RECAP: RELATIVE EVALUATION WITH CONSENSUS AUGMENTED POLICY SEARCH FOR CAUSAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning can search for causal graphs by scoring sampled structures, but learning a graph policy from limited candidate graphs is often unstable. Rewards vary across inputs, while promising edges found in earlier updates may be lost as the policy changes. We propose RECAP, Relative Evaluation with Consensus Augmented Policy Search, which couples within-input relative evaluation with cross-update structural memory. RECAP uses group-relative advantages to distinguish candidate graphs for the same input and guide the current policy update. To carry structural evidence across updates, it aggregates edges from high-reward graphs into an evolving consensus teacher, whose selective, annealed guidance favors recurring edges without permanently constraining the search. A reference-policy KL term moderates policy changes, while the graph policy samples candidates from contextual variable representations and scores them using a BIC-based reward with an acyclicity penalty. Across four benchmarks spanning Bayesian networks, simulated fMRI, and protein signaling, RECAP achieves the lowest SHD and highest precision, with the best or tied-best F1 on three. Ablations confirm the contributions of both components, while equal-budget comparisons further demonstrate RECAP's F1-SHD advantage.

1 INTRODUCTION

Causal structure learning aims to recover directed relationships among variables from observational data. The resulting graphs support intervention-effect estimation, policy evaluation, and mechanistic studies in biology and medicine Pearl (2009); Hernán & Robins (2014); Xiong et al. (2025). Yet a useful structure score alone does not solve the search problem: among the combinatorially many possible graphs, an algorithm must decide which candidates to explore and how to learn from the few it can evaluate. This distinction between scoring a graph and learning to search for one is central to reinforcement-learning-based causal discovery.

A stochastic graph policy offers a way to search directly over discrete structures using a score such as BIC Zhu et al. (2019); Wang et al. (2021). Its effectiveness, however, depends on how a limited set of scored proposals shapes subsequent sampling. Each policy update observes only a small group of candidate graphs for a given input. Their absolute rewards can reflect input-specific score levels as well as differences in graph quality, making it useful to compare candidates within the same input. At the same time, graph search extends over many updates: an edge repeatedly present in high-reward proposals may disappear from the next sampled group. A policy that learns only from the current group can therefore overlook structural evidence accumulated earlier in the search. The challenge is to use current reward differences without discarding recurring high-reward structure.

We address this challenge with **RECAP**, **R**elative **E**valuation with **C**onsensus **A**ugmented **P**olicy Search. RECAP treats graph-policy learning as a search process informed by two timescales of evidence. Within each input, it samples a group of graphs, scores them using a BIC-based reward with an acyclicity penalty, and updates the policy according to their relative advantages. Across updates, it accumulates edge patterns from high-reward proposals to guide later sampling. The first signal distinguishes stronger candidates in the current group; the second carries recurring structural information beyond that group.

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.

The cross-update signal must be used carefully: a pattern found early in the search should inform, rather than permanently determine, future proposals. RECAP therefore maintains an evolving edge-consensus teacher and distills its predictions only on selected, consensus-supported edges. This guidance begins after a warm-up period and decays during training, allowing later policy updates to remain responsive to newly scored graphs. A reference-policy KL term moderates changes in graph probabilities, while entropy maintains exploration. Together, these components let the policy compare current proposals, retain recurring edge evidence, and continue revising its search distribution. Contextual variable representations and shared pairwise scoring parameterize the underlying stochastic graph policy. Figure 1 illustrates the complete search loop.

Experiments on Asia and Cancer Bayesian networks, simulated fMRI networks (NetSim), and the Sachs protein-signaling network evaluate whether this search design improves the recovered structures. Across these four benchmarks, RECAP recovers graphs that are closest to the truth and most selective in the edges they include, achieving the lowest structural Hamming distance (SHD) and the highest precision on every benchmark and the best or tied-best F1 on three of the four. On NetSim, where absolute rewards vary most across inputs, it achieves F1 of 0.56 and SHD of 10.91, compared with a next-highest F1 of 0.49 and a next-lowest SHD of 11.10. Removing either relative evaluation or consensus guidance lowers F1 and increases SHD in all four ablation blocks, and the two removals hurt most on different benchmarks, consistent with their distinct roles. A matched-budget configuration-search comparison on NetSim retains the strongest F1–SHD pair for RECAP. These results connect the two search signals to the quality of the final graphs, rather than relying on a claim about optimization stability alone.

Our contributions are as follows:

- We propose RECAP, a stochastic causal graph-search framework that unifies within-input relative evaluation and cross-update edge consensus in a single policy-learning process. By jointly using the quality differences among current graph proposals and recurring structure from earlier proposals, RECAP turns limited score feedback into guidance for subsequent graph search.
- We develop a selective, transient structural-memory mechanism that turns cross-update edge consensus into soft guidance without making early search decisions permanent. We instantiate it within a contextual graph policy and characterize the properties of the relative update, reference-policy regularization, and consensus guidance.
- We evaluate RECAP across graph families using baseline comparisons, component ablations, and a matched-budget configuration-search comparison. The results show a strong F1–SHD balance and identify the distinct contributions of within-input evaluation and cross-update structural memory.

2 Related Work

2.1 Classical and Functional Causal Discovery

Constraint-based methods recover adjacencies and orientations through conditional-independence tests and orientation rules. PC assumes causal sufficiency; FCI allows confounding Spirtes et al. (2000; 1995). Extensions improve test reliability, introduce kernel-based criteria, and remove order dependence Kalisch & Bühlman (2007); Zhang et al. (2011); Colombo & Maathuis (2014).

Score-based methods search DAGs or Markov equivalence classes by balancing fit and complexity, with greedy equivalence search and BIC providing established search and scoring formulations Chickering (1996; 2002); Schwarz (1978). Functional causal models exploit structural-equation or noise assumptions, as in linear non-Gaussian and additive-noise models Shimizu et al. (2006); HOYER (2009). RECAP learns a distribution over graph proposals from structure scores, emphasizing how policy updates use these evaluations and retain useful edge information across updates.

2.2 Continuous and Ordering-based Structure Learning

Continuous formulations make graph search differentiable through smooth acyclicity constraints Zheng et al. (2018). Later formulations address optimization stability or use log-determinant

acyclicity criteria Nazaret et al. (2024); Bello et al. (2022). Ordering-based methods change the search space instead: diffusion- and score-informed approaches construct causal orderings from which edges can be derived Sanchez et al. (2022); Kang et al. (2025). Analyses of Transformers in causal structure learning motivate contextual, variable-wise representations Nichani et al. (2024). RECAP likewise uses contextual representations, but compares sampled graphs under the original structure reward and carries high-reward edge patterns across policy updates.

2.3 Reinforcement Learning for Graph Search

Policy-gradient methods optimize expected structure scores even when graph proposals and their rewards are discrete Williams (1992); SUTTON (2000). RL-BIC applies this principle directly to BIC-scored graphs Zhu et al. (2019). Other RL methods construct or refine topological orderings Wang et al. (2021); Yang et al. (2023), while tree search and continuous action-to-DAG mappings provide alternative proposal mechanisms Darvariu et al. (2025); Duong et al. (2025). These choices determine which structures are explored; policy learning also determines how finite, noisy reward observations affect later proposals Wang et al. (2021); Darvariu et al. (2025).

RECAP combines within-input relative evaluation Shao et al. (2024) with cross-update structural memory. Within-input comparisons emphasize differences among contemporaneous candidates, whereas an annealed edge-consensus teacher carries recurring structure forward. Reference-policy KL and entropy regulate distributional change and exploration. The resulting optimization mechanism complements advances in graph parameterization without requiring a new causal score.

3 Method

3.1 Problem Formulation

Let $\mathbf{X} \in \mathbb{R}^{n \times N}$ contain observations of N variables. RECAP learns a stochastic policy $\pi_\theta(\mathbf{A} \mid \mathbf{X})$ over zero-diagonal adjacency matrices $\mathbf{A} \in \{0,1\}^{N \times N}$. Sampled graphs receive a BIC-based reward with an acyclicity penalty. Within-input comparisons identify strong candidates in each group; edge consensus retains recurring high-reward structure between groups.

3.2 Method Overview

Figure 1 connects the sampling and learning steps. A variable-wise encoder supplies contextual node representations, and a Bernoulli sampler proposes a group of graphs. Their rewards yield relative advantages for a group-relative policy-gradient update, while reference-policy KL moderates policy change and entropy maintains exploration. High-reward proposals also refresh an edge-consensus teacher, whose annealed distillation signal favors recurring edges. Configuration selection varies the encoder or scorer while preserving this optimization loop.

3.3 Graph Policy Parameterization

3.3.1 Variable-wise Context Encoder

Each variable is represented by a token constructed from its observations or a fixed-dimensional embedding. For B data blocks, gated attention mixes $\mathbf{U} \in \mathbb{R}^{B \times N \times d_{\text{in}}}$ across variables, producing contextual node representations. Identity embeddings are reindexed with their variables; Appendix A collects the notation. The initial projection is

$$\mathbf{Z}^{(0)} \;=\; \mathrm{LN}(\mathbf{U})\,\mathbf{W}_{\text{in}} \;+\; \mathbf{E}_{\text{id}}, \tag{1}$$

where $\mathbf{W}_{\text{in}} \in \mathbb{R}^{d_{\text{in}} \times d}$ and $\mathbf{E}_{\text{id}} \in \mathbb{R}^{N \times d}$.

Each of L attention blocks mixes the N variable tokens and gates its H head outputs $\{\mathbf{O}_h\}_{h=1}^{H}$, where $d_h = d/H$. Given normalized states $\bar{\mathbf{Z}}^{(\ell)} = \mathrm{LN}(\mathbf{Z}^{(\ell)})$, the input-dependent gate is

$$\mathbf{G}_h \;=\; \sigma\Big(\bar{\mathbf{Z}}^{(\ell)}\mathbf{W}_h^G + \mathbf{b}_h^G\Big), \tilde{\mathbf{O}}_h \;=\; \mathbf{G}_h \odot \mathbf{O}_h, \tag{2}$$

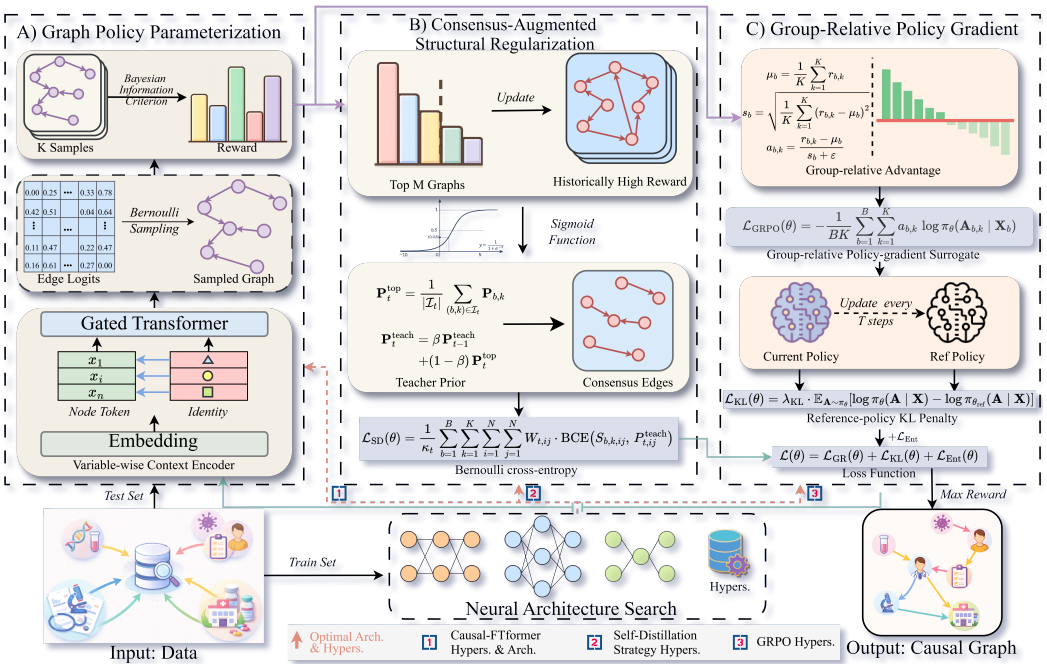


Figure 1: RECAP graph-search loop. Sampled graphs provide group-relative rewards for policy updates and high-reward edges for an annealed consensus teacher.

where $\sigma(\cdot)$ is the sigmoid function, $\odot$ denotes element-wise multiplication, and $\mathbf{G}_h \in \mathbb{R}^{B\times N\times d_h}$. Concatenating gated heads with residual and feed-forward layers yields $\mathbf{Z}^{(L)} \in \mathbb{R}^{B\times N\times d}$, the contextual representations used for edge scoring.

Permutation Equivariance. Without variable-axis positional encodings, shared attention and reindexed identity embeddings permute node representations along with the variables. Shared pairwise scoring then permutes the edge logits and factorized graph policy in the same way (Appendix B.6).

3.3.2 Graph Sampler

Given contextual representations $\mathbf{Z}^{(L)} = [\mathbf{z}_1; \ldots; \mathbf{z}_N]$, a shared pairwise scorer produces a logit for each directed edge $i \to j$. The additive option computes

$$\mathbf{s}_{ij} = \tanh\big(\mathbf{z}_i\mathbf{W}_1 + \mathbf{z}_j\mathbf{W}_2 + \mathbf{b}\big)\mathbf{u}, \tag{3}$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d\times d}$, $\mathbf{b} \in \mathbb{R}^d$, and $\mathbf{u} \in \mathbb{R}^d$ are learnable parameters. The vector $\mathbf{s}_{ij} \in \mathbb{R}^B$ contains the logits for edge $i \to j$ across the mini-batch. Bilinear and neural-tensor alternatives use the same shared pairwise interface and feed the same factorized Bernoulli policy.

For each input b, the logits form $\mathbf{S}_b \in \mathbb{R}^{N\times N}$. We draw K candidates by independently sampling off-diagonal edges; for a deterministic scorer, $S_{b,k,ij} = S_{b,ij}$:

$$A_{b,k,ij} \sim \text{Bernoulli}(\sigma(S_{b,k,ij}))\,, \qquad A_{b,k,ii} = 0, \tag{4}$$

for $b = 1, \ldots, B$ and $k = 1, \ldots, K$.

3.4 Group-Relative Policy Gradient

For input $\mathbf{X}_b$, we sample K graphs $\{\mathbf{A}_{b,k}\}_{k=1}^K \sim \pi_\theta(\cdot \mid \mathbf{X}_b)$ and assign rewards $r_{b,k} = -\text{Score}(\mathbf{A}_{b,k})$. The score combines BIC and an acyclicity penalty; higher rewards favor better-fitting, simpler, more nearly acyclic structures (Appendix E.3). Comparing candidates for the same input emphasizes structural differences rather than shifts in its reward level. Letting $\mu_b = K^{-1}\sum_{j=1}^K r_{b,j}$, we compute

$$a_{b,k} = \frac{r_{b,k} - \mu_b}{\sqrt{(K-1)^{-1}\sum_{j=1}^K (r_{b,j} - \mu_b)^2 + \epsilon_{\text{adv}}}}, \tag{5}$$

The square root is the sample standard deviation of the K rewards, and ϵ_{adv} is added to that deviation.

Centering removes input-specific reward offsets, and the unstandardized centered update aligns with the reward gradient in expectation. Optional standardization controls advantage magnitudes and is invariant to positive reward rescaling when $\epsilon_{\text{adv}} = 0$. Appendix B.2 derives these properties and the exact leave-one-out relation.

