Acquire only decision-relevant information. Let $V_0(B)$ be the largest expected effect in Eq. (19) among recommendation candidates. For depth $d \le b$,

$$\begin{aligned} Q_d(B, e) &= -c_e + \mathbb{E}_{o_e|B,e}[V_{d-1}(B \mid e, o_e)], \\ V_d(B) &= \max\{V_0(B), \max_{e \in \mathcal{A}(B)} Q_d(B, e)\}. \end{aligned} \tag{22}$$

The child state also consumes the operator's support and one budget unit. The evaluated configuration uses at most two queries, one- or two-query lookahead, and nine-node Gauss–Hermite quadrature. Therapeutic and functional-contrast queries incur declared effect-utility penalties of 0.01 and 0.03, respectively. STOP is selected when no admissible query exceeds terminal value after cost. This bounded value-of-information calculation is related to correlated-normal knowledge-gradient policies (Frazier et al., 2009).

Information gain instead rewards uncertainty reduction over a declared outcome partition (Lindley, 1956; Foster et al., 2020). Here, response-contrast regions are distinct from molecular-mechanism classes; the separate PRISM/DepMap study supplies its own mechanism likelihoods. The matched no-query and random-query comparisons in Table 21 distinguish transferred decision value from acquisition cost, without pooling these objectives.

D.4 Verified operational dynamics and selective foresight

OWM models consequences before execution (Ha and Schmidhuber, 2018; Hafner et al., 2023), using operational features rather than next-text prediction. Its state records uncertainty, evidence and contradiction coverage, outstanding obligations, completed/failed actions, memory availability, and remaining resources. Action type, tool contract, source partitions, expected outputs, and disconfirmation semantics define an exact routing regime: different tool semantics cannot become interchangeable through textual similarity.

For feature vector x and action a, compatible verified transitions j receive Gaussian distance weights $w_j = \exp[-d(x, x_j)^2/(2\ell^2)]$ inside the registered neighborhood. The outcome predictor is

$$\widehat{p}(o \mid x, a) = \frac{\eta + \sum_j w_j \mathbf{1}[o_j = o]}{|\mathcal{O}|\eta + \sum_j w_j}, \tag{23}$$

where η is the Dirichlet pseudocount and $\mathcal{O}$ contains support, contradiction, unresolved evidence, tool failure, budget pressure, and context mismatch. Weighted observations estimate information gain and cost; outcome-conditioned feature deltas project only permitted operational coordinates. Imagined transitions remain planning states, not admitted scientific evidence.

Forecast use requires sufficient neighbors, task-family diversity, effective support $n_{\text{eff}} = (\sum_j w_j)^2 / \sum_j w_j^2$, and bounded epistemic uncertainty. Once enough residuals exist, excessive recent Brier or projected-state error blocks selective foresight. Unsupported forecasts therefore cannot supply fictitious observations. After execution, the parent model, input state, action, verified outcome, and numerical errors form a residual record for eligible later fitting. Operational transition/residual memory is distinct from the source-evidence memory in Appendix B. Evaluation keeps forecasting error, memory-augmented decision value, and useful execution separate (Appendix G).

D.5 Comparator-preserving evidence-horizon control

The real-data realization isolates evidence-acquisition efficiency: the endpoint-predictor bank and reference panel sequence remain frozen. Training-only nested cross-fitting chooses a horizon $H_r \le K$ for objective route r, using lower-confidence-bound and harmful/catastrophic-stop criteria. Within a supported route and exact reference prefix of length k, the controller is

$$\pi_{\text{EH}}(r, k) = \begin{cases} \text{STOP}, & k = H_r < K, \\ \pi_0(r, k), & \text{otherwise}, \end{cases} \tag{24}$$

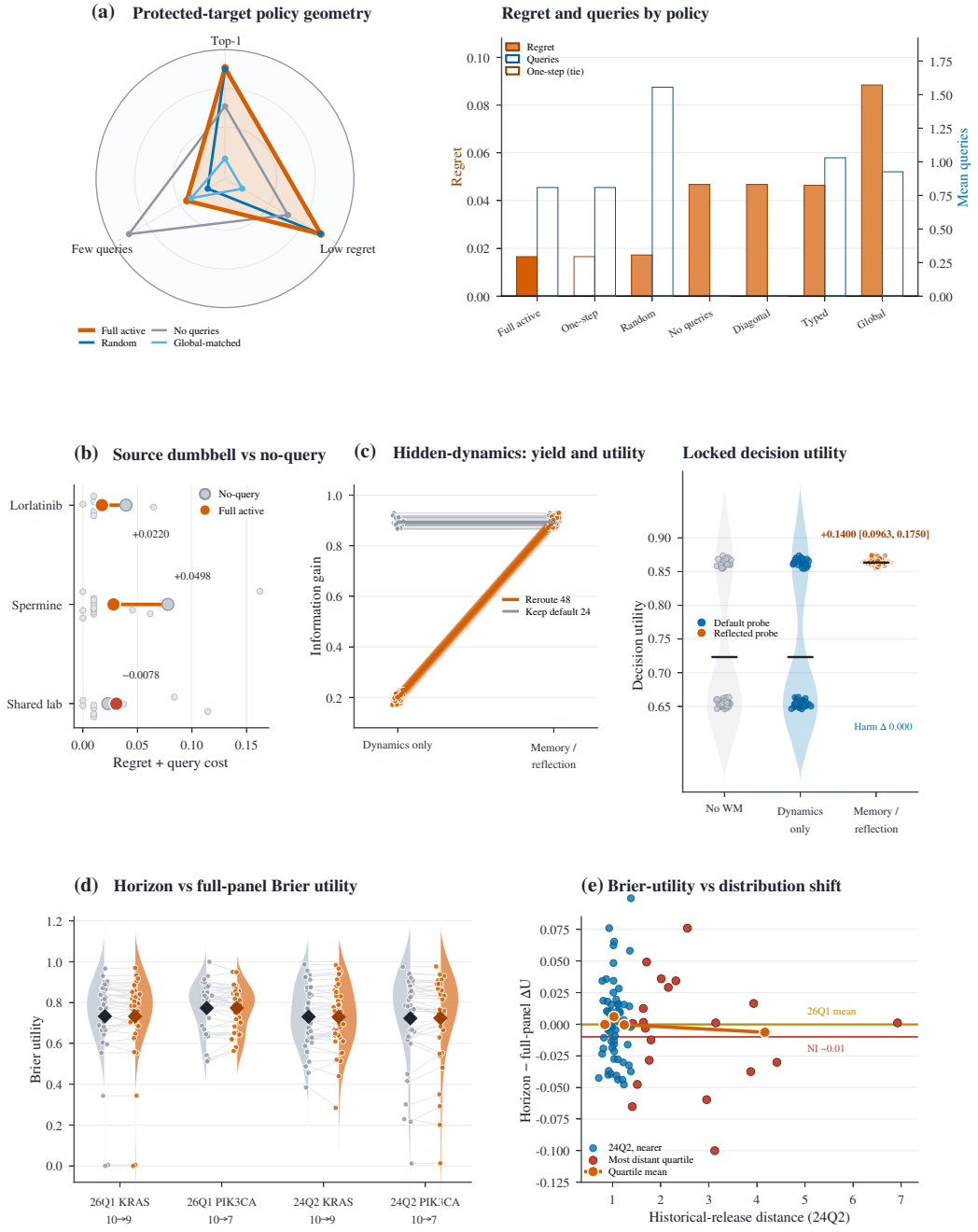

Figure 5: **World models improve decision value and reduce evidence requirements.** (a) Protected-target source-macro geometry and seven-policy regret/query bars; radar axes are rescaled within the displayed policies, full active and one-step coincide, and random achieves similar top-1 with more queries. (b) Source-macro dumbbells from no-query to full active (31 overlapping views in the background). (c) Locked hidden-dynamics confirmation: 48/72 cases reroute from the default probe and jump in information gain, while 24 already sit high; dynamics-only matches no world model, and memory/reflection raises paired utility by +0.1400 [0.0963, 0.1750] over 24 groups with harm $\Delta = 0$. (d) Model-level Brier utilities for frozen evidence-horizon prefixes versus the length-10 reference (KRAS $10 \to 9$, PIK3CA $10 \to 7$; 80 models $\times$ two releases). (e) 24Q2 model-level Brier-utility difference versus historical-release distance, with quartile means and the 26Q1 mean; both releases satisfy the -0.01 noninferiority margin. Biological recommendation and operational utility retain separate scales.

where π_0 is the frozen reference policy, including completion at K. Portfolio, prefix, or next-action mismatches preserve reference continuation. The only authorized deviation is omitting a suffix; panels are neither reordered nor substituted.

This controller is objective-conditioned, not adapted online to individual panel values. The selected horizons are nine for KRAS, seven for PIK3CA, and ten for the other supported routes ($K = 10$). Endpoint quality uses Brier utility $1 - (p - y)^2$ against full-prefix acquisition using the same predictor bank and a 0.01 noninferiority margin. The 26Q1 confirmation and exposure-disjoint 24Q2 historical-release stress are reported separately in Table 18; their acquisition savings concern the evaluated routes, not wet-lab expenditure. Appendix F specifies analysis units and Appendix I binds implementations. Predictions precede action commitment; observed outcomes inform subsequent decisions, never the already committed forecast.

Figure 5 summarizes the measured decision consequences of biological and operational foresight, separating recommendation quality, acquisition cost, and retained endpoint utility.

E Adaptive research execution and experience-driven learning

E.1 An observation-conditioned WHAT–HOW hierarchy

The planner owns WHAT: a scientific objective and its completion contract. HOW compiles the next permitted operation from verified state, rather than expanding an initial plan into an immutable sequence. This separates temporal abstraction (Sutton et al., 1999) from observation-conditioned execution (Yao et al., 2023). The intention graph links goals, procedures, capabilities, and tools (Yang et al., 2024; Xiao et al., 2026); an inner executor cannot change the objective or enlarge the authorized toolbox.

The controlled recovery route is explicit: audit the candidate chain; on an invalid audit, build a repair, convert the shortest returned path into typed edges, and re-audit; then derive stop/continue arguments from the verified audit. Empty repairs and invalid re-audits terminate without claiming success. In the integrated implementation, each round licenses one selected experiment, waits for its matching observation, and checks agreement between licensed and executed sequences. Appendix F specifies matched budgets; Table 22 separates clean and recovery outcomes.

Algorithm 2: observation-conditioned execution with explicit terminal states

Input: goal g, verified state S_0, contract $\mathcal{C}$, permitted tools $\mathcal{A}$, budget b_0.
$(S, \mathcal{T}, b) \leftarrow (S_0, \varnothing, b_0)$.
while $b > 0$ and no hard stop **do**
  Reconcile any pending effect; **if** unsettled, **return** PENDING.
  $(s, v) \leftarrow \mathrm{Certificates}(S, \mathcal{C})$ using Equation 25.
  if $s \wedge \neg v$, **return** (COMPLETE, $S, \mathcal{T}$);
  if $\neg v$, **return** (BOUNDEDINCOMPLETE, $S, \mathcal{O}(S)$).
  $c \leftarrow \mathrm{Project}(S, g)$; $(u, x, k) \leftarrow \mathrm{CompileHOW}(g, S, c)$.
  Require $u \in \mathcal{A}$, typed arguments x, and completion contract k.
  $o \leftarrow \mathrm{ExecuteAndSettle}(u, x, k)$; $\mathcal{T} \leftarrow \mathcal{T} \oplus (u, x, o)$; $b \leftarrow b - \mathrm{Cost}(o)$.
  if verified(o), $S \leftarrow \mathrm{Apply}(S, o)$; otherwise reconcile or take the contracted recovery route.
  Return requests to revise g to the planner; recompile HOW from the resulting verified state.
return terminal status with $S, \mathcal{T}$ and unresolved obligations $\mathcal{O}(S)$.

E.2 Progressive context and outcome-grounded skills

Persistent scientific state, the model-visible action context, and reopenable evidence pointers have different roles. The context compiler projects action-relevant fields and converts missing requirements into acquisition obligations; compaction preserves structured state and pointers rather than retaining every transcript token (Chen et al., 2026). Disclosure changes visibility, not permissions. Evidence-memory admission and reopening are specified in Appendix B.

A skill binds applicability, required tools, artifact predicates, minimum verified calls, and recovery conditions. Its runtime completion contract restricts execution to open obligations and disallows discretionary calls after completion. For chain repair, a protocol description is insufficient: the required build and valid re-audit must appear in the execution record. Reusable procedures retain

these conditions (Wang et al., 2023). Protocol conformance and useful completion are scored separately; corrected token and call counts accompany completion in Table 22.

E.3 Scientific sufficiency, residual opportunity, and adaptive horizons

Sufficiency is a conjunction of contract-specific evidence checks, not a language-model confidence statement. The endpoint-qualified study requires an axis core, an immune bridge, a compartment-correct functional endpoint, a cross-compartment chain, at least two independent supports, and three proposition families. T-cell tasks require cytotoxic effector or direct-killing evidence; other axes can also close through immune-dependent tumor control. Population abundance or an intervention name cannot substitute for function.

The runtime separately checks non-bypassable operational obligations. As in Section 3.6, let s_t denote joint scientific/operational sufficiency, k_t reachable critical work, and r_t a reachable high-value residual question. After pending effects settle, $\nu_t = k_t \vee (s_t \wedge r_t)$ gives

$$d_t = \begin{cases} \text{CONTINUE}, & \nu_t = 1 \text{ and budget permits}, \\ \text{COMPLETE}, & s_t = 1,\ \nu_t = 0, \\ \text{BOUNDED_INCOMPLETE}, & \text{otherwise}. \end{cases} \tag{25}$$

The historical terminal labels $\text{STOP}_{\text{sufficient}}$ and $\text{STOP}_{\text{bounded}}$ map to COMPLETE and BOUNDED_INCOMPLETE, respectively; stored records are unchanged. Algorithm 2 uses $v = \nu_t$ locally. Resolvable contradictions can request reformulation, and continuing can complete a causal chain, obtain independent corroboration, or preserve a two-stage route whose functional closure appears only at the later checkpoint.

