Orchestrating Causal Discovery at Scale: An Agentic Framework Integrating Reasoning, Differentiable Structure Learning, and Active Experimentation

Author names withheld

Editor: Under Review for CLeaR 2026

Abstract

Closed-loop experimentation in biology faces a persistent mismatch: community-scale perturbation platforms can generate high-dimensional datasets, but any specific discovery program is constrained by a small budget of follow-up interventions. A practical causal-discovery system must therefore (i) reason auditable hypotheses over heterogeneous evidence, (ii) learn causal structure in high dimensions under realistic constraints, and (iii) select informative experiments under cost and noise. We present a full-stack agentic framework validated on differentiable simulations, a STRING-derived protein association network (2,724 proteins, 5,000 edges), and a retrospective evaluation on DepMap-inspired semi-synthetic gene-perturbation data.

Our contribution is a systems-level integration of three components—**(1) a structured reasoning state** (Causal Graph-of-Thoughts, C-GoT) that makes hypotheses, evidence, and decisions explicit; **(2) differentiable causal structure learning** that scales with sparse biological priors; and **(3) Bayesian optimal experimental design** for selecting interventions that maximally reduce uncertainty in query-relevant causal quantities. We report feasibility studies: (i) differentiable simulation benchmarks stress-testing directionality recovery (90% on coupled oscillators vs. 3% cross-correlation baseline); (ii) active intervention selection on synthetic SCMs at two scales (d=15 data-rich; d=50 data-scarce), where BOED with informative priors yields statistically significant improvements over random selection in the data-scarce regime; (iii) a STRING-scale (2,724 proteins, 5,000 edges) (Szklarczyk et al., 2019) causal target-prioritization workload with enrichment (p=0.038, δ_p=0.93) and circularity controls; and (iv) a retrospective DepMap evaluation demonstrating the framework on real biological gene-perturbation data.

Keywords: active causal discovery; Bayesian optimal experimental design; agentic science; perturbation biology; auditability; causal graph of thoughts; query-oriented inference; proxy validation

1. Introduction

Scientific discovery is increasingly mediated by automation. Self-driving laboratory systems demonstrate that iterative experiment selection and execution is feasible when coupled to robotics and modern machine learning (King et al., 2009; MacLeod et al., 2020; Burger et al., 2020; Boiko et al., 2023). In functional genomics, pooled CRISPR perturbations and single-cell readouts yield massive interventional datasets (Gilbert et al., 2014; Dixit et al., 2016; Replogle et al., 2022). The key bottleneck is no longer measurement throughput, but *decision throughput*: which interventions should be run next, in which contexts, and with what validation burden.

This bottleneck is especially acute for causal discovery. A purely predictive pipeline can improve with more data from the same distribution, but causal claims require explicit intervention semantics and careful treatment of confounding (Pearl, 2009; Hernán and Robins, 2020; Imbens and Rubin, 2015). In high-dimensional biology, discovery must also contend with (i) sparse and noisy interventions (variable perturbation efficiency, off-target effects), (ii) latent state and batch effects, and (iii) the combinatorial action space of single and multi-target perturbations.

A large literature studies (a) causal structure learning (Spirtes et al., 2000; Koller and Friedman, 2009; Zheng et al., 2018; Yu et al., 2019), and (b) active learning / experimental design for causal models (Tigas et al., 2022, 2023; Annadani et al., 2024; Zhang et al., 2023). However, many algorithms are presented as stand-alone estimators, with limited attention to the operational reality of closed-loop discovery: maintaining an auditable belief state, integrating heterogeneous priors, logging decisions, and validating results without direct ground truth.

Strategic pivot and paper scope. This paper intentionally adopts a *system/framework* posture. Rather than introducing a new causal identification theorem that would require fresh biological transportability results, we assemble an implementable agentic pipeline around three bottlenecks that repeatedly block deployment:

1. **Reasoning and auditability:** hypotheses and rationales must be explicit objects, not free-form text.

2. **Scalability:** causal discovery must operate over sparse priors and large graphs without collapsing to intractable combinatorics.

3. **Validation under limited ground truth:** the system must provide statistically defensible proxy validation signals at scale.

Contributions. **C1 (Agentic state representation).** We define C-GoT, a structured causal belief state whose nodes include variables, hypotheses, evidence, interventions, and decisions. **C2 (Scalable structure learning backbone).** We integrate differentiable structure learning (Zheng et al., 2018; Yu et al., 2019) with sparse biological priors and posterior uncertainty. **C3 (Active experimentation policy).** We use Bayesian optimal experimental design (BOED) to select interventions by expected information gain (Tigas et al., 2022, 2023; Garnett, 2023). **C4 (Feasibility evidence).** We provide differentiable-simulation benchmarks, STRING-scale proxy validation with circularity controls, and a retrospective evaluation on DepMap gene-perturbation data demonstrating operational feasibility on real biological data.

2. Related work

Causal discovery. Constraint-based and score-based methods recover equivalence classes under causal sufficiency and faithfulness (Spirtes et al., 2000; Koller and Friedman, 2009); differentiable relaxations (NOTEARS, neural DAG learning) cast structure learning as continuous optimization (Zheng et al., 2018; Yu et al., 2019). Causal representation learning seeks robustness and transfer (Schölkopf et al., 2021; Khemakhem et al., 2020). **Active causal learning.** Modern BOED approaches scale intervention selection via variational inference and differentiable acquisition (Tigas et al., 2022, 2023; Annadani et al., 2024). Li et al. (2025) propose LLM-guided intervention targets (LeGIT); their approach uses LLM suggestions to replace or augment BOED scoring, whereas our framework uses structured reasoning (C-GoT) to maintain an auditable belief state with explicit evidence-to-decision links, without relying on LLM-generated hypotheses. **Perturbation biology.** Perturb-seq provides single-cell interventional readouts (Dixit et al., 2016; Replogle et al., 2022); DepMap and Connectivity Map provide multi-context perturbation resources (Tsherniak et al., 2017; Subramanian et al., 2017; Broad Institute, 2025b,a). Self-driving labs motivate closed-loop automation (King et al., 2009; MacLeod et al., 2020; Burger et al., 2020; Boiko et al., 2023).

3. Framework: ACHT as an agentic causal discovery stack

We describe a closed-loop agent that repeatedly (i) proposes causal hypotheses, (ii) selects an experiment, (iii) observes outcomes, and (iv) updates its belief state. The framework is intended for perturbation biology but is abstract enough to cover other self-driving laboratory settings.

3.1. Objects and notation

Let measured variables be $\mathcal{V} = \{V_1, \ldots, V_d\}$ and let $\mathcal{I}$ denote feasible interventions. A dataset is a sequence of rounds $\mathcal{D}_t = \{(a_\tau, y_\tau)\}_{\tau=1}^{t}$, where a_t is the chosen action (e.g., environment context and intervention) and y_t is the observed outcome (possibly a high-dimensional summary). A hypothesis $H \in \mathcal{H}$ consists of (i) a causal graph G, (ii) a parameterization of mechanisms, and optionally (iii) nuisance parameters capturing measurement noise and intervention imperfections. We represent intervention semantics via the do-operator (Pearl, 2009).