Let $\log \pi_\theta(\mathbf{A}_{b,k} \mid \mathbf{X}_b)$ denote the log-probability of a sampled adjacency matrix under the current policy. We update the graph policy using the following group-relative policy-gradient surrogate:

$$\mathcal{L}_{\text{GR}}(\theta) = -\frac{1}{BK} \sum_{b=1}^{B} \sum_{k=1}^{K} a_{b,k} \, \log \pi_\theta(\mathbf{A}_{b,k} \mid \mathbf{X}_b). \tag{6}$$

A reference policy $\pi_{\theta_{\text{ref}}}$, periodically synchronized with the current policy, moderates changes in graph probabilities. Bernoulli edge factorization permits a closed-form reference-policy KL:

$$\mathcal{L}_{\text{KL}}(\theta) = \frac{\lambda_{\text{KL}}}{B} \sum_{b=1}^{B} \frac{1}{N(N-1)} \sum_{i \neq j} \text{KL}\Big(\text{Bern}(p_{b,ij}^{\theta}) \,\|\, \text{Bern}(p_{b,ij}^{\theta_{\text{ref}}})\Big), \tag{7}$$

where $p_{b,ij}^{\theta} = \sigma(S_{b,ij}^{\theta})$ and $p_{b,ij}^{\theta_{\text{ref}}} = \sigma(S_{b,ij}^{\theta_{\text{ref}}})$. For the factorized policy, the joint graph-policy KL equals the sum of the edgewise terms. The coefficient λ_{KL} controls the strength of reference-policy regularization. When the achieved joint KL is small, it bounds changes in expected bounded reward; the edge-averaged penalty introduces an explicit graph-size factor (Appendix B.3).

An entropy bonus keeps the edge policy from concentrating prematurely. With $\mathbb{H}(\pi_\theta(\cdot \mid \mathbf{X}_b))$ denoting the average Bernoulli entropy of its edge probabilities, the loss includes

$$\mathcal{L}_{\text{Ent}}(\theta) = -\lambda_{\text{Ent}} \cdot \frac{1}{B} \sum_{b=1}^{B} \mathbb{H}(\pi_\theta(\cdot \mid \mathbf{X}_b)), \tag{8}$$

with $\lambda_{\text{Ent}} \geq 0$.

3.5 Consensus-Augmented Structural Regularization

The rewards used for within-input policy updates also identify candidates that can inform later updates. Consensus augmentation accumulates their edge information across iterations: an evolving teacher summarizes high-reward proposals and supplies selective soft targets for the policy.

For input $\mathbf{X}_b$ and candidate k, let $\mathbf{S}_{b,k} \in \mathbb{R}^{N \times N}$ denote the edge logits. Candidates can have different sampled adjacencies even when their logits coincide. Their soft edge probabilities are

$$\mathbf{P}_{b,k} = \sigma(\mathbf{S}_{b,k}) \odot (\mathbf{J}_N - \mathbf{I}_N), \tag{9}$$

where $\mathbf{J}_N = \mathbf{1}\mathbf{1}^\top$ is the all-one matrix and $\mathbf{I}_N \in \mathbb{R}^{N \times N}$ is the identity matrix. The teacher matrix $\mathbf{P}_t^{\text{teach}} \in (0,1)^{N \times N}$ tracks edge-wise consensus across iterations. At refresh iteration t, let $\mathcal{I}_t$ index the top-$\min(M_{\text{top}}, BK)$ graphs by reward among the BK current candidates. The teacher aggregates their sampled adjacencies or edge probabilities. With adjacency inputs, the batch consensus is

$$\mathbf{P}_t^{\text{top}} = \frac{1}{|\mathcal{I}_t|} \sum_{(b,k) \in \mathcal{I}_t} \mathbf{A}_{b,k}. \tag{10}$$

With probability inputs, the sum instead uses $\sigma(\mathbf{S}_{b,k})$. At each refresh, an exponential moving average (EMA) updates the teacher:

$$\mathbf{P}_t^{\text{teach}} = \beta \, \mathbf{P}_{t-1}^{\text{teach}} + (1-\beta) \, \mathbf{P}_t^{\text{top}}, \tag{11}$$

Here $\beta \in [0,1)$ determines how long past proposals influence the teacher. Its probabilities are clamped to $[p_{\min}, p_{\max}] \subset (0,1)$ before distillation. A soft Bernoulli cross-entropy aligns current logits with these targets. The binary mask $\mathbf{W}_t \in \{0,1\}^{N \times N}$ restricts that alignment to selected edges. By default, an edge is selected when its teacher probability exceeds $\tau \in (0,1)$:

$$\mathbf{C}_t = \mathbb{I}\big[\mathbf{P}_t^{\text{teach}} \geq \tau\big], \quad \mathbf{W}_t = (\mathbf{J}_N - \mathbf{I}_N) \odot \mathbf{C}_t, \tag{12}$$

where $\mathbb{I}[\cdot]$ is the element-wise indicator function. Given candidate-specific logits $\{\mathbf{S}_{b,k}\}$ at iteration t, the distillation loss is

$$\mathcal{L}_{\mathrm{SD}}(\theta) = \frac{1}{B\kappa_t} \sum_{b=1}^{B} \frac{1}{K} \sum_{k=1}^{K} \sum_{i \neq j} W_{t,ij} \cdot \mathrm{BCE}\left(S_{b,k,ij}, P_{t,ij}^{\mathrm{teach}}\right), \tag{13}$$

where $\mathrm{BCE}(\cdot,\cdot)$ denotes Bernoulli cross-entropy with logits, and $\kappa_t = \max\left(1, \sum_{i \neq j} W_{t,ij}\right)$ normalizes by the number of constrained off-diagonal entries.

Selective, transient guidance. Stopping gradients through the teacher makes Eq. (13) a masked Bernoulli KL up to a teacher-dependent constant. Selected edges receive a directed update; other edges receive no direct teacher gradient. Guidance begins after warm-up and decays thereafter, so later reward-driven updates can adapt. Appendices B.4 and B.5 derive the KL identity and analyze annealing for a fixed base objective.

Self-distillation begins after T_w warm-up iterations. Its nonnegative weight $\lambda_{\mathrm{SD}}(t)$ follows

$$\lambda_{\mathrm{SD}}(t) = \begin{cases} 0, & t < T_w, \\ \lambda_{\mathrm{SD}}^0 \gamma^{t - T_w}, & t \geq T_w, \end{cases} \tag{14}$$

Here λ_{SD}^0 (`distill_coef` in the implementation) sets the initial weight and $\gamma \in (0,1)$ controls its decay, giving $\lambda_{\mathrm{SD}}(t) \to 0$. This annealed loss strengthens recurring high-reward edges while leaving unmasked edges without direct teacher guidance.

3.6 Combined Optimization Objective

The full training objective combines the group-relative policy-gradient surrogate, reference-policy KL, entropy, and the annealed consensus loss:

$$\mathcal{L}_{\mathrm{full}}(\theta, t) = \mathcal{L}_{\mathrm{GR}}(\theta) + \mathcal{L}_{\mathrm{KL}}(\theta) + \mathcal{L}_{\mathrm{Ent}}(\theta) + \lambda_{\mathrm{SD}}(t)\mathcal{L}_{\mathrm{SD}}(\theta). \tag{15}$$

The four objective terms act at different timescales. Relative evaluation favors stronger candidates within a group, and consensus transfers recurring high-reward edges between updates. Reference-policy KL moderates drift, while entropy sustains exploration. Warm-up and annealing bound the period of direct teacher guidance. Appendix B analyzes these component-level properties under stated assumptions; Appendix C specifies the training algorithms.

4 Experiments

4.1 Experimental Setup

Datasets and protocol. The main benchmarks comprise Asia and Cancer Bayesian networks (Scutari, 2010), NetSim simulated fMRI (Smith et al., 2011), and Sachs protein signaling (Sachs et al., 2005). Configuration selection and final evaluation use disjoint sets of 2,000 samples. Main comparisons and ablations report the mean and sample standard deviation over ten runs unless stated otherwise. Appendix E details the protocols and additional datasets. Source code is provided in the Supplementary Material.

Baselines and metrics. Comparators span classical, functional, differentiable, ordering-based, and RL graph-search methods. SHD and precision/FDR are the primary evidence, since the search design targets selective inclusion of recurring edges; F1, recall, FPR, and accuracy complete the error profile. Appendix E defines the metrics and the Holm-adjusted run-paired tests.

4.2 Structural Recovery Across Benchmarks

We first ask whether learning from within-input reward differences and from recurring high-reward edges leads to better final graphs than existing search strategies. Tables 1 and 2 and Figure 2 answer this across four graph families. RECAP consistently recovers the graphs closest to the ground truth: its mean SHD is the lowest on every benchmark (matched only by DAGMA on Cancer), and no

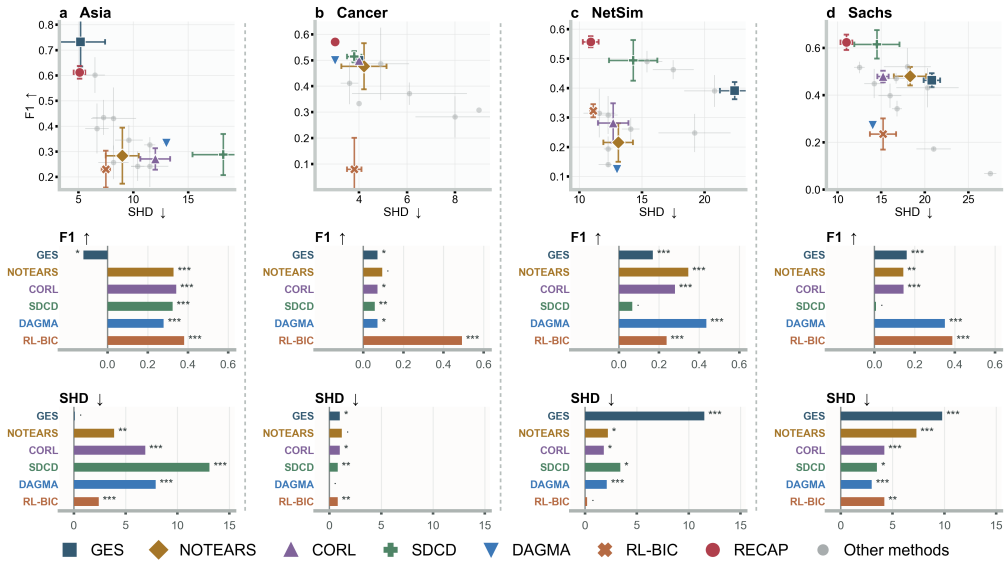


Figure 2: F1 versus SHD on Asia (a), Cancer (b), NetSim (c), and Sachs (d). Top: ten-run means with 95% confidence intervals. Bottom: run-paired advantage of RECAP in F1 and SHD (positive favors RECAP); $*$, $**$, $***$ mark Holm-adjusted $p < 0.05$, 0.01, 0.001, and $\cdot$ non-significance.

Table 1: Asia and Cancer results (ten-run mean $\pm$ standard deviation). Best values are bold; RECAP is shaded.

Dataset	Methods	FDR ↓	FPR ↓	ACC ↑	Precision ↑	Recall ↑	F1 ↑	SHD ↓
Asia	PC (2000)	0.42 ± 0.07	0.07 ± 0.08	0.90 ± 0.02	0.58 ± 0.07	0.62 ± 0.13	0.60 ± 0.10	6.50 ± 1.27
	GES (2002)	0.35 ± 0.15	0.06 ± 0.09	**0.92 ± 0.05**	0.65 ± 0.15	**0.85 ± 0.11**	**0.73 ± 0.14**	5.20 ± 3.12
	ICALiNGAM (2006)	0.51 ± 0.20	0.17 ± 0.08	0.87 ± 0.04	0.49 ± 0.20	0.39 ± 0.15	0.43 ± 0.17	8.20 ± 2.86
	NOTEARS (2018)	0.56 ± 0.28	0.14 ± 0.07	0.86 ± 0.03	0.44 ± 0.28	0.22 ± 0.13	0.28 ± 0.15	9.00 ± 2.11
	GraNDAG (2019)	0.43 ± 0.27	0.08 ± 0.09	0.87 ± 0.03	0.57 ± 0.27	0.18 ± 0.06	0.26 ± 0.09	8.20 ± 1.93
	RL-BIC (2019)	0.73 ± 0.12	0.09 ± 0.02	0.81 ± 0.03	0.27 ± 0.12	0.20 ± 0.09	0.23 ± 0.10	7.50 ± 0.53
	GOLEM (2020)	0.73 ± 0.14	0.26 ± 0.09	0.82 ± 0.04	0.27 ± 0.14	0.22 ± 0.10	0.24 ± 0.11	11.50 ± 2.37
	CORL (2021)	0.73 ± 0.08	0.31 ± 0.09	0.81 ± 0.03	0.27 ± 0.08	0.28 ± 0.05	0.27 ± 0.06	12.00 ± 1.89
	DiffAN (2022)	0.71 ± 0.08	0.20 ± 0.04	0.84 ± 0.01	0.29 ± 0.08	0.21 ± 0.08	0.24 ± 0.08	10.40 ± 0.84
	DAGMA (2022)	0.75 ± 0.00	0.25 ± 0.00	0.71 ± 0.00	0.25 ± 0.00	0.50 ± 0.00	0.33 ± 0.00	13.00 ± 0.00
	CaPS (2024)	0.61 ± 0.12	0.20 ± 0.07	0.85 ± 0.03	0.39 ± 0.12	0.31 ± 0.07	0.35 ± 0.08	9.60 ± 1.78
	COSMO (2024)	0.24 ± 0.25	**0.02 ± 0.02**	0.90 ± 0.02	0.76 ± 0.25	0.28 ± 0.11	0.39 ± 0.14	6.70 ± 1.49
	SDCD (2024)	0.79 ± 0.09	0.25 ± 0.05	0.72 ± 0.06	0.21 ± 0.09	0.45 ± 0.16	0.29 ± 0.11	18.20 ± 3.91
	ALIAS (2025)	0.69 ± 0.04	0.11 ± 0.01	0.82 ± 0.01	0.31 ± 0.04	0.35 ± 0.05	0.33 ± 0.04	11.50 ± 0.71
	ABCNAS-RL (2025)	0.42 ± 0.14	0.11 ± 0.05	0.89 ± 0.02	0.59 ± 0.14	0.35 ± 0.08	0.43 ± 0.10	7.30 ± 1.42
	RECAP (Ours)	**0.20 ± 0.12**	**0.02 ± 0.01**	**0.92 ± 0.01**	**0.80 ± 0.12**	0.50 ± 0.00	0.61 ± 0.03	**5.10 ± 0.74**
Cancer	PC (2000)	0.50 ± 0.00	0.33 ± 0.00	0.84 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	4.00 ± 0.00
	GES (2002)	0.50 ± 0.00	0.33 ± 0.00	0.84 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	4.00 ± 0.00
	ICALiNGAM (2006)	0.50 ± 0.00	0.33 ± 0.00	0.84 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	4.00 ± 0.00
	NOTEARS (2018)	0.45 ± 0.26	0.10 ± 0.06	0.83 ± 0.05	0.55 ± 0.26	0.48 ± 0.14	0.48 ± 0.12	4.20 ± 1.32
	GraNDAG (2019)	0.76 ± 0.11	0.92 ± 0.34	0.68 ± 0.09	0.24 ± 0.11	0.38 ± 0.13	0.28 ± 0.11	8.00 ± 2.31
	RL-BIC (2019)	0.80 ± 0.42	**0.05 ± 0.03**	0.77 ± 0.04	0.20 ± 0.42	0.05 ± 0.11	0.08 ± 0.17	3.80 ± 0.42
	GOLEM (2020)	0.78 ± 0.00	0.33 ± 0.00	0.64 ± 0.00	0.22 ± 0.00	0.50 ± 0.00	0.31 ± 0.00	9.00 ± 0.00
	CORL (2021)	0.50 ± 0.00	0.33 ± 0.00	0.84 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	4.00 ± 0.00
	DiffAN (2022)	0.50 ± 0.00	0.17 ± 0.00	0.84 ± 0.00	0.50 ± 0.00	0.25 ± 0.00	0.33 ± 0.00	4.00 ± 0.00
	DAGMA (2022)	0.50 ± 0.00	0.13 ± 0.00	0.80 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	**3.00 ± 0.00**
	CaPS (2024)	0.38 ± 0.40	0.63 ± 0.67	0.76 ± 0.13	0.62 ± 0.40	0.42 ± 0.21	0.37 ± 0.06	6.10 ± 3.35
	COSMO (2024)	0.58 ± 0.18	0.16 ± 0.06	0.80 ± 0.07	0.42 ± 0.18	**0.60 ± 0.27**	0.49 ± 0.20	4.90 ± 1.66
	SDCD (2024)	0.47 ± 0.07	0.09 ± 0.02	0.85 ± 0.02	0.53 ± 0.07	0.50 ± 0.00	0.51 ± 0.03	3.80 ± 0.42
	ALIAS (2025)	0.40 ± 0.16	**0.05 ± 0.02**	0.86 ± 0.02	0.60 ± 0.16	0.32 ± 0.12	0.41 ± 0.11	3.60 ± 0.52
	ABCNAS-RL (2025)	0.50 ± 0.00	0.33 ± 0.00	0.84 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	4.00 ± 0.00
	RECAP (Ours)	**0.33 ± 0.00**	**0.05 ± 0.00**	**0.88 ± 0.00**	**0.67 ± 0.00**	0.50 ± 0.00	**0.57 ± 0.00**	**3.00 ± 0.00**

baseline achieves a significantly lower SHD in any run-paired comparison (Figure 2, bottom row). The edges it recovers are also the most reliable, with the highest precision and lowest FDR on every benchmark (shared with ABCNAS-RL on Sachs), and its F1 is the best or tied-best on Cancer, NetSim, and Sachs. This combination of few false edges and few graph edits is what selective, consensus-supported guidance is designed to produce: an edge enters the recovered graph once it recurs among high-reward candidates, rather than whenever it raises the score of a single proposal. The margin is widest on NetSim, where absolute rewards vary most across inputs and within-input comparison matters most.

