SVI-DAG: A STRUCTURED VARIATIONAL INFERENCE APPROACH TO BAYESIAN CAUSAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Bayesian causal discovery seeks to determine the posterior distribution of causal theories, which are interpreted as directed acyclic graphs (DAGs) that explain the observed data. The resulting posterior allows systematic reasoning regarding epistemic uncertainty within these theories. Nonetheless, finding such graphs is difficult due to identifiability problems and limited observational data. Furthermore, precisely approximating posterior over graphs is challenging given the vast range of potential DAGs. Recent Bayesian approaches have addressed some of these challenges, yet they remain limited as they fail to encode dependencies between edges, and lack principled ways to incorporate domain knowledge as inductive biases during the search process. To overcome these limitations, we propose `SVI-DAG`, a structured variational inference approach to Bayesian causal discovery using observational data and prior beliefs that uses normalizing flows to model dependencies between edges, supporting expressive and multimodal posterior learning over DAGs. To mitigate mode seeking behaviour in evidence lower bound optimization and promote mode coverage, we use stein variational gradient descent to update the node potentials using a kernel in acyclicity space. We evaluate `SVI-DAG` against 5 state-of-the-art Bayesian DAG learning methods and demonstrate competitive performance in terms of both accuracy and uncertainty quantification.

1 INTRODUCTION

Understanding causal relationships from data (Schölkopf et al., 2021) remains a central problem across science and engineering. Causal discovery—the task of inferring which variables directly influence others—allows researchers to move beyond mere associations and reason about the consequences of interventions in complex systems. Such a capability holds the potential to accelerate progress across a broad range of domains, including epidemiology (Tennant et al., 2021), climate science (Runge et al., 2019), economics (Imbens, 2020), psychology (Foster, 2010), and beyond. Causal relationships are commonly represented as directed acyclic graphs (DAGs) (Peters et al., 2017), where nodes correspond to variables and directed edges encode direct causal influences. Learning the DAG from data, however, constitutes a formidable combinatorial problem (Heinze-Deml et al., 2018), as the number of admissible DAGs grows super-exponentially with the number of variables and with purely observational data, the true DAG may be identifiable only up to its Markov equivalence class (MEC)[1]. This fundamental identifiability limitation motivates a Bayesian treatment (Friedman & Koller, 2003), wherein the goal is to approximate the posterior distribution over all DAGs that explain the data. Recent work on safe and trustworthy AI (Bengio et al., 2026) further argues that epistemic uncertainty over causal theories is essential for building non-agentic Scientist AI (Bengio et al., 2025) systems that provide reliable predictions and can serve as guardrails against overconfident decision making.

Existing Bayesian approaches to DAG learning have made substantial progress, yet several limitations remain. A predominant assumption is that edges are modeled as independent random variables, thereby disregarding the rich dependencies that exist between edges. Failing to capture these dependencies leads to difficulty in capturing the multimodal structure of the posterior. Moreover, Bayesian methods naturally incorporate domain knowledge through structural priors, including edge-specific Bernoulli probabilities. However, fixing these probabilities does not explicitly represent uncertainty

[1]The Markov equivalence class is the set of DAGs that encode the same conditional independencies.

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

about the values—a distinction that becomes relevant when domain knowledge is incomplete or potentially misspecified. Strongly informative but incorrect probabilities can favor implausible structures, particularly when the observational data provide limited evidence to revise these beliefs.

Our Approach. To address these shortcomings, we introduce `SVI-DAG`, a structured variational inference (SVI) approach to Bayesian causal discovery. Our **core contributions** are as follows:

- We propose a differentiable Bayesian approach to causal discovery that encodes dependencies between edge logits using conditional normalizing flows to support expressive and multimodal learning over DAGs.
- Our approach provides a principled way to incorporate domain knowledge as an inductive bias using a new type of prior that adapts based on how strong or weak the prior beliefs are using a Beta-Bernoulli distribution. We further show that this construction induces a tractable Logistic-Beta prior over the edge logits.
- To mitigate mode seeking behavior in evidence lower bound (ELBO) optimization, we use stein variational gradient descent (SVGD) to update node potentials using kernel in acyclicity space. This improves mode coverage and quantification of epistemic uncertainty.

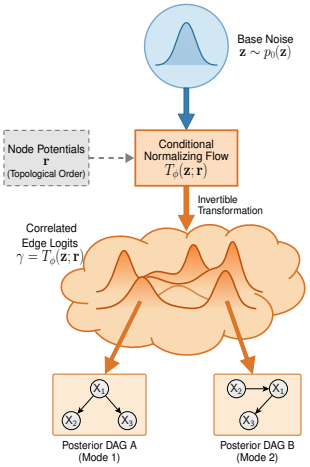

Figure 1: Illustration of SVI-DAG using conditional normalizing flows to capture edge dependencies and support multimodal learning over DAGs.

2 Related Works

Given the vast literature on causal discovery (Squires & Uhler, 2023), we focus exclusively on approaches that pose causal discovery as an optimization problem (Vowels et al., 2022).

Causal Discovery. A foundational shift in causal discovery began with NOTEARS (Zheng et al., 2018), which reformulated combinatorial DAG search as a continuous optimization problem. This idea was extended to nonlinear settings using graph neural networks (Yu et al., 2019), gradient based neural DAG learning (Lachapelle et al., 2019), and graph autoencoder architectures (Ng et al., 2019). Further improvements include DAGs with No Curl (Yu et al., 2021) which improves efficiency via a curl free condition, and methods using acyclicity such as DAGMA (Bello et al., 2022). Complementary approaches introduced masked gradients (Ng et al., 2022) or analyzed the interaction of sparsity and acyclicity penalties (Ng et al., 2020) while NoDAGs-Flow (Sethuraman et al., 2023) relaxed the acyclicity assumption entirely, learning nonlinear cyclic causal structures via normalizing flows. On the combinatorial side, BOSS (Andrews et al., 2023) demonstrated that efficient order score search can rival continuous relaxation methods in speed and accuracy, while Ban et al. (2024) proposed differentiable structure learning over partial topological orderings that flexibly interpolates between unconstrained and fully ordered search. Bridging constraint based and score based paradigms, Zhou et al. (2025) reformulated conditional independence testing differentiably, allowing end-to-end gradient based learning with statistical guarantees of constraint based methods.

Bayesian Causal Discovery. For Bayesian approaches, early differentiable methods include D-VAE (Zhang et al., 2019), which encodes DAGs in a latent space via variational autoencoders. VCN (Annadani et al., 2021) and BCD Nets (Cundy et al., 2021) introduced variational inference (VI) formulations over graph structures, while DiBS (Lorch et al., 2021) proposed a differentiable particle based framework. Concurrent lines of work explored differentiable DAG sampling (Charpentier et al., 2022), tractable marginal uncertainty computation (Wang et al., 2022), and amortized inference for both static (Lorch et al., 2022) and temporal (Löwe et al., 2022) settings. GFlowNets emerged as a powerful alternative, allowing diverse posterior sampling over structures (Deleu et al., 2022) and joint inference over graphs and parameters (Nishikawa-Toomey et al., 2022; Deleu et al., 2023). More recently, BayesDAG (Annadani et al., 2023) introduced gradient based posterior inference with improved scalability, ProDAG (Thompson et al., 2024) proposed projected VI to enforce acyclicity, and meta learning formulations (Dhir et al., 2024) have been explored to generalize across tasks.

3 METHODOLOGY

3.1 PRELIMINARIES AND NOTATIONS

3.1.1 CAUSAL GRAPH AND STRUCTURAL CAUSAL MODEL

Let $\mathcal{G} = (V, E)$ be a DAG on the variable set $V = \{X_1, \ldots, X_m\}$ where X_i represents a random variable. Let $A \in \{0,1\}^{m \times m}$ be the adjacency matrix encoding the causal relationships among the m variables with entries $A_{ij} = 1$ if there is a directed edge $X_j \to X_i$ in $\mathcal{G}$ and 0 otherwise. Let $\mathcal{X}_i \subseteq \mathbb{R}$ denote the sample space for node X_i, and define the joint sample space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$. Let $\mathcal{D} = \{x^{(n)}\}_{n=1}^N$ be the dataset, where each observation is a tuple $x^{(n)} = (x_1^{(n)}, \ldots, x_m^{(n)}) \in \mathcal{X}$ with $x_i^{(n)} \in \mathcal{X}_i$. A structural causal model (SCM) states that each variable is generated as a function of its parents and an exogenous noise. We denote by $X_{\mathrm{Pa}(i)}$ the set of parental nodes of X_i so that there is an edge from $X_j \in X_{\mathrm{Pa}(i)}$ to X_i in DAG $\mathcal{G}$. Formally, for each node $i \in \{1, \ldots, m\}$,

$$X_i = f_i\big(X_{\mathrm{Pa}(i)}; \theta_i\big) + \epsilon_i,$$

where $f_i(\cdot; \theta_i)$ is a deterministic function with parameters θ_i and and ϵ_i's are independent noise variables with strictly positive densities w.r.t Lebesgue measure. If noise variables are Gaussian and the functions f_i are neither linear nor constant, then the additive noise model (ANM) is structurally identifiable (Peters et al., 2014; Hoyer et al., 2008). Causality typically assumes structural assignments do not form cycles and they induce a DAG (Pearl, 2009).

3.1.2 BAYESIAN STRUCTURE LEARNING

Our goal is to approximate the posterior distribution over all possible DAGs A and associated parameters $\theta = \{\theta_1, \ldots, \theta_m\}$ in a neural network based SCM (NCM). Specifically, we target the posterior over structures and parameters $p(A, \theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid A, \theta)\, p(\theta \mid A)\, p(A)$ with a prior over parameters $p(\theta \mid A)$ and graphs $p(A)$ (Friedman & Koller, 2003). The likelihood is:

$$p(\mathcal{D} \mid A, \theta) = \prod_{n=1}^{N} \prod_{i=1}^{m} p\big(x_i^{(n)} \mid x_{\mathrm{Pa}(i)}^{(n)}; \theta_i\big),$$

where the dependency on A is implicit in the parent index sets $\mathrm{Pa}(i)$. Concretely, assuming the noise is Gaussian, $p(X_i \mid X_{\mathrm{Pa}(i)}, \theta_i) = \mathcal{N}\big(X_i \mid f_i(X_{\mathrm{Pa}(i)}; \theta_i), \sigma_i^2\big)$. Each θ_i determines conditional distribution of X_i given its parents respecting the DAG structure A.

3.2 OUR APPROACH

In this section, we detail each component of the `SVI-DAG` algorithm which is specifically designed to address some of the critical challenges in causal structure learning.

3.2.1 DOMAIN INFORMED PRIOR OVER ADJACENCY MATRIX

We start by defining a prior over the DAGs. A common choice is $B_{ij} \sim \mathrm{Bernoulli}(p_{ij})$. The parameters p_{ij} encode prior beliefs about edge presence. To improve robustness to misspecified prior beliefs, we introduce a hierarchical prior over B. For each ordered pair (i, j) with $i \neq j$, we introduce $\pi_{ij} \in (0, 1)$ such that $p(B_{ij} \mid \pi_{ij}) = \pi_{ij}^{B_{ij}} (1 - \pi_{ij})^{1 - B_{ij}}$. Instead of fixing π_{ij} to the prior value p_{ij}, we place a hyper prior on π_{ij} using a Beta distribution $\pi_{ij} \sim \mathrm{Beta}\Big(\alpha_{ij}, \beta_{ij}\Big)$. We choose hyperparameters so that Beta-density mode is p_{ij} given by

$$\alpha_{ij} = \nu_{ij} p_{ij} + 1, \quad \beta_{ij} = \nu_{ij}(1 - p_{ij}) + 1, \quad \nu_{ij} = \nu_{\min} + \kappa\, |p_{ij} - 0.5|^{\eta} \text{ for } i \neq j,$$

where $\nu_{ij} > 0$ is a concentration parameter that is strictly increasing in $|p_{ij} - 0.5|$ when $\kappa > 0$ and $\nu_{\min}$, κ, and η are constants. This choice assigns higher concentration to nominal beliefs farther from 0.5 (see Figure 2). As $\nu_{ij} \to \infty$, the marginal Beta–Bernoulli prior over each free edge B_{ij} converges to a $\mathrm{Bernoulli}(p_{ij})$ prior (see Proposition B.1 in appendix B). Let $\Pi = \{\pi_{ij}\}_{i \neq j}$. We sample edges as $B_{ij} \sim \mathrm{Bernoulli}(\pi_{ij})$, which gives the joint prior as

$$p(B, \Pi) \propto \prod_{i \neq j} \mathrm{Bernoulli}(B_{ij} \mid \pi_{ij})\, \mathrm{Beta}\Big(\pi_{ij} \mid \nu_{ij}\, p_{ij} + 1,\ \nu_{ij}\, (1 - p_{ij}) + 1\Big).$$

Marginalizing over π_{ij} recovers a Beta-Bernoulli model over B_{ij}. Since causal graphs are acyclic, we enforce acyclicity in the prior by construction (Annadani et al., 2023). We introduce a binary free edge matrix $B \in \{0,1\}^{m\times m}$ with zero diagonal and a vector of node potentials $r = (r_1, \ldots, r_m)^T \in \mathbb{R}^m$. Let's define acyclic mask and induced adjacency by

$$M(r) = P(r)\, L\, P(r)^T, \qquad A = B \odot M(r),$$

where $\odot$ denotes the Hadamard product. For any (B, r), A is a DAG (see Theorem B.3 in appendix). Moreover, for any DAG A, there always exists a corresponding pair (B, r) such that $A = B \odot M(r)$ (Annadani et al., 2023; Yu et al., 2021). We therefore place joint prior on the free edges and potentials $p(B, \Pi, r) \propto p(B, \Pi)\mathcal{N}(r \mid 0, \sigma_r^2 I_m)$. This construction enforces acyclicity, so no additional acyclicity indicator or penalty is required. The permutation $P(r)$ is well defined when entries of r are pairwise distinct. Since $p(r)$ is Gaussian, this condition holds almost surely. By analytically marginalizing over Π (see Derivation A.1 in appendix A), we have

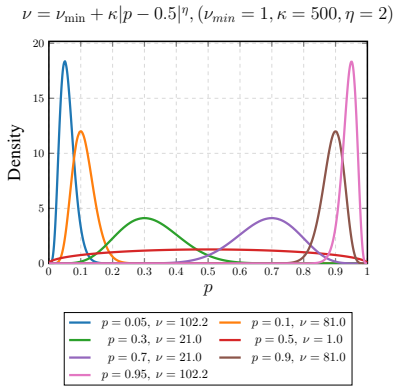

Figure 2: Adaptive domain informed prior over adjacency matrix

$$p(B, r) = \left(\prod_{i \neq j} \frac{\mathcal{B}(B_{ij} + \alpha_{ij},\, \beta_{ij} + 1 - B_{ij})}{\mathcal{B}(\alpha_{ij}, \beta_{ij})} \right) \mathcal{N}\big(r \mid 0, \sigma_r^2 I_m\big)\,,$$

where $\mathcal{B}$ denotes the Beta function. Under this formulation, p_{ij} is the nominal mode of the latent edge probability π_{ij}, not the exact finite ν_{ij} marginal probability of B_{ij}. The latter is $\alpha_{ij}/(\alpha_{ij} + \beta_{ij})$ which approaches p_{ij} as $\nu_{ij} \to \infty$.

