CAUSALITY FOR TABULAR DATA SYNTHESIS: A HIGH-ORDER STRUCTURE CAUSAL BENCHMARK FRAMEWORK

Anonymous authors
Paper under double-blind review

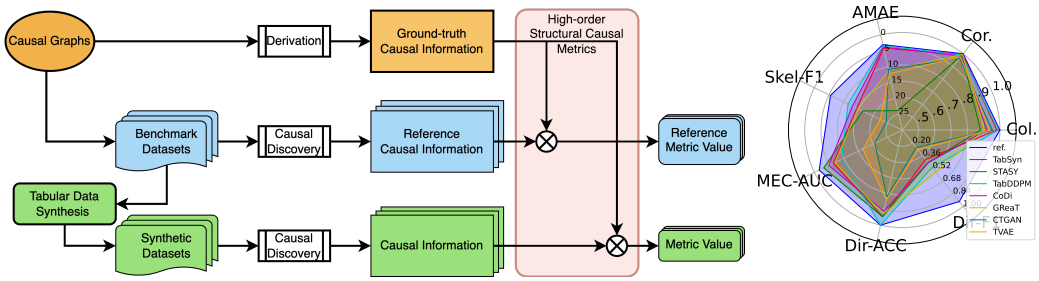
ABSTRACT

Tabular synthesis models remain ineffective in capturing complex dependencies, and the quality of synthetic data is still insufficient for comprehensive downstream tasks, such as prediction under distribution shifts, automated decision-making, and understanding cross-table relationships. A major challenge present in tabular data is the lack of prior knowledge about high-order relationships, which is defined as multivariate structural causal dependencies beyond pairwise correlations. We argue that a systematic evaluation on high-order structural information is a crucial first step in addressing this issue in tabular data synthesis. In this paper, we present high-order structural causal information as a natural form of prior knowledge and introduce a benchmark framework to evaluate tabular synthesis models. This framework allows us to generate benchmark datasets through a flexible range of data generation processes, allowing for the training of tabular synthesis models using these datasets for further evaluation. We propose multiple benchmark tasks, high-order metrics, and causal inference tasks as downstream tasks for evaluating the quality of synthetic data generated by the trained models. Our experiments demonstrate the effectiveness of the benchmark framework in evaluating the model's ability to capture high-order structural causal information. Furthermore, our benchmarking results provide an initial assessment of state-of-the-art tabular synthesis models. These results reveal significant gaps between ideal and actual performance and highlight how baseline methods differ. We open source the benchmark framework, including both code and data along with documentation, to support further research and development in this area.

1 INTRODUCTION

Tabular data are widely used in both industry and natural sciences, yet tabular data remain underexplored in machine learning research (van Breugel & van der Schaar, 2024). Among the various tasks in the tabular domain, data synthesis is particularly important due to its many applications, such as data augmentation to mitigate data scarcity (Choi et al., 2017), pretraining for downstream tasks (Hollmann et al., 2023), and privacy preservation (Hernandez et al., 2022). Recently, the quality of synthetic tabular data has significantly improved with deep diffusion models (DFMs) (Ho et al., 2020) and large language models (LLMs) (Brown et al., 2020).

Nevertheless, tabular data synthesis continues to face several challenges. These challenges fall into three broad categories: (i) handling practical issues, such as mixed data types (Ma et al., 2020) and missing data; (ii) capturing structural information inherent to the nature of tabular data, particularly high-order instance and feature dependencies (Li et al., 2023), where "high-order" refers to multivariate structural causal information that captures dependencies beyond pairwise relationships, typically represented through causal graphs or skeletons; (iii) synthesizing data in a cross-table context, such as capturing dependencies across different tables (Scetbon et al., 2024). While many studies have focused on addressing practical issues (Kotelnikov et al., 2023; Kim et al., 2023; Lee et al., 2023; Zhang et al., 2024) that are necessary steps for training synthesis models, fewer have addressed high-order information, which is crucial for complex real-world applications such as in-context prediction (Zhu et al., 2023b; Hollmann et al., 2023) and generalization and analysis of multiple tables (Wang & Sun, 2022; Zhu et al., 2023a).



(a) Benchmark framework. (b) Benchmark results.

Figure 1: High-order structural causal benchmark framework and results. Benchmark datasets are generated from sampled causal graphs and used to train tabular synthesis models. Causal information is extracted from both benchmark and synthetic datasets using causal discovery and compared to the ground truth using high-order structural causal metrics. The results highlight model performance across multiple metrics, using benchmark-derived values as references.

One reason for the lack of studies may be the absence of a systematic evaluation of synthesis models on high-order information. This absence not only creates a limited and misleading impression of the performance of synthesis models, but also impedes the development and application of high-order information-aware synthesis models. The evaluation of tabular synthesis models is an active research area. Currently, evaluations primarily focus on the performance of using synthetic data for downstream tasks, known as *extrinsic evaluation* (Bommasani et al., 2021). Extrinsic evaluation offers a limited understanding of tabular synthesis models restricted by the downstream tasks. In contrast, *intrinsic evaluation* directly evaluates the quality of synthetic data using metrics derived from lower-order statistics such as pairwise correlation-based scores. Intrinsic evaluation with high-order metrics is challenging, as it depends on informative prior knowledge that remains underexplored in the tabular domain (van Breugel & van der Schaar, 2024).

To benchmark models on high-order structural information, we propose to leverage the study of causal graphical models (Spirtes et al., 2000; Peters et al., 2017). These models apply causal prior knowledge for a wide range of machine learning applications (Pearl et al., 2016; Schölkopf et al., 2021), yet they are still underexplored in the tabular domain. We view causal graphs as a natural and compact representation of high-order structural information about causal dependencies in tabular data. A few works (Choi et al., 2020; Liu et al., 2023; Yan et al., 2023) have attempted to use general graph properties, such as adjacency matrices and directed acyclicity, for representation learning and tabular data synthesis, but they have all overlooked high-order structural causal information. In contrast, our aim is to establish a foundation that covers the essential aspects required for benchmarking tabular synthesis models on high-order structural causal information. We also demonstrate how to utilize benchmarking results to guide future improvements in tabular synthesis models. Specifically, we

1. introduce high-order structural causal information as prior knowledge for modeling dependencies among the variables of interest in tabular data. We characterize the information into three levels for benchmarking tabular synthesis models (Section 3).
2. propose a benchmark framework for evaluating tabular synthesis models on high-order structural causal information, which is summarized in Figure 1. Specifically, we illustrate how to generate benchmark datasets and how high-order structural causal tasks and downstream causal-inference tasks can be used for evaluation (Section 4).
3. demonstrate using our framework to evaluate state-of-the-art DFM and LLM-based tabular synthesis models on benchmark and real-world datasets (Section 5 and Appendix D). Our experimental results show a clear gap between the ideal and actual performance of the baseline methods, and strong performance on low-order metrics does not guarantee good performance on high-order causal metrics. These results highlight their shortcomings from various perspectives.

2 Related Work

To understand tabular data models from different perspectives and to make progress towards better real-world performance, a suite of benchmarks with different purposes is needed. In prior benchmarking efforts, Grinsztajn et al. (2022) uses diverse tabular datasets to investigate the performance of tree-

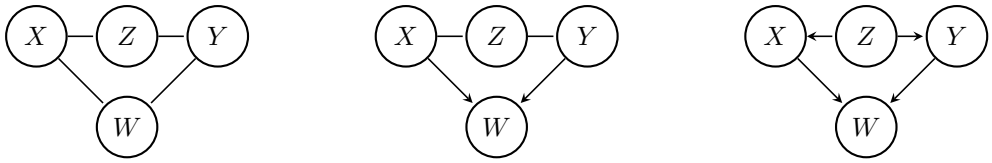
based methods; Bommert et al. (2020), Passemiers et al. (2023) and Cherepanova et al. (2023) contribute benchmark datasets to the studies of feature selection; Malinin et al. (2021) and Gardner et al. (2023) focus on the robustness to distribution shifts. Moreover, to bridge the gap between real-world applications and synthetic data, Jesus et al. (2022) provides a larger-scale tabular dataset in finance, including practical challenges. In contrast, SynthCity (Qian et al., 2023) provides a number of synthetic data generators for better model development by avoiding practical issues of real-world data, such as selection bias and missing value issues. Furthermore, in the context of tabular data synthesis, Hansen et al. (2023) introduces data-centric AI techniques that can provide data profiles, and then propose an evaluation framework to show the importance of integrating data profiles into synthesis models. Unlike previous efforts, our work provides a new perspective for understanding and improving tabular synthesis models by benchmarking on high-order structural causal information. Related to this, recent work has also discussed the challenges of causal benchmarking and the risk of overly simplistic simulated DAGs (Reisach et al., 2021).

Tabular data synthesis was initially based on classical deep generative models (Jordon et al., 2018; Xu et al., 2019; Morales-Alvarez et al., 2022), but has recently flourished with Transformer-based, LLM-based, and DFM-based methods. For example, to generate synthetic tabular data, (Gulati & Roysdon, 2023) applies a masked Transformer (Vaswani et al., 2017); *GReaT* (Borisov et al., 2022) formulates tabular data as sentences and finetunes Generative Pre-trained Transformer 2 (GPT-2) (Radford et al., 2019); *TabDDPM* (Kotelnikov et al., 2023), and *CoDi* (Lee et al., 2023) apply denoising diffusion probabilistic models (Ho et al., 2020); *STASY* (Kim et al., 2023) uses a score-based diffusion model (Song & Ermon, 2019); *TabSyn* leverages a diffusion model within a transformer-based variational autoencoder (Zhang et al., 2024); *TabPFN* Hollmann et al. (2025) is a transformer-based tabular foundation model pretrained on synthetic datasets generated from a wide range of causal graphs; *Forest-VP* and *Forest-Flow* (Jolicoeur-Martineau et al., 2024) use tree-based (Chen et al., 2015) deep diffusion and flow matching models (Lipman et al., 2023); *CTGAN* utilizes a conditional generative adversarial network (Xu et al., 2019); *TVAE* utilizes a variational autoencoder (Xu et al., 2019); . Additionally, *GOGGLE* (Liu et al., 2023) uses an encoder-decoder model to generate tabular data, where the decoder is a graph neural network to capture the dependencies between features. Recent work on causal normalizing flows (Javaloy et al., 2024) also advances the modeling of complex causal relationships in generative settings. Given the increasing number of tabular synthesis models, we mainly mentioned the representative ones for each category and benchmarked most of them in the experiments. However, we highly recommend survey papers (Borisov et al., 2022; Li et al., 2023; Fang et al., 2024) for more details on tabular synthesis models.

3 High-Order Structural Causal Information

A fundamental problem in modeling tabular data is the lack of prior knowledge about their structures and high-order information (Borisov et al., 2022; Fang et al., 2024). Natural and common prior knowledge in tabular domain can be causal dependencies in forms of causal graphs (Peters et al., 2017; Glymour & Zhang, 2019). Real-world data are generated by certain underlying mechanisms, which can be qualitatively described using causal graphs. As for tabular data whose columns are variables of interests, *causal graphs* are directed graphs where the nodes are variables and the directed edges represent causal relationships between the columns. Different from mere pair-wise information (e.g., correlations), such causal relationships represent high-order structural information. Capturing this type of high-order information requires methods that go beyond pair-wise reasoning. In our framework, we assume the causal graphs are directed acyclic graphs (DAGs) without unknown confounders. These are common assumptions in causal machine learning (Peters et al., 2017; Schölkopf et al., 2021) under which the studies are significantly supported by well-studied properties and theories as well as reliable methods and considerable applications. Given a causal DAG, the causal information is the high-order statistical information which under proper assumptions has asymmetric properties implying direct causes and effects in the data generation process. We categorize causal information into three hierarchical levels: causal skeleton, Markov equivalence class, and causal DAG, each capturing progressively richer structural details. While lower levels are easier to obtain, they convey less causal information than the higher ones. These levels correspond to different forms of causal graphs, as illustrated in Figure 2.

Level 1: Causal skeleton. *Causal skeletons* are the undirected graphs of causal DAGs (Spirtes et al., 2000), and describe the connectivity of the nodes. At this level, causal information represents



(a) Causal skeleton level: Undirected causal graphs show the connectivity entailing whether pairs of nodes are d-separated or not.

(b) Markov equivalent class level: Completed partially oriented directed acyclic graphs provide all d-separation information.

(c) Causal DAG (directed acyclic graph) level: It provides all d-separation and causal-direction information.

Figure 2: Three levels of high-order structural causal information.

d-separation relationships as a type of high-order information. Given a causal DAG, two nodes X and Y are d-separated by a set $\boldsymbol{Z}$ if and only if for each path between X and Y, there is a chain $\cdot \to Z \to \cdot$ or a fork $\cdot \leftarrow Z \to \cdot$ such that Z is in $\boldsymbol{Z}$; or it contains a collider $\cdot \to W \leftarrow \cdot$ such that W and its descendants are not in $\boldsymbol{Z}$ (Peters et al., 2017). Therefore, given a causal skeleton, we know the nodes that are d-separated; and further under the *causal sufficiency assumption*, i.e., there are no unmeasured confounders (the same common parent of children nodes), the connectivity of two nodes in causal skeletons infers whether a pair of variables are causally dependent or not.

Level 2: Markov equivalent class. At this level, we not only know the d-separation relationships of node pairs, but we also know by which nodes set they are d-separated. Two DAGs are (Markov) equivalent if and only if they have the same d-separation relationships. The Markov equivalent class can be uniquely represented by a completed partially directed acyclic graph (CPDAG) under proper assumptions (Meek, 1995; Andersson et al., 1997), which can provide complete d-separation information alongside certain causal directions, enriching the information of the causal skeleton.

Level 3: Causal DAG. At this level, all information about connectivity and causal directions is summarized in causal DAGs. This causal information goes beyond knowing all the d-separation relationships. Given the nodes in a causal DAG, we know all of their asymmetric causal relationships, i.e., their direct causes and direct effects.

4 High-Order Structural Causal Benchmark Framework

To support the development of reliable tabular synthesis models, especially those based on deep generative frameworks, a rigorous evaluation is essential. In particular, intrinsic evaluation, assessing the quality of synthetic data independently of specific downstream tasks, plays a critical role (van Breugel & van der Schaar, 2023). As discussed in Section 3, high-order structural causal information provides a principled basis for intrinsic evaluation by directly measuring a model's ability to capture complex causal dependencies (Schölkopf et al., 2021).

To enable such evaluation, our benchmark framework (Figure 1a) generates synthetic benchmark datasets using predefined causal graphs and diverse data generation processes (Section 4.1). These causal graphs provide the ground-truth structural labels required for metric computation. We then apply causal discovery methods to extract different levels of causal information[1] from both benchmark and synthetic datasets, and we compare each of them to the ground truth labels, allowing us to define high-order evaluation metrics at multiple levels (Section 4.2). We also include causal inference tasks as complementary downstream evaluations. For clarity, we use the term *benchmark datasets* to refer to synthetic data generated from causal graphs for training synthesis models, and *synthetic datasets* to refer to the outputs of these trained models.