Bayesian networks. Asia and Cancer are small graphs on which most methods recover a similar set of true edges and differ mainly in which additional edges they admit; they therefore test whether selective guidance keeps spurious edges out. On both networks RECAP has the best precision, FDR, and FPR in Table 1 together with the lowest SHD, so its graphs are simultaneously the cleanest and

Table 2: Complete NetSim and Sachs results (ten-run mean ± standard deviation). Best values are bold; RECAP is shaded.

Dataset	Methods	FDR ↓	FPR ↓	ACC ↑	Precision ↑	Recall ↑	F1 ↑	SHD ↓
NetSim	PC (2000)	0.61 ± 0.04	0.33 ± 0.04	0.85 ± 0.02	0.39 ± 0.04	0.67 ± 0.08	0.49 ± 0.05	15.40 ± 1.71
	GES (2002)	0.72 ± 0.03	0.47 ± 0.03	0.78 ± 0.02	0.28 ± 0.03	0.65 ± 0.07	0.39 ± 0.04	22.40 ± 1.65
	ICALiNGAM (2006)	0.61 ± 0.08	0.12 ± 0.02	0.88 ± 0.01	0.39 ± 0.08	0.23 ± 0.05	0.29 ± 0.06	12.50 ± 0.97
	NOTEARS (2018)	0.68 ± 0.13	0.11 ± 0.03	0.87 ± 0.02	0.32 ± 0.13	0.16 ± 0.07	0.22 ± 0.09	13.10 ± 1.66
	GraNDAG (2019)	0.67 ± 0.10	0.07 ± 0.03	0.88 ± 0.01	0.33 ± 0.10	0.09 ± 0.00	0.14 ± 0.01	12.30 ± 1.06
	RL-BIC (2019)	0.58 ± 0.03	**0.05 ± 0.00**	0.87 ± 0.00	0.42 ± 0.03	0.26 ± 0.03	0.32 ± 0.03	11.10 ± 0.32
	GOLEM (2020)	0.66 ± 0.07	0.08 ± 0.01	0.88 ± 0.00	0.34 ± 0.07	0.14 ± 0.05	0.19 ± 0.06	12.30 ± 0.48
	CORL (2021)	0.62 ± 0.13	0.12 ± 0.03	0.87 ± 0.02	0.38 ± 0.13	0.23 ± 0.08	0.28 ± 0.09	12.70 ± 1.70
	DiffAN (2022)	0.71 ± 0.06	0.48 ± 0.07	0.79 ± 0.03	0.29 ± 0.06	0.60 ± 0.10	0.39 ± 0.08	20.80 ± 3.39
	DAGMA (2022)	0.80 ± 0.00	**0.05 ± 0.00**	0.84 ± 0.00	0.20 ± 0.00	0.09 ± 0.00	0.13 ± 0.00	13.00 ± 0.00
	CaPS (2024)	0.65 ± 0.04	0.41 ± 0.06	0.82 ± 0.02	0.35 ± 0.04	**0.68 ± 0.05**	0.46 ± 0.05	17.50 ± 2.27
	COSMO (2024)	0.77 ± 0.09	0.13 ± 0.04	0.81 ± 0.04	0.23 ± 0.09	0.28 ± 0.10	0.25 ± 0.09	19.20 ± 4.02
	SDCD (2024)	0.60 ± 0.08	0.12 ± 0.02	0.86 ± 0.03	0.40 ± 0.08	0.64 ± 0.13	0.49 ± 0.10	14.30 ± 2.71
	ALIAS (2025)	0.60 ± 0.10	**0.05 ± 0.01**	0.88 ± 0.01	0.40 ± 0.10	0.25 ± 0.08	0.31 ± 0.09	12.30 ± 1.06
	ABCNAS-RL (2025)	0.69 ± 0.06	0.08 ± 0.02	0.86 ± 0.01	0.31 ± 0.06	0.23 ± 0.05	0.26 ± 0.05	14.10 ± 0.99
	RECAP (Ours)	**0.49 ± 0.05**	0.08 ± 0.02	**0.89 ± 0.01**	**0.51 ± 0.05**	0.63 ± 0.09	**0.56 ± 0.03**	**10.91 ± 0.94**
Sachs	PC (2000)	0.53 ± 0.09	0.23 ± 0.07	0.85 ± 0.03	0.47 ± 0.09	0.58 ± 0.14	0.52 ± 0.11	18.00 ± 3.77
	GES (2002)	0.59 ± 0.03	0.30 ± 0.02	0.83 ± 0.01	0.41 ± 0.03	0.53 ± 0.07	0.46 ± 0.04	20.80 ± 1.32
	ICALiNGAM (2006)	0.45 ± 0.10	0.11 ± 0.03	0.87 ± 0.02	0.55 ± 0.10	0.31 ± 0.06	0.40 ± 0.07	16.00 ± 1.83
	NOTEARS (2018)	0.53 ± 0.07	0.26 ± 0.05	0.85 ± 0.02	0.47 ± 0.07	0.49 ± 0.04	0.48 ± 0.05	18.30 ± 2.58
	GraNDAG (2019)	0.72 ± 0.05	0.16 ± 0.08	0.83 ± 0.02	0.28 ± 0.05	0.13 ± 0.02	0.17 ± 0.01	21.00 ± 2.75
	RL-BIC (2019)	0.58 ± 0.19	0.05 ± 0.02	0.83 ± 0.03	0.42 ± 0.19	0.16 ± 0.06	0.24 ± 0.09	15.20 ± 2.10
	GOLEM (2020)	0.92 ± 0.01	0.30 ± 0.03	0.77 ± 0.01	0.08 ± 0.01	0.06 ± 0.00	0.07 ± 0.00	27.50 ± 0.97
	DiffAN (2022)	0.58 ± 0.12	0.29 ± 0.09	0.83 ± 0.04	0.42 ± 0.12	0.45 ± 0.11	0.43 ± 0.12	20.30 ± 4.99
	DAGMA (2022)	0.40 ± 0.00	**0.02 ± 0.00**	0.85 ± 0.00	0.60 ± 0.00	0.18 ± 0.00	0.27 ± 0.00	14.00 ± 0.00
	CaPS (2024)	0.48 ± 0.08	0.19 ± 0.06	0.86 ± 0.02	0.52 ± 0.08	0.44 ± 0.03	0.47 ± 0.03	16.70 ± 2.00
	CORL (2021)	0.43 ± 0.04	0.14 ± 0.02	0.87 ± 0.01	0.57 ± 0.04	0.41 ± 0.04	0.48 ± 0.03	15.20 ± 0.92
	COSMO (2024)	0.33 ± 0.09	0.03 ± 0.01	0.88 ± 0.01	0.67 ± 0.09	0.35 ± 0.09	0.45 ± 0.09	14.20 ± 1.62
	SDCD (2024)	0.43 ± 0.10	0.09 ± 0.03	0.88 ± 0.03	0.57 ± 0.10	**0.68 ± 0.08**	**0.62 ± 0.08**	14.50 ± 3.63
	ALIAS (2025)	0.49 ± 0.05	0.04 ± 0.00	0.86 ± 0.01	0.51 ± 0.05	0.26 ± 0.04	0.34 ± 0.05	16.80 ± 0.79
	ABCNAS-RL (2025)	**0.25 ± 0.05**	0.06 ± 0.01	0.90 ± 0.01	**0.75 ± 0.05**	0.39 ± 0.03	0.52 ± 0.03	12.50 ± 0.85
	RECAP (Ours)	**0.25 ± 0.07**	0.03 ± 0.01	**0.91 ± 0.01**	**0.75 ± 0.07**	0.54 ± 0.08	**0.62 ± 0.05**	**11.00 ± 0.94**

the closest to the truth. The one method with a higher F1 on Asia, GES, obtains it by admitting more edges at lower precision, and its graphs are statistically indistinguishable from RECAP's in SHD; every other method trails RECAP significantly in SHD or F1. Cancer is especially telling: several baselines converge to an identical graph, and RECAP is the only method that recovers the same true edges with fewer spurious ones, doing so with zero variance across runs. The gain here is entirely a matter of leaving spurious edges out, which is what a consensus teacher that only reinforces recurring edges is designed to do. Appendix G.3 provides the metric-wise comparisons and graph examples; the CSuite results in Appendix G.4 show the same pattern on a further benchmark.

fMRI and protein signaling. NetSim and Sachs contain denser and noisier dependencies, where absolute BIC values vary widely across inputs and a policy trained on raw rewards is easily misled. This is the regime that within-input relative evaluation targets, and it is where RECAP's margin is largest. On NetSim, the baselines fall into two groups in Table 2: conservative methods such as RL-BIC that keep SHD low by predicting few edges and therefore recover few true ones, and permissive methods such as PC and SDCD that recover more true edges at the cost of a majority of false ones. RECAP sits outside this trade-off. It attains the highest F1 and the lowest SHD (0.56 and 10.91) while also having the highest precision and a recall close to the permissive methods, so its improvement comes from finding more correct edges, not from predicting fewer edges. Every run-paired F1 and SHD advantage in Figure 2(c) is positive, and nearly all are significant after Holm correction.

On Sachs, the strongest competitors each excel on a single axis: SDCD matches RECAP's F1 but needs many more graph edits, and ABCNAS-RL matches its precision but recovers fewer true edges. RECAP is the only method that leads on SHD, precision, and F1 at once, which is what a downstream user of the graph needs: recovered edges that can be trusted, in a graph that is close to the truth as a whole. The teacher–policy evolution in Figure 4 shows how recurring edges accumulate into this guidance during training; Section 4.3 quantifies its contribution. Appendix G.5 reports the metric-wise comparisons and inferred graphs.

4.3 Ablation Study

W/O Relative replaces the group-relative update with K-sample REINFORCE, W/O Consensus removes teacher guidance, W/O Encoder uses a standard Transformer, and Fixed Config. omits configuration search. The full model attains the highest F1 and lowest SHD against every variant on every benchmark (Figure 3), and each search signal matters most where its role predicts. Relative evaluation corrects for reward levels that vary across inputs; its removal is most damaging on NetSim, where F1 roughly halves as the policy admits more false edges and finds fewer true ones. Consensus

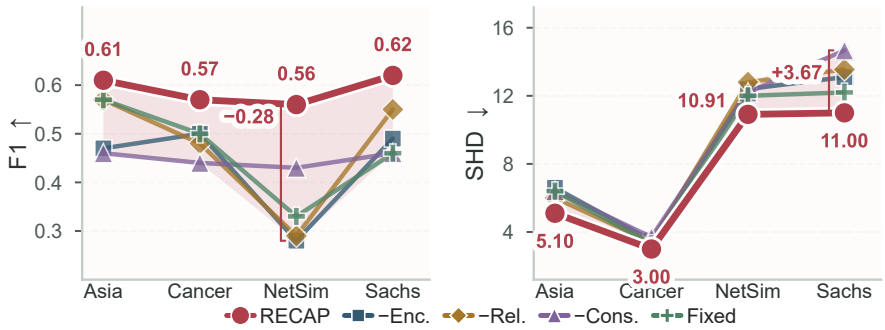


Figure 3: Ten-run mean F1 and SHD. The pale band is RECAP's margin over the weakest removal; the bracket marks the largest gap.

guidance carries recurring edges across updates; its removal is most damaging on Sachs, where edges the teacher would have retained are not rediscovered from current-group rewards, giving the largest SHD increase in the study. Both removals lower F1 and raise SHD on the remaining benchmarks as well, so neither effect is an artifact of one dataset. Replacing the gated encoder or freezing the configuration also weakens recovery, but only the two search signals line up with the conditions they were designed to handle. The full model is also the most stable variant across runs. Appendix G.1 gives the complete analysis.

4.4 Matched-budget NAS Comparison

Table 3: Matched-budget NAS results on NetSim (mean ± standard deviation). Best values are bold; RECAP is shaded.

Metric	NOTEARS NAS	GraNDAG NAS	GOLEM NAS	DiffAN NAS	CaPS NAS	CORL NAS	COSMO NAS	SDCD NAS	ALIAS NAS	ABCNAS RL	RECAP (Ours)
F1 ↑	0.22±0.11	0.44±0.02	0.36±0.13	0.45±0.08	0.51±0.03	0.29±0.06	0.32±0.17	0.46±0.07	0.34±0.10	0.26±0.05	**0.56±0.03**
SHD ↓	12.90±1.60	15.40±1.35	11.60±2.41	18.10±3.14	15.50±1.58	12.50±1.35	15.73±3.23	12.70±1.06	12.00±1.12	14.10±0.99	**10.91±0.94**

Table 3 tests whether configuration search alone explains these results by giving every eligible learned baseline the same outer-search protocol on NetSim (ABCNAS-RL keeps its own NAS). Search helps several baselines, but they still split into recall-oriented methods with higher SHD and conservative methods with lower F1; RECAP alone pairs the highest F1 with the lowest SHD (Appendix G.5.1). Its leading precision shows that the higher F1 is not bought by predicting more edges. The same permissive-versus-conservative split seen in the main comparison therefore survives an equal search budget, so the advantage lies in how the policy learns from scored candidates, not in the configuration budget. Nonlinear 100- and 200-node results and resource profiles are in Appendix G.2.

5 Conclusion

RECAP improves stochastic causal graph search by combining two sources of structural evidence: relative quality within each sampled group and recurring high-reward edges across iterations. Group-relative advantages weight the policy-gradient update, while a transient consensus teacher transfers useful edge patterns into later updates. On Bayesian-network, CSuite, simulated fMRI, and protein-signaling benchmarks, RECAP consistently recovers graphs that are closest to the truth and most selective in the edges they include, with the lowest SHD and highest precision on every main benchmark: a recovered edge can be trusted, and the graph as a whole needs few corrections. The two signals are complementary rather than redundant. Relative evaluation matters most where reward scales vary across inputs, and consensus guidance matters most where useful edges are easily lost; their largest effects fall on different benchmarks, and the full model is also the most reproducible across runs. A matched-budget comparison rules out configuration search as the explanation. The main limitation is computational efficiency: each update scores a group of graphs, and repeated reward evaluation dominates the runtime as the graph grows. Taken together, the recovery and ablation results show that the two search signals improve the recovered graphs across graph families, while the framework retains structural evidence across updates.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, language polishing, and improving the visual presentation of figures and tables. We did not use generative AI tools for other tasks requiring mandatory disclosure.

All AI-assisted content was reviewed and verified by the authors to ensure accuracy and preserve the original scientific meaning. We take full responsibility for the final content of this work.

REFERENCES

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239, 2022.

David Maxwell Chickering. Learning bayesian networks is np-complete. In *Learning from data: Artificial intelligence and statistics V*, pp. 121–130. Springer, 1996.

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

Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15:3921–3962, 2014.

Victor-Alexandru Darvariu, Stephen Hailes, and Mirco Musolesi. Tree search in dag space with model-based reinforcement learning for causal discovery. In *Proceedings A*, volume 481, pp. 20240450. The Royal Society, 2025.

Bao Duong, Hung Le, Biwei Huang, and Thin Nguyen. Reinforcement learning for causal discovery without acyclicity constraints. *Transactions on Machine Learning Research*, 2025.

Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, et al. Deep end-to-end causal inference. *Transactions on Machine Learning Research*, 2024.

Miguel A Hernán and James M Robins. Causal inference. 2014.

P HOYER. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems (NIPS), 2009*, 689, 2009.

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Jiyeon Kang, Songseong Kim, Chanhui Lee, Doyeong Hwang, Joanie Chung, Yunkyung Ko, Sumin Lee, Sungwoong Kim, and Sungbin Lim. Score-informed neural operator for enhancing ordering-based causal discovery. In *Advances in Neural Information Processing Systems*, volume 38, pp. 113109–113151. Curran Associates, Inc., 2025.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

Riccardo Massidda, Francesco Landolfi, Martina Cinquini, and Davide Bacciu. Differentiable causal discovery with smooth acyclic orientations. In *ICML 2023 Workshop on Differentiable Almost Everything: Differentiable Relaxations, Algorithms, Operators, and Simulators*, 2023.

Achille Nazaret, Justin Hong, Elham Azizi, and David Blei. Stable differentiable causal discovery. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 37413–37445, 2024.

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.

Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. In *International Conference on Machine Learning*, pp. 38018–38070. PMLR, 2024.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

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.

Pedro Sanchez, Xiao Liu, Alison Q O'Neil, and Sotirios A Tsaftaris. Diffusion models for causal discovery via topological ordering. *arXiv preprint arXiv:2210.06201*, 2022.

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

Marco Scutari. Learning bayesian networks with thebnlearnrpackage. *Journal of Statistical Software*, 35(3):1, 2010.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *NeuroImage*, 54(2):875–891, 2011.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 499–506, 1995.

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

RS SUTTON. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems (NIPS 1999)*, 12:1057–1063, 2000.

Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, 7 (1):43, 2006.

Xiaoqiang Wang, Yali Du, Shengyu Zhu, Liangjun Ke, Zhitang Chen, Jianye Hao, and Jun Wang. Ordering-based causal discovery with reinforcement learning. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*, pp. 3566–3573. International Joint Conferences on Artificial Intelligence, 2021.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Wen Xiong, Junzhong Ji, and Jinduo Liu. Brainec-llm: Brain effective connectivity estimation by multiscale mixing llm. In *Advances in Neural Information Processing Systems*, volume 38, pp. 123110–123143. Curran Associates, Inc., 2025.

Zhuopeng Xu, Yujie Li, Cheng Liu, and Ning Gui. Ordering-based causal discovery for linear and nonlinear relations. *Advances in Neural Information Processing Systems*, 37:4315–4340, 2024.

Dezhi Yang, Guoxian Yu, Jun Wang, Zhengtian Wu, and Maozu Guo. Reinforcement causal structure learning on order graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10737–10744, 2023.

Wei Yuan, Zixuan Shao, and Shuhui Wang. Global directional priors with local statistical validation for scalable causal discovery. In *Forty-third International Conference on Machine Learning*, 2026.

Jihao Zhai, Junzhong Ji, and Jinduo Liu. Inferring causal protein signaling networks with reinforcement learning via artificial bee colony neural architecture search. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 8996–9004. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/1000. Main Track.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 804–813, 2011.

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*, 2019.

A Notations

The key symbols used in this paper and their corresponding descriptions are summarized in Table 4.

B Theoretical Analysis

This section analyzes the group-relative policy-gradient estimator, graph-policy KL, and consensus guidance under stated assumptions.

B.1 Scope and Assumptions

Fix an input $\mathbf{X}$ and a factorized Bernoulli policy $\pi_\theta(\mathbf{A} \mid \mathbf{X})$ over $M = N(N-1)$ admissible directed edges. Within each group, graphs are sampled independently from the current policy. The derivations use valid score-function identities, rewards conditionally independent of θ given a sampled graph, and the required expectations and derivatives. Teacher-related results stop gradients through the EMA state. Annealing results require $\lambda_t \to 0$, obtained when $0 < \gamma < 1$ or by a final switch-off. These conditions specify the component updates analyzed below.

B.2 Relative Evaluation: Offset Invariance and Direction Alignment

Define $J(\theta; \mathbf{X}) = \mathbb{E}_{\mathbf{A}\sim\pi_\theta}[r(\mathbf{A}, \mathbf{X})]$ and $\mathbf{g}_k = \nabla_\theta \log \pi_\theta(\mathbf{A}_k \mid \mathbf{X})$. The score-function identity gives $\mathbb{E}[r_k \mathbf{g}_k] = \nabla_\theta J$ and $\mathbb{E}[\mathbf{g}_k] = \mathbf{0}$.

Theorem B.1 (Group centering is offset invariant and direction aligned)**.** *For $K \geq 2$ i.i.d. samples from a fixed conditional graph policy, let $\bar{r} = K^{-1}\sum_{k=1}^{K} r_k$ and*

$$\widehat{\mathbf{g}}_{\mathrm{ctr}} = \frac{1}{K}\sum_{k=1}^{K}(r_k - \bar{r})\mathbf{g}_k. \tag{16}$$

Replacing every r_k by $r_k + c(\mathbf{X})$ leaves the realized estimator unchanged, and

$$\mathbb{E}[\widehat{\mathbf{g}}_{\mathrm{ctr}}] = \frac{K-1}{K}\nabla_\theta J(\theta; \mathbf{X}). \tag{17}$$

Proof. The samplewise claim follows because both r_k and $\bar{r}$ increase by $c(\mathbf{X})$. For the expectation, expand $\widehat{\mathbf{g}}_{\mathrm{ctr}} = K^{-1}\sum_k r_k \mathbf{g}_k - K^{-2}\sum_{j,k} r_j \mathbf{g}_k$. Cross terms with $j \neq k$ vanish by independence and $\mathbb{E}[\mathbf{g}_k] = 0$; the K diagonal terms each equal $\nabla_\theta J$. The stated factor follows. □

Corollary B.2 (Exact leave-one-out relation)**.** *Let $b_k = (K-1)^{-1}\sum_{j\neq k} r_j$ and $\widehat{\mathbf{g}}_{\mathrm{LOO}} = K^{-1}\sum_k (r_k - b_k)\mathbf{g}_k$. Then, for every sampled group,*

$$\widehat{\mathbf{g}}_{\mathrm{LOO}} = \frac{K}{K-1}\widehat{\mathbf{g}}_{\mathrm{ctr}}, \qquad \mathbb{E}[\widehat{\mathbf{g}}_{\mathrm{LOO}}] = \nabla_\theta J. \tag{18}$$

Table 4: Notation summary used in the main text.

Symbol	Meaning	Shape / Domain
Variable-wise context encoder		
$\mathbf{X}_b$	empirical data block for instance b	$\mathbb{R}^{n_b \times N}$
$\mathbf{U}_b$	variable-wise token features for encoder input	$\mathbb{R}^{N \times d_{\text{in}}}$
$\mathbf{U}$	mini-batch of variable-wise token features	$\mathbb{R}^{B \times N \times d_{\text{in}}}$
n_b	number of observations in data block b	$\mathbb{N}$
B	batch size	$\mathbb{N}$
N	number of variables or nodes	$\mathbb{N}$
d_{in}	input feature dimension per variable	$\mathbb{N}$
d	latent embedding dimension	$\mathbb{N}$
$\mathbf{E}_{\text{id}}$	learnable node-identity embedding	$\mathbb{R}^{N \times d}$
$\mathbf{Z}$	model output representations, used to parameterize the structure policy.	$\mathbb{R}^{B \times N \times d}$
L	number of stacked variable-wise encoder blocks	$\mathbb{N}$
$\mathrm{LN}(\cdot)$	layer normalization operator	operator
H	number of heads	$\mathbb{N}$
d_h	per-head dimension	d/H
$\sigma(\cdot)$	sigmoid function	function
$f(\cdot; \mathbf{E}_{\text{id}})$	variable-wise encoder mapping from $\mathbf{U}$ to $\mathbf{Z}^{(L)}$	mapping
Graph sampler		
$\mathbf{s}_{ij}$	edge logit for $i \to j$	scalar or batch-wise
$\tanh(\cdot)$	hyperbolic tangent	function
$\mathbf{A}$	sampled adjacency matrix with zero diagonal	$\{0,1\}^{N \times N}$
$\mathrm{Bernoulli}(\cdot)$	Bernoulli distribution	distribution
Self-distillation		
t	iteration index	$\mathbb{N}$
$\mathbf{S}_{b,k}$	edge-logit matrix for block b, candidate k	$\mathbb{R}^{N \times N}$
$\mathbf{P}_{b,k}$	edge-probability matrix derived from candidate logits	$[0,1]^{N \times N}$
$\mathbf{I}_N$	identity matrix	$\mathbb{R}^{N \times N}$
$\mathbf{P}_t^{\text{teach}}$	teacher edge-probability matrix at iteration t	$(0,1)^{N \times N}$
$\mathcal{I}_t$	index set of selected top candidates in the mini-batch	set
M_{top}	number of top candidates used to form consensus	$\mathbb{N}$
$\mathbf{P}_t^{\text{top}}$	hard-edge frequency or mean soft edge probability among selected candidates	$[0,1]^{N \times N}$
β	EMA coefficient controlling memory length	$[0,1)$
τ	threshold for selecting consensus edges	$(0,1)$
$\mathrm{BCE}(\cdot,\cdot)$	Bernoulli cross-entropy with logits	function
κ_t	normalizer counting constrained entries	$\mathbb{N}$
$p_{b,ij}$	predicted Bernoulli parameter for edge (i,j)	$(0,1)$
$q_{t,ij}$	teacher Bernoulli parameter for edge (i,j)	$(0,1)$
$\mathrm{KL}(\cdot\|\cdot)$	Kullback–Leibler divergence	function
$H(\cdot)$	entropy	function
T_w	warm-up iterations before activating distillation	$\mathbb{N}$
Relative evaluation and policy gradients		
K	number of sampled graphs per instance	$\mathbb{N}$
$\pi_\theta(\mathbf{A} \mid \mathbf{X}_b)$	policy distribution over adjacency matrices	distribution
θ	policy parameters	parameters
$r_{b,k}$	scalar reward for $(\mathbf{X}_b, \mathbf{A}_{b,k})$	$\mathbb{R}$
$\pi_{\theta_{\text{ref}}}$	reference policy	distribution
θ_{ref}	reference policy parameters	parameters
$\mathbb{H}(\pi_\theta(\cdot \mid \mathbf{X}_b))$	averaged Bernoulli entropy induced by edge probabilities	$\mathbb{R}_{\geq 0}$
$\mathcal{L}_{\text{Ent}}(\theta)$	entropy bonus term	scalar
λ_{Ent}	coefficient for entropy regularization	$\mathbb{R}_{\geq 0}$
$\mathcal{L}(\theta)$	full RECAP training objective	scalar

Proof. Since $r_k - b_k = K(r_k - \bar{r})/(K-1)$, the samplewise relation is immediate; unbiasedness follows from Theorem B.1. □

Proposition B.3 (Properties of standardized advantages). *Let* $d_k = r_k - \bar{r}$, $s = ((K-1)^{-1}\sum_k d_k^2)^{1/2}$, *and* $a_k = d_k/(s+\epsilon_{\text{adv}})$ *with* $\epsilon_{\text{adv}} \geq 0$ *and* $s + \epsilon_{\text{adv}} > 0$. *Then* $\sum_k a_k = 0$ *and the vector* $\mathbf{a}$ *is invariant to a common additive reward shift. If* $\epsilon_{\text{adv}} = 0$ *and* $s > 0$, *it is also invariant to positive reward rescaling and* $|a_k| \leq (K-1)/\sqrt{K}$.

Proof. Centering gives $\sum_k d_k = 0$, proving the first two claims. With $\epsilon_{\text{adv}} = 0$, positive rescaling multiplies both numerator and denominator equally. For any k, $\sum_{j\neq k} d_j = -d_k$; Cauchy–Schwarz gives $d_k^2 \leq (K-1)\sum_{j\neq k} d_j^2$, hence $Kd_k^2 \leq (K-1)\sum_j d_j^2$. With $s^2 = (K-1)^{-1}\sum_j d_j^2$, this yields $|d_k|/s \leq (K-1)/\sqrt{K}$. □

The centering and leave-one-out results establish expected direction alignment for their respective estimators. Standardization introduces a denominator computed from the same rewards, so Proposition B.3 instead establishes its shift and scale properties.

B.3 Conservative Policy Updates in Graph Space

Let p_e and p'_e denote current and reference Bernoulli probabilities for edge e, and define

$$\bar{D}_{\text{edge}} = \frac{1}{M}\sum_{e=1}^{M} D_{\text{KL}}(\text{Bern}(p_e)\|\text{Bern}(p'_e)). \tag{19}$$

Lemma B.4 (Joint graph-policy KL decomposition). *For factorized Bernoulli graph policies* $\pi = \prod_e \text{Bern}(p_e)$ *and* $\pi' = \prod_e \text{Bern}(p'_e)$,

$$D_{\text{KL}}(\pi\|\pi') = \sum_{e=1}^{M} D_{\text{KL}}(\text{Bern}(p_e)\|\text{Bern}(p'_e)) = M\bar{D}_{\text{edge}}. \tag{20}$$

Proof. Insert the product densities into the log likelihood ratio, expand the logarithm into a sum, and marginalize every term under π. □

Theorem B.5 (Reward drift from average edge KL). *If* $r(\mathbf{A}, \mathbf{X}) \in [r_{\min}, r_{\max}]$ *and* $\Delta_r = r_{\max} - r_{\min}$, *then*

$$|\mathbb{E}_\pi r - \mathbb{E}_{\pi'} r| \leq \Delta_r\sqrt{\frac{M\bar{D}_{\text{edge}}}{2}}. \tag{21}$$

Proof. For $f = (r - r_{\min})/\Delta_r \in [0,1]$, the variational characterization of total variation yields $|\mathbb{E}_\pi f - \mathbb{E}_{\pi'} f| \leq D_{\text{TV}}(\pi, \pi')$. Pinsker's inequality and Lemma B.4 give $D_{\text{TV}} \leq \sqrt{D_{\text{KL}}(\pi\|\pi')/2} = \sqrt{M\bar{D}_{\text{edge}}/2}$. □

Corollary B.6 (Graph-event drift). *For any graph event* $\mathcal{E}$, *including acyclicity,*

$$|\pi(\mathcal{E}) - \pi'(\mathcal{E})| \leq \sqrt{\frac{M\bar{D}_{\text{edge}}}{2}}. \tag{22}$$

Both bounds use the achieved joint KL. Expressing the regularizer as an edge average introduces the explicit graph-size factor $M = N(N-1)$.

B.4 Consensus Augmentation as Directed Structural Memory

Proposition B.7 (Reverse Bernoulli KL form). *For a student logit* s, $p = \sigma(s)$, *and fixed teacher target* $q \in (0,1)$,

$$\text{BCE}(s, q) = D_{\text{KL}}(\text{Bern}(q)\|\text{Bern}(p)) + H(\text{Bern}(q)). \tag{23}$$

Thus the masked consensus loss is an average reverse Bernoulli KL plus a teacher-dependent constant.

Proof. Expand both Bernoulli terms: the $q\log q$ and $(1-q)\log(1-q)$ contributions cancel, leaving $-q\log p - (1-q)\log(1-p)$. □

Theorem B.8 (Direction and marginal optimum of consensus guidance)**.** *For each masked edge,*

$$\frac{\partial \mathrm{BCE}(s,q)}{\partial s} = p - q, \qquad \frac{\partial^2 \mathrm{BCE}(s,q)}{\partial s^2} = p(1-p) > 0. \tag{24}$$

Hence the unique finite marginal optimum is $p = q$, and a negative-gradient step moves the student probability toward the teacher. An unmasked edge receives no direct teacher gradient.

