BIOCAUSALX: AN ENVIRONMENT FOR CAUSAL DISCOVERY IN BIOLOGICAL SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) agents combine extensive prior knowledge with limited experimental evidence when investigating biological systems. Evaluating and training these agents requires repeatable cycles of intervention, observation, and hypothesis revision. Biological experiments are often slow and costly; simulated environments offer repeatable feedback but also require scientific descriptions and interventions tied to documented mechanisms. We introduce BIOCAUSALX, an environment for causal discovery across 43 executable biological systems and 168 source-supported intervention axes. Variable descriptions and permitted operations are reconciled with publications and database records. Each operation maps to an executable mechanism hidden from the agent, with supporting evidence retained for audit. Across evaluated LLMs, biological context supports recovery, while alignment with the underlying system shapes the benefits of experimentation. Within tested budgets, feedback partly compensates for missing descriptions but does not reliably overcome misleading ones. Larger experimental allowances can remain underused without improving recovery, identifying limitations in both resource allocation and evidence integration. GRPO training improves recovery on held-out systems with and without biological context. Together, these results demonstrate that BIOCAUSALX supports reproducible evaluation and training of scientific agents that combine biological knowledge with experimental feedback.

1 INTRODUCTION

Large language model (LLM) agents are increasingly being developed for scientific discovery in biology (Huang et al., 2026; Gottweis et al., 2026). Recent systems connect reasoning capabilities with experimental feedback, allowing agents to propose biological perturbations, interpret their outcomes, and refine subsequent decisions (Roohani et al., 2025). Such closed-loop interaction raises the prospect of agents that can investigate mechanisms through sequential experimentation.

Evaluating and training such agents remains challenging because biological experiments are often slow, costly, and technically demanding, limiting repeated cycles of hypothesis testing at scale. Interactive environments provide a practical alternative by supporting sequential experimentation and mechanism discovery through simulated feedback (Duan et al., 2025; Kabra et al., 2026; Yang et al., 2026). However, a gap remains between simulated discovery tasks and real biological research, where experiments are informed by extensive knowledge about biological entities and constrained by the interventions that can actually be performed. Bringing these environments closer to this setting requires incorporating biological knowledge into both the scientific context presented to agents and the experimental operations available to them. This motivates literature-grounded environments in which variable descriptions and permitted interventions are supported by documented scientific evidence and connected to executable biological mechanisms.

Literature and databases provide knowledge of biological entities and their relationships (Malik-Sheriff et al., 2020), which agents can use to formulate hypotheses and guide experimentation (Toth et al., 2022; Li et al., 2025). Studying this process requires scientific context, interpretable interventions, executable feedback, and a fixed structural reference aligned within the same system. This alignment supports systematic evaluation of hypothesis formation, experiment selection, and causal recovery, while providing repeatable feedback and measurable outcomes for agent training.

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.

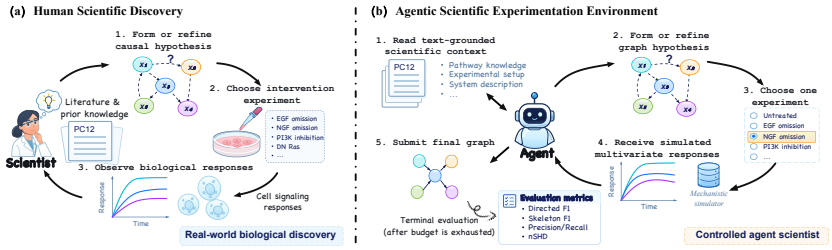

Figure 1: **From human scientific discovery to an agent experimentation environment.** (a) Scientists combine literature and prior knowledge with iterative hypothesis formation, intervention, and observation. (b) BIOCAUSALX provides an environment: agents read scientific context, select supported experiments, observe simulated trajectories, and revise hypotheses under a limited budget.

We introduce BIOCAUSALX, an interactive environment for causal discovery in biological systems, built from 43 published biological models. Each system includes descriptions of biological variables and permitted experimental operations supported by scientific publications and database records. These operations are mapped to mechanisms within the executable model. Agents select experiments under a limited budget, observe simulated trajectories, and use this feedback to guide subsequent experiments (Figure 1). The submitted causal graph is evaluated against a fixed, hidden reference derived from the model's governing equations, enabling repeatable evaluation of causal recovery within each model under fixed mechanisms, permitted operations, and scoring rules.

Our experiments show that LLM agents with biological context recover more causal structure than the evaluated numerical baselines under limited evidence, while gains from experimental interaction remain uneven. Context alignment is particularly consequential: within the tested budgets, experimental feedback partly compensates for missing biological descriptions but does not reliably overcome misleading ones. Further analyses show that larger experimental budgets can remain underused without improving recovery, and that agents recover less complete causal structures on larger biological systems under a fixed budget. We also demonstrate that GRPO training in BIOCAUSALX improves causal recovery on held-out systems both with and without biological context. In these matched runs, the policy trained and evaluated with context achieves more than twice the final Directed F1 of the no-context policy. Together, these findings show that BIOCAUSALX can effectively support the evaluation and development of scientific agents. Our code is available at `https://anonymous.4open.science/r/BioCausalX-submit-467B/`.

Our contributions are threefold:

- **A biological discovery environment.** We propose BIOCAUSALX, which has 43 executable systems with source-supported variable descriptions and 168 biologically interpretable intervention axes, retaining evidence and execution mappings for auditability.
- **A precise interactive discovery protocol.** Agents acquire trajectories under explicit experiment and observation budgets and are evaluated against fixed direct-dependency graphs extracted from symbolic model equations.
- **A controlled study of semantics, experimentation, and agent learning.** We evaluate classical methods and LLM agents, isolate the effect of semantic alignment, compare active and random experiment selection, assess numerical tools and output-validation harnesses, and demonstrate cross-system policy improvement through agentic reinforcement learning.

2 RELATED WORK

2.1 INTERACTIVE SCIENTIFIC DISCOVERY

Scientific-agent benchmarks study experimental and explanatory discovery (Wang et al., 2022; Jansen et al., 2024), research on supplied datasets (Majumder et al., 2025; Chen et al., 2025b), and research workflows (Bragg et al., 2026). BoxingGym, MaD Physics, and DiscoverPhysics study experiment selection and budgeted acquisition (Gandhi et al., 2025; Jain et al., 2026; Wiemann et al., 2026).

PhysGym systematically varies the metadata disclosed to agents (Chen et al., 2025a), whereas NewtonBench changes governing laws so that otherwise correct textbook priors become locally inconsistent and require experimental revision (Zheng et al., 2026). Prior–environment alignment is especially consequential in biology: entity names can be ambiguous (Gyori et al., 2022), and model annotations incomplete, misleading, or incorrect (Shin et al., 2023). These issues motivate BIOCAUSALX's distinct controls for missing and misleading context during iterative experimentation.

Among mechanistic environments, SciGym reconstructs BioModels reaction and kinetic models using concentration interventions and context controls (Duan et al., 2025); CausaLab studies prediction and mechanism recovery in synthetic structural causal models with restricted interventions (Yang et al., 2026); ActiveSciBench supports active regulatory-graph recovery in network motifs (Kabra et al., 2026); and Model Discovery Agent combines LLM hypotheses, Bayesian inference, and experimental design for neuronal dynamics (Murphy, 2026). BIOCAUSALX links source-supported descriptions and operations to explicit execution mappings across heterogeneous published systems, enabling controlled tests of prior–environment alignment and its effects on discovery. Differences in recovery targets preclude direct comparison of scores across these environments.

2.2 ACTIVE CAUSAL AND DYNAMICAL STRUCTURE LEARNING

Active causal discovery reduces structural or query uncertainty (Hauser & Bühlmann, 2014; Toth et al., 2022; Tigas et al., 2022; Olko et al., 2023); LeGIT incorporates LLM knowledge and intervention samples (Li et al., 2025). LLMs supply semantic causal judgments (Kıcıman et al., 2024); Causal Modelling Agents integrate metadata and statistics, COAT refines factors through causal feedback, and CausalOrder derives ordering constraints (Abdulaal et al., 2024; Liu et al., 2024; Vashishtha et al., 2025). CauScientist scores graphs on fixed observational/interventional data, RealTCD uses existing interventional time series (Peng et al., 2026; Li et al., 2024), and IGDA queries binary edges (Havrilla et al., 2025). BIOCAUSALX instead acquires trajectories through permitted biological operations. Their execution mappings specify how an action perturbs a represented mechanism, rather than assuming access to arbitrary graph interventions at unrestricted targets or strengths.

SINDy identifies equations, neural graphical modelling recovers vector-field dependencies, and DYNOTEARS/PCMCI+ estimate contemporaneous and lagged relations (Brunton et al., 2016; Bellot et al., 2022; Pamfil et al., 2020; Runge, 2020). CausalTime and CausalDynamics provide controlled trajectories for structural evaluation (Cheng et al., 2024; Herdeanu et al., 2025); Curious Causality-Seeking Agents learn state-dependent models for exploration, and CausalGame evaluates experimental decisions and causal reasoning (Zhao et al., 2025; Chen et al., 2026). BIOCAUSALX enables adaptive acquisition toward a fixed graph of direct dynamical dependencies, distinct from reaction incidence, total effects, or lagged associations; identifiability guarantees for acyclic graphs need not hold for the biological feedback cycles represented in these models.

2.3 BIOLOGICAL DISCOVERY AND GROUNDING

GeneNetWeaver, DREAM, SERGIO, and BEELINE provide simulated or curated regulatory-inference tasks (Schaffter et al., 2011; Marbach et al., 2012; Dibaeinia & Sinha, 2020; Pratapa et al., 2020); CausalBench uses real single-cell perturbations with incomplete reference networks (Chevalley et al., 2025). BioDiscoveryAgent selects perturbations for phenotype hits (Roohani et al., 2025), PerturbQA aligns knowledge-graph and experimental entities for response prediction and explanation (Wu et al., 2025), and PertReason evaluates cell-state-conditioned mechanistic paths (Kim et al., 2026). BIOCAUSALX instead tests budgeted recovery of a whole model's direct dependencies from context and acquired trajectories; plausibility alone cannot establish individual causal edges.

BioModels, JWS Online, and PEtab support publication-linked models, reproducible simulation, and parameter-estimation specifications, respectively (Malik-Sheriff et al., 2020; Peters et al., 2017; Schmiester et al., 2021). BIOCAUSALX connects 43 executable systems and 168 intervention axes through source-supported descriptions, permitted operations, execution mappings, trajectory feedback, and fixed reference graphs. This supports reproducible structural scoring within a specified model, whose mechanisms and variables delimit its biological interpretation.

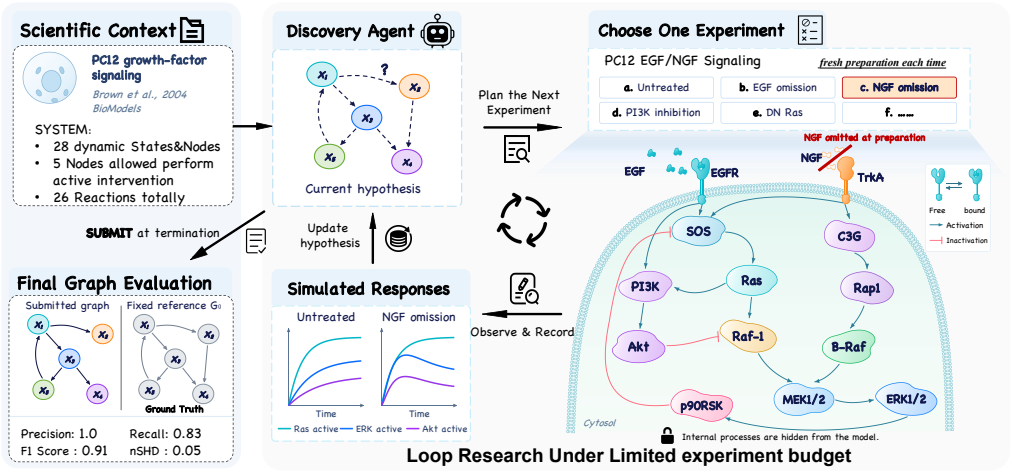

Figure 2: **The BIOCAUSALX interaction and evaluation workflow.** The PC12 EGF/NGF model (Brown et al., 2004) (28 states, 26 reactions, and five intervention axes) illustrates budgeted experimentation. Agents use scientific context and simulated responses to revise graph hypotheses. The mechanism and reference G_0 are hidden; scoring occurs only at termination. The protein-level network, response curves, five-node graphs, and scores are schematic.

3 BIOCAUSALX ENVIRONMENT DESIGN

3.1 TASK FORMULATION

Given a biological system with state variables $\mathbf{x} \in \mathbb{R}^n$ and an unperturbed model M_0, the agent receives scientific descriptions $\mathcal{C}$ and permitted interventions $\mathcal{A}$. It selects experiments under a limited budget to recover the directed dependency graph G_0, defined from the governing equations:

$$\dot{\mathbf{x}} = \mathbf{f}_0(\mathbf{x}, \tau; \boldsymbol{\theta}), \qquad A^\star_{ij} = \mathbf{1}\left[\frac{\partial f_{0,i}}{\partial x_j} \not\equiv 0\right]. \tag{1}$$

Thus, $A^\star_{ij} = 1$ denotes $j \to i$: variable j directly affects the rate of change of variable i. Dependencies are tested symbolically after combining all terms in each equation. The graph is directed and unsigned and may contain feedback cycles; scoring excludes self-edges. Interventions provide evidence about this fixed reference without changing the target, requiring agents to distinguish direct dependencies from propagated effects. The target describes the published model, with biological interpretation bounded by its represented mechanisms and variables.

Table 1: Environment scope. Intervention axes are deduplicated at the represented-mechanism level; edge counts exclude self-dependencies.

Systems	43	State variables	950
Intervention axes	168	Directed edges	5,073
Variables/system	5–94	Axes/system	2–13

3.2 LITERATURE-GROUNDED SYSTEMS AND INTERVENTIONS

We curate 43 systems from BioModels (Malik-Sheriff et al., 2020), JWS Online (Peters et al., 2017), KiMoSys (Costa et al., 2014), and PEtab model collections (Schmiester et al., 2021), spanning signaling, metabolism, biological oscillators, and physiological regulation, with 5–94 state variables per system (Table 1). Each is converted to a common executable representation. We reconcile variable annotations with source publications and database records to identify biological entities and model-specific states, retaining both source passages and evidence locations for audit.

Operations are admitted after checks of source evidence, target identity, biological context, and compatibility with the model's mechanisms; unresolved candidates are excluded.

Each accepted operation is mapped to existing inputs, states, or mechanism terms. For example, changing enzyme activity modifies its reaction terms rather than directly setting downstream states. Supporting passages and private execution mappings are retained, with benchmark approximations in numerical strength recorded separately from experimental support.

The resulting catalog provides 168 *intervention axes*, each corresponding to a distinct represented mechanism; multiple doses or publications do not create additional axes. Each system has at least two axes, validated for finite trajectories and nontrivial responses. Agents see biological identities, non-directional scientific context, and permitted conditions, while equations, exact bindings, and the reference graph remain hidden. Models, graphs, actions, and audit records are frozen together before evaluation. These checks establish supported, executable choices but do not guarantee that every edge is identifiable. Figure 3 summarizes coverage; Appendix A details construction and validation procedures.

