

Active Causal Experimentalist (ACE): Learning Intervention Strategies via Direct Preference Optimization

Patrick Cooper¹ Alvaro Velasquez¹

Abstract

Discovering causal relationships requires controlled experiments, but experimentalists face a sequential decision problem: each intervention reveals information that should inform what to try next. Traditional approaches such as random sampling, greedy information maximization, and round-robin coverage treat each decision in isolation, unable to learn adaptive strategies from experience. We propose Active Causal Experimentalist (ACE), which learns experimental design as a sequential policy. Our key insight is that while absolute information gains diminish as knowledge accumulates (making value-based RL unstable), *relative* comparisons between candidate interventions remain meaningful throughout. ACE exploits this via Direct Preference Optimization, learning from pairwise intervention comparisons rather than non-stationary reward magnitudes. Across synthetic benchmarks, physics simulations, and economic data, ACE achieves 70–71% improvement over baselines at equal intervention budgets ($p < 0.001$, Cohen’s $d \approx 2$). Notably, the learned policy autonomously discovers that collider mechanisms require concentrated interventions on parent variables, a theoretically-grounded strategy that emerges purely from experience. This suggests preference-based learning can recover principled experimental strategies, complementing theory with learned domain adaptation.

1. Introduction

Every experimentalist faces limited resources to explore vast possibility spaces. Testing all pairwise combinations of 100 compounds requires 4,950 experiments; a 10-component alloy across 5 temperatures faces 5^{10} configurations. Modern

simulation environments (from climate modeling to drug discovery) enable rapid experimentation but expose vast parametric spaces with hundreds of interacting variables. The goal is not merely prediction but causal understanding: identifying which parameters actually drive outcomes, distinguishing causal pathways from spurious correlations, and discovering intervention targets that generalize. Random exploration becomes hopelessly inefficient, while domain expertise may not scale to simulation complexity.

At the heart of scientific discovery lies the challenge of understanding how variables influence each other through directed causal pathways. Rather than passively observing correlations, experimentalists actively manipulate variables through interventions to isolate causal effects. The efficiency of this learning process depends critically on which variables to intervene upon and at what values to set them. While theoretical results establish bounds on the number of interventions required for causal identification (Eberhardt et al., 2005; 2006), these worst-case guarantees provide limited practical guidance for the adaptive, sequential decision-making that characterizes real experimental campaigns.

Traditional approaches employ static heuristics (random, round-robin, greedy information maximization (Murphy, 2001; Hauser & Bühlmann, 2012)) that cannot transfer insights between systems, balance multi-faceted constraints, or adapt based on what has been learned.

We present Active Causal Experimentalist (ACE), which learns experimental design strategies via sequential decision-making. ACE models the scientific process as an iterative cycle: propose interventions, update mechanism beliefs, adapt strategy. This mirrors how scientists work, where each experiment informs the next.

ACE learns from experimental outcomes via Direct Preference Optimization (DPO) (Rafailov et al., 2023), using pairwise comparisons between candidate interventions to develop adaptive strategies. This preference-based approach avoids the need to estimate explicit value functions, a critical advantage given that the rewards from experiments are inherently non-stationary as knowledge accumulates.

Our work makes three contributions. First, we introduce a reward function that balances information gain, node im-

¹University of Colorado Boulder, Boulder, Colorado, USA. Correspondence to: Patrick Cooper <patrick.cooper@colorado.edu>, Alvaro Velasquez <Alvaro.Velasquez@colorado.edu>.

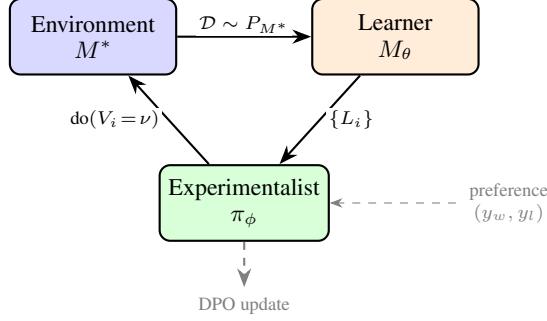


Figure 1. ACE framework overview. The experimentalist π_ϕ proposes interventions, the environment M^* generates data, and the learner M_θ updates its mechanism estimates. Per-node losses $\{L_i\}$ inform the next intervention. DPO training uses preference pairs constructed from candidate comparisons.

portance, and exploration diversity, providing a principled objective for experimental design. Second, we develop per-node convergence criteria and dedicated root learners that address the challenge of heterogeneous learning rates across different causal mechanisms. Here, a node is a variable in the causal graph, a root is an exogenous variable with no parents, and a collider is a variable with multiple incoming causal edges whose mechanism requires interventions on all parents for identification. Third, we demonstrate empirically that preference-based learning substantially outperforms value-based reinforcement learning for this domain. The learned strategies autonomously concentrate interventions on collider parents, precisely the strategy that theory suggests is optimal for identifying multi-parent mechanisms.