Proof. Differentiate the logistic cross-entropy and use $\partial p/\partial s = p(1-p)$. Strict positivity for finite s gives strict convexity and the unique stationary point $p = q$; multiplication by a zero mask removes the derivative. □

Proposition B.9 (EMA teacher as discounted candidate memory)**.** *Index teacher-refresh events by t. For the unclamped EMA state $\widetilde{\mathbf{P}}_t = \beta \widetilde{\mathbf{P}}_{t-1} + (1-\beta)\mathbf{P}_t^{\mathrm{top}}$, $0 \le \beta < 1$, we have*

$$\widetilde{\mathbf{P}}_t = \beta^t \widetilde{\mathbf{P}}_0 + (1-\beta) \sum_{u=1}^{t} \beta^{t-u} \mathbf{P}_u^{\mathrm{top}}. \tag{25}$$

If a fixed policy and selection rule make $\{\mathbf{P}_t^{\mathrm{top}}\}$ identically distributed with finite mean $\boldsymbol{\mu}_{\mathrm{top}}$, then $\mathbb{E}[\widetilde{\mathbf{P}}_t] \to \boldsymbol{\mu}_{\mathrm{top}}$ for the unclamped state.

Proof. Repeated substitution gives the expansion. Taking expectations yields $\beta^t \widetilde{\mathbf{P}}_0 + (1-\beta^t)\boldsymbol{\mu}_{\mathrm{top}}$, whose first term vanishes. □

This expansion indexes refresh events and applies to the unclamped EMA. Clamping changes the limiting expectation; with soft inputs, $\mathbf{P}_t^{\mathrm{top}}$ averages probabilities rather than sampled-edge frequencies. Because selection favors high-reward candidates, the resulting teacher provides a search signal rather than a calibrated causal posterior.

B.5 Transient Rather Than Permanent Regularization

Annealing can be analyzed by fixing the base objective L, including its reward and reference policy, while allowing the stopped-gradient consensus target R_t to vary. Define $J_t = L + \lambda_t R_t$, where $R_t \ge 0$ and $\lambda_t \to 0$. During actual training, rewards and the reference policy can also change; the results below characterize the fixed-L case.

Theorem B.10 (Accumulation consistency of exact global minimizers)**.** *Assume L is lower semicontinuous and has a global minimizer $\theta^\star$ satisfying $\sup_t R_t(\theta^\star) < \infty$. If $\theta_t \in \arg\min J_t$ and a subsequence converges to $\bar{\theta}$, then $\bar{\theta} \in \arg\min L$.*

Proof. Optimality and nonnegativity give $L(\theta_t) \le L(\theta^\star) + \lambda_t R_t(\theta^\star)$, so $L(\theta_t) \to \inf L$. Lower semicontinuity along the convergent subsequence gives $L(\bar{\theta}) \le \inf L$, hence equality. □

Uniformly bounded student logits at a base minimizer suffice for the assumed bound on $R_t(\theta^\star)$.

Proposition B.11 (First-order neutrality)**.** *Suppose L and R_t are differentiable, $\theta_t \to \bar{\theta}$, $\|\nabla J_t(\theta_t)\| \to 0$, $\lambda_t \|\nabla R_t(\theta_t)\| \to 0$, and ∇L is continuous at $\bar{\theta}$. Then $\nabla L(\bar{\theta}) = 0$.*

Proof. By the triangle inequality, $\|\nabla L(\theta_t)\| \le \|\nabla J_t(\theta_t)\| + \lambda_t \|\nabla R_t(\theta_t)\| \to 0$; continuity completes the proof. □

As its coefficient vanishes, consensus has no direct effect on limiting stationary points under these conditions, although its earlier updates can shape the search trajectory.

B.6 End-to-End Permutation Equivariance

Let $\mathbf{P}$ permute variables and simultaneously permute their identity embeddings. Shared tokenwise maps commute with $\mathbf{P}$. For self-attention, shared projections give $\mathbf{Q}' = \mathbf{PQ}$, $\mathbf{K}' = \mathbf{PK}$, and $\mathbf{V}' = \mathbf{PV}$; row-wise softmax satisfies $\mathrm{softmax}(\mathbf{PSP}^\top) = \mathbf{P}\mathrm{softmax}(\mathbf{S})\mathbf{P}^\top$. Shared gated heads, residual connections, and positionwise feed-forward maps preserve this property.

Theorem B.12 (End-to-end graph-policy equivariance)**.** *Assume no variable-axis positional encoding and a shared pairwise edge scorer $S_{ij} = h(\mathbf{z}_i, \mathbf{z}_j)$ with diagonal entries excluded. Then*

$$f(\mathbf{PU}; \mathbf{PE}_{\mathrm{id}}) = \mathbf{P}f(\mathbf{U}; \mathbf{E}_{\mathrm{id}}), \qquad \mathbf{S}(\mathbf{PZ}) = \mathbf{PS}(\mathbf{Z})\mathbf{P}^\top, \tag{26}$$

and the factorized Bernoulli policy satisfies

$$\pi_\theta(\mathbf{PAP}^\top \mid \mathbf{PU}, \mathbf{PE}_{\mathrm{id}}) = \pi_\theta(\mathbf{A} \mid \mathbf{U}, \mathbf{E}_{\mathrm{id}}). \tag{27}$$

Proof. The preceding commutation identities compose across encoder blocks, proving encoder equivariance. Sharing h over ordered pairs maps each relabeled pair to the corresponding entry of $\mathbf{PSP}^\top$. Applying the sigmoid preserves this permutation, and reindexing the product of Bernoulli edge probabilities leaves its value unchanged. □

For an entire training update to share this equivariance, the reward, reference state, teacher state, masks, and tie-breaking rule must also transform under the same relabeling.

C Training Algorithms

Algorithms 1 and 2 specify the complete training loop, from group-relative policy-gradient updates to teacher refresh and consensus distillation.

Algorithm 1 RECAP: Relative Evaluation with Consensus Augmented Policy Search

Input dataset $\mathcal{D}$, batch size B, group size K, total iterations T
Hyperparameters ϵ_{adv}, standardization switch, λ_{KL}, λ_{Ent}, T_{ref}, T_λ, T_{teach}, T_w, λ_{SD}^0, γ
Initialize policy parameters θ, reference parameters $\theta_{\mathrm{ref}} \leftarrow \theta$
Initialize penalty multipliers λ_1, λ_2 and teacher prior $\mathbf{P}^{\mathrm{teach}} \leftarrow \frac{1}{2}\mathbf{1}\mathbf{1}^\top$
for $t = 1$ **to** T **do**
  Sample a mini-batch of data blocks $\{\mathbf{X}_b\}_{b=1}^B \sim \mathcal{D}$
  Construct variable-wise token features $\{\mathbf{U}_b\}_{b=1}^B$ from $\{\mathbf{X}_b\}_{b=1}^B$
  Compute contextual representations $\mathbf{Z}^{(L)} \leftarrow f(\mathbf{U}; \mathbf{E}_{\mathrm{id}})$; use the NAS-selected pairwise scorer
  for each b **do**
    Sample K candidate graphs $\{\mathbf{A}_{b,k}\}_{k=1}^K \sim \pi_\theta(\cdot \mid \mathbf{X}_b)$ with $A_{ii} = 0$
    Evaluate rewards $r_{b,k} \leftarrow -\mathrm{Score}(\mathbf{A}_{b,k}, \mathbf{X}_b; \lambda_1, \lambda_2)$
    Compute $\mu_b \leftarrow \frac{1}{K}\sum_{k=1}^K r_{b,k}$ and $s_b \leftarrow \sqrt{\frac{1}{K-1}\sum_{k=1}^K (r_{b,k} - \mu_b)^2}$
    Set advantages $a_{b,k} \leftarrow \frac{r_{b,k} - \mu_b}{s_b + \epsilon_{\mathrm{adv}}}$ if standardizing; otherwise $a_{b,k} \leftarrow r_{b,k} - \mu_b$
  Compute
    $L_{\mathrm{GR}}(\theta) \leftarrow -\frac{1}{BK}\sum_{b=1}^B \sum_{k=1}^K a_{b,k} \log \pi_\theta(\mathbf{A}_{b,k} \mid \mathbf{X}_b)$
    $\mathcal{L}_{\mathrm{KL}}(\theta) \leftarrow \frac{\lambda_{\mathrm{KL}}}{B}\sum_{b=1}^B \frac{1}{N(N-1)}\sum_{i \neq j} \mathrm{KL}\Big(\mathrm{Bern}(p_{b,ij}^\theta) \,\|\, \mathrm{Bern}(p_{b,ij}^{\theta_{\mathrm{ref}}})\Big)$.
    $L_{\mathrm{Ent}}(\theta) \leftarrow -\frac{\lambda_{\mathrm{Ent}}}{B}\sum_{b=1}^B H(\pi_\theta(\cdot \mid \mathbf{X}_b))$
  Call Algorithm 2 to update $\mathbf{P}^{\mathrm{teach}}$ and obtain $L_{\mathrm{SD}}(\theta)$
  Set $\lambda_{\mathrm{SD}}(t) \leftarrow 0$ if $t < T_w$, and $\lambda_{\mathrm{SD}}(t) \leftarrow \lambda_{\mathrm{SD}}^0 \gamma^{t - T_w}$ otherwise
  Update θ by minimizing $L_{\mathrm{GR}}(\theta) + L_{\mathrm{KL}}(\theta) + L_{\mathrm{Ent}}(\theta) + \lambda_{\mathrm{SD}}(t) L_{\mathrm{SD}}(\theta)$
  if $t \bmod T_\lambda = 0$ **then** update λ_1, λ_2 and optionally prune the current best graph **end if**
  if $t \bmod T_{\mathrm{ref}} = 0$ **then** $\theta_{\mathrm{ref}} \leftarrow \theta$ **end if**
end for
Return a final DAG estimate selected by reward and pruning

Algorithm 2 Self-Distillation Algorithm

Input step t, sampled graphs $\{\mathbf{A}_{b,k}\}_{b,k}$, rewards $\{r_{b,k}\}_{b,k}$, logits $\{\mathbf{S}_b\}_b$, current $\mathbf{P}^{\text{teach}}$
if $t \bmod T_{\text{teach}} = 0$ **then**
  Select top candidates $\mathcal{I}_t \leftarrow \text{TopM}\big(\{r_{b,k}\}_{b,k}, \min(M_{\text{top}}, BK)\big)$
  Compute consensus $\mathbf{P}^{\text{top}} \leftarrow \frac{1}{|\mathcal{I}_t|}\sum_{(b,k)\in\mathcal{I}_t} \mathbf{Q}_{b,k}$, $\mathbf{Q}_{b,k} \in \{\mathbf{A}_{b,k}, \sigma(\mathbf{S}_{b,k})\}$
  Update teacher $\mathbf{P}^{\text{teach}} \leftarrow \beta\mathbf{P}^{\text{teach}} + (1-\beta)\mathbf{P}^{\text{top}}$
  Clamp $\mathbf{P}^{\text{teach}} \leftarrow \text{clip}(\mathbf{P}^{\text{teach}}, p_{\min}, p_{\max})$
end if
Mask $\mathbf{W} \leftarrow (\mathbf{J}_N - \mathbf{I}_N) \odot \mathbb{I}\big[\mathbf{P}^{\text{teach}} \geq \tau\big]$
$\kappa \leftarrow \max\Big(1, \sum_{i\neq j} W_{ij}\Big)$
if $t < T_w$ **then** $L_{\text{SD}}(\theta) \leftarrow 0$ **else**
  $L_{\text{SD}}(\theta) \leftarrow \frac{1}{BK\kappa}\sum_{b=1}^{B}\sum_{k=1}^{K}\sum_{i\neq j} W_{ij} \cdot \text{BCE}\big(S_{b,k,ij}, P_{ij}^{\text{teach}}\big)$
end if Return updated $\mathbf{P}^{\text{teach}}$ and $L_{\text{SD}}(\theta)$

D Computational Complexity

Per-iteration cost comprises variable encoding, edge-logit computation and graph sampling, and reward evaluation.

D.1 Per-iteration Time Complexity

The L-block encoder costs

$$\mathcal{T}_{\text{enc}} = \mathcal{O}\big(BL\left(N^2 d + Nd^2\right)\big). \tag{28}$$

Representation Learning and Graph Sampling. The independent Bernoulli sampler used in the main experiments computes edge logits once per input block. Its decoding cost is

$$\mathcal{T}_{\text{dec}} = \mathcal{O}(BN^2 d + BKN^2), \tag{29}$$

where the first term computes edge logits and the second term samples K adjacency matrices and evaluates their log-probabilities.

Reward Computation. Given a sampled graph $\mathbf{A}$, the BIC component fits N regressions on n_{sample} observations. Let p_i denote the number of selected parents for node i. Forming the normal equations costs $\mathcal{O}(n_{\text{sample}}p_i^2)$ and solving costs $\mathcal{O}(p_i^3)$. Summed over all nodes, the regression term is

$$\mathcal{T}_{\text{reg}} = \mathcal{O}\left(\sum_{i=1}^{N}\left(n_{\text{sample}}p_i^2 + p_i^3\right)\right). \tag{30}$$

The acyclicity penalty evaluates a matrix exponential on an $N \times N$ matrix, which costs $\mathcal{O}\big(N^3\big)$. Therefore, the reward cost over a mini-batch and K samples is

$$\mathcal{T}_{\text{rew}} = \mathcal{O}\left(BK\left(\sum_{i=1}^{N}\left(n_{\text{sample}}p_i^2 + p_i^3\right) + N^3\right)\right). \tag{31}$$

If $p_i = \Theta(N)$ for all i, then

$$\mathcal{T}_{\text{rew}} = \mathcal{O}\big(BK\left(n_{\text{sample}}N^3 + N^4\right)\big). \tag{32}$$

If the learned graphs are sparse, then p_i is small and the regression term decreases accordingly.

Overall Complexity. The per-iteration total combines encoding, decoding, and reward evaluation. Graph size determines pairwise operations and the acyclicity calculation, while the observation count enters the reward regressions.

D.2 Memory Complexity

Edge-level tensors for the K sampled graphs dominate training memory. Storing their adjacency matrices, logits, and edge probabilities scales as

$$\mathcal{M} = \mathcal{O}\big(BKN^2\big)\,. \tag{33}$$

Additional buffers such as teacher statistics also scale with N^2 and do not change the leading order.

E Experimental Details

E.1 Baselines

The comparison spans constraint- and score-based PC (Spirtes et al., 2000) and GES (Chickering, 2002); functional ICALiNGAM (Shimizu et al., 2006); and continuous NOTEARS (Zheng et al., 2018), GraNDAG (Lachapelle et al., 2019), GOLEM (Ng et al., 2020), SDCD (Nazaret et al., 2024), COSMO (Massidda et al., 2023), and DAGMA (Bello et al., 2022). Ordering-based comparators are DiffAN (Sanchez et al., 2022) and CaPS (Xu et al., 2024); RL comparators are CORL (Wang et al., 2021), ALIAS (Duong et al., 2025), ABCNAS-RL (Zhai et al., 2025), and RL-BIC (Zhu et al., 2019). We use official or open-source implementations with documented settings. DAGMA and RL-BIC in Tables 1 and 2 use fixed benchmark inputs across method-specific seeds; ten-run comparisons among the remaining methods are paired by run index. Deterministic methods yield identical outputs on fixed inputs.

E.2 Datasets

Bayesian networks. Asia, Cancer, and Sachs are each evaluated on 2,000 observations drawn from the corresponding Bayesian network with bnlearn. Asia and Sachs use the networks provided by the package, and Cancer uses the network released with the bnlearn repository Scutari (2010). The observations are the input, and the network adjacency is the reference graph. Sachs is the protein-signaling network of Sachs et al. (2005).

Additional benchmarks. NetSim uses simulated fMRI Sim#11 Smith et al. (2011). CSuite (Geffner et al., 2024) uses its predefined splits. The SynTReN gene network Van den Bulcke et al. (2006) has 15 nodes and 14 directed edges; its generation disables background nodes and otherwise uses default settings.

E.3 Implementation Details

E.3.1 BIC Reward and acyclicity constraint

For each mini-batch input $\mathbf{X}_b$, the policy samples K candidate adjacencies $\{\mathbf{A}_{b,k}\}_{k=1}^{K}$. A scalar reward favors low BIC and penalizes cyclicity. We first compute

$$\mathrm{Score}\big[\mathbf{A}\big] \;=\; \widetilde{\mathrm{BIC}}\big[\mathbf{A}\big] \;+\; \lambda_1\,\mathbb{I}\big[\mathcal{H}\big(\mathbf{A}\big) > \varepsilon\big] \;+\; \lambda_2\,\mathcal{H}\big(\mathbf{A}\big), \tag{34}$$

and define the reward as its negation, so that larger rewards correspond to better structures.