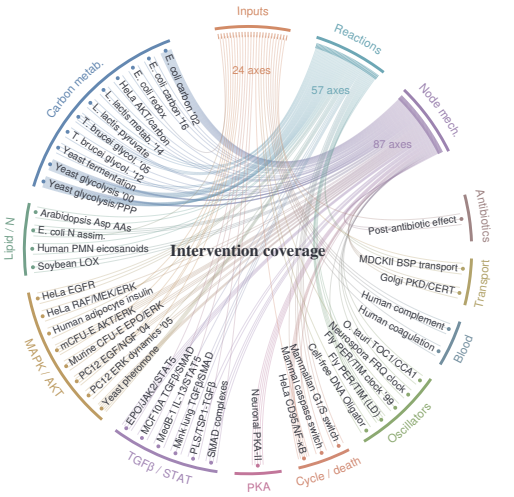

Figure 3: **Coverage of BIOCAUSALX.** The 43 systems are grouped by biological function and linked to 168 intervention axes: node mechanisms (87), reactions and interactions (57), and exogenous inputs (24). Ribbon widths encode axis counts for each system and intervention family.

3.3 INTERACTION PROTOCOL

An episode alternates between experiment selection and hypothesis revision (Figure 2). At each round, the agent chooses the untreated condition or the single frozen condition of a permitted axis. Each experiment starts from a fresh preparation and returns m_t observations over a prescribed horizon:

$$Y_t = \left(\left(\tau_k, \mathbf{x}^{(a_t)}(\tau_k) + \boldsymbol{\epsilon}_{t,k} \right) \right)_{k=1}^{m_t}, \qquad T \le B_{\text{exp}}, \qquad \sum_{t=1}^{T} m_t \le B_{\text{obs}}. \tag{2}$$

Here a_t is the selected condition, τ_k is local experiment time, and $\boldsymbol{\epsilon}_{t,k}$ denotes measurement noise. The budgets limit experiments and total generated observation time points, respectively. All integrated states are observable in both untreated and intervention trajectories.

The standard LLM interface presents sampled values and per-variable summaries together with the remaining budget. The agent finally submits a directed graph, scored only at termination without reference feedback during discovery. Section 4 specifies the evaluated interfaces and budgets; Appendix A.3 details the execution and logging protocol for reproducible evaluation.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Observational and interventional discovery. We compare classical methods and LLM agents on all 43 systems with at most three simulations of 101 time points each. Observational runs use untreated replicates; interactive runs begin with an untreated baseline and permit two sequentially selected experiments. Classical methods receive full trajectories; LLMs receive biological descriptions, permitted operations, and sampled trajectories with summary statistics. Simulation ceilings are matched, while side information and input representations differ between the method families.

Classical methods and acquisition policies. We evaluate eight classical causal-discovery methods: Linear SINDy–STLSQ (Brunton et al., 2016), Sparse VAR–BIC (Granger, 1969), PCMCI+ (Runge, 2020), VAR–LiNGAM (Hyvärinen et al., 2010), DYNOTEARS (Pamfil et al., 2020), TSCI (Butler et al., 2024), CEDAR (Fesanghary, 2026), and GRACE (Fesanghary & Havaldar, 2026).

Table 2: **Causal discovery results on 43 systems.** System-macro averages excluding self-edges. Numerical Interactive entries use the prespecified policy with highest Directed F1. Bold marks the best value within each method family and setting, including ties; lower nSHD is better. Inputs differ between classical methods and LLMs under matched simulation ceilings.

Method	Observational			Interactive		
	Dir. F1↑	Skel. F1↑	nSHD↓	Dir. F1↑	Skel. F1↑	nSHD↓
Classical discovery						
Linear SINDy–STLSQ (Brunton et al., 2016)	**0.358**	**0.443**	0.572	**0.381**	**0.452**	0.579
Sparse VAR–BIC (Granger, 1969)	0.323	0.393	0.429	0.361	0.437	0.468
PCMCI+ (Runge, 2020)	0.219	0.297	0.259	0.268	0.354	0.273
VAR–LiNGAM (Hyvärinen et al., 2010)	0.319	0.416	0.467	0.333	0.437	0.455
DYNOTEARS (Pamfil et al., 2020)	0.295	0.391	0.261	0.301	0.396	0.260
TSCI (Butler et al., 2024)	0.171	0.214	0.344	0.169	0.200	0.318
CEDAR (Fesanghary, 2026)	0.066	0.092	0.244	0.068	0.096	**0.238**
GRACE (Fesanghary & Havaldar, 2026)	0.052	0.072	**0.243**	0.054	0.076	**0.238**
LLM agents						
GPT-5.6 Terra	**0.746**	**0.794**	**0.098**	**0.765**	**0.808**	**0.092**
Claude Sonnet 4.6	0.530	0.629	0.165	0.546	0.656	0.160
DeepSeek V4.1 Flash	0.719	0.780	0.106	0.693	0.765	0.111
Gemini 2.5 Flash Thinking	0.606	0.679	0.131	0.551	0.638	0.144

We pair SINDy, Sparse VAR, TSCI, CEDAR, and GRACE with ten prespecified acquisition policies adapted to the environment; the other estimators use Random under the same simulation ceilings.

LLM agents and component configurations. We evaluate GPT-5.6 Terra, Claude Sonnet 4.6, DeepSeek V4.1 Flash, and Gemini 2.5 Flash Thinking with action and submission validation. For each model, we also assess numerical tools and harness support on fixed trajectories and context. Tools provide numerical causal reports; the harness validates and repairs submissions. Component effects are compared within this separate evidence interface. LLM scores average three API trials per system and condition. Implementation and evaluation details appear in Appendix B.

4.2 Evaluation Metrics

We report Directed F1, Skeleton F1, precision, recall, and normalized structural Hamming distance (nSHD), excluding self-edges. For predicted adjacency $\widehat{A}$, true positives, false positives, and false negatives are counted over ordered pairs $i \neq j$ against the reference adjacency $A^\star$:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{Directed F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{3}$$

Skeleton F1 applies the same definition to $S(A)_{ij} = A_{ij} \vee A_{ji}$ over unordered pairs $i < j$, counting reciprocal edges once. Our nSHD counts directed adjacency errors over the same node set:

$$\text{nSHD}(\widehat{A}, A^\star) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{1}[\widehat{A}_{ij} \neq A^\star_{ij}]. \tag{4}$$

An isolated edge reversal counts as two errors; ratios with zero denominators return zero.

4.3 Main Results

Biological priors support discovery under limited evidence. Agents recover substantial causal structure before selecting interventions (Table 2). All four LLMs outperform every numerical baseline in the observational setting across all three metrics; Terra achieves roughly twice the Directed F1 of the strongest classical method. This advantage is consistent with the value of pretrained knowledge and grounded descriptions when few trajectories are available. Because these methods receive different inputs, this comparison does not isolate the effect of pretraining.

Access to interventions does not ensure better discovery. Replacing untreated replicates with interventions improves Directed F1 for seven of the eight numerical estimators under the displayed

acquisition configurations; PCMCI+ gains about 22% relative to its observational score. Yet only half of the LLMs improve, and Gemini loses roughly 9%. Interventions can therefore improve recovery within the same simulation budget, but their benefit depends on the discovery method. Experiment selection must be evaluated together with the inference procedure that interprets its evidence.

4.4 Tools and Experiment Selection

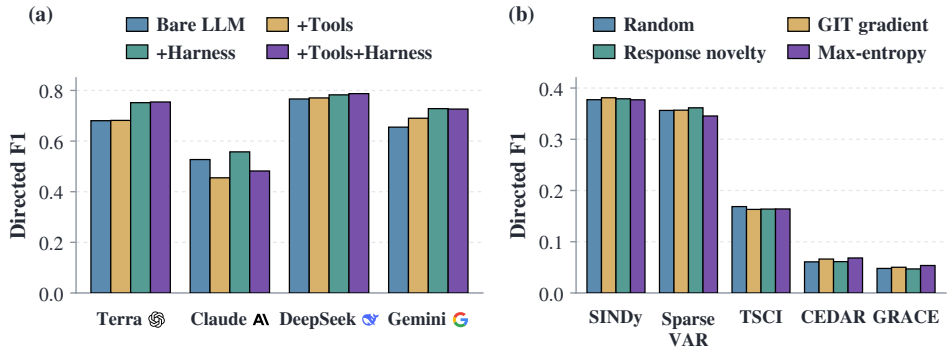

Figure 4: **Tool integration and acquisition effects.** (a) Four LLMs on shared trajectories and biological context, with three trials per condition. (b) Random and the three active policies selected in Table 2, with two interventions. Both panels report system-macro Directed F1 over 43 systems.

Numerical tools can weaken biological causal inference. At fixed trajectories and context (Figure 4a), adding numerical reports to the harness lowers Claude's Directed F1 by about 14%, while changes for Terra, DeepSeek, and Gemini stay below 0.005 in magnitude. These reports interpret the same displayed evidence. Their inconsistent benefits show that tool access alone does not ensure better recovery; integration with biological reasoning shapes how agents use the evidence.

Acquisition gains depend on the downstream estimator. Across ten policies, gains over Random stay below 0.8 percentage points for all five estimators (Figure 4b). Policy rankings reverse: maximum-entropy orientation helps CEDAR and GRACE but hurts Sparse VAR, while TSCI peaks with Random. For SINDy and Sparse VAR, Random already captures over 80% of the improvement from interventions. Selection gains remain small and estimator-dependent at this two-intervention budget.

5 Further Analysis

5.1 Context Alignment Shapes Interaction

Biological descriptions support discovery beyond explicit source identification (Figure 5). With the simulator and three-trajectory ceiling fixed, we compare reviewed context (*Biological*), removed source identifiers (*Deidentified*), removed biological content (*Anonymous*), and reassigned descriptions and action references (*Permuted*) across 43 systems. Deidentification changes interactive Directed F1 by less than 0.02 for every model; removing biological content causes much larger losses.

Experiments partly compensate for missing context but do not reliably correct misleading descriptions. All four models improve under Anonymous, raising mean Directed F1 by 72%, whereas three decline under Permuted and Gemini remains nearly unchanged despite experimental feedback.

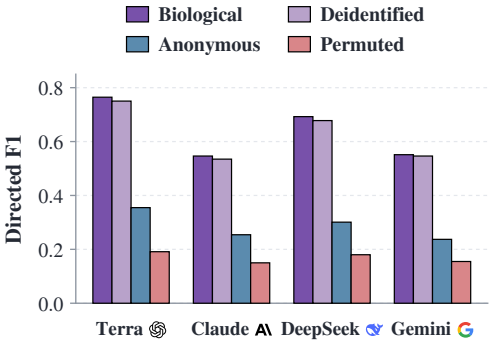

Figure 5: **Context shapes recovery.** Interactive Directed F1 over 43 systems, with three trials per system; all systems receive equal weight.

5.2 Current Agents Struggle to Exploit Experimental Budgets

Larger allowances do not yield sustained recovery gains for Terra (Figure 6). On three large systems, we increase the trajectory ceiling K and total observation budget Q from $(1, 100)$ to $(20, 2000)$ across seven settings. Both constrain acquisition; the agent chooses experiments, trajectory lengths, and when to submit. At twenty times the initial resource ceilings, Directed F1 is about 18% lower, with declines across all three large systems.

At the largest setting, the agent uses only 27% of its trajectory budget and 34% of its observation budget. Acquisition levels off as ceilings rise, and all rollouts submit with both resources remaining. Additional capacity stays unused, while acquired evidence does not improve recovery. Resource allocation and evidence integration remain limiting for this agent across the three large systems.

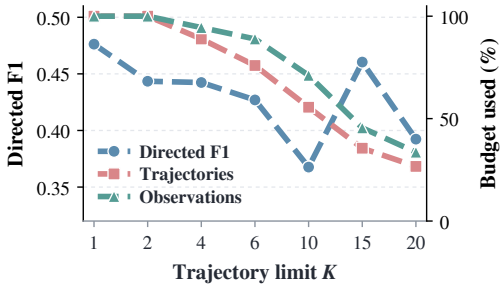

Figure 6: **Budget use and recovery.** Terra means across three systems, with three rollouts per system and setting. Both ceilings apply with $Q = 100K$; budget settings are evenly spaced.

5.3 Causal Coverage Declines with System Scale

Recovery declines with system size despite biological context and interventions (Figure 7). Across 43 systems under the same three-trajectory ceiling, interactive Directed F1 falls by 24–50% across the four agents from small (5–11 variables) to large (26–94) systems. Experimental access does not close this gap under the shared evidence ceiling, leaving less complete graphs for larger systems.

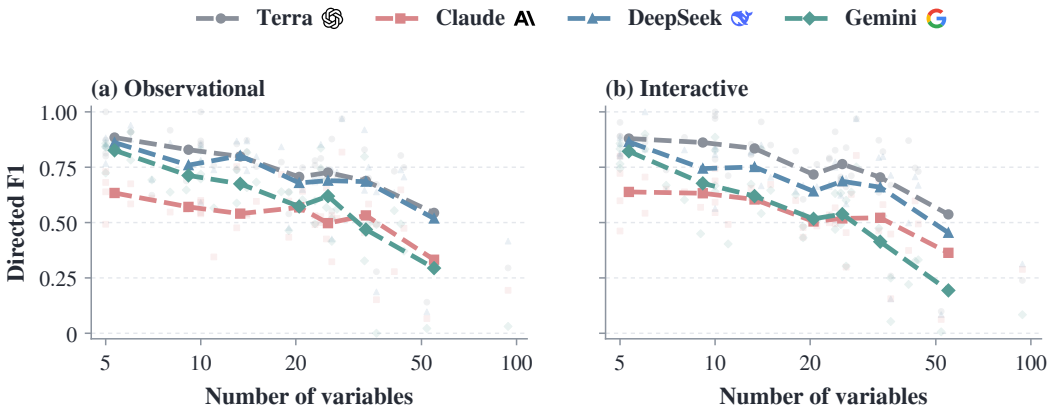

Figure 7: **Recovery across system scales.** System-macro scores within seven size groups under (a) observational and (b) interactive acquisition. Faint points show individual systems; group means weight systems equally after averaging three trials in each system and evaluation mode.

The sharper deterioration is in coverage: recall falls more than precision for every agent in both modes, leaving fewer true relations recovered as systems grow. These comparisons hold the evidence budget fixed; node count also covaries with topology, density, and biological domain.

5.4 Agentic RL Improves Discovery on Held-Out Systems

Agentic RL improves discovery on systems excluded from optimization (Figure 8). From identical initial weights, we train Qwen3.5-35B-A3B (Qwen Team, 2026) with GRPO (Shao et al., 2024) on 32 systems, independently with and without biological descriptions and meaningful action names; the latter uses opaque aliases. Both policies are evaluated under their training interfaces on the same 11 held-out systems. Both improve after 100 updates under identical resource ceilings.

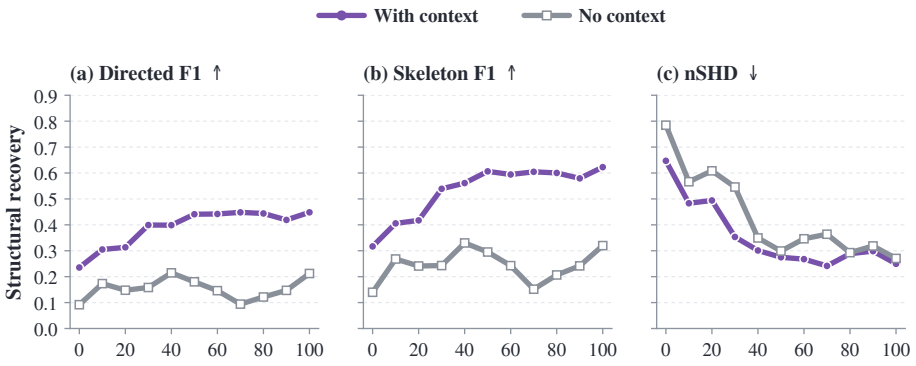

Figure 8: **Agentic RL with and without biological context.** Purple and gray curves show policies trained independently with and without context. Scores average four rollouts within each of the same 11 held-out systems, then weight systems equally regardless of graph size.

