Generative Causal Structure Learning with Dual Latent Spaces and Annealing

Anonymous authors
Paper under double-blind review

Abstract

In this work, we address causal structure learning in the presence of unobserved confounders. Such causal structures can be represented by Acyclic Directed Mixed Graphs (ADMGs), where observed cause-effect relations are depicted by directed edges and unobserved confounded relations by bidirected edges. Prior methods for causal structure learning with unobserved common causes have primarily focused on search-based approaches, and more recently on flow-based generative models. We propose a novel generative method based on a variant of the Variational Autoencoder (VAE) with dual latent spaces to represent the directed cause-effect relations and the bidirected unobserved confounded relations, associating two trainable adjacency matrices. To enhance the learning process, we introduce a causality constraint combined with the concept of a causal annealing strategy during training, guiding the learning toward meaningful causal structures. Experimental results show that our method achieves competitive performance in identifying both observed and latent causal relationships on synthetic datasets. Furthermore, we demonstrate that the learned causal structure significantly improves downstream causal inference performance on real-world data.

1 Introduction

Learning cause-effect relationships with unknown causal structures is an important need for causal reasoning. In many realistic scenarios, cause-effect relationships among the different features are unavailable. The presence of unobserved common causes makes this learning further challenging. For example, in human mobility planning, weather conditions are a common influencer to the start time and next location to be visited, If weather is not part of the data, it acts as an unobserved confounder that affects the cause-effect relationship.

In our work, we aim to learn cause-effect relationships (causal structure) among the various features in the presence of unobserved (latent) confounders, without considering any prior knowledge of causal structure. Additionally, we apply our method to perform a causal inference task using a real-world dataset, leveraging the learned cause-effect relationships using the proposed method.

Established prior works on causal structure identification considering unobserved confounders primarily exploit score based approaches, using scoring functions like the Bayesian Information Criterion (BIC), and constraint-based methods applying conditional independence tests.

Subsequently, the causal structure discovery has been formulated as a continuous optimization problem enforcing acyclicity constraints (Zheng et al. (2018)) via differentiable function with gradient-based techniques, excluding the need for a combinatorial search. However, these continuous optimization based methods assume causal sufficiency and do not consider unobserved variables.

Recently, the **bow-free** constraint (Bhattacharya et al. (2021)) applied to Acyclic Directed Mixed Graphs (ADMGs) has been utilized to identify unobserved confounding relationships. A **bow-free** constraint requires that there are no overlapping directed and bidirected edges between the same pair of variables, i.e., if there is a directed edge $V_i \rightarrow V_j$, there cannot also be a bidirected edge $Vi \leftrightarrow V_j$ (Bhattacharya et al.

(2021)). This concept has been incorporated into the structural causal model (SCM) Pearl (1998) and has been further examined in the context of neural ADMG models. The authors have established the identifiability criteria for ADMGs (Ashman et al. (2023)) under a nonlinear SCM framework, using a nonlinear additive noise model. In this model, the effect variable is generated by adding an independent noise term to a nonlinear function of its causes. We follow this identifiability assumptions for ADMG learning.

ADMG is capable of representing both direct causal relationships, and unobserved confounding relations. The observed cause-effect relations are represented by directed edges. The unobserved confounders are represented by bidirected edges between the variables which have unobserved common cause that influences both variables. We represent the directed cause effect relationship by the directed adjacency matrix $A^{(\text{Observed})}$ or A_D, which captures **directed, asymmetric, observed cause-effect** relationships. The unobserved common cause association with different pair of variables is represented by the bidirected adjacency matrix $A^{(\text{Unobserved})}$ or A_B which captures **bidirected, symmetric, unobserved confounded** associations.

Our method employs a VAE based **causally-constrained** generative learning with the prime aim to identify the unobserved causal structure following the ADMG. We refer to our model as **G-ADMG-CL** (Generative Acyclic Mixed Graph based Causal Structure Learning). We assume that detecting unobserved confounders is possible with the assumption of bow-free ADMGs (Ashman et al. (2023)), using the nonlinear additive noise structural causal model. Our method is capable of identifying causal structures having nonlinear cause-effect relationships with **unobserved confounders** and also when all features are observed, with the **trainable adjacency matrices,** A_D**,** A_B representing the causal structure. It has the following characteristics:

- The proposed VAE based method disentangles observed/unobserved causal relations and confounders by learning **dual latent spaces**.
- We use dual latent spaces to separately cater to direct cause-effect relationships and unobserved confounded relations associated with the directed adjacency matrix (A_D), and the bidirected adjacency matrix (A_B) as discussed earlier.
- The proposed method introduces a **causally aware objective function**. This ensures acyclic constraint for observed cause-effect relationships, bow-free constraint for unobserved confounded relations, entropy constraint along with maintaining sparsity to balance between randomness and sparsity reduction, while maintaining the symmetric/asymmetric characteristics of A_B and A_D respectively.
- We propose **causal annealing** where we gradually increase the causal regularization weight, from 0 to 1 until the transition epoch **causal transition epoch (CTE)** to prioritize reconstruction and KL divergence components in the initial learning phase, before activating causality constraint.

Our method yields interpretable adjacency matrices that capture both observed and unobserved causal relationships. We validate our method on real-world and synthetic datasets generated using nonlinear structural causal models (SCM Pearl (2009)), demonstrating its effectiveness in learning causal structures under confounding.

2 Related Work

Causal structure learning is a fundamental problem with wide applications across scientific and real-world domains. Identifying causal relationships in the presence of *unobserved confounders* is both more realistic and more challenging. We review key existing approaches, with a particular focus on methods that attempt to model latent confounding and capture both directed and bidirected dependencies.

Score-based and constraint-based methods: Classical approaches to causal discovery (Hasan et al. (2025); Zanga & Stella (2023)) include constraint-based and score-based methods. Constraint-based methods, such as the PC and FCI algorithms, rely on conditional independence testing to infer causal graphs, while score-based methods search for graph structures that optimize predefined scoring criteria such as the Bayesian Information Criterion (BIC). However, these methods typically assume access to a fully observed variable set or struggle with indistinguishability in equivalence classes under hidden confounding.

Differentiable DAG learning (no latent confounding). The NOTEARS framework, Zheng et al. (2018) introduced a differentiable acyclicity constraint based on the trace of a matrix exponential, enabling gradient-based optimization for DAG structure learning. This approach has been extended in several neural network-based methods such as DAG-GNN Yu et al. (2019) and N-DAG-G Geffner et al. (2022), which integrate deep architectures for end-to-end causal discovery and inference. While effective for fully observed systems, these methods operate under the assumption of causal sufficiency and are unable to represent latent confounding or bidirected edges.

Handling unobserved confounders: Several methods explicitly attempt to model latent confounding. Fast Causal Inference (FCI) Spirtes et al. (2000) is a constraint-based method capable of detecting latent variables through conditional independence tests, but it cannot distinguish between equivalence classes with identical independencies. Repetitive Causal Discovery (RCD) Maeda & Shimizu (2020) handles linear non-Gaussian models and introduces bidirected edges to represent unobserved common causes. CAM-UV Maeda & Shimizu (2021) extends the Causal Additive Model (CAM) to latent-variable settings by using HSIC-based independence tests Gretton et al. (2008) combined with a scoring procedure. More recently, Gonzales & Valizadeh (2024) proposed a score-based global search over augmented DAGs that can handle latent confounding, but is restricted in scalability.

Ancestral graphs and bow-free constraints: To more explicitly model both directed and bidirected edges under latent confounding, recent works have extended the causal discovery framework to ADMGs. The approach in Bhattacharya et al. (2021) introduces a differentiable objective incorporating acyclicity, ancestral constraints, a c-tree penalty, and the bow-free constraint, targeting linear Gaussian additive noise models. These constraints help ensure identifiability and interpretability in the learned graphs. None of these methods utilizes a generative model.

