A1 THEORETICAL GUARANTEE

In this section, we delve into the theoretical analysis concerning the asymptotic correctness of our proposed model with respect to the sample size. Sec. A1.1 lays out the essential definitions and assumptions pertinent to the problem under study. Following this, from Sec. A1.2 to A1.3, we rigorously demonstrate the asymptotic correctness of the neural network model. Finally, in Sec. A1.4, we engage in a detailed discussion about the practical advantages and superiority of neural network models.

A1.1 DEFINITIONS AND ASSUMPTIONS

As outlined in Sec. 3, a Causal Graphical Model is defined by a joint probability distribution P over d random variables $X_1, X_2, \cdots, X_d$, and a DAG G with d vertices representing the d variables. An observational dataset D consists of n records and d columns, which represents n instances drawn i.i.d. from P. In this work, we assume causal sufficiency:

Assumption A1.1 (Causal Sufficiency)**.** *There are no latent common causes of any of the variables in the graph.*

Moreover, we assume the data distribution P is Markovian to the DAG G:

Assumption A1.2 (Markov Factorization Property)**.** *Given a joint probability distribution P and a DAG G, P is said to satisfy Markov factorization property w.r.t. G if $P := P(X_1, X_2, \cdots, X_d) = \prod_{i=1}^{d} P\left(X_i \mid \mathrm{pa}_i^G\right)$, where pa_i^G is the parent set of X_i in G.*

It is noteworthy that the Markov factorization property is equivalent to the Global Markov Property (GMP) (Lauritzen, 1996), which is

Definition A1.1 (Global Markov Property (GMP))**.** *P is said to satisfy GMP (or Markovian) w.r.t. a DAG G if $X \perp_G Y | Z \Rightarrow X \perp Y | Z$. Here $\perp_G$ denotes d-separation, and $\perp$ denotes statistical independence.*

GMP indicates that any d-separation in graph G implies conditional independence in distribution P. We further assume that P is faithful to G by

Assumption A1.3 (Faithfulness)**.** *Distribution P is faithful w.r.t. a DAG G if $X \perp Y | Z \Rightarrow X \perp_G Y | Z$.*

Definition A1.2 (Canonical Assumption)**.** *We say our settings satisfy the canonical assumption if the Assumptions A1.1 - A1.3 are all satisfied.*

We restate the definitions of skeletons, Unshielded Triples (UTs) and v-strucutres as follows.

Definition A1.3 (Skeleton)**.** *A skeleton E defined over the data distribution P is an undirected graph where an edge exists between X_i and X_j if and only if X_i and X_j are always dependent in P, i.e., $\forall Z \subseteq \{X_1, X_2, \cdots, X_d\} \setminus \{X_i, X_j\}$, we have $X_i \not\perp\!\!\!\perp X_j | Z$.*

Under our assumptions, the skeleton is the same as the corresponding undirected graph of G Spirtes et al. (2000).

Definition A1.4 (Unshielded Triples (UTs) and V-structures)**.** *A triple of variables X, T, Y is an Unshield Triple (UT) denoted as $\langle X, T, Y \rangle$, if X and Y are both adjacent to T but not adjacent to each other in the DAG G or the corresponding skeleton. It becomes a v-structure denoted as $X \rightarrow T \leftarrow Y$, if the directions of the edges are from X and Y to T in G.*

We introduce the definition of separation set as:

Definition A1.5 (Separation Set)**.** *For a node pair X_i and X_j, a vertex set Z is a separation set if $X_i \perp X_j | Z$. Under faithfulness assumption, a separation set Z is a subset of variables within the vicinity that d-separates X_i and X_j.*

Finally, we assume a neural network can be used as a universal approximator in our settings.

Assumption A1.4 (Universal Approximation Capability)**.** *A neural network model can be trained to approximate a function under our settings with arbitary accuracy.*

A1.2 SKELETON LEARNING

In this section, we prove the asymptotic correctness of neural networks on the skeleton prediction task by constructing a perfect model and then approximating it with neural networks. For the sake of convenience and brevity in description, we define the skeleton predictor as follows.

Definition A1.6 (Skeleton Predictor)**.** *Given observational data D, a skeleton predictor is a predicate function with domain as observational data D and predicts the adjacency between each pair of the vertices.*

Now we restate the Remark from Ma et al. (2022) as the following proposition. It proves the existence of a perfect skeleton predictor by viewing the skeleton prediction step of PC (Spirtes et al., 2000) as a skeleton predictor, which is proved to be sound and complete.