4.1 Generation of benchmark datasets

The validity of benchmark datasets requires that (i) data are generated according to causal directed acyclic graphs and (ii) causal information (causal skeletons or directions) is identifiable from the data by causal discovery methods under proper assumptions. Condition (i) requires that each benchmark dataset has a corresponding causal DAG and a causal DAG can be used to generate a dataset for

[1]Causal skeletons and Markov equivalence classes can be identified by constraint-based methods (Spirtes et al., 2000), while causal directions in DAGs require functional causal model-based methods (Glymour & Zhang, 2019). See Appendix A for an overview of causal discovery and identifiability assumptions.

evaluation. Regarding Condition (i), the data generation process according to a causal DAG is

$$X_i = f_i(X_i^{\texttt{prt}}, E_{X_i}), \tag{1}$$

where X_i is a variable in the tabular dataset and a node in a causal DAG, $X_i^{\texttt{prt}}$ is the parents of X_i in the graph, f_i is the causal functional relationship between parent and child variables, and E_{X_i} is noise that is independent of $X_i^{\texttt{prt}}$. Condition (ii) requires that causal discovery methods validly recover causal information from a benchmark dataset. We limit data generation processes, i.e., the functional relationships and noise distributions in Equation equation 1, to the identifiable ones for causal discovery methods under proper assumptions (Glymour & Zhang, 2019). Therefore, we categorize the data generation processes by their functional relationships and noise distributions. More specifically, our data generation processes use three types of functional relationships – linear (L), sigmoid (S), and neural network-based (N) – denoted by f_i, combined with two types of noise distributions for E_{X_i}: Gaussian (G) and Uniform (U). For example, we denote the benchmark dataset generated with a linear functional relationship and Gaussian variables by "LG". We modify `CausalDiscoveryToolbox` to randomly generating benchmark datasets with continuous values according to a given causal DAG. Details of data generation processes are in Appendix C.

4.2 Benchmark metrics

Benchmark datasets are generated from predefined DAGs and used to train tabular synthesis models, and their corresponding causal DAGs are used to derive ground-truth labels of causal information for evaluation. We define high-order structural causal metrics by applying different causal discovery methods to infer causal information (typically unavailable) at different levels from synthetic data. Roughly speaking, our metrics compare how well causal information can be recovered by measuring the deviation of discovered DAGs - from both benchmark and synthetic data - from the ground-truth causal labels; cf. Figure 1a. Besides causal information at different levels, metrics also indicate model capabilities to capture joint or individual information, depending on the task. For example, individual causal information can be d-separations and causal directions. Joint causal information is based on the aggregation and integration of individual causal information.

Metrics on causal skeletons. Causal skeletons can be determined by constraint-based causal discovery methods. In our experiments, we apply PC algorithm (Spirtes et al., 2000) to benchmark datasets and synthetic datasets and then get the adjacency matrices of causal skeletons. Furthermore, given the resulting adjacency matrices and the adjacency matrices derived from ground-truth causal DAGs, structural Hamming distance (SHD), recall, precision, and F1 score can be used to measure the differences between the resulted and the ground-truth adjacency matrices. Such metrics also indicate model capability of capturing joint causal information, because causal skeletons are constructed by summarizing multiple d-separations.

Metrics on conditional independence relationships. Under causal sufficiency, faithfulness, and causal Markov assumptions, conditional independence in data implies d-separation in a causal graph (Spirtes et al., 2000). We use conditional independence relationships for benchmarking on the Markov equivalent level. The task based on individual causal information without requiring integrating d-separations. We first select a d-separation and d-connection set with the same set sizes denoted by $\mathbf{D} = \{(X_i, Y_i, \mathbf{S}_i)\}_{i=1:N}$, where X_i and Y_i are either d-connected or d-separated conditioning on the set $\mathbf{S}_i$. We then apply conditional independence tests to the selected subsets of benchmark and synthetic datasets and get results $\mathbf{C}^{\texttt{ref}} = \{c_i^{\texttt{ref}} : 0 \text{ or } 1\}_{i\in\mathbf{D}}$ and $\mathbf{C}^{\texttt{syn}} = \{c_i^{\texttt{syn}} : 0 \text{ or } 1\}_{i\in\mathbf{D}}$, where 0 and 1 represent conditional dependence and independence respectively; and derive the ground-truth conditional independence relationships from ground-truth causal DAGs denoted by $\mathbf{C}^{\texttt{gt}} = \{c_i^{\texttt{gt}} : 0 \text{ or } 1\}_{i\in\mathbf{D}}$. Considering the evaluation on this level as the evaluation of a binary classification problem, Area Under the Curve (AUC) scores of Receiver Operating Characteristic (ROC) curves are used as a metric. Specifically, we are comparing the discovered conditional dependence and independence labels $\mathbf{C}^{\texttt{ref}}$ and $\mathbf{C}^{\texttt{syn}}$ against the ground truth labels $\mathbf{C}^{\texttt{gt}}$.

Metrics on causal directions. As for methods identifying causal directions, bivariate causal discovery methods (Hoyer et al., 2008; Janzing et al., 2012) are commonly available. Different from the other metrics, the metric using bivariate causal discovery methods is based on the bivariate setting. We first select a set of edges from the ground-truth causal DAGs denoted by $\mathbf{E} = \{(X_i, Y_i)\}_{i=i:N}$, of which X_i and Y_i are d-separated after removing the edges between them. In this way, we can apply bivariate causal discovery methods to the data of X_i and Y_i without the impact of the other paths

between them on the causal direction of the edge between them. We apply bivariate causal discovery methods on the selected subsets of benchmark and synthetic datasets and get the results denoted by $\mathbf{E}^{\texttt{ref}} = \{e_i^{\texttt{ref}} : 0 \text{ or } 1\}_{i \in \mathbf{E}}$ and $\mathbf{E}^{\texttt{syn}} = \{e_i^{\texttt{syn}} : 0 \text{ or } 1\}_{i \in \mathbf{E}}$; and derive the ground-truth conditional independence relationships from ground-truth causal DAGs denoted by $\mathbf{E}^{\texttt{gt}} = \{e_i^{\texttt{gt}} : 0 \text{ or } 1\}_{i \in \mathbf{E}}$, where 0 and 1 represent different causal directions. The metric on this level is the accuracy of the predicted results, $\mathbf{E}^{\texttt{ref}}$ and $\mathbf{E}^{\texttt{syn}}$, compared to the ground-truth labels $\mathbf{E}^{\texttt{gt}}$. As a result, the evaluation with bivariate causal discovery methods is based on individual causal information. In addition to bivariate methods, LiNGAM (Linear Non-Gaussian Acyclic Model)-based approaches (Shimizu et al., 2006; 2011) can identify causal directions in the multivariate linear non-Gaussian setting. In this case, SHD, precision, recall, and F1 score are calculated by comparing the resulting fully oriented causal graphs to the ground truth DAGs, similar to the metrics used at the causal skeleton level. This evaluation reflects joint causal information.

Metrics on downstream tasks. Evaluating tabular synthesis models on downstream tasks helps assess how well the generated data supports causal reasoning and decision-making. In our framework, this involves training Structural Causal Models (SCMs) on synthetic data and evaluating their performance on downstream causal inference tasks using held-out benchmark data (Zhang et al., 2024; Fang et al., 2024). We focus on interventional and counterfactual inference tasks, as their performance directly reflects a model's ability to capture essential causal information, which is crucial for our benchmarking objective. Our evaluation and metrics are inspired by Chen et al. (2023). Firstly, benchmark and synthetic data are used for training SCMs given corresponding causal graphs. In the interventional inference task, we perform a series of interventions on each variable in the causal graph, one at a time, and utilize the trained SCM to compute the resulting interventional distributions over the remaining variables. Furthermore, we compute the average differences between the expectation of interventional distributions generated by SCM models trained on synthetic data and benchmark data. For the counterfactual inference task, we generate new observations with the ground-truth SCM for each causal graph. We then compute their counterfactual values with trained SCMs by imposing interventions on each variable individually. The metric is based on the average differences of average counterfactual values between SCM models trained on synthetic data and benchmark data. These metrics are detailed in Section 5.1.

5 Experiments

We begin by outlining the evaluation procedures and experimental settings. Section 5.2 presents benchmarking results of state-of-the-art tabular synthesis models on high-order structural causal tasks using synthetic datasets generated from known causal DAGs. We further evaluate these methods on downstream causal inference tasks and examine their performance on a widely adopted real-world dataset, highlighting the capabilities and limitations of each baseline method. Additional experimental details, benchmark configurations, implementation specifics, and supplementary results (including metrics such as α-precision, β-recall, single-variable density estimation, and pairwise correlation scores (Alaa et al., 2022; Zhang et al., 2024)) are provided in Appendix D, alongside further insights into the high-order metrics introduced in Section 4.

5.1 Experimental settings

Our baseline methods cover LLM-based and DFM-based methods, which are TabSyn (Zhang et al., 2024), STASY (Kim et al., 2023), TabDDPM (Kotelnikov et al., 2023), CoDi (Lee et al., 2023), GReaT (Borisov et al., 2022), CTGAN (Xu et al., 2019), and TVAE (Xu et al., 2019). Firstly, to benchmark baseline methods on high-order structural causal information, N_g causal DAGs $\mathcal{G}^{\texttt{gt}} = \{G_g^{\texttt{gt}}\}_{g=1:N_g}$ are randomly generated and each causal DAG is used for generating benchmark datasets $\mathcal{D}_g^{\texttt{gt}} = \{D_{g,m}^{\texttt{gt}}\}_{m \in \omega}$ with different causal mechanisms $\omega = \{\text{LG, LU, SG, NG}\}$. We then train baseline methods on benchmark datasets $D_{g,m}^{\texttt{gt}}$ and generate synthetic datasets $\mathcal{D}_g^{\texttt{syn}} = \{D_{g,m}^{\texttt{syn}}\}_{m \in \omega}$ for each $G_g^{\texttt{gt}}$. Secondly, with the causal DAGs, benchmark datasets, and synthetic datasets, causal information is identified by causal discovery methods. And causal information, such as adjacency matrices, conditional independence relationships, and predicted edge directions, is denoted by $Q_{g,m}^{\texttt{ref}} := \texttt{CD}(D_{g,m}^{\texttt{gt}})$; $Q_{g,m}^{\texttt{syn}} := \texttt{CD}(D_{g,m}^{\texttt{syn}})$, where $\texttt{CD}$ are causal discovery methods, and ground-truth causal information is derived from causal DAGs, denoted by $Q_g^{\texttt{gt}}$. We include the results on benchmark datasets as a reference ceiling to contextualize performance on synthetic data. In all cases or causal

levels, causal information discovered from both benchmark and synthetic data is evaluated against ground-truth DAGs, we do not compare one discovered DAG to another. Formally, for each metric $M_i \in \mathcal{M} = \{\text{F1, SHD, others in Section 4.2}\}$, we compute $R^{\texttt{ref}}_{g,m} = M_i(Q^{\texttt{ref}}_{g,m}, Q^{\texttt{gt}}_{g})$; $R^{\texttt{syn}}_{g,m} = M_i(Q^{\texttt{syn}}_{g,m}, Q^{\texttt{gt}}_{g})$, to evaluate all baseline methods. For different evaluation purposes, we can aggregate the metric values along different indices and make conclusions. In our experiments, we compute average metric values over all causal DAGs for each causal mechanism:

$$R^{\texttt{ref}}_{m} = \texttt{AVE}_g(R^{\texttt{ref}}_{g,m}) \quad \text{and} \quad R^{\texttt{syn}}_{m} = \texttt{AVE}_g(R^{\texttt{syn}}_{g,m}).$$

Benchmark on the level of causal skeleton. For each causal mechanism, we generate 10 benchmark datasets (according to 10 causal DAGs) with N variables, a flexible parameter in our framework, and around 17,000 samples. For computing metric values, we get 10 bootstrapping datasets with sample size 15,000 for each benchmark and synthetic dataset, and apply PC algorithm to obtain causal information quantities, denoted by $Q^{\texttt{ref}}_{g,m,b}$ and $Q^{\texttt{syn}}_{g,m,b}$ where the bootstrapping index is $b = \{1, ..., 10\}$. The metric value on each causal DAG is the average SHD, recall, precision and F1 scores over 10 bootstrapping datasets, $R^{\texttt{ref}}_{g,m} = \texttt{AVE}_{b=1:10}(R^{\texttt{ref}}_{g,m,b})$; $R^{\texttt{syn}}_{g,m} = \texttt{AVE}_{b=1:10}(R^{\texttt{syn}}_{g,m,b})$. The average and standard deviation of metric values reported in Table 1 are computed based on 10 benchmark datasets with 10 continuous variables for each causal mechanism. Additional experiment results with more continuous variables are in the appendix D.

Benchmark on the level of Markov equivalent class. To evaluate causal information on the Markov equivalent class level, we find all d-separations with minimal conditional sets between each pair of nodes in a causal graph and then apply conditional independence tests to the corresponding sets on synthetic datasets. We use the same procedures as the experiments on causal skeleton level to train baseline methods and generate synthetic datasets. Since the experimental results show minimal variation across different bootstrapped datasets, we omit bootstrapping at this level and use 15,000 samples for the tests.

Benchmark on the level of causal direction. To evaluate causal directions using only the data from a pair of variables, we select edges between nodes that become d-separated once the edge connecting them is removed. This ensures that other paths in the graph do not confound the pairwise causal relationship. We then apply bivariate causal discovery methods to these pairs to infer causal directionality. Each evaluation is conducted using 15,000 samples. In addition to bivariate methods, we also apply multivariate causal discovery algorithms to recover the full causal DAG structure and assess directionality at a more global level.

Benchmark on downstream tasks. Our evaluation is largely inspired by Chen et al. (2023). As mentioned in Section 4.2, same SCM models are trained on benchmark data and synthetic data together given the underlying causal DAGs. In the interventional inference task, for each causal graph, we take 10 interventions on each variable and approximate the estimated interventional distributions with 1,000 samples. In the counterfactual inference task, for each causal graph, we impose 10 interventions on each variable and generate 1,000 benchmark data as the new observations for computing counterfactual values. Next, we compare the results on interventional and counterfactual tasks subject to the same interventions. Considering the results of the models trained on benchmark data as the ground-truth, the metric is the average mean absolute errors (AMAE) over all variables,

$$\text{AMAE-}\texttt{syn} = \frac{1}{V}\sum_{v\in \boldsymbol{V}} \text{MAE-}\texttt{syn}(v); \text{MAE-}\texttt{syn}(v) = \frac{1}{(V-1)\times K\times N}\sum_{i\in \boldsymbol{V}\setminus v}\sum_{d\in \boldsymbol{S}_r}\left|\sum_{s=0}^{N-1} x^{\texttt{ref}}_{s,d,i} - \sum_{s=0}^{N-1} x^{\texttt{syn}}_{s,d,i}\right|,$$

where $x^{\texttt{ref}}_{s,d,i}$ is the reference ground-truth result and $x^{\texttt{syn}}_{s,d,i}$ is the result of the models trained on synthetic data; s denotes different samples; v denotes that the interventions are imposed on variable v chosen from the variable set $\boldsymbol{V}$ with V variables; d denotes the intervention value that is taken from a set $\boldsymbol{S}_r$ with sample size K; and i denotes the variable of which the interventional distribution or counterfactual value is computed for evaluation. We compute the mean and standard deviation of the average differences with 10 random seeds.

