Tabular Data Generation with Probabilistic Circuits

Dylan Ponsford

THE UNIVERSITY *of* EDINBURGH

Dagstuhl Seminar 26102: Tensor Factorizations Meet Probabilistic Circuits

6th March 2026

Prelude: BPD and image sample quality of PCs

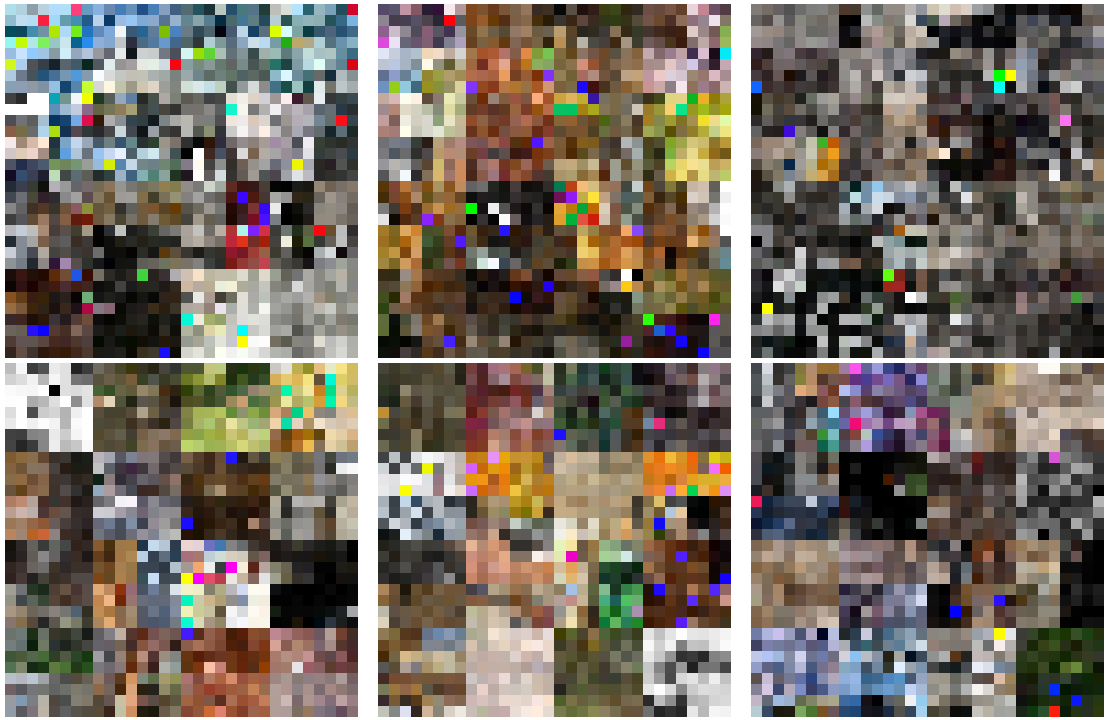


Figure: PC samples on ImageNet-32

On image data, PC samples are **not good!**
Not so clear what improving the BPD (log-likelihood) is really doing

So how good are circuit samples in domains with lower dimensionality? $\rightarrow$ Tabular data!

What is tabular data generation?

Tabular data:

- Structured **low- to medium-dimensionality** data
- (Potentially) **heterogeneous data types**
- Organised in **rows** (samples) and **columns** (features)

$\rightarrow$ Tabular data generation! (TDG)

Some motivations:

- Privacy-preserving dataset sharing
- Dataset augmentation

Example for concreteness

e.g. Train on table of different people's physical attributes

Weight (kg)	Height (cm)	Hair Colour
75.3	180	Brown
62.5	171	Blonde
⋮	⋮	⋮

Generate

Weight (kg)	Height (cm)	Hair Colour	Realistic?
67.1	183	Black	✓
40.2	190	Green	**?**

Current SotA approaches for TDG

- Current SotA approaches for TDG: **diffusion-based**[1]
- Here: we will compare against **TabDiff** [3] and **TabSyn** [6] as SotA models
- Other models we will see are CTGAN [4], TVAE [4], CoDi [2], STaSy [1].

[1] Now: a few flow-based models too but no (working) code

Why try circuits for TDG?

- Motivation 1: **Tractability**
- Motivation 2: Circuits can be seen as the **generative equivalent of decision trees / forests** which work well for discriminative tasks on tabular data

As it turns out, circuits will work quite well for TDG!

How do we build a circuit for TDG?

This is mostly out of the box!

Input layers: Gaussian for numerical features, and categorical for categorical features.