Proposition A1.1 (Existence of a Perfect Skeleton Predictor)**.** *There exists a skeleton predictor that always yields the correct skeleton with sufficient samples in D.*

Proof. We construct a skeleton predictor SP consisting of two parts by viewing PC (Spirtes et al., 2000) as a skeleton predictor. In the first part, it extracts a pairwise feature $\boldsymbol{x}_{ij}$ for each pair of nodes X_i and X_j:

$$\boldsymbol{x}_{ij} = \min_{Z \subseteq V \setminus \{X_i, X_j\}} \left\{ X_i \sim X_j \mid Z \right\}, \tag{1}$$

where $\left\{ X_i \sim X_j \mid Z \right\} \in [0, 1]$ is a scalar value that measures the conditional dependency between X_i and X_j given a vertex subset Z. Consequently, $\boldsymbol{x}_{ij} > 0$ indicates the persistent dependency between the two nodes.

In the second part, it predicts the adjacency based on $\boldsymbol{x}_{ij}$:

$$\left(X_i, X_j\right) = \begin{cases} 1 \text{ (adjacent)} & \boldsymbol{x}_{ij} \neq 0 \\ 0 \text{ (non-adjacent)} & \boldsymbol{x}_{ij} = 0 \end{cases} \tag{2}$$

Now we prove that SP always yields the correct skeleton by proving the absence of false positive predictions and false negative predictions. Here, false positive prediction denotes SP predicts a non-adjacent node pair as adjacent and false negative predictions denote SP predicts an adjacent node pair as non-adjacent.

- **False Positive.** Suppose X_i, X_j are non-adjacent. Under the Markovian assumption, there exists a set of nodes Z such that $\left\{ X_i \sim X_j \mid Z \right\} = 0$ and hence $\boldsymbol{x}_{ij} = 0$. According to Equation (2), SP will always predicts them as non-adjacent.
- **False Negative**. Suppose X_i, X_j are adjacent. Under the faithfulness assumption, for any $Z \in V \setminus \left\{X_i, X_j\right\}, \left\{ X_i \sim X_j \mid Z \right\} > 0$, which implies $\boldsymbol{x}_{ij} > 0$. Therefore, SP always predicts them as adjacent.

Therefore, SP never yields any false positive predictions or false negative predictions under the Markovian assumption and faithfulness assumption, i.e., it always yields the correct skeleton. □

With the existence of a perfect skeleton predictor, we prove the correctness of neural network models with sufficient samples under our assumptions.

Theorem A1.1. *Under the canonical assumption and the assumption that neural network can be used as a universal approximator (Assumption A1.4), there exists a neural network model that always predicts the correct skeleton with sufficient samples in D.*

Proof. From Proposition A1.1, there exists a perfect skeleton predictor that predicts the correct skeleton. Thus, according to the Assumption A1.4, a neural network model can be trained to approximate the perfect skeleton prediction hence predicts the correct skeleton. □

A1.3 ORIENTATION LEARNING

Similarly to the overall thought process in Sec. A1.2, in this section we prove the asymptotic correctness of neural networks on the v-structure prediction task by constructing a perfect model and then approximating it with neural networks.

Definition A1.7 (V-structure Predictor)**.** *Given observational data D with sufficient samples from a BN with vertices $V = \{X_1, \ldots, X_p\}$, a v-structure predictor is a predicate function with domain as observational data D and predicts existence of the v-structure for each unshielded triple.*

The following proposition proves the existence of a perfect v-structure predictor by viewing the orientation step of PC (Spirtes et al., 2000) as a v-structure predictor.

