Dual-Latent Generative Causal Structure Learning with Causal Annealing

Anonymous Author(s)
Affiliation
Address
email

Abstract

We explore causal structure learning with unobserved confounders, represented by Acyclic Directed Mixed Graphs, where directed edges indicate observed cause-effect relationships and bidirected edges capture unobserved confounding. Previous methods have focused on search-based approaches or flow-based generative models. In contrast, we propose a novel variational autoencoder framework with dual latent spaces, each associated with a trainable adjacency matrix to capture directed and bidirected edges, respectively. We propose a causality constraint and introduce a causal annealing strategy during training to obtain meaningful causal graph structures. Experiments show competitive identification of both relationship types on synthetic data, with learned structures enhancing downstream causal inference in a real-world task.

1 Introduction

Learning cause-effect [8] relationships from observational data becomes particularly challenging in the presence of *unobserved confounders*. Classical approaches, including score-based methods (e.g., BIC) and constraint-based algorithms (e.g., conditional independence tests), often fail when latent variables are present. While recent continuous optimization methods such as NOTEARS [11] provide scalable solutions under acyclicity constraints, they typically assume causal sufficiency. The framework of Acyclic Directed Mixed Graphs (ADMGs) and bow-free constraints [3] extends identifiability with hidden confounding, but generative models in this space remain underexplored.

In this work, we propose a **causally constrained** variational autoencoder (VAE) framework that disentangles observed and latent confounding via *dual latent spaces*, each linked to trainable adjacency matrices for directed and bidirected edges. To guide the model toward learning faithful causal structures, we introduce a structured objective that enforces acyclicity and bow-free constraints. Additionally, we propose a novel training strategy called *causal annealing*, which delays the application of causal regularization, allowing the model to first focus on reconstruction and KL divergence. Our approach recovers interpretable causal graphs with unobserved confounding and demonstrates improved performance on downstream causal inference tasks in both synthetic and real-world dataset.

Key Contributions.

- **Dual latent spaces for causal disentanglement:** We design a VAE framework that separates observed and unobserved causal relations by learning two distinct latent spaces, capturing directed and bidirected dependencies via adjacency matrices A_D and A_B, respectively.
- **Causally aware objective:** We propose a causality-aware loss that enforces acyclicity for A_D, bow-free constraints for A_B, and sparsity-entropy trade-offs to ensure structural interpretability and enable meaningful edge selection.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

- **Causal annealing:** We introduce a novel training strategy that gradually activates causal constraints through a *causal transition epoch (CTE)*, enabling the model to prioritize reconstruction and KL divergence in early training before focusing on causal structure learning.

2 Related Work

Causal structure learning has been widely studied through constraint-based and score-based approaches, including FCI [9], which can detect latent confounders via conditional independence tests, but often fail under hidden confounding and do not scale well to large graphs. Score-based methods such as CAM-UV [7], which uses HSIC [5] for independence testing, and RCD [6], which assumes linear non-Gaussian models, extend causal discovery to partially address latent confounding.

Differentiable optimization methods like NOTEARS [11] and its neural extensions DAG-GNN [10] and N-DAG-G [4] enable continuous DAG learning with end-to-end inference but assume full observability and cannot represent bidirected edges.

To handle latent confounding more explicitly, ADMG-based methods have been proposed. The framework in [3] introduces bow-free and ancestral constraints to support interpretable structure learning under linear Gaussian assumptions. Flow-based models such as N-ADMG-G and N-BF-ADMG-G [2] enable nonlinear causal structure learning via autoregressive generative modeling, but do not incorporate latent-variable disentanglement or generation-based optimization.

3 Methodology

We consider two tasks. The first, causal structure learning, **G-ADMG-CL**, is designed to identify causal relationships under latent confounding by identifying directed and bidirected edges considering an ADMG framework. The extended variant, **G-ADMG-CL+P**, builds on the learned structure to perform prediction and causal inference (e.g., estimating treatment effects).

Our model, **G-ADMG-CL**, is a causally constrained variational autoencoder that learns interpretable causal graphs under latent confounding by leveraging dual latent spaces and a causality-aware loss $\mathcal{L}_{\text{Causal\_ADMG}}$. We follow the identifiability assumptions established for ADMG structure learning in prior work [2].