Region graph (RG): created from Chow-Liu tree learned on data e.g. on the *Magic* UCI dataset

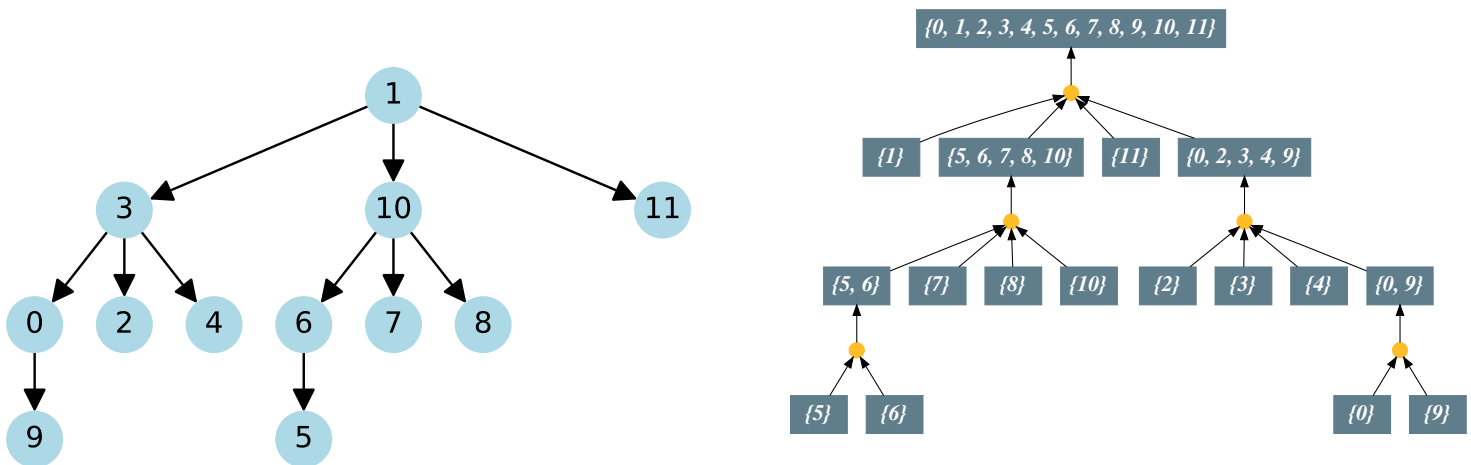


Figure: Region graph *(right)* compiled from Chow-Liu tree *(left)*

From RG to PC

We use the region graph as a template to construct the PC:

