Comparison of Single Cell RNA Synthetic Data Generators: A CAMDA Health Challenge Analysis

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

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

1 Introduction

Single-cell RNAseq and its importance in cancer research: In recent years, single-cell RNA sequencing (scRNA-seq) has dramatically transformed cancer research by providing unprecedented resolution into tumor biology. Unlike bulk RNA sequencing, which masks cellular differences by averaging gene expression, scRNA-seq profiles individual cells, enabling the identification of distinct subpopulations and rare cell types within tumors [9]. This technology has surmounted issues of tumor heterogeneity and impurity that plagued earlier studies, allowing researchers to dissect the complex cellular composition of cancers and their microenvironments. As a result, a vast body of data now illuminates intratumoral diversity at the transcriptional level, revealing new insights into how cancers develop, evade treatment, and interact with their environment. In particular, scRNA-seq has opened new frontiers in understanding intratumor heterogeneity, profiling the tumor microenvironment (TME), uncovering mechanisms of treatment resistance, and discovering biomarkers for diagnostics and prognostics.

The need for synthetic scRNAseq data: An exciting development in the single-cell field has been the generation of synthetic scRNA-seq data to augment research. As powerful as scRNA-seq is, studies often face practical limitations: the number of cells that can be profiled from rare populations or precious patient samples may be low, and certain cell states (especially those relevant to disease) might be exceedingly rare or even missing in a given dataset [12]. Additionally most available datasets are heavily skewed toward healthy control samples, while studies focusing on rare diseases, specialized tissues, and uncommon cell types are underrepresented. Both of these scarcities can limit statistical power and impede the training of robust computational models [12]. In the past few years (2020–2025), researchers have turned to deep learning and generative modeling techniques to create artificial single-cell transcriptomes that mimic real cells, thereby expanding datasets in silico. These synthetic data approaches aim to model rare cell states, improve algorithm performance, and provide testbeds for method development without needing infinite lab experiments [12, 8, 2, 15]. Synthetic scRNA-seq data generated through these methods can advance use-cases like method development and benchmarking. Researchers can generate a large synthetic single-cell dataset with known "ground truth" (since they control the generative process) and then test how well clustering or trajectory inference algorithms perform. This was difficult to do with limited real data, but now frameworks like ACTIVA [7] or scGFT [12] can produce realistic benchmarks. Additionally, synthetic cells can aid in experiment planning: one can simulate an experiment's outcome (e.g., if I sequence 10,000 cells from this tumor, will I be able to detect a hypothesized rare cell type?) and adjust the design accordingly. This helps optimize the use of expensive sequencing resources. As of 2025, these synthetic data generation techniques are still mostly in the research phase, but they are rapidly maturing. They represent a convergence of genomics with advanced computational modeling, exemplifying how cross-disciplinary innovation is pushing the boundaries of cancer data analysis.

In this abstract, we explore and compare multiple types of synthetic data generators (SDGs) to generate single-cell RNA-seq (scRNA-seq) data using the OneK1K dataset provided by the CAMDA Healthcare Challenge[1]. Specifically, we evaluate both the statistical methods scDesign2 [14] and Private-PGM [11] (which also provides formal differential privacy guarantees) as well as the recent diffusion-based model cfDiffusion [19]. Our analysis follows the evaluation pipeline and metrics defined by the challenge organizers.

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

2 Methodology

Dataset description. The OneK1K dataset is a large scRNA-seq dataset containing expression quantitative trait loci (EQTL) for peripheral blood mononuclear cells (PBMCs) from 982 North European donors, with individual donors contributing between 544 and 3401 cells each [17]. Cells were assigned to donors using reference-based demultiplexing. Following doublet and empty droplet removal, counts data were normalized by the library size, sex, and age. Cell type (label) assignment involved an initial stage of "supervised clustering" of cell expression data based on a published PBMC dataset (yielding distinct clusters for Myeloid, NK, CD4, CD8, and B cells) followed by additional unsupervised clustering identifying fourteen cell types in total [17]. In total, this produced 1.2 million cells, which were aligned to 25,834 genes.

