SyntheRela: A Benchmark For Synthetic Relational Database Generation

Anonymous authors
Paper under double-blind review

Abstract

Synthesizing relational databases has started to receive more attention from researchers, practitioners, and industry. The task is more difficult than synthesizing a single table due to the added complexity of relationships between tables. For the same reasons, benchmarking methods for synthesizing relational databases introduces new challenges. Our work is motivated by a lack of an empirical evaluation of state-of-the-art methods and by gaps in the understanding of how such an evaluation should be done. We review related work on relational database synthesis, common benchmarking datasets, and approaches to measuring the fidelity and utility of synthetic data. We combine best practices, a novel robust detection metric, and a novel approach to evaluating utility with graph neural networks into a benchmarking tool. We use this benchmark to compare 6 open-source methods over 8 real-world databases, with a total of 39 tables. The open-source SyntheRela benchmark is available on GitHub with a public leaderboard.
Data & Code: `Anonymized`
Leaderboard: `Anonymized`

1 Introduction

Synthesizing relational databases - generating relational databases that preserve the characteristics of the original databases - is an emerging field. It promises several benefits, from protecting privacy to addressing data scarcity, while preserving the complexity and dependencies present in the original databases. This makes it attractive for healthcare (Appenzeller et al., 2022), finance (Assefa et al., 2020), and education (Bonnéry et al., 2019), where accessing and utilizing data can be challenging due to privacy concerns, data scarcity, or biases (Ntoutsi et al., 2020; Rajpurkar et al., 2022).

The foundations of synthesizing relational databases were laid by the Synthetic Data Vault (Patki et al., 2016). Recently, several deep learning methods have been proposed (Gueye et al., 2023; Li et al., 2024; Mami et al., 2022; Xu et al., 2023; Canale et al., 2022; Solatorio & Dupriez, 2023; Pang et al., 2024; Hudovernik, 2024). The field has also received attention from the industry, with several commercial tools now available and with Google, Amazon, and Microsoft integrating them into their cloud services (Gretel.ai, 2024).

Although there are several packages for evaluating the quality of synthetic tabular data, only the SDMetrics package (Patki et al., 2016) provides some support for the evaluation of synthetic relational databases. As such, the field lacks not only an empirical comparison of available methods but also an understanding of how such an evaluation should be done. We address this gap with an evaluation methodology that combines established evaluation metrics (Section 2.2), best practices, sampling procedures, and real-world relational databases (Table 1). We also propose a **detection-based metric C2ST-Agg** (Section 3.1) specialized for relational databases and use ML explainability to diagnose issues with synthetic data generation methods, and a novel approach to evaluate the utility of synthetic relational databases with **relational deep learning utility** (Section 3.2).

We implement the methodology in **SyntheRela**, a benchmark and evaluation tool that is available as an open source package and can be easily extended with new metrics and datasets (see Appendix B). Finally, we

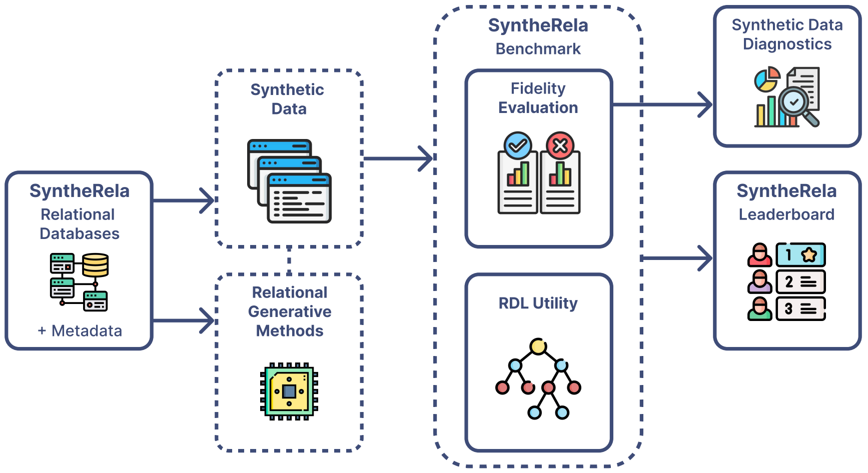


Figure 1: **SyntheRela** enables the evaluation of synthetic relational database generation methods using real-world relational databases. SyntheRela supports the evaluation of fidelity using established metrics and a relational detection-based metric, C2ST-Agg. It also assesses utility via relational deep learning (RDL) utility. The benchmark provides synthetic data diagnostics using ML explainability and a leaderboard for comparing method performance.

use the benchmark to evaluate current state-of-the-art methods (Section 2.1) over several relational databases. This is the first comprehensive evaluation and comparison of methods for the synthesis of relational databases and provides valuable insights into their ability to synthesize relational aspects of the data (Section 4). We release the code of our benchmark under an open source license alongside a public leaderboard.

2 Related Work

2.1 Methods for Synthesizing Relational Databases

In this work, we focus on relational databases — a collection of tables linked by primary and foreign keys. We distinguish this from synthesizing tabular data (a single-table), which is a special case and an even more active field (Borisov et al., 2022; Qian et al., 2023b). Existing methodologies for generating synthetic relational data can be broadly categorized into two main approaches: neural network–based methods and marginal-based methods. We provide a concise summary of each below.

Patki et al. (2016) are the first to propose a method for relational database synthesis - The Synthetic Data Vault (SDV), which employs Gaussian copulas with a hierarchical modeling algorithm and does not fit into either of these two main categories.

Neural network–based methods aim to preserve data fidelity and utility across arbitrary schemas. Row Conditional-TGAN (RCTGAN) (Gueye et al., 2023) and Incremental Relational Generator (IRG) (Li et al., 2024) are based on GANs. The Realistic Relational and Tabular Transformer (REaLTabFormer) (Solatorio & Dupriez, 2023) and Composite Generative Models (Canale et al., 2022) are based on transformers. The work of Mami et al. (2022) is based on graph variational autoencoders, while Xu et al. (2023) propose a framework for synthesizing many-to-many datasets using random graphs. Pang et al. (2024) propose ClavaDDPM, a method based on classifier-guided diffusion models. Hudovernik (2024) proposes RGCLD, a method based on graph neural networks (GNNs) and conditional diffusion models. Recently, Tiwald et al. (2025) propose TabularARGN - a framework for tabular and relational data synthesis based on autoregressive tabular models.

Marginal-based methods, by contrast, involve taking low-dimensional measurements of the dataset, such as marginal queries, and adding calibrated statistical noise to ensure DP. While well established in the single-table setting (Zhang et al., 2017; McKenna et al., 2022; Cai et al., 2021), extensions to relational data are more recent. PrivLava (Cai et al., 2023) models inter-table dependencies using graphical models with latent variables. MARE (Kapenekakis et al., 2024) targets medical relational datasets, such as electronic health records, with temporal and sequential characteristics. More recently, Alimohammadi et al. (2025) proposed a method for adapting single-table DP generators to relational data which models foreign keys by learning a bi-adjacency matrix. Finally, PrivPetal(Cai et al., 2025) synthesizes relational data by modeling flattening relational databases in an efficient manner. These approaches generally prioritize privacy and query accuracy over flexibility and general fidelity.

Our evaluation in this work will focus on neural network-based methods for the generation of synthetic relational data. This focus is driven by the fact that their implementations are generally applicable to generating synthetic data for arbitrary relational schemas. In contrast, marginal-based methods often include implementations tailored towards specific datasets with predefined query workloads. We provide a more detailed overview of related work in Appendix A.

2.2 Metrics for Evaluating Synthetic Data

The quality of synthetic tabular data and relational databases is evaluated primarily based on two criteria, *fidelity* and *utility.* Fidelity measures the degree of similarity between synthetic and real data in terms of its properties, while utility measures how well synthetic data can be used in place of real data when the data are part of a downstream task, for example, predictive modeling (Hansen et al., 2023).

We further divide fidelity metrics into *statistical*, *distance-based*, and *detection-based* metrics. The utility of synthetic data is typically assessed with the train-on-synthetic evaluate-on-real methods (Beaulieu-Jones et al., 2019).

Another dimension of evaluation metrics for relational data is granularity. The most common are *single-column* metrics that evaluate the marginal distributions, *two-column* metrics that evaluate bivariate distributions, *single-table* metrics that evaluate tables, and *multi-table* metrics that evaluate the relational aspects.

Statistical fidelity methods are typically used to assess marginal distributions, sometimes bivariate distributions. The most commonly used methods are the Kolmogorov-Smirnov test and the $\mathcal{X}^2$ test for numerical and categorical variables, respectively. For relational data, cardinality shape similarity is used, where for each parent row the number of child rows is calculated. This yields a numerical distribution for both real and synthetic data, on which a Kolmogorov-Smirnov test is performed.

Similar to statistical fidelity, **distance-based fidelity** is typically used to assess the quality of marginal distributions. However, some distance metrics also assess entire tables. Commonly used distance-based methods are total variation distance, Kullback-Leibler divergence, Jensen-Shannon distance, Wasserstein distance, maximum mean discrepancy, and pairwise correlation difference. To evaluate inter-table relationships Pang et al. (2024) use k-hop similarity, where they compute correlations between tables at distance k (0-hop refers to columns within the same table, 1-hop refers to columns in tables directly connected via a foreign key, etc.). Unlike statistical methods, reports of distance-based fidelity do not include hypothesis testing or any other quantification of uncertainty. This is an issue both when evaluating a method and when comparing two methods. In the former, a method can achieve a seemingly high distance that is in a high probability region when taking into account the sampling distribution. In the latter, a seemingly large difference between the two methods can be explained away by the variance of the sampling distribution.

The basic idea of **detection-based fidelity** is to learn a model that can discriminate between real and synthetic data. The detection-based metric can be interpreted as a null-hypothesis test for comparing two distributions (two sample testing) with classification accuracy as a proxy Kim et al. (2021). The classifier serves as a map from high-dimensional data to a one-dimensional test statistic. In machine learning literature, this is referred to as a classifier two sample test (C2ST) (Lopez-Paz & Oquab, 2017). If the model can achieve better-than-random predictive performance, this indicates that there are some patterns that identify synthetic data. Zein & Urvoy (2022) show that using discriminative models can highlight the differences between real and synthetic tabular data.

The most common detection-based metric is logistic detection (LD) (Gueye et al., 2023; Solatorio & Dupriez, 2023; Li et al., 2024; Pang et al., 2024), where a logistic regression model is used for discrimination. An extended version of LD known as parent-child logistic detection (P-C LD) is used to evaluate relational databases. P-C LD applies LD to denormalized pairs of synthetic parent and child tables, assessing the preservation of parent-child relationships. A serious issue with denormalization is that it may introduce correlation between rows, breaking the *i.i.d.* assumption. This results in an over-performance of the discrimintative model and in underestimating the quality of the method for synthesizing relational data. It also makes it impossible to set a detection threshold for testing fidelity (for example, accuracy would be greater than 50% even if both datasets were from the same data generating process). For these reasons, we do not consider P-C detection.

Note that logistic regression is unable to capture interactions between columns unless these interactions are explicitly included as features. This implies a lenient evaluation of the state-of-the-art methods (we demonstrate this empirically in Appendix D.2). Tree-based ensemble models are a better alternative, which is also suggested by the findings of Zein & Urvoy (2022) for tabular data.

The **utility** of synthetic data is most commonly measured with machine learning efficacy (ML-E) - comparing the hold-out performance of a predictive model trained on the original data with a predictive model trained on synthetic data (Canale et al., 2022; Li et al., 2024; Mami et al., 2022; Solatorio & Dupriez, 2023; Pang et al., 2024). Patki et al. (2016) measured utility with a user study and Hansen et al. (2023) with the ability to retain model or feature importance ranking (measured with rank correlation) in the train-on-synthetic evaluate-on-real paradigm. Note that all of these studies evaluated utility on a single-table, even those that investigated synthetic relational databases.

3 Evaluating Synthetic Relational Databases

3.1 Multi-table Fidelity Using Aggregation

We build our approach on existing detection-based approaches. First, we address the lenient evaluation by replacing logistic regression commonly used in related work with tree-based ensemble methods (Borisov et al., 2023; Zein & Urvoy, 2022).

Current fidelity metrics fail to thoroughly evaluate relationships between tables: (i) cardinality similarity assesses only the cardinality of foreign key relationships, (ii) k-hop similarity evaluates only linear relationships between columns in related tables, and (iii) denormalization breaks the *i.i.d.* assumption required for C2ST, leading to unprincipled inference.

To address these limitations, we introduce a classifier two-sample test with aggregation (C2ST-Agg). Instead of denormalizing tables, we preserve the multi-table structure by augmenting both real and synthetic parent tables with aggregated features derived from their child tables.

Aggregation is an established technique in the field of relational reasoning (Getoor et al., 2007; Džeroski, 2010) and C2ST-Agg can be thought of as a propositionalization (Kramer et al., 2001) approach to the C2ST on relational databases. By summarizing child-table columns and relationship cardinality through aggregation functions (e.g., mean, count, max), C2ST-Agg maintains the *i.i.d.* assumption while enhancing classifier-based fidelity assessment.

Our approach addresses the issues of current fidelity metrics: it accounts both for relationship cardinality (i) and high-level interactions across all columns in related tables (ii), while maintaining the *i.i.d.* assumption for each table (iii).

In practice, users can select aggregation functions based on the specific aspects of relational data they wish to evaluate. We provide general guidelines for choosing these aggregations, along with a detailed explanation of the C2ST-Agg metric, in Appendix B.1.1

A benefit of using predictive models to evaluate fidelity is that we can use ML explainability methods to diagnose issues with our data generation. By examining the features that the discriminative model uses to distinguish between synthetic and real data, we can identify which aspects of the data the generative method fails to model well. We can employ standard ML interpretability methods such as Shapley values (Strumbelj & Kononenko, 2010; Lundberg & Lee, 2017), partial dependence plots (Friedman, 2001), accumulated local effects (Apley & Zhu, 2020), or model-specific methods such as relative variable importance in boosted trees or coefficients of linear regression models. See Section 4.2 for an example.

3.2 Relational Deep Learning Utility

Relational deep learning (Fey et al., 2024) has emerged as an alternative to traditional ML methods by transforming relational databases into graphs (Robinson et al., 2024; Papamarkou et al., 2024) and utilizing graph neural networks (GNNs). Subsequently, Robinson et al. (2024) proposed RelBench, a benchmark for relational deep learning. The authors of RelBench show that the performance of a GNN pipeline is comparable to a traditional ML pipeline approach done by a data scientist.

Up until now, ML utility has been evaluated only on single-tables (see Section 2.2), which means inter-table relationships have not been taken into account even when evaluating multi-table synthetic data. We could create a feature engineering pipeline for each database, but that would require domain knowledge, is labor-intensive, and is not easily extensible. Instead, we incorporate the graph approach of RelBench into SyntheRela. We adopt the RelBench pipeline to fit GNN models on real and synthetic data respectively, and then evaluate them on the test set consisting entirely of real data (see Appendix E for more details).

Splitting the data into train and test sets is challenging for relational databases with complex inter-table dependencies. To address this, we adopt a time-based splitting strategy, consistent with the methodology proposed by RelBench. This allows us to evaluate data generation performance for any relational database with a time component, without the need for manual data processing and is easily extensible to new databases. Notably, this is the only approach to relational utility that includes the entire relational database, ensuring a more comprehensive evaluation. We call this approach **relational deep learning utility (RDL-utility)**.

4 Benchmarking and Results

We combine our findings into a synthetic relational database benchmark. We report existing metrics for evaluating single column, single-table, and multi-table fidelity to which we add a new metric for measuring multi-table fidelity C2ST-Agg and a novel approach to evaluating the utility of relational synthetic data, RDL-utility.

We compare the following methods for synthesizing relational data: **SDV**, **RCTGAN**, **REaLTabFormer**, **ClavaDDPM**, **RGCLD**, and **TabularARGN**. Other related work does not have available source code, does not have an API or we were not able to run the source code on the selected databases.

We include 6 datasets that feature in related work (**Airbnb**, **Rossmann**, **Walmart**, **Biodegradability**, **MovieLens**), the **Cora** dataset by McCallum et al. (2000), a popular dataset in graph representation learning, and the **F1** dataset from the relational deep learning benchmark Robinson et al. (2024), for a total of **8 benchmark datasets**. The datasets vary in types of relationships and number of tables and columns, which are summarized in Table 1 (see Appendix B.2 for details and Appendix H for schema).

Table 1: **Summary of the 8 benchmark datasets.** The number of columns represents the number of non-id columns. The collection is diverse and covers all types of relational structures.