5.2 Benchmark results and discussion

We first evaluate the baseline methods on synthetic data, where ground-truth DAGs provide labels for high-order causal structure. We then assess performance on real-world data. Detailed ablation and additional results (including different sample sizes, more variables, discretized variables, and extended graphs) are presented in the Appendix D.

Table 1: Benchmark results under linear Gaussian (LG) and linear uniform (LU) causal mechanisms. The LU case is also visualized in Figure 1b. Complete results are detailed in Appendix D.

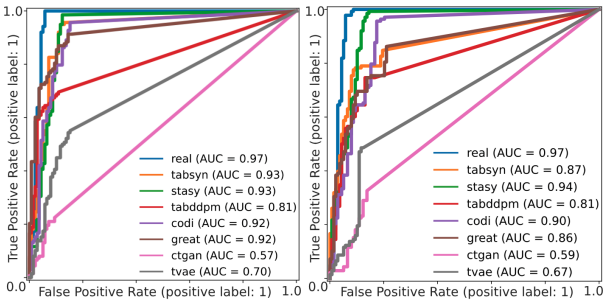
	Model	Low-order Col. ER (↓)	Low-order Pair. ER (↓)	Skeleton F1 (↑)	MEC AUC (↑)	Causal direction level SHD (↓)	Causal direction level ACC (↑)	Causal direction level F1 (↑)	Intervention AMAE (↓)	Counterfact AMAE (↓)
LG: linear Gaussian	ref.	0.00 ± 0.00	0.00 ± 0.00	0.90 ± 0.06	0.972	15.42 ± 7.01	0.500	0.38 ± 0.06	3.16 ± 0.2	0.04 ± 0.0
	TabSyn	2.11 ± 1.08	0.61 ± 0.21	0.71 ± 0.12	0.927	27.74 ± 5.14	0.500	0.26 ± 0.07	4.73 ± 1.8	0.56 ± 0.4
	STASY	12.42 ± 3.27	1.31 ± 0.81	0.68 ± 0.17	0.930	31.92 ± 4.55	**0.696**	0.21 ± 0.07	26.66 ± 7.1	0.65 ± 0.4
	TabDDPM	**0.69 ± 0.12**	0.62 ± 0.54	0.70 ± 0.12	0.814	26.64 ± 10.34	0.536	0.24 ± 0.09	4.13 ± 1.2	1.12 ± 1.3
	CoDi	4.42 ± 0.83	0.80 ± 0.43	0.72 ± 0.10	0.917	29.66 ± 5.12	0.589	0.24 ± 0.08	5.25 ± 1.6	0.67 ± 0.8
	GReaT	8.41 ± 0.73	0.76 ± 0.26	0.75 ± 0.04	0.921	18.18 ± 8.97	0.554	0.36 ± 0.06	9.77 ± 0.7	0.93 ± 0.5
	CTGAN	4.70 ± 0.78	3.76 ± 0.63	0.46 ± 0.06	0.566	36.00 ± 6.40	0.554	0.20 ± 0.06	15.02 ± 8.9	8.58 ± 7.7
	TVAE	4.51 ± 1.64	1.93 ± 0.64	0.58 ± 0.06	0.703	29.60 ± 8.47	0.536	0.20 ± 0.05	9.52 ± 5.2	5.22 ± 3.9
	TabPFN	1.63 ± 0.17	**0.30 ± 0.09**	**0.88 ± 0.07**	**0.970**	**14.66 ± 6.16**	0.375	**0.38 ± 0.07**	**3.35 ± 0.3**	**0.21 ± 0.1**
LU: linear uniform	ref.	0.00 ± 0.00	0.00 ± 0.00	0.89 ± 0.07	0.967	1.04 ± 0.85	1.000	0.94 ± 0.04	3.07 ± 0.2	0.03 ± 0.0
	TabSyn	1.88 ± 0.87	0.45 ± 0.28	0.71 ± 0.08	0.871	21.66 ± 5.75	0.946	0.41 ± 0.06	4.21 ± 1.1	0.45 ± 0.7
	STASY	11.90 ± 5.72	1.18 ± 0.59	0.67 ± 0.10	**0.936**	20.66 ± 4.99	0.946	0.45 ± 0.10	23.86 ± 11.1	0.44 ± 0.3
	TabDDPM	**0.98 ± 0.63**	1.07 ± 2.27	0.77 ± 0.07	0.815	15.86 ± 8.74	**1.000**	0.51 ± 0.14	**3.48 ± 0.6**	0.46 ± 0.7
	CoDi	8.01 ± 1.55	1.75 ± 1.52	0.74 ± 0.09	0.902	18.54 ± 5.08	0.911	0.45 ± 0.10	3.82 ± 0.6	0.58 ± 0.4
	GReaT	9.56 ± 0.73	0.56 ± 0.22	0.70 ± 0.05	0.860	13.62 ± 9.24	0.929	0.57 ± 0.11	11.20 ± 1.2	0.58 ± 0.4
	CTGAN	4.71 ± 0.50	3.52 ± 0.50	0.51 ± 0.08	0.588	33.64 ± 5.02	0.821	0.26 ± 0.07	10.89 ± 3.2	4.96 ± 3.3
	TVAE	6.33 ± 1.69	2.52 ± 1.17	0.55 ± 0.07	0.669	26.62 ± 8.15	0.839	0.29 ± 0.08	12.22 ± 3.4	5.40 ± 3.6
	TabPFN	2.04 ± 0.40	**0.30 ± 0.10**	**0.88 ± 0.06**	0.928	**2.44 ± 2.62**	0.964	**0.88 ± 0.11**	3.87 ± 0.5	**0.18 ± 0.1**

Notes: Low-order metrics (Zhang et al., 2024): column/pairwise error rates of density or correlation estimation. Skeleton: causal skeleton level. MEC: Markov equivalence class. Causal direction level: SHD (structural Hamming distance), bivariate accuracy (ACC), and multivariate F1 with LiNGAM. Shaded cells: causal discovery methods are theoretically inapplicable for LG case. Interventional task: AMAE = 100 × average mean absolute error.

Table 2: Benchmark results under LG and LU causal mechanisms, avoiding variable ordering bias.

	Model	Low-order Col. ER (↓)	Low-order Pair. ER (↓)	Skeleton F1 (↑)	MEC AUC (↑)	Causal direction level SHD (↓)	Causal direction level ACC (↑)	Causal direction level F1 (↑)	Intervention AMAE (↓)	Counterfact AMAE (↓)
LG	ref.	0.00 ± 0.00	0.00 ± 0.00	0.89 ± 0.05	0.966	15.80 ± 7.15	0.571	0.36 ± 0.05	3.19 ± 0.2	0.05 ± 0.0
	TabSyn	1.68 ± 0.55	1.40 ± 2.86	0.67 ± 0.17	0.840	23.40 ± 7.53	0.446	0.31 ± 0.08	5.02 ± 2.8	1.49 ± 2.6
	STASY	10.61 ± 5.95	1.85 ± 2.99	0.64 ± 0.17	0.844	29.95 ± 4.87	0.464	0.24 ± 0.08	24.66 ± 12.3	2.31 ± 3.4
	TabDDPM	**0.82 ± 0.15**	0.52 ± 0.38	0.72 ± 0.11	0.864	23.54 ± 7.92	0.464	0.28 ± 0.08	3.87 ± 1.2	**0.96 ± 1.4**
	CoDi	5.26 ± 2.41	1.77 ± 2.89	0.66 ± 0.17	0.819	30.23 ± 5.38	0.482	0.20 ± 0.06	9.14 ± 10.8	2.57 ± 4.4
	GReaT	8.74 ± 0.27	0.99 ± 0.39	0.69 ± 0.09	0.844	19.31 ± 7.95	0.518	0.31 ± 0.08	11.50 ± 2.3	1.49 ± 1.7
	CTGAN	5.67 ± 0.57	4.17 ± 0.91	0.48 ± 0.07	0.528	34.71 ± 6.34	0.643	0.20 ± 0.08	16.57 ± 5.2	9.54 ± 5.6
	TVAE	4.26 ± 1.70	2.37 ± 1.23	0.58 ± 0.06	0.658	27.87 ± 8.31	**0.661**	0.23 ± 0.09	9.10 ± 5.6	4.48 ± 3.0
	TabPFN	1.59 ± 0.17	**0.43 ± 0.34**	**0.82 ± 0.10**	**0.897**	**18.02 ± 9.96**	0.482	**0.32 ± 0.12**	**3.81 ± 1.7**	1.01 ± 2.4
LU	ref.	0.00 ± 0.00	0.00 ± 0.00	0.88 ± 0.04	0.972	1.54 ± 1.58	0.982	0.91 ± 0.09	2.96 ± 0.1	0.02 ± 0.0
	TabSyn	1.73 ± 0.92	0.45 ± 0.27	0.65 ± 0.09	0.749	27.58 ± 6.32	0.875	0.34 ± 0.09	4.19 ± 1.4	0.52 ± 0.8
	STASY	8.82 ± 3.25	1.17 ± 0.55	0.62 ± 0.15	0.857	23.81 ± 3.74	**0.964**	0.40 ± 0.11	17.43 ± 6.6	0.42 ± 0.6
	TabDDPM	**0.87 ± 0.14**	**0.37 ± 0.36**	0.69 ± 0.06	0.759	21.45 ± 8.84	0.911	0.39 ± 0.13	3.96 ± 1.4	0.96 ± 1.6
	CoDi	9.19 ± 2.52	1.65 ± 1.42	0.73 ± 0.10	0.839	21.57 ± 4.93	**0.964**	0.37 ± 0.11	**3.80 ± 0.6**	0.59 ± 0.5
	GReaT	9.41 ± 0.83	0.44 ± 0.11	0.69 ± 0.05	0.808	14.76 ± 9.20	0.839	0.54 ± 0.13	11.09 ± 0.8	**0.36 ± 0.2**
	CTGAN	5.75 ± 0.97	4.54 ± 2.29	0.50 ± 0.09	0.536	32.23 ± 6.08	0.768	0.27 ± 0.07	15.19 ± 7.9	7.92 ± 5.4
	TVAE	4.98 ± 1.52	2.35 ± 0.75	0.54 ± 0.08	0.547	26.62 ± 10.30	0.714	0.34 ± 0.11	10.70 ± 3.8	5.14 ± 3.2
	TabPFN	1.91 ± 0.16	0.48 ± 0.14	**0.81 ± 0.09**	**0.917**	**8.20 ± 6.35**	0.911	**0.68 ± 0.20**	4.13 ± 2.0	0.39 ± 0.6

Results on synthetic data. Table 1 shows a representative subset of results under the two data-generation regimes (LG and LU). Across low-order metrics (e.g., single-column density and pairwise correlations), Transformer-based (TabSyn and TabPFN), LLM-based (GReaT) and DFM-based (TabDDPM and CoDi) models often significantly outperform older approaches like TVAE and CTGAN. However, none of the experimented models fully match reference-level high-order structures as measured by SHD, d-separation tests, or causal directionality. ROC curves in Figure 3 visualize the performance differences of these methods on the d-separation conditional independence tests under LG and LU settings. Notably, TabSyn achieves the second lowest pairwise correlation error (LU setting) but does not perform as well w.r.t. the causal-structure scores, confirming that low-order metrics do not reliably capture multi-variate structural fidelity.



(a) ROC curve (LG) (b) ROC curve (LU)

Figure 3: Benchmark on d-separations: ROC curves of the conditional independence test results for Markov equivalent class level evaluation.

Table 3: Imputation results for the $P38$ feature in the Sachs dataset (Sachs et al., 2005), evaluated using low-order metrics from Zhang et al. (2024), mean absolute error (MAE) and mean squared error (MSE), across XGBoost (Chen & Guestrin, 2016), MLP, and TabPFN (Hollmann et al., 2025).

	Low-order		XGBoost		MLP		TabPFN (i)	
	Col. ER (↓)	Pair. ER (↓)	MAE (↓)	MSE(↓)	MAE (↓)	MSE(↓)	MAE (↓)	MSE(↓)
TabSyn	2.54±0.51	2.30±0.71	181.46±43.16	2.38e+05±1.24e+05	31.84±4.93	2.00e+04±2.09e+03	223.84±33.52	4.39e+05±1.51e+05
STASY	12.54±4.72	3.08±1.43	**106.84±18.03**	7.89e+04±3.96e+04	28.12±2.96	1.78e+04±1.02e+03	160.28±34.25	2.18e+05±1.15e+05
TabDDPM	**2.66±0.28**	**1.42±0.44**	108.56±11.35	**7.51e+04±1.82e+04**	**28.04±1.32**	**1.70e+04±1.01e+03**	151.84±14.24	1.88e+05±4.08e+04
CoDi	37.47±3.57	10.03±2.04	315.56±184.39	4.96e+05±4.26e+05	46.89±16.63	2.36e+04±1.18e+04	**131.71±82.44**	**9.34e+04±8.30e+04**
GReaT	16.07±0.23	3.27±0.49	275.58±43.68	4.67e+05±1.35e+05	43.54±7.21	3.75e+04±5.86e+03	217.61±22.92	3.39e+05±8.66e+04
CTGAN	14.57±1.45	8.54±0.64	335.08±32.11	7.35e+05±8.74e+04	45.83±7.13	4.41e+04±7.97e+03	363.49±29.36	8.71e+05±8.33e+04
TVAE	13.39±3.00	5.69±1.31	283.05±43.09	6.42e+05±1.46e+05	28.11±1.80	2.93e+04±9.70e+03	228.73±39.58	4.55e+05±1.03e+05
TabPFN (g)	4.52±0.21	8.17±0.73	261.23±79.17	8.88e+05±7.00e+05	72.51±9.86	6.23e+04±2.19e+04	181.81±33.51	3.94e+05±1.79e+05

Among the DFM-based methods, TabDDPM stands out for strong single-column density estimation and moderate high-order performance. GReaT sometimes recovers more joint structure (e.g., smaller SHD), but it can struggle on some conditional independence tasks. CoDi also consistently improves over classical generators but exhibits variability on certain causal direction metrics. Taken together, these results highlight that the best model for low-order statistics is not always the best for capturing high-order causal dependencies. TabPFN consistently performs well w.r.t. the high order metrics, which possibly could be attributed to the synthetic data it being pretrained on Hollmann et al. (2025) having similarities in terms of causal structures or generation mechanisms, with the benchmarking data.

To ensure the reported causal metrics are not driven by simple variable ordering biases in the data, we randomly permute the benchmark dataset columns before training and then restore their original order for evaluation. Table 2 presents these *reordered* results under LG and LU settings. The overall rankings of methods remain largely consistent, implying that state-of-the-art LLM-based or DFM-based approaches provide better, but still imperfect, recovery of high-order structures relative to older baselines. Nonetheless, a clear performance gap persists between the reference data and any synthetic generator on tasks involving multivariate causal metrics.

Results on real-world data. To examine how well the benchmarked models generalize beyond synthetic settings, we evaluate them on the well-known Sachs proteomic dataset (Sachs et al., 2005), a widely used benchmark in causal inference with an established ground-truth DAG. Following common practices in evaluating counterfactual inference and data imputation (Almond et al., 2005; Hill, 2011; Geffner et al., 2022), we design a missing-at-random scenario by withholding values of variable `P38`, which is known to be causally influenced by `PKC`. The baseline models are trained on the observed training partition and then used to generate synthetic datasets for imputing `P38`.