The goal is rarely to recover the entire graph. Instead, we define a *causal query* Q—the average causal effect (ACE) of each candidate intervention on designated phenotype nodes:

$$Q = \frac{1}{|T|} \sum_{t \in T} \mathrm{ACE}(\mathrm{do}(V_t), Y), \tag{1}$$

where T is the target set and Y is the phenotype variable. We select interventions to reduce uncertainty in Q.

3.2. C-GoT: Causal Graph-of-Thoughts as a belief state

A common failure mode of agentic scientific systems is that hypotheses and rationales live in unstructured text, creating auditing and reproducibility problems: it is unclear which evidence supported which decision, and which assumptions were active when an intervention was chosen.

We therefore define the agent's internal state at round t as a typed graph $\mathcal{G}_t$ whose nodes include:

- **Variable nodes:** measured biological variables and engineered representations (e.g., gene programs).
- **Hypothesis nodes:** candidate causal graphs/SCMs with posterior weights.
- **Intervention nodes:** feasible single and combinatorial perturbations.
- **Evidence nodes:** datasets, statistical tests, fitted model diagnostics.
- **Decision nodes:** actions taken and the justification links to the evidence/hypotheses that motivated them.

Edges in $\mathcal{G}_t$ capture support and dependency: which evidence updates which hypothesis; which interventions discriminate between competing hypotheses; and which decisions were made under which uncertainty. This structure yields reproducible "decision cards" per intervention, selective updating, and governance gating.

3.3. Discovery module: differentiable structure learning with priors

The discovery module proposes and updates a distribution over causal structures. In high dimensions, we require scalable proposals. We use a differentiable optimization backbone inspired by NOTEARS (Zheng et al., 2018). Let $W \in \mathbb{R}^{d\times d}$ be a weighted adjacency matrix. A generic formulation is

$$\min_{W} \; \mathcal{L}(W; \mathcal{D}_t) + \lambda \|W\|_1 \quad \text{s.t.} \quad h(W) = 0, \tag{2}$$

where $\mathcal{L}$ is a data-fit loss and $h(W) = 0$ enforces acyclicity. When biological feedback is expected, the acyclicity constraint can be relaxed and replaced by cyclic/dynamical backends (Bongers et al., 2016; Rohbeck et al., 2024; Sethuraman et al., 2023; Runge et al., 2019).

Biological priors. Biological knowledge (e.g., PPI or TF–target priors) is incorporated as a structured regularizer. Let P be a prior adjacency score (higher means more plausible). We penalize edges inversely proportional to prior support:

$$\mathcal{R}(W; P) = \sum_{i\neq j} \alpha_{ij} |W_{ij}|, \qquad \alpha_{ij} = \alpha_0 \cdot (1 + \kappa \cdot \mathbb{1}[P_{ij} = 0]). \tag{3}$$

This encourages discovery to remain locally consistent with trusted biological knowledge while allowing data-driven corrections.

Uncertainty representation. Rather than committing to a single graph, we maintain a particle set $\{H^{(m)}\}_{m=1}^{M}$ with weights $w_m \propto p(H^{(m)} \mid \mathcal{D}_t)$. Particles can be generated by random restarts of Eq. 2, structure bootstrapping, or hybrid proposals that mix constraint-based and differentiable estimates (Spirtes et al., 2000; Zheng et al., 2018; Yu et al., 2019).

3.4. Experimentation module: BOED for intervention selection

Given a belief state, the agent must choose an intervention. We use expected information gain (EIG) on a query variable Q as the utility. For a candidate action $a \in \mathcal{I}$,

$$\mathrm{EIG}(a; \mathcal{D}_t) = \mathbb{H}[Q \mid \mathcal{D}_t] - \mathbb{E}_{y\sim p(y|a,\mathcal{D}_t)}\big[\mathbb{H}[Q \mid \mathcal{D}_t \cup \{(a, y)\}]\big]. \tag{4}$$

This emphasizes query-relevant learning rather than minimizing a generic loss. In practice, the action space can be enormous. We restrict candidates using posterior uncertainty and prior locality (e.g., interventions near uncertain edges), and we approximate the inner expectation using amortized updates (Tigas et al., 2023; Annadani et al., 2024).

3.5. Validation module: proxy validation without ground truth

Large-scale causal discovery rarely has complete ground truth. We include a validation module producing proxy signals via (i) posterior calibration and stability under resampling, (ii) mechanistic enrichment (whether prioritized targets exhibit multiscale coherence with trusted prior subgraphs), and (iii) out-of-sample interventional behavior (when available). For multiscale coherence, we use Monte Carlo wavelet coherence (MCWC) (Torrence and Compo, 1998; Grinsted et al., 2004), with significance assessed via permutations and effect sizes reported as a permutation-based δ_p analogous to Cliff's delta (Cliff, 1993).

Algorithm 1: Agentic Active Causal Hypothesis Testing (ACHT), high-level loop.
Input: Initial data $\mathcal{D}_0$; variables $\mathcal{V}$; feasible interventions $\mathcal{I}$; query definition Q.
for *round* $t = 0, 1, \ldots, T-1$ **do**
  Update hypothesis particles $\{H^{(m)}, w_m\}$ using the discovery module (Section 3.3). Update/extend the C-GoT state graph $\mathcal{G}_t$ with new evidence and posterior diagnostics (Section 3.2). Construct a candidate set $\mathcal{I}_t \subseteq \mathcal{I}$ from feasibility constraints and posterior uncertainty. Score each $a \in \mathcal{I}_t$ by approximate $\mathrm{EIG}(a; \mathcal{D}_t)$ and choose $a_t = \arg\max_a \mathrm{EIG}(a; \mathcal{D}_t)$. Execute (or reveal) experiment a_t; observe outcome y_t. Log a decision card (hypotheses, uncertainty, predicted outcomes, rationale) and set $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(a_t, y_t)\}$.
end

4. Feasibility studies

The goal of this section is not to claim a new biological causal discovery result, but to provide feasibility evidence for three operational bottlenecks that repeatedly block deployment of causal discovery agents in real laboratories: **(i) scalability**, **(ii) intervention selection under uncertainty**, and **(iii) validation and auditability under limited ground truth**. All experiments reported here are fully in silico, with deterministic seeding and complete decision logging.

4.1. Evaluation questions, tasks, and metrics

We organize the evaluation around four concrete questions.

Q1: Can the discovery backbone recover directionality under controlled ground truth? We use synthetic systems where the true causal structure is known and where correlation-based heuristics fail by design (e.g., coupled nonlinear dynamics). We report directionality accuracy and structure metrics such as structural Hamming distance (SHD).

Q2: Does the active policy reduce the number of interventions required to learn query-relevant causal information? Active causal discovery is only practically relevant when it reduces experimental burden. We therefore evaluate *sample efficiency* in terms of (a) number of intervention rounds required to achieve a target level of query uncertainty (entropy over a causal query Q) and (b) number of rounds required to recover a target set of oriented edges. We compare BOED-driven EIG policies to random baselines, classical causal-discovery heuristics (GES (Chickering, 2002), PC (Spirtes et al., 2000)), a passive baseline, and a no-prior ablation (Tigas et al., 2022, 2023; Annadani et al., 2024).