Dataset Name	# Tables	# Rows	# Columns	# Relationships	Hierarchy Type
Rossmann	2	59,085	16	1	Linear
Airbnb	2	57,217	20	1	Linear
Walmart	3	15,317	17	2	Multi Child
Cora	3	57,353	2	3	Multi Child
Biodegradability	5	21,895	6	5	Multi Child & Parent
IMDB MovieLens	7	1,249,411	14	6	Multi Child & Parent
Berka	8	757,722	37	8	Multi Child & Parent
F1	9	74,063	33	13	Multi Child & Parent

We evaluate all three levels of synthetic relational databases, with a focus on multi-table evaluation (see Appendix B.1 for all benchmark metrics). Most methods are non-deterministic, so we report results for three different replications. However, all results are stable across replications. Four of the methods are capable of synthesizing all of the datasets, irrespective of their relational structure. REALTABFORMER is only capable of generating databases with linear structure and ClavaDDPM is unable to model datasets with two or more foreign keys between a pair of tables (*Biodegredability* and *CORA*).

For single-column metrics (see Table 12 in the Appendix for results), we report the complement of the Kolmogorov-Smirnov statistic and the Total Variation Distance (the complement to KS/TV distance between two distributions P and Q is $1 - D_{\text{KS/TV}}(P||Q)$). For single-table metrics, we report the complements of the KS and TV distances between column pair correlations. For multi-table metrics, we report the average cardinality shape similarity and the k-hop ($k > 1$) correlations between tables. At all levels, we report the C2ST using XGBoost (Chen & Guestrin, 2016) as the discriminative model. We use 5-fold stratified cross-validation to estimate detection accuracy. For C2ST-Agg, we augment the rows with (a) counts of child rows for each row in each parent table, (b) the mean values of the numeric columns in the child table corresponding to the parent row; and (c) the number of unique categories in related rows.

Neural-based single-table generators evaluate privacy with metrics that use a held-out test set, as opposed to marginal-based generators, that provide provable privacy guarantees (DP). As representative sampling for relational databases is non-trivial, evaluating privacy in multi-table data remains an open problem. Recently, the MIDST Challenge (Vector Institute, 2025; Wu et al., 2025) provided the first analysis of the privacy of neural-based relational data generators, focusing on membership inference attacks (MIA) (Shokri et al., 2017). These attacks are tailored to individual generative models and are therefore not generally applicable. Instead, we performed a privacy sanity check against the SMOTE baseline, following related work (Pang et al., 2024; Tiwald et al., 2025). We use the *distance to closest record* metric and perform our analysis on the *Airbnb* dataset as we can re-sample a hold-out set of equal size without a temporal shift. The results of the privacy check show that all the methods outperform SMOTE, providing no evidence of systematic data copying (see Appendix D). We also explore how C2ST can be used as a data copying diagnostic in Appendix D.3.

4.1 Single-Table Performance

We first evaluate the fidelity of individual tables. We focus on the detection metric and how well column pairs (bivariate distributions) are modeled. Table 2 summarizes the results. On the databases that it is able to generate, the diffusion-based ClavaDDPM performs best, followed by the autoregressive TabularARGN specializing in sequential (linear structured) databases.

Table 2: **Single-table results**. For each dataset and metric we report the average and standard error of C2ST accuracy (lower is better) and column pair trends (Pairs - higher is better) across all tables for three independent samples. The best result for each dataset is **in bold**.

Dataset	Metric	TARGN	RGCLD	ClavaDDPM	RCTGAN	REALTABF.	SDV
Airbnb	C2ST (↓)	**64.23**$_{\pm 0.20}$	70.25$_{\pm 3.31}$	78.10$_{\pm 0.03}$	88.37$_{\pm 0.14}$	83.97$_{\pm 4.36}$	99.75$_{\pm 5e-3}$
	Pairs (↑)	**93.48**$_{\pm 0.33}$	88.51$_{\pm 0.60}$	87.78$_{\pm 0.12}$	79.37$_{\pm 0.29}$	53.90$_{\pm 1.26}$	49.03$_{\pm 0.08}$
Rossmann	C2ST (↓)	**56.07**$_{\pm 0.58}$	68.97$_{\pm 3.85}$	66.77$_{\pm 0.14}$	88.02$_{\pm 0.50}$	74.70$_{\pm 0.55}$	96.90$_{\pm 0.21}$
	Pairs (↑)	**91.34**$_{\pm 0.08}$	90.08$_{\pm 0.58}$	84.78$_{\pm 0.80}$	84.38$_{\pm 0.40}$	84.58$_{\pm 0.88}$	67.77$_{\pm 0.25}$
Walmart	C2ST (↓)	83.54$_{\pm 0.84}$	66.76$_{\pm 1.82}$	**53.50**$_{\pm 1.95}$	76.40$_{\pm 0.55}$	70.87$_{\pm 1.07}$	87.02$_{\pm 0.81}$
	Pairs (↑)	83.89$_{\pm 0.21}$	91.74$_{\pm 0.99}$	**94.02**$_{\pm 0.14}$	86.60$_{\pm 0.25}$	83.10$_{\pm 0.46}$	87.61$_{\pm 0.23}$
Berka	C2ST (↓)	72.31$_{\pm 0.17}$	64.17$_{\pm 2.04}$	**54.48**$_{\pm 0.11}$	68.12$_{\pm 0.44}$	-	82.40$_{\pm 0.33}$
	Pairs (↑)	70.43$_{\pm 0.25}$	73.77$_{\pm 1.75}$	**88.54**$_{\pm 1.19}$	74.22$_{\pm 0.28}$		64.01$_{\pm 0.11}$
F1	C2ST (↓)	81.93$_{\pm 0.49}$	**69.63**$_{\pm 1.23}$	71.42$_{\pm 0.46}$	80.67$_{\pm 0.31}$	-	89.84$_{\pm 0.22}$
	Pairs (↑)	81.31$_{\pm 0.45}$	**92.39**$_{\pm 0.77}$	84.65$_{\pm 0.05}$	90.17$_{\pm 0.03}$		73.05$_{\pm 0.19}$
IMDB	C2ST (↓)	50.92$_{\pm 0.22}$	55.01$_{\pm 2.36}$	**49.83**$_{\pm 0.07}$	55.38$_{\pm 0.11}$	-	TLE
	Pairs (↑)	97.80$_{\pm 0.13}$	93.89$_{\pm 2.42}$	**98.66**$_{\pm 0.10}$	81.65$_{\pm 0.03}$		
Biodeg.	C2ST (↓)	58.79$_{\pm 0.19}$	63.45$_{\pm 2.26}$	-	**58.26**$_{\pm 0.15}$	-	68.59$_{\pm 0.14}$
	Pairs (↑)	74.82$_{\pm 0.35}$	90.89$_{\pm 5.05}$		85.44$_{\pm 2.19}$		**97.58**$_{\pm 0.50}$
Cora	C2ST (↓)	50.94$_{\pm 0.37}$	52.98$_{\pm 1.03}$	-	**48.97**$_{\pm 0.14}$	-	75.45$_{\pm 0.16}$

The rankings of methods for single column metrics are similar to those of single-tables. As expected, the methods model individual columns better. See Appendix D.1 for details. Interestingly, TabularARGN performs best on modeling marginal distributions, indicating their discretization approach is a robust preprocessing step.

4.2 Multi-Table Performance

Multi-table metrics examine how well the relationship cardinality and the relationships between different tables are preserved. Cardinality shape similarity examines only the former, while k-HOP similarity evaluates the latter; C2ST-Agg examines both. We report the results in Table 3. Similar to single-table fidelity, the diffusion-based methods perform best.

Table 3: **Multi-table results.** For each dataset we report the average and standard error of C2ST-Agg accuracy (lower is better), cardinality similarity (higher is better) and k-hop correlation similarity (higher is better) across all tables for three independent samples. "-" denotes a method is unable to generate the dataset, and TLE timeout. The best result is in **bold** and results within one standard are <u>underlined</u>.

Dataset	Metric	TARGN	RGCLD	ClavaDDPM	RCTGAN	REALTABF.	SDV
	C2ST-Agg (↓)	**63.47**$_{\pm 0.88}$	79.53$_{\pm 5.39}$	≈ 100.0	98.22$_{\pm 0.08}$	99.13$_{\pm 0.02}$	99.94$_{\pm 0.01}$
Airbnb	Cardinality (↑)	98.59$_{\pm 0.32}$	99.36$_{\pm 0.10}$	**99.65**$_{\pm 0.06}$	95.45$_{\pm 0.62}$	76.38$_{\pm 0.47}$	26.36$_{\pm 0.03}$
	1-HOP (↑)	79.66$_{\pm 0.36}$	84.50$_{\pm 2.74}$	**86.69**$_{\pm 0.14}$	68.78$_{\pm 0.54}$	33.99$_{\pm 5.76}$	24.58$_{\pm 0.03}$
	C2ST-Agg (↓)	**60.43**$_{\pm 0.63}$	75.56$_{\pm 0.74}$	85.77$_{\pm 0.07}$	86.11$_{\pm 1.01}$	85.90$_{\pm 1.33}$	98.37$_{\pm 0.23}$
Rossmann	Cardinality (↑)	94.17$_{\pm 1.84}$	<u>98.95</u>$_{\pm 0.50}$	**99.19**$_{\pm 0.29}$	82.69$_{\pm 1.95}$	41.82$_{\pm 10.29}$	<u>99.16</u>$_{\pm 0.15}$
	1-HOP (↑)	**92.95**$_{\pm 0.78}$	88.30$_{\pm 0.38}$	82.81$_{\pm 0.47}$	87.02$_{\pm 0.17}$	80.25$_{\pm 0.84}$	73.84$_{\pm 0.34}$
	C2ST-Agg (↓)	94.81$_{\pm 1.68}$	88.89$_{\pm 1.05}$	**73.33**$_{\pm 2.92}$	94.81$_{\pm 1.68}$	90.00$_{\pm 0.91}$	88.52$_{\pm 1.60}$
Walmart	Cardinality (↑)	65.93$_{\pm 1.98}$	**94.81**$_{\pm 0.30}$	93.33$_{\pm 2.28}$	88.15$_{\pm 1.51}$	85.56$_{\pm 4.57}$	86.30$_{\pm 1.09}$
	1-HOP (↑)	75.40$_{\pm 1.49}$	82.05$_{\pm 2.77}$	**86.40**$_{\pm 1.73}$	79.02$_{\pm 0.15}$	74.99$_{\pm 0.20}$	76.64$_{\pm 1.07}$
	C2ST-Agg (↓)	80.56$_{\pm 1.86}$	72.71$_{\pm 2.33}$	**69.12**$_{\pm 0.63}$	76.86$_{\pm 2.22}$		77.43$_{\pm 0.14}$
	Cardinality (↑)	85.17$_{\pm 0.84}$	≈ **100.0**	96.43$_{\pm 0.36}$	81.28$_{\pm 1.07}$		80.53$_{\pm 0.72}$
Berka	1-HOP (↑)	72.82$_{\pm 0.38}$	80.74$_{\pm 1.63}$	**87.92**$_{\pm 1.66}$	78.87$_{\pm 0.91}$	-	59.09$_{\pm 0.49}$
	2-HOP (↑)	65.51$_{\pm 0.31}$	73.77$_{\pm 1.78}$	**84.41**$_{\pm 2.46}$	77.98$_{\pm 0.95}$		23.09$_{\pm 0.21}$
	3-HOP (↑)	59.34$_{\pm 0.62}$	64.81$_{\pm 4.96}$	**80.67**$_{\pm 2.18}$	<u>78.65</u>$_{\pm 0.69}$		58.23$_{\pm 0.58}$
	C2ST-Agg (↓)	95.90$_{\pm 0.94}$	**74.18**$_{\pm 2.68}$	82.52$_{\pm 0.25}$	91.23$_{\pm 0.39}$		94.55$_{\pm 0.24}$
F1	Cardinality (↑)	58.17$_{\pm 3.71}$	≈ **100.0**	88.45$_{\pm 3.05}$	56.82$_{\pm 1.55}$	-	71.88$_{\pm 0.12}$
	1-HOP (↑)	77.37$_{\pm 0.26}$	**88.35**$_{\pm 2.01}$	79.35$_{\pm 0.03}$	79.14$_{\pm 0.72}$		68.45$_{\pm 0.20}$
	2-HOP (↑)	76.25$_{\pm 0.32}$	**87.53**$_{\pm 1.56}$	84.18$_{\pm 0.12}$	83.50$_{\pm 0.82}$		76.93$_{\pm 0.24}$
	C2ST-Agg (↓)	73.76$_{\pm 1.78}$	65.37$_{\pm 5.54}$	**65.00**$_{\pm 0.34}$	81.56$_{\pm 2.00}$		
IMDB	Cardinality (↑)	81.19$_{\pm 0.80}$	≈ **100.0**	98.95$_{\pm 0.03}$	79.53$_{\pm 1.27}$	-	TLE
	1-HOP (↑)	88.64$_{\pm 0.70}$	85.66$_{\pm 5.25}$	**91.57**$_{\pm 1.25}$	81.76$_{\pm 0.20}$		
	C2ST-Agg (↓)	88.86$_{\pm 0.26}$	**70.51**$_{\pm 4.08}$		83.82$_{\pm 3.35}$		98.02$_{\pm 0.06}$
Biodeg.	Cardinality (↑)	79.53$_{\pm 0.24}$	**97.95**$_{\pm 0.01}$	-	85.22$_{\pm 0.50}$	-	61.17$_{\pm 0.36}$
	1-HOP (↑)	61.36$_{\pm 0.47}$	**78.46**$_{\pm 4.65}$		<u>75.80</u>$_{\pm 1.46}$		49.09$_{\pm 0.59}$
	2-HOP (↑)	60.54$_{\pm 0.44}$	<u>75.12</u>$_{\pm 4.02}$		**77.04**$_{\pm 1.96}$		47.80$_{\pm 2.16}$
	C2ST-Agg (↓)	68.80$_{\pm 0.67}$	**62.33**$_{\pm 1.60}$		73.74$_{\pm 0.47}$		99.59$_{\pm 0.03}$
Cora	Cardinality (↑)	96.27$_{\pm 0.13}$	≈ **100.0**	-	90.48$_{\pm 2.16}$	-	68.82$_{\pm 0.29}$
	1-HOP (↑)	**80.42**$_{\pm 0.34}$	63.81$_{\pm 3.06}$		68.39$_{\pm 0.08}$		4.95$_{\pm 0.12}$

C2ST with aggregation uses the same features for single-table and multi-table evaluation (with the exception of the aggregation attributes). This allows us to directly compare single and multi-table performance. For most methods, we observe a significant drop in fidelity when adding aggregations. Figure 2 shows how C2ST detection accuracy increases when we add information about relationships between tables. We investigate this further using explainability methods and show how these can be used to debug generative methods.

ML explanation with feature importance reveals that methods struggle with preserving the relationships between columns across tables. Figure 3a shows an example of how aggregation attributes summarizing information about child table rows are the most discriminative features. We further examine two such

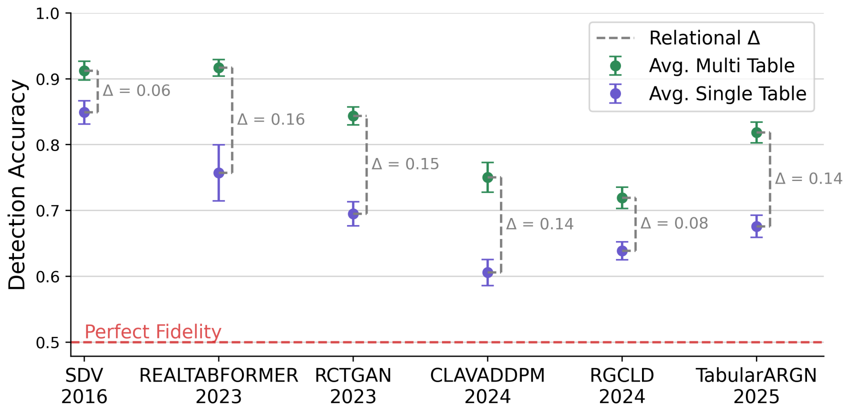
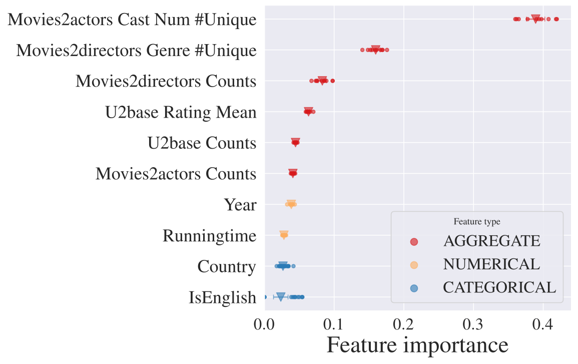
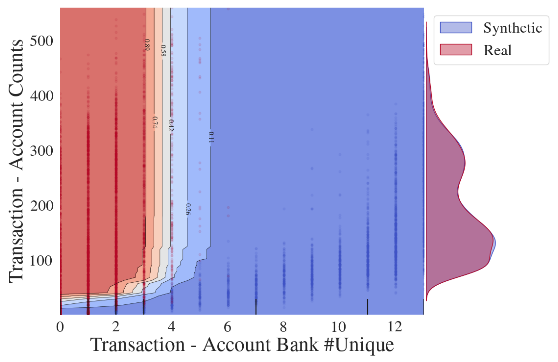


Figure 2: **Comparing single and multi-table performance**. While there is an overall trend of improvement over time of both single-table and multi-table fidelity, most methods still exhibit a significant gap between single-table and multi-table fidelity (relational Δ).

attributes in Figure 4. The partial dependence plots of the first and fourth most important features from Figure 3 show how subsets of both categorical (Fig. 4a) and numerical (Fig. 4b) features' conditional distributions are informative to the discriminative model. In Figure 3b, we illustrate how a simple interaction between two aggregation attributes—specifically, relationship cardinality and the number of unique categories in a child table—renders the synthetic data nearly separable from the original. This highlights how C2ST-Agg captures both structural and attribute-based discrepancies, including their interactions, offering a more comprehensive evaluation than simpler metrics like k-hop similarity. We include the interpretability methods directly into our implementation of the detection metric, allowing users to immediately investigate where their method is underperforming after evaluating the synthetic data.

