

From Graphical Models to Foundation Models: Synthetic Bulk RNA-seq Data Generation

Daniil Filienko[†], Elias Chaibub Neto[‡] Sikha Pentyala[†],
Jineta Banerjee[‡], Luca Foschini[‡], Martine De Cock[†]

[†] University of Washington Tacoma, [‡] Sage Bionetworks, Seattle

1 Introduction

AI has revolutionized computational biology and precision medicine, enabling the creation of versatile tools for tasks ranging from personalized diagnosis [4] to disease prevention [17]. However, data-driven AI has also created significant privacy concerns. Genomic data in particular contains sensitive information and is subject to strict privacy regulations [13, 14, 2]. As a result, such data routinely ends up siloed within research centres and health systems. Even when genomic data is nominally “open”, researchers who wish to use it must navigate substantial access barriers: cumbersome data-use agreements with the data holders, institutional ethics approvals, or working within restricted computational enclaves [16]. These constraints directly undermine the *Findable, Accessible, Interoperable, and Reusable* (FAIR) principles that guide open science, effectively rendering such data “dark data” and impeding reproducibility in biomedical research [6].

Synthetic Data Generation (SDG) offers a promising route toward greater data access without the additional risks to patient privacy. SDG models can capture and replicate the properties of the real data without reproducing individual records [9]. Synthetically generated datasets can be tailored to specific privacy requirements and distributed by the data holders at will, making it particularly attractive for enabling broad, asynchronous data reuse.

Chen et al. [1] and Oztürk et al. [15] compared the performance of various tabular SDG algorithms for bulk RNA sequencing (RNA-seq) data, measured by downstream classification performance, fidelity, and biological utility of the generated data. The landscape of tabular SDG itself, however, keeps evolving, with new deep learning methods based on tabular foundation models and lightweight, non-parametric alternatives demonstrating strong utility and fidelity on structured data. However, these newer methods have been largely evaluated on general-purpose tabular benchmarks, and their effectiveness for high-dimensional biological data, such as bulk RNA-seq, remains underexplored. To investigate the extent to which their success carries over to genomics data, in this paper we evaluate two recent tabular SDG methods on the bulk RNA-seq datasets provided by the CAMDA’26 Health Privacy Challenge:¹ the deep learning based MIAV [12] and the non-parametric TabSDS [11]. These offer two different visions of an effective SDG: MIAV extends the recently introduced TabPFN [5] while TabSDS is a non-parametric shuffle-based approach, allowing us to see the advantage offered by the more computationally expensive deep learning models.

CAMDA Health Privacy Challenge 2026; Track I. The challenge invited participants to generate synthetic bulk gene expression data preserving both the data donors’ privacy and the biological signal of the real data. The platform for this challenge provided pre-processed versions of two TCGA RNA-seq datasets: (a) TCGA-BRCA consisting of 1,089 samples and 978 gene expression features with 5 cancer subtypes for prediction and (b) TCGA-Combined consisting of 4,323 samples and 978 features labeled for prediction of 10 different cancer tissue origins.

2 Methodology

Synthetic data generation for tabular data encompasses a diverse set of methodological paradigms that differ in how they represent and reproduce the joint distribution of the original dataset. We use

¹<https://benchmarks.elsa-ai.eu/?ch=4&com=introduction>

3 methods that occupy distinct positions in this design space: MIAV-TabPFN [12], a deep-learning-based generative algorithm built on tabular foundation models; TabSDS [11], a lightweight, model-free method that operates via rank-based transformations and controlled shuffling; and Private-PGM [7], a statistics-based, marginal-driven approach that serves as an established baseline in differentially private SDG. These methods differ in their modeling assumptions and computational characteristics, as well as in how they trade off data fidelity, scalability, and privacy. Our evaluation highlights these contrasts, with MIAV-TabPFN as the primary method of interest, TabSDS as an efficient and practical alternative, and Private-PGM as a strong DP baseline grounded in prior CAMDA work [8, 15].