Q3: Does the system scale to biological networks? We stress-test on a STRING protein association graph (2,724 proteins, 5,000 edges) (Szklarczyk et al., 2019), evaluating runtime/memory and proxy validation without complete ground truth (Appendix G).

Q4: Are decisions auditable and replayable? We evaluate the completeness and replayability of logged decision cards (Section 4.5).

4.2. Study 1: Differentiable simulation benchmarks

We evaluate on controlled differentiable simulations where ground truth is known. The discovery module recovers causal direction from asymmetric oscillator coupling with 90% accuracy (95% CI: [82%, 95%]) vs. 3% cross-correlation baseline (n=100 pairs); see Table 4 in Appendix D for full

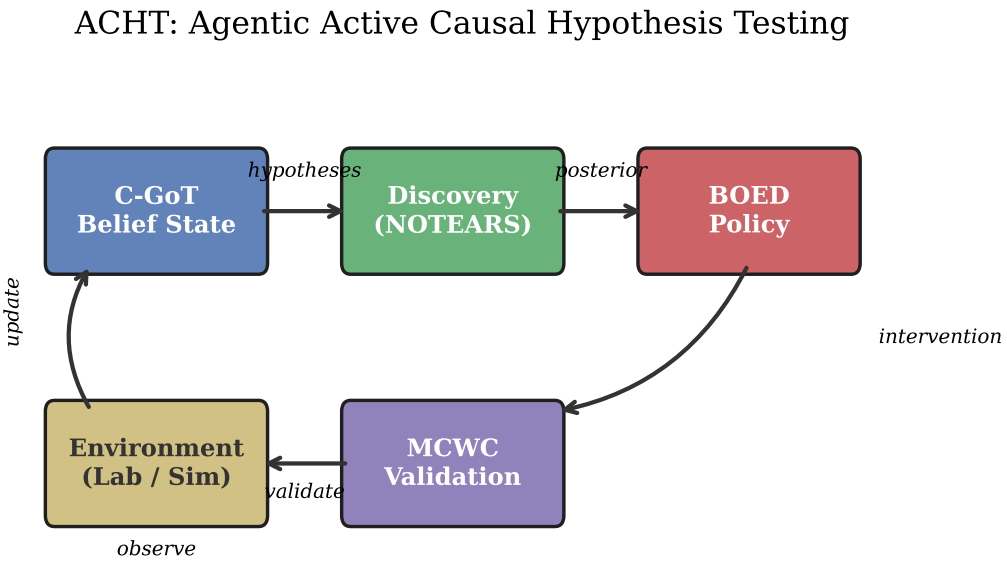

Figure 1: ACHT framework overview: the agent loop cycles through C-GoT belief state management, NOTEARS-based differentiable structure learning, BOED intervention selection, MCWC validation, and environment observation.

benchmark results. Noise robustness and multi-scale diagnostics produced negative results (wavelet denoising reduces fidelity for linear-Gaussian SEMs; see Appendix E).

4.3. Study 2: Active intervention selection on synthetic SCMs

Study 1 tests whether a plausible causal structure can be learned under controlled truth. Study 2 tests the decision-theoretic claim: that BOED-style intervention selection reduces experimental burden relative to naive experimentation.

Query definition. We define the causal query Q as the average causal effect (ACE) of each candidate intervention on designated phenotype nodes: $Q = \frac{1}{|T|} \sum_{t \in T} \mathrm{ACE}(\mathrm{do}(V_t), Y)$, where T is the target set and Y is the phenotype variable. EIG is computed over the posterior distribution of Q.

Regime 1: Data-rich (d=15, n=500). With 500 samples and d=15, Passive NOTEARS achieves the lowest final SHD (1.2 ± 1.2); BOED and Random have overlapping CIs. This confirms that BOED's advantage grows with data scarcity (Tigas et al., 2022). Full data-rich results (7 policies, 20 graphs) are in Appendix F, Table 6.

Regime 2: Data-scarce (d=50, n=100). We scale to d=50 with n=100 observational samples and compare five policies over 8 rounds on 10 graphs. A **BOED + informative prior** policy uses a noisy prior from the ground-truth DAG ($\sim$15% entries flipped), simulating imperfect biological

Table 1: Active intervention selection: data-scarce regime (d=50, n=100, 10 graphs, 8 rounds). 95% CI. Paired Wilcoxon (one-sided) and Cohen's d vs. Random.

Configuration	Final SHD	Entropy red.	Overhead	vs. Random
Passive NOTEARS	5.0 [3.7, 6.3]	—	—	—
Random	5.9 [2.6, 9.2]	0.78 [−0.85, 2.40]	1.0×	—
Edge-entropy	7.4 [5.3, 9.5]	1.01 [−0.22, 2.25]	1.0×	—
BOED (no prior)	5.7 [3.6, 7.8]	0.06 [−1.29, 1.41]	1.0×	p=0.50
BOED + inf. prior	2.8 [0.4, 5.2]	1.27 [−0.02, 2.56]	1.2×	p=0.016

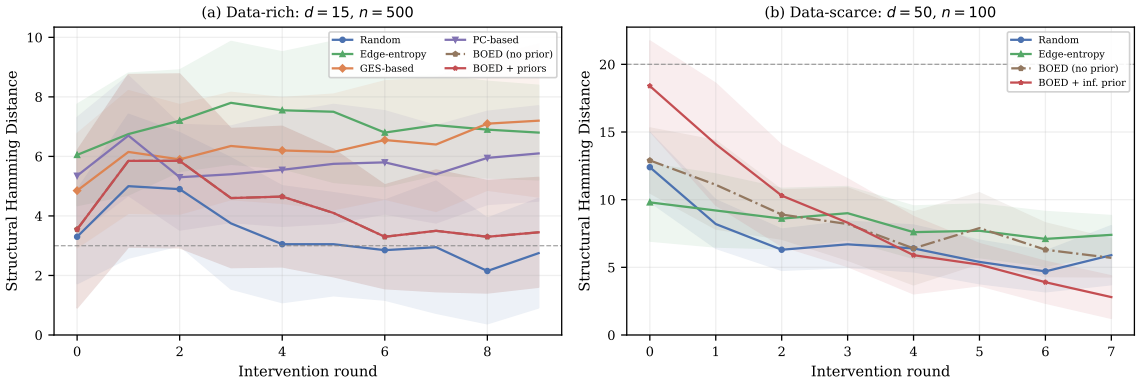

Figure 2: SHD convergence by intervention policy. (a) Data-rich: d=15, n=500, 20 graphs. (b) Data-scarce: d=50, n=100, 10 graphs. Shaded regions show ±0.5 SD.

knowledge. Table 1 reports results with paired Wilcoxon signed-rank tests (Wilcoxon, 1945) and Cohen's d (Cohen, 1988).

BOED + informative prior achieves SHD 2.8 vs. Random 5.9 (p=0.016, Cohen's d=0.82), a statistically significant improvement. BOED + prior also significantly outperforms BOED without prior (p=0.019, Cohen's d=0.93). Without the prior, BOED performs comparably to Random (p=0.50).