BIC Term. Given $\mathbf{A}$, each variable is regressed on its selected parents, yielding residual sum of squares $\mathrm{RSS}_i\big[\mathbf{A}\big]$. BIC aggregates the node-wise residual terms and penalizes the selected parent sets:

$$\mathrm{BIC}[\mathbf{A}] = \sum_{j=1}^{N}\left[n_{\mathrm{sample}}\log\left(\frac{\mathrm{RSS}_j[\mathbf{A}]}{n_{\mathrm{sample}}}+\epsilon\right) + |\mathrm{Pa}_j(\mathbf{A})|\log n_{\mathrm{sample}}\right]. \tag{35}$$

In the implementation, $\widetilde{\mathrm{BIC}}$ is obtained by linearly rescaling BIC using precomputed bounds, which stabilizes the reward scale across datasets and training stages.

Acyclicity Term. We adopt a smooth acyclicity measure

$$\mathcal{H}\big(\mathbf{A}\big) \;=\; \mathrm{tr}\Big[\exp\big[\mathbf{A}\big]\Big] - N, \tag{36}$$

and apply a hard barrier with threshold ε together with a soft penalty weighted by λ_2.

Table 5: Architecture and optimization search space for RECAP benchmark experiments.

Parameter	Search space
Encoder width and heads	choice over compatible (d, H) pairs
Encoder depth	integers $[2, 9]$
Sampler hidden-width ratio	$\{0.25, 0.5, 1.0\}$
Pairwise scorer	additive, bilinear, neural tensor
Mini-batch size	$\{32, 64, 128\}$
Initial learning rate	log-uniform $[10^{-4}, 3 \cdot 10^{-3}]$
Training iterations	integers $[2000, 8000]$
Candidate group size K	$\{4, 16, 32\}$
Advantage stabilizer ϵ_{adv}	log-uniform $[0.01, 0.2]$
Reference-policy KL weight	log-uniform $[0.001, 0.9]$
Entropy weight	log-uniform $[0.001, 0.9]$
Standardize group advantages	$\{\mathrm{true}, \mathrm{false}\}$
Reference refresh period	integers $[1, 500]$
Gradient clipping	log-uniform $[0.1, 2.0]$
Initial consensus weight	log-uniform $[0.01, 0.2]$
Teacher top-set size	$\{16, 32, 64, 128, 256, 512\}$
Consensus threshold	log-uniform $[0.6, 0.9]$
Teacher refresh period	integers $[1, 500]$

E.3.2 Configuration selection

An NNI outer search selects encoder and pairwise-scorer families, group size, and optimization hyperparameters. Every trial trains the relative-evaluation and consensus-augmented objective, so selection varies policy capacity and training settings while keeping the two search mechanisms. Table 5 lists the benchmark search space. When standardization is disabled, the advantage in Eq. (5) becomes $r_{b,k} - \mu_b$; Appendix B.2 analyzes this centered estimator separately from the standardized one.

E.3.3 Graph sampler instantiations

Additive, bilinear, and neural-tensor pairwise scorers produce off-diagonal Bernoulli logits. Equation (3) illustrates the additive case; factorization applies to each scorer, and permutation equivariance requires shared pairwise weights and consistent identity-embedding reindexing.

E.3.4 Compute Resources

Experiments used one NVIDIA GeForce RTX 5090 GPU and 128 GB system memory. Appendix D details the costs of sampling and reward evaluation.

E.4 Evaluation Metrics

Seven metrics describe edge recovery and graph structure: false discovery rate (FDR), false positive rate (FPR), structural Hamming distance (SHD), precision, recall, F1, and accuracy. TP, FP, TN, and FN count true positives, false positives, true negatives, and false negatives over ordered pairs (i, j). A reversed edge contributes one FP and one FN to the edge metrics but one reversal to SHD. Precision, recall, F1, and accuracy are higher-is-better; FDR, FPR, and SHD are lower-is-better. Rate-based metrics lie in $[0, 1]$. SHD counts the minimum edge additions, deletions, and reversals needed to

transform the predicted graph into the ground truth, with each reversal counted once.

$$\begin{aligned}
\mathrm{FDR} &= \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}}, \\
\mathrm{FPR} &= \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}, \\
\mathrm{Precision} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \\
\mathrm{Recall} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\
\mathrm{F1} &= \frac{2\times\mathrm{TP}}{2\times\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \\
\mathrm{ACC} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}},
\end{aligned} \tag{37}$$

Statistical comparisons. The main run-paired two-sided comparisons use Holm adjustment across baselines within each dataset and metric, with exact sign-flip tests for constant differences.

F Training Dynamics of Consensus Guidance

F.1 Stage-wise Teacher–Policy Evolution

The Sachs training trajectory shows how cross-update edge consensus becomes guidance for the graph policy. In Figure 4, the teacher concentrates on recurring high-reward edges as the policy aligns with it; their measured ℓ_1 distance decreases from 0.099 to 0.045 to 0.021. This transfer has a corresponding structural-recovery effect: removing consensus lowers Sachs F1 from 0.62 to 0.46 and raises SHD from 11.00 to 14.67 (Table 6). The trajectory and ablation together connect the proposed structural memory to improved final graph recovery.

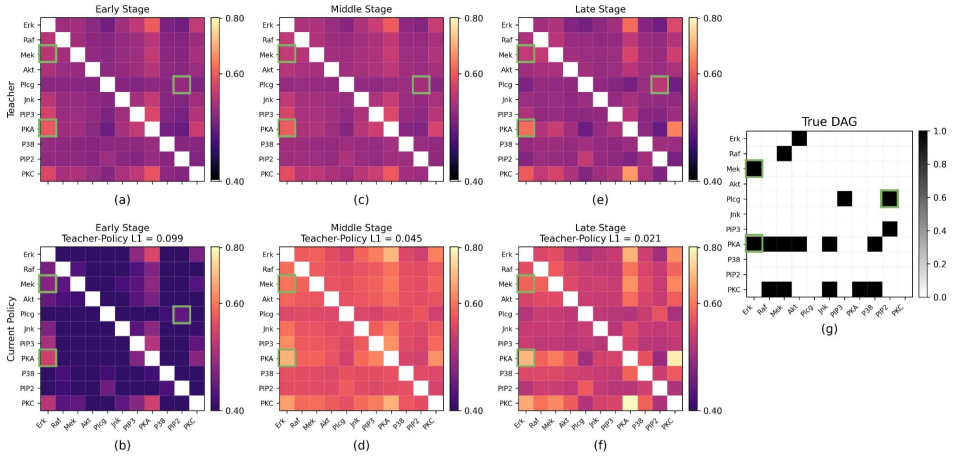


Figure 4: Teacher–policy evolution on Sachs. Teacher panels (a), (c), (e) and policy panels (b), (d), (f) show early, middle, and late stages; (g) shows the reference DAG. Green boxes mark true-edge locations. Teacher–policy ℓ_1 distances are 0.099, 0.045, and 0.021.

F.2 Stage-wise Loss-Term Decomposition

Figure 5 tracks how the objective's components change alongside the teacher–policy trajectories. Let $m \in \{\mathrm{GR}, \mathrm{KL}, \mathrm{Ent}, \mathrm{SD}\}$ index the four contributions to Eq. (15), with $\mathcal{L}_{\mathrm{SD}}^{\mathrm{weighted}}(t) =$

$\lambda_{\mathrm{SD}}(t)\mathcal{L}_{\mathrm{SD}}(t)$. Each component's share of the sum of absolute weighted magnitudes is

$$\rho_m(t) = \frac{|\mathcal{L}_m^{\mathrm{weighted}}(t)|}{|\mathcal{L}_{\mathrm{GR}}(t)| + |\mathcal{L}_{\mathrm{KL}}(t)| + |\mathcal{L}_{\mathrm{Ent}}(t)| + |\lambda_{\mathrm{SD}}(t)\mathcal{L}_{\mathrm{SD}}(t)| + \epsilon}, \tag{38}$$

The other three terms already include their weights in Eq. (15), and $\epsilon > 0$ prevents division by zero. Before consensus activation, entropy and reference-policy KL have the largest measured shares in Figure 5. After warm-up, the weighted consensus term becomes visible alongside a stronger group-relative policy-gradient contribution: reward comparisons continue while the teacher supplies additional edge guidance. Its relative magnitude later recedes as $\lambda_{\mathrm{SD}}(t)$ decays, whereas the reward-driven and entropy terms remain active. These are shares of loss magnitudes rather than gradient contributions, so the plot describes when each term is present and how its measured scale changes; the recovery effect of removing a term is tested separately by Table 6.

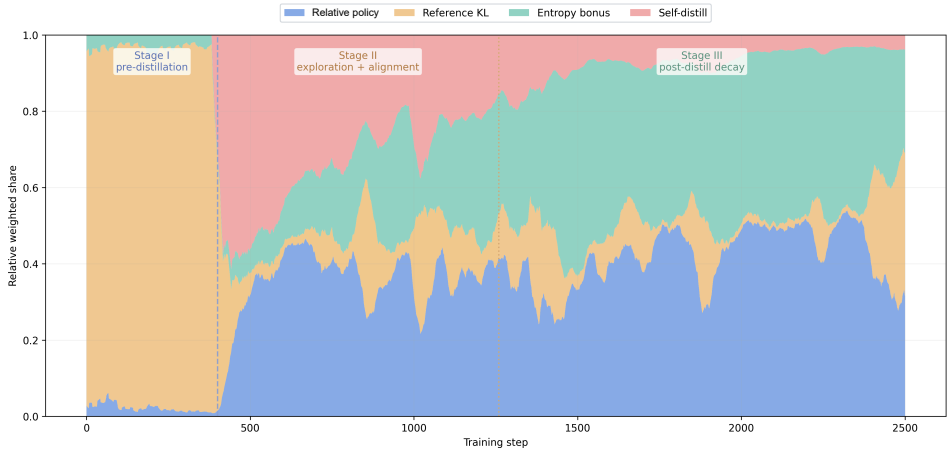


Figure 5: Relative weighted magnitudes of the group-relative, KL, entropy, and consensus loss terms during Sachs training. Consensus appears after warm-up and declines as its weight decays.

G Supplementary Experimental Results

G.1 Full Component Ablation Results

Table 6 records the full metric profile for the variants in Figure 3, with ten-run means and sample standard deviations. This section expands the summary in Section 4.3.

Within-input relative evaluation. Relative evaluation addresses reward levels that vary across inputs, so its removal should hurt most where that variation is largest. NetSim is that case: replacing the relative update with REINFORCE drops F1 from 0.56 to 0.29 and raises SHD from 10.91 to 12.80, the largest F1 change in the study (Figure 3, bracket), while precision falls from 0.51 to 0.38 and recall from 0.63 to 0.25. The policy trained on raw rewards both admits more false edges and finds fewer true ones. The other three benchmarks show the same direction with smaller magnitude (Asia F1 0.61 $\rightarrow$ 0.57, Cancer 0.57 $\rightarrow$ 0.48, Sachs 0.62 $\rightarrow$ 0.55).

Cross-update structural memory. Consensus guidance retains edges that recur among high-reward candidates, so its removal should hurt most where useful edges are hard to find and easily lost. Sachs is that case: removing the teacher lowers F1 from 0.62 to 0.46 and raises SHD from 11.00 to 14.67, the largest SHD change in the study, with recall dropping from 0.54 to 0.38; the edges the teacher would have carried forward are not rediscovered from current-group rewards alone. F1 decreases and SHD increases on Asia, Cancer, and NetSim as well. Together with the previous paragraph, this places each signal's largest effect on a different benchmark, in line with the two-timescale design.

Representation, configuration, and run-to-run stability. The gated encoder and configuration search contribute on top of the two search signals: on NetSim, a standard Transformer gives F1

0.28 and SHD 12.40, and a fixed configuration gives F1 0.33 and SHD 12.00. The full model is also the most stable variant. Its F1 standard deviation is the smallest or tied-smallest in every block (0.03 on Asia versus 0.15–0.20 for the search-signal ablations; 0.00 on Cancer versus 0.09–0.12; 0.03 on NetSim versus 0.07–0.12), so the two signals improve not only the mean graph but its reproducibility across runs, the instability that motivated the design. Section 4.4 tests whether configuration search alone explains the gains, and Appendix G.7 examines coefficient response and teacher–policy trajectories.

Table 6: Full component ablations across four benchmarks. Entries are mean ± sample standard deviation over ten runs; bold marks the best value per dataset, and gray highlights RECAP. Figure 3 traces F1 and SHD.

Dataset	Method	FDR ↓	FPR ↓	ACC ↑	Precision ↑	Recall ↑	F1 ↑	SHD ↓
Asia	W/O Encoder	0.28 ± 0.30	0.03 ± 0.04	0.90 ± 0.05	0.72 ± 0.30	0.38 ± 0.18	0.47 ± 0.20	6.60 ± 2.88
	W/O Relative	0.26 ± 0.31	0.03 ± 0.04	0.91 ± 0.06	0.74 ± 0.31	0.48 ± 0.14	0.57 ± 0.20	6.00 ± 3.54
	W/O Consensus	0.27 ± 0.22	0.03 ± 0.02	0.90 ± 0.03	0.73 ± 0.22	0.35 ± 0.14	0.46 ± 0.15	6.40 ± 1.67
	Fixed Config.	0.32 ± 0.18	0.05 ± 0.04	0.90 ± 0.03	0.68 ± 0.18	**0.54 ± 0.06**	0.57 ± 0.04	6.40 ± 1.90
	RECAP	**0.20 ± 0.12**	**0.02 ± 0.01**	**0.92 ± 0.01**	**0.80 ± 0.12**	0.50 ± 0.00	**0.61 ± 0.03**	**5.10 ± 0.74**
Cancer	W/O Encoder	0.42 ± 0.09	0.06 ± 0.02	0.86 ± 0.02	0.58 ± 0.09	0.45 ± 0.11	0.50 ± 0.09	3.50 ± 0.53
	W/O Relative	0.40 ± 0.09	**0.05 ± 0.00**	0.86 ± 0.02	0.60 ± 0.09	0.40 ± 0.13	0.48 ± 0.12	3.40 ± 0.52
	W/O Consensus	0.45 ± 0.08	0.06 ± 0.02	0.85 ± 0.02	0.55 ± 0.08	0.38 ± 0.13	0.44 ± 0.11	3.70 ± 0.48
	Fixed Config.	0.38 ± 0.08	**0.05 ± 0.00**	0.87 ± 0.02	0.62 ± 0.08	0.42 ± 0.12	0.50 ± 0.12	3.30 ± 0.48
	RECAP	0.33 ± 0.00	**0.05 ± 0.00**	**0.88 ± 0.00**	**0.67 ± 0.00**	**0.50 ± 0.00**	**0.57 ± 0.00**	**3.00 ± 0.00**
NetSim	W/O Encoder	0.60 ± 0.11	0.05 ± 0.01	0.88 ± 0.02	0.40 ± 0.11	0.29 ± 0.15	0.28 ± 0.07	12.40 ± 1.52
	W/O Relative	0.62 ± 0.14	0.05 ± 0.02	0.87 ± 0.02	0.38 ± 0.14	0.25 ± 0.10	0.29 ± 0.12	12.80 ± 1.92
	W/O Consensus	0.56 ± 0.04	0.07 ± 0.02	0.88 ± 0.01	0.44 ± 0.04	0.44 ± 0.15	0.43 ± 0.09	12.20 ± 0.84
	Fixed Config.	0.57 ± 0.13	**0.04 ± 0.01**	0.88 ± 0.02	0.43 ± 0.13	0.27 ± 0.10	0.33 ± 0.11	12.00 ± 1.70
	RECAP	**0.49 ± 0.05**	0.08 ± 0.02	**0.89 ± 0.01**	**0.51 ± 0.05**	**0.63 ± 0.09**	**0.56 ± 0.03**	**10.91 ± 0.94**
Sachs	W/O Encoder	0.28 ± 0.08	**0.02 ± 0.01**	0.89 ± 0.01	0.72 ± 0.08	0.38 ± 0.10	0.49 ± 0.09	13.10 ± 1.45
	W/O Relative	0.37 ± 0.06	0.05 ± 0.01	0.89 ± 0.01	0.63 ± 0.06	0.50 ± 0.05	0.55 ± 0.05	13.55 ± 1.63
	W/O Consensus	0.40 ± 0.14	0.04 ± 0.01	0.88 ± 0.02	0.60 ± 0.14	0.38 ± 0.15	0.46 ± 0.15	14.67 ± 2.66
	Fixed Config.	0.35 ± 0.23	0.03 ± 0.01	0.89 ± 0.02	0.65 ± 0.23	0.35 ± 0.12	0.46 ± 0.15	12.20 ± 0.92
	RECAP	**0.25 ± 0.07**	0.03 ± 0.01	**0.91 ± 0.01**	**0.75 ± 0.07**	**0.54 ± 0.08**	**0.62 ± 0.05**	**11.00 ± 0.94**

G.2 Large-Graph Performance and Resource Usage

Table 7 examines directed-edge F1 and reversal-aware SHD on nonlinear Erdős–Rényi graphs with 100 nodes/1,000 observations and 200 nodes/2,000 observations. At 100 nodes, we compare RECAP's previously reported five-run result with the literature results of Yuan et al. (2026) under the same graph setting; the graph realizations are not paired. At 200 nodes, RECAP and OCMB-SciNO use five runs on the same seed-indexed generated graphs. All uncertainties below are across-seed sample standard deviations.