MIAV-TabPFN. MIAV-TabPFN builds on recent advances in tabular foundation models, in particular TabPFN, which demonstrated that probabilistic transformers trained via in-context learning can match or exceed classical machine learning methods for tabular prediction tasks [5]. Extending such models to SDG, however, introduces challenges when feature–feature associations are weak or sparse, as limited predictive signal is available in the context examples alone.

Maximal Information Auxiliary Variables (MIAV) address this limitation by augmenting the model’s input with carefully constructed auxiliary variables [12]. These variables are generated by sampling independent noise and then rank-matching it to the empirical distributions of the real features. The resulting auxiliary variables are maximally informative and distribution-preserving, providing a strong synthetic signal that the foundation model can exploit during generation. In our experiments (Tables 1 and 2), we report the results for MIAV instantiated with different magnitudes of noise to demonstrate the tradeoff in utility and relative distance that can be achieved with MIAV.

In MIAV-TabPFN, the auxiliary variables are supplied as the context to TabPFN, and each variable in the synthetic dataset is generated in a single forward pass. This design improves performance in settings with weakly associated variables, preserves strict invariance to feature ordering, and retains the training-free nature of in-context learning.

TabSDS. Tabular Synthetic Data Generation via Sequential Data Shuffling (TabSDS) [11] represents a fundamentally different approach. Rather than learning or sampling from an explicit generative model, TabSDS is a non-parametric, model-free procedure that reconstructs dependency structure through rank-based transformations and controlled perturbations of the original data.

The method proceeds in 3 conceptual stages. First, synthetic marginal distributions are generated independently for each feature using an order-statistics-based interpolation scheme. Second, the original dataset is perturbed using sequential joint probability-preserving data shuffling (SJP-PDS) [10], producing a shuffled version of the data in which associations are preserved to a degree controlled by a user-defined parameter n_c . Finally, rank-matching is used to transfer the dependency structure of the shuffled data to the synthetic marginal samples.

This pipeline allows TabSDS to generate synthetic data extremely efficiently, with computational costs far below those of deep generative or foundation-model-based approaches. The parameter n_c provides explicit control over the fidelity–privacy trade-off: larger values preserve more of the original association structure at the cost of increased disclosure risk. In our experiments (Tables 1 and 2), we follow [11] and set $n_c = 20$ with a sampling proportion $prop = 0.5$.

Private-PGM. Private-PGM [7] is a representative statistics-based SDG method that constructs a synthetic dataset by estimating low-dimensional marginals (typically one- and two-way feature distributions) and representing them as a probabilistic graphical model. Synthetic samples are then drawn from this estimated joint distribution.

A defining feature of Private-PGM is its explicit use of differential privacy (DP) [3]: calibrated noise is added to the estimated marginals to ensure formal privacy guarantees before model fitting and sampling. This approach has proven competitive in the prior CAMDA’25 Health Privacy