Proposition A1.2 (Existence of a Perfect V-structure Predictor)**.** *Under the Markov assumption and faithfulness assumption, there exists skeleton predictor that always yields the correct skeleton.*

Proof. We construct a v-structure predictor VP consisting of two parts by viewing PC (Spirtes et al., 2000) as a v-structure predictor. In the first part, it extracts a feature $\boldsymbol{z}_{ijk}$ for each UT $\langle X_i, X_k, X_j \rangle$:

$$\boldsymbol{z}_{ijk} = \frac{\left|\{(X_k, Z) | \{X_i \sim X_j | Z\} = 0 \wedge X_k \in Z\}\right|}{\left|\{Z | \{X_i \sim X_j | Z\} = 0\}\right|}, \tag{3}$$

□

where $\{X_i \sim X_j \mid Z\} \in [0, 1]$ is a scalar value that measures the conditional dependency between X_i and X_j given a vertex subset Z, and $|\cdot|$ represents the cardinality of a set. Note that the denominator is always positive because the separation set of a UT always exists (See Lemma 4.1 in Dai et al. (2023)). Intuitively, $\boldsymbol{z}_{ijk}$ represents the proportion of supsets of X_i and X_j that include X_k.

In the second part, it predicts the v-structures based on $\boldsymbol{z}_{ijk}$:

$$\langle X_i, X_k, X_j \rangle = \begin{cases} 0 \text{ (not v-structure)} & \boldsymbol{z}_{ijk} \neq 0 \\ 1 \text{ (v-structure)} & \boldsymbol{z}_{ijk} = 0 \end{cases} \tag{4}$$

Now we prove that VP always yields the correct predictions of v-structures. According to Theorem 5.1 on p.410 of Spirtes et al. (2000), assuming faithfulness and sufficient samples, if a UT $\langle X_i, X_k, X_j \rangle$ is a v-structure, then X_k does not belong to any separation sets of (X_i, X_j); if a UT $\langle X_i, X_k, X_j \rangle$ is not a v-structure, then X_k belongs to every separation sets of (X_i, X_j). Therefore, we have $\boldsymbol{z}_{ijk} = 0$ if and only if X_k is not in any separation set of X_i and X_j, i.e., $\langle X_i, X_k, X_j \rangle$ is a v-structure.

With the existence of a perfect v-structure predictor, we prove the correctness of neural network models with sufficient samples under our assumptions.

Theorem A1.2. *Under the canonical assumption and the assumption that neural network can be used as a universal approximator (Assumption A1.4), there exists a neural network model that always predicts the correct v-structures with sufficient samples in D.*

Proof. From Proposition A1.1, there exists a perfect skeleton predictor that predicts the correct v-structures. Thus, according to the Assumption A1.4, a neural network model can be trained to approximate the perfect v-structure predictions hence predicts the correct v-structures. □

A1.4 DISCUSSION

In the sections above, we prove the asymptotic correctness of neural network models by constructing theoretically perfect predictors. These predictors both consist of two parts: feature extractors providing features $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ijk}$, and final predictors of adjacency and v-structures. Even though they have a theoretical guarantee of the correctness with sufficient samples, it is noteworthy that they are hard to be applied practically. For example, to obtain $\boldsymbol{x}_{ij}$ in Equation (1), we need to calculate the conditional dependency between X_i and X_j given every vertex subset $Z \subseteq V \setminus \{X_i, X_j\}$. Leaving

aside the fact that the number of Zs itself presents factorial complexity, the main issue is that when Z is relatively large, due to the curse of dimensionality, it becomes challenging to find sufficient samples to calculate the conditional dependency. This difficulty significantly hampers the ability to apply the constructed prefect predictors in practical scenarios.

Some existing methods can be interpreted as constructing more practical predictors. Majority-PC (MPC) (Colombo et al., 2014) achieves better performance on finite samples by modifying Equation 4 as:

$$\langle X_i, X_k, X_j \rangle = \begin{cases} 0 \text{ (not v-structure)} & z_{ijk} > 0.5 \\ 1 \text{(v-structure)} & z_{ijk} \leq 0.5 \end{cases} \tag{5}$$

Due to its more complex classification mechanism, it achieves better performance empirically. However, from the machine learning perspective, features from both the PC and MPC predictors are relatively simple. As supervised causal learning methods, ML4S (Ma et al., 2022) and ML4C (Dai et al., 2023) provide more systematic featurizations by manual feature engineering and utilization of powerful machine learning models for classification. While these methods show enhanced practical efficacy, their manual feature engineering processes are complex. In our paper, we utilize neural networks as universal approximators for learning the prediction of identifiable causal structures. It not only simplifies the procedure but also potentially uncovers more nuanced and complex patterns within the data that manual methods might overlook. It is noteworthy that the benefits of supervised causal learning using neural networks are also discussed elsewhere, as mentioned in CSIvA (Ke et al., 2023).

