Synthetic Data Generation for bulk RNA-seq Data: A CAMDA Health Challenge Analysis

Shane Menzies[†], Sikha Pentyala[†], Daniil Filienko[†], Steven Golob[†],
Jineta Banerjee[‡], Luca Foschini[‡], Martine De Cock[†]

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

1 Introduction

Advanced analytical tools in computational biology, including those based on artificial intelligence (AI), are heavily data-driven. This data often includes sensitive individual-level information such as genomics data which are subject to strict privacy regulations. Due to this, such data commonly requires tight access control and, as a result, ends up siloed in research centers and health systems. In many cases, even when health data is advertised as "open", researchers who want to use it must navigate long and cumbersome processes to access each silo, such as establishing onerous contractual agreements with data controllers, completing lengthy training processes, or working within data enclaves [17]. This significantly hinders the data's ability to be *Findable, Accessible, Interoperable, and Reusable*, effectively making it "dark data". Dark data is a widespread issue among biomedical communities and is one of the major barrier to progress in biomedical research and limited reproducibility in healthcare research [9].

A promising direction to address the challenge of dark data is *Synthetic Data Generation (SDG)*. SDG generates artificial data using a synthesizer trained on real data, thus offering an appealing solution for making data available while mitigating privacy concerns [6, 13]. When done well, synthetic data has the same characteristics as the original data but, crucially, without replicating personal information.

While there are a number of methods available for synthetic data generation, including approaches based on generative adversarial networks like CTAB-GAN [18], variational autoencoders like CVAE [1], and more recently, diffusion-based SDGs like TabDDPM [7], only a subset of these approaches offer formal privacy guarantees. Differential Privacy (DP) is a rigorous privacy framework that limits the risk of reidentifying individuals in a dataset, even when an adversary has access to auxiliary information [4]. Among SDG methods that provide DP guarantees, statistical-based SDGs – including marginals-based methods such as MST [11] and AIM [12] – are reported to achieve higher utility for synthetic tabular data generation than their above mentioned neural network-based counterparts [3, 14].

In the genomics domain, several works have proposed synthetic data generators tailored to specific biological data types, such as for bulk RNA-Seq data [15]. However, only a few of them incorporate DP when generating synthetic tabular bulk RNA-Seq data. This raises a key research question – "*Do state-of-the-art, DP SDG methods for tabular data perform well on genomics data?*". To address this, Chen et al. investigated how DP generative models perform on gene expression data in the context of acute myeloid leukemia (AML) datasets [2]. Building on this, in this abstract, we study the performance of the marginal-based DP SDG method Private-PGM [10] on two gene expression (bulk RNA-Seq) datasets provided in the CAMDA Health Privacy Challenge.[1]

CAMDA Health Privacy Challenge – Track I The goal of this challenge was to "develop privacy preserving generative methods that can mitigate privacy risks while preserving biological insights for bulk gene expression datasets". The platform for this challenge provided pre-processed versions of two open-access 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

[1] `https://benchmarks.elsa-ai.eu/?ch=4&com=introduction`

consisting of 4,323 samples and 97 features labeled for prediction of 10 different cancer tissue origins. The challenge also provided an evaluation pipeline to assess the quality and privacy of the generated synthetic data. In the following sections, we briefly describe our submission as a blue team to the challenge, the SDGs we experimented with, and our findings.

2 Methodology

Differential Privacy (DP). DP guarantees plausible deniability regarding an instance being present in a dataset, hence providing privacy guarantees [4]. Consider two neighboring datasets D and D' that differ by a single record. A randomized algorithm F is said to be (ϵ, δ)-DP if, for any pair of neighboring datasets D and D' and for all subsets O in the range of the output of F, it holds that $\Pr(F(D) \in O) \leq e^{\epsilon} \cdot \Pr(F(D') \in O) + \delta$. ϵ and δ are the privacy parameters that represent the privacy budget ϵ (measure of privacy loss), and the probability δ of the privacy being compromised, respectively [4]. DP ensures that the inclusion of any entry in the real dataset is obscured, in the sense that any output (synthetic dataset) obtained from computations over the real dataset would have been similarly likely to be reached whether the entry was present or not. The smaller the values of ϵ and δ, the stronger the privacy guarantees. A DP algorithm F is usually created out of an algorithm F^* by adding noise proportional to the sensitivity of F^*, where sensitivity is the maximum possible change in F^*'s output when comparing outputs on D and D'. The *Gaussian Mechanism* achieves this by adding noise drawn from a Gaussian (normal) distribution with 0 mean and with a standard deviation scaled to the l_2-sensitivity of F^*.