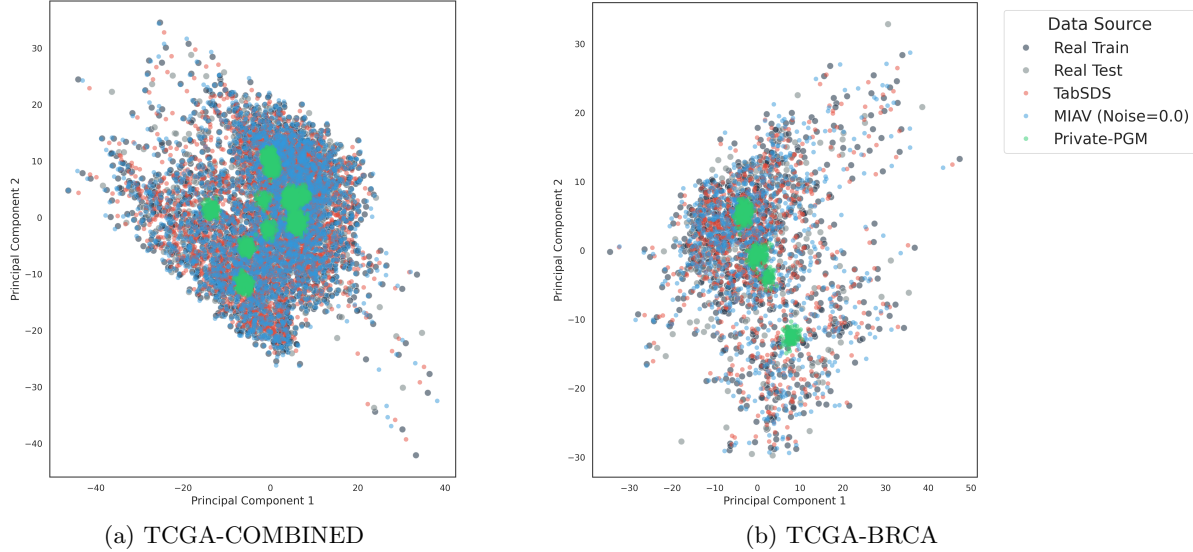


Figure 1: Global PCA projection of real and generated synthetic RNA-seq data

challenge [15]. Since our use of Private-PGM in this work closely mirrors its application in the CAMDA’25 challenge, we refer the reader to Menzies et al. [8] for details on the DP formulation, parameterization, and data preprocessing, including the number of bins for discretization, iterations, and other hyperparameters.

3 Analysis

Figure 1 contains PCA plots of the real data as well as synthetic data generated with the three SDG methods. As can be observed in the plots, TabSDS and MIAV mirror the real data distribution more closely while Private-PGM generates a few tightly organized clusters. This is corroborated by the results in Tables 1 and 2, where the distance between the synthetic and real data is greater for Private-PGM than for the other methods.

From a privacy perspective, large distances between synthetic records and their closest records in the real data (as with Private-PGM) are advantageous, but of course this needs to be looked at in tandem with the utility and fidelity of the synthetic data. Table 1 and Table 2 report a variety of metrics to this end. The majority of these metrics were adapted from Öztürk et al. [15]. To obtain the results, we split each dataset TCGA-BRCA and TCGA-COMBINED into 80% for training and 20% for testing.

Utility. The first column, TSTR AUC-ROC (Train on Synthetic, Test on Real), evaluates downstream utility by training a one-vs-rest logistic regression model for cancer type classification on the synthetic data and testing it on the real hold-out test data, reported as the Area Under the ROC Curve. Higher values indicate better utility. The results demonstrate high AUC across the SDG methods, even for Private-PGM which mirrors the real distribution the least as shown in Fig. 1. We furthermore observe an expected empirical privacy-utility tradeoff of MIAV at varying degrees of noise, with higher amounts of noise resulting in lower TSTR AUC.

Fidelity. The value in the second column is the inverted maximum mean discrepancy ($1/(1+\text{MMD})$) between the synthetic data and the 20% test subset of the real data. Higher values for this inverted discriminative score metric indicate that the synthetic data is more similar to the real data in its distribution, which, from a fidelity perspective, is a strength. The Feature Overlap column reports

predictive biological fidelity, measured by the proportion of overlapping top-10 important features (genes) extracted from the respective real and synthetic downstream classification models (logistic regression) *per class*; higher proportions indicate better agreement in predictive feature importance. Inverted Kullback-Leibler (KL) divergence measures how well the synthetic data’s per-feature value distributions match the real (held-out test) data’s distributions, averaged across all features. As expected from Fig. 1, Private-PGM scores much lower on this metric than MIAV and TabSDS.

Privacy. The remaining two columns assess privacy risks via distance metrics. Dist. Closest measures the absolute Distance to Closest Record (DCR), calculated as the average raw, unscaled Euclidean distance from each synthetic data point to its nearest neighbor in the real training dataset. Higher values suggest less memorization and better privacy preservation. For comparison, the baseline Dist. Closest (the average distance from the real test data to the real training data) is 24.04 for TCGA-BRCA and 23.27 for TCGA-Combined. Rel. DCR (Relative Distance to Closest Real sample) normalizes this metric by expressing it as a ratio of the synthetic distance to the real test baseline. Values greater than 1.0 indicate that synthetic points are, on average, farther from the training records than unseen real test patients are, demonstrating strong privacy guarantees against membership inference. Lower relative distances for the TabSDS and MIAV methods compared to Private-PGM are consistent with our expectations, with Private-PGM adding DP noise when producing synthetic data.