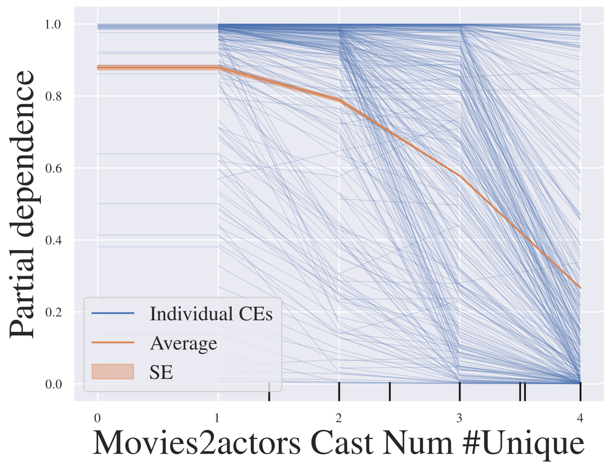


(a)

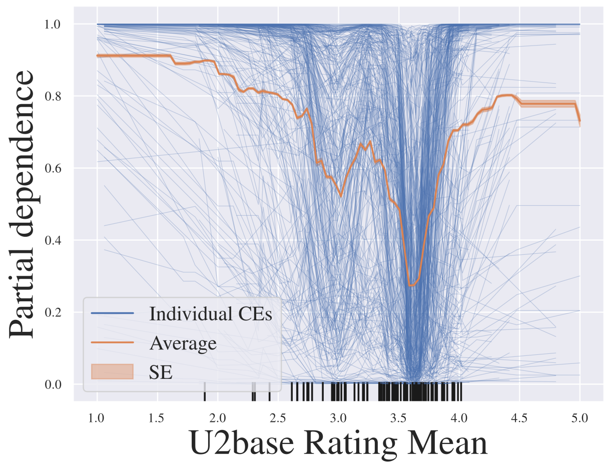


(b)

Figure 3: **ML Interpretability Diagnostics of C2ST.** (a) Feature importance for C2ST-Agg using XGBoost. Results are for the best-performing method. The added features that incorporate relational information (red) are the most important for discriminating between real and synthetic data. (b) Two-way PDP showing the interaction between the relationship cardinality of the Account–Transaction key (with its marginal distribution shown on the right) and the number of unique banks associated with an account's transactions. The plot illustrates how a combination of structural and informational aggregation attributes can render the synthetic data nearly separable from the original.



(a)



(b)

Figure 4: **Partial dependence plots of C2ST**. Results are for the 1st and 4th most important feature from Figure 3. With ideally generated synthetic data, features would not differ between synthetic and original data and every partial dependence plot would be a horizontal line at 50% probability. We can observe that (a) the synthetic data have too many unique actor cast numbers (higher probability of being synthetic when feature value is larger than 4) and (b) the mean movie ratings in the original data vary more than in the synthetic data, where they are more concentrated around 3.5.

4.3 Relational Deep Learning Utility Performance

We run extensive experiments for RDL-utility across 5 databases that contain a temporal feature, using 6 GNN architectures with hyperparameter tuning for each architecture and database pair on original data (see Appendix E for more details). Table 4 summarizes the RDL-utility results (Section 3.2). Details of the tasks are provided in the Appendix E. The utility scores follow the same trend as the multi-table results, with models achieving high fidelity also performing best in utility tasks. The best performing models are the diffusion based approaches ClavaDDPM and RGCLD, with the autoregressive approach TabularARGN performing best on one simpler two-table database. The drop in utility performance for databases with more complex relational structures is also consistent with fidelity results and shows that utility of relational databases should be evaluated using multi-table approaches.

To be able to estimate the value and validity of RDL-utility in measuring the utility performance of the generated synthetic databases, we implement two baseline metrics for measuring utility (see Appendix E.3 for more details). The first one is a simple baseline, where we train a LightGBM (Ke et al., 2017) model directly on the entity table with no feature engineering. This corresponds to the typical machine learning utility evaluation used in related work (Solatorio & Dupriez, 2023; Pang et al., 2024). The second baseline is an automated feature engineering baseline using Deep Feature Synthesis (DFS) (Kanter & Veeramachaneni, 2015), based on calculating aggregated features of the child tables and appending them to the entity table, also trained with a LightGBM model.

Entity table LightGBM results (Table 5) again show that diffusion based models are able to best generate a single table overall. The DFS feature engineering baseline scores (Table 6) show a similar story. The drop in performance in the RDL-utility compared to the entity-table baseline stems from the noise introduced by poorly generating the relational structure and features of child tables.

RDL-utility, alongside naive baselines, demonstrates that the models can learn patterns from synthetic databases that generalize to held-out test sets even in cases of non-perfect fidelity.

Table 4: **RDL-utility results.** We report mean ± SE of ROC-AUC (classification) and MAE (regression). Naive baseline (mean/majority) is in parentheses. "-" indicates invalid time columns. Best is **bold**; within one SE is $\underline{\text{underlined}}$.

Dataset	Model		ORIG.	RGCLD	TARGN	CLAVA	RCTGAN	REALTABF.	SDV
Rossmann	GAT	MAE (↓)	211 (324)	$252_{\pm 21}$	$304_{\pm 18}$	$\mathbf{230}_{\pm 0.55}$	$289_{\pm 5}$	$564_{\pm 59}$	$963_{\pm 40}$
	GATv2		205 (324)	$\mathbf{215}_{\pm 9}$	$330_{\pm 11}$	$239_{\pm 2}$	$251_{\pm 8}$	$531_{\pm 91}$	$1212_{\pm 154}$
	GIN		178 (324)	$197_{\pm 5}$	$224_{\pm 8}$	$\mathbf{193}_{\pm 1}$	$219_{\pm 2}$	$257_{\pm 27}$	3648
	G-Conv		178 (324)	$206_{\pm 9}$	$229_{\pm 1}$	$\mathbf{192}_{\pm 0.74}$	$216_{\pm 4}$	$253_{\pm 28}$	3430
	G-SAGE		178 (324)	$204_{\pm 6}$	$229_{\pm 3}$	$\mathbf{193}_{\pm 0.70}$	$224_{\pm 4}$	$256_{\pm 35}$	3428
	RelGNN		178 (324)	$210_{\pm 11}$	$223_{\pm 12}$	$\mathbf{192}_{\pm 1}$	$219_{\pm 3}$	$254_{\pm 21}$	3440
Walmart	GAT	MAE (↓)	13118 (14.7k)	$13626_{\pm 147}$	$13846_{\pm 50}$	$\mathbf{13496}_{\pm 67}$	$13813_{\pm 38}$	$14364_{\pm 151}$	$13795_{\pm 76}$
	GATv2		13114 (14.7k)	$14081_{\pm 291}$	$13841_{\pm 41}$	$\mathbf{13497}_{\pm 35}$	$13750_{\pm 44}$	$14362_{\pm 146}$	$13760_{\pm 163}$
	GIN		13056 (14.7k)	$15165_{\pm 1420}$	$13841_{\pm 19}$	$\mathbf{13423}_{\pm 74}$	$13755_{\pm 71}$	$14305_{\pm 138}$	$13539_{\pm 144}$
	G-Conv		13033 (14.7k)	$15268_{\pm 1540}$	$13850_{\pm 56}$	$\mathbf{13329}_{\pm 23}$	$13849_{\pm 81}$	$14267_{\pm 131}$	$13581_{\pm 103}$
	G-SAGE		13052 (14.7k)	$15090_{\pm 1204}$	$13838_{\pm 43}$	$\mathbf{13421}_{\pm 17}$	$13738_{\pm 43}$	$14270_{\pm 130}$	$13761_{\pm 189}$
	RelGNN		13097 (14.7k)	$15243_{\pm 1682}$	$13846_{\pm 18}$	$\mathbf{13408}_{\pm 29}$	$13791_{\pm 51}$	$14311_{\pm 125}$	$13722_{\pm 161}$
Airbnb	GAT	AUC (↑)	0.72 (0.5)	$0.64_{\pm 0.01}$	$\mathbf{0.67}_{\pm 0.00}$	$0.64_{\pm 0.01}$	$0.48_{\pm 0.01}$	-	$0.46_{\pm 0.01}$
	GATv2		0.74 (0.5)	$0.62_{\pm 0.01}$	$\mathbf{0.65}_{\pm 0.01}$	$0.56_{\pm 0.07}$	$0.50_{\pm 0.03}$		$0.51_{\pm 0.00}$
	GIN		0.75 (0.5)	$0.62_{\pm 0.04}$	$\mathbf{0.67}_{\pm 0.01}$	$0.66_{\pm 0.02}$	$0.56_{\pm 0.01}$		$0.56_{\pm 0.01}$
	G-Conv		0.74 (0.5)	$0.63_{\pm 0.02}$	$\mathbf{0.69}_{\pm 0.02}$	$0.65_{\pm 0.01}$	$0.54_{\pm 0.00}$		$0.54_{\pm 0.00}$
	G-SAGE		0.78 (0.5)	$\underline{0.68}_{\pm 0.04}$	$\mathbf{0.70}_{\pm 0.02}$	$\underline{0.68}_{\pm 0.01}$	$0.63_{\pm 0.01}$		$0.63_{\pm 0.00}$
	RelGNN		0.77 (0.5)	$0.66_{\pm 0.03}$	$\mathbf{0.71}_{\pm 0.01}$	$0.67_{\pm 0.01}$	$0.56_{\pm 0.03}$		$0.60_{\pm 0.00}$
Berka	GAT	AUC (↑)	0.92 (0.5)	$0.56_{\pm 0.11}$	$\underline{0.71}_{\pm 0.07}$	$\mathbf{0.73}_{\pm 0.05}$	-		-
	GATv2		0.96 (0.5)	$0.44_{\pm 0.09}$	$0.52_{\pm 0.08}$	$\mathbf{0.64}_{\pm 0.05}$			
	GIN		0.96 (0.5)	$\underline{0.61}_{\pm 0.05}$	$0.53_{\pm 0.12}$	$\mathbf{0.68}_{\pm 0.13}$			
	G-Conv		0.98 (0.5)	$\mathbf{0.68}_{\pm 0.05}$	$0.35_{\pm 0.09}$	$0.56_{\pm 0.15}$			
	G-SAGE		0.93 (0.5)	$0.56_{\pm 0.03}$	$\mathbf{0.70}_{\pm 0.12}$	$\underline{0.61}_{\pm 0.14}$			
	RelGNN		0.95 (0.5)	$\underline{0.55}_{\pm 0.17}$	$0.31_{\pm 0.08}$	$\mathbf{0.65}_{\pm 0.15}$			
F1	GAT	AUC (↑)	0.85 (0.5)	$0.55_{\pm 0.05}$	$\underline{0.66}_{\pm 0.04}$	$\underline{0.66}_{\pm 0.04}$	$\mathbf{0.68}_{\pm 0.04}$		$0.49_{\pm 0.09}$
	GATv2		0.86 (0.5)	$0.73_{\pm 0.04}$	$0.54_{\pm 0.10}$	$\mathbf{0.74}_{\pm 0.01}$	$0.50_{\pm 0.15}$		$0.56_{\pm 0.06}$
	GIN		0.87 (0.5)	$\mathbf{0.77}_{\pm 0.01}$	$0.56_{\pm 0.05}$	$\underline{0.76}_{\pm 0.02}$	$0.42_{\pm 0.14}$		$0.64_{\pm 0.12}$
	G-Conv		0.87 (0.5)	$\mathbf{0.78}_{\pm 0.01}$	$0.43_{\pm 0.11}$	$0.75_{\pm 0.01}$	$0.46_{\pm 0.12}$		$0.75_{\pm 0.02}$
	G-SAGE		0.86 (0.5)	$\mathbf{0.78}_{\pm 0.01}$	$0.44_{\pm 0.14}$	$0.76_{\pm 0.02}$	$0.49_{\pm 0.10}$		$0.68_{\pm 0.05}$
	RelGNN		0.84 (0.5)	$\mathbf{0.77}_{\pm 0.01}$	$0.44_{\pm 0.08}$	$0.69_{\pm 0.03}$	$0.46_{\pm 0.12}$		$0.60_{\pm 0.15}$

Table 5: **Entity table LightGBM baseline.** We report mean ± SE of ROC-AUC (classification) and MAE (regression). Naive baseline (mean/majority) is in parentheses. "-" indicates invalid time columns. Best is **bold**; within one SE is $\underline{\text{underlined}}$.

Dataset		ORIG.	TARGN	RGCLD	CLAVADDPM	RCTGAN	REALTABF.	SDV
Rossmann	MAE (↓)	225 (324)	$227_{\pm 0.77}$	$225_{\pm 0.96}$	$225_{\pm 0.06}$	$\mathbf{218}_{\pm 6}$	$224_{\pm 2}$	$4286_{\pm 2}$
Walmart	MAE (↓)	13664 (14.7k)	$13841_{\pm 26}$	$13704_{\pm 45}$	$\mathbf{13666}_{\pm 36}$	$13818_{\pm 78}$	$14433_{\pm 129}$	$13949_{\pm 84}$
Airbnb	AUC (↑)	0.73 (0.5)	$\underline{0.73}_{\pm 0.01}$	$\mathbf{0.74}_{\pm 0.01}$	$0.62_{\pm 0.03}$	$0.71_{\pm 0.01}$	-	$0.51_{\pm 0.02}$
Berka	AUC (↑)	0.71 (0.5)	$0.53_{\pm 0.10}$	$0.62_{\pm 0.10}$	$\mathbf{0.73}_{\pm 0.04}$	-		-
F1	AUC (↑)	0.61 (0.5)	$0.53_{\pm 0.01}$	$\mathbf{0.63}_{\pm 0.04}$	$0.57_{\pm 0.02}$	$0.58_{\pm 0.04}$		$\underline{0.60}_{\pm 0.01}$

Table 6: **DFS feature engineering LightGBM baseline.** We report mean ± SE of ROC-AUC (classification) and MAE (regression). Naive baseline (mean/majority) is in parentheses. "-" indicates invalid time columns. Best is **bold**; within one SE is $\underline{\text{underlined}}$.

