Effective Bayesian Causal Inference via Structural Marginalisation and Autoregressive Orders

Christian Toth
TU Graz

Christian Knoll
Levata GmbH

Franz Pernkopf
TU Graz

Robert Peharz
TU Graz

Abstract

The traditional two-stage approach to causal inference first identifies a *single* causal model (or equivalence class of models), which is then used to answer causal queries. However, this neglects any epistemic model uncertainty. In contrast, *Bayesian* causal inference does incorporate epistemic uncertainty into query estimates via Bayesian marginalisation (posterior averaging) over *all* causal models. While principled, this marginalisation over entire causal models, i.e., both causal structures (graphs) and mechanisms, poses a tremendous computational challenge. In this work, we address this challenge by decomposing structure marginalisation into the marginalisation over (i) causal orders and (ii) directed acyclic graphs (DAGs) given an order. We can marginalise the latter in closed form by limiting the number of parents per variable and utilising Gaussian processes to model mechanisms. To marginalise over orders, we use a sampling-based approximation, for which we devise a novel auto-regressive distribution over causal orders (ARCO). Our method outperforms state-of-the-art in structure learning on simulated non-linear additive noise benchmarks, and yields competitive results on real-world data. Furthermore, we can accurately infer interventional distributions and average causal effects.

1 INTRODUCTION

Few topics in science and philosophy have been as controversial as the nature of causality. Interestingly, the discussion becomes relatively benign, from a philosophical perspective, as soon as one agrees on a well-defined mathematical model of causality, such as a *structural causal model* (SCM) (Pearl, 2009). Assuming that the data comes from *some* model within a considered class of SCMs, causal questions reduce, in principle, to *epistemic* questions, i.e., questions about what and how much is known about the model.

In principle, knowing 'just' the causal structure of the true model, e.g., provided in the form of a directed acyclic graph (DAG), already permits the identification and estimation of causal quantities. When lacking such structural knowledge, one typically infers a *single* causal structure, which is then used to estimate the desired causal quantities (e.g., an average treatment effect). However, committing to a single model neglects any epistemic model uncertainty stemming from the finite amount and/or quality of available data. This is problematic, as a mismatch between the inferred causal structure and the true structure may severely affect the quality and truthfulness of the subsequent causal estimates. Reflecting the epistemic uncertainty about the underlying causal structure in downstream causal estimates is thus of central importance.

These considerations naturally invite a Bayesian treatment, i.e., specifying a prior distribution over *entire* SCMs (including causal structure, mechanisms and exogenous variables) and a likelihood model to infer the posterior over SCMs given collected data. *Bayesian causal inference* (BCI) then naturally incorporates epistemic uncertainty about the true causal model into downstream causal estimates via Bayesian marginalisation (posterior averaging) over *all* causal models: the causal estimate of each model is weighted with the models' posterior score.

Although BCI is conceptually appealing and principled, in practice, it becomes computationally intractable even for small problem instances due to the prohibitive number of possible causal structures to marginalise over. For example, a model with 20 variables would require

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

Code available at:
https://www.github.com/chritoth/bci-arco-gp

to compute and marginalise over posterior SCMs with more than 10^{72} candidate causal DAGs (OEIS Foundation Inc., 2024). Therefore, any practical approach to BCI must rely on approximations and/or assumptions to limit the number of considered candidate structures.

In this work, we improve upon existing BCI approaches by decomposing structure marginalisation into (i) the marginalisation over causal orders and (ii) the marginalisation over DAGs given a causal order. We perform the latter by restricting the number of parents per variable to a fixed maximal number K, permitting *exact* marginalisation over all possible DAGs for a given order over d variables in polynomial ($\mathcal{O}(d^K)$) time. This *structural marginalisation* over DAGs given causal orders, in combination with the analytic treatment of the causal mechanisms via Gaussian processes (GPs) as in (Toth et al., 2022), reduces the inference problem to the marginalisation w.r.t. causal orders. To this end, we propose a neural *A*uto-*R*egressive distribution over *C*ausal *O*rders (ARCO), that we utilise for a sampling-based approximation.

These techniques form a novel and effective approach to BCI, where our main contributions are:

- We propose ARCO, a neural auto-regressive distribution over causal orders, and devise a gradient-based learning scheme for it.
- We combine ARCO and exact structural marginalisation over DAGs with GPs into *ARCO-GP*, an effective BCI framework, reducing the intractable BCI problem to the marginalisation over causal orders.
- We demonstrate in experiments that ARCO-GP sets state-of-the-art in causal structure learning against a wide range of baselines when our model assumption are exactly matched. When the model assumptions are violated, we still outperform or are competitive with all baselines.
- We illustrate that our method accurately infers interventional distributions, which allows us to estimate posterior average causal effects and many other causal quantities of interest.

2 RELATED WORK

Rather than Bayesian *causal inference*, most existing literature addresses Bayesian approaches to causal *structure learning*, which dates back to work as early as (Heckerman, 1995; Heckerman et al., 1997; Madigan et al., 1995; Murphy, 2001; Tong and Koller, 2001). Often based on MCMC techniques, many works utilise (causal) orders, e.g., (Koller and Friedman, 2003; Koivisto and Sood, 2004; Teyssier and Koller, 2012; Ellis and Wong, 2008; Niinimäki et al., 2016; Kuipers and Moffa, 2017; Viinikka et al., 2020; Giudice et al., 2023), for a better exploration of the posterior space. The technique of restricting the maximum number of parents per node is well-established in Markov chain Monte Carlo (MCMC)-based structure learning (Koller and Friedman, 2003; Koivisto and Sood, 2004; Viinikka et al., 2020), but has—to the best of our knowledge—not been exploited in gradient-based structure learning so far. Additionally, MCMC inference comes with its own set of challenges, and none of these works implement non-linear mechanism models. In contrast, our work focuses on non-linear additive noise models and utilises gradient-based learning of a generative model over causal orders. Besides sampling based inference, orders can also facilitate exact optimization schemes, e.g. (Cussens, 2010; De Campos and Ji, 2011; Peharz and Pernkopf, 2012).

In a different stream of work utilising gradient-based (Bayesian) DAG structure learning methods (Zheng et al., 2018; Yu et al., 2019; Brouillard et al., 2020; Lachapelle et al., 2020; Lorch et al., 2021; Annadani et al., 2021; Tigas et al., 2022; Deleu et al., 2022; Wang et al., 2022; Rittel and Tschiatschek, 2023), inference via orders recently gained interest in the gradient-based causal structure learning community as a vehicle to sample DAGs without the need of utilising soft acyclicity constraints during optimisation (Cundy et al., 2021; Charpentier et al., 2022; Wang et al., 2022; Annadani et al., 2023; Rittel and Tschiatschek, 2023). Wang et al. (2022) utilise probabilistic circuits to enable tractable uncertainty estimates for causal structures. In contrast to these works, we stress the importance of an expressive gradient-based model for causal orders and we utilise causal orders for structural marginalisation in Bayesian causal inference.

Finally, existing BCI approaches are mostly restricted to linear Gaussian models (Geiger and Heckerman, 1994; Viinikka et al., 2020; Pensar et al., 2020; Horii, 2021) or binary variables (Moffa et al., 2017; Kuipers et al., 2019). Only recently, works on a Bayesian treatment of entire *non-linear* SCMs (i.e., including mechanisms and exogenous noise) have been proposed by Toth et al. (2022); Giudice et al. (2024). Toth et al. (2022) focus on an active learning scenario using DIBS (Lorch et al., 2021) for inferring posterior causal graphs and GPs for mechanism inference. Similarly, in a recent pre-print, Giudice et al. (2024) follow Toth et al. (2022) in using GPs for BCI but use MCMC for sampling posterior causal graphs. None of these approaches features gradient-based inference of causal orders, nor utilises causal orders for structure marginalisation.

3 BACKGROUND

Structural Causal Models. An SCM $\mathcal{M}$ over observed endogenous variables $\mathbf{X} = \{X_1, \dots, X_d\}$ and unobserved exogenous variables $\mathbf{U} = \{U_1, \dots, U_d\}$ with joint distribution $p(\mathbf{U})$, consists of structural equations, or mechanisms,

$$X_i := f_i(\mathbf{Pa}_i, U_i), \qquad \text{for} \quad i \in \{1, \dots, d\}, \quad (1)$$

which assign the value of each X_i as a deterministic function f_i of its direct causes, or causal parents, $\mathbf{Pa}_i \subseteq \mathbf{X} \setminus \{X_i\}$ and an exogenous variable U_i. In this paper we assume that the exogenous variables are independent Gaussian and enter in additive fashion, i.e. $f_i(\mathbf{Pa}_i, U_i) = f_i(\mathbf{Pa}_i) + U_i$, thus implying causal sufficiency. Associated with each SCM is a directed graph G induced by the set of parent sets $\mathbf{Pa} = \{\mathbf{Pa}_i\}_{i=1}^d$ with vertices $\mathbf{X}$ and edges $X_j \to X_i$ if and only if $X_j \in \mathbf{Pa}_i$. Any SCM with an *acyclic* directed graph (DAG) then induces a unique observational distribution $p(\mathbf{X} \,|\, \mathcal{M})$ over the endogenous variables $\mathbf{X}$, which is obtained as the pushforward measure of $p(\boldsymbol{U})$ through the causal mechanisms in Equation (1).

A (hard) intervention $do(\mathbf{W} = \mathbf{w})$ on a set of endogenous variables $\mathbf{W} \subset \mathbf{X}$ replaces the targeted mechanisms with constants $\mathbf{w}$, resulting in a modified SCM. The entailed modified causal graph lacks the incoming edges into any intervention target. The pushforward through the modified SCM yields an interventional distribution $p(\mathbf{X} \,|\, do(\mathbf{W} = \mathbf{w}), \mathcal{M})$.

Causal Orders. A permutation $L = \langle L_1, \dots, L_d \rangle$ of the endogenous variables $\mathbf{X} = \bigcup_{i=1}^d \{L_i\}$, where $L_i \neq L_j$ for all $i \neq j$, entails a strict total order $L_1 \prec L_2 \prec \cdots \prec L_d$ among the variables. Henceforth, we refer to such a permutation L as a *causal order*. A causal order L constrains the possible causal interactions between the variables, i.e., X_i can be a (direct) cause of X_j if, and only if, $X_i \prec X_j$ in L. We define $L_{<k} = \langle L_1, \dots, L_{k-1} \rangle$ to be the first $k-1$ elements in L. Finally, let $\lambda^L \colon \mathbf{X} \mapsto \{1, \dots, d\}$ be the bijective mapping between $\mathbf{X}$ and indices in L, i.e., $\lambda^L(X_i) = k \iff L_k = X_i$. We then denote by $Q^L \in \{0,1\}^{d \times d}$ the *permutation matrix* representing L, where $Q^L_{ij} = 1$ iff $\lambda^L(X_i) = j$.

4 BAYESIAN CAUSAL INFERENCE VIA STRUCTURAL MARGINALISATION

Within the Bayesian causal inference (BCI) framework, we refer to the causal quantity of interest as the *causal query* Y, which is a function of the SCM $\mathcal{M}$ (see (Toth et al., 2022)). This causal query could be, for example, an endogenous variable under some intervention, (features of) the true causal graph, or even the entire SCM. Since we are following a Bayesian approach, $\mathcal{M}$ is a random variable equipped with a prior distribution $p(\mathcal{M})$, and hence also Y is a random variable with distribution $p(Y \,|\, \mathcal{M})$. In this work, we focus our practical implementation on acyclic, non-linear additive Gaussian noise models, which informs our prior $p(\mathcal{M})$ and likelihood $p(\mathcal{D} \,|\, \mathcal{M})$ accordingly (see Sections 4.2 to 4.4). Under our assumptions, the true graph and causal effects are identifiable (see discussion in Appendix A).

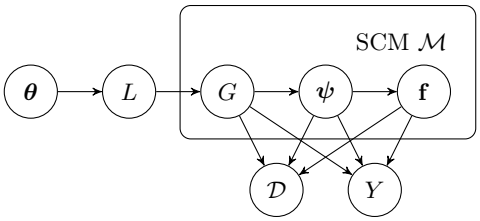

Figure 1: **Generative model of ARCO-GP.** We characterise a Structural Causal Model (SCM) $\mathcal{M} = (G, \mathbf{f}, \boldsymbol{\psi})$ by a causal graph G, causal mechanisms $\mathbf{f}$ and parameters $\boldsymbol{\psi}$ of a joint distribution over mechanisms and exogenous variables $p(\mathbf{f}, \mathbf{U} \,|\, \boldsymbol{\psi})$. We model the mechanisms $\mathbf{f}$ using Gaussian Processes (GPs) and $\mathbf{U}$ as additive Gaussian noise, implying that $\boldsymbol{\psi}$ is a set of GP hyper-parameters. The SCM gives rise to the data-generating likelihood $p(\mathcal{D} \,|\, \mathbf{f}, \boldsymbol{\psi}, G)$ and determines the (distribution over the) causal query Y. To sample a SCM, we first sample a causal order L from a neural auto-regressive distribution over causal orders (ARCO) $p(L \,|\, \boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$. Given a causal order and assuming a limited maximum cardinality of parent sets, we can then sample or marginalise causal graphs G and mechanisms $\mathbf{f}$ in closed form.

Given a set of (observational) data $\mathcal{D} = \{\mathbf{x}_n \overset{\text{i.i.d}}{\sim} p(\mathbf{X} \,|\, \mathcal{M}^*)\}_{n=1}^N$ collected from the true underlying SCM $\mathcal{M}^*$, BCI aims at inferring the posterior $p(Y \,|\, \mathcal{D})$ of the causal query. Since the causal query Y is determined by the SCM, we obtain the query posterior

$$\begin{aligned} p(Y \,|\, \mathcal{D}) &= \int p(\mathcal{M} \,|\, \mathcal{D}) \cdot p(Y \,|\, \mathcal{M}) \, d\mathcal{M} \\ &= \mathbb{E}_{\mathcal{M} \,|\, \mathcal{D}}[\, p(Y \,|\, \mathcal{M})], \end{aligned} \quad (2)$$

by marginalising over posterior SCMs. Simply put, to practically perform BCI, we first learn a generative model over SCMs $\mathcal{M} = (G, \mathbf{f}, \boldsymbol{\psi})$ which we parametrise by a causal graph G, causal mechanisms $\mathbf{f}$ and parameters $\boldsymbol{\psi}$ of a joint distribution over mechanisms and exogenous variables $p(\mathbf{f}, \mathbf{U} \,|\, \boldsymbol{\psi})$ (see description in Figure 1). This allows us to sample from and/or evaluate $p(Y \,|\, \mathcal{M})$. We then use our model to (approximately) marginalise over posterior SCMs as in Equation (2).

4.1 Learning and Inference

Our inference procedure, described in Algorithm 1, divides into a parameter learning and an inference phase. In the learning phase (Algorithm 1, lines 1-5), we infer posterior parameters $p(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D})$ of our generative model outlined in Figure 1. In the inference phase (Algorithm 1, lines 6-11), we use samples drawn from the learned generative model to approximate the query posterior. We first provide an overview of the individual phases starting with the inference phase, as it motivates our learning objective. We elaborate on the more technical details in Sections 4.2 to 4.4.

Inference Phase. The query posterior in Equation (2) can be written as the following importance weighted expectation w.r.t. our generative model in Figure 1 (for a derivation see Appendix E.1):

$$\begin{aligned} p(Y \mid \mathcal{D}) &= \mathbb{E}_{\mathcal{M} \mid \mathcal{D}} \left[p(Y \mid \mathcal{M}) \right] \\ &= \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D}} \left[\mathbb{E}_{L \mid \boldsymbol{\theta}} \left[w^L \cdot \mathbb{E}_{G \mid L, \boldsymbol{\psi}, \mathcal{D}} \left[\mathbb{E}_{\mathbf{f} \mid \boldsymbol{\psi}, \mathcal{D}} \left[p(Y \mid \mathcal{M}) \right] \right] \right] \right] \end{aligned} \tag{3}$$

with importance weights

$$w^L := \frac{\mathbb{E}_{G \mid L} \left[p(\mathcal{D} \mid \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \mid G) \right]}{\mathbb{E}_{L' \mid \boldsymbol{\theta}} \left[\mathbb{E}_{G' \mid L'} \left[p(\mathcal{D} \mid \boldsymbol{\psi}, G') \cdot p(\boldsymbol{\psi} \mid G') \right] \right]}. \tag{4}$$

In the inference phase, we assume that we have already learned a set of posterior parameters $\boldsymbol{\theta}, \boldsymbol{\psi}$ approximating $p(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D})$ (see the learning phase). On a high level, the query posterior $p(Y \mid \mathcal{D})$ in Eq. (3) is then estimated by (i) sampling several candidate SCMs from the learned generative model in a nested manner given $\boldsymbol{\theta}, \boldsymbol{\psi}$—first sampling an order L, then a graph G (parent sets) conditional on L, and mechanisms $\mathbf{f}$ conditional on G, (ii) sampling queries given the candidate SCM from $p(Y \mid \mathcal{M})$, and (iii) weighting the sampled queries with their corresponding importance weight w^L. See Appendix B.4 for details.

Learning Phase. To estimate the query posterior in Equation (3) we need to approximate $p(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D})$ (see the inference phase). By ensuring that the generative distributions $p(L \mid \boldsymbol{\theta})$ and $p(\mathbf{f}, \mathbf{U} \mid \boldsymbol{\psi})$ of our model are sufficiently expressive, it suffices to infer a maximum a posteriori (MAP) estimate of the parameters $\boldsymbol{\theta}, \boldsymbol{\psi}$ via gradient-based optimisation of (see Appendix E.2 for a derivation)

$$\begin{aligned} \nabla \log p(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D}) &= \nabla \log p(\boldsymbol{\theta}) \\ &\quad + \nabla \log \mathbb{E}_{L \mid \boldsymbol{\theta}} \left[\mathbb{E}_{G \mid L} \left[p(\mathcal{D} \mid \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \mid G) \right] \right]. \end{aligned} \tag{5}$$

A major issue for the training of such a model is the quality of the estimated gradients in the face of the high-dimensional problem space and the coupling of the parameters $\boldsymbol{\theta}, \boldsymbol{\psi}$. Specifically, updating all model parameters simultaneously in a single gradient step is prone to yield very noisy gradients:

1. The gradient w.r.t. $\boldsymbol{\theta}$ (causal order model) depends on $\boldsymbol{\psi}$ (mechanism and noise parameters) through the quality of the estimated importance weights w^L in Equations (4) and (6). A bad estimate of $\boldsymbol{\psi}$ will result in poor estimates of the importance weights and consequently the gradient w.r.t. $\boldsymbol{\theta}$.

2. The gradient w.r.t. $\boldsymbol{\psi}$ likewise depends on $\boldsymbol{\theta}$ via the quality of the sampled orders and their induced parent sets. In general, the more often a parent set occurs in the sampled orders relative to other parent sets, the stronger its gradient w.r.t. $\boldsymbol{\psi}$.[1]

To mitigate these issues, we propose a nested optimisation procedure as laid out in Algorithm 1 (lines 1-5). In an outer loop, we learn the parameters $\boldsymbol{\theta}$ of our generative distribution $p(L \mid \boldsymbol{\theta})$ using gradient-based optimisation of Equation (6) (see Section 4.2). However, the gradient estimates w.r.t. $\boldsymbol{\theta}$ in Equation (6), depend on the mechanism and noise parameters $\boldsymbol{\psi}$ needed to compute the importance weights w^L. Therefore, *before* computing the importance weights, we optimise the GP hyper-parameters in an inner loop (see Section 4.4) for each yet unseen parent set compatible with some sampled order (Algorithm 1 (lines 12-19)). This ensures truthful estimates of the importance weights for the sampled orders and, consequently, provides good gradient estimates.

4.2 Marginalising over Causal Orders

Computing the expectation w.r.t. causal orders in Equation (3) poses a hard combinatorial problem, as there are $d!$ possible causal orders over d variables. The involved distribution over causal orders $p(L \mid \boldsymbol{\theta})$ appearing in the BCI estimator in Equation (3) can be interpreted as proposal distribution with importance weights as defined in Equation (4), where the optimal proposal is the true posterior over causal orders. Hence, we require an expressive representation over causal orders which can account for *multi-modal distributions* over orders. In particular, consider the example of a Markov equivalence class (MEC) including a chain graph. Since the chain is contained in the MEC, also the reverse chain graph must be in the MEC, and thus, the proposal over causal orders must be able to represent both orders with equal probability.