Synthetic data generation using multiple methods. The OneK1K data can be represented as tabular data in which the rows are cells and the columns are genes (or vice versa). While the landscape of synthetic data generation (SDG) algorithms is substantial and expanding, nearly all SDG algorithms for tabular data belong to one of two prominent classes. The first class consists of statistics based algorithms that (roughly speaking) learn a probability distribution over the real data and subsequently sample from it. This class includes SDG algorithms such as Private-PGM [11], PrivBayes [18], and RAP [1]. The second class consists of neural network-based methods. These rely on training neural networks that generate data resembling the training data. This class includes algorithms such as CTGAN [16], and diffusion based SDG algorithms such as TabDDPM [10]. Many, though not all, of these SDG algorithms offer formal privacy guarantees through the use of differential privacy (see below). In addition to these generic SDG algorithms for tabular data, recently several SDG methods were proposed specifically for RNA-seq data. These include the statistics based SDG algorithm scDesign2 [13] and the neural network based SDG algorithms cfDiffusion [19]. In this abstract we perform an empirical comparison of the statistics based scDesign2 technique with the neural network-based cfDiffusion method, evaluating the fidelity of each approach for the generation of single cell RNA-seq data. We further compare their performance with Private-PGM, an SDG technique known to achieve good performance for bulk RNA-seq data [5].

scDesign2: scDesign2 [13], a marginals-based method, was used for generation of synthetic scRNA-seq data. For each gene and each cell type, it fits an individual marginal distribution, either using Maximum Likelihood Estimates of Negative Binomial or Zero-Inflated Negative Binomial (ZINB) distributions for overdispersed genes and Poisson or Zero-Inflated Poisson distributions for all others [13]. For the majority of genes, scDesign2 samples from these individual distributions to create synthetic data. For some subset of genes exceeding a threshold of non-zero entries, scDesign2 incorporates individual distributions into a Gaussian copula by way of the Probability Integral Transform. For each gene i, a uniform distribution u_i is computed over the cumulative distribution function (CDF) of the fitted distribution; the final multivariate distribution follows the following formula, where $\Phi_p(...; R)$ denotes the CDF of a multivariate Gaussian distribution with covariance matrix R.

$$F(x_{1j}, ..., x_{pj}) = \Phi_p(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_p); R) \tag{1}$$

This enables scDesign2 to fit multiple types of distributions over individual genes while preserving correlations between them. scDesign2 used *highly variable genes* (HVGs) to generate the synthetic data. Following library-size and log1p-normalization, HVGs were selected according to the default parameters of scanpy, specifying a maximum dispersion of 0.5, a maximum mean of 3, and a minimum mean of 0.0125. This yielded 1,118 HVGs of 25,834 total genes.

Given the substantial memory and time requirements of scDesign2, we employed two strategies for synthetic data generation. Our first method trained scDesign2 copulae models for each of the fourteen cell types over the HVGs. Our second method used these copulae models to generate HVGs while fitting Poisson distributions individually over the normalized counts distributions of all non-HVG genes. We ran three versions of scDesign2 (see Table 1): one using the 1118 HVGs computed according to the above mentioned method; another using the top 5000 most variable HVGs; and a final method which used scDesign2 to generate counts for the 1118 HVGs while fitting Poisson distributions over the counts of the remaining genes following library-size and log1p normalization.

Figure 1 shows the UMAP representation of real and synthetic data generated via the scDesign2 (1118 HVGs) method.

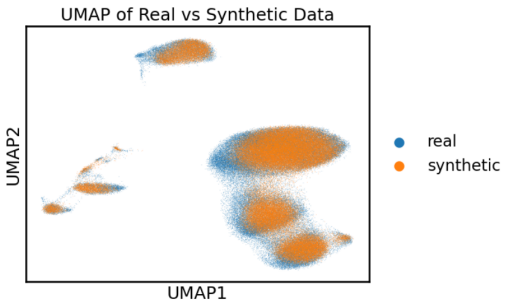

Figure 1: UMAP of real vs synthetic data generated via scDesign2 over 1,118 HVGs.

