CAUSAL STRUCTURE REFINEMENT WITH WARM-START REINFORCEMENT LEARNING AND FINITE-SAMPLE GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal structure learning from observational data is challenging due to the combinatorial search over directed acyclic graphs and the difficulty of reliable model selection under finite samples. Reinforcement learning (RL) offers a flexible framework for exploring the space of graph structures, but remains largely heuristic and lacks statistical guarantees. We propose a warm-start RL framework that recasts causal structure learning as refinement of an existing graph. Starting from an initial graph, the method performs a sequence of local edge modifications guided by a learned policy, while maintaining a dynamically constructed candidate set for selection. We adopt a copula-based BIC score that extends BIC to data with non-Gaussian marginals through a rank-based Gaussian-copula construction. We establish two results. First, under a monotone progress-potential condition, we bound the expected number of proposals required to reach a 1-optimal graph, with the bound scaling linearly in the warm-start depth. Second, the population-best fitted candidate in the explored archive is selected, using independent data, with high probability under a variance-adaptive Bernstein bound. Empirically, refinement improves the composite score of its initializer on all six benchmarks without increasing SHD, including the 100- and 223-node benchmarks where most baselines do not converge. Compared to RL without warm start, our approach achieves over 200% improvement in Score, up to $4.3\times$ gains in TPR, and up to $\sim$50% reduction in structural error (SHD) on the two networks where this ablation was run.

1 INTRODUCTION

Causal discovery aims to uncover the graphical structure encoding causal relationships among variables. In observational data, different directed acyclic graphs (DAGs) can encode the same conditional independences; without additional assumptions, the target is a Markov equivalence class (MEC), not a uniquely oriented DAG. This statistical ambiguity accompanies a difficult search problem: even a useful graph returned by an established algorithm can contain errors, yet restarting discovery from scratch discards its structural information.

We therefore ask a narrower, operational question: *how should we refine an existing graph, and what can be guaranteed about the search and final selection?* Our framework takes a valid DAG from an initializer such as GES or GraN-DAG and learns which local edge edits to propose. Figure 1 summarizes the three ingredients and their roles. A rank-based copula transform followed by Gaussian BIC provides the search score. The current graph changes only when this score improves, and each episode resumes from the best graph found so far. All proposals, including rejected ones, enter a candidate archive. Because the initializer is retained, maximizing the same score over this archive cannot return a lower-scoring graph.

This construction does not eliminate the hard parts of causal discovery. A score-monotone trajectory can stop at a spurious local optimum; a Gaussianizing marginal transform does not make every joint distribution Gaussian; and selecting candidates on the same observations that generated them creates statistical dependence. Rather than conflating these issues, we separate *search*, *selection*, and *causal identification*. The search analysis applies to any persistently exploring proposal policy, including randomized hill climbing. The statistical analysis applies to frozen candidate score functions evaluated

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

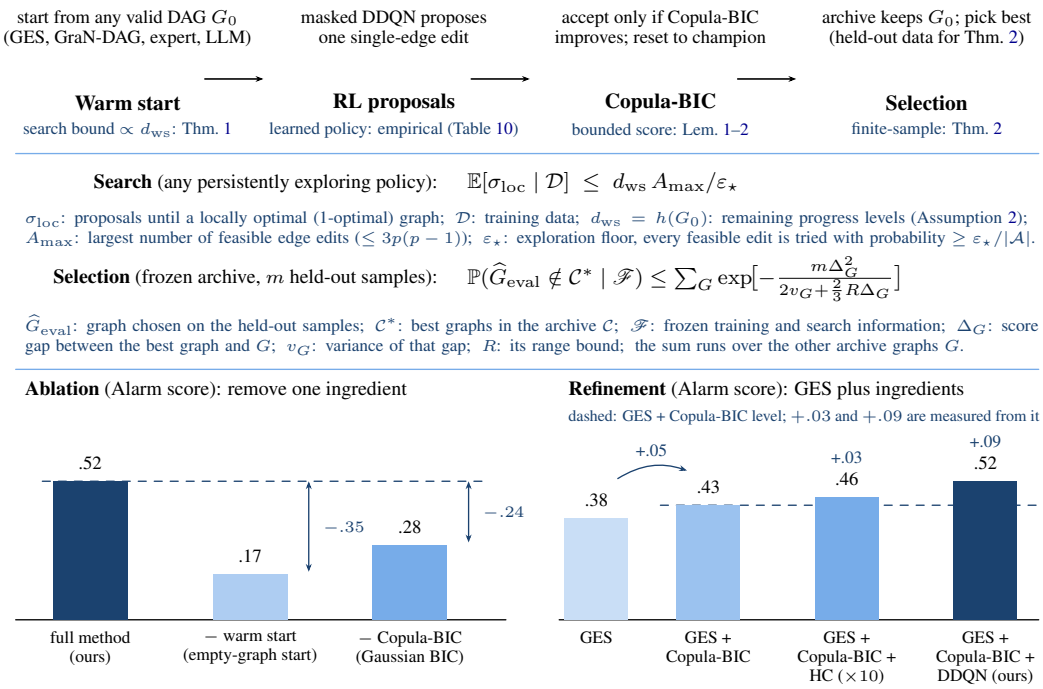

Figure 1: **Our method in a nutshell: three ingredients, one job each.** The search bound scales with the warm-start depth (Theorem 1); Copula-BIC gives the bounded score behind selection (Lemmas 1–2, Theorem 2); DDQN proposes the edits, and its benefit is empirical. *Left:* removing the warm start or Copula-BIC lowers the Alarm score (Table 2). *Right:* Alarm scores for GES, GES + Copula-BIC, and two alternative search rules on top of it (Table 10).

on independent data. DDQN is the adaptive proposal mechanism whose benefit must be established empirically, not an assumption that makes a generic concentration argument RL-specific.

Contributions. **(1) A specified refinement procedure.** We give the state and action representations, proposal reward, monotone transition rule, masked off-policy DDQN updates, and champion resets, together with an algorithm-to-theorem map. **(2) Scoped search guarantees.** We prove a hitting-time bound in terms of a progress potential and state the additional landscape assumption required to turn local optimality into MEC recovery. A short improving path alone is insufficient. **(3) Finite-candidate selection.** We establish a variance-adaptive Bernstein selection bound for a frozen candidate archive scored on independent data. The bound concerns the population-best *fitted candidate*, including under model misspecification. **(4) Component-level evidence.** We consolidate six-network comparisons, GES with copula scoring, warm-start hill climbing at two budgets, five-run uncertainty estimates, sensitivity results, and non-Gaussian-copula experiments.

2 Related Work

Discovery and refinement. Constraint-based procedures such as PC reason from conditional independences, whereas GES searches equivalence classes using a decomposable score (Spirtes et al., 2001; Chickering, 2002). Functional models such as LiNGAM and nonlinear additive-noise models exploit stronger distributional assumptions (Shimizu et al., 2011; Hoyer et al., 2008); NOTEARS, GOLEM, and GraN-DAG instead use continuous optimization (Zheng et al., 2018; Ng et al., 2020; Lachapelle et al., 2020). Our objective is not to replace these methods, but to reuse their outputs. A CPDAG initializer requires a consistent DAG extension; a partially ancestral graph from a latent-variable method cannot generally be substituted without changing the model class.

Structural priors and prior-guided search. Using an initial graph is related to, but different from, placing a Bayesian prior over graphs. VaMSL incorporates expert information into variational inference for mixtures of DAGs (Björkman et al., 2025). ILS-CSL iteratively uses language-model feedback to refine structural constraints (Ban et al., 2023). Cluster-PC and Cluster-FCI exploit cluster-level background knowledge to warm-start constraint-based discovery (de Vargas et al., 2025). Our warm start is an algorithmic initialization, not a posterior distribution or an elicited prior; the learned Q-function estimates return, not posterior probability. These precedents rule out a claim that graph refinement itself is new. Our emphasis is the combination of scored local proposals, incumbent retention, and explicitly separated search and selection guarantees.

RL and statistical guarantees. RL-BIC learns graph-generating policies, CORL searches variable orderings, and KCRL incorporates prior knowledge into RL-based discovery (Zhu et al., 2020; Wang et al., 2021; Hasan & Gani, 2022); CORE emphasizes scalable RL search (Sauter et al., 2024). We use Double DQN (Van Hasselt et al., 2016) to prioritize local proposals. Our hitting-time argument does not prove convergence of its neural Q-function or superiority over other proposal policies. Similarly, finite-sample causal learning has a substantial literature (Kalisch & Bühlmann, 2007; Gao et al., 2020; Wadhwa & Dong, 2021). Our finite-class bound is a selection result, not a full-space recovery theorem. Rank Gaussianization also builds on established Gaussian-copula modeling (Liu et al., 2009); we do not claim the transform itself as a new estimator.

3 Methodology

Let $\mathcal{D} = \{x^{(i)}\}_{i=1}^{n}$ contain independent observations of p variables. We write G for a DAG, E_G for its edge set, $A_G \in \{0,1\}^{p\times p}$ for its adjacency matrix, and $\mathrm{pa}_G(j)$ for the parents of node j. Superscripts index observations and subscripts index variables. The admissible class $\mathcal{G}$ may impose an in-degree bound b and background constraints. The feasible actions $\mathcal{A}(G)$ are single-edge additions, deletions, or reversals that remain in $\mathcal{G}$; $\mathcal{T}(G,a)$ denotes the proposed graph. The one-edit neighborhood is $\mathcal{N}_1(G) = \{\mathcal{T}(G,a) : a \in \mathcal{A}(G)\}$. Table 3 in Appendix A.1 collects all notation.

3.1 Copula-BIC and its scope

For continuous measurements, transform each marginal using ranks:

$$U_j^{(i)} = \frac{\mathrm{rank}(x_j^{(i)})}{n+1}, \qquad z_j^{(i)} = \Phi^{-1}(U_j^{(i)}), \tag{1}$$

where Φ is the standard normal distribution function. Ties receive midranks. Fit nodewise Gaussian regressions with intercept $\widehat{\alpha}_{Gj}$, coefficients $\widehat{\beta}_{Gj}$, and positive residual variances $\widehat{v}_{Gj}$. The fitted per-observation log score and total BIC are

$$s_G(z) = -\frac{1}{2}\sum_{j=1}^{p}\left[\log(2\pi\widehat{v}_{Gj}) + \frac{(z_j - \widehat{\alpha}_{Gj} - \widehat{\beta}_{Gj}^{\mathsf{T}} z_{\mathrm{pa}_G(j)})^2}{\widehat{v}_{Gj}}\right], \tag{2}$$

$$B_n(G) = \sum_{i=1}^{n} s_G(z^{(i)}) - \frac{k(G)}{2}\log n, \qquad S_n(G) = \frac{B_n(G)}{n}. \tag{3}$$

Here $k(G) = |E_G| + 2p$ counts regression coefficients, intercepts, and variances in this parameterization (Schwarz, 1978). Total B_n and normalized S_n have the same maximizer, but their scales must not be interchanged in the reward.

For fixed fitted parameters, Eq. (1) bounds every transformed training coordinate by $M_n = \Phi^{-1}(n/(n+1))$. This yields a finite Lipschitz constant for the quadratic score on the resulting box, provided fitted variances are bounded away from zero (Lemma 1). It does *not* make the rank-transformed training rows independent, or bound sensitivity to refitting an entire dataset. For causal interpretation, the intended model is a Gaussian copula with continuous invertible marginal transforms and a faithful, causally sufficient DAG representation. Marginal normalization cannot remove non-Gaussian dependence or arbitrary nonlinear mechanisms; it is a working score outside this regime.