3.2.2 Prior over NCM Parameters conditioned on the Adjacency Matrix

Next, conditional on A, we place a prior on NCM parameters $\theta = \{\theta_1, \ldots, \theta_m\}$. We consider a prior

$$p(\theta \mid A) \;=\; \mathcal{N}\Big(\theta \mid\; \mu_p(A),\; \mathrm{diag}\big(\sigma_p^2(A)\big)\Big) \;=\; \prod_{k=1}^{d_\theta} \mathcal{N}\big(\theta_k \mid\; \mu_{p,k}(A),\; \sigma_{p,k}^2(A)\big),$$

where d_θ is the dimensionality and $\mu_p(A) \in \mathbb{R}^{d_\theta}$ and $\sigma_p(A) \in (0, \infty)^{d_\theta}$ are deterministic functions of A. In the simplest case, these functions are constant (for example, zero mean and fixed variance), but the notation permits dependence on A. The joint prior factorizes as $p(A, \theta) \;=\; p(\theta \mid A)\; p(A)$.

3.2.3 ELBO without Normalizing Flow

By marginalizing Π in the prior to get $p(B, r)$ (see Derivation A.1 in appendix A) and rearranging terms (see Derivation A.2 in appendix A), we have the ELBO as

$$\mathrm{ELBO}(\phi) = \mathbb{E}_{q_\phi(B,r)}\Big[\underbrace{\mathbb{E}_{q_\phi(\theta|A)}[\log p(\mathcal{D} \mid A, \theta)]}_{\text{expected log likelihood}} - \underbrace{D_{\mathrm{KL}}(q_\phi(\theta \mid A) \,\|\, p(\theta \mid A))}_{\text{KL divergence for } \theta} \Big]$$
$$- \underbrace{D_{\mathrm{KL}}(q_\phi(B, r) \,\|\, p(B, r))}_{\text{KL divergence for } B \text{ and } r}.$$

where $q_\phi(\theta \mid A)$ and $q_\phi(B, r)$ denotes the conditional and structural guides.

Remark 3.1. Note that the above ELBO is not equivalent to the ELBO defined using $q_\phi(A)$ under deterministic map $A = B \odot M(r)$ because the pushforward from (B, r) to A is many-to-one. If $q_\phi(A)$ and $p(A)$ are defined as the pushforwards of $q_\phi(B, r)$ and $p(B, r)$, then Proposition B.2 implies $D_{\mathrm{KL}}\big(q_\phi(B, r) \,\|\, p(B, r)\big) \geq D_{\mathrm{KL}}\big(q_\phi(A) \,\|\, p(A)\big)$. Consequently, an ELBO written directly in terms of A is at least as tight as the ELBO in (B, r) space. However, the induced prior $p(A)$ is generally intractable, whereas $p(B, r)$ admits a simple factorized form. We therefore work with Eq. A.1 for computational tractability, even though ELBO in A space would be tighter.

3.3 Soft Relaxation for Adjacency Matrix and Acyclicity

Since B is discrete, we cannot differentiate with respect to B_{ij}. We therefore introduce a continuous relaxation for B via the Gumbel-Softmax trick (Jang et al., 2016; Maddison et al., 2016; Ng et al., 2022). For each (i, j) with $i \neq j$, we introduce a learnable edge logit $\gamma_{ij} \in \mathbb{R}$ and a temperature $T > 0$. We draw logistic noise $R_{ij} \overset{\text{iid}}{\sim} \text{Logistic}(0, 1)$ implemented as $R_{ij} = \log U_{ij} - \log(1 - U_{ij})$, $U_{ij} \overset{\text{iid}}{\sim} \text{Uniform}(0, 1)$. Equivalently, R_{ij} can be generated as the difference of two i.i.d. $\texttt{Gumbel}(0, 1)$ variables. We use the same noise to define both a hard Bernoulli sample and its differentiable relaxation:

$$B_{ij} = \mathbf{1}\{\gamma_{ij} + R_{ij} > 0\}, \qquad \tilde{B}_{ij} = \sigma\Big((\gamma_{ij} + R_{ij})/T\Big), \qquad \sigma(x) = 1/(1 + \mathrm{e}^{-x}), \tag{1}$$

where $B_{ii} = \tilde{B}_{ii} = 0$. Then $B_{ij} \sim \text{Bernoulli}(\sigma(\gamma_{ij}))$, while $\tilde{B}_{ij} \in (0, 1)$ converges pointwise to B_{ij} as $T \to 0$, except on the probability-zero event $\gamma_{ij} + R_{ij} = 0$ (Ng et al., 2022). For the order, let $P(r)$ be the hard permutation obtained by sorting r. We also define a backward surrogate through the Sinkhorn operator $\mathcal{S}$ (Adams & Zemel, 2011). For $\tau > 0$, let $S_0(r) = r\,\mathbf{o}^T$, where $\mathbf{o} = (m, m-1, \ldots, 1)^T$, and applying K_S alternating row and column normalizations to $\exp(S_0(r)/\tau)$ we have

$$P_{\tau,K_S}(r) = \mathcal{S}_{K_S}(\exp(S_0(r)/\tau)), \qquad M_{\tau,K_S}(r) = P_{\tau,K_S}(r)\, L\, P_{\tau,K_S}(r)^T.$$

Double stochasticity is guaranteed in the $K_S \to \infty$ limit. More importantly, even an exactly doubly stochastic nonpermutation matrix generally makes $M_{\tau,K_S}(r)$ dense, so this relaxed order mask is not generally acyclic and can have positive entries in both directions. Lemma B.4 (see appendix B) states that $P_{\tau,K_S}(r) \to P(r)$ and $M_{\tau,K_S}(r) \to M(r)$ as $\tau \to 0$ and $K_S \to \infty$. We distinguish the hard DAG from its soft backward surrogate as

$$A = B \odot M(r), \qquad \tilde{A} = \tilde{B} \odot M_{\tau,K_S}(r) \odot (\mathbf{1}_{m\times m} - I_m), \qquad A_{\text{ST}} = \tilde{A} + \text{sg}(A - \tilde{A}), \tag{2}$$

where $M(r) = P(r)LP(r)^T$ and sg is stop-gradient. Thus $A_{\text{ST}} = A$ in the forward pass and every evaluated graph is a binary DAG by Theorem B.3 (see appendix B). The diagonal factor removes self-loops from $\tilde{A}$, but does not make the finite-temperature relaxation acyclic.

3.4 Forward Model

We now define forward model using Bayesian MLP for node i that uses the parent $X_{\text{Pa}(i)}$ as input and outputs a prediction $\hat{X}_i$. Formally, $\hat{X}_i = f_i\big(X_{\text{Pa}(i)}; \theta_i\big)$. To incorporate adjacency, let $X = (X_1, \ldots, X_m)^T$ and let $(A_{\text{ST}})_i$ denote the ith row of the straight-through matrix in Eq. 2. We define

$$g_i(x; \theta_i) = f_i\big((A_{\text{ST}})_i \odot x; \theta_i\big).$$

Since $A_{\text{ST}} = A$ in the forward pass, the evaluated input zeros every nonparent exactly and all likelihood values correspond to a hard DAG. During backpropagation, derivatives are routed through the generally cyclic relaxation $\tilde{A}$. These straight through derivatives are biased surrogate gradients.

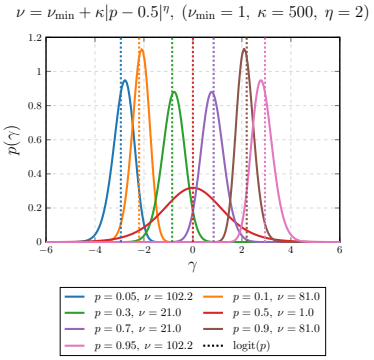

Figure 3: Induced prior over edge logits

3.5 ELBO with Normalizing Flow

In many applications, the assumption that $q_\phi(B)$ factorizes independently across edges can be restrictive. SVI (Rezende & Mohamed, 2015) addresses this limitation by constructing a more expressive $q_\phi(B)$. We define a standard base distribution $p_0(z)$ for edge latent variables. We introduce an invertible conditional map $T_\phi : \mathbb{R}^{m(m-1)} \times \mathbb{R}^m \to \mathbb{R}^{m(m-1)}$, parameterized by ϕ. The transformation gives γ as a function of z and r given by $\gamma = T_\phi(z; r)$. By rearrangement (see Derivation A.3 in appendix A), we have the revised ELBO as

$$\begin{aligned}\text{ELBO}(\phi, \psi) = \mathbb{E}_{r\sim q_\psi(r)}\, \mathbb{E}_{z\sim p_0}\, \mathbb{E}_{B\sim p(\cdot|T_\phi(z;r))} \Big[\underbrace{\mathbb{E}_{q_\phi(\theta|A)}\big[\log p(\mathcal{D} \mid A, \theta)\big]}_{\text{expected log likelihood}} - \underbrace{D_{\text{KL}}\big(q_\phi(\theta \mid A) \,\|\, p(\theta \mid A)\big)}_{\text{KL divergence for } \theta} \\ + \log\Big|\det \frac{\partial T_\phi(z; r)}{\partial z}\Big| - \log p_0(z) + \log p\big(T_\phi(z; r)\big) \Big] - \underbrace{D_{\text{KL}}\big(q_\psi(r) \,\|\, p(r)\big)}_{\text{KL divergence for potentials}}.\end{aligned} \tag{3}$$

We define T_ϕ as either a single flow transformation or a composition $T_{\phi_K}^{(K)} \circ \cdots \circ T_{\phi_1}^{(1)}$. In compositional case, the log determinant term in above equation decomposes as

$$\log \left| \det \frac{\partial T_\phi(z;r)}{\partial z} \right| = \sum_{k=1}^{K} \log \left| \det \frac{\partial T_{\phi_k}(z^{(k-1)};r)}{\partial z^{(k-1)}} \right|, \qquad z^{(k)} = T_{\phi_k}(z^{(k-1)};r),\ z^{(0)} = z.$$

Under Derivation A.4 (see appendix A) and the modeling choice that model and guide share the hard conditional $p(B \mid \gamma)$, the domain informed prior now enters through prior over edge logits $p(\gamma)$ (see Figure 3) as

$$p(\gamma) = \prod_{i \neq j} \frac{1}{\mathcal{B}(\alpha_{ij}, \beta_{ij})}\, \sigma(\gamma_{ij})^{\alpha_{ij}} \left(1 - \sigma(\gamma_{ij})\right)^{\beta_{ij}}, \qquad \sigma(x) = 1/(1+e^{-x}).$$

Remark 3.2. Had we instead placed a Bernoulli prior $B_{ij} \sim \mathrm{Bernoulli}(p_{ij})$ on each edge, the edge probability would be deterministic (i.e., $\pi_{ij} = p_{ij}$). Equivalently, $p(\pi_{ij}) = \delta(\pi_{ij} - p_{ij})$, and the logit transformation would induce $p(\gamma_{ij}) = \delta(\gamma_{ij} - \mathrm{logit}(p_{ij}))$. Hence, a fixed Bernoulli prior induces a degenerate prior over γ rather than an absolutely continuous density. This is not suitable to our formulation, since $q_\phi(\gamma \mid r)$ is continuous—so the KL divergence $D_{\mathrm{KL}}(q_\phi(\gamma \mid r) \,\|\, p(\gamma))$ to this degenerate prior diverges unless the flow collapses to the same point mass. The Beta prior over π_{ij} avoids this issue and gives the Logistic-Beta density. We further note that although the prior over π_{ij} is parameterized to have mode p_{ij}, this mode is not preserved under the nonlinear logit transformation. In particular, $\mathrm{logit}(p_{ij})$ is the image of the mode in π_{ij} space, whereas the mode of the Logistic-Beta density over γ_{ij} is $\log(\alpha_{ij}/\beta_{ij})$. Consequently, the peaks of $p(\gamma)$ need not align exactly with $\mathrm{logit}(p_{ij})$, as seen in Figure 3.