A2 Illustration of the Case Study in Sec. 4.3

Figure A4 presents an illustration for the case study of the Bernoulli-sampling adjacency matrix approach in Sec. 4.3. It clearly shows that observational data with the two different parametrized forms follow the same joint distribution:

$$P([X,Y,T]) = \mathcal{N}\left([0,0,0], \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}\right). \tag{6}$$

Therefore, the observational datasets coming from the two DAGs are inherently indistinguishable.

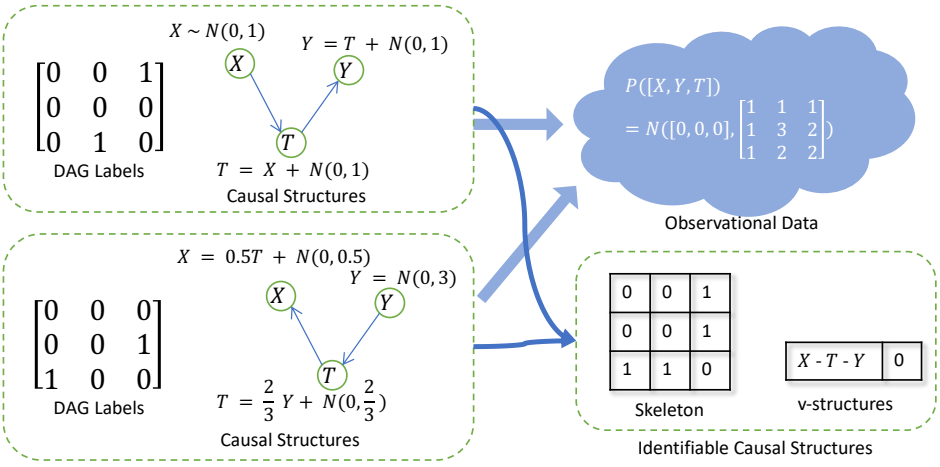


Figure A4: The problem setting to emphasize the limitations of the Bernoulli-sampling adjacency matrix approach. *Best viewed in color.*

A3 Experimental Settings

Metrics. In all tables, ± indicates that the mean value and maximum deviation of three runs with different random seeds are reported. In the field of skeleton prediction tasks, the F1 score has

emerged as a widely adopted metric due to its ability to effectively balance precision and recall (Ding et al., 2020; Ma et al., 2022). This metric provides a comprehensive evaluation of the model's performance, particularly in cases where the data distribution is imbalanced. Accuracy, another commonly used metric, offers a direct measure of the proportion of misclassified edges within the graph. It can also be interpreted as a normalized version of the Structural Hamming Distance (SHD), which has gained popularity in recent years (Ma et al., 2022; Lorch et al., 2022; Ke et al., 2023).

Considering that deep learning models typically output probabilities rather than discrete labels, the Area Under the Receiver Operating Characteristic Curve (AUC) and the Area Under the Precision-Recall Curve (AUPRC) are also employed as more robust metrics. These metrics take into account all possible decision thresholds, providing a comprehensive evaluation of the model's performance across various operating points.

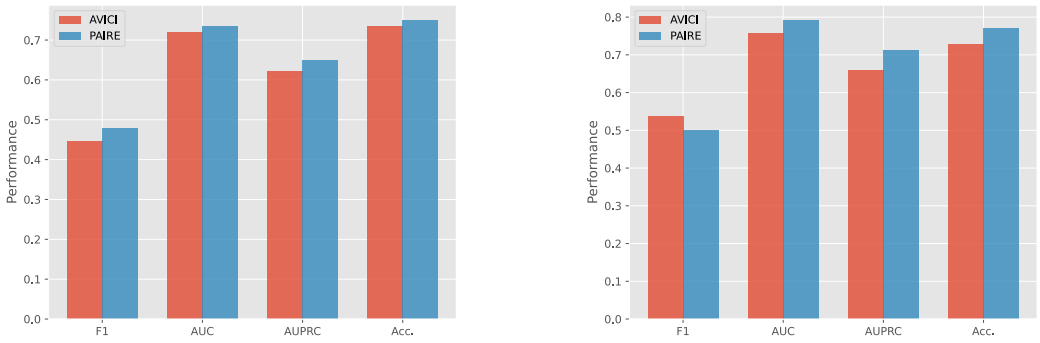
For the CPDAG prediction task, accuracy is used as a comparison metric, which measures the ratio of misclassified edges in the predicted CPDAG. Following the previous paper (Dai et al., 2023), the F1-scores calculated for identifiable edges and v-structures are also provided for a more comprehensive comparison.