Biological grounding remains consequential after training: the context-trained policy reaches 0.448 Directed F1 versus 0.212 without context, and the gap widens during optimization. Similar final nSHD scores thus coexist with markedly different recovery of true relations. On sparse graphs, low error over all possible directed pairs can conceal missing causal edges. Training can therefore improve the optimized graph-error objective without eliminating reliance on biological context for recovering the underlying dependencies in biological systems excluded from optimization.

6 Discussion and Limitations

BioCausalX defines each discovery task through a published mechanistic model. Its reference graph captures the modeled dependencies; real systems may involve additional variables, context-dependent mechanisms, and measurement uncertainty. Literature evidence grounds the intervention choices, while their quantitative effects follow the simulator's assumptions. The interface also provides fully observed trajectories and repeatable experiments, whereas laboratory studies often involve partial measurements, variability, and differing costs and delays. These abstractions enable controlled comparisons against known structural targets, with transfer to experimental data remaining an important direction for validation.

Future work could explore reinforcement learning algorithms for long-horizon experiment planning and efficient exploration, together with agent systems with persistent memory, scientific tools, and collaboration among specialized agents. BioCausalX provides fixed structural targets and traceable interaction histories for comparing how these designs select experiments, integrate evidence, and revise hypotheses under shared resource limits. Extensions to partial observations and intervention-dependent costs and delays would introduce further challenges for planning. Comparisons with real perturbation data could then help assess which learned strategies remain useful under more realistic biological conditions.

7 Conclusion

We presented BioCausalX, a literature-grounded environment for interactive causal discovery across 43 executable biological systems. The environment aligns scientific descriptions, source-supported interventions, simulated trajectories, and fixed direct-dependency targets. Our results show that accurate biological semantics are central to current agent performance, while active experiment selection and numerical tool support provide model- and budget-dependent benefits rather than uniform improvements. A GRPO study further improves held-out structural recovery across systems absent from optimization, showing that the environment can support policy learning in addition to evaluation. By exposing these factors separately, BioCausalX enables more precise development and evaluation of scientific agents that must combine prior knowledge with newly acquired evidence.

AI USE STATEMENT

Generative AI systems were used as evaluated agents, as components of the environment-construction workflow, and for manuscript drafting, result interpretation, figure preparation, and code assistance. Within environment construction, model outputs assisted biological interpretation, model mapping, experimental-method assessment, contradiction checking, verification, and adjudication. Supporting passages were retained, unresolved candidates were excluded, and admitted operations were reviewed and validated through executable checks. AI-generated text was reviewed by the authors, and generative AI outputs were not treated as authoritative scientific evidence. The authors take responsibility for the final content, claims, code, and released artifacts.

ETHICS STATEMENT

This work introduces a simulation environment built from published mechanistic models and scientific records; it does not collect new human-subject or animal-subject data. The environment is intended for evaluating scientific reasoning and causal-discovery methods, not for clinical decision making. Source provenance and model scope are retained because simulated mechanisms may be incomplete, context specific, or inconsistent with later biological evidence. Released materials should be used in accordance with the licenses and attribution requirements of their source repositories.

REPRODUCIBILITY STATEMENT

Our code is available at `https://anonymous.4open.science/r/BioCausalX-submit-467B/`. Sections 3–4.2 define the task, interaction budgets, and metrics. The supplementary material details environment construction, evaluation interfaces and scoring, training protocols, and complete result tables. Source models, executable systems, reference graphs, intervention catalogs, evidence records, prompts, and task configurations are versioned together. Figure scripts and their data provenance are included in the repository.

REFERENCES

Ahmed Abdulaal, Adamos Hadjivasiliou, Nina Montaña-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel Castro, and Daniel Alexander. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://proceedings.iclr.cc/paper_files/paper/2024/hash/fe90657b12193c7b52a3418bdc351807-Abstract-Conference.html`.

Alexis Bellot, Kim Branson, and Mihaela van der Schaar. Neural graphical modelling in continuous-time: consistency guarantees and algorithms. In *International Conference on Learning Representations*, 2022. URL `https://arxiv.org/abs/2105.02522v3`.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, Chloe Anastasiades, Stefan Candra, Jason Dunkelberger, Dan Emery, Rob Evans, Malachi Hamada, Regan Huff, Rodney Kinney, Matt Latzke, Jaron Lochner, Ruben Lozano-Aguilera, Cecile Nguyen, Smita Rao, Amber Tanaka, Brooke Vlahos, Peter Clark, Doug Downey, Yoav Goldberg, Ashish Sabharwal, and Daniel S. Weld. AstaBench: Rigorous Benchmarking of AI Agents with a Scientific Research Suite. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://proceedings.iclr.cc/paper_files/paper/2026/file/b2ce9568dbb559aefc8c98ca5b5314ce-Paper-Conference.pdf`.

Kevin S. Brown, Colin C. Hill, Guillermo A. Calero, Christopher R. Myers, Kelvin H. Lee, James P. Sethna, and Richard A. Cerione. The statistical mechanics of complex signaling networks: Nerve growth factor signaling. *Physical Biology*, 1(3–4):184–195, 2004. doi: 10.1088/1478-3967/1/3/006. URL `https://doi.org/10.1088/1478-3967/1/3/006`.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Kurt Butler, Daniel Waxman, and Petar M. Djurić. Tangent Space Causal Inference: Leveraging Vector Fields for Causal Discovery in Dynamical Systems. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d9253fba38ed8a140f86fa22d89344ec-Abstract-Conference.html.

Yimeng Chen, Piotr Piękos, Mateusz Ostaszewski, Firas Laakom, and Jürgen Schmidhuber. PhysGym: Benchmarking LLMs in Interactive Physics Discovery with Controlled Priors. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. URL https://papers.nips.cc/paper_files/paper/2025/hash/42abcf9dffa48ca44fea1c497539a914-Abstract-Datasets_and_Benchmarks_Track.html.

Zhenhao Chen, Yongqiang Chen, Chenxi Liu, Junchi Yu, Xiangchen Song, Zijian Li, Jialin Li, Philip Torr, Bo Han, and Kun Zhang. CausalGame: Benchmarking Causal Thinking of LLM Agents in Games. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://causalgame.github.io/. Oral presentation.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=6z4YKr0GK6.

Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. CausalTime: Realistically Generated Time-series for Benchmarking of Causal Discovery. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=iad1yyyGme.

Mathieu Chevalley, Yusuf H. Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. A large-scale benchmark for network inference from single-cell perturbation data. *Communications Biology*, 8:412, 2025. doi: 10.1038/s42003-025-07764-y. URL https://www.nature.com/articles/s42003-025-07764-y.

Rafael S Costa, André Veríssimo, and Susana Vinga. KiMoSys: a web-based repository of experimental data for KInetic MOdels of biological SYStems. *BMC Systems Biology*, 8(1):85, 2014. doi: 10.1186/s12918-014-0085-3. URL https://link.springer.com/article/10.1186/s12918-014-0085-3.

Payam Dibaeinia and Saurabh Sinha. SERGIO: A Single-Cell Expression Simulator Guided by Gene Regulatory Networks. *Cell Systems*, 11(3):252–271.e11, 2020. doi: 10.1016/j.cels.2020.08.003. URL https://doi.org/10.1016/j.cels.2020.08.003.

Haonan Duan, Stephen Zhewen Lu, Caitlin F. Harrigan, Nishkrit Desai, Jiarui Lu, Michał Koziarski, Leonardo Cotta, and Chris J. Maddison. Measuring scientific capabilities of language models with a systems biology dry lab. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3733. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a23760ba51036d530a2656c3835f826c-Abstract-Datasets_and_Benchmarks_Track.html.

Mohammad Fesanghary. CEDAR: Causal edge discovery for autoregressive processes, 2026. URL https://arxiv.org/abs/2607.20696v1. Version 1, July 22, 2026.

Mohammad Fesanghary and Abhinav Havaldar. GRACE: Gated refinement for accurate causal edge discovery in high-dimensional time series, 2026. URL https://arxiv.org/abs/2606.23880.

Kanishk Gandhi, Michael Y. Li, Lyle Goodyear, Agam Bhatia, Louise Li, Aditi Bhaskar, Mohammed Zaman, and Noah D. Goodman. BoxingGym: Benchmarking progress in automated experimental design and model discovery, 2025. URL https://arxiv.org/abs/2501.01540.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, 2026. doi: 10.1038/s41586-026-10644-y. URL https://www.nature.com/articles/s41586-026-10644-y.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. doi: 10.2307/1912791.

Benjamin M. Gyori, Charles Tapley Hoyt, and Albert Steppi. Gilda: biomedical entity text normalization with machine-learned disambiguation as a service. *Bioinformatics Advances*, 2(1):vbac034, 2022. doi: 10.1093/bioadv/vbac034.

Alain Hauser and Peter Bühlmann. Two Optimal Strategies for Active Learning of Causal Models from Interventional Data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014. doi: 10.1016/j.ijar.2013.11.007. URL https://doi.org/10.1016/j.ijar.2013.11.007.

Alex Havrilla, David Alvarez-Melis, and Nicolo Fusi. IGDA: Interactive graph discovery through large language model agents. In *ICLR 2025 Workshop on Reasoning and Planning for Large Language Models*, 2025. URL https://openreview.net/forum?id=cHV3Iw84AC.

Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(84):2523–2547, 2008. URL https://www.jmlr.org/papers/v9/he08a.html.

Benjamin Herdeanu, Juan Nathaniel, Carla Roesch, Jatan Buch, Gregor Ramien, Johannes Haux, and Pierre Gentine. CausalDynamics: A large-scale benchmark for structural discovery of dynamical causal models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/39a3f7dc00e4723a5ca7808b04fc4d4f-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

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://www.science.org/doi/10.1126/science.adz4351.

M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang, and SBML Forum. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003. doi: 10.1093/bioinformatics/btg015. URL https://doi.org/10.1093/bioinformatics/btg015.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11:1709–1731, 2010. URL https://www.jmlr.org/papers/v11/hyvarinen10a.html.

Moksh Jain, Mehdi Bennani, Johannes Bausch, Yuri Chervonyi, Bogdan Georgiev, Simon Osindero, and Nenad Tomašev. MaD Physics: Evaluating information seeking under constraints in physical environments, 2026. URL https://arxiv.org/abs/2605.10820.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. DiscoveryWorld: A Virtual Environment for Developing and Evaluating Automated Scientific Discovery Agents. In *Advances in Neural Information Processing Systems*, volume 37, pp. 10088–10116, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/13836f251823945316ae067350a5c366-Abstract-Datasets_and_Benchmarks_Track.html.

Sanchit Kabra, Nikhil Abhyankar, Saaketh Desai, Prasad Iyer, and Chandan K. Reddy. LLM-AutoSciLab: Closed-loop scientific discovery via active experimentation with LLMs. *arXiv preprint arXiv:2605.24043*, 2026. doi: 10.48550/arXiv.2605.24043. URL https://arxiv.org/abs/2605.24043.

Dongkwan Kim, Yiming Gao, Yining Yang, and Yang Shen. PertReason: A Knowledge-Grounded Benchmark and Framework for Cell-State-Conditioned Mechanistic Reasoning of Perturbation Effects. arXiv preprint, 2026. URL https://arxiv.org/abs/2607.18777.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2305.00050v3.

Nicolas Le Novère, Andrew Finney, Michael Hucka, Upinder S Bhalla, Fabien Campagne, Julio Collado-Vides, Edmund J Crampin, Matt Halstead, Edda Klipp, Pedro Mendes, Poul Nielsen, Herbert Sauro, Bruce Shapiro, Jacky L Snoep, Hugh D Spence, and Barry L Wanner. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnology*, 23(12):1509–1515, 2005. doi: 10.1038/nbt1156. URL https://www.nature.com/articles/nbt1156.

David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–12, 1994.

Junyi Li, Yongqiang Chen, Chenxi Liu, Qianyi Cai, Tongliang Liu, Bo Han, Kun Zhang, and Hui Xiong. Can Large Language Models Help Experimental Design for Causal Discovery? In *ICML 2025 Workshop on Scaling up Intervention Models*, 2025. URL https://arxiv.org/abs/2503.01139.

Peiwen Li, Xin Wang, Zeyang Zhang, Yuan Meng, Fang Shen, Yue Li, Jialong Wang, Yang Li, and Wenwu Zhu. RealTCD: Temporal Causal Discovery from Interventional Data with Large Language Model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024. doi: 10.1145/3627673.3680042. URL https://doi.org/10.1145/3627673.3680042.

Zi-Rong Li, Si-Yang Liu, Tian-Zuo Wang, and Han-Jia Ye. CausalArena: Benchmarking causal discovery in the foundation model era, 2026. URL https://arxiv.org/abs/2609.11897. Preprint, September 2026.

Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. Discovery of the hidden world with large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 102307–102365, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b99a07486702417d3b1bd64ec2cf74ad-Abstract-Conference.html.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards Data-Driven Discovery with Large Language Models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0d70af566e69f1dfb687791ecf955e28-Abstract-Conference.html.

Rahuman S Malik-Sheriff, Mihai Glont, Tung V N Nguyen, Krishna Tiwari, Matthew G Roberts, Ashley Xavier, Manh T Vu, Jinghao Men, Matthieu Maire, Sarubini Kananathan, Emma L Fairbanks, Johannes P Meyer, Chinmay Arankalle, Thawfeek M Varusai, Vincent Knight-Schrijver, Lu Li, Corina Dueñas-Roca, Gaurhari Dass, Sarah M Keating, Young M Park, Nicola Buso, Nicolas Rodriguez, Michael Hucka, and Henning Hermjakob. BioModels—15 years of sharing computational models in life science. *Nucleic Acids Research*, 48(D1):D407–D415, 2020. doi: 10.1093/nar/gkz1055. URL https://doi.org/10.1093/nar/gkz1055.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, DREAM5 Consortium, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804, 2012. doi: 10.1038/nmeth.2016. URL https://www.nature.com/articles/nmeth.2016.

Kevin Murphy. Model Discovery Agent: LLM-assisted Bayesian Experiment Design for Data-efficient Discovery of Mechanistic World Models. *arXiv preprint arXiv:2608.09696*, 2026. URL https://arxiv.org/abs/2608.09696v4. Version 4, revised August 25, 2026.

Mateusz Olko, Michał Zając, Aleksandra Nowak, Nino Scherrer, Yashas Annadani, Stefan Bauer, Łukasz Kuciński, and Piotr Miłoś. Trust Your ∇: Gradient-based Intervention Targeting for Causal Discovery. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2211.13715.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. DYNOTEARS: Structure Learning from Time-Series Data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1595–1605. PMLR, 2020. URL https://proceedings.mlr.press/v108/pamfil20a.html.

