HOW WELL DOES YOUR TABULAR GENERATOR LEARN THE STRUCTURE OF TABULAR DATA?

Xiangjian Jiang[1], **Nikola Simidjievski**[1,2] **& Mateja Jamnik**[1]
[1]Department of Computer Science and Technology
[2]PBCI, Department of Oncology
University of Cambridge
Cambridge, UK
{xj265,ns779,mj201}@cam.ac.uk

ABSTRACT

Heterogeneous tabular data poses unique challenges in generative modelling due to its fundamentally different underlying data structure compared to homogeneous modalities, such as images and text. Although previous research has sought to adapt the successes of generative modelling in homogeneous modalities to the tabular domain, defining an effective generator for tabular data remains an open problem. One major reason is that the evaluation criteria inherited from other modalities often fail to adequately assess whether tabular generative models effectively capture or utilise the unique structural information encoded in tabular data. In this paper, we carefully examine the limitations of the prevailing evaluation framework and introduce **TabStruct**, a novel evaluation benchmark that positions structural fidelity as a core evaluation dimension. Specifically, TabStruct evaluates the alignment of causal structures in real and synthetic data, providing a direct measure of how effectively tabular generative models learn the structure of tabular data. Through extensive experiments using generators from eight categories on seven datasets with expert-validated causal graphical structures, we show that structural fidelity offers a task-independent, domain-agnostic evaluation dimension. Our findings highlight the importance of tabular data structure and offer practical guidance for developing more effective and robust tabular generative models. Code is available at https://github.com/SilenceX12138/TabStruct.

1 INTRODUCTION

Tabular data generation is a cornerstone of many real-world machine learning tasks (Borisov et al., 2022; Fang et al., 2024), ranging from training data augmentation (Margeloiu et al., 2024; Cui et al., 2024) to missing data imputation (Zhang et al., 2023). These applications highlight the importance of building powerful models capable of generating high-quality synthetic tabular data, which necessitates an appropriate understanding of the underlying data structure. For instance, textual data conforms to the distributional hypothesis, and thus the autoregressive process can be a natural and effective approach for text generation (Zhao et al., 2023; Sahlgren, 2008). In contrast, tabular data poses unique challenges due to its heterogeneity – the features within a dataset typically have varying types and semantics, with feature sets that can differ across datasets (Grinsztajn et al., 2022; Shi et al., 2024). Recent work in tabular foundation predictors demonstrates that (causal) structure can be an effective prior for tabular data structure (Hollmann et al., 2025), which is fundamentally different to homogeneous modalities like text or images. As such, it is important to investigate how effectively existing tabular generative models capture and leverage the tabular data structure.

Prior work (Hansen et al., 2023; Zhang et al., 2023; Margeloiu et al., 2024) has proposed tabular generative models spanning multiple categories for high-quality synthetic data. However, a fair and comprehensive benchmarking framework remains absent. Specifically, existing benchmarks exhibit three primary limitations: **(i) Lack of evaluating the tabular data structure.** The mainstream benchmarks primarily adopt evaluation dimensions from homogeneous modalities, including density estimation (Alaa et al., 2022), downstream utility (Xu et al., 2019), and privacy preservation (Kotelnikov et al., 2023). While these metrics have proven effective for other modalities, they

fail to fully assess whether tabular generative models capture the unique structural information of tabular data. **(ii) Potentially biased evaluation.** Beyond overlooking structural information, certain conventional evaluation metrics may introduce bias (see Section 2.4 for more details). For instance, evaluating synthetic data based on downstream utility depends heavily on the choice of the performance metric as well as the downstream models and tasks (Hansen et al., 2023; Margeloiu et al., 2024), which may obscure the true capabilities of tabular generative models. **(iii) Limited coverage of tabular generative models.** Existing benchmarks often evaluate a narrow range of tabular generative models, limiting their ability to provide a comprehensive comparison of model performance across the broader landscape of tabular generative modelling. Appendix A further summarises the scope of the evaluation metrics and generators in TabStruct and existing benchmarks. In this paper, we aim to address these gaps by developing a systematic and comprehensive evaluation framework for existing tabular generative models.

We introduce **TabStruct** (Figure 1), a novel benchmark framework designed to comprehensively evaluate tabular generative models across diverse metrics and model categories. TabStruct is characterised by three core concepts. Firstly, TabStruct positions structural fidelity as a core evaluation dimension, and quantifies it through the alignment of feature independence relationships between real and synthetic data. Secondly, TabStruct retains the conventional evaluation metrics and investigates their interplay with structural fidelity. Thirdly, TabStruct includes eight generator categories, ensuring holistic and robust benchmarking results.

Our contributions can be summarised as follows: ① **Conceptual** (Section 2): We propose TabStruct, a novel benchmark framework that integrates structural fidelity as a core evaluation dimension for tabular generative models. ② **Empirical** (Section 3): We quantitatively analyse the model capabilities across four dimensions and provide actionable insights for designing more robust tabular generative models. ③ **Technical**: We will release TabStruct, including the benchmark suite, the associated codebase, and all raw experimental results. This open-source library will enable researchers and practitioners to evaluate their models efficiently and comprehensively with a standardised framework.

2 TabStruct Benchmark Framework

Figure 1 provides an overview of the TabStruct framework. We first describe our problem setup (Section 2.1). Then we discuss the empirically effective structural prior of tabular data (Section 2.2), and the proposed methodology for quantifying structural fidelity (Section 2.3). Next, we detail the conventional evaluation dimensions employed in TabStruct (Section 2.4). Finally, we introduce the benchmark datasets (Section 2.5) and the benchmark generators (Section 2.6).

2.1 Problem Setup

We address the task of tabular data generation. Let $\mathcal{D} \coloneqq \{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, y^{(N)})\}$ represent a labelled tabular dataset consisting of N samples. For the i-th sample $\boldsymbol{x}^{(i)}$, $x_d^{(i)}$ denotes its d-th feature, and $y^{(i)}$ denotes the corresponding target. To simplify notation, we refer to the training split of the full dataset $\mathcal{D}$ as the reference data, denoted by $\mathcal{D}_{\text{ref}}$. The synthetic data produced by tabular data generators is denoted by $\mathcal{D}_{\text{syn}}$. The evaluation of tabular generative models is conducted by assessing the quality of $\mathcal{D}_{\text{syn}}$ across multiple dimensions. We further illustrate the setup in Appendix B.2.

