A Sobering Look at Tabular Data Generation via Probabilistic Circuits

[1]University of Edinburgh
[2]University of Trieste
[3]Aindo

Davide Scassola[*2,3] **Dylan Ponsford**[*1] **Adrián Javaloy**[1] **Sebastiano Saccani**[3] **Luca Bortolussi**[2] **Henry Gouk**[1] **Antonio Vergari**[1]

TLDR: ***Metrics*** *for* ***tabular data generation*** *are* ***problematic****, and* ***simple & fast models*** *such as* ***probabilistic circuits*** *can* ***compete with or outperform*** *sophisticated diffusion-based models.*

1 Tabular data generation (TDG)

Task: learn from a training set to generate ***realistic synthetic tabular data*** (without just copying!)

Uses: Privacy protection [1], data augmentation [2]

Age	Height	Weight	Smoker	Diagnosis
55	183	80.2	Yes	Healthy
32	150	52.5	Yes	Heart Disease

Diffusion-based models, e.g. TabSyn [3] and TabDiff [4], are SotA... ***but are they***?

2 Metrics are distorting progress!

A fully factorized **(FF)** model scores as well as diffusion on:

Trend: similarity of "**pairwise dependencies**". Averages:

- Continuous column pairs: *correlation similarity*
- Discrete / mixed column pairs: *contingency similarity*

Method	Adult	Beijing	Default	Diabetes	Magic	News	Shoppers
TabDiff	0.985	0.974	0.973	0.967	0.992	0.984	0.981
TabPFN-3	0.915	0.924	0.770	*	0.894	0.949	0.810
FF	**0.925**	**0.910**	**0.888**	**0.968**	**0.870**	**0.955**	**0.937**

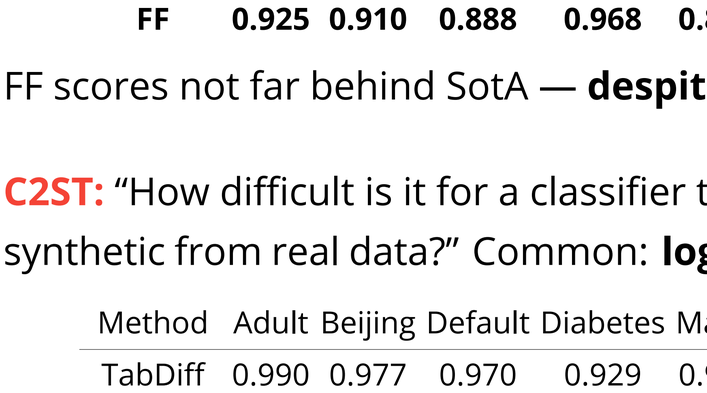
FF scores not far behind SotA — **despite independence**!

C2ST: "How difficult is it for a classifier to distinguish synthetic from real data?" Common: **logistic regression** (LR)

Method	Adult	Beijing	Default	Diabetes	Magic	News	Shoppers
TabDiff	0.990	0.977	0.970	0.929	0.996	0.906	0.976
TabPFN-3	0.886	0.996	0.882	*	0.972	0.838	0.717
FF	**1.000**	**0.998**	**0.995**	**1.000**	**0.989**	**0.980**	**1.000**

FF achieves perfect scores on some datasets!

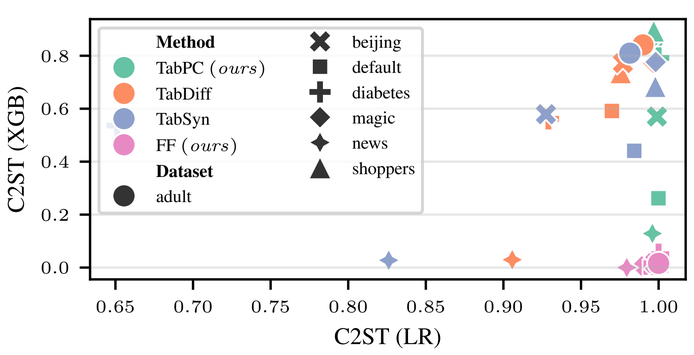
3 Better measuring dependencies



Instead of Trend:

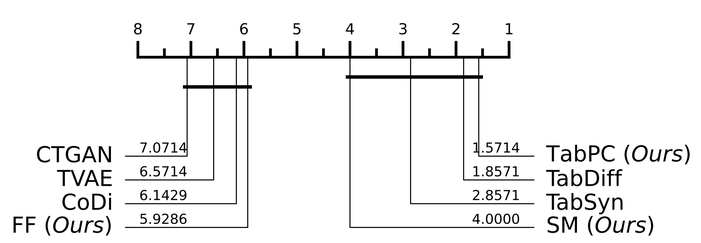
weighted
Normalised
Mutual
Information
Similarity!

4 Better classifier, clearer fidelity



We prove **a fitted FF model can perfectly fool LR!** → compute **C2ST with XGBoost**!