Model Overview: We use a VAE with **dual latent spaces** $\mathbf{z}_D$ for directed cause-effect and $\mathbf{z}_B$ for bidirected latent confounding associated with trainable adjacency matrices A_D and A_B. These spaces guide both reconstruction and causal structure estimation. The functional compoenets are depicted in Figure 1.

Learning Causal Structure: Our proposed method, G-ADMG-CL, proceeds in the following

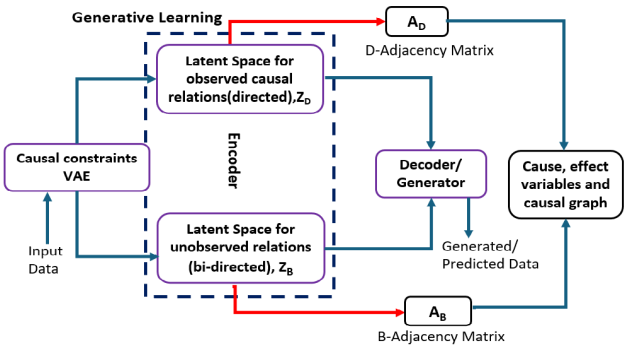


Figure 1: Functional components of the proposed G-ADMG-CL method.

stages. First, we initialize trainable graph parameters: W_1 and W_2, corresponding to the directed and bidirected adjacency matrices A_D and A_B, respectively. The encoder then maps the input data $\mathbf{X}$ into two sets of latent variables (μ_D, σ_D^2) and (μ_B, σ_B^2), each updated by the respective adjacency weights during every training epoch, to produce structure-aware latents. Using the standard reparameterization trick, we sample latent vectors $\mathbf{z}_D$ and $\mathbf{z}_B$, which are concatenated to form a joint

Algorithm 1 G-ADMG-CL: Causal Relationships Learning

Input: data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$
Output: reconstructed $\hat{\mathbf{X}}$, directed A_D, bidirected A_B
Initialize W_1, W_2; encode $\mathbf{X}$ into $(\boldsymbol{\mu}_D, \log \sigma_D^2)$ and $(\boldsymbol{\mu}_B, \log \sigma_B^2)$
Compute $\mu_{DA_D} = \boldsymbol{\mu}_D W_1$, $\mu_{BA_B} = \boldsymbol{\mu}_B W_2$
Structure-aware latents: $\mathbf{z}_D \leftarrow \mu_{DA_D} + \boldsymbol{\epsilon}_D \odot \exp(0.5 \cdot \log \sigma_D^2)$, $\mathbf{z}_B \leftarrow \mu_{BA_B} + \boldsymbol{\epsilon}_B \odot \exp(0.5 \cdot \log \sigma_B^2)$ where $\boldsymbol{\epsilon}_D, \boldsymbol{\epsilon}_B \sim \mathcal{N}(0, I)$
Sample $\mathbf{z}_D$, $\mathbf{z}_B$; form $\mathbf{z} = [\mathbf{z}_D, \mathbf{z}_B]$; decode $\hat{\mathbf{X}}$
Estimate $A_D = f(\mathbf{z}_D)$, $A_B = f(\mathbf{z}_B)$
Minimize total loss $\mathcal{L}_{\text{total}}$ with annealing schedules
return A_D, A_B, $\hat{\mathbf{X}}$

latent representation $\mathbf{z} = [\mathbf{z}_D, \mathbf{z}_B]$. This combined representation is passed through a decoder to reconstruct the input as $\hat{\mathbf{X}}$. Simultaneously, the soft adjacency matrices A_D and A_B are estimated from $\mathbf{z}_D$ and $\mathbf{z}_B$ through trainable functions. The model is trained by minimizing the following total loss, the **training objective**.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reconstruction}} + \lambda_{\text{KL}}(\mathcal{L}_{\text{KL\_directed}} + \mathcal{L}_{\text{KL\_bidirected}}) + \lambda_{\text{causal}}\mathcal{L}_{\text{Causal\_ADMG}},$$