2.2 Tabular Data Structure

The underlying structure of tabular data has long been an open research question (Kitson et al., 2023; Hollmann et al., 2025; Müller et al., 2022). For other modalities like textual data, it is natural to characterise their structure as autoregressive, guided by human knowledge (Yang, 2019). Therefore, pretraining paradigms aligned with the autoregressive structure, such as next-token prediction (Achiam et al., 2023), have proven successful in textual generative modelling. In contrast, heterogeneous tabular data does not naturally lend itself to human interpretation, making a structural prior for such data generally elusive.

Recent studies (Hollmann et al., 2025; Müller et al., 2022) on tabular foundation predictors have begun to shed light on the underlying structure of tabular data. Hollmann et al. (2025) introduces

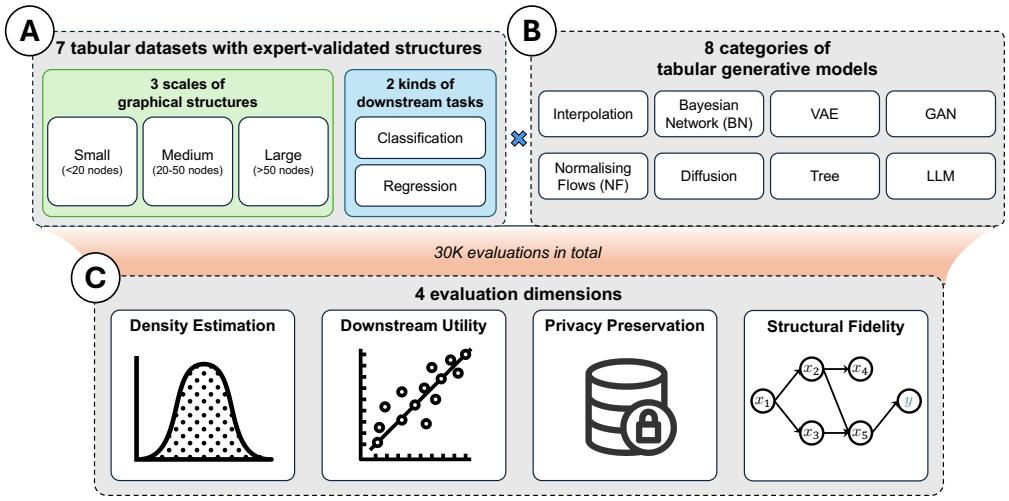

Figure 1: **The overview of the TabStruct evaluation framework. (A)** Given the graphical structures (i.e., structural causal models) validated by domain experts, we perform prior sampling on these graphs to generate a full dataset $\mathcal{D}$. **(B)** We train tabular generative models on the training split $\mathcal{D}_{\text{ref}} \subset \mathcal{D}$. We then generate synthetic data $\mathcal{D}_{\text{syn}}$ with the fitted models. **(C)** We evaluate the quality of synthetic data by comparing $\mathcal{D}_{\text{ref}}$ and $\mathcal{D}_{\text{syn}}$ across four dimensions.

TabPFN, a tabular foundation predictor pretrained on 100 million "synthetic" tabular datasets. These datasets are "synthetic" because they do not incorporate real-world semantics: they are produced with randomly constructed structural causal models (SCM). Remarkably, despite not being explicitly trained on any real-world dataset, TabPFN is able to outperform an ensemble of strong baseline predictors, which have been fine-tuned on each individual classification task. The exceptional performance of TabPFN suggests that the SCMs used to construct the pretraining datasets, despite lacking real-world semantics, effectively reflect the structural information encoded in real-world tabular data.

However, it is important to note that this does not imply that SCMs can fully represent the underlying structures of all tabular data. Instead, TabPFN demonstrates that the causal relationships between features, as modelled by SCMs, act as an empirically effective structural prior for a great proportion of real-world tabular data.

As the success of LLMs primarily stems from their ability to leverage the autoregressive nature of textual data, we argue that a robust tabular data generation process should be able to capture the unique causal structures within the tabular data. More specifically, generating data aligned with the causal structures in reference data could provide valuable insights into the open research question of how to effectively leverage the structural information inherent in tabular data.

2.3 Structural Fidelity

Using causal relationships as the structural prior for tabular data, we define the *structural fidelity* of a tabular generative model as the alignment between the causal structures in the reference data $\mathcal{D}_{\text{ref}}$ and the synthetic data $\mathcal{D}_{\text{syn}}$. Following prior benchmarks on causal discovery and inference (Spirtes et al., 2001; Tu et al., 2024), TabStruct evaluates structural fidelity at the level of the Markov equivalent class. At this level, causal structures are represented by completed partially directed acyclic graphs (CPDAGs). The causal structures of $\mathcal{D}_{\text{ref}}$ and $\mathcal{D}_{\text{syn}}$ are considered equivalent as long as they encode the same set of conditional independence relationships between features.

Fine-grained quantification of structural fidelity. Given the ground-truth causal structure of $\mathcal{D}_{\text{ref}}$, we can derive all conditional independence relationships between features programmatically (Figure 2). These conditional independence relationships are then tested on $\mathcal{D}_{\text{syn}}$ to investigate whether the synthetic data exhibits a Markov equivalent causal structure to the reference data. For each pair of features, the conditional independence test is formulated as a binary classification task,

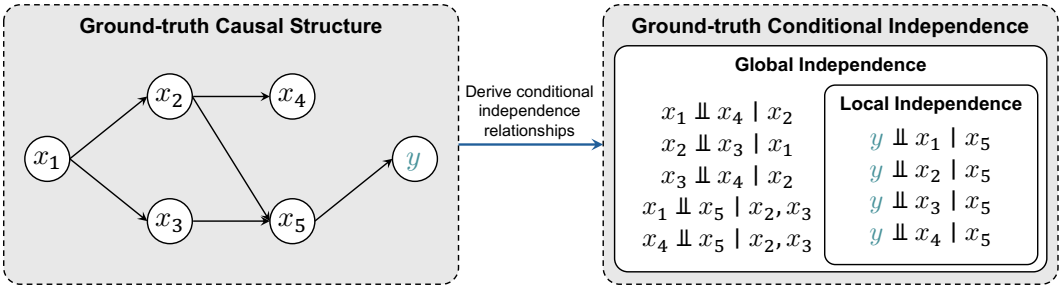

Figure 2: **An illustrative example for the quantification of structural fidelity.** Given the ground-truth causal structure, we first derive the conditional independence relationships between features. These relationships are then divided into two levels of granularity: global and local. The global set encompasses all conditional independence relationships across the entire feature set, whereas the local set includes only those relationships that are directly relevant to the target variable y. Next, we apply conditional independence tests on $\mathcal{D}_{\text{syn}}$ to examine the alignment of conditional independence relationships between features.

where 1 indicates independence and 0 indicates dependence. The balanced accuracy of the conditional independence tests on $\mathcal{D}_{\text{syn}}$ is then computed in order to quantify structural fidelity.