The eight-task development evaluation preconfigures early/middle/late evidence-completion strata; the runner receives neither these assignments nor evaluator targets. H30/H60/H100 denote registered campaign horizons, not counts of neural calls, with 6/9/12 deeply read records from each twelve-record packet. Decisions use cumulative, clause-grounded state; the scored object combines model output with that structured state. Forced-long execution supplies the matched counterfactual, including reformulation costs. Table 23 reports every task's stage coverage and accounting for selective continuation.

E.4 Offline policy learning from action-time information

The masked recurrent policy selects `search`, `verify`, `reject`, or `expand`. Logs originate from three evidence-research task families under five frozen, exploratory behavior-policy snapshots; full legal-action probabilities are retained. Human-adjudicated claim support and source links supply evaluator-side supervision, not expert action demonstrations. After execution, a deterministic evaluator assigns `gold_action`: forced search first; verify supported/partial claims when source and coverage checks pass; otherwise expand once when permitted or reject. Unsupported claims with a recovered source and at least three retrieved records are rejected directly. Supported/partial verification requires query coverage 0.45/0.35 and source diversity; supported claims additionally require abstract evidence. Labels and current-step outcomes are excluded from policy inputs.

Each 29-dimensional input is constructed entirely from pre-action state: a bias, four legal-action indicators, eight evidence-state features, four prior-outcome features, four progress indicators, and eight MedCPT similarity summaries. The summaries comprise minimum/maximum/mean/standard deviation, top-two margin, and fractions above 0.50/0.60/0.70. Progress uses the declared budget: at zero-based step t, `declared_horizon_fraction` is $t/\max(H-1,1)$ for the task's predeclared maximum H. Completed trajectory length and future-step counts are forbidden. Current labels and current-action outcomes remain evaluator-only.

The linear reference, supervised GRU, and AWR GRU are fitted on the same 337 trajectories from 74 groups and frozen before execution. The two GRUs share architecture, feature standardization, training seed, supervision, optimizer, and behavior regularization; only return weighting differs. Evaluation uses 150 tasks in 106 groups, with zero source-component and group overlap against all 420 reference trajectories, including the separate 83-trajectory holdout. The present evidence comes from separately recorded policy executions.

Table 8: **Training recipe for the three frozen execution policies.** All features are reconstructed before fitting; executive language-model weights remain fixed. The separate 83-trajectory holdout is excluded from the 337-trajectory fit.

Component	Recorded setting
Architecture	29 inputs; one 32-unit GRU; four masked actions; 6,180 parameters.
Optimization	Full-batch Adam; learning rate 0.003; weight decay 10^{-4}; 100 epochs; gradient norm cap 1.
AWR / supervision	Temperature 1; weight cap 12; task-family/target-action mean-return baseline; inverse-frequency balance exponent 0.5; supervised weights equal one.
Fixed recipe	Both GRUs use the same 100-epoch recipe; no confirmation-dependent refitting or epoch selection.
Final fit / separate holdout	234 train + 103 dev = 337 trajectories / 74 groups; separate 83 trajectories / 16 groups.
Source-disjoint execution	150 tasks / 106 groups per arm; 450 greedy episodes; zero source-component and group overlap with the earlier 420 trajectories.
Determinism / reference	Seed 20260711; linear reference refitted on the same action-time features, with 900 epochs, learning rate 0.08, and $L_2 = 10^{-4}$.

The training return is defined as:

$$R(\tau) = Q(\tau) - 0.2\widetilde{C}(\tau) + 0.02\min(N_{\text{info}}, 5) - 0.1N_{\text{err}}. \tag{26}$$

Here $\widetilde{C} = \min(\max(c, 0)/60, 1)$ for recorded wall seconds c; N_{info} counts retained counterevidence, nulls, or failed hypotheses, whereas N_{err} counts avoidable operational failures. The evaluator score is $Q = 0.45A_{\text{step}} + 0.20A_{\text{terminal}} + 0.30O_{\text{recovery}} + 0.05E_{\text{cost}}$, combining logged-action agreement with evaluator targets, terminal agreement, objective recovery, and $E_{\text{cost}} = \max(0, 1 - c/10)$. These are registered operational proxies.

Equation 6 uses $w_t = 1$ for supervised learning. The AWR-inspired variant (Peng et al., 2019) sets $w_t = \min\{\exp[(R(\tau) - b_t)/\beta], w_{\max}\}$, with a training-only task-family/target-action mean-return baseline, $\beta = 1$, and $w_{\max} = 12$. Writing the weighted supervision term separately makes the shared objective explicit:

$$\begin{aligned} \mathcal{L}_{\text{sup},w}(\theta) &= -\frac{\sum_{t\in\mathcal{V}} q_t w_t \log \pi_\theta(a_t^\star \mid x_{\le t})}{\max\{1, \sum_{t\in\mathcal{V}} q_t w_t\}}, \\ \mathcal{L}(\theta) &= \mathcal{L}_{\text{sup},w}(\theta) + \frac{0.35}{\max\{1, |\mathcal{V}|\}} \sum_{t\in\mathcal{V}} D_{\text{KL}}(\mu_t \,\|\, \pi_\theta(\cdot \mid x_{\le t})) \,. \end{aligned} \tag{27}$$

Here $\mathcal{V}$ contains valid steps, μ_t is the logged behavior distribution, and q_t is the fitting-split action-balance weight defined in the main text. The KL term appears once in the complete objective, identically for both GRUs. Masks precede normalization. Forced steps have zero action-choice supervision weight but retain their cost and history; the final fit contains 410 informative decisions and 337 forced steps. Table 8 records the policy architecture, optimization recipe, and separated selection/refitting/evaluation partitions.

E.5 Controlled policy execution and diagnostic separation

Evaluate the learned policy through its actions. Each policy greedily selects a legal action from its current state. A fixed Qwen3.8-27B backend selects a query from the shared bounded search/expansion menu; a local pre-cutoff PubMed index returns evidence. Search retrieves five records and expansion ten; verify/reject terminate execution using the retrieved evidence. The study evaluates active computational execution through a constrained tool interface with fixed policy parameters. All arms use the same action library, task budgets, tools, and backend.

For an executed episode, the registered score is $Q_{\text{exec}} = 0.45A_{\text{step}} + 0.20A_{\text{terminal}} + 0.30O_{\text{recovery}} + 0.05E_{\text{retrieval}}$, where $E_{\text{retrieval}} = \max\{0, 1 - \max(N_{\text{call}} - 1, 0)/2\}$. Step and terminal agreement use the evaluator's action targets; the forced initial search is included in step accuracy. Recovery

requires verify for a supported/partial target whose source was retrieved, and reject otherwise. The recorded `supported_valid_output` field implements the same disposition predicate as recovery, including correct rejection. Nonnecessary calls are expansions when the evaluator already prescribes verify/reject, divided by $\max(T-1,1)$. The recorded tool-call count includes terminal decisions; model requests are counted separately. These operational metrics quantify computational execution quality and agreement with evaluator targets.

Task means describe 150 executions per arm. Paired effects first average within each of 106 source groups and then weight groups equally; intervals use 10,000 paired group-bootstrap draws. Consequently, these effects need not equal differences between task means (Table 24).

Logged-sequence estimation with completed-trajectory features. An earlier logged-sequence analysis is retained for comparison with the original records. Completed trajectory length is included among the features, defining the estimand as completed-sequence value. Figure 6(e) and Table 24 instead report the three frozen policies acting on pre-action information. For historical trajectory i, the estimator uses logged actions and behavior probabilities (Mahmood et al., 2014; Thomas and Brunskill, 2016):

$$\begin{aligned} \ell_{it} &= \sum_{s \le t} \operatorname{clip}\left(\log \pi(a_{is} \mid x_{\le s}) - \log \mu_{is}(a_{is}), -4, 4\right), \\ \widehat{V}(\pi) &= \sum_t \frac{\sum_i e^{\ell_{it}} r_{it}}{\sum_i e^{\ell_{it}}}. \end{aligned} \tag{28}$$

The terminal step receives $R(\tau_i)$ and other rewards are zero; final cumulative weights are carried forward after termination. Paired resampling uses 10,000 draws over 106 groups. The logged-sequence values are 0.6952 and 0.7864, with difference $+0.0912$ [0.0404, 0.1391]. Minimum time-step ESS is 51.57 and 71.93; the declared reliability thresholds are time-step ESS fraction 0.20, action-level ESS fraction 0.15, and clipped-ratio fraction at most 0.10. The historical non-forced action diagnostic is 136/179 correct (verify 117/124, expand 13/43, reject 6/12). These estimates and completed-sequence features belong to the logged-sequence analysis; the actual-execution scorecard uses pre-action information.

Procedural consolidation updates reusable workflows, while offline fitting updates action preferences. The evaluated policies act only within the controlled execution profile; neither training nor execution changes evidence-admission rules, evaluator definitions, or foundation-model weights. Mechanism revision, evidence reuse, and learned action selection consequently retain separately testable consequences.

Figure 6 connects observation-conditioned recovery and adaptive effort allocation to actual frozen-policy execution. Panel (e) separates equal-task outcome means from equal-source-group paired effects; Table 24 reports the full comparison.

F Benchmark construction and statistical analysis

F.1 Human-reviewed literature resources

The foundational release contains a **469-paper human-reviewed resource** linking paper qualification to experiment-grounded mechanistic claims. Form A contains 238 paper-level assessments for future-mechanism qualification; Form B contains 872 adjudicated claim records from 231 historical papers. PMID deduplication confirms that the two paper sets do not overlap. Table 9 reports the three foundational batches and the separate September Form A release, rather than counting reviewer copies, intermediate claim slots, or repeated downloads as additional papers. The resource supports both evidence preparation and benchmark construction; its size is distinct from any individual evaluation denominator.

Separate September Form A release. The later release resolves 281 review rows into 278 unique papers, with 125 included, 37 deferred, and 116 excluded. Its records retain review/adjudication provenance, metadata resolutions, duplicate mappings, and nonblocking uncertainties. It adds no Form B claims or KG ingestion. The 278-paper batch and 469-paper foundational resource are

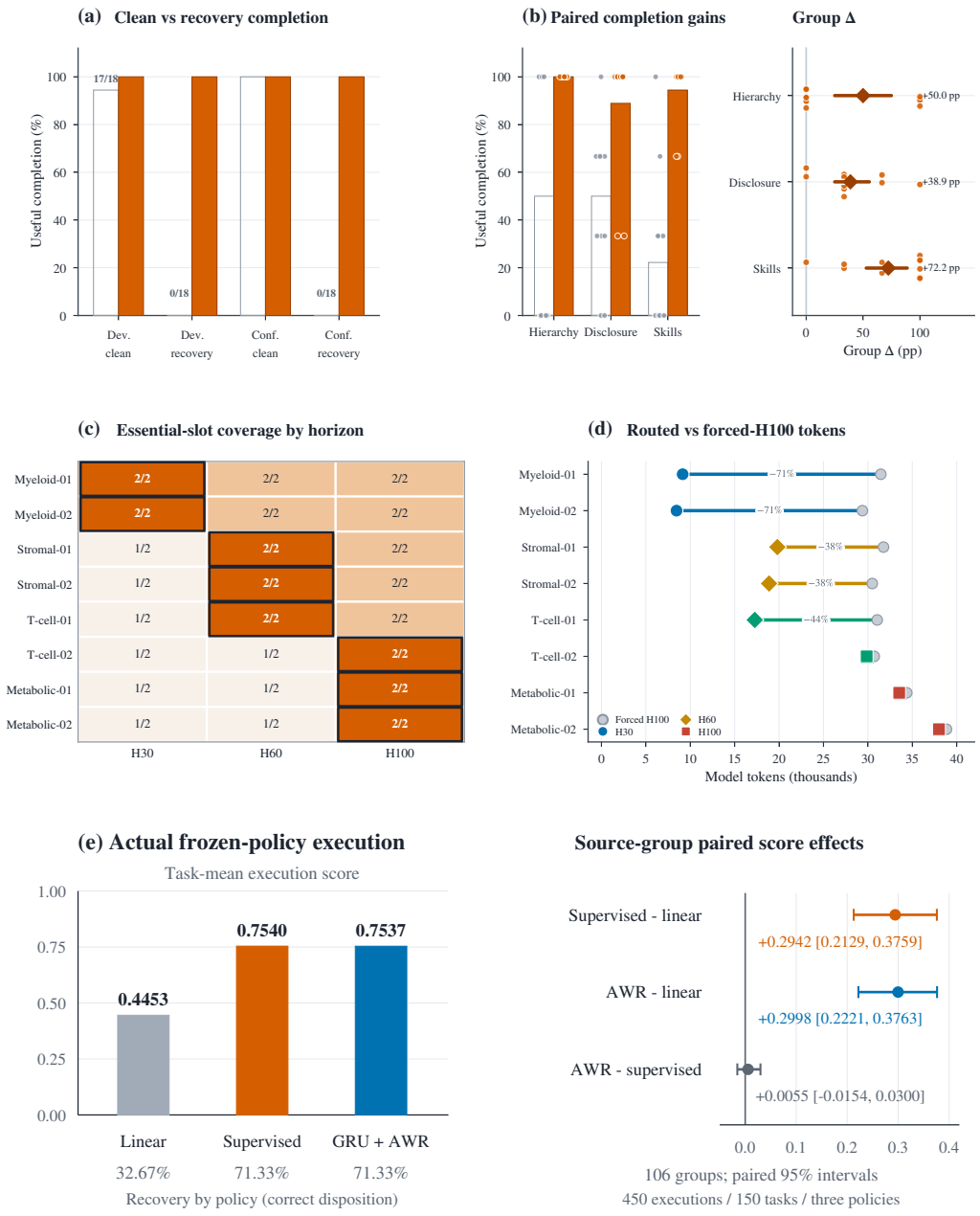

Figure 6: **Observation-conditioned execution improves recovery while adaptive routing concentrates effort on unresolved tasks.** (a) Useful completion under clean and recovery campaigns in development and confirmation (12 groups × 3 seeds); hierarchical control recovers 18/18 confirmation failures that the flat loop misses. (b) Three separately frozen paired gains, with 12 task-group means: hierarchical control +50.0 pp [25.0, 75.0], progressive disclosure +38.9 pp [25.0, 55.6], and outcome-grounded skills +72.2 pp [52.8, 88.9]. (c) Registered essential-slot coverage at H30/H60/H100; ink boxes mark the selected horizon (2/3/3 routes, no premature stop). (d) Matched routed versus forced-H100 token counts on the eight development tasks (31.2% aggregate saving). (e) Actual frozen-policy execution: equal-task scores and recovery over 150 tasks per arm (450 episodes), alongside equal-source-group score differences with registered paired 95% intervals over 106 groups. Supervised GRU and GRU+AWR improve on the linear reference; their matched return-weighting contrast includes zero. Recovery includes correct rejection; executive weights remain fixed (Table 24).

