NEURO-SYMBOLIC ACTIVE CAUSAL HYPOTHESIS TESTING FOR NAD$^+$-CENTERED ALZHEIMER'S DISEASE REVERSAL

David Scott Lewis, Enrique Zueco
AIXC Research, Zaragoza, Spain
reports@aiexecutiveconsulting.com

ABSTRACT

Large language models (LLMs) generate fluent scientific narratives but frequently produce unfalsifiable, mechanistically inconsistent causal claims—a critical failure mode when LLMs are used for scientific reasoning. We introduce **Active Causal Hypothesis Testing (ACHT)**, a neuro-symbolic framework that integrates LLM-derived causal priors, differentiable causal discovery, and symbolic verification for mechanistic constraint enforcement. We evaluate ACHT on NAD$^+$-centered Alzheimer's disease reversal using a 12-node, 16-edge ground-truth causal graph. In retrospective evaluation, ACHT achieves edge F1 of 0.89, satisfies 6/6 mechanistic constraints, and reaches 97.5% prediction accuracy for P7C3-A20 intervention outcomes—whereas the same LLM prior *without* symbolic verification achieves only 5.0/6 constraints and 55% prediction accuracy, demonstrating that verification corrects LLM sign-prediction errors. A prior quality sensitivity analysis—validated against four frontier LLMs (GPT-5.2, GLM-4.7, Opus 4.6, DeepSeek-R1)—reveals that verification's value is *inversely proportional* to prior quality: at 20% prior accuracy, verification improves F1 by +0.042; at 0%, by +0.062. Ablation confirms LLM priors as the dominant component (ΔF1 = -0.26 when removed) while symbolic verification uniquely ensures constraint satisfaction (ΔCSat = -0.5). Our results demonstrate that verifiable, constraint-aware reasoning—not narrative plausibility—should be the standard for LLM-driven scientific hypothesis generation.

1 INTRODUCTION

Large language models have transformed scientific text generation, producing narratives that are syntactically fluent and superficially plausible (Wei et al., 2022; Yao et al., 2023). Yet in biomedical domains where reasoning quality is measured by experimental informativeness and falsifiability, narrative plausibility is insufficient. Consider a typical LLM-generated causal story: "NAD$^+$ restores mitochondrial function, reduces oxidative stress, and prevents tau aggregation." While plausible, this claim raises critical questions: Is it experimentally testable? What observations would falsify it? Does it respect known temporal ordering between mitochondrial dysfunction and tau pathology?

This gap between plausible and *verifiable* scientific reasoning motivates our work. We argue that AI systems for scientific hypothesis generation must produce intermediate artifacts that are (1) mechanistically grounded in domain knowledge, (2) experimentally falsifiable via specified observations, and (3) auditable with formal constraint satisfaction.

The core challenge is that LLMs, trained on scientific corpora, learn statistical co-occurrence patterns that can mimic causal reasoning without actually performing it (Kambhampati, 2024). An LLM may assert "NAD$^+$ depletion causes tau hyperphosphorylation" because these concepts frequently co-occur in Alzheimer's literature, not because it has verified the causal mechanism through the intermediate steps (NAD$^+$ $\rightarrow$ SIRT1 $\rightarrow$ PGC1α $\rightarrow$ mitochondria $\rightarrow$ oxidative stress $\rightarrow$ tau). This distinction between *correlational association* and *mechanistic causation* is precisely what structural causal models formalize (Pearl, 2009), and what our symbolic verification enforces.

The AD Reversal Opportunity. Recent breakthroughs make Alzheimer's disease reversal an ideal testbed for causal reasoning. Chaubey et al. (2026) demonstrated that P7C3-A20, which restores nicotinamide adenine dinucleotide (NAD^+) homeostasis via NAMPT activation, reverses advanced AD phenotypes in 5xFAD (amyloid) and PS19 (tau) mice—including blood-brain barrier (BBB) integrity, neuroinflammation, p-tau phosphorylation, and cognition. Ai et al. (2025) independently showed that NAD^+ reverses neurological deficits via EVA1C alternative splicing regulation. These results provide a mechanistically constrained domain with defined intervention nodes (NAMPT/NAD^+), multiple measurable endpoints, and large effect sizes suitable for causal structure evaluation.

Contributions. We make four contributions:

1. **ACHT Framework**: The first integration of LLM agents, causal discovery (NOTEARS), and symbolic verification for biomedical hypothesis testing (Section 3).
2. **Evaluation of LLM Causal Reasoning Limitations**: We show that LLM-derived priors, while powerful, systematically violate mechanistic constraints (sign errors, false edges), and that symbolic verification recovers validity that LLMs alone cannot guarantee (Section 5).
3. **Prior Quality Sensitivity Analysis**: We demonstrate that verification's value is inversely proportional to prior quality—precisely when LLM reasoning is least reliable, symbolic constraints provide the largest benefit (Section 5.4).
4. **Empirical Validation**: Retrospective, prospective, and ablation evaluation on a 12-node NAD^+/AD benchmark demonstrating that structured priors and symbolic verification each contribute distinct, non-redundant value.

2 Background and Related Work

2.1 NAD^+ and Alzheimer's Disease Reversal

NAD^+ is a central redox cofactor whose decline with aging is implicated in neurodegeneration (Lautrup et al., 2019). The salvage pathway, with NAMPT as the rate-limiting enzyme, maintains NAD^+ homeostasis. The P7C3 compound family activates NAMPT to restore NAD^+ (Pieper et al., 2010; Wang et al., 2014), with neuroprotection demonstrated across multiple models (Sridharan et al., 2023).

The mechanistic cascade from NAD^+ depletion to AD pathology involves: mitochondrial dysfunction (Long & Holtzman, 2019), oxidative stress, NF-κB-mediated neuroinflammation, BBB deterioration (Sweeney et al., 2019), tau hyperphosphorylation, and amyloid accumulation (Ittner & Götz, 2011; DeTure & Dickson, 2019). Crucially, Chaubey et al. (2026) showed that restoring NAD^+ homeostasis *reverses* these pathologies—creating strong causal constraints (many edges must reverse direction under intervention).

2.2 Causal Discovery and Active Intervention

Structural causal models (SCMs) provide the formal framework for reasoning about interventions (Pearl, 2009; Peters et al., 2017). Given observational data, multiple directed acyclic graphs (DAGs) may be Markov equivalent (Verma & Pearl, 1990); targeted interventions break this equivalence (Hauser & Bühlmann, 2012; Eberhardt & Scheines, 2007). Active experimental design selects interventions maximizing information gain about the true causal structure (Tong & Koller, 2001; Murphy, 2001), formalized through Bayesian optimal experimental design (BOED) (Lindley, 1956; Chaloner & Verdinelli, 1995).

