

Balancing Utility and Privacy: Evaluating Synthetic Gene Expression Data with a Novel Dilution Metric for Feature Stability

Setareh Akbari

setareh-akbari@utulsa.edu

The University of Tulsa

Jamie Li

Stanford University

Brett McKinney

University of Tulsa

Research Article

Keywords:

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9707995/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Balancing Utility and Privacy: Evaluating Synthetic Gene Expression Data with a Novel Dilution Metric for Feature Stability

Setareh Akbari^{1,*}, Yijie (Jamie) Li², Brett McKinney³

¹Oklahoma Cyber Innovation Institute, The University of Tulsa, Tulsa, OK, USA (setareh-akbari@utulsa.edu).

²Stanford University, Palo Alto, California, USA (jamieyl@stanford.edu).

³Tandy School of Computer Science, University of Tulsa, Tulsa, OK, USA (brett-mckinney@utulsa.edu).

*Corresponding author: S. Akbari (setareh-akbari@utulsa.edu).

Abstract

Background

Background: Synthetic data generation has emerged as a promising approach to address privacy and data-sharing constraints in biomedical research. In high-dimensional omics settings, however, evaluating synthetic data remains challenging, as existing metrics often fail to capture whether biologically meaningful signals are preserved. In particular, synthetic data may achieve high predictive utility or statistical similarity while altering feature importance rankings, which are critical for biomarker discovery.

Results

We present a reproducible evaluation framework for synthetic gene-expression data and apply it to a peripheral blood RNA-seq dataset for major depressive disorder. Two generative models, Adversarial Random Forest and Conditional Tabular GAN, are evaluated across six complementary dimensions: predictive utility, statistical fidelity, privacy with a domain-separator model summarized by the area under the receiver operating characteristic curve,

feature-importance alignment, bootstrap stability, and a novel metric termed dilution. Dilution quantifies the extent to which synthetic data perturb feature-importance rankings derived from real data, providing a direct measure of interpretability preservation. Across these metrics, Adversarial Random Forest demonstrates a more favorable overall balance by maintaining predictive performance, preserving gene-wise distributions, and exhibiting lower real–synthetic separability. In contrast, Conditional Tabular GAN introduces substantial distributional shifts and produces greater disruption of feature rankings, reflected in higher dilution and weaker feature alignment. As a proof-of-concept extension, we include a minimal quantum–adversarial robustness checks on the real dataset: a 6-qubit parameterized quantum classifier achieves accuracy comparable to that of classical baselines and shows modest transfer robustness under fast gradient sign method perturbations.

Conclusions

Our results highlight that commonly used evaluation metrics may overlook important distortions in biologically meaningful signals. The proposed dilution metric provides a simple and interpretable measure of feature-ranking stability after synthetic data augmentation, complementing existing assessments of utility and fidelity. Together, this framework supports a more reliable evaluation of synthetic omics data and may help guide the development of privacy-preserving approaches that better retain biological interpretability.

1. Introduction

Machine learning has become an important tool in biomedical informatics for identifying molecular disease signatures, with gene-expression profiling and genome-wide association studies (GWAS) supporting disease classification and biomarker discovery in complex neuropsychiatric conditions such as major depressive disorder (MDD). However, privacy concerns and data-sharing restrictions often limit access to the large, diverse cohorts needed

for robust model development. Because high-throughput omics data can reveal sensitive clinical information and pose re-identification risks, ethical and regulatory considerations frequently restrict broad data sharing [1-3]. As a result, available cohorts are often smaller and less diverse, reducing reproducibility and limiting model generalizability. Synthetic data generation has emerged as a promising strategy for addressing privacy and data-sharing challenges in biomedical research. Its goal is to generate artificial datasets that preserve the statistical and biological structure of real data while reducing the risk of identifying individual participants. In practice, however, this requires balancing three competing objectives: utility, meaning support for downstream analytical tasks; fidelity, meaning agreement with the statistical properties of the original data; and privacy, meaning protection against the disclosure of identifiable information. Achieving this balance is particularly difficult in high-dimensional omics settings with limited sample sizes, where subtle dependency patterns may carry important biological information [4]. As a result, synthetic data that appear statistically realistic may still weaken predictive performance or distort biologically meaningful signals, whereas stronger privacy protection may remove relevant structure.