Flow-based methods: Despite these advances, most existing methods are either search-based or rely on structural constraints but do not leverage a generative modeling perspective. A notable exception is Ashman et al. (2023), which introduces a flow-based auto-regressive model for learning ADMG structures with non-linear relationships. Their method supports both bow-free (N-BF-ADMG-G) and general ADMG settings (N-ADMG-G), and is capable of modeling latent confounders. However, it lacks latent-variable disentanglement and does not optimize for structure recovery through generation-based objectives.

3 Methodology

In this section, we elaborate the two tasks. The first, **causal structure learning** in the presence of both observed cause-effect and unobserved confounded relations, using the proposed model **G-ADMG-CL**.

The second task expands on this framework to focus on **prediction and causal inference**, using the proposed model. We refer to this second task as G-ADMG-CL+P. Once a causal graph is learned, we utilize it to support downstream applications, such as causal inference and the estimation of treatment effects (e.g., Average Treatment Effect, ATE). This extension integrates learned structural knowledge into the decoder, enabling the model to perform prediction and causal inference (e.g., estimating treatment effects).

We propose a novel generative framework based on a variational autoencoder (VAE) to jointly learn both directed and bidirected causal relationships in the presence of latent confounding. Our approach leverages dual latent spaces to disentangle observed cause-effect interactions from unobserved confounded associations. To guide structure learning, we introduce a causality-aware objective that incorporates acyclicity, bow-free constraints, and structural regularization. Furthermore, we introduce a novel *causal annealing* strategy that gradually enforces causal constraints during training, enabling more stable and accurate graph recovery. The goal of our method is to simultaneously learn both edge types while ensuring that the resulting structure is identifiable, interpretable, and aligned with the true underlying causal mechanisms.

Causal Structure Learning: We present the proposed G-ADMG-CL, a method for learning causal relationships that includes both observed cause-effect relations and unobserved confounding relationships, exploiting VAE based generative model, under the ADMG framework.

VAE is a popular generative model for learning disentangled latent representation Higgins et al. (2017). The probabilistic formulation of VAE allows incorporating constraints in the latent space while learning and

optimizing the loss function. Previously, VAE has been applied to learn observed linear causal relations Yang et al. (2023) in the absence of unobserved confounders.

The key innovations in our method are the introduction of dual latent spaces, causality constraint optimization, and causal annealing to guide causality learning. The dual latent spaces separately encode the directed causal effects and bidirected confounded relationships, each associated with the two distinct trainable adjacency matrices A_D (directed) and A_B (bidirected) to embed structural information into the representation. The causality constraint optimization regularizes the learning to capture causal structure accurately. Additionally, the proposed causal annealing approach further guides the causal structure learning to enforce causal constraints during training.

The functional components of the proposed method are illustrated in Figure 1. The input data is fed into the model. The encoder encodes the input data into two disentangled latent spaces, the directed latent space Z_D representing observed causal relations, and the bidirected latent space Z_B representing unobserved confounded relations. These latent representations are used to learn two corresponding trainable adjacency matrices, A_D, and A_B. Causal constraints are incorporated during the training to guide the learning process. The learned adjacency matrices are fed into the decoder/generator to reconstruct or generate the predicted data, and to obtain the inferred causal structure integrating both observed cause-effect and unobserved confounded relations. The predicted data is further utilized for causal inference task.

We present our detailed methodology as follows:

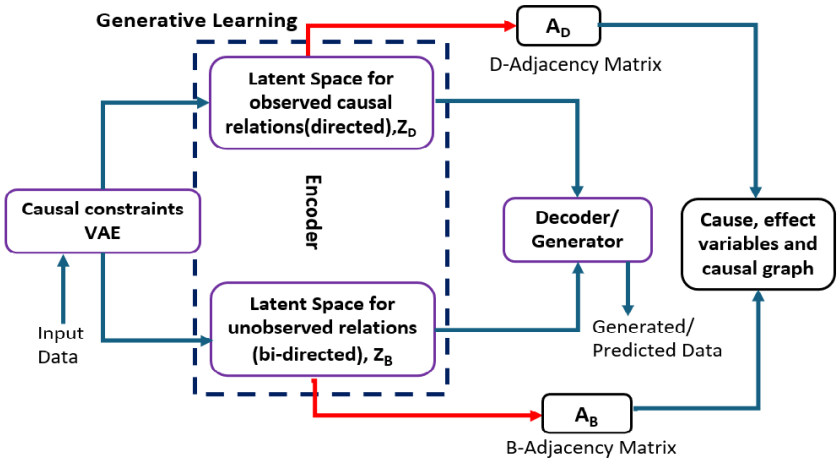


Figure 1: Functional Components: Dual latent spaces Z_D and Z_B, guided by trainable adjacency matrices A_D and A_B, enable causal structure learning.

Encoder: We use Variational Autoencoder (Doersch (2021), Kingma & Welling (2013)) based generative learner with the dual latent spaces to learn the observed and unobserved cause-effect relationships. The encoder component comprises two fully connected layers. The first dense layer has 128 units, and the 2nd layer has 64 units with ELU (exponential linear unit) as an activation function. Dropout is applied to prevent overfitting. The encoder comprises two latent spaces, the directed latent space z_{directed} to learn the observed causal relationship, and the other is the bi-directed latent space $z_{\text{bidirected}}$ which addresses unobserved or confounded relationship. The latent space representation for both the latent spaces obtained by computing mean $\mu_{\text{directed}}(\mu_D)$, $\mu_{\text{bidirected}}(\mu_B)$, and log-variance σ_{directed}, $\sigma_{\text{bidirected}}$ respectively with the reparameterization trick to sample from the latent space, first ϵ is sampled from $N(0,1)$ and then are computed as by the Eq. 2, where we compute $\mu_{\text{directed_guided_Adjacency}}(\mu_{DA_D})$ and $\mu_{\text{bidirected_guided_Adjacency}}(\mu_{BA_B})$ using Eq. 1 to guide z_{directed}, and $z_{\text{bidirected}}$ by the cause effect relationships learned by A_D, and unobserved confounded effect learned by A_B. The aim is to obtain the causally structured latent variables, improving reconstruction and making the trained model more interpretable by following the cause-effect relationships.

$$\mu_{DA_D} = \mu_{\text{directed}} \cdot \mathbf{A}_D, \mu_{BA_B} = \mu_{\text{bidirected}} \cdot \mathbf{A}_B. \tag{1}$$

$$
\begin{aligned}
z_{\text{directed}} &= \mu_{DA_D} + \sigma^{1/2}_{\text{directed}}(Y|X) \cdot \epsilon_{\text{directed}}, \\
z_{\text{bidirected}} &= \mu_{BA_B} + \sigma^{1/2}_{\text{bidirected}}(Y|X) \cdot \epsilon_{\text{bidirected}}.
\end{aligned} \tag{2}
$$

Decoder: The decoder component comprises two fully connected layers. The first dense layer has 64 units, and the 2nd layer has 128 units with LeakyReLU (rectified linear unit) as an activation function. Dropout is applied to obtain better generalization. **Decoder combines the causally guided dual latent spaces** $[z_{\textbf{directed}}, z_{\textbf{bidirected}}]$ from the Encoder and reconstructs the given input from the combined latent spaces retaining the causal and unobserved confounded relationship among the variables. This reconstructed outcome is further exploited for the **prediction task**.

Generator: The decoder is employed to generate new data samples by passing the latent samples sampled from a Gaussian distribution.

