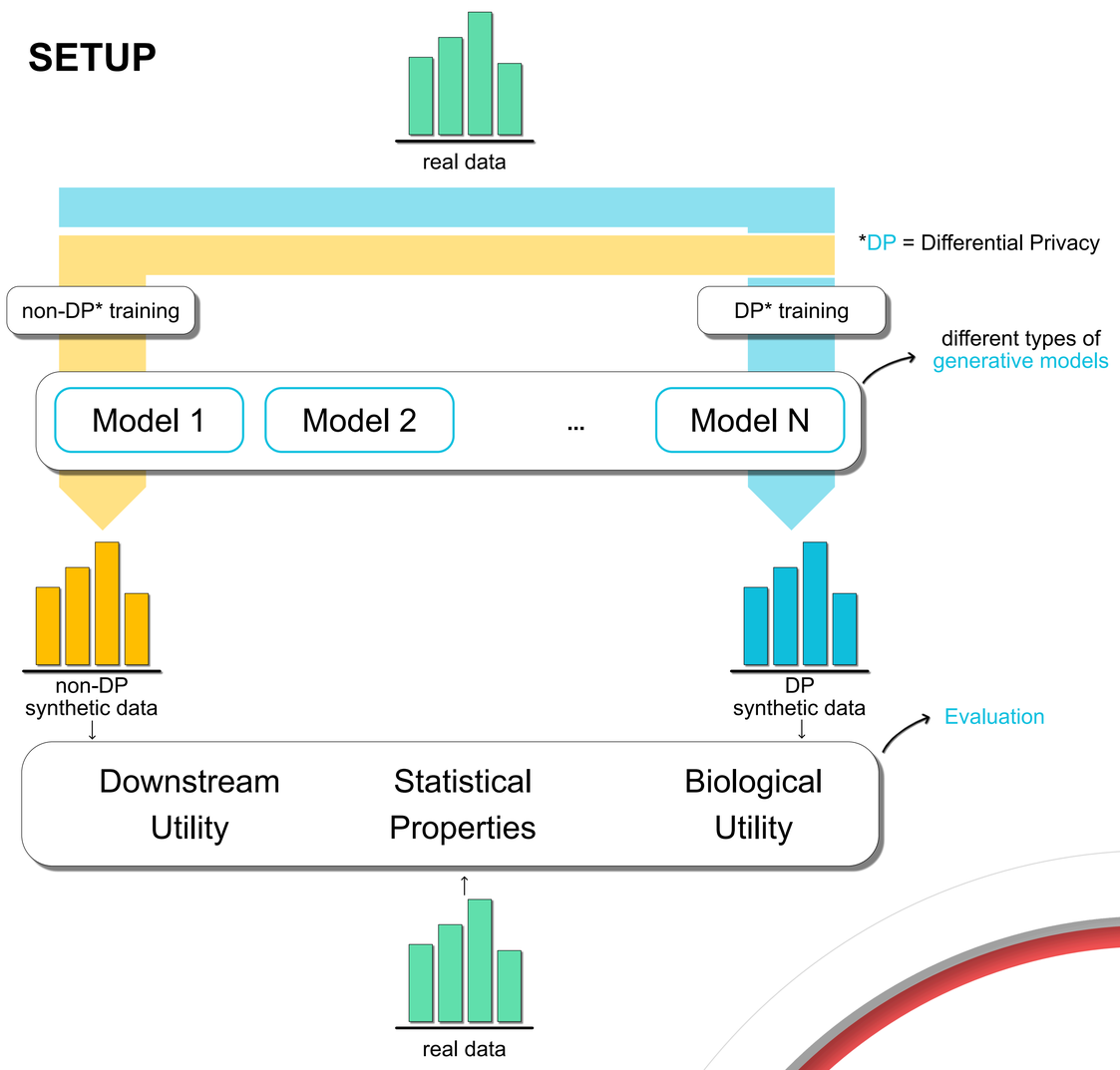
Towards Biologically Plausible and Private Gene Expression Data Generation