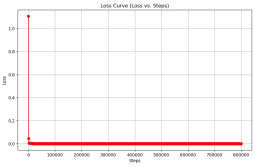

Figure 2: Loss in diffusion model training

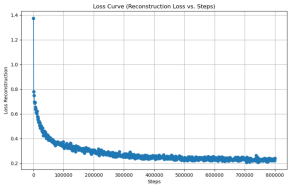

Figure 3: Loss for fine-tuning cfDiffusion AE from the pretrained SCimilarity model

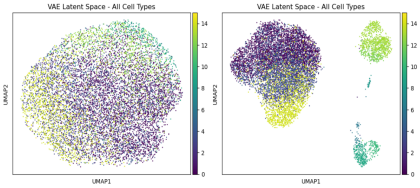

Figure 4: AE latent representations of the subsampled training dataset before and after fine-tuning AE

cfDiffusion: Deep learning based cfDiffusion model [19] was also used to generate synthetic data. cfDiffusion consists of a combination of an Autoencoder (AE) followed by a diffusion model. First, the Autoencoder (AE) based on SCimilarity foundation model [6] as its backbone, trained on 22.7 million cell corpus, is used to obtain n-dimensional representations of the raw scRNA-seq counts, transforming the discrete gene expression into a distribution. This latent representation is then used for predicting noise during the diffusion process based on U-Net model comprising of several Multilayer Perceptron (MLP) layers. So, given an initial cell embedding $\mathbf{x}_0 \sim q(\mathbf{x})$ from the training data distribution, the diffusion process gradually transforms it into a noisy embedding $\mathbf{x}_T$ by iteratively adding Gaussian noise over T time steps. The goal of training is to learn the reverse diffusion process $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$, which is also modeled as a Gaussian distribution. The resulting diffusion model is then used as a generative model by sampling random Gaussian noise and transforming it into realistic single-cell latent representations, such as ones produced by the AE. The AE is then used as a decoder to transform the representations back to scRNA-seq counts. cfDiffusion incorporates the cell types labels directly via a classifier-free mechanism during diffusion model training, to align the latent features to the target class simplifying training. In Figure 3, we can see the fine-tuning process, and gradual improvement of the model, with the loss decreasing as steps increase. In Figure 4 we can see the result of fine-tuning AE on our dataset. While initially the latent representations seem to demonstrate little to no separation between various cell types, after 800k epochs of fine-tuning the model on our dataset, the model learns to distinguish different cell types more clearly in the latent space, which should lead to better performance.

Private-PGM. Private-PGM [11] 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 differential privacy (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. The smaller the value of ε, the higher the formal privacy guarantees.

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 captured univariate distributions for each gene, while the latter included gene-label pairwise marginals to preserve relationships between features and class labels. Next, these marginals are 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. The learned PGM is then used to sample synthetic data samples. We used an ε value of $\varepsilon = 10$ for all single-cell experiments with Private-PGM.

3 Analysis

Fidelity & Utility We compared these methods across several metrics included in the baseline CAMDA code. The Spearman and Pearson correlation coefficients are computed between the means of each gene in real and synthetic data, indicating the extent to which average expression levels are preserved in the synthetic data generation process. The Maximum Mean Discrepancy measures the distance of feature means from the real and synthetic data in a high-dimensional space; real and synthetic data are first subjected to Principal Component Analysis, and the MMD computed over their PCA representations using the Radial Basis Function (RBF) as a kernel. The Adjusted Rand Index (ARI) measures how well two datasets cluster together, with higher values (close to 1) indicating a greater degree of clustering, 0 values indicating randomness, and negative values indicating worse-than-random clustering. A principal-component representation of the real and synthetic data, with corresponding indices having identical cell types, are clustered using the Louvain method [3] and the ARI computed over the corresponding clusters. The Local Inverse Simpson's Index likewise indicates the degree to which real and synthetic data cluster together. The integration LISI of an individual data point is the inverse of the sum of squared proportions of each category (real vs synthetic) over the 10 closest cells, yielding a value between 1 and C where C is the number of categories; the CAMDA code computes such a metric over all datapoints, takes its mean, and normalizes this by $\frac{1}{C}$, yielding a value between 0 and 1, with high values suggesting high mixture between real and synthetic data.