To provide a more fine-grained assessment of structural fidelity, we decompose structural fidelity into two complementary metrics: *global independence* and *local independence*. As illustrated in Figure 2, global independence evaluates all conditional independence relationships in the dataset, whereas local independence focuses only on those relationships relevant to the target variable y. Intuitively, local independence assesses how well the generator models the relationships between the target and the features, while global independence provides a comprehensive evaluation of the generator's ability to capture the overall structure of tabular data.

Rationales for CPDAG-level evaluation. TabStruct does not evaluate causal fidelity at the directed acyclic graph (DAG) level, as this would require an additional causal discovery method to determine the causal directions between features. Recovering causal directions is an inherently challenging task, and no existing causal discovery methods can guarantee perfect identification of causal directions (Zanga et al., 2022; Kaddour et al., 2022). This limitation is further illustrated in Section 3. Additionally, evaluating at the DAG level can introduce biases, as the results depend on the specific causal discovery method used. This issue is similar to the key limitation of "downstream utility", which is inherently biased by the choice of downstream tasks and predictors. To address this, TabStruct evaluates causal structures at the CPDAG level, reducing the risks associated with inaccurate or biased identification of causal directions.

2.4 Conventional Evaluation Approaches

Density estimation evaluates the mismatch between the marginal (i.e., low-order) or joint (i.e., high-order) distributions of reference and synthetic data (Hansen et al., 2023). A generator can trivially achieve high performance on low-order metrics by independently sampling from each feature's marginal distribution. While high-order metrics measure sample-level similarity, they still fail to explicitly demonstrate whether the synthetic data presents the same causal structures as reference data.

Following prior studies (Hansen et al., 2023; Shi et al., 2024; Zhang et al., 2023), we evaluate density estimation using four metrics of two categories: (i) Low-order: *Shape* and *Trend* (Wüst, 2011). Shape measures the synthetic data's ability to replicate each column's marginal density. Trend assesses its capacity to capture correlations between different columns. (ii) High-order: α-*precision* and β-*recall* (Alaa et al., 2022). α-*precision* quantifies the similarity between the reference and synthetic data, and β-*recall* assesses the diversity of the synthetic data.

Downstream utility measures the performance gap when substituting reference data with synthetic data in downstream tasks. This metric is inherently task-specific and susceptible to bias from the choice of downstream models and tasks. A parallel can be drawn to image generation, where Mixup (Psaroudakis & Kollias, 2022) augments training data with synthetic samples by interpolat-

ing between real samples. While Mixup improves downstream performance, it disrupts the spatial structure of images, resulting in synthetic samples that are generally visually unrealistic (Mumuni & Mumuni, 2022). This example shows that downstream utility, while useful for specific tasks, cannot serve as a holistic measure of a tabular data generator.

For all downstream tasks, we adopt the "train-on-synthetic, test-on-real" strategy (Xu et al., 2019). To mitigate the bias from downstream models, we evaluate downstream utility by averaging the performance of six representative downstream predictors, including three standard baselines: Logistic Regression (LR) (Cox, 1958), KNN (Fix, 1985) and MLP (Gorishniy et al., 2021); two tree-based methods: Random Forest (RF) (Breiman, 2001) and XGBoost (Chen & Guestrin, 2016); and a PFN method: TabPFN (Hollmann et al., 2025).

Privacy preservation primarily focuses on the trade-off between specific downstream tasks and privacy leakage (Margeloiu et al., 2024). Similar to downstream utility, this dimension is also highly task-dependent, making it susceptible to bias, and limiting its ability to provide a comprehensive assessment of the capability of tabular generative models.

We measure privacy preservation using two metrics: (i) *median Distance to Closest Record* (DCR) (Zhao et al., 2021), where a higher DCR indicates that synthetic data is less likely to be directly copied from the reference data; (ii) *Authenticity* (Alaa et al., 2022), where a higher score indicates that the generated samples are less likely to be mere replicas of the reference data.

2.5 Benchmark Datasets with Ground Truth Causal Structures

To accurately quantify structural fidelity, the reference data should be paired with ground-truth causal structures. Therefore, we construct benchmark datasets by leveraging structural causal models (SCMs) that have been validated by human experts (Scutari, 2011). Human validation ensures that the causal structures are realistic, increasing the likelihood that TabStruct's benchmark results can generalise to other real-world datasets without known causal structures (i.e., where structural fidelity cannot be directly evaluated). We note that this is a core distinction between TabStruct and prior studies (Tu et al., 2024; Hollmann et al., 2025): instead of relying on datasets without real-world semantics, TabStruct utilises reference data with expert-validated, realistic causal structures and mixed feature types.

We outline the process of building the reference datasets as follows. Firstly, we use ground-truth SCMs with realistic and expert-validated structures. Secondly, we perform prior sampling on these SCMs: root nodes are randomly initialised, and their values are propagated through the causal graph. A single sample is generated by recording the node values after propagation, with each propagation producing one sample. Thirdly, this process is repeated until sufficient samples are obtained. By following this procedure, we construct full datasets $\mathcal{D}$ with accessible and well-defined causal structures. We include both classification (Table 4) and regression (Table 5) datasets, and the detailed descriptions are in Appendix B.1.

2.6 Benchmark Generators

TabStruct includes nine existing tabular data generation methods of eight different categories: (i) a standard interpolation method SMOTE (Chawla et al., 2002); (ii) a structure learning method Bayesian Network (Qian et al., 2024); (iii) two Variational Autoencoders (VAE) based methods TVAE (Xu et al., 2019) and GOGGLE (Liu et al., 2023); (iv) a Generative Adversarial Networks (GAN) method CTGAN (Xu et al., 2019); (v) a normalising flow model Neural Spine Flows (NFLOW) (Durkan et al., 2019); (vi) a diffusion model TabDDPM (Kotelnikov et al., 2023); (vii) a tree-based method Adversarial Random Forests (ARF) (Watson et al., 2023); and (viii) a Large Language Model (LLM) based method GReaT (Borisov et al., 2023). In addition, we include $\mathcal{D}_{\text{ref}}$, where the reference data is directly used for evaluation. We provide full implementation details of benchmark generators in Appendix B.5.