D. Chen[2], M. Oestreich[1], **T. Afonja**[2], R. Kerkouche[2], M. Becker[1], M. Fritz[2]

[1] Modular High Performance Computing and Artificial Intelligence, Systems Medicine, German Center for Neurodegenerative Diseases (DZNE), Bonn, North Rhine-Westphalia, Germany
[2] CISPA Helmholtz Center for Information Security, Saarbrücken, Saarland, Germany

1 SETUP



2 MODELS

RON-GAUSS[1]

Category: Density estimation

Attribute Type: Continuous only

DP Sanitization: One-shot

[1] Thee Chanyaswad, Changchang Liu, and Prateek Mittal. 2019. RON-Gauss: Enhancing Utility in Non-Interactive Private Data Release. Proceedings on Privacy Enhancing Technologies 1 (2019), 26–46.

VAE[2]

Category: Deep generative model

Attribute Type: Continuous

DP Sanitization: iterative

[2] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In 2nd International Conference on Learning Representations (ICLR).

GAN[3]

Category: Deep generative model

Attribute Type: Continuous

DP Sanitization: Iterative

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. Advances in neural information processing systems 27 (2014).

Private-PGM[4]

Category: Graphical model

Attribute Type: Discrete only

DP Sanitization: One-shot

[4] Ryan McKenna, Daniel Sheldon, and Gerome Miklau. 2019. Graphical-model based estimation and inference for differential privacy. In International Conference on Machine Learning. PMLR, 4435–4444.

PrivSyn[5]

Category: Marginal

Attribute Type: Discrete only

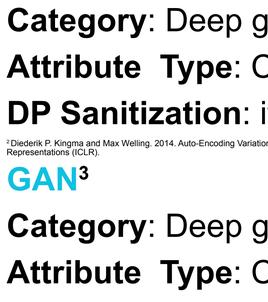
DP Sanitization: One-shot

[5] Zhikun Zhang, Tianhao Wang, Ninghui Li, Jean Honorio, Michael Backes, Shibo He, Jiming Chen, and Yang Zhang. 2021. {PrivSyn}: Differentially Private Data Synthesis. In 30th USENIX Security Symposium (USENIX Security 21). 929–946.

3 METRICS

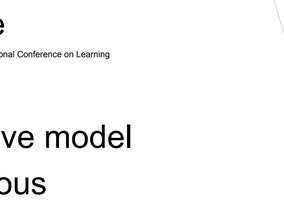
Downstream Utility

Classification

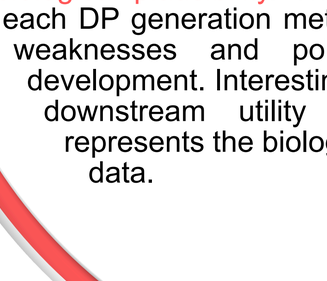


Statistical Properties

Histogram Intersection

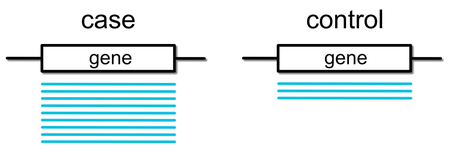


Distance to Closest Record

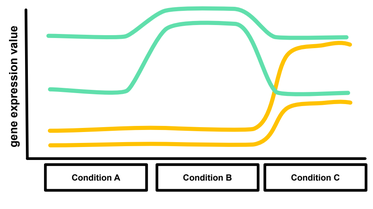


Biological Utility

Differential Gene Expression



Gene Co-Expression



Abstract

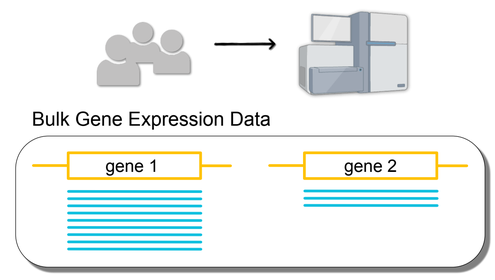
The use of generative models trained with differential privacy (DP) for creating synthetic data in various applications has gained significant attention. However, previous studies have mainly focused on basic benchmarking datasets, reporting positive results primarily for fundamental metrics and straightforward data distributions. This work presents a systematic analysis of the performance of DP generative models in real-world scenarios, specifically focusing on gene expression data.
We examine five prominent DP generation methods and evaluate them based on downstream utility, statistical properties, and biological plausibility. The findings reveal unique features of each DP generation method, highlighting their strengths and weaknesses and pointing out directions for future development. Interestingly, most methods show promising downstream utility results, but none accurately represents the biological characteristics of real-world data.