We evaluate the imputation quality using three regression models: XGBoost (Chen & Guestrin, 2016), a shallow MLP, and the tabular foundation model TabPFN (Hollmann et al., 2025) both for generation (g) and for imputation (i); they are trained solely on the synthetic data and tested on held-out real observations. As shown in Table 3, performance varies considerably. TabDDPM consistently ranks among the top performers, achieving the lowest MSE on both XGBoost and MLP regressors. STASY yields the best MAE with XGBoost, and CoDi provides the strongest TabPFN results. TabPFN (g) performs poorly in comparison to it's strong performance in the fully synthetic setting 1, indicating that high-order metrics alone are not sufficient to address generalisability to real-world settings, but rather the high- and low-order metrics should be used in conjunction when evaluating tabular synthesis models, especially when real-world data is concerned. Notably, these results correlate with models' downstream counterfactual inference scores under synthetic benchmarks (Appendix D, Table 12), indicating that high-order causal fidelity does translate into real-world utility for decision-support tasks.

Despite these observations, no model dominates across all metrics and learners, suggesting that current methods remain limited in consistently modeling the full structural complexity of tabular data. This highlights the importance of high-order benchmarking as a tool to guide improvements in causal-aware synthetic data generation.

6 Conclusion

We present a benchmark framework designed to systematically evaluate tabular synthesis models with respect to their capability of capturing high-order structural causal information. By introducing

causal graphs as explicit and meaningful prior knowledge, we categorize causal dependencies into three hierarchical levels, facilitating precise benchmark tasks and metrics. To overcome challenges associated with ground-truth availability, we develop synthetic benchmark datasets complemented by causal discovery techniques, enabling robust evaluation across multiple causal dimensions. Our framework effectively distinguishes the strengths and limitations of contemporary synthesis models, providing insights for model improvement. Notably, our evaluations reveal that current state-of-the-art methods, although excelling in capturing lower-order statistical properties, still exhibit significant gaps in modeling high-order causal structures. These results highlight the necessity for developing synthesis methods that inherently incorporate structural priors or enforce causal constraints during training. Ultimately, our benchmark advances methodological rigor in tabular data synthesis and highlights its relevance to high-stakes domains where causal fidelity is essential. Future work and limitations are discussed in Appendix B.

Ethics Statement

The proposed benchmark framework focuses exclusively on synthetic and publicly available datasets and does not involve human subjects, personal data, or sensitive information. All benchmark datasets used in this study were either generated via predefined causal graphs under controlled simulation processes or drawn from established open datasets that are licensed for research purposes (e.g., UCI Machine Learning Repository, DELVE repository). No private or proprietary data were used.

Potential negative societal impacts include the possibility of misinterpretation of benchmark results if applied without consideration of the assumptions underlying causal discovery methods. As discussed in our limitations (Appendix B), violations of these assumptions may affect transferability to real-world scenarios, and fairness-related concerns are beyond the current scope of this framework. We emphasize that our benchmark is intended as a research tool to advance the methodological rigor of tabular data synthesis, rather than as a direct application pipeline. We provide full access to all experimental code, data, and documentation (shared anonymously as supplemental material during the review phase and publicly released afterward) to foster transparency, accountability, and responsible research practices.

Reproducibility Statement

We have taken extensive measures to ensure the reproducibility of our results. The benchmark framework (including code, data generation procedures, and documentation) will be made publicly available after the double-blind review period. **For now, we provide the code together with an anonymized link to the dataset in the supplementary materials.** Detailed descriptions of benchmark dataset generation, causal mechanisms, and noise models are presented in Section 4.1 and Appendix C. Evaluation metrics for high-order structural causal information, conditional independence, and causal direction are formally defined in Section 4.2, with additional details in Appendix A.

Our experimental settings, including baseline implementations, training configurations, and evaluation procedures, are described in Section 5.1 and Appendix D. Results are reported with mean and standard deviation across multiple random seeds and bootstrapped datasets, reducing variance due to randomness. To mitigate variable-ordering bias, we conducted experiments with randomized column orders (Appendix D.3). We also include experiments on real-world datasets to assess external validity (Appendix D.4). Together, these measures ensure that independent researchers can reproduce and verify our findings with the provided resources.

References

Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. *International Conference on Machine Learning*, 2022.

Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

Steen A Andersson, David Madigan, and Michael D Perlman. A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. *International Conference on Artificial Intelligence and Statistics*, 2018.

Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A. Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *arXiv preprint arXiv:2206.06821*, 2022.

R. Bock. MAGIC Gamma Telescope. UCI Machine Learning Repository, 2007. DOI: https://doi.org/10.24432/C52C8B.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Andrea Bommert, Xudong Sun, Bernd Bischl, Jörg Rahnenführer, and Michel Lang. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, 2020.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020.

Asic Chen, Ruian Ian Shi, Xiang Gao, Ricardo Baptista, and Rahul G Krishnan. Structured neural networks for density estimation and causal inference. *Advances in Neural Information Processing Systems*, 2023.

Song Chen. Beijing PM2.5. UCI Machine Learning Repository, 2017. DOI: https://doi.org/10.24432/C5JS49.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016. URL `http://arxiv.org/abs/1603.02754`.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

Valeriia Cherepanova, Roman Levin, Gowthami Somepalli, Jonas Geiping, C Bayan Bruss, Andrew G Wilson, Tom Goldstein, and Micah Goldblum. A performance-driven benchmark for feature selection in tabular deep learning. *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, 2023.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 2002.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. *Machine Learning for Healthcare Conference*, 2017.

Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. Learning the graphical structure of electronic health records with graph convolutional transformer. In *Proceedings of the AAAI conference on artificial intelligence*, 2020.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. Large language models on tabular data–a survey. *arXiv preprint arXiv:2402.17944*, 2024.

José AR Fonollosa. Conditional distribution variability measures for causality detection. *Cause Effect Pairs in Machine Learning*, pp. 339–347, 2019.

Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with tableshift. *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, 2023.

Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, et al. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022.

Clark Glymour and Kun Zhang. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:418407, 2019.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 2022.

Manbir Gulati and Paul Roysdon. Tabmt: Generating tabular data with masked transformers. *Advances in Neural Information Processing Systems*, 2023.

Lasse Hansen, Nabeel Seedat, Mihaela van der Schaar, and Andrija Petrovic. Reimagining synthetic tabular data generation through data-centric ai: A comprehensive benchmark. *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, 2023.

Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 2022.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. *International Conference on Learning Representations*, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6. URL https://doi.org/10.1038/s41586-024-08328-6.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 2008.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018.

Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 2012.

Adrián Javaloy, Pablo Sánchez-Martín, and Isabel Valera. Causal normalizing flows: from theory to practice. *Advances in Neural Information Processing Systems*, 36, 2024.

Sérgio Jesus, José Pombal, Duarte Alves, André Cruz, Pedro Saleiro, Rita Ribeiro, João Gama, and Pedro Bizarro. Turning the tables: Biased, imbalanced, dynamic tabular datasets for ml evaluation. *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, 2022.

Alexia Jolicoeur-Martineau, Kilian Fatras, and Tal Kachman. Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees. *International Conference on Artificial Intelligence and Statistics*, 2024.

James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. *International conference on learning representations*, 2018.

Jayoung Kim, Chaejeong Lee, and Noseong Park. STaSy: Score-based tabular data synthesis. *International Conference on Learning Representations*, 2023.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDMP: Modelling tabular data with diffusion models. *International Conference on Machine Learning*, 2023.

Chaejeong Lee, Jayoung Kim, and Noseong Park. CoDi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis. *International Conference on Machine Learning*, 2023.

Cheng-Te Li, Yu-Che Tsai, and Jay Chiehen Liao. Graph neural networks for tabular data learning. *International Conference on Data Engineering (ICDE)*, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. *International Conference on Learning Representations*, 2023.

Tennison Liu, Zhaozhi Qian, Jeroen Berrevoets, and Mihaela van der Schaar. GOGGLE: Generative modelling for tabular data by learning relational structure. *International Conference on Learning Representations*, 2023.

Chao Ma, Sebastian Tschiatschek, Richard Turner, José Miguel Hernández-Lobato, and Cheng Zhang. VAEM: a deep generative model for heterogeneous mixed type data. *Advances in Neural Information Processing Systems*, 2020.

Andrey Malinin, Neil Band, Yarin Gal, Mark Gales, Alexander Ganshin, German Chesnokov, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 1995.

Pablo Morales-Alvarez, Wenbo Gong, Angus Lamb, Simon Woodhead, Simon Peyton Jones, Nick Pawlowski, Miltiadis Allamanis, and Cheng Zhang. Simultaneous missing value imputation and structure learning with groups. *Advances in Neural Information Processing Systems*, 2022.

Ignavier Ng, Sébastien Lachapelle, Nan Rosemary Ke, Simon Lacoste-Julien, and Kun Zhang. On the convergence of continuous constrained optimization for structure learning. *International Conference on Artificial Intelligence and Statistics*, 2022.

Antoine Passemiers, Pietro Folco, Daniele Raimondi, Giovanni Birolo, Yves Moreau, and Piero Fariselli. How good neural networks interpretation methods really are? a quantitative benchmark. *arXiv preprint arXiv:2304.02383*, 2023.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Zhaozhi Qian, Rob Davis, and Mihaela van der Schaar. Synthcity: a benchmark framework for diverse use cases of tabular synthetic data. *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, April 2005.

Meyer Scetbon, Joel Jennings, Agrin Hilmkil, Cheng Zhang, and Chao Ma. FiP: a fixed-point approach for causal generative modeling. *arXiv preprint arXiv:2404.06969*, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. In *Proceedings of the IEEE*, 2021.

Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 2011.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 2019.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Luis Torgo. house_16H. DELVE repository of data, 2014. https://www.openml.org/search?type=data&id=821&sort=runs&status=active.

Athanasios Tsanas and Max Little. Parkinsons Telemonitoring. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C5ZS3N.

Boris van Breugel and Mihaela van der Schaar. Beyond privacy: Navigating the opportunities and challenges of synthetic data. *arXiv preprint arXiv:2304.03722*, 2023.

Boris van Breugel and Mihaela van der Schaar. Why tabular foundation models should be a research priority. *International Conference on Machine Learning*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Zifeng Wang and Jimeng Sun. TransTab: Learning transferable tabular transformers across tables. *Advances in Neural Information Processing Systems*, 2022.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 2019.

Jiahuan Yan, Jintai Chen, Yixuan Wu, Danny Z Chen, and Jian Wu. T2g-former: organizing tabular features into relation graphs promotes heterogeneous feature interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Hengrui Zhang, Jiani Zhang, Balasubramaniam Srinivasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. *International Conference on Learning Representations*, 2024.

K Zhang and A Hyvärinen. On the identifiability of the post-nonlinear causal model. *Conference on Uncertainty in Artificial Intelligence*, 2009.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 2018.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 2024.

Bingzhao Zhu, Xingjian Shi, Nick Erickson, Mu Li, George Karypis, and Mahsa Shoaran. XTab: Cross-table pretraining for tabular transformers. *International Conference on Machine Learning*, 2023a.

Max Zhu, Katarzyna Kobalczyk, Andrija Petrovic, Mladen Nikolic, Mihaela van der Schaar, Boris Delibasic, and Petro Lio. Tabular few-shot generalization across heterogeneous feature spaces. *arXiv preprint arXiv:2311.10051*, 2023b.

Appendix of Causality for Tabular Data Synthesis: A High-Order Structure Causal Benchmark Framework

CONTENTS

A BRIEF INTRODUCTION OF CAUSAL DISCOVERY

Causal discovery aims to determine causal relationships purely based on observational data by leveraging their statistical properties under proper assumptions (Spirtes et al., 2000; Peters et al., 2017). The methodology of causal discovery can be characterized into constraint-based, score-based, and Functional Causal Model (FCM)-based methods (Glymour & Zhang, 2019). *Constraint-based methods*, such as PC and FCI algorithms, apply conditional independence tests to each pair of variables and infer causal skeletons and causal directions based on certain rules (Spirtes et al., 2000). *Score-based methods* Chickering (2002); Huang et al. (2018) formulate causal discovery as an optimization problem and optimize score functions by searching in the space of DAGs. Many deep learning-based methods (Zheng et al., 2018; Ng et al., 2022) can be considered as score-based ones. Under Markov and causal faithfulness assumptions, constraint-based and score-based methods identify causal graphs up to certain equivalence classes. For example, suppose that there are no unknown confounders, PC algorithm identifies the Markov equivalence classes of causal DAGs as CPDAGs, and the results of score-based method GES (Chickering, 2002) also converge to Markov equivalence classes. To further identify causal relationships in the same equivalence class, *FCM-based methods* impose additional assumptions of functional classes and distributions on the data generation processes. The most flexible and identifiable FCM is proposed by Hoyer et al. (2008). Common FCM-based methods (Shimizu et al., 2006; Hoyer et al., 2008) are based on the bivariate case determining the cause and the effect between two variables. For example, one can fit two FCMs in different directions and select the one with large likelihood on the observational data. There are also multivariate FCM-based methods, such as LiNGAM (Shimizu et al., 2006) that assumes that the FCM is a linear non-Gaussian model.

Assumptions of causal discovery methods. Throughout the paper, we frequently mention proper assumptions without explicitly stating them. This is because assumptions for different causal discovery methods are different and some of them involve specific and technical details such that consistently stating them would make the paper unnecessarily complicated and more difficult to follow. For the sake of brevity and clarity, we use "under proper assumptions" in general. The involved assumptions on the causal skeleton level:

- The identifiability conditions of PC algorithm (Spirtes et al., 2000; Peters et al., 2017).

The involved assumptions on the causal direction level:

- The identifiability conditions of additive noise models (Hoyer et al., 2008);
- The identifiability conditions of post-nonlinear models (Zhang & Hyvärinen, 2009);
- The identifiability conditions of LiNGAM (Shimizu et al., 2006).

Potential negative societal impacts. Since real-world scenarios are always violating the proper assumptions in different ways, the benchmarking results need to be carefully interpreted together with the causal discovery assumptions. Considering the potential violation of the assumptions for specific applications is necessary for a proper usage of the benchmark framework. Although we conduct experiments on real-world data, as shown in Tables 3 and 20 that show an indication of the transferability of the framework to real-world settings to some extent, we advice against concluding that this transferability always holds, as our experiments on this are not exhaustive and merely a first attempt at bridging this gap. While the proposed framework accounts for certain biases (e.g., variable ordering bias) in evaluating baseline models, broader concerns such as fairness are beyond the current scope and represent valuable directions for future work. We suggest using separate evaluation and mitigation techniques to compliment the proposed framework in order to address this, prior to any application of the baseline models.

Best Practices in using causal discovery methods for benchmarking on high-order causal structural information. Benchmark datasets with linear relationships and simple distributions (e.g., Gaussian and uniform distributions) are good enough for distinguishing models in terms of capturing causal information. Moreover, the conditional independence tests in constraint-based causal discovery methods can be efficiently applied to datasets with linear relationships; whereas, the kernel-based tests for the datasets with nonlinear relationships are only feasible when the sample size is around 1000. As for the application of bivariate causal discovery methods, we find that the methods without requiring training models are more efficient for the purpose of evaluation and can provide a reasonable evaluation results. And in general, their results do not vary a lot, and do not require bootstrapping for an error bar.