which combines reconstruction error, KL divergence (with annealing), and a structured causal loss that enforces acyclicity, bow-free constraints, and regularization terms. To ensure stable training, we introduce *causal annealing* (Appendix C), where the causal regularization weight λ_{causal} is gradually increased until a designated transition epoch, allowing the model to first focus on data reconstruction before enforcing structural constraints. Finally, the learned soft adjacency matrices are thresholded to yield interpretable causal graphs (A_D, A_B) alongside the reconstructed data $\hat{\mathbf{X}}$. The learned soft adjacency matrices (A_D, A_B) are thresholded to obtain binary graphs used for evaluation. The pseudocode is presented in Algorithm 1. The role of causal annealing is detailed in the ablation study (Appendix D).

Causal inference: For this task the extended model G-ADMG-CL+P leverages the learned graph structure to estimate treatment effects from partially observed covariates.

4 Experiments

Datasets. We evaluate on diverse datasets: (i) Fork Collider (FC), (ii) Erdős–Rényi (ER) synthetic graphs, and (iii) IHDP [1] real-world causal inference dataset, simulating unobserved confounding by excluding treated individuals with non-white mothers and generating outcomes using log-linear response surfaces. The SCMs of the first two data are given in Appendix B.

Table 1: Performance Comparison: F1 Scores (F1D for Directed, F1B for BiDirected Edges) on FC, ER(4,6,4), and ER(12,50,10)

Method	FC		ER(4,6,4)		ER(12,50,10)	
	F1D	F1B	F1D	F1B	F1D	F1B
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**

Results. G-ADMG-CL: Table 1 shows that our method achieves superior F1 scores (Appendix A) for directed edges (F1D) on both FC and ER datasets, outperforming FCI, RCD, CAM-UV, and neural ADMG variants. While the F1B score on FC is lower due to approximate confounding estimation, our method achieves competitive bidirected performance on ER graphs. Learned causal graphs are shown in Appendix F. Thresholding used to binarize the learned causal graphs is detailed in Appendix G.

Table 2: 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
Proposed (G-ADMG-CL+P)	0.031

G-ADMG-CL+P: On the IHDP dataset, Table 2 shows that our model achieves the lowest RMSE-ATE (Appendix A), the most reliable metric in the absence of a ground-truth causal graph. Training hyperparameters are summarized for all datasets in Appendix E.

5 Conclusion

Prior methods do not utilize a latent-variable generative model. In contrast, our work proposes a VAE-based framework that learns disentangled latent spaces for directed and bidirected relations. With a structured causal loss and a novel causal annealing schedule, our approach enables interpretable and robust causal structure discovery under latent confounding. The proposed method achieves strong performance on synthetic graphs, and improves causal inference on real data. Future work will explore the impact of causal annealing and CTE across varied structural setups and causal dynamics.

References

[1] `https://github.com/AMLab-Amsterdam/CEVAE/tree/master/datasets/IHDP`.

[2] 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.

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

[4] 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.

[5] 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, pages 585–592, 2008.

[6] 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)*, pages 735–745. PMLR, 2020.

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

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

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

[10] 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.

[11] 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.

A Performance Metrics

F1 Score. To evaluate the accuracy of structure recovery, we report the F1 score for both directed (F1D) and bidirected (F1B) edges. F1 is computed as the harmonic mean of precision and recall:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

where true positives are correctly predicted edges, and precision/recall are calculated separately for directed and bidirected adjacency matrices.

RMSE-ATE. For causal inference performance on the IHDP dataset, we report the Root Mean Squared Error of the Average Treatment Effect (RMSE-ATE). This is computed between the true and estimated ATE across test samples:

$$\text{RMSE-ATE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}_i - \tau_i)^2},$$

where τ_i and $\hat{\tau}_i$ denote the ground-truth and estimated treatment effect for individual i, respectively.

B SCM of Datasets (FC and ER)

The following Structural Causal Model defines the data-generating process used for the Fork Collider (FC).

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

The following Structural Causal Model defines the data-generating process used for the ER (d,e,m). Here, d denotes the number of observed variables (nodes), e the number of directed edges (cause-effect relations), and m the number of bidirected edges (latent confounders). For instance, ER(4,6,4) represents a graph with 4 variables, 6 directed edges, and 4 bidirected edges.

