TabStruct: Measuring Structural Fidelity of Tabular Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Evaluating tabular generators remains a challenging problem, as the unique causal structural prior of heterogeneous tabular data does not lend itself to intuitive human inspection. Recent work has introduced structural fidelity as a tabular-specific evaluation dimension to assess whether synthetic data complies with the causal structures of real data. However, existing benchmarks often neglect the interplay between structural fidelity and conventional evaluation dimensions, thus failing to provide a holistic understanding of model performance. Moreover, they are typically limited to toy datasets, as quantifying existing structural fidelity metrics requires access to ground-truth causal structures, which is rarely available for real-world datasets. In this paper, we propose a novel evaluation framework that jointly considers structural fidelity and conventional evaluation dimensions. We introduce a new evaluation metric, *global utility*, which enables the assessment of structural fidelity even in the absence of ground-truth causal structures. In addition, we present *TabStruct*, a comprehensive evaluation benchmark offering large-scale quantitative analysis on 13 tabular generators from nine distinct categories, across 29 datasets. Our results demonstrate that global utility provides a task-independent, domain-agnostic lens for tabular generator performance. We release the TabStruct benchmark suite, including all datasets, evaluation pipelines, and raw results.

1 Introduction

Tabular data generation is a cornerstone of many real-world machine learning tasks [10, 29], ranging from training data augmentation [61, 23] to missing data imputation [98, 80]. These applications underscore the importance of generative modelling, which necessitates an appropriate understanding of the underlying data structure [50, 35, 9]. For instance, textual data conforms to the distributional hypothesis, and thus the autoregressive models are a natural workhorse for the text generation process [100, 77]. In contrast to the homogeneous modalities like text, tabular data can pose a different structural prior due to its heterogeneity – the features within a dataset typically have varying types and semantics, with feature sets that can differ across datasets [37, 80]. Recent work [41] on tabular foundation predictors has empirically demonstrated that the Structural Causal Model (SCM) is an effective structural prior of tabular data. As such, it is important to investigate how effectively existing generative models capture and leverage the causal structures of tabular data.

Prior work [40, 76, 25, 89, 59, 47] has attempted to assess tabular data generators by evaluating the synthetic data they produce. However, the prevailing evaluation paradigms still exhibit three primary limitations, which are summarised in Table 1: *(i) Insufficient tabular-specific fidelity assessments.* Current benchmarks largely adopt evaluation dimensions from homogeneous data modalities, such as density estimation [4], machine learning (ML) efficacy [94], and privacy preservation [53]. While effective in other modalities, they exhibit conceptual limitations when applied to tabular data – they do not explic-

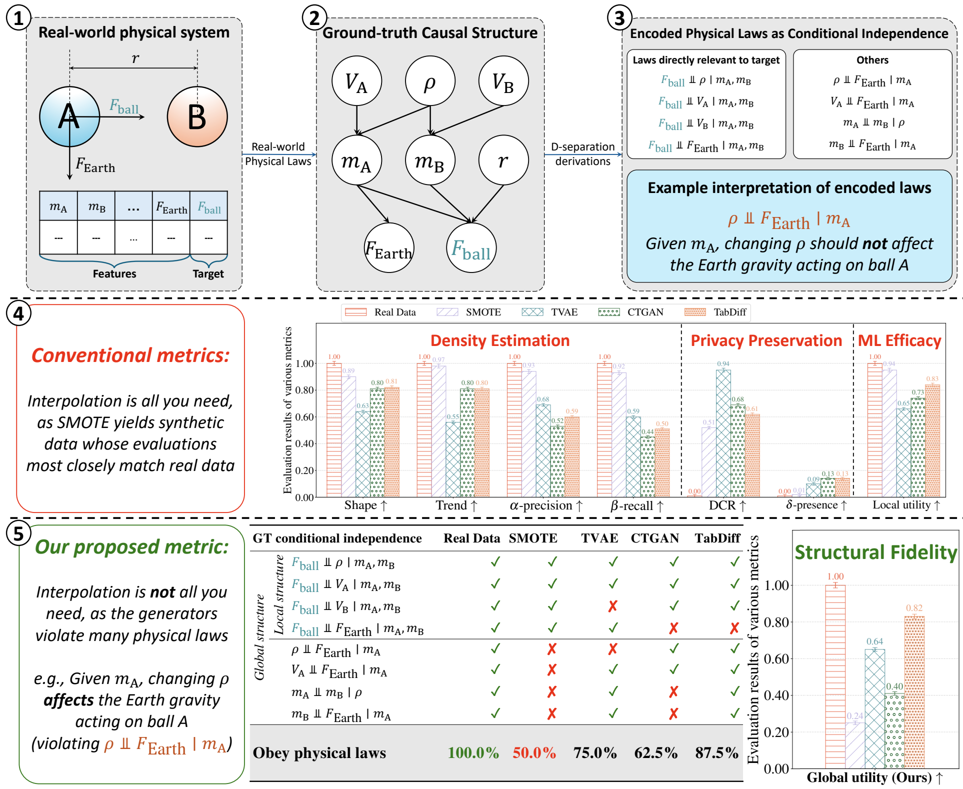

Figure 1: **Illustrative example highlighting the importance of fidelity check for tabular data structure.** ①**:** A real-world physical system showing the gravitational forces acting on ball A. The system is described by ball density (ρ), volume (V), masses (m_{A} & m_{B}), distance (r), and gravitational forces (F_{ball} & F_{Earth}). For simplicity, we assume both balls share identical density. ②**:** We derive the ground-truth (GT) causal structure of the system based on Newton's law of universal gravitation. ③**:** We interpret the encoded physical laws of the system as the conditional independence (CI) across variables. ④**:** We evaluate four generators by conventional metrics. ⑤**:** We assess the structural fidelity by CI tests and the proposed *global utility* metric. We note that the *global structure* reflects full conditional independence across all variables, while the *local structure* includes only those directly relevant to a specific prediction task at hand (F_{ball}). Results demonstrate that conventional metrics are insufficient: for instance, while SMOTE is able to outperform other generators on conventionally used dimensions (e.g., ML efficacy) – the generated synthetic data only preserves local structure and violates most physical laws. For tabular data, where the truthfulness and authenticity of synthetic data is hard to verify, global utility provides an effective mechanism for evaluating the alignment of the synthetic data to the likely ground-truth causal structure.

itly assess the unique structural prior of tabular data. A notable example is that many generators (e.g., SMOTE) can produce synthetic data with similar density estimation as real data, yet still violate underlying causal structures – such as physical laws illustrated in Figure 1(③). Although CauTabBench [89] takes a step forward to assess the structural fidelity of synthetic data, it remains confined to toy SCM datasets (i.e., synthetic datasets derived from random SCMs). Thus, CauTabBench offers limited insight into generative modelling performance on real-world tabular data, where the ground-truth SCMs are unavailable. *(ii) Potential evaluation biases.* Many benchmarks [40, 76] and model studies [94, 61, 98] prioritise ML efficacy as the principal dimension for assessing generator performance. For instance, in a classification setting, a generator is often considered effective if its synthetic data allows downstream models to achieve high predictive accuracy. However, while useful, ML efficacy can be highly sensitive to the choice of prediction task and target (Section 3.2.1). The reliance on ML efficacy can lead to biased conclusions: it tends to favour generators that are well-fitted for a specific prediction target, while obscuring their capacity to capture the global data structure (Figure 1(⑤)). *(iii) Limited evaluation scope.* Existing benchmarks mainly consider only a narrow range of datasets and generative models (Table 1), which restricts their ability to provide a thorough and generalisable comparison of model performance across the broader landscape of tabular generative modelling.

Table 1: **Evaluation scope comparison between TabStruct and prior tabular generative modelling benchmarks.** TabStruct presents a comprehensive evaluation framework for tabular generative models, incorporating a wide range of evaluation dimensions, datasets, and generator categories.