B Limitations and Future Works

This work aims at establishing the foundation of benchmarking tabular synthesis models on causal information and future works can consider to further address some limitations. We conducted initial experimentation with respect to discrete data in Table 5, but we suggest to expand the benchmark framework with more generation and discretisation processes in addition to mixed data types while satisfying Conditions (i) and (ii) while including more baseline methods. Current benchmark datasets are based mainly on causal mechanisms LG, LU, SG, and NG, and our experimental setup uses $N = 10 or 20$ variables which could be easily expanded as N is a configurable parameter in our benchmark framework. Although for the purpose of this paper, it is sufficient with our current setting, it may not be suitable for specific applications or other use cases. Additionally, since this work is based on a general consideration of causal relationships without a specific downstream application, we only choose general benchmark tasks and metrics based on causal discovery algorithms. It is worthwhile to design more task-specific benchmark tasks and metrics when there are specific downstream applications of synthetic tabular data. Moreover, the current experiment setting is too ideal compared with real-world scenarios. More factors can be considered so that benchmark datasets are close to real-world cases. For example, we currently assume that all the variables of a causal graph are known; however, it is not always the case in real-world scenario. And future works can consider to include unknown confounders. Last but not least, our evaluation is limited by the assumptions of causal discovery algorithms, because causal information used for evaluation is not the direct outputs of tabular synthesis models but is extracted by causal discovery methods. For example, we use causal DAGs instead of causal graphs because common causal discovery methods rely on the DAG assumption and the purpose of the work is to evaluate synthesis models, for which the validity is more important than the flexibility of structural causal models. And the mixed data type of continuous and discrete data is not included in the benchmark because there is still lack of studies in causal discovery on such data. This also motivates the research studies in causal discovery domain, which have potentials for modelling tabular data and evaluating synthesis models.

Furthermore, as demonstrated in (Reisach et al., 2021), existing graph simulation methodologies, including those applied in this study, exhibit notable limitations. In particular, the widely used Erdős-Rényi model for generating random DAGs does not produce a uniform distribution over the space of possible graph structures. This highlights the importance of further examining the transferability of the framework to real-world scenarios and supports the investigation of alternative graph simulation strategies that address the limitations identified in (Reisach et al., 2021).

C Data Generation Processes of Benchmark Data

The causal mechanisms for linear, sigmoid, and neural network-based functional relationships are

$$X_i = W_i \cdot X_i^{\texttt{prt}} + E_{X_i}; \tag{2}$$

$$X_i = W_i \cdot \sigma(X_i^{\texttt{prt}}) + E_{X_i}; \tag{3}$$

$$X_i = W_{1i} \cdot \sigma(W_{2i} \cdot (X_i^{\texttt{prt}} \oplus E_{X_i})); \tag{4}$$

where $\oplus$ is concatenation, and W_i, W_{1i}, and W_{2i} are weight matrices.

To generate a discrete value with K categories of $X_i^{\texttt{disc}\cdot}$, we first generate a continuous value, then compute the probability of each category as Equation equation 5, and sample from the categorical distribution as Equation equation 6,

$$\texttt{prob}_k \coloneqq \texttt{softmax}(\sigma(W_{i,k} \cdot X_i)); \tag{5}$$

$$X_i^{\texttt{disc}\cdot} \sim \mathcal{C}(\texttt{prob}_1, ..., \texttt{prob}_K), \tag{6}$$

where X_i is a continuous-valued variable, $X_i^{\texttt{disc}\cdot}$ is a discrete-valued variable, σ is the sigmoid function, $W_{i,k}$ is a random weight for each category of $X_i^{\texttt{disc}\cdot}$, $\mathcal{C}$ denotes a categorical distribution with parameters $\texttt{prob}_k$ for $k = 1, ..., K$. The results of applying this procedure for the linear gaussian setting are available in Table 5.

D More Experiments

D.1 Hardware, datasets, software, and implementation

Hardware. We used one NVIDIA RTX 2080 Ti for the benchmarking results for the non-reordered benchmarking using 10 variables. A Google Cloud g2-standard-32 virtual machine with an NVIDIA-L4 accelerator was used for the discretized, additional variables and Sachs benchmarking.

Implementation of the benchmark framework. Our benchmark framework is available at URL `https://github.com/TURuibo/CauTabBench`.

Baseline methods. Baseline methods are implemented based on the repository `https://github.com/amazon-science/tabsyn` of (Zhang et al., 2024). As shown in Table 4, we evaluate the synthetic datasets with the metrics in (Zhang et al., 2024) as the reproduced results for a sanity check and a reference for other works. We used the training configurations provided for the "Magic" dataset (Bock, 2007) to train baseline methods.

Benchmark dataset generation. We modify `CausalDiscoveryToolbox` for generating benchmark datasets with randomly generated causal DAGs. For demonstration, we used the configuration, variable types, number of variables, and sample size of a real-world dataset (Bock, 2007), which is also used for our evaluation of real-world datasets. 10 causal DAGs are randomly generated and each has 10 nodes representing continuous variables and 1 node representing a binary variable. The binary variable in (Bock, 2007) is the classification target variable. For each causal DAG, we generate benchmark datasets of which the sample size is 17,117. We find that there is a lack of implementation for causal discovery methods in the presence of mixed data types; hence, we generate the binary variable independent of all other variables. In this way, we train the baseline methods on 11 variables and evaluate on 10 continuous variables by dropping the binary one, reducing the binary variable's influence on the evaluation. We used random seeds from 100 to 109 to generate the 10 causal graphs and their corresponding benchmark datasets.

Table 4: Benchmark on low-order statistics. Values are mean and standard deviation of metric values (error rate (%) of single column density, error rate (%) of pair-wise correlation score, $\alpha-$precision, $\beta-$recall) over 10 random causal DAGs.

(a) Linear Gaussian

	Col.	Pair.	α-precision	β-recall
TabSyn	2.11 ± 1.08	0.61 ± 0.21	98.41 ± 1.39	49.34 ± 9.09
STASY	12.42 ± 3.27	1.31 ± 0.81	93.93 ± 3.92	43.67 ± 5.49
TabDDPM	0.69 ± 0.12	0.62 ± 0.54	99.34 ± 0.14	50.00 ± 0.40
CoDi	4.42 ± 0.83	0.80 ± 0.43	84.60 ± 3.13	58.50 ± 2.02
GReaT	8.41 ± 0.73	0.76 ± 0.26	81.55 ± 6.78	51.49 ± 1.18
CTGAN	4.70 ± 0.78	3.76 ± 0.63	88.29 ± 4.09	13.23 ± 9.09
TVAE	4.51 ± 1.64	1.93 ± 0.64	87.17 ± 9.68	30.77 ± 10.16
TabPFN	1.63 ± 0.17	0.30 ± 0.09	93.94 ± 1.55	51.97 ± 0.24

(b) Linear uniform

	Col.	Pair.	α-precision	β-recall
TabSyn	1.88 ± 0.87	0.45 ± 0.28	98.40 ± 1.01	49.81 ± 0.95
STASY	11.90 ± 5.72	1.18 ± 0.59	93.71 ± 5.70	47.37 ± 2.62
TabDDPM	0.98 ± 0.63	1.07 ± 2.27	99.24 ± 0.35	41.06 ± 18.73
CoDi	8.01 ± 1.55	1.75 ± 1.52	66.48 ± 2.37	59.84 ± 2.56
GReaT	9.56 ± 0.73	0.56 ± 0.22	96.22 ± 0.84	46.17 ± 1.09
CTGAN	4.71 ± 0.50	3.52 ± 0.50	92.35 ± 2.37	8.06 ± 8.69
TVAE	6.33 ± 1.69	2.52 ± 1.17	92.78 ± 3.83	14.99 ± 10.87
TabPFN	2.04 ± 0.40	0.30 ± 0.10	98.25 ± 0.74	50.10 ± 0.43

(c) Sigmoid Gaussian

	Col.	Pair.	α-precision	β-recall
TabSyn	1.83 ± 0.84	0.49 ± 0.20	98.64 ± 1.05	49.37 ± 0.64
STASY	12.48 ± 5.13	1.52 ± 0.43	92.77 ± 6.81	44.83 ± 3.78
TabDDPM	0.85 ± 0.14	0.40 ± 0.30	99.18 ± 0.40	49.83 ± 0.69
CoDi	5.96 ± 2.53	1.14 ± 0.29	92.86 ± 4.46	61.66 ± 2.28
GReaT	8.51 ± 2.03	1.45 ± 0.72	86.12 ± 5.82	50.65 ± 2.58
CTGAN	4.75 ± 0.51	3.16 ± 0.35	89.44 ± 4.16	23.01 ± 6.96
TVAE	4.86 ± 0.93	1.46 ± 0.50	89.70 ± 5.98	41.38 ± 6.49
TabPFN	1.61 ± 0.12	0.37 ± 0.10	94.16 ± 1.05	52.14 ± 0.64

(d) Neural network Gaussian

	Col.	Pair.	α-precision	β-recall
TabSyn	1.91 ± 0.62	0.57 ± 0.23	98.41 ± 1.19	48.93 ± 0.75
STASY	10.25 ± 2.80	1.74 ± 0.87	92.57 ± 6.11	47.55 ± 2.99
TabDDPM	0.75 ± 0.11	0.30 ± 0.11	99.16 ± 0.32	50.27 ± 0.87
CoDi	6.72 ± 2.73	1.16 ± 0.59	86.91 ± 6.90	58.71 ± 3.76
GReaT	7.36 ± 1.35	1.34 ± 0.44	87.40 ± 5.86	48.15 ± 4.07
CTGAN	4.79 ± 0.59	3.92 ± 1.86	89.49 ± 3.95	11.13 ± 10.92
TVAE	5.27 ± 1.61	1.78 ± 0.70	90.38 ± 6.74	22.04 ± 15.71
TabPFN	1.60 ± 0.11	0.46 ± 0.15	94.97 ± 1.40	51.81 ± 0.57

Causal discovery and inference methods for evaluation. `causal-learn`(Zheng et al., 2024) is used to evaluate the causal skeleton level and Markov equivalent class level. The conditional independence test is Fisher's-Z tests for datasets with linear relationships and is kernel conditional independence tests for datasets with nonlinear relationships. `CausalDiscoveryToolbox` is used for the causal direction level. And `DoWhy` (Sharma & Kiciman, 2020; Blöbaum et al., 2022) is used for the evaluation on the causal inference downstream tasks with additive noise models.

20 node and discrete and discrete experiments To investigate how the size of the DAG impacts the framework, we extend the experiments to 20 nodes (see Table 6). Furthermore, to investigate whether the framework generalizes and can be employed for discrete data, we apply the proposed discretization procedure (see Appendix C) to generated datasets prior to training the baseline models and evaluating them using the framework (Table 5).

Real-world datasets used for the evaluation. We used 4 real-world datasets for the evaluation, which are suitable for our evaluation on the high-order structural causal information. Because they are based on linear relationships and continuous variables with a causal semantic context. They are Beijing (Chen, 2017), Magic (Bock, 2007), House (Torgo, 2014), and Parkinsons (Tsanas & Little, 2009). These datasets are licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

D.2 More details and benchmarking results

The results in Table 4 show that CoDi stands out for its robust β-recall. However, this model tends to lag in α-precision when compared to state-of-the-art results. On the other hand, TabDDPM and Tabsyn achieve the lowest single column density estimation and pair-wise correlation errors. This suggests their ability to handle also the joint probability distributions between columns. Furthermore, STASY shows limitations in terms of single column density estimation task, having the highest error rates across all datasets and causal mechanisms.

As for benchmarking on the causal skeleton and Markov equivalent class level, we use Fisher's Z-test and the kernel independence test for datasets with linear and nonlinear functional relationships, respectively. Because it is not feasible to apply kernel independence tests to datasets with large sample sizes. Our experiments only include the results with small sample sizes, e.g., 1500. Since such conditional independence tests are more reliable with larger sample sizes, the performance of baseline methods on the nonlinear datasets is less distinguishable and not as informative as the results on the linear datasets as shown in Table 8 and Table 9. As for the causal direction level, we apply 3 bivariate causal discovery methods: RECI (Blöbaum et al., 2018), IGCI (Janzing et al., 2012), and CDS (Fonollosa, 2019) and use the best result of them as the final metric value in Table 10.

Table 5: Discretized data (see Appendix C) Results. 5a Values are mean and standard deviation of metric values (error rate (%) of single column density, error rate (%) of pair-wise correlation score, $\alpha-$precision, $\beta-$recall) over 10 random causal DAGs. 5b Benchmark on causal skeletons. Values are mean and standard deviation of metric values (SHD, F1 score, recall, and precision) over 10 random causal DAGs. Each metric value on a causal graph is the average value over 5 bootstrapping datasets. 5c Causal direction level and Interventional downstream tasks.

(a) Low-Order

	Col.	Pair.	α-precision	β-recall
TabSyn	2.14 ± 0.66	4.94 ± 0.78	99.00 ± 0.57	70.60 ± 0.65
STASY	24.29 ± 5.34	34.83 ± 7.33	88.49 ± 5.35	55.99 ± 4.80
TabDDPM	1.05 ± 0.09	4.62 ± 0.05	99.33 ± 0.27	68.82 ± 0.45
CoDi	23.14 ± 1.47	37.60 ± 1.57	54.33 ± 8.70	47.66 ± 1.35
GReaT	1.17 ± 0.10	4.64 ± 0.08	98.47 ± 0.46	69.22 ± 0.44
CTGAN	22.06 ± 0.83	35.04 ± 0.89	92.91 ± 3.97	50.85 ± 1.35
TVAE	83.96 ± 16.89	95.38 ± 10.63	18.02 ± 7.09	1.61 ± 4.68
TabPFN	18.11 ± 0.61	25.76 ± 0.90	90.22 ± 5.22	60.30 ± 1.05

(b) Skel

	Adj	F1	Precision	Recall
ref.	26.30 ± 5.99	0.23 ± 0.05	0.15 ± 0.06	0.21 ± 0.09
TabSyn	27.46 ± 4.82	0.22 ± 0.06	0.17 ± 0.06	0.20 ± 0.10
STASY	35.30 ± 4.08	0.23 ± 0.04	0.30 ± 0.05	0.19 ± 0.06
TabDDPM	26.46 ± 3.41	0.21 ± 0.03	0.14 ± 0.06	0.19 ± 0.11
CoDi	30.96 ± 5.15	0.26 ± 0.10	0.30 ± 0.11	0.23 ± 0.12
GReaT	25.96 ± 4.76	0.21 ± 0.02	0.16 ± 0.05	0.21 ± 0.08
CTGAN	53.82 ± 6.68	0.30 ± 0.06	0.67 ± 0.16	0.20 ± 0.05
TVAE	69.04 ± 9.87	0.33 ± 0.09	0.96 ± 0.13	0.20 ± 0.06
TabPFN	28.94 ± 4.29	0.22 ± 0.07	0.21 ± 0.08	0.20 ± 0.10

(c) High-Order

	MEC AUC	RECI	IGCI	CDS	Intv
ref.	0.500	0.518	0.482	0.464	9.50 ± 0.3
TabSyn	0.500	0.607	0.393	0.589	10.37 ± 0.6
STASY	0.500	0.643	0.357	0.679	55.01 ± 15.3
TabDDPM	0.501	0.500	0.500	0.429	9.77 ± 0.3
CoDi	0.505	0.482	0.429	0.429	58.12 ± 17.4
GReaT	0.499	0.482	0.536	0.589	9.80 ± 0.3
CTGAN	0.503	0.589	0.393	0.482	47.72 ± 9.7
TVAE	N/A^*	0.036	0.089	0.071	204.75 ± 61.8
TabPFN	0.499	0.446	0.536	0.554	38.44 ± 4.1

*Note: TVAE suffers from severe mode-collapse for the discrete data, collapsing to a single repeated row across all samples, causing MEC AUC to be incalculable using the framework and largely explains the poor performance for the other metrics.