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

C Causal Annealing Schedule

We introduce a **causal annealing**, a mechanism designed to systematically control the influence of the causal regularization term within the total loss. In early stage of training it is beneficial to prioritize

reconstruction and latent representation learning before enforcing strong causal constraints. To this end, we gradually increase the causal weight λ_{causal} over training epochs using using either a hard or linear annealing schedule. **Algorithm 2** details the annealing procedure. Given total epochs E, causal transition epoch CTE, and an optional warm-up start epoch e_t, the algorithm updates λ_{causal} at each epoch. In `"hard"` mode, the causal weight is kept at 0 until epoch CTE, after which it is set to 1. In `"linear"` mode, λ_{causal} increases gradually from 0 (starting at e_t) to 1 (at CTE), following a linear ramp-up schedule. This delayed enforcement of the causal loss prevents early convergence to poor graph structures and promotes better structure recovery and generalization.

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

D Ablation Results

Causal annealing is a key training strategy that stabilizes structure learning by delaying the influence of causal regularization.

Effect of Causal Annealing: Table 3 presents an ablation comparing G-ADMG-CL trained with and without causal annealing (hard mode). On the FC dataset, the F1 score for directed edges (F1D) improves from 0.50 to 1.00 when annealing is applied, while maintaining F1B. Similarly, on ER(4,6,4), F1D improves from 0.75 to 0.92. This shows that causal annealing significantly improves structure recovery, in the presence of unobserved confounding.

Table 3: Impact of causal annealing on structure recovery (F1).

Method	FC		ER(4,6,4)	
	F1D	F1B	F1D	F1B
G-ADMG-CL (with annealing)	1.00	0.50	0.92	0.89
G-ADMG-CL (no annealing)	0.50	0.50	0.75	0.80

E Training Configuration Summary

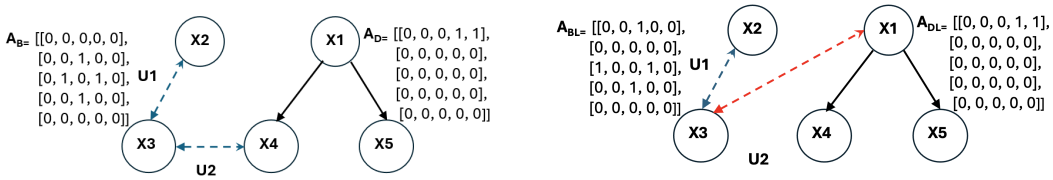
We summarize the key training hyperparameters for the synthetic datasets (FC and ER variants), and IHDP given in Table 4.

Table 4: Key training parameters for synthetic and real-world datasets used in experiments.

Parameter	FC	ER(4,6,4)	ER(12,50,10)	IHDP
KL Annealing Epoch	50	100	800	20
Causal Transition Epoch (CTE)	1200	150	1000	1000
Latent Dim (z, (z_D, z_B))	24	24	36	50
λ_{cycle}	1	7	5	1
$\lambda_{\text{symmetry}}(A_B)$	0.5	1.5	1.75	4.75

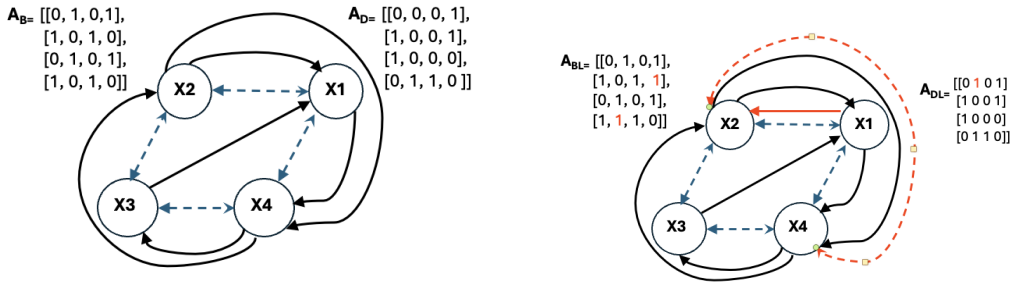
F Additional Figures: Learned Causal Graph