Bo Peng, Sirui Chen, Lei Xu, and Chaochao Lu. CauScientist: Teaching LLMs to Respect Data for Causal Discovery, 2026. URL https://arxiv.org/abs/2601.13614. arXiv preprint; publication venue not independently verified.

Martin Peters, Johann J Eicher, David D van Niekerk, Dagmar Waltemath, and Jacky L Snoep. The JWS online simulation database. *Bioinformatics*, 33(10):1589–1590, 2017. doi: 10.1093/bioinformatics/btw831. URL https://academic.oup.com/bioinformatics/article/33/10/1589/2960005.

Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, 2020. doi: 10.1038/s41592-019-0690-6. URL https://www.nature.com/articles/s41592-019-0690-6.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5. Model card: https://huggingface.co/Qwen/Qwen3.5-35B-A3B.

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.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1388–1397. PMLR, 2020. URL https://proceedings.mlr.press/v124/runge20a.html.

Thomas Schaffter, Daniel Marbach, and Dario Floreano. GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16): 2263–2270, 2011. doi: 10.1093/bioinformatics/btr373. URL https://pubmed.ncbi.nlm.nih.gov/21697125/.

Nino Scherrer, Olexa Bilaniuk, Yashas Annadani, Anirudh Goyal, Patrick Schwab, Bernhard Schölkopf, Michael C. Mozer, Yoshua Bengio, Stefan Bauer, and Nan Rosemary Ke. Learning Neural Causal Models with Active Interventions, 2021. URL https://arxiv.org/abs/2109.02429. arXiv preprint; publication venue not independently verified.

Leonard Schmiester, Yannik Schälte, Frank T. Bergmann, Tacio Camba, Erika Dudkin, Janine Egert, Fabian Fröhlich, Lara Fuhrmann, Adrian L. Hauber, Svenja Kemmer, Polina Lakrisenko, Carolin Loos, Simon Merkt, Wolfgang Müller, Dilan Pathirana, Elba Raimúndez, Lukas Refisch, Marcus Rosenblatt, Paul L. Stapor, Philipp Städter, Dantong Wang, Franz-Georg Wieland, Julio R. Banga, Jens Timmer, Alejandro F. Villaverde, Sven Sahle, Clemens Kreutz, Jan Hasenauer, and Daniel Weindl. PEtab—Interoperable specification of parameter estimation problems in systems biology. *PLOS Computational Biology*, 17(1):e1008646, 2021. doi: 10.1371/journal.pcbi.1008646. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008646.

Alexandra M. Schnoes, Shoshana D. Brown, Igor Dodevski, and Patricia C. Babbitt. Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLOS Computational Biology*, 5(12):e1000605, 2009. doi: 10.1371/journal.pcbi.1000605. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000605.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. doi: 10.1214/aos/1176344136.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1708.00489.

Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, Department of Computer Sciences, 2009. URL https://burrsettles.com/pub/settles.activelearning.pdf.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024. URL https://arxiv.org/abs/2409.19256.

Woosub Shin, John H. Gennari, Joseph L. Hellerstein, and Herbert M. Sauro. An automated model annotation system (AMAS) for SBML models. *Bioinformatics*, 39(11):btad658, 2023. doi: 10.1093/bioinformatics/btad658.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Panagiotis Tigas, Yashas Annadani, Andrew Jesson, Bernhard Schölkopf, Yarin Gal, and Stefan Bauer. Interventions, Where and How? Experimental Design for Causal Models at Scale. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2203.02016.

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://proceedings.neurips.cc/paper_files/paper/2022/hash/675e371eeeea99551ce47797ed6ed33e-Abstract-Conference.html.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. Causal Order: The Key to Leveraging Imperfect Experts in Causal Inference. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/edcbd38b2509b57954d38de6cd9c05b3-Abstract-Conference.html.

Julius von Kügelgen, Paul K. Rubenstein, Bernhard Schölkopf, and Adrian Weller. Optimal experimental design via bayesian optimization: Active causal structure learning for gaussian process networks. In *NeurIPS 2019 Workshop: Do the Right Thing—Machine Learning and Causal Inference for Improved Decision Making*, 2019. URL https://arxiv.org/abs/1910.03962.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your Agent Smarter than a 5th Grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, 2022. doi: 10.18653/v1/2022.emnlp-main.775. URL https://aclanthology.org/2022.emnlp-main.775/.

Matt L. Wiemann, Lindsay M. Smith, Peter Melchior, Siddharth Mishra-Sharma, Andrew Gordon Wilson, Pavel Izmailov, and Carolina Cuesta-Lázaro. DiscoverPhysics: Benchmarking LLMs for Out-of-the-Box Scientific Thinking, 2026. URL https://arxiv.org/abs/2605.26087.

Menghua Wu, Russell Littman, Jacob Levine, Lin Qiu, David Richmond, Tommaso Biancalani, and Jan-Christian Hütter. Contextualizing biological perturbation experiments through language. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b6e67ae290635d0874c4cb43ba2a2cfb-Paper-Conference.pdf.

Junlin Yang, Dylan Zhang, Xiangchen Song, Qirun Dai, Xiao Liu, Yuen Chen, Aniket Vashishtha, Jing Shi, Chenhao Tan, and Hao Peng. CausaLab: A scalable environment for interactive causal discovery toward AI scientists. *arXiv preprint arXiv:2605.26029*, 2026. doi: 10.48550/arXiv.2605.26029. URL https://arxiv.org/abs/2605.26029.

Zhiyu Zhao, Haoxuan Li, Haifeng Zhang, Jun Wang, Francesco Faccio, Jürgen Schmidhuber, and Mengyue Yang. Curious causality-seeking agents in open-ended worlds. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e21c92a1661709c3caf21fdd958bac99-Abstract-Conference.html.

Tianshi Zheng, Kelvin Kiu-Wai Tam, Newt Hue-Nam K. Nguyen, Baixuan Xu, Zhaowei Wang, Jiayang Cheng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Tianqing Fang, Yangqiu Song, Ginny Y. Wong, and Simon See. NewtonBench: Benchmarking Generalizable Scientific Law Discovery in LLM Agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/a0e1c2c40fc245b5fe7251ea33fbb045-Paper-Conference.pdf.

A Environment Construction and Interaction Details

This section details how source models, biological descriptions, and permitted operations are compiled into the environment. Section B specifies evaluation interfaces and baseline results; Sections C and D report training and additional analyses. Section E expands the related-work comparison.

A.1 Model Collection and Reference Graphs

We collect published models and documentation from BioModels (Malik-Sheriff et al., 2020), JWS Online (Peters et al., 2017), KiMoSys (Costa et al., 2014), and collections distributed in PEtab format (Schmiester et al., 2021). The 43 systems span signaling, metabolism, biological oscillators, and physiological regulation, with 5–94 state variables each (Table 1). Source models, including SBML representations (Hucka et al., 2003), are converted to a common executable form while preserving state identities and governing mechanisms. Numerical execution and reference-graph consistency are checked before experimental operations are added. Figure 3 also shows uneven operation coverage: 87 of the 168 axes affect node mechanisms, 57 affect reactions or interactions, and 24 affect exogenous inputs. The catalog spans these operation types without balancing their counts across systems.

Reference construction. Equation 1 tests symbolic dependence after combining all terms in each governing equation and retaining parameters symbolically. Simulations use the frozen numerical parameterization. The directed, unsigned graph can contain feedback cycles; self-dependencies are retained in the model but excluded from scoring. Its target remains the unperturbed model throughout an episode, including when an intervention modifies a represented mechanism. Thus, a downstream response provides evidence to interpret rather than a direct-edge label.

A.2 Scientific Context and Grounded Operations

Scientific descriptions. Model annotations are reconciled with publications and database records to identify biological entities and distinguish model-specific states, such as molecular forms or compartmental pools. Evidence locations are retained. Agents receive these identities and non-directional context, including the states associated with permitted operations. Governing equations, exact parameter or reaction bindings, and reference adjacency remain hidden.

Candidate retrieval and review. Candidate operations are retrieved from the model's associated publications and studies of the relevant entities. Review checks whether the experiment was performed, which entity and mechanism it targets, and whether it is compatible with the modeled context. The construction tooling assigns LLM roles to biological interpretation, model mapping, experimental-method assessment, contradiction checking, verification, and adjudication. Supporting passages and assessments are retained; unresolved candidates are excluded.

Executable mappings. Each accepted operation is mapped to inputs, states, or mechanisms represented in M_0. Modifying enzyme activity acts through its reaction terms; changing an external stimulus acts through its represented input or initial preparation. An operation may therefore affect several terms without directly setting downstream states. Its public description specifies the biological manipulation and allowed condition, while the execution mapping remains private. Approximate numerical strengths are recorded separately from the source evidence supporting the operation.

Intervention axes and release validation. Realizations controlling the same represented mechanism form one *intervention axis*; doses, schedules, and supporting publications do not independently increase this count. The released catalog contains 168 axes, with at least two per system. Admission requires source support, compatible context, an executable mapping, and successful graph and trajectory checks. Numerical validation checks finite trajectories and nontrivial responses under the declared conditions. Models, reference graphs, actions, and audit records are versioned together. These checks establish executable experimental choices without guaranteeing that the available evidence identifies every edge.

A.3 Interaction and Observation Protocol

In the standard discovery protocol, each experiment selects the untreated condition or the single frozen condition of a permitted axis and starts from a fresh preparation. The environment validates the request, simulates the system over its prescribed observation horizon, and returns observations with the remaining budget. All integrated states are observable, and time is local to each experiment. Equation 2 specifies the experiment ceiling B_{exp} and the total generated-point ceiling B_{obs}.

Complete trajectories are recorded, while the standard LLM interface presents sampled values and per-variable statistics. Agents submit a directed graph at termination; reference-based scores are withheld during discovery. Requests, observations, and predictions are retained with the task configuration. Section B.1 defines the evaluated displays and scoring rules. The stateful budget study in Section D.2 additionally permits continuation and online intervention within a live trajectory, with its own explicit accounting rules.

B Evaluation Protocols and Complete Results

Standard discovery compares inference from untreated and agent-selected trajectories. The component study holds evidence fixed to examine numerical reports and submission validation. The protocols below specify what each comparison can establish.

B.1 Evaluation Interfaces and Scoring

Standard discovery interface. Both observational and interventional discovery allow at most three simulations of 101 generated time points each, beginning with an untreated baseline. Observational runs permit only untreated replicates; interventional runs permit two subsequent experiments selected sequentially from the system's documented action space. Classical methods receive the full numerical trajectories. LLM agents receive reviewed system and node descriptions, permitted experimental operations, and a compact trajectory representation: at most eight uniformly sampled, endpoint-inclusive values per requested window, together with per-node summary statistics. The standard interactive agent validates actions and submissions, allows up to five interaction turns, and provides a final submission opportunity. The comparisons therefore match simulation ceilings while retaining different side information and input representations.

Fixed-evidence component comparison. For each of the four LLMs, Bare LLM, +Tools, +Harness, and +Tools+Harness use three fixed untreated trajectories with 12 displayed samples each and identical biological context. Tools add Linear SINDy and Sparse VAR reports plus a fixed top-three rank consensus, fitted to those displayed samples. Raw and tool-augmented inputs are each assessed before and after the validation harness, yielding four configurations. Valid initial graphs are retained at the harness endpoint. This study measures numerical-report and submission-recovery effects on shared evidence; its prompt and evidence representation differ from standard discovery.

Trial counts and resource ceilings. Primary LLM results average three API trials per system and condition under frozen simulator settings; the providers do not expose an API seed parameter. Standard-discovery completion limits are 6,000 tokens for Terra and Claude and 32,000 for DeepSeek and Gemini; the latter two use 64,000 in the component study. Classical results use one or three frozen local replicates as available, and each acquisition-policy configuration uses one local run. Simulation ceilings are shared across methods, whereas inference allowances differ.

Three Claude observational trials used two simulations; all other primary LLM trials used three. The fixed-evidence component prompt also retains a legacy sentence referring to six trajectories, although the recorded panels contain three. This wording is identical across raw and tool-augmented inputs and is retained when reproducing those results.

Aggregation. Each metric is computed per trial, averaged within a system, and then averaged equally across systems. Macro F1 is therefore an average of F1 values, not a value reconstructed from macro precision and recall. For strict-first component rows, F1 includes invalid responses as specified below, while precision, recall, and nSHD condition on validity. Acquisition-policy tables cover all 43 systems; paired policy diagnostics use the 30 systems with more candidate actions than

the two-experiment budget. Differences and relative changes use unrounded values, so they may differ slightly from arithmetic on displayed scores.

Invalid outputs and component scoring. Bare LLM and +Tools rows evaluate the initial response after removing at most one complete outer JSON code fence and applying strict graph validation. Invalid first responses receive zero Directed F1 and Skeleton F1; precision, recall, and nSHD are averaged only over valid outputs. Consequently, these columns describe different response sets when validity is below 100%.

Harness rows allow tolerant local parsing and structural validation, followed when necessary by at most two schema-only repair turns without reference-graph feedback. Valid first graphs are unchanged, and recovered graphs are scored normally. Missing experiment records are treated as errors rather than empty predictions. The common fence normalization recovers five Gemini responses without tools and two with tools without changing graph content.

Submission reliability and tool effects. Table 3 reports valid-submission counts for the input conditions in Figure 4a. All harness endpoints recover 129/129 submissions while leaving valid first graphs unchanged. For Terra, Directed F1 rises from 0.681 to 0.752 (about 10%) without new observations, identifying submission reliability as a contributor to end-to-end recovery. Both Claude harness conditions reach 100% validity, yet tools reduce Directed F1 from 0.557 to 0.482. The tool penalty therefore persists when every submitted graph is valid.

Table 3: **Submission reliability on fixed evidence.** Valid graphs out of 129 trials per input condition (43 systems, three trials each). Harness endpoints include the parsing and schema-only recovery described above.

Backbone	Bare LLM	+Tools	+Harness	+Tools+Harness
Terra	117	116	129	129
Claude	121	122	129	129
DeepSeek	126	126	129	129
Gemini	112	121	129	129

Native Codex comparison. Native Codex uses the same three-trajectory panel but changes the provider, system instructions, available tools, prompt, and inference allowances. Its scores therefore compare complete systems. Directed F1 is 0.757, alongside 0.752 for Terra with validation and 0.754 with tools and validation. Native Codex yields 124/129 valid submissions, compared with 129/129 at both validation-harness endpoints.

B.2 Numerical Baselines and Acquisition Policies

Estimator configurations. Linear SINDy–STLSQ uses a linear state library (Brunton et al., 2016). Sparse VAR–BIC combines lag-one regression, Lasso regularization, and BIC selection (Granger, 1969; Tibshirani, 1996; Schwarz, 1978). The ten-policy sweep covers SINDy, Sparse VAR, TSCI, CEDAR, and GRACE; PCMCI+, VAR–LiNGAM, and DYNOTEARS use Random acquisition. All methods are scored against the same direct-dependency target.

Acquisition policies. The ten policy implementations are adapters for the environment, motivated by prior work: Random (Settles, 2009), Coverage (Sener & Savarese, 2018), Uncertainty (Lewis & Gale, 1994), maximum-entropy orientation design (He & Geng, 2008), AIT disagreement (Scherrer et al., 2021), D-optimality and BOED-EIG (von Kügelgen et al., 2019), ABCI graph-EIG (Toth et al., 2022), GIT gradient (Olko et al., 2023), and the environment's Response novelty rule. The metrics below are system-macro averages over all 43 systems; the 30-system choice-eligible subset is used only for the paired policy diagnostics, not for the displayed aggregate scores. For each estimator in the sweep, Table 2 displays the policy with the highest unrounded system-macro Directed F1 on this same 43-system cohort. The candidate policies are prespecified, while the displayed choice is made from their evaluation scores. One policy is selected per estimator across all systems, and every metric in that row comes from the same configuration. This selection summarizes the evaluated policies

Table 4: **Ten-policy Interactive sweep for Linear SINDy, Sparse VAR, and TSCI.** System-macro Precision, Recall, Directed F1, Skeleton F1, and nSHD over 43 systems. Bold rows are the configurations reported in the main table.