Recent studies on bulk RNA sequencing (RNA-Seq) data have shown that deep generative models can produce realistic transcriptomic profiles. In addition, augmenting real datasets with generated samples has been reported to improve downstream classification performance, suggesting that synthetic augmentation may offer a practical strategy for addressing limited cohort sizes [5]. Recent approaches to synthetic tabular data generation include both deep generative models and ensemble-based probabilistic learners. Conditional Tabular GAN (CTGAN) is a generative adversarial model developed for heterogeneous tabular data and can capture complex nonlinear dependencies and multimodal distributions [6]. In contrast, Adversarial Random Forest (ARF) is an ensemble-based method that uses tree-based modeling to approximate joint structure more stably in small-sample settings, including those commonly

encountered in omics research [7, 8]. In practice, CTGAN provides flexible and expressive synthesis but may be sensitive to limited sample sizes and mixed data types, while ARF often favors stability and preservation of marginal and interaction structure over generative flexibility [8].

Despite the growing interest in privacy-preserving synthetic data generation for biomedical applications, systematic quantitative comparisons of these approaches in omics settings remain limited [9]. To address this gap, we propose a comprehensive multi-metric framework for evaluating synthetic omics data. Specifically, we assess synthetic datasets across six complementary dimensions: (i) utility, measured by downstream predictive performance on real tasks; (ii) fidelity, assessed through gene-level distributional similarity using Kolmogorov-Smirnov (KS) and Wilcoxon tests with Benjamini-Hochberg false discovery rate (BH-FDR) control; (iii) privacy, evaluated by the ability of a domain-separator model to distinguish real from synthetic samples, summarized by the area under the receiver operating characteristic curve (AUC); (iv) feature-importance alignment, examined using Random Forest (RF) permutation importance and Nearest Neighbor Projected Distance Regression (NPDR), with agreement quantified by Top-K overlap and Jaccard similarity between feature rankings derived from real and synthetic data [10]; (v) dilution, a novel metric that quantifies the extent to which synthetic augmentation perturbs feature rankings derived from real data; and (vi) bootstrap stability of feature selection, defined as the reproducibility of Top-K features across bootstrap resamples and summarized by Jaccard similarity [8, 11].

In this study, we develop a unified and reproducible framework for evaluating synthetic gene-expression data with respect to utility, fidelity, privacy, and interpretability. We also introduce dilution, a novel metric for quantifying signal conservation under synthetic data augmentation by capturing how generated samples shift the ranking of biologically relevant features. To provide a more precise assessment of biological signal preservation, we incorporate a dual-

perspective feature-alignment analysis that integrates global and local interpretability [12]. Using an MDD gene-expression dataset, we evaluate each method under matched conditions and script all analyses for full reproducibility. Our direct comparison shows that ARF provides a more balanced trade-off among privacy, utility, and fidelity than CTGAN, with stronger predictive performance, more stable feature-importance patterns, and better preservation of key statistical properties of the real data. As a proof-of-concept extension, we include a minimal quantum-adversarial robustness experiment to illustrate that the pipeline can be adapted to emerging computational models, while leaving a more complete quantum analysis for future work [13].

2. Methods

To evaluate synthetic omics data, we use six complementary quantitative metrics: (1) downstream predictive utility; (2) gene-wise statistical fidelity, assessed using KS and Wilcoxon tests with BH-FDR control; (3) privacy, quantified by domain-separator AUC; (4) feature-importance alignment, evaluated using RF permutation importance together with NPDR overlap and Jaccard similarity; (5) dilution of real feature rankings under synthetic data augmentation; and (6) bootstrap stability of Top-K features.

2.1 Utility.

Utility assesses whether synthetic data preserve the feature-label relationships needed for downstream prediction. To evaluate utility, we train a classifier, such as RF, on either real or synthetic data and then assess its performance on a held-out real test set. Synthetic data are considered to have high utility when models trained on them achieve performance comparable to that obtained using real data alone. Utility can be summarized using accuracy, F1-score, or related predictive metrics, consistent with prior synthetic-data evaluation frameworks [6-8]. In

bulk RNA-Seq settings, synthetic augmentation with deep generative models has also been reported to improve downstream classification performance [5].