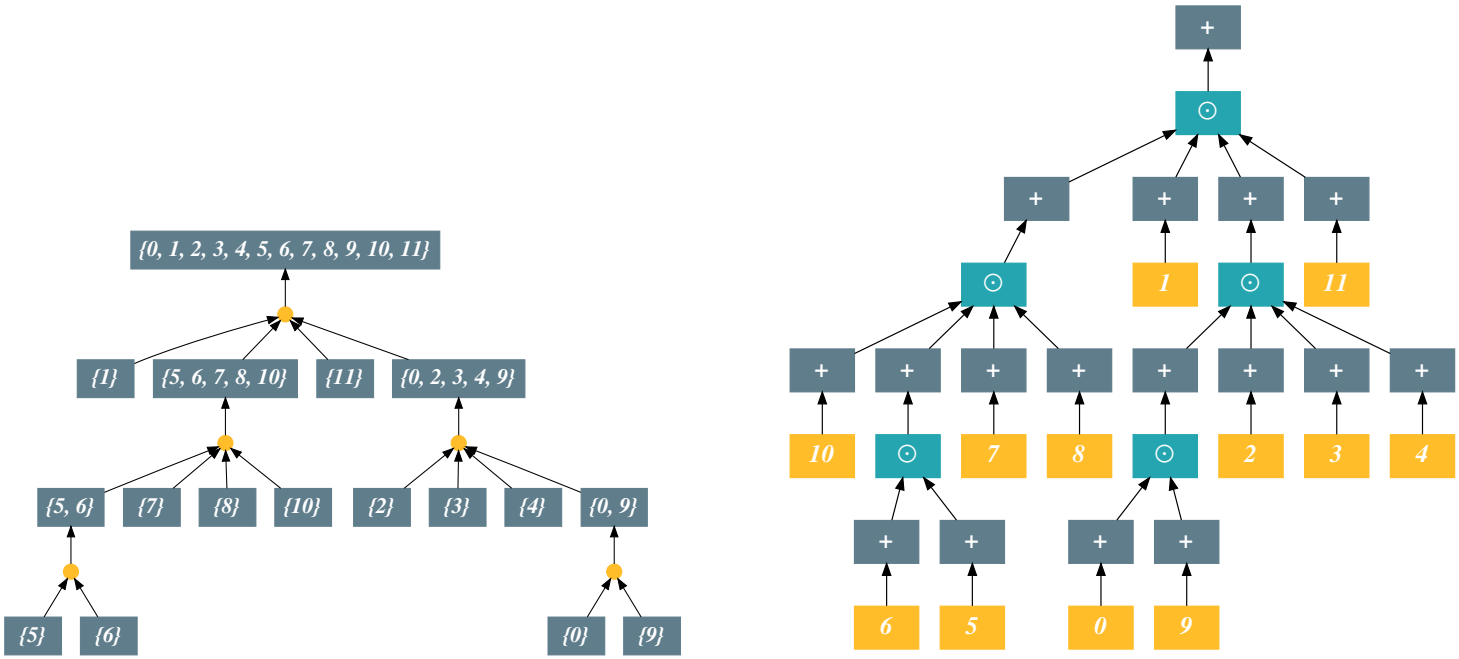


Figure: PC *(right)* created from the learned region graph *(left)* on the *Magic* UCI dataset

Key consideration: we highly overparameterise our circuits (e.g. for *Magic* we have $\sim 10^7$ parameters)

Once we have a circuit: how do we evaluate our generated data?

How do we evaluate synthetic data?

(Some) typical axes of classification:

1. **Fidelity metrics**
2. Utility metrics
3. Privacy metrics[2]

[2] The metric generally used for this (*Distance to Closest Record; DCR*) is terrible [5]

How do we compute these metrics?

- Train $N = 5$ models
- Generate 1 synthetic dataset per model
- Compute the metric for each synthetic dataset, comparing against the real data
- Then report mean $\pm$ standard deviation over model seeds

Creating indistinguishable data (C2ST)

Desideratum:

Generated data should be indistinguishable from real data.

- So train a classifier to determine this!
- Classifier performance $\uparrow$ dataset quality $\downarrow$
- *C2ST = Classifier 2-Sample Test*

$$1 - (2 \cdot \max(\mathrm{AUROC}(c, \mathcal{D}^{(\mathrm{test})}), 0.5) - 1) \in [0, 1] \qquad \text{(C2ST)}$$

We use XGBoost as the classifier.[3]

[3] Some previous works use logistic regression—not sufficient!

Modelling dependencies (wNMIS)

Desideratum:

Generated data should accurately model the dependencies seen between features in real data.

- So let's measure how well we model pairwise normalised mutual information (NMI)!
- Add a weighting to emphasise pairs which *actually* have some dependency

$\rightarrow$ weighted Normalised Mutual Information Similarity (wNMIS)[4]

[4] Details in upcoming arXiv preprint

So how good are PCs at TDG?