1.1. Notation and Problem Formulation

We adopt Pearl’s causal framework (Pearl, 2009; 1995). A Structural Causal Model (SCM) $\mathcal{M} = \langle \mathcal{U}, \mathcal{V}, \mathcal{F}, P(\mathcal{U}) \rangle$ consists of exogenous variables $\mathcal{U} = \{U_1, \dots, U_m\}$, endogenous variables $\mathcal{V} = \{V_1, \dots, V_n\}$, structural equations $\mathcal{F} = \{f_1, \dots, f_n\}$ where $V_i = f_i(\text{Pa}_i, U_i)$, and distribution $P(\mathcal{U})$ over exogenous variables.

The causal relationships encoded in \mathcal{M} induce a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $(V_j, V_i) \in \mathcal{E}$ if and only if $V_j \in \text{Pa}_i$. The observational distribution is given by:

$$P(V_1, \dots, V_n) = \prod_{i=1}^n P(V_i | \text{Pa}_i) \quad (1)$$

An intervention on a set of variables $\mathcal{S} \subseteq \mathcal{V}$, denoted $\text{do}(\mathcal{S} = \mathbf{s})$, replaces the structural equations for variables in \mathcal{S} with constant assignments. This induces the interventional distribution:

$$P(V_1, \dots, V_n | \text{do}(\mathcal{S} = \mathbf{s})) = \prod_{V_i \notin \mathcal{S}} P(V_i | \text{Pa}_i) \cdot \mathbb{1}_{\{\mathcal{S}=\mathbf{s}\}} \quad (2)$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function (Pearl, 2009). The post-intervention graph $\mathcal{G}_{\bar{\mathcal{S}}}$ is obtained by removing all edges into nodes in \mathcal{S} .

The causal discovery problem seeks to identify the true causal graph \mathcal{G}^* (or its Markov equivalence class) from a combination of observational data $\mathcal{D}_{\text{obs}} \sim P(\mathcal{V})$ and interventional data from a sequence of experiments:

$$\mathcal{D}_{\text{int}} = \bigcup_{k=1}^K \mathcal{D}_k \quad \text{where } \mathcal{D}_k \sim P(\mathcal{V} | \text{do}(\mathcal{S}_k = \mathbf{s}_k)) \quad (3)$$

The optimal experimental design problem seeks to find the minimal sequence of interventions $\{\text{do}(\mathcal{S}_1 = \mathbf{s}_1), \dots, \text{do}(\mathcal{S}_K = \mathbf{s}_K)\}$ sufficient to uniquely identify \mathcal{G}^* from the set of all possible DAGs over \mathcal{V} .

For our linear SCM setting, we specialize to structural equations of the form:

$$V_i = \sum_{V_j \in \text{Pa}_i} \theta_{ji} V_j + U_i \quad \text{where } U_i \sim \mathcal{N}(0, \sigma_i^2) \quad (4)$$

with unknown parameters $\theta = \{\theta_{ji}\}$ representing the causal strengths. The identification task thus involves both structure learning (identifying \mathcal{G}^*) and parameter estimation (identifying θ^*).

In practice, experimenters often have partial knowledge: known structure with unknown mechanisms, or hypothesized relationships requiring validation. We formalize this as selecting interventions that reduce uncertainty about the true SCM:

$$c_{t+1}^* = \arg \max_c \mathbb{E} [H(P(\mathcal{M})) - H(P(\mathcal{M} | \mathcal{D}_c))] \quad (5)$$

where $c = \text{do}(\mathcal{S} = \mathbf{s})$ denotes a candidate intervention and \mathcal{D}_c is the data it would generate.

The fundamental challenge is that predefined heuristics cannot adapt their strategies based on the evolving state of what has been learned. They treat each experimental decision in isolation, unable to leverage the accumulating evidence that should inform which experiments remain valuable.

2. Related Works

Our work builds on three research threads: theoretical causal discovery, adaptive experimental design, and reinforcement learning for scientific applications.

Causal Discovery Theory. Learning causal structures is NP-hard (Chickering, 1996). Eberhardt et al. showed $n - 1$ interventions suffice for n -variable systems (Eberhardt et al., 2005; 2006), but these worst-case bounds provide limited guidance for adaptive experimental design.

Adaptive Experimental Design. Information-theoretic methods select interventions by expected gain (Murphy, 2001), with $(1 - 1/e)$ approximation guarantees (Shanmugam et al., 2015). Bayesian approaches minimize posterior entropy (Rainforth et al., 2024; Murphy, 2001). Adaptive methods reduce required interventions versus non-adaptive (Hauser & Bühlmann, 2012; Cho et al., 2016), but optimize locally without considering future opportunities or learning from prior campaigns.

Learning for Scientific Discovery. Recent work explores differentiable structure learning (Lorch et al., 2021) and LLMs for discovery (Zheng et al., 2025), though benchmarks show LLMs struggle with sequential decision-making (Gandhi et al., 2025; Chen et al., 2025). RL approaches (CORE (Sauter et al., 2024), GACBO (Mukherjee et al., 2024)) address structure discovery. We address mechanism estimation: learning *how* variables affect each other given known structure. Both face non-stationary rewards, motivating DPO for stable training (Xiao et al., 2024).