100-node structural recovery. RECAP achieves the lowest mean SHD in the reported comparison, 93.8 ± 13.3, compared with 105.0 ± 19.5 for GRaSP and 109.0 ± 21.0 for PC. This 11.2-edit improvement over the strongest published SHD comparator highlights RECAP's structural recovery on larger nonlinear graphs. Its directed-edge F1 is 0.506 ± 0.039; the main advantage at this scale is graph-edit accuracy.

200-node directed-edge recovery. On the five seed-indexed graphs, RECAP attains F1 0.345 ± 0.043, compared with 0.284 ± 0.038 for the official OCMB-SciNO implementation. Its precision is 0.808 ± 0.078 versus 0.507 ± 0.072, and recall is 0.221 ± 0.036 versus 0.198 ± 0.026; mean SHD is similar (176.6 ± 6.9 versus 177.4 ± 13.0). Together, the results highlight graph-edit accuracy at 100 nodes and stronger directed-edge recovery at 200 nodes.

The fixed 30-epoch profile in Table 8 uses synthetic graphs with 2,000 samples, covers 10–200 nodes, and averages measurements over three runs on the hardware specified in Appendix E.3.4. Time per epoch increases from 0.381 seconds at 10 nodes to 7.998 seconds at 200 nodes; peak allocated GPU memory rises from 24.48 to 375.09 MiB over the same range. The intermediate entries show how these resource measures change with graph size rather than only at the endpoints. These measurements describe a fixed training window, while Appendix D relates sampled-edge storage to $O(BKN^2)$ and reward evaluation to regression and matrix-exponential costs.

Table 7: Directed-edge F1 and SHD on nonlinear ER graphs (mean ± sample standard deviation). The 100-node RECAP result is from the previously reported five runs; comparator results are from Yuan et al. (2026). The 200-node comparison uses the same generated graphs across methods.

Nodes	Method	Source	F1 ↑	SHD ↓
100	PC	OCMB paper (Yuan et al., 2026)	0.663 ± 0.050	109.0 ± 21.0
	GRaSP	OCMB paper	0.656 ± 0.040	105.0 ± 19.5
	H2PC	OCMB paper	0.250 ± 0.010	140.3 ± 15.5
	MMHC	OCMB paper	0.250 ± 0.010	140.3 ± 15.5
	IAMB	OCMB paper	0.400 ± 0.010	168.3 ± 11.8
	OCMB	OCMB paper	0.236 ± 0.030	167.3 ± 17.6
	OCMB-HITON	OCMB paper	0.355 ± 0.020	175.3 ± 19.3
	DAGMA-linear	OCMB paper	0.300 ± 0.040	192.0 ± 27.2
	CAM	OCMB paper	0.519 ± 0.050	284.0 ± 35.1
	NOTEARS-MLP	OCMB paper	0.275 ± 0.030	371.0 ± 18.5
	CaPS-alone	OCMB paper	0.177 ± 0.010	540.0 ± 16.5
	RECAP (ours)	Previously reported (5 runs)	0.506 ± 0.039	93.8 ± 13.3
200	OCMB-SciNO (Yuan et al., 2026)	Official implementation; 5 seeds	0.284 ± 0.038	177.4 ± 13.0
	RECAP (ours)	Five-seed rerun	0.345 ± 0.043	176.6 ± 6.9

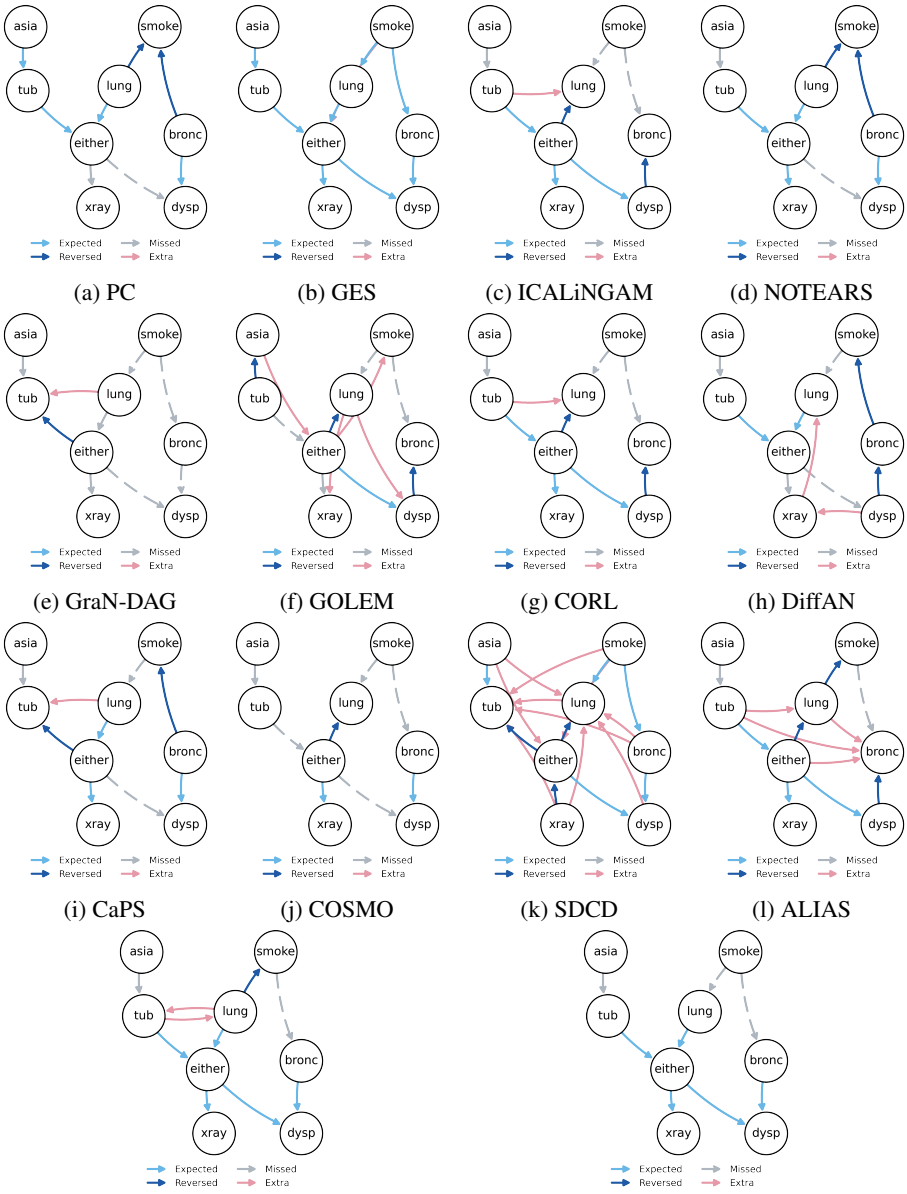
Table 8: Fixed-window resource profile on synthetic graphs (three-run averages). Runtime is per epoch; GPU memory is peak allocated or reserved memory in MiB.

Nodes	Time/epoch (s)	Peak allocated GPU (MiB)	Peak reserved GPU (MiB)
10	0.381	24.48	30.00
20	0.781	27.28	34.00
50	2.032	47.05	62.00
100	3.592	115.66	130.00
200	7.998	375.09	410.00

G.3 Asia and Cancer Benchmark Results

Figure 7 complements the main-text results in Table 1 with metric-wise comparisons of precision, recall, F1, and SHD on Asia and Cancer.

The Asia graph example in Figure 6 makes structural selectivity concrete: correctly recovered links can be inspected separately from extra, missing, and reversed links. Together with the ten-run metrics in Table 1, the example shows why fewer erroneous directed commitments translate into a more faithful final graph; the image illustrates the pattern while the repeated-run comparison establishes its scope.



(a) PC (b) GES (c) ICALiNGAM (d) NOTEARS

(e) GraN-DAG (f) GOLEM (g) CORL (h) DiffAN

(i) CaPS (j) COSMO (k) SDCD (l) ALIAS

(m) ABCNAS-RL (n) RECAP

Figure 6: Asia Bayesian-network graphs inferred by 14 methods. Edges are marked as correctly recovered, missing, reversed, or extra relative to the reference graph.

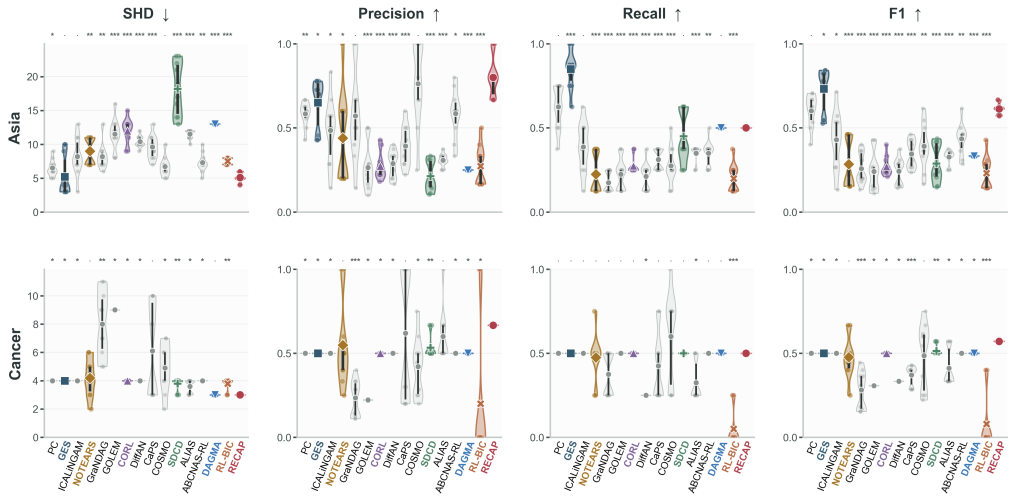


Figure 7: Run-level SHD (lower is better), precision, recall, and F1 (higher is better) on Asia and Cancer. Violins show the run distributions; points show runs, bars show interquartile ranges, and larger symbols show means. Colored symbols denote representative baselines; gray denotes other methods. Stars indicate Holm-adjusted paired comparisons with RECAP ($***$: $p < 0.001$; $**$: $p < 0.01$; $*$: $p < 0.05$); $\cdot$ indicates non-significance.

G.4 CSuite Benchmark Results

CSuite tests whether the precision–edit relationship seen on the Bayesian networks persists in another benchmark (Table 9). RECAP matches the best recall at 0.78 while reaching precision 1.00; the resulting F1 0.88 and SHD 2.00 are the best reported values. Here the improvement cannot be attributed to discarding true edges: at the same recall, selecting fewer incorrect directed edges produces a closer final graph. This complements the Asia comparison, where the advantage is primarily structural fidelity rather than maximum recall.

Table 9: CSuite results (ten-run mean ± standard deviation). Best values are bold; RECAP is shaded.

Metric	CaPS (2024)	COSMO (2024)	SDCD (2024)	ALIAS (2025)	ABCNAS-RL (2025)	RECAP (Ours)
Precision ↑	0.67 ± 0.03	0.52 ± 0.13	0.75 ± 0.00	0.88 ± 0.00	0.54 ± 0.00	**1.00 ± 0.00**
ACC ↑	0.93 ± 0.01	0.89 ± 0.03	0.94 ± 0.00	0.96 ± 0.00	0.90 ± 0.00	**0.98 ± 0.00**
Recall ↑	**0.78 ± 0.00**	0.56 ± 0.12	0.67 ± 0.00	**0.78 ± 0.00**	**0.78 ± 0.00**	**0.78 ± 0.00**
F1 ↑	0.72 ± 0.02	0.54 ± 0.12	0.71 ± 0.00	0.82 ± 0.00	0.64 ± 0.00	**0.88 ± 0.00**
SHD ↓	5.50 ± 0.53	8.70 ± 2.36	5.00 ± 0.00	3.00 ± 0.00	8.00 ± 0.00	**2.00 ± 0.00**

G.5 NetSim and Sachs Benchmark Results

Figure 8 presents the metric-wise comparisons of precision, recall, F1, and SHD on NetSim and Sachs. Figures 9 and 10 show the corresponding inferred graphs, distinguishing correctly recovered, missing, reversed, and extra edges.

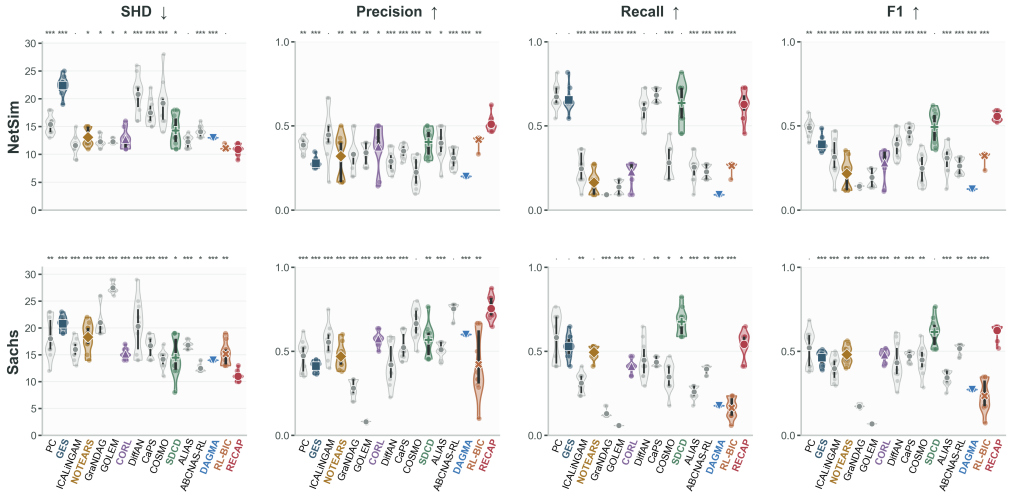
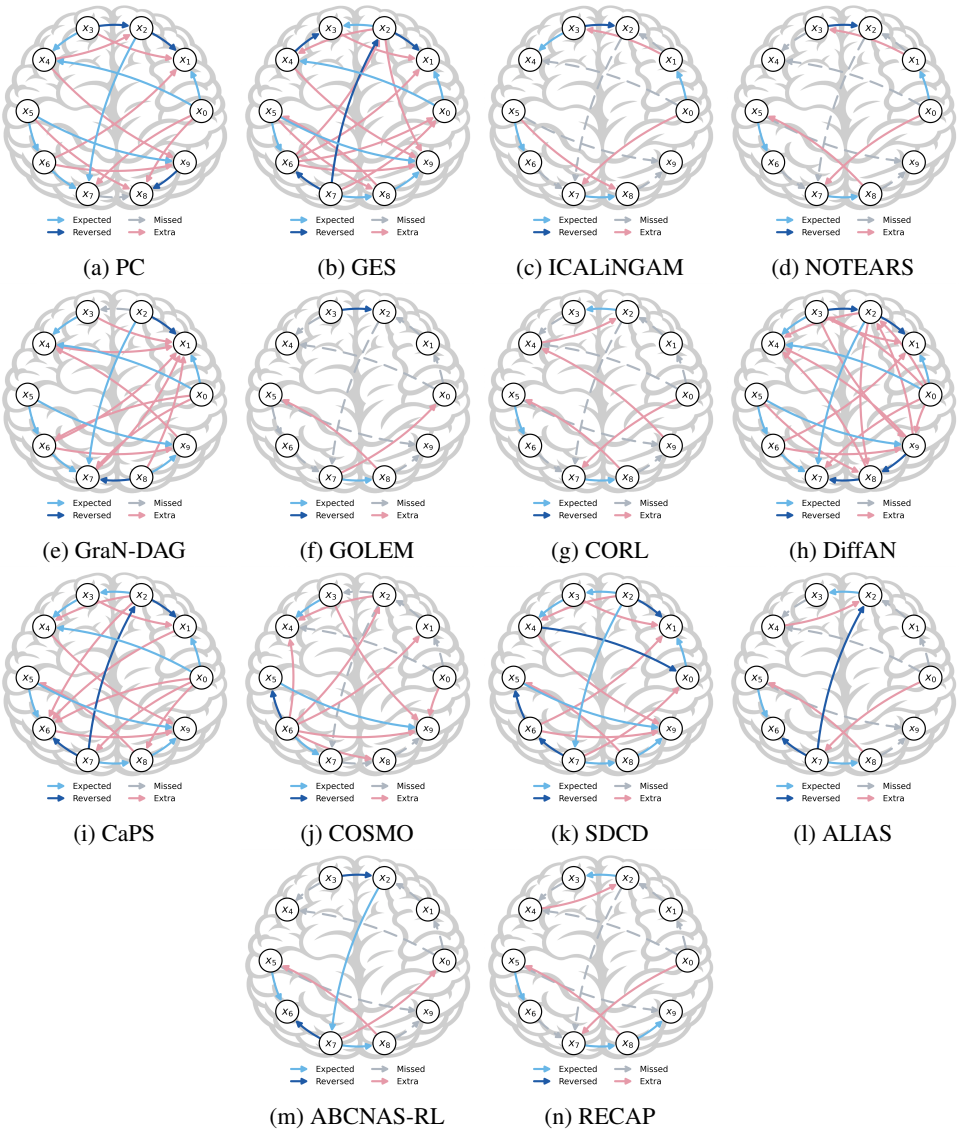