[1] Arguably, this may be especially problematic when sharing parameters between mechanisms with different parent sets, as is the case, e.g., when modeling the mechanisms with a single, masked neural network.

Algorithm 1: BCI with ARCO-GP

Input: (Observational) data $\mathcal{D}$.
Output: Posterior parameters $\boldsymbol{\theta}$, $\boldsymbol{\psi}$.
Estimated (posterior over the) causal query Y and importance weights $\mathbf{w}^L$.

/* Learning Phase */
1 **repeat**
2 sample causal orders $\mathbf{L} \leftarrow \{L_m \sim p(L \mid \boldsymbol{\theta})\}$ ▷Section 4.2
3 $\mathbf{w}^L, \boldsymbol{\psi} \leftarrow$ `ComputeIW`$(\mathbf{L}, \mathcal{D})$ ▷Equation (4)
4 estimate the gradients for $\boldsymbol{\theta}$ and perform a gradient step ▷Equation (6)
5 **until** convergence
/* Inference Phase */
6 sample causal orders $\mathbf{L} \leftarrow \{L_m \sim p(L \mid \boldsymbol{\theta})\}$ ▷Section 4.2
7 $\mathbf{w}^L, \boldsymbol{\psi} \leftarrow$ `ComputeIW`$(\mathbf{L}, \mathcal{D})$ ▷Equation (4)
8 *[optional]* sample graphs $\mathbf{G} \leftarrow \{G_k \sim p(G \mid L, \boldsymbol{\psi}, \mathcal{D})\}$ ▷Section 4.3
9 *[optional]* sample mechanisms $\mathbf{f} \leftarrow \{\mathbf{f}_j \sim p(\mathbf{f} \mid \boldsymbol{\psi}, \mathcal{D})\}$ ▷Section 4.4
10 sample candidate queries $\mathbf{Y} \leftarrow \{Y_i \sim p(Y \mid \mathbf{f}, \boldsymbol{\psi}, G)\}$ ▷Section 4.4
11 **return** $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, $\mathbf{Y}$, $\mathbf{w}^L$
/* Subroutine: compute importance weights and update mechanisms. */
12 **Subroutine** `ComputeIW`$(\mathbf{L}, \mathcal{D})$
13 **foreach** causal order $L_m \in \mathbf{L}$ **do**
14 **foreach** parent set $\mathbf{Pa}_i$ compatible with L_m **do**
15 Learn (or retrieve) the corresponding GP hyper-parameters $\boldsymbol{\psi}_i$ ▷Equation (12)
16 **end**
17 Compute the importance weight w^{L_m} ▷Equation (4)
18 **end**
19 **return** $\mathbf{w}^L, \boldsymbol{\psi}$

A simple parameterisation of orders as proposed in (Charpentier et al., 2022) is not able to represent the true posterior over causal orders in this case. Specifically, they sample orders using the Gumbel Top-k trick (Kool et al., 2019) by perturbing d logits (corresponding to the d variables) with Gumbel noise and sorting these perturbed logits, yielding an order over variables. In essence, learning such a model boils down to ordering (and spreading) the d logits on the real line. Now, to sample a chain graph and a reverse chain graph, some variable is the first element in the causal order in one case and must thus have the highest (perturbed) logit, and it is the last element in the causal order in the other case where it must have the lowest (perturbed) logit, which is contradictory. In practice, we observe that when trying to learn a multi-modal distribution over causal orders with this model, the logits cluster together, resulting approximately in a uniform distribution over causal orders.

We therefore propose an expressive, auto-regressive distribution (Larochelle and Murray, 2011) $p(L \mid \boldsymbol{\theta}) = p(L_1 \mid \boldsymbol{\theta}) \cdot \prod_{k=2}^{d} p(L_k \mid L_{<k}, \boldsymbol{\theta})$ over causal orders (ARCO) that is amenable to gradient-based optimisation and can represent multi-modal distributions over orders, avoiding the shortcoming described above.[2]

Sampling Causal Orders. The ordering of variables naturally implies a sequential sampling procedure as listed in Algorithm 2. In each step of the sampling procedure, we sample the next variable in the order from a categorical distribution $p(L_k \mid L_{<k}, \boldsymbol{\theta})$ over the set of yet unassigned variables, conditional on all preceding variables in the order (Algorithm 2, line 6). To account for the dependence on the preceding order $L_{<k}$, we compute the logits of the categorical distribution using a differentiable function $g_{\boldsymbol{\theta}} : \mathbb{R}^{d \times d} \mapsto \mathbb{R}^d$ (Algorithm 2, line 4) and re-normalise them to exclude the elements in $L_{<k}$ (Algorithm 2, line 5). We implement $g_{\boldsymbol{\theta}}$ as feed-forward neural network, taking as input a suitable encoding of the so-far sampled order $L_{<k}$. To this end, we encode $L_{<k}$ using its induced permutation matrix $Q^{L_{<k}}$ (see Section 3) and mask the rows corresponding to elements $L_{>=k}$ with zeros.

Training ARCO. Training ARCO amounts to learning the parameters $\boldsymbol{\theta}$ of the neural network $g_{\boldsymbol{\theta}}$ by performing gradient ascent on Equation (5). As Equation (5) is not differentiable w.r.t. $\boldsymbol{\theta}$ because of the discrete nature of causal orders, we use the score-function estimator (Williams, 1992) to estimate the gradients via (see Appendices E.1 and E.2)

$$\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D}) = \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \mathbb{E}_{L \mid \boldsymbol{\theta}} \left[w^L \cdot \nabla_{\boldsymbol{\theta}} \log p(L \mid \boldsymbol{\theta}) \right], \tag{6}$$

with w^L as defined in Equation (4). To evaluate $\log p(L \mid \boldsymbol{\theta})$ for a given causal order L, we simply need to sum the log-probabilities of the categorical distributions $\log p(L_k \mid L_{<k}, \boldsymbol{\theta})$ for the respective elements L_k. The necessary log-probabilities (i.e., the normalised logits) are computed as described in Algorithm 2. Note that, although we need to compute the logits sequentially in the sampling procedure, we can compute them in parallel during evaluation.

[2] ARCO can be understood as an auto-regressive distribution estimator (Larochelle and Murray, 2011) constrained to causal orders.

Algorithm 2: Sample Causal Order (ARCO)

Input: Logit function $g_{\boldsymbol{\theta}} : \mathbb{R}^{d\times d} \mapsto \mathbb{R}^d$
Output: Causal order L sampled from $p(L \mid \boldsymbol{\theta})$

$\mathbf{R} \leftarrow \mathbf{X}$ ▷set of unassigned elements
$L \leftarrow \varnothing$ ▷causal order
for $k = 1 \dots d$ **do**
  $\boldsymbol{\phi} \leftarrow g_{\boldsymbol{\theta}}(Q^{L_{<k}})$ ▷compute logits
  $\phi_i \leftarrow \begin{cases} \phi_i - \log \sum_{j \mid X_j \in \mathbf{R}} \exp \phi_j & \text{if} \quad X_i \in \mathbf{R} \\ -\infty \quad \text{otherwise} \end{cases}$ ▷normalise logits
  $l \sim \text{CATEGORICAL}(\boldsymbol{\phi})$ ▷sample next element
  $L \leftarrow L \cup \langle X_l \rangle$ ▷update causal order
  $\mathbf{R} \leftarrow \mathbf{R} \setminus \{X_l\}$
end

4.3 Marginalising over Causal Graphs

The marginalisation w.r.t. causal graphs G in Equations (3) to (5) is in general intractable, as the number of DAGs consistent with any given (causal) order is $2^{\frac{d\cdot(d-1)}{2}}$. Although this is significantly smaller than the total number of DAGs with d nodes (which grows super-exponentially in d, see e.g. (OEIS Foundation Inc., 2024)), an exhaustive enumeration is still infeasible.

In this work, we tackle this problem by restricting the number of parents per variable. By restricting the maximum size of any admissible parent set to some integer K, the number of distinct parent sets consistent with any causal order L is in $\mathcal{O}(d^K)$. Although the exhaustive enumeration of all DAGs with restricted parent set size is still infeasible, it turns out that, given a causal order, the expectation w.r.t. graphs in Equation (3) can be tractably computed under certain assumptions. Rewriting the relevant expectation yields

$$\mathbb{E}_{G \mid L, \boldsymbol{\psi}, \mathcal{D}} \left[Y(G) \right] = \mathbb{E}_{G \mid L} [w(G) \cdot Y(G)] \tag{7}$$

by letting $Y(G) := \mathbb{E}_{\mathbf{f} \mid \boldsymbol{\psi}, \mathcal{D}} \left[p(Y \mid \mathcal{M}) \right]$ to avoid clutter, and

$$w(G) = \frac{p(\mathcal{D} \mid \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \mid G)}{\mathbb{E}_{G \mid L} \left[p(\mathcal{D} \mid \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \mid G) \right]}. \tag{8}$$

For $w(G)$ and the prior over graphs $p(G \mid L) = \prod_i p(\mathbf{Pa}_i \mid L)$ factorising, we can compute queries $Y(G)$ that decompose over the parent sets by the following two propositions (proofs in Appendix E.3).[3][4]

Proposition 4.1. *Let* $Y(G) = \prod_i Y_i(\boldsymbol{Pa}_i^G)$ *and* $w(G) = \prod_i w(\boldsymbol{Pa}_i^G)$ *be factorising over the parent sets, then*

$$\mathbb{E}_{G \mid L} \left[w(G) Y(G) \right] = \prod_{i=1}^{d} \sum_{\boldsymbol{Pa}_i} p(\boldsymbol{Pa}_i \mid L) w_i(\boldsymbol{Pa}_i) Y_i(\boldsymbol{Pa}_i). \tag{9}$$

Proposition 4.2. *Let* $Y(G) = \sum_i Y_i(\boldsymbol{Pa}_i^G)$ *be summing and* $w(G) = \prod_i w(\boldsymbol{Pa}_i^G)$ *be factorising over the parent sets, then*

$$\mathbb{E}_{G \mid L} \left[w(G) Y(G) \right] = \tag{10}$$
$$\sum_{i=1}^{d} \left(\prod_{j \neq i} \alpha_j(L) \right) \cdot \sum_{\boldsymbol{Pa}_i} p(\boldsymbol{Pa}_i \mid L) w_i(\boldsymbol{Pa}_i) Y_i(\boldsymbol{Pa}_i),$$

where

$$\alpha_j(L) = \sum_{\boldsymbol{Pa}_i} p(\boldsymbol{Pa}_i \mid L) w_i(\boldsymbol{Pa}_i).$$

Furthermore, for any causal query that does not decompose over parent sets, we can compute a Monte-Carlo estimate of the query by sampling DAGs from the true posterior $p(G \mid L, \boldsymbol{\psi}, \mathcal{D}) = w(G) \cdot p(G \mid L)$ in Equation (3). Employing Proposition 4.1, we can compute $\mathbb{E}_{G \mid L} \left[p(\mathcal{D} \mid \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \mid G) \right]$ and consequently $w(G)$ in closed-form by choosing $w(G) = p(\mathcal{D} \mid \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \mid G)$ and $Y(G) = 1$. Since $p(G \mid L, \boldsymbol{\psi}, \mathcal{D})$ factorises, we can sample DAGs by sampling parent sets individually for each node.

4.4 Mechanism Inference and Marginalisation

To compute the importance weights in Equations (4) and (8) we need to compute the marginal log-likelihood $p(\mathcal{D} \mid \boldsymbol{\psi}, G) = \prod_i p(\mathcal{D}_i \mid \boldsymbol{\psi}_i, \mathbf{Pa}_i^G)$ which is intractable for general models.[5]. As we focus on non-linear additive noise models in this work, we follow (von Kügelgen et al., 2019; Toth et al., 2022) and model each mechanism via a distinct GPs, assuming homoscedastic Gaussian noise and causal sufficiency. Under these assumptions, we can compute the marginal likelihood $p(\mathcal{D}_i \mid \boldsymbol{\psi}_i, \mathbf{Pa}_i) = \mathbb{E}_{\mathbf{f} \mid \boldsymbol{\psi}_i} \left[p(\mathcal{D}_i \mid \mathbf{f}_i, \boldsymbol{\psi}_i, \mathbf{Pa}_i) \right]$ and the GP predictive posterior in closed form. The GP hyperparameter prior and likelihood factorise over parent sets, i.e.,

$$p(\mathcal{D} \mid \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \mid G) = \prod_i p(\mathcal{D}_i \mid \boldsymbol{\psi}_i, G) \cdot p(\boldsymbol{\psi}_i \mid \mathbf{Pa}_i^G). \tag{11}$$

[3] In practice, we assume a uniform prior over parent sets consistent with a given causal order.

[4] These propositions generalise the results presented by Koller and Friedman (2003); Koivisto and Sood (2004) on how to compute the posterior probabilities of edges or parent sets, which is a special case of Proposition 4.1, to our setting of Bayesian causal inference.

[5] We abuse notation to denote by $p(\mathcal{D}_i \mid \boldsymbol{\psi}_i, \mathbf{Pa}_i^G) := p(\mathbf{x}_i \mid \mathbf{x}_{\mathbf{Pa}_i^G}, \boldsymbol{\psi}_i, \mathbf{Pa}_i^G)$

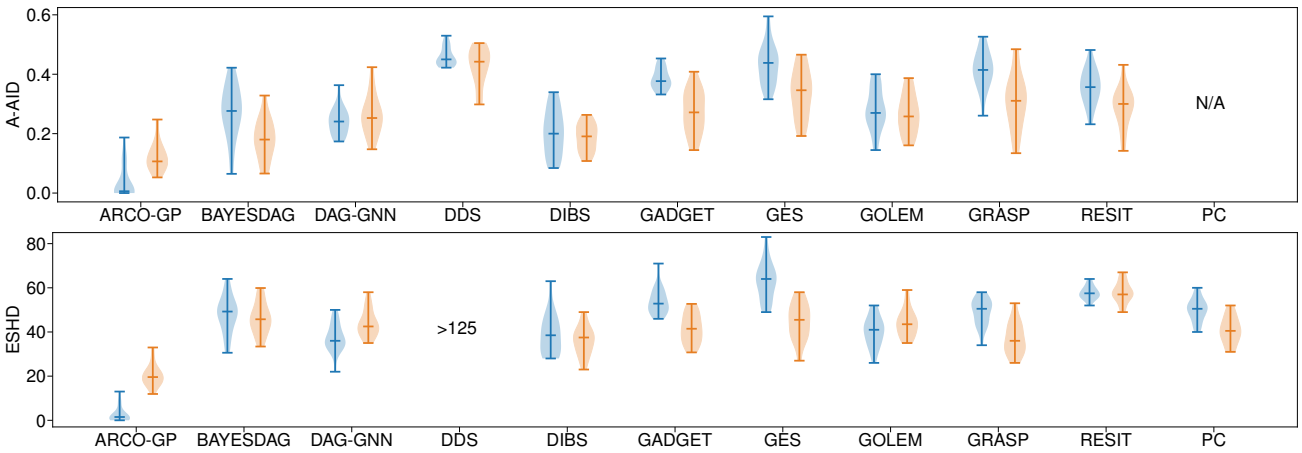

Figure 2: **Causal discovery on nonlinear additive noise models.** Structure learning results in terms of *expected Hamming distance* (ESHD) and *ancestor adjustment identification distance* (A-AID) on simulated non-linear models with scale-free (left, blue) and Erdös-Renyi (right, orange) graphs, each with 20 nodes and 200 data samples. Whiskers indicate maximum, minimum and median values across 20 simulated ground truth instances. For both metrics lower is better. Range for ESHD is set for better readability, omitting the result for DDS (> 125).

In practice, we thus use a separate GP for each unique parent set (in contrast to having a separate set of GPs for each individual graph), allowing for efficient computation by caching intermediate results. For the individual GP we infer a MAP-Type II estimate of its hyper-parameters $\boldsymbol{\psi}$ by performing gradient-ascent on[6]

$$\nabla_{\boldsymbol{\psi}_i} \log p(\boldsymbol{\psi}_i \,|\, \mathcal{D}, G) = \nabla_{\boldsymbol{\psi}_i} \log p(\mathcal{D}_i \,|\, \boldsymbol{\psi}_i, G) + \nabla_{\boldsymbol{\psi}_i} \log p(\boldsymbol{\psi}_i \,|\, G). \tag{12}$$

For root nodes, i.e., nodes without parents, we place a conjugate normal-inverse-gamma prior on the mean and variance of that node, which also allows for closed-form inference.

5 EXPERIMENTS

As a truthful (posterior over) causal structure is paramount for downstream causal inferences, Task I evaluates structure learning. To illustrate ARCO-GP's capability of performing integrated causal structure learning and reasoning, Task II evaluates the capability to infer interventional distributions from observational data.

Task I: Causal Structure Learning. For our results in Figure 2, we sample a fixed set of 200 training samples from the observational distribution of non-linear additive noise ground truth SCMs with Erdös-Rényi (ER) (Erdös and Rényi, 1959) and scale-free (SF) (Barabási and Albert, 1999) graph structures with 20 nodes. The small sample size emulates the setting of significant uncertainty relevant to the Bayesian inference scenario. We report the *expected structural Hamming distance* (ESHD) as a standard structure learning metric. Additionally, to asses the inferred structures in terms of causal implications, we report the *ancestor adjustment identification distance* (A-AID) recently proposed by Henckel et al. (2024) (see Appendix D for details and additional metrics).

We compare our inference model (ARCO-GP) with maximum parent set size $K = 2$ to a diverse set of ten different structure learning methods (see Appendix D for descriptions): BAYESDAG (Annadani et al., 2023), DAG-GNN (Yu et al., 2019), DDS (Charpentier et al., 2022), DIBS-GP (Toth et al., 2022; Lorch et al., 2021), GADGET (Viinikka et al., 2020), GES (Chickering, 2003), GOLEM (Ng et al., 2020), GRASP (Lam et al., 2022), PC (Spirtes et al., 2000) and RESIT (Peters et al., 2014).

The results in Figure 2 demonstrate that ARCO-GP is highly effective in terms of causal structure learning. In particular, for scale-free graphs, where the assumption of a restricted number of parents matches our modelling assumption, ARCO-GP clearly outperforms all baselines. Moreover, even when the assumptions are violated (in our case for Erdos-Renyi graphs), we demonstrate that ARCO-GP still outperforms the baselines, albeit with a smaller margin. Additional experiments on real-world data from Sachs et al. (2005), invertible non-linear models, and ablations varying the number of variables $d \in \{11, 20, 50\}$, sample sizes, as well as investigating the influence of the parent size restriction are provided in Appendix D.

[6]Importantly, in Appendix E.2 we show that this not an ad-hoc choice, but a consequence of optimising Equation (5).

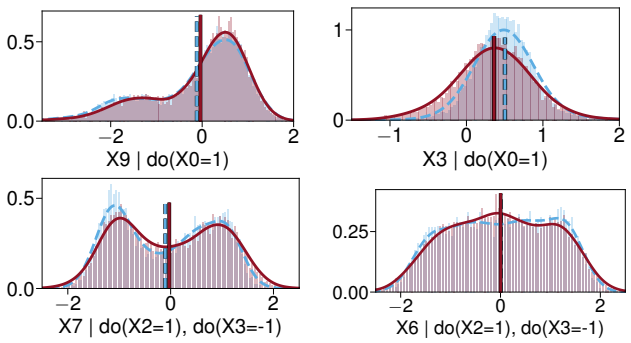

Figure 3: **Posterior interventional distributions.** Several interventional distributions as inferred by ARCO-GP (red, solid) and the corresponding ground truth (blue, dashed). Specifically, we sampled full SCMs (orders, graphs given orders, mechanisms, exogeneous variables) and performed the indicated intervention to produce a sample from the corresponding distribution, which effectively marginalises over the posterior over SCMs. Vertical lines indicate the estimated distribution means (average causal effects). See Appendix C for details.

Task II: Inferring Interventional Distributions and Average Causal Effects. We illustrate ARCO-GP's causal reasoning capabilities by visualising a selected set of estimated posterior interventional distributions $p(X_i \,|\, do(\mathbf{W} = \mathbf{w}), \mathcal{D})$ in Figure 3. To have access to ground truth interventional distributions, we simulate a non-linear additive noise model on the consensus protein interaction graph reported by Sachs et al. (2005) (see Appendix, Figure 4). We generate samples from the inferred interventional distributions with the procedure laid out in Algorithm 1 and Section 4, smoothing the empirical distribution with a kernel density estimate (see Appendix B for details). Importantly, the multi-modality of the inferred distributions illustrates the benefits of a Bayesian approach to causal inference, as we can represent structural uncertainty via full distributions instead of single point estimates (see especially Figure 3, lower left).