Table 9: **Literature resources by release.** Paper counts are unique PMIDs; Form B claims are final adjudicated records. Batch identifiers are retained for source tracing, not interpreted as publication dates.

	Reviewed papers		Released resource	
Batch	Form A	Form B	All papers	Form B claims
6.26	20	16	36	79
6.30	73	73	146	263
7.11	145	142	287	530
Foundational total	**238**	**231**	**469**	**872**
September Form A release	278	—	278	—

reported separately; no unverified cross-batch sum or model-assisted-to-human-only relabeling is made.

Form A: qualify the paper and its scientific use. Review records distinguish original research, direct PDAC relevance, the supported mechanism, and its permissible use. Temporal qualification records the earliest verified availability, its precision and source, rather than relying on issue year. Prior-work searches retain queries, sources, candidate predecessors, and mechanism-equivalence decisions. The final disposition separates primary future candidates, additional candidate inclusions, continuity evidence, intervention support, context-only material, deferrals, and exclusions. Thus, a follow-up study can remain useful evidence without being relabeled as an independent new mechanism.

Form B: adjudicate the claim and its experimental support. Two reviewer-source records are reconciled with the final claim and its adjudication rationale. The record retains upstream/downstream entities, relation and causal status, evidence strength, species/model, intervention context, primary and supporting evidence locations, and limitations or counterevidence. Reconciliation aligns semantically corresponding claims, splits composite claims, merges duplicate endpoints, and rebinds evidence to the appropriate result or figure panel. Adjudication can narrow causal language or restrict transfer across models instead of discarding otherwise useful evidence. Source-link tables preserve the many-to-many relation between original annotations and final claims.

Quality control preserves usable evidence and its boundaries. The release workflow checks source traceability, dates, evidence location, atomicity, causal labels, model scope, and overclaiming. All 872 final claims have a nonempty primary-evidence-location field. Among their QC labels, 585 explicitly retain scope limits and 97 restrict use to association rather than a positive causal edge; the other 190 have standard or minor-edit pass labels. These counts summarize release composition and evidence traceability. The first batch illustrates the distinction between adjudication history and released content: its 80-row intermediate claim table retains a merged duplicate for lineage, whereas the final export contributes 79 claims. Human review, model-assisted proposals, deterministic compilation, and evaluator judgments retain separate provenance. Evidence use by MEC is specified separately in Appendix B.

F.2 From curated papers to fixed evaluation tasks

Current 125-task collection. The Form A release dated September 22, 2026 resolves 281 source-review rows into 278 unique PMID records. Its final disposition includes 125 papers, defers 37, and excludes 116. The current comparison uses all 125 included records, with nonempty public titles and curated mechanism summaries, passing or nonblocking-uncertainty QC, and unique blind task identities. Six included records retain predecessor-equivalence uncertainty; none has a confirmed equivalent precursor in the release. The 125 curated paper-context problems are the evaluation units.

Reference labels and visible evidence. The reference bottleneck and experiment ID are compiled from the curated Form A mechanism summary using the frozen text-to-menu mapping. Pattern counts and deterministic tie-breaking select a bottleneck, with a declared checkpoint/dysfunction fallback;

Table 10: **Current task construction and label coverage.** The left column separates review rows, unique papers, and mutually exclusive dispositions; the right gives the nine reference classes in the 125-task collection. The task inventory is fixed by the September release.

Paper flow / disposition	Count	Reference bottleneck	Count
Input review rows	281	Checkpoint / T-cell dysfunction	41
Duplicate groups resolved	3	Myeloid reprogramming	35
Unique released papers	278	Antigen presentation / IFN	27
Included tasks	125	Stromal / TGF-beta exclusion	13
Deferred papers	37	Oncogene–immune crosstalk	4
Excluded papers	116	Metabolic / hypoxia relief	2
Execution seeds per task	2	Antigen / vaccine / neoantigen	1
Records per complete configuration	250	Lymphoid organization / TLS	1
Complete configurations	7	Vascular / trafficking	1

ordered intervention patterns and a bottleneck-dependent fallback select among ten experiment IDs. Curated scientific content and its deterministic label projection are separate provenance layers. The nine benchmark classes, five evidence-question axes, and seven functional diagnosis axes are not interchangeable taxonomies.

Agents receive the public title and declared PDAC/immune-cold predicates, task-conditioned pre-cutoff evidence, and the closed menus. The target PMID is excluded from the retrieval surface; the private mechanism summary, gold choices, and final outcomes are withheld. The initial retrieval policy selects up to fourteen evidence records from its eligible pool and constructs source-pointer memory. Prompt construction uses bounded evidence retrieval; the 872-claim resource remains the wider evidence collection. Worker-specific truncation and output projection remain part of each assessed configuration.

The evidence cutoff of December 31, 2024 is distinct from the target-paper availability window and September 2026 resource release. Public target titles are intentionally visible. Outcome masking controls private information supplied at decision time. Public titles retain their scientific content, and foundation-model pretraining exposure lies outside this input-control protocol. Internal encoded outcomes in the task wrapper are also distinct from the archival functional assessment in Appendix H.

F.3 Native-code comparison and decision-time information

The complete-cohort comparison includes Mechanova and six public native-code configurations: Biomni, BioDiscoveryAgent, OpenScientist, AI Scientist-v2, SPARK, and Jataware Open Co-Scientist (Huang et al., 2026; Roohani et al., 2025; Roberts et al., 2026; Yamada et al., 2025; Trost et al., 2026). The last is a public replica, not Google's proprietary Co-Scientist (Gottweis et al., 2026). Seeds 1203 and 1211 yield 250 records per complete configuration. At the manuscript snapshot (23 September 2026, 17:24:13 UTC), SciAgents has 208 scored records across 106 tasks and 42 blocked records. It remains partial and outside the seven-configuration, 1,750-record comparison. Appendix I.1 binds this dated snapshot independently of the evolving runtime leaderboard.

Model proposals, policy decisions, and scoring. The initial Mechanova package contains model-generated hypotheses. The authority policy then combines task-local title/context signals, quoted support, retrieved evidence, and deterministic fallback. Its title parser counts the short token `tan` only as a whole word and omits the bare `spatial`/`metabolic` bonuses, before the inherited window-filtered decision. The mechanistic selection and experiment ranking are separately recorded. The complete scored configuration includes the explicit rule-assisted policy, with the standalone C6 MEC override disabled.

External hypotheses and native text are projected by a separate Qwen3.8-27B call into normalized bottleneck weights and an experiment ranking. The projection uses temperature zero, disabled thinking, an 8,192-token response ceiling, and bounded native text. Projected probabilities reflect the separate projector's interpretation of native outputs and public context. The comparison evaluates

Table 11: **Mechanova improves both initial-decision endpoints over frozen title-only routing.** Accuracy is shown as correct/total (%); Δ is the percentage-point difference. C and H count paired corrections and harms, respectively; net is C−H. The 125-task cohort contributes two seeds per task; dev105 is a separate development analysis.

Cohort / endpoint	`title_v1`	Mechanova	Δ (pp)	C / H / net
125-task, mechanism	138/250 (55.2)	**152/250 (60.8)**	**+5.6**	26 / 12 / +14
125-task, intervention	126/250 (50.4)	**130/250 (52.0)**	**+1.6**	22 / 18 / +4
dev105, mechanism	57/105 (54.3)	**73/105 (69.5)**	**+15.2**	16 / 0 / +16
dev105, intervention	58/105 (55.2)	**61/105 (58.1)**	**+2.9**	11 / 8 / +3

bounded native-code configurations through this projection stage, with configuration-specific compute budgets.

Mechanova is scored from the pre-observation bottleneck distribution and original experiment preference order. Later posteriors and downstream reordered lists do not replace these choices. The decision policy is re-executed against the same previously generated hypothesis packages, with cache reuse included in model-cost accounting.

Decision gains over frozen title-only routing. The `title_v1` control applies the frozen `infer_bottleneck_from_context` rule to the same public task context and maps its selected bottleneck to an experiment using `BOTTLENECK_TO_EXPERIMENT`. This comparator remains fixed; it is not Mechanova's internal `strict_title_rule`. Mechanism and intervention endpoints are scored against the same cohort-specific references, using decisions committed before observation. Table 11 reports the complete policy's incremental performance over this reference.

On 125 tasks, mechanism choices agree with `title_v1` in 204/250 records and differ in 46/250; the differing choices span 23 papers and are consistent across both seeds. Experiment choices agree in 190/250 records and differ in 60/250. Mechanova alone is correct in 26 mechanism records versus twelve for the control, and in 22 intervention records versus eighteen for the control. Thus, decision changes yield net improvements in both endpoints, with 125 tasks—not 250 independent observations—as the sampling units.

On dev105, mechanism choices agree in 86/105 cases and differ in nineteen: sixteen corrections, no harms, and three wrong-to-wrong changes. Mechanism accuracy increases from 57/105 to 73/105; intervention accuracy increases from 58/105 to 61/105, with eleven corrections and eight harms. These development results remain separate from the 125-task comparison. The table summarizes descriptive paired gains for the complete policy. Graph-edit attribution is evaluated separately under matched downstream rules (Section 4.2).

F.4 Public evidence questions and target-family evaluation

Five axes—myeloid, stromal/TGF-beta, antigen-presentation/IFN, vascular/trafficking, and metabolic/hypoxia—cross nine predefined question lenses to produce 45 public tasks. Each matched task–seed comparison uses the same twelve-record pre-cutoff packet. The objective, biological context, assay/intervention classes, and focus terms are public; target-family identities and reference propositions are not. The target frame contains 186 positive, 89 conflict-negative, and 47 unresolved families, with canonical propositions chosen by fixed identifier ordering. Lenses share their axis's target frame rather than supplying nine independent biological corpora.

Eight parent configurations yield $45 \times 3 \times 8 = 1{,}080$ outputs; the frozen compiler adds 135 outputs from regenerated Core parents. Seeds are 501, 509, and 521. Primary matched references are Core, unconditional atomic repair, and Full-reference; the other parent outputs support execution-integrity checks. The compiler uses public spans and parent content, never evaluator target tuples. Its budget includes the parent and permits at most two logical model calls and 24,000 approximate tokens. The completed runs need no additional editor-model call. The successor preserves the public tasks and packets while rejecting malformed primary claims before envelope construction; it does not repair

Table 12: **Evaluation inventory and analysis units.** Counts include all displayed factors; repeated seeds, views, or model executions do not add independent source papers. Results are detailed in Appendix G.

Study	Analysis unit	Execution / data scale	Endpoint
Native-agent comparison	125 paper-context tasks	$125 \times 7 \times 2 = 1{,}750$	Initial top-1
Evidence licensing	45 tasks; 5 axes	1,080 parent + 135 compiler	Recovery; support
Selective memory	12 task groups	$12 \times 3 \times 4 \times 3 = 432$	Paired utility
MEC	12 task groups	$12 \times 4 \times 3 = 144$	Grounding; conflict
Hierarchical execution	12 groups per split	$12 \times 2 \times 3 = 72$ per split	Useful completion
Environment / skills	12 groups per split	72 per contrast per split	Useful completion
Adaptive horizons	8 tasks	2 / 3 / 3 short/medium/long routes	Tokens; coverage
OWM forecasting	4 held-out campaigns	72 train + 8 held-out transitions	Forecast errors
Hidden dynamics	24 locked groups	$24 \times 3 \times 3 = 216$	Decision utility
DepMap acquisition	80 models per release	$80 \times 2 = 160$ executions	Brier; acquisition
Active screens	30 models	36 interventions; 5 queries	Class prediction
Protected responses	3 source groups; 8 panels	31 protected-pair views; at most 2 queries	Regret + cost
Response representations	Paired perturbations	4 few-shot contexts	Correlation
Matched edit permission	88 source families	528 two-stage trials; 376 graph-substitution calls	Brier; graph sensitivity
Archival programs	10 contexts / 7 families	$10 \times 3 = 30$ assessments	Functional criterion
Policy execution	106 source groups	$150 \times 3 = 450$ greedy episodes	Score; recovery; cost
Logged-sequence diagnostic	106 lineage groups	150 logged trajectories; 2 policies	Logged-sequence value

them into acceptance, replace tasks, or change the licensing rule. Freeze and terminal records precede private family matching.