Baselines. To demonstrate the effectiveness and superiority of the proposed framework, several strong baselines representing multiple categories are selected for comparison. These baselines include:

1. PC: A classic constraint-based causal discovery algorithm based on conditional independence tests. The version with parallelized optimization is selected (Le et al., 2016).
2. GES: A classic score-based greedy equivalence search algorithm (Chickering, 2002).
3. NOTEARS: A gradient-based algorithm for linear data models (Zheng et al., 2018).
4. DAG-GNN: A continuous optimization algorithm based on graph neural networks (Yu et al., 2019).
5. AVICI: A powerful deep learning-based supervised causal learning method (Lorch et al., 2022).
6. NOTEARS-MLP: A gradient-based algorithm for non-linear data models (Zheng et al., 2018).
7. GOLEM: A more efficient version of NOTEARS (Ng et al., 2020).
8. GraNDAG: A gradient-based algorithm using neural network modeling for non-linear additive noise data (Lachapelle et al., 2020).

The implementation from gCastle (Zhang et al., 2021) is utilized for the first four baselines. Note that the CSIvA model (Ke et al., 2023) is also a closely related method, but it is not compared due to the unavailability of its relevant codes and its requirement for interventional data as input. The original AVICI model (Lorch et al., 2022) does not support discrete data. Therefore, we use an embedding layer to replace its first linear layer when using AVICI on discrete data. All classic algorithms are run on an AMD EPYC 7V13 CPU, and DNN-based methods are run on an Nvidia A100 GPU. Our codes can be accessed at https://anonymous.4open.science/r/paire-C05D.

Synthetic Data. We randomly generate random graphs from multiple random graph models. For continuous data, following previous work (Lorch et al., 2022), Erdős-Rényi (ER) and Scale-free (SF) are utilized as the training graph distribution $p(G)$. The degree of training graphs in our experiments varies randomly among 1, 2, and 3. For testing graph distributions, Watts-Strogatz (WS) and Stochastic Block Model (SBM) are used, with parameters consistent with those in the previous paper (Lorch et al., 2022). All synthetic graphs for continuous data contain 30 nodes. The lattice dimension of Watts-Strogatz (WS) graphs is sampled from $\{2, 3\}$, yielding an average degree of about 4.92. The average degrees of Stochastic Block Model (SBM) graphs are set at 2, following the settings in the aforementioned paper. For discrete data, 11-node graphs are used. SF is utilized as the training graph distribution $p(G)$ and ER is used for testing. The synthetic training data is generated in real-time, and the training process does not use the same data repeatedly. All synthetic test datasets contain 100 graphs, and the average values of the metrics on the 100 graphs are reported to comprehensively reflect the performance.



(a) Linear Gaussian

(b) General Nonlinear

Figure A5: Compairson between AVICI and PAIRE of Skeleton prediction task on WS Graphs.

For the forward sampling process from graph to continuous data, both the linear Gaussian mechanism and general nonlinear mechanism are applied. Concretely, the Random Fourier Function mechanism is used for the general nonlinear data following the previous paper (Lorch et al., 2022). In synthesizing discrete datasets, the Bernoulli distribution is used following previous papers (Dai et al., 2023; Ma et al., 2022).

Post-processing. Although our method theoretically guarantees asymptotic correctness, in practice, conflicts in predicted v-structures might occasionally occur in practice. Therefore, in the post-processing stage, we apply a straightforward heuristic to resolve the potential conflicts among predicted v-structures following previous work (Dai et al., 2023). After that, we use an improved version of Meek rules (Meek, 1995a; Tsagris, 2019) to obtain other identifiable edges without introducing extra cycles. Combining the skeleton from the skeleton predictor model with all identifiable edge directions, we get the final output of the CPDAG.

A4 Extra Experimental Results

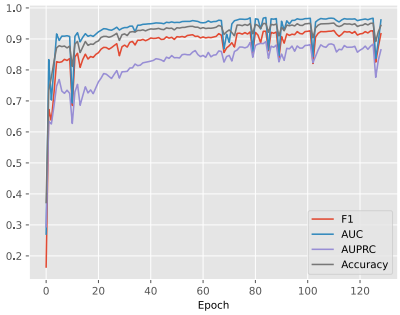
Table A6: Comparison of skeleton prediction task on ER random graphs of discrete data.