3.2 PROPOSAL POLICY, TRANSITIONS, AND LEARNING

The state is $s_t = \mathrm{vec}(A_{G_t}) \in \mathbb{R}^{p^2}$. Each output of Q_θ indexes an operation type and ordered node pair; the architecture is specified after Algorithm 1. Invalid operations, including cycle-creating edits, are masked to $-\infty$ before any maximization. With deterministic tie breaking and $a_\theta(s) = \arg\max_{a \in \mathcal{A}(G)} Q_\theta(s, a)$, the behavior policy is

$$\pi_\theta(a \mid s) = (1 - \varepsilon)\mathbf{1}\{a = a_\theta(s)\} + \frac{\varepsilon}{|\mathcal{A}(G)|}, \qquad a \in \mathcal{A}(G), \tag{4}$$

and zero for infeasible actions. Thus every feasible action has probability at least $\varepsilon/|\mathcal{A}(G)|$. If there are no feasible actions, the episode terminates.

For proposal $G' = \mathcal{T}(G_t, a_t)$, the reward and accepted transition are

$$r_t = \frac{B_n(G') - B_n(G_t)}{p} - \lambda|E_{G'}| - c, \tag{5}$$

$$G_{t+1} = \begin{cases} G', & B_n(G') > B_n(G_t), \\ G_t, & \text{otherwise.} \end{cases} \tag{6}$$

The reward uses the *current graph*, not the empty graph, as the score baseline. The density penalty applies to every proposal: relative to otherwise equal proposals, deleting an edge reduces this penalty and adding one increases it. It shapes learning in addition to BIC's complexity term, but does not change the acceptance criterion or final selection score. A rejected proposal is a self-loop with its proposal reward, and is still stored in replay and in the candidate archive. Strict acceptance means the method does not cross a score-decreasing barrier merely because it uses RL.

Training is off-policy, using minibatches of q replay transitions. For a transition (s, a, r, s', d), where d marks termination or the episode horizon, the masked Double DQN target and minibatch loss are

$$y = r + \gamma(1 - d)Q_{\theta^-}\left(s', \arg\max_{a' \in \mathcal{A}(G_{s'})} Q_\theta(s', a')\right), \quad \mathcal{L}(\theta) = \frac{1}{q}\sum_{u=1}^{q}(Q_\theta(s_u, a_u) - y_u)^2. \tag{7}$$

Here $G_{s'}$ is the graph encoded by s'. The bootstrap term is set to zero before computing a maximization when $d = 1$ or the next action set is empty. Adam updates θ every four environment steps once the replay buffer contains q transitions; target weights are updated as $\theta^- \leftarrow (1 - \tau)\theta^- + \tau\theta$. Here θ denotes network weights, θ^- target-network weights, and γ the discount. For comparison, $Q^\pi(s, a) = \mathbb{E}_\pi[\sum_{u \geq 0} \gamma^u r_{t+u} \mid s_t = s, a_t = a]$ is a *definition* for an arbitrary fixed policy π, not the DDQN training rule. Algorithm 1 gives the full update sequence; teal annotations mark the steps used by each theoretical result, and Table 4 lists the mechanism each result requires.

Policy architecture. $Q_\theta(s, a)$ is a three-layer MLP (hidden width 256, ReLU). Its input is $s = \mathrm{vec}(A_G) \in \mathbb{R}^{p^2}$ and its output is $3p(p-1)$ Q-values, one per edge-operation slot; infeasible actions are masked to $-\infty$. Here θ denotes the MLP weights; all training settings are in Table 5.

3.3 WHAT WARM STARTS GUARANTEE

Algorithm 1 resets to the *current champion*, not to G_0. Its finite budget gives $|\mathcal{C}| \leq 1 + NT$, and monotone acceptance restricts the realized trajectory. However, valid edge edits alone do not define a small search space: with unrestricted additions and deletions, all DAGs can be connected through the empty graph. Warm-starting may shorten refinement; it is not by itself a polynomial bound on the size of that graph space.

For any fixed score J used consistently in final selection, $G_0 \in \mathcal{C}$ implies

$$J\left(\arg\max_{G \in \mathcal{C}} J(G)\right) \geq J(G_0). \tag{8}$$

This deterministic statement holds for the training score or for an independent evaluation score. It does not guarantee improved SHD, TPR, FDR, or population score. The archive also enables the independent evaluation of Section 4.4.

Algorithm 1 Warm-start DDQN refinement with champion resets

Require: Data $\mathcal{D}$, initializer G_0, episode budget N, learning rate μ
Require: $T = 100$, $\varepsilon = .10$, $\gamma = .99$, replay capacity 10,000, $q = 64$
Require: Target mixing $\tau = .01$, density coefficient $\lambda = .5$, step cost $c = .001$

Construct the training rank transform and cache local graph scores
Initialize θ; set $\theta^- \leftarrow \theta$, $\mathcal{B} \leftarrow \varnothing$, step counter $u \leftarrow 0$
$\mathcal{C} \leftarrow \{G_0\}$; $\bar{G} \leftarrow G_0$
for $e = 1, \ldots, N$ **do**
  $G \leftarrow \bar{G}$; $s \leftarrow \mathrm{vec}(A_G)$
  ▷ Champion reset: preserves progress in Theorem 1.
  for $t = 1, \ldots, T$ **do**
    Compute feasible mask $\mathcal{A}(G)$; break if empty
    Sample a using Eq. (4)
  ▷ Persistent exploration (Definition 2): $\varepsilon_\star = \varepsilon$.
    $G' \leftarrow \mathcal{T}(G, a)$; fit/cache its local score terms
    $r \leftarrow [B_n(G') - B_n(G)]/p - \lambda|E_{G'}| - c$
  ▷ Copula-BIC score: Lemmas 1–2; current-graph reward.
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{G'\}$
  ▷ Archive all proposals, including rejections; freeze before Theorem 2 evaluation.
    if $B_n(G') > B_n(G)$ **then**
      $G \leftarrow G'$
  ▷ Monotone accepted update: required by the hitting-time proof.
    $s' \leftarrow \mathrm{vec}(A_G)$; $d \leftarrow \mathbf{1}\{t = T \text{ or } \mathcal{A}(G) = \varnothing\}$
    Store (s, a, r, s', d) in FIFO replay buffer $\mathcal{B}$; $u \leftarrow u + 1$
    if u is divisible by 4 and $|\mathcal{B}| \geq q$ **then**
      Sample q transitions uniformly from $\mathcal{B}$
      for each sampled $(s_i, a_i, r_i, s_i', d_i)$ **do**
        if $d_i = 1$ or $\mathcal{A}(G_{s_i'}) = \varnothing$ **then**
          $y_i \leftarrow r_i$
        else
          $a_i^* \leftarrow \arg\max_{a' \in \mathcal{A}(G_{s_i'})} Q_\theta(s_i', a')$; $y_i \leftarrow r_i + \gamma Q_{\theta^-}(s_i', a_i^*)$
      Take an Adam step of size μ on $q^{-1} \sum_i (Q_\theta(s_i, a_i) - y_i)^2$
  ▷ Off-policy DDQN: the policy is learned here, every four environment steps.
      $\theta^- \leftarrow (1 - \tau)\theta^- + \tau\theta$
    if $B_n(G) > B_n(\bar{G})$ **then**
      $\bar{G} \leftarrow G$
    $s \leftarrow s'$; break if $d = 1$
return $\mathcal{C}$ and $\widehat{G}_{\mathrm{train}} \in \arg\max_{G \in \mathcal{C}} B_n(G)$

4 Theoretical Guarantees

4.1 Setup and Assumptions

Throughout this section the training data $\mathcal{D}$, the score B_n of Eq. (3), the admissible class $\mathcal{G}$, the feasible actions $\mathcal{A}(G)$, and the one-edit neighborhood $\mathcal{N}_1(G)$ of Section 3 are fixed. $\mathcal{R}(G_0)$ denotes the set of graphs reachable from G_0 through accepted moves (6), and $\mathcal{C}$ is the candidate archive returned by Algorithm 1.

Definition 1 (1-optimal DAG)**.** *A DAG $G \in \mathcal{G}$ is* 1-optimal *if $B_n(G') \leq B_n(G)$ for every $G' \in \mathcal{N}_1(G)$.*

Definition 2 (Persistent exploration)**.** *A proposal policy satisfies* persistent exploration *with parameter $\varepsilon_\star \in (0, 1]$ if, conditional on the entire search history, it selects each feasible action $a \in \mathcal{A}(G)$ at the current graph G with probability at least $\varepsilon_\star/|\mathcal{A}(G)|$. The policy in Eq. (4) satisfies this with $\varepsilon_\star = \varepsilon$.*

Assumption 1 (Well-conditioned frozen regressions)**.** *Let $M > 0$ be such that every transformed observation lies in the box $[-M, M]^p$; for the rank transform (1), $M = M_n$. There are constants $C_\beta \geq 0$, $\alpha_{\max} \geq 0$, and $v_{\min} > 0$ such that, for every candidate $G \in \mathcal{C}$ and every node j, the fitted quantities in Eq. (2) satisfy $\|\widehat{\beta}_{Gj}\|_1 \leq C_\beta$, $|\widehat{\alpha}_{Gj}| \leq \alpha_{\max}$, and $\widehat{v}_{Gj} \geq v_{\min}$. The fits and the transform are fixed before any evaluation data are scored.*

4.2 Helping Lemmas

For $G \in \mathcal{C}$ and node j, let $u_{Gj} \in \mathbb{R}^p$ have jth entry 1, entries $-\widehat{\beta}_{Gj}$ at the coordinates $\mathrm{pa}_G(j)$, and zeros elsewhere, so that the jth residual in Eq. (2) is $u_{Gj}^{\mathsf{T}} z - \widehat{\alpha}_{Gj}$. Define the score Hessian $H_G \in \mathbb{R}^{p\times p}$ and gradient offset $\omega_G \in \mathbb{R}^p$ by

$$H_G = -\sum_{j=1}^{p} \frac{u_{Gj} u_{Gj}^{\mathsf{T}}}{\widehat{v}_{Gj}}, \qquad \omega_G = \sum_{j=1}^{p} \frac{\widehat{\alpha}_{Gj}\, u_{Gj}}{\widehat{v}_{Gj}}. \tag{9}$$

$\|\cdot\|_{\mathrm{op}}$ denotes the spectral norm and $\|\cdot\|_2$ the Euclidean norm.

Lemma 1 (Lipschitz per-sample Copula-BIC score)**.** *Under Assumption 1, each s_G is quadratic with $\nabla s_G(z) = H_G z + \omega_G$, and for all $G \in \mathcal{C}$ and $z, z' \in [-M, M]^p$,*

$$|s_G(z) - s_G(z')| \le L\|z - z'\|_2, \qquad L = \max_{G\in\mathcal{C}} \left(\|H_G\|_{\mathrm{op}}\sqrt{p}\, M + \|\omega_G\|_2\right). \tag{10}$$

Proof sketch (full proof in Appendix B.2; explicit bound on L in Proposition 1, Appendix B.5). Differentiating Eq. (2) with the fitted parameters held fixed gives $\nabla s_G(z) = H_G z + \omega_G$. On the box, $\|z\|_2 \le \sqrt{p}M$, so the triangle inequality bounds the gradient norm by L. Integrating the gradient along the segment between z and z', which stays in the box, proves Eq. (10). Empirical values of L are in Table 8. □

Lemma 2 (Sub-exponential score differences)**.** *Let Assumption 1 hold, let $G_1, G_2 \in \mathcal{C}$, and let Z be a random vector in $[-M, M]^p$ that is independent of the frozen fits (e.g., one transformed evaluation observation). Let $Y = s_{G_1}(Z) - s_{G_2}(Z)$, $v = \mathrm{Var}(Y)$, and let $R > 0$ satisfy $|Y - \mathbb{E}Y| \le R$ almost surely. Both $R = 4\sqrt{p}\, ML$, with L from Lemma 1, and $R = p[M(1 + C_\beta) + \alpha_{\max}]^2/v_{\min}$ (Lemma 3) are valid under Assumption 1. Then for every real t with $|t| < 3/R$,*

$$\log \mathbb{E}e^{t(Y-\mathbb{E}Y)} \le \frac{t^2 v}{2(1 - |t|R/3)}. \tag{11}$$