Synthetic data generation. Motivated by advances and the demonstrated success of deep learning models in biomedical applications, we began our study by exploring deep learning–based SDGs, even SDG algorithms that do not provide DP guarantees. Our initial experiments included GAN-based and diffusion-based SDGs, both of which were specifically designed for bulk RNA-seq data[2]. These approaches did not perform well on the CAMDA datasets. They were computationally intensive, time-consuming to adapt, produced synthetic data of generally low quality (reflected in utility metrics such as a low accuracy of 20% in cell type prediction with the TCGA-Combined data generated with diffusion model) lacking in meaningful biological correlations, leading to poor clustering across cell types. We also explored the use of large language models (LLMs) such as GPT-4, inspired by recent work on using LLMs for tabular and genomic data generation [16, 8]. Limitations of the model's context window led to poor scalability for long gene expression profiles, resulting in high generation costs and low quality of generated data. Given these challenges, and motivated by [2], we turned our focus to Private-PGM [10], an SDG algorithm with the additional advantage of offering formal DP guarantees.

Private-PGM. Private-PGM is a marginals based SDG that constructs an undirected graphical model from measurements over low-dimensional marginals, which facilitates the generation of new synthetic samples via sampling from the learned graphical model. Private-PGM provides DP guarantees by perturbing the measured marginals with Gaussian noise. The scale of the noise is, among other things, controlled by a privacy budget parameter ε. As mentioned above, the smaller the value of ε, the higher the formal privacy guarantees.

Private-PGM works by modeling and sampling from a DP approximation of the data's marginal probability distributions. Private-PGM requires categorical or discretized inputs. To make the pre-processed CAMDA dataset compatible with Private-PGM, we discretized the continuous gene expression values using quantile binning, transforming each feature into 4 equidepth discrete bins. We then computed a set of 1-way and 2-way marginals: the former capture univariate distributions for

[2] `https://forge.ibisc.univ-evry.fr/alacan/rna-diffusion`

each gene, while the latter include gene-label pairwise marginals to preserve relationships between features and class labels. These marginals are then made DP (by adding calibrated noise to the marginals) and these DP marginals are used to construct a probabilistic graphical model (PGM) as per the Private-PGM algorithm [10]. The learned PGM is then used to sample synthetic data samples.

3 Analysis

We present an analysis of the synthetic datasets generated using Private-PGM on two TCGA bulk RNA-seq datasets from the CAMDA challenge. Table 1 reports utility, fidelity, and privacy metrics according to the evaluation pipeline provided by the CAMDA Health Challenge. The first two columns, Accuracy (Acc.) and Area Under the Precision-Recall Curve (AUPRC), correspond to training a one-vs-rest logistic regression model on the synthetic data and testing it on the real data. Higher values indicate better utility. For reference, when the model is trained and tested on real data, the metrics for TCGA-BRCA are: Accuracy = 0.8650, AUPRC = 0.8656; and for TCGA-Combined: Accuracy = 0.9783, AUPRC = 0.9919. The third column reports statistical utility, measured by the number of overlapping important features (genes) between real and synthetic data; higher values indicate better agreement in feature importance. The next two columns, MMD and discriminative score, assess the fidelity of the generated synthetic data. While MMD (Maximum Mean Discrepancy) measures the distance between the probability distributions of synthetic and real datasets, discriminative score is the F1 score of a classifier trained to distinguish synthetic from real samples. Lower values for MMD and discriminative score are better. DCR (Distance to the Closest Real sample) is a privacy metric indicating the average distance from each synthetic data point to its nearest neighbor in the real dataset. Higher values suggest better privacy preservation. For comparison, the DCR on real data is 24.04 for TCGA-BRCA and 23.27 for TCGA-Combined.

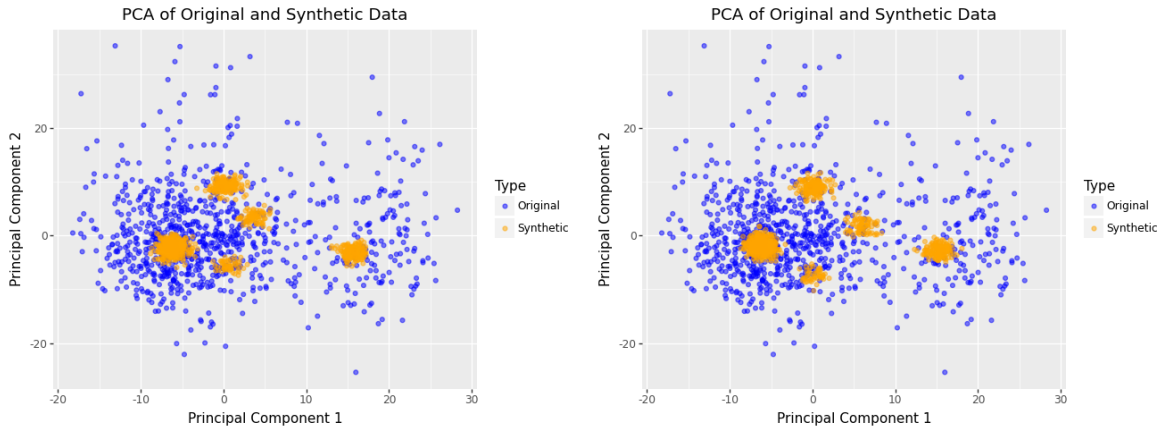
Table 1: **Evaluation of synthetic data generated with Private-PGM with different values of privacy loss (ε) for CAMDA Challenege datasets.** Baseline refers to the non-private SDG (best among all the baselines) provided by the competition organizers.