To quantify performance, we use our ground truth model to simulate five random interventional distributions $p(\mathbf{X} \,|\, \mathrm{do}(X_i = t))$ per node with $t \in [-1, 1]$, drawing 1000 samples per intervention. We then compute averages of the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) and L1 and L2 distances between the distribution means of inferred vs. true interventional distributions across all interventions. We compare against two BCI baselines, i.e., DIBS-GP (Toth et al., 2022; Lorch et al., 2021) and BEEPS (Viinikka et al., 2020). To emulate a baseline using the traditional two-stage approach, we first infer a causal graph using RESIT, that we then use to fit a GP model and infer interventional distributions (RESIT-GP). Note that our other baselines used in Task I do not infer interventional distributions, as they are geared towards structure learning. The results in Table 1 show that ARCO-GP outperforms the baselines with a significant gap to the emulated two-stage baseline RESIT-GP.

Table 1: **Inferring Interventional Distributions.** We report the MMD, as well as L1 and L2 distances between the distribution means of inferred and true interventional distributions. We generate samples from five interventional distributions $p(\mathbf{X} \,|\, \mathrm{do}(X_i = t))$ per node with random intervention values $t \sim \mathcal{U}(-1, 1)$. The metrics are averaged across all interventions. The reported numbers are averages and 95% confidence intervals (CIs) obtained from simulations on 10 different ground truth models with fixed graph (see Figure 4) and simulated nonlinear mechanisms.

	MMD	L1	L2
ARCO-GP	0.15 ± 0.04	0.93 ± 0.16	0.37 ± 0.06
BEEPS	-	1.43 ± 0.16	0.62 ± 0.07
DIBS-GP	0.17 ± 0.05	1.13 ± 0.26	0.47 ± 0.11
RESIT-GP	0.31 ± 0.12	1.57 ± 0.19	0.68 ± 0.07

6 DISCUSSION

We demonstrated that our proposed ARCO-GP method for BCI, leveraging structural marginalisation, yields superior structure learning performance on non-linear additive noise models against a set of ten state-of-the-art baseline methods. Moreover, we illustrate the capability of ARCO-GP to accurately infer posterior interventional distributions and average causal effects. The capabilities our method relies upon the following assumptions and limitations.

Assumptions. Our assumptions on the data generating process include causal sufficiency and additive, homoscedastic Gaussian noise. An extension to heteroscedastic (non-Gaussian) noise could be achieved by utilising GP extensions as, e.g., proposed by Wang and Neal (2012); Dutordoir et al. (2018), without making substantial changes to the ARCO-GP framework.

We further assume a limited maximum parent set size. Presumably, for very dense graphs the performance of ARCO-GP will deteriorate in comparison to other methods. However, our ablations in Appendix D show that ARCO-GP is still superior when violations are only moderate.

While our framework and implementation can handle training from interventional data, we do not evaluate this scenario experimentally because not all baselines support interventional input data, and the observa-

tional case is the more difficult problem from the perspective of model identifiability.

Scalability and Computation. The main driver of complexity is the exact inference using GPs, which grows with N^3 in the number of available data points. Although we used only CPUs for running our experiments, scaleable GPU inference techniques for GPs were proposed, e.g., by Gardner et al. (2018); Pleiss et al. (2018). Additionally, the training of the GPs could be straightforwardly parallelised. Conceptually, our framework is flexible and modular, allowing to use alternative mechanism models like normalising flows as in Brouillard et al. (2020); Pawlowski et al. (2020). A second driver of computational complexity is the exhaustive enumeration of parent sets, which may be prohibitive on larger problem instances and bigger parent sets. Note, however, that the individual parent set contributions necessary to compute the importance weights in Equation (4) could be pre-computed in parallel.

Acknowledgements

This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/COE12. This research has been conducted in whole or in part within the project VENTUS (project number FO999910263), which has received funding in the framework "AI4Green, Call 2023", a research and technology program of the Austrian Ministry of Climate Action and Energy, granted by the Austrian Research Promotion Agency (FFG, www.ffg.at). The financial support by the Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged. The computational results presented have in part been achieved using the Vienna Scientific Cluster (VSC).

References

Annadani, Y., Pawlowski, N., Jennings, J., Bauer, S., Zhang, C., and Gong, W. (2023). BayesDAG: Gradient-Based Posterior Inference for Causal Discovery. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Annadani, Y., Rothfuss, J., Lacoste, A., Scherrer, N., Goyal, A., Bengio, Y., and Bauer, S. (2021). Variational Causal Networks: Approximate Bayesian Inference over Causal Structures. *arXiv:2106.07635*.

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C. K., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M. Y., Suk, H., Zhang, S., Suo, M., Tillet, P., Zhao, X., Wang, E., Zhou, K., Zou, R., Wang, X., Mathews, A., Wen, W., Chanan, G., Wu, P., and Chintala, S. (2024). PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, New York, NY, USA. ACM.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. (2020). Differentiable Causal Discovery from Interventional Data. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Bühlmann, P., Peters, J., and Ernest, J. (2014). CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42.

Charpentier, B., Kibler, S., and Günnemann, S. (2022). Differentiable {DAG} Sampling. In *International Conference on Learning Representations*.

Chickering, D. M. (2003). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3.

Cundy, C., Grover, A., and Ermon, S. (2021). BCD Nets: Scalable Variational Approaches for Bayesian Causal Discovery. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Cussens, J. (2010). Maximum likelihood pedigree reconstruction using integer programming. In *WCB@ ICLP*, pages 8–19.

De Campos, C. P. and Ji, Q. (2011). Efficient structure learning of bayesian networks using constraints. *The Journal of Machine Learning Research*, 12:663–689.

Deleu, T., Góis, A., Emezue, C., Rankawat, M., Lacoste-Julien, S., Bauer, S., and Bengio, Y. (2022). Bayesian structure learning with generative flow networks. In Cussens, J. and Zhang, K., editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, Proceedings of Machine Learning Research. PMLR.

Dutordoir, V., Salimbeni, H., Hensman, J., and Deisenroth, M. (2018). Gaussian Process Conditional Density Estimation. In Bengio, S., Wallach, H., Larochelle,

H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Ellis, B. and Wong, W. H. (2008). Learning Causal Bayesian Network Structures from Experimental Data. *Journal of the American Statistical Association*.

Erdös, P. and Rényi, A. (1959). On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290.

Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. (2018). Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Geiger, D. and Heckerman, D. (1994). Learning gaussian networks. In *Uncertainty in Artificial Intelligence*, pages 235–243. Elsevier.

Giudice, E., Kuipers, J., and Moffa, G. (2023). A Bayesian Take on Gaussian Process Networks. In *Advances in Neural Information Processing Systems*.

Giudice, E., Kuipers, J., and Moffa, G. (2024). Bayesian causal inference with gaussian process networks. *arXiv preprint arXiv:2402.00623*.

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13.

Heckerman, D. (1995). A bayesian approach to learning causal networks. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann.

Heckerman, D., Meek, C., and Cooper, G. (1997). A Bayesian Approach to Causal Discovery. *Computation, Causation, and Discovery*.

Heinze-Deml, C., Maathuis, M. H., and Meinshausen, N. (2018). Causal Structure Learning. *Annual Review of Statistics and Its Application*.

Henckel, L., Würtzen, T., and Weichwald, S. (2024). Adjustment Identification Distance: A gadjid for Causal Structure Learning. *arXiv:2402.08616*.

Hinton, G. (2012). Neural networks for machine learning - lecture 6a - overview of mini-batch gradient descent. https://www.cs.toronto.edu/ tijmen/csc321/slides/lecture_slides_lec6.pdf.

Horii, S. (2021). Bayesian Model Averaging for Causality Estimation and its Approximation based on Gaussian Scale Mixture Distributions. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In And, Y. B. and LeCun, Y., editors, *3rd International Conference on Learning Representations, Conference Track Proceedings*, San Diego, CA, USA.

Koivisto, M. and Sood, K. (2004). Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*.

Koller, D. and Friedman, N. (2003). Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*.

Kool, W., Van Hoof, H., and Welling, M. (2019). Stochastic Beams and Where To Find Them: The {G}umbel-Top-k Trick for Sampling Sequences Without Replacement. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97. PMLR.

Kuipers, J. and Moffa, G. (2017). Partition MCMC for Inference on Acyclic Digraphs. *Journal of the American Statistical Association*.

Kuipers, J., Moffa, G., Kuipers, E., Freeman, D., and Bebbington, P. (2019). Links between psychotic and neurotic symptoms in the general population: an analysis of longitudinal british national survey data using directed acyclic graphs. *Psychological Medicine*, 49(3):388–395.

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. (2020). Gradient-Based Neural DAG Learning. In *International Conference on Learning Representations*.

Lam, W.-Y., Andrews, B., and Ramsey, J. (2022). Greedy relaxations of the sparsest permutation algorithm. In Cussens, J. and Zhang, K., editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, Proceedings of Machine Learning Research. PMLR.

Larochelle, H. and Murray, I. (2011). The neural autoregressive distribution estimator. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 29–37. JMLR Workshop and Conference Proceedings.

Lorch, L., Rothfuss, J., Schölkopf, B., and Krause, A. (2021). DiBS: Differentiable Bayesian Structure Learning. *Advances in Neural Information Processing Systems*.

Madigan, D., York, J., and Allard, D. (1995). Bayesian Graphical Models for Discrete Data. *International Statistical Review / Revue Internationale de Statistique*, 63.

Moffa, G., Catone, G., Kuipers, J., Kuipers, E., Freeman, D., Marwaha, S., Lennox, B. R., Broome, M. R., and Bebbington, P. (2017). Using directed acyclic graphs in epidemiological research in psychosis: an analysis of the role of bullying in psychosis. *Schizophrenia bulletin*, 43(6):1273–1279.

Murphy, K. P. (2001). Active learning of causal Bayes net structure. Technical report, Department of Computer Science, U.C. Berkeley.

Murphy, K. P. (2007). Conjugate Bayesian Analysis of the Gaussian Distribution. Technical report, University of British Columbia.

Murphy, K. P. (2023). *Probabilistic Machine Learning: Advanced Topics*. MIT Press.

Ng, I., Ghassami, A., and Zhang, K. (2020). On the Role of Sparsity and DAG Constraints for Learning Linear DAGs. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc.

Niinimäki, T., Parviainen, P., and Koivisto, M. (2016). Structure Discovery in Bayesian Networks by Sampling Partial Orders. *Journal of Machine Learning Research*.

OEIS Foundation Inc. (2024). Number of acyclic digraphs (or dags) with n labeled nodes. Entry A003024 in The On-Line Encyclopedia of Integer Sequences, https://oeis.org/A003024.

Pawlowski, N., Castro, D. C., and Glocker, B. (2020). Deep Structural Causal Models for Tractable Counterfactual Inference. In *Advances in Neural Information Processing Systems*.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peharz, R. and Pernkopf, F. (2012). Exact maximum margin structure learning of bayesian networks. *arXiv preprint arXiv:1206.6431*.

Pensar, J., Talvitie, T., Hyttinen, A., and Koivisto, M. (2020). A bayesian approach for estimating causal effects from observational data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5395–5402.

Peters, J. and Bühlmann, P. (2015). Structural Intervention Distance for Evaluating Causal Graphs. *Neural Computation*, 27.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research*.

Pleiss, G., Gardner, J., Weinberger, K., and Wilson, A. G. (2018). Constant-time predictive distributions for Gaussian processes. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Raskutti, G. and Uhler, C. (2018). Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7.

Reisach, A. G., Seiler, C., and Weichwald, S. (2021). Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy To Game. In Vaughan, M. R., Beygelzimer, A., Dauphin, Y., Liang, P., and Wortman, J., editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Rittel, S. and Tschiatschek, S. (2023). Specifying Prior Beliefs over DAGs in Deep Bayesian Causal Structure Learning. In Kobi, G., Ann, N., J., N. G., Roy, F., and Roxana, R., editors, *Frontiers in Artificial Intelligence and Applications*.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York, 2nd edition.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, 308.

Solus, L., Wang, Y., and Uhler, C. (2021). Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. The MIT Press, second edition.

Squires, C. and Uhler, C. (2022). Causal Structure Learning: A Combinatorial Perspective. *Foundations of Computational Mathematics*.

Teyssier, M. and Koller, D. (2012). Ordering-Based Search: A Simple and Effective Algorithm for Learning Bayesian Networks. *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence, UAI 2005*.

Tigas, P., Annadani, Y., Jesson, A., Schölkopf, B., Gal, Y., and Bauer, S. (2022). Interventions, Where and How? Experimental Design for Causal Models at Scale. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Tong, S. and Koller, D. (2001). Active learning for structure in Bayesian networks. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Toth, C., Lorch, L., Knoll, C., Krause, A., Pernkopf, F., Peharz, R., and von Kügelgen, J. (2022). Active Bayesian Causal Inference. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Viinikka, J., Hyttinen, A., Pensar, J., and Koivisto, M. (2020). Towards Scalable Bayesian Learning of Causal DAGs. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

von Kügelgen, J., Rubenstein, P. K., Schölkopf, B., and Weller, A. (2019). Optimal experimental design via Bayesian optimization: active causal structure learning for Gaussian process networks. *arXiv:1910.03962*.

Vowels, M. J., Camgoz, N. C., and Bowden, R. (2022). D'ya Like DAGs? A Survey on Structure Learning and Causal Discovery. *ACM Computing Surveys*.

Wang, B., Wicker, M. R., and Kwiatkowska, M. (2022). Tractable Uncertainty for Structure Learning. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Wang, C. and Neal, R. M. (2012). Gaussian Process Regression with Heteroscedastic or Non-Gaussian Residuals. *arXiv:1212.6246*.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8.

Yu, Y., Chen, J., Gao, T., and Yu, M. (2019). DAG-GNN: DAG Structure Learning with Graph Neural Networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Zhang, K., Zhu, S., Kalander, M., Ng, I., Ye, J., Chen, Z., and Pan, L. (2021). gcastle: A python toolbox for causal discovery. *arXiv:2111.15155*.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). DAGs with NO TEARS: Continuous Optimization for Structure Learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc.

Zheng, Y., Huang, B., Chen, W., Ramsey, J., Gong, M., Cai, R., Shimizu, S., Spirtes, P., and Zhang, K. (2024). Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25:1–8.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

Effective Bayesian Causal Inference via Structural Marginalisation and Autoregressive Orders

A IDENTIFIABILITY OF CAUSAL QUERIES

In Bayesian causal inference, our goal is to infer a query posterior $p(Y \mid \mathcal{D}) = \mathbb{E}_{\mathcal{M} \mid \mathcal{D}}\left[p(Y \mid \mathcal{M})\right]$ given collected (observational) data $\mathcal{D}$ from the underlying true SCM $\mathcal{M}^\star$ (see Section 4). As a consequence of this formulation, the query posterior $p(Y \mid \mathcal{D})$ will match the true query distribution $p(Y \mid \mathcal{M}^\star)$, if the posterior over causal models $p(\mathcal{M} \mid \mathcal{D})$ converges to a point mass on the true SCM $\mathcal{M}^\star$ in the limit on infinitely many data.[7] Therefore, whether a causal query Y (or a query distribution $p(Y \mid \mathcal{M})$) is identifiable (in a statistical rather than a causal sense), depends on our choices of likelihood $p(\mathcal{D} \mid \mathcal{M})$ and prior $p(\mathcal{M})$ over SCMs.

Relating to causal identifiability in our problem setup, recall that we focus on non-linear additive noise models, i.e., we place a GP prior with a Gaussian likelihood on the mechanisms. Identifiability of non-linear additive noise models has been established under mild conditions in (Hoyer et al., 2009; Peters et al., 2014; Bühlmann et al., 2014). Further note that, as we preclude unobserved confounding, causal effects are identifiable given the causal DAG. Thus, in the limit of infinitely many data, our posterior would collapse onto the true SCM and causal effects could be truthfully estimated.

[7] In case the query is fully determined by the SCM, then $p(Y \mid \mathcal{M})$ has a point mass on $Y(\mathcal{M})$ and $p(Y \mid \mathcal{D})$ will converge to a point mass on $Y(\mathcal{M}^\star)$. For example, let $Y(\mathcal{M}) = G$ be the causal graph, then $p(Y \mid \mathcal{D})$ will asymptotically concentrate all its mass on the true graph $G^\star$.

B IMPLEMENTATION AND ESTIMATION DETAILS

Our model is implemented in Python mainly relying on PyTorch (Ansel et al., 2024), and GPyTorch (Gardner et al., 2018) for GP inference. Our code is available at https://github.com/chritoth/bci-arco-gp.

B.1 Mechanism Model

We follow Toth et al. (2022) and use two types of models for our mechanism.

For root nodes, i.e., causal variables without parents, we place a conjugate normal-inverse-gamma prior N-Γ^{-1} on the mean and variance of that node. Specifically,

$$\begin{aligned} p(f_i, \sigma_i^2 \,|\, \mathbf{Pa}_i = \varnothing) &= \text{N-}\Gamma^{-1}(f_i, \sigma_i^2 \,|\, \mu_0, \kappa_0^{-1}, \alpha_0, \beta_0) \\ &= \mathrm{N}(f_i \,|\, \mu_0, \sigma_i^2 \cdot \kappa_0^{-1}) \cdot \Gamma^{-1}(\sigma_i^2 \,|\, \alpha_0, \beta_0) \end{aligned}$$

where $\mu_0, \kappa_0^{-1}, \alpha_0, \beta_0$ are fixed hyper-parameters. When sampling ground-truth SCMs we set $\mu_0 = 0$, $\kappa_0 = 1$, $\alpha_0 = 5$ and $\beta_0 = 10$, sample a mean and variance from the prior and keep them fixed thereafter. For the inference with ARCO-GP and DIBS-GP, we use $\mu_0 = 0$, $\kappa_0 = 1$, $\alpha_0 = 10$ and $\beta_0 = 10$. This yields a prior mean of 1 for the variance σ^2, which is sufficiently broad considering that we standardise all training data to zero mean and unit variance. Analytic expressions for the posterior marginal likelihood are found in (Murphy, 2007).

For non-root nodes we place a GP prior on the mechanisms. Specifically, we use rational quadratic (RQ) kernel

$$k_{\mathrm{RQ}}(\mathbf{x_1}, \mathbf{x_2}) = \delta \cdot \left(1 + \frac{1}{2\gamma}(\mathbf{x_1} - \mathbf{x_2})^\top \lambda^{-2} (\mathbf{x_1} - \mathbf{x_2})\right)^{-\gamma}$$

with scaling parameter δ, lengthscale parameter λ and mixing parameter γ to model non-linear mechanisms. We model the additive noise with a Gaussian likelihood with variance σ_i^2. We place Gamma priors $\Gamma(\delta \,|\, \alpha_\delta, \beta_\delta)$, $\Gamma(\lambda \,|\, \alpha_\lambda, \beta_\lambda)$, $\Gamma(\gamma \,|\, \alpha_\gamma, \beta_\gamma)$ and $\Gamma(\sigma_i^2 \,|\, \alpha_\sigma, \beta_\sigma)$ on these parameters.

When generating ground-truth models we set $\alpha_\delta = 100$, $\beta_\delta = 10$, $\alpha_\lambda = 30 \cdot |\mathbf{Pa}_i|$, $\beta_\lambda = 30$, $\alpha_\gamma = 20$, $\beta_\gamma = 10$ and $\alpha_\sigma = 50$, $\beta_\sigma = 50$. For each GP we sample a set of hyper-parameters from their priors and keep them fixed thereafter. Additionally, we sample a function from the GP prior with 50 support points sampled uniform random in range [-10, 10]. For the inference with ARCO-GP and DIBS-GP, we set $\alpha_\delta = 100$, $\beta_\delta = 10$, $\alpha_\lambda = 30 \cdot |\mathbf{Pa}_i|$, $\beta_\lambda = 30$, $\alpha_\gamma = 20$, $\beta_\gamma = 10$ and $\alpha_\sigma = 2$, $\beta_\sigma = 8$ again considering that we standardise all training data to zero mean and unit variance. We train GP hyper-parameters for a maximum of 100 steps with the RMSprop (Hinton, 2012) optimiser with learning rate 0.05.