Target recovery and source support answer different questions. For candidate c and target family f, the frozen bidirectional local NLI rule is

$$M(c, f) = \mathbf{1}\{\max(e_{c\to f}, e_{f\to c}) \geq 0.60, \max(d_{c\to f}, d_{f\to c}) < 0.25\}\,, \tag{29}$$

where e, d are entailment and contradiction outputs. With submitted length $K \leq 10$ and $m_f = \sum_{c=1}^{K} M(c, f)$,

$$R_f@10 = \mathbf{1}\{m_f > 0\}, \qquad P_f@10 = m_f/10, \qquad Y_f = \begin{cases} m_f/K, & K > 0, \\ 0, & K = 0. \end{cases} \tag{30}$$

Family-averaged precision and yield are registered target-recovery quantities, not precision against one globally deduplicated answer set. Negative and unresolved strata use false promotion $m_f/10$ and nonpromotion $\mathbf{1}\{m_f = 0\}$; their results remain separate (Table 15). Source support instead measures A_u/A for unsupported assertions among A audited assertions, assigning zero when $A = 0$. Licensed coverage is fully licensed cards divided by ten. Nonempty-output frequency and card count qualify zero unsupported-assertion rates; empty output alone is not successful recovery.

F.5 Layered evaluation and matched component designs

Table 12 separates resource scale, execution work, and the unit supporting inference, following task-oriented agent evaluation (Liu et al., 2024; Ruan et al., 2024). No sum of these heterogeneous rows is reported as a biological sample size. Providers and learning procedures are specified in Appendices I.1, D, and E; this section records their evaluation design.

Memory crosses three arms with clean, stale, conflicting, and poisoned-pointer conditions; the main contrast is provenance-aware versus episodic memory. MEC compares full coupling, no MEC, no counterevidence, and no formal knowledge. Seeds and conditions remain within their twelve task

groups. Typed-grounding and contradiction endpoints are evaluator proxies, and the two targeted removals are distinct comparisons.

Hierarchical execution uses twelve development and twelve confirmation groups. Both arms receive the same model, task, seed, tools, 32,768-token context capacity, 2,048-token response ceiling, twelve turns, and eight tool calls; seeds are 13, 31, and 53. Clean and recovery campaigns remain in both arms. Environment and skill contrasts separately use six turns, eight tool calls, and seeds 11, 29, and 47. Adaptive horizons retain their eight-task development status. Policy execution uses 150 tasks per arm in 106 source groups, disjoint from the 337-trajectory fit and all 420 earlier trajectories. Training partitions and repeated policy executions are not added to the source-group denominator.

F.6 Source-grouped structural-revision evaluation

Design and information boundary. The matched two-stage study uses 88 source families, edit-on and edit-off arms, and seeds 1701, 1707, and 1713: 528 factual records, with 264 per arm. Both arms share context, public evidence, the initial mechanism, an opened trigger observation, follow-up event descriptions, and the intervention panel. Stage one varies edit permission; stage two uses identical probability and action rules (Appendix C.5). The study estimates the contribution of edit permission under matched conditions using newly generated completions.

Score predictions and intervention choices separately. Binary event outcomes and panel scores remain evaluator-side. Unknown outcomes are excluded; missing or abstained forecasts for known events receive probability 0.5. Candidate panels share a two-action cap. For known events K_t and evaluable panel P_t,

$$B_t = \frac{1}{|K_t|} \sum_{e \in K_t} (p_{te} - y_{te})^2, \qquad r_t = \frac{\max_{a \in P_t} u_{ta} - u_{t\hat{a}}}{\max_{a \in P_t} u_{ta} - \min_{a \in P_t} u_{ta}}. \tag{31}$$

An unmapped or unmeasured recommendation receives the worst evaluable panel score; a zero score range yields zero regret. Selection accuracy requires an evaluable choice with maximal encoded score. Panel scores are curated numerical encodings, so regret measures agreement with the registered panel rather than a clinical treatment effect. Brier measures probabilistic predictive quality, not calibration alone. Seeds are averaged within task and family before computing paired differences; 10,000 family-bootstrap draws use analysis seed 20260920.

Primary endpoints and fixed acceptance criteria. The protocol orders three primary endpoints: native Brier difference, child-versus-parent output change, and edit occurrence when needed. Its acceptance conjunction requires a negative Brier interval, a positive graph-replacement change interval, edit occurrence when needed at least 0.70, unnecessary-edit rate at most 0.25, verified prompt isolation, and unchanged source artifacts. Intervals are pointwise rather than simultaneous. Native regret, selection accuracy, frozen graph-readout metrics, and source-quote overlap remain secondary; they do not replace a primary endpoint.

Graph interventions retain the family-level denominator. The 188 edit-on trials with a changed structural signature span 64 families. Each receives a parent-graph substitution and a same-seed cross-family child substitution, giving 376 additional stage-two calls. Output change is a changed recommendation or rounded probability/abstention output. Indicators are first averaged over eligible seeds within each family and then over the 64 families. These rates concern eligible graph-changing trials, not all 264 edit-on trials; cross-family substitution is a sensitivity probe rather than a feasible biological intervention.

Necessary and unnecessary revision are evaluated on different strata. Edit occurrence when needed averages structural-change indicators over the 75 evaluator-labeled `should_edit=true` families. Unnecessary editing averages over the thirteen `should_edit=false` families. These labels are hidden from the model. This endpoint records whether a required structural change occurred; relation correctness and operation-type agreement require separate assessment. Appendix G.8 reports both stratum-specific and all-record diagnostics, and Appendix I.1 binds the prompts, outputs, and scorer.

F.7 Paired estimands, dependence, and outcome accounting

For unit g, arm a, and seed s, average repeated scores within the unit to obtain $\bar{x}_{g,a}$. The equal-unit paired estimand is

$$\widehat{\Delta} = \frac{1}{G}\sum_{g=1}^{G}(\bar{x}_{g,\text{Mechanova}} - \bar{x}_{g,\text{reference}}). \tag{32}$$

Resampling retains both arms and all nested observations of each sampled unit. Marginal intervals resample an arm's unit means; paired intervals resample within-unit differences. Overlap of marginal intervals is not a paired test. For the 125-task comparison, marginal intervals use 10,000 draws with seed 20260922, and the twelve external-comparator contrasts use seed 20260923. These paired reporting analyses provide pointwise intervals. The frozen title-only table reports observed paired counts and gains, without a statistical-significance claim.

Task-level and axis/family sensitivity are separate analyses. For the evidence benchmark, family scores are first averaged within the task's axis and then across three seeds. Primary intervals use 10,000 paired draws over 45 task means, with analysis seed 330033 and deterministic comparison-specific offsets. Common support/cost scores omit the family average. The stored sensitivity first averages each family's scores across its nine lenses and three seeds, then resamples five axes and families within sampled axes in paired hierarchical draws. Table 15 reports the resulting intervals without substituting them for primary intervals. This checks dependence on the shared target frame; it does not create additional independent biological sources.

Study-specific aggregation and scoring. Component effects resample task groups with their nested seeds; memory also retains the reliability conditions within each group. Active-screen random orders are averaged within model before model-paired comparison, with likelihood-calibration and evaluation lineages separated. Protected-pair scores average views within source and then weight the three source groups equally; eight panels are nested within those groups. Archival assessment reports complete counts for ten contexts from seven source families, with overall and evaluable denominators distinguished. Actual policy-execution effects average task scores within each source group before paired group resampling; descriptive task means retain equal task weights. The completed-sequence analysis uses logged action probabilities and lineage-group resampling (Appendix E.5).

Operational probability forecasts use the declared class-averaged Brier score, calibration summary and numerical errors; binary endpoint prediction uses $1-(p-y)^2$ Brier utility (Gneiting and Raftery, 2007). Evidence-acquisition noninferiority requires the paired utility interval's lower bound to exceed -0.01. Useful completion requires verified task outcomes, not protocol conformance alone. Terminal failures remain in their registered execution denominators; absent archival class matches remain non-evaluable without nearest-arm substitution. Frozen scoring definitions, gate outcomes, and source records are unchanged by this presentation.

G Extended experimental results

This section expands Section 4 with paired effects, condition-specific attribution, predictive diagnostics, and learning support. Each table retains its analysis unit and evaluation stage; protocols and estimators are specified in Appendix F. Archival outcomes and graph-level trajectories remain in Appendix H.

G.1 Initial scientific decisions across native agents

Complete-cohort decisions and paired gains. Mechanova leads both initial-decision endpoints across the seven complete configurations: **60.8% mechanism identification and 52.0% pre-observation intervention agreement**. Table 13 combines absolute performance with all twelve Mechanova-minus-comparator contrasts on the same 125 paper-context tasks. Each configuration contributes two seeds per task, giving 1,750 records overall. Biomni is the strongest completed mechanism comparator, and OpenScientist the strongest intervention comparator. The majority reference class contains 41/125 tasks (32.8%). Incomplete configurations are excluded from complete-cohort comparisons rather than scored as zero.

Table 13: **Initial scientific decisions and paired gains on 125 paper-context tasks.** A: Top-1 accuracy (%) with marginal task-bootstrap 95% intervals; each configuration contributes 250 records. B: Mechanova-minus-comparator differences in percentage points, paired over 125 tasks after averaging two seeds. All intervals use 10,000 draws; paired intervals are pointwise. Open Co-Scientist denotes the public replica.

A. Initial decisions: 250 records per configuration

Configuration	Mechanism Top-1 ↑	Intervention Top-1 ↑
Mechanova	**60.8 [52.0, 68.8]**	**52.0 [43.2, 60.8]**
Biomni	41.2 [33.2, 49.2]	37.6 [30.0, 45.2]
OpenScientist	40.8 [33.2, 48.4]	39.6 [32.0, 47.6]
BioDiscoveryAgent	40.0 [32.4, 47.6]	36.8 [29.2, 44.8]
AI Scientist-v2	35.2 [27.6, 42.8]	33.2 [25.6, 40.8]
Open Co-Scientist replica	34.4 [27.2, 41.6]	30.4 [23.2, 37.6]
SPARK	22.0 [15.6, 28.8]	25.2 [18.4, 32.4]

B. Paired gains on the same 125 tasks

Reference	Mechanism Δ [95% CI]	Intervention Δ [95% CI]
Biomni	+19.6 [9.6, 29.2]	+14.4 [4.4, 24.0]
OpenScientist	+20.0 [9.6, 30.0]	+12.4 [2.4, 22.4]
BioDiscoveryAgent	+20.8 [10.4, 30.8]	+15.2 [4.8, 25.6]
AI Scientist-v2	+25.6 [14.8, 36.0]	+18.8 [8.8, 28.8]
Open Co-Scientist replica	+26.4 [15.6, 36.8]	+21.6 [11.2, 32.0]
SPARK	+38.8 [29.6, 47.6]	+26.8 [16.8, 36.8]

The paired contrasts score the initial, pre-observation decisions on a shared task inventory. All twelve pointwise intervals favor Mechanova. The frozen title-only comparison in Table 11 separately quantifies the complete policy's observed incremental gains over a fixed rule. Section 4.2 isolates edit permission under matched downstream conditions. Together, these analyses distinguish system performance, rule-relative improvement, and structural-revision attribution.

G.2 Evidence licensing: recovery, support, and robustness

The frozen compiler improves positive-family recovery without another editor-model call (Table 14). All 135 outputs are nonempty, with one to six cards and mean 2.52, alongside zero measured unsupported assertions. Approximate generation tokens, including the parent, decrease by 63.6% against the full-reference pipeline. The complete 1,080-parent execution lattice and three registered comparator arms are distinguished in Appendix F.4.

All six task-paired recall/yield contrasts are positive; none of the precision intervals excludes zero. The separate five-axis/family sensitivity analysis preserves recall and yield gains despite shared target frames (Table 15). All seventeen registered gates, including the noninferiority quality gates, pass under their original definitions. Secondary nonpromotion differences are negative against all three references, whereas FP@10 difference intervals include zero. Source licensing supports grounded submission, while the negative-target results characterize a distinct discrimination endpoint.

G.3 Attributing evidence quality and selective reuse

MEC's typed representation and opposing evidence have distinguishable effects: removing formal knowledge reduces typed grounding, whereas removing counterevidence reduces contradiction recall. Table 16 retains all primary-comparison endpoints and component-specific contrasts. These twelve-group development metrics are lexical/structural proxies. Full MEC has an unsupported-claim rate of 0.0306 versus 0.0000 for no MEC, with paired difference +0.0306 [0.0000, 0.0861]. This source-support trade-off accompanies the utility and grounding gains.

Provenance-aware memory is most useful when reusable evidence becomes unreliable (Table 17). Its aggregate utility advantage concentrates in stale, conflicting, and poisoned pointer conditions.

Table 14: **All registered comparator means on 45 public questions.** Each arm contributes 135 outputs. Recall, precision, and yield use the target-family definitions in Appendix F.4; source-support checks have a different denominator. Atomic denotes unconditional atomic repair. Tokens measure generation, including the parent, and exclude evaluator computation. Negative and unresolved targets remain separate from positive-family recovery.

Endpoint	Core	Atomic	Full-reference	Compiler
Recall@10 ↑	0.2631	0.2673	0.2516	**0.3683**
Precision@10 ↑	0.0422	0.0438	0.0392	0.0507
Supported yield ↑	0.1149	0.1141	0.1082	**0.1902**
Unsupported assertions ↓	0.5394	0.5325	0.5160	**0.0000**
Licensed submission coverage ↑	0.0800	0.0867	0.0793	0.1911
Citation resolution ↑	0.8815	0.8815	0.8444	1.0000
Approximate tokens ↓	8,859.8	13,697.1	24,306.9	8,859.8
Model calls ↓	1	2	3	1
Conflict-negative FP@10 ↓	0.0453	0.0470	0.0431	0.0533
Conflict-negative nonpromotion ↑	0.7190	0.7163	0.7303	0.6103
Unresolved FP@10 ↓	0.0470	0.0483	0.0456	0.0506
Unresolved nonpromotion ↑	0.7524	0.7477	0.7360	0.6252

Table 15: **Paired recovery, target strata, and dependence sensitivity.** A: compiler-minus-reference task-paired effects after family/seed averaging. B: descriptive full-reference contrasts across nine lenses per axis. C: secondary full-reference comparisons; other references have the same nonpromotion direction. D: nested 95% sensitivity intervals over five axes and their positive families after lens/seed averaging, using the same point estimates as A. Sensitivity complements rather than replaces the registered 45-task analysis.