Table 6: 20 nodes/variables Results 6a Mean and standard deviation of metric values (error rate (%) of single column density, error rate (%) of pair-wise correlation score, $\alpha-$precision, $\beta-$recall) over 10 random causal DAGs. 6b Benchmark on causal skeletons. Values are mean and standard deviation of metric values (SHD, F1 score, recall, and precision) over 10 random causal DAGs. Each metric value on a causal graph is the average value over 5 bootstrapping datasets. 6c Causal direction level and Interventional downstream tasks.

(a) Low-Order

	Col.	Pair.	α-precision	β-recall
TabSyn	2.10 ± 1.08	0.58 ± 0.28	98.45 ± 1.09	49.54 ± 0.54
STASY	8.77 ± 2.24	1.14 ± 0.19	95.89 ± 2.13	51.64 ± 3.51
TabDDPM	0.99 ± 0.21	1.77 ± 1.13	98.13 ± 0.53	49.84 ± 0.88
CoDi	5.78 ± 0.73	1.06 ± 0.25	72.12 ± 4.48	70.37 ± 1.83
GReaT	8.60 ± 0.47	1.29 ± 0.99	75.89 ± 5.88	58.70 ± 2.17
CTGAN	5.42 ± 0.84	3.55 ± 0.52	81.14 ± 4.65	10.49 ± 6.56
TVAE	4.29 ± 0.72	2.23 ± 0.77	88.65 ± 4.28	24.65 ± 10.89
TabPFN	1.67 ± 0.14	0.43 ± 0.10	91.78 ± 1.57	53.15 ± 0.73

(b) Skel

	Adj	F1	Precision	Recall
ref.	13.18 ± 3.90	0.83 ± 0.04	0.86 ± 0.09	0.82 ± 0.07
TabSyn	28.14 ± 6.67	0.68 ± 0.06	0.81 ± 0.10	0.60 ± 0.09
STASY	43.72 ± 8.64	0.58 ± 0.07	0.80 ± 0.10	0.46 ± 0.09
TabDDPM	56.14 ± 15.46	0.51 ± 0.09	0.78 ± 0.15	0.39 ± 0.07
CoDi	48.24 ± 14.64	0.55 ± 0.10	0.77 ± 0.09	0.44 ± 0.12
GReaT	31.64 ± 7.45	0.64 ± 0.06	0.77 ± 0.09	0.56 ± 0.06
CTGAN	102.64 ± 13.57	0.37 ± 0.05	0.83 ± 0.11	0.24 ± 0.03
TVAE	74.26 ± 15.61	0.46 ± 0.06	0.85 ± 0.11	0.32 ± 0.04
TabPFN	14.47 ± 4.89	0.81 ± 0.05	0.87 ± 0.07	0.77 ± 0.07

(c) High-Order

	MEC AUC	RECI	IGCI	CDS	Intv
ref.	0.980	0.454	0.378	0.487	3.30 ± 0.0
TabSyn	0.915	0.597	0.487	0.496	4.66 ± 1.7
STASY	0.843	0.521	0.546	0.496	19.51 ± 5.5
TabDDPM	0.720	0.513	0.597	0.496	5.33 ± 2.6
CoDi	0.823	0.563	0.479	0.538	7.01 ± 2.2
GReaT	0.812	0.395	0.597	0.588	11.48 ± 0.9
CTGAN	0.550	0.630	0.471	0.462	13.46 ± 1.8
TVAE	0.599	0.605	0.487	0.546	9.04 ± 2.8
TabPFN	0.963	0.455	0.505	0.525	3.46 ± 0.1

D.3 More details and benchmarking results avoiding variables ordering bias

Tables 13, 14, 16, 15, 17, and 18 present a more detailed view of benchmarking results avoiding ordering bias. To ensure unbiased modeling of causal information evaluation and avoid information leakage from data ordering as per the DAG, we randomly shuffle the variables of the benchmark data and train the synthetic tabular data model. This is done per seed such that the order of the benchmark dataset's columns is randomized, this ensures all synthesis models have consistent training data with a different order for every seed evaluted. This prevents the synthesis models from exploiting

Table 7: Imputation of the $P38$ feature for the Sachs dataset Sachs et al. (2005). Metrics are Low-order metrics from Zhang et al. (2024), r^2 coefficient of determination, mean absolute error (MAE) and mean squared error (MSE), the models evaluated are XGBoost Chen & Guestrin (2016), a simple MLP and TabPFN Hollmann et al. (2025).

(a) Low-Order

	Col.	Pair.	α-precision	β-recall
TabSyn	2.54 ± 0.51	2.30 ± 0.71	98.17 ± 0.96	48.54 ± 1.06
STASY	12.54 ± 4.72	3.08 ± 1.43	91.78 ± 2.97	48.25 ± 7.58
TabDDPM	2.66 ± 0.28	1.42 ± 0.44	93.26 ± 0.92	77.48 ± 1.90
CoDi	37.47 ± 3.57	10.03 ± 2.04	57.39 ± 4.89	32.21 ± 4.69
GReaT	16.07 ± 0.23	3.27 ± 0.49	77.10 ± 0.46	43.66 ± 0.59
CTGAN	14.57 ± 1.45	8.54 ± 0.64	91.55 ± 4.08	23.33 ± 1.80
TVAE	13.39 ± 3.00	5.69 ± 1.31	85.22 ± 3.22	36.46 ± 1.93
TabPFN	4.52 ± 0.21	8.17 ± 0.73	94.66 ± 0.89	50.46 ± 0.48

(b) XGBoost

	r^2	MAE	MSE
TabSyn	-0.02 ± 1.81	181.46 ± 43.16	$2.38e+05 \pm 1.24e+05$
STASY	0.87 ± 0.08	106.84 ± 18.03	$7.89e+04 \pm 3.96e+04$
TabDDPM	0.88 ± 0.03	108.56 ± 11.35	$7.51e+04 \pm 1.82e+04$
CoDi	0.57 ± 0.31	315.56 ± 184.39	$4.96e+05 \pm 4.26e+05$
GReaT	0.23 ± 0.24	275.58 ± 43.68	$4.67e+05 \pm 1.35e+05$
CTGAN	-22.57 ± 27.91	335.08 ± 32.11	$7.35e+05 \pm 8.74e+04$
TVAE	-10.98 ± 12.03	283.05 ± 43.09	$6.42e+05 \pm 1.46e+05$
TabPFN	-5.05 ± 5.29	261.23 ± 79.17	$8.88e+05 \pm 7.00e+05$

(c) MLP

	r^2	MAE	MSE
TabSyn	0.14 ± 0.56	31.84 ± 4.93	$2.00e+04 \pm 2.09e+03$
STASY	0.47 ± 0.08	28.12 ± 2.96	$1.78e+04 \pm 1.02e+03$
TabDDPM	0.47 ± 0.06	28.04 ± 1.32	$1.70e+04 \pm 1.01e+03$
CoDi	0.45 ± 0.08	46.89 ± 16.63	$2.36e+04 \pm 1.18e+04$
GReaT	-1.13 ± 0.70	43.54 ± 7.21	$3.75e+04 \pm 5.86e+03$
CTGAN	-15.53 ± 8.47	45.83 ± 7.13	$4.41e+04 \pm 7.97e+03$
TVAE	-4.29 ± 8.27	28.11 ± 1.80	$2.93e+04 \pm 9.70e+03$
TabPFN	-2.38 ± 1.19	72.51 ± 9.86	$6.23e+04 \pm 2.19e+04$

(d) TabPFN

	r^2	MAE	MSE
TabSyn	-4.73 ± 7.33	223.84 ± 33.52	$4.39e+05 \pm 1.51e+05$
STASY	-0.07 ± 1.71	160.28 ± 34.25	$2.18e+05 \pm 1.15e+05$
TabDDPM	0.57 ± 0.18	151.84 ± 14.24	$1.88e+05 \pm 4.08e+04$
CoDi	0.89 ± 0.10	131.71 ± 82.44	$9.34e+04 \pm 8.30e+04$
GReaT	-1.07 ± 0.96	217.61 ± 22.92	$3.39e+05 \pm 8.66e+04$
CTGAN	-3885.74 ± 6312.95	363.49 ± 29.36	$8.71e+05 \pm 8.33e+04$
TVAE	-4.24 ± 4.71	228.73 ± 39.58	$4.55e+05 \pm 1.03e+05$
TabPFN	-2.92 ± 4.00	181.81 ± 33.51	$3.94e+05 \pm 1.79e+05$

Table 8: Benchmark on causal skeletons. The values represent mean and standard deviation of metric values (SHD, F1 score, Recall, and Precision) over 10 random causal DAGs. Each metric value on a causal graph is averaged over 5 bootstrapping datasets.

(a) Linear Gaussian

	Adj	F1	Precision	Recall
ref.	3.48 ± 1.69	0.90 ± 0.06	0.92 ± 0.07	0.89 ± 0.11
TabSyn	12.96 ± 4.92	0.71 ± 0.12	0.88 ± 0.09	0.61 ± 0.16
STASY	16.44 ± 9.19	0.68 ± 0.17	0.94 ± 0.07	0.56 ± 0.19
TabDDPM	14.16 ± 8.28	0.70 ± 0.12	0.86 ± 0.11	0.61 ± 0.14
CoDi	12.32 ± 3.29	0.72 ± 0.10	0.92 ± 0.09	0.60 ± 0.13
GReaT	10.96 ± 3.53	0.75 ± 0.04	0.93 ± 0.08	0.64 ± 0.05
CTGAN	35.04 ± 5.81	0.46 ± 0.06	0.88 ± 0.13	0.32 ± 0.06
TVAE	23.32 ± 7.00	0.58 ± 0.06	0.89 ± 0.08	0.43 ± 0.07
TabPFN	4.28 ± 2.12	0.88 ± 0.07	0.92 ± 0.09	0.85 ± 0.09

(b) Linear uniform

	Adj	F1	Precision	Recall
ref.	4.04 ± 2.50	0.89 ± 0.07	0.91 ± 0.09	0.88 ± 0.10
TabSyn	13.16 ± 5.39	0.71 ± 0.08	0.89 ± 0.10	0.60 ± 0.10
STASY	14.80 ± 4.02	0.67 ± 0.10	0.89 ± 0.11	0.56 ± 0.12
TabDDPMm	10.52 ± 5.37	0.77 ± 0.07	0.91 ± 0.09	0.67 ± 0.10
CoDi	11.12 ± 4.22	0.74 ± 0.09	0.93 ± 0.09	0.64 ± 0.11
GReaT	13.48 ± 5.70	0.70 ± 0.05	0.87 ± 0.09	0.60 ± 0.06
CTGAN	29.60 ± 6.69	0.51 ± 0.08	0.89 ± 0.13	0.37 ± 0.07
TVAE	25.28 ± 6.56	0.55 ± 0.07	0.89 ± 0.12	0.41 ± 0.06
TabPFN	4.84 ± 3.76	0.88 ± 0.06	0.92 ± 0.11	0.86 ± 0.07

(c) Sigmoid Gaussian (sample size: 1500)

	Adj	F1	Precision	Recall
ref.	2.04 ± 1.46	0.95 ± 0.03	0.94 ± 0.06	0.95 ± 0.04
TabSyn	3.00 ± 1.20	0.92 ± 0.02	0.96 ± 0.04	0.89 ± 0.05
STASY	4.32 ± 2.12	0.88 ± 0.07	0.95 ± 0.06	0.84 ± 0.11
TabDDPM	2.88 ± 1.26	0.92 ± 0.03	0.95 ± 0.06	0.91 ± 0.06
CoDi	5.16 ± 2.29	0.87 ± 0.06	0.96 ± 0.05	0.80 ± 0.11
GReaT	4.88 ± 3.21	0.87 ± 0.07	0.92 ± 0.10	0.84 ± 0.06
CTGAN	20.40 ± 3.58	0.61 ± 0.05	0.94 ± 0.07	0.46 ± 0.06
TVAE	11.12 ± 4.07	0.75 ± 0.08	0.96 ± 0.05	0.63 ± 0.11
TabPFN	2.70 ± 1.07	0.93 ± 0.02	0.94 ± 0.06	0.92 ± 0.04

(d) Neural network Gaussian (sample size: 1500)

	Adj	F1	Precision	Recall
ref.	5.28 ± 3.15	0.85 ± 0.07	0.81 ± 0.15	0.93 ± 0.07
TabSyn	6.12 ± 3.36	0.84 ± 0.06	0.89 ± 0.08	0.82 ± 0.09
STASY	6.32 ± 3.79	0.83 ± 0.07	0.81 ± 0.14	0.87 ± 0.07
TabDDPM	5.48 ± 3.21	0.85 ± 0.07	0.81 ± 0.14	0.91 ± 0.05
CoDi	6.68 ± 2.75	0.82 ± 0.05	0.86 ± 0.11	0.82 ± 0.09
GReaT	5.92 ± 2.28	0.84 ± 0.05	0.86 ± 0.09	0.84 ± 0.07
CTGAN	23.28 ± 5.63	0.58 ± 0.04	0.92 ± 0.08	0.43 ± 0.04
TVAE	14.72 ± 3.83	0.69 ± 0.08	0.93 ± 0.08	0.56 ± 0.09
TabPFN	5.42 ± 3.73	0.85 ± 0.09	0.80 ± 0.15	0.93 ± 0.05

Table 9: Benchmark on d-separations: Area under the curve scores (AUC) of ROC curves. sz represents the sample size.

		ref.	TabSyn	STASY	TabDDPM	CoDi	GReaT	CTGAN	TVAE	TabPFN
AUC	LG (sz 15000)	0.972	0.927	0.930	0.814	0.917	0.921	0.566	0.703	0.970
	LU (sz 15000)	0.967	0.871	0.936	0.815	0.902	0.860	0.588	0.669	0.928
	SG (sz 5000)	0.982	0.974	0.963	0.982	0.956	0.964	0.559	0.826	0.977
	NN (sz 5000)	0.986	0.890	0.972	0.957	0.909	0.943	0.566	0.752	0.967

the original causal order. After synthetic data generation, we reverse the shuffling of the columns to restore the original order corresponding to the data generated from the DAG. This ensures the evaluation of high-order causal discovery metrics remains consistent.