Dataset		ORIG.	TARGN	RGCLD	CLAVADDPM	RCTGAN	REALTABF.	SDV
Rossmann	MAE (↓)	100 (324)	$215_{\pm 5}$	$210_{\pm 7}$	$\mathbf{203}_{\pm 3}$	$210_{\pm 2}$	$230_{\pm 5}$	$2290_{\pm 67}$
Walmart	MAE (↓)	13251 (14.7k)	$13826_{\pm 21}$	$14560_{\pm 830}$	$\mathbf{13392}_{\pm 48}$	$13741_{\pm 32}$	$14130_{\pm 127}$	$13486_{\pm 63}$
Airbnb	AUC (↑)	0.74 (0.5)	$\underline{0.72}_{\pm 0.02}$	$\mathbf{0.73}_{\pm 0.01}$	$0.60_{\pm 0.06}$	$0.71_{\pm 0.02}$	-	$0.54_{\pm 0.01}$
Berka	AUC (↑)	0.90 (0.5)	$0.64_{\pm 0.08}$	$0.65_{\pm 0.05}$	$\mathbf{0.71}_{\pm 0.01}$	-		-
F1	AUC (↑)	0.74 (0.5)	$0.45_{\pm 0.01}$	$\mathbf{0.72}_{\pm 0.03}$	$0.68_{\pm 0.01}$	$0.65_{\pm 0.08}$		$0.51_{\pm 0.02}$

4.4 Privacy Evaluation

Our primary focus is on evaluating the **fidelity** and **utility** of synthetic relational data. However, high data quality should not come at the expense of **privacy**. Following prior work (Kotelnikov et al., 2023; Pang et al., 2024; Zhang et al., 2024), we conduct a privacy comparison against SMOTE (Chawla et al., 2002), an interpolation-based method that synthesizes new data points via convex combinations of real samples and their nearest neighbors.

To assess privacy risks, we compute the distance to closest record (DCR)(Zhao et al., 2021), a common metric that quantifies how close synthetic records are to real data. Specifically, we report the DCR score—the probability that a synthetic sample is closer to a real training point than to any sample from a held-out test set. We conduct our privacy check on the *Airbnb* dataset containing anonymised information of real users.

However, recent studies have highlighted limitations of distance-based privacy metrics (Yao et al., 2025; Ganev & De Cristofaro, 2025; Ward et al., 2024), showing that such measures can fail to capture information leakage in high-dimensional complex datasets and adversarial scenarios. To provide a more robust and comprehensive evaluation, we therefore complement DCR with attack-based privacy assessment via membership inference attacks (MIA).

In an MIA, an adversary attempts to infer whether a given record from the real data was part of the training set used to generate the synthetic data (El Emam et al., 2022). We adopt the implementation of Lautrup et al. (2024), which assumes a black-box adversary with knowledge of the population but no insight into the synthesis model. The attacker is modeled as a LightGBM classifier (Ke et al., 2017) trained to distinguish between synthetic samples and real records from an external holdout set. The classifier's discriminative performance serves as an estimate of membership disclosure risk. Analogous to the DCR evaluation, MIA is conducted at the table level, as no multi-table variant currently exists.

We report DCR results in Table 7 and MIA AUC scores in Table 8. Across both tables, all methods—except for ClavaDDPM on the Users table—maintain reasonable DCR values close to the ideal 50% range, suggesting no systematic data copying or excessive proximity between synthetic and real samples. However, the MIA evaluation tells a different story. As shown in Table 8, most generators expose the training data to a non-negligible membership inference risk. With the exception of SDV and ReaLTabFormer, whose MIA AUC scores remain close to random guessing (≈ 50%) but correspond to low data quality (see Tables 2, 3, 4), the adversarial classifier achieves above-random discrimination for nearly all methods. We include detailed results of our privacy experiments in Appendix D.

Table 7: **DCR Scores** on the Airbnb dataset represent the probability of a random sample being closer to a training sample rather than a randomly drawn sample from a held-out dataset. Values near 50% incidate no systematic data copying, while larger values indicate privacy risks.

Method	Sessions	Users
TARGN	50.40 ± 0.20	49.87 ± 0.50
RGCLD	51.77 ± 0.21	50.30 ± 0.50
ClavaDDPM	51.77 ± 0.21	84.83 ± 0.36
RCTGAN	49.48 ± 0.22	48.93 ± 0.50
REALTABF.	46.65 ± 1.56	51.38 ± 0.50
SDV	46.16 ± 0.09	48.01 ± 0.50
SMOTE	88.86 ± 0.10	99.82 ± 0.04

Table 8: **MIA AUC Scores** on the Airbnb dataset measure the overall ability of an attacker to distinguish between training records (members) and held-out records (non-members). Values near 50% indicate the attack is equivalent to random guessing, suggesting high privacy protection.

Method	Sessions	Users
TARGN	55.18 ± 0.10	59.77 ± 0.36
RGCLD	50.94 ± 0.06	78.52 ± 0.72
CLAVADDPM	64.43 ± 0.25	68.88 ± 0.32
RCTGAN	62.35 ± 0.62	98.76 ± 0.18
REALTABF.	50.98 ± 0.52	50.11 ± 0.04
SDV	50.00 ± 0.00	50.02 ± 0.01
SMOTE	94.03 ± 0.11	92.35 ± 0.15

Our privacy evaluation shows that despite not achieving perfect fidelity nor systematically copying the original data, current methods still pose measurable privacy risks to their training data. These findings further confirm recent results highlighting the limitations of distance-based privacy metrics and underscore the importance of attack-based evaluations such as MIA. More broadly, they motivate future research into privacy evaluation protocols for synthetic relational data and the development of privacy-preserving generative models that balance fidelity and privacy protection more effectively (Izzo et al., 2024; Jiang et al., 2025).

5 Conclusion

To address the need for standardized evaluation in the emerging field of synthetic relational database generation, the primary contribution of this work is SyntheRela, the first open-source benchmark specifically designed for this task. In developing SyntheRela, we critically reviewed existing methods for evaluating the fidelity and utility of synthetic data and introduced two novel contributions: C2ST-Agg, a robust detection-based metric for assessing multi-table fidelity, and RDL-Utility, a general framework for evaluating the utility of synthetic relational databases. SyntheRela is designed to offer practical insights for synthetic data users and features a public leaderboard to serve as a baseline for future methods.

We also applied the benchmark to current state-of-the-art synthetic data generation methods. The best methods generate marginal distributions well and are on par with single-table methods. Diffusion-based approaches also perform well on single-table fidelity. However, all methods decline in performance when evaluated on multi-table fidelity, highlighting the added complexity of modeling relational databases. Although no method perfectly preserves inter-table relationships, when evaluated on RDL-utility, most outperform naive baselines, and some achieve model performance scores comparable to those trained on the original database. This suggests that despite not achieving perfect fidelity, these synthetic data remain valuable for downstream machine learning tasks.

5.1 Limitations

Our benchmark, SyntheRela, evaluates neural-based generative methods focused on fidelity and utility on real-world relational databases. This excludes marginal-based methods focused on differential privacy and query error metrics, which are often tailored to specific datasets and queries. The current benchmark datasets lack predefined queries for evaluating these differential privacy-centric approaches. Future work could incorporate datasets with predefined queries and incorporate marginal-based methods to bridge the gap between the two classes of methods.

Several aspects of the evaluation of synthetic data are limited by the difficulty of representative sampling. We circumvent this in RDL-utility by using a temporal-based split, which limits us to databases with a temporal component. More work needs to be done to understand the limitations and prepare new benchmark datasets or dataset generators.

References

Airbnb. Airbnb new user bookings. `https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings`, 2015.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

Kaveh Alimohammadi, Hao Wang, Ojas Gulati, Akash Srivastava, and Navid Azizan. Differentially private synthetic data generation for relational databases, 2025. URL `https://arxiv.org/abs/2405.18670`.

Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1059–1086, 2020.

Arno Appenzeller, Moritz Leitner, Patrick Philipp, Erik Krempel, and Jürgen Beyerer. Privacy and utility of private synthetic data for medical data analyses. *Applied Sciences*, 12(23):12320, 2022.

Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Ro(Beaulieu-Jones et al., 2019).bert E Tillman, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, pp. 1–8, 2020.

Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.

Petr Berka et al. Guide to the financial data set. *PKDD2000 discovery challenge*, 2000.

Hendrik Blockeel, Sašo Džeroski, Boris Kompare, Stefan Kramer, Bernhard Pfahringer, and Wim Laer. Experiments in predicting biodegradability. *Applied Artificial Intelligence*, 18, 06 1999.

Daniel Bonnéry, Yi Feng, Angela K Henneberger, Tessa L Johnson, Mark Lachowicz, Bess A Rose, Terry Shaw, Laura M Stapleton, Michael E Woolley, and Yating Zheng. The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data. *Journal of Research on Educational Effectiveness*, 12(4):616–647, 2019.

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.

Teodora Sandra Buda, Thomas Cerqueus, John Murphy, and Morten Kristiansen. Cods: A representative sampling method for relational databases. In *Database and Expert Systems Applications: 24th International Conference, DEXA 2013, Prague, Czech Republic, August 26-29, 2013. Proceedings, Part I 24*, pp. 342–356. Springer, 2013.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

Kuntai Cai, Xiaoyu Lei, Jianxin Wei, and Xiaokui Xiao. Data synthesis via differentially private markov random fields. *Proceedings of the VLDB Endowment*, 14(11):2190–2202, 2021.

Kuntai Cai, Xiaokui Xiao, and Graham Cormode. Privlava: Synthesizing relational data with foreign keys under differential privacy. *Proc. ACM Manag. Data*, 1(2), jun 2023.

Kuntai Cai, Xiaokui Xiao, and Yin Yang. Privpetal: Relational data synthesis via permutation relations, 2025. URL `https://arxiv.org/abs/2503.22970`.

Luca Canale, Nicolas Grislain, Grégoire Lothe, and Johan Leduc. Generative modeling of complex data, 2022. URL `https://arxiv.org/abs/2202.02145`.

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*, KDD '16, pp. 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322.

Inc. DataCebo. *SDMetrics*, 10 2022. URL `https://docs.sdv.dev/sdmetrics/`. Version 0.8.0.

Asim Kumar Debnath, Rosa L. Lopez de Compadre, Gargi Debnath, Alan J. Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of Medicinal Chemistry*, 34(2):786–797, 1991.

Sašo Džeroski. *Relational data mining*. Springer, 2010.

Khaled El Emam, Lucy Mosquera, Xi Fang, and Alaa El-Hussuna. Utility metrics for evaluating synthetic health data generation methods: validation study. *JMIR medical informatics*, 10(4):e35734, 2022.

F1. F1 db, 2021. URL `https://github.com/f1db/f1db`.

Matthias Fey, Weihua Hu, Kexin Huang, Jan Eric Lenssen, Rishabh Ranjan, Joshua Robinson, Rex Ying, Jiaxuan You, and Jure Leskovec. Position: relational deep learning-graph representation learning on relational databases. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 13592–13607, 2024.

Will Cukierski FlorianKnauer. Rossmann store sales, 2015. URL `https://kaggle.com/competitions/rossmann-store-sales`.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Georgi Ganev and Emiliano De Cristofaro. The inadequacy of similarity-based privacy metrics: Privacy attacks against "truly anonymous" synthetic datasets. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 4007–4025. IEEE, 2025.

Rainer Gemulla, Philipp Rösch, and Wolfgang Lehner. Linked bernoulli synopses: Sampling along foreign keys. In *Scientific and Statistical Database Management: 20th International Conference, SSDBM 2008, Hong Kong, China, July 9-11, 2008 Proceedings 20*, pp. 6–23. Springer, 2008.

Lise Getoor, Nir Friedman, Daphne Koller, Avi Pfeffer, and Ben Taskar. Probabilistic relational models. In *Introduction to Statistical Relational Learning*. The MIT Press, 08 2007.

Gretel.ai. Gretel blog. `https://gretel.ai/blog`, 2024. Accessed on March 24th, 2024.

Mohamed Gueye, Yazid Attabi, and Maxime Dumas. Row conditional-tgan for generating synthetic relational databases. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

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

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), dec 2015. ISSN 2160-6455. doi: 10.1145/2827872.

Valter Hudovernik. Relational data generation with graph neural networks and latent diffusion models. In *NeurIPS 2024 Third Table Representation Learning Workshop*, 2024.

Zachary Izzo, Jinsung Yoon, Sercan O Arik, and James Zou. Provable membership inference privacy. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Zhanhong Jiang, Md Zahid Hasan, Nastaran Saadati, Aditya Balu, Chao Liu, and Soumik Sarkar. Balancing utility and privacy: Dynamically private SGD with random projection. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

James Max Kanter and Kalyan Veeramachaneni. Deep feature synthesis: Towards automating data science endeavors. In *2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Paris, France, October 19-21, 2015*, pp. 1–10. IEEE, 2015.

Antheas Kapenekakis, Daniele Dell'Aglio, Charles Vesteghem, Laurids Poulsen, Martin Bøgsted, Minos Garofalakis, and Katja Hose. Synthesizing accurate relational data under differential privacy. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 433–439. IEEE, 2024.

Shingo Kato, suharay, Will Cukierski, and haisland0909. Coupon purchase prediction, 2015. URL `https://kaggle.com/competitions/coupon-purchase-prediction`.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Ilmun Kim, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classification accuracy as a proxy for two-sample testing. *Annals of Statistics*, 49(1):411–434, 2021.

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.

Stefan Kramer, Nada Lavrač, and Peter Flach. Propositionalization approaches to relational data mining. *Relational data mining*, pp. 262–291, 2001.

Anton D. Lautrup, Tobias Hyrup, Arthur Zimek, and Peter Schneider-Kamp. Syntheval: a framework for detailed utility and privacy evaluation of tabular synthetic data. *Data Mining and Knowledge Discovery*, 39(1), 2024. doi: 10.1007/s10618-024-01081-4.

Jiayu Li, Zilong Zhao, Vikram Chundawat, Biplab Sikdar, and Y. C. Tay. Irg: Generating synthetic relational databases using deep learning with insightful relational understanding, 2024. URL `https://arxiv.org/abs/2312.15187`.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Ciro Antonio Mami, Andrea Coser, Eric Medvet, Alexander T. P. Boudewijn, Marco Volpe, Michael Whitworth, Borut Svara, Gabriele Sgroi, Daniele Panfilo, and Sebastiano Saccani. Generating realistic synthetic relational data through graph variational autoencoders, 2022. URL `https://arxiv.org/abs/2211.16889`.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3:127–163, 2000.

Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. Aim: an adaptive and iterative mechanism for differentially private synthetic data. *Proceedings of the VLDB Endowment*, 15(11), 2022.

Anna Montoya, LizSellier, Meghan O'Connell, Wendy Kan, and alokgupta. Airbnb new user bookings, 2015. URL `https://kaggle.com/competitions/airbnb-recruiting-new-user-bookings`.

Jan Motl and Oliver Schulte. The ctu prague relational learning repository, 2024. URL `https://arxiv.org/abs/1511.03086`.

Beata Nowok, Gillian M Raab, and Chris Dibben. synthpop: Bespoke creation of synthetic data in r. *Journal of statistical software*, 74:1–26, 2016.

Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

Wei Pang, Masoumeh Shafieinejad, Lucy Liu, Stephanie Hazlewood, and Xi He. ClavaDDPM: Multi-relational data synthesis with cluster-guided diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Theodore Papamarkou et al. Position: Topological deep learning is the new frontier for relational learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39529–39555. PMLR, 2024.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410, 2016.

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.

Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. Synthcity: facilitating innovative use cases of synthetic data in different data modalities, 2023a. URL `https://arxiv.org/abs/2301.07573`.

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

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.

Joshua Robinson, Rishabh Ranjan, Weihua Hu, Kexin Huang, Jiaqi Han, Alejandro Dobles, Matthias Fey, Jan Eric Lenssen, Yiwen Yuan, Zecheng Zhang, Xinwei He, and Jure Leskovec. Relbench: A benchmark for deep learning on relational databases. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Andrey Sidorenko, Michael Platzer, Mario Scriminaci, and Paul Tiwald. Benchmarking synthetic tabular data: A multi-dimensional evaluation framework, 2025. URL `https://arxiv.org/abs/2504.01908`.

Aivin V. Solatorio and Olivier Dupriez. Realtabformer: Generating realistic relational and tabular data using transformers, 2023. URL `https://arxiv.org/abs/2302.02041`.

Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18, 2010.

C Task, K Bhagat, and G Howarth. Sdnist v2: Deidentified data report tool. *National Institute of Standards and Technology*, 2023. doi: 10.18434/mds2-2943.

Paul Tiwald, Ivona Krchova, Andrey Sidorenko, Mariana Vargas-Vieyra, Mario Scriminaci, and Michael Platzer. Tabulararqn: A flexible and efficient auto-regressive framework for generating high-fidelity synthetic data, 2025. URL `https://arxiv.org/abs/2501.12012`.

Jan Tönshoff, Martin Ritzert, Eran Rosenbluth, and Martin Grohe. Where did the gap go? reassessing the long-range graph benchmark. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Vector Institute. Announcing the Winners of the MIDST Challenge! `https://vectorinstitute.github.io/MIDST/`, 2025. Accessed: 2025-05-15.

Will Cukierski Walmart. Walmart recruiting - store sales forecasting, 2014. URL `https://kaggle.com/competitions/walmart-recruiting-store-sales-forecasting`.