3 Experiments

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

Table 1: **Benchmark results of nine tabular data generators on seven datasets with varying feature scales.** The results are grouped based on the tasks. For each group, we report the normalised mean ± std metric values across datasets. We also highlight the **First**, **Second** and **Third** best performances for each metric. Existing tabular generative models, including advanced neural networks, struggle to accurately capture the underlying structure of tabular data.

Generator	Density Estimation				Downstream Utility		Privacy Preservation		Structural Fidelity	
	Shape ↑	Trend ↑	α-precision ↑	β-recall ↑	Accuracy ↑	RMSE ↓	DCR ↑	Authenticity ↑	Local independence ↑	Global independence ↑
Classification tasks										
$\mathcal{D}_{\text{ref}}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	–	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
SMOTE	**0.94**$_{\pm 0.00}$	**0.93**$_{\pm 0.00}$	$0.82_{\pm 0.01}$	**1.00**$_{\pm 0.00}$	**95.71**$_{\pm 1.54}$	–	$0.11_{\pm 0.01}$	$0.05_{\pm 0.01}$	**74.02**$_{\pm 3.13}$	$35.39_{\pm 0.85}$
BN	**0.97**$_{\pm 0.00}$	**0.91**$_{\pm 0.00}$	**0.95**$_{\pm 0.01}$	**0.75**$_{\pm 0.01}$	**89.88**$_{\pm 1.01}$	–	$0.79_{\pm 0.03}$	$0.50_{\pm 0.01}$	$35.49_{\pm 3.42}$	$45.31_{\pm 0.79}$
TVAE	$0.83_{\pm 0.00}$	$0.75_{\pm 0.00}$	$0.70_{\pm 0.02}$	**0.66**$_{\pm 0.02}$	**94.57**$_{\pm 1.50}$	–	**0.94**$_{\pm 0.04}$	$0.60_{\pm 0.02}$	**65.96**$_{\pm 3.64}$	**64.29**$_{\pm 0.77}$
GOGGLE	$0.08_{\pm 0.02}$	$0.05_{\pm 0.01}$	$0.01_{\pm 0.01}$	$0.24_{\pm 0.02}$	$18.57_{\pm 1.16}$	–	**0.98**$_{\pm 0.03}$	**0.86**$_{\pm 0.01}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$
CTGAN	$0.72_{\pm 0.02}$	$0.74_{\pm 0.02}$	**0.89**$_{\pm 0.04}$	$0.52_{\pm 0.06}$	$76.74_{\pm 2.80}$	–	$0.82_{\pm 0.03}$	$0.74_{\pm 0.05}$	**55.58**$_{\pm 0.47}$	**50.95**$_{\pm 1.02}$
NFlow	$0.73_{\pm 0.01}$	$0.66_{\pm 0.01}$	$0.79_{\pm 0.03}$	$0.20_{\pm 0.03}$	$23.79_{\pm 3.02}$	–	$0.84_{\pm 0.04}$	**0.92**$_{\pm 0.01}$	$20.88_{\pm 4.40}$	$40.74_{\pm 1.14}$
TabDDPM	$0.39_{\pm 0.01}$	$0.37_{\pm 0.01}$	$0.24_{\pm 0.01}$	$0.22_{\pm 0.01}$	$33.90_{\pm 1.07}$	–	$0.77_{\pm 0.04}$	**0.83**$_{\pm 0.01}$	$5.63_{\pm 2.62}$	$23.69_{\pm 0.68}$
ARF	**0.97**$_{\pm 0.00}$	**0.90**$_{\pm 0.00}$	**0.92**$_{\pm 0.01}$	$0.61_{\pm 0.02}$	$59.21_{\pm 2.15}$	–	**0.85**$_{\pm 0.03}$	$0.65_{\pm 0.01}$	$32.59_{\pm 3.72}$	**46.40**$_{\pm 0.94}$
GReaT	$0.74_{\pm 0.01}$	$0.71_{\pm 0.01}$	$0.65_{\pm 0.02}$	$0.56_{\pm 0.02}$	$48.34_{\pm 1.57}$	–	$0.69_{\pm 0.03}$	$0.62_{\pm 0.01}$	$38.56_{\pm 3.15}$	$45.20_{\pm 0.77}$
Regression datasets										
$\mathcal{D}_{\text{ref}}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	–	$0.00_{\pm 0.01}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
SMOTE	**0.85**$_{\pm 0.00}$	**0.89**$_{\pm 0.00}$	$0.71_{\pm 0.00}$	**0.96**$_{\pm 0.00}$	–	**0.16**$_{\pm 0.02}$	$0.29_{\pm 0.01}$	$0.08_{\pm 0.01}$	**49.66**$_{\pm 6.67}$	**69.21**$_{\pm 5.36}$
BN	**0.86**$_{\pm 0.00}$	**0.77**$_{\pm 0.00}$	**0.90**$_{\pm 0.01}$	**0.73**$_{\pm 0.01}$	–	**0.10**$_{\pm 0.02}$	$0.22_{\pm 0.01}$	$0.30_{\pm 0.01}$	**71.05**$_{\pm 5.00}$	**77.51**$_{\pm 2.96}$
TVAE	$0.70_{\pm 0.01}$	$0.61_{\pm 0.01}$	$0.61_{\pm 0.03}$	$0.63_{\pm 0.02}$	–	$0.20_{\pm 0.27}$	**0.66**$_{\pm 0.04}$	$0.50_{\pm 0.02}$	**44.10**$_{\pm 4.15}$	**52.52**$_{\pm 2.31}$
GOGGLE	$0.30_{\pm 0.06}$	$0.24_{\pm 0.01}$	$0.30_{\pm 0.09}$	$0.28_{\pm 0.03}$	–	$0.83_{\pm 0.29}$	$0.32_{\pm 0.06}$	**0.80**$_{\pm 0.02}$	$21.31_{\pm 1.47}$	$22.33_{\pm 0.74}$
CTGAN	$0.54_{\pm 0.04}$	$0.51_{\pm 0.02}$	**0.73**$_{\pm 0.09}$	$0.44_{\pm 0.08}$	–	$0.37_{\pm 0.55}$	$0.31_{\pm 0.03}$	**0.76**$_{\pm 0.05}$	$9.21_{\pm 6.09}$	$11.96_{\pm 4.11}$
NFlow	$0.74_{\pm 0.01}$	$0.62_{\pm 0.01}$	$0.67_{\pm 0.04}$	$0.52_{\pm 0.05}$	–	$0.31_{\pm 0.08}$	**0.58**$_{\pm 0.04}$	$0.72_{\pm 0.04}$	$37.62_{\pm 3.62}$	$23.68_{\pm 3.70}$
TabDDPM	$0.20_{\pm 0.02}$	$0.25_{\pm 0.01}$	$0.31_{\pm 0.00}$	$0.18_{\pm 0.01}$	–	**0.02**$_{\pm 0.03}$	**0.68**$_{\pm 0.03}$	**0.87**$_{\pm 0.00}$	$38.97_{\pm 0.72}$	$10.20_{\pm 5.33}$
ARF	**0.84**$_{\pm 0.00}$	**0.79**$_{\pm 0.00}$	**0.94**$_{\pm 0.01}$	$0.53_{\pm 0.02}$	–	$0.18_{\pm 0.26}$	$0.36_{\pm 0.01}$	$0.66_{\pm 0.02}$	$33.64_{\pm 3.50}$	$31.56_{\pm 2.32}$
GReaT	$0.67_{\pm 0.01}$	$0.67_{\pm 0.01}$	$0.69_{\pm 0.03}$	**0.64**$_{\pm 0.03}$	–	$0.21_{\pm 0.25}$	$0.39_{\pm 0.03}$	$0.51_{\pm 0.02}$	$38.42_{\pm 5.09}$	$38.66_{\pm 3.38}$