Method	Policy	Directed F1	Skeleton F1	Precision	Recall	nSHD
Linear SINDy–STLSQ	Random	0.377	0.454	0.269	0.823	0.583
	Coverage	0.377	0.451	0.269	0.820	0.583
	Uncertainty	0.377	0.451	0.269	0.817	0.578
	Max-entropy orientation	0.377	0.451	0.268	0.817	0.578
	AIT disagreement	0.378	0.452	0.269	0.818	0.577
	D-optimality	0.377	0.451	0.269	0.815	0.577
	BOED-EIG	0.376	0.452	0.267	0.816	0.580
	ABCI graph-EIG	0.375	0.453	0.266	0.817	0.584
	GIT gradient	**0.381**	**0.452**	**0.271**	**0.831**	**0.579**
	Response novelty	0.379	0.453	0.270	0.827	0.582
Sparse VAR–BIC	Random	0.356	0.432	0.281	0.667	0.468
	Coverage	0.355	0.430	0.281	0.664	0.465
	Uncertainty	0.354	0.430	0.279	0.659	0.466
	Max-entropy orientation	0.345	0.428	0.274	0.635	0.463
	AIT disagreement	0.357	0.432	0.276	0.667	0.466
	D-optimality	0.351	0.436	0.273	0.644	0.466
	BOED-EIG	0.354	0.438	0.274	0.652	0.466
	ABCI graph-EIG	0.356	0.438	0.275	0.661	0.467
	GIT gradient	0.357	0.433	0.280	0.662	0.462
	Response novelty	**0.361**	**0.437**	**0.279**	**0.677**	**0.468**
TSCI (2024)	**Random**	**0.169**	**0.200**	**0.180**	**0.238**	**0.318**
	Coverage	0.165	0.195	0.169	0.235	0.319
	Uncertainty	0.158	0.197	0.174	0.228	0.323
	Max-entropy orientation	0.164	0.198	0.171	0.236	0.327
	AIT disagreement	0.163	0.200	0.169	0.231	0.321
	D-optimality	0.162	0.197	0.172	0.227	0.325
	BOED-EIG	0.165	0.199	0.173	0.235	0.326
	ABCI graph-EIG	0.158	0.193	0.169	0.216	0.317
	GIT gradient	0.163	0.193	0.159	0.232	0.319
	Response novelty	0.164	0.200	0.170	0.230	0.321

without a separate validation cohort. Bold rows below identify the displayed choices; all alternatives are retained for comparison.

Table 4 shows different preferred policies: GIT gradient for SINDy, Response novelty for Sparse VAR, and Random for TSCI. The policy ranking therefore depends on the estimator; none is uniformly best across these three methods.

B.3 CEDAR AND GRACE IMPLEMENTATION DETAILS

CEDAR and GRACE were selected before aggregate BIOCAUSALX scores were inspected, based on compatibility with the dynamical target, fidelity to the released evaluation path, official-code quality, reproducibility, feasibility across all 43 systems, and methodological diversity. We use the official Causal-TS implementation (`https://github.com/bloomberg/causal-ts`), version 0.27.0 at commit `050e54a6ac5bd1c86d307c75852adb6524dc7c29`.

The source methods supply estimators; the Interactive condition adds the same ten external acquisition policies used in this environment. Each episode contains an untreated 101-point trajectory and two sequentially selected 101-point experiments. All 860 method–policy–system episodes complete, giving 43/43 coverage and zero failures in every cell. CEDAR returns lagged edges, and GRACE also includes a contemporaneous component. Their scores evaluate these temporal estimates against the common direct-dependency target; the estimands are not identical.

In Table 5, maximum-entropy orientation gives the highest Directed F1 for both methods. Across policies, recall remains 0.033–0.041 for CEDAR and 0.024–0.029 for GRACE. Both complete all runs but recover only a small fraction of the common target edges.

B.4 COMPLETE DISCOVERY AND COMPONENT RESULTS

Table 6 expands Table 2 with precision and recall and reports the separate fixed-evidence component study. Standard-discovery rows compare observational and interactive evidence; component rows compare configurations within the shared fixed panel.

Table 5: **Complete CEDAR and GRACE Interactive policy sweep.** System-macro metrics over the 43-system cohort. Coverage and failures are reported for each method-policy cell. Bold rows are repeated in the main table.

Method	Policy	Directed F1	Skeleton F1	Precision	Recall	nSHD	Cov./fails
CEDAR (2026)	Random	0.061	0.092	0.347	0.036	0.240	43/43; 0
	Coverage	0.062	0.091	0.346	0.037	0.240	43/43; 0
	Uncertainty	0.055	0.077	0.315	0.033	0.241	43/43; 0
	Max-entropy orientation	**0.068**	**0.096**	**0.344**	**0.041**	**0.238**	43/43; 0
	AIT disagreement	0.061	0.086	0.379	0.036	0.240	43/43; 0
	D-optimality	0.058	0.085	0.302	0.035	0.241	43/43; 0
	BOED-EIG	0.062	0.090	0.329	0.037	0.241	43/43; 0
	ABCI graph-EIG	0.059	0.085	0.324	0.036	0.240	43/43; 0
	GIT gradient	0.066	0.093	0.363	0.040	0.239	43/43; 0
	Response novelty	0.061	0.087	0.373	0.036	0.240	43/43; 0
GRACE (2026)	Random	0.048	0.066	0.423	0.026	0.239	43/43; 0
	Coverage	0.050	0.075	0.425	0.027	0.240	43/43; 0
	Uncertainty	0.053	0.072	0.428	0.029	0.238	43/43; 0
	Max-entropy orientation	**0.054**	**0.076**	**0.445**	**0.029**	**0.238**	43/43; 0
	AIT disagreement	0.047	0.062	0.380	0.025	0.239	43/43; 0
	D-optimality	0.045	0.062	0.398	0.024	0.241	43/43; 0
	BOED-EIG	0.045	0.062	0.400	0.024	0.241	43/43; 0
	ABCI graph-EIG	0.047	0.063	0.382	0.025	0.240	43/43; 0
	GIT gradient	0.050	0.077	0.421	0.027	0.239	43/43; 0
	Response novelty	0.047	0.062	0.391	0.026	0.239	43/43; 0

Interaction gains and remaining structural errors. Under the standard evaluation, Terra's Directed F1 rises from 0.746 to 0.765 when untreated replicates are replaced with agent-selected experiments. Its interactive Skeleton F1 is 0.808, indicating residual adjacency errors even when edge directions are ignored. These aggregate scores do not distinguish limitations in experiment choice from limitations in interpreting the resulting observations.

Precision and structural coverage. The interactive LLMs retain precision between 0.842 and 0.904, while recall ranges from 0.431 to 0.715. Their predictions therefore omit many reference edges despite being mostly correct when an edge is reported. Interaction also changes the two error types differently: DeepSeek's precision rises from 0.879 to 0.904 while recall falls from 0.647 to 0.601; Gemini follows the same pattern. Their lower interactive F1 thus accompanies narrower structural coverage, even as the reported edges become more selective.

Different errors across numerical estimators. In interactive discovery, SINDy combines 0.831 recall with 0.271 precision; CEDAR and GRACE recover few target edges, with recall of 0.041 and 0.029. Low Directed F1 thus reflects different error profiles across estimators. The lower nSHD of CEDAR and GRACE also needs this context: nSHD counts mismatches over all possible directed pairs. On sparse reference graphs, conservative predictions can obtain a low nSHD while missing many true relations. Precision and recall expose this trade-off.

Tool effects after submission recovery. In the fixed-evidence block, adding numerical reports to Claude's harness reduces both precision (0.861 to 0.743) and recall (0.453 to 0.407), despite complete submission validity in both conditions. The loss therefore extends to graph content. With the harness enabled, adding tools changes Directed F1 by just 0.002 for Terra, 0.004 for DeepSeek, and -0.001 for Gemini. These matched comparisons show that reliable submission alone does not make the current numerical reports useful for structural inference.

C Agentic Reinforcement Learning Details

The main RL study trains separate policies with and without biological context. Section C.3 instead removes context from fixed policy weights, using a policy trained under a different reward schedule.

C.1 Training and Evaluation Protocol

Split and optimization. The experiment uses the frozen Gold-v31 catalog of 43 systems. The effective split contains 32 training systems and 11 evaluation systems with no optimizer exposure. Seven evaluation systems belong to the original source-family holdout. Four additional systems (KIMOSYS_MODEL_40, Lucarelli_CellSystems2018, SRIMANI2017, and Weber_BMC2015) were

Table 6: **Detailed causal-discovery results on 43 systems and fixed-evidence component study.** System-macro averages excluding self-edges. Standard-discovery and fixed-evidence blocks use the interfaces in Section B.1. Numerical Interactive rows use the selection rule in Section B.2. For initial component responses, invalid graphs receive zero F1 while precision, recall, and nSHD condition on validity. Bold marks the highest Directed F1 within each numerical setting or matched component group, and the higher value in each LLM observational–interactive pair. Native Codex is a separate complete-system comparison.

Method	Condition	Dir. F1↑	Skel. F1↑	Prec.↑	Recall↑	nSHD↓
Standard Discovery Evaluation						
Numerical baselines — observational trajectories						
Linear SINDy–STLSQ	Observational	**0.358**	0.443	0.263	0.751	0.572
Sparse VAR–BIC	Observational	0.323	0.393	0.265	0.558	0.429
PCMCI+	Observational	0.219	0.297	0.369	0.182	0.259
VAR–LiNGAM	Observational	0.319	0.416	0.284	0.520	0.467
DYNOTEARS	Observational	0.295	0.391	0.420	0.267	0.261
TSCI	Observational	0.171	0.214	0.161	0.268	0.344
CEDAR (2026)	Observational	0.066	0.092	0.365	0.040	0.244
GRACE (2026)	Observational	0.052	0.072	0.371	0.029	0.243
Numerical baselines — selected Interactive configurations						
Linear SINDy–STLSQ	GIT gradient	**0.381**	0.452	0.271	0.831	0.579
Sparse VAR–BIC	Response novelty	0.361	0.437	0.279	0.677	0.468
PCMCI+	Random	0.268	0.354	0.386	0.254	0.273
VAR–LiNGAM	Random	0.333	0.437	0.284	0.535	0.455
DYNOTEARS	Random	0.301	0.396	0.414	0.280	0.260
TSCI	Random	0.169	0.200	0.180	0.238	0.318
CEDAR (2026)	Max-entropy orientation	0.068	0.096	0.344	0.041	0.238
GRACE (2026)	Max-entropy orientation	0.054	0.076	0.445	0.029	0.238
LLM agents — observational vs. agent-selected experiments						
GPT-5.6 Terra	Observational	0.746	0.794	0.900	0.678	0.098
	Agent-selected	**0.765**	0.808	0.893	0.715	0.092
Claude Sonnet 4.6	Observational	0.530	0.629	0.854	0.412	0.165
	Agent-selected	**0.546**	0.656	0.863	0.431	0.160
DeepSeek V4.1 Flash	Observational	**0.719**	0.780	0.879	0.647	0.106
	Agent-selected	0.693	0.765	0.904	0.601	0.111
Gemini 2.5 Flash Thinking	Observational	**0.606**	0.679	0.831	0.521	0.131
	Agent-selected	0.551	0.638	0.842	0.460	0.144
Fixed-Evidence Component Study						
Numerical baselines on the shared evidence panel						
Linear SINDy–STLSQ	Numerical only	**0.353**	0.447	0.249	0.797	0.621
Sparse VAR–BIC	Numerical only	0.227	0.314	0.206	0.328	0.368
Top-3 rank consensus (ours)	Numerical only	0.261	0.349	0.272	0.293	0.313
LLM system configurations						
GPT-5.6 Terra	Bare LLM (first)	0.681	0.718	0.903	0.682	0.094
	+ Tools (first)	0.682	0.719	0.910	0.688	0.094
	+ Harness	0.752	0.792	0.904	0.685	0.094
	+ Tools + Harness	**0.754**	0.798	0.911	0.686	0.094
	Native Codex (full)	0.757	0.794	0.911	0.725	0.089
Claude Sonnet 4.6	Bare LLM (first)	0.527	0.630	0.860	0.452	0.153
	+ Tools (first)	0.455	0.573	0.743	0.405	0.188
	+ Harness	**0.557**	0.669	0.861	0.453	0.153
	+ Tools + Harness	0.482	0.606	0.743	0.407	0.188
DeepSeek V4.1 Flash	Bare LLM (first)	0.766	0.805	0.877	0.736	0.090
	+ Tools (first)	0.771	0.804	0.867	0.754	0.090
	+ Harness	0.783	0.820	0.878	0.736	0.090
	+ Tools + Harness	**0.787**	0.821	0.867	0.755	0.090
Gemini 2.5 Flash Thinking	Bare LLM (first)	0.655	0.690	0.847	0.716	0.095
	+ Tools (first)	0.690	0.725	0.794	0.707	0.108
	+ Harness	**0.728**	0.768	0.850	0.698	0.100
	+ Tools + Harness	0.727	0.765	0.794	0.710	0.108

added after an exposure audit showed that they had entered no gradient update; this membership was fixed before their evaluation scores were inspected. We train two policies independently from the same Qwen3.5-35B-A3B weights. One sees the reviewed system and node descriptions plus meaningful experiment names. The other sees no descriptions and deterministic opaque action aliases during both training and evaluation; the aliases still invoke the same executable actions. This is a joint ablation of descriptions and action naming, not a separation of the two mechanisms.

We implement GRPO with verl/HybridFlow (Sheng et al., 2024). Each GRPO update samples eight systems and four agent rollouts per system. Both runs use 100 updates, actor learning rate 10^{-6}, eight A100 GPUs, rollout tensor parallel size four, six simulations and 600 cumulative time steps per episode, and the same decoding, feedback, tools, and verifier. We evaluate step 0 and every 10 updates through step 100, using four rollouts for each of the 11 held-out systems: 44 episodes per checkpoint, 484 per condition, and 968 total. The no-context run is direct. The context run resumes complete checkpoints at steps 20, 40, and 80, restoring model, optimizer, scheduler, RNG, and data state. A hardware-localized router fault required replacing top-k selection with descending sort for the final continuation; this preserves the selected top eight experts and does not change the task, reward, or optimization data.

Reward and aggregation. Both policies use the fixed objective

$$r = 0.1f + 0.9(1 - \mathrm{nSHD}), \tag{5}$$

where f indicates a valid, strictly formatted submission and nSHD is the number of mismatched directed off-diagonal entries divided by $N(N-1)$. A malformed or missing final graph receives $f = 0$, reward and both F1 scores zero, and nSHD 1. For every checkpoint, we first average the four agent rollouts of each system and then macro-average the 11 systems; thus a system contributes equally regardless of graph size. The separate inference-time diagnostic uses a different training and reward schedule, specified in Section C.3.