Joshua Ward, Chi-Hua Wang, and Guang Cheng. Data plagiarism index: Characterizing the privacy risk of data-copying in tabular generative models, 2024. URL `https://arxiv.org/abs/2406.13012`.

gdantel Wendy Kan. Telstra network disruptions, 2015. URL `https://kaggle.com/competitions/telstra-recruiting-network`.

Sohier Dane World Bank. World development indicators, 2019. URL `https://www.kaggle.com/datasets/theworldbank/world-development-indicators/data`.

Xiaoyu Wu, Yifei Pang, Terrance Liu, and Steven Wu. Winning the midst challenge: New membership inference attacks on diffusion models for tabular data synthesis, 2025. URL `https://arxiv.org/abs/2503.12008`.

Kai Xu, Georgi Ganev, Emile Joubert, Rees Davison, Olivier Van Acker, and Luke Robinson. Synthetic data generation of many-to-many datasets via random graph generation. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Zexi Yao, Nataša Krčo, Georgi Ganev, and Yves-Alexandre de Montjoye. The dcr delusion: Measuring the privacy risk of synthetic data. In *European Symposium on Research in Computer Security*, pp. 469–487. Springer, 2025.

EL Hacen Zein and Tanguy Urvoy. Tabular data generation: Can we fool XGBoost ? In *NeurIPS 2022 First Table Representation Workshop*, 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*, 2024.

Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):1–41, 2017.

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

A A Survey of Synthetic Relational Database Generation Methods

The **Synthetic Data Vault (SDV)** (Patki et al., 2016) introduced the first learning-based method for generating relational databases. The method is based on the Hierarchical Modeling Algorithm (HMA) synthesizer, which is a multivariate version of the Gaussian Copula method. The method converts all columns to a predefined set of distributions and selects the best-fitting one. To learn dependencies, columns are converted to a standard normal before calculating the covariances. Tables are modeled with a recursive conditional parameter aggregation technique, which incorporates child table covariance and column distribution information into the parent table. The method requires the relational structure or metadata, which has since become a common practice.

The work of Mami et al. (2022) leverages the graph representation of relational database using **Graph Variational Autoencoders**. They focus on the case of one primary table connected by an identifier to an arbitrary number of secondary tables. The approach begins by transforming categorical, datetime, and numeric attributes into a normalised numeric format using an invertible function. Subsequently, all tables' attributes are merged into a single-table, where rows from each table are vertically concatenated. This merged table, along with an adjacency matrix based on foreign key relations, forms a homogeneous graph representation of the dataset. Message passing is then applied to this graph representation using gated recurrent units (GRU). Following the message passing phase, the data is processed through a variational autoencoder, which encodes the joined table and random samples are taken from its latent space. These samples are then decoded back to the data space.

Composite Generative Models (Canale et al., 2022) propose a generative framework based on codecs for modeling complex data structures, such as relational databases. They define a codec as a quadruplet: C = (E,D,S,L), consisting of an encoder E producing embeddings and intermediate contexts, a decoder D for distribution representation, a sampler S and loss function L. The authors define the following codecs: Categorical and Numerical Codecs for individual columns, while composite data types are encoded using Struct and List Codecs, allowing for relational database synthesis. They also propose a specific implementation using causal transformers as generative models.

The **Row Conditional-TGAN (RCTGAN)** (Gueye et al., 2023) extends the conditional tabular GAN model (Xu et al., 2019) to relational databases. RCTGAN incorporates data from parent rows into the child table GAN model, allowing it to synthesise data conditionally on the connected parent table rows. The ability for conditional synthesis allows the method to handle various relationship schemas without additional processing. They enhance RCTGAN to capture the influence of grandparent rows on their grandchild rows, preserving this connection even when the relationship information is not transferred by the parent table rows. Database synthesis is based on the row conditional generator of RCTGAN model trained for each table. First, all parent tables are synthesised, followed by sampling the tables for which parents are already sampled. This allows using the synthesised parent rows as features when synthesizing child table rows.

The **Incremental Relational Generator (IRG)** (Li et al., 2024) uses GANs to incrementally fit and sample the relational dataset. They first define a topologically ordered sequence of tables in the dataset. Parent tables are modeled individually, while child tables undergo a three-step generation process. First, a potential context table is constructed by combining data from all related tables through join operations and aggregation. Then, the model predicts the number of child rows to be generated for each parent row, which they call its degree. They then extend the context table with corresponding degrees. Taking this table as context, they use a conditional synthetic tabular data generation model to generate the child table.

The **Realistic Relational and Tabular Transformer (REaLTabFormer)** (Solatorio & Dupriez, 2023) focuses on synthesizing single parent relational data and employs a GPT-2 encoder with a causal language model head to independently model the parent table. The encoder is frozen after training and used to conditionally model the child tables. Each child table requires a new conditional model, implemented as a sequence-to-sequence (Seq2Seq) transformer. The GPT-2 decoder with a causal language model head is trained to synthesise observations from the child table, accommodating arbitrary-length synthetic data conditioned on an input. While this method supports conditional synthesis of child rows, only one level is supported by this method.

Xu et al. (2023) propose a method for modeling many-to-many (M2M) datasets via random graph generation. They leverage a heterogeneous graph representation of the relational data and propose a factorization for modeling the graph representation incrementally. First, the edges of the graph are generated unconditionally using a random graph model. Second, one of the tables is generated conditionally on the topology of edges. One way to achieve such conditioning is by using a node embedding. Lastly, the remaining tables are generated using the conditional table model, which requires the generation of each node of the table based on the currently generated tables and all connections. They achieve this by using set embeddings to conditionally generate connected tables. The authors propose two variants using different conditional table models **BayesM2M** and **NeuralM2M**.

The **Cluster Latent Variable guided Diffusion Probabilistic Models (ClavaDDPM)** (Pang et al., 2024) utilizes classifier-guided diffusion models, integrating clustering labels as intermediaries between tables connected by foreign-key relations. The authors first propose a model for generating a single parent-child relationship. The connection between the tables is modeled by a latent variable obtained using Gaussian Mixture Model clustering. ClavaDDPM learns a diffusion process on the joint parent and latent variable distribution, followed by training a latent variable classifier on the child table to guide the diffusion model for the child table. Additionally, it includes a model to estimate child group sizes, to preserve relation cardinality. The authors then extend this to more parent-child constraints through bottom-up modeling and address multi-parent scenarios by employing majority voting to mitigate potential clustering inconsistencies.

Hudovernik (2024) adapts the tabular latent diffusion-based model TabSyn (Zhang et al., 2024) for conditional generation of relational databases. The method **Relational Graph-Conditioned Latent Diffusion**

(RGCLD) utilizes a heterogenous graph representation of a relational database. The rows of each table are represented as nodes of a particular type, and the foreign keys between tables are represented by edges connecting the nodes. The method trains a graph neural network for each table to encode the relationships between connected tables. The embeddings of the GNN are used to guide the diffusion process in the latent space. During sampling, the method generates tables sequentially based on a topological order defined by the dataset's schema.

Tiwald et al. (2025) propose **TabularARGN**, an auto-regressive model for generating synthetic data for flat and sequential tables. TabularARGN is a shallow any-order auto-regressive network architecture. The method homogenizes diverse datatypes by discretizing them and is trained by minimizing the categorical cross-entropy loss across the discretized attributes. TabularARGN is designed to generate tabular synthetic data by learning the full set of conditional probabilities across features in a dataset. The sequential table TabularARGN model can handle sequences of arbitrary lengths. TabularARGN's sequential table model can utilize a flat table as context during training and generation, enabling the synthesis of two-table setups, such as a flat table containing time-independent information (e.g., bank customer data) and a sequential table containing time-dependent data (the bank customer transaction histories). The method was open-sourced by the commercial provider Mostly.ai , while the method is specialized for sequential data, they also support generating multi-parent schemas. For these datasets, the method will retain the context for one of the parent tables and retain the referential integrity for the rest[1].

In contrast to neural-based approaches described above, marginal-based approaches focus on preserving low-dimensional measurements of the dataset, such as marginal queries, often under differential privacy (DP), and are typically evaluated by the accuracy of synthesized marginal queries. While marginal-based methods for single-table synthesis are well-established (Zhang et al., 2017; McKenna et al., 2022; Cai et al., 2021), their extension to relational data is a more recent focus. For instance, PrivLava(Cai et al., 2023) achieves DP in relational data through graphical models with latent variables to represent inter-table dependencies. Following this, (Kapenekakis et al., 2024) proposed a method for synthesizing relational data, with a focus on data with sequential aspects like electronic health records. More recently, (Alimohammadi et al., 2025) adapt single-table DP generators to relational data by modeling relationships with a learned bi-adjacency matrix, and PrivPetal (Cai et al., 2025) synthesizes DP relational data by modeling a flattened version of the table using permutation relations.

B Synthetic Relational Database Generation Benchmark

We provide our work as a Python package **SyntheRela**. The main goal of the package is the evaluation of the quality of synthetic relational databases. We can compare multiple methods across multiple databases with the *Benchmark* class or evaluate a single method on a single database with the *Report* class. All of the results of the benchmark are saved as JSON files and then parsed by our package for results summarization and visualization. The package is open source under the MIT license and can be easily extended with new methods, evaluation metrics, or databases.

B.1 Evaluation Metrics

We list the evaluation metrics for data fidelity and utility currently supported in our benchmark in Table 9, based on the granularity of the data they evaluate. For single column fidelity, we use the Kolmogorov-Smirnov and $\mathcal{X}^2$ statistical tests, total variation, Hellinger, Jensen-Shannon, and Wasserstein distances, alongside a single column C2ST. We include column pair correlation similarity, maximum mean discrepancy, pairwise correlation difference, and C2ST for single-table fidelity. We evaluate multi-table fidelity with cardinality shape similarity and k-hop correlation similarity alongside C2ST-Agg. We also implement the Parent-Child C2ST for comparison with related work. Single-column utility is generally covered by fidelity metrics and not evaluated in related work. For single-table utility, we implement tabular machine learning utility metrics, and for multi-table utility, we include RDL-utility.

[1]For details on multi-table generation see `https://mostly.ai/docs/generators/configure/set-table-relationships/multi-table`.

Table 9: **Evaluation metrics supported in the benchmark.**

	Single Column	**Single-Table**	**Multi-Table**
Statistical	KS Test, $\mathcal{X}^2$ Test	Column pair correlations	cardinality shape similarity
Distance	Total Variation, Hellinger, Jensen-Shannon, Wasserstein	Maximum Mean Discrepancy, Pairwise Correlation Difference	k-hop correlation similarity
Detection	C2ST	C2ST	C2ST-Agg, Parent-Child C2ST
Utility	/	Tabular ML-Utility	Relational DL-Utility

B.1.1 C2ST with Aggregation

Fidelity methods are concerned with measuring the similarity between two databases with the same schema but different data. Typically, these will be the real database $\mathbb{D}_{\text{REAL}}$ and a synthetic database $\mathbb{D}_{\text{SYN}}$, with the goal of detecting if, to what extent, and where the synthetic data differ from the real data.

Let a relational database $\mathbb{D}$ be a collection of tables $\mathcal{T} = \{T_1, ..., T_n\}$ and a schema $\mathcal{S} = (\mathcal{R}, \mathcal{A})$, where $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{T}$ are the relations between the tables and $A_{T_i} = \{a_1^{T_i}, \dots, a_l^{T_i}\} \in \mathcal{A}$ define the tables' attributes. Each table is a set $T = \{v_1, ..., v_{n_T}\}$ consisting of elements v_i called rows. Each row $v \in T$ has three components $v = (p_v, \mathcal{K}_v, x_v)$. A **primary key** p_v that uniquely identifies the row v; the set of **foreign keys** $\mathcal{K}_v = \{p_{v'} : v' \in T' \text{ and } (T, T') \in \mathcal{R}\}$, where $p_{v'}$ is the primary key of the row v'; and the set of **values** $x_v = \{(a, x) : a \in A_T\}$ corresponding to attributes of table T.

Algorithm 1 describes how we add aggregations to the target table. For each child table, we add *CountRows*, a count of the number of child rows corresponding to a parent row. For each attribute (i.e. column) in each child table, we compute an aggregation attribute (*mean*, *count*, etc.). The aggregation attributes are added to the target table. In practice, different aggregation functions may be applied, as long as they maintain the i.i.d. assumption of the data. In our benchmark, we use column means for numerical columns and the number of distinct categories in categorical columns along with the child row count.

Algorithm 1 Relational Aggregation.

```
Require: relational database $\mathbb{D}$ with tables $\mathcal{T}$ and relational schema $\mathcal{S} = \{\mathcal{R}, \{A_{T_1} \dots A_{T_n}\}\}$
Require: target table $T$
 1: aggregationAttributes ← [ ]
 2: for each $C_i \in \{C : (C, T) \in \mathcal{R}\}$ do
 3:     Add $CountRows(C_i, T)$ to aggregationAttributes      ▷ Count number of child rows for each row in $T$.
 4:     for each $a_j^{C_i} \in A_{C_i}$ do                      ▷ Iterate through child table columns.
 5:         Add $Agg(C_i, a_j^{C_i}, T)$ to aggregationAttributes   ▷ Compute aggregation (e.g., mean, distinct...).
 6:     end for
 7: end for
 8: for each $v \in T$ do
 9:     for each $\mathbf{a} \in$ aggregationAttributes do
10:         Add ($\mathbf{a}$.name, $\mathbf{a}$.value) to $v$           ▷ Append computed aggregation attributes.
11:     end for
12: end for
13: return $T_i$                                              ▷ Return table with aggregations applied.
```

Our implementation allows us to directly control how many levels of aggregations we add (similarly to k-hop correlation similarity (Pang et al., 2024)). Each level accounts for one application of Algorithm 1. When aggregating for 1 level, only the information directly from the table's children is aggregated; at the second level, the aggregated grandchild columns are also added to the table. In our benchmark, we choose to add only one level of aggregation, as most methods struggle to preserve the relationships at distance $k = 1$.

It has been shown that the accuracy-based approach to two-sample testing is consistent and controls for Type I error and (asymptotically) Type II error (see Kim et al. (2021) for theoretical results and a summary of empirical results). In practice, we are also interested in finite sample behavior. Experiment-based recommendations show that the approach should have an advantage in power when the data are well-structured or we have a lot of data, or when it is difficult to specify a test statistic, which is very common for high-dimensional data. Therefore, the large, higher-dimensional and structured nature of relational data is a perfect fit for C2ST.

B.2 Datasets

B.2.1 Relational Databases and Sampling Procedures

An important issue with evaluating relational data is that representative sampling is difficult (Buda et al., 2013; Gemulla et al., 2008). If the dataset does not include a time component or if the relationships are non-linear, the sampling becomes non-trivial and directly impacts the performance of the generative method. Even if the data have a strict hierarchy between tables, the rows in a child table are related via their parent, which breaks the assumption of *i.i.d.* sampling. This makes splitting the dataset into a representative train and test set difficult. Consequently, the methods for synthesizing relational databases are typically trained using the entire original dataset.

We organize the datasets used in related work based on the structure of their relational schema. Datasets using only linear relationships (one parent and one child table) include Airbnb (Montoya et al., 2015) and Rossmann Store Sales (FlorianKnauer, 2015). While this structure may be sufficient for some practical applications, Gueye et al. (2023) and Xu et al. (2023) highlight the need for methods supporting more complex, multiple-parent relational structures found in datasets like *MovieLens* (Harper & Konstan, 2015) and *World Development Indicators* (World Bank, 2019). Datasets including multiple child tables include *Telstra Network Disruptions* (Wendy Kan, 2015), *Walmart Recruiting - Store Sales Forecasting* (Walmart, 2014), and *Mutagenesis* (Debnath et al., 1991). Datasets with multiple children and parents include *Coupon Purchase Prediction* (Kato et al., 2015), *World Development Indicators* (World Bank, 2019), *MovieLens* (Harper & Konstan, 2015), *Biodegradability* (Blockeel et al., 1999) and *Berka* Berka et al. (2000).

B.2.2 Benchmark Datasets

Table 1 summarizes the relational datasets used in our benchmark. Six datasets are from related work and we add the *Cora* dataset by McCallum et al. (2000), which contains a simple yet challenging relational schema, and **F1** F1 (2021) from the RelBench benchmark. We include 2 datasets per hierarchy type to progressively add complexity in generation. The datasets used in our evaluation are diverse in terms of the number of columns, tables and relationships.

The **Airbnb** (Airbnb, 2015) dataset includes user demographics, web session records, and summary statistics. It provides data about users' interactions with the platform, with the aim of predicting the most likely country of the users' next trip. See Figure 9a for schema.