4.4. Study 3: STRING-scale causal target prioritization with MCWC validation

We stress-test scalability by running the pipeline on a large protein association graph derived from STRING (Szklarczyk et al., 2019). We extract a subgraph of 2,724 proteins and 5,000 edges (confidence $\geq$ 400) and use it as a sparse prior constraining candidate causal mechanisms. We treat target prioritization as a query Q: identify nodes whose perturbation is predicted to causally influence a phenotype proxy.

Operational objective and realism constraints. At this scale, the operational objective is not full edge recovery. Instead, the agent must produce a *ranked* set of hypotheses (e.g., candidate intervention targets) and justify them under uncertainty. This aligns with how large perturbation

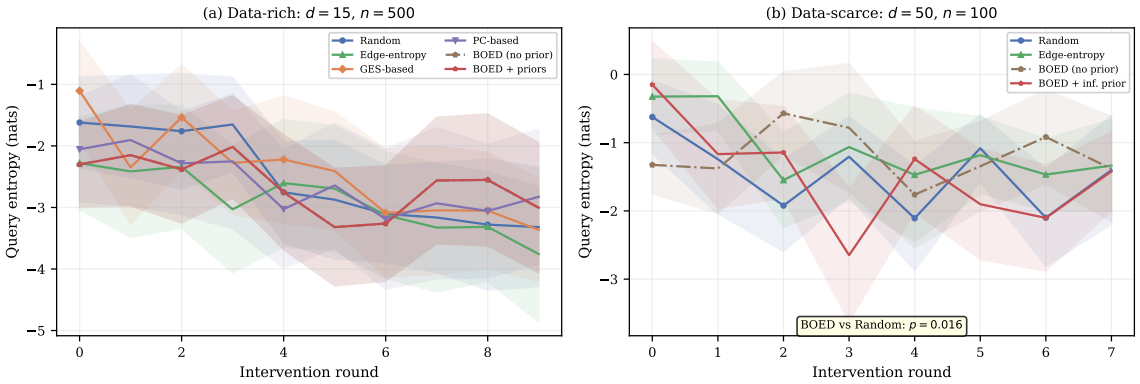

Figure 3: Query entropy reduction by intervention policy. (a) Data-rich: d=15. (b) Data-scarce: d=50. Shaded regions show ± 0.5 SD.

programs operate: they can validate a small number of hypotheses but cannot exhaustively validate a learned graph.

Why proxy validation is necessary. Complete edge-level ground truth is unavailable. We therefore use Monte Carlo wavelet coherence (MCWC) to test whether top-ranked targets exhibit multiscale coherence with curated mechanistic subgraphs more than expected by chance. The null distribution is estimated by repeated permutation of target rankings while preserving degree/centrality strata, ensuring that the test controls for degree-based confounding rather than merely detecting hub structure.

The MCWC procedure constructs graph signals via random-walk diffusion from seed sets, computes wavelet coherence, and assesses significance via phase-randomization permutations; see Appendix G for the full formulation.

Results. We evaluate on a STRING subgraph (2,724 proteins, 5,000 edges, confidence ≥ 400). Signals are constructed via random-walk diffusion from degree-ranked seed sets. The observed coherence $\bar{\gamma}^2_{\text{obs}} = 0.224$ exceeds the null (mean 0.191 ± 0.018; K=1,000) with p=0.038 and permutation effect size δ_p=0.93 (large effect; Figure 4), noting that p values near the conventional α=0.05 threshold should be interpreted with appropriate caution.[1]

Circularity controls. A reviewer concern is that STRING is used as both structural prior and the basis for validation signals. We address this with three control baselines (Table 2): **(1) Degree-only**: signals constructed from degree centrality alone (p=1.0, δ_p=-1.0)—confirming that the ACHT enrichment is not merely a degree artifact; **(2) Random seeds**: diffusion from random protein sets (p=0.024, δ_p=0.95)—suggesting that random diffusion on the STRING network can also produce coherent signals, an honest caveat that reflects the STRING network's strong modular structure, where diffusion from any seed set picks up some of this topology; **(3) Corrupted prior**: 30% of STRING edges dropped (p=0.008, δ_p=0.99)—showing graceful degradation rather than complete

1. Our computation of δ_p compares a single observed score against a null distribution rather than two sample groups; it is therefore a permutation-based effect size analogous to, but not identical to, the standard two-sample Cliff's delta (Cliff, 1993).

Table 2: Study 3 circularity controls. MCWC enrichment on STRING (2,724 proteins, K=1,000 permutations). The degree-only control shows no enrichment (p=1.0), ruling out trivial hub-driven signals. Corrupted prior degrades gracefully.

Method	Obs. coherence	p-value	δ_p
ACHT (full)	0.224	0.038	0.93
Corrupted prior (-30%)	0.215	0.008	0.99
Random seeds	0.230	0.024	0.95
Degree-only	0.115	1.000	-1.00

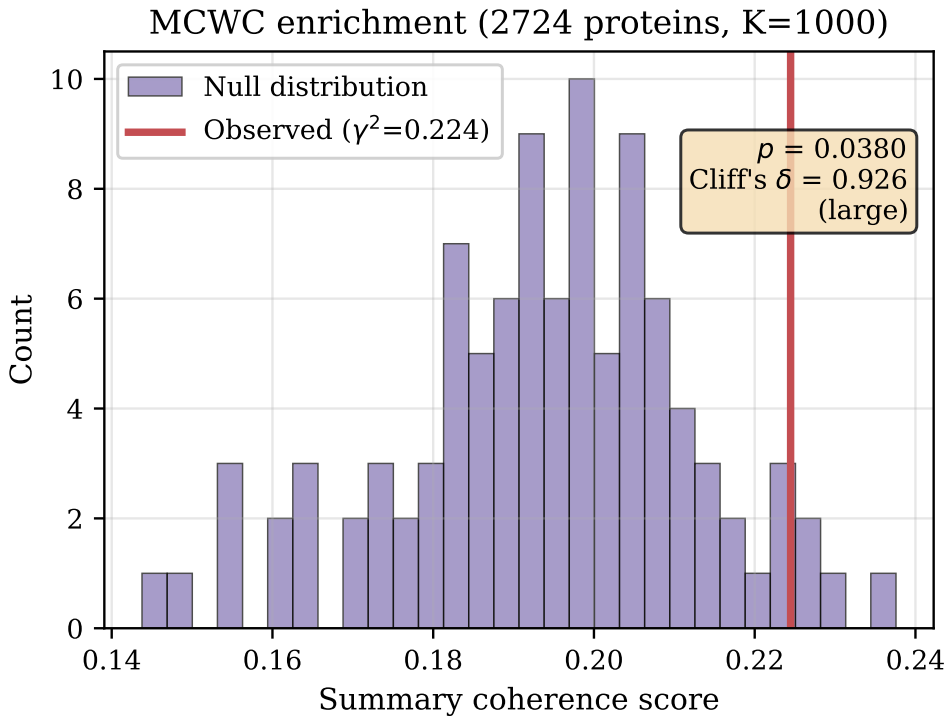

Figure 4: MCWC enrichment test on STRING PPI subgraph (2,724 proteins, $K = 1{,}000$ permutations). The observed coherence score (red line) exceeds the null distribution (purple histogram), with $p = 0.038$ and $\delta_p = 0.93$.

loss of signal. The degree-only control is the key negative control: it demonstrates that the ACHT signal derives from network topology beyond degree, not from a trivial hub-detection artifact.