2.2 Fidelity.

Fidelity assesses how well synthetic data preserve the statistical characteristics and multivariate structure of the original dataset. High-fidelity synthetic data maintain similar marginal distributions, variances, and correlation patterns, thereby supporting statistical comparability with real observations. This metric reflects both feature-level similarity and the preservation of higher-order gene dependencies relevant to downstream biological analyses. In this study, fidelity is evaluated using per-gene nonparametric tests, specifically KS and Wilcoxon rank-sum tests, with BH-FDR correction applied to account for multiple testing [14]. Together, these tests quantify how well a generative model reproduces the structure of real omics data without overfitting to individual samples [7].

2.3 Privacy.

Privacy quantifies how distinguishable synthetic samples are from real data. In this framework, privacy is assessed using a domain-separator test in which a supervised classifier, typically RF, is trained to distinguish real from synthetic instances. Discriminative performance is summarized by AUC: values close to 0.5 indicate performance near random guessing and therefore stronger privacy preservation, whereas values farther from 0.5 indicate greater separability between real and synthetic data. Because AUC is symmetric under label inversion, we additionally report a normalized measure $AUC_{\text{norm}} = \max(AUC, 1 - AUC)$, so that interpretation remains consistent across experiments. Under this convention, values closer to 0.5 indicate stronger privacy, while values closer to 1 indicate weaker privacy. This provides a quantitative, model-agnostic proxy for privacy risk by linking privacy directly to the

statistical separability of real and synthetic distributions. Importantly, the measure captures both explicit leakage, in which real samples are reproduced, and implicit leakage, in which latent feature dependencies reveal sensitive structure [8].

2.4 Feature Importance Alignment.

Feature-importance alignment assesses how consistently key predictive variables, such as genes or biomarkers, are identified in real and synthetic datasets. In the context of privacy-preserving data synthesis, this metric evaluates whether synthetic data preserve the discriminative structure underlying classification in the real data. To quantify alignment, we use two complementary approaches: (i) RF permutation importance, which estimates each gene's contribution to classification performance by measuring the increase in prediction error after permuting that gene's values [7]; and (ii) NPDR, a neighborhood-based method that tests whether differences in feature values are associated with phenotype differences, providing a robust nonparametric measure of feature relevance in high-dimensional omics data. Agreement between real and synthetic importance profiles is summarized using Top-K overlap and the Jaccard index computed over the Top-K feature sets. Greater overlap indicates better preservation of biological signal and discriminative structure, whereas lower overlap suggests distortion or sensitivity to noise. By combining global model-based interpretability from RF with local distance-based relevance from NPDR, this analysis provides a more complete view of how faithfully each generative model preserves biologically meaningful transcriptomic features [7, 8, 10, 11].

2.5 Dilution

Dilution is a novel stability metric that quantifies how the addition of synthetic data perturbs feature-importance rankings derived from real data. Specifically, an RF model is first trained

on the real dataset alone and then retrained on the combined dataset containing both real and synthetic samples. For each gene i , the dilution score is defined as the change in its importance rank between the two models: $D(i) = rank_{real+synthetic}(i) - rank_{real}(i)$. Positive values indicate a loss of importance, meaning that the feature moves to a lower rank, whereas negative values indicate amplification, meaning that the feature shifts to a higher rank. Overall dilution is summarized using the mean and median rank shift, together with the percentage of real Top-K genes that drop out of the Top-K list after augmentation. Building on prior work in feature reproducibility and model perturbation analysis[7, 8], and provides a transparent way to evaluate whether synthetic samples preserve the feature hierarchy present in real biomedical data.