Table 1 presents these metrics across the aforementioned methods. For each method, we trained a model on the 1118 HVGs (or 5000 HVGs) identified according to the methodology above and evaluated against the test set distributed by the CAMDA organizers over the same set of genes.

	scDesign2 (1118 HVGs)	scDesign2 (5000 HVGs)	scDesign2 + Poisson	cfDiffusion (1118 HVGs)	Private PGM (1118 HVGs)
Spearman CC	0.997	0.988	0.997	0.841	0.955
Pearson CC	0.997	0.985	0.960	0.713	0.667
MMD	0.0001	0.001	0.0002	0.001	0.0001
LISI	0.819	0.548	0	0	0
ARI Real vs Synthetic	0.512	0.374	0.676	0.15	−0.006
ARI GT vs Combined	0.567	0.312	0.439	0	0.236

Table 1: Quality metrics for different methods of synthetic data generation. Higher Spearman Correlation Coefficients, Pearson Correlation Coefficients, Local Inverse Simpson's Indices, and ARIs indicate higher-quality data, while lower MMDs indicate better quality.

	scDesign2 (1118 HVGs)	cfDiffusion (1118 HVGs)	Private PGM (1118 HVGs)
Accuracy	0.50	0.50	0.50
AUCROC	0.499	0.50	0.50
Average Precision	0.5005	0.50	0.50

Table 2: Accuracy of DOMIAS membership inference attack against various method.

Privacy Vulnerabilites We consider the vulnerability of each of the aforementioned data generation methods to Membership Inference Attacks (MIA). The baseline code for the CAMDA challenge implements a GAN-leaks attack [4], a distance-based attack. Given samples $S_G^k = \{x_i\}_{i=1}^k$ from some generative model G and some test datapoint $x*$ whose membership in the train set is unknown, the membership attack score $A(x*; G)$, scaling negatively with the probability of membership, is assigned based on the distance to the nearest point in the synthetic dataset.

$$A(x*; G) = min_{x_i \in S_G^k} ||x* - x_i||_2 \tag{2}$$

In the baseline implementation, this the probability of membership is calculated as $e^{-A(x*;G)}$. We ran this attack across all synthetic data generation methods using a subset of the OneK1K dataset provided by the organizers which contained 131,636 records of train (member) and test (non-member) data in equal quantities.

Table 2 shows the accuracy and AUROC of the GAN-leaks attack in predicting membership. On all methods, the DOMIAS attack scores almost exactly as one would expect from random guessing; this is because the distance between the real and synthetic datasets is so great that it assigns each datapoint as a non-member, corresponding to exactly 50% of the synthetic dataset. This can be interpreted either to indicate the poorness of the attack or that of the synthetic data.

Discussion: Our results indicate that scDesign2 best preserves the clustering within the additional data, as indicated by high ARI and LISI scores; cfDiffusion and Private-PGM perform very poorly along these metrics, achieving ARIs and LISIs close to 0, the former indicating nearly random clustering and the latter indicating almost no mixture between the neighborhoods of real and synthetic data. Figure 1, showing the UMAP distribution of real vs synthetic data generated by scDesign2 over the 1118 HVGs, further demonstrates the quality of scDesign2's clusters, with synthetic data accurately preserving cell type clusters present in the training data. The poor performance of cfDiffusion likely stems from an issue during diffusion model training, where the diffusion model does not seem to be able to use latent representation meaningfully to generate realistic looking data, leading to mode collapse, as can be seen in Figure 2, where the model stops learning early. This may be due to significant class imbalance (from 228797 cells for cell type 0 to 1815 cells for cell type 9), leading to poorly separated latent representations, or due to limited size of the diffusion model. However, our results also show successful generation of synthetic data which has high fidelity to original data and also presents high utility. scDesign2 was our method of choice which showed the best performance due to its strong reliance on highly variable genes to capture and recapitulate the strongest patterns in the original data. The degradation of scDesign2's performance as more genes are added is unsurprising; as the definition of HVG becomes more permissive, the model will begin to fit increasingly noisy genes, leading to degradation in model quality.

