Identifying GANs Blind Spots in Transcriptomic Data Generation

Assmaa Alsamadi[1], Alice Lacan[1], Blaise Hanczar[1], and Michèle Sebag[2]

[1] IBISC, University Paris-Saclay (Univ. Evry), Evry 91000, France
[2] TAU, CNRS-INRIA-LISN, University Paris-Saclay, Gif-sur-Yvette 91190, France

Abstract. Generative Adversarial Networks (GANs) show promise for data augmentation. However, they can overfit high-dimensional transcriptomic data, failing to capture biological diversity. This paper introduces a new metric that reveals discrepancy patterns between real and generated data, often overlooked by standard metrics. Our evaluation opens new perspectives for GAN-based data augmentation, balancing a priori the diverse modes of the real data and aligning a posteriori the generated data.

Keywords: Generative Modeling · Tabular Data · Transcriptomics

1 Introduction

Transcriptomics data (e.g., microarrays [1]) provide complex information on individual gene expression levels, enabling cancer diagnosis and prognosis predictions in cancer research [2]. Deep learning (DL) models excel at extracting features from complex high-dimensional data. However, these models require large training datasets to avoid overfitting. The largest public transcriptomics datasets are smaller by several orders of magnitude than the standards of DL models, making their implementation challenging.
To address data scarcity, model-based data augmentation methods have been proposed [3]. Deep generative models (e.g., Generative Adversarial Networks (GANs) [4]) can generate realistic synthetic data, offering an alternative to traditional augmentation techniques unsuitable for tabular data (e.g., cropping). Despite their widespread adoption, GANs often struggle to cover the data diversity. This phenomenon is harder to detect in tabular data, where human perceptual analysis is not feasible. A systematic evaluation of GANs is of the utmost importance, especially when the synthetic data is used for downstream medical applications.

2 Context & Contribution

Context. Following the GAN-based augmentation strategy of [5], we focus on the Wasserstein GAN with Gradient Penalty (WGAN-GP) [4]. This model trains simultaneously a generator to create realistic data and a critic evaluator.

General Challenges. Intrinsic biases in training data and GANs can cause reduced diversity in generated data, known as mode dropping and mode collapse. The former happens when the GAN misses specific modes in the data, while the latter results in overly similar samples within a mode.
State-of-the-Art-Metrics. The widely adopted Precision/Recall (PR) metrics [6] evaluate the realism-diversity trade-off. Conversely to their classification counterparts, they rely on manifold approximation and local neighborhoods to measure overlap between generated and true samples. However, these metrics can underestimate the lack of diversity in generated data, as such approximations are very sensitive to outliers and high dimensionality.
Contribution. We suggest a new method to assess data diversity by estimating the distance to the underlying manifold. Our tractable manifold approximation is based on extracting the real data principal components and using the reconstruction error as a distance proxy. The proposed PCA-based **reconstruction error metric** (REM) is defined as:

$$REM(\mathbf{x_i}) = \|\mathbf{x}_i - PP^T\mathbf{x}_i\|_2 \tag{1}$$

Where $\mathbf{x_i}$ is an original input sample, P is the matrix of the 2,000 first principal components (PCs) [3] resulting from the PCA performed on the real data, and $\|.\|_2$ is the Euclidean norm. A lower REM distribution in the generated data compared to true data should thus account for the diversity loss.

3 Results & Experimental validation

Experimental setting. We used a benchmark microarray dataset [7], retaining 32,043 genes after mapping preprocessing[4]. The data was standardized to zero mean and unit variance. Our feature selection approach achieved 98.77% $\pm$ 0.08 test multiclass accuracy[5], outperforming [2] by 4% (train-test split is 10,643-1,313 samples). Our best WGAN-GP was optimized over 40 runs using Bayesian optimization and maximizing the F_1 score over PR. The optimal model was trained with 500 epochs, batch normalization, batch size of 64, 1e-3 (generator) and 5e-4 (discriminator) learning rates, and Adam optimizer.

Results. In Fig. 1B, the UMAP visualization of real (colored) and data generated by our best WGAN-GP (black) shows a compelling data manifold coverage. This model achieved a relatively good precision of 93.33% $\pm$ 0.31 and a recall of 69.81% $\pm$ 0.41. However, a closer look in Fig. 1C shows that recall performance varies across tissue types regardless of their dataset representation. For instance, the 'ovary' cohort has poor recall (78.57% 0 recall) and low representation (3.31%), while 'bone' is undercovered (49.53% 0 recall) despite being the second most prevalent tissue (25.4%). We analyze this mode collapse through

[3] The PCs that explain the maximum variance.
[4] Mapping was performed using: http://www.ensembl.org/biomart
[5] All results are averaged over five runs on a 48GB A40 NVIDIA GPU.

additional covariates showing 43.54% of cancerous data and 23.69% healthy data had a 0 recall. This indicates that average PR metrics can conceal GAN failures.