C.2 Learning Curves and Per-System Outcomes

The complete curves in Table 7 are non-monotone, but both endpoint changes are positive. The context-trained policy gains 0.213 Directed F1 and 0.306 Skeleton F1; the no-context policy gains 0.120 and 0.180. Directed F1 rises from 0.235 to 0.448 with context and from 0.092 to 0.212 without it; using unrounded scores, the context advantage widens from 0.144 to 0.236. Each curve is one training trajectory, so the comparison does not estimate variation across independent seeds. Figure 9 complements the structural metrics with the optimized reward and submission validity.

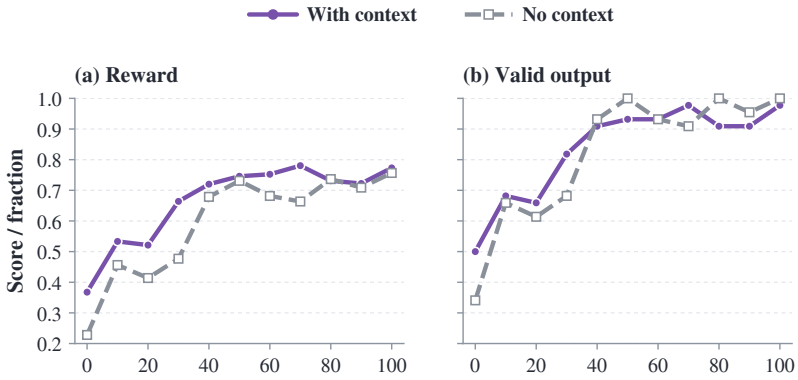

Figure 9: **Reward and execution reliability during matched Agentic RL.** Solid curves train and evaluate with context; dashed curves train and evaluate without it. Reward is $0.1f + 0.9(1 - \mathrm{nSHD})$, and validity is the accepted-submission fraction. Aggregation and held-out systems match Figure 8.

The context-trained policy improves Directed F1 on 9 of 11 held-out systems, and the no-context policy improves on 10 of 11 from its lower baseline (Figure 10; Table 9).

Reward and structural recovery. Table 8 separates structural metrics from reward and submission validity. Final rewards are close (0.773 with context and 0.757 without it), as are nSHD scores (0.250 and 0.270). The no-context policy reaches 100% validity yet only 0.212 Directed F1. Thus, improved execution and optimized reward do not by themselves imply comparable recovery of true relations.

The objective helps explain why these endpoints can differ: nSHD weights errors over all directed non-self pairs. For example, an empty graph has nSHD equal to the reference edge density while

Table 7: Complete learning curves for independently trained context and no-context policies. Each cell averages four rollouts within each of the same 11 systems excluded from optimization, then weights systems equally.

	With context					Without context				
Step	Reward	Dir. F1	Skel. F1	nSHD	Valid	Reward	Dir. F1	Skel. F1	nSHD	Valid
0	0.368	0.235	0.317	0.647	50.0%	0.228	0.092	0.140	0.785	34.1%
10	0.533	0.305	0.406	0.483	68.2%	0.456	0.173	0.268	0.566	65.9%
20	0.521	0.313	0.417	0.494	65.9%	0.414	0.148	0.241	0.608	61.4%
30	0.664	0.399	0.539	0.353	81.8%	0.477	0.158	0.243	0.546	68.2%
40	0.720	0.399	0.561	0.301	90.9%	0.679	0.215	0.330	0.349	93.2%
50	0.746	0.441	0.606	0.275	93.2%	0.731	0.180	0.295	0.299	100%
60	0.752	0.442	0.594	0.268	93.2%	0.682	0.146	0.243	0.346	93.2%
70	0.780	0.448	0.605	0.241	97.7%	0.664	0.094	0.152	0.364	90.9%
80	0.731	0.444	0.600	0.289	90.9%	0.737	0.122	0.207	0.293	100%
90	0.722	0.419	0.579	0.299	90.9%	0.709	0.147	0.242	0.319	95.5%
100	0.773	0.448	0.623	0.250	97.7%	0.757	0.212	0.320	0.270	100%

Table 8: Endpoint changes from step 0 to 100 under strict invalid-as-zero scoring. *Improved* counts held-out systems with a favorable change.

Training/eval context	Metric	Step 0	Step 100	Δ	Improved
With context	Reward	0.368	0.773	$+0.405$	10/11
	Directed F1	0.235	0.448	$+0.213$	9/11
	Skeleton F1	0.317	0.623	$+0.306$	11/11
	nSHD	0.647	0.250	-0.398	10/11
	Valid output	50.0%	97.7%	$+47.7$ pp	8/11
Without context	Reward	0.228	0.757	$+0.529$	11/11
	Directed F1	0.092	0.212	$+0.120$	10/11
	Skeleton F1	0.140	0.320	$+0.180$	10/11
	nSHD	0.785	0.270	-0.514	11/11
	Valid output	34.1%	100%	$+65.9$ pp	11/11

Table 9: Per-system Directed-F1 and validity changes from step 0 to 100 for the independently trained policies, under strict invalid-as-zero scoring and ordered by node count.

System	N	Δ Directed F1		Valid$_0 \rightarrow {}_{100}$	
		Context	No context	Context	No context
BioModels 0197	5	$+0.121$	$+0.050$	100%→100%	75%→100%
SRIMANI2017	5	-0.053	$+0.415$	100%→100%	50%→100%
Weber_BMC2015	7	-0.030	$+0.049$	100%→100%	75%→100%
BioModels 0212	8	$+0.170$	$+0.168$	75%→100%	50%→100%
BioModels 0447	13	$+0.160$	$+0.116$	75%→100%	25%→100%
BioModels 0064	14	$+0.447$	$+0.315$	25%→100%	25%→100%
BioModels 0051	18	$+0.491$	$+0.103$	0%→100%	0%→100%
BioModels 0106	19	$+0.342$	-0.011	25%→100%	50%→100%
BioModels 0426	21	$+0.222$	$+0.048$	50%→100%	25%→100%
Lucarelli_CellSystems2018	33	$+0.368$	$+0.047$	0%→100%	0%→100%
KIMOSYS_MODEL_40	36	$+0.105$	$+0.025$	0%→75%	0%→100%

its Directed F1 is zero. Low all-pairs error can therefore coexist with poor edge recovery on sparse graphs.

C.3 Inference-Time Context Removal

Protocol. This experiment holds policy weights fixed when changing the semantic interface. It evaluates step 0 and step 100 of a context-trained policy that received 20 legacy-reward updates followed by 80 balanced-reward updates, with an optimizer reset at the boundary. These checkpoints are separate from the independently trained policies in Figure 8. All reported rewards use the same balanced evaluation objective: for a valid graph, $r_{\text{bal}} = 0.1 + 0.5F_{1,\text{dir}} + 0.2F_{1,\text{skel}} + 0.2(1 - \text{bSHD})$.

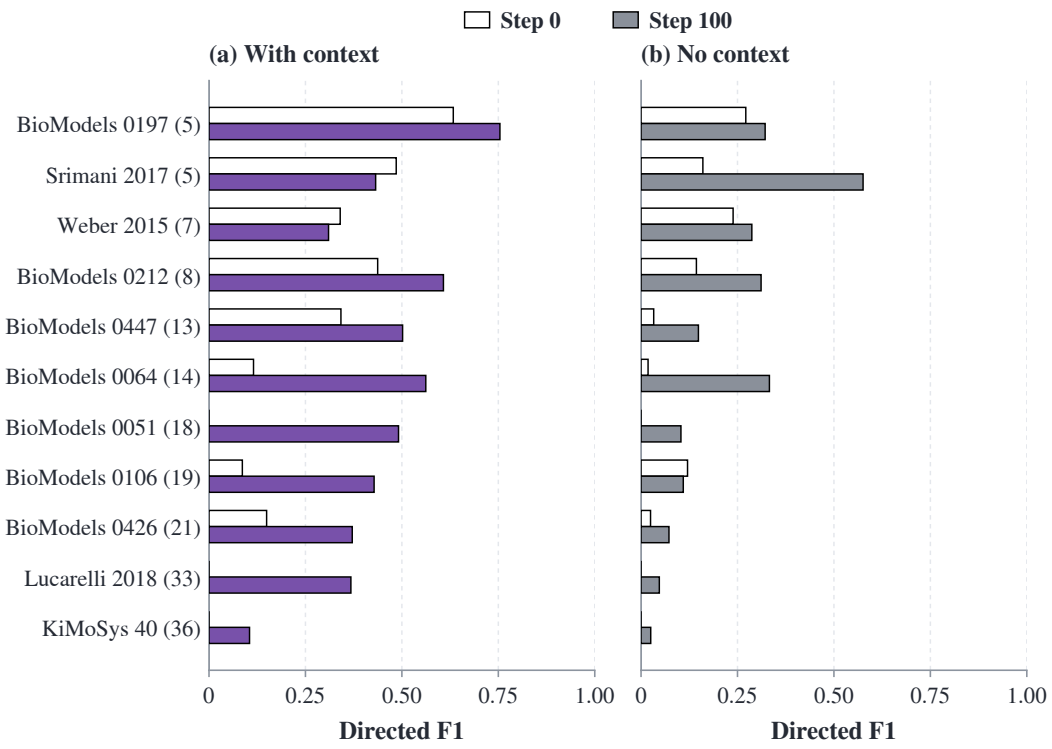

Figure 10: **RL outcomes across held-out systems.** Outlined and filled bars show Directed F1 at updates 0 and 100 for independently trained context and no-context policies. Each bar averages four rollouts, including invalid graphs as zero. Systems are ordered by variable count, shown in parentheses.

Here bSHD averages the false-negative and false-positive rates over nonempty reference classes of directed non-self pairs. Invalid submissions receive zero reward.

Context removal hides node identities, system descriptions, action names, and operation text; opaque aliases still execute the same experiments. Numerical feedback, tools, simulation budgets, verifier, and the 11 held-out systems remain fixed within this diagnostic. Each cell contains four rollouts per system, aggregated as above. Only the two endpoints were evaluated.

Table 10: Inference-time context removal for a separate context-trained policy using the 20+80 reward schedule. With/Without pairs use identical weights. Values are system-macro averages over 11 held-out systems under strict invalid-as-zero scoring; rewards use the balanced evaluation objective.

Policy	Ctx.	Reward	Dir. F1	Skel. F1	nSHD↓	Valid
Pre-GRPO	With	0.314	0.254	0.341	0.617	52.3%
Pre-GRPO	Without	0.189	0.128	0.200	0.745	40.9%
Final GRPO	With	0.587	0.479	0.624	0.250	95.5%
Final GRPO	Without	0.382	0.253	0.384	0.437	86.4%

Table 10 shows that removing context from the final policy reduces Directed F1 from 0.479 to 0.253 and Skeleton F1 from 0.624 to 0.384. Final weights still outperform initial weights without context (0.128 Directed F1), indicating partial transfer under context removal. At the final checkpoint, context improves Directed F1 on all 11 systems, reward and Skeleton F1 on 10, and nSHD on 9. This diagnostic measures reliance on context at inference; independent no-context learning is assessed by the main RL study.

The same split, defined by audited gradient exposure before inspecting endpoint scores, is used in both RL studies.

D Additional Analyses

The analyses use different cohorts and resource budgets. Each subsection specifies its protocol so that comparisons remain tied to the evidence supplied in that experiment.

D.1 Semantic Context Controls

Conditions and controlled quantities. The semantic analysis evaluates all 43 systems with three API trials per model, semantic condition, and discovery mode. Biological supplies the reviewed system and node descriptions. Deidentified removes the system title and explicit source identifiers while preserving biological identities and mechanistic content. Anonymous removes biological identities and mechanistic descriptions, retaining numerical observations and non-directional action–node context. Permuted reassigns node descriptions and their associated action references within a system, so the displayed semantics no longer correspond to the underlying variables.

The numerical simulator, frozen reference graph, available executable operations, observation format, and three-simulation ceiling remain unchanged. Observational episodes use untreated evidence; interactive episodes permit agent-selected experiments after the initial baseline. These conditions vary the semantic interface while preserving the simulated mechanism. Anonymous and Permuted differ in both content and correspondence to the system, and their prompts are not token-matched. Figure 11 compares observational and interactive recovery within each condition; the contrast between Anonymous and Permuted is not an alignment-only intervention at fixed information quantity.

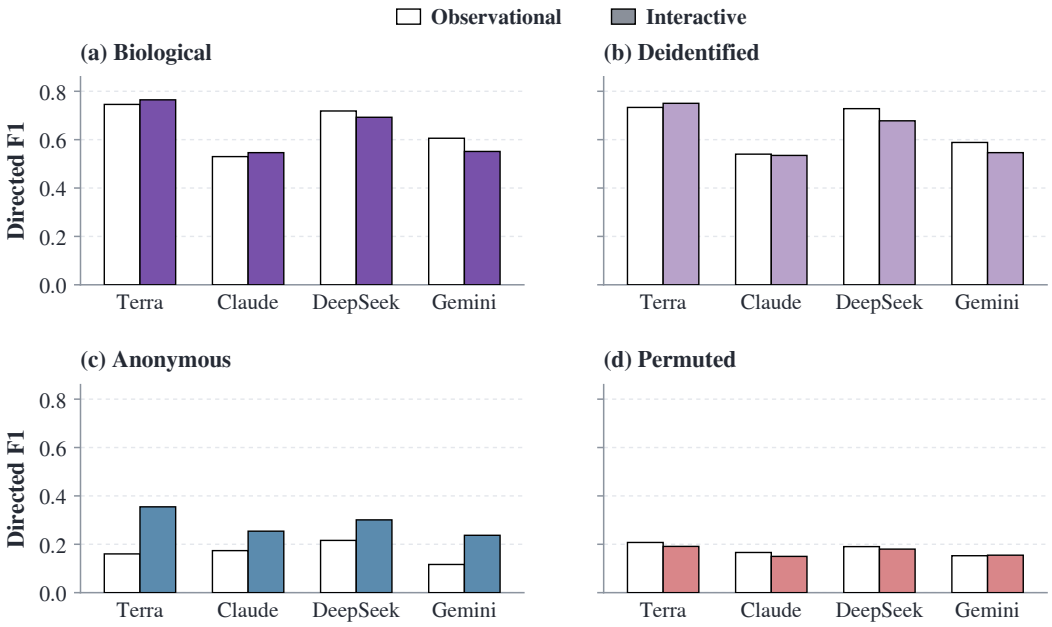

Figure 11: **Experimental interaction under four context conditions.** Outlined bars show observational discovery and filled bars show interactive discovery. Scores average three trials within each of 43 systems, then weight systems equally. All panels share the same vertical scale.

Missing and misleading context. Table 11 shows that interaction raises the four-model Anonymous mean from 0.167 to 0.287, approximately 72%, while the Permuted mean changes from 0.179 to 0.169. All four models improve with Anonymous context; with Permuted context, three decline and Gemini changes by only 0.002.

Removing explicit source identifiers changes Directed F1 by at most 0.018 across model–mode combinations. This limits an explanation based solely on those identifiers, while leaving recognition from the retained biological descriptions possible. Section D.1.1 examines the same distinction within one system.

Table 11: **Complete semantic-context results.** Directed F1 is averaged over trials within each system and then over all 43 systems. Δ denotes interventional minus observational F1.