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] S. Aydore, W. Brown, M. Kearns, K. Kenthapadi, L. Melis, A. Roth, and A. A. Siva. Differentially private query release through adaptive projection. In *International Conference on Machine Learning*, pages 457–467. PMLR, 2021.

[2] S. Bej, A.-M. Galow, R. David, M. Wolfien, and O. Wolkenhauer. Automated annotation of rare-cell types from single-cell RNA-sequencing data through synthetic oversampling. *BMC bioinformatics*, 22:1–17, 2021.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[4] D. Chen, N. Yu, Y. Zhang, and M. Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, page 343–362. ACM, Oct. 2020. doi: 10.1145/3372297.3417238. URL `http://dx.doi.org/10.1145/3372297.3417238`.

[5] 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.

[6] G. Heimberg, T. Kuo, D. DePianto, T. Heigl, N. Diamant, O. Salem, G. Scalia, T. Biancalani, S. Turley, J. Rock, et al. Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages. *BioRxiv*, pages 2023–07, 2023.

[7] A. A. Heydari, O. A. Davalos, L. Zhao, K. K. Hoyer, and S. S. Sindi. Activa: realistic single-cell rna-seq generation with automatic cell-type identification using introspective variational autoencoders. *Bioinformatics*, 38(8): 2194–2201, 02 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac095. URL `https://doi.org/10.1093/bioinformatics/btac095`.

[8] A. A. Heydari, O. A. Davalos, L. Zhao, K. K. Hoyer, and S. S. Sindi. ACTIVA: realistic single-cell RNA-seq generation with automatic cell-type identification using introspective variational autoencoders. *Bioinformatics*, 38 (8):2194–2201, 2022.

[9] D. Huang, N. Ma, X. Li, Y. Gou, Y. Duan, B. Liu, J. Xia, X. Zhao, X. Wang, Q. Li, et al. Advances in single-cell RNA sequencing and its applications in cancer research. *Journal of Hematology & Oncology*, 16(1):98, 2023.

[10] 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.

[11] 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.
[12] N. Nouri. Single-cell RNA-seq data augmentation using generative Fourier transformer. *Communications Biology*, 8(1):113, 2025.
[13] T. Sun, D. Song, W. V. Li, and J. J. Li. scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured, 2021.
[14] T. Sun, D. Song, W. V. Li, and J. J. Li. scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome biology*, 22(1):163, 2021.
[15] M. Treppner, A. Salas-Bastos, M. Hess, S. Lenz, T. Vogel, and H. Binder. Synthetic single cell rna sequencing data from small pilot studies using deep generative models. *Scientific reports*, 11(1):9403, 2021.
[16] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems*, 32, 2019.
[17] S. Yazar, J. Alquicira-Hernandez, K. Wing, A. Senabouth, M. G. Gordon, S. Andersen, Q. Lu, A. Rowson, T. R. P. Taylor, L. Clarke, K. Maccora, C. Chen, A. L. Cook, C. J. Ye, K. A. Fairfax, A. W. Hewitt, and J. E. Powell. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science*, 376(6589): eabf3041, 2022.
[18] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. PrivBayes: Private data release via Bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):1–41, 2017.
[19] T. Zhang, Z. Zhao, J. Ren, Z. Zhang, H. Zhang, and G. Wang. cfDiffusion: diffusion-based efficient generation of high quality scRNA-seq data with classifier-free guidance. *Briefings in Bioinformatics*, 26(1):bbaf071, 2025.