The **Loss function** ($\mathcal{L}_{\textbf{total}}$) as shown in Eq. (3) comprises three components, where the first two components are already present in the literature and are exploited by various VAE-based models to execute different tasks. We introduce a novel third component, the **causal mixed graph loss** $\left(\mathcal{L}_{\text{Causal_ADMG}}\right)$ weighted by the λ_{causal}.

The loss function follows the structure of ADMG. The dual latent spaces capture directed causal relationships (via a trainable matrix W_1) and bidirected (confounded) relationships (via W_2). Using this causality constraint, $\left(\mathcal{L}_{\text{Causal_ADMG}}\right)$ the training process enforces the structural regularization aligned properties of ADMGs, and learn causal relationships in the presence of unobserved confounders.

Total Loss:

$$
\begin{aligned}
\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reconstruction}} &+ \lambda_{\text{KL}}\left(\mathcal{L}_{\text{KL_directed}} + \mathcal{L}_{\text{KL_bidirected}}\right) \\
&+ \lambda_{\text{causal}}\left(\mathcal{L}_{\text{Causal_ADMG}}\right)
\end{aligned} \tag{3}
$$

1. **Reconstruction loss:** This is for reconstruction of the input data accurately from the latent spaces.

2. **KL divergence loss:** This is for latent space regularization, where, $\mathcal{L}_{\text{KL_directed}}$ (Eq. (4)) for the directed, and $\mathcal{L}_{\text{KL_bidirected}}$ (Eq. (5)) for the bidirected KL divergence. We apply KL annealing Li et al. (2019), kl_weight (λ_{KL}) (a gradually increasing weight) to multiply the KL divergence term to counter KL-vanishing during the initial training phase and focus primarily on minimizing the reconstruction error.

$$
\lambda_{\text{KL}} \cdot L_{\text{KL_directed}} = \lambda_{\text{KL}} \cdot \text{KL}\left(q_\phi(\mathbf{z}_{\text{directed}}|\mathbf{X}) \parallel p_\theta(\mathbf{z}_{\text{directed}}|\mathbf{X})\right), \tag{4}
$$

$$
\lambda_{\text{KL}} \cdot L_{\text{KL_bidirected}} = \lambda_{\text{KL}} \cdot \text{KL}\left(q_\phi(\mathbf{z}_{\text{bidirected}}|\mathbf{X}) \parallel p_\theta(\mathbf{z}_{\text{bidirected}}|\mathbf{X})\right) \tag{5}
$$

3. **Causal mixed graph loss** ($\mathcal{L}_{\text{Causal_ADMG}}$) : This is for causal structure regularization, with the causal regularization weight λ_{causal} as shown in Eq. (6).

$$
\begin{aligned}
\mathcal{L}_{\text{Causal_ADMG}} = \lambda_{\text{cycle}} \cdot \mathcal{L}_{\text{cycle}}(A_D) &+ \lambda_{\text{bow}} \cdot \mathcal{L}_{\text{bow}}(A_D, A_B) \\
&+ \lambda_{\text{entropy}}(A_D) \cdot \mathcal{L}_{\text{entropy}}(A_D) + \lambda_{\text{entropy}}(A_B) \cdot \mathcal{L}_{\text{entropy}}(A_B) \\
&+ \lambda_{\text{asymmetry}}(A_D) \cdot \mathcal{L}_{\text{asymmetry}}(A_D) + \lambda_{\text{symmetry}}(A_B) \cdot \mathcal{L}_{\text{symmetry}}(A_B) \\
&+ \lambda_{\text{sparsity}}(A_D) \cdot \mathcal{L}_{\text{sparsity}}(A_D).
\end{aligned} \tag{6}
$$

This comprises the following components:

- Acyclicity: $\mathcal{L}_{\textbf{cycle}}$ as shown in Eq. (7) to apply acyclic constraints on A_D and the directed latent space to prevent cyclic causal structures among the directed latent variables.
- Bow-free: $\mathcal{L}_{\textbf{bow}}$ as shown in Eq. (8) to penalize simultaneous directed and bidirected edges for restricting the same pair of variables having both cause-effect and unobserved confounded relationships.
- Entropy: $\mathcal{L}_{\textbf{entropy}}(A_D)$ as shown in Eq. (9), $\mathcal{L}_{\textbf{entropy}}(A_B)$ as shown in Eq. (9) to maintain variability in the latent spaces, to enforce steady learning.
- Symmetry-Asymmetry: The $\mathcal{L}_{\text{asymmetry}}(A_D)$ ensures the non-existence of both $A_D[i,j]$ and $A_D[j,i]$ for any i, j and $\mathcal{L}_{\text{symmetry}}(A_B)$ ensures simultaneous existence of $A_B[i,j]$ and $A_B[j,i]$ for any i, j, as A_D is asymmetric and bidirectional edges in A_B are symmetric.

- Sparsity: This reduces noise in both the directed A_D and bidirected A_B adjacency matrices.

$$\mathcal{L}_{\text{cycle}} = \left|\text{trace}(e^{A_D}) - d\right|, (d : dimension A_D, A_B) \tag{7}$$

$$\mathcal{L}_{\text{bow}} = \alpha \cdot \left(\frac{1}{n}\sum_{i,j} A_D[i,j]^2\right) \cdot \beta \cdot \left(\frac{1}{n}\sum_{i,j} A_B[i,j]^2\right) \tag{8}$$

$$\begin{gathered}\mathcal{L}_{\text{entropy}}(A_D) = -\sum_i p_i^{(A_D)} \log p_i^{(A_D)}, \quad \mathcal{L}_{\text{entropy}}(A_B) = -\sum_i p_i^{(A_B)} \log p_i^{(A_B)} \\ p_i^{(A_D)} = \frac{|A_D[i]|}{\sum_j |A_D[j]|}, \quad p_i^{(A_B)} = \frac{|A_B[i]|}{\sum_j |A_B[j]|}\end{gathered} \tag{9}$$

Algorithm 1 Causal Relationships Learning: G-ADMG-CL

Input: Data: $\mathbf{X} \in \mathbb{R}^{n \times d}$ with unobserved confounder
Output: Reconstructed data $\hat{\mathbf{X}}$, Directed (A_D), Bidirected (A_B) adjacency matrices
— **Encoder Block** —
Initialize trainable adjacency matrices:
Directed adjacency matrix: $W_1 \sim \mathcal{N}(0,1)^{d \times d}$
Bidirected adjacency matrix: $W_2 \sim \mathcal{N}(0,1)^{d \times d}$
Compute directed latent statistics:
    $\boldsymbol{\mu}_D, \log \boldsymbol{\sigma}_D^2 \leftarrow$ FC layers on $\mathbf{X}$
    $\mu_{DA_D} \leftarrow \boldsymbol{\mu}_D \cdot W1$
Compute bidirected latent statistics:
    $\boldsymbol{\mu}_B, \log \boldsymbol{\sigma}_B^2 \leftarrow$ FC layers on $\mathbf{X}$
    $\mu_{BA_B} \leftarrow \boldsymbol{\mu}_B \cdot W2$
— **Reparameterization** —
Sample structure aware latent variables using reparameterization:
    $\mathbf{z}_D \leftarrow \mu_{DA_D} + \boldsymbol{\epsilon}_D \odot \exp(0.5 \cdot \log \boldsymbol{\sigma}_D^2), \quad \boldsymbol{\epsilon}_D \sim \mathcal{N}(0,1)$
    $\mathbf{z}_B \leftarrow \mu_{BA_B} + \boldsymbol{\epsilon}_B \odot \exp(0.5 \cdot \log \boldsymbol{\sigma}_B^2), \quad \boldsymbol{\epsilon}_B \sim \mathcal{N}(0,1)$
Concatenate latent representations: $\mathbf{z} \leftarrow [\mathbf{z}_D, \mathbf{z}_B]$
— **Decoder Block** —
    $\hat{\mathbf{X}} \leftarrow \text{Decoder}(\mathbf{z})$
— **Adjacency Matrix Estimation** —
    $A_D \leftarrow f(\mathbf{z}_D), \quad A_B \leftarrow f(\mathbf{z}_B)$
— **Learning and Optimization** —
Optimize with causal constraints: as defined in Eq. (3) during training

$$\begin{aligned}\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reconstruction}} &+ \lambda_{\text{KL}}\big(\mathcal{L}_{\text{KL_directed}} + \mathcal{L}_{\text{KL_bidirected}}\big) \\ &+ \lambda_{\text{causal}}\big(\mathcal{L}_{\text{Causal_ADMG}}\big)\end{aligned}$$