Table 1: Performance on TCGA-BRCA (higher values are better)

Method	Utility	Fidelity			Privacy	
	TSTR AUC	Inv. MMD	Feat. Overlap	Inv. KL	Rel. DCR	Dist. Closest
MIAV (noise=0.00)	0.948	0.999	0.306	0.772	0.196	4.643
MIAV (noise=0.25)	0.858	0.999	0.125	0.566	0.401	9.507
MIAV (noise=0.50)	0.711	0.993	0.100	0.430	0.580	13.739
MIAV (noise=0.75)	0.664	0.979	0.035	0.345	0.716	16.972
TabSDS	0.802	0.999	0.100	0.798	0.325	7.691
Private-PGM ($\epsilon = 10$)	0.953	0.992	0.080	0.051	1.145	27.114

Table 2: Performance on TCGA-COMBINED (higher values are better)

Method	Utility	Fidelity			Privacy	
	TSTR AUC	Inv. MMD	Feat. Overlap	Inv. KL	Rel. DCR	Dist. Closest
MIAV (noise=0.00)	0.988	0.999	0.214	0.939	0.130	3.008
MIAV (noise=0.25)	0.783	0.999	0.129	0.674	0.471	10.855
MIAV (noise=0.50)	0.669	0.993	0.100	0.515	0.722	16.638
MIAV (noise=0.75)	0.650	0.975	0.050	0.410	0.907	20.912
TabSDS	0.961	0.999	0.194	0.950	0.432	9.965
Private-PGM ($\epsilon = 10$)	0.996	0.994	0.263	0.054	1.297	29.900

MIAV and TabSDS shine where Private-PGM falls short, which is in the fidelity of the synthetic data, more precisely the preservation of the data distribution. Private-PGM, on the other hand, offers much better privacy protection and maintains high utility (better than even the best results with MIAV). The version of Private-PGM, as used as a baseline for CAMDA’26 and as reported in this paper, measures a one-way marginal distribution for each gene, and a two-way marginal distribution for each gene-cancer label pair. An interesting future direction to improve fidelity would be to measure and preserve two-way marginal distributions for gene-gene pairs in Private-PGM.

Acknowledgements. This material is based upon work supported by the National Science Foundation under Grant No. 2451163 and No. 2523406, by NSF NAIRR 240485 (Cloudbank AWS, TACC Vista) and NSF NAIRR 240091 (TACC Frontera), the UW Tacoma Founders Endowment, and the UW eScience Institute. This research was, in part, funded by the National Institutes of Health (NIH)

Agreement No. 1OT2OD032581. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH.