The **Berka** (also known as the Financial dataset) Berka et al. (2000) dataset contains 606 successful and 76 unsuccessful loans along with their information and transactions. The standard task is to predict the loan outcome for finished loans (A vs B) at the time of the loan start. See Figure 12 for schema.

The **Biodegradability** dataset (Blockeel et al., 1999) comprises a collection of chemical structures, specifically 328 compounds, each labeled with its half-life for aerobic aqueous biodegradation. This dataset is intended for regression analysis, aiming to predict the biodegradation half-live activity based on the chemical features of the compounds. See Figure 13 for schema.

The **Cora** dataset (McCallum et al., 2000) is a widely-used benchmark dataset in the field of graph representation learning. It consists of academic papers from various domains. The dataset consists of 2708 scientific publications classified into one of seven classes and their contents. The citation network consists of 5429 links. See Figure 10b for schema.

The **F1** dataset (F1, 2021) contains Formula 1 racing data and statistics dating back to 1950. It contains information on drivers, constructors, race results, and standings covering every season in F1 history. See Figure 11 for schema.

The **IMDB MovieLens** dataset (Harper & Konstan, 2015) comprises information on movies, actors, directors, and users' film ratings. The dataset consists of seven tables, each containing at least one additional feature besides the primary and foreign keys. See Figure 14 for schema.

The **Rossmann Store Sales** (FlorianKnauer, 2015) features historical sales data for 1115 Rossmann stores. The dataset consists of two tables connected by a single foreign key. This makes it the simplest type of relational dataset. The first table contains general information about the stores and the second contains sales-related data. See Figure 9b for schema.

The **Walmart** dataset (Walmart, 2014) includes historical sales data for 45 Walmart stores across various regions. It includes numerical, date-time and categorical features across three connected tables *store*, *features* and *depts*. The dataset is from a Kaggle competition, with the task of predicting department-wide sales. See Figure 10a for schema.

B.3 Comparison with Existing Evaluation Tools

Several packages for evaluating synthetic tabular data exist (Nowok et al., 2016; Task et al., 2023; Lautrup et al., 2024; Sidorenko et al., 2025) with the most popular and comprehensive package being Synthcity (Qian et al., 2023a;b). It supports many statistical, privacy and detection-based (with several different models) metrics.

The only package that supports multi-table evaluation is SDMetrics (DataCebo, 2022). It includes multi-table metrics cardinality shape similarity and parent-child detection with logistic detection and support vector classifier. The package is not easy to extend and limits the adaptation of metrics. We re-implement detection metrics (discriminative detection, aggregation detection, and parent-child detection) to be used with an arbitrary classifier supporting the Scikit-learn (Pedregosa et al., 2011; Buitinck et al., 2013) classifier API. In SDMetrics, the results of different metrics are aggregated into a single-value, which limits the comparison of individual metrics between the methods and datasets. We re-implement the distance and statistical metrics so that each statistic, p-value, and confidence interval is easy to access.

Our benchmark package can be easily extended with new methods, metrics, and datasets. The process for adding custom metrics and new datasets is described in `Anonymized`.

B.4 License and Privacy

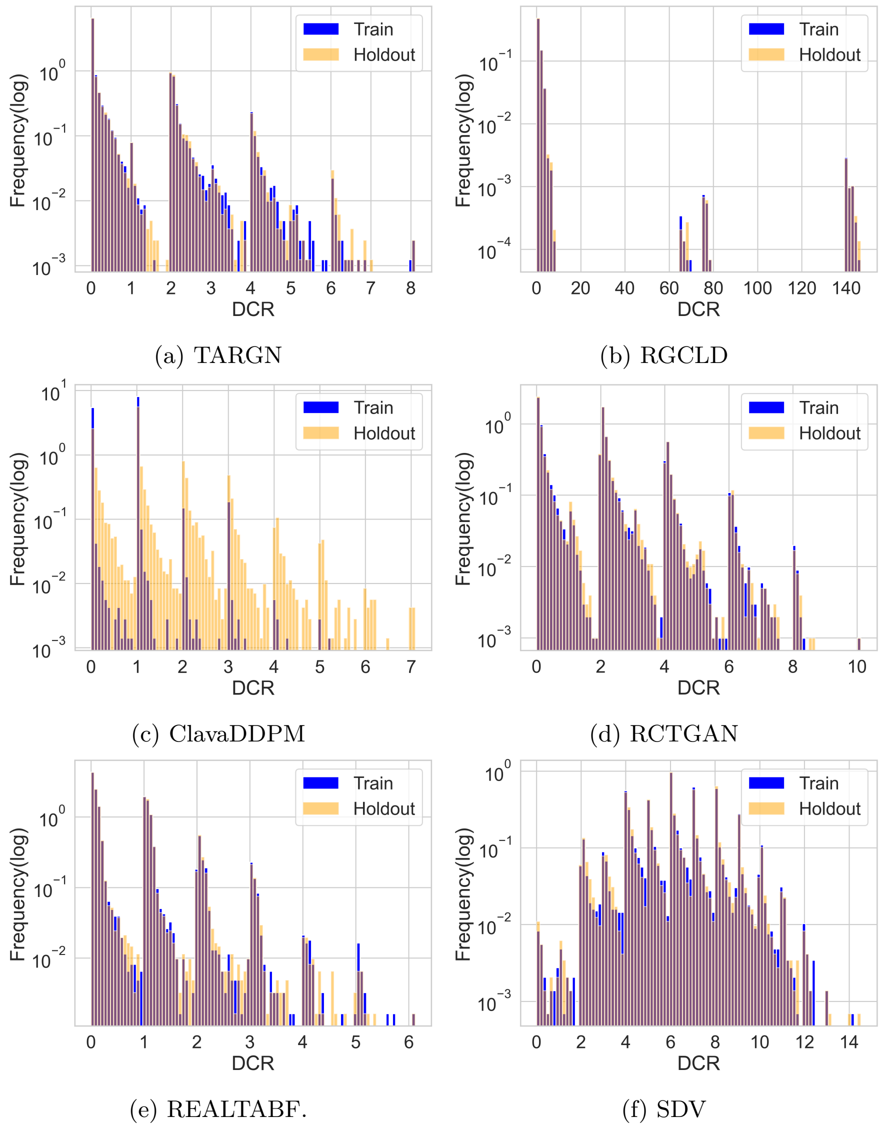
We obtain the *Airbnb, Biodegradability, Cora, IMDB MovieLens, Rossmann and Walmart* datasets from the public SDV relational demo datasets repository (`https://docs.sdv.dev/sdv/single-table-data/data-preparation/loading-data`, accessed June 6th, 2024.). The SDV project is licensed under the Business Source License 1.1 (`https://github.com/sdv-dev/SDV?tab=License-1-ov-file#readme`), which allows use for research purposes. The Berka dataset was obtained from the CTU Relational Dataset Repository (Motl & Schulte, 2024) an open-access repository of relational databases. The F1 dataset was obtained using the official RelBench implementation Robinson et al. (2024). It is released under the CC-BY-4.0 license (`https://github.com/f1db/f1db?tab=CC-BY-4.0-1-ov-file`). We remove all textual columns from the F1 dataset as relational generative methods are generally unable to generate those. We manually check all of the datasets to ensure they do not include any personally identifiable information. Some of the datasets contain processed columns, including aggregations of numerical values and connected table rows (e.g., nb_rows_in_{related table}). The authors of SDV (Patki et al., 2016) confirmed that these aggregations are not part of the original datasets, so we post-process all of the datasets to include only the columns found in their original form and update the metadata accordingly. We adapt some of the metrics from the SDMetrics (DataCebo, 2022) (MIT License) and Synthcity (Qian et al., 2023a;b) (Apache-2.0 License) synthetic data evaluation tools. We acknowledge Flaticon.com for providing the icons used in Figure 1.

C Additional Experiments

D Privacy Evaluation

In this section, we present additional results from our privacy evaluation experiments, providing a more granular view of the behavior of each generative model. We visualize the full DCR distributions for all evaluated methods, and report the precision and recall metrics of the membership inference attacks.

As shown in Figures 5 and 6, the DCR distributions for most methods exhibit substantial overlap between synthetic–training and synthetic–holdout distances, indicating no systematic data copying. However, ClavaDDPM on the *users* table shows a noticeable skew toward lower distances between synthetic and training records, consistent with the elevated DCR score reported in Table 7, suggesting potential data copying and privacy issues.

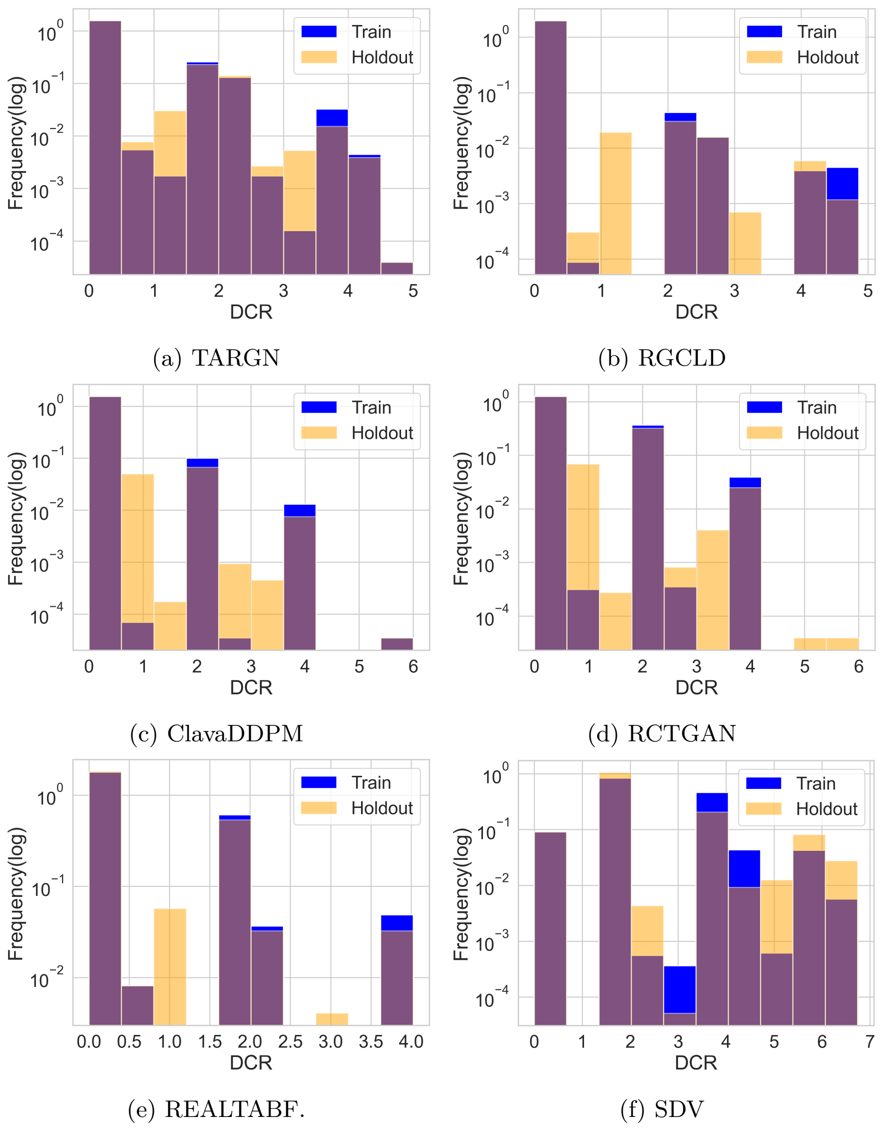


(a) TARGN (b) RGCLD

(c) ClavaDDPM (d) RCTGAN

(e) REALTABF. (f) SDV

Figure 5: **DCR Distributions** for the *users* table in the anonymized *Airbnb* dataset. The y-axis is log-scaled for a better comparison.



(a) TARGN (b) RGCLD

(c) ClavaDDPM (d) RCTGAN

(e) REALTABF. (f) SDV

Figure 6: **DCR Distributions** for the *sessions* table in the anonymized *Airbnb* dataset. The *sessions* table consists primarily of categorical columns, resulting in highly discretized DCR distributions. Despite this, most methods maintain a reasonable level of privacy. The y-axis is log-scaled for a better comparison.

In Tables 11 and 10 we report the *recall* and *precision* of an adversarial classifier, respectively. Recall reflects the probability that rows from the training data are correctly identified as such, and therefore serves as a direct measure of membership disclosure risk. Precision, on the other hand, reflects the probability that a record predicted as a member truly originates from the training set, and serves as a measure of the reliability of the attack's positive predictions (a scores of 50% is analogous to random guessing).

We also examine an additional aspect of privacy: the potential of the C2ST test to detect data copying, similar to the DCR score. To explore this, we conduct a data copying simulation across the benchmark datasets. The experimental setup and a discussion on how common implementations of C2ST metrics may inadvertently mask such privacy risks are provided in Appendix D.3.

Table 10: **MIA Recall** denotes the probability that the attack predicts a record as a member given that it is truly a member of the training set. Values significantly above 50% indicate privacy concerns.

Method	Sessions	Users
TARGN	11.27 ± 0.15	17.58 ± 2.06
RGCLD	2.14 ± 0.17	61.31 ± 2.21
CLAVA	30.03 ± 2.74	37.49 ± 0.71
RCTGAN	28.16 ± 0.85	98.55 ± 0.35
REALTABF.	3.14 ± 0.38	0.19 ± 0.07
SDV	0.03 ± 0.00	0.03 ± 0.02
SMOTE	91.02 ± 0.17	85.24 ± 0.34

Table 11: **MIA Precision** denotes the probability that a record is truly a member of the training set given that the attack predicted it as a member. We estimate uncertainty with standard errors.

Method	Sessions	Users
TARGN	92.14 ± 0.48	99.49 ± 0.12
RGCLD	86.04 ± 0.58	99.78 ± 0.05
CLAVADDPM	92.81 ± 0.71	99.94 ± 0.03
RCTGAN	93.11 ± 0.21	99.72 ± 0.09
REALTABF.	70.76 ± 6.15	62.67 ± 17.62
SDV	85.00 ± 10.00	35.00 ± 21.79
SMOTE	97.14 ± 0.18	99.86 ± 0.03

D.1 Single Column Performance

We evaluate the marginal distributions of individual columns by evaluating the column shapes and using the detection metric. Results are summarized in Table 12. TabularARGN performs best on modeling marginal distributions. We hypothesize this is a result of their preprocessing, which removes outliers and discretizes all columns. The method is closely followed by ClavaDDPM, which models individual tables best.

Table 12: **Single Column Results.** We report the detection accuracy for C2ST (lower is better), and the KS/TV complement for column shapes (higher is better). For each dataset and metric, we report the average across all tables for three independent samples. SDV exceeds the sampling time limit on IMDB (TLE) and "-" denotes a method is unable to generate the dataset. The best result is **in bold** and results within one standard deviation underlined.

Dataset	Metric	TARGN	RGCLD	ClavaDDPM	RCTGAN	REALTABF.	SDV	MARE
Airbnb	C2ST (↓)	**51.88**± 0.09	52.40± 0.36	55.34± 0.02	56.37± 0.02	61.68± 0.61	70.36± 0.03	58.57$\pm$5e-3
	Shapes (↑)	**95.70**± 0.05	95.57± 0.63	94.42± 0.01	89.18± 0.17	71.66± 0.92	59.37± 0.04	83.82$\pm$3e-3
Rossmann	C2ST (↓)	**51.07**± 0.19	52.76± 0.33	54.79± 0.08	54.98± 0.23	53.62± 0.17	59.99± 0.10	-
	Shapes (↑)	**96.96**± 0.19	95.83± 0.55	94.05± 0.07	91.31± 0.04	90.65± 0.38	81.05± 0.19	
Walmart	C2ST (↓)	59.25± 0.29	59.34± 0.80	**52.60**± 0.55	64.12± 0.40	59.08± 0.31	66.71± 0.58	-
	Shapes (↑)	89.09± 0.33	88.43± 1.00	**92.21**± 0.52	82.31± 0.51	81.71± 0.40	81.80± 0.10	
Berka	C2ST (↓)	58.64± 0.26	60.00± 1.30	**47.25**± 0.12	57.35± 0.20	-	70.77± 0.17	-
	Shapes (↑)	82.20± 0.26	78.91± 2.42	**91.62**± 0.10	81.90± 0.38		56.27± 0.29	
F1	C2ST (↓)	61.61± 0.04	**57.69**± 0.67	60.76± 0.20	63.71± 0.22	-	79.85± 0.22	-
	Shapes (↑)	84.71± 1.15	**91.04**± 1.51	84.63± 0.28	89.68± 0.40		52.62± 0.57	
IMDB	C2ST (↓)	50.36± 0.12	53.02± 1.50	**49.86**± 0.06	53.74± 0.08	-	-	-
	Shapes (↑)	98.40± 0.14	93.61± 2.75	**99.01**± 0.05	92.70± 0.09			
Biodegradability	C2ST (↓)	**55.42**± 0.48	61.85± 1.57	-	57.84± 0.40	-	64.24± 0.37	-
	Shapes (↑)	90.85± 0.14	84.53± 3.42		**90.91**± 0.40		79.46± 0.47	
Cora	C2ST (↓)	50.94± 0.37	52.98± 1.03	-	**48.97**± 0.14	-	75.45± 0.16	-
	Shapes (↑)	92.65± 0.12	88.27± 1.94		**96.38**± 0.15		50.24± 0.17	

D.2 Shortcomings of Logistic Detection

As explained in Section 2.2 a significant limitation of LD is its inability to capture interactions between columns. It can thus assign a perfect fidelity score to a dataset that is completely corrupted. In this section, we empirically show this shortcoming. We conduct the experiment by selecting a table from each dataset (with the exception of CORA in which no table has two columns, which are not primary or foreign keys). We first select the table and split it in half to simulate the original table and a perfectly generated (by the underlying DGP) synthetic table. We then copy the "generated" table and randomly shuffle values in each column, completely ruining the fidelity of the dataset while keeping the marginal distributions intact. We then evaluate the perfectly generated and shuffled datasets using LD and C2ST using XGBoost. The results are visualized in Figure 7.