A single sample Monte Carlo (MC) estimator of revised ELBO is unbiased and practical in high dimension. The straight through backward pass is generally a biased gradient estimator. If both $q_\phi(\theta \mid A)$ and $p(\theta \mid A)$ are diagonal Gaussians with strictly positive component scales, their KL divergence has the closed form (see Derivation A.5 in appendix A) as

$$D_{\mathrm{KL}}\big(q_\phi(\theta \mid A) \,\|\, p(\theta \mid A)\big) = \tfrac{1}{2} \sum_{k=1}^{d_\theta} \left[\frac{\sigma_{\phi,k}^2(A)}{\sigma_{p,k}^2(A)} + \frac{\big(\mu_{\phi,k}(A) - \mu_{p,k}(A)\big)^2}{\sigma_{p,k}^2(A)} - 1 + 2 \log \frac{\sigma_{p,k}(A)}{\sigma_{\phi,k}(A)} \right]. \quad (4)$$

3.6 Updating Node Potentials with SVGD using a Kernel defined in Relaxed Acyclicity Space

Note that, even with a dependent guide, standard VI maximizes the ELBO, which is equivalent to minimizing the reverse KL divergence $D_{\mathrm{KL}}(q_\phi \| p)$. For multimodal posteriors, this reverse KL objective can be mode-seeking (Blei et al., 2017; Zhang et al., 2018). In particular, it encourages the approximation to concentrate on a single mode, largely to avoid allocating mass to low-density posterior regions. Consequently, although a mixture-of-Gaussians guide for r could, in principle, represent multiple modes, standard ELBO optimization would still not guarantee recovery of all posterior modes. To promote mode coverage, we replace the parametric guide over r with a non-parametric particle approximation and update the particles using SVGD (Liu & Wang, 2016). If we condition the revised ELBO (see Equation 3) on a fixed r, we have

$$\begin{aligned} \mathcal{L}(\phi; r) = \mathbb{E}_{z \sim p_0}\, \mathbb{E}_{B \sim p(\cdot \mid T_\phi(z;r))} \Big[& \underbrace{\mathbb{E}_{q_\phi(\theta \mid A)}\big[\log p(\mathcal{D} \mid A, \theta)\big]}_{\text{expected log likelihood}} - \underbrace{D_{\mathrm{KL}}\big(q_\phi(\theta \mid A) \,\|\, p(\theta \mid A)\big)}_{\text{KL divergence for } \theta} \\ & + \log \Big| \det \frac{\partial T_\phi(z;r)}{\partial z} \Big| - \log p_0(z) + \log p\big(T_\phi(z;r)\big) \Big]. \end{aligned} \quad (5)$$

The SVGD algorithm updates particles $\{r^{(k)}\}_{k=1}^{K_{\text{part}}}$ using gradient of log target density. The target score is $\nabla_r \log \tilde{p}_\phi(r \mid \mathcal{D}) = \nabla_r [\mathcal{L}(\phi; r) + \log p(r)]$. Let's define the continuously differentiable, positive-definite Gaussian kernel $k_M(\psi, \psi')$ and let $\psi(r) = \mathrm{vec}(M_{\tau, K_S}(r))$. Its pullback $k_r(r, r') = k_M(\psi(r), \psi(r'))$ is, in turn, a continuously differentiable, positive-definite kernel on potential space. Applying the standard smooth-target SVGD formula to this score gives the particle update rule

$$r^{(k)} \leftarrow r^{(k)} + \eta_r \frac{1}{K_{\text{part}}} \sum_{j=1}^{K_{\text{part}}} \Big[k_r(r^{(j)}, r^{(k)}) \nabla_{r^{(j)}} \log \tilde{p}_\phi(r \mid \mathcal{D}) + \nabla_{r^{(j)}} k_r(r^{(j)}, r^{(k)}) \Big]. \quad (6)$$

Here, $\nabla_{r^{(j)}} k_r$ differentiates through $\psi(r^{(j)})$. The second term in the bracket is the repulsive force that helps in mode coverage.

3.7 Reparameterization Trick for Estimating Gradients

In order to calculate gradients, let f_i be an L-layer MLP with parameters θ that contain weights and biases $\{W^{(l)}, b^{(l)}\}_{l=1}^{L}$ such that $W^{(l)} \in \mathbb{R}^{d_{\text{out}}^{(l)} \times d_{\text{in}}^{(l)}}$ and $b^{(l)} \in \mathbb{R}^{d_{\text{out}}^{(l)}}$. We approximate the conditional posterior over network parameters with a Gaussian guide. In software the hypernetwork H_ϕ receives A_{ST} since its forward value is exactly A, this evaluates the same density as $q_\phi(\theta \mid A)$ while routing a surrogate derivative through $\tilde{A}$:

$$q_\phi(\theta \mid A_{\text{ST}}) = \mathcal{N}\Big(\theta \mid \mu_\phi(A_{\text{ST}}),\ \text{diag}\big(\sigma_\phi^2(A_{\text{ST}})\big)\Big) = \prod_{k=1}^{d_\theta} \mathcal{N}\big(\theta_k \mid \mu_{\phi,k}(A_{\text{ST}}),\ \sigma_{\phi,k}^2(A_{\text{ST}})\big),$$

where $(\mu_\phi, \rho_\phi) = H_\phi(A_{\text{ST}})$ and $\sigma_\phi = \text{softplus}(\rho_\phi) = \log(1 + \exp(\rho_\phi))$, which enforces positivity. The conditioning is therefore explicit and differentiable. Changing A_{ST} changes the full variational posterior over θ through H_ϕ.

$$\theta = \mu_\phi(A_{\text{ST}}) + \sigma_\phi(A_{\text{ST}}) \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_{d_\theta}),$$

which preserves differentiability through both H_ϕ and the predictors f_i. $\tilde{B}$ is reparameterized using the Gumbel-Sigmoid relaxation as shown in Eq. 1. We refer the reader to Appendix A.6 for the pseudo code and computational complexity of the `SVI-DAG` algorithm.

4 Results and Discussion

We evaluate `SVI-DAG` against five state-of-the-art (SOTA) Bayesian DAG learning methods on linear synthetic data, nonlinear synthetic data and real data using brier score, expected structural hamming distance (SHD), expected F1 score, and area under the receiver operating characteristic curve (AUROC). The rationale underlying the selection of these metrics is provided in appendix C.

Datasets. To isolate effect of domain informed prior, we use a two node graph with data generated from ANM with Gaussian noise, where small graph size makes role of prior tractable to analyze. For synthetic experiments at scale, we generate linear and nonlinear structural equation models over Erdős–Rényi (ER) DAGs with 25 and 50 nodes. We report the 25 node results in the main text and defer the 50 node results to appendix F.1. For real data, we evaluate on Sachs dataset.

Baselines. We focus our evaluation on Bayesian methods to DAG learning and compare with `ProDAG` (Thompson et al., 2024), `BayesDAG` (Annadani et al., 2023), `DDS` (Charpentier et al., 2022), `BCD` Nets (Cundy et al., 2021), and `DiBS` (Lorch et al., 2021). Appendix E describes the hyperparameters used for benchmarking these algorithms.

4.1 Effect of Domain Informed Prior

Figure 4 illustrates effect of prior on a 2 node graph using 1000 samples, chosen to balance influence of likelihood and prior. After optimization, we draw 10000 hard posterior DAG samples $A = B \odot M(r)$, r sampled uniformly from SVGD particles, γ from $q(\gamma \mid r)$, hard Bernoulli edges $B = \mathbf{1}\{\gamma + R > 0\}$ and the hard permutation mask $M(r) = P(r)\, L\, P(r)^\top$, so every sample is a binary DAG and report the proportion of samples assigned to $x_1 \to x_2$, $x_1 \leftarrow x_2$, and $x_1 \perp\!\!\!\perp x_2$. With ground truth $x_1 \to x_2$, linear Gaussian setting (non-identifiable) shows that prior dominates orientation: an incorrect prior (p = 0.01 correct, 0.99 incorrect) concentrates mass on the wrong direction (84.7%), a noninformative prior (p = 0.5) gives balanced posterior, and correct prior (p = 0.99 correct, 0.01 incorrect) shifts mass to true direction (83.2%). In nonlinear Gaussian setting (identifiable), prior acts primarily as regularizer: the incorrect prior suppresses both edges, concentrating 86.4% on $x_1 \perp\!\!\!\perp x_2$, while a correct prior sharpens posterior and assigns highest mass to $x_1 \to x_2$ (95.2%).

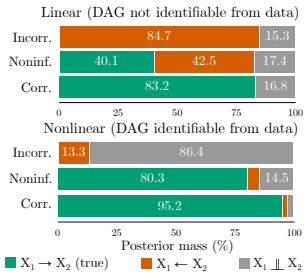

Figure 4: Effect of prior on 2 node graph

4.2 LINEAR SYNTHETIC DATA

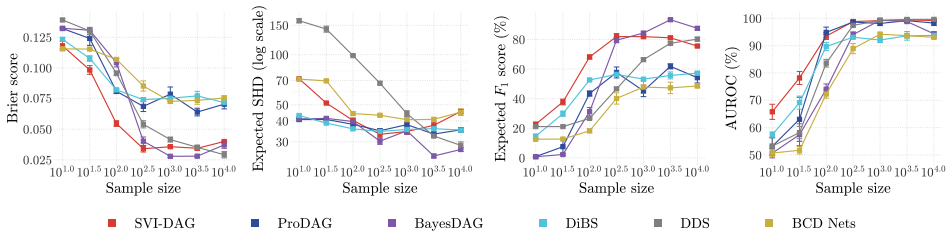

Figure 5: Results on data from linear ER DAGs with $p = 25$ nodes, $s = 40$ edges under Gaussian noise. Each point denotes mean, and error bars indicate standard error across 5 independent dataset.

Figure 5 and Figure F.1 (see appendix F.1) present results on datasets generated from linear ER DAGs with $p = 25$, $s = 40$ and $p = 50$, $s = 80$ across increasing sample sizes. Since the underlying structure is identifiable only up to its MEC, all metrics are evaluated at CPDAG level. `SVI-DAG` achieves lower Brier scores and higher E-F1 score across most of the sample sizes while remaining competitive with baselines in terms of E-SHD and AUROC.

4.3 NONLINEAR SYNTHETIC DATA

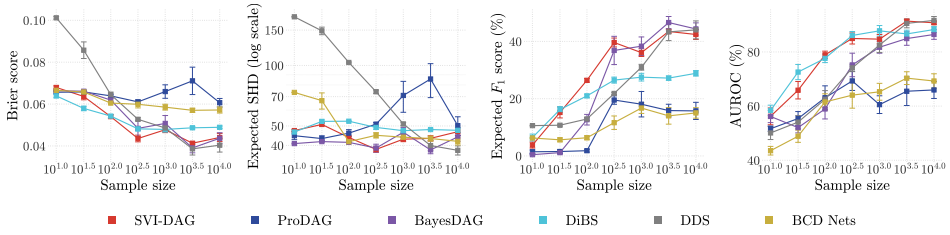

Figure 6: Results on synthetic data generated from nonlinear ER DAGs with $p = 25$ nodes and $s = 40$ edges under Gaussian noise.

Figure 6 and Figure F.2 (see appendix F.1) present results on datasets generated from nonlinear ER DAGs with $p = 25$, $s = 40$ and $p = 50$, $s = 80$ across increasing sample sizes. In this setting, the DAG is identifiable, and all metrics are therefore evaluated at the DAG level. We can observe that `SVI-DAG` attains lower Brier score and higher E-F1 across most sample sizes, while also staying competitive with other algorithms in terms of other metrics.

4.4 REAL DATA

The flow cytometry dataset (Sachs et al., 2005) is a standard benchmark for causal discovery in biological systems, comprising 7466 samples over 11 variables and 18 edges corresponding to phosphoproteins and phospholipids. Tables 2 and 3 report the performance on Sachs dataset under DAG and CPDAG metrics, respectively. We observe that `SVI-DAG` consistently outperforms the algorithms in F1 score for DAG metrics and in AUROC for CPDAG metrics. Table 1 reports the computational time.

Table 1: Wall-clock time per train/evaluate on the Sachs dataset on an A100-40GB GPU using JAX.

	Time per run (s)
`SVI-DAG`	134.2 ± 30.9
`ProDAG`	100.2 ± 4.1
`BayesDAG`	266.5 ± 13.0
`DDS`	65.1 ± 1.9
`BCD Nets`	151.1 ± 2.5
`DiBS`	$\mathbf{5.6 \pm 0.7}$

We note that all results reported for `SVI-DAG` in this paper correspond to a single choice of normalizing flow—neural spline flows. Our algorithm is deliberately flow-agnostic, and hence any normalizing flow family can be substituted for the spline flow without modifying the rest of the

framework. The reported results should therefore be read as the performance of one particular instantiation of SVI−DAG, rather than an upper bound on what the approach can ultimately achieve.

Table 2: Performance on Sachs dataset. The average and standard errors are measured over 10-fold cross-validation splits of data. The best value of each DAG metric is indicated in bold.