References

- [1] D. Chen, M. Oestreich, T. Afonja, R. Kerkouche, M. Becker, and M. Fritz. Towards Biologically Plausible and Private Gene Expression Data Generation. In *Proceedings in Privacy-Enhancing Technologies*, volume 2, pages 531–554, 2024.
- [2] R. Collins. A Message to Our Participants: UK Biobank Data Security Update, 2026. URL <https://www.ukbiobank.ac.uk/news/a-message-to-our-participants-uk-biobank-data-security-update/>.
- [3] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.
- [4] G. Heimberg, T. Kuo, D. J. DePianto, O. Salem, T. Heigl, N. Diamant, G. Scalia, T. Biancalani, S. J. Turley, J. R. Rock, et al. A cell atlas foundation model for scalable search of similar human cells. *Nature*, 638(8052):1085–1094, 2025.
- [5] N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister, and F. Hutter. Accurate Predictions on Small Data with a Tabular Foundation Model. *Nature*, 637(8045): 319–326, 2025.
- [6] M. McDermott, S. Wang, N. Marinsek, R. Ranganath, M. Ghassemi, and L. Foschini. Reproducibility in Machine Learning for Health Research: Still a Ways to Go. *Science Translational Medicine*, 13(586): eabb1655, 2021.
- [7] R. McKenna, D. Sheldon, and G. Miklau. Graphical-Model Based Estimation and Inference for Differential Privacy. In *International Conference on Machine Learning*, pages 4435–4444. PMLR, 2019.
- [8] S. Menzies, D. Filenko, S. Golob, S. Pentyala, J. Banerjee, L. Foschini, and M. de Cock. Synthetic Data Generation for bulk RNA-seq Data: A CAMDA Health Challenge Analysis. In *Annual International Conference on Critical Assessment of Massive Data Analysis, collocated with ISMB/ECCB*, 2025.
- [9] P. Myles, C. Mitchell, E. Redrup Hill, L. Foschini, and Z. Wang. High-fidelity synthetic patient data applications and privacy considerations. *Journal of Data Protection & Privacy*, 6.4, 2024.
- [10] E. C. Neto. Statistical disclosure control for numeric microdata via sequential joint probability preserving data shuffling. *Transactions on Data Privacy*, 17(3):147–179, 2024.
- [11] E. C. Neto. TabSDS: a Lightweight, Fully Non-Parametric, and Model Free Approach for Generating Synthetic Tabular Data. In *International Conference on Machine Learning*, pages 7131–7195. PMLR, 2025.
- [12] E. C. Neto. Using Maximal Information Auxiliary Variables to Improve Synthetic Data Generation Based on TabPFN Foundation Models. In *International Conference on Learning Representations*, 2026.
- [13] NIH. Implementation Update for Data Management and Access Practices Under the Genomic Data Sharing Policy. Available at: grants.nih.gov/grants/guide/notice-files/NOT-OD-24-157.html, 2024.
- [14] M. Oestreich, D. Chen, J. L. Schultze, M. Fritz, and M. Becker. Privacy Considerations for Sharing Genomics Data. *EXCLI Journal*, 20:1243–1260, 2021.
- [15] H. Öztürk, T. Afonja, J. Jätkö, R. Binkyte, P. Rodriguez-Mier, S. Lobentanzer, A. Wicks, J. Kreuer, S. Ouaari, N. Pfeifer, et al. Towards Useful and Private Synthetic Omics: Community Benchmarking of Generative Models for Transcriptomics Data. *bioRxiv*, pages 2026–03, 2026.
- [16] H. Watson, J. Gallifant, Y. Lai, A. P. Radunsky, C. Villanueva, N. Martinez, J. Gichoya, U. K. Huynh, and L. A. Celi. Delivering on NIH Dsata Sharing Requirements: Avoiding Open Data in Appearance Only. *BMJ Health & Care Informatics*, 30(1), 2023.
- [17] D. Zeevi, T. Korem, N. Zmora, D. Israeli, D. Rothschild, A. Weinberger, O. Ben-Yacov, D. Lador, T. Avnit-Sagi, M. Lotan-Pompan, et al. Personalized Nutrition by Prediction of Glycemic Responses. *Cell*, 163(5):1079–1094, 2015.

Table 3: COMB results for MIAV noise levels 0–0.75.

Experiment	Accuracy	Feature Overlap	MMD Score	Dist. Closest
miav_0	0.8563	15.4	−0.000239	3.073
miav_0.25	0.6297	6.4	0.000902	10.866
miav_0.5	0.5392	4.4	0.008139	16.611
miav_0.75	0.4619	4.8	0.027025	20.872

Table 4: BRCA results for MIAV noise levels 0–0.75.

Experiment	Accuracy	Feature Overlap	MMD Score	Dist. Closest
miav_0	0.7897	12.6	−0.000605	4.729
miav_0.25	0.7135	4.6	0.000652	9.484
miav_0.5	0.6456	2.0	0.007324	13.697
miav_0.75	0.6271	2.2	0.022233	16.906