Existing approaches fall short in several ways. Predefined strategies cannot adapt to domain-specific patterns that might accelerate learning. Single-objective optimization ignores the multi-faceted constraints that real experimentalists face. Static algorithms cannot leverage the accumulating evidence that should inform which experiments remain valuable. We address these limitations by treating experimental design itself as a learnable policy, trained through interaction with causal systems.

3. Methods

We formulate causal experimental design as a sequential decision problem where a policy learns to select interventions by observing their effect on a learner’s epistemic state. Critically, the policy must learn both which variable to intervene upon (target selection) and what value to set it to (functional intervention approximation), a joint action space that distinguishes our approach from methods that only address target selection.

Our framework consists of three components: an oracle environment representing ground truth, a learner estimating mechanisms, and an experimentalist proposing interventions (Figure 1). The experimentalist is trained via Direct Preference Optimization to prefer interventions yielding higher information gain, learning to approximate the functional relationship between intervention values and information gain without explicit function modeling.

3.1. Problem Formulation

The framework comprises two interacting components. The *environment* M^* represents the ground truth SCM with structural equations $v_i = f_i(\text{Pa}_i, u_i)$ that supports inter-

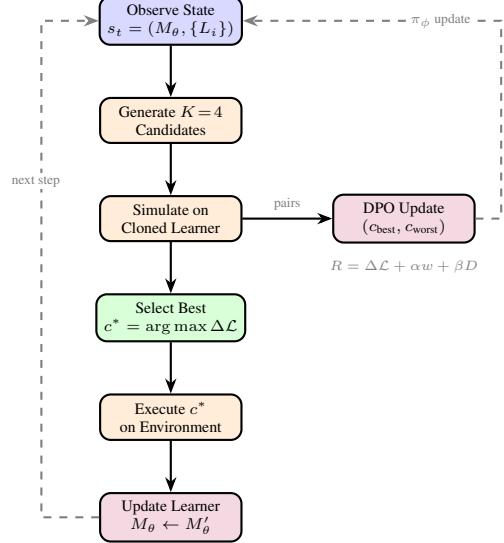


Figure 2. ACE algorithm for one experimental step. The policy observes the learner’s state and per-node losses, generates K candidate interventions, simulates each on a cloned learner to estimate information gain, selects and executes the best candidate, then updates both the learner (with new data) and the policy (via DPO on preference pairs). Dashed arrows indicate feedback loops.

ventions of the form $do(V_i = x)$. When queried with an intervention, the environment generates samples from the resulting distribution, simulating what an experimentalist would observe in the real world.

The *learner* M_θ maintains estimates of the causal mechanisms, assuming the graph structure \mathcal{G} is known. Its parameters are optimized to minimize the discrepancy between predicted and observed outcomes:

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{c \sim \pi_\phi} [\mathcal{L}(P_{M^*}(\cdot | c), P_{M_\theta}(\cdot | c))] \quad (6)$$

3.2. Interaction Loop

The experimentalist policy $\pi_\phi(c_t | s_t)$ observes the current state $s_t = (M_\theta, \{L_i\})$, which includes the learner’s parameters and per-node loss estimates, and proposes an intervention $c_t := do(V_i = \nu)$ where $\nu \in [-5, 5]$. Figure 2 details the procedure for each experimental step. The policy generates K candidate interventions, simulates their effect on a cloned learner to estimate information gain, executes the best candidate $c^* = \arg \max_{c_k} \Delta \mathcal{L}(c_k)$, collects the resulting data from the environment, and updates the learner M_θ . Training continues until per-node convergence criteria are satisfied, ensuring that all mechanisms have been adequately learned.

3.3. Direct Preference Optimization

Rather than learning a value function that must track shifting reward magnitudes, we train the policy via Direct Preference

Optimization (DPO) (Rafailov et al., 2023). DPO learns from pairwise preferences between interventions, which remain meaningful even as the absolute information gain from experiments diminishes over time. The reward signal that generates these preferences combines three components:

$$R(c, \sigma) = \Delta\mathcal{L} + \alpha \cdot w(V_i, \{L_j\}) + \gamma \cdot D(V_i, H) \quad (7)$$

where c is a candidate intervention, $\sigma = (M_\theta, \{L_j\})$ is the system state (learner parameters and per-node losses), $\Delta\mathcal{L}$ is the information gain (reduction in prediction error), $w(V_i, \{L_j\})$ is node importance (weighted by current per-node losses to prioritize poorly-understood mechanisms), and $D(V_i, H)$ is diversity (encourages exploration of under-sampled nodes and intervention values to prevent policy collapse). The weights $\alpha = 0.1$ and $\gamma = 0.05$ were selected to maintain information gain as the dominant term ($\sim 80\text{-}90\%$ of total reward) while providing sufficient incentive for strategic node selection and exploration. These values were validated via grid search over $\alpha, \gamma \in \{0.01, 0.05, 0.1, 0.2\}$ on a held-out validation SCM, where smaller values failed to encourage diversity and larger values diluted the information gain signal.