Proof sketch (full proof in Appendix B.4; bound on R in Appendix B.3). The range bound follows from Lemma 1: s_G varies by at most $L \cdot \mathrm{diam}([-M, M]^p) = 2\sqrt{p}\, ML$ on the box, so Y has range at most $4\sqrt{p}\, ML$. Let $V = Y - \mathbb{E}Y$, so $\mathbb{E}V = 0$, $\mathbb{E}V^2 = v$, and $|V| \le R$. For every integer moment order $\ell \ge 2$, $\mathbb{E}|V|^\ell \le R^{\ell-2}\mathbb{E}V^2 = vR^{\ell-2}$. Expanding $\mathbb{E}e^{tV} = 1 + \sum_{\ell\ge 2} t^\ell \mathbb{E}V^\ell/\ell!$ and using the factorial bound $\ell! \ge 2 \cdot 3^{\ell-2}$ gives the geometric series $\mathbb{E}e^{tV} \le 1 + \frac{t^2 v}{2}\sum_{\ell\ge 2}(|t|R/3)^{\ell-2} = 1 + \frac{t^2 v}{2(1-|t|R/3)}$; finally $\log(1 + x) \le x$. □

4.3 Hitting-Time Guarantee

Assumption 2 (Monotone progress potential)**.** *There is a function $h : \mathcal{R}(G_0) \to \{0, 1, 2, \ldots\}$ such that (i) $h(G) = 0$ if and only if G is 1-optimal; (ii) every accepted move $G \to G'$ satisfies $h(G') \le h(G)$; and (iii) at every $G \in \mathcal{R}(G_0)$ with $h(G) = k > 0$, $|\mathcal{A}(G)| \le A_k$ and at least $\phi_k \ge 1$ feasible actions are accepted and lead to $h(G') \le k - 1$. We call $d_{\mathrm{ws}} = h(G_0)$ the* warm-start depth *and write $A_{\max} = \max_{1\le k\le d_{\mathrm{ws}}} A_k$.*

Theorem 1 (Champion-reset hitting time)**.** *Suppose Assumption 2 holds, Algorithm 1 resets each episode to the current champion (line 5), and its proposal policy satisfies persistent exploration with parameter $\varepsilon_\star$ (Definition 2). Let σ_{loc} be the number of proposals until the current graph is first 1-optimal. Then*

$$\mathbb{E}[\sigma_{\mathrm{loc}} \mid \mathcal{D}] \le \sum_{k=1}^{d_{\mathrm{ws}}} \frac{A_k}{\varepsilon_\star \phi_k} \le \frac{d_{\mathrm{ws}} A_{\max}}{\varepsilon_\star}, \qquad A_{\max} \le 3p(p-1). \tag{12}$$

Since every episode contains at least one proposal, the expected number of episodes obeys the same bound, which is polynomial in p whenever d_{ws} and $1/\varepsilon_\star$ are.

Corollary 1 (When local optimality implies MEC membership)**.** *Let $G^\dagger$ be the data-generating DAG and $\mathrm{MEC}(G^\dagger)$ its Markov equivalence class. If every $G \in \mathcal{R}(G_0) \setminus \mathrm{MEC}(G^\dagger)$ has some $G' \in \mathcal{N}_1(G)$ with $B_n(G') > B_n(G)$, then every 1-optimal $G \in \mathcal{R}(G_0)$ belongs to $\mathrm{MEC}(G^\dagger)$.*

Proof. A 1-optimal graph outside $\mathrm{MEC}(G^\dagger)$ would have a strictly improving neighbor by the stated condition, contradicting Definition 1. □

Proof sketch of Theorem 1 (full proof in Appendix B.1). At potential level k, persistent exploration selects one of the ϕ_k progress actions with conditional probability at least $\varepsilon_\star \phi_k / A_k$, so the waiting time at level k is dominated by a geometric variable with this success probability. Monotone acceptance and champion resets prevent a return to a higher level, so summing the expected waiting times over $k = 1, \ldots, d_{\mathrm{ws}}$ proves Eq. (12). □

The potential condition is essential: the policy need not follow a shortest improving path, and d_{ws} is not automatically the SHD to the true DAG. Faithfulness and BIC consistency alone do not exclude spurious 1-optima, since a score-neutral covered reversal can be necessary; hence the extra condition in Corollary 1. The theorem applies equally to persistently randomized hill climbing; it is not a DDQN-specific convergence claim.

4.4 Finite-Sample Selection Guarantee

To give a valid finite-candidate guarantee for an adaptive search, we separate training from evaluation. Let $\mathscr{F}$ collect the training data, fitted marginal transform, fitted regressions, search randomness, and the resulting finite archive $\mathcal{C}$. After freezing $\mathscr{F}$, draw evaluation observations $X^{(1)}, \ldots, X^{(m)}$ i.i.d. from the data distribution, independently of $\mathscr{F}$. Each is transformed by the frozen clipped map $g_{\mathscr{F}}(x) = \big(\Phi^{-1}(\min\{1-\eta, \max\{\eta, \widehat{\Psi}_{n,j}(x_j)\}\})\big)_{j=1}^{p}$, where $\widehat{\Psi}_{n,j}$ is the training empirical distribution function of variable j and $\eta = 1/(n+1)$; ranks are not recomputed. Define

$$f_G(X) = s_G(g_{\mathscr{F}}(X)) - \mathrm{pen}_G, \quad \widehat{J}_m(G) = \frac{1}{m}\sum_{i=1}^{m} f_G(X^{(i)}), \quad J(G) = \mathbb{E}[f_G(X) \mid \mathscr{F}], \tag{13}$$

where X is a generic evaluation observation and $\mathrm{pen}_G = k(G)\log n/(2n)$ is the training-determined penalty. This is a held-out penalized log score of a frozen fit, *not* BIC refitted on evaluation data. Select $\widehat{G}_{\mathrm{eval}} \in \arg\max_{G\in\mathcal{C}} \widehat{J}_m(G)$.

Theorem 2 (Finite-candidate Bernstein selection)**.** *Conditional on $\mathscr{F}$, let $\mathcal{C}^* = \arg\max_{G\in\mathcal{C}} J(G)$ be the set of population maximizers and fix $G^* \in \mathcal{C}^*$. For each $G \in \mathcal{C}\setminus\mathcal{C}^*$, let $Y_G = f_{G^*}(X) - f_G(X)$, with gap $\Delta_G = \mathbb{E}[Y_G \mid \mathscr{F}] > 0$ and variance $v_G = \mathrm{Var}(Y_G \mid \mathscr{F})$, and suppose $|Y_G - \Delta_G| \le R$ almost surely for a constant $R > 0$ (under Assumption 1, the R of Lemma 2 is valid). Then*

$$\mathbb{P}(\widehat{G}_{\mathrm{eval}} \notin \mathcal{C}^* \mid \mathscr{F}) \le \sum_{G\in\mathcal{C}\setminus\mathcal{C}^*} \exp\left[-\frac{m\Delta_G^2}{2v_G + (2/3)R\Delta_G}\right]. \tag{14}$$

Let $\Delta_{\min} = \min_{G\in\mathcal{C}\setminus\mathcal{C}^*}\Delta_G$ and $v_{\max} = \max_{G\in\mathcal{C}\setminus\mathcal{C}^*} v_G$. For any $\delta \in (0,1)$, the conditional error probability is at most δ whenever

$$m \ge \left(\frac{2v_{\max}}{\Delta_{\min}^2} + \frac{2R}{3\Delta_{\min}}\right)\log\frac{|\mathcal{C}|}{\delta}. \tag{15}$$

If $\mathcal{C}^* = \mathcal{C}$, the error probability is zero.

Proof sketch (full proof in Appendix B.6). If a suboptimal candidate wins, its empirical score is at least that of the fixed population maximizer G^*. For each competitor G, center the per-observation difference Y_G at Δ_G and apply the Bernstein tail of Lemma 2; independent evaluation makes these terms i.i.d. conditional on $\mathscr{F}$. A union bound over competitors yields Eq. (14); replacing each gap and variance by $\Delta_{\min}$ and $v_{\max}$ gives Eq. (15). □

Theorem 2 uses no property of the RL search: it applies to any finite frozen archive, for example graphs produced by GES with Copula-BIC.

What can be combined. Theorem 1 concerns local optimality under the *training* score; Theorem 2 concerns the best frozen candidate under a *population evaluation* score. If an event of probability at least $1-\kappa$ ensures that the archive covers the target and $\mathcal{C}^* \subseteq \mathrm{MEC}(G^\dagger)$, the overall MEC error is at most $\kappa + \delta$ (Appendix B.6); neither theorem establishes that event by itself.

Table 1: Results on all benchmarks. Each cell: TPR ↑/FDR ↓ (first line) and SHD ↓/Score ↑ (second line); gray sublines give 95% CI half-widths over five runs (zero for methods that are deterministic on fixed data). Bold: best value, ties included. NC: no result within the compute budget; NA: initializer unavailable; –: not run. GES+C: GES with Copula-BIC; WS+HC: hill climbing from GES with Copula-BIC. WS-GD/WS-GES: ours with GraN-DAG/GES warm start.

Size	Dataset	Baselines and controls												Ours	
		KCRL	NT	GLM	RL-B	ICA-L	D-L	PC	CORL	GD	GES	GES+C	WS+HC	WS-GD	WS-GES
		TPR/FDR and SHD/Score; gray sublines: corresponding 95% CI half-widths													
Small	Sachs	.35/.45 .050/.062 15/.32 1.86/.025	.30/.59 .012/.025 12/.26 1.24/.012	.18/.83 .012/.037 24/.13 1.86/.012	.24/.67 .062/.074 14/.21 2.48/.037	.22/.50 .012/.012 14/.26 1.24/.012	.12/.50 .000/.000 15/.23 0.00/.000	.33/.77 .000/.000 30/.20 0.00/.000	.47/.77 .074/.087 26/.25 2.48/.037	**.55**/.30 .037/.050 12/.43 1.24/.012	.48/.64 .000/.000 28/.29 0.00/.000	.50/.61 .000/.000 26/.30 0.00/.000	.50/.60 .025/.037 26/.31 1.24/.012	**.55/.28** .025/.037 **10/.45** 0.62/.012	.52/.62 .025/.037 28/.31 1.24/.012
	Child	.15/.80 .037/.062 28/.13 1.86/.012	.12/.62 .012/.025 22/.18 1.24/.012	.10/.78 .012/.025 24/.12 1.24/.012	.44/**.39** .074/.025 **21**/.36 1.86/.025	.24/.55 .012/.012 **21**/.24 1.24/.012	.12/.82 .000/.000 28/.11 0.00/.000	.24/.86 .000/.000 43/.13 0.00/.000	NC	.50/.67 .037/.050 25/.29 2.48/.012	.38/.89 .000/.000 34/.17 0.00/.000	.40/.87 .000/.000 33/.18 0.00/.000	–	**.72**/.55 .037/.050 **21/.41** 1.86/.025	.32/.81 .025/.037 33/.18 1.86/.012
Medium	Alarm	.33/.63 .050/.062 49/.24 3.72/.025	.17/.43 .012/.025 41/.26 2.48/.012	.15/.40 .012/.025 43/.26 2.48/.012	.30/.74 .074/.087 56/.20 6.21/.037	.57/.32 .037/.050 29/.42 3.10/.025	.39/.50 .000/.000 40/.30 0.00/.000	.67/.60 .000/.000 55/.36 0.00/.000	NC	.24/.73 .050/.062 60/.18 4.97/.025	.74/.61 .000/.000 56/.38 0.00/.000	.78/.47 .000/.000 40/.43 0.00/.000	.77/.49 .025/.037 42/.42 2.48/.012	.40/.65 .037/.050 49/.26 3.72/.025	**.82/.30** .025/.037 **26/.52** 2.48/.012
	Hepar2	NC	.02/.99 .012/.012 157/.01 8.69/.012	NC	NC	.19/.49 .025/.037 112/.24 6.21/.012	.10/**.07** .000/.000 110/.35 0.00/.000	.35/.75 .000/.000 172/.20 0.00/.000	NC	.28/.39 .025/.050 96/.30 7.45/.012	.50/.23 .000/.000 **70**/.42 0.00/.000	.54/.22 .000/.000 71/.44 0.00/.000	–	.54/.51 .025/.062 84/.35 6.21/.012	**.61**/.23 .025/.037 **70/.46** 4.97/.012
Large	DREAM	NC	.07/.97 .012/.012 293/.03 13.66/.012	NC	NC	NC	NC	NC	NC	.09/.97 .025/.012 251/.04 9.93/.012	NC	–	–	**.15/.82** .025/.037 **184/.11** 8.69/.012	NA
	Andes	NC	.07/.97 .012/.012 316/.03 18.62/.012	NC	NC	NC	NC	NC	NC	.09/.97 .025/.025 314/.04 12.42/.012	NC	–	–	**.18/.80** .025/.037 **254/.13** 11.17/.012	NA

5 Experiments

5.1 Protocol and primary comparisons

We evaluate on Sachs ($p = 11$, $n = 7{,}446$), Child (20, 1,000), Alarm (37, 10,000), Hepar2 (70, 5,000), DREAM (100, 100), and Andes (223, 10,000). Dataset details are in Appendix F. The primary initializers are GES and GraN-DAG; additional baselines are PC, ICA-LiNGAM, DirectLiNGAM, NOTEARS, GOLEM, RL-BIC, CORL, and KCRL.