A. Positive-family effects with task-bootstrap 95% intervals

Reference	Recall Δ [95% CI]	Yield Δ [95% CI]	Precision Δ [95% CI]
Core	+0.1052 [0.0230, 0.1888]	+0.0753 [0.0354, 0.1176]	$+0.0086\ [-0.0095, 0.0265]$
Atomic	+0.1010 [0.0167, 0.1871]	+0.0760 [0.0353, 0.1177]	$+0.0070\ [-0.0112, 0.0257]$
Full-reference	+0.1167 [0.0391, 0.1959]	+0.0820 [0.0451, 0.1200]	$+0.0115\ [-0.0069, 0.0300]$

B. Biological-axis heterogeneity

Axis	Recall Δ	Supported-yield Δ
Antigen/IFN	+0.0417	+0.0426
Metabolic/hypoxia	+0.1275	+0.0584
Myeloid	+0.1580	+0.1183
Stromal/TGF-beta	+0.2634	+0.1267
Vascular/trafficking	−0.0071	+0.0640

C. Secondary target errors: compiler minus full-reference

Target stratum	FP@10 Δ [95% CI]	Nonpromotion Δ [95% CI]
Conflict-negative	$+0.0102\ [-0.0110, 0.0315]$	$-0.1200\ [-0.2212, -0.0191]$
Unresolved	$+0.0050\ [-0.0145, 0.0233]$	$-0.1107\ [-0.2086, -0.0164]$

D. Paired five-axis/family sensitivity intervals

Reference	Recall 95% CI	Yield 95% CI
Core	[0.0227, 0.2007]	[0.0398, 0.1150]
Atomic	[0.0221, 0.1982]	[0.0417, 0.1144]
Full-reference	[0.0333, 0.2045]	[0.0515, 0.1168]

Table 16: **MEC quality, cost, and component attribution.** Four arms, twelve groups, and three seeds yield 144 executions. Primary differences are full MEC minus no MEC, with task-group 95% intervals; the final two rows use their explicitly named references. Utility, grounding, source support, and cost are reported together.

Endpoint	Full MEC	No MEC	Δ [95% CI]
Utility ↑	0.6907	0.6108	+0.0799 [0.0549, 0.1079]
Typed grounding ↑	0.8344	0.2322	**+0.6022 [0.5129, 0.6842]**
Contradiction recall ↑	0.3053	0.1428	+0.1625 [0.0771, 0.2490]
Unsupported claims ↓	0.0306	0.0000	+0.0306 [0.0000, 0.0861]
Exact-support proxy ↑	0.9694	1.0000	−0.0306 [−0.0889, 0.0000]
Target-relevance proxy ↑	0.3499	0.3517	−0.0018 [−0.0447, 0.0423]
Calibration score ↑	0.5890	0.5761	+0.0128 [−0.1104, 0.1635]
Falsifiable-hypothesis rate ↑	0.6019	0.5648	+0.0370 [−0.1343, 0.1898]
Normalized cost ↓	0.6461	0.5432	+0.1029 [0.0862, 0.1235]
Latency (s) ↓	233.08	223.51	+9.57 [−3.66, 25.01]
Component-specific effects			
Reference removed	Endpoint		Δ [95% CI]
Counterevidence	Contradiction recall		+0.3053 [0.2332, 0.3813]
Formal knowledge	Typed grounding		+0.5652 [0.4747, 0.6479]

Table 17: **Reliability-conditioned memory effects.** Full provenance-aware minus episodic memory; 432 executions cover three arms, four conditions, twelve groups, and three seeds. Intervals resample groups, not the 144 matched executions per arm. The registered aggregate utility/safety conjunction passes; subgroup effects remain separately interpretable.

A. Utility by pointer condition

Condition	Δ	95% CI
Clean	−0.0046	[−0.0090, −0.0005]
Conflicting	+0.0776	[0.0737, 0.0817]
Poisoned	+0.0675	[0.0587, 0.0765]
Stale	+0.0689	[0.0622, 0.0760]

B. All registered aggregate endpoints

Endpoint	Δ	95% CI
Utility ↑	**+0.05237**	**[0.04957, 0.05504]**
Task success ↑	0.00000	[0.00000, 0.00000]
Memory-induced hallucination ↓	+0.00099	[0.00000, 0.00298]
Negative transfer ↓	−0.00238	[−0.00476, −0.00026]
Repeated failure ↓	0.00000	[0.00000, 0.00000]
Counterevidence retention ↑	+0.25000	[0.25000, 0.25000]
Rollback recovery ↑	+0.50000	[0.50000, 0.50000]
Maintenance cost ↓	0.00000	[0.00000, 0.00000]

Counterevidence retention and rollback recovery describe different recovery routes; their constant group contrasts are bookkeeping properties of this controlled test. Against no memory, utility improves by 0.0690 [0.0587, 0.0808] with maintenance-cost increase 3.4167 [2.5000, 4.3333].

G.4 Operational forecasting, action value, and evidence efficiency

Three tests isolate different consequences of operational foresight (Table 18). The family-disjoint replay evaluation covers eighty unresolved transitions and measures numerical forecast error within that observed class. An always-unresolved predictor would have zero Brier loss on these labels. Outcome-set coverage measures inclusion of the observed outcome in the predicted set. The hidden-

Table 18: **OWM evidence at three levels.** A: controlled replay, 72 training transitions and eight held-out transitions from four campaigns in a disjoint family. B: 24 locked groups, three seeds, and three arms; effects compare memory/reflection-enhanced with dynamics-only OWM, using 4,000 grouped bootstrap draws. C: eighty model-level units per release; $U = 1 - (p - y)^2$. Forecast error, decision utility, and endpoint utility are different estimands.

A. Held-out transition diagnostic

Brier ↓	ECE ↓	Coverage ↑	IG / cost MAE	State MAE
0.00283	0.1190	1.000	0.000 / 0.000	0.000

Observed labels: unresolved 80; the other five outcome classes have zero observations.

B. Committed actions in hidden dynamics

Endpoint	Δ	95% CI
Decision utility ↑	**+0.1400**	**[0.0963, 0.1750]**
Harmful-outcome rate ↓	0.0000	[0.0000, 0.0000]

C. Frozen evidence-horizon policy on real data

Release / policy	Brier utility ↑	Accuracy (%)	Panels	Saving (%)
26Q1 / full-panel reference	0.753015	63.75	10	—
26Q1 / evidence horizon	0.752787	65.00	8	20
24Q2 / full-panel reference	0.726720	55.00	10	—
24Q2 / evidence horizon	0.726475	53.75	8	20

26Q1: $\Delta U = -0.000228$ $[-0.007151, 0.007174]$; noninferiority and efficiency pass.
24Q2: $\Delta U = -0.000244$ $[-0.007637, 0.007469]$; historical stress, same criteria pass.

dynamics comparison instead establishes decision value from memory and selective reflection, with choices committed before transition materialization and evaluation episodes excluded from operational memory.

The real-data controller retains the reference endpoint predictors and fixed panel order, stopping at train-selected objective-specific horizons: nine panels for KRAS and seven for PIK3CA. Both release evaluations meet the aggregate 0.01 Brier-utility noninferiority criterion with 20% fewer panels. An exploratory analysis performed after the historical stress evaluation places three Brier-utility differences below -0.05 in the most distant quartile (3/20 versus 0/60 elsewhere). This observation motivates prospective testing of a shift-warning rule.

G.5 Biological prediction and informative experimental decisions

Response representation across cellular backgrounds. The four-context few-shot evaluation contains 4,058 paired perturbation–context records. ST-HVG denotes the provider checkpoint `ST-HVG-Replogle` with representation `X_hvg`; ST-SE denotes `ST-SE-Replogle` with `X_state`. Both use the provider's frozen few-shot evaluation, non-targeting controls, and perturbation-identity pairing; checkpoint revisions and metric directions are bound in Appendix I.2. ST-HVG improves Pearson transition fidelity, differential-expression direction agreement, and mean-squared transition error over ST-SE in every frozen context (Table 19). The combined representation record also retains three Tahoe pharmacological zero-shot and four Replogle genetic zero-shot contexts: eleven evaluated context–regime combinations and 33 favorable metric tests. These correlated tests characterize the supplied checkpoint configurations across the evaluated cellular contexts and biological endpoints.

Selecting informative measurements across screens. Likelihood calibration and evaluation use separate screens and disjoint lineages. Table 20 exposes the full selection ladder on thirty balanced models and a thirty-six-intervention menu, including all recorded budgets. At the registered five-query endpoint, expected information gain (EIG) improves reference-label log score over random, uncertainty-only, and metadata-family cycling. Accuracy need not increase monotonically with budget: the five-query endpoint is retained rather than replaced by the highest observed accuracy.

Table 19: **Complete Replogle few-shot representation comparison.** Effects have paired-perturbation 95% intervals. Pearson and differential-expression (DE) direction gains are ST-HVG minus ST-SE; MSE reduction reverses the subtraction and is displayed in 10^{-3} units. Positive values consistently favor ST-HVG. Counts sum to 4,058, not 4,058 independent source studies.

Context	Pairs	Pearson gain [95% CI]	DE-direction gain [95% CI]	MSE reduction [95% CI]
HepG2	945	0.0416 [0.0335, 0.0498]	0.0336 [0.0250, 0.0423]	0.846 [0.661, 1.039]
Jurkat	1,084	0.0844 [0.0772, 0.0915]	0.0611 [0.0521, 0.0699]	1.336 [1.199, 1.465]
K562	968	0.0424 [0.0364, 0.0485]	0.0276 [0.0190, 0.0361]	0.620 [0.542, 0.698]
RPE1	1,061	0.0209 [0.0112, 0.0304]	0.0084 [0.0015, 0.0155]	0.769 [0.479, 1.052]

Table 20: **Cross-screen active mechanism identification.** Top-1 values are percentages; random-order repetitions are averaged within each model. The same thirty models support every budget, so additional queries are not additional samples. Paired intervals use model-level log-score differences at the registered five-query endpoint; other budgets are descriptive.

	Top-1 by query budget ↑				**Five-query log score ↑**
Policy	1	3	5	8	Mean
EIG	43.3	63.3	**56.7**	60.0	−1.0153
Random order	37.5	40.9	42.9	45.6	−1.0788
Uncertainty only	36.7	43.3	33.3	46.7	−1.0968
Metadata-family cycle	30.0	36.7	40.0	53.3	−1.0840

Reference	EIG log-score gain	95% paired CI
Random order	+0.0635	[0.0262, 0.1011]
Uncertainty only	+0.0815	[0.0332, 0.1312]
Metadata-family cycle	+0.0687	[0.0059, 0.1294]

Transfer to protected, unmeasured intervention targets. The protected-pair regime uses 31 overlapping views of eight panels from three source groups; 51 is the view count across all four decision regimes. Permitted non-target observations improve recommendations relative to the same prior without queries (Table 21). Against random querying, the main advantage is fewer measurements at similar top-1 agreement. Removing off-diagonal covariance eliminates protected-target updating in this comparison, while one- and two-step planning tie. The result supports decision-relevant transfer and acquisition efficiency, not an advantage of greater lookahead depth.

Source-specific cost-adjusted gains over no-query are 0.02204, 0.04985, and -0.00778; leave-one-source-out averages remain positive (0.00713–0.03595). The advantage is regime-specific: diagonal covariance is slightly better in original-full and matched-count views, and typed active is better in diagnostic transfer (0.07124 versus 0.11420). These comparisons quantify predictive decision

Table 21: **Protected-target prediction and acquisition-cost decomposition.** All values are source-macro development summaries; overlapping views and 7,650 simulations across the full study do not support episode-level confidence intervals. The final target is never queried. Lower regret and cost are better; top-1 is higher-better.

Policy	Regret	Queries	Query cost	Top-1 (%)	Regret + cost
Full active	0.01652	0.809	0.00886	83.41	**0.02538**
Full, no queries	0.04675	0.000	0.00000	74.37	0.04675
Full, random queries	0.01722	1.556	0.02542	83.07	0.04264
Typed active	0.04642	1.030	0.01397	73.67	0.06039
Diagonal covariance	0.04675	0.000	0.00000	74.37	0.04675
Global-matched active	0.08841	0.926	0.01070	62.11	0.09911
Full, one-step	0.01652	0.809	0.00886	83.41	0.02538

Table 22: **Where execution gains arise.** Each panel is a separately frozen two-arm study with twelve groups and three seeds per split. Counts denote executions, not independent tasks; differences are percentage points (pp), except cost/exposure rows. Intervals use paired group resampling; clean/recovery counts are descriptive. Skill tokens retain corrected accounting for terminal failures.

Endpoint	Reference	Component	Δ	95% CI
A. Flat loop versus observation-conditioned hierarchy				
Development completion	17/36	**36/36**	+52.8 pp	[25.0, 77.8]
Confirmation completion	18/36	**36/36**	**+50.0 pp**	[25.0, 75.0]
Clean confirmation	18/18	18/18	0.0 pp	—
Recovery confirmation	0/18	**18/18**	+100.0 pp	—
B. Full-inventory versus progressive disclosure: confirmation				
Useful completion	18/36	**32/36**	+38.9 pp	[25.0, 55.6]
Visible context characters	19,832	12,261	$-7,571$	$[-10,435, -4,367]$
C. Protocol-only versus outcome-grounded skills: confirmation				
Useful completion	8/36	**34/36**	**+72.2 pp**	[52.8, 88.9]
Protocol conformance	100%	100%	0.0 pp	—
Mean tool calls	4.33	**2.00**	-2.33	$[-2.67, -2.03]$
Mean corrected model tokens	33,563	**21,117**	$-12,446$	$[-15,173, -9,678]$

quality and acquisition cost; Appendix H reports the proxy-assisted assessment of immune-dependent functional outcomes.