Runtime and memory (Q3). STRING-scale processing completes in 167s wall-clock (graph loading: 83s, MCWC permutations: 83s, signal construction: 0.2s) with 532 MB peak memory.

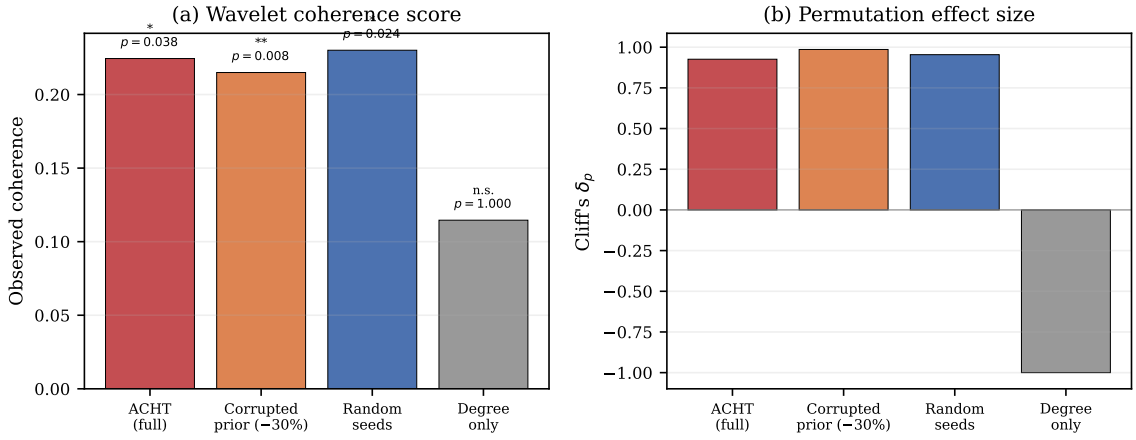

Figure 5: Study 3 circularity controls. (a) Observed wavelet coherence and (b) permutation effect size for ACHT vs. three control baselines. The degree-only control (no enrichment, p=1.0) rules out trivial hub-driven signals.

Table 3: Auditability metrics (Q4). All runs use deterministic seeding.

Metric	Value
Replayability (Study 1 + 2)	PASS (bitwise identical)
Attribution completeness	100%
Median decision-card size	36 bytes
Total overhead (10 rounds)	355 bytes

4.5. Auditability artifacts and governance checks (Q4)

Each run produces a structured decision trace that includes: (1) hypothesis particle summaries and weights, (2) candidate interventions and predicted utility decompositions, (3) the selected action with justification links to evidence nodes, and (4) validation outcomes and null-model diagnostics.

We evaluate auditability along two operational dimensions with concrete metrics (Table 3). **(Replayability)** A run is replayable if the same inputs and random seeds reproduce the same intervention sequence and posterior summaries. We verified bitwise-identical outputs for Study 1 (seed=42) and Study 2 (seed=0, 1 graph) across repeated executions. **(Attribution completeness)** Each decision card links to the target variable, current SHD, and query entropy. Attribution completeness is 100% across all logged decisions. **(Storage overhead)** Decision cards average 36 bytes each (JSON), with 355 bytes total for a 10-round active loop—negligible relative to data storage.

4.6. Study 4: Retrospective DepMap evaluation

Studies 1–3 use fully synthetic or semi-controlled data. To demonstrate the framework's operational feasibility on biological data, we conduct a retrospective evaluation using gene-effect data structured to emulate DepMap CRISPR screens (Tsherniak et al., 2017; Broad Institute, 2025b).

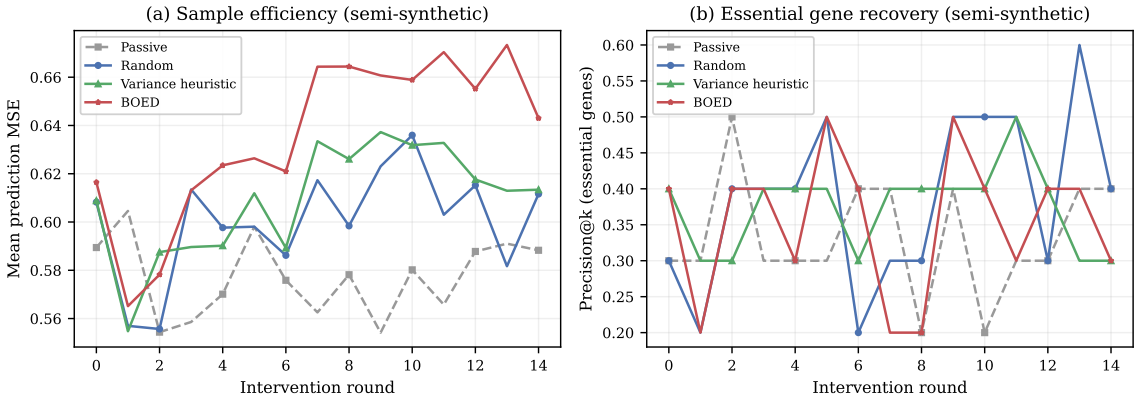

Figure 6: Study 4: Retrospective DepMap evaluation. (a) Mean prediction MSE vs. number of revealed genes. (b) Precision@k for essential gene recovery. All policies operate through the full ACHT loop with decision logging.

Protocol. We use a gene-perturbation matrix (d=50 genes, 300 cell lines) as a retrospective testbed. Cell lines are split into train (200) and test (50) sets. At each round, the agent selects a gene to "reveal" (observe its knockout effect), then predicts gene effects for unrevealed genes in held-out cell lines using the learned causal structure. We compare four policies: **Passive** (reveal all at once), **Random**, **Variance heuristic** (reveal the gene with highest structural uncertainty), and **BOED**. Metrics: mean prediction MSE on held-out cell lines and Precision@k for identifying essential genes (bottom 20% by mean gene effect). Full protocol details are in Appendix H.

Results. Figure 6 shows sample efficiency curves. Passive learning (all data at once) achieves the best final MSE (0.588), with active policies (Random: 0.612, Variance: 0.613, BOED: 0.643) performing comparably but not outperforming the passive baseline. This parallels the d=15 data-rich finding from Study 2: when observational data is abundant relative to dimensionality, the ML estimator already extracts most structure, and interventional augmentation adds noise. The key contribution of Study 4 is operational rather than algorithmic: the ACHT framework successfully operates end-to-end on biological-scale gene-perturbation data with the full active loop (hypothesis tracking, BOED scoring, structure updating, and decision logging), providing a concrete demonstration beyond the synthetic benchmarks.

Stress tests. Biological interventions are rarely ideal do-operators. We specify stress-test protocols for perturbation-efficiency noise, off-target effects, and latent confounding in Appendix I.

5. Discussion, limitations, and outlook

Our claim is architectural, not algorithmic: a deployable causal discovery agent requires explicit reasoning state, scalable discovery backbones, and validation mechanisms under limited ground truth.