We report true positive rate (TPR), false discovery rate (FDR), structural Hamming distance (SHD), and the composite score

$$\text{Score} = \tfrac{1}{3}\left[\text{TPR} + (1 - \text{FDR}) + (1 + \text{SHD})^{-1}\right]. \tag{16}$$

This structural metric is distinct from the training Copula-BIC objective (Appendix F.1). All reported graphs are $\widehat{G}_{\text{train}}$ (training-score selection, Algorithm 1); the held-out selector of Theorem 2 is not used. Table 1 reports all four metrics with approximate 95% CI half-widths over five runs (Table 12 and Appendix E); GES, PC, DirectLiNGAM, and GES+C are deterministic on fixed data, so their zero half-widths denote repeatability only. Bold marks all best-value ties.

Refinement of the corresponding initializer. By Eq. (8), every warm-start run returns a graph whose training Copula-BIC is at least that of its initializer; the structural metrics show how this translates into TPR, FDR, and SHD. On Sachs, WS-GD reduces SHD from 12 to 10 and FDR from .30 to .28 with unchanged TPR of .55; the larger comparison against GES is *not* a same-initializer improvement. On Child, WS-GD increases TPR from .50 to .72 (22 percentage points), reduces FDR from .67 to .55, and reduces SHD from 25 to 21. On Alarm, WS-GES increases TPR from .74 to .82 (8 percentage points), reduces FDR from .61 to .30 (31 percentage points), and reduces SHD from 56 to 26 (53.6% relative). On Hepar2, WS-GES improves TPR from .50 to .61 while tying GES at

Table 2: Ablations, each changing one component of the full method. "No WS": $G_0 = \emptyset$ instead of GES/GraN-DAG. "Gaussian": Gaussian BIC instead of Copula-BIC. Full: GES on Alarm, GraN-DAG on Sachs. $\pm$: 95% CI half-width (five runs). Bold: best.

Dataset	Variant	TPR ↑	FDR ↓	SHD ↓	Score ↑
Alarm	No WS	$.19 \pm .04$	$.70 \pm .06$	50 ± 4.2	$.17 \pm .03$
	Gaussian	$.39 \pm .05$	$.55 \pm .05$	35 ± 3.1	$.28 \pm .04$
	Full	$\mathbf{.82 \pm .03}$	$\mathbf{.30 \pm .04}$	$\mathbf{26 \pm 2.0}$	$\mathbf{.52 \pm .03}$
Sachs	No WS	$.24 \pm .05$	$.89 \pm .07$	21 ± 2.8	$.13 \pm .03$
	Gaussian	$.25 \pm .04$	$.88 \pm .06$	17 ± 2.2	$.15 \pm .03$
	Full	$\mathbf{.55 \pm .03}$	$\mathbf{.28 \pm .04}$	$\mathbf{10 \pm 1.2}$	$\mathbf{.45 \pm .02}$

FDR .23 and SHD 70. WS-GD on Hepar2 increases FDR from .39 to .51, illustrating why score non-degradation is not metric-wise non-degradation.

On DREAM and Andes, WS-GD reduces SHD from 251 to 184 (26.7%) and from 314 to 254 (19.1%), respectively. Eight of the ten baselines do not converge on these two networks. The absolute FDR remains .82 and .80, so the large-scale results show improvement over the initializer rather than accurate recovery of most edges.

5.2 Separating scoring, search, and initialization

On Alarm, GES+C improves SHD from 56 to 40, while GES-initialized HC gives 42 and HC with ten times the search budget gives 37. WS-GES reaches 26. On Sachs, GES+C and HC both reach SHD 26, HC$\times$10 reaches 25, and WS-GES remains at 28. The best Sachs result (SHD 10) uses a *GraN-DAG* initializer. Tables 9 and 10 in Appendix D.2 give the copula-scoring and hill-climbing controls in full.

Removing warm starts reduces the Alarm score from .52 to .17 and Sachs from .45 to .13 (Table 2). The corresponding SHD reductions from No WS to Full are 48% and 52.4%. RL is thus most effective as a refiner rather than an end-to-end learner. Zero-shot GPT-4o graphs as alternative initializers are compared in Table 11 (Appendix D.3).

Sensitivity and uncertainty. The defaults are $\varepsilon = .10$, $T = 100$, $\lambda = .5$, and $c = .001$. On Alarm, varying ε over $.05, .10, .20, .30$ gives scores $.48, .52, .50, .47$, a .05 span against the .35 No-WS gap; shortening T to 50 lowers the score to .44. Tables 14 and 15 give all sweeps with five-run CIs.

Compute. Mean wall-clock times on one A100 GPU (three runs) are about 4, 12, 28, 55, 90, and 120 minutes for Sachs through Andes (Appendix F.2).

5.3 Misspecification and the meaning of selection

Appendix C evaluates selection under Frank, Gumbel, and Clayton copulas (40 trials per n; Table 6): mis-selection falls from .100–.250 at $n = 500$ to .000–.050 at $n = 4{,}000$. Under Clayton(5) the population-best candidate still has SHD 2, so selection is consistent without being causally correct. None of these experiments involves RL; see also Figure 2 and Table 7.

6 Conclusion and Limitations

We recast RL-based causal discovery as refinement of an existing graph: masked DDQN proposes single-edge edits, Copula-BIC accepts only improvements, and an archive that retains the initializer guarantees score non-degradation. We proved a hitting-time bound linear in the warm-start depth, a condition under which a reached local optimum lies in the true MEC, and a variance-adaptive finite-sample selection guarantee. Across six benchmarks, refinement improves the composite score over its initializer in all ten initializer–dataset pairs without increasing SHD. **Limitations.** We assume causal sufficiency; latent-variable (PAG) extensions are future work.

AI USE STATEMENT

Large language models were used to assist with editing and formatting the manuscript. The authors reviewed all content and take full responsibility for it.

REPRODUCIBILITY STATEMENT

Algorithm 1 and Appendix A specify the complete DDQN procedure, architecture, and training settings; Appendix B contains complete proofs; Appendices C–F describe the synthetic experiments, additional results, datasets, and metrics. The full implementation (DDQN agent, Copula-BIC scorer, warm-start initializers, dataset loaders, and reproduction scripts) will be released under an open-source license upon publication.

ETHICS STATEMENT

Observationally inferred graphs should not be treated as established intervention targets, especially under unmeasured confounding or model misspecification. Protein-signaling and diagnostic-network benchmarks motivate the method, but the reported results do not establish clinical validity. Scientific use requires domain expertise, uncertainty assessment, appropriate data permissions, and independent validation of consequential causal claims.

REFERENCES

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, and Huanhuan Chen. Causal structure learning supervised by large language model. *arXiv preprint arXiv:2311.11689*, 2023.

Ingo A Beinlich, Henri Jacques Suermondt, R Martin Chavez, and Gregory F Cooper. The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In *AIME 89: Second European Conference on Artificial Intelligence in Medicine, London, August 29th–31st 1989. Proceedings*, pp. 247–256. Springer, 1989.

Zachris Björkman, Jorge Loría, Sophie Wharrie, and Samuel Kaski. Incorporating expert knowledge into bayesian causal discovery of mixtures of directed acyclic graphs. *arXiv preprint arXiv:2510.06735*, 2025.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Cristina Conati, Abigail S Gertner, Kurt VanLehn, and Marek J Druzdzel. On-line student modeling for coached problem solving using bayesian networks. In *User Modeling: Proceedings of the Sixth International Conference UM97 Chia Laguna, Sardinia, Italy June 2–5 1997*, pp. 231–242. Springer, 1997.

Jan Marco Ruiz de Vargas, Kirtan Padh, and Niki Kilbertus. Cluster-dags as powerful background knowledge for causal discovery. *arXiv preprint arXiv:2512.10032*, 2025.

Ming Gao, Yi Ding, and Bryon Aragam. A polynomial-time algorithm for learning nonparametric causal graphs. *Advances in Neural Information Processing Systems*, 33:11599–11611, 2020.

Uzma Hasan and Md Osman Gani. Kcrl: A prior knowledge based causal discovery framework with reinforcement learning. In *Machine Learning for Healthcare Conference*, pp. 691–714. PMLR, 2022.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Diviyan Kalainathan, Olivier Goudet, and Ritik Dutta. Causal discovery toolbox: Uncovering causal relationships in python. *Journal of Machine Learning Research*, 21(37):1–5, 2020.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. Mach. Learn. Res.*, 8:613–636, May 2007. ISSN 1532-4435.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *International Conference on Learning Representations*, 2020.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(10), 2009.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

Agnieszka Onisko. Probabilistic causal models in medicine: Application to diagnosis of liver disorders. In *Ph. D. dissertation, Inst. Biocybern. Biomed. Eng., Polish Academy Sci., Warsaw, Poland*, 2003.

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(5721):523–529, 2005.

Andreas WM Sauter, Nicolò Botteghi, Erman Acar, and Aske Plaat. Core: Towards scalable and efficient causal discovery with reinforcement learning. *arXiv preprint arXiv:2401.16974*, 2024.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pp. 461–464, 1978.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12 (Apr):1225–1248, 2011.

David J Spiegelhalter, A Philip Dawid, Steffen L Lauritzen, and Robert G Cowell. Bayesian analysis in expert systems. *Statistical science*, pp. 219–247, 1993.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Samir Wadhwa and Roy Dong. On the sample complexity of causal discovery and the value of domain expertise. *arXiv preprint arXiv:2102.03274*, 2021.

Xiaoqiang Wang, Yali Du, Shengyu Zhu, Liangjun Ke, Zhitang Chen, Jianye Hao, and Jun Wang. Ordering-based causal discovery with reinforcement learning. *arXiv preprint arXiv:2105.06631*, 2021.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1g2skStPB.

A Algorithm, notation, and implementation details

A.1 Notation and observational model

Table 3: Notation. Sample indices are superscripts, variable indices are subscripts, and graph-dependent quantities carry the subscript G.