	ε	Acc. (↑)	AUPR (↑)	Feature Overlap (↑)	MMD (↓)	Discriminative Score (↓)	DCR (↑)
BRCA	Baseline	0.8476	0.8415	19.6	0.0180	0.5496	28.5348
	1	0.7163	0.5535	12	0.0524	1.0000	28.138
	2	0.7594	0.6669	12.4	0.0260	0.9996	27.869
	5	0.8118	0.7535	14	0.0106	0.9234	27.247
	7	0.8154	0.7630	15.4	0.0086	0.8638	27.219
	10	0.8292	0.7757	15	0.0074	0.7769	24.042
COMBINED	Baseline	0.9755	0.9908	65.6	0.00938	0.57832	28.87758
	1	0.8276	0.8190	42.2	0.0207	0.9831	33.70426
	2	0.8804	0.9213	43.4	0.0121	0.8742	31.6821
	5	0.9190	0.9587	44.6	0.0068	0.7248	29.9346
	7	0.9288	0.9671	43.8	0.0061	0.6666	29.5171
	10	0.9392	0.9656	47.8	0.0055	0.6357	29.2343

Table 1 summarizes the performance of Private-PGM on the CAMDA Health Challenge datasets across a range of privacy budgets $\varepsilon = (1, 2, 5, 7, 10)$. The baseline corresponds to the best-performing non-private SDG provided by the challenge organizers. As the baseline does not provide formal privacy guarantees via DP, it performs the best for utility and fidelity metrics. For both TCGA-BRCA and TCGA-Combined datasets, we observe that increasing the privacy loss ε improves utility (accuracy, AUPR, and feature overlap) and fidelity (MMD, discriminative score), gradually approaching the performance of the non-private baseline. The TCGA-Combined dataset consistently yields higher

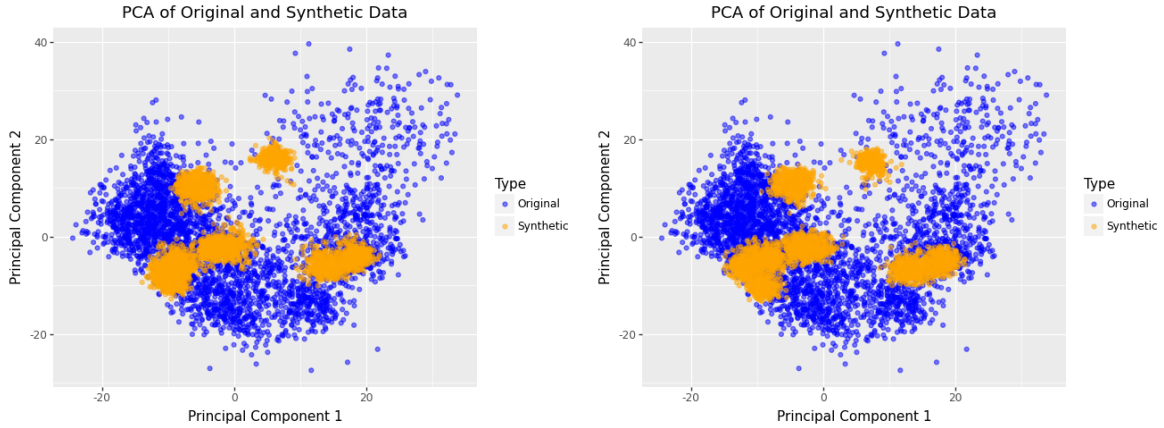
utility scores than BRCA, even at lower ε values. This may be due to its larger sample size and lower feature dimensionality, which make the data more amenable to marginal-based modeling. Interestingly, the non-private baseline exhibits higher DCR than the privacy-preserving methods. Moreover the trends in DCR with respect to ε do not follow the expected pattern as the other metrics did. However, a higher DCR does not imply stronger privacy in the absence of formal guarantees. Since DCR is only a heuristic measure of privacy, more rigorous evaluations, such as membership inference attacks (MIA) [5], are needed to assess the actual privacy risks associated with synthetic data generation methods.