— **Learned Causal Structure** —
    Return A_D, A_B for evaluation and $\hat{\mathbf{X}}$

We introduce **causal annealing**, a learning strategy designed to systematically control the influence of the causal regularisation within the total loss during training, where we keep λ_{causal} a gradually increasing between 0 to 1 until the **causal transition epoch (CTE)**. At CTE λ_{causal} becomes 1 and remains 1 for the rest of the learning cycle. It helps to learn the data characteristics associated with the reconstruction and KL divergence components, and then apply the causality constraints. Algorithm 2 describes the causal annealing. Proposed causal annealing has two schedules, one is the `default hard mode`, where λ_{causal} remains 0 for $e < CTE$, and switches to 1 at CTE. To avoid abrupt changes in the gradient behavior and prevent abrupt instability at transition a relaxed schedule is introduced. In this relaxed schedule, we consider a relaxed transition epoch e_t, where λ_{causal} starts linearly increasing from 0 at e_t to 1 at the CTE.

Algorithm 2 Causal Annealing During Training

Input: Total epochs E, causal transition epoch CTE, linear transition start epoch e_t, anneal mode (`"hard"` or `"linear"`)
Output: Causal regularization schedule λ_{causal} for each epoch
Initialize $\lambda_{\text{causal}} \leftarrow 0$
for epoch $e = 1$ to E **do**
  if `anneal_mode` == `"hard"` **then**
    if $e < CTE$ **then**
      $\lambda_{\text{causal}} \leftarrow 0$
    else
      $\lambda_{\text{causal}} \leftarrow 1$
    end if
  else
    if $e < e_t$ **then**
      $\lambda_{\text{causal}} \leftarrow 0$
    else if $e < CTE$ **then**
      $\lambda_{\text{causal}} \leftarrow \dfrac{e - e_t}{CTE - e_t}$
    else
      $\lambda_{\text{causal}} \leftarrow 1$
    end if
  end if
  Update model parameters using λ_{causal}
end for

The proposed method G-ADMG-CL as presented in Algorithm 1 exploits a VAE-based generative model to learn the causal structures with unobserved confounders following the ADMG. The Proposed G-ADMG-CL uses dual latent spaces to represent the directed cause-effect relationships and the bidirected unobserved confounded relations, which are associated with the trainable adjacency matrices A_D, A_B respectively. The causally aware objective function is constrained to enforce the acyclicity which is a prime property of DAGs, along with the bow-free property. It ensures a proper balance between randomness and sparsity reduction while holding the symmetric property of A_B and asymmetric property of A_D. Our method is applicable in various contexts such as conditional, and counterfactual data generation with altered cause-effect relationships, and offers interpretable outcomes employing its dual latent spaces and learned adjacency matrices. Algorithm 1 presents G-ADMG-CL.

Causal Inference: We further employ our method to the prediction task as G-ADMG-CL+P to conduct the causal inference using learned dual latent spaces, $[z_{\text{directed}}, z_{\text{bidirected}}]$. The learned directed cause-effect and unobserved confounded relationships from the latent spaces reduce the variability in the average treatment effect (ATE) (Eq. 10), defined as (Eq. 10) the difference in the expected effect when the treatment or cause is applied depending on both observed and unobserved confounders.

$$\text{ATE} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)], \tag{10}$$

where $Y(1)$, $Y(0)$ represent the outcomes with and without applying the treatment (or cause), respectively.

4 Data

To evaluate our model under varying levels of structural complexity and confounding we consider synthetic datasets Fork Collider (FC) and Erdős–Rényi (ER) and IHDP data for causal structure identification and causal inference tasks.

Fork Collider (FC):

We synthesize 4000 samples of this data using the Eq. 11 of the SCM following Ashman et al. (2023). The causal graph and adjacency matrices are depicted in Figure 2a. Variables X2, X3 and X3, X4 have bidirected

edges, where u1 and u2 are the respective confounders influencing these pair of variables. The bidirected edges are represented by the dotted line, and the directed edges are shown by the solid line.

$$\begin{aligned}
\mathbf{T} &= [u_1, u_2, \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5]^T \sim \mathcal{N}(0,1), \\
x_1 &= \epsilon_1, \\
x_2 &= \sqrt{6}\exp(-u_1^2) + 0.1\epsilon_2, \\
x_3 &= \sqrt{6}\exp(-u_1^2) + \sqrt{6}\exp(-u_2^2) + 0.2\epsilon_3, \\
x_4 &= \sqrt{6}\exp(-u_2^2) + \sqrt{6}\exp(-x_1^2) + 0.1\epsilon_4, \\
x_5 &= \sqrt{6}\exp(-x_1^2) + 0.1\epsilon_5.
\end{aligned} \tag{11}$$

Erdős–Rényi (ER): We evaluate our method on two Erdős–Rényi (ER) synthetic graph configurations: ER(4,6,4) and ER(12,50,10), where each tuple represents (d, e, m) with unobserved confounders. Here, d denotes the number of observed variables (nodes), e the number of directed edges (representing causal relationships), and m the number of bidirected edges (capturing latent confounding).

The first configuration, ER(4,6,4), contains $d = 4$ variables, $e = 6$ directed edges, and $m = 4$ bidirected edges representing a compact graph with moderate confounding. The second, ER(12,50,10), includes $d = 12$ variables, $e = 50$ directed edges, and $m = 10$ bidirected edges, modelling a more complex and densely connected structure. These settings are designed to evaluate the scalability and robustness of our model in recovering both observed and latent causal relationships under increasing graph density and confounding.

We generate data from these graphs using the Eq. (16) (Appendix A), following the ER-based ADMG simulation procedure described in Ashman et al. (2023). The corresponding causal graphs and adjacency matrices are illustrated in Figure 3a. Notably, in these ER settings, directed (solid) and bidirected (dotted) edges often overlap reflecting realistic scenarios where a pair of variables possessing direct cause-effect relationship may be influenced by the unobserved common cause.

IHDP: IHDP (Infant Health and Development Program) is a real-world dataset originally introduced by (Hill, 2011) comprising measurements of infants and their mothers, collected during a randomized experiment. We adopt the IHDP dataset (Louizos et al., 2017), for the causal inference task, where the goal is to estimate the effect of home visits by specialists (cause) on the cognitive development of infants (effect). The dataset is used with 10 replicates, each consisting of a 70% train (IHDP_train) and 30% test (IHDP_test) split, with a fixed size of 747 individuals per replicate. It includes 19 binary covariates and 6 continuous covariates (e.g., birth weight, head circumference), all normalized before training. To incorporate unobserved confounding, treated individuals with non-white mothers are excluded. Outcomes are simulated based on Setting B (log-linear response surfaces), and Gaussian exogenous noise is applied.