2.6 Stability.

Stability assesses the reproducibility of feature selection across repeated resampling of the data. It quantifies how consistently a model identifies the same top-ranked features under small perturbations of the training set, providing an empirical measure of robustness. In this framework, bootstrap resampling is applied separately to the real and synthetic datasets. For each bootstrap replicate, we extract the Top-K feature set and compute the Jaccard index across pairs of replicates. Higher mean Jaccard similarity indicates that the model consistently selects similar predictive features, reflecting a stable and generalizable signal, whereas lower values suggest instability or overfitting to sample-specific noise. This measure complements fidelity and feature-importance alignment by emphasizing internal reproducibility rather than similarity between datasets. Bootstrap-based stability analysis is well established in the machine learning and computational genomics literature as a criterion for reliable feature selection and biological interpretability [8, 15].

2.7 Quantum-Adversarial Evaluation

As a proof-of-concept quantum extension, we performed a focused quantum-adversarial robustness evaluation on the real gene-expression dataset. We identified the label column, removed zero-variance features, and generated a stratified 75/25 train-test split. To prevent data leakage, features were standardized using a StandardScaler fit on the training set and then applied to the test set, while labels were encoded consistently for samples from the healthy control (HC) and MDD groups. We evaluated three classifiers under this pipeline: (i) an RF classifier with 1,000 trees, with the number of candidate features at each split set to the square root of the total number of input features; (ii) a ReLU-based multilayer perceptron (MLP) trained with cross-entropy loss; and (iii) a 6-qubit re-uploading parameterized quantum circuit (PQC). Because the input space was high-dimensional (1783 genes), the PQC was driven by a train-set-only linear projection from 1783 features to six latent inputs, rather than six individual genes, before quantum embedding. A small linear classification head then produced the output logits. Adversarial vulnerability was assessed using fast gradient sign method (FGSM), a one-step gradient-sign attack constructed on the MLP surrogate, where ε denotes the L_∞ perturbation budget applied to standardized inputs. We evaluated $\varepsilon \in \{0.00, 0.03, 0.06, 0.10, 0.20\}$. These adversarial examples were then tested on the PQC in a black-box transfer setting, meaning that attacks were optimized on the classical surrogate and transferred to the quantum model without using PQC gradients [13, 16].

2.8 Data

We evaluated the framework on a publicly available peripheral blood RNA-Seq gene-expression dataset for MDD [17]. The dataset includes transcriptomic profiles from 64 individuals with MDD and 64 matched healthy controls, enabling supervised classification and biomarker discovery. Gene-expression measurements were obtained using high-

throughput sequencing and preprocessed to remove low-variance features and standardize expression levels across samples.

After preprocessing, the dataset contained 1783 gene-expression features, reflecting the small-n, large-p setting commonly encountered in transcriptomic studies. This setting provides a stringent test for synthetic data generation methods, because preserving both marginal distributions and higher-order feature relationships is essential for maintaining biological signal. Prior analyses of this dataset have identified differentially expressed genes associated with MDD, including candidates linked to neural development and synaptic function, supporting its suitability for studying the preservation of biologically meaningful signal under synthetic data generation [17].

3. Results

3.1. Utility of synthetic data

Across 100 out-of-bag (OOB) runs, augmenting the real data with ARF-generated synthetic samples resulted in a slight improvement in predictive utility, whereas augmentation with CTGAN-generated samples led to a substantial decline (Fig.1). These results suggest that ARF more effectively preserves diagnostically relevant structure during augmentation, while CTGAN introduces distributional shifts that impair downstream predictive performance.