Figure 8: Run-level SHD (lower is better), precision, recall, and F1 (higher is better) on NetSim and Sachs. Violins show the run distributions; points show runs, bars show interquartile ranges, and larger symbols show means. Colored symbols denote representative baselines; gray denotes other methods. Stars indicate Holm-adjusted paired comparisons with RECAP ($***$: $p < 0.001$; $**$: $p < 0.01$; $*$: $p < 0.05$); $\cdot$ indicates non-significance.

G.5.1 Matched-budget NAS on NetSim

A stronger graph policy should remain useful even when competing learned methods receive configuration search. Table 10 expands the NetSim matched-budget comparison to the full error profile. Eligible learned baselines receive the same NAS procedure; ABCNAS-RL retains its original search procedure. RECAP combines F1 0.56 with SHD 10.91, versus the strongest competing F1 of 0.51 and the next-lowest SHD of 11.60. Its leading precision (0.51) and FDR (0.49) show that this combination is not achieved by simply adding more edges. Since the strongest alternative differs by metric, the joint F1–SHD advantage is the relevant observation: configuration search alone does not account for the observed balance of edge recovery and structural fidelity.



(a) PC (b) GES (c) ICALiNGAM (d) NOTEARS

(e) GraN-DAG (f) GOLEM (g) CORL (h) DiffAN

(i) CaPS (j) COSMO (k) SDCD (l) ALIAS

(m) ABCNAS-RL (n) RECAP

Figure 9: NetSim fMRI graphs inferred by 14 methods. Edges are marked as correctly recovered, missing, reversed, or extra relative to the reference graph.

Table 10: Full matched-budget NAS results on NetSim. Best values are bold; RECAP is shaded.

Dataset	Methods	FDR ↓	FPR ↓	ACC ↑	Precision ↑	Recall ↑	F1 ↑	SHD ↓
NetSim	NOTEARS-NAS (2018)	0.68 ± 0.14	0.08 ± 0.04	0.87 ± 0.02	0.32 ± 0.14	0.16 ± 0.08	0.22 ± 0.11	12.90 ± 1.60
	GraNDAG-NAS (2019)	0.63 ± 0.03	0.12 ± 0.02	0.85 ± 0.01	0.37 ± 0.03	0.55 ± 0.00	0.44 ± 0.02	15.40 ± 1.35
	GOLEM-NAS (2020)	0.54 ± 0.16	0.04 ± 0.02	0.88 ± 0.02	0.46 ± 0.16	0.30 ± 0.11	0.36 ± 0.13	11.60 ± 2.41
	DiffAN-NAS (2022)	0.66 ± 0.06	0.16 ± 0.02	0.82 ± 0.03	0.34 ± 0.06	0.67 ± 0.10	0.45 ± 0.08	18.10 ± 3.14
	CaPS-NAS (2024)	0.61 ± 0.03	0.37 ± 0.05	0.84 ± 0.02	0.39 ± 0.03	**0.73 ± 0.00**	0.51 ± 0.03	15.50 ± 1.58
	CORL-NAS (2021)	0.61 ± 0.10	0.04 ± 0.01	0.88 ± 0.01	0.39 ± 0.10	0.23 ± 0.05	0.29 ± 0.06	12.50 ± 1.35
	COSMO-NAS (2024)	0.73 ± 0.14	0.10 ± 0.06	0.84 ± 0.03	0.27 ± 0.14	0.40 ± 0.23	0.32 ± 0.17	15.73 ± 3.23
	SDCD-NAS (2024)	0.64 ± 0.06	0.16 ± 0.01	0.82 ± 0.02	0.36 ± 0.06	0.65 ± 0.11	0.46 ± 0.07	12.70 ± 1.06
	ALIAS-NAS (2025)	0.58 ± 0.10	**0.05 ± 0.01**	0.88 ± 0.01	0.42 ± 0.10	0.28 ± 0.10	0.34 ± 0.10	12.00 ± 1.12
	ABCNAS-RL (2025)	0.69 ± 0.06	0.08 ± 0.02	0.86 ± 0.01	0.31 ± 0.06	0.23 ± 0.05	0.26 ± 0.05	14.10 ± 0.99
	RECAP (Ours)	**0.49 ± 0.05**	0.08 ± 0.02	**0.89 ± 0.01**	**0.51 ± 0.05**	0.63 ± 0.09	**0.56 ± 0.03**	**10.91 ± 0.94**

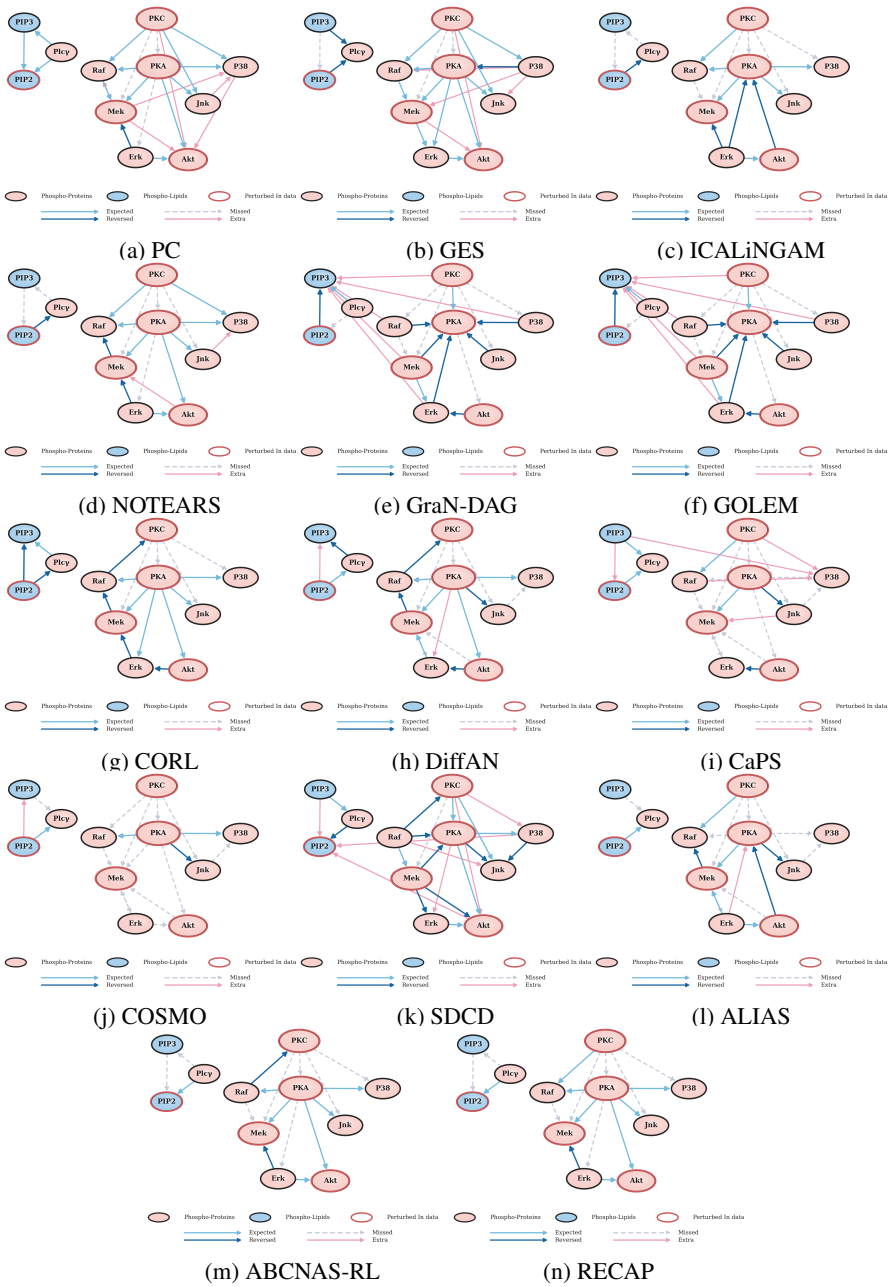


Figure 10: Causal graphs inferred by 14 algorithms on the Sachs protein-signaling network. Edges are distinguished as correctly recovered, missing, reversed, or extra relative to the reference network.

G.6 SynTReN Benchmark Results

SynTReN probes the cost of unreliable edge commitments in a gene-network setting (Table 11). RECAP has the highest precision (0.42) and the lowest FDR (0.58), FPR (0.03), and SHD (15.00) among the reported methods. The principal result here is a precision–structural-error profile: at F1 0.18, RECAP prioritizes a cleaner inferred graph and fewer edits, rather than optimizing for edge coverage (the highest reported F1 is 0.23). This is the distinction motivating relative evaluation and cross-iteration consensus: the search should preserve useful proposals without converting uncertain associations into extra directed edges. The component ablations, rather than this benchmark comparison, test the contribution of the two mechanisms.

Table 11: SynTReN results (mean ± standard deviation over repeated runs). Best values are bold; RECAP is shaded.

Dataset	Methods	FDR ↓	FPR ↓	ACC ↑	Precision ↑	Recall ↑	F1 ↑	SHD ↓
SynTReN	PC (2000)	0.81 ± 0.06	0.07 ± 0.02	0.90 ± 0.01	0.19 ± 0.06	0.19 ± 0.06	0.19 ± 0.06	22.60 ± 1.90
	GES (2002)	0.81 ± 0.02	0.05 ± 0.01	0.91 ± 0.00	0.19 ± 0.02	0.14 ± 0.02	0.16 ± 0.02	20.20 ± 0.63
	ICALiNGAM (2006)	0.80 ± 0.01	0.05 ± 0.00	0.92 ± 0.00	0.20 ± 0.01	0.07 ± 0.00	0.10 ± 0.00	17.10 ± 0.32
	NOTEARS (2018)	0.91 ± 0.02	0.19 ± 0.04	0.87 ± 0.01	0.09 ± 0.02	0.13 ± 0.03	0.11 ± 0.02	29.40 ± 3.06
	GraNDAG (2019)	0.90 ± 0.05	0.20 ± 0.06	0.87 ± 0.02	0.10 ± 0.05	0.14 ± 0.07	0.11 ± 0.05	29.90 ± 5.40
	GOLEM (2020)	0.95 ± 0.01	0.24 ± 0.01	0.85 ± 0.00	0.05 ± 0.01	0.08 ± 0.02	0.06 ± 0.02	34.40 ± 0.70
	DiffAN (2022)	0.59 ± 0.16	0.04 ± 0.01	**0.93 ± 0.01**	0.41 ± 0.16	0.16 ± 0.06	**0.23 ± 0.09**	15.10 ± 1.85
	CORL (2021)	0.88 ± 0.05	0.25 ± 0.02	0.85 ± 0.01	0.12 ± 0.05	0.21 ± 0.10	0.15 ± 0.06	33.80 ± 2.44
	CaPS (2024)	0.92 ± 0.03	0.18 ± 0.01	0.87 ± 0.00	0.08 ± 0.03	0.11 ± 0.04	0.09 ± 0.03	28.80 ± 0.42
	COSMO (2024)	0.96 ± 0.04	0.07 ± 0.03	0.88 ± 0.03	0.04 ± 0.04	0.04 ± 0.04	0.03 ± 0.04	27.60 ± 6.02
	SDCD (2024)	0.96 ± 0.03	0.07 ± 0.01	0.87 ± 0.01	0.04 ± 0.03	0.04 ± 0.04	0.04 ± 0.03	28.60 ± 2.17
	ALIAS (2025)	0.85 ± 0.04	0.10 ± 0.01	0.86 ± 0.01	0.15 ± 0.04	**0.26 ± 0.08**	0.19 ± 0.05	31.10 ± 2.08
	ABCNAS-RL (2025)	0.79 ± 0.07	0.10 ± 0.02	0.84 ± 0.01	0.21 ± 0.07	0.15 ± 0.06	0.18 ± 0.07	15.70 ± 1.16
	RECAP (Ours)	**0.58 ± 0.08**	**0.03 ± 0.01**	**0.93 ± 0.01**	**0.42 ± 0.08**	0.12 ± 0.07	0.18 ± 0.07	**15.00 ± 1.15**

G.7 Hyperparameter Sensitivity Analysis

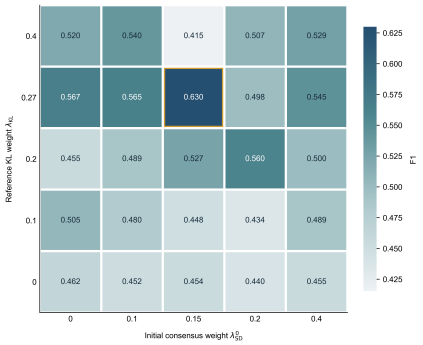


Figure 11: Recorded F1 by reference-policy KL weight (vertical) and initial consensus weight (horizontal). Cell labels show the 25 recorded values; the gold outline marks the maximum.

The coefficient grid asks whether policy stabilization and consensus guidance require a narrow tuning coincidence, while the teacher–policy and loss trajectories in Appendix F show how guidance enters and eventually leaves the search. These are complementary diagnostics, not independent proofs of better recovery: Table 6 directly tests the outcome of removing the search mechanisms.

Optimization coefficient scope. Two implementation coefficients control the regularizers: `grpo_kl_coef` weights the reference-policy KL term $\mathcal{L}_{\mathrm{KL}}$, while `distill_coef` sets the initial consensus weight λ_{SD}^0. The group-relative policy-gradient surrogate $\mathcal{L}_{\mathrm{GR}}$ has no separate coefficient. Consensus activates after warm-up and decays with `distill_coef_decay` on subsequent updates. Figure 11 plots all 25 archived F1 entries with axes interpreted according to the accompanying plotting code. The largest recorded F1 is 0.630, at KL weight 0.27 and initial distillation weight 0.15. Reading the two axes together shows how the relative strengths of policy stabilization and teacher guidance correspond to the recorded recovery scores; the peak lies at an interior combination rather than at a zero-regularization setting. The grid thus identifies a recorded coefficient response and its best observed combination, without estimating uncertainty across runs.