Figure 1 visualizes the 2D PCA projections of original and synthetic data. We observe that Private-PGM-generated synthetic data (in orange) forms tight clusters aligned with dense regions of the real data (in blue). We think this could be because the model relies on lower-dimensional marginal statistics (e.g. 1-way and 2-way marginals) to reconstruct the joint distribution, which tends to concentrate generation in regions with stronger statistical support. Moreover, we do not see a particular trend with varying ε. One would expect a loosening of the clusters with increased values of ε. We believe a more rigorous study is needed to capture the trends.



(a) TCGA-BRCA, $\epsilon = 7$

(b) TCGA-BRCA, $\epsilon = 10$



(c) TCGA-COMBINED, $\epsilon = 7$

(d) TCGA-COMBINED, $\epsilon = 10$

Figure 1: **PCA Plots.** Visual comparison between generated synthetic data and real data

While Private-PGM seemed to be offering better performance based on several statistical utility and fidelity metrics in Table 1, the visualizations reveal that Private-PGM does not capture the biological correlations well. This is in line with findings by [2] on a different dataset.

Based on our study, we conclude that there is a gap in the current literature for DP SDGs for

genomics, i.e. for SDGs that offer formal differential privacy guarantees while preserving utility and fidelity along with biological patterns.

Acknowledgements. This research was supported by NSF #2451163, NIH 1OT2OD032581, NSF NAIRR 240485 (Cloudbank AWS), NSF NAIRR 240091 (TACC Frontera), and the UW Tacoma Founders Endowment. Shane Menzies is supported by a Mary Gates Scholarship. Steven Golob is supported by an NSF CSGrad4US Fellowship.

References

[1] P. A. Apellániz, J. Parras, and S. Zazo. An improved tabular data generator with VAE-GMM integration. In *32nd European Signal Processing Conference (EUSIPCO)*, pages 1886–1890. IEEE, 2024.

[2] 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.

[3] Y. Du and N. Li. Systematic assessment of tabular data synthesis algorithms. *arXiv preprint arXiv:2402.06806*, 2024.

[4] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[5] S. Golob, S. Pentyala, A. Maratkhan, and M. De Cock. Privacy vulnerabilities in marginals-based synthetic data. *IEEE Secure and Trustworthy Machine Learning Conference (SaTML)*, 2025.

[6] Y. Hu, F. Wu, Q. Li, Y. Long, G. Garrido, C. Ge, B. Ding, D. Forsyth, B. Li, and D. Song. SoK: Privacy-preserving data synthesis. In *IEEE Symposium on Security and Privacy (SP)*, pages 4696–4713, 2024.

[7] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko. TabDDPM: modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR, 2023.

[8] D. Levine, S. A. Rizvi, S. Lévy, N. Pallikkavaliyaveetil, D. Zhang, X. Chen, S. Ghadermarzi, R. Wu, Z. Zheng, I. Vrkic, et al. Cell2Sentence: teaching large language models the language of biology. *BioRxiv*, pages 2023–09, 2024.

[9] M. McDermott, S. Wang, N. Marinsek, R. Ranganath, M. Ghassemi, and t. **Foschini**. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, 13(586): eabb1655, 2021.

[10] 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.

[11] R. McKenna, G. Miklau, and D. Sheldon. Winning the NIST contest: A scalable and general approach to differentially private synthetic data. *Journal of Privacy and Confidentiality*, 11(3), 2021.

[12] R. McKenna, B. Mullins, D. Sheldon, and G. Miklau. AIM: an adaptive and iterative mechanism for differentially private synthetic data. *Proceedings of the VLDB Endowment*, 15(11):2599–2612, 2022.

[13] 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.

[14] Y. Tao, R. McKenna, M. Hay, A. Machanavajjhala, and G. Miklau. Benchmarking differentially private synthetic data generation algorithms. In *AAAI-22 Workshop on Privacy-Preserving Artificial Intelligence*, 2022.

[15] Y. Wang, Q. Chen, H. Shao, R. Zhang, and H. Shen. Generating bulk rna-seq gene expression data based on generative deep learning models and utilizing it for data augmentation. *Computers in Biology and Medicine*, 169:107828, 2024.

[16] Y. Wang, D. Feng, Y. Dai, Z. Chen, J. Huang, S. Ananiadou, Q. Xie, and H. Wang. HARMONIC: Harnessing LLMs for tabular data synthesis and privacy protection. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[17] 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 data sharing requirements: avoiding open data in appearance only. *BMJ Health & Care Informatics*, 30(1), 2023.

[18] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen. CTAB-GAN: Effective table data synthesizing. In *Asian Conference on Machine Learning*, pages 97–112. PMLR, 2021.