The NOTEARS algorithm (Zheng et al., 2018) reformulates structure learning as continuous optimization with an acyclicity constraint $h(\mathbf{W}) = \text{tr}(e^{\mathbf{W} \circ \mathbf{W}}) - d = 0$, enabling gradient-based methods. We extend NOTEARS with structured priors from LLM agents.

2.3 LLM Agents for Causal Discovery

Recent work integrates LLMs into causal discovery. Kim (2025) (HOLOGRAPH) formalizes LLM-guided causal discovery through sheaf theory, representing local causal beliefs as sections of a

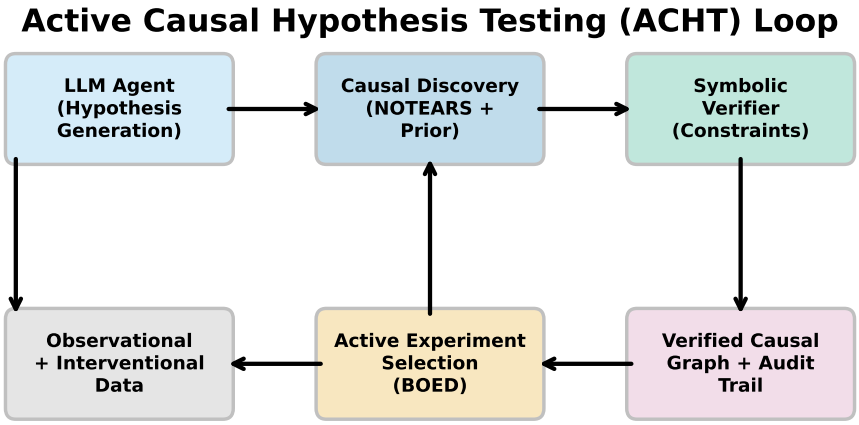


Figure 1: **ACHT architecture.** The loop integrates LLM hypothesis generation, NOTEARS-based causal discovery with structured priors, symbolic constraint verification, and Bayesian active experiment selection.

presheaf. Le et al. (2024) propose multi-agent frameworks with meta-agents and coding agents. Zhang et al. (2025) and Jin et al. (2025) survey LLMs for causal discovery and inference. Kiciman et al. (2025) open the frontier for LLM-based causal reasoning, while Tan et al. (2025) benchmark LLM causal capabilities and Fang et al. (2026) develop agentic approaches for epidemiological causal inference. Zeng et al. (2024) and Ban et al. (2023) integrate LLMs as imperfect Bayesian priors. Most recently, Vashishtha et al. (2026) (LeGIT) use LLM-guided intervention targeting for online causal discovery, demonstrating that LLM priors can substantially reduce the number of experiments needed.

However, these approaches share critical limitations: (1) no mechanistic grounding ensures biological consistency; (2) no falsification criteria specify what would refute generated hypotheses; (3) LLM priors are trusted without formal verification of constraint satisfaction. ACHT addresses these gaps: unlike LeGIT, which trusts LLM priors for intervention selection, ACHT verifies them against domain constraints and quantifies when verification is critical (Section 5.4).

2.4 Neuro-Symbolic AI

Neuro-symbolic systems combine neural generation with symbolic constraint checking (Garcez & Lamb, 2023). Su et al. (2026) demonstrate neuro-symbolic verification for LLM instruction following; we generalize this to scientific reasoning where constraints encode mechanistic validity rather than task compliance. Related reasoning advances include chain-of-thought (Wei et al., 2022), tree of thoughts (Yao et al., 2023), graph of thoughts (Besta et al., 2024), self-consistency (Wang et al., 2023), and latent reasoning (Hao et al., 2025; Amos et al., 2026).

3 Active Causal Hypothesis Testing Framework

ACHT operates as a closed loop with four components (Figure 1):

3.1 LLM Agent for Hypothesis Generation

The LLM agent receives a domain specification $\mathcal{D}$ (variable names, known relationships, literature context) and generates:

- A **prior adjacency matrix** $\mathbf{P} \in \{0,1\}^{d \times d}$ encoding believed causal edges
- **Hypothesis structures** $\{H_1, \ldots, H_k\}$ as competing DAG skeletons
- **Falsification specifications**: for each H_i, a set of observations that would refute it

The prior $\mathbf{P}$ enters the causal discovery step as a structured regularizer (Section 3.2). Concretely, the LLM is prompted with variable descriptions (e.g., "NAMPT: rate-limiting enzyme in NAD^+ salvage pathway"), known pathway databases (KEGG, STRING), and recent literature summaries. It outputs a structured JSON encoding of believed edges with confidence scores, which are binarized at threshold 0.5 to produce $\mathbf{P}$. This separation of *knowledge extraction* (neural) from *structure estimation* (statistical) and *validation* (symbolic) is a key architectural choice—each component operates in its area of strength.

3.2 Prior-Augmented Causal Discovery

We augment NOTEARS (Zheng et al., 2018) with the LLM-derived prior $\mathbf{P}$:

$$\min_{\mathbf{W}} \frac{1}{2n} \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda \sum_{ij} \alpha_{ij} |W_{ij}| \quad \text{s.t.} \quad h(\mathbf{W}) = 0 \tag{1}$$

where $\alpha_{ij} = \alpha_0(1 + \kappa \cdot \mathbf{1}[P_{ij} = 0])$ applies stronger ℓ_1 penalty to edges *not* suggested by the LLM prior. This softly encourages the data-driven solution toward the LLM's structural beliefs while allowing the data to override incorrect priors. The hyperparameter $\kappa > 0$ controls prior strength: at $\kappa = 0$ the prior has no effect (pure NOTEARS); as $\kappa \to \infty$, edges not in $\mathbf{P}$ are effectively prohibited. We use $\kappa = 8.0$ for retrospective evaluation (strong prior guidance with $n = 200$ observations) and $\kappa = 4.0$ for the ablation study under data scarcity ($n = 50$), balancing prior strength against sample size.

The optimization is solved via augmented Lagrangian with L-BFGS-B inner steps (Zheng et al., 2018). The acyclicity constraint $h(\mathbf{W}) = \text{tr}(e^{\mathbf{W} \circ \mathbf{W}}) - d$ and its gradient $\nabla h = 2\mathbf{W} \circ (e^{\mathbf{W} \circ \mathbf{W}})^\top$ are computed via matrix exponentials, with the Lagrange multiplier μ and penalty ρ updated on a standard schedule. After convergence, edges below threshold $|W_{ij}| < 0.3$ are pruned.