G.6 Observation-conditioned execution and selective scientific effort

Execution gains concentrate in recovery: both controllers complete all eighteen clean confirmation trials, but only hierarchical control completes the eighteen recovery trials. Table 22 retains development/confirmation separation, context exposure, and corrected skill costs. Progressive disclosure and outcome-grounded skills test different execution interventions; both skill arms have perfect protocol conformance despite sharply different useful completion.

The eight-task routing study evaluates selective continuation (Table 23). It saves 31.2% of tokens while retaining every registered typed-evidence slot. This score evaluates the declared model-plus-clause-grounded state; legacy target recall remains 0.7967 alongside complete essential-slot coverage (1.0). Across sixteen evaluated checkpoints, including forced-counterfactual checkpoints after routed stopping, the primary sufficiency criterion records nine productive continuations and seven safe stops. The separate target-increment criterion has precision/recall 0.667/0.857. The earlier fixed-horizon confirmation evaluates a different estimand: its 100-minus-30 utility interval includes zero.

G.7 Experience-driven policy improvement across investigations

The action-time policy study tests learned research-action selection through controlled execution after offline fitting and freezing. Each of three policies executes the same 150 source-disjoint tasks in 106 groups. Both recurrent configurations improve task-mean terminal score from 0.4453 to approximately 0.754 and objective recovery from 32.7% to 71.3% (Table 24). Equal-group score gains are $+0.2942$ [0.2129, 0.3759] for supervised GRU and $+0.2998$ [0.2221, 0.3763] for AWR GRU relative to the retrained linear reference. These contrasts compare the full trained architectures and their associated capacity and optimization settings.

Isolating the contribution of return weighting. The matched AWR-minus-supervised score interval includes zero, leaving the difference unresolved in this setting. Their action sequences differ on only seven tasks: AWR wins three, supervision four, and 143 tie in score. The reference terminates with reject on 128/150 tasks, compared with 34/150 for supervised GRU and 36/150 for AWR. Across three task families, recovery is respectively 42/72/74% (reference/supervised/AWR) for claim–evidence discovery, 34/58/58% for support verification, and 22/84/82% for counterevidence stress. These task-family results describe subgroups within the same evaluation cohort.

Table 23: **Complete task-level adaptive-horizon diagnostics.** Entries count matched/required essential slots in the registered model-plus-clause-grounded state. Completion strata are preconfigured but hidden from the runner; H30/H60/H100 expose 6/9/12 deeply read records. All rows belong to the same eight-task development study, with no premature stop under its registered sufficiency criterion.

	Essential-slot coverage				**Model tokens**	
Task	H30	H60	H100	Selected	Routed	Forced H100
Myeloid-01	2/2	2/2	2/2	H30	9,143	31,452
Myeloid-02	2/2	2/2	2/2	H30	8,452	29,379
Stromal-01	1/2	2/2	2/2	H60	19,809	31,754
Stromal-02	1/2	2/2	2/2	H60	18,875	30,492
T-cell-01	1/2	2/2	2/2	H60	17,264	31,049
T-cell-02	1/2	1/2	2/2	H100	30,713	30,713
Metabolic-01	1/2	1/2	2/2	H100	34,375	34,375
Metabolic-02	1/2	1/2	2/2	H100	38,854	38,854
Total: 2 / 3 / 3 routes; 31.2% token saving					**177,485**	258,068

Table 24: **Frozen policies in source-disjoint actual execution.** A: equal-task means over 150 tasks per arm; recorded calls include terminal decisions, and valid-disposition recovery includes correct rejection. B: equal-group paired effects with 10,000 bootstrap draws over 106 groups, not differences between the task means in A. The two GRUs differ only in return weighting. All entries are obtained from actual policy execution.

A. Task-mean execution outcomes

Policy	Score ↑	Recovery ↑	Nonnecessary ↓	Calls ↓	Errors
Linear reference	0.4453	0.3267	0.0800	2.35	0
Supervised GRU	0.7540	0.7133	0.0500	2.23	0
GRU + AWR	0.7537	0.7133	0.0533	2.24	0

B. Equal-group paired effects [95% CI]

Contrast	Score Δ	Recovery Δ
Supervised − reference	+0.2942 [0.2129, 0.3759]	+0.3632 [0.2406, 0.4858]
AWR − reference	+0.2998 [0.2221, 0.3763]	+0.3726 [0.2547, 0.4906]
AWR − supervised	$+0.0055$ $[-0.0154, 0.0300]$	$+0.0094$ $[-0.0189, 0.0425]$

The 450 episodes use 573 Qwen3.8-27B requests and record zero execution errors. Nonnecessary-call group differences are -0.0259 $[-0.0448, -0.0094]$ for supervised-minus-reference and -0.0212 $[-0.0377, -0.0071]$ for AWR-minus-reference. Execution errors measure software reliability; agreement and recovery metrics assess scientific decision quality. The reported comparison uses greedy execution and one frozen training seed.

Scope of experience-driven learning. Evidence memory changes reusable information, graph revision changes an explanation, and offline fitting changes action selection. The controlled executions establish higher scores and recovery for the fitted recurrent policies relative to the linear reference. The return-weighting comparison is reported above. Policies and executive-model weights remain fixed throughout this separately evaluated action-time execution study.

G.8 Structural revision improves prediction under matched decision rules

Permission to revise the working explanation improves native probabilistic prediction. Across 88 source families, edit-on reduces native Brier from 0.111875 to 0.083521, a 25.3% relative reduction and a paired difference of -0.028354 $[-0.043869, -0.014555]$. The two arms share scientific inputs and downstream probability and action rules, so the comparison tests edit permission within a

Table 25: **Complete matched-study predictive and structural diagnostics.** Each arm has 264 factual records from 88 families. Paired effects are edit-on minus edit-off; intervals are pointwise family-bootstrap intervals. Selection and change rates are descriptive endpoints.

Endpoint	Edit-off	Edit-on	Paired effect [95% CI]
Native Brier ↓	0.111875	0.083521	-0.028354 $[-0.043869, -0.014555]$
Native regret ↓	0.183070	0.154086	-0.028985 $[-0.088588, 0.026831]$
Native selection (%) ↑	77.65	80.68	Descriptive
Frozen graph Brier ↓	0.242919	0.261482	$+0.018563$ $[-0.011114, 0.047079]$
Structural change (all records)	0.0000	0.7121	Descriptive
Unnecessary edit (all records)	0.0000	0.0341	Descriptive
Missed needed edit (all records)	0.8523	0.1742	Descriptive

common decision interface (Appendices C.5 and F.6). Table 1 reports the primary outcomes; Table 25 retains all corresponding predictive and structural diagnostics.

The revised mechanism is an operative downstream input. Replacing the factual child with its parent changes the recommendation or event predictions in 52.6% of eligible family-averaged comparisons [40.6, 64.1]; replacing it with a cross-family child changes outputs in 50.0% [38.0, 61.5]. These 376 substitution calls show that the working graph is consumed by the common decision procedure, rather than merely emitted as an explanatory artifact. Prediction improvement and graph dependence answer complementary questions: the former measures the benefit of edit permission; the latter tests whether downstream outputs depend on the supplied mechanism.

Revision selectivity and relation-level evidence. The edit-on arm changes 188/264 structural signatures. Edit occurrence when needed is 0.7956 [0.7022, 0.8800] over 75 eligible families; unnecessary editing is 0.2308 [0.0000, 0.4615] over 13 no-edit families. The latter is not the 0.0341 all-record rate in Table 25. Added-edge source-quote overlap is 0.6565. Its scorer removes generic words from the edge's source/relation/target tokens and returns positive when any remaining token occurs in the concatenated hidden source quotes. This score measures token-level association with source quotes; full directed-relation entailment, blinded expert review, and biological confirmation are distinct evidence levels. Recorded acceptable-operation overlap is assessed separately from edit occurrence.

Prediction and action remain distinct endpoints. The native-regret difference is -0.028985 $[-0.088588, 0.026831]$, and the frozen graph-Brier difference is $+0.018563$ $[-0.011114, 0.047079]$; both intervals include zero. Native selection rises descriptively from 77.65% to 80.68%. The demonstrated contribution is therefore improved native probabilistic prediction under matched rules, accompanied by measured downstream dependence on the revised graph. Graph-substitution sensitivity measures output dependence; output improvement and added-relation correctness require separate endpoint assessments.

G.9 Executable invariants and version-bound verification

Independent semantic reconstruction tests the structural kernel's scientific prerequisites, not only replay identity. The saved acceptance snapshot contains 91 passing tests, with zero failures, errors, or skips: 37 existing cases and 54 added regressions. The prepatch subset produces 27 failures while retaining two positive controls, demonstrating sensitivity to the repaired conditions. One test additionally checks 250 observation/confirmation combinations against the independent semantic implementation; these combinations are not counted as additional tests.

The covered interfaces include incorrect endpoints, reversed or incompatible relations, context self-expansion, untrusted admission flags, receipt rebinding, normalized ID collisions, different payloads sharing one ID, partial replacements, disconnected splits, readout-level mismatches, missing proximal/terminal observations, and invalid downstream program promotion. Valid multi-edit and supported-observation controls remain. These tests establish the executable constraints of

`structural_revision_contract_v2`; source-paper reannotation and other scientific modules retain separate validation requirements.

Replay and reconstruction answer different questions. Archived source reproduces the topology of forty lineages from ten saved Full trials. Complete original parent, prediction, and proposal inputs are unavailable in the saved records; the semantic consistency assessment uses reconstructed inputs. A separately identified reconstruction-based sensitivity analysis preserves projected diagnoses, edge changes, and child topology in all forty cases. These results establish replay and reconstruction consistency within their recorded scopes; the original trials remain bound to their executed implementation.

Scope of the recorded acceptance. The same saved verification snapshot reports 432 passes, two failures, and one collection error in the expanded C2H suite, alongside full-repository collection results and a timed-out unittest run. These outcomes characterize the saved source snapshot. The 91-test result applies to the specified diagnosis, admission, application, and downstream-eligibility interfaces; the language-model scoring path and source-literature adjudication retain separate validation requirements.

H Archival functional assessment and extended scientific traces

H.1 Cohort, supplied information, and program policies

The archival study evaluates whether a challenged explanation produces an assessable intervention program. It uses the deterministic kernel and source bindings recorded with its frozen trials; Appendix G.9 separates executable-contract verification from archival replay. Ten PDAC contexts span seven source families (Table 27); an eleventh ex-vivo organoid resource is excluded by functional-cohort schema qualification. One deterministic execution per context and policy (seed 1201) yields thirty program assessments. The Full loop composes four candidates with at most three initial edges from supplied mechanism and action libraries, selects and opens one mechanistic probe, updates the candidate set, and compiles a program. Its two information-opening stages are

$$(\mathcal{H}, \widehat{Y}_{\mathrm{mech}}) \xrightarrow{\text{open } Y_{\mathrm{mech}}} (\mathcal{H}', P, \widehat{Y}_{\mathrm{func}}) \xrightarrow{\text{freeze}} \mathrm{lookup}(P) \xrightarrow{\text{open } Y_{\mathrm{func}}} \mathrm{assess}. \tag{33}$$

Public inputs include curated baseline states, source-context descriptions, and probe/action menus; final functional ordinals are withheld until program commitment. This is an outcome-separated, proxy-assisted computational assessment.

The static bottleneck expert optimizes a single action from the same library; the checkpoint-only policy always selects effector rescue. Neither uses the opened probe or observation-conditioned multi-action synthesis. Thus, the comparison evaluates the complete mechanism-to-program procedure against static policies with restricted action spaces and different combination-search budgets. General synthesis rules are specified in Appendix C.4; the separate 88-family matched edit-permission comparison is in Appendix G.8.

H.2 Class correspondence and source-aware adjudication

Let $\phi(P)$ project a program to an unordered perturbation-class set. An evaluable match requires

$$|\{a \in \mathcal{A}_c : \phi(P) = \phi(a)\}| = 1. \tag{34}$$

Zero or multiple matches remain non-evaluable; no nearest-arm substitution is permitted. The projection merges antigenicity and antigen-presentation actions, merges trafficking and stromal remodeling, and omits maintenance. Correspondence is defined on unordered perturbation classes; drug identity, dose, chemotherapy background, and administration sequence are outside the matching fields.

The five benefit fields represent priming (DC activation, antigen/IFN recovery, or adoptive effector provision), penetration (intratumoral immune accumulation), cytotoxicity (killing or cytotoxic-effector readouts), immune-dependent control (tumor burden/survival with the registered dependence flag), and persistence (durability, relapse, rechallenge, or continued survival). Immune accumulation,

Table 26: **Program success and prediction agreement answer different questions.** Panel A retains every context in the overall denominator. Panel B descriptively re-tabulates the ten frozen Full predictions against the curated ordinal payload within the same assessment.

A. Complete three-policy assessment

Policy	Contexts	Evaluable	Success / all	Success / evaluable
Mechanova	10	10	**10/10**	**10/10**
Static bottleneck expert	10	9	0/10	0/9
Checkpoint-only policy	10	10	0/10	0/10

B. Full-program ordinal forecasts

Endpoint	Exact agreement	Disagreement
Priming	3/10	7/10
Intratumoral penetration	10/10	0/10
Cytotoxicity	10/10	0/10
Immune-dependent control	10/10	0/10
Persistence	10/10	0/10

depletion, molecular cytotoxicity, and direct-killing assays provide distinct evidence types; each field retains the assay and biological context of its supporting source.