Circuits are fast to train and competitive

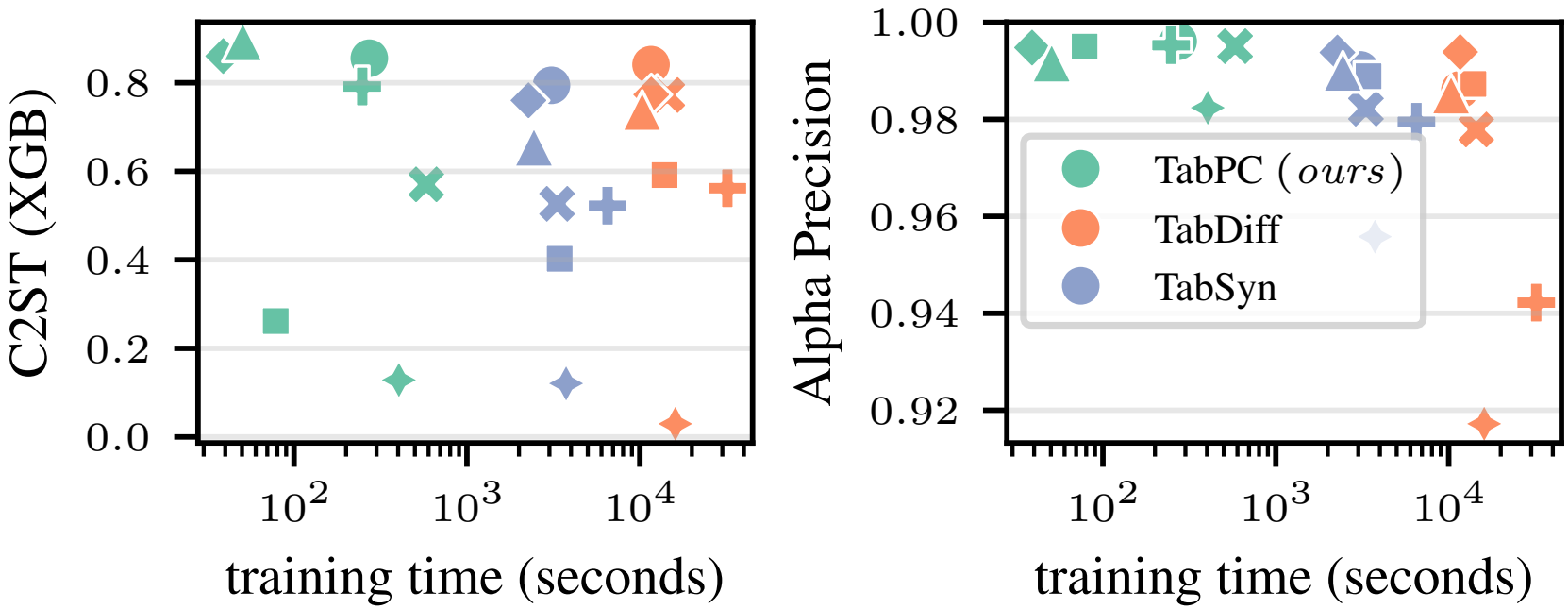


Figure: **PCs for tabular data (TabPC) compete with diffusion-based approaches yet train an order of magnitude faster**

Critical difference diagrams (CDDs)

These summarise whether we can statistically distinguish performance.
Scale: average rank across datasets.
Solid line connecting methods: cliques (can't distinguish performance)