Method	F1	AUC	AUPRC	Accuracy
AVICI	0.833	0.961	0.925	0.914
PAIRE	**0.862**	**0.976**	**0.952**	**0.921**

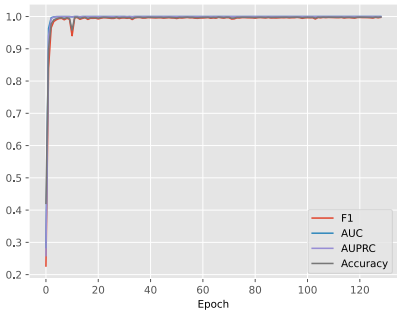
More Comparisons on Continuous Data. Table A7 - A8 presents additionally comparison between PAIRE and more baseline methods on the WS dataset for both skeleton prediction task and CPDAG prediction task. PAIRE consistently demonstrates superior performance in comparison with these methods. Figure A5 presents an experimental comparison between AVICI and PAIRE on WS random graphs for skeleton prediction. These results reinforce the analysis in Sec. 5.3 and demonstrate the effectiveness of the proposed pairwise encoder module.

Figure A6 illustrates the test performance trends of the v-structure prediction model on SBM and WS random graphs during the training process. In this model, the feature extractor FE is fine-tuned from the skeleton prediction model. The performance increases rapidly and achieves a relatively high level after just a few initial epochs. This suggests that the v-structure task is relatively straightforward, and the pre-trained pairwise features from the skeleton prediction model are both effective and generalizable.

More Comparisons on Synthetic Discrete Data. Since neural network model outputs range from 0 to 1 as probabilities rather than single predictions, AUC and AUPRC are more appropriate metrics



(a) Variation trends of test performance on WS graph.



(b) Variation trends of test performance on SBM graph.

Figure A6: Variation trends of the test performance of v-structure prediction model on WS and SBM random graphs during training.

Table A7: More comparison of skeleton prediction results on linear Gaussian data and WS graphs.

Method	Skeleton F1	Skeleton AUC
GRANDAG	0.163	0.502
NOTEARS-MLP	0.256	0.513
GOLEM	0.293	0.539
PAIRE	$\mathbf{0.479 \pm 0.015}$	$\mathbf{0.750 \pm 0.003}$

for comparing DNN-based SCL methods. The comparison between AVICI and PAIRE for the skeleton prediction task on discrete data is provided in Table A6. By adjusting the classification threshold of DNN-based SCL methods, we can also compare them with traditional methods. These results are presented in Table A9. It is evident that DNN-based SCL methods outperform their counterparts, with PAIRE consistently achieving the best performance under these conditions.

More Results on Sachs. We present the comparison on F1 score and accuracy on the Sachs dataset in Figure A7. The results demonstrate that DNN-based SCL methods consistently outperform classical approaches, thereby confirming their effectiveness and superiority in this context.

Table A9: Comparison of skeleton prediction task on ER random graphs of discrete data.

Method	F1	Accuracy
PC	0.822	0.830
GES	0.821	0.818
NOTEARS	0.164	0.747
AVICI	0.833	0.914
PAIRE	**0.862**	**0.921**

Training Data Diversity and Model Generalization. We present experimental evidence that highlights the significant contribution of training data diversity to the model's generalization capabilities, even when applied to out-of-distribution (OOD) datasets. To illustrate this, we trained one PAIRE model on a combined dataset of both SF and ER, and another solely on the SF dataset. The comparative performance of these models is detailed in Table A10. The model trained on the combined ER and SF datasets exhibited markedly better per-

Table A8: More comparison of CPDAG prediction results on linear Gaussian data and WS graphs.