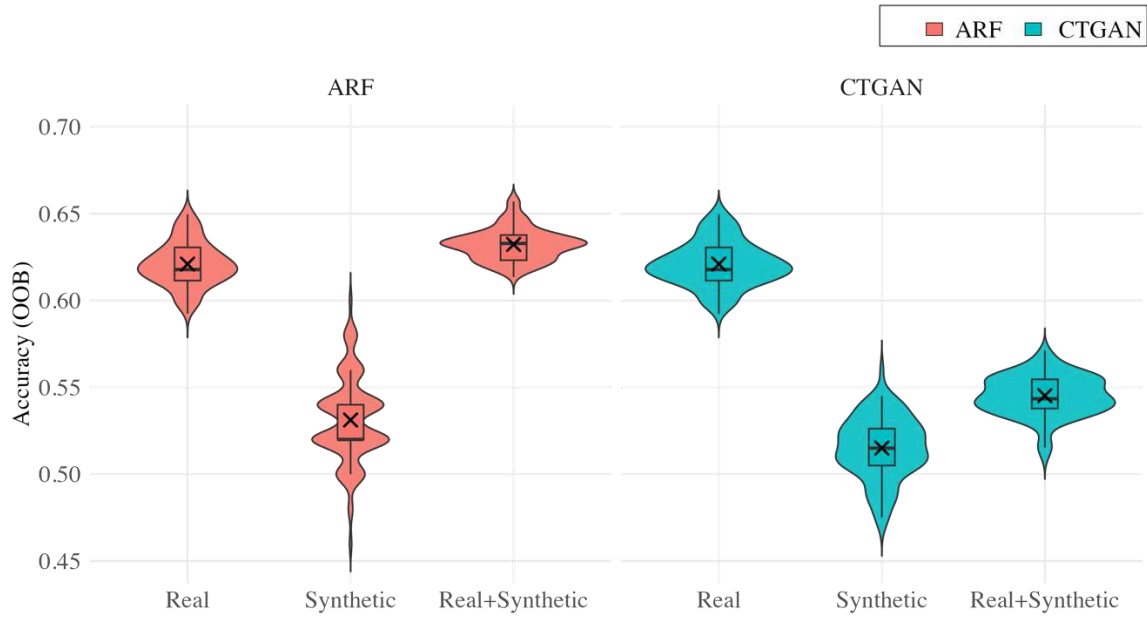


Fig.1. Distribution of RF OOB accuracy across 100 runs for real, synthetic-only, and real plus synthetic training conditions. The left panel shows ARF and the right panel shows CTGAN. Higher values indicate greater predictive utility.

3.2. Low-dimensional structure (UMAP)

Although UMAP is a stochastic embedding, we use it here as a qualitative visualization of low-dimensional structure and class separation rather than as a formal measure of fidelity (Fig.2). In the overlaid views, ARF-generated samples largely co-locate with the real-data embedding and preserve the overall HC/MDD organization (Fig.2a), whereas CTGAN-generated samples show a less coherent correspondence with the real-data manifold (Fig.2c). When viewed in isolation, the ARF-generated embedding remains broadly compatible with the geometry observed in the real data (Fig.2b), while the CTGAN-generated embedding shows a more shifted and diffuse configuration relative to the real-data layout (Fig.2d). Overall, these visual patterns are consistent with ARF generating synthetic data that more closely preserve the low-dimensional transcriptomic structure of the original dataset.