5 Results

We evaluate the proposed G-ADMG-CL model across both synthetic and real-world datasets to assess its ability to identify causal structure and perform causal inference in the presence of latent confounders.

5.1 Performance Measure

We evaluate model performance using the following metrics:

F1 Score. The harmonic mean of precision and recall, computed over predicted vs. ground-truth edges (directed and bidirected separately). It is defined as:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{12}$$

RMSE-ATE. Measures the error in causal inference by computing the root mean squared error between the predicted average treatment effect (ATE_P) and ground-truth ATE over n samples:

$$\text{RMSE-ATE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{ATE_P}_i - \text{ATE}_i)^2}. \tag{13}$$

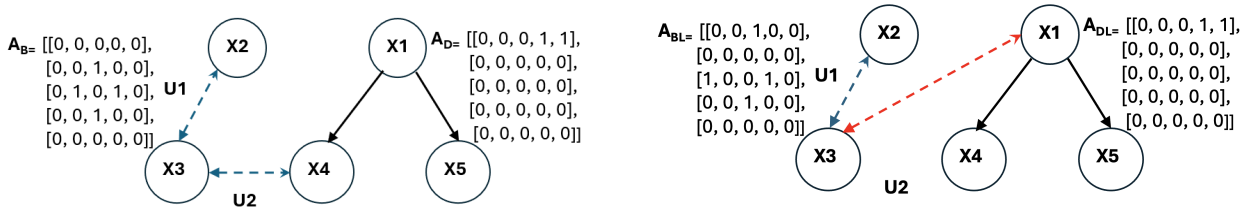
5.2 Evaluation of G-ADMG-CL

We evaluate the performance of the learned causal structure across different datasets by comparing it with various established state-of-the-art methods. F1 scores are used to measure the performance. The $F1_{\mathrm{D}}$, $F1_{\mathrm{B}}$ represent the F1 scores obtained for directed cause-effect relationships and for the bidirected unobserved confounded relations among the variables, respectively.

We compare our method against the following state-of-the-art approaches:

- FCI: Constraint-based method using independence tests to detect confounders.
- CAM-UV: Additive noise model with HSIC tests for hidden variables.
- RCD: Learns linear non-Gaussian models with bidirected edges.
- DCD: Additive causal discovery with structural regularization.
- N-DAG-G: Neural DAG learner assuming no hidden confounding.
- N-ADMG-G: Flow-based ADMG learner capturing latent confounders.
- N-BF-ADMG-G: Flow-based model for ADMGs with bow-free constraints.

Unlike these approaches, our method leverages learning of *dual latent space* with causality constraints to jointly model directed and bidirected causal relations, while *introducing a causal annealing strategy* that stabilizes training and improves the identification of latent confounding structures. This combination enables better disentanglement of causal directions and latent confounders, resulting in consistent performance gains in both $F1_{\mathrm{D}}$ and $F1_{\mathrm{B}}$ scores. The comparative results are summarized in Table 1, demonstrating that our method delivers consistently competitive performance and, in certain scenarios, outperforms state-of-the-art approaches across multiple datasets. We identify the significant edges by applying a best thresholding technique presented in the Appendix B (Algorithm 3), for the binarizing of the learned adjacency matrices A_D and A_B. We consider the default causal annealing during training.



(a) Ground truth causal graph and adjacency matrices.

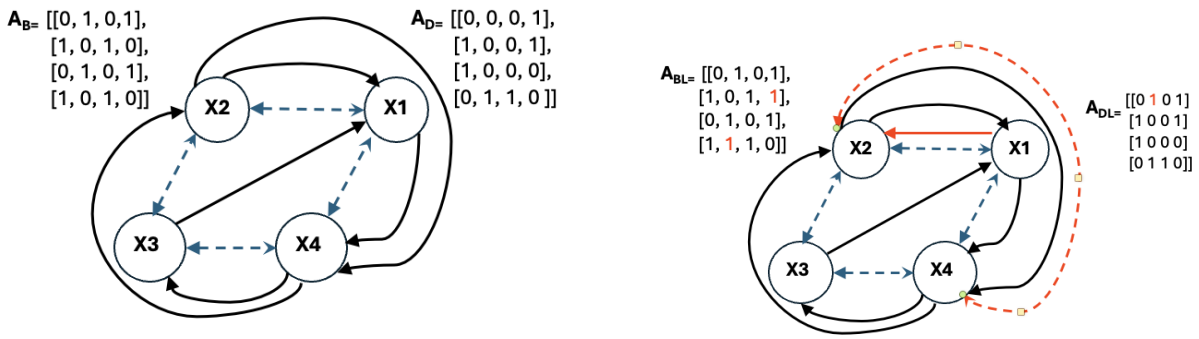
(b) Learned causal graph and adjacency matrices.

Figure 2: Comparison of ground-truth (left) and learned (right) causal graphs for FC. Solid: directed; dashed: bidirected; red: incorrect edges. Adjacency matrices: ground-truth: A_D, A_B; learned: A_{DL}, A_{BL}.

In case of **FC**, we train the model with 5000 epochs using ADAM optimizer with `ExponentialDecay` (Goodfellow et al. (2016)) learning rate scheduling. We apply KL annealing with *kl_weight* (a gradually increasing weight) until epoch 50 to multiply the KL divergence term to counter KL-vanishing during the initial training phase and focus only on reconstruction error minimization. The causal transition epoch (CTE) is set to 1200 to learn the data distribution and input characteristics well before applying the causal regularization. The complete configuration details are provided in the Appendix D (Table 7).

The learned causal graph and adjacency matrices are shown in Figure 2b. **All directed** cause-effect relations are correctly identified by the proposed method. For bidirected confounding, the model correctly detects the edge between X_2 and X_3, but replaces the ground-truth bidirected edge $X_3 \leftrightarrow X_4$ with $X_3 \leftrightarrow X_1$, representing a close confounding approximation and demonstrating the model's ability to infer approximate latent structures even in challenging scenarios. Red edges indicate relationships absent in the ground truth.

For **ER**, each dataset ER(4,6,4), and ER(12,50,10) we train the proposed model using the ADAM optimizer with appropriate learning rate scheduling (either `ExponentialDecay` or `CosineDecay`), KL annealing to gradually increase the *kl_weight* to 1.5, and dataset-specific regularization coefficients. The complete configuration details, including learning schedules, latent dimensions, and regularization parameters, are provided in the Appendix D (Table 7). The learned causal graph and adjacency matrices are depicted in



(a) Ground truth causal graph and adjacency matrices.

(b) Learned causal graph and adjacency matrices.

Figure 3: Comparison of ground-truth (left) and learned (right) causal graphs for ER. Solid: directed; dashed: bidirected; red: incorrect edges. Adjacency matrices: ground-truth: A_D, A_B; learned: A_{DL}, A_{BL}.

Figure 3b. CTE plays an important role in the training process. The proposed method achieves competitive performance (depicted in Table 1) compared to various state-of-the-art methods.

Table 1: F1 Score Comparison ($F1_D$: directed, $F1_B$: bidirected) across FC, ER dataset.