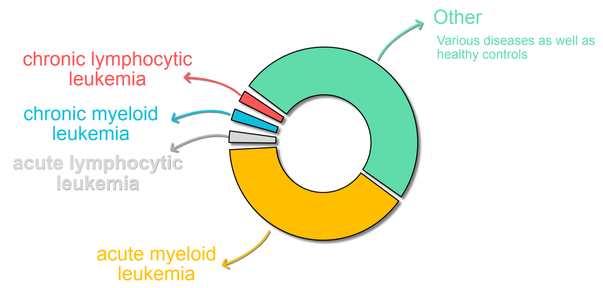
Paper accepted @ PoPETS 2024

GitHub

Scan Me !

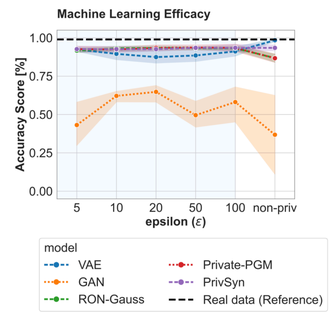
4 DATA



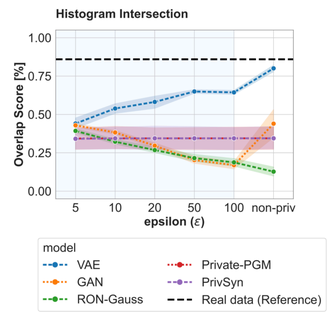


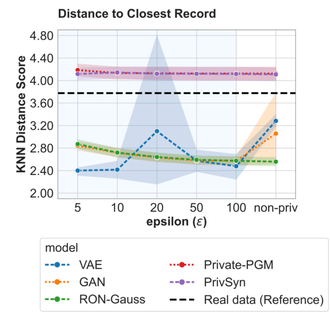
5 RESULTS

Downstream Utility

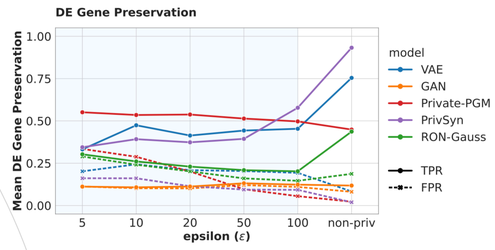


Statistical Properties





Biological Utility

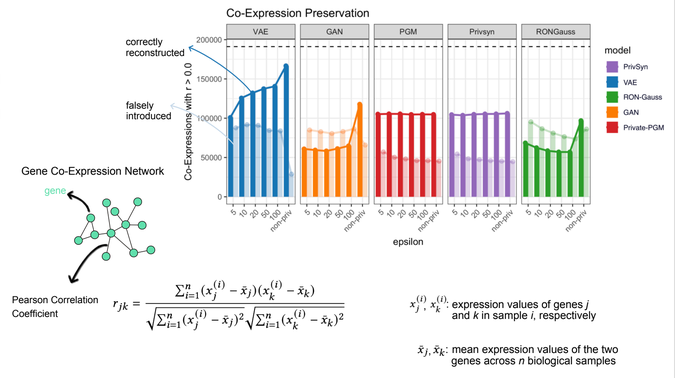


$$\mathrm{TPR}_{\mathcal{M},\varepsilon} = \frac{\sum_{(a_i,a_j)\subset\mathcal{A},a_i\neq a_j} \left(\mathrm{TPR}^{\mathrm{up}}_{a_i,a_j,\mathcal{M},\varepsilon} + \mathrm{TPR}^{\mathrm{down}}_{a_i,a_j,\mathcal{M},\varepsilon}\right)}{2\cdot\binom{|\mathcal{A}|}{2}}$$