4 More than just means

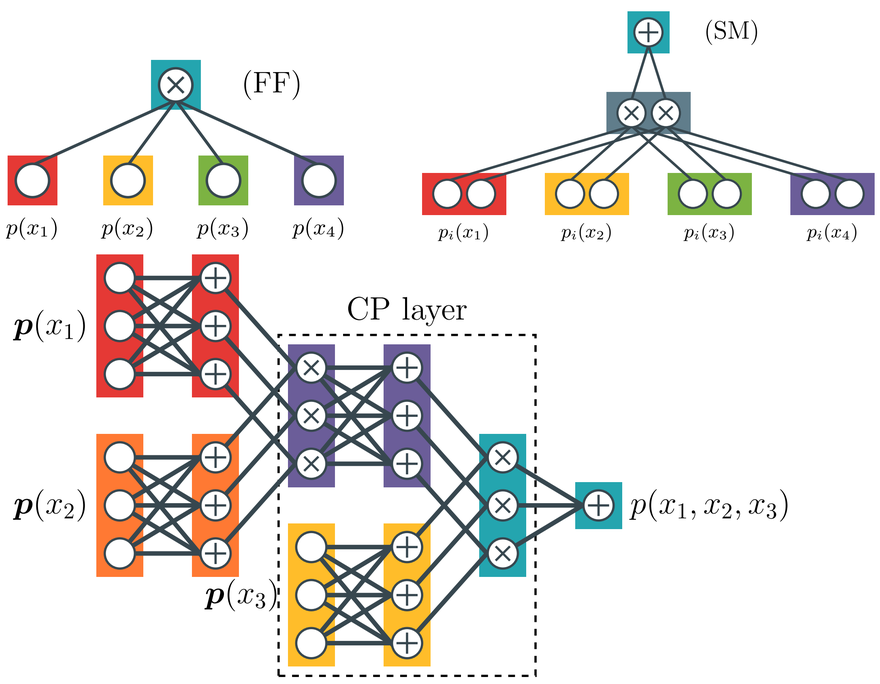


Critical difference diagrams display statistical comparisons, e.g. for C2ST (XGB) - better results are not always significant!

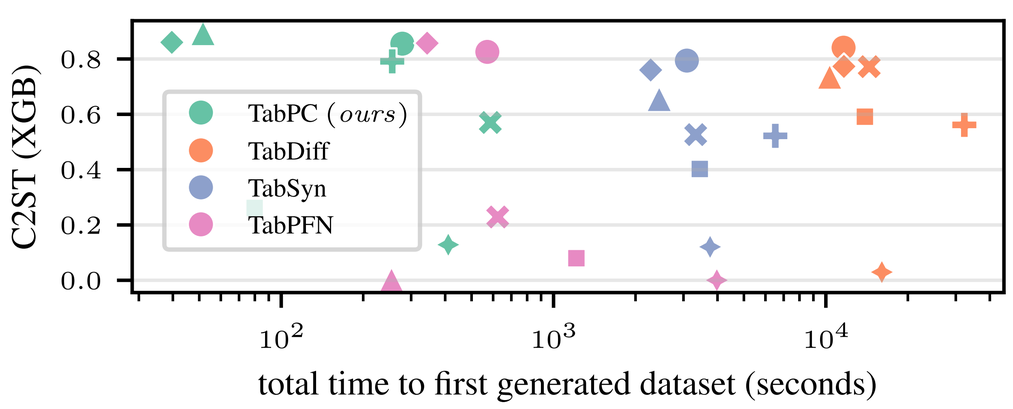
5 PCs to the rescue

PCs ≈ **deep, hierarchical mixture models**

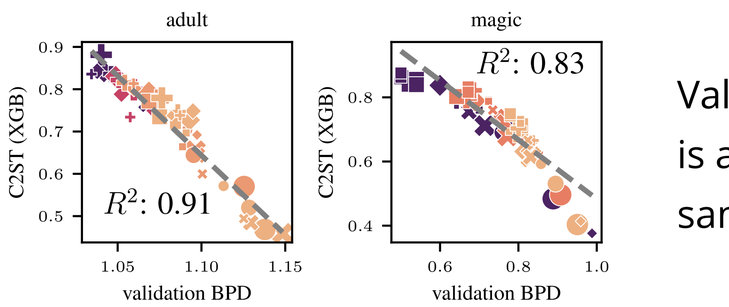
→ ***Key advantage:*** ***expressive*** *yet* ***tractable!***



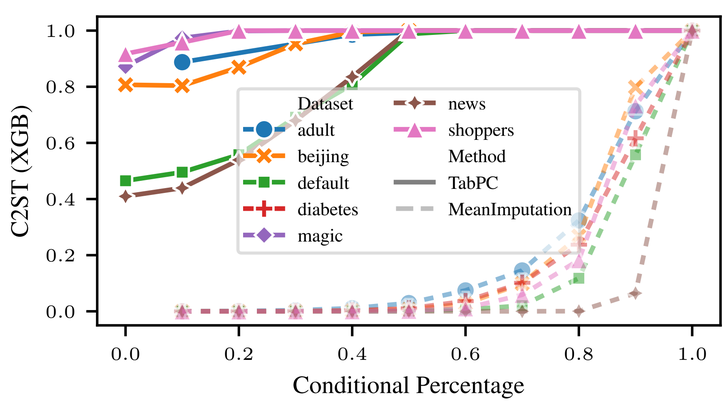
6 PCs are fast and competitive!



6 PCs enable tractable inference!



Validation log-likelihood is a strong signal for sample quality



TabPC produces high-fidelity conditional samples out of the box

PAPER

REPO

[1] M. Hernandez et al. "Synthetic data generation for tabular health records: A systematic review". Neurocomputing 493 (2022), pp. 28–45. ISSN: 0925-2312.
[2] L. Cui et al. "Tabular Data Augmentation for Machine Learning: Progress and Prospects of Embracing Generative AI". ACM Comput. Surv. 58.12 (June 2026). ISSN: 0360-0300.
[3] H. Zhang et al. "Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space". *The Twelfth International Conference on Learning Representations*.
[4] J. Shi et al. "TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation". *The Thirteenth International Conference on Learning Representations*.