Alternatively, for ablations in Appendix D we use invertible (sigmoidal) non-linear mechanisms to generate ground-truth models, as proposed by Bühlmann et al. (2014). Concretely, each mechanisms has the form

$$f_i(\mathbf{Pa}_i) = \sum_{X_j \in \mathbf{Pa}_i} \delta_j \cdot \frac{\gamma_j (X_j + \mu_j)}{1 + |\gamma_j (X_j + \mu_j)|} + \epsilon,$$

i.e., the mechanisms are additive and invertible over the individual parents. In our experiments, we sample the parameters from $\delta_j \sim \Gamma(50, 10)$, $\gamma_j \sim \mathcal{U}([-2, -0.5] \cup [0.5, 2])$, $\mu_j \sim \mathcal{U}([-2, 2])$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ where $\sigma^2 \sim \Gamma(50, 50)$.

B.2 ARCO Model

We use a normal prior $\mathcal{N}(\boldsymbol{\theta} \,|\, 0, \sigma^2 I)$ with $\sigma = 10$ over ARCO's neural network parameters $\boldsymbol{\theta}$. The neural network $g_{\boldsymbol{\theta}}$ uses a single hidden layer with 30 neurons and ReLU activation functions (see Section 4.2). We train ARCO for a maximum of 400 gradient steps, using the ADAM (Kingma and Ba, 2015) optimiser with learning rate of 0.01. For the score-function gradient estimator in Equation (6) we use an exponential moving average baseline with decay rate 0.9. We limit the parent set size to a maximum of two parents per node unless stated otherwise. To estimate causal queries (Equation (3)) or ARCO's gradients (Equation (6)), we sample 100 causal orders to approximate the expectation w.r.t. $p(L \,|\, \boldsymbol{\theta})$ and to compute the necessary importance weights respectively (see Appendix B.3).

B.3 Estimation of Importance Weights w^L

To compute the importance weights (re-stated here for convenience; defined in Equation (4))

$$w^L := \frac{\mathbb{E}_{G \mid L}\left[p(\mathcal{D} \mid \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \mid G)\right]}{\mathbb{E}_{L' \mid \boldsymbol{\theta}}\left[\mathbb{E}_{G' \mid L'}\left[p(\mathcal{D} \mid \boldsymbol{\psi}, G') \cdot p(\boldsymbol{\psi} \mid G')\right]\right]}$$

needed in the estimation of queries and gradients (in Equations (3) and (6)), note that $\mathbb{E}_{G \mid L}\left[p(\mathcal{D} \mid \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \mid G)\right]$ can be computed in closed-form by restricting the maximum parent set size (see Section 4.3). As the expectation w.r.t. $p(L' \mid \boldsymbol{\theta})$ in the denominator is intractable, we use the asymptotically unbiased ratio estimator (see e.g., (Robert and Casella, 2004)), i.e., we normalise the weights such that they sum to 1.

B.4 Estimation of Causal Queries

To estimate causal queries $Y = Y(\mathcal{M})$[8], we either

1. estimate the query posterior $p(Y \mid \mathcal{D}) = \mathbb{E}_{\mathcal{M} \mid \mathcal{D}}\left[p(Y \mid \mathcal{M})\right]$ directly via Equation (3),

2. estimate the predictive posterior mean

$$\begin{aligned}\mathbb{E}_{Y \mid \mathcal{D}}[Y] &= \mathbb{E}_{\mathcal{M} \mid \mathcal{D}}\left[\mathbb{E}_{Y \mid \mathcal{M}}[Y]\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D}}\left[\mathbb{E}_{L \mid \boldsymbol{\theta}}\left[w^L \cdot \mathbb{E}_{G \mid L, \boldsymbol{\psi}, \mathcal{D}}\left[\mathbb{E}_{\mathbf{f} \mid \boldsymbol{\psi}, \mathcal{D}}\left[\mathbb{E}_{Y \mid \mathcal{M}}[Y]\right]\right]\right]\right], \end{aligned} \tag{13}$$

3. or we sample from the query posterior $p(Y \mid \mathcal{D})$ as outlined in Section 4 and compute a sample-based metric or estimate an empirical distribution as in Figure 3.

Below, we provide concrete examples for all three cases for better comprehensibility.

Example 1: Marginal Edge Probability. To estimate the probability that the edge $X_i \to X_j$ is present in the true graph, our query $Y(\mathcal{M}) = \mathbb{1}_{\mathcal{M}}\left(X_i \to X_j\right)$ indicates whether the edge is present $Y(\mathcal{M}) = 1$ or absent $Y(\mathcal{M}) = 0$ in the graph of $\mathcal{M}$. Consequently,

$$p(Y = 1 \mid \mathcal{M}) = \mathbb{1}_{\mathcal{M}}\left(X_i \to X_j\right) = \begin{cases} 1 & \text{if} \quad X_i \to X_j \in G \\ 0 & \text{otherwise.} \end{cases}$$

and, starting from Equation (3), we get

$$\begin{aligned} p(Y \mid \mathcal{D}) &= \mathbb{E}_{\mathcal{M} \mid \mathcal{D}}\left[p(Y \mid \mathcal{M})\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D}}\left[\mathbb{E}_{L \mid \boldsymbol{\theta}}\left[w^L \cdot \mathbb{E}_{G \mid L, \boldsymbol{\psi}, \mathcal{D}}\left[\mathbb{E}_{\mathbf{f} \mid \boldsymbol{\psi}, \mathcal{D}}\left[p(Y \mid \mathcal{M})\right]\right]\right]\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D}}\left[\mathbb{E}_{L \mid \boldsymbol{\theta}}\left[w^L \cdot \mathbb{E}_{G \mid L, \boldsymbol{\psi}, \mathcal{D}}\left[\mathbb{E}_{\mathbf{f} \mid \boldsymbol{\psi}, \mathcal{D}}\left[\mathbb{1}_{\mathcal{M}}\left(X_i \to X_j\right)\right]\right]\right]\right] \end{aligned}$$

and since the query $\mathbb{1}_{\mathcal{M}}\left(X_i \to X_j\right)$ does not depend on the mechanisms $\mathbf{f}$

$$= \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D}}\left[\mathbb{E}_{L \mid \boldsymbol{\theta}}\left[w^L \cdot \mathbb{E}_{G \mid L, \boldsymbol{\psi}, \mathcal{D}}\left[\mathbb{1}_{\mathcal{M}}\left(X_i \to X_j\right)\right]\right]\right].$$

As $\mathbb{1}_{\mathcal{M}}\left(X_i \to X_j\right)$ clearly factorises over parent sets, Proposition 4.1 applies to compute the expectation w.r.t. G in closed-form. To approximate the expectation w.r.t. L we compute a Monte-Carlo estimate by sampling causal orders from ARCO with importance weights w^L (computed as in Appendix B.3), given our learned MAP parameters $\boldsymbol{\theta}, \boldsymbol{\psi}$.

[8] We write $Y(\mathcal{M})$ to make explicit that the query is a function of the SCM.

Example 2: Expected Structural Hamming Distance. To compute the expected structural Hamming distance w.r.t. the true graph $G^\star$, the query is $Y(\mathcal{M}) = \mathrm{SHD}(G, G^\star)$ and thus

$$p(Y = y \,|\, \mathcal{M}) = \begin{cases} 1 & \text{if} \quad y = \mathrm{SHD}(G, G^\star) \\ 0 & \text{otherwise.} \end{cases}$$

Note that in this case, $p(Y \,|\, \mathcal{M})$ is a point mass, as $Y(\mathcal{M}) = \mathrm{SHD}(G, G^\star)$ is fully determined by the SCM. Therefore,

$$\mathbb{E}_{Y \,|\, \mathcal{M}}\left[Y\right] = Y(\mathcal{M}) = \mathrm{SHD}(G, G^\star)$$

and the query posterior mean in Equation (13) reduces to

$$\mathbb{E}_{Y \,|\, \mathcal{D}}\left[Y\right] = \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\psi} \,|\, \mathcal{D}}\left[\mathbb{E}_{L \,|\, \boldsymbol{\theta}}\left[w^L \cdot \mathbb{E}_{G \,|\, L, \boldsymbol{\psi}, \mathcal{D}}\left[\mathbb{E}_{\mathbf{f} \,|\, \boldsymbol{\psi}, \mathcal{D}}\left[\mathrm{SHD}(G, G^\star)\right]\right]\right]\right]$$

and since again the query does not depend on the mechanisms $\mathbf{f}$

$$= \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\psi} \,|\, \mathcal{D}}\left[\mathbb{E}_{L \,|\, \boldsymbol{\theta}}\left[w^L \cdot \mathbb{E}_{G \,|\, L, \boldsymbol{\psi}, \mathcal{D}}\left[\mathrm{SHD}(G, G^\star)\right]\right]\right].$$

As the SHD additively decomposes over the parent sets, we can use Proposition 4.2 to compute the inner expectation w.r.t. G in closed form. We approximate the outer expectations as described in Example 1. The above applies analogously for our other structure learning metrics (see Appendix C.2), using either Proposition 4.1, Proposition 4.2 or a sampling-based approximation of the expectation w.r.t. G, depending on the given metric (see Section 4.3).

We use a sampling-based approximation w.r.t. as described in Example 1.

Example 3: Maximum Mean Discrepancy. When evaluating the ability to infer interventional distributions, we compute an empirical estimate of the MMD (see (Gretton et al., 2012)) between an inferred posterior interventional distribution $p(\mathbf{X} \,|\, \mathrm{do}(\mathbf{w}), \mathcal{D})$ and the corresponding ground truth $p(\mathbf{X} \,|\, \mathrm{do}(\mathbf{w}), \mathcal{M}^*)$, i.e.,

$$\mathrm{MMD}\left(p(\mathbf{X} \,|\, \mathrm{do}(\mathbf{w}), \mathcal{D}), p(\mathbf{X} \,|\, \mathrm{do}(\mathbf{w}), \mathcal{M}^*)\right) \approx \mathrm{MMD}\left(\mathbf{x}, \mathbf{x}^*\right),$$

where $\mathbf{x} \sim p(\mathbf{X} \,|\, \mathrm{do}(\mathbf{w}), \mathcal{D})$ and $\mathbf{x}^* \sim p(\mathbf{X} \,|\, \mathrm{do}(\mathbf{w}), \mathcal{M}^*)$ are samples drawn from the respective distributions (see Appendix C.2). In this case, the causal query $Y(\mathcal{M}) = \mathbf{X} \,|\, \mathrm{do}(\mathbf{W} = \mathbf{w})$ is the set of endogenous (random) variables under intervention. Accordingly, the query posterior reads

$$\begin{aligned} p(Y \,|\, \mathcal{D}) &= p(\mathbf{X} \,|\, \mathrm{do}(\mathbf{w}), \mathcal{D}) \\ &= \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\psi} \,|\, \mathcal{D}}\left[\mathbb{E}_{L \,|\, \boldsymbol{\theta}}\left[w^L \cdot \mathbb{E}_{G \,|\, L, \boldsymbol{\psi}, \mathcal{D}}\left[\mathbb{E}_{\mathbf{f} \,|\, \boldsymbol{\psi}, \mathcal{D}}\left[p(\mathbf{X} \,|\, \mathrm{do}(\mathbf{w}), \mathbf{f}, \boldsymbol{\psi}, \mathcal{D})\right]\right]\right]\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\psi} \,|\, \mathcal{D}}\left[\mathbb{E}_{L \,|\, \boldsymbol{\theta}}\left[w^L \cdot \mathbb{E}_{G \,|\, L, \boldsymbol{\psi}, \mathcal{D}}\left[p(\mathbf{X} \,|\, \mathrm{do}(\mathbf{w}), \boldsymbol{\psi}, \mathcal{D})\right]\right]\right]. \end{aligned}$$

We can draw samples from this expression as per Algorithm 1, i.e., for our learned MAP parameters $\boldsymbol{\theta}, \boldsymbol{\psi}$, we sample causal orders from $p(L \,|\, \boldsymbol{\theta})$, then graphs given orders from $p(G \,|\, L, \boldsymbol{\psi}, \mathcal{D})$ and finally samples from $p(\mathbf{X} \,|\, \mathrm{do}(\mathbf{w}), \boldsymbol{\psi}, \mathcal{D})$ by sampling from the respective GP predictive posteriors (see Section 4.4). The samples corresponding to each sampled order L are weighted with their respective importance weights w^L, computed as in Appendix B.3.

B.5 Kernel Density Estimates of Interventional Distributions.

To illustrate empirical distributions in Figure 3, we compute a kernel density estimate (KDE) given (weighted) samples drawn from the (learned) distributions. We draw 100 causal orders, with 10 graphs per order and 10 samples per graph, equalling 10000 samples from the posterior interventional distribution in total. From the ground-truth model, we draw 10000 samples as well. We use the KDE implementation provided by scikit-learn (Pedregosa et al., 2011), with a Gaussian kernel and a bandwidth of 0.2.

C EXPERIMENTAL SETUP

C.1 Sampling of Ground-Truth SCMs.

To sample ground truth SCMs we draw graph structures from either the Erdös-Rényi (ER) (Erdös and Rényi, 1959) or the scale-free (SF) (Barabási and Albert, 1999) random models commonly used in DAG structure learning benchmarks. Per default, we follow the setup of Lorch et al. (2021) and generate scale-free and Erdös-Rényi graphs with an expected degree of 2. Specifically, this will yield SF graphs with a maximum parent set size of 2, whereas no such restriction applies to ER graphs in general. For ablations in Appendix D we also generate graphs with expected degree in $\{4, 8\}$. We instantiate the causal mechanisms for non-linear additive noise models using GPs as described in Appendix B.

For each ground truth SCM, we sample a fixed set of N training data from the observational distribution. Importantly, Reisach et al. (2021) argue, that the causal order may strongly correlate with increasing marginal variance in simulated data, and therefore, benchmarks may be easy to game by exploiting this property. As this may be especially relevant to order-based inference methods, we follow their recommendation and normalise the training data for each endogenous variable to zero mean and unit variance.

C.2 Metrics

As metrics for learning the causal structure we report the *expected structural Hamming distance* (ESHD), as well as the *area under the receiver operating characteristic* (AUROC) and the *area under the precision recall curve* (AUPRC) w.r.t. posterior edge prediction as commonly reported metrics. To provide additional insight into the methods' strengths and weaknesses, we also report the expected number of edges (#Edges), the *true positive rate* (TPR) and the *true negative rate* (TNR) for the edge prediction task. We do **not** report the log-likelihood on held-out test data as is sometimes reported, because the evaluated methods implement different noise models and approximate inference schemes, which leads to the (marginal) log-likelihoods being uncalibrated and thus incomparable (cf. (Murphy, 2023)[Sec. 7.5] and references therein).

To evaluate the inferred causal structures w.r.t. their causal implications, Henckel et al. (2024) recently proposed a family of causal distances called *Adjustment Identification Distance* (AID). Briefly summarised, the AID counts the number of correctly identified adjustment sets for pairwise causal effects w.r.t. a target graph, where the adjustment sets are determined according to a chosen identification strategy. We adopt the three variants proposed by Henckel et al. (2024), namely *Ancestor* adjustment (A-AID), *Parent* adjustment (P-AID), and *Optimal* adjustment (OSET-AID). These metrics can also be applied between different graph types, e.g., comparing the AID between a predicted CPDAG and a reference DAG. Note that the (P-AID) computed between two DAGs is equivalent to the well-known Structural Identification Distance proposed by Peters and Bühlmann (2015).

To evaluate the inference of interventional distributions, we report the *maximum mean discrepancy* MMD (Gretton et al., 2012) between an inferred posterior interventional distribution $p(\mathbf{X} \mid \mathrm{do}(\mathbf{w}), \mathcal{D})$ and the corresponding ground truth $p(\mathbf{X} \mid \mathrm{do}(\mathbf{w}), \mathcal{M}^*)$. As the MMD is intractable in general, we compute an empirical estimate

$$\mathrm{MMD}^2(\mathbf{x}, \mathbf{x}^*) = \sum_i^N \sum_{j \neq i}^N \frac{w_i w_j}{\sum_{k \neq i} w_k} k(\mathbf{x}_i, \mathbf{x}_j) - \frac{2}{M} \sum_i^N \sum_j^M w_i \cdot k(\mathbf{x}_i, \mathbf{x}_j^*) + \frac{1}{M(M-1)} \sum_i^M \sum_{i \neq j}^M k(\mathbf{x}_i^*, \mathbf{x}_j^*)$$

where $\mathbf{x} = \{\mathbf{x}_n \sim p(\mathbf{X} \mid \mathrm{do}(\mathbf{w}), \mathcal{D})\}_{n=1}^N$ are samples from our learned model with associated weights w_n s.t. $\sum_n w_n = 1$, and $\mathbf{x}^* = \{\mathbf{x}_m \sim p(\mathbf{X} \mid \mathrm{do}(\mathbf{w}), \mathcal{M}^*)\}_{m=1}^M$ are samples from the true model. We use a Gaussian kernel

$$k(\mathbf{x}, \mathbf{x}') = \delta \cdot \exp\left(-\frac{(\mathbf{x} - \mathbf{x}')^T (\mathbf{x} - \mathbf{x}')}{2\gamma}\right)$$

with scaling parameter $\delta = 1000$ and lengthscale $\gamma = 0.2$. The parameters where choosen empirically, see also Appendix B.5.

Additionally, we compute the L1 and L2 error norms

$$||\mathbb{E}_{\mathbf{X} \mid \mathrm{do}(\mathbf{w}), \mathcal{D}}[\mathbf{X}] - \mathbb{E}_{\mathbf{X} \mid \mathrm{do}(\mathbf{w}), \mathcal{M}^*}[\mathbf{X}]||_p \quad \text{for} \quad p \in \{1, 2\},$$

between the inferred posterior mean and the true mean of an interventional distribution.

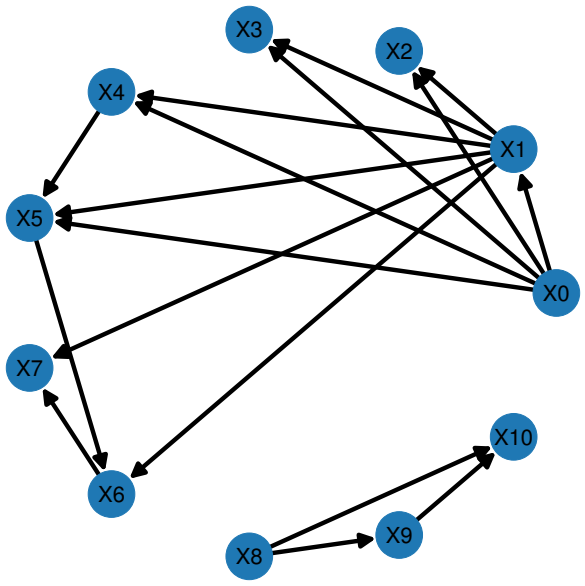

Figure 4: **Sachs Graph.** Consensus protein interaction graph from (Sachs et al., 2005). We relabeled nodes to avoid misinterpretation of our simulation results. Nodes X0 to X10 correspond to the original labels ['PKC', 'PKA', 'Jnk', 'P38', 'Raf', 'Mek', 'Erk', 'Akt', 'Plcg', 'PIP3', 'PIP2'].

C.3 Baselines.

Structure learning baselines. In the past, a plethora of methods for *causal structure learning* (a.k.a. *causal discovery*) have been proposed (Squires and Uhler, 2022; Glymour et al., 2019; Vowels et al., 2022; Heinze-Deml et al., 2018). To evaluate our method in a wider context, we compare our inference model (ARCO-GP) to a diverse set of ten different structure learning methods from different streams of research. We describe our baselines briefly in the following.