	Brier score (↓)	Exp. SHD (↓)	Exp. F1 score (↑)	AUROC (↑)
SVI−DAG	0.185 ± 0.007	25.70 ± 0.92	$\mathbf{28.95 \pm 1.73}$	58.16 ± 2.27
ProDAG	0.159 ± 0.003	$\mathbf{19.68 \pm 0.35}$	14.32 ± 1.54	52.95 ± 1.81
BayesDAG	0.174 ± 0.008	28.66 ± 0.78	26.75 ± 2.08	62.23 ± 1.93
DDS	0.168 ± 0.004	24.62 ± 0.56	23.21 ± 1.52	51.35 ± 2.18
BCD Nets	$\mathbf{0.153 \pm 0.004}$	20.94 ± 0.50	10.79 ± 2.16	55.91 ± 2.07
DiBS	0.180 ± 0.004	40.74 ± 0.50	27.78 ± 0.74	$\mathbf{65.63 \pm 2.07}$

Table 3: Performance on Sachs dataset. The average and standard errors are measured over 10-fold cross-validation splits of data. The best value of each CPDAG metric is indicated in bold.

	Brier score (↓)	Exp. SHD (↓)	Exp. F1 score (↑)	AUROC (↑)
SVI−DAG	0.237 ± 0.016	23.08 ± 0.87	54.05 ± 2.07	$\mathbf{70.17 \pm 2.28}$
ProDAG	0.267 ± 0.001	$\mathbf{17.36 \pm 0.07}$	31.55 ± 0.64	62.70 ± 0.90
BayesDAG	0.247 ± 0.010	26.38 ± 0.42	50.10 ± 1.36	62.88 ± 1.42
DDS	0.240 ± 0.005	23.02 ± 0.29	43.46 ± 0.91	68.91 ± 1.30
BCD Nets	0.259 ± 0.002	18.68 ± 0.16	26.30 ± 1.61	56.65 ± 1.62
DiBS	$\mathbf{0.215 \pm 0.003}$	35.49 ± 0.22	$\mathbf{54.46 \pm 0.29}$	66.80 ± 1.10

4.5 Ablation Studies

Table 4: Component ablation. Nonlinear ER $p = 20$, $s = 40$, mean $\pm$ s.e. over 10 seeds. MEC-cov: expected fraction of the true MEC receiving posterior mass from linear study ($p = 10$, $s = 10$).

Variant	Brier score ↓	Exp. SHD ↓	Exp. F1 (%) ↑	AUROC (%) ↑	MEC-cov ↑
SVI−DAG	$\mathbf{0.077 \pm 0.002}$	69.3 ± 1.1	26.9 ± 0.8	$\mathbf{81.7 \pm 1.2}$	$\mathbf{0.06 \pm 0.02}$
(a) w/o flow	0.078 ± 0.002	66.0 ± 0.8	25.4 ± 0.7	78.9 ± 1.2	0.00 ± 0.00
(b) w/o SVGD	0.083 ± 0.006	$\mathbf{62.0 \pm 2.5}$	$\mathbf{34.8 \pm 2.1}$	73.7 ± 3.2	0.05 ± 0.04
(c) w/o domain prior	0.208 ± 0.003	164.4 ± 1.5	18.2 ± 0.4	57.0 ± 2.4	0.00 ± 0.00
(d) w/o flow & SVGD	0.083 ± 0.005	62.2 ± 2.0	29.1 ± 2.1	70.3 ± 3.7	0.00 ± 0.00

Table 4 ablates each component of SVI−DAG. We observe that no ablation improves edge ranking—AUROC is highest for full model and degrades in every ablated variant. The prior is the dominant component in this regime: with $p(p-1) = 380$ edges informed by only 300 observations, removing it more than doubles the expected SHD and degrades calibration. Ablating SVGD, i.e., replacing the particle ensemble over order potentials with a reparameterized Gaussian guide, allows the resulting unimodal guide to commit to a single ordering mode, which lowers AUROC and raises Brier. The flow requires more care. Since every posterior DAG factorizes as $A = B \odot M(r)$, Brier, AUROC, and expected SHD are functionals of the posterior edge marginals alone, and a mean-field $q(\gamma)$ can match any marginal vector the conditional flow induces. Hence, the close agreement between the first two rows in the calibration column is largely expected and should not be read as evidence that the flow has no effect. Its contribution surfaces in the final column: in a companion study on linear graphs small enough to enumerate the Markov equivalence class exactly, only the flow-bearing variants attain nonzero coverage of the true class, so removing the flow eliminates exact structure recovery entirely. Intuitively, the probability assigned to a specific graph is a product over hundreds of per-edge events, and only a correlated posterior can concentrate mass on coherent whole graphs.

5 Conclusion

We presented SVI−DAG, a differentiable Bayesian approach to causal discovery that uses normalizing flows to encode dependencies between edges for expressive and multimodal learning over DAGs.

USE OF LARGE LANGUAGE MODELS

We used large language models to help implement and debug the code for `SVI-DAG` and its experiments, and to polish the writing of a manuscript drafted by the authors. All AI-assisted code and text were reviewed and verified by us.

REFERENCES

Ryan Prescott Adams and Richard S Zemel. Ranking via sinkhorn propagation. *arXiv preprint arXiv:1106.1925*, 2011.

Bryan Andrews, Joseph Ramsey, Ruben Sanchez Romero, Jazmin Camchong, and Erich Kummerfeld. Fast scalable and accurate discovery of dags using the best order score search and grow shrink trees. *Advances in neural information processing systems*, 36:63945–63956, 2023.

Yashas Annadani, Jonas Rothfuss, Alexandre Lacoste, Nino Scherrer, Anirudh Goyal, Yoshua Bengio, and Stefan Bauer. Variational causal networks: Approximate bayesian inference over causal structures. *arXiv preprint arXiv:2106.07635*, 2021.

Yashas Annadani, Nick Pawlowski, Joel Jennings, Stefan Bauer, Cheng Zhang, and Wenbo Gong. Bayesdag: Gradient-based posterior inference for causal discovery. *Advances in Neural Information Processing Systems*, 36:1738–1763, 2023.

Taiyu Ban, Lyuzhou Chen, Xiangyu Wang, Xin Wang, Derui Lyu, and Huanhuan Chen. Differentiable structure learning with partial orders. *Advances in Neural Information Processing Systems*, 37: 117426–117455, 2024.

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

Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, et al. Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? *arXiv preprint arXiv:2502.15657*, 2025.

Yoshua Bengio, Stephen Clare, Carina Prunkl, Maksym Andriushchenko, Ben Bucknall, Malcolm Murray, Rishi Bommasani, Stephen Casper, Tom Davidson, Raymond Douglas, et al. International ai safety report 2026. *arXiv preprint arXiv:2602.21012*, 2026.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Bertrand Charpentier, Simon Kibler, and Stephan Günnemann. Differentiable DAG Sampling. *International Conference on Learning Representations*, 2022.

Chris Cundy, Aditya Grover, and Stefano Ermon. Bcd nets: Scalable variational approaches for bayesian causal discovery. *Advances in Neural Information Processing Systems*, 34:7095–7110, 2021.

Tristan Deleu, António Góis, Chris Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, and Yoshua Bengio. Bayesian structure learning with generative flow networks. In *Uncertainty in Artificial Intelligence*, pp. 518–528. PMLR, 2022.

Tristan Deleu, Mizu Nishikawa-Toomey, Jithendaraa Subramanian, Nikolay Malkin, Laurent Charlin, and Yoshua Bengio. Joint bayesian inference of graphical structure and parameters with a single generative flow network. *Advances in Neural Information Processing Systems*, 36:31204–31231, 2023.

Anish Dhir, Matthew Ashman, James Requeima, and Mark van der Wilk. A meta-learning approach to bayesian causal discovery. *arXiv preprint arXiv:2412.16577*, 2024.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

E Michael Foster. Causal inference and developmental psychology. *Developmental psychology*, 46 (6):1454, 2010.

Nir Friedman and Daphne Koller. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine learning*, 50(1):95–125, 2003.

Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391, 2018.

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

Guido W Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–1179, 2020.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. DiBS: Differentiable Bayesian Structure Learning. *Advances in Neural Information Processing Systems*, 2021.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. *Advances in Neural Information Processing Systems*, 35: 13104–13118, 2022.

Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, pp. 509–525. PMLR, 2022.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Ignavier Ng, Shengyu Zhu, Zhitang Chen, and Zhuangyan Fang. A graph autoencoder approach to causal structure learning. *arXiv preprint arXiv:1911.07420*, 2019.

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

Ignavier Ng, Shengyu Zhu, Zhuangyan Fang, Haoyang Li, Zhitang Chen, and Jun Wang. Masked gradient-based causal structure learning. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pp. 424–432. SIAM, 2022.

Mizu Nishikawa-Toomey, Tristan Deleu, Jithendaraa Subramanian, Yoshua Bengio, and Laurent Charlin. Bayesian learning of causal structure and mechanisms with gflownets and variational bayes. *arXiv preprint arXiv:2211.02763*, 2022.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

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

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Muralikrishnna G Sethuraman, Romain Lopez, Rahul Mohan, Faramarz Fekri, Tommaso Biancalani, and Jan-Christian Hütter. Nodags-flow: Nonlinear cyclic causal structure learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 6371–6387. PMLR, 2023.

Chandler Squires and Caroline Uhler. Causal structure learning: A combinatorial perspective. *Foundations of Computational Mathematics*, 23(5):1781–1815, 2023.

Peter WG Tennant, Eleanor J Murray, Kellyn F Arnold, Laurie Berrie, Matthew P Fox, Sarah C Gadd, Wendy J Harrison, Claire Keeble, Lynsie R Ranker, Johannes Textor, et al. Use of directed acyclic graphs (dags) to identify confounders in applied health research: review and recommendations. *International journal of epidemiology*, 50(2):620–632, 2021.

Ryan Thompson, Edwin V Bonilla, and Robert Kohn. Prodag: Projected variational inference for directed acyclic graphs. *arXiv preprint arXiv:2405.15167*, 2024.

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

Benjie Wang, Matthew R Wicker, and Marta Kwiatkowska. Tractable uncertainty for structure learning. In *International Conference on Machine Learning*, pp. 23131–23150. PMLR, 2022.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International conference on machine learning*, pp. 7154–7163. PMLR, 2019.

Yue Yu, Tian Gao, Naiyu Yin, and Qiang Ji. Dags with no curl: An efficient dag structure learning approach. In *International Conference on Machine Learning*, pp. 12156–12166. Pmlr, 2021.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

Muhan Zhang, Shali Jiang, Zhicheng Cui, Roman Garnett, and Yixin Chen. D-vae: A variational autoencoder for directed acyclic graphs. *Advances in neural information processing systems*, 32, 2019.

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

Jincheng Zhou, Mengbo Wang, Anqi He, Yumeng Zhou, Hessam Olya, Murat Kocaoglu, and Bruno Ribeiro. Differentiable constraint-based causal discovery. *arXiv preprint arXiv:2510.22031*, 2025.

APPENDIX

This appendix includes additional details for the paper, "SVI-DAG: A Structured Variational Inference Approach to Bayesian Causal Discovery".

A THEORY

A.1 DERIVATION OF $p(B, r)$

We start from the construction that enforces acyclicity by mapping the free edge matrix $B \in \{0,1\}^{m\times m}$ and the potentials $r \in \mathbb{R}^m$ to the adjacency $A = B \odot M(r)$ where $M(r)$ is the acyclic mask induced by the order defined by r. The joint prior over B, the latent edge probabilities $\Pi = \{\pi_{ij}\}_{i\neq j}$, and r is

$$p(B, \Pi, r) \;\propto\; \left(\prod_{i\neq j} \text{Bernoulli}(B_{ij} \mid \pi_{ij})\ \text{Beta}(\pi_{ij} \mid \alpha_{ij}, \beta_{ij})\right) \mathcal{N}\big(r \mid 0, \sigma_r^2 I_m\big)\,,$$

with $B_{ii} = 0$ for all i and with independence between r and the set of free edges in the prior. The factors are

$$\text{Bernoulli}(B_{ij} \mid \pi_{ij}) = \pi_{ij}^{B_{ij}}\,(1-\pi_{ij})^{1-B_{ij}}\,, \qquad \text{Beta}(\pi_{ij} \mid \alpha_{ij}, \beta_{ij}) = \frac{\pi_{ij}^{\alpha_{ij}-1}\,(1-\pi_{ij})^{\beta_{ij}-1}}{\mathcal{B}(\alpha_{ij}, \beta_{ij})},$$

where $\mathcal{B}$ denotes the Beta function. Our goal is to integrate out π_{ij} to get the marginal prior on (B, r). For a fixed ordered pair (i, j) with $i \neq j$ we compute

$$p(B_{ij}) = \int_0^1 \text{Bernoulli}(B_{ij} \mid \pi_{ij})\ \text{Beta}(\pi_{ij} \mid \alpha_{ij}, \beta_{ij})\ d\pi_{ij}.$$

Substituting the explicit forms gives

$$p(B_{ij}) = \frac{1}{\mathcal{B}(\alpha_{ij}, \beta_{ij})} \int_0^1 \pi_{ij}^{B_{ij}+\alpha_{ij}-1}\,(1-\pi_{ij})^{(1-B_{ij})+\beta_{ij}-1}\ d\pi_{ij}.$$

The integral equals a Beta function with updated parameters, which gives the standard Beta Bernoulli marginal

$$p(B_{ij}) = \frac{\mathcal{B}(B_{ij}+\alpha_{ij},\ \beta_{ij}+1-B_{ij})}{\mathcal{B}(\alpha_{ij}, \beta_{ij})}.$$

Using independence across ordered pairs in the prior and independence of r from B in the prior, we get the full marginal as

$$p(B, r) = \left(\prod_{i\neq j} \frac{\mathcal{B}(B_{ij}+\alpha_{ij},\ \beta_{ij}+1-B_{ij})}{\mathcal{B}(\alpha_{ij}, \beta_{ij})}\right) \mathcal{N}\big(r \mid 0, \sigma_r^2 I_m\big)\,.$$

The diagonal constraint $B_{ii} = 0$ excludes self loops and is implemented by masking the diagonal of B. Note that instead of $p(B_{ij} = 1) = p_{ij}$ and $p(B_{ij} = 0) = 1 - p_{ij}$ we have

$$p(B_{ij} = 1) = \frac{\alpha_{ij}}{\alpha_{ij}+\beta_{ij}}, \qquad p(B_{ij} = 0) = \frac{\beta_{ij}}{\alpha_{ij}+\beta_{ij}},$$

for a chosen finite ν_{ij}. This completes the derivation.