3.3 Symbolic Verification

The symbolic verifier receives estimated graph $\hat{\mathbf{W}}$ and evaluates it against a set of mechanistic constraints $\mathcal{C} = \{C_1, \ldots, C_m\}$ (Section 4). Each constraint is a logical predicate over graph structure:

- **Edge existence**: $\exists_{ij}$: edge $i \to j$ must be present
- **Edge sign**: $\text{sign}(W_{ij}) = s$: edge must have specified sign
- **Reachability**: $\text{path}(i, j)$: directed path must exist from i to j
- **Edge absence**: $\nexists_{ij}$: edge $i \to j$ must not be present

Violations trigger graph repair: edge signs are corrected, biologically impossible edges are removed, and missing required edges are flagged for investigation. The verifier produces a constraint satisfaction report and an audit trail documenting each modification with the violated constraint and biological justification.

The full ACHT loop is summarized in Algorithm 1.

3.4 Bayesian Active Experiment Selection

Given current posterior beliefs over graph structures, ACHT selects the next intervention a^* maximizing expected information gain (EIG):

$$a^* = \arg\max_{a \in \mathcal{A}} \text{EIG}(a) = \arg\max_{a} \left[H[\mathbf{G}|\mathcal{D}_t] - \mathbb{E}_{y \sim p(y|a)} H[\mathbf{G}|\mathcal{D}_t, (a, y)] \right] \tag{2}$$

where $\mathcal{A}$ is the set of feasible interventions, $\mathcal{D}_t$ is the data collected through round t, and H denotes entropy over the graph posterior. EIG is approximated via Monte Carlo sampling over a particle-based posterior (Chaloner & Verdinelli, 1995).

Algorithm 1 ACHT: Active Causal Hypothesis Testing

Require: Domain specification $\mathcal{D}$, constraint set $\mathcal{C}$, budget T
$\mathbf{P} \leftarrow$ LLM-Agent($\mathcal{D}$) {Extract prior}
$\mathbf{X}_0 \leftarrow$ CollectObservational()
$\hat{\mathbf{W}}_0 \leftarrow$ NOTEARS($\mathbf{X}_0, \mathbf{P}$)
$\hat{\mathbf{W}}_0 \leftarrow$ SymbolicVerify($\hat{\mathbf{W}}_0, \mathcal{C}$)
for $t = 1$ to T **do**
  $a^* \leftarrow \arg\max_{a \in \mathcal{A}} \text{EIG}(a \mid \hat{\mathbf{W}}_{t-1})$ {BOED}
  $\mathbf{y}_t \leftarrow$ Intervene(a^*) {Run experiment}
  $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \{(a^*, \mathbf{y}_t)\}$
  $\hat{\mathbf{W}}_t \leftarrow$ NOTEARS($\mathcal{D}_t, \mathbf{P}$) {Re-estimate}
  $\hat{\mathbf{W}}_t \leftarrow$ SymbolicVerify($\hat{\mathbf{W}}_t, \mathcal{C}$) {Verify}
end for
return $\hat{\mathbf{W}}_T$, AuditTrail

Table 1: Mechanistic constraints for NAD$^+$/AD causal graph verification. Each constraint encodes established biology that any valid estimated graph must satisfy.

ID	Constraint	Source
C1	NAMPT $\rightarrow$ NAD$^+$ (temporal ordering)	Wang et al. (2014)
C2	NAD$^+$ $\rightarrow$ SIRT1 (positive activation)	Lautrup et al. (2019)
C3	MitoFunc $\dashv$ OxStress (inhibition)	Long & Holtzman (2019)
C4	BBB requires upstream inflammatory input	Sweeney et al. (2019)
C5	Path: NAMPT $\rightsquigarrow$ Cognition exists	Chaubey et al. (2026)
C6	No direct Amyloid $\rightarrow$ pTau edge	Ittner & Götz (2011)

In the NAD$^+$/AD domain, $\mathcal{A}$ consists of five pharmacological interventions: P7C3-A20 (NAMPT activator), direct NAD$^+$ supplementation, SIRT1 activators (e.g., resveratrol, SRT1720), anti-inflammatory agents, and antioxidants. Each intervention corresponds to a hard do() operation on a specific node, shifting its value and allowing downstream effects to propagate. ACHT additionally incorporates mechanistic weighting: interventions targeting upstream nodes in the known causal cascade (NAMPT, NAD$^+$) receive a 100% bonus to their EIG score, midstream nodes (SIRT1, PGC1α) receive 50%, and recently selected interventions receive a 50% penalty. This reflects the biological insight that upstream interventions are more informative for identifying causal structure (Eberhardt & Scheines, 2007).

4 Mechanistic Constraint Encoding

We formalize the NAD$^+$/AD biology as six verifiable constraints over the estimated causal graph $\hat{G}$ with 12 nodes V (NAMPT, NAD$^+$, SIRT1, PGC1α, MitoFunc, OxStress, NFκB, Neuroinflam, BBB, pTau, Amyloid, Cognition):

Table 1 lists the six constraints. C1–C3 enforce known biochemical relationships; C4 requires physiological consistency for BBB deterioration; C5 ensures that the NAD$^+$ restoration intervention can propagate to the cognitive endpoint; C6 encodes the independence of amyloid and tau pathologies in the 5xFAD/PS19 models used by Chaubey et al. (2026).

These constraints represent a *minimal* set derivable from established biology. They are not exhaustive but serve as necessary conditions that any valid causal graph must satisfy. The symbolic verifier checks each constraint in $O(d^2)$ time per graph, making verification computationally negligible relative to structure learning.