- **BAYESDAG** (Annadani et al., 2023). BayesDAG utilises a mixture of MCMC to infer permutations and mechanism parameters, and Variational Inference (VI) to infer DAG edges given the permutations. We use the implementation provided by Annadani et al. (2023) at https://github.com/microsoft/Project-BayesDAG. We needed to adapt the sparsity regularisation hyper-parameter in order to get meaningful results and ran our experiments with the configurations in Listings 2-3. As the implementation runs a number of MCMC chains and only evaluates the best chain afterwards, we use only one MCMC chain to enable a fair comparison, as multiple chains would correspond to multiple runs of the other methods.
- **DAG-GNN** (Yu et al., 2019). DAG-GNN is a gradient-based structure learning approach combining graph neural networks an acyclicity constraint similar to Zheng et al. (2018). We use the implementation provided by Zhang et al. (2021)[Version 1.0.3] using default settings.
- **DDS** (Charpentier et al., 2022). DDS builds upon the permutation-based approach of Cundy et al. (2021) and utilises differentiable permutation sampling and VI to infer a posterior over DAGs. We use the implementation provided by Charpentier et al. (2022) at https://github.com/sharpenb/Differentiable-DAG-Sampling. We needed to adapt the default hyper-parameters in order to get meaningful results and ran our experiments with the configuration displayed in Listing 1.
- **DIBS-GP** (Toth et al., 2022; Lorch et al., 2021). DIBS-GP is a Bayesian causal inference framework recently proposed by Toth et al. (2022) utilising the differential structure learning method by Lorch et al. (2021) employing a soft acyclicity constraint in line with Zheng et al. (2018); Yu et al. (2019). We use the implementation provided by Toth et al. (2022) at https://github.com/chritoth/active-bayesian-causal-inference. We use their standard parameters with 10 latent particles and constant hyper-parameters $\alpha = \beta = 1$

as described in (Lorch et al., 2021). For each latent particle we sample 100 graphs to estimate gradients and causal quantities.

- **GADGET** (Viinikka et al., 2020). GADGET is a Bayesian structure learning method for linear Gaussian models based on MCMC and structure marginalisation over capped-size parent sets. We use the implementation provided by (Viinikka et al., 2020) at https://github.com/Sums-of-Products/sumu and parameters as shown in Listing 4.

- **GES** (Chickering, 2003). GES is a well-known greedy score-based method for causal discovery using the BIC score. We use the implementation provided by Zhang et al. (2021)[Version 1.0.3] using default settings.

- **GOLEM** (Ng et al., 2020). GOLEM is a differentiable DAG structure learning similar to Zheng et al. (2018) but with a likelihood-based score function for linear models. We use the implementation provided by Zhang et al. (2021)[Version 1.0.3] using default settings.

- **GRASP** (Lam et al., 2022). We compare against GRASP as a state-of-the-art representative of a recent line of work of permutation-based approaches to causal discovery building upon the sparsest permutation principle (Raskutti and Uhler, 2018; Solus et al., 2021). We use the implementation provided by Zheng et al. (2024) using default settings.

- **RESIT** (Peters et al., 2014). RESIT is designed for causal discovery on additive noise models. It combines regression with independence tests on the regression residuals to first construct a causal order resp. a fully connected DAG, which it then uses to interatively remove edges to obtain the causal graph. As a regression model for RESIT, we use GPs with the same set of hyper-parameters and training procedure as ARCO-GP to provide better comparability.

- **PC** (Spirtes et al., 2000). The PC algorithm is another standard causal discovery methods based on conditional independence tests. We use the implementation provided by Zhang et al. (2021)[Version 1.0.3] using default settings. It happens quite frequently that PC return a possibly cyclic PDAG, for which the AID metrics cannot computed and are thus omited in our experimental results.

BCI baselines. There are few prior works on BCI and baselines for non-linear additive noise models that learn a full generative model (i.e., graph, mechanisms and distribution parameters of the exogenous variables) and that provide the ability to sample from interventional distributions or estimate average causal effects are scarce (see Section 2). As this is the case for all our structure learning baselines except DIBS-GP, we need to use a different set of baselines when evaluating ARCO-GPs ability to infer interventional distributions and (average) causal effects.

We compare to

- **DIBS-GP** as a baseline operating in the same problem scenario (see description above),

- **RESIT-GP**, as a baseline emulating a traditional two-stage approach to causal inference: we first infer a causal DAG using RESIT (see above) and then fit GP mechanisms with the same setup as for ARCO-GP to provide better comparability, and

- **GADET+BEEPS** ((Viinikka et al., 2020)), that we describe below.

GADGET+BEEPS is a BCI method for linear Gaussian models that samples from the posterior over causal graphs using GADGET and infers causal effects as the path-wise accumulation of the linear mechanism weights (BEEPS). As the implementation of GADGET+BEEPS does not naturally support sampling from interventional distributions or the estimation of interventional distribution means, we extended the implementation in the following way.

A linear SCM can be written as

$$\mathbf{X} = (\mathbf{G} \odot \mathbf{B})^T \cdot \mathbf{X} + \boldsymbol{\mu} + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ are the endogenous variables, $\mathbf{G}$ is the adjacency matrix of the causal DAG, $\mathbf{B}$ is the weight matrix of the linear edge weights, $\odot$ denotes the element-wise multiplication, $\boldsymbol{\mu}$ denotes the mean vector of the variables and $\boldsymbol{\epsilon}$ denotes the exogenous Gaussian noise. Solving for $\mathbf{X}$ yields

$$\mathbf{X} = (\mathbf{I} - (\mathbf{G} \odot \mathbf{B})^T)^{-1}(\boldsymbol{\mu} + \boldsymbol{\epsilon}),$$

and taking the expectation w.r.t. ϵ yields

$$\mathbb{E}_\epsilon[\mathbf{X}] = (\mathbf{I} - (\mathbf{G} \odot \mathbf{B})^T)^{-1}\boldsymbol{\mu}. \tag{14}$$

We then estimate the interventional distribution mean $\mathbb{E}_{\mathbf{X} \mid do(X_j = x_j)}[\mathbf{X}]$ by (i) drawing posterior DAGs G and weight matrices $\mathbf{B}$ using GADGET+BEEPS, (ii) performing the intervention by removing all parents of X_j in the sampled DAGs and setting the posterior mean $\mu_j = x_j$, and (iii) finally solving Equation (14) to get the desired estimate. We do not extend GADGET+BEEPS to be able to sample from interventional distributions.

Listing 1: DDS Parameter Set.

```
# Architecture parameters
'seed_model': 123,  # Seed to init model.
'ma_hidden_dims': [32, 32, 32],  # Output dimension.
'ma_architecture': 'linear',  # Output dimension.
'ma_fast': False,  # Output dimension.
'pd_initial_adj': 'Learned',  # Output dimension.
'pd_temperature': 1.0,  # Output dimension.
'pd_hard': True,  # Output dimension.
'pd_order_type': 'topk',  # Output dimension.
'pd_noise_factor': 1.0,  # Hidden dimensions.

# Training parameters
'max_epochs': 500,  # Maximum number of epochs for training
'patience': 150,  # Patience for early stopping.
'frequency': 2,  # Frequency for early stopping test.
'batch_size': 16,  # Batch size.
'ma_lr': 1e-3,  # Learning rate.
'pd_lr': 1e-2,  # Learning rate.
'loss': 'ELBO',  # Loss name. string
'regr': 1e-1,  # Regularization factor in Bayesian loss.
'prior_p': 1e-6  # Regularization factor in Bayesian loss.
```

Listing 2: BAYESDAG Non-Linear ER Parameter Set.

```
"model_hyperparams": {
    "num_chains": 1,
    "sinkhorn_n_iter": 3000,
    "scale_noise_p": 0.001,
    "scale_noise": 0.001,
    "VI_norm": true,
    "input_perm": false,
    "lambda_sparse": 10,
    "sparse_init": false
},
"training_hyperparams": {
    "learning_rate": 1e-3,
    "batch_size": 512,
    "stardardize_data_mean": false,
    "stardardize_data_std": false,
    "max_epochs": 500
}
```

Listing 3: BAYESDAG Non-linear SF Parameter Set.

```
"model_hyperparams": {
    "num_chains": 1,
    "sinkhorn_n_iter": 3000,
    "scale_noise_p": 0.001,
    "scale_noise": 0.001,
    "VI_norm": true,
    "input_perm": false,
    "lambda_sparse": 10,
    "sparse_init": false
},
"training_hyperparams": {
    "learning_rate": 1e-3,
    "batch_size": 512,
    "stardardize_data_mean": false,
    "stardardize_data_std": false,
    "max_epochs": 500
}
```

Listing 4: GADGET Parameter Set.

```
"scoref": 'bge',
"max_id": -1,
"K": min(self.num_nodes - 1, 16),
"d": 2,
"cp_algo": 'greedy-lite',
"mc3_chains": 16,
"burn_in": 1000,
"iterations": 1000,
"thinning": 10,
```

D EXTENDED EXPERIMENTAL RESULTS

In this appendix, we present additional experimental results and ablations further evaluating our ARCO-GP method. For these extended experiments we report an extended set of eight additional structure learning metrics (described in Appendix C) compared to our results in Section 5. For details on our experimental setup, e.g., data generation, baselines, etc., we refer to Appendices B and C.

Real-world dataset from Sachs et al. (2005). The dataset consists of 853 observational data. The target consensus graph has 11 nodes and 17 edges. We compare our ARCO-GP method in variants with maximum parent set size $k \in 2, 3$. As the results in Table 2 show, both ARCO-GP variants are competitive with the baselines.

Simulations on larger models with $d = 50$ variables. We report structure learning results for non-linear models with ER and SF graphs over $d = 50$ variables in Table 3. The results of ARCO-GP w.r.t. the baselines are qualitatively similar to the ones achieved on 20 node models, i.e., we outperform all baselines for both graph types. Especially on scale-free graphs for that our assumption on a maximum parent set size of $K = 2$ holds, we achieve almost flawless graph identification, resulting in a significant performance gap to the baselines.

Influence of the maximum parent set size. We first evaluate the influence of the maximum parent set size restriction on the performance of our ARCO-GP method in our default benchmark setup (models with 20 nodes and expected degree of 2). We distinguish the models variants by labels ARCO-Kk-GP for $k \in \{1, 2, 3, 4\}$, meaning we allow a maximum parent set size of k. The results in Table 4 show, that $k = 1$ performs worst as expected, since each node can only have one parent. Not surprisingly, $k = 2$ performs best on scale-free graphs as for these graphs no node has more than two parents. On Erdös-Renyi graphs, $k = 4$ performs best, as the these graphs can have an arbitrary number of parents, which is best reflected by the largest k.

In a second experiment, we keep the maximum parent set size of ARCO-GP fixed to $k \in \{2, 4\}$ and we generate ground-truth models with expected degree in $\{4, 8\}$. Note again, that nodes in a scale-free graph cannot have more than 4 parents, whereas nodes in an Erdös-Renyi graph can have an arbitrary number of parents. The results are listed in Tables 5 and 6. As expected, the performance of ARCO-GP deteriorates as the true graphs get denser and our assumptions are violated. However, we still outperform our baselines on most metrics, albeit with smaller margin. It is interesting to observe, that ARCO-GP maintains a high true negative rate in all cases. We conjecture, that due to the restriction on the number of parents, ARCO-GP infers sub-graphs of the true model.

Invertible non-linear models. We test ARCO-GP on invertible non-linear models with sigmoidal mechanisms (see Appendix C), for which structure identification is more challenging (Bühlmann et al., 2014). The results are listed in Table 7. Although performance decreases compared to models with non-invertible GP mechanisms, ARCO-GP still outperforms the baselines with a clear margin.

Influence of the number of available training data. We investigate the influence of the number of training data $N \in \{100, 200, 500, 1000\}$ for non-linear scale-free models in Tables 8 and 9 and for non-linear Erdös-Renyi models in Tables 10 and 11. For scale-free models, ARCO-GP outperforms all baselines on all metrics in all configurations. For Erdös-Renyi models, ARCO-GP outperforms the baselines on most metrics across the configurations or is competitive.

Table 2: **Real-world dataset from Sachs et al. (2005)).** The target graph has 11 nodes and 17 edges. We report means and 95% confidence intervals (CIs) across 20 different ground truth models. Arrows next to metrics indicate lower is better (↓) and higher is better (↑).

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-K2-GP	7 ± 0	0.22 ± 0.01	0.48 ± 0.02	0.23 ± 0.01
ARCO-K3-GP	7 ± 1	0.22 ± 0.00	0.47 ± 0.02	0.22 ± 0.00
BAYESDAG	24 ± 1	0.33 ± 0.02	0.48 ± 0.02	0.33 ± 0.02
DAG-GNN	7 ± 0	0.23 ± 0.00	0.45 ± 0.00	0.23 ± 0.00
DDS	39 ± 1	0.36 ± 0.01	0.41 ± 0.01	0.36 ± 0.01
DIBS-GP	7 ± 1	0.22 ± 0.01	0.43 ± 0.04	0.22 ± 0.01
GADGET	9 ± 0	0.25 ± 0.00	0.50 ± 0.01	0.25 ± 0.00
GES	8 ± 0	0.28 ± 0.00	0.56 ± 0.00	0.28 ± 0.00
GOLEM	17 ± 0	0.40 ± 0.00	0.49 ± 0.00	0.39 ± 0.00
GRASP	8 ± 0	0.28 ± 0.00	0.56 ± 0.00	0.28 ± 0.00
PC	8 ± 0	–	–	–

(a) **AID Metrics.**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-K2-GP	17 ± 1	0.56 ± 0.02	0.30 ± 0.02	0.20 ± 0.03	0.96 ± 0.01
ARCO-K3-GP	17 ± 0	0.55 ± 0.03	0.31 ± 0.02	0.21 ± 0.02	0.96 ± 0.00
BAYESDAG	34 ± 2	0.49 ± 0.04	0.18 ± 0.03	0.22 ± 0.06	0.78 ± 0.01
DAG-GNN	18 ± 0	0.57 ± 0.00	0.37 ± 0.00	0.18 ± 0.00	0.96 ± 0.00
DDS	45 ± 2	0.48 ± 0.01	0.19 ± 0.02	0.33 ± 0.00	0.64 ± 0.02
DIBS-GP	16 ± 1	0.61 ± 0.02	0.36 ± 0.04	0.24 ± 0.04	0.97 ± 0.01
GADGET	20 ± 0	0.63 ± 0.03	0.27 ± 0.02	0.19 ± 0.01	0.94 ± 0.00
GES	19 ± 0	0.63 ± 0.00	0.44 ± 0.00	0.35 ± 0.00	0.91 ± 0.00
GOLEM	26 ± 0	0.55 ± 0.00	0.29 ± 0.00	0.24 ± 0.00	0.86 ± 0.00
GRASP	19 ± 0	0.63 ± 0.00	0.44 ± 0.00	0.35 ± 0.00	0.91 ± 0.00
PC	19 ± 0	0.63 ± 0.00	0.44 ± 0.00	0.35 ± 0.00	0.91 ± 0.00

(b) **Edge prediction metrics.**

Table 3: **Benchmarks on systems with 50 variables.** Ablation studies on simulated non-linear ground truth models with 50 nodes and different DAG structures. We report means and 95% confidence intervals (CIs) across 20 different ground truth models. Arrows next to metrics indicate lower is better (↓) and higher is better (↑).

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-GP	91 ± 1	0.03 ± 0.01	0.06 ± 0.02	0.04 ± 0.02
BAYESDAG	237 ± 5	0.25 ± 0.02	0.57 ± 0.05	0.25 ± 0.02
DAG-GNN	37 ± 4	0.14 ± 0.01	0.94 ± 0.02	0.14 ± 0.01
DDS	1082 ± 26	0.39 ± 0.01	0.42 ± 0.01	0.39 ± 0.01
DIBS-GP	70 ± 7	0.14 ± 0.01	0.82 ± 0.04	0.14 ± 0.01
GADGET	176 ± 5	0.31 ± 0.02	0.75 ± 0.02	0.31 ± 0.02
GES	220 ± 11	0.38 ± 0.05	0.71 ± 0.05	0.39 ± 0.05
GOLEM	56 ± 5	0.16 ± 0.02	0.88 ± 0.03	0.17 ± 0.02
GRASP	154 ± 8	0.30 ± 0.02	0.71 ± 0.03	0.31 ± 0.02
PC	108 ± 3	–	–	–

(a) **Scale-free Nonlinear (Part 1/2).**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-GP	6 ± 2	0.98 ± 0.01	0.95 ± 0.03	0.94 ± 0.02	1.00 ± 0.00
BAYESDAG	220 ± 11	0.77 ± 0.03	0.40 ± 0.03	0.59 ± 0.04	0.92 ± 0.00
DAG-GNN	90 ± 4	0.61 ± 0.02	0.42 ± 0.03	0.22 ± 0.03	0.99 ± 0.00
DDS	1037 ± 25	0.79 ± 0.03	0.52 ± 0.05	0.73 ± 0.03	0.57 ± 0.01
DIBS-GP	123 ± 7	0.61 ± 0.02	0.27 ± 0.03	0.22 ± 0.03	0.98 ± 0.00
GADGET	191 ± 7	0.79 ± 0.02	0.30 ± 0.03	0.42 ± 0.02	0.94 ± 0.00
GES	229 ± 13	0.69 ± 0.02	0.34 ± 0.03	0.46 ± 0.05	0.92 ± 0.00
GOLEM	94 ± 6	0.65 ± 0.02	0.43 ± 0.04	0.30 ± 0.03	0.99 ± 0.00
GRASP	160 ± 9	0.72 ± 0.01	0.41 ± 0.02	0.49 ± 0.02	0.95 ± 0.00
PC	156 ± 5	0.61 ± 0.01	0.25 ± 0.02	0.25 ± 0.02	0.96 ± 0.00

(b) **Scale-free Nonlinear (Part 2/2).**

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-GP	48 ± 3	0.09 ± 0.01	0.26 ± 0.02	0.09 ± 0.01
BAYESDAG	195 ± 4	0.17 ± 0.03	0.48 ± 0.06	0.18 ± 0.03
DAG-GNN	34 ± 3	0.14 ± 0.01	0.57 ± 0.04	0.14 ± 0.01
DDS	928 ± 30	0.37 ± 0.03	0.45 ± 0.03	0.38 ± 0.03
DIBS-GP	33 ± 8	0.12 ± 0.01	0.43 ± 0.04	0.13 ± 0.01
GADGET	121 ± 4	0.25 ± 0.02	0.50 ± 0.05	0.26 ± 0.02
GES	134 ± 7	0.25 ± 0.03	0.45 ± 0.04	0.27 ± 0.03
GOLEM	52 ± 5	0.16 ± 0.02	0.54 ± 0.03	0.16 ± 0.01
GRASP	92 ± 5	0.18 ± 0.02	0.40 ± 0.05	0.19 ± 0.02
PC	92 ± 3	–	–	–

(c) **Erdös-Rényi Nonlinear (Part 1/2).**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-GP	56 ± 4	0.77 ± 0.02	0.57 ± 0.04	0.46 ± 0.03	1.00 ± 0.00
BAYESDAG	213 ± 9	0.68 ± 0.02	0.29 ± 0.02	0.41 ± 0.04	0.93 ± 0.00
DAG-GNN	110 ± 4	0.55 ± 0.01	0.25 ± 0.03	0.12 ± 0.01	0.99 ± 0.00
DDS	900 ± 26	0.75 ± 0.03	0.44 ± 0.05	0.64 ± 0.04	0.63 ± 0.01
DIBS-GP	105 ± 5	0.56 ± 0.02	0.30 ± 0.04	0.13 ± 0.03	0.99 ± 0.00
GADGET	130 ± 4	0.84 ± 0.01	0.46 ± 0.03	0.45 ± 0.02	0.97 ± 0.00
GES	122 ± 9	0.77 ± 0.02	0.51 ± 0.03	0.58 ± 0.03	0.97 ± 0.00
GOLEM	111 ± 4	0.59 ± 0.01	0.31 ± 0.03	0.20 ± 0.02	0.99 ± 0.00
GRASP	80 ± 6	0.79 ± 0.01	0.61 ± 0.03	0.60 ± 0.03	0.98 ± 0.00
PC	112 ± 5	0.69 ± 0.01	0.43 ± 0.02	0.40 ± 0.02	0.98 ± 0.00

(d) **Erdös-Rényi Nonlinear (Part 2/2).**

Table 4: **Varying the maximal parent set size.** Ablation studies on simulated (non-)linear ground truth models with 20 nodes. We report means and 95% confidence intervals (CIs) across 20 different ground truth models. Arrows next to metrics indicate lower is better (↓) and higher is better (↑).