$\mathcal{A}$: set of condition pairs (each pair representing different disease types distinguished by unique label classes in our case)

$\mathrm{TPR}^{\mathrm{up\,(down)}}_{a_i,a_j,\mathcal{M},\varepsilon}$: true positive rate for identifying up- (or down-) regulated DE-genes within synthetic data generated by the model $\mathcal{M}$ under a given privacy budget ε, in comparison to the actual DE-genes observed in the real dataset for conditions a_i and a_j.

$\binom{|\mathcal{A}|}{2}$: count of all possible unordered condition pairs



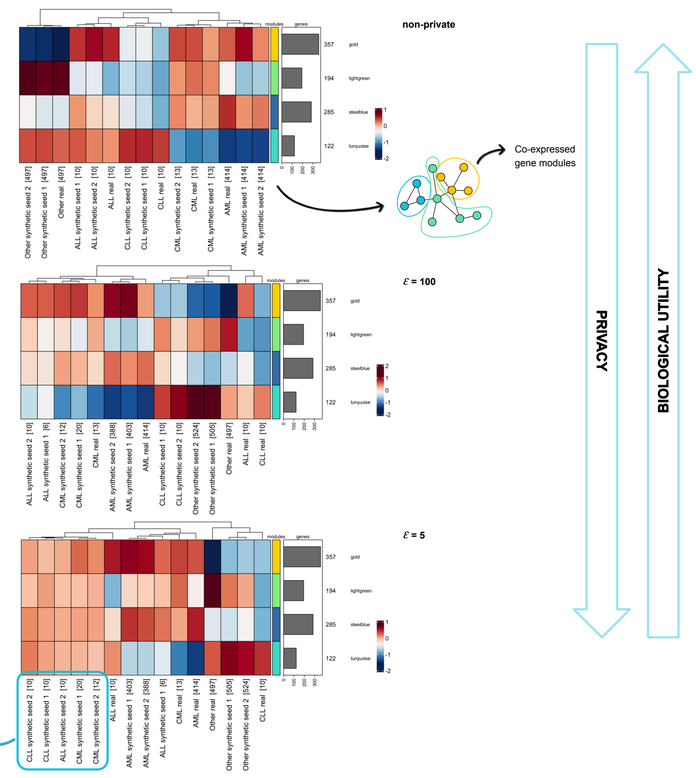
Gene Co-Expression Network

Pearson Correlation Coefficient

$$r_{jk} = \frac{\sum_{i=1}^{n}(x_j^{(i)} - \bar{x}_j)(x_k^{(i)} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_j^{(i)} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_k^{(i)} - \bar{x}_k)^2}}$$

$x_j^{(i)}, x_k^{(i)}$: expression values of genes j and k in sample i, respectively

$\bar{x}_j, \bar{x}_k$: mean expression values of the two genes across n biological samples



If biological utility is maintained in the synthetic data, **synthetic and real instances of each condition should cluster together**, i.e. have similar expression patterns across gene modules.

Especially classes with low sample size loose their gene-expression patterns.

6 CONCLUSION

We provide the first systematic analysis of non-private and differentially private generation of gene expression data that covers five diverse modeling approaches. Our analysis encompasses a diverse set of metrics that shed light on the quality of the generated data in terms of statistical and biological properties as well as down-stream utility.

Key Messages:

- A broad evaluation is necessary in order to understand the limitations of current generators.
- Simple estimators fall behind in performance but equally very complex models like GAN are suffering from the low sample regime typically encountered in bio-medical applications.
- While downstream utility can be strong, the synthetic data itself might not retain statistical nor biological properties.
- Adding privacy guarantees amplifies these problems.

7 ACKNOWLEDGEMENTS

Dingfan Chen was partially supported by Qualcomm Innovation Fellowship Europe.