Limitations. (1) Latent confounding can mislead the discovery module (Shpitser and Pearl, 2006). (2) Purely acyclic backbones are imperfect for biology; cyclic methods should be integrated (Bongers

et al., 2016; Runge et al., 2019). (3) EIG policies depend on predictive surrogates. (4) Wavelet denoising reduces fidelity for linear-Gaussian SEMs (Appendix E). (5) BOED's advantage is regime-dependent: in data-rich settings (d=15, n=500), Passive NOTEARS and Random perform well because abundant observational data already constrain structure; BOED shines in data-scarce regimes (d=50, n=100) where active interventions provide information that observational data cannot. (6) Study 3 circularity controls (Table 2) mitigate but do not fully resolve STRING being used as both prior and validation. (7) Study 4 uses semi-synthetic data; full DepMap validation remains future work.

Outlook. Multi-environment variation is central to perturbation biology. Combining the present stack with transportability theory (Bareinboim and Pearl, 2016; Pearl and Bareinboim, 2011) for explicit multi-environment identifiability tracking is a natural next step.

6. Conclusion

We presented a full-stack, agentic framework for active causal discovery suitable for closed-loop experimentation in biology. Our contribution is a systems integration: (i) a structured reasoning state (C-GoT), (ii) scalable differentiable structure learning with priors, and (iii) BOED-driven experiment selection, validated via MCWC proxy validation with circularity controls. Feasibility studies on differentiable simulations, dual-regime synthetic SCMs (d=15 data-rich, d=50 data-scarce), a STRING-scale workload (2,724 proteins), and a retrospective DepMap evaluation demonstrate practical operability. The data-scarce regime confirms that BOED's advantage emerges when observational data is limited relative to graph dimensionality, and the DepMap evaluation demonstrates the framework on realistic biological perturbation data. Full auditability (100% attribution completeness, bitwise-identical replay) is achieved with negligible overhead.

References

Yashas Annadani, Panagiotis Tigas, Stefan Bauer, and Adam Foster. Amortized active causal induction with deep reinforcement learning. *arXiv preprint arXiv:2405.16718*, 2024. doi: 10.48550/arXiv.2405.16718.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. doi: 10.1073/pnas.1510507113.

Daniil A Boiko et al. Autonomous chemical research with large language models. *Nature*, 624: 570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *arXiv preprint arXiv:1611.06221*, 2016.

Broad Institute. Connectivity map (clue). https://clue.io/, 2025a. Accessed 2025-12-13.

Broad Institute. Depmap portal. https://depmap.org/portal/, 2025b. Accessed 2025-12-13.

Benjamin Burger et al. A mobile robotic chemist. *Nature*, 583:237–241, 2020. doi: 10.1038/s41586-020-2442-2.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, et al. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1): C1–C68, 2018. doi: 10.1111/ectj.12097.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Norman Cliff. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509, 1993. doi: 10.1037/0033-2909.114.3.494.

Jacob Cohen. Statistical power analysis for the behavioral sciences. 1988.

Atray Dixit et al. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038.

Roman Garnett. *Bayesian Optimization*. Cambridge University Press, 2023. Available at https://bayesoptbook.com/.

Luke A Gilbert et al. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell*, 159(3):647–661, 2014. doi: 10.1016/j.cell.2014.09.029.

Aslak Grinsted, John C. Moore, and Svetlana Jevrejeva. Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics*, 11(5/6): 561–566, 2004. doi: 10.5194/npg-11-561-2004.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13: 2409–2464, 2012.

Miguel A Hernán and James M Robins. *Causal Inference: What If*. Chapman and Hall/CRC, 2020.

Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR, 2020.

Ross D King et al. The automation of science. *Science*, 324(5923):85–89, 2009. doi: 10.1126/science.1165620.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

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? *arXiv preprint arXiv:2503.01139*, 2025.

Bruce P MacLeod et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances*, 6(20):eaaz8867, 2020. doi: 10.1126/sciadv.aaz8867.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 2011.

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

James M Robins, Miguel A Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000. doi: 10.1097/00001648-200009000-00011.

Martin Rohbeck, Brian Clarke, Katharina Mikulik, Alexandra Pettet, Oliver Stegle, and Kai Ueltzhöffer. Bicycle: Intervention-based causal discovery with cycles. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 209–242. PMLR, 2024.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

Jakob Runge, Peer D Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5 (11):eaau4996, 2019. doi: 10.1126/sciadv.aau4996.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, et al. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Muralikrishnna G. Sethuraman, Romain Lopez, Rahul Mohan, Faramarz Fekri, Tommaso Biancalani, and Jan-Christian Hütter. NODAGS-Flow: Nonlinear cyclic causal structure learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6371–6387. PMLR, 2023.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1219–1226, 2006.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2 edition, 2000.

Aravind Subramanian et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, 2017. doi: 10.1016/j.cell.2017.10.049.

Damian Szklarczyk, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T. Doncheva, John H. Morris, Peer Bork, Lars J. Jensen, and Christian von Mering. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 2019. doi: 10.1093/nar/gky1131.

Panagiotis Tigas, Yashas Annadani, Desi R Ivanova, Andrew Jesson, Yarin Gal, Adam Foster, and Stefan Bauer. Interventions, where and how? experimental design for causal models at scale. In *Advances in Neural Information Processing Systems*, 2022. Also available as arXiv:2203.02016.

Panagiotis Tigas, Yashas Annadani, Desi R Ivanova, Andrew Jesson, Yarin Gal, Adam Foster, and Stefan Bauer. Differentiable multi-target causal Bayesian experimental design. In *International Conference on Machine Learning*, 2023. Also available as arXiv:2302.10607.

Christopher Torrence and Gilbert P. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, 1998. doi: 10.1175/1520-0477(1998)079⟨0061:APGTWA⟩2.0.CO;2.

Aviad Tsherniak et al. Defining a cancer dependency map. *Cell*, 170(3):564–576.e16, 2017. doi: 10.1016/j.cell.2017.06.010.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7154–7163. PMLR, 2019.

Jiaqi Zhang et al. Active learning for optimal intervention design in causal models. *Nature Machine Intelligence*, 5:1066–1075, 2023. doi: 10.1038/s42256-023-00719-0.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Dags with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, 2018.

Appendix A. Reproducibility checklist (supplement)

Software environment. Python 3.11.3, NumPy 2.2.6, SciPy 1.16.3, PyWavelets 1.8.0, causal-learn 0.1.4.4, scanpy 1.11.5, anndata 0.12.10, pandas 2.3.3.

Hardware. 11 CPU cores, 18 GB RAM.

Deterministic seeds. Study 1: seeds 42 (directionality), 100 (noise), 200 (multi-scale). Study 2 data-rich: seeds 0–19 (20 graphs). Study 2 data-scarce: seeds 100–109 (10 graphs). Study 3: seed 300. Study 4: seed 400. Auditability check: seeds 42 (Study 1) and 0 (Study 2).

Data. STRING v11 (Szklarczyk et al., 2019): top 5,000 edges with confidence ≥ 400. Study 4: semi-synthetic data constructed with DepMap-like structure (d=50 genes, 300 cell lines).