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-K1-GP	10 ± 1	0.17 ± 0.02	0.82 ± 0.02	0.19 ± 0.02
ARCO-K2-GP	32 ± 1	0.06 ± 0.02	0.13 ± 0.04	0.07 ± 0.02
ARCO-K3-GP	35 ± 1	0.09 ± 0.04	0.15 ± 0.04	0.10 ± 0.04
ARCO-K4-GP	37 ± 1	0.08 ± 0.03	0.16 ± 0.05	0.10 ± 0.03

(a) **Scale-free Nonlinear (Part 1/2).**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-K1-GP	26 ± 1	0.86 ± 0.02	0.67 ± 0.02	0.28 ± 0.03	1.00 ± 0.00
ARCO-K2-GP	5 ± 2	0.97 ± 0.02	0.94 ± 0.03	0.88 ± 0.04	1.00 ± 0.00
ARCO-K3-GP	9 ± 2	0.94 ± 0.03	0.89 ± 0.04	0.87 ± 0.04	0.99 ± 0.00
ARCO-K4-GP	11 ± 2	0.94 ± 0.02	0.90 ± 0.04	0.86 ± 0.04	0.98 ± 0.00

(b) **Scale-free Nonlinear (Part 2/2).**

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-K1-GP	8 ± 1	0.21 ± 0.03	0.50 ± 0.03	0.21 ± 0.02
ARCO-K2-GP	20 ± 2	0.15 ± 0.02	0.32 ± 0.04	0.18 ± 0.03
ARCO-K3-GP	29 ± 2	0.12 ± 0.02	0.23 ± 0.03	0.16 ± 0.03
ARCO-K4-GP	32 ± 3	0.10 ± 0.03	0.21 ± 0.04	0.15 ± 0.04

(c) **Erdös-Rényi Nonlinear (Part 1/2).**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-K1-GP	32 ± 3	0.79 ± 0.03	0.54 ± 0.04	0.19 ± 0.03	1.00 ± 0.00
ARCO-K2-GP	21 ± 3	0.84 ± 0.03	0.68 ± 0.05	0.49 ± 0.05	1.00 ± 0.00
ARCO-K3-GP	17 ± 2	0.87 ± 0.02	0.76 ± 0.04	0.65 ± 0.05	0.99 ± 0.00
ARCO-K4-GP	17 ± 3	0.88 ± 0.03	0.79 ± 0.04	0.70 ± 0.05	0.99 ± 0.00

(d) **Erdös-Rényi Nonlinear (Part 2/2).**

Table 5: **Benchmarks on models with average degree 4.** Ablation studies on simulated non-linear ground truth models with 20 nodes scale-free and Erdös-Renyi DAG structures with average degree 4. We report means and 95% confidence intervals (CIs) across 20 different ground truth models. Arrows next to metrics indicate lower is better (↓) and higher is better (↑).

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-K2-GP	24 ± 1	0.30 ± 0.02	0.75 ± 0.03	0.34 ± 0.02
ARCO-K4-GP	53 ± 1	0.23 ± 0.04	0.41 ± 0.06	0.34 ± 0.03
BAYESDAG	65 ± 2	0.48 ± 0.02	0.81 ± 0.02	0.48 ± 0.02
DAG-GNN	22 ± 4	0.40 ± 0.02	0.92 ± 0.01	0.41 ± 0.02
DDS	165 ± 5	0.63 ± 0.01	0.68 ± 0.01	0.63 ± 0.01
DIBS-GP	38 ± 5	0.36 ± 0.02	0.71 ± 0.04	0.39 ± 0.02
GADGET	54 ± 1	0.54 ± 0.02	0.86 ± 0.01	0.55 ± 0.02
GES	63 ± 2	0.59 ± 0.03	0.85 ± 0.02	0.60 ± 0.03
GOLEM	32 ± 2	0.45 ± 0.02	0.90 ± 0.02	0.45 ± 0.02
GRASP	44 ± 1	0.54 ± 0.03	0.89 ± 0.02	0.55 ± 0.03
RESIT	21 ± 1	0.48 ± 0.03	0.96 ± 0.01	0.47 ± 0.03
PC	31 ± 1	–	–	–

(a) **Scale-free Nonlinear (Part 1/2).**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-K2-GP	47 ± 2	0.76 ± 0.02	0.57 ± 0.04	0.32 ± 0.02	0.99 ± 0.00
ARCO-K4-GP	26 ± 4	0.87 ± 0.02	0.80 ± 0.04	0.71 ± 0.04	0.98 ± 0.01
BAYESDAG	77 ± 3	0.67 ± 0.02	0.41 ± 0.03	0.41 ± 0.03	0.88 ± 0.01
DAG-GNN	62 ± 2	0.58 ± 0.01	0.44 ± 0.03	0.18 ± 0.02	0.97 ± 0.01
DDS	159 ± 5	0.64 ± 0.03	0.37 ± 0.04	0.55 ± 0.04	0.59 ± 0.01
DIBS-GP	65 ± 4	0.62 ± 0.03	0.40 ± 0.04	0.29 ± 0.05	0.94 ± 0.01
GADGET	75 ± 3	0.72 ± 0.02	0.43 ± 0.03	0.33 ± 0.02	0.90 ± 0.01
GES	86 ± 5	0.60 ± 0.02	0.38 ± 0.03	0.33 ± 0.03	0.86 ± 0.01
GOLEM	61 ± 3	0.61 ± 0.01	0.47 ± 0.03	0.27 ± 0.03	0.95 ± 0.01
GRASP	68 ± 4	0.63 ± 0.01	0.45 ± 0.03	0.33 ± 0.02	0.92 ± 0.01
RESIT	80 ± 2	0.49 ± 0.01	0.16 ± 0.03	0.04 ± 0.01	0.94 ± 0.00
PC	73 ± 3	0.56 ± 0.01	0.34 ± 0.03	0.18 ± 0.02	0.94 ± 0.01

(b) **Scale-free Nonlinear (Part 2/2).**

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-K2-GP	20 ± 2	0.33 ± 0.02	0.65 ± 0.04	0.37 ± 0.02
ARCO-K4-GP	44 ± 2	0.29 ± 0.04	0.52 ± 0.05	0.39 ± 0.03
BAYESDAG	55 ± 2	0.44 ± 0.04	0.77 ± 0.04	0.46 ± 0.03
DAG-GNN	19 ± 4	0.46 ± 0.03	0.93 ± 0.02	0.46 ± 0.03
DDS	166 ± 5	0.67 ± 0.01	0.72 ± 0.01	0.67 ± 0.01
DIBS-GP	30 ± 5	0.42 ± 0.02	0.75 ± 0.03	0.45 ± 0.02
GADGET	53 ± 3	0.55 ± 0.02	0.85 ± 0.02	0.56 ± 0.02
GES	62 ± 4	0.60 ± 0.04	0.85 ± 0.03	0.61 ± 0.04
GOLEM	29 ± 3	0.52 ± 0.02	0.93 ± 0.02	0.51 ± 0.02
GRASP	44 ± 3	0.57 ± 0.04	0.87 ± 0.02	0.59 ± 0.03
RESIT	20 ± 1	0.51 ± 0.02	0.91 ± 0.03	0.52 ± 0.02
PC	32 ± 1	–	–	–

(c) **Erdös-Rényi Nonlinear (Part 1/2).**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-K2-GP	65 ± 4	0.72 ± 0.03	0.54 ± 0.03	0.23 ± 0.02	0.99 ± 0.00
ARCO-K4-GP	50 ± 4	0.79 ± 0.03	0.69 ± 0.04	0.46 ± 0.04	0.98 ± 0.00
BAYESDAG	82 ± 5	0.65 ± 0.02	0.44 ± 0.03	0.34 ± 0.03	0.91 ± 0.01
DAG-GNN	84 ± 4	0.53 ± 0.01	0.34 ± 0.04	0.10 ± 0.03	0.96 ± 0.01
DDS	157 ± 6	0.63 ± 0.03	0.39 ± 0.03	0.55 ± 0.04	0.60 ± 0.01
DIBS-GP	85 ± 6	0.55 ± 0.02	0.33 ± 0.04	0.16 ± 0.03	0.94 ± 0.01
GADGET	81 ± 5	0.72 ± 0.02	0.50 ± 0.04	0.33 ± 0.03	0.91 ± 0.01
GES	86 ± 6	0.62 ± 0.02	0.48 ± 0.03	0.36 ± 0.03	0.89 ± 0.01
GOLEM	84 ± 4	0.55 ± 0.01	0.39 ± 0.03	0.16 ± 0.02	0.95 ± 0.01
GRASP	79 ± 6	0.62 ± 0.02	0.49 ± 0.03	0.31 ± 0.03	0.93 ± 0.01
RESIT	97 ± 3	0.48 ± 0.01	0.17 ± 0.02	0.03 ± 0.01	0.94 ± 0.00
PC	84 ± 4	0.56 ± 0.01	0.41 ± 0.03	0.19 ± 0.02	0.94 ± 0.01

(d) **Erdös-Rényi Nonlinear (Part 2/2).**

Table 6: **Benchmarks on models with average degree 8.** Ablation studies on simulated non-linear ground truth models with 20 nodes scale-free and Erdös-Renyi DAG structures with average degree 8. We report means and 95% confidence intervals (CIs) across 20 different ground truth models. Arrows next to metrics indicate lower is better (↓) and higher is better (↑).

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-K2-GP	20 ± 2	0.37 ± 0.02	0.61 ± 0.04	0.38 ± 0.02
ARCO-K4-GP	42 ± 3	0.32 ± 0.02	0.58 ± 0.03	0.41 ± 0.03
BAYESDAG	62 ± 3	0.54 ± 0.02	0.83 ± 0.02	0.54 ± 0.02
DAG-GNN	21 ± 4	0.43 ± 0.02	0.87 ± 0.03	0.43 ± 0.01
DDS	169 ± 6	0.69 ± 0.01	0.72 ± 0.01	0.69 ± 0.01
DIBS-GP	45 ± 7	0.36 ± 0.02	0.55 ± 0.04	0.42 ± 0.01
GADGET	61 ± 3	0.62 ± 0.02	0.87 ± 0.01	0.62 ± 0.02
GES	71 ± 3	0.65 ± 0.02	0.85 ± 0.02	0.65 ± 0.02
GOLEM	33 ± 3	0.47 ± 0.03	0.85 ± 0.03	0.46 ± 0.02
GRASP	52 ± 3	0.61 ± 0.02	0.87 ± 0.02	0.61 ± 0.02
RESIT	22 ± 1	0.54 ± 0.02	0.96 ± 0.01	0.54 ± 0.02
PC	31 ± 2	–	–	–

(a) **Scale-free Nonlinear (Part 1/2).**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-K2-GP	83 ± 2	0.62 ± 0.05	0.47 ± 0.04	0.17 ± 0.02	0.99 ± 0.00
ARCO-K4-GP	78 ± 4	0.72 ± 0.04	0.57 ± 0.05	0.31 ± 0.03	0.96 ± 0.01
BAYESDAG	103 ± 4	0.59 ± 0.02	0.41 ± 0.02	0.28 ± 0.02	0.88 ± 0.01
DAG-GNN	95 ± 2	0.54 ± 0.01	0.42 ± 0.03	0.11 ± 0.03	0.96 ± 0.01
DDS	173 ± 8	0.56 ± 0.03	0.35 ± 0.03	0.48 ± 0.04	0.57 ± 0.02
DIBS-GP	87 ± 4	0.64 ± 0.03	0.50 ± 0.04	0.28 ± 0.05	0.94 ± 0.01
GADGET	111 ± 4	0.62 ± 0.02	0.37 ± 0.03	0.24 ± 0.02	0.87 ± 0.01
GES	117 ± 5	0.55 ± 0.01	0.41 ± 0.02	0.27 ± 0.02	0.84 ± 0.01
GOLEM	95 ± 2	0.56 ± 0.01	0.45 ± 0.02	0.18 ± 0.02	0.94 ± 0.01
GRASP	104 ± 4	0.56 ± 0.02	0.43 ± 0.03	0.24 ± 0.03	0.89 ± 0.01
RESIT	112 ± 2	0.48 ± 0.01	0.20 ± 0.02	0.03 ± 0.01	0.93 ± 0.00
PC	106 ± 2	0.52 ± 0.01	0.34 ± 0.02	0.12 ± 0.01	0.93 ± 0.00

(b) **Scale-free Nonlinear (Part 2/2).**

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-K2-GP	20 ± 2	0.45 ± 0.02	0.81 ± 0.02	0.48 ± 0.02
ARCO-K4-GP	46 ± 4	0.46 ± 0.03	0.73 ± 0.04	0.53 ± 0.03
BAYESDAG	60 ± 2	0.59 ± 0.02	0.87 ± 0.01	0.60 ± 0.02
DAG-GNN	22 ± 4	0.53 ± 0.01	0.94 ± 0.02	0.53 ± 0.01
DDS	49 ± 7	0.58 ± 0.03	0.85 ± 0.02	0.60 ± 0.02
DIBS-GP	161 ± 7	0.73 ± 0.01	0.78 ± 0.01	0.73 ± 0.00
GADGET	64 ± 3	0.69 ± 0.01	0.91 ± 0.01	0.69 ± 0.01
GES	74 ± 2	0.71 ± 0.02	0.89 ± 0.01	0.71 ± 0.02
GOLEM	34 ± 4	0.56 ± 0.02	0.93 ± 0.02	0.56 ± 0.02
GRASP	54 ± 2	0.70 ± 0.03	0.92 ± 0.01	0.70 ± 0.02
RESIT	20 ± 1	0.58 ± 0.02	0.97 ± 0.02	0.58 ± 0.02
PC	33 ± 1	–	–	–

(c) **Erdös-Rényi Nonlinear (Part 1/2).**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-K2-GP	147 ± 3	0.61 ± 0.04	0.58 ± 0.04	0.10 ± 0.01	0.98 ± 0.00
ARCO-K4-GP	137 ± 5	0.64 ± 0.04	0.61 ± 0.04	0.22 ± 0.03	0.95 ± 0.01
BAYESDAG	149 ± 5	0.58 ± 0.02	0.54 ± 0.03	0.22 ± 0.02	0.89 ± 0.01
DAG-GNN	156 ± 3	0.52 ± 0.01	0.53 ± 0.04	0.08 ± 0.02	0.96 ± 0.01
DDS	166 ± 5	0.47 ± 0.03	0.43 ± 0.04	0.14 ± 0.03	0.88 ± 0.02
DIBS-GP	172 ± 11	0.56 ± 0.04	0.50 ± 0.04	0.47 ± 0.04	0.61 ± 0.03
GADGET	162 ± 4	0.55 ± 0.02	0.48 ± 0.02	0.19 ± 0.01	0.85 ± 0.01
GES	165 ± 7	0.52 ± 0.02	0.51 ± 0.03	0.22 ± 0.02	0.82 ± 0.02
GOLEM	152 ± 4	0.54 ± 0.01	0.56 ± 0.02	0.13 ± 0.02	0.94 ± 0.01
GRASP	163 ± 4	0.52 ± 0.01	0.50 ± 0.02	0.17 ± 0.01	0.87 ± 0.01
RESIT	169 ± 5	0.49 ± 0.01	0.37 ± 0.05	0.04 ± 0.01	0.94 ± 0.01
PC	164 ± 4	0.50 ± 0.01	0.46 ± 0.03	0.09 ± 0.01	0.92 ± 0.01

(d) **Erdös-Rényi Nonlinear (Part 2/2).**

Table 7: **Benchmarks on invertible non-linear models.** Ablation studies on simulated non-linear ground truth models with 20 nodes scale-free and Erdös-Renyi DAG structures with average degree 2 and invertible (sigmoidal) non-linear mechanisms. We report means and 95% confidence intervals (CIs) across 20 different ground truth models. Arrows next to metrics indicate lower is better (↓) and higher is better (↑).

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-GP	32 ± 1	0.11 ± 0.04	0.40 ± 0.07	0.15 ± 0.04
BAYESDAG	64 ± 3	0.38 ± 0.04	0.62 ± 0.05	0.38 ± 0.04
DAG-GNN	24 ± 3	0.32 ± 0.03	0.88 ± 0.03	0.31 ± 0.03
DDS	166 ± 5	0.50 ± 0.02	0.54 ± 0.02	0.50 ± 0.02
DIBS-GP	60 ± 5	0.31 ± 0.05	0.61 ± 0.06	0.32 ± 0.05
GADGET	66 ± 4	0.44 ± 0.03	0.70 ± 0.03	0.45 ± 0.03
GES	67 ± 3	0.49 ± 0.05	0.74 ± 0.04	0.51 ± 0.05
GOLEM	41 ± 5	0.39 ± 0.04	0.78 ± 0.04	0.41 ± 0.04
GRASP	49 ± 2	0.47 ± 0.05	0.77 ± 0.04	0.48 ± 0.04
RESIT	30 ± 3	0.36 ± 0.03	0.89 ± 0.02	0.33 ± 0.04
PC	28 ± 1	–	–	–

(a) **Scale-free Nonlinear (Part 1/2).**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-GP	16 ± 4	0.91 ± 0.03	0.79 ± 0.07	0.72 ± 0.06	0.98 ± 0.00
BAYESDAG	65 ± 5	0.71 ± 0.04	0.36 ± 0.04	0.49 ± 0.04	0.86 ± 0.01
DAG-GNN	45 ± 2	0.58 ± 0.02	0.29 ± 0.03	0.20 ± 0.04	0.95 ± 0.01
DDS	164 ± 5	0.61 ± 0.04	0.26 ± 0.05	0.53 ± 0.05	0.57 ± 0.01
DIBS-GP	60 ± 6	0.77 ± 0.04	0.42 ± 0.06	0.50 ± 0.05	0.88 ± 0.02
GADGET	72 ± 4	0.76 ± 0.03	0.30 ± 0.04	0.41 ± 0.03	0.85 ± 0.01
GES	77 ± 5	0.61 ± 0.03	0.32 ± 0.04	0.38 ± 0.05	0.84 ± 0.01
GOLEM	54 ± 5	0.62 ± 0.03	0.34 ± 0.04	0.32 ± 0.05	0.91 ± 0.01
GRASP	60 ± 5	0.64 ± 0.03	0.36 ± 0.05	0.39 ± 0.06	0.89 ± 0.01
RESIT	60 ± 3	0.51 ± 0.01	0.14 ± 0.02	0.09 ± 0.02	0.92 ± 0.01
PC	51 ± 2	0.57 ± 0.01	0.27 ± 0.02	0.21 ± 0.02	0.93 ± 0.00

(b) **Scale-free Nonlinear (Part 2/2).**

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-GP	27 ± 2	0.16 ± 0.05	0.39 ± 0.08	0.18 ± 0.05
BAYESDAG	50 ± 2	0.28 ± 0.04	0.53 ± 0.08	0.29 ± 0.05
DAG-GNN	28 ± 4	0.32 ± 0.04	0.71 ± 0.06	0.34 ± 0.04
DDS	56 ± 6	0.27 ± 0.04	0.46 ± 0.07	0.33 ± 0.04
DIBS-GP	158 ± 5	0.46 ± 0.04	0.51 ± 0.05	0.46 ± 0.04
GADGET	59 ± 5	0.28 ± 0.06	0.46 ± 0.08	0.30 ± 0.06
GES	61 ± 6	0.34 ± 0.06	0.49 ± 0.09	0.35 ± 0.06
GOLEM	37 ± 7	0.39 ± 0.05	0.68 ± 0.06	0.40 ± 0.04
GRASP	45 ± 4	0.34 ± 0.07	0.54 ± 0.09	0.36 ± 0.07
RESIT	25 ± 2	0.32 ± 0.04	0.66 ± 0.07	0.33 ± 0.04
PC	27 ± 1	–	–	–

(c) **Erdös-Rényi Nonlinear (Part 1/2).**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-GP	24 ± 5	0.83 ± 0.04	0.64 ± 0.06	0.54 ± 0.05	0.98 ± 0.01
BAYESDAG	56 ± 6	0.71 ± 0.04	0.35 ± 0.05	0.43 ± 0.05	0.90 ± 0.01
DAG-GNN	56 ± 4	0.54 ± 0.01	0.23 ± 0.02	0.15 ± 0.02	0.94 ± 0.01
DDS	59 ± 5	0.74 ± 0.04	0.41 ± 0.05	0.45 ± 0.07	0.89 ± 0.01
DIBS-GP	154 ± 6	0.64 ± 0.05	0.31 ± 0.06	0.55 ± 0.06	0.60 ± 0.01
GADGET	54 ± 7	0.84 ± 0.04	0.54 ± 0.06	0.56 ± 0.06	0.89 ± 0.01
GES	58 ± 9	0.73 ± 0.04	0.49 ± 0.05	0.57 ± 0.07	0.88 ± 0.02
GOLEM	59 ± 6	0.57 ± 0.02	0.28 ± 0.03	0.22 ± 0.04	0.92 ± 0.02
GRASP	47 ± 8	0.73 ± 0.04	0.52 ± 0.06	0.54 ± 0.07	0.92 ± 0.01
RESIT	61 ± 3	0.49 ± 0.01	0.12 ± 0.02	0.06 ± 0.02	0.93 ± 0.00
PC	45 ± 5	0.63 ± 0.03	0.40 ± 0.05	0.31 ± 0.06	0.95 ± 0.00

(d) **Erdös-Rényi Nonlinear (Part 2/2).**

Table 8: **Varying the number of data (scale-free non-linear models, table 1/2).** Ablation studies on simulated non-linear ground truth models with 20 nodes. We report means and 95% confidence intervals (CIs) across 20 different ground truth models. Arrows next to metrics indicate lower is better (↓) and higher is better (↑).