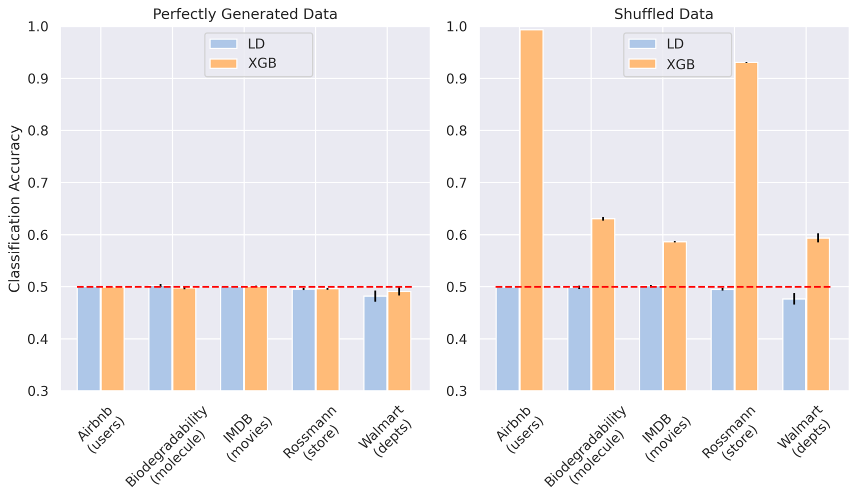


Figure 7: **Issues with logistic detection.** For each dataset, we simulate a perfectly generated table by splitting the original table in half. We copy one part of the table and shuffle the values in each column and thus completely ruin the fidelity of the table. While the XGBoost classifier can almost perfectly segment the corrupted rows, logistic regression assigns both of the datasets the same score.

Notably LD assigns both versions of the dataset the same score, labeling them indistinctive from the original data. If the fidelity aspect of interest would be solely the marginal distributions, the LD results would be more appropriate than those of C2ST using XGBoost (as marginals are identical in both datasets). However, given that we are interested in single-table fidelity, our experiment showcases a fundamental shortcoming of LD as a measure of single-table fidelity.

D.3 Classifier Accuracy as a Data Copying Diagnostic

We investigate how the accuracy of the discriminative model in classifier two sample testing can be used to diagnose data copying in Figure 8. We also demonstrate how the classifier performance commonly reported in LD ($2 \cdot \max(\text{AUC}, \frac{1}{2}) - 1$) masks this issue.

As in the previous experiment, we simulate a perfect synthetic generating a dataset by splitting the original table in half. However, instead of introducing corruption into the second half, we create an exact copy of the original data (i.e., the first half). The commonly used LD implementation fails to detect data copying and assigns the copied data a perfect score. In contrast, C2ST (with XGBoost) accuracy successfully detects data copying as accuracy drops significantly below 50%. We then examine the behaviour of C2ST when only a portion of the data is copied. We keep a portion of the dataset as an identical copy and sample the rest of the values from the "perfectly generated" half. For most of the datasets, even when a relatively low percentage of the data is copied, the model detects the duplication.

E Relational Deep Learning Utility

We adopt the RelBench (Robinson et al., 2024) relational deep learning benchmark to support synthetic data into the relational deep learning utility framework. It is designed to support training GNN models both on real-world and synthetic relational datasets, and evaluating on a real held-out test set, in a *train-on-synthetic-evaluate-on-real* paradigm. RDL-utility allows integration of new databases and models.

The framework creates two separate dataset-task pairs. One pair includes the synthetic dataset with training and validation splits used for model training. The other pair contains the real dataset with an additional test

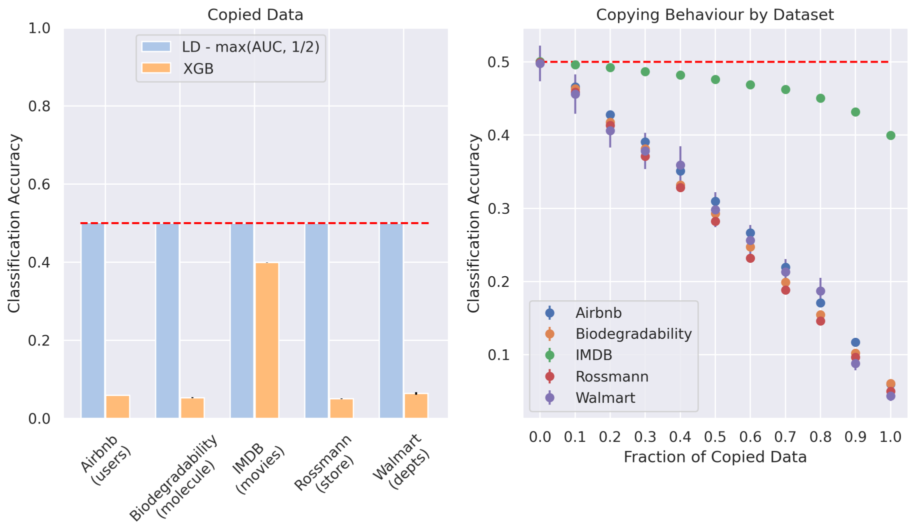


Figure 8: **Detecting data copying with C2ST.** The left plot demonstrates how the error estimation of LD $(2 \cdot \max(\text{AUC}, \frac{1}{2}) - 1)$ masks data copying, while C2ST with XGBoost detects it across all datasets. In the right plot, we observe how copying only a fraction of the original data affects discriminative model accuracy, with accuracy consistently decreasing as more data is duplicated.

split used for final evaluation. Both pairs follow the same schema but are kept separate to avoid data leakage. Graph construction converts relational tables into heterogeneous graphs for both synthetic and real datasets. The process maps tables to node types and foreign key relations to edge types. Temporal splits are applied consistently to ensure only past data is available during training and testing. GNN models are trained with temporal awareness, restricting access to data strictly before the test time.

E.1 Utility Tasks

Table 13: **Task statistics for relational datasets used in evaluation.**

Dataset	Task name	Task type	#Rows of Task Table			#Unique Entities	%Train-Val/Test Entity Overlap
			Train	Validation	Test		
rossmann	Autocomplete	Regression	46,750	9,350	28,985	46,750	0.0
Walmart	Autocomplete	Regression	5,925	2,952	11,946	5,925	0.0
Airbnb	Autocomplete	Classification	1,661	217	999	1,661	0.0
berka	Autocomplete	Classification	328	196	158	328	0.0
f1	Driver-top3	Classification	1,353	588	726	92	50.0

The summary of the RDL-utility tasks is displayed in Table 13. We define the AutoComplete task for the Rossmann, Airbnb, Walmart and Berka datasets based on the predefined tasks by the dataset authors from Kaggle competitions or other dataset information. The task objectives per datasets are as defined as follows:

- **Rossmann** [Regression]: We predict the daily number of customers for each store.
- **Walmart** [Regression]: We predict the weekly sales for each department of each store.
- **Airbnb** [Classification]: We predict whether a user has already made a booking or not.
- **Berka** [Classification]: We predict the binary status of the loan (successful or unsuccessful).
- **F1** [Classification]: We predict whether a driver will qualify in the top-3 for a race in the next month, which is one of the original RelBench predictive tasks for this dataset.

E.2 Hyperparameter Tuning

As GNN models are sensitive to hyperparameters Tönshoff et al. (2024), we perform an exhaustive hyperparameter search across all datasets and GNN variants. We use the `Optuna` library Akiba et al. (2019) to create a hyperparameter search for each dataset and GNN variant pair on the original data. We run 100 experiments for each dataset-GNN pair and optimize the parameters specified in Table 14.

Table 14: **Hyperparameter search specification for GNN training.** We run 100 experiments for each dataset-GNN variant pair.

Hyperparameter	Type	Search Space / Values
Learning rate	Categorical	{0.0001, 0.001, 0.01, 0.1, 0.5, 1.0}
Number of layers	Categorical	{1, 2, 3}
Number of neighbors	Categorical	{-1, 128}
Weight decay	Continuous (log scale)	[1e-9, 0.01]
Number of MLP layers	Categorical	{1, 2, 3}
Aggregation method	Categorical	{sum, mean, max, min}

The best hyperparameters used in RDL-utility are specified in the Reproducibility section of the Appendix in Table 15.

Table 15: **RDL-utility hyperparameter specification.** We list the best hyperparameters from the hyperparameter search of 100 runs on original data.

GNN	Dataset	LR	Layers	Neighbors	Weight Decay	MLP Layers	Aggr
Hetero-GAT	Rossmann	0.001	2	128	5.4e-05	1	sum
	Walmart	1.0	2	128	2.6e-06	1	sum
	Airbnb	0.0001	1	128	2.7e-07	1	min
	Berka	0.01	2	128	1.3e-07	2	mean
	F1	0.5	1	128	6.4e-05	2	mean
Hetero-GATv2	Rossmann	0.001	2	128	2.1e-08	3	sum
	Walmart	0.5	2	128	1.6e-07	2	max
	Airbnb	0.0001	1	-1	1.8e-09	2	max
	Berka	0.1	2	128	8.7e-06	2	min
	F1	0.1	1	-1	3.5e-07	3	sum
Hetero-GIN	Rossmann	1.0	2	-1	1.3e-05	1	sum
	Walmart	0.1	2	128	1.3e-08	2	mean
	Airbnb	0.001	1	-1	3.4e-04	2	max
	Berka	0.1	2	128	1.6e-05	2	max
	F1	0.5	3	-1	1.5e-06	1	mean
Hetero-GraphConv	Rossmann	0.5	1	128	1.0e-04	3	min
	Walmart	1.0	2	128	2.5e-08	2	sum
	Airbnb	0.001	1	-1	3.0e-04	1	sum
	Berka	0.01	2	128	2.3e-09	3	max
	F1	0.1	2	128	9.9e-07	3	sum
Hetero-GraphSAGE	Rossmann	0.1	1	-1	0.00152	3	sum
	Walmart	0.5	3	128	2.1e-09	2	mean
	Airbnb	0.001	2	-1	8.5e-07	1	max
	Berka	0.001	3	-1	8.5e-07	3	max
	F1	0.1	1	128	4.0e-07	3	sum
RelGNN	Rossmann	1.0	1	128	2.3e-05	3	sum
	Walmart	0.1	2	128	5.2e-05	2	sum
	Airbnb	0.001	2	-1	4.4e-06	1	sum
	Berka	0.01	2	128	5.6e-06	2	min
	F1	0.01	2	-1	6.5e-05	1	mean

E.3 Utility Baselines

To assess the effectiveness of RDL-utility in quantifying the utility performance of synthetic relational databases, we implement two baseline approaches that measure utility on single tables.

The single-table baseline evaluates utility using only the entity table. As the predictive model, we use LightGBM Ke et al. (2017), a gradient boosting decision tree implementation. The model is trained with ten iterations of automated hyperparameter tuning on the training and validation splits, and the best configuration is evaluated on the test set.

The second baseline implements the automated feature generation deep feature synthesis (Kanter & Veeramachaneni, 2015). Here, relational structure is incorporated through automated feature generation: for categorical columns, we apply `count` aggregation; for numerical columns, we use `mean` and `sum`; and for datetime columns, we extract the `month` and `year`. Aggregations are applied recursively throughout the database to produce a flattened feature table. This table is then used to train LightGBM with the same procedure as in the single-table baseline, allowing a direct comparison between learning from the entity table alone and learning with additional relational features.

F Experimental Setup

F.1 Computational Resources

The generative methods were trained on NVIDIA 32GB V100S GPUs and H100 80GB GPUs. The total number of GPU hours spent across all experiments is approximately 500. Results which do not require a GPU were run on machines running AMD EPYC 7702P 64-Core Processor with 256GB of RAM. All experiments were performed on an internal HPC cluster.

F.2 Reproducibility

F.2.1 Datasets and Data Splitting

Scripts for downloading the datasets and their metadata in the SDV format (Patki et al., 2016) are available in the project repository, as well as the corresponding synthetic data samples for all methods to enable the reproduction of the benchmark results.

We opt not to split the datasets into train, test, and validation sets for generative model training. When no temporal information is included and the structure is non-linear the representative sampling in relational datasets is non-trivial. We delegate this to future work.

Due to computational limits (also reported by Solatorio & Dupriez (2023)), we subsample the Rossmann Store Sales, Airbnb, and Walmart datasets. Additionally, we subsample the Berka and F1 datasets for the purposes of obtaining a held-out test set for our utility experiments.

- **Rossmann Store Sales**: Subsampled on table *historical*, column *Date* by taking the rows of a two month period from *2014-07-31* to *2014-09-30*, similarly to Solatorio & Dupriez (2023).
- **Airbnb**: Subsampled the dataset by only including the users who have less than 50 sessions and then sampled 10k users, as done by Solatorio & Dupriez (2023).
- **Walmart**: Subsampled on tables *departments* and *features* on the column *Date* by taking the rows from January 2012.
- **Berka**: We use the data prior to *1998-01-01* and use the *1998* data as the test set.
- **F1**: We use the data prior to *2010-01-01* for training. This corresponds to the test set timestamp in the official RelBench implementation [2].

[2]`https://relbench.stanford.edu/datasets/rel-f1/`

F.2.2 Experimental Details and Hyperparameters

To provide some quantification of the variability from the non-deterministic nature of the methods, we generated synthetic data for each of the methods for each of the datasets 3 times with different fixed random seeds. We ran the benchmark for each replication.

Scripts for reproducing the generative model training and instructions for training commercial methods are included in the project repository.

It is possible that better performance could be achieved by investing more effort into parameter tuning. However, due to our choice to not split the data, it was not clear how to optimize hyperparameters; therefore, we selected default hyperparameters for all methods (see Tables 16 and 17).

G Broader impacts

The research introduces a benchmark for evaluating the quality of synthetic relational databases. This could improve the development of synthetic data, which may be beneficial in fields with privacy restrictions, for example, healthcare (Appenzeller et al., 2022), finance (Assefa et al., 2020), education (Bonnéry et al., 2019), and in cases of limited or biased data (Ntoutsi et al., 2020; Rajpurkar et al., 2022). However, there are possible negative aspects. Synthesis processes might have weaknesses that affect privacy and amplify biases already present in the original data. Also, synthetic data that resembles real data could be misused. Therefore, further work on privacy protection, bias reduction, and detection of synthetic data is needed for its proper use.