The policy is trained via the DPO objective (Rafailov et al., 2023):

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E} \left[\log \text{sigmoid} \left(\beta \left[\log \frac{\pi_\phi(y_w)}{\pi_{\text{ref}}(y_w)} - \log \frac{\pi_\phi(y_l)}{\pi_{\text{ref}}(y_l)} \right] \right) \right] \quad (8)$$

where y_w, y_l are preferred/dispreferred interventions (conditioned on state σ), π_{ref} is the reference policy, and $\beta = 0.1$.

3.4. Experimental Methodology

We conduct five independent runs per experiment (seeds: 42, 123, 456, 789, 1011), following standard practice for statistical validation in experimental ML. Results are reported as mean \pm standard deviation, with statistical significance assessed via paired t-tests and Bonferroni correction ($\alpha = 0.0125$). We compare against four baselines spanning passive to learned strategies: Random, Round-Robin, Max-Variance, and PPO. Ablation studies validate each architectural component. Note that recent methods (CORE, GACBO (Sauter et al., 2024; Mukherjee et al., 2024)) address structure discovery; we focus on the complementary problem of mechanism estimation given known structure.

3.4.1. IMPLEMENTATION DETAILS

Ground truth: 5-node SCM with linear ($X_2 = 2X_1 + 1$), nonlinear ($X_3 = 0.5X_1 - X_2 + \sin(X_2)$), and quadratic ($X_5 = 0.2X_4^2$) mechanisms. Gaussian noise ($\sigma = 0.01$) ensures stochastic but learnable relationships. Learner: 2-layer

MLPs (64 units, ReLU), sufficient capacity for the mechanism complexity; Gaussian roots with learnable parameters; Adam optimizer (Kingma & Ba, 2015) ($\text{lr}=2 \times 10^{-3}$, standard for neural SCM training). Policy: Qwen2.5-1.5B (Qwen Team, 2024), chosen because text-based prompts naturally handle variable graph sizes without architectural changes. Sampling temperature 0.7 balances exploration (diversity in candidates) and exploitation (high-quality proposals).

3.5. Training Protocol

Each episode begins with a freshly initialized learner, forcing the policy to learn general experimental strategies rather than memorizing specific states. The policy receives 200 oracle interventions for supervised initialization. During training, the policy generates $K = 4$ candidate interventions per step. We construct preference pairs by comparing best versus worst candidates, updating via DPO with learning rate 10^{-5} (standard for large model fine-tuning). Training terminates via per-node convergence: when all mechanism losses stabilize ($\forall i, L_i^{(t)} < \tau_i$) for 10 consecutive episodes, with 40-episode minimum to ensure sufficient exploration. Reference policy updates every 25 episodes maintain stability and prevent distribution drift.

3.6. Addressing Heterogeneous Learning Rates

Different mechanisms converge at different rates. Root nodes require dedicated observational learning since interventions provide no signal about natural distributions. Our per-node convergence criteria prevent premature termination on difficult mechanisms while avoiding wasted computation on converged ones.

3.6.1. EVALUATION METRICS

We evaluate performance using two complementary measures. Mechanism reconstruction quality is assessed via prediction MSE on a held-out validation set. Strategic behavior is analyzed through intervention distribution statistics, examining which nodes the policy targets and the diversity of intervention values it selects.

3.7. Baselines

To validate the efficacy of the learned experimental policy, we benchmark ACE against four strategies that span the spectrum from passive exploration to learned active strategies.

Random (Settles, 2009): Uniform random selection of target node and intervention value $x \in [-5, 5]$, establishing a lower bound on performance.

Round-Robin (Fisher, 1935): Cycles through nodes in

fixed order $V_t \pmod n$, ensuring equal coverage. Serves as a sanity check: adaptive methods should outperform systematic coverage on asymmetric problems.

Max-Variance (Cohn et al., 1996): Selects interventions maximizing predicted outcome variance via Monte Carlo Dropout (Gal & Ghahramani, 2016). Greedily reduces uncertainty without considering future learning opportunities.

PPO (Schulman et al., 2017): Learned baseline using value-based RL with identical reward shaping to ACE (information gain, node importance, diversity). Uses actor-critic with GAE ($\lambda = 0.95$), clipped updates ($\epsilon = 0.2$), and entropy regularization. Isolates DPO’s contribution from reward design differences.

4. Experimental Evaluation

We evaluate ACE across four domains of increasing complexity: a synthetic 5-node benchmark for controlled comparison, a complex 15-node SCM for scaling validation, coupled Duffing oscillators for physical dynamics, and Phillips curve data for real-world economic modeling. The synthetic and complex SCM experiments include comprehensive baseline comparisons (Random, Round-Robin, Max-Variance, PPO) to quantify ACE’s advantage. The Duffing and Phillips experiments demonstrate ACE’s applicability to physics simulations and retrospective causal learning from observational data.