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-GP	32 ± 1	0.06 ± 0.02	0.13 ± 0.04	0.07 ± 0.02
BAYESDAG	55 ± 2	0.28 ± 0.02	0.50 ± 0.03	0.28 ± 0.02
DAG-GNN	15 ± 2	0.23 ± 0.02	0.87 ± 0.03	0.23 ± 0.02
DDS	167 ± 4	0.48 ± 0.02	0.52 ± 0.02	0.47 ± 0.02
DIBS-GP	26 ± 4	0.23 ± 0.03	0.65 ± 0.06	0.23 ± 0.03
GADGET	35 ± 2	0.37 ± 0.02	0.82 ± 0.03	0.37 ± 0.02
GES	44 ± 3	0.44 ± 0.03	0.78 ± 0.03	0.45 ± 0.02
GOLEM	30 ± 2	0.27 ± 0.03	0.77 ± 0.04	0.29 ± 0.03
GRASP	29 ± 2	0.38 ± 0.02	0.82 ± 0.04	0.39 ± 0.02
PC	22 ± 1	–	–	–

(a) **N = 100 data.**

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-GP	34 ± 1	0.04 ± 0.03	0.07 ± 0.04	0.04 ± 0.03
BAYESDAG	62 ± 2	0.27 ± 0.04	0.39 ± 0.06	0.27 ± 0.04
DAG-GNN	17 ± 4	0.24 ± 0.02	0.89 ± 0.03	0.24 ± 0.02
DDS	169 ± 5	0.46 ± 0.01	0.49 ± 0.02	0.46 ± 0.01
DIBS-GP	32 ± 4	0.20 ± 0.04	0.61 ± 0.07	0.22 ± 0.03
GADGET	44 ± 2	0.38 ± 0.02	0.75 ± 0.03	0.39 ± 0.02
GES	52 ± 3	0.43 ± 0.03	0.75 ± 0.03	0.45 ± 0.03
GOLEM	26 ± 2	0.28 ± 0.03	0.85 ± 0.03	0.28 ± 0.03
GRASP	37 ± 2	0.41 ± 0.03	0.79 ± 0.04	0.43 ± 0.03
PC	29 ± 1	–	–	–

(b) **N = 200 data.**

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-GP	35 ± 1	0.03 ± 0.02	0.05 ± 0.03	0.04 ± 0.02
BAYESDAG	77 ± 2	0.33 ± 0.04	0.41 ± 0.05	0.33 ± 0.03
DAG-GNN	14 ± 2	0.26 ± 0.03	0.91 ± 0.02	0.26 ± 0.03
DDS	178 ± 3	0.46 ± 0.02	0.48 ± 0.02	0.46 ± 0.02
DIBS-GP	32 ± 3	0.23 ± 0.02	0.61 ± 0.05	0.24 ± 0.02
GADGET	59 ± 3	0.40 ± 0.03	0.70 ± 0.03	0.41 ± 0.03
GES	66 ± 4	0.42 ± 0.04	0.67 ± 0.03	0.44 ± 0.03
GOLEM	23 ± 2	0.28 ± 0.03	0.85 ± 0.03	0.29 ± 0.02
GRASP	50 ± 3	0.45 ± 0.04	0.73 ± 0.04	0.46 ± 0.03
PC	37 ± 2	–	–	–

(c) **N = 500 data.**

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-GP	35 ± 1	0.03 ± 0.02	0.05 ± 0.02	0.04 ± 0.02
BAYESDAG	83 ± 4	0.32 ± 0.03	0.39 ± 0.05	0.32 ± 0.03
DAG-GNN	18 ± 2	0.26 ± 0.02	0.88 ± 0.02	0.26 ± 0.02
DDS	178 ± 3	0.46 ± 0.02	0.47 ± 0.02	0.46 ± 0.02
DIBS-GP	37 ± 3	0.24 ± 0.03	0.59 ± 0.05	0.26 ± 0.03
GADGET	73 ± 3	0.41 ± 0.03	0.62 ± 0.03	0.41 ± 0.02
GES	81 ± 4	0.43 ± 0.04	0.59 ± 0.05	0.44 ± 0.04
GOLEM	24 ± 2	0.27 ± 0.03	0.84 ± 0.03	0.27 ± 0.02
GRASP	61 ± 2	0.40 ± 0.04	0.65 ± 0.05	0.42 ± 0.03
PC	42 ± 2	–	–	–

(d) **N = 1000 data.**

Table 9: **Varying the number of data (scale-free non-linear models, table 2/2).** Ablation studies on simulated non-linear ground truth models with 20 nodes. We report means and 95% confidence intervals (CIs) across 20 different ground truth models. Arrows next to metrics indicate lower is better (↓) and higher is better (↑).

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-GP	5 ± 2	0.97 ± 0.02	0.94 ± 0.03	0.88 ± 0.04	1.00 ± 0.00
BAYESDAG	46 ± 3	0.81 ± 0.02	0.50 ± 0.03	0.61 ± 0.03	0.91 ± 0.01
DAG-GNN	36 ± 2	0.59 ± 0.02	0.39 ± 0.05	0.20 ± 0.04	0.98 ± 0.01
DDS	157 ± 5	0.71 ± 0.03	0.37 ± 0.05	0.65 ± 0.04	0.58 ± 0.01
DIBS-GP	36 ± 4	0.69 ± 0.03	0.42 ± 0.05	0.36 ± 0.05	0.96 ± 0.01
GADGET	51 ± 3	0.72 ± 0.02	0.30 ± 0.04	0.28 ± 0.03	0.93 ± 0.01
GES	58 ± 4	0.61 ± 0.02	0.32 ± 0.03	0.32 ± 0.03	0.90 ± 0.01
GOLEM	38 ± 3	0.67 ± 0.02	0.45 ± 0.04	0.38 ± 0.04	0.95 ± 0.01
GRASP	45 ± 3	0.63 ± 0.02	0.37 ± 0.04	0.32 ± 0.04	0.94 ± 0.01
PC	46 ± 2	0.57 ± 0.02	0.27 ± 0.04	0.18 ± 0.03	0.95 ± 0.00

(a) **N = 100 data.**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-GP	3 ± 2	0.98 ± 0.02	0.96 ± 0.03	0.94 ± 0.04	1.00 ± 0.00
BAYESDAG	48 ± 4	0.85 ± 0.03	0.52 ± 0.04	0.70 ± 0.04	0.89 ± 0.01
DAG-GNN	37 ± 3	0.59 ± 0.03	0.38 ± 0.06	0.22 ± 0.05	0.97 ± 0.01
DDS	158 ± 4	0.73 ± 0.04	0.46 ± 0.05	0.66 ± 0.06	0.58 ± 0.01
DIBS-GP	39 ± 4	0.72 ± 0.04	0.43 ± 0.05	0.40 ± 0.06	0.95 ± 0.01
GADGET	54 ± 3	0.74 ± 0.03	0.34 ± 0.04	0.36 ± 0.03	0.91 ± 0.01
GES	63 ± 4	0.63 ± 0.02	0.34 ± 0.03	0.37 ± 0.04	0.88 ± 0.01
GOLEM	40 ± 3	0.63 ± 0.03	0.39 ± 0.05	0.30 ± 0.05	0.96 ± 0.01
GRASP	49 ± 3	0.65 ± 0.02	0.38 ± 0.04	0.37 ± 0.04	0.92 ± 0.01
PC	50 ± 3	0.57 ± 0.02	0.27 ± 0.04	0.21 ± 0.03	0.94 ± 0.00

(b) **N = 200 data.**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-GP	2 ± 1	0.98 ± 0.01	0.98 ± 0.01	0.95 ± 0.03	1.00 ± 0.00
BAYESDAG	61 ± 5	0.84 ± 0.03	0.45 ± 0.04	0.71 ± 0.05	0.85 ± 0.01
DAG-GNN	36 ± 2	0.59 ± 0.02	0.39 ± 0.05	0.20 ± 0.04	0.98 ± 0.00
DDS	164 ± 4	0.76 ± 0.02	0.52 ± 0.05	0.70 ± 0.03	0.56 ± 0.01
DIBS-GP	38 ± 3	0.71 ± 0.03	0.43 ± 0.04	0.41 ± 0.04	0.95 ± 0.01
GADGET	64 ± 5	0.75 ± 0.03	0.34 ± 0.05	0.43 ± 0.03	0.87 ± 0.01
GES	69 ± 5	0.66 ± 0.02	0.39 ± 0.03	0.48 ± 0.03	0.85 ± 0.01
GOLEM	37 ± 2	0.64 ± 0.02	0.44 ± 0.04	0.32 ± 0.03	0.97 ± 0.01
GRASP	57 ± 5	0.67 ± 0.02	0.41 ± 0.04	0.45 ± 0.04	0.89 ± 0.01
PC	54 ± 2	0.60 ± 0.02	0.30 ± 0.03	0.27 ± 0.03	0.92 ± 0.00

(c) **N = 500 data.**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-GP	2 ± 1	0.99 ± 0.01	0.98 ± 0.01	0.95 ± 0.02	1.00 ± 0.00
BAYESDAG	68 ± 6	0.84 ± 0.02	0.43 ± 0.04	0.72 ± 0.04	0.83 ± 0.02
DAG-GNN	35 ± 1	0.62 ± 0.01	0.43 ± 0.03	0.26 ± 0.03	0.98 ± 0.00
DDS	163 ± 5	0.76 ± 0.04	0.52 ± 0.07	0.71 ± 0.05	0.56 ± 0.01
DIBS-GP	40 ± 3	0.73 ± 0.02	0.46 ± 0.03	0.45 ± 0.04	0.94 ± 0.01
GADGET	74 ± 4	0.77 ± 0.03	0.33 ± 0.04	0.48 ± 0.03	0.84 ± 0.01
GES	82 ± 7	0.66 ± 0.04	0.39 ± 0.04	0.51 ± 0.06	0.81 ± 0.01
GOLEM	36 ± 2	0.65 ± 0.01	0.46 ± 0.03	0.34 ± 0.03	0.97 ± 0.01
GRASP	63 ± 5	0.69 ± 0.03	0.42 ± 0.04	0.51 ± 0.05	0.87 ± 0.01
PC	59 ± 3	0.59 ± 0.02	0.28 ± 0.03	0.27 ± 0.04	0.91 ± 0.01

(d) **N = 1000 data.**

Table 10: **Varying the number of data (Erdös-Renyi non-linear models, table 1/2).** Ablation studies on simulated non-linear ground truth models with 20 nodes. We report means and 95% confidence intervals (CIs) across 20 different ground truth models. Arrows next to metrics indicate lower is better (↓) and higher is better (↑).

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-GP	20 ± 2	0.15 ± 0.02	0.32 ± 0.04	0.18 ± 0.03
BAYESDAG	42 ± 1	0.27 ± 0.04	0.56 ± 0.06	0.29 ± 0.04
DAG-GNN	13 ± 2	0.27 ± 0.03	0.71 ± 0.06	0.26 ± 0.03
DDS	159 ± 7	0.48 ± 0.03	0.54 ± 0.04	0.48 ± 0.04
DIBS-GP	18 ± 4	0.23 ± 0.03	0.48 ± 0.04	0.25 ± 0.03
GADGET	28 ± 3	0.29 ± 0.04	0.62 ± 0.05	0.31 ± 0.04
GES	36 ± 3	0.39 ± 0.05	0.67 ± 0.05	0.42 ± 0.04
GOLEM	27 ± 2	0.32 ± 0.04	0.71 ± 0.05	0.32 ± 0.04
GRASP	23 ± 2	0.35 ± 0.05	0.65 ± 0.07	0.36 ± 0.05
PC	21 ± 1	–	–	–

(a) **N = 100 data.**

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-GP	21 ± 2	0.12 ± 0.02	0.28 ± 0.04	0.14 ± 0.02
BAYESDAG	48 ± 1	0.18 ± 0.03	0.38 ± 0.05	0.20 ± 0.03
DAG-GNN	16 ± 3	0.26 ± 0.03	0.65 ± 0.05	0.26 ± 0.03
DDS	159 ± 5	0.43 ± 0.03	0.48 ± 0.03	0.43 ± 0.03
DIBS-GP	20 ± 3	0.19 ± 0.02	0.42 ± 0.05	0.20 ± 0.02
GADGET	34 ± 2	0.27 ± 0.03	0.54 ± 0.05	0.29 ± 0.03
GES	39 ± 3	0.34 ± 0.04	0.55 ± 0.05	0.35 ± 0.03
GOLEM	23 ± 2	0.27 ± 0.03	0.62 ± 0.06	0.27 ± 0.03
GRASP	28 ± 2	0.31 ± 0.04	0.55 ± 0.06	0.32 ± 0.04
PC	26 ± 1	–	–	–

(b) **N = 200 data.**

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-GP	20 ± 2	0.14 ± 0.02	0.28 ± 0.03	0.16 ± 0.02
BAYESDAG	61 ± 2	0.22 ± 0.04	0.38 ± 0.06	0.24 ± 0.04
DAG-GNN	16 ± 3	0.29 ± 0.03	0.68 ± 0.06	0.28 ± 0.03
DDS	161 ± 7	0.43 ± 0.04	0.47 ± 0.04	0.43 ± 0.04
DIBS-GP	22 ± 5	0.21 ± 0.03	0.42 ± 0.07	0.23 ± 0.03
GADGET	43 ± 4	0.27 ± 0.04	0.46 ± 0.07	0.29 ± 0.04
GES	48 ± 5	0.33 ± 0.05	0.50 ± 0.07	0.35 ± 0.05
GOLEM	21 ± 2	0.30 ± 0.04	0.65 ± 0.06	0.30 ± 0.04
GRASP	36 ± 4	0.32 ± 0.05	0.50 ± 0.07	0.33 ± 0.05
PC	32 ± 2	–	–	–

(c) **N = 500 data.**

Model	#Edges	↓ A-AID	↓ P-AID	↓ OSET-AID
ARCO-GP	22 ± 2	0.13 ± 0.03	0.31 ± 0.04	0.16 ± 0.03
BAYESDAG	65 ± 2	0.25 ± 0.05	0.42 ± 0.06	0.27 ± 0.05
DAG-GNN	18 ± 2	0.29 ± 0.02	0.73 ± 0.04	0.29 ± 0.03
DDS	164 ± 5	0.45 ± 0.03	0.49 ± 0.03	0.45 ± 0.03
DIBS-GP	22 ± 3	0.22 ± 0.03	0.47 ± 0.05	0.25 ± 0.03
GADGET	53 ± 4	0.30 ± 0.05	0.50 ± 0.06	0.32 ± 0.04
GES	58 ± 5	0.36 ± 0.05	0.54 ± 0.06	0.38 ± 0.05
GOLEM	22 ± 2	0.30 ± 0.03	0.70 ± 0.03	0.30 ± 0.03
GRASP	44 ± 3	0.35 ± 0.04	0.55 ± 0.05	0.36 ± 0.04
PC	38 ± 2	–	–	–

(d) **N = 1000 data.**

Table 11: **Varying the number of data (Erdös-Renyi non-linear models, table 2/2).** Ablation studies on simulated non-linear ground truth models with 20 nodes. We report means and 95% confidence intervals (CIs) across 20 different ground truth models. Arrows next to metrics indicate lower is better (↓) and higher is better (↑).

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-GP	21 ± 3	0.84 ± 0.03	0.68 ± 0.05	0.49 ± 0.05	1.00 ± 0.00
BAYESDAG	51 ± 3	0.71 ± 0.03	0.34 ± 0.04	0.39 ± 0.03	0.92 ± 0.01
DAG-GNN	45 ± 3	0.54 ± 0.01	0.25 ± 0.03	0.10 ± 0.02	0.97 ± 0.01
DDS	157 ± 7	0.63 ± 0.04	0.29 ± 0.05	0.53 ± 0.06	0.59 ± 0.02
DIBS-GP	37 ± 5	0.64 ± 0.03	0.38 ± 0.05	0.26 ± 0.05	0.98 ± 0.01
GADGET	43 ± 4	0.76 ± 0.02	0.43 ± 0.05	0.31 ± 0.03	0.95 ± 0.01
GES	47 ± 5	0.66 ± 0.03	0.43 ± 0.05	0.39 ± 0.05	0.93 ± 0.01
GOLEM	48 ± 3	0.59 ± 0.01	0.32 ± 0.02	0.23 ± 0.02	0.95 ± 0.01
GRASP	40 ± 4	0.64 ± 0.03	0.44 ± 0.05	0.33 ± 0.05	0.96 ± 0.01
PC	42 ± 4	0.61 ± 0.02	0.39 ± 0.04	0.25 ± 0.03	0.96 ± 0.00

(a) **N = 100 data.**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-GP	20 ± 2	0.87 ± 0.03	0.71 ± 0.05	0.51 ± 0.04	1.00 ± 0.00
BAYESDAG	46 ± 3	0.76 ± 0.02	0.45 ± 0.04	0.52 ± 0.04	0.92 ± 0.01
DAG-GNN	44 ± 3	0.56 ± 0.01	0.30 ± 0.04	0.14 ± 0.02	0.97 ± 0.01
DDS	152 ± 6	0.68 ± 0.04	0.40 ± 0.05	0.59 ± 0.05	0.60 ± 0.01
DIBS-GP	36 ± 3	0.65 ± 0.02	0.39 ± 0.03	0.30 ± 0.04	0.97 ± 0.01
GADGET	41 ± 3	0.80 ± 0.02	0.49 ± 0.06	0.40 ± 0.04	0.95 ± 0.01
GES	45 ± 4	0.69 ± 0.02	0.47 ± 0.03	0.44 ± 0.03	0.93 ± 0.01
GOLEM	44 ± 3	0.59 ± 0.01	0.35 ± 0.03	0.23 ± 0.03	0.96 ± 0.01
GRASP	37 ± 3	0.69 ± 0.02	0.51 ± 0.04	0.43 ± 0.04	0.96 ± 0.01
PC	41 ± 3	0.65 ± 0.02	0.43 ± 0.03	0.33 ± 0.03	0.96 ± 0.00

(b) **N = 200 data.**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-GP	20 ± 2	0.80 ± 0.03	0.67 ± 0.05	0.50 ± 0.04	1.00 ± 0.00
BAYESDAG	50 ± 4	0.80 ± 0.03	0.48 ± 0.04	0.64 ± 0.04	0.89 ± 0.01
DAG-GNN	43 ± 3	0.56 ± 0.02	0.29 ± 0.05	0.15 ± 0.04	0.97 ± 0.01
DDS	150 ± 6	0.73 ± 0.03	0.50 ± 0.05	0.65 ± 0.04	0.60 ± 0.02
DIBS-GP	36 ± 5	0.67 ± 0.03	0.41 ± 0.06	0.31 ± 0.06	0.97 ± 0.01
GADGET	40 ± 4	0.84 ± 0.03	0.60 ± 0.06	0.53 ± 0.05	0.94 ± 0.01
GES	46 ± 7	0.73 ± 0.03	0.51 ± 0.05	0.55 ± 0.06	0.92 ± 0.01
GOLEM	43 ± 4	0.59 ± 0.02	0.34 ± 0.05	0.22 ± 0.04	0.96 ± 0.00
GRASP	37 ± 6	0.74 ± 0.03	0.56 ± 0.05	0.54 ± 0.06	0.94 ± 0.01
PC	40 ± 5	0.68 ± 0.03	0.48 ± 0.05	0.41 ± 0.05	0.95 ± 0.01