Table 2: **Correlation between ranks of different metrics.** The relatively higher correlation between Accuracy/RMSE and local independence demonstrates that a generator can achieve high downstream utility by prioritising the conditional independence relationships directly relevant to the target variable while overlooking the global structure.

	Local independence	Global independence
Accuracy	**0.90**	0.57
RMSE	**0.77**	0.33

(90% training and 10% validation). For classification datasets, stratification is preserved during data splitting. We provide detailed descriptions of data splitting in Appendix B. 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_{syn} samples. For classification, the synthetic data preserves the stratification of reference data. Since a small N_{syn} may not yield robust results of model performance (Margeloiu et al., 2024), we conduct a proof-of-concept experiment (see Appendix C for more details) and empirically set $N_{\text{syn}} = 3N_{\text{ref}}$ as the saturation point where further increases in N_{syn} have negligible impact on evaluation results.

Aggregation of evaluation results. The reported results are averaged by default over 10 runs on the test sets. When aggregating results across datasets, we use the average distance to the minimum (ADTM) metric via affine renormalisation between the top-performing and worse-performing models (Grinsztajn et al., 2022; McElfresh et al., 2024; Hollmann et al., 2025; Margeloiu et al., 2024; Jiang et al., 2024). To aggregate different metrics within the same evaluation dimension, we compute their average. For downstream utility, evaluation results are averaged over six downstream predictors to mitigate the bias from specific predictors.

3.1 Generator Performance in Learning Tabular Data Structure

Downstream utility is not the golden standard for tabular generative modelling. In prior studies (Appendix A), downstream utility is often considered as the core evaluation dimension. From this perspective, a generator is considered effective if its synthetic data achieves high performance in downstream tasks. However, as discussed in Section 2.4, downstream utility inherently biases evaluation towards relationships between the target variable and the features, thus overlooking the

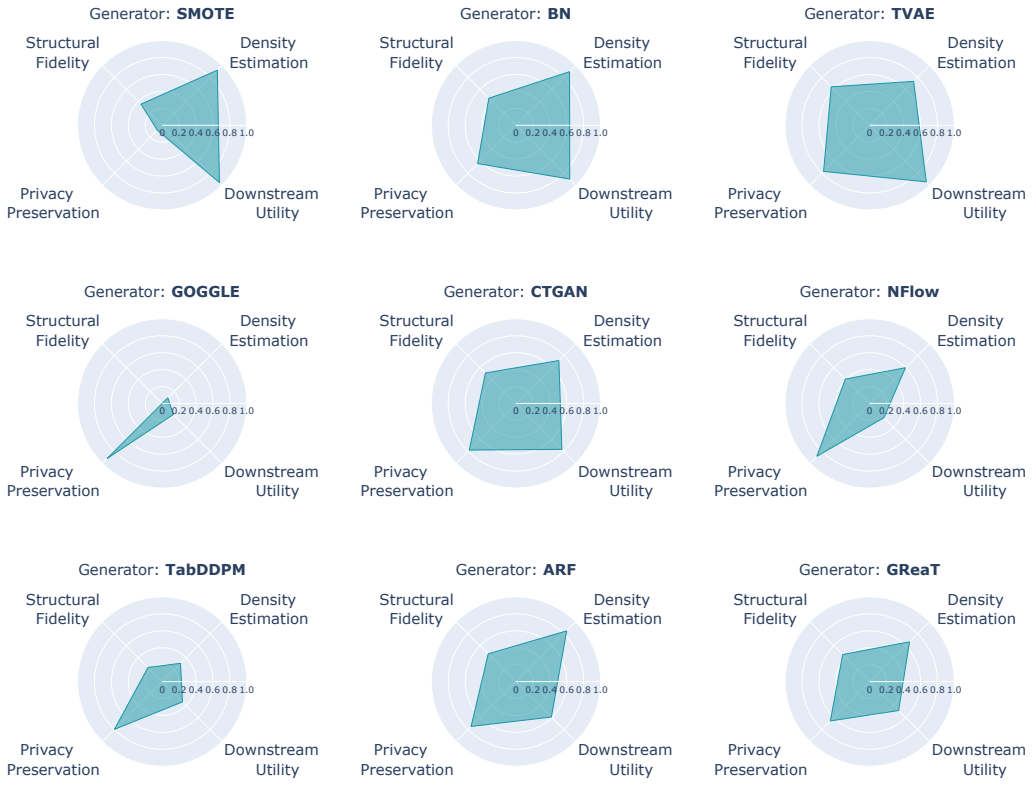

Figure 3: **Summarised comparison of nine tabular data generators across four evaluation dimensions.** The results reveal that excelling in conventional evaluation dimensions does not ensure the model's ability to capture the underlying data structure. Learning the underlying data structure remains challenging for tabular generative modelling.

relationships between features. Table 1 and Table 2 quantitatively demonstrate this limitation. The rankings of downstream utility are strongly correlated with local independence but exhibit much weaker correlation with global independence. This indicates that a generator can achieve high downstream utility by prioritising local independence at the expense of global independence. For instance, SMOTE achieves the highest downstream utility and local independence in classification tasks, but it performs poorly in global independence. This suggests that SMOTE focuses narrowly on structures relevant to the target variable while neglecting inter-feature relationships. Therefore, a generator should not be deemed effective solely based on downstream utility, as it can overlook the broader structural information encoded in the data.