The frozen encoding maps detrimental, null, minimal, partial, marked, and eradication to $(0.12, 0.28, 0.38, 0.52, 0.78, 0.90)$, respectively. Success requires all five benefit floors $(0.60, 0.60, 0.60, 0.60, 0.55)$, the stored immune-dependence flag, and risk ceilings: pseudo-hot 0.20, direct toxicity 0.25, normal-context toxicity 0.25, rebound 0.30, and predictive uncertainty 0.35. Defaults for the first four risk fields are $0.08, 0.12, 0.10, 0.14$; rebound ordinals can override the last, and evaluator uncertainty defaults to 0.22. These values specify the benchmark's computational encoding and default settings. Required benefit information must be available for a record to satisfy the success predicate.

Bibliographic identity and evidence scope. Table 27 identifies the source studies directly. Bibliographic identifiers and the frozen computational payload retain separate provenance. In particular, CD40 survival in spontaneous tumors and depletion/rechallenge in transplanted tumors retain different contexts (Winograd et al., 2015); FAK depletion and combination-response experiments likewise retain separate anchors (Jiang et al., 2016). The replay verifies mapping and adjudication for all thirty submissions, including the non-evaluable static case. It uses existing curated outcomes and shared mapping rules through an execution path separate from agent/BWM prediction (Appendix I.5).

H.3 Complete programs and frozen-prediction agreement

All ten Full programs match a combination and satisfy the registered predicate; neither static policy succeeds (Table 26). Each successful mapping also has a stored failing proper-subset comparator and concordant control/killing direction. These subset checks establish the archived combination-versus-subset contrast. Counts describe a finite cohort of ten contexts from seven source families; replays and endpoint fields describe those same contexts.

The frozen forecasts agree on 43 of 50 benefit fields. Both MED12 contexts and CD40 agree on all five; the remaining seven predict minimal rather than the archived marked priming. All matched Full outcomes encode the five benefits as marked. This comparison measures exact agreement with the curated ordinal encoding.

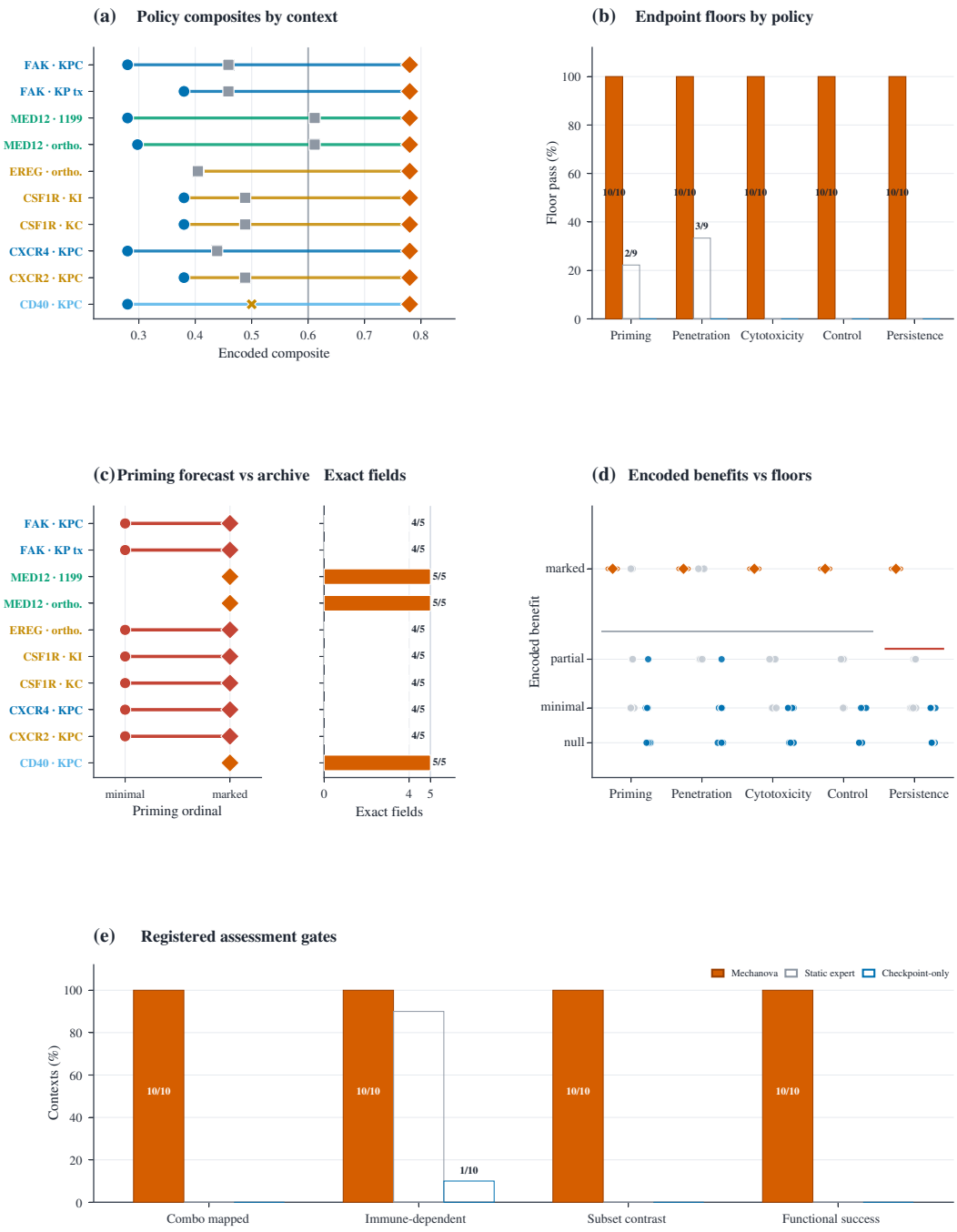

Figure 7: **Mechanism-derived programs receive archival functional support, with prediction agreement assessed separately.** (a) Encoded composites for the three policies; Mechanova is 10/10 while static expert is 0/10 (one non-evaluable) and checkpoint-only is 0/10. MED12 static exceeds 0.60 yet still fails the registered predicate. (b) Registered benefit-floor pass rates; static counts use the nine evaluable contexts. (c) Frozen priming forecasts versus archival encodings, and per-context exact-field counts (43/50); the seven disagreements are priming (minimal vs marked). (d) Encoded benefit values against the declared floors (0.60, persistence 0.55). (e) Combo mapping, immune dependence, stored failing proper-subset comparators (21/21), and functional success remain separate gates. Results cover ten contexts from seven source families and assess computational perturbation-class correspondence against archival outcomes after program commitment.

Table 27: **Every Full program, its archival correspondence, and forecast agreement.** The intervention column retains distinct actions before projection. Archive arm labels summarize stored correspondence at the intervention-class level. References identify the originating source families.

Context	Frozen intervention actions	Matched archive arm / source	Exact
FAK–KPC	Stromal remodeling + effector rescue	FAK inhibition + PD-1 (Jiang et al., 2016)	4/5
FAK–KP transplant	Stromal remodeling + effector rescue	FAK inhibition + gemcitabine + PD-1 (Jiang et al., 2016)	4/5
MED12–KPC1199	Antigenicity + antigen presentation + effector rescue	MED12 loss + PD-1 (Tang et al., 2024)	5/5
MED12–orthotopic	Antigenicity + antigen presentation + effector rescue	MED12 loss + PD-1 (Tang et al., 2024)	5/5
EREG–orthotopic	Myeloid reprogramming + effector rescue + maintenance	EREG targeting + CAR-T (McClellan et al., 2026)	4/5
CSF1R–KI	Myeloid reprogramming + effector rescue	CSF1R blockade + checkpoint (Zhu et al., 2014)	4/5
CSF1R–orthotopic KC	Myeloid reprogramming + effector rescue	CSF1R blockade + checkpoint (Zhu et al., 2014)	4/5
CXCR4–KPC	Trafficking restoration + effector rescue	AMD3100 + PD-L1 (Feig et al., 2013)	4/5
CXCR2–KPC	Myeloid reprogramming + effector rescue	CXCR2 inhibition + PD-1 (Steele et al., 2016)	4/5
CD40–KPC	DC priming + effector rescue	CD40/chemotherapy + checkpoint (Winograd et al., 2015)	5/5

CD40 illustrates why mapping is program-specific: Full projects to priming plus effector rescue and matches the combination, whereas the static expert's antigen-presentation-only action has no exact arm. That static record remains non-evaluable because an exact archive arm is absent. Across Full, nine selected graphs retain their parent structure; EREG adds a library edge.

Figure 7 brings together complete-cohort outcomes, endpoint-level assessment, and frozen-prediction agreement; the following cases expose the recorded transformations behind individual programs.

H.4 FAK–KPC: OBSERVATION-DRIVEN RIVAL ELIMINATION

The selected checkpoint-rescue probe opens the signed readouts $(0.86, 0.10, 0.05, 0.04, 0.02)$ for engagement, immediate effector response, killing, control, and persistence. These are benchmark encodings. Engagement without downstream rescue retires two of four rivals (Table 28, A). The selected survivor retains its CAF-exclusion and priming-deficit edges and the same structural identity. The compiler addresses the operative stromal barrier and adds effector rescue, freezing the two-action program before lookup. Its exact projected match is the archived FAK-plus-PD-1 class combination; four forecasts agree, while priming does not. The episode demonstrates candidate elimination and program composition on the retained graph.

H.5 EREG: A STRUCTURAL ADDITION WITH SEPARATELY TRACKED ACTION CONSEQUENCES

The EREG probe opens $(0.90, 0.72, 0.22, 0.10, -0.20)$ on the same five mechanistic readouts. Proximal agreement but distal failure yields a secondary-lock diagnosis: immediate support is 0.85 versus functional support 0.41. The selected parent contains priming-deficit, effector-killing, and myeloid-suppression edges; its child adds `edge-metabolic-lock`, changing structural identity without retiring a rival. The added metabolic-lock edge is a library-level modeling proposal. The cited study supports EREG-associated myeloid remodeling and CAR-T sensitization (McClellan et al., 2026); the metabolic interpretation remains a hypothesis requiring direct evidence.

Four regenerated contrast objects expose the action consequence (Table 28, B). Axis-level compression leaves no standalone residual action, so first-plus-residual equals first-only. Full adds observation-supported effector rescue and rule-triggered maintenance. After maintenance is omitted, its two classes match EREG-plus-CAR-T. Graph-conditioned numerical readouts, action-composition

Table 28: **Executed graph, prediction, and program transformations.** Panel A gives readable local labels for the four frozen FAK candidates. Panel B reports EREG's stored program-level forecasts; the empty residual and duplicate action sets are retained. Only the final program receives the reported archival assessment.

A. FAK candidate set after the checkpoint probe

Candidate	Parent edge inventory	Recorded disposition
h_1	CAF exclusion; priming deficit; effector killing	Retired
h_2	CAF exclusion; priming deficit; PD-1 dysfunction	Retired
h_3	Antigen silence; CAF exclusion; priming deficit	Retained, weakened
h_4	CAF exclusion; priming deficit	Retained, selected

B. EREG contrasts after structural revision

Contrast	Actual action set	Control	Persistence
First lock alone	Myeloid reprogramming	minimal	null
Residual lock alone	$\varnothing$	null	null
First + residual	Myeloid reprogramming	minimal	null
Full program	Myeloid reprogramming; effector rescue; maintenance	marked	marked

ordinal forecasts, and archival observations are tracked as distinct objects. The trace exposes an inspectable graph–prediction–program transformation, and the functional comparison scores the final projected program as a whole.

I Reproducibility and implementation map

I.1 Result identity and reproduction materials

A reproducible result binds its input and source scope, provider, configuration, permitted information, trajectory, scorer, and aggregation. The current reporting entry point, `reproduce_v4.py`, uses the hash-bound `result_index_v4.json`: a dated leaderboard snapshot, seven complete prediction files, matched-edit records, actual policy executions, and the component results used in this paper. Each index entry gives its path scope, SHA-256 digest, and relevant JSON pointers. Historical 111-task native scores and completed-sequence OPE are catalogued separately and excluded from the current export. Table 29 maps the reported studies to these computation records. A current source filename does not replace the executed version and configuration recorded in an experiment manifest.

The actual policy-execution study is bound under `runs/frontier_closure_20260822/e4_causal_policy_execution_v1_20260917/`: `training_report.json`, `execution_scorecard.json`, `execution_analysis.json`, `execution_records.jsonl`, and `policies/`. These artifacts identify the fitted policies, action-time inputs, executed decisions, and their task/group-level scoring.

Initial-decision result binding. The 125-task comparison is bound under `runs/frontier_closure_20260822/top_tier_experiments_20260827/c2h_mechanova_forma920_flagship_v1/forma920_native_comparison/`. The companion snapshot `reporting/leaderboard_snapshot_v4.json` is a byte-exact copy of the runtime leaderboard dated 23 September 2026, 17:24:13 UTC, with SHA-256 `8f07b4a15455eefe25fdc097c03d540112662694e0ad555a8c75c069c1b5b31b`. Reporting reads this fixed copy rather than the mutable `leaderboard_v1.json`; `predictions/mechanova_v20_full_v1.json` has SHA-256 `674207f9678f6dcefc89fb49de7cb6a34ef0c3bda26001ac660225e4a82da53a`. Per-configuration prediction files contain the complete task/seed records. `cohort_public_v1.json`, private evaluator records, visible bundles, hypothesis packages, trials, and projection receipts retain their distinct access roles.

The task release is `Expert_review_9.20/final_release_2026-09-22` within the curator-only annotation root; its `FORM_A_RELEASE_RECEIPT_20260922.json` identifies 278 unique released papers and 125 included tasks. The evaluated policy implementation is located in `c2h_v20_title_token_v1.py`, with inherited task-conditioning and window-filter modules. Request, cache, and trial bindings separately record the original hypothesis-generation calls and subsequent decision-policy execution.