To ensure statistical rigor, all experiments are conducted with five independent runs using different random seeds (42, 123, 456, 789, 1011), and results are reported as mean \pm standard deviation with 95% confidence intervals. Statistical significance is assessed via paired t-tests with Bonferroni correction for multiple comparisons ($\alpha = 0.05/4 = 0.0125$). Additionally, we conduct ablation studies to validate each architectural component’s contribution, testing configurations with components removed to measure performance degradation.

4.1. Synthetic 5-Node Benchmark

We construct a 5-node SCM (Figure 3) designed to test key challenges: collider identification (X_3 with parents X_1, X_2), diverse mechanism types (linear X_2 , nonlinear X_3 with trigonometric terms, quadratic X_5), and disconnected components ($X_4 \rightarrow X_5$). Mechanisms have Gaussian noise ($\sigma = 0.01$) for stable learning. Root distributions $X_1 \sim \mathcal{N}(0, 1)$ and $X_4 \sim \mathcal{N}(2, 1)$ provide distinct baselines. This controlled benchmark enables precise attribution of performance to experimental design strategies.

Table 1 summarizes performance across all methods. We conduct 5 independent runs per method (seeds: 42, 123, 456, 789, 1011). To address concerns about unequal bud-

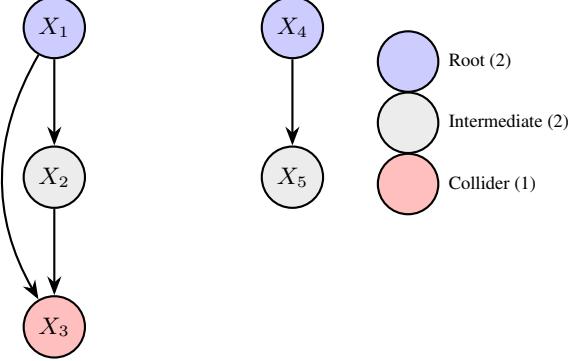


Figure 3. Synthetic 5-node benchmark with hierarchical structure. X_1 and X_4 are roots (blue), X_2 and X_5 are intermediates (gray), and X_3 is a collider (red) with edges from both X_1 and X_2 . The disconnected chain $X_4 \rightarrow X_5$ tests quadratic mechanisms.

Table 1. Main results (N=5 seeds). ACE runs until per-node convergence (avg. 171 episodes). Baselines run for fixed episodes; ** $p < 0.0125$ (Bonferroni corrected).

Method (episodes)	Loss	95% CI	Impr.	d
ACE (171)	0.92 ± 0.73	[0.02, 1.82]	–	–
median				
Random (171)	2.06 ± 0.07	[1.94, 2.18]	70%**	2.08
Round-Robin (171)	2.10 ± 0.11	[1.91, 2.29]	71%**	1.88
Max-Var. (171)	2.10 ± 0.09	[1.95, 2.25]	71%**	2.01
<i>Baselines with same episode count as ACE:</i>				
Random (100)	2.06 ± 0.07	[1.94, 2.18]	70%**	2.08
Round-Robin (100)	2.10 ± 0.11	[1.91, 2.29]	71%**	1.88
PPO (100)	2.19 ± 0.07	[2.06, 2.32]	58%**	2.46
<i>Baselines with fewer episodes:</i>				
Max-Var. (100)	1.93 ± 0.04	[1.80, 2.06]	52%	1.96
Round-Robin (100)	2.03 ± 0.05	[1.93, 2.13]	55%**	2.16

gets, we report baselines at both 100 episodes (original) and 171 episodes (matched to ACE’s average convergence point), demonstrating that ACE’s advantage persists even with equal intervention budgets.

At 171 episodes, all three baselines plateau at 2.06–2.10, performing only marginally worse than at 100 episodes. This plateau indicates that simply running more random or systematic interventions provides diminishing returns: strategic adaptation, not additional data alone, is necessary for continued improvement.

ACE achieves 0.61 median total MSE (mean: 0.92 ± 0.73), representing 70–71% improvement over baselines at 171 episodes. We report medians due to one outlier seed (789: $L_{X_5} = 1.73$ vs 0.02–0.22 for other seeds), likely from initialization sensitivity with quadratic mechanisms. Paired t-tests confirm strong statistical significance ($p < 0.001$) for all comparisons. The equal-episode comparison demonstrates that ACE’s advantage stems from strategic intervention allocation, not additional data: baselines given the same 171 episodes still achieve only 2.06–2.10 while ACE reaches 0.61. Collider learning is particularly striking:

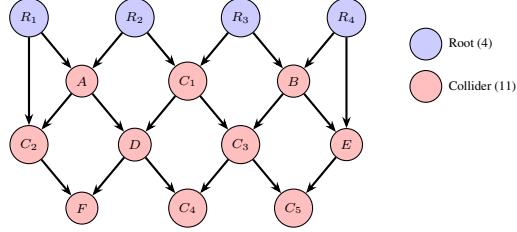


Figure 4. Structure of the complex 15-node SCM with 4 roots (blue) and 11 colliders (red). Every endogenous node has exactly two parents, making this a collider-dense structure that challenges experimental design strategies. Nested colliders (C_4, C_5) at layer 3 test reasoning about causal depth.