Method	V-structure F1	Identfiable edges F1	SHD
GRANDAG	0.116	0.115	169.48
NOTEARS-MLP	0.128	0.126	192.59
GOLEM	0.158	0.191	172.63
PAIRE	$\mathbf{0.298 \pm 0.076}$	$\mathbf{0.370 \pm 0.062}$	$\mathbf{116.797 \pm 7.253}$

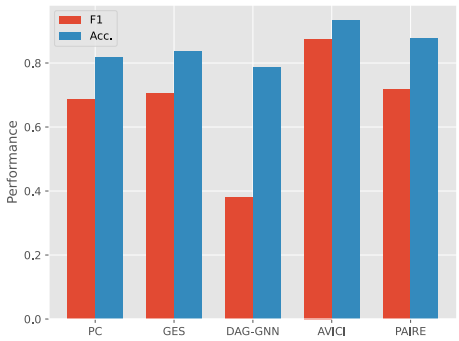


Figure A7: Skeleton prediction results on the Sachs dataset.

Table A10: Comparison of PAIRE models with different training data diversity on skeleton prediction.

(a) Model trained on both ER and SF

Test Dataset	F1	AUC	AUPRC	Accuracy
WS	0.3631	0.7056	0.6061	0.7329
SBM	0.7811	0.9675	0.8809	0.9483
ER	0.8069	0.9603	0.8918	0.9473
SF	0.8473	0.9845	0.9355	0.9551

(b) Model trained on SF

Test Dataset	F1	AUC	AUPRC	Accuracy
WS	0.4011	0.6300	0.4609	0.6353
SBM	0.6427	0.9165	0.7287	0.9087
ER	0.6706	0.9042	0.7393	0.9080
SF	0.8783	0.9886	0.9529	0.9611

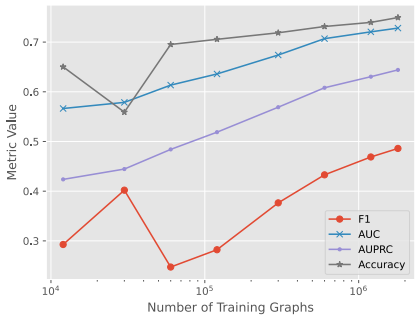
formance, not only on the ER dataset but also on the other two OOD datasets, with only a marginal decrease in performance on the SF dataset. These findings suggest that enhancing the diversity of the training data correspondingly improves the model's ability to generalize and maintain robust performance across novel OOD datasets.

Varying Amount of Training Graphs. We present an analysis of how varying the amount of the training graphs influences performance on the skeleton prediction task. The results, depicted in Figure A8, illustrate a clear trend: model performance improves in tandem with the expansion of the training dataset. This trend underscores the potential of our method to achieve even greater accuracy given a more extensive dataset.

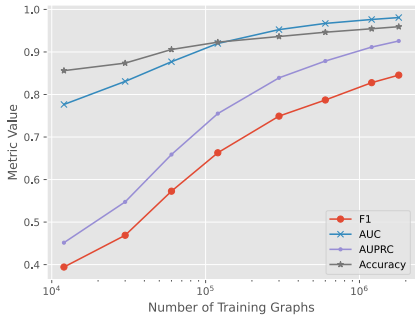
Varying Sample Size. We assessed PAIRE across various quantities of observational samples per graph during testing (100, 200, ..., 1000). The outcomes for both the skeleton prediction task and the CPDAG prediction task are depicted in Figure A9. It is evident that the model's performance enhances with the augmentation of sample size. These consistent upward trends suggest that PAIRE exhibits stability and is not overly sensitive to changes in sample size.

Varying Edge Density. We evaluate PAIRE over a range of edge densities in the test graphs, utilizing the SBM dataset, as it allows for the direct setting of average edge densities. The findings are presented in Figure A10. It's apparent that the task is becomes more difficult as edge densities increase. However, the performance decline is not abrupt, indicating that PAIRE's performance remains relatively stable across various edge densities, thereby confirming its versatility.

Acyclicity. We provide an empirical evidence supporting of the rarity of cycles in the final predictions. The experimental data presented in Table A11 corroborates that cycles are infrequently observed in the final predicted CPDAGs.