Structural fidelity presents consistent challenges across tasks and generators. Table 1 shows a notable gap in structural fidelity between reference data ($\mathcal{D}_{\text{ref}}$) and synthetic data ($\mathcal{D}_{\text{syn}}$) across both classification and regression tasks. For instance, in classification tasks, the highest local independence achieved is 74.02% (SMOTE), indicating the smallest performance gap relative to $\mathcal{D}_{\text{ref}}$ is over 25%. Global independence shows an even large performance gap of 35% between $\mathcal{D}_{\text{ref}}$ and $\mathcal{D}_{\text{syn}}$. In contrast, the smallest gaps between $\mathcal{D}_{\text{ref}}$ and $\mathcal{D}_{\text{syn}}$ in statistical fidelity and downstream utility remain consistently below 10%. The underperformance in structural fidelity also exists in regression datasets. These results underline the consistent challenges faced by existing tabular generative models in capturing the underlying structure of tabular data.

Existing structure learning methods struggle with tabular data generation. While Bayesian Network (BN) exhibit relatively strong performance in structural fidelity, their success is unsurprising – the reference datasets are constructed with SCMs that perfectly align with the required

assumptions of the causal discovery methods employed in BN (i.e., causal Markov assumption, causal sufficiency and causal faithfulness). Despite this advantage, the gap between $\mathcal{D}_{\text{ref}}$ and $\mathcal{D}_{\text{syn}}$ remains notable for BN. For instance, in classification tasks, the global independence gap exceeds 50% compared to $\mathcal{D}_{\text{ref}}$. This demonstrates the limitations of existing structure learning methods in recovering perfect causal structures from observed data alone. Such findings are consistent with previous research (Tu et al., 2024), which reveals that current causal discovery methods struggle with datasets containing more than 10 features. In TabStruct, we employ realistic SCMs, with the number of features ranging from 7 to 223. Consequently, BNs perform less effectively despite having an objective function for explicit structure learning. This further justifies our choice to evaluate structural fidelity at the CPDAG level rather than the DAG level.

Baseline models can outperform complex models in structural fidelity. Interestingly, simple baseline models such as SMOTE and TVAE exhibit competitive performance in structural fidelity. For global independence, Table 1 shows that TVAE consistently ranks among the top-3 across both classification and regression datasets. We note that TVAE does not possess explicit advantages as a structure learning method, indicating that variational autoencoders remain effective models for capturing feature relationships in tabular data.

All evaluation dimensions are complementary, rather than interchangeable. As demonstrated in Figure 3, no single metric is fully indicative of all other metrics. This highlights the necessity for researchers and practitioners to select evaluation metrics that are aligned with the specific objectives of their tasks, rather than relying on a single dimension to evaluate the performance of tabular generative models. For instance, in regression tasks, BN excels in capturing the underlying tabular data structure, suggesting that its synthetic data can facilitate more accurate causal inference compared to TabDDPM. However, if a practitioner's primary concern is downstream performance, TabDDPM would be the preferred choice. Similarly, SMOTE consistently achieves competitive results in downstream utility across tasks. Nevertheless, SMOTE introduces high risks of privacy leakage, which may be unacceptable for certain sensitive scenarios.

Limitations and future work. While TabStruct provides valuable insights, we acknowledge several directions for future exploration. One primary limitation of TabStruct is the scope of datasets. As discussed in Section 2.3, due to the limitations of existing causal discovery methods, TabStruct relies on datasets with expert-validated causal graphs. However, most real-world tabular datasets lack ground truth causal graphs, making it challenging to assess structural fidelity in such cases. To address this, we plan to develop new evaluation metrics that enable more flexible assessment of structural fidelity in real-world datasets, where ground truth causal structures are unavailable.

4 Conclusion

We introduce TabStruct, a novel benchmark framework for the holistic evaluation of tabular generative models. TabStruct positions structure fidelity as a core aspect of model performance, and quantifies it at the Markov equivalent class level by evaluating the conditional independence relationships between features. Additionally, TabStruct provides conventional evaluation metrics while considering their interplay between structural fidelity. Our experimental results demonstrate that conventional evaluation dimensions fail to provide a holistic view of model performance, and the existing tabular generative models still struggle to effectively capture the underlying structure of tabular data. The insights from TabStruct and the open-source library can guide researchers in developing next-generation tabular data generators, and help practitioners select appropriate models for their tasks.

Acknowledgements

The authors would like to express their gratitude to Prof. Carl Henrik Ek for insightful discussions on structure learning, and to Prof. Ferenc Huszár and Dr. Ruibo Tu for their enlightening perspectives on causal machine learning. NS and MJ acknowledge the support of the U.S. Army Medical Research and Development Command of the Department of Defense; through the FY22 Breast Cancer Research Program of the Congressionally Directed Medical Research Programs, Clinical Research Extension Award GRANT13769713. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense.

REFERENCES

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.

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*, pp. 290–306. PMLR, 2022.

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, pp. 12180–12188, 2023.

Ankur Ankan and Johannes Textor. pgmpy: A python toolkit for bayesian networks. *Journal of Machine Learning Research*, 25(265):1–8, 2024. URL `http://jmlr.org/papers/v25/23-0487.html`.

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.

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.

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.

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

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.

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*, pp. 785–794, 2016.

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

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

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.

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

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*, pp. arXiv–2402, 2024.

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

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

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.

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.

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.

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.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007. doi: 10.1109/MCSE.2007.55.

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.

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.

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.

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

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

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. URL `http://jmlr.org/papers/v18/16-365.html`.

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.

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.

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.

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.

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.

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

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.

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

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

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.

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*, pp. 2367–2375, 2022.

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.

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

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

Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. Tabdiff: a multi-modal diffusion model for tabular data generation. *arXiv preprint arXiv:2410.20626*, 2024.

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

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.

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*, pp. 5357–5375. PMLR, 2023.

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

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

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

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

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.

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.

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

Appendix: How Well Does Your Tabular Generator Learn the Structure of Tabular Data?

Table of Contents

A SUMMARY OF EXISTING BENCHMARKS

Table 3 presents a comparative analysis of TabStruct against existing benchmarks for evaluating tabular generative models. TabStruct is the only benchmark that covers four evaluation metrics, including density estimation, downstream utility, privacy preservation, and structural fidelity. Moreover, it is the only comprehensive benchmark, supporting all eight generator types and offering a more holistic overview of existing tabular generative models.

Table 3: **Comparison of evaluation scopes between TabStruct and existing benchmarks.** (a) TabStruct introduces a novel benchmark for the holistic evaluation of tabular generative models, with a particular emphasis on capturing the underlying structure of tabular data. (b) TabStruct stands out as the only benchmark that covers eight generator categories.