Benchmark	Conventional dimensions			Structural fidelity		Data		Generator	
	Density Estimation	Privacy Preservation	ML Efficacy	SCM data	Real-world data	# Datasets	Contamination-free	# Models	# Categories
Hansen et al. [40]	✔	✗	✔	✗	✗	11	✔	5	5
Synthcity [76]	✔	✔	✔	✗	✗	18	✗	6	4
SynMeter [25]	✔	✔	✔	✗	✗	12	✗	8	4
CauTabBench [89]	✔	✗	✗	✔	✗	10	✔	7	4
SynthEval [59]	✗	✔	✔	✗	✗	1	✔	5	3
Karpar et al.[47]	✔	✗	✔	✗	✗	2	✔	6	4
TabStruct (Ours)	✔	✔	✔	✔	✔	**29**	✔	**13**	**9**

In this paper, we aim to bridge these gaps by introducing a systematic and comprehensive evaluation framework for existing tabular generative models, with a particular focus on the structural prior of tabular data. Our proposed framework is characterised by five key concepts: (i) We explicitly incorporate **structural fidelity** of synthetic data as a core evaluation dimension for tabular generative models. Structural fidelity can directly reflect model capability in learning the structure of tabular data, without biasing towards a specific prediction target. In addition, we retain the three conventional evaluation dimensions (density estimation, privacy preservation, and ML efficacy) and investigate their interplay with structural fidelity, offering customised guidance for selecting suitable generators across diverse use cases. (ii) We evaluate structural fidelity on expert-validated SCM datasets. To ensure alignment with ground-truth causal structures, we avoid using toy SCMs and instead select SCM datasets with expert-validated causal structures. With ground-truth SCMs, we can derive the **conditional independence (CI)** of features. We then quantify structural fidelity through the difference in CI between real and synthetic data as shown in Figure 1(5)) (iii) We further extend the evaluation of structural fidelity to real-world datasets, where the ground-truth SCMs are unavailable. To this end, we propose a novel evaluation metric, **global utility**, which treats each variable as a prediction target and measures how well it can be predicted using other variables. Importantly, global utility does not require ground-truth causal structures, thus enabling the evaluation of structural fidelity in real-world scenarios. (iv) We conduct an extensive empirical study on the performance of **13 tabular generators spanning nine categories on 29 datasets**, resulting in a total of **over 150,000 evaluations**. The large evaluation scope can ensure holistic and robust benchmarking results. (v) We introduce **TabStruct**, the benchmark suite developed for this work. TabStruct features a well-structured system design and consistent APIs for building and evaluating various tabular generative models. This open-source library aims to help the research community explore tabular generative modelling within a standardised framework.

Across both SCM and real-world datasets, our primary finding is:

Structural fidelity, as quantified by the proposed global utility, should be a core dimension when evaluating tabular generative models.

The benchmark results suggest the prevailing paradigm (i.e., optimising tabular generators primarily for improved density estimation and ML efficacy) is insufficient. In contrast, global utility offers a complementary perspective – tabular-specific fidelity assessments. Finally, we find that diffusion-based generators can be considered as a reliable approach for tabular data generation, given their consistent performance in capturing high-fidelity global data structures.

Our contributions can be summarised as follows:

- **Conceptual** (Section 3): We propose a unified evaluation framework for tabular generators that integrates structural fidelity with conventional dimensions, and introduce *global utility*, a novel metric that measures structural fidelity without requiring access to ground-truth causal structures.
- **Technical** (Section 3): We release the *TabStruct* benchmark suite[1], including datasets, generator implementations, evaluation pipelines, and all raw results.
- **Empirical** (Section 4): We conduct a large-scale quantitative study of 13 tabular generators on 29 datasets. The results offer actionable insights into model performance and can inspire the design of more effective tabular generators by attending to the unique structural prior of tabular data.

[1]Code is available at `https://anonymous.4open.science/r/TabStruct-E4E4`.

2 Related Work

Tabular Generator Benchmarks. An extensive line of benchmarks [86, 40, 76, 25, 49, 82, 60] has been proposed for tabular data generation, conventionally established around three dimensions: *density estimation*, *privacy preservation*, and *ML efficacy*. Density estimation [40, 4, 80, 98] assesses the divergence between real and synthetic data distributions. However, it fails to capture inter-feature interactions and, as a result, cannot evaluate whether synthetic data preserves the causal structures present in the real data. ML efficacy [94, 76, 79, 87] evaluates the performance difference when real data is replaced with synthetic data in downstream tasks, which primarily focuses on $p(y \mid \mathbf{x})$, thus inherently prioritising feature-target relationships over inter-feature interactions. Privacy preservation [25, 53, 42, 28], although essential in privacy-sensitive scenarios, is generally task-specific and usually does not necessitate high structural fidelity [20, 59, 67]. Recent efforts such as Synthcity [76] and SynMeter [25] have aimed to standardise the evaluation of tabular data generators by incorporating the three conventional dimensions. Nonetheless, they omit explicit assessment of tabular data structure. To the best of our knowledge, CauTabBench [89] is the only other benchmark to explicitly evaluate structural fidelity, but it is limited to toy SCM datasets, as existing metrics [16, 84] typically assume access to the ground-truth SCMs – a condition that is seldom satisfied and arguably infeasible for most real-world datasets [46, 34, 102]. We further provide a detailed summary of prior studies on tabular data generation in Appendix B. As shown in Table 1, despite the ongoing progress, existing benchmarks neither comprehensively cover all evaluation dimensions nor provide a broad evaluation scope across datasets and generators. To bridge these gaps, we introduce *global utility*, an SCM-free metric that quantifies how well a generator preserves the causal structure of real data. Our *TabStruct* benchmark provides a comprehensive evaluation framework for tabular generators.

3 Methods

3.1 Problem Setup

Dataset and tabular generator. Let $\mathcal{D}_{\text{full}} \coloneqq \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N} \sim p(\mathbf{x}, y)$ represent a labelled tabular dataset with $\mathbf{x}^{(i)} \in \mathbb{R}^D$. We refer to the d-th feature (i.e., a column/variable) as $\boldsymbol{x}_d$, and the d-th feature of the i-th sample (i.e., a cell) as $x_d^{(i)}$. For notational simplicity, we define $\boldsymbol{x}_{D+1} \coloneqq \{y^{(i)}\}_{i=1}^{N}$, so that the full collection of variables, including both features and target, can be written as $\mathcal{X} \coloneqq \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_D, \boldsymbol{x}_{D+1}\}$. We denote the training split of $\mathcal{D}_{\text{full}}$ as the reference dataset ($\mathcal{D}_{\text{ref}}$), and test data as $\mathcal{D}_{\text{test}}$. A tabular generator is trained on $\mathcal{D}_{\text{ref}}$ and aims to generate synthetic data $\mathcal{D}_{\text{syn}} \sim p(\tilde{\mathbf{x}}, \tilde{y})$ close to $p(\mathbf{x}, y)$. We evaluate the quality of $\mathcal{D}_{\text{ref}}$ wrt. all the metrics, thus providing a benchmark performance against which $\mathcal{D}_{\text{syn}}$ is compared. We refer to any dataset being assessed as "evaluation dataset $\mathcal{D}$", thus, both $\mathcal{D}_{\text{ref}}$ and $\mathcal{D}_{\text{syn}}$ may serve as evaluation datasets.

Structural causal models (SCM). Under the assumptions of causal sufficiency, the Markov property, and faithfulness, an SCM is defined by the quadruple $M \coloneqq \langle \mathcal{X}, \mathcal{G}, \mathcal{F}, \mathcal{E} \rangle$. $\mathcal{G}$ is the causal graph that encodes the causal relationships among the variables. $\mathcal{E} \coloneqq \{\boldsymbol{\epsilon}_j\}_{j=1}^{D+1}$ denotes the exogenous noise, and $\mathcal{F} \coloneqq \{f_j\}_{j=1}^{D+1}$ is the set of structural functions. Each variable $\boldsymbol{x}_j$ is determined by a function f_j of its parents and its exogenous noise, that is, $\boldsymbol{x}_j = f_j\left(\text{pa}(\boldsymbol{x}_j), \boldsymbol{\epsilon}_j\right)$, where $\text{pa}(\boldsymbol{x}_j) \subseteq \mathcal{X} \setminus \{\boldsymbol{x}_j\}$ denotes the parent set of $\boldsymbol{x}_j$ in the graph $\mathcal{G}$.