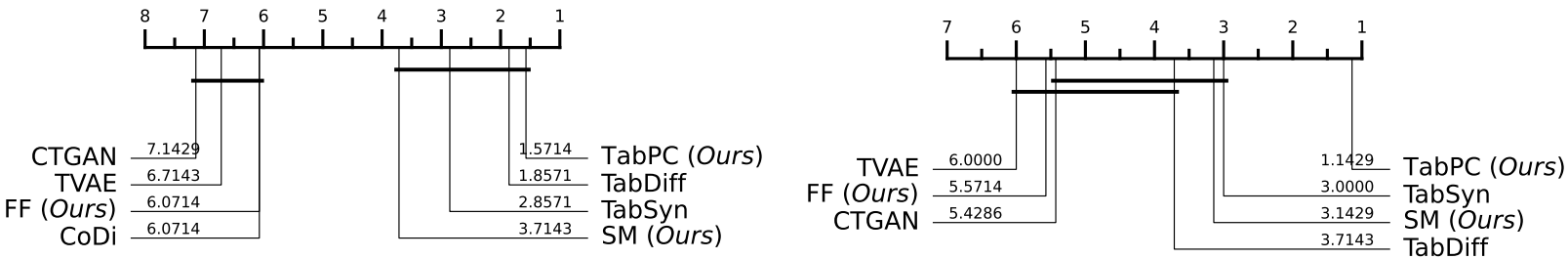


Figure: CDDs for (L) C2ST (XGB) and (R) Alpha Precision

More CDDs

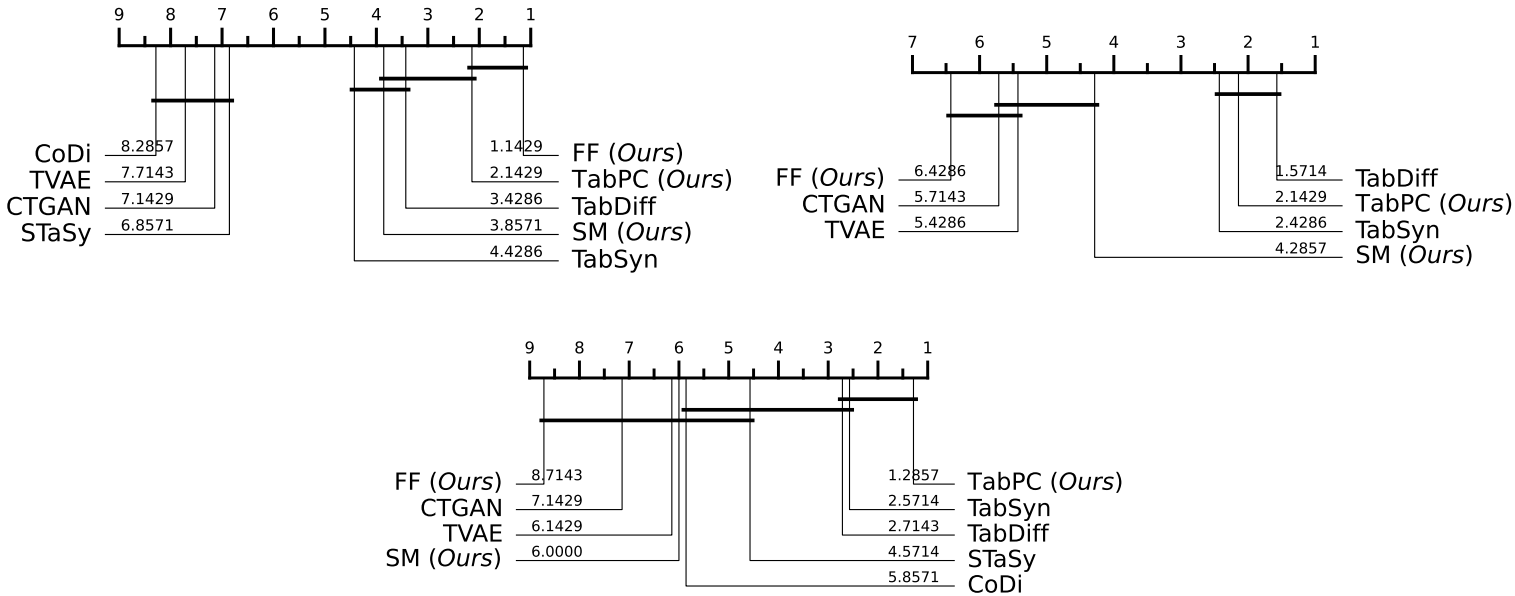


Figure: *Top Left:* Shape, *Top Right:* Beta Recall, *Bottom:* wNMIS (weighted Normalized Mutual Information Similarity)

Based on these: **TabPC is at least competitive with SotA approaches and sometimes better**!

What else do we get with PCs?

For 'free' (with smoothness and decomposability):

- **Exact likelihood computation**
- **Marginalisation / missing values**
- **Conditional sampling**

$\rightarrow$ Let's try to highlight these!

What does the likelihood tell us?