(a) Evaluation Metrics

Benchmark source	**Density Estimation**		**Downstream Utility**		**Privacy Preservation**	**Structural Fidelity**
	Low-order	**High-order**	**Classification**	**Regression**		
Xu et al. (2019)	✔	✔	✔	✔	✗	✗
Durkan et al. (2019)	✔	✗	✗	✗	✗	✗
Watson et al. (2023)	✗	✗	✔	✗	✗	✗
Liu et al. (2023)	✔	✔	✔	✗	✗	✗
Borisov et al. (2023)	✔	✔	✔	✔	✔	✗
Kotelnikov et al. (2023)	✔	✗	✔	✔	✔	✗
Hansen et al. (2023)	✔	✔	✔	✗	✗	✗
Zhang et al. (2023)	✔	✔	✔	✔	✔	✗
Tu et al. (2024)	✔	✔	✗	✗	✗	✔
Shi et al. (2024)	✔	✔	✔	✔	✔	✗
TabStruct (Ours)	✔	✔	✔	✔	✔	✔

(b) Generator Category Coverage

Benchmark source	Interpolation	BN	GAN	VAE	NF	Tree	Diffusion	LLM	# Generators
Xu et al. (2019)	✗	✔	✔	✔	✗	✗	✗	✗	7
Durkan et al. (2019)	✗	✗	✗	✔	✔	✗	✗	✗	10
Watson et al. (2023)	✗	✗	✔	✔	✗	✔	✗	✗	6
Liu et al. (2023)	✗	✔	✔	✔	✔	✗	✗	✗	7
Borisov et al. (2023)	✗	✗	✔	✔	✗	✗	✗	✔	4
Kotelnikov et al. (2023)	✔	✗	✔	✔	✗	✗	✔	✗	6
Hansen et al. (2023)	✗	✔	✔	✔	✔	✗	✔	✗	5
Zhang et al. (2023)	✔	✗	✔	✔	✗	✗	✔	✔	9
Tu et al. (2024)	✗	✗	✔	✔	✗	✗	✔	✔	7
Shi et al. (2024)	✗	✗	✔	✔	✗	✗	✔	✔	9
TabStruct (Ours)	✔	✔	✔	✔	✔	✔	✔	✔	**9**

B REPRODUCIBILITY

B.1 REFERENCE DATASETS

To ensure that the causal structures of reference datasets are realistic, we select seven publicly available datasets from bnlearn (Scutari, 2011). Each dataset is accompanied by a ground-truth structural causal model (SCM) validated by human experts. Furthermore, to obtain generalisable benchmark results, we select datasets from diverse domains, and they are across three different levels of structure scales (i.e., small, medium and large).

In contrast, the only prior benchmark that addresses structural fidelity, CauTabBench (Tu et al., 2024), does not utilise SCMs validated by human experts. Additionally, the dimensionality of their datasets is fixed at 10 numerical features. In summary, TabStruct is one of the first to offer a comprehensive benchmark for tabular generative models, leveraging datasets with realistic causal structures, mixed feature types, and more than 10 features.

Table 4: Details of four classification datasets with realistic structures.

Dataset	Domain	Structure scale	# Samples	# Features	# Numerical	# Categorical	# Classes	# Samples per class (Min)	# Samples per class (Max)
Sangiovese	Agriculture	Small (<20 nodes)	2,000	15	14	1	16	108	146
Insurance	Economics	Medium (20–50 nodes)	2,000	27	0	27	4	38	1,122
Hailfinder	Meteorology	Large (>50 nodes)	2,000	56	0	56	3	519	880
ANDES	Education	Large (>50 nodes)	2,000	223	0	223	2	830	1,170

Table 5: Details of three regression datasets with realistic structures.

Dataset	Domain	Structure scale	# Samples	# Features	# Numerical	# Categorical
Healthcare	Medicine	Small (<20 nodes)	2,000	7	7	3
MEHRA	Meteorology	Medium (20–50 nodes)	2,000	24	20	4
ARTH150	Life Science	Large (>50 nodes)	2,000	107	107	0

B.2 DATA SPLITTING

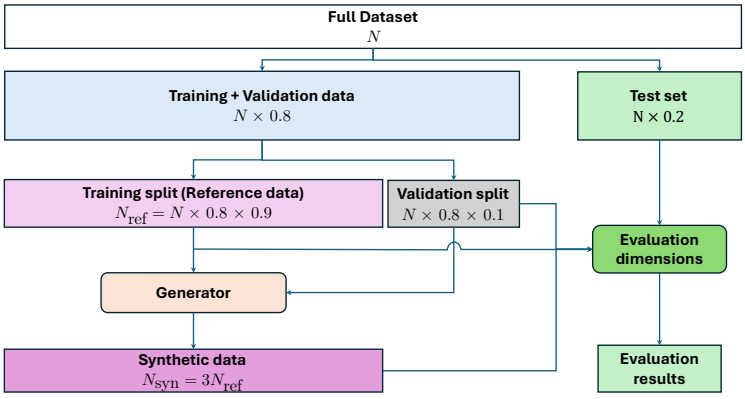

Figure 4: Data splitting strategies for benchmarking tabular data generators.

B.3 DATA PREPROCESSING

Following the procedures presented in prior work (McElfresh et al., 2024; Grinsztajn et al., 2022), we perform preprocessing in two steps. Firstly, we impute the missing values with the mean value for numerical features and the most mode value for categorical features. We then compute the required statistics with training data and then transform it. For categorical features, we convert them into one-hot encodings. For numerical features, we perform Z-score normalisation. We compute each feature's mean and standard deviation in the training data and then transform the training samples to

have a mean of zero and a variance of one for each feature. Finally, we apply the same transformation to the validation and test data before conducting evaluations.

B.4 Implementation of conditional independence test

For categorical datasets, we employ the chi-square independence test (McHugh, 2013); for numerical datasets, we use partial correlation based on the Pearson correlation coefficient (Baba et al., 2004); and for mixed datasets, we utilise a residualisation-based conditional independence test (Ankan & Textor, 2023; Li & Shepherd, 2010; Muller & Peterson, 1984). We implement these conditional independence tests using pgmpy (Ankan & Textor, 2024), an open-source Python library for causal and probabilistic inference. The significance level is set to 0.01 by default (i.e., the p-value is 0.01).