3.2 Structural Fidelity

As an empirically effective structural prior for tabular data, SCM provides a formal framework for the underlying generative processes of tabular data [41, 89]. Therefore, we define the *structural fidelity* of a tabular generator as the alignment between the SCMs in its synthetic data and the ground-truth causal structures. Next, we introduce the quantifications of structural fidelity on SCM (Section 3.2.1) and real-world (Section 3.2.2) datasets. We further discuss the rationales behind using causal structural prior for tabular data in Appendix D.

3.2.1 Conditional Independence Score: Quantifying Structural Fidelity with SCM

Motivation. We begin by quantifying structural fidelity under the assumption that the ground-truth SCM is available. Following established benchmarks in causal discovery and inference [84, 46, 89], we evaluate structural fidelity at the level of the Markov equivalence class. At this level, causal structures are represented as completed partially directed acyclic graphs (CPDAGs). The SCMs of $\mathcal{D}_{\text{ref}}$ and $\mathcal{D}_{\text{syn}}$ are equivalent if they entail the same set of conditional independence (CI) statements (see Figure 1(② & ③) for an illustration). This implies that both SCMs serve as minimal I-MAPs [2] of the joint distribution factorisation $p(\mathcal{X}) = \prod_{j=1}^{D+1} p(\boldsymbol{x}_j \mid \mathrm{pa}(\boldsymbol{x}_j))$, and no causal directions can be further removed. Therefore, the CPDAG-level evaluation provides a lens to interpret the fidelity of the tabular data. Further discussion on the rationale for CPDAG-level evaluation is provided in Appendix D.

CI scores at various granularities. Following prior work [84, 89], the full set of CI statements implied by the ground-truth SCM on $\mathcal{D}_{\text{ref}}$ is defined as

$$\mathcal{C}_{\text{global}} := \left\{(\boldsymbol{x}_j \perp\!\!\!\perp \boldsymbol{x}_k \mid S_{j,k}) \mid S_{j,k} \subseteq \mathcal{X} \setminus \{\boldsymbol{x}_j, \boldsymbol{x}_k\}\right\} \cup \left\{(\boldsymbol{x}_j \not\!\perp\!\!\!\perp \boldsymbol{x}_k \mid \hat{S}_{j,k}) \mid \hat{S}_{j,k} \subsetneq S_{j,k}\right\} \quad (1)$$

where $S_{j,k}$ and $\hat{S}_{j,k}$ are the d-separation and d-connection sets for $(\boldsymbol{x}_j, \boldsymbol{x}_k)$, respectively. The derivations of CI statements are fully programmatic [85, 24, 21]. More details are in Appendix C.

For each CI statement, we assess whether it holds in the evaluation dataset $\mathcal{D}$ (i.e., $\mathcal{D}_{\text{ref}}$ or $\mathcal{D}_{\text{syn}}$) by conducting a CI test at the significance level $\alpha = 0.01$ via

$$\widehat{\mathcal{I}}_\alpha(\boldsymbol{x}_j, \boldsymbol{x}_k \mid S_{j,k}, \hat{S}_{j,k}; \mathcal{D}) = \begin{cases} 1, & \text{if the CI statement is } \textit{not} \text{ rejected on } \mathcal{D} \text{ at level } \alpha, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

To quantify structural fidelity at varying levels of granularity, we define the CI score for any subset of CI statements $\mathcal{C} \subseteq \mathcal{C}_{\text{global}}$ as:

$$\mathrm{CI}(\mathcal{C}, \mathcal{D}) = \frac{1}{|\mathcal{C}|} \sum_{\mathcal{C}} \mathbb{1}\left[\widehat{\mathcal{I}}_\alpha(\boldsymbol{x}_j, \boldsymbol{x}_k \mid S_{j,k}, \hat{S}_{j,k}; \mathcal{D}) = 1\right] \quad (3)$$

where $\mathrm{CI}(\mathcal{C}, \mathcal{D}) \in [0,1]$ measures the fidelity of selected CI statements in $\mathcal{D}$, and $\mathbb{1}(\cdot)$ denotes the indicator function. A higher CI score indicates that the evaluation dataset more closely aligns with the structure of the ground-truth SCM. Implementation details for the CI scores are in Appendix C.

Local structure vs. Global structure. We assess structural fidelity at two levels of granularity: local and global. For local structural fidelity, we define the **local CI** score, $\mathrm{CI}(\mathcal{C}_{\text{local}}, \mathcal{D})$, by considering only the CI statements that directly involve the prediction target y of a given dataset and predictive task. Specifically, we compute the local CI score using Equation (3) with $\mathcal{C}_{\text{local}} = \left\{(\boldsymbol{x}_j \perp\!\!\!\perp \boldsymbol{x}_{D+1} \mid S_{j,D+1}) \mid j \in [D]\right\} \cup \left\{(\boldsymbol{x}_j \not\!\perp\!\!\!\perp \boldsymbol{x}_{D+1} \mid \hat{S}_{j,D+1}) \mid j \in [D]\right\}$ (see Figure 1(3) for an illustration). $\mathcal{C}_{\text{local}}$ highlights which features are uninformative for predicting y when conditioned on the corresponding d-separation sets. Therefore, matching the local CI set indicates which features should be ignored when learning $p(y \mid \mathbf{x})$. A higher local CI score suggests the generator faithfully captures the local structure around the target, implying the potentially high utility of $\mathcal{D}$ for downstream predictive tasks. We empirically observe a strong correlation between the local CI score and the predictive performance on y (Section 4.2).

For global structural fidelity, we define the **global CI** score as the CI score computed over the full set of CI statements, that is, $\mathrm{CI}(\mathcal{C}_{\text{global}}, \mathcal{D})$. Global CI provides a comprehensive assessment of the entire causal structure encoded in the dataset, mitigating potential bias towards any particular variable.

3.2.2 Global Utility: SCM-free Metric for Global Structural Fidelity

Motivation. The CI scores introduced in Section 3.2.1 require access to a ground-truth SCM to enumerate the CI statements $\mathcal{C}_{\text{global}}$. However, for real-world datasets, such an SCM is typically unavailable or even non-identifiable, thereby precluding direct evaluation of structural fidelity. To address this limitation, we propose *global utility* as an SCM-free proxy for global CI.

Utility per variable. Given an evaluation dataset $\mathcal{D}$, we treat each variable $\boldsymbol{x}_j \in \mathcal{X}$ as a prediction target. An ensemble of multiple downstream predictors is trained to predict $\boldsymbol{x}_j$ using the remaining variables $\mathcal{X} \setminus \{\boldsymbol{x}_j\}$ as inputs, following a standard supervised learning setup. The predictive performance on $\mathcal{D}_{\text{test}}$ is denoted as $\mathrm{Perf}_j(\mathcal{D})$, measured using *balanced accuracy* for categorical variables and *root mean square error (RMSE)* for numerical variables. We define the utility of variable $\boldsymbol{x}_j$ as the relative performance achieved on evaluation data compared to reference data:

$$\mathrm{Utility}_j(\mathcal{D}) := \begin{cases} \mathrm{Perf}_j(\mathcal{D}_{\text{ref}})^{-1}\mathrm{Perf}_j(\mathcal{D}), & \text{if } \boldsymbol{x}_j \text{ is categorical,} \\ \mathrm{Perf}_j(\mathcal{D})^{-1}\mathrm{Perf}_j(\mathcal{D}_{\text{ref}}), & \text{if } \boldsymbol{x}_j \text{ is numerical.} \end{cases} \quad (4)$$

Utility offers a unified perspective for interpreting downstream performance across mixed variable types: $\mathrm{Utility}_j \geq 1$ indicates that downstream predictors trained on $\mathcal{D}$ perform on par with or better than those trained on $\mathcal{D}_{\text{ref}}$ for predicting $\boldsymbol{x}_j$, whereas $\text{Utility}_j < 1$ implies a loss in predictive power. To mitigate the potential bias from a specific downstream predictor, we ensemble nine different predictors with AutoGluon [27]. Full technical details are in Appendix C.