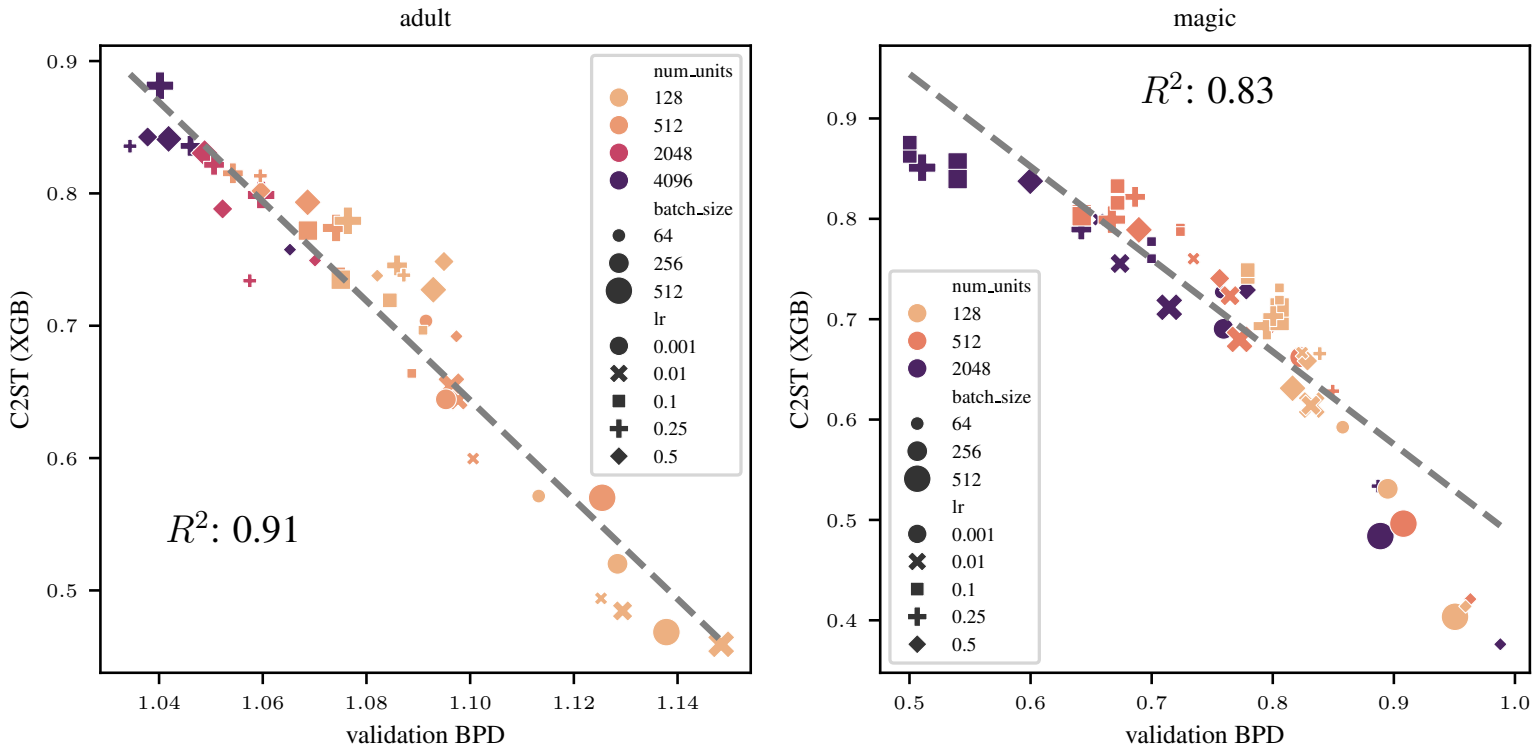


Figure: Validation set BPD vs sample C2ST (XGB) for the (L) Adult and (R) Magic datasets

Here, BPD is a strong signal for sample quality! We also use validation BPD for model selection

Across all datasets...