Method	**FC**		**ER(4,6,4)**		**ER(12, 50, 10)**	
	$F1_{\mathrm{D}}$	$F1_{\mathrm{B}}$	$F1_{\mathrm{D}}$	$F1_{\mathrm{B}}$	$F1_{\mathrm{D}}$	$F1_{\mathrm{B}}$
FCI	0.00	0.75	0.50	0.40	0.25	0.33
CAM-UV	0.80	0.67	0.30	0.25	0.38	0.36
RCD	0.00	0.54	0.35	0.35	0.45	0.20
DCD	0.00	0.67	0.25	0.20	0.32	0.18
N-DAG-G	0.50	0.00	0.60	0.00	0.55	0.00
N-ADMG-G	0.49	**0.99**	0.75	0.60	0.60	0.38
N-BF-ADMG-G	0.64	0.93	0.78	0.80	**0.60**	0.40
Proposed (G-ADMG-CL)	**1.0**	0.50	**0.92**	**0.89**	0.51	**0.45**

For the FC dataset, the proposed method achieves the highest F1 score for $F1_{\mathbf{D}} = 1.0$ to learn the directed causal relationship. The lower $F1_{\mathrm{B}}$ score 0.5 for the FC dataset results from a close confounding approximation, where the bidirected edge between X3 and X4 is approximated as a bidirected edge between X3 and X1, reflecting a close latent confounding relationship, this supports the model's ability to infer approximate latent causal structures even in challenging scenarios with subtle confounding signals.

For the ER dataset, our method achieves the highest $F1_{\mathbf{D}}$ score of 0.92 and the highest $F1_{\mathbf{B}}$ score of 0.89 on the ER(4,6,4) graph, effectively recovering both directed and bidirected causal relationships. On the more complex ER(12,50,10) graph, the model attains a competitive $F1_{\mathbf{D}}$ score of 0.51 and achieves the highest $F1_{\mathbf{B}}$ score of 0.45, outperforming all baselines in identifying unobserved confounded relationships. These results demonstrate the robustness of our method in recovering causal structures under dense and high-dimensional settings. All reported results are obtained using a fixed random seed; note that baselines

may produce different graphs even with identical configurations (i.e., same number of nodes, directed edges, and bidirected edges). To assess robustness across random structural variations, we evaluate our method on multiple randomly sampled ER graphs (Appendix C). It achieves average scores of $F1_{\mathrm{D}} = 0.84 \pm 0.051$, $F1_{\mathrm{B}} = 0.89 \pm 0.031$ for ER(4,6,4), and $F1_{\mathrm{D}} = 0.53 \pm 0.049$, $F1_{\mathrm{B}} = 0.41 \pm 0.052$ for ER(12,50,10).

Ablation study G-ADMG-CL: We perform ablation study for the causal relationships learning task. $\mathcal{L}_{\mathbf{Causal_ADMG}}$: The ablation study on causal mixed graph loss $\mathcal{L}_{\mathrm{Causal_ADMG}}$, presents the significance of the proposed causal constraint applied to the loss function of proposed generative model (G-ADMG-CL) using causally constrained VAE.

FC: We observe that removing causal constraint $\mathcal{L}_{\mathrm{Causal_ADMG}}$ causes a significant reduction in the performance in FC. As a next step to establish the importance of the different components of causal mixed graph loss like bow-free constraint λ_{bow}, symmetry/asymmetry constraints of adjacency matrices, we set their value to zero to remove their influence. We alter the value of λ_{cycle} to a high value to judge its sensitivity, as depicted in Table 2. This study establishes that all these components significantly impact performance and play a significant role in causal structure learning.

ER: We observe that removing $\mathcal{L}_{\mathrm{Causal_ADMG}}$ causes a significant reduction in the performance as depicted in Table 2. As a next step to establish the importance of the different components of this causal mixed graph loss we perform the same steps as performed in FC. We notice that the symmetry constraint of A_B plays an important role. ER data has overlapping directed and bidirected edges where bow-free constraint aims to ensure non overlapping directed and bidirected edges which reduces the impact of the bow effect. This study also reflects that our method performs well in the presence of overlapping directed and bidirected edges, that is, when variables have both a direct cause-effect relationship and are influenced by a common unobserved confounder.

Table 2: Ablation Study of $\mathcal{L}_{\mathrm{Causal_ADMG}}$ and its components using F1 Score

Method	**FC**		**ER(4,6,4)**	
	$F1_{\mathbf{D}}$	$F1_{\mathbf{B}}$	$F1_{\mathbf{D}}$	$F1_{\mathbf{B}}$
Proposed with $\mathcal{L}_{\mathrm{total}}$	1.0	0.50	0.92	0.89
$\mathcal{L}_{\mathrm{total}} - \mathcal{L}_{\mathrm{Causal_ADMG}}$	0.50	0.31	0.86	0.78
$\mathcal{L}_{\mathrm{Causal_ADMG}} - \mathcal{L}_{\mathrm{bow}}$	0.44	0.5	0.86	0.88
$\mathcal{L}_{\mathrm{Causal_ADMG}} - \mathcal{L}_{\mathrm{symmetry}}(A_B)$	0.66	0.25	0.86	0.82
$\mathcal{L}_{\mathrm{Causal_ADMG}} - \mathcal{L}_{\mathrm{asymmetry}}(A_D)$	0.66	0.5	0.92	0.89
$\mathcal{L}_{\mathrm{Causal_ADMG}} : \lambda_{\mathrm{cycle}} = 10$	0.57	0.40	0.92	0.89

Causal Annealing: The ablation study on causal annealing presents the sensitivity of the proposed causal annealing applied to the proposed G-ADMG-CL, the causally constrained VAE. We compare the baseline G-ADMG-CL without annealing, ($\lambda_{\mathrm{causal}} = 0$) against G-ADMG-CL with hard causal annealing ($\lambda_{\mathrm{causal}}$ =1). Table 3 presents an ablation study to evaluate the impact of causal annealing.

FC: In this dataset, we observe that using baseline G-ADMG-CL reduces the performance for directed relationships on the FC dataset ($F1_{\mathrm{D}} = 0.50$) significantly. In contrast, G-ADMG-CL with annealing achieves the highest directed $F1_{\mathrm{D}} = 1.00$ performance.

Table 3: Ablation Study of Causal Annealing using F1 Score

Method	**FC**		**ER(4,6,4)**	
	$F1_{\mathbf{D}}$	$F1_{\mathbf{B}}$	$F1_{\mathbf{D}}$	$F1_{\mathbf{B}}$
G-ADMG-CL (with Causal Annealing)	**1.0**	0.50	**0.92**	**0.89**
G-ADMG-CL (without Causal Annealing)	0.50	0.50	0.75	0.80

ER: For ER, we notice that using baseline G-ADMG-CL reduces the performance for directed relationships on the ER dataset ($F1_{\mathrm{D}} = 0.75$) significantly, and ($F1_{\mathrm{B}} = 0.80$). In contrast, G-ADMG-CL with annealing achieves the highest directed $F1_{\mathrm{D}}$ on ER (0.92) and maintains strong bidirected performance (0.89).