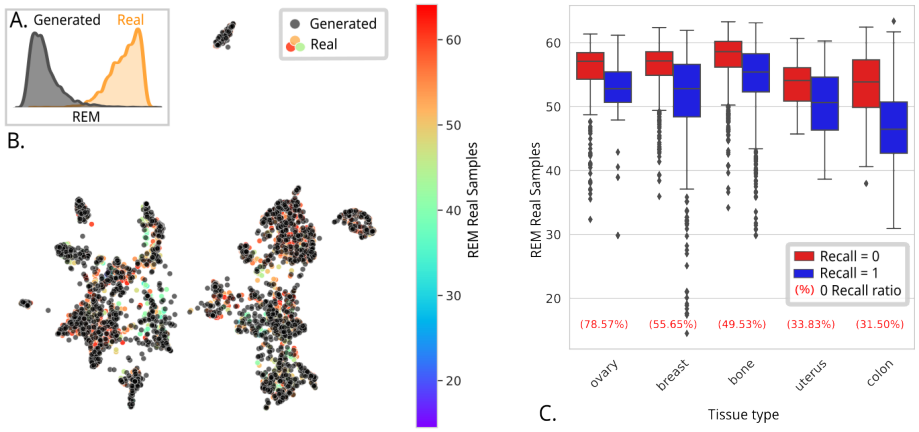

Fig. 1: Multifaceted evaluation of the realism-diversity trade-off: A. Reconstruction Error Metric (REM) distribution for generated (black) and real (orange) data; B. UMAP of generated (black) and real samples colored by their RE; C. REM of the five worst tissue types according to the real samples with recall=1 and recall=0.

In contrast, our REM metric provides deeper insights into failure modes. Fig. 1A shows a great discrepancy between real and generated data REM, clearly highlighting **mode collapse**: generated samples closely resemble a few originals on specific dimensions with uninformative noise in others. In Fig. 1B, some high-REM real samples (orange) are uncovered by generated data, suggesting they are in hard-access regions. Our REM metric also confirms the cancerous data heterogeneity assumption: i) Fig. 1C demonstrates that low recall (red) correlates with higher REM; ii) on average, cancerous data REM is higher by 4.6 than healthy data per tissue type.

4 Conclusion

Our multifaceted evaluation of GANs on gene expression data revealed key insights beyond PR metrics. First, our metric shows that noise-corrupted original samples can mimic realism. Secondly, high REM values underscore how data heterogeneity affects diversity learning. Lastly, dimensionality reduction visualizations can falsely suggest sample proximity. These findings highlight the GANs' imitation bias that some approaches address with outlier-robust indicators [8] or GANs adapted to disconnected manifolds [9] at increased computational cost. We propose a data-centric solution: smoothing the generated manifold through Optimal Transport [10] alignment between collapsed samples (low REM) and diversified samples (high REM). Similar balancing between the diverse modes of the real data could be conducted a priori.

References

1. Govindarajan, R., Jeyapradha, D., et al.: Microarray and its applications. Journal of Pharmacy and Bioallied Sciences **4**(2) (2012). https://doi.org/10.4103/0975-7406.100283
2. Hanczar, B., Bourgeais, V., et al.: Assessment of deep learning and transfer learning for cancer prediction based on gene expression data. BMC Bioinformatics **23**, 262 (2022), https://doi.org/10.1186/s12859-022-04807-7
3. Viñas, R., Andrés-Terré, H., Liò, P., Bryson, K.: Adversarial generation of gene expression data. Bioinformatics **38**(3), 730–737 (2021). https://doi.org/10.1093/bioinformatics/btab035
4. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein gans. In: NeurIPs (2017)
5. Lacan, A., Sebag, M., Hanczar, B.: GAN-based data augmentation for transcriptomics: survey and comparative assessment. Bioinformatics **39**(Supplement 1), i111–i120 (06 2023). https://doi.org/10.1093/bioinformatics/btad239
6. Kynkäänniemi, T., Karras, T.: Improved precision and recall metric for assessing generative models. In: NeurIPS (2019)
7. Torrente, A.: A comprehensive human expression map. https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-3732 (2015)
8. Alaa, A., van Breugel, B., et al.: How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In: ICML (2022)
9. Khayatkhoei, M., Elgammal, A., et al.: Disconnected manifold learning for generative adversarial networks. In: NeurIPS (2018)
10. Vilani, C.: Optimal Transport: Old and New. Springer (2009)