A.2 DERIVATION OF ELBO WITHOUT THE NORMALIZING FLOW

Given $A = B \odot M(r)$, which is a deterministic function of (B, r) alone, we have the full joint model as $p(B, \Pi, r, \theta, \mathcal{D}) = p(B, \Pi, r)\ p(\theta \mid A)\ p(\mathcal{D} \mid A, \theta)$. The posterior is

$$p(B, \Pi, r, \theta \mid \mathcal{D}) = \frac{p(B, \Pi, r, \theta, \mathcal{D})}{\int p(B, \Pi, r, \theta, \mathcal{D})\, dB\, d\Pi\, dr\, d\theta}.$$

Since this posterior is intractable, we approximate it with a variational distribution

$$q_\phi(B, \Pi, r, \theta) = q_\phi(B, \Pi, r)\; q_\phi(\theta \mid A), \qquad A = B \odot M(r).$$

Here ϕ denotes the collection of variational parameters for B, Π, r, and θ. We begin by writing the evidence as $\log p(\mathcal{D}) = \log \int p(B, \Pi, r, \theta, \mathcal{D})\, dB\, d\Pi\, dr\, d\theta$. By definition, the ELBO is

$$\text{ELBO}(\phi) = \mathbb{E}_{q_\phi(B,\Pi,r,\theta)} \left[\log p(B, \Pi, r, \theta, \mathcal{D}) - \log q_\phi(B, \Pi, r, \theta)\right].$$

Using $p(B, \Pi, r, \theta, \mathcal{D}) = p(\mathcal{D} \mid A, \theta)\; p(\theta \mid A)\; p(B, \Pi, r)$, we have

$$\text{ELBO}(\phi) = \mathbb{E}_{q_\phi(B,\Pi,r,\theta)}[\log p(\mathcal{D} \mid A, \theta) + \log p(\theta \mid A) + \log p(B, \Pi, r) - \log q_\phi(B, \Pi, r, \theta)].$$

If we assume independence between structural variables and parameter variables in the guide, we can factorize:

$$q_\phi(B, \Pi, r, \theta) = q_\phi(B, \Pi, r)\, q_\phi(\theta \mid A),$$

so that

$$\log q_\phi(B, \Pi, r, \theta) = \log q_\phi(B, \Pi, r) + \log q_\phi(\theta \mid A).$$

By linearity of expectation and the definition of the KL divergence ($D_{\mathrm{KL}}(q\|p) = \mathbb{E}_q[\log q - \log p]$), the ELBO can then be rearranged as:

$$\begin{aligned}
\text{ELBO}(\phi) = \mathbb{E}_{q_\phi(B,\Pi,r)}\Big[&\mathbb{E}_{q_\phi(\theta|A)}[\log p(\mathcal{D} \mid A, \theta) + \log p(\theta \mid A)] \\
&- \mathbb{E}_{q_\phi(\theta|A)}[\log q_\phi(\theta \mid A)] + \log p(B, \Pi, r) - \log q_\phi(B, \Pi, r)\Big] \\
= \mathbb{E}_{q_\phi(B,\Pi,r)}&\Big[\mathbb{E}_{q_\phi(\theta|A)}[\log p(\mathcal{D} \mid A, \theta)]\Big] \\
&- \mathbb{E}_{q_\phi(B,\Pi,r)}\Big[D_{\mathrm{KL}}(q_\phi(\theta \mid A) \,\|\, p(\theta \mid A))\Big] - D_{\mathrm{KL}}(q_\phi(B, \Pi, r) \,\|\, p(B, \Pi, r)).
\end{aligned}$$

If we analytically marginalize Π in the prior to get $p(B, r)$ (see Derivation A.1) we have

$$p(B, r) = \left(\prod_{i \neq j} \frac{\mathcal{B}\left(B_{ij} + \alpha_{ij}, \beta_{ij} + 1 - B_{ij}\right)}{\mathcal{B}\left(\alpha_{ij}, \beta_{ij}\right)}\right) \mathcal{N}(r \mid 0, \sigma_r^2 I_m),$$

where $\mathcal{B}$ represents the Beta function. Since Π is an auxiliary latent variable, marginalizing it out does not change the evidence $p(\mathcal{D})$. We therefore perform variational inference directly on the collapsed model

$$p(\mathcal{D}, B, r, \theta) = p(\mathcal{D} \mid A, \theta)\, p(\theta \mid A)\, p(B, r), \qquad A = B \odot M(r),$$

and the corresponding guide $q_\phi(B, r)\, q_\phi(\theta \mid A)$. The ELBO of this collapsed model is

$$\begin{aligned}
\text{ELBO}(\phi) = \mathbb{E}_{q_\phi(B,r)}\Big[&\underbrace{\mathbb{E}_{q_\phi(\theta|A)}[\log p(\mathcal{D} \mid A, \theta)]}_{\text{expected log likelihood}} - \underbrace{D_{\mathrm{KL}}(q_\phi(\theta \mid A) \| p(\theta \mid A))}_{\text{KL divergence for } \theta}\Big] \\
&- \underbrace{D_{\mathrm{KL}}(q_\phi(B, r) \| p(B, r))}_{\text{KL divergence for } B \text{ and } r}.
\end{aligned} \tag{A.1}$$

A.3 Derivation of ELBO with Normalizing Flow

We derive the ELBO when the edge logits γ are generated by a conditional normalizing flow T_ϕ and acyclicity is enforced by construction. The construction uses a free edge matrix B with zero diagonal and a vector of node potentials $r \in \mathbb{R}^m$. The adjacency matrix used inside the model is given by

$$A = B \odot M(r), \qquad M(r) = P(r)\, L\, P(r)^T,$$

where $P(r)$, L, and $M(r)$ are as defined in the main text. Note that a factorization where $q_\phi(A)$ assumes independence between r and γ prevents the model from capturing multiple modes in the DAG posterior. For example, if two DAGs fit the data equally well but imply different topological orders, an independent guide $q_\phi(r)\, q_\phi(\gamma)$ forces the model to average edge probabilities across incompatible orders, which can give a dense graph that does not coincide with any high posterior DAG. So to represent multiple modes, we model dependence between r and γ. Let $z \sim p_0(z)$

and $\gamma = T_\phi(z; r)$, where for each r, $T_\phi(\cdot; r)$ is a C^1 diffeomorphism. Assume $0 < p(\mathcal{D}) < \infty$, $q_\psi(r) \ll p(r)$, $q_\phi(\gamma \mid r) \ll p(\gamma)$, $q_\phi(\theta \mid A) \ll p(\theta \mid A)$, and integrability under $q_{\phi,\psi}$ of all displayed log-density terms, including the log likelihood. The induced conditional guide over γ is given by the change of variables

$$q_\phi(\gamma \mid r) = p_0\big(T_\phi^{-1}(\gamma; r)\big) \left|\det \frac{\partial T_\phi^{-1}(\gamma; r)}{\partial \gamma}\right| \;=\; p_0(z) \left|\det \frac{\partial T_\phi(z; r)}{\partial z}\right|^{-1}.$$

The hard DAG adjacency is a deterministic function of B and r. The joint model is

$$p(\mathcal{D}, B, r, \theta, \gamma) = p(\mathcal{D} \mid A, \theta)\, p(\theta \mid A)\, p(B \mid \gamma)\, p(r)\, p(\gamma), \tag{A.2}$$

where $p(B \mid \gamma)$ is the Bernoulli conditional and $p(r)$ is the potential prior. Note that while the guide q_ϕ conditions γ on r, the prior $p(r)p(\gamma)$ typically factorizes unless specific domain knowledge suggests otherwise. We have the joint guide as

$$q_{\phi,\psi}(\gamma, B, r, \theta) = q_\phi(\gamma \mid r)\, p(B \mid \gamma)\, q_\psi(r)\, q_\phi(\theta \mid A), \tag{A.3}$$

to avoid the collapse to a single mode over DAGs. By definition,

$$\mathrm{ELBO}(\phi, \psi) = \mathbb{E}_{q_{\phi,\psi}(\gamma, B, r, \theta)}\Big[\log p(\mathcal{D}, B, r, \theta, \gamma) - \log q_{\phi,\psi}(\gamma, B, r, \theta)\Big].$$

Using Eq. A.2 and Eq. A.3 we expand

$$\begin{aligned}\mathrm{ELBO}(\phi, \psi) = \mathbb{E}_{q_\psi(r)\, q_\phi(\gamma|r)\, p(B|\gamma)\, q_\phi(\theta|A)}\Big[&\log p(\mathcal{D} \mid A, \theta) + \log p(\theta \mid A) + \log p(B \mid \gamma) \\ &+ \log p(r) + \log p(\gamma) - \log q_\phi(\theta \mid A) - \log p(B \mid \gamma) \\ &- \log q_\psi(r) - \log q_\phi(\gamma \mid r)\Big].\end{aligned} \tag{A.4}$$

We first integrate over θ conditionally on the hard adjacency A. The terms that depend on θ are $\log p(\mathcal{D} \mid A, \theta) + \log p(\theta \mid A) - \log q_\phi(\theta \mid A)$. Therefore

$$\begin{aligned}\mathrm{ELBO}(\phi, \psi) = \mathbb{E}_{q_\psi(r)\, q_\phi(\gamma|r)\, p(B|\gamma)}\Big[&\mathbb{E}_{q_\phi(\theta|A)}\big[\log p(\mathcal{D} \mid A, \theta)\big] - D_{\mathrm{KL}}\big(q_\phi(\theta \mid A) \,\|\, p(\theta \mid A)\big) \\ &+\; \log p(B \mid \gamma) - \log p(B \mid \gamma) + \;\log p(r) - \log q_\psi(r)\Big] \\ &- \mathbb{E}_{q_\psi(r)}\big[D_{\mathrm{KL}}\big(q_\phi(\gamma \mid r) \,\|\, p(\gamma)\big)\big].\end{aligned} \tag{A.5}$$

The model and guide share the same hard Bernoulli conditional $p(B \mid \gamma)$, so its two log-density terms cancel exactly. Separating the remaining terms for r gives

$$\begin{aligned}\mathrm{ELBO}(\phi, \psi) = \mathbb{E}_{q_\psi(r)\, q_\phi(\gamma|r)\, p(B|\gamma)}\Big[&\mathbb{E}_{q_\phi(\theta|A)}\big[\log p(\mathcal{D} \mid A, \theta)\big] - D_{\mathrm{KL}}\big(q_\phi(\theta \mid A) \,\|\, p(\theta \mid A)\big)\Big] \\ &-\; D_{\mathrm{KL}}\big(q_\psi(r) \,\|\, p(r)\big) \;-\; \mathbb{E}_{q_\psi(r)}\big[D_{\mathrm{KL}}\big(q_\phi(\gamma \mid r) \,\|\, p(\gamma)\big)\big].\end{aligned} \tag{A.6}$$

The domain informed prior is now incorporated in $p(\gamma)$ (see Derivation A.4) as

$$p(\gamma) = \prod_{i \neq j} \frac{1}{\mathcal{B}(\alpha_{ij}, \beta_{ij})}\, \sigma(\gamma_{ij})^{\alpha_{ij}} \big(1 - \sigma(\gamma_{ij})\big)^{\beta_{ij}}, \qquad \sigma(x) = \frac{1}{1 + e^{-x}},$$

Using Eq. A.3, we reparameterize the expectation over $q_\phi(\gamma \mid r)$ by $z \sim p_0(z)$ and $\gamma = T_\phi(z; r)$. The conditional KL divergence for the edge logits becomes

$$\begin{aligned}D_{\mathrm{KL}}\big(q_\phi(\gamma \mid r) \,\|\, p(\gamma)\big) &= \mathbb{E}_{z \sim p_0}\Big[\log q_\phi\big(T_\phi(z; r) \mid r\big) - \log p\big(T_\phi(z; r)\big)\Big] \\ &= \mathbb{E}_{z \sim p_0}\Big[\log p_0(z) - \log\Big|\det \frac{\partial T_\phi(z; r)}{\partial z}\Big| - \log p\big(T_\phi(z; r)\big)\Big].\end{aligned}$$

Substituting above equation into Eq. A.6 and changing variables in the inner expectations we get the reparameterized bound

$$\begin{aligned}\mathrm{ELBO}(\phi, \psi) = \mathbb{E}_{r \sim q_\psi(r)}\, \mathbb{E}_{z \sim p_0}\, \mathbb{E}_{B \sim p(\cdot \mid T_\phi(z; r))}\Big[&\underbrace{\mathbb{E}_{q_\phi(\theta|A)}\big[\log p(\mathcal{D} \mid A, \theta)\big]}_{\text{expected log likelihood}} - \underbrace{D_{\mathrm{KL}}\big(q_\phi(\theta \mid A) \,\|\, p(\theta \mid A)\big)}_{\text{KL divergence for } \theta} \\ &+\; \log\Big|\det \frac{\partial T_\phi(z; r)}{\partial z}\Big| - \log p_0(z) + \log p\big(T_\phi(z; r)\big)\Big] \;-\; \underbrace{D_{\mathrm{KL}}\big(q_\psi(r) \,\|\, p(r)\big)}_{\text{KL divergence for potentials}}.\end{aligned} \tag{A.7}$$

This is our revised ELBO.