Global utility. The theoretical (Section 3.2.1) and empirical (Section 4.2) analysis showcases a strong correlation between the local CI score ($\text{CI}\left(\mathcal{C}_{\text{local}}, \mathcal{D}\right)$) and the predictive performance of y ($\text{Utility}_{D+1}(\mathcal{D})$). Therefore, we hypothesise that aggregating the utility across all features can approximate the global CI score ($\text{CI}\left(\mathcal{C}_{\text{global}}, \mathcal{D}\right)$), and we define the global utility as: $\text{Global Utility}(\mathcal{D}) \coloneqq \frac{1}{D+1}\sum_{j=1}^{D+1} \text{Utility}_j(\mathcal{D})$. Global utility is grounded in the observation that a high-fidelity generator should enable accurate conditional prediction of each variable from the others – an idea closely tied to the Markov blanket in SCMs [32, 33]. Our experiments reveal a strong correlation between global CI and global utility (Section 4.2), supporting that global utility serves as an effective and practical metric for evaluating global structural fidelity in the absence of ground-truth SCMs.

ML efficacy and local utility. The utility of the prediction target, $\text{Utility}_{D+1}(\mathcal{D})$, commonly referred to as local utility, aligns with the standard metric for assessing ML efficacy of tabular data generators. However, both theoretically (Section 3.2.1) and empirically (Section 4.2), we demonstrate that local utility can be biased, and even fail to reflect the model's ability to capture the full causal structure. In contrast, our proposed global utility mitigates this limitation by treating each feature fairly, thereby enabling a more robust and comprehensive evaluation of structural fidelity.

3.3 TabStruct Benchmark Suite

To address the limited evaluation scope of existing benchmarks, we propose **TabStruct**, a novel benchmark suite that jointly considers structural fidelity alongside conventional evaluation dimensions, and offers practical insights into real-world scenarios. Detailed descriptions are in Appendix F.

SCM datasets. To reduce the gap between causal structures in SCM and real-world data, we select six expert-validated SCM datasets from `bnlearn` [78], containing 7-64 features. Full dataset descriptions are provided in Appendix E.

Real-world datasets. We observe that many existing generators achieve near-perfect performance on commonly used benchmark datasets [80, 98], suggesting that these datasets offer limited discriminative power. To address this, we select 14 classification datasets from the hard TabZilla suite [26], containing 846-98,050 samples and 6-145 features. We further select nine challenging regression datasets, containing 345-22,784 samples and 6-82 features. Following prior work [61], we exclude any datasets employed for meta-validation of TabPFN to prevent data contamination, as TabPFN is used to compute utility scores. Full dataset descriptions are available in Appendix E.

Benchmark generators. TabStruct includes 13 existing tabular data generation methods of nine different categories: (i) a standard interpolation method SMOTE [15]; (ii) a structure learning method Bayesian Network (BN) [76]; (iii) two Variational Autoencoders (VAE) based methods TVAE [94] and GOGGLE [58]; (iv) a Generative Adversarial Networks (GAN) method CTGAN [94]; (v) a normalising flow model Neural Spine Flows (NFLOW) [26]; (vi) a tree-based method Adversarial Random Forests (ARF) [92]; (vii) three diffusion models: TabDDPM [53], TabSyn [98], TabDiff [80]; (viii) two energy-based models: TabEBM [61] and NRGBoost [12]; and (ix) a Large Language Model (LLM) based method GReaT [11]. In addition, we include $\mathcal{D}_{\text{ref}}$, where the reference data is used directly for evaluation. Full implementation details of benchmark generators are in Appendix E.

4 Experiments

We evaluate 13 tabular generators on 29 datasets by focusing on four research questions, and we further provide promising directions and practical guidance for developing tabular generative models across various use cases in Appendix G.

- **Validity of Benchmark Framework (Q1)**: Can the proposed evaluation framework, including the selected datasets and metrics, yield valid evaluation results regarding generator performance?
- **Validity of Global Utility (Q2)**: Can global utility serve as an effective metric for structural fidelity when ground-truth causal structures are unavailable?
- **Structural Fidelity of Generators (Q3)**: Can existing tabular generators accurately capture the underlying data structures across both SCM and real-world datasets?
- **Practicability of Global Utility (Q4)**: Can global utility provide stable and computationally feasible evaluation results for structural fidelity?

Table 2: **Benchmark results of 13 tabular generators on 29 datasets.** We report the normalised mean ± std metric values across datasets. "N/A" denotes that a specific metric is not applicable. We highlight the **First**, **Second** and **Third** best performances for each metric. For visualisation, we abbreviate "conditional independence" as "CI". The results show that the Top-3 methods in *Global CI* and *Global utility* are largely consistent between SCM and real-world datasets. This alignment suggests that the selected SCM datasets represent real-world causal structure, and global utility can serve as an effective proxy for global CI to evaluate global structural fidelity.