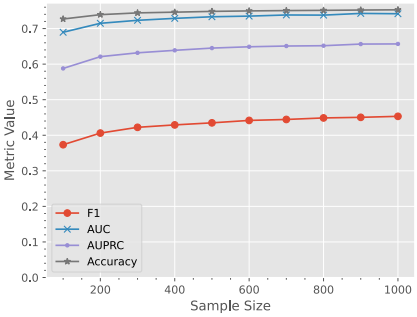


(a) WS dataset

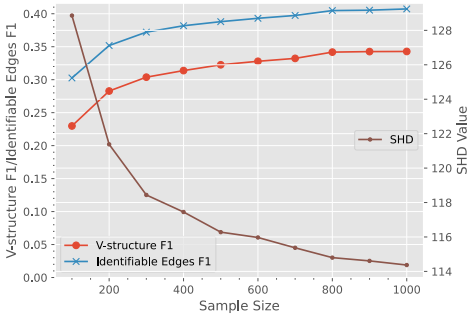


(b) SBM dataset

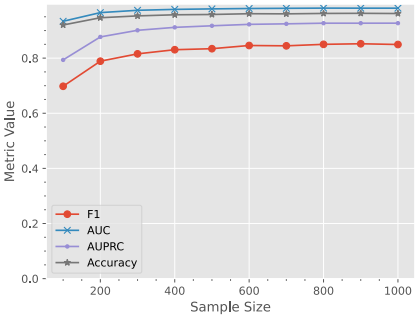
Figure A8: Model performance with varying amount of training graphs.



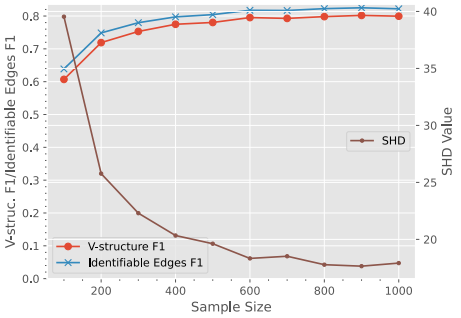
(a) Variation trends of skeleton predicton task performance on WS graph with varying sample sizes.



(b) Variation trends of CPDAG predicton task performance on WS graph with varying sample sizes.

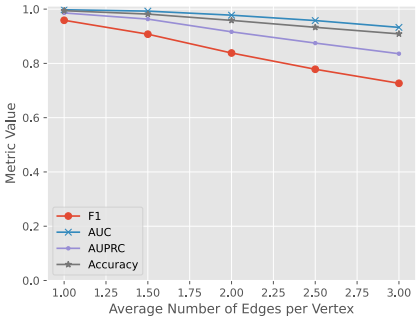


(c) Variation trends of skeleton predicton task performance on SBM graph with varying sample sizes.

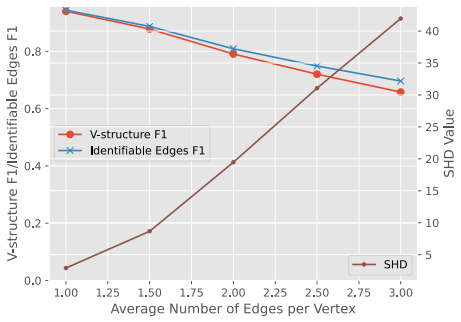


(d) Variation trends of CPDAG predicton task performance on SBM graph with varying sample sizes.

Figure A9: Variation trends of performance with varying sample sizes.



(a) Variation trends of skeleton predicton task performance on SBM graph with varying edge densities.



(b) Variation trends of CPDAG predicton task performance on SBM graph with varying edge densities.

Figure A10: Variation trends of performance with varying edge densities.

Table A11: Count of cycles in the final CPDAG predictions.

Dataset	WS	SBM
Rate of Graphs with Cycles	$0.66 \pm 0.66\%$	$0.00 \pm 0.00\%$

Generality on Testing Graph Sizes. We offer an analytical perspective on the performance of the PAIRE model when applied to larger WS graphs. It is important to highlight that the models were initially trained on graphs comprising 30 vertices, positioning this task within an out-of-distribution setting in terms of graph size. To establish a point of reference, we have included results from the PC algorithm as a baseline comparison. These findings can be examined in Table A12. Despite the OOD conditions, PAIRE maintains robust performance, reinforcing its scalability and the model's general applicability across varying graph sizes.

Table A12: Performance comparison with varying amounts of graph sizes.

Metric	F1 Score			V-structure-F1			Identifiable-Edges F1		
Graph Size	50	70	100	50	70	100	50	70	100
PC	0.1767	0.1484	0.1060	0.0635	0.0504	0.0366	0.0699	0.0559	0.0403
PAIRE	0.4156	0.3738	0.2834	0.3494	0.3070	0.2261	0.3793	0.3369	0.2479