A.4 Logistic–Beta Prior on Edge Logits induced by the Domain Informed Prior

For each ordered pair (i, j) with $i \neq j$, let the edge probability $\pi_{ij} \in (0, 1)$ carry a domain informed Beta prior

$$\pi_{ij} \sim \mathrm{Beta}(\alpha_{ij}, \beta_{ij}), \qquad p(\pi_{ij}) = \frac{1}{\mathcal{B}(\alpha_{ij}, \beta_{ij})}\, \pi_{ij}^{\alpha_{ij}-1}(1-\pi_{ij})^{\beta_{ij}-1},$$

with hyperparameters constructed from our beliefs $p_{ij} \in (0, 1)$ and concentration $\nu_{ij} > 0$ via

$$\alpha_{ij} = \nu_{ij} p_{ij} + 1, \qquad \beta_{ij} = \nu_{ij}(1-p_{ij}) + 1.$$

In particular $\alpha_{ij} > 0$ and $\beta_{ij} > 0$ (indeed $\alpha_{ij}, \beta_{ij} \geq 1$), so the Beta density is proper. Let's define the edge logit

$$\gamma_{ij} = \mathrm{logit}(\pi_{ij}) = \log \frac{\pi_{ij}}{1-\pi_{ij}} \quad \Longleftrightarrow \quad \pi_{ij} = \sigma(\gamma_{ij}) = \frac{1}{1+\mathrm{e}^{-\gamma_{ij}}}.$$

The map $\gamma_{ij} \mapsto \sigma(\gamma_{ij})$ is a C^1 diffeomorphism from $\mathbb{R}$ onto $(0, 1)$, with strictly positive derivative, so the change of variables formula for densities applies. The Jacobian factor is

$$\left|\frac{d\pi_{ij}}{d\gamma_{ij}}\right| = \sigma(\gamma_{ij})\big(1-\sigma(\gamma_{ij})\big), \tag{A.8}$$

which is positive, so the absolute value may be dropped. By the change of variables formula, the prior density of edge logit γ_{ij} induced by the Beta prior on π_{ij} is

$$\begin{aligned} p(\gamma_{ij}) &= p(\pi_{ij})\Big|_{\pi_{ij}=\sigma(\gamma_{ij})} \left|\frac{d\pi_{ij}}{d\gamma_{ij}}\right| \\ &= \frac{1}{\mathcal{B}(\alpha_{ij}, \beta_{ij})} \underbrace{\sigma(\gamma_{ij})^{\alpha_{ij}-1}}_{\text{from } p(\pi_{ij})} \underbrace{\big(1-\sigma(\gamma_{ij})\big)^{\beta_{ij}-1}}_{\text{from } p(\pi_{ij})} \underbrace{\sigma(\gamma_{ij})\big(1-\sigma(\gamma_{ij})\big)}_{\text{Jacobian in equation A.8}} \\ &= \frac{1}{\mathcal{B}(\alpha_{ij}, \beta_{ij})}\, \sigma(\gamma_{ij})^{\alpha_{ij}} \big(1-\sigma(\gamma_{ij})\big)^{\beta_{ij}}. \end{aligned}$$

This is the Logistic-Beta density on edge logits. It is a proper density since $\gamma_{ij} \mapsto \sigma(\gamma_{ij})$ is a C^1 diffeomorphism of $\mathbb{R}$ onto $(0, 1)$ with Jacobian equation A.8, the substitution $\pi_{ij} = \sigma(\gamma_{ij})$ gives

$$\int_{\mathbb{R}} p(\gamma_{ij})\, d\gamma_{ij} = \int_0^1 p(\pi_{ij})\, d\pi_{ij} = 1.$$

Assuming independence across ordered pairs in the prior,

$$p(\gamma) = \prod_{i \neq j} p(\gamma_{ij}) = \prod_{i \neq j} \frac{1}{\mathcal{B}(\alpha_{ij}, \beta_{ij})}\, \sigma(\gamma_{ij})^{\alpha_{ij}} \big(1-\sigma(\gamma_{ij})\big)^{\beta_{ij}}.$$

A.5 Closed Form KL for Gaussian $p(\theta \mid A)$

Let d_θ be the dimension of θ. Suppose

$$q_\phi(\theta \mid A) \;=\; \mathcal{N}\big(\mu_q(A),\, \Sigma_q(A)\big), \qquad p(\theta \mid A) \;=\; \mathcal{N}\big(\mu_p(A),\, \Sigma_p(A)\big),$$

where $\Sigma_q(A) = \mathrm{diag}\big(\sigma_\phi^2(A)\big)$ and $\Sigma_p(A) = \mathrm{diag}\big(\sigma_p^2(A)\big)$ are diagonal with strictly positive entries $\sigma_{\phi,k}(A) > 0$ and $\sigma_{p,k}(A) > 0$ for all k (so that Σ_q, Σ_p are positive definite and invertible), and $\mu_q(A) = \mu_\phi(A)$. These conditions hold in our model, since $\sigma_\phi = \mathrm{softplus}(\rho_\phi) > 0$ and σ_p is a fixed positive prior scale. We derive a closed form for

$$D_{\mathrm{KL}}\big(q_\phi(\theta \mid A) \,\|\, p(\theta \mid A)\big) \;=\; \mathbb{E}_{q_\phi(\theta|A)}\big[\log q_\phi(\theta \mid A) - \log p(\theta \mid A)\big].$$

For any mean μ and positive definite covariance Σ,

$$\log \mathcal{N}(\theta; \mu, \Sigma) = -\frac{1}{2}\Big(d_\theta \log(2\pi) + \log|\Sigma| + (\theta-\mu)^\top \Sigma^{-1}(\theta-\mu)\Big).$$

Therefore,

$$\log q_\phi(\theta \mid A) = -\frac{1}{2}\Big(d_\theta \log(2\pi) + \log|\Sigma_q| + (\theta-\mu_q)^\top \Sigma_q^{-1}(\theta-\mu_q)\Big),$$
$$\log p(\theta \mid A) = -\frac{1}{2}\Big(d_\theta \log(2\pi) + \log|\Sigma_p| + (\theta-\mu_p)^\top \Sigma_p^{-1}(\theta-\mu_p)\Big),$$

where we suppress the explicit dependence on A inside $\mu_q, \mu_p, \Sigma_q, \Sigma_p$ for readability. Now taking the expectation under $q_\phi(\theta \mid A)$, the $d_\theta \log(2\pi)$ terms cancel and we have

$$\begin{aligned} D_{\mathrm{KL}}(q\|p) &= \mathbb{E}_q[\log q - \log p] \\ &= \frac{1}{2}\Big(\log|\Sigma_p| - \log|\Sigma_q|\Big) + \frac{1}{2}\mathbb{E}_q\big[(\theta-\mu_p)^\top \Sigma_p^{-1}(\theta-\mu_p)\big] \\ &\quad - \frac{1}{2}\mathbb{E}_q\big[(\theta-\mu_q)^\top \Sigma_q^{-1}(\theta-\mu_q)\big]. \end{aligned} \tag{A.9}$$

Let $\delta = \mu_q - \mu_p$. Since $(\theta-\mu_p) = (\theta-\mu_q) + \delta$ and Σ_p^{-1} is symmetric, we have

$$(\theta-\mu_p)^\top \Sigma_p^{-1}(\theta-\mu_p) = (\theta-\mu_q)^\top \Sigma_p^{-1}(\theta-\mu_q) + 2\,\delta^\top \Sigma_p^{-1}(\theta-\mu_q) + \delta^\top \Sigma_p^{-1}\delta.$$

Taking $\mathbb{E}_q$ and using the Gaussian moments $\mathbb{E}_q[\theta-\mu_q] = 0$ and $\mathbb{E}_q\big[(\theta-\mu_q)(\theta-\mu_q)^\top\big] = \Sigma_q$ (so that $\mathbb{E}_q[(\theta-\mu_q)^\top M(\theta-\mu_q)] = \mathrm{tr}(M\Sigma_q)$ for any fixed M) gives the standard identity:

$$\mathbb{E}_q\big[(\theta-\mu_p)^\top \Sigma_p^{-1}(\theta-\mu_p)\big] = \mathrm{tr}\big(\Sigma_p^{-1}\Sigma_q\big) + \delta^\top \Sigma_p^{-1}\delta. \tag{A.10}$$

Similarly,

$$\mathbb{E}_q\big[(\theta-\mu_q)^\top \Sigma_q^{-1}(\theta-\mu_q)\big] = \mathrm{tr}\big(\Sigma_q^{-1}\Sigma_q\big) = d_\theta. \tag{A.11}$$

Now plugging Eq. A.10 and Eq. A.11 into Eq. A.9 gives

$$D_{\mathrm{KL}}(q\|p) = \frac{1}{2}\Big(\log|\Sigma_p| - \log|\Sigma_q| - d_\theta + \mathrm{tr}\big(\Sigma_p^{-1}\Sigma_q\big) + (\mu_q-\mu_p)^\top \Sigma_p^{-1}(\mu_q-\mu_p)\Big). \tag{A.12}$$

Since $\Sigma_q = \mathrm{diag}(\sigma_\phi^2)$ and $\Sigma_p = \mathrm{diag}(\sigma_p^2)$ are diagonal with strictly positive entries, we have $\Sigma_p^{-1} = \mathrm{diag}(1/\sigma_p^2)$ and

$$\log|\Sigma_p| - \log|\Sigma_q| = \sum_{k=1}^{d_\theta} \log \sigma_{p,k}^2 - \sum_{k=1}^{d_\theta} \log \sigma_{\phi,k}^2 = 2\sum_{k=1}^{d_\theta} \log \frac{\sigma_{p,k}}{\sigma_{\phi,k}},$$

where the last equality uses $\log \sigma^2 = 2\log\sigma$ (valid since $\sigma_{p,k}, \sigma_{\phi,k} > 0$), together with

$$\mathrm{tr}\big(\Sigma_p^{-1}\Sigma_q\big) = \sum_{k=1}^{d_\theta} \frac{\sigma_{\phi,k}^2}{\sigma_{p,k}^2}, \qquad (\mu_q-\mu_p)^\top \Sigma_p^{-1}(\mu_q-\mu_p) = \sum_{k=1}^{d_\theta} \frac{(\mu_{\phi,k}-\mu_{p,k})^2}{\sigma_{p,k}^2}.$$

Substituting the diagonal expressions into Eq. A.12 and restoring the explicit A dependence we get,

$$D_{\mathrm{KL}}\big(q_\phi(\theta \mid A)\,\|\,p(\theta \mid A)\big) = \frac{1}{2}\sum_{k=1}^{d_\theta}\left[\frac{\sigma_{\phi,k}^2(A)}{\sigma_{p,k}^2(A)} + \frac{\big(\mu_{\phi,k}(A)-\mu_{p,k}(A)\big)^2}{\sigma_{p,k}^2(A)} - 1 + 2\log\frac{\sigma_{p,k}(A)}{\sigma_{\phi,k}(A)}\right].$$

This is the desired closed form.