Symbol	Meaning
Data and score	
p, n, m	Variables; training observations; independent evaluation observations
$x_j^{(i)}, U_j^{(i)}, z_j^{(i)}$	Raw, rank-transformed, and Gaussianized value of variable j in observation i
M, Φ	Box bound of the transformed data, $M_n = \Phi^{-1}(n/(n+1))$; standard normal distribution function
$\widehat{\alpha}_{Gj}, \widehat{\beta}_{Gj}, \widehat{v}_{Gj}$	Fitted intercept, coefficients, and residual variance of node j under G
$s_G(z), B_n(G), S_n(G)$	Fitted per-observation log score; total Copula-BIC; Copula-BIC per observation
$k(G)$	Number of fitted parameters, $\lvert E_G\rvert + 2p$
Graphs and moves	
$G^\dagger, \mathrm{MEC}(G^\dagger)$	Data-generating DAG and its Markov equivalence class
$G_0, G_t, G', \bar{G}$	Initializer, current graph, proposal, current champion
$A_G, E_G, \mathrm{pa}_G(j)$	Adjacency matrix, edge set, and parent set of G
b	Maximum allowed in-degree
$\mathcal{G}, \mathcal{R}(G_0)$	Admissible DAGs; states reachable from G_0 by accepted moves
$\mathcal{A}(G), \mathcal{N}_1(G), \mathcal{T}(G, a)$	Feasible actions; one-edit neighbors; proposal map for edge operation a
$\mathcal{C}$	Candidate archive, $\lvert\mathcal{C}\rvert \le 1 + NT$
DDQN	
s_t, r_t	Vectorized adjacency state and proposal reward; λ, c are the reward density coefficient and step cost
π, Q^π	Arbitrary fixed policy and its expected discounted-return function
$\theta, \theta^-, \pi_\theta$	Online weights, target weights, ε-greedy behavior policy
$\gamma, \varepsilon, \varepsilon_\star$	Discount; exploration weight; exploration lower bound (Definition 2)
$\mathcal{B}, q, \tau, \mu$	Replay buffer; minibatch size; target mixing rate; learning rate
y, d	DDQN target and termination indicator, Eq. (7)
N, T	Episode budget; proposals per episode
Hitting time (Theorem 1)	
h, d_{ws}	Progress potential; warm-start depth $h(G_0)$
A_k, ϕ_k	Bound on $\lvert\mathcal{A}(G)\rvert$ and number of progress actions at level k
$\rho_k, \nu_k, \sigma_{\mathrm{loc}}$	Per-proposal progress probability $\varepsilon_\star \phi_k / A_k$; proposals spent at level k; proposals to a 1-optimal graph
$h^\star$	Length of the longest improving path
Selection (Theorem 2)	
$\mathscr{F}$	Frozen training and search information
$\widehat{\Psi}_{n,j}, \eta, g_{\mathscr{F}}$	Training empirical distribution function of variable j; clipping level; frozen marginal map
pen_G, f_G	Training penalty $k(G)\log n/(2n)$; frozen evaluation score $s_G \circ g_{\mathscr{F}} - \mathrm{pen}_G$
$J, \widehat{J}_m$	Conditional population score and its empirical mean on m evaluation observations
$\mathcal{C}^*, \Delta_G, v_G$	Population maximizers; gap and gap variance of candidate G
R, W, δ	Centered range bound; range width of score differences; target error probability
$\mathcal{E}, \kappa$	Event of training/search coverage and its failure probability
Regularity (Lemma 1, Proposition 1)	
$C_\beta, \alpha_{\max}, v_{\min}$	Bounds on $\lVert\widehat{\beta}_{Gj}\rVert_1$, $\lvert\widehat{\alpha}_{Gj}\rvert$, and below on $\widehat{v}_{Gj}$ (Assumption 1)
u_{Gj}, H_G, ω_G, L	Residual vector, score Hessian, gradient offset (Eq. (9)), Lipschitz constant
Π_G, Γ_G, D_G	Parent-coefficient matrix, $I - \Pi_G$, and $\mathrm{diag}(\widehat{v}_{Gj})$
$\beta_{\max}, K_\beta$	Bound on $\lvert\widehat{\beta}_{Gji}\rvert$; $1 + \sqrt{pb}\,\beta_{\max}$

A causal DAG model can be written $X_j = f_j(X_{\mathrm{pa}_{G^\dagger}(j)}, \xi_j)$ in a topological order, with mutually independent exogenous variables ξ_j. Causal sufficiency excludes unobserved common causes of the modeled variables. The Markov property implies that d-separation in the graph gives conditional

independence in the distribution; faithfulness assumes the converse. DAGs with the same skeleton and unshielded colliders form an MEC under this conditional-independence interpretation (Spirtes et al., 2001). A CPDAG represents the equivalence class rather than selecting a unique orientation for every edge. Neither benchmark-directed SHD nor a best-scoring DAG should be mistaken for an identifiable orientation under these assumptions alone.

Any valid DAG can initialize the procedure: a LiNGAM or additive-noise output, a consistent DAG extension of a PC/GES CPDAG (preserving compelled directions and introducing no new unshielded colliders), or a graph supplied by an expert or a language model. A PAG from a latent-variable method has no causally sufficient DAG counterpart and is outside our model class.

A.2 Implementation remarks on Algorithm 1

Operation masks are used both for behavior-policy selection and for DDQN bootstrapping. After a rejected proposal, the stored next state is the unchanged current graph, and the rejected graph is still added to the archive. Episodes terminate at the horizon T; the state-only Q-network is a time-homogeneous approximation of the finite-horizon value function. None of our proofs relies on convergence of the neural value function.

Table 4: Algorithm–theorem and experiment mapping. The statistical theorem concerns a separate evaluation pass after the archive is frozen.

Component	Required mechanism and scope
Theorem 1	Reset to the champion (line 5), persistent exploration (line 8), monotone acceptance (line 13), and Assumption 2; the DDQN update itself is not required.
Corollary 1	Requires an improving neighbor outside the true MEC, in addition to reaching a local optimum.
Theorem 2	Finite archive (line 11); freeze transform, fitted parameters, candidates, and penalties; evaluate independent observations. No RL assumption.
Lemma 1	Quadratic fitted score, bounded transformed input, bounded coefficients, and positive variance floor; no policy-learning assumption. Gives the range $R = 4\sqrt{p}\,ML$ used by Lemma 2 and Theorem 2.
Non-degradation	Retain G_0 and maximize the same declared selection score over the archive.
No-warm-start ablation	Replace G_0 by the empty graph; retain the proposal loop.
Gaussian-score ablation	Omit the marginal transform; change the score everywhere, not only in the reward.
Hill-climbing control	Replace the learned proposal policy by ε-greedy hill climbing; keep the warm start, Copula-BIC, and admissible edits.
Sensitivity	Vary ε, T, λ, or c individually.
Synthetic selection	Finite candidate sets built by MEC enumeration; no DDQN training.

A.3 Training settings

Table 5: Training settings used in all experiments.

Setting	Value
Architecture	$p^2 \to 256 \to 256 \to 3p(p-1)$; two hidden ReLU layers
Behavior policy	ε-greedy with masked actions, $\varepsilon = .10$; deterministic argmax tie breaking
Discount	$\gamma = .99$
Horizon / episodes	$T = 100$ proposals per episode / $N = 1{,}000$ episodes
Replay	FIFO buffer of 10,000 transitions, uniform minibatches of $q = 64$
Optimizer	Adam, learning rate $\mu = .001$, one update every 4 environment steps
Target network	Polyak averaging with $\tau = .01$
Reward	Eq. (5) with $\lambda = .5$ and $c = .001$
Copula transform	Midranks for ties

All experiments use the current-graph reward of Eq. (5), in which the total Copula-BIC difference $B_n(G') - B_n(G)$ is normalized by p.

B Detailed Proofs

B.1 Proof of Theorem 1 (Champion-reset hitting time)

Proof. Condition on the training data and on the full search history, including all network updates. While the potential equals $k > 0$, each of at least ϕ_k improving, potential-reducing actions is sampled with conditional probability at least $\varepsilon_\star / A_k$. Thus the probability of leaving level k downward at any proposal is at least $\rho_k = \varepsilon_\star \phi_k / A_k$. The conditional probability of remaining at that level for at least w additional proposals is at most $(1 - \rho_k)^w$, by repeated conditioning. No independence between action draws is needed.

Every accepted move is non-increasing in h, and rejections preserve the graph. A champion reset likewise does not undo progress: under strict acceptance with a fixed score, the terminal accepted graph is the champion, up to irrelevant repeated representations of the same graph. Therefore a level cannot be revisited after it is left downward. If ν_k is the total number of proposals spent at level k, with $\nu_k = 0$ for skipped levels, then $\mathbb{E}[\nu_k \mid \mathcal{D}] \le 1/\rho_k$, and $\sigma_{\mathrm{loc}} = \sum_k \nu_k$. Summing over $k = 1, \dots, d_{\mathrm{ws}}$ proves Eq. (12). Because every nonterminal episode includes at least one proposal, the number of episodes to first hit the target is no larger than the number of proposals. A finite computational budget can of course terminate the implementation before the hitting event.

□

An unconditional finite-depth version. Orient an arc $G \to G'$ whenever an admissible edit strictly improves B_n. The resulting directed graph on the finite DAG class is acyclic, because the score strictly increases along every arc. Let $h^\star(G)$ be the length of the longest directed path starting at G. At any nonterminal G, at least one improving action exists, and every improving successor satisfies $h^\star(G') \le h^\star(G) - 1$. Thus $h = h^\star$ satisfies Assumption 2 with $\phi_k \ge 1$, giving

$$\mathbb{E}[\sigma_{\mathrm{loc}} \mid \mathcal{D}] \le h^\star(G_0) A_{\max} / \varepsilon_\star, \qquad h^\star(G_0) \le |\mathcal{R}(G_0)| - 1.$$

This gives finite expected termination, not a generic polynomial result: the number of reachable DAGs can be very large. Replacing $h^\star(G_0)$ by the length of a shortest improving path is invalid if the policy can take other improving branches. A cold-versus-warm comparison using the same action bound and exploration rate has an upper-bound ratio given by the corresponding progress depths; comparing upper bounds alone does not compare actual expected runtimes.

B.2 Proof of Lemma 1 (Lipschitz per-sample score)

Let u_{Gj}, H_G, and ω_G be as in Eq. (9). Since $s_G(z) = -\frac{1}{2} \sum_j [\log(2\pi \widehat{v}_{Gj}) + (u_{Gj}^{\mathsf{T}} z - \widehat{\alpha}_{Gj})^2 / \widehat{v}_{Gj}]$ with all fitted parameters fixed, differentiation gives $\nabla s_G(z) = H_G z + \omega_G$. These quantities depend on the fitted graph model, not on the particular evaluation observation. Centering the data and using zero intercepts gives $\omega_G = 0$.

Well-conditioned least squares or ridge regression can control fitted coefficients for bounded designs; ridge alone does not enforce a positive residual-variance floor. Such a floor must be assumed or imposed separately, especially if a candidate interpolates the training observations. Any numerical regularization changes the fitted scoring rule and must be declared. The statistical results apply to that declared frozen score rather than claiming that its maximizer is automatically the same as unregularized BIC's.

Proof. The line segment between any z, z' in the box remains in the box. Along that segment, Eq. (9) implies $\|\nabla s_G\|_2 \le \|H_G\|_{\mathrm{op}} \sqrt{p} M + \|\omega_G\|_2$. Integrating the directional derivative along the segment yields $|s_G(z) - s_G(z')| \le L \|z - z'\|_2$. □

The training rank transform has $M = M_n = O(\sqrt{\log n})$. This does not imply an unconditional $L = O(\sqrt{\log n})$ uniformly in dimension or model fits: the Hessian and intercept norms can also depend on p, n, conditioning, and residual variances. Moreover, $\max_{G,i} \|\nabla s_G(z^{(i)})\|_2$ is an *observed-point diagnostic*, not a box-wide upper bound; the maximum gradient can occur between or beyond observed points. Evaluating a general dense quadratic gradient costs $O(p^2)$ per observation, or $O(pb)$ using sparse node residuals, not universally $O(p)$.