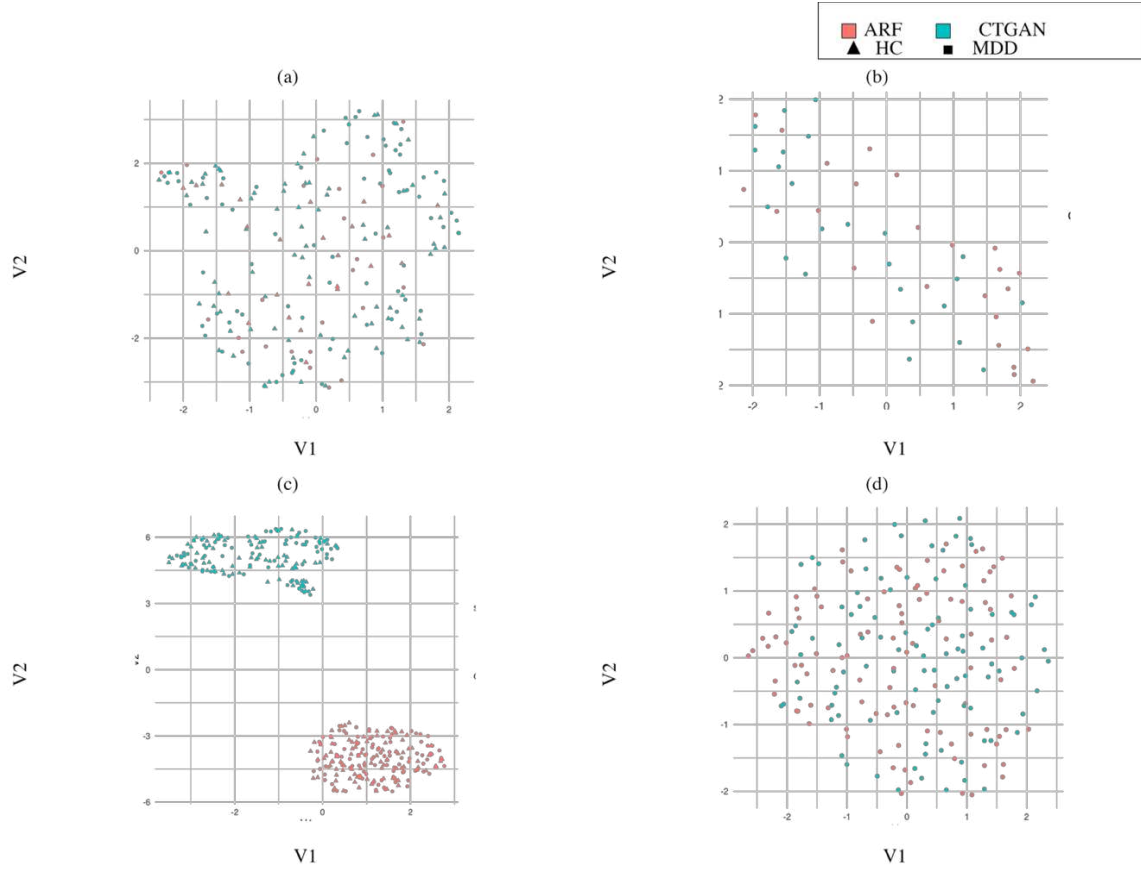


Fig.2. UMAP visualization of real and synthetic transcriptomic data. (a) Real + ARF, (b) ARF only, (c) Real + CTGAN, and (d) CTGAN only.

3.3.Privacy (domain-separator test)

We evaluated privacy preservation by training a domain-separator classifier to distinguish real from synthetic samples and summarizing separability using AUC under 5×5 cross-validation (Fig.3). Values closer to 0.5 indicate stronger privacy, reflecting greater difficulty in discriminating real from synthetic data, whereas values closer to 1 indicate poor privacy preservation. Because AUC is symmetric under label inversion, we also report a normalized value, $AUC_{\text{norm}} = \max(AUC, 1 - AUC)$. ARF yielded a raw AUC of 0.409, corresponding to $AUC_{\text{norm}} = 0.591$, suggesting that real and ARF-generated samples were relatively difficult to distinguish. In contrast, CTGAN was effectively perfectly separable from the real data ($AUC \approx 0$; $AUC_{\text{norm}} \approx 1$) indicating substantial real-synthetic distinguishability and poor privacy preservation.