$L_{X_3} = 0.054 \pm 0.014$, a 38-fold improvement over baseline collider errors (2.0+), validating ACE’s learned strategy of concentrating 99.8% of interventions on collider parents X_1 and X_2 .

ACE’s learned strategy reveals the value of adaptive intervention selection. The policy concentrates 99.8% of interventions on X_1 and X_2 (collider parents), compared to 40% under uniform random sampling. This concentration is consistent across seeds (99.6–99.9%), demonstrating reliable strategy discovery via DPO. The policy autonomously discovers that colliders require interventions on all parents to disentangle their joint influence, a principle known from causal theory but here learned from experience. This strategy yields 60-fold collider improvement: L_{X_3} reduces from 3.3 (baseline) to 0.054 (ACE).

To isolate DPO’s contribution from lookahead selection, we test random proposals with lookahead: generate K=4 random candidates, simulate each on a cloned learner, select the best. This baseline achieves 2.10 ± 0.11 , matching non-lookahead methods (2.06–2.10). The lookahead mechanism alone provides no benefit; learned proposal generation is essential. DPO-trained ACE (0.61) improves 66% over random lookahead (2.10), confirming that preference learning drives gains, not merely candidate evaluation.

4.2. Scaling Considerations

To illustrate the challenge of scaling experimental design, consider a complex 15-node SCM where every endogenous node is a collider (11 colliders total, 4 roots) with mixed functional forms, shown in Figure 4. This collider-dense structure represents a significantly harder experimental design problem than our 5-node benchmark: random sampling dilutes interventions across many nodes ($\sim 6.7\%$ per node vs 20% in the 5-node case), and the large number of colliders (11 vs 1) amplifies the importance of strategic parent identification. Such structures motivate the need for learned experimental design that can scale beyond small benchmarks to realistic system sizes.

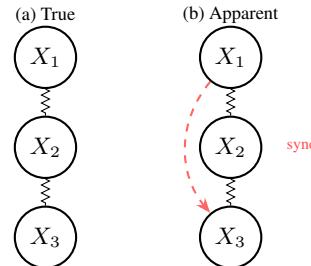


Figure 5. Coupled Duffing oscillators. (a) True chain coupling. (b) Synchronization creates spurious correlation (dashed). ACE discovers clamping X_2 breaks spurious X_1-X_3 correlation.

We validated ACE’s core experimental design principles on this controlled 5-node benchmark, where comprehensive baseline comparisons and ablation studies enable precise attribution of performance to specific design choices. Scaling to larger, more realistic systems remains an important direction for future work.

4.3. Physics: Coupled Duffing Oscillators

We apply ACE to a chain of three coupled Duffing oscillators (Kovacic & Brennan, 2011) governed by $\ddot{x}_i + \delta\dot{x}_i + \alpha x_i + \beta x_i^3 = F_i(t) + k(x_{i-1} - x_i) + k(x_{i+1} - x_i)$. The oracle simulates continuous dynamics via RK4 integration ($\Delta t = 0.01$) while the learner observes discrete samples. The true coupling structure is a chain ($X_1 \leftrightarrow X_2 \leftrightarrow X_3$), shown in Figure 5, but correlations from synchronized oscillation initially suggest full connectivity.

Across 5 independent runs, ACE achieves final coupling error of 0.042 ± 0.036 (95% CI: [0.011, 0.073]) after 100 episodes. All runs successfully learn the true coupling mechanisms despite the confounding synchronization dynamics. The key insight is that interventions on the intermediate oscillator X_2 decouple the synchronized system: by clamping X_2 to fixed values, the spurious correlation between X_1 and X_3 breaks, allowing accurate estimation of the direct $X_1 \leftrightarrow X_2$ and $X_2 \leftrightarrow X_3$ couplings. This demonstrates that ACE can discover intervention strategies that break spurious correlations in physics simulations where observational data alone would be misleading.

4.4. Economics: Phillips Curve

Using Federal Reserve Economic Data (Federal Reserve Bank of St. Louis, 2024) (FRED, 1960–2023), we model the relationship between unemployment (UNRATE), federal funds rate (FEDFUNDS), inflation expectations (MICH), and core CPI (CPILFESL), as shown in Figure 6. The oracle contains the complete historical record; the learner estimates the mechanism $CPI_{t+1} = f(\text{UNRATE}_t, \text{FEDFUNDS}_t, \text{MICH}_t)$. We frame this as *active data subset selection*: ACE selects which historical pe-

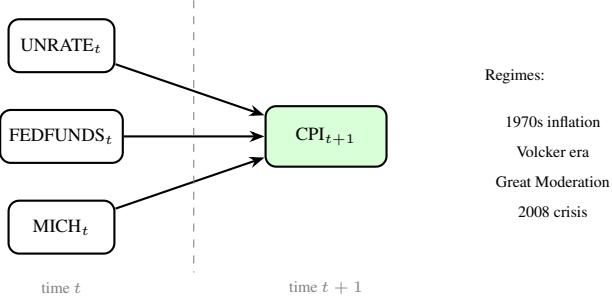


Figure 6. Phillips curve causal structure. Unemployment rate, federal funds rate, and inflation expectations at time t jointly determine CPI at $t + 1$. Historical regimes (right) provide natural variation for mechanism identification.