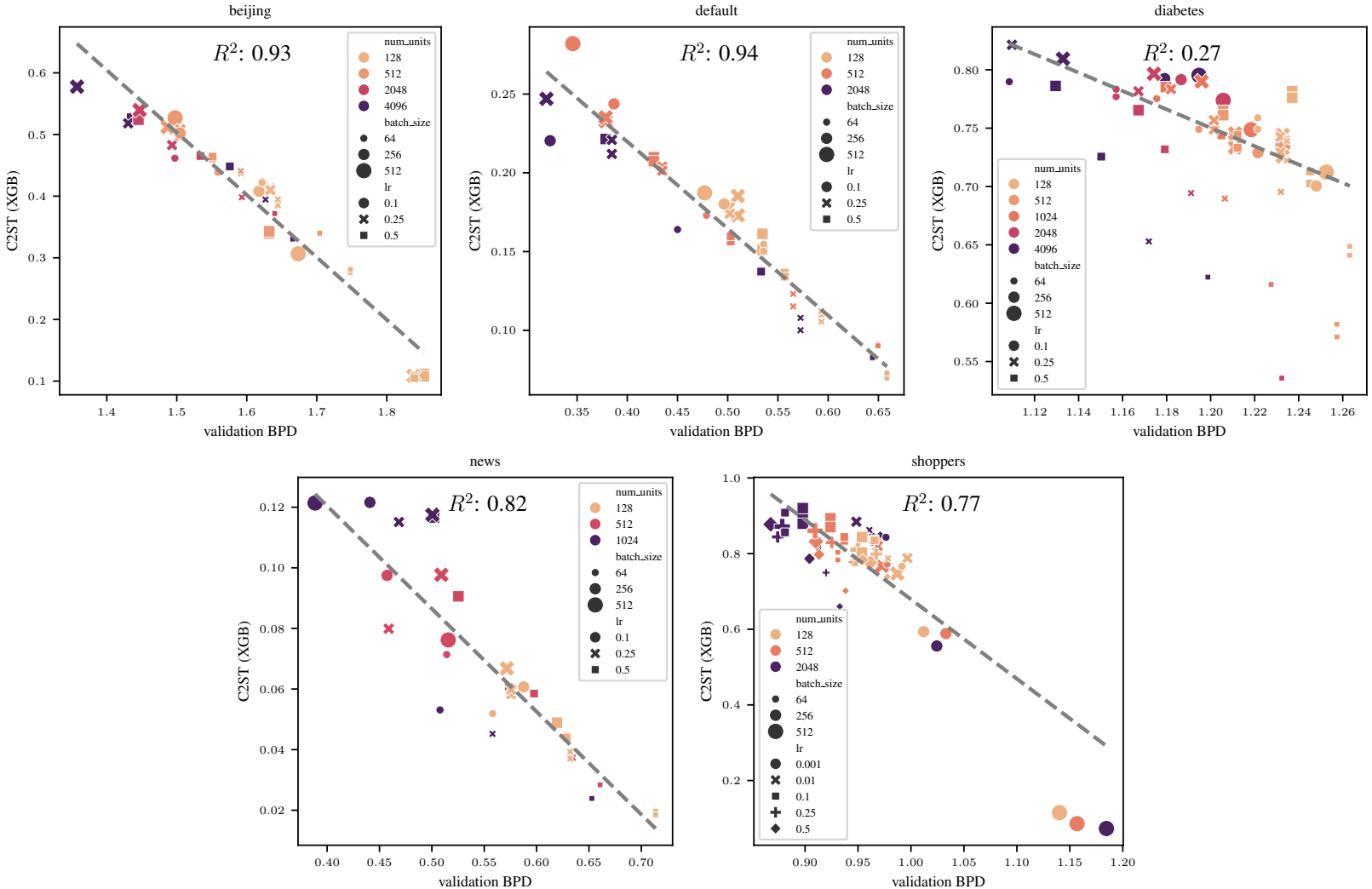


Figure: Validation set BPD vs sample C2ST (XGB) for other datasets: Beijing, Default, Diabetes, News, and Shoppers

Conditional sampling

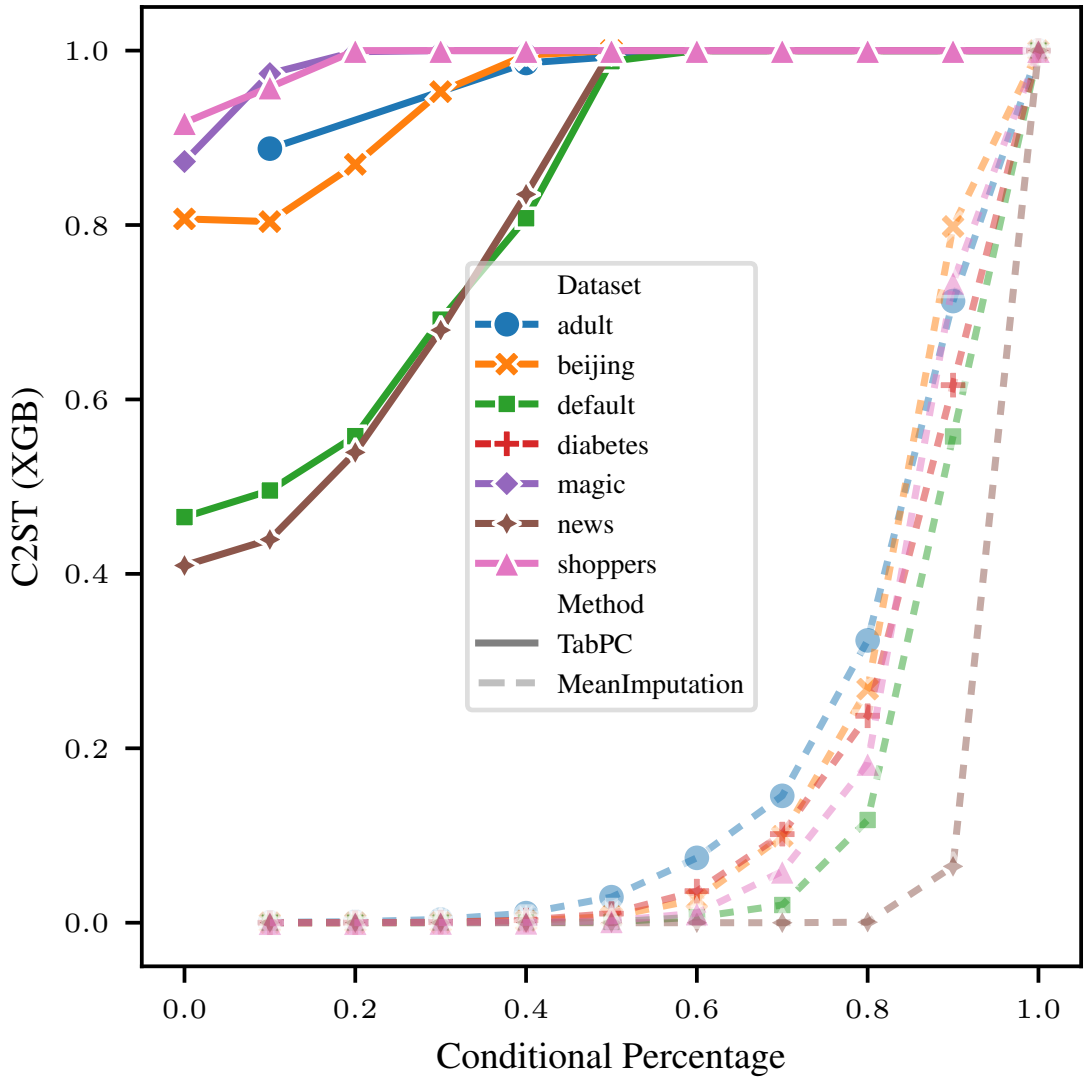
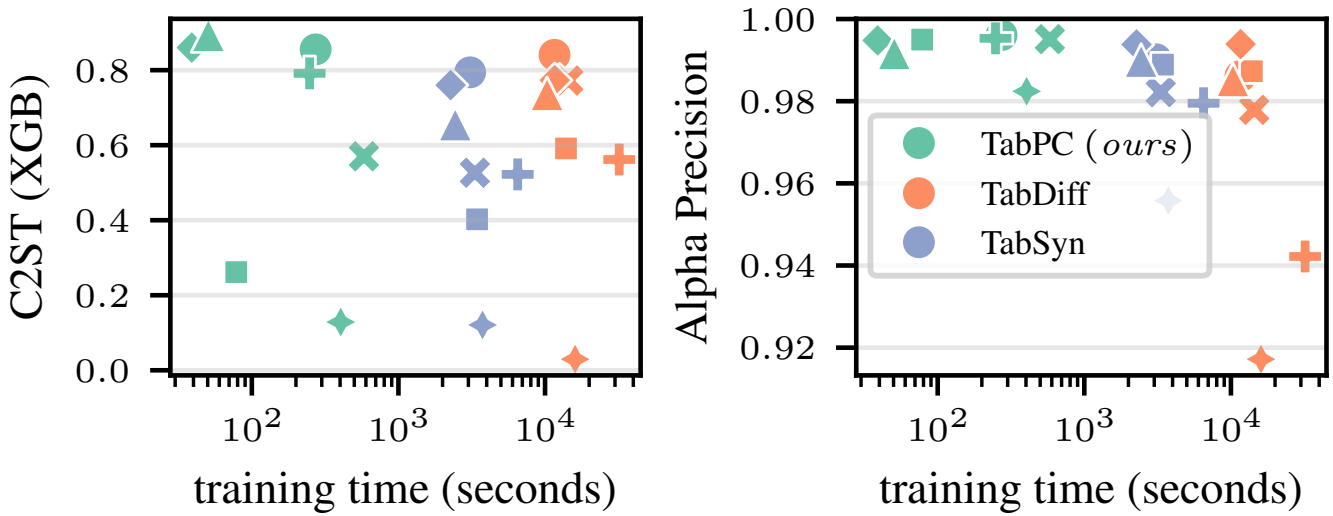