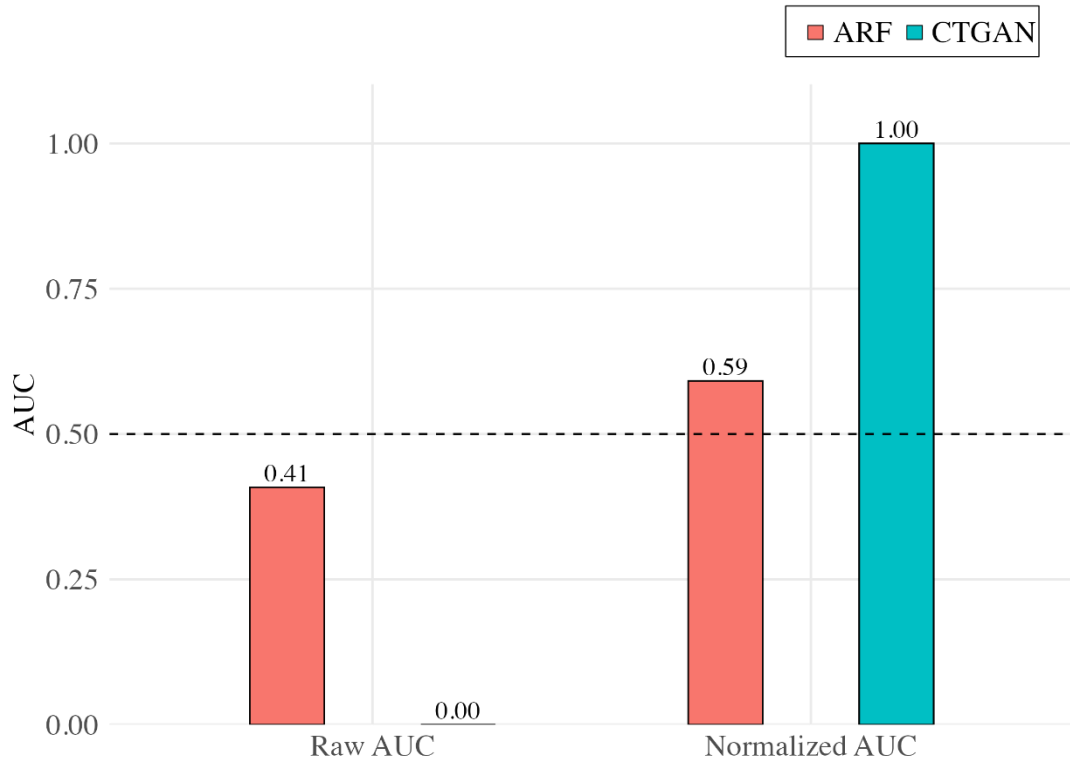


Fig.3. Domain-separator AUC under 5×5 cross-validation, shown as raw and normalized values for ARF and CTGAN. Values closer to 0.5 indicate lower separability and better privacy preservation.

3.4. Fidelity

Fidelity was assessed using per-gene KS and Wilcoxon tests, with and without BH-FDR correction (Table 1). ARF showed minimal marginal distortion, with only 1.6% of genes significant before correction and 0% significant after BH-FDR correction in both tests, indicating close agreement with the real gene-wise distributions. In contrast, CTGAN showed substantial distributional shift, with 94.4% of genes remaining significant by KS and 88.1% by Wilcoxon after BH-FDR correction, suggesting widespread divergence from the real marginal gene-expression profiles.

Table 1. Percentage of genes identified as significantly different between real and synthetic data by K-S and Wilcoxon tests, before and after BH-FDR correction.

Model	KS raw (%)	Wilcoxon raw (%)	KS BH-FDR (%)	Wilcoxon BH-FDR (%)
ARF	1.6	1.6	0	0

CTGAN	94.6	88.3	94.4	88.1
-------	------	------	------	------

273

274 3.5. Feature importance alignment

275 Feature-importance alignment was examined using the Top-20 genes ranked by Random
276 Forest permutation importance (Fig.4) .Panels (a) and (b) present the real-data and ARF-based
277 rankings, respectively, whereas panels (c) and (d) present the real-data and CTGAN-based
278 rankings. Across both generators, the ranked gene lists show only partial agreement, indicating
279 that the most discriminative genes identified in the real data were not fully reproduced in the
280 synthetic data. Nevertheless, the ARF-based comparison shows somewhat greater overlap in
281 high-ranking genes than the CTGAN-based comparison. By contrast, the CTGAN panels
282 display a more substantial shift in both gene identity and ranking order. Taken together, these
283 results suggest that ARF preserves feature-importance structure more effectively than CTGAN,
284 although overall agreement remains modest.