5.3 Causal Inference task

We demonstrate the applicability of our method to a **causal inference task** using the IHDP dataset. While we refer to the causal inference extension of our model as G-ADMG-CL+P, we report results and comparisons using the core model name **G-ADMG-CL**, consistent with baseline naming conventions. We first train the model using IHDP_train to obtain the learned directed and bidirected latent representations. The model is trained for 5000 epochs using the ADAM optimizer, with ExponentialDecay as the learning scheduler. The initial learning rate is set to $initial_learning_rate = 0.001$, with $decay_steps$=1000, and $decay_rate = 0.90$. We apply KL annealing, with kl_weight (a gradually increasing weight) until epoch 20 to the $kl_weight = 1.5$ to multiply the KL divergence term to counter KL-vanishing during the early training phase. This allows us to focus only on reconstruction error minimization. We consider the length of z_{directed}, and the length of $z_{\text{bidirected}}$ as 50 as this data is relatively complex. We consider $\lambda_{\text{cycle}} = 1$, $\lambda_{\text{bow}} = 5$, $\lambda_{\text{entropy}}(A_D) = 0.01$, $\lambda_{\text{entropy}}(A_B) = 0.001$, $\lambda_{\text{asymmetry}}(A_D) = 0.05$, $\lambda_{\text{symmetry}}(A_B) = 4.75$. The causal transition epoch (CTE) is set to 1000 to ensure that the G-ADMG-CL learns the data distribution and input characteristics well before applying the causal regularization. We apply the trained model, G_ADMG_CL, to the IHDP_test for obtaining the predicted outcome $y_{\text{treated_P}}, y_{\text{control_P}}$ as shown in Eq. (14), Eq. (15) using the decoder reconstructed output as described in the methodology.

$$model = \text{G_ADMG_CL(IHDP_train)} \tag{14}$$

$$y_{\text{treated_P}}, y_{\text{control_P}} = model.predict(\text{IHDP_test}) \tag{15}$$

The proposed method outperforms compared to the various state-of-the-arts methods by achieving the lowest RMSE. Since the IHDP does not provide a ground-truth causal graph, the ATE serves as the most reliable metric for comparing causal relationship learning performance.

Table 4: Causal Inference Results using IHDP Dataset

Method	**RMSE-ATE**
FCI	0.13
CAM-UV	0.15
RCD	0.14
DCD	0.16
N-DAG-G	0.12
N-BF-ADMG-G	0.10
G-ADMG-CL	**0.031**

6 Conclusion

We propose a novel generative method based on VAE for learning causal relationships with dual latent spaces. Our neural causal method identifies the nonlinear cause-effect relationships that involve unobserved common causes, following the ADMG framework. The causal structures are represented by the learned dual latent representations, z_{directed} and $z_{\text{bidirected}}$. These representations are associated with the trainable directed adjacency matrix A_D and the bidirected adjacency matrix A_B, which correspond to directed cause-effect relationships and bidirected latent confounded influences, respectively. We propose the causal mixed graph loss $\mathcal{L}_{\text{Causal_ADMG}}$ as a causality constraint to enforce the acyclicity, and to learn bidirected unobserved influence by the common causes. We introduce the concept of *causal annealing* to guide the learning of causal structure. We establish the importance of the causality constraint and causal annealing through ablation studies. Our method, **G-ADMG-CL**, is evaluated and compared against the established prior works, where it achieves competitive performance and outperforms several settings. Furthermore, we perform a causal inference task using the learned causal structure with real data, achieving significantly improved performance. As part of the future scope, we wish to further investigate the causal annealing and effects of CTE using different setups and diverse cause-effect relations. We also aim to further study to address the co-existence of the directed edges representing cause-effect relationships and the bidirected edges that present the unobserved confounded influence and the underlying graph structure.

References

M. Ashman, C. Ma, A. Hilmkil, J. Jennings, and C. Zhang. Causal reasoning in the presence of latent confounders via neural admg learning. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 2023.

R. Bhattacharya, T. Nagarajan, D. Malinsky, and I. Shpitser. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, pp. 2314–2322. PMLR, 2021.

C. Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:2111.10846*, 2021.

T. Geffner, J. Antoran, A. Foster, W. Gong, C. Ma, E. Kiciman, A. Sharma, A. Lamb, M. Kukla, N. Pawlowski, et al. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022.

Christophe Gonzales and Amir-Hosein Valizadeh. A full dag score-based algorithm for learning causal bayesian networks with latent confounders. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, 2024.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Sch"olkopf, and A. J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20, pp. 585–592, 2008.

U. Hasan, E. Hossain, and M. O. Gani. A survey on causal discovery methods for i.i.d. and time series data. Technical report, Causal AI Lab, Dept. of Information Systems, Univ. of Maryland, Baltimore County, Baltimore, MD, USA, 2025.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Chenliang Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, USA, 2019. URL `https://openreview.net/forum?id=Syx4wnEtvH`.

C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6446–6456, Long Beach, CA, USA, 2017. `https://github.com/AMLab-Amsterdam/CEVAE/tree/master/datasets/IHDP`.

T. N. Maeda and S. Shimizu. Rcd: Repetitive causal discovery of linear non-gaussian acyclic models with latent confounders. In *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, pp. 735–745. PMLR, 2020.

T. N. Maeda and S. Shimizu. Causal additive models with unobserved variables. In *Proc. Uncertainty in Artificial Intelligence (UAI)*, pp. 97–106. PMLR, 2021.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2nd edition, 2009.

Judea Pearl. Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2): 226–284, 1998.

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, Prediction, and Search*. MIT Press, 2000.

M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang. Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*, 2023. `https://arxiv.org/abs/2004.08697`.

Yue Yu, Cheng Zhang, Zhifeng Kong, Yanfeng Wang, and Zhaowen Wang. DAG-GNN: Dag structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

A. Zanga and F. Stella. A survey on causal discovery: Theory and practice. *arXiv preprint arXiv:2305.10032*, 2023. `https://arxiv.org/abs/2305.10032`.

X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Appendix

A ER Data Synthesis

ER Configuration. The ER-based graphs are generated using three parameters: d, e, and m.

- d: the number of observed variables or nodes in the causal graph,
- e: the number of directed edges representing direct cause-effect relationships,
- m: the number of bidirected edges modeling unobserved confounding between variables.

For example, ER(4,6,4) denotes a 4-variable graph with 6 directed edges and 4 bidirected edges. Similarly, ER(12,50,10) corresponds to a larger graph with 12 variables, 50 directed edges, and 10 bidirected edges. These settings allow us to evaluate how the model performs under increasing structural complexity and the presence of latent confounders. We use the Eq. 16 to generate ER data.

$$
\begin{aligned}
A_D &\sim \mathrm{ER}\left(d, \frac{e}{d(d-1)}\right), \quad \mathrm{diag}(A_D) = 0 \\
A_D[i,j] &= \begin{cases} 1 & \text{if there is a directed edge from } i \text{ to } j, \\ 0 & \text{otherwise.} \end{cases} \\
A_B[i,j] &\sim \mathrm{Bernoulli}\left(\frac{m}{d(d-1)}\right), \quad \mathrm{diag}(A_B) = 0 \\
A_B &= \mathrm{triu}(A_B, 1) + \mathrm{triu}(A_B, 1)^\top, \\
&: \text{triu extracts the elements above the diagonal,} \\
\epsilon &\sim \mathcal{N}(0, 0.1^2), \quad u \sim \mathcal{N}(0, 0.1^2), \\
X_i &= \sum_{p \in \mathrm{Pa}_D(i)} f(X_p) + \sum_{q \in \mathrm{Pa}_B(i)} g(u_q) + \epsilon_i.
\end{aligned} \tag{16}
$$

B Threshold Selection

In this section, we present the threshold selection method. We also describe the procedure for identifying the optimal threshold to obtain the best F1 score for the learned adjacency matrices.