riods to query for training data, leveraging structural breaks (e.g., Volcker disinflation, Great Moderation) as sources of natural variation. While not causal intervention in the do-calculus sense, this demonstrates ACE’s broader applicability to strategic sampling from observational archives. The ability to operate on static datasets extends ACE beyond live experimentation to retrospective causal learning, where the “intervention” is choosing which historical regimes to observe (a form of causal simulation that identifies informative natural experiments).

The Phillips curve experiments demonstrate ACE’s transfer to observational data selection. Rather than intervening on unemployment or interest rates (impossible in historical data), ACE selects which time periods to query for training. Across five runs, ACE consistently prioritizes high-volatility regimes (1970s stagflation, Volcker disinflation, Great Recession) that expose nonlinear inflation dynamics. This validates that ACE’s learned experimental design principles transfer beyond controlled interventions to strategic sampling from historical archives, extending applicability to domains where live experimentation is infeasible.

5. Discussion

5.1. Ablation Studies

We systematically ablate each architectural component across three to four independent seeds. Table 2 shows that removing any component causes substantial performance degradation (130–156%), validating the necessity of each design choice.

The consistent degradation (130–156%) demonstrates that ACE’s advantage derives from the synergistic combination of all components. Each ablation reduces performance to baseline levels (2.06–2.10), confirming that preference learning (DPO), adaptive convergence, specialized root handling, and exploration diversity are all essential for ACE’s strategic behavior.

Table 2. Ablation study results. Each component contributes substantially; removing any degrades performance to near-baseline levels.

Component Removed	Loss	Degrad.
Per-Node Convergence (N=4, fixed 100 ep)	2.36±0.24	+156%
Root Learner (N=3)	2.18±0.05	+137%
DPO Training (N=3, custom transformer)	2.12±0.10	+130%
Diversity Reward (N=3)	2.12±0.10	+130%
ACE (Full): 0.92±0.73, median 0.61		

One seed (789) exhibited persistent X_5 mechanism failure, achieving loss 1.73 compared to 0.02–0.22 for other seeds, indicating sensitivity to initialization or optimization challenges with quadratic mechanisms. This outlier motivates our use of median statistics and highlights the value of multi-seed validation.

5.2. Why Preference Learning Outperforms Value-Based RL

DPO-trained ACE (median: 0.61) consistently outperforms PPO (2.19 ± 0.07) despite receiving identical reward signals, a 68% median improvement that is statistically significant ($p=0.0046$) and practically large (Cohen’s $d=-2.46$). The reason illuminates a key insight about experimental design.

The core challenge is that information gain is inherently non-stationary. Early in training, when the learner knows little, a single well-chosen experiment can produce dramatic loss reductions ($\Delta\mathcal{L} > 50$). As the learner improves, the same quality of experimental design yields diminishing absolute returns ($\Delta\mathcal{L} < 0.1$). This shifting reward scale poses fundamental challenges for value-based methods: PPO’s critic must learn to predict expected returns, but the magnitude of those returns changes by orders of magnitude over training. The critic’s value estimates become unreliable, leading to unstable policy updates.

DPO sidesteps this problem entirely by learning from rankings rather than magnitudes. Preferences depend only on reward differences $r_0(a) - r_0(b)$ between candidate interventions, which remain meaningful even as absolute rewards shrink. Formally, if rewards scale by a time-varying factor $f(t)$, preferences are invariant: the better intervention remains better regardless of the current scale (Rafailov et al., 2023). This provides inherent robustness to the diminishing returns that characterize scientific discovery.

The advantage is particularly pronounced for collider learning, where ACE achieves $L_{X_3} = 0.054$ through strategic concentration on collider parents (99.8% of interventions on

X_1 and X_2). PPO, hampered by its unstable value estimates, fails to discover this strategy and distributes interventions more uniformly.

6. Conclusion

Efficient causal discovery requires strategic experimentation, but which intervention should come next? We addressed this by treating experimental design as a learnable policy. ACE learns to propose interventions via Direct Preference Optimization, using pairwise comparisons that remain meaningful even as absolute information gains diminish over the course of discovery.

ACE achieves 70–71% improvement over baselines at equal intervention budgets (171 episodes), with $p < 0.001$ and Cohen’s $d \approx 2$. The comparison is informative: baselines plateau at 2.06–2.10 regardless of additional episodes, while ACE reaches 0.61 median loss, indicating that strategic adaptation—not simply more data—drives continued improvement. Ablation studies validate all four architectural components, with 130–156% degradation when any is removed.

A key finding is the policy’s emergent behavior. Without explicit instruction, ACE learns to concentrate 99.8% of interventions on collider parents, matching the strategy that causal theory identifies as optimal for multi-parent mechanisms. This suggests that preference learning can recover principled experimental strategies through experience, complementing theoretical understanding with learned domain adaptation.