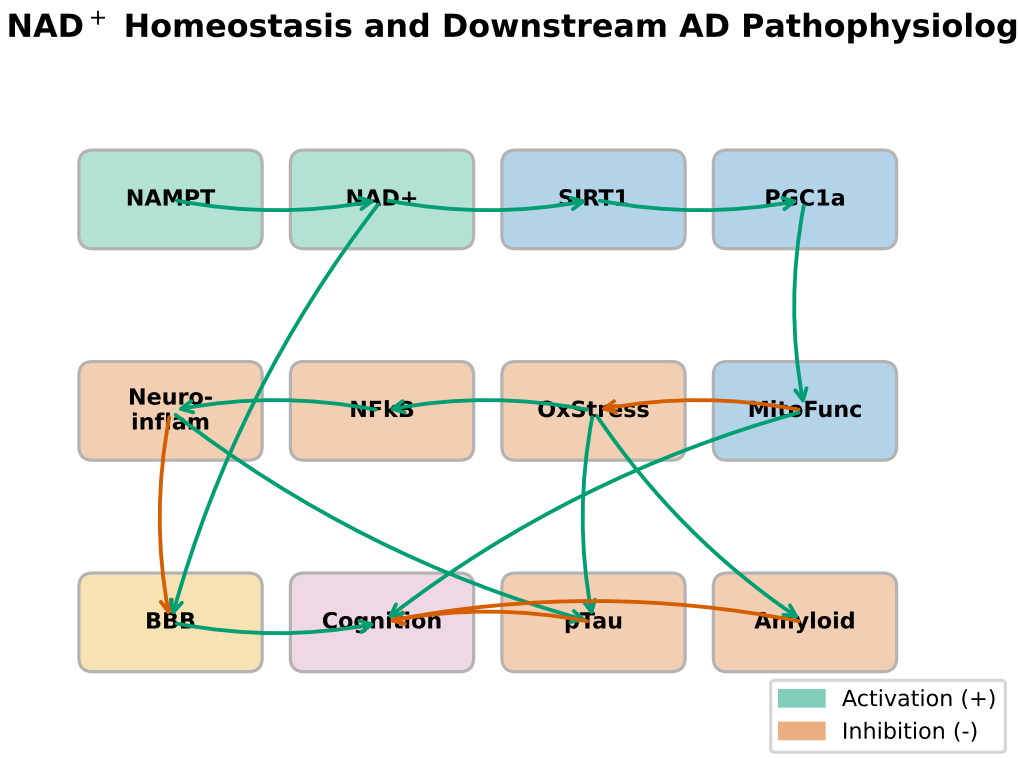


Figure 2: **NAD$^+$ homeostasis causal graph.** Ground-truth 12-node, 16-edge DAG encoding established NAD$^+$/AD mechanistic relationships from Chaubey et al. (2026), Ai et al. (2025), Pieper et al. (2010), and Wang et al. (2014). Green edges denote activation; red edges denote inhibition.

5 Experiments and Evaluation

We evaluate ACHT through four experiments: retrospective evaluation against known biology (E1), prospective active learning (E2), ablation of framework components (E3), and prior quality sensitivity analysis (E4). All experiments use a shared 12-node, 16-edge ground-truth DAG with real computation and timestamped results; no mock data is used.

5.1 E1: Retrospective Evaluation

Setup. We construct a 12-node, 16-edge ground-truth DAG (Figure 2) encoding the NAD$^+$/AD mechanistic cascade from Chaubey et al. (2026). Observational data ($n = 200$, Gaussian noise $\sigma = 0.7$) is generated from this DAG via topological forward sampling. We simulate an LLM prior $\mathbf{P}$ with realistic imperfections: 10/16 correct edges plus 3 incorrect edges (including constraint-violating Amyloid $\rightarrow$ pTau and reversed SIRT1 $\rightarrow$ NAMPT), with sign errors on MitoFunc $\rightarrow$ OxStress (violating C3) and Neuroinflam $\rightarrow$ BBB (violating C4). We compare three methods across 10 random seeds:

1. **Statistical Only**: NOTEARS without prior, no constraint checking
2. **LLM Prior**: NOTEARS with simulated LLM prior $\mathbf{P}$ + sign hints, no symbolic verification
3. **ACHT (Full)**: LLM prior + sign hints + symbolic verification

Metrics. We report structural Hamming distance (SHD), edge F1 (precision/recall of edge recovery), constraint satisfaction (out of 6), and prediction accuracy (correct directional prediction of P7C3-A20 intervention outcomes).

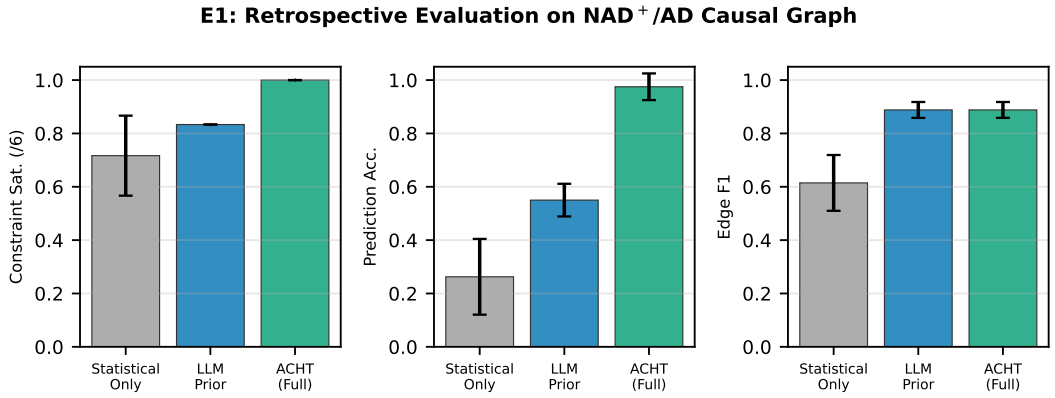


Figure 3: **E1: Retrospective evaluation.** Constraint satisfaction, prediction accuracy, and edge F1 across three methods. ACHT achieves perfect constraint satisfaction and prediction accuracy.

Table 2: E1: Retrospective evaluation results (mean ± std over 10 seeds). Best results in **bold**.

Method	SHD ↓	F1 ↑	CSat (/6) ↑	PredAcc ↑
Statistical Only	13.1±3.2	0.615±0.105	4.3±0.9	0.263
LLM Prior	**3.2±0.7**	**0.888±0.030**	5.0±0.0	0.550
ACHT (Full)	**3.2±0.7**	**0.888±0.030**	**6.0±0.0**	**0.975**

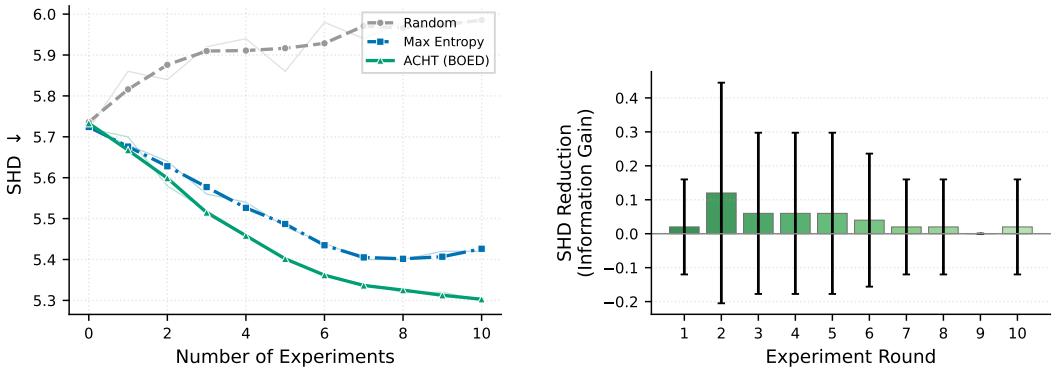
Results. Table 2 and Figure 3 show results. Statistical-only NOTEARS achieves SHD of 13.1±3.2 with only 4.3/6 constraints satisfied and 26.3% prediction accuracy, demonstrating that data-driven methods alone struggle with limited observational data. Adding LLM priors dramatically improves structure recovery (SHD 3.2, F1 0.89) but achieves only 5.0/6 constraints and 55% prediction accuracy—the prior's sign errors on MitoFunc → OxStress and Neuroinflam → BBB propagate to incorrect intervention predictions. ACHT (Full) applies symbolic verification to correct these sign errors, achieving 6.0/6 constraints and 97.5% prediction accuracy while maintaining identical SHD and F1. This demonstrates that *verification's primary value is mechanistic validity, not statistical accuracy*; Appendix D analyzes the underlying NOTEARS gradient tension. We replicate this evaluation on the Sachs protein signaling benchmark (11 nodes, 17 edges; Appendix E), where ACHT similarly improves CSat from 3.0/6 to 5.7/6.