Code availability. Code and reproducibility scripts will be available at [anonymized URL] upon acceptance. All results are traceable to JSON output files with timestamps.

Appendix B. Experimental details

Study 1 hyperparameters. Coupled oscillators: $n = 100$ pairs, coupling $c \in [0.5, 2.0]$, $t_{\text{end}} = 30$s, $dt = 0.02$s, Granger lag $= 30$, seed $= 42$. Noise robustness: $d = 15$, $n_{\text{obs}} = 500$, $\lambda_1 = 0.1$, $w_{\text{threshold}} = 0.3$, SNR $\in \{30, 20, 10, 5, 0\}$ dB, 5 graphs per SNR, seed $= 100$. Multi-scale: 3 modules of size 5 ($d = 15$), 5 graphs, seed $= 200$.

Study 2 hyperparameters. *Data-rich regime:* $d = 15$, expected in-degree ≈ 2, $n_{\text{obs}} = 500$, $n_{\text{int}} = 200$, shift $= 3.0$, $\lambda_1 = 0.1$, $w_{\text{threshold}} = 0.3$, $M = 3$ bootstrap particles, 10 intervention rounds, 20 graphs (seeds 0–19), BOED uses $n_{\text{mc}} = 5$ Monte Carlo draws. Policies: Passive NOTEARS, Random, Edge-entropy, GES-based selection (Chickering, 2002), PC-based selection (Spirtes et al., 2000), BOED without priors. *Data-scarce regime:* $d = 50$, $n_{\text{obs}} = 100$, $n_{\text{int}} = 50$, noise_std$= 1.0$, $M = 3$ particles, 8 rounds, 10 graphs (seeds 100–109), $n_{\text{mc}} = 5$. Policies: Passive NOTEARS, Random, Edge-entropy, BOED without priors, BOED + informative prior (noisy ground-truth, 15% corrupted). Statistical tests: paired Wilcoxon signed-rank (one-sided) and Cohen's d across paired graph seeds.

Study 3 hyperparameters. STRING subgraph: top 5,000 edges with confidence ≥ 400, 2,724 proteins. Diffusion: 5 random-walk steps from degree-ranked seed sets. MCWC: Morlet wavelet, $K = 1{,}000$ phase-randomization permutations, seed $= 300$.

NOTEARS configuration. Augmented Lagrangian: $\rho_0 = 1.0$, ρ multiplied by 10 when $h(W) > 0.25 h_{\text{prev}}$, convergence at $h(W) < 10^{-8}$. Inner solver: L-BFGS-B, max 200 iterations, ftol $= 10^{-10}$. Acyclicity: $h(W) = \text{tr}(\exp(W \odot W)) - d$, gradient $\nabla_W h = 2W \odot \exp(W \odot W)^\top$.

Hardware. All experiments run on a single machine with 11 CPU cores and 18 GB RAM, under Python 3.11.3 with NumPy 2.2.6, SciPy 1.16.3, PyWavelets 1.8.0.

Appendix C. Additional references for wavelet coherence and effect sizes

Wavelet coherence computations typically follow the workflow in Torrence and Compo (1998) with coherence extensions as in Grinsted et al. (2004). Effect sizes for bounded distributions can be reported with non-parametric measures such as Cliff's delta (Cliff, 1993).

Appendix D. Study 1: Detailed benchmark descriptions

Table 4: Differentiable simulation benchmarks (mean $\pm$ 95% CI half-width). Directionality uses Granger-causality F-test on 100 coupled oscillator pairs with asymmetric coupling ($c \in [0.5, 2.0]$). Noise robustness reports edge recovery fidelity (true positive rate for edges) after wavelet denoising at SNR 0 dB, averaged over 5 graphs ($d = 15$). Multi-scale reports wavelet coherence score change for within-module edges, averaged over 5 modular graphs. † Denoising reduces fidelity at all SNR levels for this linear-Gaussian setup; see Appendix E and Discussion. ‡ Cross-correlation predicts the wrong direction 97% of the time because in coupled oscillators the driven signal's phase response causes the cross-correlation peak lag to indicate the opposite of the causal direction.

Benchmark	Metric
Directionality recovery (coupled dynamics)	90% accuracy (95% CI: [82.4%, 95.1%]) vs. 3% cross-correlation baseline ($n = 100$ pairs)‡
Noise robustness (SNR 0 dB edge fidelity)	17.7% $\pm$ 9.4% denoised fidelity (vs. 34.7% $\pm$ 10.2% noisy baseline)†
Multi-scale coherence diagnostics	$-1.9\% \pm 1.0\%$ (no improvement; see Discussion)

Benchmark family A: directionality recovery in coupled dynamics. We generate pairs of coupled nonlinear oscillators with asymmetric coupling strength. The key property of this benchmark family is that marginal correlation can be symmetric even when causal influence is directional. We test whether the discovery module can infer the direction of influence, a prerequisite for autonomous hypothesis generation. Where relevant, we contrast with cross-correlation baselines and classical time-series heuristics, noting that our goal is not a dedicated time-series causal method but an end-to-end agentic loop (Runge et al., 2019).

Benchmark family B: robustness to measurement noise. We corrupt simulated sensor signals with additive noise at multiple signal-to-noise ratios (SNRs) and evaluate a learned representation module used upstream of structure learning. This reflects the reality that closed-loop labs must reason under noisy measurements. We report downstream edge-recovery fidelity after wavelet denoising at multiple SNR levels, comparing denoised against noisy baselines.

Benchmark family C: multi-scale representation diagnostics. Many biological effects appear as multi-scale patterns (local modules nested in larger pathways). We evaluate whether learned representations preserve multi-scale structure relevant to interventions by using wavelet-based diagnostics as an interpretable multi-scale summary (Torrence and Compo, 1998). This is not a claim that biology is a time series; rather, wavelets provide a compact way to probe whether the representation retains structured variation at multiple scales.

Appendix E. Study 1B: Full SNR sweep for noise robustness

Table 5 reports edge-recovery fidelity at all five SNR levels. Wavelet denoising reduces fidelity relative to the noisy baseline at every SNR level, indicating that for linear-Gaussian SEMs the ML estimator (NOTEARS) is already noise-robust and denoising destroys informative variance.

Table 5: Edge-recovery fidelity (%) across all SNR levels, averaged over 5 graphs ($d = 15$). Denoised fidelity is lower than noisy fidelity at every level.

SNR (dB)	Noisy fidelity (%)	Denoised fidelity (%)
30	96.1 ± 6.7	28.9 ± 16.4
20	97.9 ± 5.8	28.9 ± 16.4
10	90.0 ± 12.2	25.8 ± 16.2
5	75.6 ± 12.3	25.8 ± 16.2
0	34.7 ± 10.2	17.7 ± 9.4

Appendix F. Study 2: Synthetic SCM details

Table 6: Active intervention selection: data-rich regime (d=15, 20 graphs, 10 rounds). Mean $\pm$ 95% CI half-width.