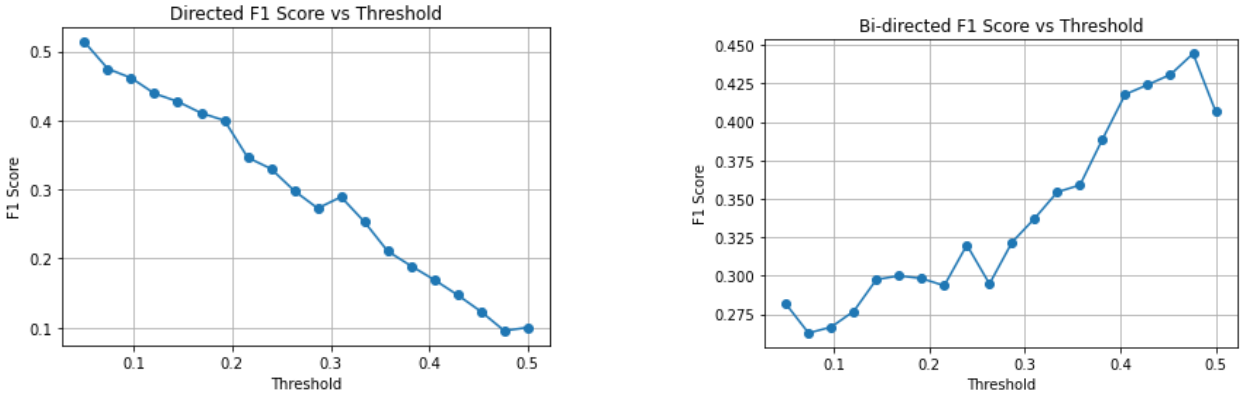
B.1 Algorithm for Optimal Threshold Selection

To obtain the best-directed ($F1_{\mathrm{D}}$) and bidirected ($F1_{\mathrm{B}}$) adjacency F1 scores, we employ the method described in Algorithm 3. Figure 4 illustrates the variation of F1 scores with respect to the threshold values to obtain the optimal threshold selection for ER(12, 50, 10). Subfigure 4a shows the best-obtained F1 score for the directed adjacency matrix A_D, while Subfigure 4b depicts the best-obtained F1 score for the bidirected adjacency matrix A_B. The optimal threshold, corresponding to the peak F1 value in each curve, is selected as described in Algorithm 3.
The optimal threshold selection plots for the directed and bidirected adjacency matrices on the FC dataset are shown in Figure 5, highlighting the threshold values that yield the highest F1 score for each edge type.

In realistic scenarios where the ground-truth causal graph is unavailable, the adaptive threshold mechanism provides a suitable alternative for threshold determination. However, state-of-the-art methods report their best-obtained performance, we also compare against the best-achieved results for fairness.

Algorithm 3 Optimal Threshold Selection for F1 Score

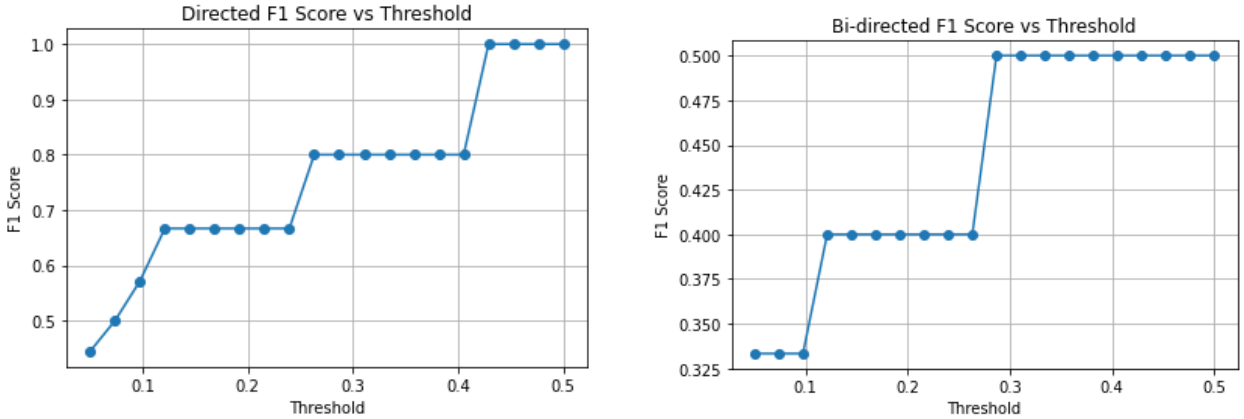
Input: Ground truth adjacency matrix A, learned adjacency matrix W, set of thresholds $\mathcal{T}$
Output: Best threshold t^*, maximum F1 score $F1_{\max}$
$F1_{\max} \leftarrow 0, \quad t^* \leftarrow 0$
for each threshold $t \in \mathcal{T}$ **do**
  $W_{\text{bin}} \leftarrow \mathbb{I}(|W| \geq t)$ ▷ Binarize W at threshold t
  $\mathbf{a} \leftarrow \text{flatten}(A)$ ▷ Convert matrix A to vector
  $\mathbf{w} \leftarrow \text{flatten}(W_{\text{bin}})$ ▷ Convert W_{bin} to vector
  $F1_t \leftarrow \text{F1_score}(\mathbf{a}, \mathbf{w})$
  if $F1_t > F1_{\max}$ **then**
    $F1_{\max} \leftarrow F1_t$
    $t^* \leftarrow t$
  end if
end for
return $t^*, F1_{\max}$



(a) Obtained best F1 score for learned A_D .

(b) Obtained best F1 score for learned A_B.

Figure 4: Optimal threshold selection for ER(12, 50, 10).



(a) Obtained best F1 score for learned for A_D.

(b) Obtained best F1 score for A_B.

Figure 5: Optimal threshold selection for FC.

C Additional Results

In addition to the best results reported in the main text, we provide reproducibility analysis for the ER(4, 6, 4) and ER(12, 50, 10) configurations. Each run corresponds to a new random graph sampled under the respective ER settings. Table 5 and Table 6 report F1 scores across five such runs.

Table 5: Additional results for ER(4,6,4) dataset with five independent graphs.

Run	$F1_{\mathrm{D}}$	$F1_{\mathrm{B}}$
1	0.800	0.933
2	0.800	0.842
3	0.857	0.880
4	0.800	0.889
5	0.920	0.889
Mean ± Std	**0.835 ± 0.051**	**0.887 ± 0.031**

Table 6: Additional results for ER(12, 50, 10) with five independent graphs.

Run	$\mathbf{F1_D}$	$\mathbf{F1_B}$
1	0.576	0.316
2	0.589	0.450
3	0.538	0.414
4	0.510	0.450
5	0.452	0.424
Mean ± Std	**0.533 ± 0.049**	**0.411 ± 0.052**

Notably, the bidirected recovery remains relatively stable across runs, highlighting the robustness of our model in consistently capturing latent confounding even in dense, high-dimensional graphs such as ER(12,50,10).

D Training Configurations for Synthetic Datasets

We present the training configuration table for the synthetic data FC and ER below Table 7.

Table 7: Training configurations for FC, ER(4,6,4), and ER(12,50,10) experiments. Hyperparameters include KL annealing, learning rate schedules, and regularization weights (λ) for entropy, symmetry, cycle, and bow-free constraints.

Setting	**FC**	**ER(4,6,4)**	**ER(12, 50, 10)**
Epochs	5000	3000	2000
Optimizer	ADAM	ADAM	ADAM
Learning Rate Schedule	ExponentialDecay	CosineDecay	ExponentialDecay
Initial Learning Rate	0.001	0.001	0.001
Decay Steps	1000	1000	64
Decay Rate	0.90	–	0.98
Minimum Learning Rate (α)	–	1×10^{-5}	–
KL Annealing End Epoch	50	100	800
Final *kl_weight*	1.5	1.5	1.5
Latent Dimension	z_{dir}: 24, z_{bidir}: 24	24	36
λ_{cycle}	1	7	5
λ_{bow}	5	5	5
$\lambda_{\text{entropy}}(A_D)$	0.005	0.001	0.01
$\lambda_{\text{entropy}}(A_B)$	0.01	0.01	0.01
$\lambda_{\text{asymmetry}}(A_D)$	0.05	0.05	0.05
$\lambda_{\text{symmetry}}(A_B)$	0.5	1.5	1.75
Causal Transition Epoch (CTE)	1200	150	1000