Model	Context	Observational	Interactive	Δ
Terra	Biological	0.746	0.765	$+0.019$
	Deidentified	0.733	0.750	$+0.017$
	Anonymous	0.160	0.355	$+0.195$
	Permuted	0.208	0.191	-0.016
Claude	Biological	0.530	0.546	$+0.016$
	Deidentified	0.540	0.535	-0.005
	Anonymous	0.174	0.254	$+0.080$
	Permuted	0.166	0.150	-0.016
DeepSeek	Biological	0.719	0.693	-0.026
	Deidentified	0.728	0.678	-0.050
	Anonymous	0.216	0.301	$+0.085$
	Permuted	0.190	0.180	-0.010
Gemini	Biological	0.606	0.551	-0.055
	Deidentified	0.588	0.546	-0.042
	Anonymous	0.117	0.237	$+0.121$
	Permuted	0.153	0.155	$+0.002$

D.1.1 PC12 Signaling Case Study

We examine the PC12 EGF/NGF signaling system (BIOMD0000000033) (Brown et al., 2004) as a concrete instance of the semantic-context comparison. It contains 28 dynamic states and five permitted intervention axes, spanning growth-factor withdrawal and perturbations of the PI3K, Ras, and Rap1 branches. Ras–Raf1 and Rap1–B-Raf routes converge on MEK–ERK alongside PI3K–Akt regulation and ERK–p90RSK–SOS negative feedback. These coupled pathways illustrate why a downstream response need not establish a direct edge: effects can pass through intermediate components or feedback loops. All graph scores use the original 28-state reference.

Table 12: **Interaction gains for Terra on PC12 signaling.** Changes in Directed F1 from observational to agent-selected interventional discovery, averaged over the existing three trials.

Context	Δ Directed F1
Biological	0.000
Anonymous	$+0.117$
Permuted	-0.001

In Table 12, interaction raises Directed F1 by 0.117 under Anonymous context, while Biological and Permuted show almost no change. Recall also rises by 0.143 under Anonymous context. The scores summarize final graph recovery and do not attribute particular edge revisions to unobserved reasoning steps.

D.2 Interactive Budget Scaling

Cohort and paired ceilings. The sweep uses the prespecified systems JWS_Adlung_2017_mCFUE (43 variables), JWS_Fujita_2014 (44), and BioModels 0049 (94). GPT-5.6 Terra receives the same reviewed biological descriptions, executable actions, stateful interaction harness, and compact feedback at every setting. The seven budget pairs are $(K, Q) \in \{(1, 100), (2, 200), (4, 400), (6, 600), (10, 1000), (15, 1500), (20, 2000)\}$. Three rollouts per system and pair yield 63 episodes. Environment seeds are matched across settings within each replicate, and the measurement-noise setting remains 0.001. Scores exclude self-edges, average rollouts within each system, and then weight the three systems equally.

Two simultaneous resource constraints. Starting a fresh trajectory consumes one unit of K. Advancing either a fresh or the current live trajectory consumes its number of new observation time points from the shared Q budget. The agent selects 1–100 new points per operation and may continue a trajectory for up to 500 elapsed steps; trajectory length is not fixed at 100. Exhausting K prevents a new trajectory but allows the live one to continue while Q remains. The separately exposed initial boundary of each fresh trajectory is not charged to Q; continuation boundaries are cached and not remeasured. Thus, the number of unique exposed time points equals charged new points plus fresh-trajectory initial boundaries. The intervention ceiling is also K, with at most one intervention per trajectory and no combinations or replacements.

The agent receives at most eight endpoint-inclusive samples per returned segment plus statistics over all variables. Raw stored trajectories retain every acquired point, so charged observations and values displayed in the conversation are distinct quantities. Generation temperature, feedback, and inference allowances are fixed: temperature zero, at most 6,000 completion tokens per call and 128,000 per episode, with 6,000 reserved for final submission. The 2,000-operation, 2,200-turn, and one-hour guards remain unchanged. All episodes terminate by graph submission rather than token or wall-time exhaustion, and no outputs are repaired after submission.

Recovery across the full sweep. Table 13 reports all seven settings, and Figure 12 separates systems and metrics. The trend is non-monotone, including a rebound at $(15, 1500)$, but endpoint Directed F1 falls from 0.476 to 0.392 (approximately 18%). All three systems decline: by 0.079 on Adlung, 0.149 on Fujita, and 0.023 on BioModels 0049. Macro recall falls by 8.9 percentage points while precision rises by 1.7, indicating reduced coverage of true edges. All 63 submissions are valid, so strict and valid-only scores coincide.

Table 13: **Complete interactive budget-scaling results.** Every row averages three rollouts within each of the same three systems, then averages systems equally. Both budget ceilings apply to each episode; all nine graphs per row are valid.

K	Q	Dir. F1	Skel. F1	Precision	Recall	nSHD↓
1	100	0.476	0.537	0.831	0.376	0.040
2	200	0.443	0.502	0.874	0.336	0.041
4	400	0.442	0.493	0.861	0.334	0.041
6	600	0.427	0.484	0.873	0.315	0.041
10	1,000	0.368	0.443	0.800	0.258	0.045
15	1,500	0.460	0.517	0.814	0.366	0.041
20	2,000	0.392	0.474	0.848	0.287	0.044

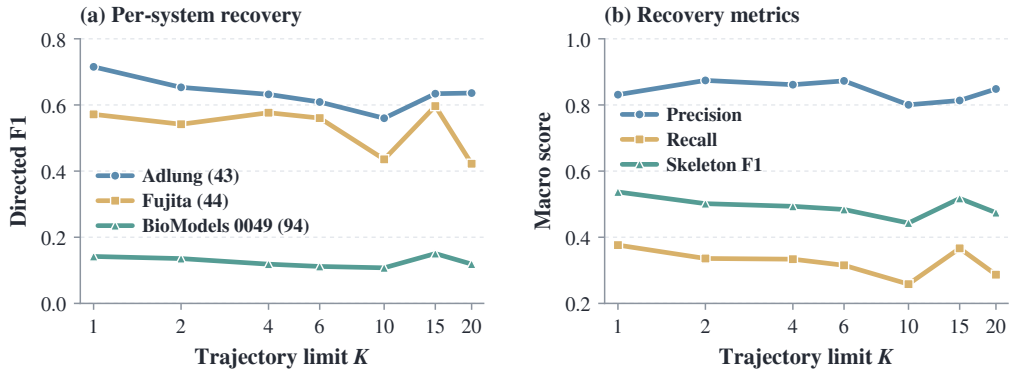

Figure 12: **Recovery under larger experimental allowances.** Left: Directed F1 for each system; parentheses give variable counts. Right: system-macro precision, recall, and Skeleton F1. Both ceilings increase together, and the logarithmic horizontal axis shows the seven measured settings.

Available versus acquired evidence. Table 14 distinguishes resource ceilings from consumption. Acquisition reaches 5.56 trajectories and 711 charged points at $(10, 1000)$ and then levels off. At

$(20, 2000)$, the means are 5.33 trajectories and 672 points, using 26.7% and 33.6% of the allowances; all nine episodes submit with both budgets remaining. Despite acquiring more evidence than at $(1, 100)$, the agent recovers fewer edges correctly. The extra evidence acquired in these episodes did not improve recovery. The scores and budget counters do not explain why the agent stopped or predict the effect of forcing further acquisition.

Table 14: **Actual experimental allocation.** Counts are rollout means with systems weighted equally. New points are charged against the episode-wide Q; initial boundaries are separate. Treated starts count fresh trajectories begun with an intervention, and continuations count extensions of the live trajectory.

K	Fresh trajectories	New points	K used (%)	Q used (%)	Treated starts	Continuations
1	1.00	100.0	100.0	100.0	0.00	0.00
2	2.00	200.0	100.0	100.0	1.00	0.00
4	3.56	377.8	88.9	94.4	2.56	0.22
6	4.56	533.3	75.9	88.9	3.56	0.78
10	5.56	711.1	55.6	71.1	4.00	1.56
15	5.33	682.2	35.6	45.5	4.00	1.67
20	5.33	672.2	26.7	33.6	4.00	1.44

At the smallest setting, every trajectory starts untreated, but five of the nine episodes intervene later within the live trajectory. Consequently, $(1, 100)$ is not an observational-only reference. Budget counters and terminal reasons were replayed from all 63 action logs, and structural metrics were recomputed from the submitted graphs over ordered non-self pairs. The completed export contains all 21 system–setting groups and no missing or infrastructure-pending episodes.

Scope of the comparison. The sweep studies one model on three large systems while increasing both ceilings together. It therefore does not separate trajectory count from observation density. The Active–Random comparison in Section D.4 instead holds total evidence fixed. The earlier resource-expansion comparison in Section D.3 additionally changes the harness and feedback, so its endpoints are analyzed separately.

D.3 System Scale and Resource Expansion

Fixed-budget comparison. The main scale figure divides the 43 systems into seven contiguous groups by the number of integrated dynamic variables: 5–6 (6 systems), 7–10 (7), 11–15 (6), 18–23 (6), 24–27 (7), 28–36 (6), and 41–94 (5). Boundaries depend on node count alone and do not split equal-sized systems. All four agents retain biological context and the standard three-trajectory ceiling. We average each metric over three trials within a system and then take each group's unweighted mean. In Figure 7, faint points retain individual system means, while the connected group means are positioned at their arithmetic mean node counts. The lines summarize these groups; they are not fitted scaling curves.

For the numerical small-to-large comparisons in the text and the precision–recall decomposition in Figure 13, we retain the original three strata: small (5–11 variables; 14 systems), medium (13–25; 15 systems), and large (26–94; 14 systems). This preserves the endpoint comparison while the main figure shows finer variation within the size range. The large-minus-small Directed F1 difference ranges from -0.137 to -0.338 in observational mode and from -0.173 to -0.363 in interactive mode. Relative to the small-system strata, interactive Directed F1 falls by 24–50% across the four agents. Group markers are placed at their mean variable count on a logarithmic horizontal axis.

Incomplete structural coverage. Figure 13 decomposes the small-to-large decline into precision and recall. Recall decreases by 12.2–35.7 percentage points in observational discovery and 16.8–36.5 points in interactive discovery. The corresponding precision losses are 1.1–16.2 and 2.7–12.1 points. Recall declines more than precision for all eight model–mode combinations. Large-system precision remains between 0.718 and 0.885, while recall ranges from 0.277 to 0.601, indicating that the reported edges cover a smaller fraction of the reference graph.

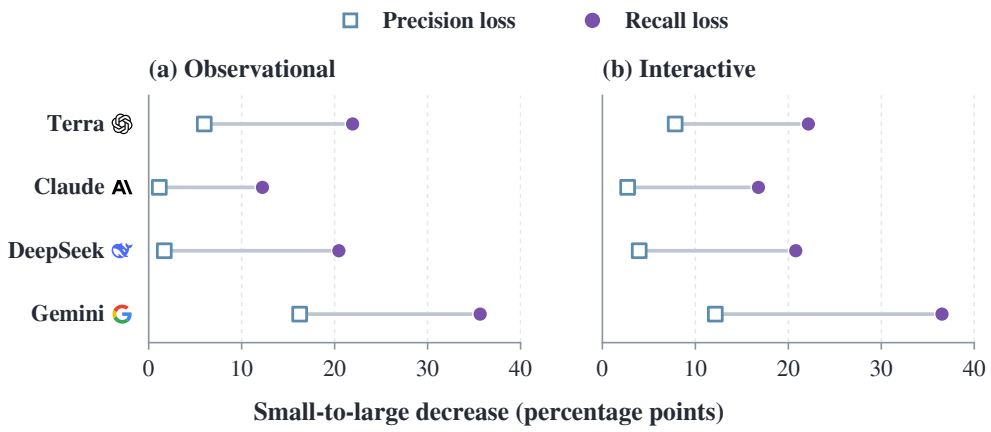

Figure 13: **Larger systems lose more recall than precision.** Each marker shows the small-group mean minus the large-group mean, in percentage points, with three trials averaged within each system and systems weighted equally. The connector pairs the precision and recall losses for the same agent and mode.

The number of candidate directed non-self pairs grows as $N(N-1)$, but size is not independently manipulated. The strata also differ in topology, density, and biological domain. Their performance differences characterize the evaluated collection and do not establish a causal scaling law. The plotting script and companion trial-level CSV reproduce both figures from the frozen evaluation records; the provenance file identifies the source commit and metric fields.

Separate resource-expansion comparison. This earlier comparison changes the harness and feedback as well as resources. Its two endpoints are separate from the controlled budget sweep in Section D.2. The diagnostic selects the three systems with at least 40 dynamic variables and at least three reviewed online-capable interventions. GPT-5.6 Terra receives 20 trajectories of 2,000 points each: two untreated baselines and 18 interventions distributed across reviewed actions and intervention times. Three native Codex rollouts per system are compared with the historical three-trajectory, 12-point regime, also with three rollouts per system. Here 3×12 denotes the displayed baseline panel, whereas the expanded regime acquires 2,000 points per trajectory. Intervention coverage, temporal resolution, and inference resources change together.

Table 15: **Resource expansion under strict and valid-only scoring.** Macro averages over the same three systems. Strict F1 retains every rollout and assigns zero to invalid outputs; valid-only metrics average accepted graphs within each system before averaging systems.

Evidence regime	Valid	Strict Dir. F1	Valid-only Dir. F1	Valid-only Precision	Valid-only Recall
3×12	9/9	0.688	0.688	0.797	0.626
$20 \times 2{,}000$	8/9	0.686	0.761	0.805	0.728

Table 16: **Per-system resource-expansion outcomes.** Baseline outputs are all valid. Expanded-resource validity and both F1 endpoints are shown without repairing or replacing the invalid rollout.

System	N	Baseline Dir. F1	Expanded Valid	Expanded Strict F1	Expanded Valid-only F1
BioModels 0049	94	0.464	2/3	0.448	0.672
JWS_Adlung_2017_mCFUE	43	0.837	3/3	0.871	0.871
JWS_Fujita_2014	44	0.763	3/3	0.740	0.740

Structural gains and execution failures. In Table 15, valid-only recall increases by 0.102 and precision by 0.007; nSHD decreases from 0.030 to 0.027. Table 16 locates the largest valid-only gain in the 94-variable system, where Directed F1 increases by 0.208. However, one of its expanded-resource rollouts contains a duplicate edge and fails validation, reducing that system's strict F1 below baseline. Across all three systems, strict Directed F1 changes by -0.002 despite the positive valid-only result.

The remaining systems also differ. The 43-variable system improves by 0.034 Directed F1. On the 44-variable system, recall increases slightly while precision falls, reducing F1 by 0.023. Richer resources therefore do not uniformly improve graph selectivity or reliable completion. The three-system cohort and simultaneous changes to evidence and inference prevent attributing these outcomes to any one resource.

D.4 Active Selection at a Fixed Evidence Budget

Cohort and evidence allocation. The fixed-cost analysis includes BioModels systems 0032, 0033, 0051, 0064, 0171, 0211, 0243, and 0247, together with Bachmann_MSB2011, Isensee_JCB2018, and JWS_Adlung_2017_mCFUE. Each has at least five documented actions. For $B \in \{0, 1, 2, 3, 5\}$, each episode contains six trajectories with 101 generated time points each: B treated and $6 - B$ untreated. All conditions retain the correct biological descriptions and the same compact feedback format. Three trials are run per system and condition. At $B = 0$, the two policies share one all-untreated reference.

Active selection chooses distinct actions sequentially using available feedback; Random samples distinct actions without replacement. Both use the declared intervention level. Increasing B replaces untreated trajectories rather than increasing the total amount of simulation evidence. For systems with exactly five actions, $B = 5$ exhausts the same action set under both policies, although ordering may differ. Accordingly, a nonzero endpoint difference can reflect how a model processes or orders the evidence as well as which actions it selects.