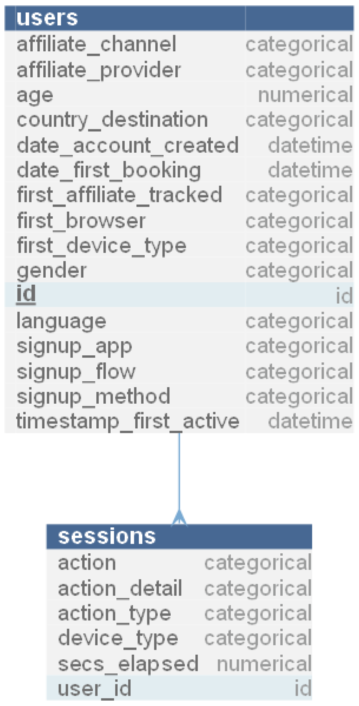
Table 16: **Hyperparameter specification** for RCTGAN (Gueye et al., 2023), SDV (Patki et al., 2016) and RealTabFormer Solatorio & Dupriez (2023).

model	hyperparameter	value
RCTGAN	embedding_dim	128
	generator_dim	(256, 256)
	discriminator_dim	(256, 256)
	generator_lr	0.0002
	generator_decay	1e-06
	discriminator_lr	0.0002
	discriminator_decay	1e-06
	batch_size	500
	discriminator_steps	1
	epochs	1000
	pac	10
	grand_parent	True
	field_transformers	None
	constraints	None
	rounding	"auto"
	min_value	"auto"
	max_value	"auto"
SDV	locales	None
	verbose	True
	table_synthesizer	"GaussianCopulaSynthesizer"
	enforce_min_max_values	True
	enforce_rounding	True
	numerical_distributions	{}
	default_distribution	"beta"
REALTABFORMER	epochs	100
	batch_size	8
	train_size	0.95
	output_max_length	512
	early_stopping_patience	5
	early_stopping_threshold	0
	mask_rate	0
	numeric_nparts	1
	numeric_precision	4
	numeric_max_len	10
	evaluation_strategy	"steps"
	metric_for_best_model	"loss"
	gradient_accumulation_steps	4
	remove_unused_columns	True
	logging_steps	100
	save_steps	100
	eval_steps	100
	load_best_model_at_end	True
	save_total_limit	6
	optim	"adamw_torch"

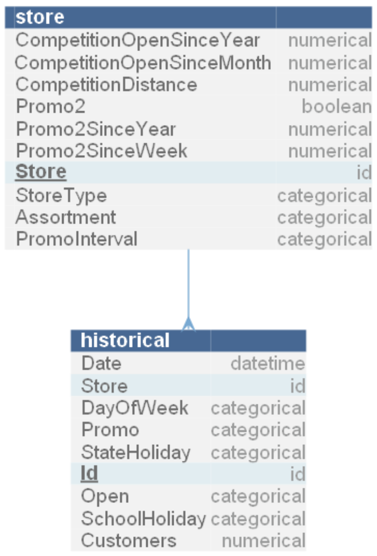
Table 17: **Hyperparameter specification** for TabularARGN (Tiwald et al., 2025), ClavaDDPM (Pang et al., 2024) and RGCLD Hudovernik (2024).

model	hyperparameter	value
TARGN (API)	Configuration presets	Accuracy
	Max sample size	100%
	Model size	Large
	Batch size	Auto
	Flexible generation	Off
	Value protection	Off
CLAVADDPM	num_clusters	50
	parent_scale	1.0
	classifier_scale	1.0
	num_timesteps	2000
	batch_size	4096
	layers_diffusion	[512, 1024, 1024, 1024, 1024, 512]
	iterations_diffusion	200000
	lr_diffusion	0.0006
	weight_decay_diffusion	1e-05
	scheduler_diffusion	"cosine"
	layers_classifier	[128, 256, 512, 1024, 512, 256, 128]
	iterations_classifier	20000
	lr_classifier	0.0001
	dim_t	128
RGCLD	GNN hidden dim	128
	GNN aggregation	sum
	GNN layers	# Tables
	GNN lr	0.008
	GNN weight decay	$1e-5$
	GNN epochs	250
	VAE layers	2
	VAE token dim	4
	VAE hidden dim	128
	VAE δ	0.7
	VAE $(\beta_{max}, \beta_{min})$	$(0.01, 1e-5)$
	VAE lr	$1e-3$
	VAE epochs	4000
	Diff model dim	1024
	Diff lr	0.001
	Diff weight decay	$1e-6$
	Diff epochs	4000

H Dataset Schema

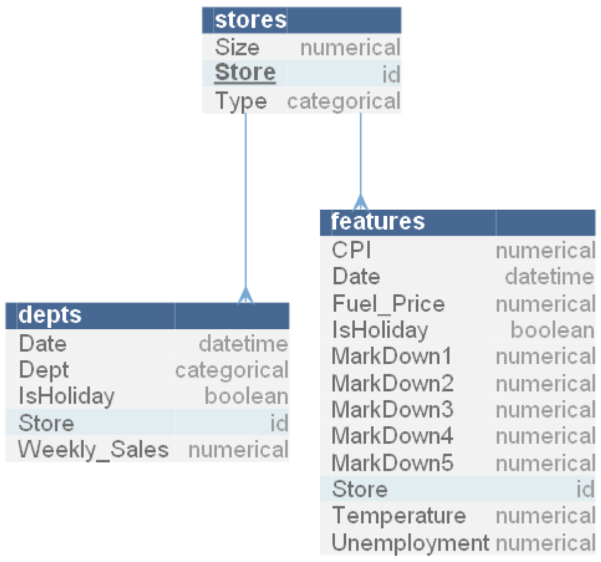

users
affiliate_channel categorical
affiliate_provider categorical
age numerical
country_destination categorical
date_account_created datetime
date_first_booking datetime
first_affiliate_tracked categorical
first_browser categorical
first_device_type categorical
gender categorical
id id
language categorical
signup_app categorical
signup_flow categorical
signup_method categorical
timestamp_first_active datetime
sessions
action categorical
action_detail categorical
action_type categorical
device_type categorical
secs_elapsed numerical
user_id id

(a) Airbnb

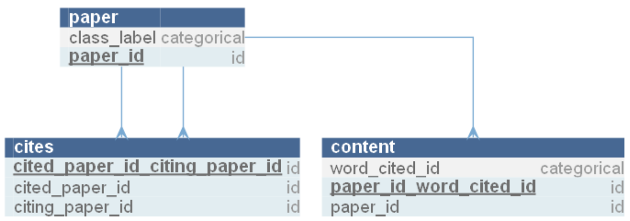


(b) Rossmann

Figure 9: Two table database diagrams.



(a) Walmart



(b) Cora

Figure 10: Walmart and CORA database diagrams.

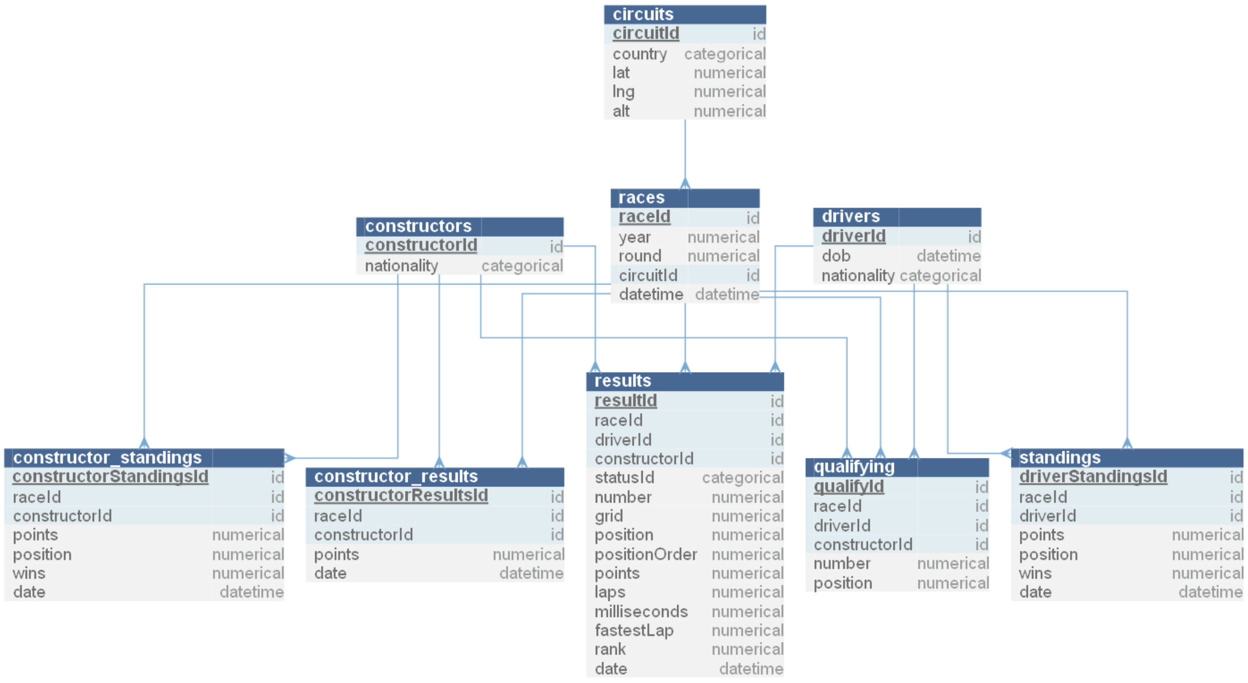


Figure 11: F1 database diagram.

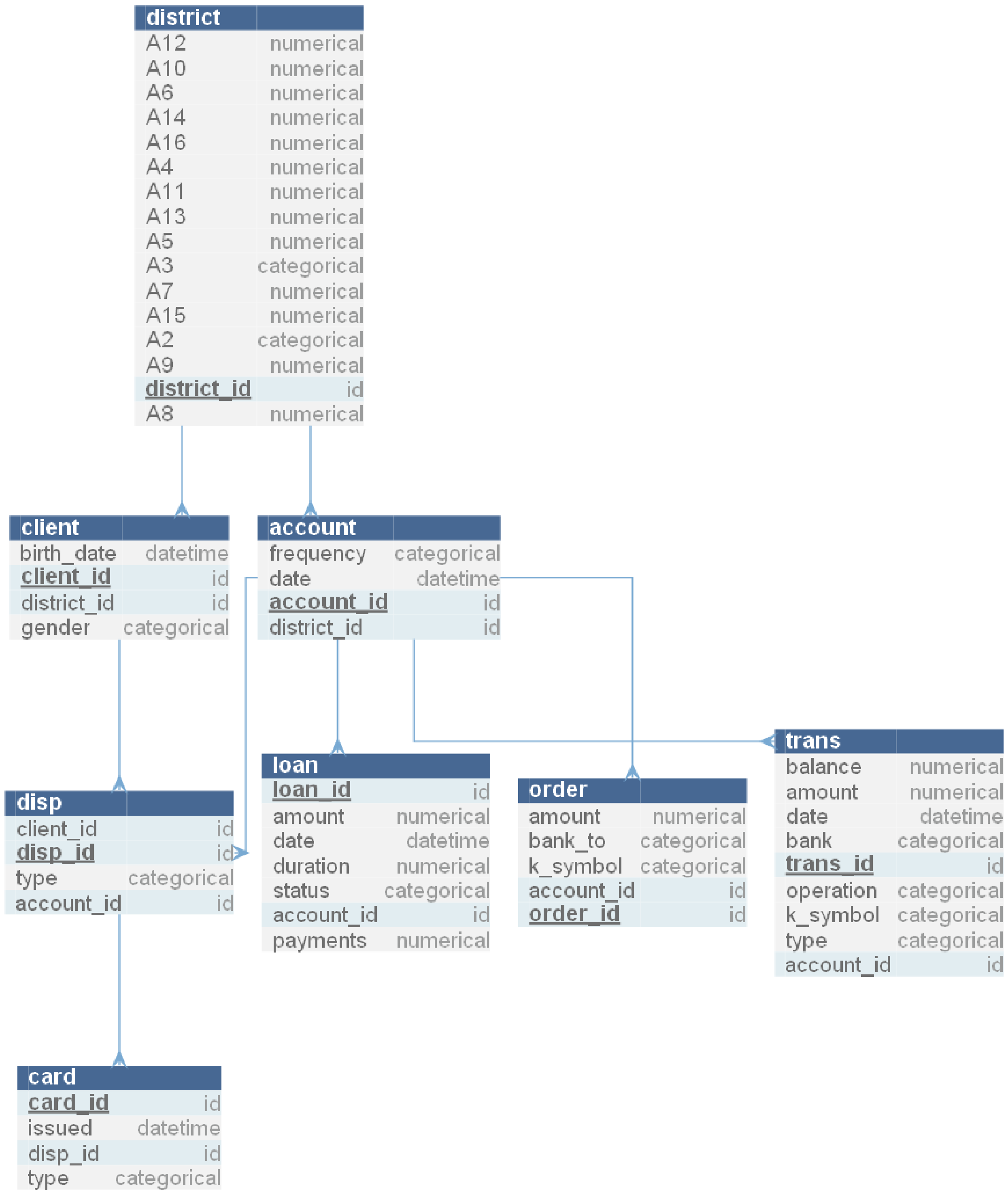


Figure 12: Berka database diagram.

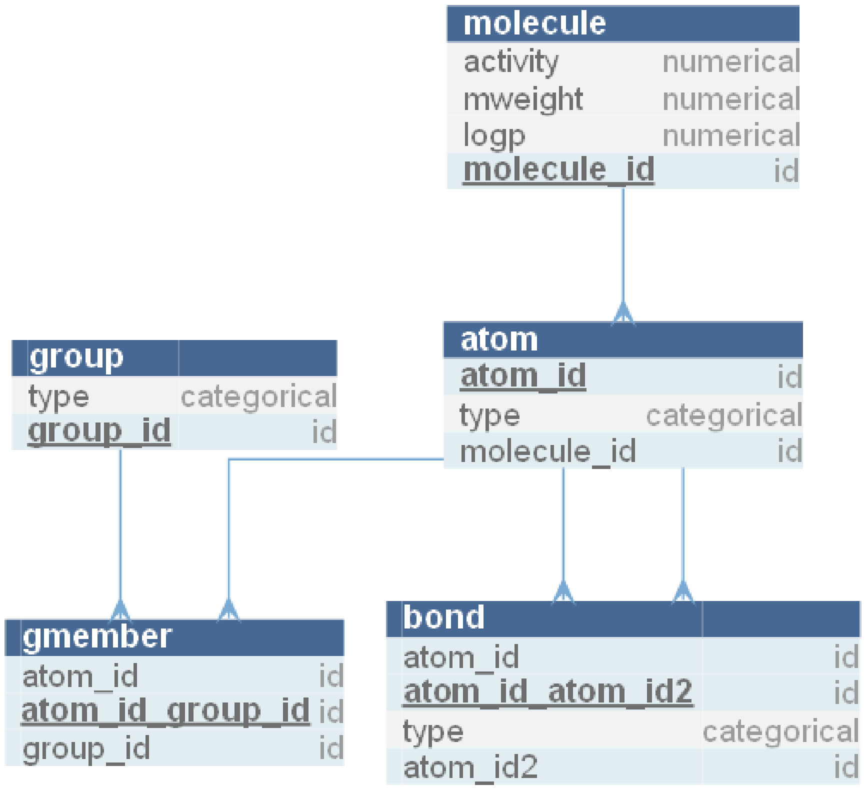


Figure 13: Biodegradability database diagram.

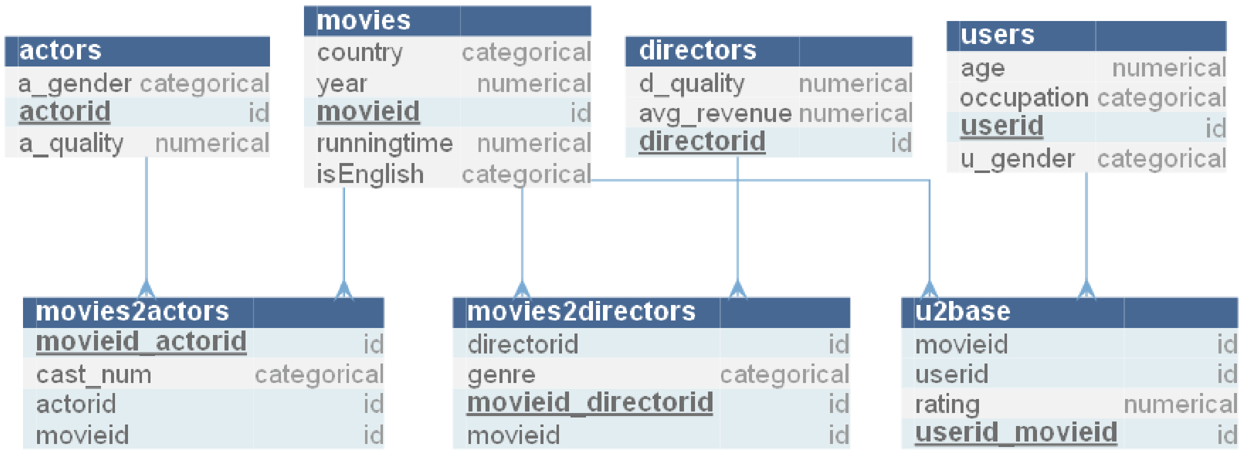


Figure 14: IMDB MovieLens database diagram.