B.3 AUXILIARY RANGE BOUND

Lemma 3 (Bounded score differences). *Let $Q = M(1 + C_\beta) + \alpha_{\max}$. Under Assumption 1, for every $G \in \mathcal{C}$, s_G restricted to $[-M, M]^p$ takes values in an interval of width at most $pQ^2/(2v_{\min})$. Consequently, each pairwise difference $f_{G_1} - f_{G_2}$, $G_1, G_2 \in \mathcal{C}$, has range width at most*

$$W = \frac{pQ^2}{v_{\min}}. \tag{17}$$

Its centered absolute value is at most W, and its variance is at most $W^2/4$.

Proof. Each residual has absolute value at most Q, so its squared, variance-normalized contribution lies in $[0, Q^2/\widehat{v}_{Gj}]$. The log-variance terms and penalties are constant with respect to the evaluation observation. Summing and multiplying by $1/2$ bounds the score's range. A difference of two scores has width at most the sum of their widths. The centered absolute-value bound follows because the mean lies inside the range; the variance bound is the elementary bounded-interval variance inequality. □

Range from the Lipschitz constant. By Lemma 1, $|s_G(z) - s_G(z')| \le L\|z - z'\|_2 \le 2\sqrt{p}\, ML$ for all $z, z' \in [-M, M]^p$, since the diameter of the box is $2\sqrt{p}\, M$. Hence each s_G has range width at most $2\sqrt{p}\, ML$, and a difference of two scores has width at most $4\sqrt{p}\, ML$. The penalties pen_G are constant and do not change this width. Either this bound or W of Eq. (17) may serve as R in Lemma 2 and Theorem 2; combined with Proposition 1, it makes R explicit in the coefficient bound $\beta_{\max}$, the in-degree b, and $v_{\min}$.

B.4 PROOF OF LEMMA 2 (SUB-EXPONENTIAL SCORE DIFFERENCES)

Proof. Let $V = Y - \mathbb{E}Y$, a centered random variable with $|V| \le R$ and variance v. For every integer moment order $\ell \ge 2$, $\mathbb{E}|V|^\ell \le R^{\ell-2}\mathbb{E}V^2 = vR^{\ell-2}$. Expanding $\mathbb{E}e^{tV} = 1 + \sum_{\ell \ge 2} t^\ell \mathbb{E}V^\ell/\ell!$ and using $\ell! \ge 2 \cdot 3^{\ell-2}$ for $\ell \ge 2$ gives $\mathbb{E}e^{tV} \le 1 + \frac{t^2 v}{2}\sum_{\ell \ge 2}(|t|R/3)^{\ell-2}$; summing the geometric series and using $\log(1 + x) \le x$ yields

$$\log \mathbb{E}e^{tV} \le \frac{t^2 v}{2(1 - |t|R/3)}, \qquad |t| < 3/R.$$

This conditional sub-exponential moment bound, applied to i.i.d. copies $V_1, \ldots, V_m$ from independent evaluation rows and optimized by a Chernoff argument, gives, for every $u > 0$, $\mathbb{P}(m^{-1}\sum_i V_i \le -u) \le \exp[-mu^2/(2v + 2Ru/3)]$. Dependence among variables within an observation is permitted; independence is required across evaluation observations. No distribution-free $O(p \log n)$ variance bound follows solely from clipping correlated coordinates.

□

B.5 EXPLICIT LIPSCHITZ UPPER BOUND

Proposition 1 (Coefficient-dependent upper bound on L). *Let Assumption 1 hold, let $\beta_{\max}$ bound the absolute value of every fitted parent coefficient (every entry of every $\widehat{\beta}_{Gj}$, $G \in \mathcal{C}$), let b be the maximum in-degree of the candidates, and set $K_\beta = 1 + \sqrt{pb}\, \beta_{\max}$. Write $\widehat{\alpha}_G = (\widehat{\alpha}_{G1}, \ldots, \widehat{\alpha}_{Gp})^\mathsf{T}$ for the intercept vector of G. Then the constant L of Lemma 1 satisfies*

$$L \le \frac{K_\beta^2 \sqrt{p} M + K_\beta \max_G \|\widehat{\alpha}_G\|_2}{v_{\min}}. \tag{18}$$

For zero-intercept centered fits and $M = \Phi^{-1}(n/(n + 1))$, this restores the bound $L \le (1 + \sqrt{pb}\, \beta_{\max})^2 \sqrt{p}\, \Phi^{-1}(n/(n + 1))/v_{\min}$.

Proof. For each candidate, let $\Pi_G \in \mathbb{R}^{p\times p}$ have row j equal to $\widehat{\beta}_{Gj}$ on the columns $\mathrm{pa}_G(j)$ and zero elsewhere, $\Gamma_G = I - \Pi_G$, and $D_G = \mathrm{diag}(\widehat{v}_{G1}, \ldots, \widehat{v}_{Gp})$. The residual vector is $\Gamma_G z - \widehat{\alpha}_G$, hence

$$s_G(z) = -\tfrac{1}{2}(\Gamma_G z - \widehat{\alpha}_G)^\mathsf{T} D_G^{-1}(\Gamma_G z - \widehat{\alpha}_G) - \tfrac{1}{2}\sum_j \log(2\pi\widehat{v}_{Gj}), \tag{19}$$

$$H_G = -\Gamma_G^\mathsf{T} D_G^{-1}\Gamma_G, \qquad \omega_G = \Gamma_G^\mathsf{T} D_G^{-1}\widehat{\alpha}_G. \tag{20}$$

Since at most pb entries of Π_G are nonzero, $\|\Pi_G\|_F^2 \le pb\beta_{\max}^2$. Therefore

$$\|\Gamma_G\|_{\mathrm{op}} \le 1 + \|\Pi_G\|_{\mathrm{op}} \le 1 + \|\Pi_G\|_F \le K_\beta.$$

Also $\|D_G^{-1}\|_{\mathrm{op}} \le 1/v_{\min}$, which gives

$$\|H_G\|_{\mathrm{op}} \le K_\beta^2/v_{\min}, \qquad \|\omega_G\|_2 \le K_\beta\|\widehat{\alpha}_G\|_2/v_{\min}.$$

Substitute these bounds in Lemma 1. For centered zero-intercept fits, $\omega_G = 0$ exactly. For fixed $b, \beta_{\max}, v_{\min}$ this centered upper bound is $O(p^{3/2}\sqrt{\log n})$; uniformity in fitted parameters requires the stated bounds, not merely finite observed gradients. □

B.6 Proof of Theorem 2 (Finite-sample selection)

An explicit frozen marginal map can use the training empirical distribution $\widehat{\Psi}_{n,j}$:

$$g_{\mathscr{F},j}(x) = \Phi^{-1}\Big(\min\{1-\eta, \max\{\eta, \widehat{\Psi}_{n,j}(x)\}\}\Big), \qquad 0 < \eta < \tfrac{1}{2}, \tag{21}$$

where η is fixed using training information (for example, $\eta = 1/(n+1)$ for $n \ge 2$). Every evaluation observation is transformed individually using this function. There is no rank recomputation, coefficient refit, candidate search, hyperparameter tuning, or choice of clipping level based on evaluation scores. Independence of evaluation observations from $\mathscr{F}$ then makes the score vectors independent and identically distributed conditional on $\mathscr{F}$. Clipping and finite training fits can change the population target, which is why the theorem targets J, not an unqualified true DAG.

Proof. Choose a training-measurable $G^* \in \mathcal{C}^*$, for example by a fixed ordering. If an empirical maximizer is outside $\mathcal{C}^*$, then at least one suboptimal G satisfies $\widehat{J}_m(G) \ge \widehat{J}_m(G^*)$, implying $m^{-1}\sum_i Y_G^{(i)} \le 0$ for $Y_G^{(i)} = f_{G^*}(X^{(i)}) - f_G(X^{(i)})$. Center each difference by its positive conditional mean Δ_G and apply the Bernstein tail from the proof of Lemma 2 (Appendix B.4) with $u = \Delta_G$. A union bound over suboptimal candidates proves Eq. (14). The ratio $\Delta^2/(2v + (2/3)R\Delta)$ is increasing in $\Delta > 0$ and decreasing in v, so replacing the gaps by $\Delta_{\min}$ and variances by $v_{\max}$ yields Eq. (15).

No independence across candidate graphs is needed for this union bound. Ties among population maximizers cause no problem because the error event is selection outside the entire maximizing set. Under misspecification, this set can contain no member of the true MEC. The proof remains a valid predictive-score selection result, without becoming a causal-identification result.

□

Why conditioning on the search archive is insufficient. When ranks, fitted coefficients, and candidates are all determined from the same sample being scored, the per-observation score terms are not fixed independent functions of independent inputs. Conditioning on the realized archive alone does not repair this dependence. One alternative to independent evaluation is a uniform bound over an appropriately controlled, predetermined function class that includes all fitted models and transformations; its complexity is not generally $\log|\mathcal{C}|$ for the realized archive. Establishing such a same-sample bound is outside the present result. Similarly, repeatedly inspecting evaluation data and extending the search would invalidate the fixed-archive analysis without a further sequential correction.

Composition with causal coverage. Let $\mathcal{E}$ be a training/search event on which the archive contains a correctly specified target and its population-optimal fitted candidates all belong to $\mathrm{MEC}(G^\dagger)$. If $\mathbb{P}(\mathcal{E}^c) \le \kappa$ and the conditional selection error is bounded by δ on $\mathcal{E}$, then

$$\mathbb{P}(\widehat{G}_{\mathrm{eval}} \notin \mathrm{MEC}(G^\dagger)) \le \mathbb{P}(\mathcal{E}^c) + \mathbb{E}[\mathbf{1}_{\mathcal{E}}\mathbb{P}(\widehat{G}_{\mathrm{eval}} \notin \mathcal{C}^* \mid \mathscr{F})] \le \kappa + \delta.$$

The corollary in the main text can help establish empirical-search coverage under its strong landscape condition, but does not prove population-score alignment or identify κ. This is the precise sense in which the two theorems cannot simply be stacked into an unconditional recovery guarantee.

C SYNTHETIC SELECTION AND MISSPECIFICATION DIAGNOSTICS

These experiments verify Theorem 2 on finite candidate sets without RL. For each data-generating process (DGP) and sample size, the candidate set $\mathcal{C}$ is built by MEC enumeration, the highest-scoring candidate is selected, and the mis-selection probability is estimated over independent trials. $\widehat{\Delta}_n$ is the empirical score gap between the population-best candidate and its closest competitor, and "best SHD" is the SHD of the population-best candidate.

Table 6: Theorem 2 under non-Gaussian copulas (copula parameter in parentheses). Mis-selection frequencies over 40 trials, with exact 95% binomial CIs in brackets, and empirical gap $\widehat{\Delta}_n$.

Copula / best SHD	Quantity	$n = 500$	1,000	2,000	4,000
Frank (3) / 0	Mis-selection	.100 [0.028,0.237]	.050 [0.006,0.169]	.025 [0.001,0.132]	.000 [0.000,0.088]
	$\widehat{\Delta}_n$	.006	.008	.010	.013
Gumbel (2) / 0	Mis-selection	.150 [0.057,0.298]	.075 [0.016,0.204]	.025 [0.001,0.132]	.000 [0.000,0.088]
	$\widehat{\Delta}_n$	.005	.007	.009	.012
Clayton (2) / 0	Mis-selection	.175 [0.073,0.328]	.100 [0.028,0.237]	.050 [0.006,0.169]	.025 [0.001,0.132]
	$\widehat{\Delta}_n$	.004	.006	.008	.011
Clayton (5) / 2	Mis-selection	.250 [0.127,0.412]	.175 [0.073,0.328]	.100 [0.028,0.237]	.050 [0.006,0.169]
	$\widehat{\Delta}_n$	.002	.003	.004	.006

Across all four copula families, mis-selection decreases monotonically with n while $\widehat{\Delta}_n$ grows, as predicted by Theorem 2. Under the strongly asymmetric Clayton copula with parameter 5, the population-best candidate has SHD 2: selection remains consistent, while the Gaussian-copula score is no longer perfectly aligned with the true graph.

Table 7: Theorem 2 under non-Gaussian noise and nonlinear mechanisms: mis-selection probability and empirical gap $\widehat{\Delta}_n$.