D.4 Benchmarking on real-world datasets

As for evaluating baseline methods on real-world data, there is no ground-truth causal DAG available $\mathcal{G}^{\mathtt{gt}}$; hence, we consider the results of causal discovery methods on all the training data as pseudo labels. In this way, conclusions should be made carefully enough, because a worse performance

Table 10: Benchmark on the causal direction level.

(a) Evaluation with the accuracy (↑) of recovering causal directions.

	LG	LU	SG	NN
ref.	0.50	1.00	0.96	0.89
TabSyn	0.50	0.95	**0.98**	0.91
STASY	**0.70**	0.95	0.96	**0.93**
TabDDPM	0.54	1.00	0.86	0.91
CoDi	0.59	0.91	0.89	0.86
GReaT	0.55	0.93	0.91	0.50
CTGAN	0.55	0.82	0.86	0.77
TVAE	0.54	0.84	0.86	0.80
TabPFN	0.66	**0.96**	0.96	0.91

(b) Evaluation with LiNGAM on linear uniform distribution data (bootstrapping times 5).

	SHD (↓)	F1 (↑)
ref.	1.04 ± 0.85	0.94 ± 0.04
TabSyn	21.66 ± 5.75	0.41 ± 0.06
STASY	20.66 ± 4.99	0.45 ± 0.10
TabDDPM	15.86 ± 8.74	0.51 ± 0.14
CoDi	18.54 ± 5.08	0.45 ± 0.10
GReaT	13.62 ± 9.24	0.57 ± 0.11
CTGAN	33.64 ± 5.02	0.26 ± 0.07
TVAE	26.62 ± 8.15	0.29 ± 0.08
TabPFN	2.44 ± 2.62	0.88 ± 0.11

Table 11: Benchmark on causal discovery directionality using LiNGAM under two different data-generating assumptions with sample size 15000 and bootstrapping 5.

(a) LiNGAM on linear **uniform** data.

	SHD	F1	Precision	Recall
ref.	1.04 ± 0.85	0.94 ± 0.04	0.98 ± 0.03	0.92 ± 0.05
TabSyn	21.66 ± 5.75	0.41 ± 0.06	0.85 ± 0.09	0.28 ± 0.05
STASY	20.66 ± 4.99	0.45 ± 0.10	0.96 ± 0.10	0.30 ± 0.08
TabDDPM	15.86 ± 8.74	0.51 ± 0.14	0.76 ± 0.08	0.39 ± 0.14
CoDi	18.54 ± 5.08	0.45 ± 0.10	0.84 ± 0.15	0.31 ± 0.08
GReaT	13.62 ± 9.24	0.57 ± 0.11	0.84 ± 0.05	0.44 ± 0.11
CTGAN	33.64 ± 5.02	0.26 ± 0.07	0.69 ± 0.22	0.16 ± 0.04
TVAE	26.62 ± 8.15	0.29 ± 0.08	0.63 ± 0.18	0.19 ± 0.05
TabPFN	2.44 ± 2.62	0.88 ± 0.11	0.95 ± 0.08	0.83 ± 0.13

(b) LiNGAM on linear **Gaussian** data.

	SHD	F1	Precision	Recall
ref.	15.42 ± 7.01	0.38 ± 0.06	0.49 ± 0.05	0.32 ± 0.07
TabSyn	27.74 ± 5.14	0.26 ± 0.07	0.51 ± 0.17	0.17 ± 0.05
STASY	31.92 ± 4.55	0.21 ± 0.07	0.45 ± 0.13	0.13 ± 0.06
TabDDPM	26.64 ± 10.34	0.24 ± 0.09	0.42 ± 0.11	0.18 ± 0.08
CoDi	29.66 ± 5.12	0.24 ± 0.08	0.50 ± 0.13	0.16 ± 0.06
GReaT	18.18 ± 8.97	0.36 ± 0.06	0.52 ± 0.08	0.28 ± 0.07
CTGAN	36.00 ± 6.40	0.20 ± 0.06	0.54 ± 0.17	0.13 ± 0.04
TVAE	29.60 ± 8.47	0.20 ± 0.05	0.42 ± 0.12	0.13 ± 0.04
TabPFN	14.66 ± 6.16	0.38 ± 0.07	0.48 ± 0.09	0.33 ± 0.07

Table 12: Benchmark on interventional and counterfactual tasks with sample size 1000. Values are $100\times$ AMAEs (average mean absolute errors).

(a) Intervention inference.

	LG	LU	SG	NN
ref.	3.16 ± 0.2	3.07 ± 0.2	3.3 ± 0.1	3.3 ± 0.2
TabSyn	4.73 ± 1.8	4.21 ± 1.1	4.6 ± 1.1	4.7 ± 0.8
STASY	26.66 ± 7.1	23.86 ± 11.1	25.7 ± 10.5	21.3 ± 5.8
TabDDPM	4.13 ± 1.2	3.48 ± 0.6	4.2 ± 0.4	3.8 ± 0.4
CoDi	5.25 ± 1.6	3.82 ± 0.6	10.2 ± 4.2	9.5 ± 3.9
GReaT	9.77 ± 0.7	11.20 ± 1.2	9.9 ± 3.0	9.3 ± 2.2
CTGAN	15.02 ± 8.9	10.89 ± 3.2	10.8 ± 2.2	13.0 ± 3.6
TVAE	9.52 ± 5.2	12.22 ± 3.4	7.1 ± 1.3	9.4 ± 2.7
TabPFN	3.35 ± 0.3	3.87 ± 0.5	3.95 ± 0.4	4.12 ± 0.7

(b) Counterfactual inference.

	LG	LU	SG	NN
ref.	0.04 ± 0.0	0.03 ± 0.0	0.32 ± 0.2	0.23 ± 0.1
TabSyn	0.56 ± 0.4	0.45 ± 0.7	0.88 ± 0.4	0.90 ± 0.5
STASY	0.65 ± 0.4	0.44 ± 0.3	1.46 ± 0.8	1.63 ± 0.9
TabDDPM	1.12 ± 1.3	0.46 ± 0.7	1.20 ± 0.6	0.75 ± 0.3
CoDi	0.67 ± 0.8	0.58 ± 0.4	0.94 ± 0.4	1.53 ± 0.7
GReaT	0.93 ± 0.5	0.58 ± 0.4	1.47 ± 0.6	2.60 ± 1.4
CTGAN	8.58 ± 7.7	4.96 ± 3.3	5.02 ± 2.2	7.56 ± 3.8
TVAE	5.22 ± 3.9	5.40 ± 3.6	2.50 ± 1.4	4.53 ± 2.5
TabPFN	0.21 ± 0.1	0.18 ± 0.1	0.56 ± 0.2	0.76 ± 0.4

Table 13: Benchmark on low-order statistics avoiding variables ordering bias. Values are mean and standard deviation of metric values (error rate (%) of single column density, error rate (%) of pair-wise correlation score, $\alpha-$precision, $\beta-$recall) over 10 random causal DAGs.

(a) Linear Gaussian

	Col.	Pair.	α-precision	β-recall
TabSyn	1.68 ± 0.55	1.40 ± 2.86	98.61 ± 1.11	46.05 ± 11.56
STASY	10.61 ± 5.95	1.85 ± 2.99	95.08 ± 5.63	44.19 ± 11.55
TabDDPM	0.82 ± 0.15	0.52 ± 0.38	99.10 ± 0.65	50.32 ± 0.38
CoDi	5.26 ± 2.41	1.77 ± 2.89	85.29 ± 3.44	57.41 ± 16.47
GReaT	8.74 ± 0.27	0.99 ± 0.39	80.43 ± 6.34	52.67 ± 1.16
CTGAN	5.67 ± 0.57	4.17 ± 0.91	86.89 ± 5.62	14.75 ± 10.47
TVAE	4.26 ± 1.70	2.37 ± 1.23	88.50 ± 7.21	34.82 ± 11.47
TabPFN	1.59 ± 0.17	0.43 ± 0.34	94.27 ± 1.52	52.10 ± 0.41

(b) Linear uniform

	Col.	Pair.	α-precision	β-recall
TabSyn	1.73 ± 0.92	0.45 ± 0.27	98.78 ± 0.78	49.61 ± 1.55
STASY	8.82 ± 3.25	1.17 ± 0.55	95.72 ± 2.90	48.71 ± 2.58
TabDDPM	0.87 ± 0.14	0.37 ± 0.36	99.38 ± 0.25	49.54 ± 0.94
CoDi	9.19 ± 2.52	1.65 ± 1.42	65.09 ± 4.18	61.27 ± 4.16
GReaT	9.41 ± 0.83	0.44 ± 0.11	96.02 ± 1.40	45.76 ± 1.02
CTGAN	5.75 ± 0.97	4.54 ± 2.29	91.99 ± 3.29	5.23 ± 6.89
TVAE	4.98 ± 1.52	2.35 ± 0.75	94.34 ± 2.35	10.43 ± 10.45
TabPFN	1.82 ± 0.62	0.35 ± 0.12	98.73 ± 0.54	50.44 ± 0.76

(c) Sigmoid Gaussian

	Col.	Pair.	α-precision	β-recall
TabSyn	1.62 ± 0.60	0.60 ± 0.33	98.64 ± 0.82	49.55 ± 0.52
STASY	10.06 ± 2.57	1.71 ± 0.83	93.42 ± 3.53	46.18 ± 5.20
TabDDPM	0.93 ± 0.13	0.46 ± 0.21	98.93 ± 0.64	49.95 ± 0.53
CoDi	7.46 ± 2.43	1.58 ± 0.90	91.96 ± 2.73	65.53 ± 3.02
GReaT	9.53 ± 2.58	1.69 ± 1.09	89.33 ± 3.57	49.42 ± 2.10
CTGAN	5.82 ± 1.14	3.84 ± 0.73	88.81 ± 3.11	18.40 ± 6.49
TVAE	4.65 ± 0.67	1.64 ± 0.74	92.06 ± 3.14	37.86 ± 6.24
TabPFN	1.91 ± 0.16	0.48 ± 0.14	95.60 ± 0.89	51.73 ± 0.66

(d) Neural network Gaussian

	Col.	Pair.	α-precision	β-recall
TabSyn	2.46 ± 2.26	3.08 ± 7.96	98.47 ± 1.72	44.87 ± 14.91
STASY	7.69 ± 2.34	3.98 ± 8.11	96.15 ± 3.12	43.47 ± 14.71
TabDDPM	0.87 ± 0.14	0.41 ± 0.24	99.07 ± 0.37	50.48 ± 0.58
CoDi	7.02 ± 2.19	3.53 ± 7.68	86.73 ± 2.74	54.64 ± 18.31
GReaT	7.46 ± 2.65	3.81 ± 7.84	87.83 ± 4.43	43.65 ± 14.84
CTGAN	6.05 ± 1.99	6.86 ± 6.95	90.15 ± 3.40	7.05 ± 5.88
TVAE	5.53 ± 1.59	4.20 ± 7.34	90.26 ± 4.45	20.56 ± 12.09
TabPFN	1.60 ± 0.32	0.47 ± 0.08	95.36 ± 1.94	51.40 ± 0.53

Table 14: Benchmark on causal skeletons avoiding variables ordering bias. Values are mean and standard deviation of metric values (SHD, F1 score, Recall, and Precision) over 10 random causal DAGs. Each metric value on a causal graph is averaged over 10 bootstrapping datasets.

(a) Linear Gaussian

	Adj	F1	Precision	Recall
ref.	4.04 ± 1.97	0.89 ± 0.05	0.91 ± 0.10	0.88 ± 0.08
TabSyn	14.66 ± 6.32	0.67 ± 0.17	0.82 ± 0.19	0.57 ± 0.17
STASY	17.62 ± 8.55	0.64 ± 0.17	0.86 ± 0.15	0.53 ± 0.19
TabDDPM	12.78 ± 6.24	0.72 ± 0.11	0.91 ± 0.09	0.62 ± 0.14
CoDi	15.12 ± 6.20	0.66 ± 0.17	0.86 ± 0.18	0.56 ± 0.17
GReaT	14.28 ± 6.75	0.69 ± 0.09	0.87 ± 0.13	0.59 ± 0.09
CTGAN	34.04 ± 5.87	0.48 ± 0.07	0.92 ± 0.10	0.33 ± 0.07
TVAE	23.52 ± 7.07	0.58 ± 0.06	0.90 ± 0.10	0.43 ± 0.06
TabPFN	7.92 ± 6.55	0.82 ± 0.10	0.89 ± 0.10	0.78 ± 0.12

(b) Linear uniform

	Adj	F1	Precision	Recall
ref.	4.24 ± 2.02	0.88 ± 0.04	0.90 ± 0.08	0.88 ± 0.08
TabSyn	17.02 ± 5.83	0.65 ± 0.09	0.88 ± 0.12	0.52 ± 0.10
STASY	18.14 ± 7.32	0.62 ± 0.15	0.86 ± 0.14	0.52 ± 0.17
TabDDPM	14.06 ± 6.52	0.69 ± 0.06	0.85 ± 0.12	0.60 ± 0.09
CoDi	12.14 ± 4.82	0.73 ± 0.10	0.91 ± 0.12	0.62 ± 0.13
GReaT	14.00 ± 4.42	0.69 ± 0.05	0.85 ± 0.09	0.58 ± 0.04
CTGAN	33.58 ± 8.31	0.50 ± 0.09	0.93 ± 0.07	0.34 ± 0.08
TVAE	27.28 ± 9.93	0.54 ± 0.08	0.88 ± 0.11	0.39 ± 0.08
TabPFN	7.82 ± 5.20	0.81 ± 0.09	0.89 ± 0.08	0.76 ± 0.12

(c) Sigmoid Gaussian (sample size: 500)

	Adj	F1	Precision	Recall
ref.	2.98 ± 2.35	0.91 ± 0.07	0.89 ± 0.08	0.94 ± 0.07
TabSyn	3.18 ± 2.35	0.91 ± 0.05	0.90 ± 0.07	0.93 ± 0.05
STASY	3.98 ± 2.75	0.88 ± 0.08	0.88 ± 0.10	0.90 ± 0.07
TabDDPM	3.96 ± 2.83	0.88 ± 0.08	0.87 ± 0.10	0.91 ± 0.07
CoDi	3.52 ± 2.37	0.89 ± 0.08	0.89 ± 0.08	0.91 ± 0.09
GReaT	5.10 ± 3.42	0.86 ± 0.08	0.85 ± 0.09	0.87 ± 0.09
CTGAN	15.32 ± 4.90	0.67 ± 0.04	0.87 ± 0.09	0.55 ± 0.04
TVAE	7.66 ± 2.59	0.80 ± 0.04	0.88 ± 0.08	0.75 ± 0.04
TabPFN	3.30 ± 2.61	0.90 ± 0.08	0.87 ± 0.10	0.94 ± 0.06

(d) Neural network Gaussian (sample size: 500)

	Adj	F1	Precision	Recall
ref.	6.00 ± 5.28	0.84 ± 0.13	0.78 ± 0.19	0.93 ± 0.05
TabSyn	6.92 ± 5.03	0.82 ± 0.10	0.80 ± 0.15	0.85 ± 0.07
STASY	6.20 ± 4.90	0.83 ± 0.11	0.78 ± 0.18	0.92 ± 0.06
TabDDPM	5.82 ± 5.20	0.84 ± 0.12	0.79 ± 0.18	0.93 ± 0.06
CoDi	5.84 ± 5.05	0.84 ± 0.12	0.79 ± 0.19	0.93 ± 0.04
GReaT	7.32 ± 4.49	0.79 ± 0.10	0.74 ± 0.15	0.88 ± 0.06
CTGAN	16.98 ± 7.08	0.64 ± 0.09	0.83 ± 0.16	0.53 ± 0.06
TVAE	11.48 ± 6.47	0.72 ± 0.12	0.79 ± 0.18	0.68 ± 0.09
TabPFN	8.10 ± 6.55	0.77 ± 0.14	0.69 ± 0.18	0.89 ± 0.11

Table 15: Benchmark on d-separations avoiding variables reordering: Area under the curve scores (AUC) of ROC curves. sz represents the sample size.

		ref.	TabSyn	STASY	TabDDPM	CoDi	GReaT	CTGAN	TVAE	TabPFN
AUC	LG (sz=13500)	0.966	0.840	0.844	0.864	0.819	0.844	0.528	0.658	0.897
	LU (sz=13500)	0.972	0.749	0.857	0.759	0.839	0.808	0.536	0.547	0.917
	SG (sz=500)	0.952	0.941	0.941	0.956	0.963	0.959	0.830	0.926	0.955
	NN (sz=500)	0.962	0.940	0.954	0.965	0.944	0.924	0.788	0.860	0.916

Table 16: Benchmark on the causal direction level avoiding variables ordering bias: Evaluation with the accuracy (↑) of recovering causal directions.