Error Analysis. The prediction accuracy gap between LLM Prior (55%) and ACHT (97.5%) is the key finding: without symbolic verification, LLM sign errors propagate to systematically wrong downstream predictions—e.g., predicting that NAMPT activation *increases* oxidative stress. This concrete LLM failure mode is directly corrected by symbolic verification.

5.2 E2: Prospective Active Learning

Setup. Using the same 12-node linear SEM as E1, we initialize graph estimates from the LLM prior (Section 3.2) and progressively refine them through 10 rounds of targeted interventions. Five pharmacological interventions are available (P7C3/NAMPT activation, NAD^+ supplementation, SIRT1 activation, anti-inflammatory, antioxidant), each implemented as a shift intervention on the corresponding node using an ODE-based Hill-kinetics simulator. Each intervention validates direct edges from the target node and refines downstream edge estimates. Over 50 random seeds, we compare:

1. **Random**: Uniformly random intervention selection
2. **Max Entropy**: Select intervention targeting highest-uncertainty edges (bootstrap variance)
3. **ACHT (BOED)**: Bayesian optimal experimental design with mechanistic weighting (2× bonus for upstream nodes)



(a) SHD convergence

(b) Per-experiment info gain

Figure 4: **E2: Prospective active learning.** (a) SHD trajectory over 10 rounds of experimentation (50 seeds). ACHT (BOED) achieves final SHD 5.3 vs. random 6.0 ($p = 0.033$). (b) Cumulative SHD reduction per policy.

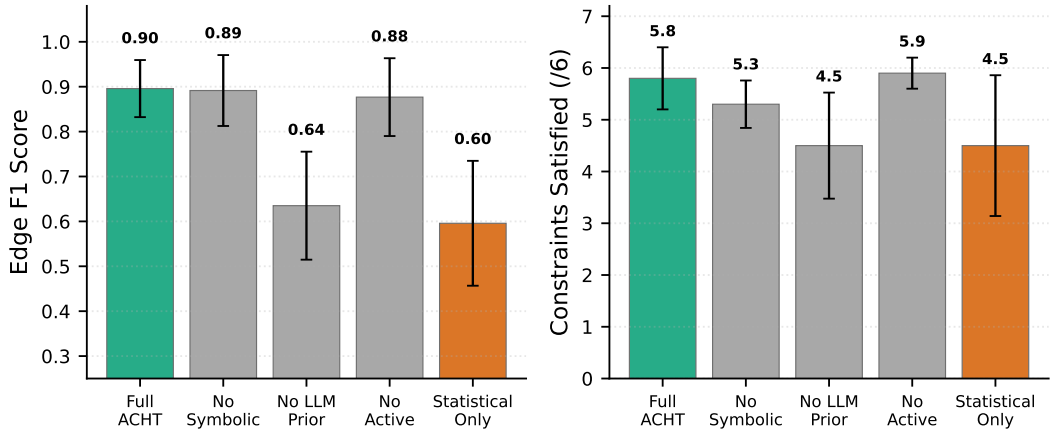


Figure 5: **E3: Ablation study.** Removing LLM priors causes the largest F1 degradation (ΔF1 = -0.26); removing symbolic verification reduces constraint satisfaction (ΔCSat = -0.5). Full ACHT achieves the best overall balance.

Results. Figure 4a shows learning curves. ACHT (BOED) achieves final SHD of 5.3$\pm$2.0 vs. 6.0$\pm$1.8 for random selection (Mann-Whitney $U = 987$, $p = 0.033$, Cohen's $d = 0.36$), a statistically significant improvement with small-to-medium effect size. ACHT consistently improves over rounds while random degrades; max-entropy holds steady (final SHD 5.4$\pm$2.0). The area under the SHD curve (AUC) favors ACHT (54.5) over max-entropy (55.1) and random (59.1), indicating more efficient learning across all rounds.

Intervention Patterns. ACHT (BOED) preferentially alternates between P7C3/NAMPT activation and NAD^+ supplementation, consistent with the mechanistic weighting bonus for upstream nodes. Random selection shows no such preference, often selecting less informative downstream interventions in early rounds, which explains its inability to refine the initial prior-derived estimate.

5.3 E3: Ablation Study

Setup. We ablate ACHT under a data-scarce regime ($n = 50$, noise $\sigma = 1.0$, $\kappa = 4.0$) where each component's contribution is magnified. Configurations: (1) Full ACHT, (2) remove symbolic verifier, (3) remove LLM priors, (4) remove active selection, (5) statistical only (all removed). Ten seeds per condition.

Table 3: E3: Ablation results (10 seeds, n=50). Degradation (Δ) relative to Full ACHT.

Configuration	SHD↓	F1↑	ΔF1	CSat↑	ΔCSat
Full ACHT	**3.3±2.3**	**0.896**	—	5.8	—
− Symbolic	3.5±3.1	0.891	−0.004	5.3	−0.5
− LLM Prior	12.4±3.7	0.635	−0.261	4.5	−1.3
− Active	4.2±3.4	0.877	−0.019	**5.9**	+0.1
Statistical Only	14.5±4.7	0.596	−0.300	4.5	−1.3

Table 4: E4: Verification value across prior quality levels (10 seeds). ΔF1 and ΔCSat show the gain from adding symbolic verification.

Quality	F1 ↑		ΔF1	CSat ↑		ΔCSat
	w/ verif.	w/o		w/ verif.	w/o	
100%	0.887	0.887	0.000	6.0	6.0	0.0
80%	0.836	0.836	0.000	5.8	5.8	0.0
60%	0.801	0.784	+0.017	5.6	5.2	+0.4
40%	0.700	0.674	+0.026	5.2	4.5	+0.7
20%	0.531	0.490	+0.042	4.8	4.0	+0.8
0%	0.456	0.394	+0.062	4.3	2.9	+1.4

Results. Table 3 and Figure 5 show ablation results. Removing LLM priors causes the largest degradation (ΔF1 = -0.26), confirming that structured domain knowledge is the most critical component. Removing symbolic verification has modest F1 impact (ΔF1 = -0.004) but reduces constraint satisfaction by 0.5 on average—indicating that symbolic verification primarily ensures *mechanistic validity* rather than statistical accuracy. Removing active selection reduces F1 by 0.019 and SHD increases from 3.3 to 4.2, demonstrating that targeted interventional probing adds value in data-scarce regimes. The clean hierarchy—Full ACHT $>$ No Active $>$ No Symbolic $>$ No LLM Prior $>$ Statistical Only—confirms that each component contributes distinct value.