Generator	Density Estimation				Privacy Preservation		ML Efficacy	Structural Fidelity		
	Shape ↑	Trend ↑	α-precision ↑	β-recall ↑	DCR ↑	δ-Presence ↑	Local utility ↑	Local CI ↑	Global CI ↑	Global utility ↑
SCM datasets										
$\mathcal{D}_{\text{ref}}$	$1.00_{\pm0.00}$	$1.00_{\pm0.00}$	$1.00_{\pm0.00}$	$1.00_{\pm0.00}$	$0.00_{\pm0.00}$	$0.00_{\pm0.00}$	$0.99_{\pm0.01}$	$0.89_{\pm0.10}$	$1.00_{\pm0.00}$	$0.99_{\pm0.01}$
SMOTE	$\mathbf{0.82}_{\pm0.09}$	$\mathbf{0.85}_{\pm0.06}$	$0.60_{\pm0.17}$	$\mathbf{0.83}_{\pm0.01}$	$0.21_{\pm0.09}$	$0.01_{\pm0.01}$	$\mathbf{0.92}_{\pm0.07}$	$\mathbf{0.82}_{\pm0.12}$	$0.30_{\pm0.11}$	$0.39_{\pm0.09}$
BN	$\mathbf{0.80}_{\pm0.09}$	$\mathbf{0.73}_{\pm0.10}$	$\mathbf{0.78}_{\pm0.10}$	$0.32_{\pm0.08}$	$\mathbf{0.65}_{\pm0.16}$	$0.07_{\pm0.05}$	$0.41_{\pm0.17}$	$0.23_{\pm0.12}$	$0.35_{\pm0.20}$	$0.49_{\pm0.24}$
TVAE	$0.59_{\pm0.10}$	$0.59_{\pm0.14}$	$0.65_{\pm0.14}$	$0.36_{\pm0.06}$	$\mathbf{0.70}_{\pm0.10}$	$0.13_{\pm0.11}$	$0.78_{\pm0.13}$	$0.50_{\pm0.21}$	$0.40_{\pm0.09}$	$0.70_{\pm0.11}$
GOGGLE	$0.46_{\pm0.16}$	$0.50_{\pm0.13}$	$0.47_{\pm0.20}$	$0.36_{\pm0.09}$	$0.55_{\pm0.13}$	$\mathbf{0.38}_{\pm0.19}$	$0.53_{\pm0.06}$	$0.42_{\pm0.27}$	$0.14_{\pm0.03}$	$0.24_{\pm0.08}$
CTGAN	$0.46_{\pm0.14}$	$0.50_{\pm0.16}$	$0.71_{\pm0.13}$	$0.34_{\pm0.08}$	$0.52_{\pm0.11}$	$0.19_{\pm0.15}$	$\mathbf{0.80}_{\pm0.11}$	$0.61_{\pm0.08}$	$0.08_{\pm0.04}$	$0.26_{\pm0.10}$
NFlow	$0.31_{\pm0.15}$	$0.26_{\pm0.10}$	$0.31_{\pm0.21}$	$0.15_{\pm0.09}$	$\mathbf{0.73}_{\pm0.16}$	$\mathbf{0.51}_{\pm0.13}$	$0.10_{\pm0.05}$	$0.09_{\pm0.07}$	$0.09_{\pm0.07}$	$0.12_{\pm0.07}$
ARF	$\mathbf{0.75}_{\pm0.14}$	$\mathbf{0.71}_{\pm0.11}$	$\mathbf{0.79}_{\pm0.09}$	$0.36_{\pm0.09}$	$0.50_{\pm0.13}$	$0.09_{\pm0.07}$	$0.57_{\pm0.04}$	$0.21_{\pm0.09}$	$0.35_{\pm0.11}$	$0.68_{\pm0.11}$
TabDDPM	$0.62_{\pm0.11}$	$0.60_{\pm0.12}$	$0.64_{\pm0.19}$	$\mathbf{0.39}_{\pm0.09}$	$0.44_{\pm0.19}$	$0.14_{\pm0.05}$	$0.29_{\pm0.06}$	$0.17_{\pm0.08}$	$\mathbf{0.69}_{\pm0.08}$	$\mathbf{0.80}_{\pm0.05}$
TabSyn	$0.50_{\pm0.16}$	$0.48_{\pm0.17}$	$0.59_{\pm0.14}$	$0.31_{\pm0.11}$	$0.45_{\pm0.14}$	$\mathbf{0.32}_{\pm0.21}$	$0.76_{\pm0.05}$	$\mathbf{0.70}_{\pm0.06}$	$\mathbf{0.70}_{\pm0.04}$	$\mathbf{0.76}_{\pm0.06}$
TabDiff	$0.69_{\pm0.11}$	$0.62_{\pm0.15}$	$0.75_{\pm0.09}$	$0.36_{\pm0.09}$	$0.50_{\pm0.14}$	$0.13_{\pm0.03}$	$\mathbf{0.80}_{\pm0.06}$	$0.58_{\pm0.14}$	$\mathbf{0.57}_{\pm0.15}$	$\mathbf{0.75}_{\pm0.07}$
TabEBM	$0.67_{\pm0.12}$	$0.57_{\pm0.15}$	$\mathbf{0.76}_{\pm0.04}$	$0.27_{\pm0.09}$	$0.55_{\pm0.22}$	$0.14_{\pm0.06}$	$0.59_{\pm0.05}$	$0.50_{\pm0.19}$	$0.26_{\pm0.11}$	$0.30_{\pm0.08}$
NRGBoost	$0.65_{\pm0.10}$	$0.50_{\pm0.15}$	$0.61_{\pm0.14}$	$0.26_{\pm0.07}$	$0.53_{\pm0.12}$	$0.28_{\pm0.21}$	$0.75_{\pm0.01}$	$\mathbf{0.64}_{\pm0.05}$	$0.11_{\pm0.05}$	$0.16_{\pm0.02}$
GReaT	$0.62_{\pm0.09}$	$0.59_{\pm0.07}$	$0.62_{\pm0.10}$	$\mathbf{0.38}_{\pm0.07}$	$0.52_{\pm0.07}$	$0.18_{\pm0.05}$	$0.27_{\pm0.09}$	$0.17_{\pm0.04}$	$0.16_{\pm0.05}$	$0.25_{\pm0.08}$
Real-world datasets										
$\mathcal{D}_{\text{ref}}$	$1.00_{\pm0.00}$	$1.00_{\pm0.00}$	$1.00_{\pm0.00}$	$1.00_{\pm0.00}$	$0.00_{\pm0.00}$	$0.00_{\pm0.00}$	$0.96_{\pm0.06}$	N/A	N/A	$0.99_{\pm0.01}$
SMOTE	$\mathbf{0.61}_{\pm0.13}$	$\mathbf{0.87}_{\pm0.05}$	$\mathbf{0.81}_{\pm0.11}$	$\mathbf{0.77}_{\pm0.01}$	$0.19_{\pm0.09}$	$0.02_{\pm0.02}$	$\mathbf{0.91}_{\pm0.07}$	N/A	N/A	$0.41_{\pm0.04}$
BN	$\mathbf{0.66}_{\pm0.11}$	$\mathbf{0.72}_{\pm0.09}$	$\mathbf{0.86}_{\pm0.09}$	$\mathbf{0.30}_{\pm0.04}$	$0.48_{\pm0.16}$	$0.07_{\pm0.08}$	$0.38_{\pm0.16}$	N/A	N/A	$0.44_{\pm0.25}$
TVAE	$0.45_{\pm0.20}$	$0.50_{\pm0.14}$	$0.55_{\pm0.20}$	$0.18_{\pm0.04}$	$\mathbf{0.68}_{\pm0.18}$	$0.29_{\pm0.18}$	$0.70_{\pm0.06}$	N/A	N/A	$0.53_{\pm0.13}$
GOGGLE	$0.41_{\pm0.15}$	$0.47_{\pm0.14}$	$0.57_{\pm0.16}$	$0.26_{\pm0.07}$	$0.50_{\pm0.11}$	$\mathbf{0.35}_{\pm0.18}$	$0.46_{\pm0.04}$	N/A	N/A	$0.21_{\pm0.06}$
CTGAN	$0.29_{\pm0.18}$	$0.53_{\pm0.14}$	$0.66_{\pm0.21}$	$0.11_{\pm0.05}$	$0.51_{\pm0.13}$	$\mathbf{0.30}_{\pm0.24}$	$0.70_{\pm0.06}$	N/A	N/A	$0.13_{\pm0.06}$
NFlow	$0.38_{\pm0.19}$	$0.28_{\pm0.16}$	$0.52_{\pm0.15}$	$0.07_{\pm0.04}$	$\mathbf{0.64}_{\pm0.14}$	$\mathbf{0.42}_{\pm0.25}$	$0.10_{\pm0.06}$	N/A	N/A	$0.14_{\pm0.12}$
ARF	$\mathbf{0.61}_{\pm0.11}$	$0.58_{\pm0.12}$	$\mathbf{0.83}_{\pm0.10}$	$0.21_{\pm0.04}$	$0.48_{\pm0.14}$	$0.05_{\pm0.04}$	$0.54_{\pm0.07}$	N/A	N/A	$0.56_{\pm0.12}$
TabDDPM	$0.43_{\pm0.16}$	$0.49_{\pm0.18}$	$0.54_{\pm0.22}$	$0.26_{\pm0.09}$	$0.42_{\pm0.19}$	$0.27_{\pm0.18}$	$0.27_{\pm0.06}$	N/A	N/A	$\mathbf{0.72}_{\pm0.08}$
TabSyn	$0.44_{\pm0.14}$	$0.51_{\pm0.16}$	$0.62_{\pm0.18}$	$0.24_{\pm0.08}$	$0.51_{\pm0.12}$	$0.24_{\pm0.14}$	$\mathbf{0.76}_{\pm0.08}$	N/A	N/A	$\mathbf{0.73}_{\pm0.07}$
TabDiff	$0.54_{\pm0.15}$	$0.52_{\pm0.16}$	$0.69_{\pm0.12}$	$0.22_{\pm0.07}$	$0.57_{\pm0.15}$	$0.20_{\pm0.13}$	$\mathbf{0.78}_{\pm0.03}$	N/A	N/A	$\mathbf{0.73}_{\pm0.07}$
TabEBM	$0.59_{\pm0.15}$	$\mathbf{0.65}_{\pm0.08}$	$0.79_{\pm0.04}$	$\mathbf{0.30}_{\pm0.10}$	$\mathbf{0.58}_{\pm0.16}$	$0.14_{\pm0.03}$	$0.63_{\pm0.11}$	N/A	N/A	$0.35_{\pm0.11}$
NRGBoost	$0.54_{\pm0.12}$	$0.49_{\pm0.13}$	$0.62_{\pm0.16}$	$0.20_{\pm0.07}$	$0.51_{\pm0.15}$	$0.22_{\pm0.13}$	$0.74_{\pm0.05}$	N/A	N/A	$0.16_{\pm0.05}$
GReaT	$0.47_{\pm0.10}$	$0.49_{\pm0.13}$	$0.57_{\pm0.14}$	$0.26_{\pm0.08}$	$0.52_{\pm0.11}$	$0.27_{\pm0.15}$	$0.23_{\pm0.07}$	N/A	N/A	$0.20_{\pm0.06}$

Experimental setup. For each dataset of N samples, we first split it into train and test sets (80% train and 20% test). We further split the train set into a training split ($\mathcal{D}_{\text{ref}}$) and a validation split (90% training and 10% validation). We repeat the splitting 10 times, summing up to 10 runs per dataset. All benchmark generators are trained on $\mathcal{D}_{\text{ref}}$, and each generator produces a synthetic dataset with N_{ref} samples. We tune the parameterised generators using Optuna [3] to minimise their average validation loss across 10 repeated runs. Each generator is given at most two hours to complete a single repeat. The reported results are averaged by default over 10 repeats. We aggregate results across all SCM or real-world datasets because the findings are consistent across classification and regression tasks. Specifically, we use the average distance to the minimum (ADTM) metric via affine renormalisation between the top-performing and worse-performing models [37, 63, 41, 61, 44]. We further provide the detailed configurations (Appendix E) and raw results (Appendix H).