	RECI				IGCI				CDS			
	LG	LU	SG	NN	LG	LU	SG	NN	LG	LU	SG	NN
ref.	0.571	0.982	0.304	0.232	0.446	0.125	0.911	0.804	0.571	0.946	0.964	0.696
TabSyn	0.446	0.875	0.357	0.196	0.518	0.054	0.911	0.786	0.536	0.964	0.929	0.750
STASY	0.464	0.964	0.321	0.232	0.536	0.089	0.857	0.804	0.500	0.857	0.946	0.786
TabDDPM	0.464	0.911	0.304	0.214	0.589	0.071	0.893	0.839	0.536	0.893	0.875	0.750
CoDi	0.482	0.964	0.321	0.214	0.589	0.250	0.857	0.821	0.536	0.893	0.964	0.732
GReaT	0.518	0.839	0.393	0.196	0.536	0.536	0.661	0.536	0.482	0.768	0.821	0.643
CTGAN	0.643	0.768	0.393	0.214	0.625	0.286	0.804	0.786	0.571	0.607	0.250	0.268
TVAE	0.661	0.714	0.339	0.232	0.571	0.179	0.893	0.857	0.500	0.839	0.464	0.321
TabPFN	0.464	0.875	0.393	0.268	0.482	0.161	0.857	0.893	0.482	0.911	0.964	0.554

Table 17: Benchmark on causal discovery directionality avoiding variables reordering using LiNGAM under two different data-generating assumptions with sample size 15000 and bootstrapping 10.

(a) LiNGAM on linear **uniform** data.

	SHD	F1	Precision	Recall
ref.	1.54 ± 1.58	0.91 ± 0.09	0.94 ± 0.08	0.89 ± 0.09
TabSyn	27.58 ± 6.32	0.34 ± 0.09	0.79 ± 0.17	0.22 ± 0.07
STASY	23.81 ± 3.74	0.40 ± 0.11	0.88 ± 0.11	0.26 ± 0.09
TabDDPM	21.45 ± 8.84	0.39 ± 0.13	0.71 ± 0.16	0.28 ± 0.12
CoDi	21.57 ± 4.93	0.37 ± 0.11	0.73 ± 0.18	0.25 ± 0.09
GReaT	14.76 ± 9.20	0.54 ± 0.13	0.84 ± 0.15	0.41 ± 0.12
CTGAN	32.23 ± 6.08	0.27 ± 0.07	0.70 ± 0.18	0.17 ± 0.05
TVAE	26.62 ± 10.30	0.34 ± 0.11	0.72 ± 0.19	0.22 ± 0.08
TabPFN	8.20 ± 6.35	0.68 ± 0.20	0.83 ± 0.20	0.59 ± 0.19

(b) LiNGAM on linear **Gaussian** data.

	SHD	F1	Precision	Recall
ref.	15.80 ± 7.15	0.36 ± 0.05	0.45 ± 0.04	0.31 ± 0.06
TabSyn	23.40 ± 7.53	0.31 ± 0.08	0.54 ± 0.16	0.21 ± 0.07
STASY	29.95 ± 4.87	0.24 ± 0.08	0.51 ± 0.13	0.16 ± 0.06
TabDDPM	23.54 ± 7.92	0.28 ± 0.08	0.49 ± 0.12	0.20 ± 0.08
CoDi	30.23 ± 5.38	0.20 ± 0.06	0.41 ± 0.11	0.13 ± 0.04
GReaT	19.31 ± 7.95	0.31 ± 0.08	0.47 ± 0.12	0.23 ± 0.06
CTGAN	34.71 ± 6.34	0.20 ± 0.08	0.50 ± 0.19	0.13 ± 0.05
TVAE	27.87 ± 8.31	0.23 ± 0.09	0.46 ± 0.19	0.15 ± 0.07
TabPFN	18.02 ± 9.96	0.32 ± 0.12	0.41 ± 0.13	0.27 ± 0.12

compared with pseudo labels does not necessarily mean that the performance is poor but only represents the relative differences. As shown in Table 20, TabSyn is in general the best model over the four real-world datasets on causal skeleton-level evaluation. Though CoDi and GReaT can perform well on synthetic data, they do not outperform TabSyn in real-world datasets.

Table 18: Benchmark on interventional and counterfactual tasks avoiding variables ordering bias with sample size 1000. Values are $100\times$ AMAEs (average mean absolute errors).

(a) Intervention inference.

	LG	LU	SG	NN
ref.	3.19 ± 0.2	2.96 ± 0.1	3.21 ± 0.2	3.17 ± 0.1
TabSyn	5.02 ± 2.8	4.19 ± 1.4	4.16 ± 1.7	9.08 ± 14.6
STASY	24.66 ± 12.3	17.43 ± 6.6	23.29 ± 6.4	25.70 ± 28.0
TabDDPM	3.87 ± 1.2	3.96 ± 1.4	3.62 ± 0.5	3.98 ± 0.9
CoDi	9.14 ± 10.8	3.80 ± 0.6	16.20 ± 7.5	19.12 ± 26.3
GReaT	11.50 ± 2.3	11.09 ± 0.8	12.96 ± 3.2	16.67 ± 24.9
CTGAN	16.57 ± 5.2	15.19 ± 7.9	17.97 ± 8.7	26.12 ± 27.2
TVAE	9.10 ± 5.6	10.70 ± 3.8	8.83 ± 2.2	22.44 ± 36.8
TabPFN	3.81 ± 1.7	4.13 ± 2.0	4.24 ± 0.5	3.92 ± 0.6

(b) Counterfactual inference.

	LG	LU	SG	NN
ref.	0.05 ± 0.0	0.02 ± 0.0	0.07 ± 0.0	0.06 ± 0.0
TabSyn	1.49 ± 2.6	0.52 ± 0.8	0.86 ± 0.9	10.74 ± 30.7
STASY	2.31 ± 3.4	0.42 ± 0.6	3.87 ± 2.4	10.12 ± 23.2
TabDDPM	0.96 ± 1.4	0.96 ± 1.6	0.65 ± 0.4	0.97 ± 1.1
CoDi	2.57 ± 4.4	0.59 ± 0.5	4.74 ± 4.0	8.26 ± 16.8
GReaT	1.49 ± 1.7	0.36 ± 0.2	3.24 ± 1.9	9.53 ± 21.4
CTGAN	9.54 ± 5.6	7.92 ± 5.4	11.02 ± 8.5	18.69 ± 25.9
TVAE	4.48 ± 3.0	5.14 ± 3.2	4.06 ± 2.2	13.16 ± 22.3
TabPFN	1.01 ± 2.4	0.39 ± 0.6	0.78 ± 0.4	0.64 ± 0.3

Table 19: Benchmark on low-order statistics on real-world datasets. Values are mean and standard deviation of metric values (error rate (%) of single column density, error rate (%) of pair-wise correlation score, $\alpha-$precision, $\beta-$recall).

(a) Beijing

	Col.	Pair.	α-precision	β-recall
TabSyn	3.69	6.59	99.31	47.96
STASY	8.22	11.10	93.61	50.12
TabDDPM	63.50*	63.29*	0.55*	0.70*
CoDi	20.74	6.79	95.35	52.96
GReaT	9.57	60.92	97.36	60.19
CTGAN	19.37	24.89	96.38	39.48
TVAE	36.53*	40.97*	66.89*	1.69*

(b) Magic

	Col.	Pair.	α-precision	β-recall
TabSyn	1.26	0.99	98.17	48.22
STASY	6.53	4.28	92.15	49.50
TabDDPM	0.79	1.33	98.66	47.32
CoDi	8.84	5.52	87.20	51.51
GReaT	15.16	9.66	85.17	39.90
CTGAN	4.43	7.33	90.29	15.13
TVAE	4.83	6.82	96.22	36.73

(c) House

	Col.	Pair.	α-precision	β-recall
TabSyn	3.76	1.60	95.53	40.41
STASY	8.27	2.46	96.82	49.31
TabDDPMm	1.80	2.11	97.24	47.69
CoDi	26.09	5.69	77.07	34.57
GReaT	18.28	6.17	91.90	37.68
CTGAN	15.71	9.58	51.34	16.08
TVAE	10.77	4.72	95.37	26.62

(d) Parkinsons

	Col.	Pair.	α-precision	β-recall
TabSyn	1.47	22.63	95.08	27.43
STASY	28.07	24.84	60.39	21.11
TabDDPMm	1.34	22.79	92.49	27.93
CoDi	11.57	26.61	91.66	38.86
GReaT	7.18	24.36	81.66	29.99
CTGAN	15.83	17.70	88.57	18.72
TVAE	7.66	6.55	88.86	33.10

Table 20: Benchmark on real-world data.

(a) Beijing (sample size: 15000)

	SHD	F1	Precision	Recall
ref.	1.20 ± 0.98	0.98 ± 0.02	1.00 ± 0.00	0.96 ± 0.03
TabSyn	9.60 ± 1.74	0.80 ± 0.04	0.74 ± 0.04	0.88 ± 0.06
STASY	$\mathbf{2.40 \pm 1.20}$	$\mathbf{0.96 \pm 0.02}$	1.00 ± 0.00	0.92 ± 0.04
TabDDPM	19.00 ± 1.00	0.71 ± 0.02	0.88 ± 0.04	0.59 ± 0.01
CoDi	13.40 ± 1.80	0.73 ± 0.03	0.69 ± 0.00	0.77 ± 0.06
GReaT	7.80 ± 1.40	0.84 ± 0.03	0.78 ± 0.04	0.91 ± 0.04
CTGAN	13.80 ± 1.89	0.76 ± 0.03	0.82 ± 0.04	0.70 ± 0.04
TVAE	21.80 ± 2.75	0.53 ± 0.06	0.47 ± 0.05	0.61 ± 0.07

(b) Magic (sample size: 15000)

	SHD	F1	Precision	Recall
ref.	4.40 ± 3.56	0.94 ± 0.05	0.93 ± 0.05	0.95 ± 0.04
TabSyn	$\mathbf{9.40 \pm 2.20}$	$\mathbf{0.88 \pm 0.03}$	0.86 ± 0.02	0.90 ± 0.04
STASY	12.20 ± 2.89	0.85 ± 0.04	0.83 ± 0.04	0.86 ± 0.05
TabDDPM	17.00 ± 3.38	0.79 ± 0.03	0.82 ± 0.02	0.78 ± 0.05
CoDi	16.40 ± 3.32	0.81 ± 0.04	0.85 ± 0.03	0.77 ± 0.04
GReaT	20.40 ± 2.50	0.75 ± 0.03	0.77 ± 0.02	0.74 ± 0.04
CTGAN	18.80 ± 2.23	0.78 ± 0.02	0.86 ± 0.03	0.73 ± 0.03
TVAE	21.60 ± 3.44	0.73 ± 0.04	0.75 ± 0.06	0.72 ± 0.04

(c) House (sample size: 15000)

	SHD	F1	Precision	Recall
ref.	15.20 ± 3.25	0.90 ± 0.02	0.87 ± 0.03	0.92 ± 0.04
TabSyn	30.80 ± 3.82	$\mathbf{0.80 \pm 0.02}$	0.81 ± 0.03	0.79 ± 0.03
STASY	$\mathbf{29.00 \pm 3.82}$	$\mathbf{0.80 \pm 0.03}$	0.78 ± 0.04	0.83 ± 0.02
TabDDPM	51.40 ± 2.97	0.65 ± 0.02	0.63 ± 0.04	0.67 ± 0.02
CoDi	62.60 ± 3.58	0.61 ± 0.03	0.66 ± 0.04	0.58 ± 0.02
GReaT	64.40 ± 3.67	0.61 ± 0.02	0.66 ± 0.03	0.57 ± 0.02
CTGAN	80.80 ± 1.83	0.58 ± 0.01	0.73 ± 0.02	0.48 ± 0.01
TVAE	43.40 ± 5.73	0.72 ± 0.04	0.75 ± 0.04	0.70 ± 0.04

(d) Parkinsons (sample size: 5000)

	SHD	F1	Precision	Recall
ref.	4.80 ± 3.37	0.91 ± 0.06	0.99 ± 0.03	0.84 ± 0.11
TabSyn	$\mathbf{6.60 \pm 2.20}$	$\mathbf{0.86 \pm 0.04}$	0.95 ± 0.04	0.80 ± 0.07
STASY	19.60 ± 1.74	0.62 ± 0.02	0.72 ± 0.03	0.54 ± 0.03
TabDDPM	8.20 ± 2.60	0.84 ± 0.04	0.98 ± 0.04	0.74 ± 0.06
CoDi	14.80 ± 2.04	0.72 ± 0.03	0.88 ± 0.04	0.62 ± 0.04
GReaT	13.00 ± 1.84	0.75 ± 0.03	0.88 ± 0.04	0.65 ± 0.04
CTGAN	33.20 ± 3.37	0.46 ± 0.05	0.65 ± 0.07	0.36 ± 0.04
TVAE	22.00 ± 1.26	0.61 ± 0.02	0.79 ± 0.04	0.50 ± 0.02

We pre-process the real-world dataset, Beijing, and remove the rows with any missing values and the date and time columns with strong correlation (almost deterministic relationship), "year", "month", "day", "hour", and "cbwd".

E USE OF LARGE LANGUAGE MODELS (LLMS)

In preparing this submission, LLMs were used solely as general-purpose writing assistants. Their role was limited to grammar checking, minor rephrasing for clarity, and ensuring consistency of style. LLMs were not involved in research ideation, experimental design, data analysis, or the generation of novel scientific content. All conceptual contributions, methodological developments, experiments, and results are entirely the work of the authors.

The authors take full responsibility for the content of this paper.