Table 17: **Complete fixed-cost intervention-budget results.** Directed F1 is a macro average over 11 systems after averaging three trials. Each row uses six total trajectories. Δ denotes Active minus Random.

Model	B	Random	Active	Δ
Terra	0	0.810 (shared)		—
	1	0.773	0.813	+0.040
	2	0.785	0.815	+0.030
	3	0.816	0.815	−0.001
	5	0.826	0.810	−0.016
Claude	0	0.607 (shared)		—
	1	0.625	0.585	−0.039
	2	0.583	0.579	−0.005
	3	0.574	0.570	−0.004
	5	0.584	0.567	−0.017
DeepSeek	0	0.726 (shared)		—
	1	0.707	0.723	+0.016
	2	0.695	0.733	+0.039
	3	0.720	0.708	−0.012
	5	0.723	0.690	−0.033
Gemini	0	0.575 (shared)		—
	1	0.584	0.523	−0.062
	2	0.535	0.552	+0.017
	3	0.549	0.589	+0.041
	5	0.591	0.603	+0.013

Table 17 and Figure 14 show positive acquisition gains in 7 of the 16 model–allocation pairs. Terra and DeepSeek lose their initial gains as B increases, while Gemini benefits at larger allocations and Claude remains below Random. The tested agents therefore show no shared budget threshold beyond which active acquisition stops helping.

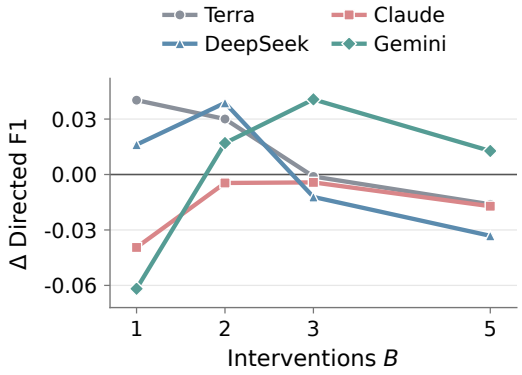

Figure 14: **Selection gains vary across agents at a fixed total budget.** Active minus Random Directed F1, averaged over 11 systems and three trials per condition. Each episode contains six trajectories; increasing B replaces untreated trajectories with treated ones.

Two distinct reference points. Active minus Random measures the gain from adaptive acquisition at a fixed intervention allocation; Active minus $B = 0$ measures the gain from replacing untreated evidence. For Terra at $B = 1$, the former is 0.040 while the latter is only about 0.004, and its active score is nearly flat across allocations. Gemini at $B = 5$ exceeds both references. Thus, an advantage over Random alone does not establish an improvement over untreated evidence, and these agents show no common intervention threshold for acquisition gains.

D.5 Conditional Response Propagation

Prediction task. We use fixed graphs recovered under biological semantics to evaluate documented actions that were not observed during the corresponding discovery episode. The evaluator observes the first 10% of each noise-free test response, excluding its initial point, identifies early responding nodes, and predicts later responding nodes by graph reachability. The predicted set includes both these initial responders and all their descendants. This is a conditional propagation test using early response observations; it is not a prediction made before any response is observed.

Affected-node definition. For each node i, let $x_i^0(t)$ be the untreated trajectory and $x_i^a(t)$ the test-action trajectory. Responses are normalized as

$$e_i(t) = \frac{|x_i^a(t) - x_i^0(t)|}{s_i}, \qquad s_i = \max\{\mathrm{range}(x_i^0),\ 0.1\,\mathrm{RMS}(x_i^0)\}\,. \tag{6}$$

For numerical stability, scales are floored at 10^{-6} times the median scale above 10^{-12}; the median is set to one if no scale exceeds this cutoff. With T sampled points, the early score is the maximum $e_i(t)$ over indices 1 through $\max(2, \lceil 0.1T \rceil) - 1$, and the late score is the maximum over the second half of the trajectory. Each window uses a threshold equal to the larger of 0.05 and one quarter of that window's largest node score. If a thresholded set is empty, it includes the highest-scoring node. Binary F1 compares the reachability prediction with the late affected-node set.

Table 18: **Conditional response propagation on unobserved actions.** Rec.–FC and Rec.–Ref. are recovered-graph response F1 minus fully connected and reference-graph response F1. Int.–Obs. is the matched interventional minus observational difference. ρ is the Spearman association between structural and response F1 across eligible episodes.

Model	Rec.–FC	Rec.–Ref.	Int.–Obs.	ρ
Terra	+0.062	+0.043	+0.013	0.275
Claude	+0.130	+0.109	−0.035	−0.014
DeepSeek	+0.084	+0.064	+0.012	0.127
Gemini	+0.123	+0.101	+0.011	0.094

Aggregation and interpretation. For each model, 219 of the 258 discovery episodes (43 systems, two modes, three trials) retain at least one unobserved action, yielding 750 action-level tests. Graph-support comparisons weight all 43 systems equally. Paired interactive–observational comparisons use 30 systems with eligible actions, while rank associations use the 219 episode-level means.

Table 18 shows that recovered graphs outperform fully connected and reference-graph reachability on this thresholded endpoint. A structural path represents a potential dependency, whereas the endpoint requires a sufficiently large response within a finite window. A sparser recovered graph can therefore improve response selectivity while missing true edges. The weak, model-dependent association with Directed F1 and mixed interaction gains support reporting this test separately from structural recovery.

E Extended Related Work and Environment Positioning

We expand Section 2 by comparing prior knowledge, available experiments, and recovery targets across scientific discovery environments, biological resources, and causal inference methods.

E.1 Interactive Scientific and Mechanistic Discovery

ScienceWorld and DiscoveryWorld evaluate scientific reasoning through actions, observations, and explanations in interactive environments (Wang et al., 2022; Jansen et al., 2024). DiscoveryBench and ScienceAgentBench organize discovery around supplied datasets and research tasks (Majumder et al., 2025; Chen et al., 2025b), while AstaBench standardizes broader research workflows and tool access (Bragg et al., 2026). BoxingGym, MaD Physics, and DiscoverPhysics further make experimental selection and resource allocation part of discovery (Gandhi et al., 2025; Jain et al., 2026; Wiemann et al., 2026). For BioCausalX, interactivity specifically means choosing new simulated experiments whose outcomes can inform a structural hypothesis.

The closest biological predecessor is SciGym, which uses published BioModels systems for reaction and kinetic-model reconstruction. Its supplementary experiments also include knockout access and changes to the domain context supplied with the same mathematical system (Duan et al., 2025). These precedents combine biological dynamics, interaction, and context controls. BioCausalX links source-supported entity descriptions and permitted operations to hidden execution mappings and a common direct-dependency target. Reaction-model recovery additionally involves reaction structure and kinetic forms (Table 19), so success on our graph target does not imply reconstructing an equivalent kinetic model.

CausaLab separates predictive success from graph and functional-mechanism recovery in synthetic structural causal models with restricted manipulators (Yang et al., 2026). ActiveSciBench systematically varies enzyme-kinetic tasks and regulatory motifs within a closed discovery loop (Kabra et al., 2026). These constructions systematically vary mechanisms; our heterogeneous published systems emphasize source-specific biological context and operations. MDA combines LLM-generated hypotheses with Bayesian inference and experimental design across several domains, including neuronal dynamics (Murphy, 2026). It supplies a discovery framework that addresses uncertainty over models, while BioCausalX supplies tasks on which different acquisition and inference policies can be evaluated.

E.2 Biological Resources and Benchmark Construction

BioModels supports sharing, curation, and reuse of computational biological models (Malik-Sheriff et al., 2020); JWS Online provides reproducible, modifiable simulation studies (Peters et al., 2017). MIRIAM connects model annotation to reference descriptions and biological entities (Le Novère et al., 2005), and PEtab specifies mappings among models, experimental conditions, measurements, and parameters (Schmiester et al., 2021). A discovery environment builds on this provenance and reproducible execution by specifying the disclosed information, permitted operations, their effects on the model, and the hidden evaluation target.

Biological network benchmarks also combine several forms of grounding. GeneNetWeaver generates dynamical data using structures informed by biological regulatory networks (Schaffter et al., 2011); DREAM5 evaluates inference across simulated and organism-derived datasets (Marbach et al., 2012).

Table 19: Closest mechanistic discovery settings. Different recovery targets preclude direct comparison of their reported scores.

Environment	Mechanisms and interaction	Recovery target
SciGym	BioModels systems; simulated concentration perturbations	Reactions, kinetic models, trajectories
CausaLab	Synthetic SCMs; restricted manipulators	Predictions, graph, functional mechanisms
ActiveSciBench	Enzyme kinetics and regulatory motifs; sequential experiment queries	Kinetic hypotheses and regulatory structure
MDA	Multiple discovery domains; Bayesian experimental design	Model hypotheses, parameters, predictions
BIOCAUSALX	Published biological systems; source-supported operations with execution mappings	Direct dynamical dependencies among declared states

SERGIO models single-cell expression and measurement-related variation from specified regulatory networks (Dibaeinia & Sinha, 2020), while BEELINE includes synthetic, literature-curated Boolean, and experimental single-cell sources (Pratapa et al., 2020). Topology, dynamics, observations, and reference interactions can thus have different origins within one benchmark. Describing these origins is more informative than a single synthetic-or-real label.

CausalBench uses large-scale single-cell perturbation data, biological reference information, and distribution-based interventional evaluation to address the absence of complete real-world ground truth (Chevalley et al., 2025). BIOCAUSALX takes a complementary route: the governing equations provide an exact structural reference for the implemented model. Its 43 systems and 168 intervention axes connect that reference to reviewed descriptions and documented operations. This permits reproducible scoring, while retaining a specific scope: source support for an operation does not validate every simulated strength or trajectory, and the model's graph does not exhaust the biology of the corresponding organism.

E.3 BIOLOGICAL SEMANTICS AND PRIOR–ENVIRONMENT ALIGNMENT

PhysGym varies the prior information disclosed to agents (Chen et al., 2025a), NewtonBench changes governing laws so that familiar physical knowledge becomes locally inappropriate (Zheng et al., 2026), and SciGym varies domain context while retaining the underlying dynamics (Duan et al., 2025). These settings distinguish prior information quantity, validity, and semantic framing. In biology, even recognizable, scientifically plausible descriptions can refer to the wrong entity or mechanism.

Gilda addresses biomedical entity normalization through contextual disambiguation, reflecting the fact that a name can refer to multiple entities (Gyori et al., 2022). AMAS documents limited annotation coverage in biochemical models and notes that existing annotations can contain typographical errors or specifications inconsistent with reference databases (Shin et al., 2023). Complementary empirical evidence comes from Schnoes et al. (2009), who audited molecular-function annotations for 37 enzyme families across four databases and found substantial errors in some database–family combinations. These studies support distinct concerns about ambiguity, completeness, and correctness. The historical enzyme-family audit is not an estimate of current error rates across all biological text, but it demonstrates why supplying an annotation need not improve the information available for discovery.

These concerns motivate our separate treatment of missing and misleading biological context. Anonymization removes biological identities and mechanistic descriptions, whereas permutation reassigns node descriptions and associated action references while preserving the executable system. Missing context limits prior guidance; misleading context supplies guidance inconsistent with the mechanism being investigated. In our interactive evaluations, the latter can be more damaging, and experimental feedback does not reliably remove this disadvantage (Section 5.1). These stress tests differ in information content and wording; they neither model natural annotation-error frequencies nor match token counts.

BioDiscoveryAgent, PerturbQA, and PertReason further demonstrate the value of biological semantics for phenotype-directed experiment selection, perturbation-outcome reasoning, and context-specific explanations, respectively (Roohani et al., 2025; Wu et al., 2025; Kim et al., 2026). BIOCAUSALX complements these objectives by evaluating whether descriptions and newly acquired evidence support recovery of the particular system's direct dependencies. A useful perturbation or a plausible pathway explanation need not identify the complete target graph.

E.4 LANGUAGE-GUIDED AND ACTIVE CAUSAL DISCOVERY

Language models can propose causal relationships from semantic knowledge (Kıcıman et al., 2024). Causal Modelling Agents combine such metadata with statistical evidence (Abdulaal et al., 2024); COAT uses causal feedback to refine high-level variables extracted from unstructured information (Liu et al., 2024); and CausalOrder supplies language-derived ordering constraints (Vashishtha et al., 2025). These approaches address different stages of discovery. Our variables are predefined, so the task evaluates relationships among declared states rather than discovery of the variables themselves. Its cyclic graph also differs from targets constrained by a global topological ordering.

Active causal discovery selects interventions to reduce uncertainty about structure or a causal query (Hauser & Bühlmann, 2014; Toth et al., 2022; Tigas et al., 2022). GIT uses gradient information for intervention targeting (Olko et al., 2023), and LeGIT incorporates LLM knowledge into this selection process (Li et al., 2025). BIOCAUSALX provides a biologically documented action space for studying this established combination of semantic priors and numerical acquisition. Available operations may affect a node mechanism, reaction, interaction, or exogenous input; their execution mappings determine the actual perturbation. An idealized intervention on an arbitrary graph node is consequently not interchangeable with every permitted biological operation.

The feedback channel also changes the inference problem. IGDA uses budgeted edge queries with binary relationship feedback (Havrilla et al., 2025), whereas BIOCAUSALX returns trajectories that must be interpreted to distinguish direct influence from downstream propagation. RealTCD and CauScientist respectively couple domain knowledge to temporal inference and statistically verified graph refinement (Li et al., 2024; Peng et al., 2026). Such evidence integration is distinct from selecting the next experiment. Our semantic controls and Active–Random comparisons examine context reliability and acquisition behavior under a shared action catalog.

E.5 DYNAMICAL TARGETS AND EVALUATION SCOPE

SINDy recovers sparse dynamical equations from candidate functions (Brunton et al., 2016), and neural graphical modeling in continuous time relates regularized vector-field estimation to local dependency recovery (Bellot et al., 2022). DYNOTEARS and PCMCI+ instead represent contemporaneous and lagged relationships at the observation timescale (Pamfil et al., 2020; Runge, 2020). These targets carry different information. For $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, our reference records $j \to i$ when f_i depends nontrivially on x_j, excluding self-edges. It retains feedback cycles but omits signs, parameters, and kinetic forms. A variable's response to an intervention can arise through an indirect path, while a direct dependency may remain difficult to detect along a trajectory confined to a narrow regime.

CausalTime produces graph-labeled time series through generative models fitted to observations (Cheng et al., 2024), and CausalDynamics uses coupled dynamical systems for structural evaluation (Herdeanu et al., 2025). BIOCAUSALX builds on the CausalDynamics/`causalx` codebase, extending this foundation with biological descriptions, grounded operations, and an agent interface. CausalArena likewise emphasizes diverse synthetic, semantic, and formula-grounded mechanisms and the importance of pretraining–evaluation overlap (Li et al., 2026). Our contribution connects published biological mechanisms, source-supported context, executable actions, and a common recovery target.

Evaluation interfaces limit these comparisons. Numerical methods receive full trajectories while LLMs receive compressed observations, so equal simulation budgets do not imply identical inputs. Recovery with biological context can combine general knowledge, source-specific recall, and experimental inference. Anonymization and held-out-system training do not rule out pretraining overlap. Throughout, the target is the frozen model's structure, not the complete biology of the corresponding system.