4.1 Validity of Benchmark Framework (Q1)

The benchmark metrics effectively evaluate data quality. Table 2 demonstrates that all metrics effectively distinguish between high- and low-quality data. Specifically, except for privacy-related metrics, the reference data ($\mathcal{D}_{\text{ref}}$) consistently achieves the highest scores. This is expected, as $\mathcal{D}_{\text{ref}}$ is the ground truth and should score highly on metrics of density estimation, ML efficacy, and structural fidelity. In contrast, privacy metrics reward greater differences from the ground truth to indicate stronger privacy preservation. Since $\mathcal{D}_{\text{ref}}$ is identical to the ground truth, it naturally scores poorly for privacy. These results show that the selected metrics provide appropriate evaluations for data quality. Therefore, we consider the evaluation results to be valid and meaningful for analysis.

The benchmark datasets present a genuine challenge for existing generators. As detailed in Section 3.3, we select challenging, contamination-free real-world datasets, ensuring that they are non-trivial for existing tabular data generators. Table 2 illustrates that, unlike prior studies [80, 98, 61], no generator can easily match $\mathcal{D}_{\text{ref}}$ on our benchmark datasets. This confirms that the selected datasets offer a more informative and realistic assessment of generator capabilities.

4.2 Validity of Global Utility (Q2)

Global utility serves as an effective metric for global structural fidelity. Table 2 and Figure 2 (left) demonstrate a strong monotonic correlation between global utility and global CI scores

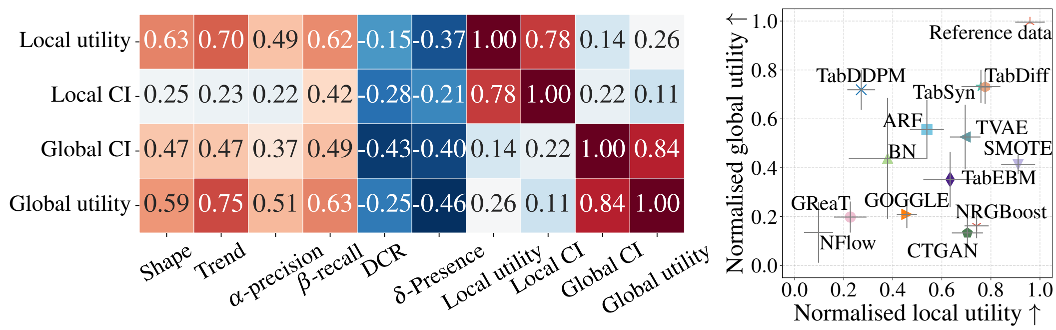

Figure 2: **Left:** Spearman's rank correlation heatmap based on metric values on six SCM datasets. Global utility correlates strongly with global CI, suggesting that global utility can effectively assess global structural fidelity as an SCM-free proxy of global CI. **Right:** Mean normalised local utility vs. mean normalised global utility on 23 real-world datasets. SMOTE prioritises local utility, whereas TabDiff and TabSyn generally achieve a balanced preservation of both global and local data structures.

($r_s = 0.84, \mathrm{p} < 0.001$). In addition, global CI exhibits weaker correlations with all other metrics, showing conventional dimensions cannot sufficiently indicate structural fidelity. Appendix G further shows that global utility more closely approximates the ranking induced by global CI than local utility. In other words, generators that excel in capturing global structure also score highly in global utility (e.g., TabSyn and TabDiff), while the ones that struggle to capture global structure (e.g., NFlow and NRGBoost) perform poorly under global utility. Although global utility does not theoretically guarantee causal alignment, we note that even state-of-the-art causal discovery methods *cannot* theoretically ensure the inferred graphs align with the ground-truth SCMs [46]. In contrast, the empirical results validate the use of global utility for measuring global structural fidelity when the ground-truth SCM is unavailable.

Local utility is not always the golden standard, due to its bias towards the local structure. We further examine the correlation between local utility and local CI, which only considers the local structure associated with the prediction target. As shown in Figure 2 (left), local utility exhibits a strong correlation with local CI ($r_s = 0.78, \mathrm{p} < 0.001$), but a much weaker correlation with global CI ($r_s = 0.14, \mathrm{p} < 0.001$). The results indicate that local utility may reward "myopic" generators while missing the holistic data structure. This underscores the necessity of global utility for a more comprehensive evaluation of structural fidelity in tabular generation.

4.3 Structural Fidelity of Generators (Q3)

Diffusion models generally capture the global structure well. As reported in Table 2 and Figure 2 (right), diffusion-based models consistently achieve the highest scores in global structural fidelity: the Top-3 methods are TabDDPM, TabSyn, and TabDiff across both SCM and real-world datasets. We attribute their strong performance to the inherent capacity of diffusion models for learning permutation-invariant conditional distributions of each feature. At the training stage, since Gaussian noise is added independently to each feature, the diffusion network is optimised at every denoising step to reconstruct each feature by conditioning on all others. Consequently, it learns the conditional distribution for every feature $p(\boldsymbol{x}_j \mid \mathcal{X} \setminus \{\boldsymbol{x}_j\})$ simultaneously. Moreover, unlike autoregressive models, which generally rely on a fixed generation order, diffusion models impose no ordering constraint. This results in efficient computation (Figure 3) and permutation-invariant conditional distributions, a property that aligns naturally with the structure of tabular data. These theoretical properties align with the conditional independence analysis in Section 3.2.1, thus confirming that diffusion models are capable of capturing global structure.

Interpolation and energy-based methods tend to prioritise local structure over global structure. Figure 2 (right) shows that the interpolation method (e.g., SMOTE) and energy-based models (e.g., TabEBM and NRGBoost) can effectively capture local structure, yet perform poorly when modelling global structure. These two families of methods share a common trait in their generation process: they generate new samples from class-specific reference data. For example, in classification tasks, SMOTE interpolates between samples of the same class, and TabEBM samples from a class-specific energy surface. As a result, the generated samples are inevitably biased towards local structure.

Structure learning methods struggle with tabular data generation. One surprising finding is that BN and GOGGLE do not demonstrate strong performance in terms of structural fidelity, despite their

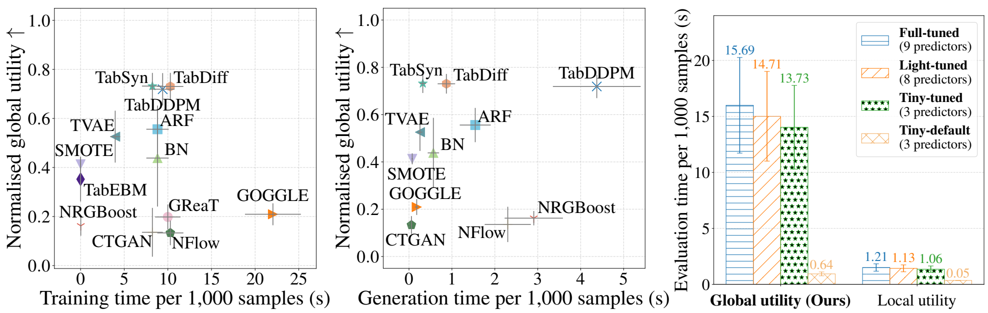

Figure 3: **Computation efficiency on 23 real-world datasets. Left:** Median training time per 1,000 samples vs. mean normalised global utility. **Middle:** Median generation time per 1,000 samples vs. mean normalised global utility. We exclude the outliers (TabEBM and GReaT) due to their long generation time (over 30s). Full results are in Appendix H. **Right:** Median evaluation time. Because global utility yields stable rankings across downstream predictors (Appendix G), computing global utility can be highly efficient with only a small ensemble of predictors (i.e., Tiny-default).

inductive bias towards learning tabular data structures. This observation aligns with prior work [89], which highlights that current causal discovery algorithms often struggle when the number of features exceeds 10. In contrast, our benchmark datasets have features from 7 up to 145. Furthermore, GOGGLE exhibits notable performance degradation when prior knowledge about the data structure is missing [58]. The results underscore the limitations of existing causal discovery methods in recovering precise causal structures from real-world data, further justifying our evaluation at the CPDAG level.