Frozen title-only control binding. Table 11 uses the frozen `title_v1` implementation of `infer_bottleneck_from_context` and its `BOTTLENECK_TO_EXPERIMENT` mapping. Reconstruction must preserve this baseline identity rather than substitute `strict_title_rule`, match task and seed identifiers, and score the committed pre-observation choices. The development and 125-task inventories remain distinct; repeated seeds are nested within tasks.

Matched graph-edit result binding. The two-stage study is stored under `ccfa-workfiles/experiments/c005-graph-edit-attribution/`. Its `C005_RESULT.json` has SHA-256 `f421018d5f0821b77a9a104f9cb8616079d25679845e86c4c4b015286c78c9a2`; `PROTOCOL.md` has SHA-256 `9f1b4bbb53a84c5a14a363ef36c92e161a80c1f62b7a23e318db5de1df972b30`. The shared stage-two system-prompt digest is `91437b201e18b1d983e325d8c4acf106736770d4583f4fdada1e6c7001b2d987`. The study retains 528 factual two-stage submissions, 188 parent substitutions, and 188 cross-family substitutions. The scoring implementation records the lexical source-quote diagnostic separately from native prediction metrics.

Reporting and statistical reconstruction. The companion checks all 30 current source files and their declared JSON fields. It recomputes complete-cohort hit counts, task means, and twelve paired external-comparator intervals, using 10,000 draws with seed 20260923; marginal intervals retain the bound scoreboard's seed 20260922. For the 450 policy executions, it checks action/terminal agreement, recorded-call and score arithmetic, task means, and equal-group paired point effects. Policy and matched-edit intervals are retained from their registered scorecards; matched Brier arithmetic is checked separately. The output `result_values_v4.json` records these scopes explicitly. The title-only table additionally requires its frozen rule implementation, cohort-specific references, and committed choices; it is not regenerated by this reporting command.

Decision providers and outcome access. The native comparison scores bounded external agents through a shared output projector and Mechanova through its frozen initial distribution and pre-observation experiment order (Appendix F.3). Evidence recovery combines Qwen3.8-27B generation with frozen public-span repair, assessed separately for target recovery and source support (Appendix F.4). The archival loop uses supplied edge/action libraries, deterministic diagnosis, structural revision, and program compilation; final functional lookup follows program commitment (Appendix H).

The protected-target study uses fitted Gaussian response beliefs and finite lookahead, whereas perturbation-representation comparisons bind their own checkpoints and held-out responses (Appendix D). Operational studies separately bind a transition predictor for forecast fidelity, forecast-guided policies for hidden-dynamics decisions, and a frozen evidence-prefix policy with a fixed endpoint predictor for acquisition efficiency. Execution studies compare observation-conditioned control, disclosure, or skill completion under their own matched profiles. Policy execution compares retrained and frozen linear, supervised-GRU, and AWR-GRU policies using only action-time information. The logged-sequence diagnostic uses its own estimator; the executive model remains fixed (Appendix E).

Field-level statistical identity. The `complete_configurations` entries of `result_values_v4.json` retain each configuration's hit count, task-mean accuracy, marginal interval, and analysis seed. The `paired_comparisons` entries retain the mean difference, paired interval, and 125-task denominator. Evidence-family sensitivity intervals remain separate from task-level intervals. For policy execution, task means use `arm_summaries.<arm>`; equal-group effects use `comparisons.<contrast>_<metric>` in `execution_scorecard.json`. The execution-record SHA-256 is `e0a77c4434cdf55746540a44`

Table 29: **Current results and the separately retained logged-sequence analysis.** Current keys resolve in `result_index_v4.json`. Reporting verifies identities and specified arithmetic; it does not rerun models, retrain policies, or replace each study's scorer.

Scientific computation	Index or entry point	Bound result or reconstruction object
Initial decisions	`native_snapshot`, `prediction_*`	Fixed leaderboard copy and seven complete prediction files; 125-task pairing.
Frozen title-only control	§F.3	Fixed rule and experiment mapping; paired task/seed choices.
Matched edit permission	`matched_edit`, `matched_protocol`	Shared prompts; 528 factual records; parent/swap interventions.
Evidence licensing	`licensing`	Matched compiler–reference scores; packet and parent-output bindings.
Selective memory / MEC	`memory`, `mec`	Registered rows selected by metric and comparator.
Hierarchical execution	`how`	Separate development/confirmation summaries and source bindings.
Environment / skills	`environment_skills`	Matched visibility and skill-completion contrasts.
Adaptive horizons	`horizon`	Eight-task routing, token totals, stopping checks, terminal chain.
Operational forecasting	`owm_transition`	Transition fit and separately frozen hidden-dynamics decisions.
Evidence acquisition	`owm_26q1`, `owm_24q2`	Release-specific Brier utility, acquisition counts, and policy bindings.
Active screen selection	`active_screen`	Primary-screen score, calibration-disjoint plan, scoring script.
Biological response	`bwm_protected`, `representation`	Protected-pair source-macro scores; separate representation comparisons.
Policy execution	`policy_execution`, `policy_records`	Three-arm scorecard and 450 records; fitting and analysis companions.
Historical diagnostic	Separate historical index	Completed-sequence estimator and support checks; Appendix E.5.
Archival programs	`archival`	Three-arm class-matched assessment; seed 1201; thirty trials.

`fb35fbe54ad9991d324f7ad17164e8482fe40ea4`, identifying the 450 recorded executions.

I.2 Implementation and parameter bindings

General experiment-field compilation and causal grammar reside in `literature_intelligence/`; the evaluated span compiler and its verifier are `e1_bpsmg_v35_2_f7_d_v12.py` and `e1_bpsmg_v35_2_f7_c_verifier.py`. Memory separates pointer selection, treatment, and campaign binding in `cold2hot_governed_memory_v1.py`, `governed_memory_treatment_v1.py`, and `governed_memory_runtime.py`. Within `scientific_planning/`, `campaign_react_v2.py` is the frozen controlled HOW compiler, whereas `campaign_react_system_v2.py` provides live-system integration. The current index binds reported results and evaluation configurations; the module locators here identify their implementation surfaces. Executed source versions remain with the study manifests. Module paths are relative to `src/se_oncoagent/` unless prefixed by `scripts/` or `configs/`.

Parameter provenance distinguishes fitted weights, development-selected recipes, declared algorithmic rules, curated encodings, and defaults. The action-time policy recipe is `configs/training/e4_causal_policy_execution_v1.yaml`; feature construction, fitting, execution, and scoring are implemented in `src/se_oncoagent/experiments/e4_causal_policy_execution_v1.py`. The separately bound run preserves fitted policies and executed settings. The completed-sequence analysis retains its own `configs/training/evo_conductor_frontier_policy_v7.yaml` recipe and `runs/training/adaptive_`

wave4_offline_ranker_repro_v1_20260822/ artifacts. Public solver packets, evaluator targets, opened probe observations, archival outcomes, and trajectory supervision have distinct access roles. Curated ordinals and defaults are not empirical response or safety estimates; model-assisted labels retain their annotation provenance.

Replogle representation bindings. The representation_protocol entry binds configs/benchmarks/bwm_replogle_fewshot_replication_v6_3.json. It specifies arcinstitute/ST-HVG-Replogle at revision bb6a9562cbbf1fd152df14cc53b4cc7517c77175 using X_hvg, and arcinstitute/ST-SE-Replogle at revision e324967ff4cea5ec199e29bcbb5c1f00e5b9d69c using X_state. The provider's eval_best.ckpt selection is fixed for HepG2, Jurkat, K562, and RPE1 under the supplied few-shot split. Effects pair by perturbation identity; Pearson and DE-direction effects are ST-HVG minus ST-SE, and MSE reduction reverses the subtraction. Registered intervals use 10,000 paired draws with seed 2026082943. The comparison evaluates the separately supplied checkpoint configurations, each with its recorded training choices.

Structural-kernel implementation. The revised diagnosis, admission, and application interfaces are implemented in c2h_counterfactual_revision_v1.py, c2h_structural_revision_contract_v1.py, and c2h_structural_mechanism_revision_v1.py, with c2h_revision_binding_v2.py and the independently implemented c2h_revision_semantic_verifier_v2.py. Public function suffixes do not override the explicit contract version. The acceptance ledger and reconstruction limits are stored under ccfa-workfiles/c002_c004_hardening_v2/; this binding is distinct from the matched language-model scorer and archival execution source.

Archival state encoding. The baseline ecosystem is supplied by the public case packet. The archival adapter maps signed stromal, infiltration, myeloid, cytotoxicity, control, priming, antigenicity, and persistence readouts to their corresponding state fields as specified in Appendix A.3. Two coupled rules supplement those direct updates: an antigenicity readout y also raises presentation to $\mathrm{clip}_{[0,1]} \max\{x_{\mathrm{presentation}}, 0.5 + 0.45y\}$; control below 0.45 with functionality above 0.55 raises rebound pressure to at least 0.62. Other fields are retained. These declared rules are implemented in c2h_fvc_schema_reserve_v1.py and c2h_fvc_structural_revision_v1.py, separately from fitted response predictions and final archival adjudication.

I.3 Execution profiles and model accounting

Generative language-model calls use Qwen3.8-27B. Table 30 summarizes the study-specific decoding and execution settings.

Native projection is a model call: it retains at most twelve hypotheses, truncates hypothesis text to 4,000 characters and the native-text prefix to 6,000, normalizes priors, and deduplicates experiment IDs. Reproduction preserves its prompt/schema, tie-breaking, fallback text handling, and completed-response cache. Generation, projection, source-support checking, and target matching remain separately identifiable stages.

Qwen3.8-27B serving aliases identify recorded endpoints; exact weight identification is a separate regeneration requirement. Regeneration additionally requires the bound weights, tokenizer, chat template, inference environment, and request configuration. The offline-policy training receipt specifies an A800 80-GB device for that run. Provider tokens, approximate tokens, evidence panels, and wall time retain their measured scopes; cached replay and fresh generation have distinct cost scopes. Foundation-model weights stay fixed. The three-arm policy study is a separately bound controlled execution.

I.4 Minimal reproduction sequence

Use Python 3.11 or later and the reporting dependency file from the manuscript directory:

```
python3 -m venv .venv-report
.venv-report/bin/python -m pip install \
  -r requirements-reporting-v4.txt
```

Table 30: **Study-specific execution settings.** Profiles are not propagated between experiments; complete prompts, configurations, and provider bindings remain with their source/run records.

Profile	Bound setting
Native-output projection	Qwen3.8-27B; temperature 0; thinking disabled; 8,192-token default response ceiling; seeds 1203/1211; fixed prompt/schema.
Evidence licensing	Qwen3.8-27B parent generation; seeds 501/509/521; twelve-record packets; at most ten cards; parent-generation cost retained; zero extra editor calls in the reported runs.
Controlled HOW	32,768-token context; 2,048-token response ceiling; twelve turns/eight tool calls; seeds 13/31/53.
Environment / skills	Separate six-turn/eight-tool profile; seeds 11/29/47; verified useful completion.
Recurrent-policy fitting	Two width-32 GRUs; 29 rebuilt inputs/four actions; 100 epochs; learning rate 0.003; weight decay 10^{-4}; supervised weights 1 versus AWR temperature 1/cap 12; KL weight 0.35; clip norm 1; seed 20260711.
Policy execution	Greedy legal actions; 150 tasks per arm; 32 workers; fixed Qwen3.8-27B serving alias on GPU4; query temperature 0, thinking disabled, response ceiling 128 tokens.
Matched edit permission	Qwen3.8-27B on GPU2; temperature 0, thinking disabled; 32 workers; seeds 1701/1707/1713; two stages per factual trial plus 376 graph-substitution stage-two calls.
Initial-decision policy	Recorded GPU0 hypothesis packages reused; rule-assisted decision execution, seeds 1203/1211. Cache replay and new model generation are counted separately.

```
.venv-report/bin/python reproduce_v4.py --self-test
.venv-report/bin/python reproduce_v4.py --inventory
.venv-report/bin/python reproduce_v4.py \
  --root /path/to/project --write result_values_v4.json
```

The dependency pin is NumPy 2.3.5; self-test and inventory modes use only the Python standard library. The checks reject changed source bytes, missing fields, duplicate task/seed pairs, incomplete configurations, invalid hit values, and writes targeting source records. Project-scoped entries resolve under `-root`; the immutable leaderboard copy resolves beside the companion. The current export excludes historical native and OPE outputs by construction.

The original `reproduce_v2.py` and `result_index_v2.json` remain unchanged for historical reproduction. Their native entry is the 111-task study and their policy entry is completed-sequence OPE; neither replaces the current 125-task comparison or actual policy execution. Current component entries preserve the same original score-file bindings.

Figure 6 is rebuilt without model calls using `python Figure/build_figure6_v4.py -root /path/to/project` in a PyMuPDF environment. The builder checks both source-figure and policy-scorecard hashes, preserves panels (a–d) as vector graphics, and constructs panel (e) from task means and registered group-paired effects. It writes a separate `Figure/figure6_v4.pdf` and plotted-value record, leaving the historical figure untouched.

Reporting reconstruction checks saved records; deterministic replay additionally restores each study's packets, caches, manifests, and scorer/verifier. Re-execution or refitting restores the original model, tokenizer, data partitions, and configuration in a new run directory. Full re-execution requires experimental dependencies obtained under their original access terms in addition to the manuscript companions.

I.5 Verification scope and scientific lineage

Verification follows the object being claimed. Evidence replay checks source support and budgets; graph verification reconstructs observation coverage, admitted operations, and the child; operational replay checks verified transitions. Policy-execution verification binds action-time features, frozen policies, source isolation, recorded actions, and the task/group aggregation distinction. Functional

replay reconstructs thirty class mappings, including the non-evaluable static program, under the recorded adjudication semantics, using unordered class membership as the matching criterion. Exact artifact replay, new stochastic generation, independent curation, and biological replication remain different operations. Every reconstruction retains its source identifiers and scoring definitions.