A.6 Pseudo Code and Computational Complexity

Algorithm 1 SVI−DAG Algorithm

Inputs: Dataset $\mathcal{D}$; priors $(\alpha_{ij}, \beta_{ij})$; flow T_ϕ; hypernetwork H_ϕ; base $p_0(z)$; iterations T_{itr}; Sinkhorn iterations K_S; objective draws S_{MC}; temperatures $T > 0$, $\tau > 0$; particles K_{part}; step sizes η_ϕ, η_r

Outputs: Variational parameters ϕ and posterior particles $\{r^{(k)}\}_{k=1}^{K_{\text{part}}}$

Initialize parameters ϕ and particles $\{r^{(k)}\}_{k=1}^{K_{\text{part}}} \sim N(0, \sigma_r^2 I_m)$
for $t = 1, \dots, T_{\text{itr}}$ **do**
  Compute particle wise objectives and gradients
  for $k = 1, \dots, K_{\text{part}}$ **do**
    Compute $P_{\tau,K_S}(r^{(k)}) \leftarrow \mathcal{S}_{K_S}\big(\exp(S_0(r^{(k)})/\tau)\big)$.
    Compute $M(r^{(k)}) \leftarrow P(r^{(k)}) L P(r^{(k)})^T$.
    Compute $M_{\tau,K_S}(r^{(k)}) \leftarrow P_{\tau,K_S}(r^{(k)})\, L\, P_{\tau,K_S}(r^{(k)})^T$.
    for $s = 1, \dots, S_{\text{MC}}$ **do**
      Sample $z^{(k,s)} \sim p_0$ and set $\gamma^{(k,s)} \leftarrow T_\phi(z^{(k,s)}; r^{(k)})$.
      Sample $B^{(k,s)} \leftarrow \mathbf{1}\{\gamma^{(k,s)} + R^{(k,s)} > 0\}$. ▷ *Eq. 1*
      Sample $\tilde{B}^{(k,s)} \leftarrow \sigma((\gamma^{(k,s)} + R^{(k,s)})/T)$. ▷ *Eq. 1*
      Set $A^{(k,s)} \leftarrow B^{(k,s)} \odot M(r^{(k)}) \odot (\mathbf{1}_{m\times m} - I_m)$.
      Set $\tilde{A}^{(k,s)} \leftarrow \tilde{B}^{(k,s)} \odot M_{\tau,K_S}(r^{(k)}) \odot (\mathbf{1}_{m\times m} - I_m)$.
      $A_{\text{ST}}^{(k,s)} \leftarrow \tilde{A}^{(k,s)} + \text{sg}(A^{(k,s)} - \tilde{A}^{(k,s)})$. ▷ *Eq. 2*
      $(\mu_\phi^{(k,s)}, \rho_\phi^{(k,s)}) \leftarrow H_\phi(A_{\text{ST}}^{(k,s)})$ and set $\sigma_\phi^{(k,s)} \leftarrow \text{softplus}(\rho_\phi^{(k,s)})$.
      Sample $\epsilon^{(k,s)} \sim \mathcal{N}(0, I_{d_\theta})$ and set $\theta^{(k,s)} \leftarrow \mu_\phi^{(k,s)} + \sigma_\phi^{(k,s)} \odot \epsilon^{(k,s)}$.
      Draw a minibatch $\mathcal{D}_b^{(k,s)} \subset \{1, \dots, N\}$ and compute

$$\widehat{\mathcal{L}}_{\text{ELL}}^{(k,s)} \leftarrow \frac{N}{|\mathcal{D}_b^{(k,s)}|} \sum_{i=1}^{m} \sum_{n \in \mathcal{D}_b^{(k,s)}} \log p_i\big(x_i^{(n)} \mid g_i(x^{(n)}; \theta_i^{(k,s)})\big).$$

      Compute $\mathcal{L}_{\text{KL}}^{(k,s)}(\phi; A_{\text{ST}}^{(k,s)})$ ▷ *Eq. 4*
      $\mathcal{L}_{\text{flow}}^{(k,s)} \leftarrow \log\left|\det \frac{\partial \gamma^{(k,s)}}{\partial z^{(k,s)}}\right| - \log p_0(z^{(k,s)}) + \log p(\gamma^{(k,s)})$.
      $\widehat{\mathcal{L}}^{(k,s)} \leftarrow \widehat{\mathcal{L}}_{\text{ELL}}^{(k,s)} - \mathcal{L}_{\text{KL}}^{(k,s)} + \mathcal{L}_{\text{flow}}^{(k,s)}$.
    $\widehat{\mathcal{L}}^{(k)} \leftarrow S_{\text{MC}}^{-1} \sum_{s=1}^{S_{\text{MC}}} \widehat{\mathcal{L}}^{(k,s)}$. ▷ *Eq. 5*
    Target score: $g^{(k)} \leftarrow \nabla_{r^{(k)}}^{\text{ST}}\big(\widehat{\mathcal{L}}^{(k)} + \log p(r^{(k)})\big)$.
    $\psi^{(k)} \leftarrow \text{vec}(M_{\tau,K_S}^{(k)})$.
  Update particles via SVGD in relaxed acyclicity space
  for $k = 1, \dots, K_{\text{part}}$ **do**
    $\Phi^{(k)} \leftarrow K_{\text{part}}^{-1} \sum_{j=1}^{K_{\text{part}}} \big[k_M(\psi^{(j)}, \psi^{(k)}) g^{(j)} + \nabla_{r^{(j)}} k_M(\psi^{(j)}, \psi^{(k)})\big]$.
    $r^{(k)} \leftarrow r^{(k)} + \eta_r \Phi^{(k)}$. ▷ *Eq. 6*
  Update ϕ with particles held fixed
  $\widehat{\mathcal{J}}_{\text{block}} \leftarrow K_{\text{part}}^{-1} \sum_k \widehat{\mathcal{L}}^{(k)}$.
  $\phi \leftarrow \phi + \eta_\phi \nabla_\phi^{\text{ST}} \widehat{\mathcal{J}}_{\text{block}}$.
return ϕ, $\{r^{(k)}\}_{k=1}^{K_{\text{part}}}$.

We include the pseudo code for SVI−DAG in Algorithm 1. We choose neural spline flows (Durkan et al., 2019) for T_ϕ as they provide expressive transformations while allowing tractable forward evaluation and log Jacobian computation. The most computationally intensive components are the formation of the relaxed acyclicity mask and particle interactions in the SVGD style update. For each particle, applying K_S Sinkhorn normalizations to an $m \times m$ matrix costs $\mathcal{O}(K_S m^2)$, giving $\mathcal{O}(K_{\text{part}} K_S m^2)$ work per training iteration. This is followed by the dense multiplication $P_{\tau,K_S} L P_{\tau,K_S}^T$ that forms the generally cyclic relaxed acyclicity mask, imposing a worst case cost

of $\mathcal{O}(K_{\text{part}}m^3)$. Forming the hard permutation mask can instead exploit permutation structure and does not create this cubic term. The SVGD style update requires pairwise kernel evaluations and gradients across the K_{part} particles. Since the kernel uses vectorized relaxed masks u of length m^2, this step adds $\mathcal{O}(K_{\text{part}}^2 m^2)$. The particle computations before the interaction step can be batched on a GPU. If we are to include the S_{MC} hard/soft edge and adjacency constructions, the overall algorithmic complexity is $\mathcal{O}(K_{\text{part}}m^3 + K_{\text{part}}K_S m^2 + K_{\text{part}}S_{\text{MC}}m^2 + K_{\text{part}}^2 m^2)$, which aligns with standard cubic complexity of modern approaches to causal discovery (Thompson et al., 2024; Lorch et al., 2021; Bello et al., 2022; Annadani et al., 2023). This excludes the architecture and minibatch dependent costs of the flow, hypernetwork, likelihood, and their reverse mode derivatives, so it is not a bound on total runtime.

B Supporting Statements and Proofs

Proposition B.1. *Let's fix $p \in (0,1)$ and set $\alpha_\nu = \nu p + 1$ and $\beta_\nu = \nu(1-p) + 1$ for $\nu > 0$. Let $\pi_\nu \sim \text{Beta}(\alpha_\nu, \beta_\nu)$ and conditionally on π_ν we draw $B_\nu \sim \text{Bernoulli}(\pi_\nu)$. Then $\pi_\nu \to p$ in probability as $\nu \to \infty$ and $\mathbb{P}(B_\nu = 1) = \mathbb{E}[\pi] \to p$. Hence the Beta Bernoulli distribution converges to $\text{Bernoulli}(p)$ in the large confidence limit.*

Proof. We have

$$\mathbb{E}[\pi_\nu] = \frac{\alpha_\nu}{\alpha_\nu + \beta_\nu} = \frac{\nu p + 1}{\nu + 2} \to p,$$

as $\nu \to \infty$. Also

$$\text{Var}(\pi_\nu) = \frac{\alpha_\nu \beta_\nu}{(\alpha_\nu + \beta_\nu)^2(\alpha_\nu + \beta_\nu + 1)} = \frac{\nu^2 p(1-p) + \nu + 1}{(\nu+2)^2(\nu+3)} \sim \frac{p(1-p)}{\nu} \to 0.$$

The bias-variance identity gives

$$\mathbb{E}\big[(\pi_\nu - p)^2\big] = \text{Var}(\pi_\nu) + \big(\mathbb{E}[\pi_\nu] - p\big)^2 = \text{Var}(\pi_\nu) + \left(\frac{1-2p}{\nu+2}\right)^2 \to 0.$$

By Markov's inequality, for every $\varepsilon > 0$,

$$\mathbb{P}\big(|\pi_\nu - p| > \varepsilon\big) \le \frac{\mathbb{E}[(\pi_\nu - p)^2]}{\varepsilon^2} \to 0.$$

Hence $\pi_\nu \to p$ in probability. Finally, by the law of iterated expectations,

$$\mathbb{P}(B_\nu = 1) = \mathbb{E}[\mathbb{P}(B_\nu = 1 \mid \pi_\nu)] = \mathbb{E}[\pi_\nu] = \frac{\nu p + 1}{\nu + 2} \to p.$$

Thus the marginal law of B_ν is $\text{Bernoulli}(\mathbb{E}[\pi_\nu])$ and converges to $\text{Bernoulli}(p)$. □

Proposition B.2. *Let $\mathcal{X}$ be a measurable space and let $T : \mathcal{X} \to \mathcal{Y}$ be a measurable map. Let q and p be probability measures on $\mathcal{X}$ with pushforwards q_T and p_T on $\mathcal{Y}$. Then*

$$D_{\text{KL}}(q \,\|\, p) \;\ge\; D_{\text{KL}}(q_T \,\|\, p_T), \tag{B.1}$$

with equality if and only if a version of the conditional distributions satisfies $q(\cdot \mid y) = p(\cdot \mid y)$ for q_T almost every y. Specifically, for the deterministic mapping $(B, r) \mapsto A = B \odot M(r)$, the divergence in the (B, r) space provides an upper bound for the divergence in the adjacency matrix space A.

Proof. Let $X \sim q$ and $Y = T(X)$. By disintegration, there exist conditional probability measures $q(\cdot \mid y)$ and $p(\cdot \mid y)$ such that

$$q(dx) = q(dx \mid y)\, q_T(dy), \qquad p(dx) = p(dx \mid y)\, p_T(dy).$$

We begin by expressing the total KL divergence on the space $\mathcal{X}$ as an integral:

$$D_{\text{KL}}(q \,\|\, p) = \int_{\mathcal{X}} \log\left(\frac{q(dx)}{p(dx)}\right) q(dx).$$

By substituting the disintegrated forms into the integrand, we have:

$$D_{\mathrm{KL}}(q \,\|\, p) = \int_{\mathcal{Y}} \int_{\mathcal{X}|y} \log \left(\frac{q_T(dy)\, q(dx \mid y)}{p_T(dy)\, p(dx \mid y)} \right) q(dx \mid y)\, q_T(dy).$$

We then use the properties of the logarithm to separate the marginal and conditional components:

$$D_{\mathrm{KL}}(q \,\|\, p) = \int_{\mathcal{Y}} \log \left(\frac{q_T(dy)}{p_T(dy)} \right) q_T(dy) + \int_{\mathcal{Y}} \left[\int_{\mathcal{X}|y} \log \left(\frac{q(dx \mid y)}{p(dx \mid y)} \right) q(dx \mid y) \right] q_T(dy).$$

The first term on the right hand side is the KL divergence between the pushforward measures, $D_{\mathrm{KL}}(q_T \,\|\, p_T)$. The second term represents the expected conditional KL divergence, which we denote as $E_{q_T}[D_{\mathrm{KL}}(q(\cdot \mid y) \,\|\, p(\cdot \mid y))]$. The Gibbs inequality states that the KL divergence between any two probability measures is non negative. Therefore, if we assume the measures are well defined, then the conditional term satisfies:

$$\int_{\mathcal{Y}} D_{\mathrm{KL}}(q(\cdot \mid y) \,\|\, p(\cdot \mid y))\, q_T(dy) \geq 0.$$

This non negativity implies that $D_{\mathrm{KL}}(q \,\|\, p) \geq D_{\mathrm{KL}}(q_T \,\|\, p_T)$ establishing the inequality. We can observe that for the inequality to become an equality, the conditional term must vanish. This occurs if and only if $D_{\mathrm{KL}}(q(\cdot \mid y) \,\|\, p(\cdot \mid y)) = 0$ for q_T almost every y, which is equivalent to the condition that $q(\cdot \mid y) = p(\cdot \mid y)$. In the context of our algorithm, this demonstrates that minimizing the divergence in the latent parameter space (B, r) minimizes a conservative upper bound on the divergence between the actual graph distributions. □

Theorem B.3. *Let $r \in \mathbb{R}^m$ have pairwise distinct entries and let $B \in \{0,1\}^{m \times m}$ have zero diagonal. Let's define $M(r) = P(r)LP(r)^T$ and $A = B \odot M(r)$. Then the directed graph with adjacency A is acyclic. Moreover for every binary DAG there exist potentials r and a free edge matrix B such that $A = B \odot M(r)$.*

Proof. See Annadani et al. (2023).

□

Lemma B.4. *Let $m \geq 2$ and let $r = (r_1, \ldots, r_m)^T \in \mathbb{R}^m$ have pairwise distinct entries. We define the score matrix $S_0(r) = r\, o^T$, where $o = (m, \ldots, 1)^T$. For $\tau > 0$ and $K_S \geq 1$, let $P_{\tau,K_S}(r)$ denote the output obtained after K_S alternating row and column normalizations applied to $\exp\big(S_0(r)/\tau\big)$, and let*

$$P^\star_\tau(r) = \lim_{K_S \to \infty} P_{\tau,K_S}(r)$$

be its Sinkhorn limit, which exists and is the unique doubly stochastic solution of the entropic assignment problem because $\exp(S_0(r)/\tau)$ has strictly positive entries (Sinkhorn–Knopp). Let $P(r)$ be the permutation matrix that sorts indices by descending potential, and set $M^\star_\tau(r) = P^\star_\tau(r)\, L\, P^\star_\tau(r)^T$ and $M(r) = P(r)\, L\, P(r)^T$, where L is the strictly lower triangular matrix of ones. Then

$$P^\star_\tau(r) \to P(r) \quad \textit{and} \quad M^\star_\tau(r) \to M(r) \quad \textit{entrywise as } \tau \to 0.$$

Equivalently, for any iteration schedule $K_S(\tau) \to \infty$ with $\big\|P_{\tau,K_S(\tau)}(r) - P^\star_\tau(r)\big\|_{\max} \to 0$ as $\tau \to 0$, one has $P_{\tau,K_S(\tau)}(r) \to P(r)$ and $P_{\tau,K_S(\tau)}(r)\, L\, P_{\tau,K_S(\tau)}(r)^T \to M(r)$ entrywise.