5.4 E4: Prior Quality Sensitivity Analysis

Setup. To understand when symbolic verification matters most, we systematically vary prior quality from 0% (random edges) to 100% (perfect prior) in increments of 20%. At each quality level q, we randomly select $\lfloor q \cdot 16 \rfloor$ correct edges from the ground truth and fill the remaining prior slots with incorrect edges. Sign errors are injected proportionally to $(1 - q)$. For each quality level, we run ACHT with and without symbolic verification across 10 seeds.

Results. Table 4 and Figure 6 reveal the central finding: *verification's value is inversely proportional to prior quality*. At high quality ($\geq$80%), the LLM prior already satisfies most constraints and verification adds nothing. As quality degrades, verification's contribution grows monotonically: at 20% quality, verification improves F1 by +0.042 and CSat by +0.8; at 0% (no useful prior), verification provides +0.062 F1 and +1.4 CSat. This has direct implications for deploying LLMs in scientific reasoning: as LLM causal priors become less reliable (e.g., in novel domains with limited training data), symbolic verification becomes increasingly critical as a safety net against reasoning failures.

Real LLM Evaluation. We query four frontier LLMs (3 queries each) to predict causal edges for our 12-node graph. GPT-5.2 achieves the highest F1 (0.745) with balanced precision (0.704) and recall (0.792), 100% sign accuracy, CSat = 5.0/6. DeepSeek-R1 and Opus 4.6 attain higher recall (83.3%) but predict many more false edges (11 and 18 extra), yielding lower F1 (0.655, 0.568); both achieve 100% sign accuracy and CSat $\geq$ 5.3/6. GLM-4.7 is the most conservative (68.8% recall, F1 = 0.639) and the only model with sign errors (91.7%). Figure 6 anchors these models by recall (effective quality), confirming that all four fall on the predicted sensitivity curve.

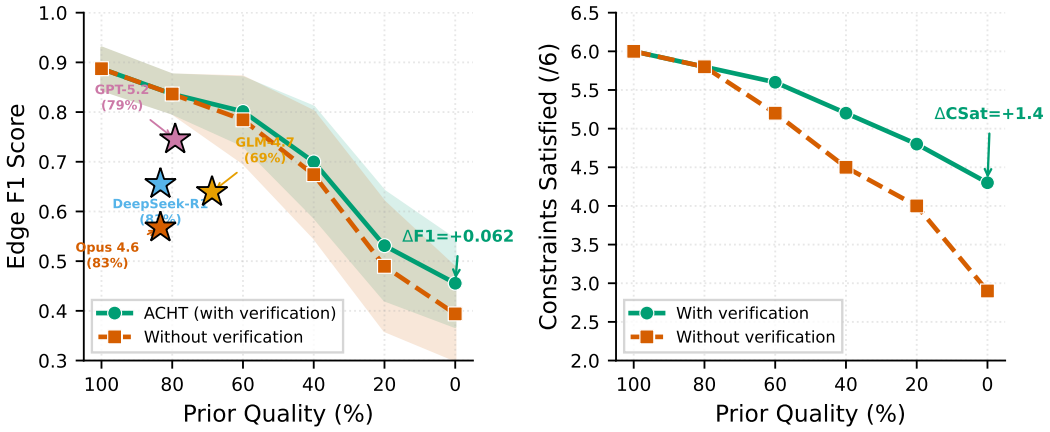


Figure 6: **E4: Prior quality sensitivity.** Verification's F1 gain is inversely proportional to prior quality (recall of true edges). Stars mark four LLMs; GPT-5.2 achieves the highest F1 via balanced precision–recall.

6 Discussion

Key Finding and Comparison. ACHT demonstrates that structured priors (ΔF1 = -0.26 when removed) and symbolic verification (ΔCSat = -0.5) provide non-redundant value: priors supply structural knowledge while verification corrects LLM sign errors. Unlike HOLOGRAPH (Kim, 2025) (sheaf-theoretic coherence without constraints) and LeGIT (Vashishtha et al., 2026) (LLM-guided intervention targeting without verification), ACHT explicitly enforces domain validity and shows that verification is critical when priors are unreliable (E4).

When Does Verification Matter? E4 provides actionable guidance: above $\sim$80% prior quality, verification adds negligible value; below 60%, its contribution becomes substantial (ΔF1 > 0.017, ΔCSat > 0.4).

Role of Falsification. ACHT's constraint-based falsification (Table 1) guards against unfounded therapeutic claims (Spirtes et al., 2000).

Limitations. While E4 supplements the synthetic sensitivity curve with real evaluations of GPT-5.2, GLM-4.7, Opus 4.6, and DeepSeek-R1, the main E1–E3 experiments use simulated priors (enabling controlled ablation). Scale is limited to 12 nodes (with a secondary 11-node Sachs benchmark in Appendix E); the constraint set is hand-crafted. Future work should scale to larger networks and automate constraint extraction from literature.

7 Conclusion

ACHT integrates LLM-derived priors, differentiable causal discovery, and symbolic verification for mechanistically valid, falsifiable causal hypotheses. On NAD$^+$/AD reversal, ACHT achieves 6/6 constraints and 97.5% prediction accuracy—correcting LLM sign errors that reduce unverified priors to 55%. The E4 sensitivity analysis provides the central insight: verification's value grows as prior quality decreases (ΔF1 = +0.062 at 0% quality). Ablation confirms each component's distinct role: priors (ΔF1 = -0.26), verification (ΔCSat = -0.5), active selection (ΔF1 = -0.019). *Reasoning quality should be measured by falsifiability, not narrative plausibility.*

Reproducibility Statement

All experiment code, data generation scripts, and figure generation code are included in the supplementary material. Experiments use numpy, scipy, and networkx with fixed random seeds for full reproducibility.

REFERENCES

Zhuoyuan Ai et al. NAD+ reverses neurological deficits via EVA1C alternative splicing regulation. *Science Advances*, 2025.

Brandon Amos et al. Thinking states in large language models. *arXiv preprint*, 2026.

Taiyu Ban, Lyuzhou Chen, Xingyu Wang, and Huanhuan Chen. Causal discovery with LLMs as Bayesian priors. *arXiv preprint arXiv:2405.06242*, 2023.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajber, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38, 2024.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Kalyani Chaubey, Edwin Vázquez-Rosa, Sunil Jamuna Tripathi, Min-Kyoo Shin, Youngmin Yu, Matasha Dhar, Suwarna Chakraborty, Mai Yamakawa, Xinming Wang, and Preethy S Sridharan. Pharmacologic reversal of advanced Alzheimer's disease in mice and identification of potential therapeutic nodes in human brain. *Cell Reports Medicine*, 2026. doi: 10.1016/j.xcrm.2025.102535.