DGP / best SHD	Quantity	400	600	800	1,000
Linear Student-t(3) / 3	Mis-selection	.200	.125	.075	.0025
	$\widehat{\Delta}_n$	.035	.038	.041	.044
Linear exponential / 3	Mis-selection	.100	.075	.025	.00139
	$\widehat{\Delta}_n$	.040	.043	.046	.049
DGP / best SHD	**Quantity**	**500**	**1,000**	**2,000**	**4,000**
Tanh + Gaussian / 0	Mis-selection	.075	.050	.025	.00304
	$\widehat{\Delta}_n$	.002	.003	.005	.007
Tanh + t(4) / 0	Mis-selection	.150	.100	.050	.0001
	$\widehat{\Delta}_n$	.0012	.004	.005	.007

Under linear non-Gaussian noise the population-best candidate has SHD 3, whereas under nonlinear tanh mechanisms it is the true DAG (SHD 0). In all settings mis-selection decreases and $\widehat{\Delta}_n$ increases with n, so the concentration predicted by Theorem 2 persists well beyond the linear-Gaussian regime.

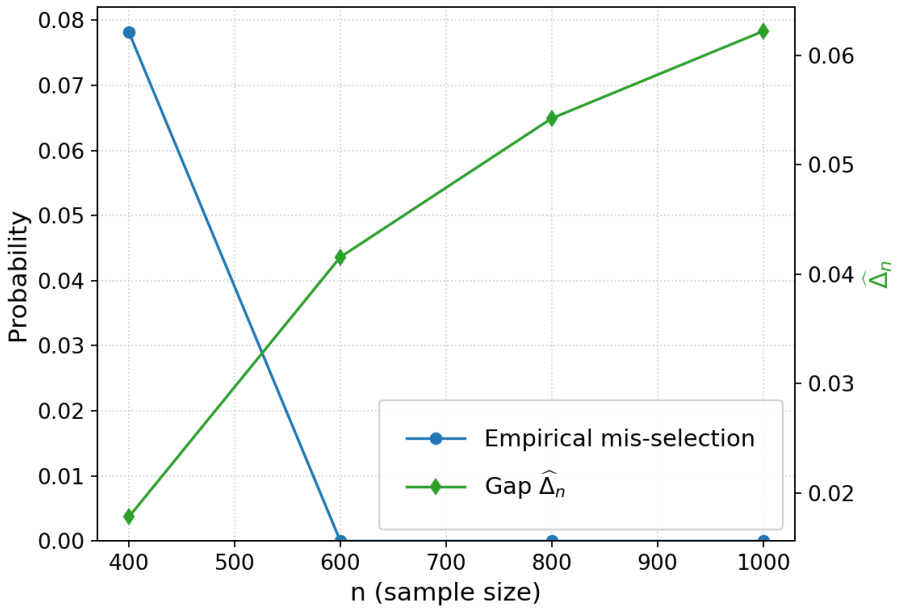

Figure 2: Theorem 2 in the linear-Gaussian setting ($p = 30$): mis-selection probability (left axis) decreases and the empirical gap $\widehat{\Delta}_n$ (right axis) increases with n.

Table 8: Empirical Lipschitz constants $\widehat{L} = \max_{G,i} \|\nabla s_G(z^{(i)})\|_2$ ($p = 20$, $n = 2{,}000$).

DGP	$\widehat{L}$
Linear Gaussian	15.9
Linear Student-t(3)	48.0
Linear exponential	29.1
Tanh + t(4)	10.3

All empirical values of $\widehat{L}$ are finite and moderate; heavy-tailed noise yields the largest.

D Complete benchmark results and component controls

D.1 Composite scores

Figure 3 shows the composite scores of Table 1 with 95% CI error bars; Table 13 lists the corresponding confidence limits.

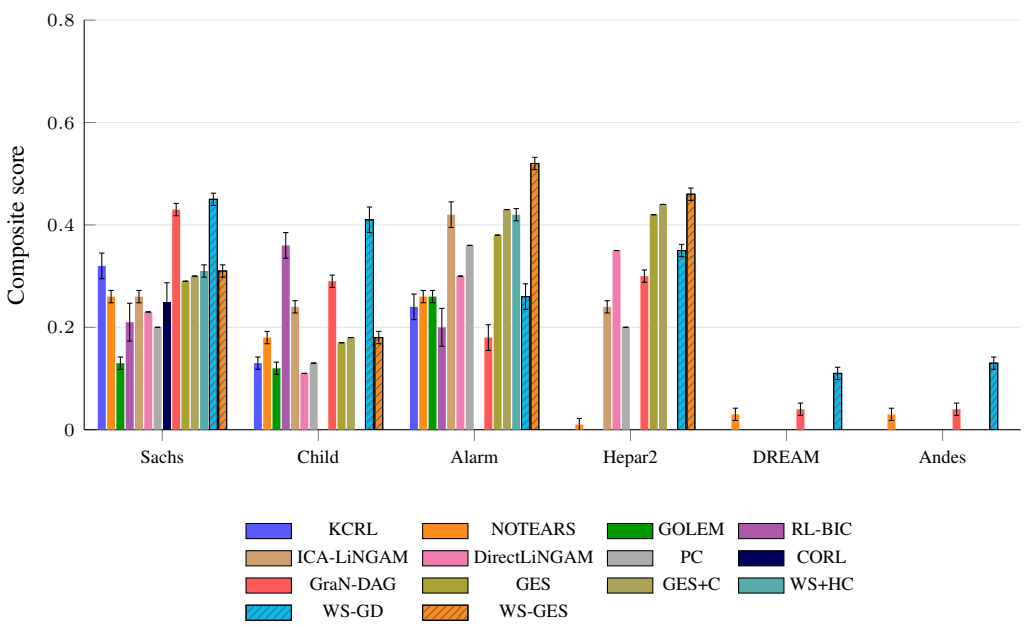

Figure 3: Composite scores on all benchmarks (Table 1) with 95% CI error bars over five runs; hatched bars: our method. Missing bars: no convergence or not run.

D.2 Copula-scoring and hill-climbing controls

Table 9: Copula-scoring control: GES with standard BIC, GES with Copula-BIC (GES+C), and our refinement (GraN-DAG warm start on Sachs/Child, GES on Alarm/Hepar2). ±: 95% CI half-width. Bold: best value, ties included.

Dataset	Method	TPR ↑	FDR ↓	SHD ↓	Score ↑
Sachs	GES	.48 ± .000	.64 ± .000	28 ± 0.00	.29 ± .000
	GES+C	.50 ± .000	.61 ± .000	26 ± 0.00	.30 ± .000
	WS-GD	**.55** ± .025	**.28** ± .037	**10** ± 0.62	**.45** ± .012
Child	GES	.38 ± .000	.89 ± .000	34 ± 0.00	.17 ± .000
	GES+C	.40 ± .000	.87 ± .000	33 ± 0.00	.18 ± .000
	WS-GD	**.72** ± .037	**.55** ± .050	**21** ± 1.86	**.41** ± .025
Alarm	GES	.74 ± .000	.61 ± .000	56 ± 0.00	.38 ± .000
	GES+C	.78 ± .000	.47 ± .000	40 ± 0.00	.43 ± .000
	WS-GES	**.82** ± .025	**.30** ± .037	**26** ± 2.48	**.52** ± .012
Hepar2	GES	.50 ± .000	.23 ± .000	**70** ± 0.00	.42 ± .000
	GES+C	.54 ± .000	**.22** ± .000	71 ± 0.00	.44 ± .000
	WS-GES	**.61** ± .025	.23 ± .037	**70** ± 4.97	**.46** ± .012

Table 9 separates the effect of the score from that of refinement.

Table 10: Hill-climbing controls. HC starts from GES with Copula-BIC; HC×10 uses ten times its search budget. WS-GD on Sachs starts from GraN-DAG. ±: 95% CI half-width over five runs. Bold: best value.

Dataset	Method	TPR ↑	FDR ↓	SHD ↓	Score ↑
Sachs	GES	$.48 \pm .000$	$.64 \pm .000$	28 ± 0.00	$.29 \pm .000$
	GES + HC	$.50 \pm .025$	$.60 \pm .037$	26 ± 1.24	$.31 \pm .012$
	GES + HC×10	.51	.58	25	.32
	WS-GES	$.52 \pm .025$	$.62 \pm .037$	28 ± 1.24	$.31 \pm .012$
	WS-GD	$\mathbf{.55} \pm .025$	$\mathbf{.28} \pm .037$	$\mathbf{10} \pm 0.62$	$\mathbf{.45} \pm .012$
Alarm	GES	$.74 \pm .000$	$.61 \pm .000$	56 ± 0.00	$.38 \pm .000$
	GES + HC	$.77 \pm .025$	$.49 \pm .037$	42 ± 2.48	$.42 \pm .012$
	GES + HC×10	.79	.43	37	.46
	WS-GES	$\mathbf{.82} \pm .025$	$\mathbf{.30} \pm .037$	$\mathbf{26} \pm 2.48$	$\mathbf{.52} \pm .012$

Copula-BIC alone improves GES only moderately (e.g., Alarm SHD $56 \rightarrow 40$), whereas refinement yields the largest gains (Alarm SHD 26).

Table 10 compares learned refinement with ε-greedy hill climbing from the same warm start, with the same score and admissible edits, on Sachs and Alarm. Theorem 1 covers both proposal rules; the comparison shows that learning the proposal policy adds value in practice, since on Alarm even ten times the hill-climbing budget does not reach the learned refinement.

D.3 Alternative language-model initializers

Table 11: Zero-shot GPT-4o graphs as alternative initial graphs, compared with the conventional initializer and our refinement. ±: 95% CI half-width.

Dataset	Method	TPR ↑	FDR ↓	SHD ↓	Score ↑
Child	GraN-DAG	$.50 \pm .037$	$.67 \pm .050$	25 ± 2.48	$.29 \pm .012$
	GPT-4o	.29	.74	38	.18
	WS-GD	$\mathbf{.72} \pm .037$	$\mathbf{.55} \pm .050$	$\mathbf{21} \pm 1.86$	$\mathbf{.41} \pm .025$
Alarm	GES	$.74 \pm .000$	$.61 \pm .000$	56 ± 0.00	$.38 \pm .000$
	GPT-4o	.17	.55	48	.21
	WS-GES	$\mathbf{.82} \pm .025$	$\mathbf{.30} \pm .037$	$\mathbf{26} \pm 2.48$	$\mathbf{.52} \pm .012$
Hepar2	GraN-DAG	$.28 \pm .025$	$.39 \pm .050$	96 ± 7.45	$.30 \pm .012$
	GPT-4o	.09	.63	131	.16
	WS-GES	$\mathbf{.61} \pm .025$	$\mathbf{.23} \pm .037$	$\mathbf{70} \pm 4.97$	$\mathbf{.46} \pm .012$

GPT-4o graphs have lower TPR and composite score than the conventional initializers on all three networks, although on Alarm they have lower SHD and FDR than GES. Because the procedure accepts any valid DAG as G_0, LLM- or expert-provided graphs can be used directly as warm starts.

E Uncertainty, sensitivity, and step-cost results

E.1 Confidence intervals

Table 12: Five-run means with 95% CI half-widths for selected methods.

Dataset	Method	TPR ↑	FDR ↓	SHD ↓	Score ↑
Alarm	GES	$.74 \pm .000$	$.61 \pm .000$	56 ± 0.00	$.38 \pm .000$
	GraN-DAG	$.24 \pm .050$	$.73 \pm .062$	60 ± 4.97	$.18 \pm .025$
	RL-BIC	$.30 \pm .074$	$.74 \pm .087$	56 ± 6.21	$.20 \pm .037$
	WS-GES	$.82 \pm .025$	$.30 \pm .037$	26 ± 2.48	$.52 \pm .012$
DREAM	GraN-DAG	$.09 \pm .025$	$.97 \pm .012$	251 ± 9.93	$.04 \pm .012$
	NOTEARS	$.07 \pm .012$	$.97 \pm .012$	293 ± 13.66	$.03 \pm .012$
	WS-GD	$.15 \pm .025$	$.82 \pm .037$	184 ± 8.69	$.11 \pm .012$

All half-widths are 95% Student-t intervals over five runs, $\bar{s} \pm t_{.975,4}\widehat{\sigma}/\sqrt{5} \approx \bar{s} \pm 1.24\,\widehat{\sigma}$. GES, PC, DirectLiNGAM, and GES+C are deterministic on fixed data and therefore have zero run-to-run variation. Table 13 lists the confidence limits for every entry of Table 1.