Discussion on more models. We further provide a detailed performance analysis of the remaining generators, such as autoregressive models, in Appendix G.

4.4 Practicability of Global Utility (Q4)

Global utility is robust, stable, and efficient. To evaluate the robustness of global utility, we examine how generator rankings vary under different configurations of downstream models. Appendix G shows that global utility yields stable rankings across both nine tuned predictors ("Full-tuned") and three untuned ones ("Tiny-default"). In contrast, local utility necessitates nine tuned predictors ("Full-tuned") for reliable results. In practice, we are often interested in identifying the most promising model before fine-tuning it for optimal performance [63]. As such, global utility offers an efficient and informative evaluation. As illustrated in Figure 3 (right), global utility ("Tiny-default") provides reliable rankings with a median runtime of just 0.64 seconds per 1,000 samples, nearly half the time required by local utility (1.21 seconds with "Full-tuned").

Limitations and future work. While our proposed global utility is a robust and effective metric for assessing global structural fidelity, it is an empirical approximation of the likely SCMs behind the data at hand. This is in line with several open challenges in the field, specifically the lack of causal discovery methods that can reliably infer the governing SCMs of real-world tabular data with strong theoretical guarantees [46, 89, 34, 71]. Addressing this gap would require advances in causal modelling, which we leave for future work. More discussion on future work is in Appendix G.

5 Conclusion

We present TabStruct, a principled benchmark for tabular data generators along with both structural fidelity and conventional dimensions. To address the challenge of assessing structural fidelity in the absence of ground-truth SCMs, we introduce global utility – a novel, SCM-free metric that enables unbiased and holistic evaluation for tabular data structure.

In our large-scale study of 13 generators across 29 datasets, we find that existing evaluation methods often favour models that capture local correlations while neglecting global structure. Our results show that diffusion models, due to their permutation-invariant generation process, offer valuable insights into the fundamental representation learning of tabular data. We further observe that the four evaluation dimensions are complementary, offering practical guidance for selecting suitable generators across diverse applications. TabStruct is an ongoing effort. As such, it will continue to evolve with additional datasets, generators, and evaluation metrics – both through our engagement and contributions from the community. We envision that the open-source nature of TabStruct will help drive progress in tabular generative modelling.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Raj Agrawal, Caroline Uhler, and Tamara Broderick. Minimal i-map mcmc for scalable structure discovery in causal dag models. In *International Conference on Machine Learning*, pages 89–98. PMLR, 2018.

[3] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[4] 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. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022.

[5] Ankur Ankan and Johannes Textor. A simple unified approach to testing high-dimensional conditional independences for categorical and ordinal data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12180–12188, 2023.

[6] Ankur Ankan and Johannes Textor. pgmpy: A python toolkit for bayesian networks. *Journal of Machine Learning Research*, 25(265):1–8, 2024.

[7] Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.

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

[9] Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F Jensen. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1608, 2022.

[10] 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.

[11] Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. In *The Eleventh International Conference on Learning Representations*, 2023.

[12] João Bravo. Nrgboost: Energy-based generative boosted trees. *International Conference on Learning Representations*, 2025.

[13] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[14] Gerlise Chan, Tom Claassen, Holger Hoos, Tom Heskes, and Mitra Baratchi. Autocd: Automated machine learning for causal discovery algorithms. *Sl: OpenReview*, 2024.

[15] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[16] 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*, 36:66438–66450, 2023.

[17] Pei Chen, Soumajyoti Sarkar, Leonard Lausen, Balasubramaniam Srinivasan, Sheng Zha, Ruihong Huang, and George Karypis. Hytrel: Hypergraph-enhanced tabular data representation learning. *Advances in Neural Information Processing Systems*, 36:32173–32193, 2023.

[18] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[19] CKCN Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

[20] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, Mukund Lahoti, and Pratik Narang. A universal metric for robust evaluation of synthetic tabular data. *IEEE Transactions on Artificial Intelligence*, 5(1):300–309, 2022.

[21] Panayiota Constantinou and A Philip Dawid. Extended conditional independence and applications in causal inference. *The Annals of Statistics*, pages 2618–2653, 2017.

[22] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232, 1958.

[23] Lingxi Cui, Huan Li, Ke Chen, Lidan Shou, and Gang Chen. Tabular data augmentation for machine learning: Progress and prospects of embracing generative ai. *arXiv preprint arXiv:2407.21523*, 2024.

[24] A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(1):1–15, 1979.

[25] Yuntao Du and Ninghui Li. Systematic assessment of tabular data synthesis algorithms. *arXiv e-prints*, pages arXiv–2402, 2024.

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

[27] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

[28] Erica Espinosa and Alvaro Figueira. On the quality of synthetic generated tabular data. *Mathematics*, 11(15):3278, 2023.

[29] 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 e-prints*, pages arXiv–2402, 2024.

[30] Evelyn Fix. *Discriminatory analysis: nonparametric discrimination, consistency properties*, volume 1. USAF school of Aviation Medicine, 1985.

[31] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine learning*, 29:131–163, 1997.

[32] Shunkai Fu and Michel C Desmarais. Markov blanket based feature selection: a review of past decade. In *Proceedings of the world congress on engineering*, volume 1, pages 321–328. Newswood Ltd. Hong Kong, China, 2010.

[33] Tian Gao and Qiang Ji. Efficient markov blanket discovery and its application. *IEEE transactions on Cybernetics*, 47(5):1169–1179, 2016.

[34] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[35] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[36] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

[37] 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*, 35:507–520, 2022.

[38] Manbir Gulati and Paul Roysdon. Tabmt: Generating tabular data with masked transformers. *Advances in Neural Information Processing Systems*, 36:46245–46254, 2023.

[39] Trygve Haavelmo. The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115, 1944.

[40] 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*, 36:33781–33823, 2023.

[41] 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.

[42] Yuzheng Hu, Fan Wu, Qinbin Li, Yunhui Long, Gonzalo Munilla Garrido, Chang Ge, Bolin Ding, David Forsyth, Bo Li, and Dawn Song. Sok: Privacy-preserving data synthesis. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 4696–4713. IEEE, 2024.

[43] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[44] Xiangjian Jiang, Andrei Margeloiu, Nikola Simidjievski, and Mateja Jamnik. Protogate: Prototype-based neural networks with global-to-local feature selection for tabular biomedical data. In *Forty-first International Conference on Machine Learning*, 2024.

[45] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.

[46] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

[47] Jan Kapar, Niklas Koenen, and Martin Jullum. What's wrong with your synthetic tabular data? using explainable ai to evaluate generative models. *arXiv e-prints*, pages arXiv–2504, 2025.

[48] Jayoung Kim, Chaejeong Lee, and Noseong Park. Stasy: Score-based tabular data synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

[49] G Charbel N Kindji, Lina Maria Rojas-Barahona, Elisa Fromont, and Tanguy Urvoy. Under the hood of tabular data generation models: the strong impact of hyperparameter tuning. *arXiv preprint arXiv:2406.12945*, 2024.

[50] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:1, 2014.

[51] Neville Kenneth Kitson, Anthony C Constantinou, Zhigao Guo, Yang Liu, and Kiattikun Chobtham. A survey of bayesian network structure learning. *Artificial Intelligence Review*, 56(8):8721–8814, 2023.

[52] Daphane Koller. Probabilistic graphical models: Principles and techniques, 2009.

[53] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR, 2023.

[54] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[55] Roman Levin, Valeriia Cherepanova, Avi Schwarzschild, Arpit Bansal, C Bayan Bruss, Tom Goldstein, Andrew Gordon Wilson, and Micah Goldblum. Transfer learning with deep tabular models. In *The Eleventh International Conference on Learning Representations*, 2023.

[56] Chun Li and Bryan E Shepherd. Test of association between two ordinal variables while adjusting for covariates. *Journal of the American Statistical Association*, 105(490):612–620, 2010.

[57] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. Pacgan: The power of two samples in generative adversarial networks. *Advances in neural information processing systems*, 31, 2018.