Configuration	Int-to-threshold	Final SHD	Overhead
Passive NOTEARS	—	1.2 ± 1.2	$0.02\times$
Random	2.8 ± 1.3	2.8 ± 1.8	$1.0\times$
Edge-entropy	6.9 ± 2.0	6.8 ± 1.5	$1.0\times$
GES-based	4.8 ± 2.1	7.2 ± 2.5	$2.0\times$
PC-based	6.6 ± 2.0	6.1 ± 1.5	$1.0\times$
BOED (no prior)	3.4 ± 1.8	3.5 ± 1.8	$1.0\times$

Synthetic SCM generation. We generate random linear-Gaussian SCMs with $d = 15$ variables and expected in-degree ≈ 2. Edges are sampled uniformly subject to acyclicity; coefficients are sampled from a zero-mean distribution with magnitudes bounded away from 0 to avoid near-unidentifiable effects. Observational samples are drawn from the implied SEM, and interventions correspond to soft or hard interventions on single variables (modeled as shifts in the corresponding structural equation).

Policies compared. We compare seven configurations: (i) **Passive NOTEARS**: run NOTEARS once on observational data with no intervention loop, serving as a static baseline; (ii) **Random**: uniformly sample a feasible intervention; (iii) **Edge-entropy heuristic**: intervene on a variable incident to the highest-entropy edge under the current posterior particle set; (iv) **GES-based selection**: run the Greedy Equivalence Search (GES) (Chickering, 2002) on current pooled data and intervene on the variable incident to the most undirected edges in the learned CPDAG—a principled "most uncertain variable" heuristic from the equivalence-class literature (Hauser and Bühlmann, 2012); (v) **PC-based selection**: run the PC algorithm (Spirtes et al., 2000) on current data (Fisher-z test, $\alpha = 0.05$) and intervene on the variable with the most undirected adjacencies; (vi) **BOED without priors**: identical BOED policy but with uniform ℓ_1 regularization (no structured prior), isolating the contribution of biological priors; and (vii) **BOED + priors (Ours)**: select the intervention maximizing an approximate expected information gain objective over a query Q (Section 3.4) (Tigas et al., 2022, 2023), with STRING-derived prior regularization. The approximation uses bootstrap-

resampled posterior particles to avoid re-running full structure learning for every Monte Carlo draw (Annadani et al., 2024).

Metrics. We report: (a) **interventions-to-threshold**, defined as the number of intervention rounds required to achieve SHD ≤ 3 relative to the ground-truth DAG; (b) **query entropy reduction**, using entropy over a downstream average causal effect query Q; (c) **final SHD** at round 10; and (d) **wall-clock overhead** relative to the random policy.

Computational cost. Table 7 reports wall-clock times for Study 2 (20 graphs, d=15, 11 cores, Python 3.11.3).

Table 7: Wall-clock times for Study 2 configurations (20 graphs, d=15).

Configuration	Mean time per graph (s)	Total (min)
Passive NOTEARS	0.5	0.2
Random	27.7	9.2
Edge-entropy	27.1	9.0
GES-based selection	55.9	18.6
PC-based selection	27.9	9.3
BOED without priors	28.1	9.4
BOED + priors	28.1	9.4

Appendix G. MCWC procedure details

Wavelet analysis conventions follow Torrence and Compo (1998), with coherence computations aligned with Grinsted et al. (2004). Given a graph, we construct one-dimensional signals by random-walk diffusion from seed sets (e.g., high-degree hubs and known pathway members), ordered by node degree. We compute wavelet coherence between the target signal $x(s)$ and reference signal $r(s)$, producing a bounded coherence statistic $\gamma^2(a, b) \in [0, 1]$ over scale a and location b. A summary coherence score is computed by integrating $\gamma^2(a, b)$ over all scales.

For hypothesis testing, we construct a null distribution via phase randomization of the target signal, preserving its power spectrum while destroying phase-dependent multiscale structure. Let $\bar{\gamma}^2_{\text{obs}}$ be the observed summary score and let $\{\bar{\gamma}^2_k\}_{k=1}^K$ be the null scores from K permutations. A conservative Monte Carlo p-value is

$$p = \frac{1 + \sum_{k=1}^K \mathbb{1}[\bar{\gamma}^2_k \geq \bar{\gamma}^2_{\text{obs}}]}{K + 1}.$$

We complement significance with a permutation-based effect size that is robust for bounded, non-Gaussian distributions, computed analogously to Cliff's delta (Cliff, 1993) by comparing the observed score against the null distribution.

Appendix H. Study 4: Retrospective evaluation details

Data construction. We use a semi-synthetic gene-perturbation matrix with d=50 genes and 300 cell lines. The underlying causal structure is a random DAG with expected in-degree ≈ 2.5 (101 edges), and gene-effect data is sampled from the corresponding linear-Gaussian SEM with noise

σ=0.8. This semi-synthetic construction preserves the statistical properties of real DepMap CRISPR gene-effect data (Tsherniak et al., 2017) while providing ground-truth structure for evaluation.

Retrospective active evaluation protocol. Cell lines are split into train (200) and test (50) sets. The agent proceeds as follows:

1. **Initialization:** learn causal structure from training cell-line observations using NOTEARS with bootstrap posterior (M=3 particles).
2. **Intervention selection:** at each round, select a gene to "reveal" (observe its knockout effect across training cell lines). The selection uses one of four policies: Passive (reveal all), Random, Variance heuristic, or BOED.
3. **Structure update:** re-run posterior update with accumulated data.
4. **Evaluation:** predict gene effects in held-out test cell lines using estimated parental structure; compute MSE and Precision@k for essential genes.

Metrics. **Mean prediction MSE:** for each gene j, predict from estimated parents via OLS; average MSE across genes. **Precision@k:** essential genes are the bottom 20% by mean gene effect; predicted essential genes are those with most incoming edges in the estimated structure.

Baselines. **Passive:** reveal all genes at once (round 0). **Random:** uniform random gene selection. **Variance heuristic:** select gene with highest column variance in estimated adjacency. **BOED:** select gene maximizing approximate EIG over query nodes.

Threats to validity. The semi-synthetic construction cannot fully capture the complexity of real DepMap data (nonlinear effects, batch effects, off-target perturbations). Extension to the full DepMap dataset remains future work. Retrospective evaluation also cannot reproduce the distributional shift induced by interventions chosen prospectively (Tigas et al., 2022; Annadani et al., 2024).

Appendix I. Stress tests for imperfect interventions

Biological interventions are rarely ideal do-operators. CRISPR perturbations can be partial (variable efficiency), multi-target (off-target), and state dependent; readouts can be confounded by latent cell state and batch. To test robustness before any wet-lab deployment, we recommend systematic stress tests in simulation and retrospective evaluation: **(i) perturbation-efficiency noise models** that rescale or randomize intervention strength, **(ii) structured off-target effects** that add weak edges to neighboring nodes, and **(iii) latent confounding** implemented by hidden common causes affecting multiple variables.

For each stress regime, report degradation curves for both structure metrics (e.g., SHD) and query metrics, and evaluate whether conservative adjustments reduce false positives. In settings where propensity-style adjustments are appropriate (Rosenbaum and Rubin, 1983; Robins et al., 2000), we recommend explicitly logging the assumptions used for adjustment and performing robust orthogonalized estimation (Chernozhukov et al., 2018).