Table 13: 95% confidence limits for every entry of Table 1.

Dataset	Method	TPR	FDR	SHD	Score
Sachs	KCRL	[0.300, 0.400]	[0.388, 0.512]	[13.14, 16.86]	[0.295, 0.345]
Sachs	NOTEARS	[0.288, 0.312]	[0.565, 0.615]	[10.76, 13.24]	[0.248, 0.272]
Sachs	GOLEM	[0.168, 0.192]	[0.793, 0.867]	[22.14, 25.86]	[0.118, 0.142]
Sachs	RL-BIC	[0.178, 0.302]	[0.596, 0.744]	[11.52, 16.48]	[0.173, 0.247]
Sachs	ICA-LiNGAM	[0.208, 0.232]	[0.488, 0.512]	[12.76, 15.24]	[0.248, 0.272]
Sachs	DirectLiNGAM	[0.120, 0.120]	[0.500, 0.500]	[15.00, 15.00]	[0.230, 0.230]
Sachs	PC	[0.330, 0.330]	[0.770, 0.770]	[30.00, 30.00]	[0.200, 0.200]
Sachs	CORL	[0.396, 0.544]	[0.683, 0.857]	[23.52, 28.48]	[0.213, 0.287]
Sachs	GraN-DAG	[0.513, 0.587]	[0.250, 0.350]	[10.76, 13.24]	[0.418, 0.442]
Sachs	GES	[0.480, 0.480]	[0.640, 0.640]	[28.00, 28.00]	[0.290, 0.290]
Sachs	GES+C	[0.500, 0.500]	[0.610, 0.610]	[26.00, 26.00]	[0.300, 0.300]
Sachs	HC	[0.475, 0.525]	[0.563, 0.637]	[24.76, 27.24]	[0.298, 0.322]
Sachs	WS-GD	[0.525, 0.575]	[0.243, 0.317]	[9.38, 10.62]	[0.438, 0.462]
Sachs	WS-GES	[0.495, 0.545]	[0.583, 0.657]	[26.76, 29.24]	[0.298, 0.322]
Child	KCRL	[0.113, 0.187]	[0.738, 0.862]	[26.14, 29.86]	[0.118, 0.142]
Child	NOTEARS	[0.108, 0.132]	[0.595, 0.645]	[20.76, 23.24]	[0.168, 0.192]
Child	GOLEM	[0.088, 0.112]	[0.755, 0.805]	[22.76, 25.24]	[0.108, 0.132]
Child	RL-BIC	[0.366, 0.514]	[0.365, 0.415]	[19.14, 22.86]	[0.335, 0.385]
Child	ICA-LiNGAM	[0.228, 0.252]	[0.538, 0.562]	[19.76, 22.24]	[0.228, 0.252]
Child	DirectLiNGAM	[0.120, 0.120]	[0.820, 0.820]	[28.00, 28.00]	[0.110, 0.110]
Child	PC	[0.240, 0.240]	[0.860, 0.860]	[43.00, 43.00]	[0.130, 0.130]
Child	GraN-DAG	[0.463, 0.537]	[0.620, 0.720]	[22.52, 27.48]	[0.278, 0.302]
Child	GES	[0.380, 0.380]	[0.890, 0.890]	[34.00, 34.00]	[0.170, 0.170]
Child	GES+C	[0.400, 0.400]	[0.870, 0.870]	[33.00, 33.00]	[0.180, 0.180]
Child	WS-GD	[0.683, 0.757]	[0.500, 0.600]	[19.14, 22.86]	[0.385, 0.435]
Child	WS-GES	[0.295, 0.345]	[0.773, 0.847]	[31.14, 34.86]	[0.168, 0.192]
Alarm	KCRL	[0.280, 0.380]	[0.568, 0.692]	[45.28, 52.72]	[0.215, 0.265]
Alarm	NOTEARS	[0.158, 0.182]	[0.405, 0.455]	[38.52, 43.48]	[0.248, 0.272]
Alarm	GOLEM	[0.138, 0.162]	[0.375, 0.425]	[40.52, 45.48]	[0.248, 0.272]
Alarm	RL-BIC	[0.226, 0.374]	[0.653, 0.827]	[49.79, 62.21]	[0.163, 0.237]
Alarm	ICA-LiNGAM	[0.533, 0.607]	[0.270, 0.370]	[25.90, 32.10]	[0.395, 0.445]
Alarm	DirectLiNGAM	[0.390, 0.390]	[0.500, 0.500]	[40.00, 40.00]	[0.300, 0.300]
Alarm	PC	[0.670, 0.670]	[0.600, 0.600]	[55.00, 55.00]	[0.360, 0.360]
Alarm	GraN-DAG	[0.190, 0.290]	[0.668, 0.792]	[55.03, 64.97]	[0.155, 0.205]
Alarm	GES	[0.740, 0.740]	[0.610, 0.610]	[56.00, 56.00]	[0.380, 0.380]
Alarm	GES+C	[0.780, 0.780]	[0.470, 0.470]	[40.00, 40.00]	[0.430, 0.430]
Alarm	HC	[0.745, 0.795]	[0.453, 0.527]	[39.52, 44.48]	[0.408, 0.432]
Alarm	WS-GD	[0.363, 0.437]	[0.600, 0.700]	[45.28, 52.72]	[0.235, 0.285]
Alarm	WS-GES	[0.795, 0.845]	[0.263, 0.337]	[23.52, 28.48]	[0.508, 0.532]
Hepar2	NOTEARS	[0.008, 0.032]	[0.978, 1.002]	[148.31, 165.69]	[−0.002, 0.022]
Hepar2	ICA-LiNGAM	[0.165, 0.215]	[0.453, 0.527]	[105.79, 118.21]	[0.228, 0.252]
Hepar2	DirectLiNGAM	[0.100, 0.100]	[0.070, 0.070]	[110.00, 110.00]	[0.350, 0.350]
Hepar2	PC	[0.350, 0.350]	[0.750, 0.750]	[172.00, 172.00]	[0.200, 0.200]
Hepar2	GraN-DAG	[0.255, 0.305]	[0.340, 0.440]	[88.55, 103.45]	[0.288, 0.312]
Hepar2	GES	[0.500, 0.500]	[0.230, 0.230]	[70.00, 70.00]	[0.420, 0.420]
Hepar2	GES+C	[0.540, 0.540]	[0.220, 0.220]	[71.00, 71.00]	[0.440, 0.440]
Hepar2	WS-GD	[0.515, 0.565]	[0.448, 0.572]	[77.79, 90.21]	[0.338, 0.362]
Hepar2	WS-GES	[0.585, 0.635]	[0.193, 0.267]	[65.03, 74.97]	[0.448, 0.472]
DREAM	NOTEARS	[0.058, 0.082]	[0.958, 0.982]	[279.34, 306.66]	[0.018, 0.042]
DREAM	GraN-DAG	[0.065, 0.115]	[0.958, 0.982]	[241.07, 260.93]	[0.028, 0.052]
DREAM	WS-GD	[0.125, 0.175]	[0.783, 0.857]	[175.31, 192.69]	[0.098, 0.122]
Andes	NOTEARS	[0.058, 0.082]	[0.958, 0.982]	[297.38, 334.62]	[0.018, 0.042]
Andes	GraN-DAG	[0.065, 0.115]	[0.945, 0.995]	[301.58, 326.42]	[0.028, 0.052]
Andes	WS-GD	[0.155, 0.205]	[0.763, 0.837]	[242.83, 265.17]	[0.118, 0.142]

E.2 HYPERPARAMETER SWEEPS

Table 14: Hyperparameter sensitivity: composite-score means with 95% CI half-widths over five runs. Exploration ε, episode horizon T, and reward density coefficient λ are varied one at a time; defaults ($\varepsilon = .10$, $T = 100$, $\lambda = .5$, $c = .001$) in bold.

Parameter	Value	Alarm	DREAM
ε	.05	$.48 \pm .025$	$.09 \pm .012$
	.10 (default)	$\mathbf{.52 \pm .012}$	$\mathbf{.11 \pm .012}$
	.20	$.50 \pm .025$	$.10 \pm .012$
	.30	$.47 \pm .037$	$.08 \pm .012$
T	50	$.44 \pm .037$	$.08 \pm .012$
	100 (default)	$\mathbf{.52 \pm .012}$	$\mathbf{.11 \pm .012}$
	200	$.53 \pm .012$	$.11 \pm .012$
λ	.1	$.46 \pm .025$	$.09 \pm .012$
	.5 (default)	$\mathbf{.52 \pm .012}$	$\mathbf{.11 \pm .012}$
	1.0	$.49 \pm .025$	$.10 \pm .012$

The defaults attain the best composite score in every sweep except T, where doubling the horizon to $T = 200$ gains only .01 on Alarm and nothing on DREAM. Varying ε or λ changes the Alarm score by at most .06, small relative to the .35 gain over the no-warm-start variant; the default score of each network coincides with its entry in Table 1.

Table 15: Step-cost sensitivity (composite score); all other settings at their defaults.

Step cost c	Alarm score	DREAM score
.0001	.50	.10
.001 (default)	**.52**	**.11**
.01	.51	**.11**
.05	.47	.09

The score varies by at most .02 on Alarm and .01 on DREAM for $c \in [.0001, .01]$, and degrades only at $c = .05$, where the step cost dominates the score signal.

F DATASETS, METRICS, AND COMPUTATION

Table 16: Benchmark dimensions, sample sizes, and mean wall-clock time of our method on one A100 GPU (three runs).

Dataset	Variables	Samples	Time (min)	Domain
Sachs	11	7,446	4	Protein signaling
Child	20	1,000	12	Pediatric diagnosis
Alarm	37	10,000	28	Intensive-care monitoring
Hepar2	70	5,000	55	Liver diagnosis
DREAM	100	100	90	Gene regulation
Andes	223	10,000	120	Physics tutoring

Sachs contains 11 signaling proteins with an experimentally validated ground-truth network (Sachs et al., 2005). **Child** is an expert-designed pediatric-diagnosis network with 20 nodes, 25 arcs, and maximum in-degree 2 (Spiegelhalter et al., 1993). **Alarm** models intensive-care patient monitoring with 37 nodes (Beinlich et al., 1989). **Hepar2** is a liver-disorder diagnosis network with 70 nodes, 123 arcs, and maximum in-degree 6 (Onisko, 2003). **DREAM** is a 100-node gene-regulatory network benchmark (Kalainathan et al., 2020). **Andes** is an intelligent-tutoring network with 223 nodes, 338 arcs, and maximum in-degree 6 (Conati et al., 1997).

F.1 Evaluation metrics

TPR is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ and FDR is $\mathrm{FP}/(\mathrm{TP}+\mathrm{FP})$, computed on directed edges. SHD counts missing, extra, and reversed edges.

The composite score in Eq. (16) weights recall, precision, and a reciprocal SHD term equally; we always report its three components alongside it.

F.2 Computational cost

At fixed hidden width 256, the Q-network has $256p^2$ input and $256 \cdot 3p(p-1)$ output weights, so its size grows as $O(p^2)$. Copula-BIC is decomposable: an addition or deletion changes one node's parent regression and a reversal changes two, so each proposal requires refitting at most two local regressions with at most b parents, at cost $O(nb^2 + b^3)$, which cached covariance statistics amortize. Wall-clock times are reported in Table 16.