[58] Tennison Liu, Zhaozhi Qian, Jeroen Berrevoets, and Mihaela van der Schaar. Goggle: Generative modelling for tabular data by learning relational structure. In *The Eleventh International Conference on Learning Representations*, 2023.

[59] Ioannis E Livieris, Nikos Alimpertis, George Domalis, and Dimitris Tsakalidis. An evaluation framework for synthetic data generation models. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 320–335. Springer, 2024.

[60] Yunbo Long, Liming Xu, and Alexandra Brintrup. Evaluating inter-column logical relationships in synthetic tabular data generation. *arXiv preprint arXiv:2502.04055*, 2025.

[61] Andrei Margeloiu, Xiangjian Jiang, Nikola Simidjievski, and Mateja Jamnik. Tabebm: A tabular data augmentation method with distinct class-specific energy-based models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[62] Calvin McCarter. Unmasking trees for tabular data. In *NeurIPS 2024 Third Table Representation Learning Workshop*, 2024.

[63] Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? *Advances in Neural Information Processing Systems*, 36, 2024.

[64] Mary L McHugh. The chi-square test of independence. *Biochemia medica*, 23(2):143–149, 2013.

[65] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the nist contest: A scalable and general approach to differentially private synthetic data. *arXiv preprint arXiv:2108.04978*, 2021.

[66] Ryan McKenna, Daniel Sheldon, and Gerome Miklau. Graphical-model based estimation and inference for differential privacy. In *International Conference on Machine Learning*, pages 4435–4444. PMLR, 2019.

[67] Scott McLachlan, Kudakwashe Dube, Thomas Gallagher, Bridget Daley, Jason Walonoski, et al. The aten framework for creating the realistic synthetic electronic health record. *International Joint Conference on Biomedical Engineering Systems and Technologies*, 2018.

[68] Keith E Muller and Bercedis L Peterson. Practical methods for computing power in testing the multivariate general linear hypothesis. *Computational Statistics & Data Analysis*, 2(2):143–158, 1984.

[69] Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022.

[70] Alhassan Mumuni and Fuseini Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258, 2022.

[71] Vivian Nastl and Moritz Hardt. Do causal predictors generalize better to new domains? *Advances in Neural Information Processing Systems*, 37:31202–31315, 2024.

[72] Wei Pang, Masoumeh Shafieinejad, Lucy Liu, Stephanie Hazlewood, and Xi He. Clavaddpm: Multi-relational data synthesis with cluster-guided diffusion models. *Advances in Neural Information Processing Systems*, 37:83521–83547, 2024.

[73] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[74] Parjanya Prajakta Prashant, Ignavier Ng, Kun Zhang, and Biwei Huang. Differentiable causal discovery for latent hierarchical causal models. In *NeurIPS 2024 Causal Representation Learning Workshop*, 2024.

[75] Andreas Psaroudakis and Dimitrios Kollias. Mixaugment & mixup: Augmentation methods for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2375, 2022.

[76] 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*, 36, 2024.

[77] Magnus Sahlgren. The distributional hypothesis. *Italian Journal of linguistics*, 20:33–53, 2008.

[78] Marco Scutari. bnlearn-an r package for bayesian network learning and inference. *UCL Genetics Institute, University College, London, London, UK*, 2011.

[79] Nabeel Seedat, Nicolas Huynh, Boris Van Breugel, and Mihaela Van Der Schaar. Curated llm: Synergy of llms and data curation for tabular augmentation in low-data regimes. In *International Conference on Machine Learning*, pages 44060–44092. PMLR, 2024.

[80] Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. Tabdiff: a multi-modal diffusion model for tabular data generation. *International Conference on Learning Representations*, 2025.

[81] Shohei Shimizu. Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

[82] Andrey Sidorenko, Michael Platzer, Mario Scriminaci, and Paul Tiwald. Benchmarking synthetic tabular data: A multi-dimensional evaluation framework. *arXiv preprint arXiv:2504.01908*, 2025.

[83] AV Solatorio and O Dupriez. Realtabformer: Generating realistic relational and tabular data using transformers. arxiv. *arXiv preprint arXiv:2302.02041*, 2023.

[84] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

[85] Wolfgang Spohn. Stochastic independence, causal independence, and shieldability. *Journal of Philosophical logic*, 9:73–99, 1980.

[86] Mihaela CÄ Stoian, Eleonora Giunchiglia, and Thomas Lukasiewicz. A survey on tabular data generation: Utility, alignment, fidelity, privacy, and beyond. *arXiv preprint arXiv:2503.05954*, 2025.

[87] Paul Tiwald, Ivona Krchova, Andrey Sidorenko, Mariana Vargas Vieyra, Mario Scriminaci, and Michael Platzer. Tabularargn: A flexible and efficient auto-regressive framework for generating high-fidelity synthetic data. *arXiv preprint arXiv:2501.12012*, 2025.

[88] Gianluca Truda. Generating tabular datasets under differential privacy. *arXiv preprint arXiv:2308.14784*, 2023.

[89] Ruibo Tu, Zineb Senane, Lele Cao, Cheng Zhang, Hedvig Kjellström, and Gustav Eje Henter. Causality for tabular data synthesis: A high-order structure causal benchmark framework. *arXiv preprint arXiv:2406.08311*, 2024.

[90] Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34:18853–18865, 2021.

[91] Zifeng Wang and Jimeng Sun. Transtab: Learning transferable tabular transformers across tables. *Advances in Neural Information Processing Systems*, 35:2902–2915, 2022.

[92] David S Watson, Kristin Blesch, Jan Kapar, and Marvin N Wright. Adversarial random forests for density estimation and generative modeling. In *International Conference on Artificial Intelligence and Statistics*, pages 5357–5375. PMLR, 2023.

[93] Jürgen Wüst. Sdmetrics. *Online: http://www. sdmetrics. com*, 2011.

[94] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

[95] Zhilin Yang. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[96] Alessio Zanga, Elif Ozkirimli, and Fabio Stella. A survey on causal discovery: theory and practice. *International Journal of Approximate Reasoning*, 151:101–129, 2022.

[97] Yan Zeng, Shohei Shimizu, Hidetoshi Matsui, and Fuchun Sun. Causal discovery for linear mixed data. In *Conference on Causal Learning and Reasoning*, pages 994–1009. PMLR, 2022.

[98] Hengrui Zhang, Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. In *The Twelfth International Conference on Learning Representations*, 2023.

[99] Zhikun Zhang, Tianhao Wang, Ninghui Li, Jean Honorio, Michael Backes, Shibo He, Jiming Chen, and Yang Zhang. {PrivSyn}: Differentially private data synthesis. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 929–946, 2021.

[100] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[101] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning*, pages 97–112. PMLR, 2021.

[102] Wei Zhou, Hong Huang, Guowen Zhang, Ruize Shi, Kehan Yin, Yuanyuan Lin, and Bang Liu. Ocdb: Revisiting causal discovery with a comprehensive benchmark and evaluation framework. *arXiv preprint arXiv:2406.04598*, 2024.

[103] Bingzhao Zhu, Xingjian Shi, Nick Erickson, Mu Li, George Karypis, and Mahsa Shoaran. Xtab: cross-table pretraining for tabular transformers. In *Proceedings of the 40th International Conference on Machine Learning*, pages 43181–43204, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 1 details our research objectives and highlights our contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Presented in Section 4.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Section 3 presents the theoretical results of our proposed metric.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Refer to Appendix E, where we provide full details on reproducing the results in the paper. We provide an open-source library of the proposed benchmark.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Refer to Appendix E. All datasets used in this paper are publicly available, and the implementations of benchmark generators are open-source. We also provide an open-source library `https://anonymous.4open.science/r/TabStruct-E4E4`.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix E provides full descriptions of the experimental setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Refer to Section 4, where we provide standard deviations for all tables. Figure 2 (right) and Figure 3 contain error bars. In Figure 2 (left), we show statistical significance tests of the correlation between different metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Refer to Appendix E, where we provide full details on the computation resources used in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We carefully check the NeurIPS Code of Ethics, and we confirm that our work follows the Code in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Refer to Appendix A, where we include the societal impacts of our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Refer to Appendix E, where we provide the open-source licenses followed by the creators or original owners of assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the implementation of our proposed metric and benchmark as a python library attached to this submission. We will make it publicly available post-publication.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.