B.5 Implementations of Benchmark Generators

SMOTE is an interpolation-based oversampling method (Chawla et al., 2002). It generates new samples by interpolating between real samples. We use the open-source implementation of SMOTE from Imbalanced-learn (Lemaître et al., 2017), setting the number of neighbours k within the range $\{1, 3, 5\}$. When applicable, we use the default value for nearest neighbours (i.e., $k = 5$).

Bayesian Network (BN) is a probabilistic graphical model used to represent and reason about the dependence relationships between features (Qian et al., 2024; Hansen et al., 2023). It consists of two main components: (i) a causal discovery model to construct a directed acyclic graph (DAG), where features and the target serve as nodes, and their dependencies are represented as edges; (ii) a parameter estimation mechanism to quantify the dependence relationships. Following Hansen et al. (2023), the causal discovery method is selected from Hill Climbing Search (Koller, 2009), the Peter-Clark algorithm (Koller, 2009), and Chow-Liu or Tree-augmented Naive Bayes (Chow & Liu, 1968; Friedman et al., 1997). We then build the parametrised BN using maximum likelihood estimation.

TVAE is a variational autoencoder (VAE) designed for tabular data (Xu et al., 2019). TVAE employs mode-specific normalisation to handle the complex distributions of numerical features. To address the class imbalance problem, TVAE conditions on specific categorical features during generation.

GOGGLE is a VAE-based tabular data generator designed to model the dependence relationships between features (Liu et al., 2023). GOGGLE proposes to learn an adjacency matrix to model the dependence relationships between features. However, TabStruct and prior benchmarks (Margeloiu et al., 2024; Zhang et al., 2023; Shi et al., 2024) all show that the downstream utility of GOGGLE is limited. We hypothesise that this is because of the challenge of learning accurate structures of tabular data. The inherent structure learning mechanism in GOGGLE fails to capture accurate conditional independence relationships between features, it could thus lead to poor-quality synthetic data, even if the model attempts to explicitly model the relationships between features like GOGGLE.

CTGAN is a conditional generative adversarial network (GAN) designed for tabular data (Xu et al., 2019). CTGAN leverages PacGAN (Lin et al., 2018) framework to mitigate mode collapse. In addition, CTGAN employs the same mode-specific normalisation technique as TVAE.

NFlow is a normalisation flow model designed for tabular data generation (Durkan et al., 2019). NFlow incorporates neural splines as a drop-in replacement for affine or additive transformations in coupling and autoregressive layers.

TabDDPM is a diffusion-based model for tabular data generation (Kotelnikov et al., 2023). TabDDPM introduces two core diffusion processes: (i) Gaussian noise for numerical features and (ii) multinomial diffusion with categorical noise for categorical features. TabDDPM directly concatenates numerical and categorical features as the input and output of the denoising function.

ARF is a tree-based model for tabular data generation (Watson et al., 2023). ARF employs a recursive adaptation of unsupervised random forests for joint density estimation by iteratively refining synthetic data distributions using adversarial training principles.

GReaT leverages large language models (LLMs) to generate synthetic tabular data (Borisov et al., 2023). GReaT converts each sample into a sentence and fine-tunes the language model to capture

the sentence-level distributions. Additionally, GReaT shuffles the order of features to mitigate the permutation variance in sentence-level distributions.

B.6 Software and Computing Resources

Software implementation. *(i) For generators:* We implemented SMOTE with Imbalanced-learn (Lemaître et al., 2017), an open-source Python library for imbalanced datasets with an MIT licence. For other benchmark generators, we used their open-source implementations in Synthcity (Qian et al., 2024), a library for generating and evaluating synthetic tabular data with an Apache-2.0 license. *(ii) For downstream predictors:* We implemented TabPFN with its open-source implementation (`https://github.com/automl/TabPFN`). We implemented the other five downstream predictors (i.e., Logistic Regression, KNN, MLP, Random Forest and XGBoost) with their open-source implementation in scikit-learn (Pedregosa et al., 2011), an open-source Python library under the 3-Clause BSD license. *(iii) For result analysis and visualisation:* All numerical plots and graphics have been generated using Matplotlib 3.7 (Hunter, 2007), a Python-based plotting library with a BSD licence. The icons for downstream tasks are from `https://icons8.com/`.

We ensure the consistency and reproducibility of experimental results by implementing a uniform pipeline using PyTorch Lightning, an open-source library under an Apache-2.0 licence. We further fixed the random seeds for data loading and evaluation throughout the training and evaluation process. This ensured that TabEBM and all benchmark models were trained and evaluated on the same set of samples. The experimental environment settings, including library dependencies, are specified in the open-source library for reference and reproduction purposes.

Computing Resources. All the experiments were conducted on a machine equipped with an NVIDIA A100 GPU with 40GB memory and an Intel(R) Xeon(R) CPU (at 2.20GHz) with six cores. The operating system used was Ubuntu 20.04.5 LTS.

C RATIONALES FOR SAMPLE SIZE OF SYNTHETIC DATA

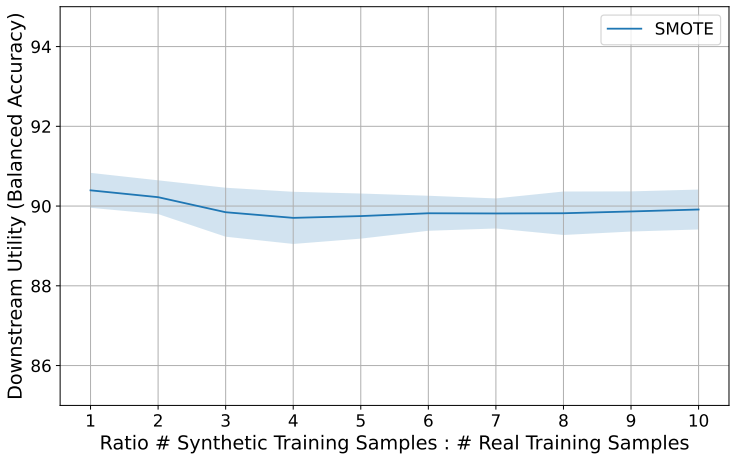

Figure 5: **Downstream utility vs. different ratios between the number of synthetic data and reference data ($N_{\text{syn}} : N_{\text{ref}}$).** On the "Hailfinder" dataset, as N_{syn} increases, the evaluation results become saturated. Specifically, the range of balanced accuracy varies by less than 0.3% when the ratio increases from $N_{\text{syn}} : N_{\text{ref}} = 3 : 1$ to $N_{\text{syn}} : N_{\text{ref}} = 10 : 1$. Therefore, we set $N_{\text{syn}} = 3N_{\text{ref}}$ in all experiments to ensure stable evaluation results.