This section visualizes the learned causal graphs and corresponding adjacency matrices produced by our method, G-ADMG-CL, compared against the ground truth graphs for both the FC and ER (4,6,4) datasets. Figure 2 shows the comparison for the FC dataset, and Figure 3 presents the results for the ER dataset. Each pair of subfigures shows the true causal structure (left) and the structure learned by our model (right). Directed edges are denoted as solid lines and bidirected edges as dashed lines. Red edges in the learned graphs indicate spurious connections not present in the ground truth, highlighting areas of overestimation or structural deviation. These figures provide a qualitative understanding of how well the model captures both observed and latent confounding relationships.



(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 structures for the FC dataset. Directed edges are solid; bidirected edges are dashed. Red edges indicate connections not present in the ground truth.



(a) Ground-truth causal graph and adjacency matrices. (b) Learned causal graph and adjacency matrices.

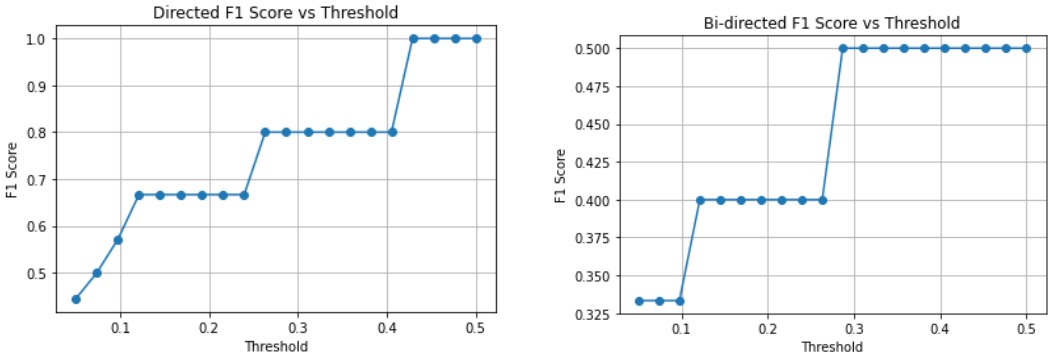
Figure 3: Comparison of ground truth (left) and learned (right) causal structures for the ER dataset. Directed edges are solid; bidirected edges are dashed. Red edges indicate connections not present in the ground truth.

We also plan to explore the interplay between directed edges (representing observable causal relations) and bidirected edges (capturing latent confounding), to better understand their co-existence within complex graph structures.

G Threshold Optimization

Algorithm 3 Optimal Threshold Selection via F1 Sweep

Input: Ground truth adjacency matrix A, learned soft matrix W, threshold set $\mathcal{T}$
Output: Optimal threshold t^*, maximum F1 score $F1_{\max}$
Initialize $F1_{\max} \leftarrow 0$
Initialize $t^* \leftarrow 0$
for each threshold $t \in \mathcal{T}$ **do**
  $W_{\text{bin}} \leftarrow (|W| \geq t)$ *Element-wise thresholding*
  $\mathbf{a} \leftarrow \text{flatten}(A)$
  $\mathbf{w} \leftarrow \text{flatten}(W_{\text{bin}})$
  $F1_t \leftarrow \text{F1\_score}(\mathbf{a}, \mathbf{w})$
  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 for A_D.

(b) Obtained best F1 score for A_B.

Figure 4: Optimal threshold selection for FC.

To convert the learned soft adjacency matrices (A_D, A_B) into interpretable binary graphs, we apply a thresholding mehod that selects the threshold maximizing F1 score. A grid search is performed over a set of candidate thresholds $\mathcal{T}$ (e.g., [0.05, 0.5]) to binarize the edges and compute F1 scores against the ground truth graph. The threshold that yields the highest F1 is selected for final evaluation. For fair comparison with prior structure learning methods, we follow the common practice of selecting the threshold that maximizes validation F1 score. The full procedure is presented in Algorithm 3. The optimal threshold selection plots for the directed and bidirected adjacency matrices on the FC dataset are shown in Figure 4, highlighting the threshold values that yield the highest F1 score for each edge type.