Figure: **TabPC produces high fidelity conditional samples**

Takeaways



- **Circuits are fast to train and effective** for TDG
- For tabular data, **BPD strongly correlates with sample quality**
- Tractability gives us **exact and efficient conditional sampling**

Many thanks to Davide Scassola, Adrián Javaloy, Sebastiano Saccani, Luca Bortolussi, Henry Gouk, and Antonio Vergari! Preprint should soon be available on arXiv (work under review at UAI).

References I

[1] Jayoung Kim, Chaejeong Lee, and Noseong Park. *STaSy: Score-based Tabular data Synthesis*. May 29, 2023. arXiv: 2210.04018[cs].

[2] Chaejeong Lee, Jayoung Kim, and Noseong Park. *CoDi: Co-evolving Contrastive Diffusion Models for Mixed-type Tabular Synthesis*. Sept. 21, 2023. arXiv: 2304.12654[cs].

[3] Juntong Shi et al. *TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation*. Feb. 16, 2025. arXiv: 2410.20626[cs].

[4] Lei Xu et al. *Modeling Tabular data using Conditional GAN*. Oct. 28, 2019. arXiv: 1907.00503[cs].

References II

[5] Zexi Yao et al. "The DCR Delusion: Measuring the Privacy Risk of Synthetic Data". In: *Computer Security – ESORICS 2025: 30th European Symposium on Research in Computer Security, Toulouse, France, September 22–24, 2025, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, Oct. 13, 2025, pp. 469–487. ISBN: 978-3-032-07883-4.

[6] Hengrui Zhang et al. *Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space*. May 11, 2024. arXiv: 2310.09656[cs].