Proof. See Annadani et al. (2023). □

C Evaluation Metrics

We assess the quality of the learned posterior distribution using four metrics, each capturing a distinct aspect of structured recovery. Let $q(\mathcal{G})$ represent the approximate posterior over graphs and let $\mathcal{G}^{\mathrm{GT}}$ represent the ground truth causal graph. For each metric, we draw N_e posterior samples $\{\mathcal{G}^{(i)}\}_{i=1}^{N_e}$ with $\mathcal{G}^{(i)} \sim q(\mathcal{G})$ and report Monte Carlo estimates.

Expected SHD (E-SHD). The $\mathrm{SHD}(\mathcal{G}, \mathcal{G}^{\mathrm{GT}})$ counts the minimum number of edge additions, deletions, and reversals required to transform the estimated graph $\mathcal{G}$ into $\mathcal{G}^{\mathrm{GT}}$. Because our method returns a distribution rather than a point estimate, we report the expected SHD under the approximate posterior:

$$\text{E-SHD} = \mathbb{E}_{\mathcal{G}\sim q(\mathcal{G})}\big[\mathrm{SHD}(\mathcal{G},\, \mathcal{G}^{\mathrm{GT}})\big] \;\approx\; \frac{1}{N_e}\sum_{i=1}^{N_e} \mathrm{SHD}\big(\mathcal{G}^{(i)},\, \mathcal{G}^{\mathrm{GT}}\big).$$

Lower values indicate that posterior samples are, on average, structurally closer to the ground truth. This metric penalizes both missing and spurious edges as well as orientation errors, providing a comprehensive measure of structural accuracy.

Expected F1 score. For each posterior sample $\mathcal{G}^{(i)}$, we treat every ordered pair (i, j) as a binary classification problem: an edge is either present or absent. Precision is the fraction of predicted edges that appear in $\mathcal{G}^{\mathrm{GT}}$, recall is the fraction of true edges recovered by $\mathcal{G}^{(i)}$, and the F1 score is their harmonic mean. The expected F1 score averages this quantity over the approximate posterior:

$$\text{E-F1} = \mathbb{E}_{\mathcal{G}\sim q(\mathcal{G})}\big[\mathrm{F1}(\mathcal{G},\, \mathcal{G}^{\mathrm{GT}})\big] \;\approx\; \frac{1}{N_e}\sum_{i=1}^{N_e} \mathrm{F1}\big(\mathcal{G}^{(i)},\, \mathcal{G}^{\mathrm{GT}}\big).$$

Higher values reflect better balance between precision and recall across posterior samples.

Brier score. The Brier score evaluates the calibration of the posterior edge probabilities by measuring the mean squared error between the marginal posterior edge probabilities and the ground truth binary edge indicators. Let $\hat{p}_{ij} = \mathbb{E}_{\mathcal{G}\sim q(\mathcal{G})}[A_{ij}]$ denote the marginal posterior probability of edge (i, j) and let $A_{ij}^{\mathrm{GT}} \in \{0, 1\}$ denote the corresponding ground truth edge indicator. The Brier score is defined as

$$\mathrm{Brier} = \frac{1}{m(m-1)}\sum_{i\neq j}\left(\hat{p}_{ij} - A_{ij}^{\mathrm{GT}}\right)^2.$$

A lower Brier score indicates superior uncertainty quantification. A perfectly calibrated posterior that assigns probability one to true edges and zero to absent edges achieves a Brier score of zero.

AUROC. The AUROC measures the discriminative ability of the posterior edge probabilities $\{\hat{p}_{ij}\}_{i\neq j}$ in distinguishing true edges from absent ones, treating the marginal probabilities as scores for a binary classifier. It equals the probability that a randomly chosen true edge receives a higher posterior probability than a randomly chosen absent edge. An AUROC of 100% indicates perfect separation, while 50% corresponds to chance level discrimination. Unlike the Brier score, the AUROC is threshold free and invariant to monotone transformations of the scores, making it a complementary measure of ranking quality.

CPDAG evaluation. In settings like linear synthetic data, where the DAG is identifiable only up to its MEC, point wise comparison against a single ground truth DAG would unfairly penalize methods that correctly recover the equivalence class but select a different representative. We therefore follow standard practice and convert both the ground truth DAG and each posterior sample to their corresponding CPDAGs (Peters et al., 2017) before computing the metrics above. A CPDAG represents a MEC by directing only those edges whose orientation is shared by every member of the class and leaving the remaining edges undirected. All four metrics are then evaluated at the CPDAG level, ensuring that methods are not penalized for orientation ambiguities that are inherently unresolvable from observational data alone.

D CODE AND LICENSE

We use the following open source repositories for comparison with the baselines:

- ProDAG: https://github.com/ryan-thompson/ProDAG.jl (MIT license).
- BayesDAG: https://github.com/microsoft/Project-BayesDAG (MIT license).
- BCD: https://github.com/ermongroup/BCD-Nets (No license included).

- DIBS: https://github.com/larslorch/dibs (MIT license).
- DDS/VI-DP-DAG https://github.com/sharpenb/Differentiable-DAG-Sampling (No license included).

E Detailed Model Specifications and Hyperparameters.

The common hyperparameters used for benchmarking the algorithms across all experiments are summarized in Tables E.1–E.6.

Table E.1: Architectural configurations used for `DiBS` algorithm

Hyperparameter	Value
No. of SVGD steps	2000
No. of particles (K_{part})	20
Optimizer	RMSProp
Step size	0.005
No. of likelihood gradient MC samples	128
No. of acyclicity gradient MC samples	32
Graph prior	Erdős–Rényi
SVGD	
Type	Additive Frobenius squared exponential
Latent bandwidth	5.0
Latent graph relaxation	
Acyclicity penalty slope	1.0
Gumbel-softmax temperature	1.0

Table E.2: Architectural configurations used for `BayesDAG` algorithm

Hyperparameter	Value
Number of epochs	150
Batch size	128
Learning rate	0.0003
Optimizer	Adam (Kingma & Ba, 2014)
Stochastic gradient MCMC sampler	
No. of chains	10
Noise scale on permutation	0.1
Noise scale on weights	0.01
Nonlinear network architecture	
No. of hidden layers	2
Neurons per hidden layer	128
Activation function	ReLU
Layer normalization	True
Residual connections	True
Sinkhorn	
Sinkhorn iterations (K_S)	100
Sparsity parameter tuning	
Grid	4 values between $\lambda_{\min}$ and $\lambda_{\max}$
$(\lambda_{\min}, \lambda_{\max})$	$(10, 10^3)$
Grid proxy fits	25 epochs, 64 posterior samples
Selection of λ	Closest to true sparsity level
Validation set size	$\lfloor 0.1n \rfloor$

Table E.3: Architectural configurations used for `SVI-DAG` algorithm

Hyperparameter	Value
Batch size	64
Optimizer	Adam (Kingma & Ba, 2014)
Gradient clipping (global norm)	1.0
Straight through mask	True
No. of ELBO MC samples	1
Temperatures (T, τ_{final})	(0.3, 0.1)
Sinkhorn iterations (K_S)	100
Domain informed prior	
ν_{min}	1
κ	500
η	2
Normalizing flow	
Type	Neural Spline Flows (coupling) (Durkan et al., 2019)
No. of flows	5
No. of bins	8
Domain boundary	5
SVGD	
Type	Radial Basis Function
σ_r	1.0
Particle gradient clip (ℓ_2 norm)	10

Table E.4: Architectural configurations used for `ProDAG` algorithm

Hyperparameter	Value
Optimizer	Adam (Kingma & Ba, 2014)
Learning rate	0.1
Max no. of epochs	1000
Early-stopping patience	5
No. of ELBO MC samples	100
Acyclicity threshold	0.1
Acyclicity projection	
Total steps	10
Initial penalty ($\mu^{(1)}$)	1
Penalty update	$\mu^{(t+1)} = \mu^{(t)}/2$
Initialization ($W^{(0)}$)	0
Sparsity parameter grid	
Number of grid values	10
λ^{min}	0
λ^{max}	Average l_1 ball with $\lambda = \infty$
Selection of λ	Held-out validation MSE ($\lfloor 0.1n \rfloor$ samples)

Table E.5: Architectural configurations used for `BCD Nets` algorithm

Hyperparameter	Value
No. of steps	15000
Batch size	256
Learning rate	0.001
Optimizer	AdaBelief
Permutation network (P)	
Type	MLP
No. of hidden layers	2
Hidden size	128
Activation function	GeLU
Gumbel–Sinkhorn relaxation	
Temperature	Annealed $30 \to 10 \to 1$
Sinkhorn doubly-stochastic tolerance	0.01
Sampling mode	Hard (straight-through)
Variational posterior over L	
L parameterisation	means | log-stds (lower triangle + noise)
Per-node noise scales	True
Priors	
Lower-triangular weight prior	Horseshoe
Horseshoe scale	van der Pas et al. (n-independent)
Expected in-degree ($\deg$)	1
Log-noise prior std (s)	3.0

Table E.6: Architectural configurations used for `DDS` algorithm

Hyperparameter	Value
Max no. of epochs	100
Early-stopping patience	20
Batch size	64
Learning rate (autoencoder)	0.001
Learning rate (DAG sampler)	0.01
Optimizer	Adam (Kingma & Ba, 2014)
Probabilistic DAG sampler	
Order type	Top-k (SoftSort)
Sampling mode	Hard (straight-through)
Gumbel temperature	1.0
Noise factor	1.0
Masked autoencoder	
Architecture	Linear
Hidden layer sizes	$(16, 16, 16)$
Loss & edge prior	
Loss	ELBO
Sparsity regulariser (Bernoulli KL) weight	0.1
Prior edge probability (p)	0.01

F Additional Results

F.1 Results on larger graphs

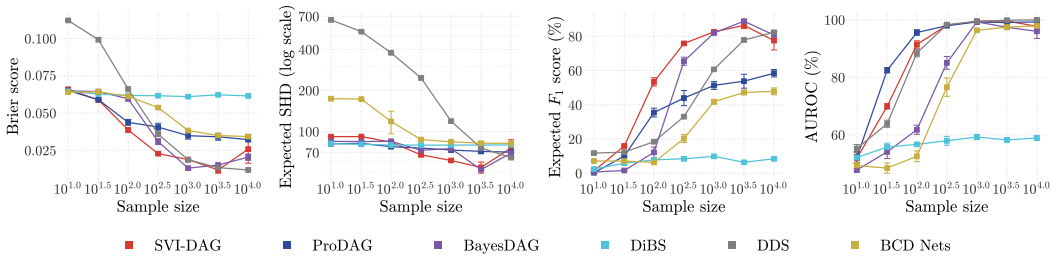

Figure F.1: Results on synthetic data generated from linear ER DAGs with $p = 50$ nodes and $s = 80$ edges under Gaussian noise. Each point denotes sample mean, and accompanying error bars indicate standard error, both computed across 5 independent draws of the dataset.

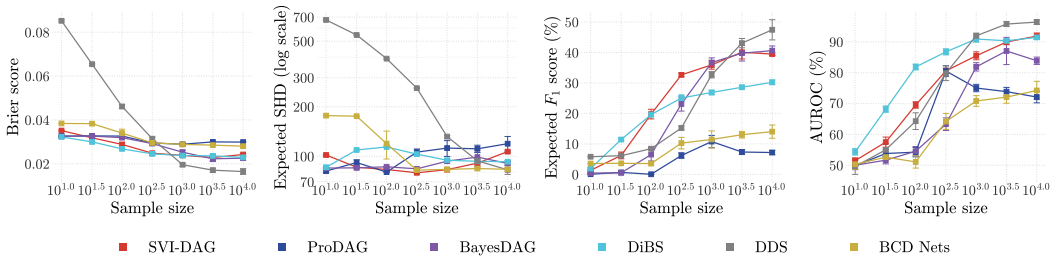

Figure F.2: Results on synthetic data generated from nonlinear ER DAGs with $p = 50$ nodes and $s = 80$ edges under Gaussian noise. Each point denotes sample mean, and accompanying error bars indicate standard error, both computed across 5 independent draws of the dataset.

G Broader Impact & Limitations

G.0.1 Broader impact statement

This work is concerned with understanding cause and effect relationships from data, with potential applications across empirical sciences, economics, epidemiology, and climate science. By explicitly quantifying epistemic uncertainty over causal structures and providing a principled mechanism to incorporate domain knowledge, our approach can support more calibrated decision making, expose biases present in the data, and allow reliable answers to causal queries in settings where data are scarce or the underlying DAG is not identifiable. As such, we envision this line of work to not have any significant negative impact.

G.0.2 Limitations

Though our approach has several strengths, it naturally has limitations. As with most Bayesian approaches that rely on variational inference, our framework lacks theoretical guarantees for exact posterior approximation. In particular, while the combination of normalizing flows and SVGD in the relaxed acyclicity space is designed to promote mode coverage, our method does not guarantee recovery of all DAGs within MEC. It can only aim to capture as many of them as possible, and some equivalent structures may remain underrepresented in the learned posterior.