The approach extends beyond controlled experiments: our Phillips curve results demonstrate that ACE’s principles transfer to strategic sampling from observational archives, where “interventions” become choices about which historical regimes to observe. Future work will extend to joint structure discovery, scale to larger systems, and deploy in domains where intervention costs make learned experimental design most valuable.

7. Limitations and Future Work

Statistical power. With $N=5$ seeds, confidence intervals remain relatively wide and one outlier seed drives the high variance in ACE’s results (0.92 ± 0.73). Larger sample sizes ($N \geq 10$) would provide more precise effect size estimates and better characterize rare failure modes.

Scope. ACE assumes known causal structure (focusing on mechanism estimation rather than joint structure-and-mechanism discovery) and faces scalability limits beyond 20 nodes from text-based graph encoding. The Phillips curve demonstrates ACE’s applicability to retrospective causal learning on observational archives, though regime selection

differs from controlled do-calculus intervention.

Future work will extend to joint structure discovery, scale to larger systems, and deploy where intervention costs make learned experimental design most valuable.

Impact Statement

This paper aims to advance machine learning for scientific discovery. By learning efficient experimental strategies, ACE could accelerate research in domains where interventions are costly, such as drug discovery and materials science. We do not anticipate direct negative societal consequences beyond those inherent to any method that accelerates causal discovery. Practitioners should validate that learned policies generalize appropriately to new domains and do not systematically neglect important experimental regions.

References

- Chen, T., Anumasa, S., Lin, B., Shah, V., Goyal, A., and Liu, D. Auto-bench: An automated benchmark for scientific discovery in LLMs. *arXiv preprint arXiv:2502.15224*, 2025.
- Chickering, D. M. Learning Bayesian networks is NP-complete. In Fisher, D. and Lenz, H.-J. (eds.), *Learning from Data: Artificial Intelligence and Statistics V*, pp. 121–130. Springer, New York, NY, 1996.
- Cho, H., Berger, B., and Peng, J. Reconstructing causal biological networks through active learning. *PLOS ONE*, 11(3):e0150611, 2016.
- Cohn, D. A., Ghahramani, Z., and Jordan, M. I. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.
- Eberhardt, F., Glymour, C., and Scheines, R. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among N variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 178–184. AUAI Press, 2005.
- Eberhardt, F., Glymour, C., and Scheines, R. $N-1$ experiments suffice to determine the causal relations among N variables. In Holmes, D. E. and Jain, L. C. (eds.), *Innovations in Machine Learning*, pp. 97–112. Springer, Berlin, Heidelberg, 2006.
- Federal Reserve Bank of St. Louis. Federal reserve economic data (FRED). <https://fred.stlouisfed.org>, 2024. Accessed: 2024.
- Fisher, R. A. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

- Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.
- Gandhi, K., Li, M. Y., Goodyear, L., Bhatia, A., Li, L., Bhaskar, A., Zaman, M., and Goodman, N. D. BoxingGym: Benchmarking progress in automated experimental design and model discovery. *arXiv preprint arXiv:2501.01540*, 2025.
- Hauser, A. and Bühlmann, P. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- Kovacic, I. and Brennan, M. J. *The Duffing Equation: Non-linear Oscillators and their Behaviour*. John Wiley & Sons, 2011. ISBN 9780470715499.
- Lorch, L., Rothfuss, J., Schölkopf, B., and Krause, A. DiBS: Differentiable Bayesian structure learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24111–24123, 2021.
- Mukherjee, S., Zhang, M., Flaxman, S., and Vollmer, S. J. Graph agnostic causal Bayesian optimisation. *arXiv preprint arXiv:2411.03028*, 2024.
- Murphy, K. P. Active learning of causal Bayes net structure. Technical report, University of California, Berkeley, 2001. URL https://www.cs.ubc.ca/~murphyk/Papers/causal_active.pdf. Technical report.
- Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.
- Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.
- Qwen Team. Qwen2.5: A party of foundation models. <https://qwenlm.github.io/blog/qwen2.5/>, September 2024.
- Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.
- Rainforth, T., Foster, A., Ivanova, D. R., and Bickford Smith, F. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024. doi: 10.1214/23-STS915.
- Sauter, A. W. M., Botteghi, N., Acar, E., and Plaat, A. CORE: Towards scalable and efficient causal discovery with reinforcement learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1664–1672, 2024.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Settles, B. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009.
- Shanmugam, K., Kocaoglu, M., Dimakis, A. G., and Vishwanath, S. Learning causal graphs with small interventions. In *Advances in Neural Information Processing Systems*, 2015.
- Xiao, W., Wang, Q., et al. A comprehensive survey of direct preference optimization. *arXiv preprint arXiv:2410.15595*, 2024.
- Zheng, T., Deng, Z., Tsang, H. T., Wang, W., Bai, J., Wang, Z., and Song, Y. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. URL <https://aclanthology.org/2025.emnlp-main.895/>.