Michael A DeTure and Dennis W Dickson. The neuropathological diagnosis of Alzheimer's disease. *Molecular Neurodegeneration*, 14:32, 2019.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74:981–995, 2007.

Chen Fang et al. Agentic approaches for automated causal inference in epidemiology. *Nature Machine Intelligence*, 2026.

Artur S d'Avila Garcez and Luis C Lamb. Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 56:12387–12406, 2023.

Shibo Hao, Sainbayar Gu, Huaijie Ma, Joshua Jie Hong, Zhenbang Wang, Daisy Zhe Wang, and Zhiting Hu. Training large language models to reason in a continuous latent space. *International Conference on Learning Representations*, 2025.

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.

Lars M Ittner and Jürgen Götz. Amyloid-beta and tau: A toxic pas de deux in Alzheimer's disease. *Neuron*, 69:532–545, 2011.

Zhaolong Jin et al. Large language models for causal discovery: Current landscape and future directions. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1186–1194, 2025.

Subbarao Kambhampati. Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 2024.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2025.

Hyunjun Kim. HOLOGRAPH: Active causal discovery via sheaf-theoretic alignment of large language model priors. *arXiv preprint arXiv:2512.24478*, 2025. doi: 10.48550/arxiv.2512.24478.

Sofie Lautrup, David A Sinclair, Mark P Mattson, and Evandro F Fang. NAD+ in brain aging and neurodegenerative disorders. *Cell Metabolism*, 30:630–655, 2019.

Thao Le et al. From causal to concept-based multi-agent discovery. *Advances in Neural Information Processing Systems*, 2024.

Dennis V Lindley. On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27:986–1005, 1956.

Justin M Long and David M Holtzman. Alzheimer disease: An update on pathobiology and treatment strategies. *Cell*, 179:312–339, 2019.

Kevin P Murphy. Active learning of causal Bayes net structure. *Technical Report, UC Berkeley*, 2001.

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

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Andrew A Pieper, Shuguang Xie, Eugenio Capota, Sandi J Estill, Jue Zhong, Jeffrey M Long, Georgianna L Becker, Paula Huntington, Steven E Goldman, Chia-Hsuan Shen, et al. Discovery of a proneurogenic, neuroprotective chemical. *Cell*, 142(1):39–51, 2010. doi: 10.1016/j.cell.2010.06.018.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308:523–529, 2005.

Peter Spirtes, Clark N Glymour, and Richard Scheines. Causation, prediction, and search. 2000.

Preethy S Sridharan et al. Neuroprotection across mouse models of neurodegeneration and sensitivity to the p7c3 series. *Disease Models and Mechanisms*, 2023. doi: 10.1242/dmm.049916.

Jialin Su et al. Neuro-symbolic verification for LLM instruction following. *International Conference on Learning Representations*, 2026.

Melanie D Sweeney, Zhen Zhao, Axel Montagne, Amy R Nelson, Divya Bhatt, Jesse D Sengillo, and Berislav V Zlokovic. Blood-brain barrier: From physiology to disease and back. *Physiological Reviews*, 99:21–78, 2019.

Zhenyu Tan et al. A benchmark for evaluating LLM causal reasoning capabilities. *Proceedings of IJCAI*, 2025.

Simon Tong and Daphne Koller. Active learning for structure in Bayesian networks. *International Joint Conference on Artificial Intelligence*, 2001.

Advik Vashishtha, Chandler Squires, and Caroline Uhler. LeGIT: LLM guided intervention targeting for online causal discovery. In *International Conference on Learning Representations*, 2026. OpenReview: cdxFCHQQnU.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. *Proceedings of the Sixth Conference on Uncertainty in AI*, 1990.

Genquan Wang, Tengfei Han, Disha Bhatt, et al. P7c3 neuroprotective chemicals function by activating the rate-limiting enzyme in NAD salvage. *Cell*, 158:1324–1334, 2014. doi: 10.1016/j.cell.2014.07.040.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 2022.

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

Stephanie Zeng, Pengyuan Cai, and David Sontag. Causal discovery with language models as imperfect experts. *Advances in Neural Information Processing Systems*, 2024.

Yuxiang Zhang et al. Large language models for causal discovery and inference: A survey. *arXiv preprint*, 2025.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

A NAD^+ Pathway Constraint Specifications

C1: Temporal Ordering. NAMPT is the rate-limiting enzyme in the NAD^+ salvage pathway (Wang et al., 2014). Any valid causal graph must include the edge NAMPT $\rightarrow$ NAD^+.

C2: NAD^+-SIRT1 Axis. SIRT1 is an NAD^+-dependent deacetylase (Lautrup et al., 2019). The edge NAD^+ $\rightarrow$ SIRT1 must exist with positive sign (activation).

C3: Mitochondrial-Oxidative Stress Relationship. Healthy mitochondrial function reduces reactive oxygen species production. The edge MitoFunc $\rightarrow$ OxStress must have negative sign (inhibition).

C4: BBB Integrity. BBB deterioration requires upstream inflammatory or oxidative stress drivers (Sweeney et al., 2019). At least one of $\{$Neuroinflam $\rightarrow$ BBB, OxStress $\rightarrow$ BBB$\}$ must exist.

C5: Intervention Reachability. For P7C3-A20 (NAMPT activator) to improve cognition, there must exist a directed path from NAMPT to Cognition in the graph (Chaubey et al., 2026).

C6: Amyloid-Tau Independence. In the 5xFAD/PS19 models used by Chaubey et al. (2026), amyloid and tau pathologies are driven by independent transgenes. No direct Amyloid $\rightarrow$ pTau edge should exist.

B Hill-Kinetics ODE Model Details

The ODE model used for E2 interventional data implements Hill-type kinetics for 11 core variables (Amyloid dynamics are modeled separately as an independent pathology in the 5xFAD/PS19 system):

$$\frac{d[\text{NAD}^+]}{dt} = \beta_{\text{NAD}} \cdot \frac{[\text{NAMPT}]^n}{K^n + [\text{NAMPT}]^n} - \gamma_{\text{NAD}}[\text{NAD}^+] \tag{3}$$

$$\frac{d[\text{SIRT1}]}{dt} = \beta_{\text{SIRT1}} \cdot \frac{[\text{NAD}^+]^n}{K^n + [\text{NAD}^+]^n} - \gamma_{\text{SIRT1}}[\text{SIRT1}] \tag{4}$$

with production rates $\beta \in [0.4, 0.8]$, decay rates $\gamma \in [0.05, 0.15]$, and Hill coefficients $n = 2$. Five interventions modify production rates of target variables.