(c) **N = 500 data.**

Model	↓ ESHD	↑ AUROC	↑ AUPRC	↑ TPR	↑ TNR
ARCO-GP	21 ± 3	0.78 ± 0.03	0.68 ± 0.05	0.52 ± 0.06	1.00 ± 0.00
BAYESDAG	51 ± 4	0.85 ± 0.02	0.51 ± 0.05	0.68 ± 0.04	0.89 ± 0.01
DAG-GNN	45 ± 3	0.57 ± 0.02	0.32 ± 0.05	0.17 ± 0.04	0.97 ± 0.00
DDS	150 ± 4	0.75 ± 0.02	0.52 ± 0.04	0.68 ± 0.03	0.60 ± 0.01
DIBS-GP	41 ± 5	0.63 ± 0.03	0.39 ± 0.06	0.28 ± 0.05	0.97 ± 0.01
GADGET	47 ± 5	0.86 ± 0.03	0.59 ± 0.05	0.57 ± 0.04	0.91 ± 0.01
GES	54 ± 8	0.73 ± 0.03	0.51 ± 0.04	0.57 ± 0.04	0.89 ± 0.02
GOLEM	44 ± 3	0.59 ± 0.01	0.37 ± 0.03	0.22 ± 0.03	0.96 ± 0.00
GRASP	42 ± 4	0.75 ± 0.02	0.55 ± 0.04	0.56 ± 0.04	0.93 ± 0.01
PC	47 ± 4	0.66 ± 0.02	0.44 ± 0.04	0.39 ± 0.04	0.94 ± 0.01

(d) **N = 1000 data.**

E PROOFS AND DERIVATIONS

E.1 Derivation of the Posterior Expectation w.r.t. SCMs

This derivation is similar in spirit to the formulation in (Lorch et al., 2021; Toth et al., 2022), albeit for different generative models. In the following, we derive the expectation w.r.t. SCMs in Equation (3) and the corresponding importance weights in Equation (4). We start with the expectation w.r.t. SCMs $\mathcal{M} = (G, \boldsymbol{\psi}, \mathbf{f})$ that we parametrise as causal graph G, GP hyper-parameters $\boldsymbol{\psi}$ and mechanisms $\mathbf{f}$

$$\mathbb{E}_{\mathcal{M} \mid \mathcal{D}} \left[p(Y \mid \mathcal{M}) \right] = \mathbb{E}_{G, \mathbf{f}, \boldsymbol{\psi} \mid \mathcal{D}} \left[p(Y \mid \mathcal{M}) \right]$$

and by marginalising over/exploiting conditional independences of our generative model in Figure 1

$$\begin{aligned}
&= \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D}} \left[\mathbb{E}_{G, \mathbf{f} \mid \boldsymbol{\theta}, \boldsymbol{\psi}, \mathcal{D}} \left[p(Y \mid \mathcal{M}) \right] \right] \\
&= \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D}} \left[\mathbb{E}_{G \mid \boldsymbol{\theta}, \boldsymbol{\psi}, \mathcal{D}} \left[\mathbb{E}_{\mathbf{f} \mid \boldsymbol{\psi}, \mathcal{D}} \left[p(Y \mid \mathcal{M}) \right] \right] \right] \\
&= \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D}} \left[\mathbb{E}_{L \mid \boldsymbol{\theta}, \boldsymbol{\psi}, \mathcal{D}} \left[\mathbb{E}_{G \mid L, \boldsymbol{\psi}, \mathcal{D}} \left[\mathbb{E}_{\mathbf{f} \mid \boldsymbol{\psi}, \mathcal{D}} \left[p(Y \mid \mathcal{M}) \right] \right] \right] \right]
\end{aligned}$$

and by rewriting $p(L \mid \boldsymbol{\theta}, \boldsymbol{\psi}, \mathcal{D})$ using Bayes' law

$$\begin{aligned}
&= \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D}} \left[\mathbb{E}_{L \mid \boldsymbol{\theta}} \left[\frac{p(\mathcal{D}, \boldsymbol{\psi} \mid L)}{p(\mathcal{D}, \boldsymbol{\psi} \mid \boldsymbol{\theta})} \cdot \mathbb{E}_{G \mid L, \boldsymbol{\psi}, \mathcal{D}} \left[\mathbb{E}_{\mathbf{f} \mid \boldsymbol{\psi}, \mathcal{D}} \left[p(Y \mid \mathcal{M}) \right] \right] \right] \right] \\
&= \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D}} \left[\mathbb{E}_{L \mid \boldsymbol{\theta}} \left[w^L \cdot \mathbb{E}_{G \mid L, \boldsymbol{\psi}, \mathcal{D}} \left[\mathbb{E}_{\mathbf{f} \mid \boldsymbol{\psi}, \mathcal{D}} \left[p(Y \mid \mathcal{M}) \right] \right] \right] \right]
\end{aligned}$$

with

$$\begin{aligned}
w^L &:= \frac{p(\mathcal{D}, \boldsymbol{\psi} \mid L)}{p(\mathcal{D}, \boldsymbol{\psi} \mid \boldsymbol{\theta})} \\
&= \frac{\mathbb{E}_{G \mid L} \left[p(\mathcal{D}, \boldsymbol{\psi} \mid G) \right]}{p(\mathcal{D}, \boldsymbol{\psi} \mid \boldsymbol{\theta})} \\
&= \frac{\mathbb{E}_{G \mid L} \left[p(\mathcal{D} \mid \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \mid G) \right]}{p(\mathcal{D}, \boldsymbol{\psi} \mid \boldsymbol{\theta})} \\
&= \frac{\mathbb{E}_{G \mid L} \left[p(\mathcal{D} \mid \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \mid G) \right]}{\mathbb{E}_{L' \mid \boldsymbol{\theta}} \left[p(\mathcal{D}, \boldsymbol{\psi} \mid L') \right]} \\
&= \frac{\mathbb{E}_{G \mid L} \left[p(\mathcal{D} \mid \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \mid G) \right]}{\mathbb{E}_{L' \mid \boldsymbol{\theta}} \left[\mathbb{E}_{G' \mid L'} \left[p(\mathcal{D} \mid \boldsymbol{\psi}, G') \cdot p(\boldsymbol{\psi} \mid G') \right] \right]}.
\end{aligned}$$

E.2 Derivation of the Gradient Estimators

In the following, we derive the gradient estimators in Equations (5), (6) and (12). We denote by $\nabla = \nabla_{\boldsymbol{\theta}, \boldsymbol{\psi}}$ to avoid clutter.

General Posterior Gradient. The general posterior gradient in Equation (5) reads as follows.

$$\begin{aligned}
\nabla \log p(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathcal{D}) &= \nabla \log \frac{p(\mathcal{D}, \boldsymbol{\psi} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\mathcal{D})} \\
&= \nabla \log p(\boldsymbol{\theta}) + \nabla \log p(\mathcal{D}, \boldsymbol{\psi} \mid \boldsymbol{\theta}) \\
&= \nabla \log p(\boldsymbol{\theta}) + \nabla \log \mathbb{E}_{L \mid \boldsymbol{\theta}} \left[p(\mathcal{D}, \boldsymbol{\psi} \mid L) \right] \\
&= \nabla \log p(\boldsymbol{\theta}) + \nabla \log \mathbb{E}_{L \mid \boldsymbol{\theta}} \left[\mathbb{E}_{G \mid L} \left[p(\mathcal{D}, \boldsymbol{\psi} \mid G) \right] \right] \\
&= \nabla \log p(\boldsymbol{\theta}) + \nabla \log \mathbb{E}_{L \mid \boldsymbol{\theta}} \left[\mathbb{E}_{G \mid L} \left[p(\mathcal{D} \mid \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \mid G) \right] \right]
\end{aligned}$$

ARCO Gradient. Using the above as starting point for the gradient in Equation (6), we get

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}, \boldsymbol{\psi} \,|\, \mathcal{D}) &= \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \log \mathbb{E}_{L \,|\, \boldsymbol{\theta}} \left[\mathbb{E}_{G \,|\, L} \left[p(\mathcal{D} \,|\, \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \,|\, G)\right]\right] \\ &= \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \frac{\nabla_{\boldsymbol{\theta}} \mathbb{E}_{L \,|\, \boldsymbol{\theta}} \left[\mathbb{E}_{G \,|\, L} \left[p(\mathcal{D} \,|\, \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \,|\, G)\right]\right]}{\mathbb{E}_{L' \,|\, \boldsymbol{\theta}} \left[\mathbb{E}_{G' \,|\, L'} \left[p(\mathcal{D} \,|\, \boldsymbol{\psi}, G') \cdot p(\boldsymbol{\psi} \,|\, G')\right]\right]}\end{aligned}$$

and using $\nabla_{\boldsymbol{\theta}} p(L \,|\, \boldsymbol{\theta}) = p(L \,|\, \boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}} \log p(L \,|\, \boldsymbol{\theta})$

$$\begin{aligned}&= \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \frac{\mathbb{E}_{L \,|\, \boldsymbol{\theta}} \left[\mathbb{E}_{G \,|\, L} \left[p(\mathcal{D} \,|\, \boldsymbol{\psi}, G) \cdot p(\boldsymbol{\psi} \,|\, G)\right] \cdot \nabla_{\boldsymbol{\theta}} \log p(L \,|\, \boldsymbol{\theta})\right]}{\mathbb{E}_{L' \,|\, \boldsymbol{\theta}} \left[\mathbb{E}_{G' \,|\, L'} \left[p(\mathcal{D} \,|\, \boldsymbol{\psi}, G') \cdot p(\boldsymbol{\psi} \,|\, G')\right]\right]} \\ &= \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \mathbb{E}_{L \,|\, \boldsymbol{\theta}} \left[w^L \cdot \nabla_{\boldsymbol{\theta}} \log p(L \,|\, \boldsymbol{\theta})\right]\end{aligned}$$

with w^L as defined in Equation (4).

GP Hyper-parameter Gradient. For the gradient of a distinct GP modeling the mechanism from parents $\mathbf{Pa}_k$ to target node X_k with corresponding hyper-parameters $\boldsymbol{\psi}_k$ as in Equation (12) we have

$$\begin{aligned}\nabla_{\boldsymbol{\psi}_k} \log p(\boldsymbol{\theta}, \boldsymbol{\psi} \,|\, \mathcal{D}) &= \nabla_{\boldsymbol{\psi}_k} \log p(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\psi}_k} \log \mathbb{E}_{L \,|\, \boldsymbol{\theta}} \left[\mathbb{E}_{G \,|\, L} \left[p(\mathcal{D}, \boldsymbol{\psi} \,|\, G)\right]\right] \\ &= \frac{\mathbb{E}_{L \,|\, \boldsymbol{\theta}} \left[\mathbb{E}_{G \,|\, L} \left[\nabla_{\boldsymbol{\psi}_k} p(\mathcal{D}, \boldsymbol{\psi} \,|\, G)\right]\right]}{\mathbb{E}_{L \,|\, \boldsymbol{\theta}} \left[\mathbb{E}_{G \,|\, L} \left[p(\mathcal{D}, \boldsymbol{\psi} \,|\, G)\right]\right]} \\ &= \frac{\mathbb{E}_{L \,|\, \boldsymbol{\theta}} \left[\mathbb{E}_{G \,|\, L} \left[\nabla_{\boldsymbol{\psi}_k} p(\mathcal{D}, \boldsymbol{\psi} \,|\, G)\right]\right]}{p(\mathcal{D}, \boldsymbol{\psi} \,|\, \boldsymbol{\theta})}\end{aligned}$$

and as the marginal likelihood and the prior over GP hyper-parameters factorise over parent sets we further get

$$\begin{aligned}&= \frac{\mathbb{E}_{L \,|\, \boldsymbol{\theta}} \left[\mathbb{E}_{G \,|\, L} \left[\nabla_{\boldsymbol{\psi}_k} \prod_{i=1}^{d} p(\mathcal{D}_i, \boldsymbol{\psi}_i \,|\, \mathbf{Pa}_i^G)\right]\right]}{p(\mathcal{D}, \boldsymbol{\psi} \,|\, \boldsymbol{\theta})} \\ &= \frac{\mathbb{E}_{L \,|\, \boldsymbol{\theta}} \left[\sum_G p(G \,|\, L) \cdot \nabla_{\boldsymbol{\psi}_k} \prod_{i=1}^{d} p(\mathcal{D}_i, \boldsymbol{\psi}_i \,|\, \mathbf{Pa}_i^G)\right]}{p(\mathcal{D}, \boldsymbol{\psi} \,|\, \boldsymbol{\theta})}.\end{aligned}$$

Now, note that for the summation over graphs G, the gradient is zero for all graphs that do not contain the parent set $\mathbf{Pa}_k$ corresponding to the GP with hyper-parameters ψ_k. Consequently, we get

$$= \frac{\mathbb{E}_{L \,|\, \boldsymbol{\theta}} \left[\sum_{G \,|\, \mathbf{Pa}_k \in G} p(G \,|\, L) \cdot \nabla_{\boldsymbol{\psi}_k} \prod_{i=1}^{d} p(\mathcal{D}_i, \boldsymbol{\psi}_i \,|\, \mathbf{Pa}_i^G)\right]}{p(\mathcal{D}, \boldsymbol{\psi} \,|\, \boldsymbol{\theta})}$$

and since $\nabla_{\boldsymbol{\psi}_k} p(\mathcal{D}_k, \boldsymbol{\psi}_k \,|\, \mathbf{Pa}_k) = p(\mathcal{D}_k, \boldsymbol{\psi}_k \,|\, \mathbf{Pa}_k) \cdot \nabla_{\boldsymbol{\psi}_k} \log p(\mathcal{D}_k, \boldsymbol{\psi}_k \,|\, \mathbf{Pa}_k)$

$$\begin{aligned}&= \frac{\mathbb{E}_{L \,|\, \boldsymbol{\theta}} \left[\sum_{G \,|\, \mathbf{Pa}_k \in G} p(G \,|\, L) \cdot \prod_{i=1}^{d} p(\mathcal{D}_i, \boldsymbol{\psi}_i \,|\, \mathbf{Pa}_i^G) \cdot \nabla_{\boldsymbol{\psi}_k} \log p(\mathcal{D}_k, \boldsymbol{\psi}_k \,|\, \mathbf{Pa}_k)\right]}{p(\mathcal{D}, \boldsymbol{\psi} \,|\, \boldsymbol{\theta})} \\ &= \frac{\mathbb{E}_{L \,|\, \boldsymbol{\theta}} \left[\sum_{G \,|\, \mathbf{Pa}_k \in G} p(G \,|\, L) \cdot p(\mathcal{D}, \boldsymbol{\psi} \,|\, G)\right]}{p(\mathcal{D}, \boldsymbol{\psi} \,|\, \boldsymbol{\theta})} \cdot \nabla_{\boldsymbol{\psi}_k} \log p(\mathcal{D}_k, \boldsymbol{\psi}_k \,|\, \mathbf{Pa}_k).\end{aligned}$$

Note that the term preceding the gradient $\nabla_{\boldsymbol{\psi}_k} \log p(\mathcal{D}_k, \boldsymbol{\psi}_k \,|\, \mathbf{Pa}_k)$ is a scalar factor that *does not* influence the *direction* of the gradient and can thus be practically omitted for gradient-based optimisation, as optimisation algorithms will scale the gradient depending on tune-able step size parameters anyways. Therefore, optimising the GP hyper-parameters w.r.t. the gradient in Equation (5) practically yields the same gradient direction as the common MAP type II gradient $\nabla_{\boldsymbol{\psi}_k} \log p(\mathcal{D}_k, \boldsymbol{\psi}_k \,|\, \mathbf{Pa}_k)$. Arguably, for mechanism models that do not decompose over individual mechanisms for each parent set and target variable, naively estimating the gradient in Equation (5) may yield very noisy gradients, as the magnitude of the estimated gradient will depend on the sampled structures used for its estimation.

E.3 Proofs regarding Exhaustive Parent Set Enumeration

Proof of Proposition 4.1

Proof.

$$\mathbb{E}_{G \mid L}\left[w(G) \cdot Y(G)\right] = \sum_{G} p(G \mid L) \cdot w(G) \cdot Y(G)$$

Since we assume that $w(G)$ and $Y(G)$ factorise over the parent sets, we have

$$= \sum_{G} \prod_{i} p(\mathbf{Pa}_i^G \mid L) \cdot w_i(\mathbf{Pa}_i^G) \cdot Y_i(\mathbf{Pa}_i^G)$$

The sum over all graphs can be represented as sum over all combinations of possible parent sets to get

$$\begin{aligned} &= \sum_{\mathbf{Pa}_1} \sum_{\mathbf{Pa}_2} \cdots \sum_{\mathbf{Pa}_d} \prod_{i} p(\mathbf{Pa}_i \mid L) \cdot w_i(\mathbf{Pa}_i) \cdot Y_i(\mathbf{Pa}_i) \\ &= \sum_{\mathbf{Pa}_1} p(\mathbf{Pa}_1 \mid L) \cdot w_1(\mathbf{Pa}_1) \cdot Y_1(\mathbf{Pa}_1) \sum_{\mathbf{Pa}_2} p(\mathbf{Pa}_2 \mid L) \cdot w_2(\mathbf{Pa}_2) \cdot Y_2(\mathbf{Pa}_2) \dots \end{aligned}$$

Since each summation over parent sets is independent of the others, we get the final result

$$= \prod_{i} \sum_{\mathbf{Pa}_i} p(\mathbf{Pa}_i \mid L) \cdot w_i(\mathbf{Pa}_i) \cdot Y_i(\mathbf{Pa}_i)$$

□

Proof of Proposition 4.2

Proof.

$$\mathbb{E}_{G \mid L}\left[w(G) \cdot Y(G)\right] = \sum_{G} p(G \mid L) \cdot w(G) \cdot Y(G)$$

Since we assume that $w(G)$ factorises and $Y(G)$ sums over the parent sets, we have

$$= \sum_{G} \prod_{i} p(\mathbf{Pa}_i^G \mid L) \cdot w_i(\mathbf{Pa}_i^G) \cdot \sum_{j} Y_j(\mathbf{pa}_j^G)$$

The sum over all graphs can be represented as sum over all combinations of possible parent sets to get

$$\begin{aligned} &= \sum_{\mathbf{Pa}_1} \sum_{\mathbf{Pa}_2} \cdots \sum_{\mathbf{Pa}_d} \prod_{i} p(\mathbf{Pa}_i \mid L) \cdot w_i(\mathbf{Pa}_i) \cdot \sum_{j} Y_j(\mathbf{Pa}_j) \\ &= \sum_{\mathbf{Pa}_1} \sum_{\mathbf{Pa}_2} \cdots \sum_{\mathbf{Pa}_d} \cdot \sum_{j} Y_j(\mathbf{Pa}_j) \prod_{i} p(\mathbf{Pa}_i \mid L) \cdot w_i(\mathbf{Pa}_i) \\ &= \sum_{\mathbf{Pa}_1} \sum_{\mathbf{Pa}_2} \cdots \sum_{\mathbf{Pa}_d} Y_1(\mathbf{Pa}_1) \prod_{i} p(\mathbf{Pa}_i \mid L) \cdot w_i(\mathbf{Pa}_i) + \sum_{\mathbf{Pa}_1} \sum_{\mathbf{Pa}_2} \cdots \sum_{\mathbf{Pa}_d} \sum_{j=2}^{d} \dots \\ &= \sum_{\mathbf{Pa}_1} Y_1(\mathbf{Pa}_1) \cdot p(\mathbf{Pa}_1 \mid L) \cdot w_1(\mathbf{Pa}_1) \sum_{\mathbf{Pa}_2} p(\mathbf{Pa}_2 \mid L) \cdot w_2(\mathbf{Pa}_2) \cdots + \sum_{\mathbf{Pa}_1} \sum_{\mathbf{Pa}_2} \cdots \sum_{\mathbf{Pa}_d} \sum_{j=2}^{d} \dots \end{aligned}$$

By abbreviating $\alpha_i(L) = \sum_{\mathbf{Pa}_i} p(\mathbf{Pa}_i \mid L) w_i(\mathbf{Pa}_i)$ we get

$$= \sum_{\mathbf{Pa}_1} Y_1(\mathbf{Pa}_1) \cdot p(\mathbf{Pa}_1 \mid L) \cdot w_1(\mathbf{Pa}_1) \cdot \prod_{k \neq 1} \alpha_k(L) + \sum_{\mathbf{Pa}_1} \sum_{\mathbf{Pa}_2} \cdots \sum_{\mathbf{Pa}_d} \sum_{j=2}^{d} Y_j(\mathbf{Pa}_j) \prod_i p(\mathbf{Pa}_i \mid L) \cdot w_i(\mathbf{Pa}_i)$$

By repeating this procedure for the remaining summands j, we get the final result

$$= \sum_i \left(\prod_{k \neq i} \alpha_k(L) \right) \cdot \sum_{\mathbf{Pa}_i} p(\mathbf{Pa}_i \mid L) \cdot w_i(\mathbf{Pa}_i) \cdot Y_i(\mathbf{Pa}_i)$$

□