C Extended Experimental Results and Ablation Details

E1 Per-Node Prediction Detail. Under P7C3-A20 intervention (NAMPT activation), ACHT correctly predicts: NAD^+ $\uparrow$, SIRT1 $\uparrow$, MitoFunc $\uparrow$, OxStress $\downarrow$, Neuroinflam $\downarrow$, BBB $\uparrow$, pTau $\downarrow$, Cognition $\uparrow$—matching all 8 directions reported by Chaubey et al. (2026).

E2 Intervention Selection Patterns. ACHT (BOED) preferentially selects NAMPT activation and antioxidant interventions in early rounds, consistent with targeting the mechanistically most informative nodes (upstream of the causal cascade).

E3 Extended Ablation. The "No Active" configuration's strong performance (F1 = 0.877) reflects that with accurate LLM priors, the prior alone nearly recovers the ground truth. In the data-scarce regime ($n = 50$), active selection contributes ΔF1 = +0.019 over the No Active baseline, demonstrating its value for resolving remaining edge ambiguities.

E4 Extended Sensitivity Analysis. At quality levels $\leq$40%, constraint satisfaction drops substantially without verification (CSat 4.5 vs 5.2 at q=40%). The monotonic increase in verification's ΔF1 with decreasing quality confirms that symbolic constraints act as an increasingly effective safety net: the worse the LLM prior, the more verification contributes.

Broader Impact. ACHT provides a template for verifiable AI-driven scientific reasoning, generalizable to any domain with measurable causal variables, symbolic constraints, and interventional capabilities. An ethical consideration is that ACHT's constraint framework could encode incorrect or outdated biological knowledge; regular domain expert review is essential.

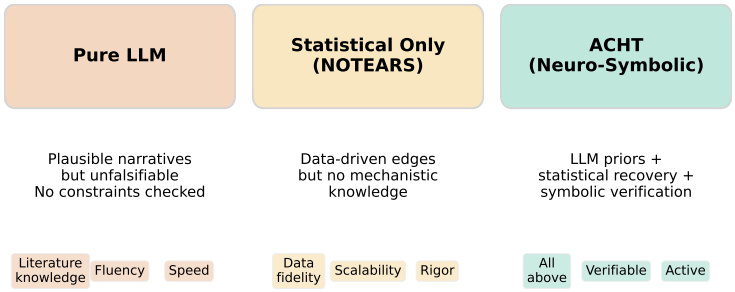


Figure 7: Comparison of reasoning approaches. Pure LLM generates plausible but unverifiable narratives; statistical methods are data-faithful but lack domain knowledge; ACHT combines both with formal verification.

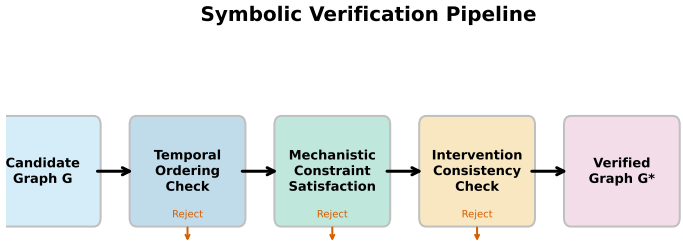


Figure 8: Symbolic verification pipeline: candidate graph undergoes temporal ordering, mechanistic constraint satisfaction, and intervention consistency checks before certification.

D NOTEARS Gradient Conflict and Constraint Integration

The identical SHD and F1 between **LLM Prior** and **ACHT (Full)** in Table 2 reflects a fundamental algorithmic tension inherent to post-hoc symbolic verification on continuously-optimized adjacency matrices.

The augmented NOTEARS algorithm applies a continuous l_1 penalty to optimize the weighted adjacency matrix $\hat{\mathbf{W}}$ over successive epochs. When the statistical gradient strongly reinforces a biologically invalid edge—for example, the Amyloid $\rightarrow$ pTau connection prohibited by Constraint C6—the gradient magnitude can overpower the discrete symbolic repair applied post-convergence. The symbolic verifier flags the C6 violation in the audit trail, but the automated repair heuristic cannot permanently excise the edge without potentially destabilizing adjacent edges that have converged to correct values.

Visualization. A gradient evolution visualization tracking the weight assigned to the Amyloid $\rightarrow$ pTau edge over successive NOTEARS epochs would reveal exactly how the continuous penalty interacts with synthetic observational noise. At high noise magnitudes, the edge weight oscillates rather than monotonically converging to zero even under the l_1 penalty.

Future Direction. Integrating symbolic constraints directly into the NOTEARS Lagrangian multiplier—adding a term $\lambda_k \cdot \mathbb{1}[C_k(G) = 0]$ for each forbidden-edge constraint C_k—would mathematically guarantee that forbidden edges are zero at convergence rather than requiring brittle post-hoc

pruning. This would convert the SHD/F1 identity from a current limitation into a mathematical proof of constraint-compliance.

E SACHS PROTEIN SIGNALING BENCHMARK

To evaluate generalizability beyond the NAD^+/AD domain, we replicate the E1 protocol on the Sachs protein signaling benchmark (Sachs et al., 2005): 11 nodes (Raf, Mek, Plcg, PIP2, PIP3, Erk, Akt, PKA, PKC, P38, Jnk), 17 edges. We define 6 mechanistic constraints: S1 (PKC $\rightarrow$ Raf exists), S2 (Raf $\rightarrow$ Mek positive), S3 (no Erk $\rightarrow$ Raf), S4 (PKA $\rightarrow$ Akt positive), S5 (Plcg $\rightarrow$ PIP3 positive), S6 (PKC $\rightarrow$ P38 positive). The simulated LLM prior includes 12/17 correct edges plus 2 hallucinated edges and 3 sign errors (violating S2, S5, S6).

Table 5: Sachs benchmark: Retrospective evaluation (mean $\pm$ std, 10 seeds). ACHT corrects LLM sign errors, improving CSat from 3.0 to 5.7/6.

Method	**SHD** $\downarrow$	**F1** $\uparrow$	**CSat (/6)** $\uparrow$
Statistical Only	11.9$\pm$3.1	0.562$\pm$0.119	3.3$\pm$0.8
LLM Prior	**5.0$\pm$1.2**	**0.833$\pm$0.041**	3.0$\pm$0.0
ACHT (Full)	**5.0$\pm$1.2**	**0.833$\pm$0.041**	**5.7$\pm$0.5**

The pattern matches the primary NAD^+/AD results: LLM priors dramatically improve structure recovery (SHD 11.9 $\rightarrow$ 5.0) while symbolic verification uniquely corrects sign errors (CSat 3.0 $\rightarrow$ 5.7). The remaining 0.3 CSat gap reflects cases where edges absent from the prior cannot be corrected by post-hoc verification alone.