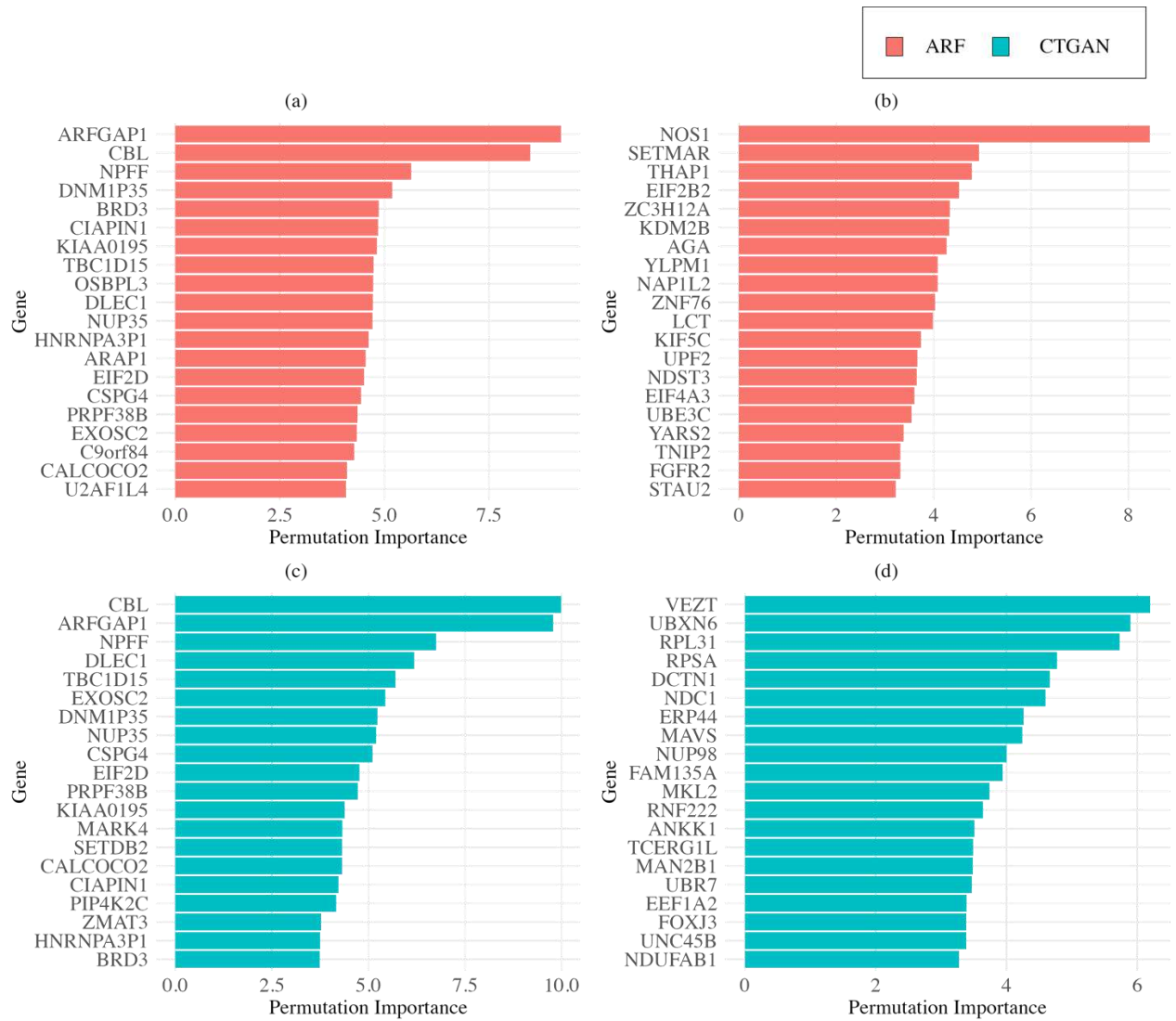


Fig.4. Top-20 genes ranked by Random Forest permutation importance. Panels (a) and (b) correspond to the real-data and ARF-synthetic rankings, respectively, whereas panels (c) and (d) correspond to the real-data and CTGAN-synthetic rankings, respectively. Greater similarity between paired panels indicates better preservation of discriminative signal.

To further evaluate feature-level agreement, NPDR-based Top-K overlaps were examined across dataset combinations using Jaccard similarity (Fig.5). For both generators, the strongest agreement was observed between the real data and the corresponding real+synthetic dataset, indicating that augmentation largely preserved the features identified in the real data alone. By contrast, direct agreement between the real and synthetic datasets was lower, while the synthetic versus real+synthetic comparison showed intermediate similarity. Overall, the ARF-based comparisons exhibited slightly stronger agreement than the CTGAN-based comparisons, in line with the Random Forest feature-importance patterns shown in (Fig.4).

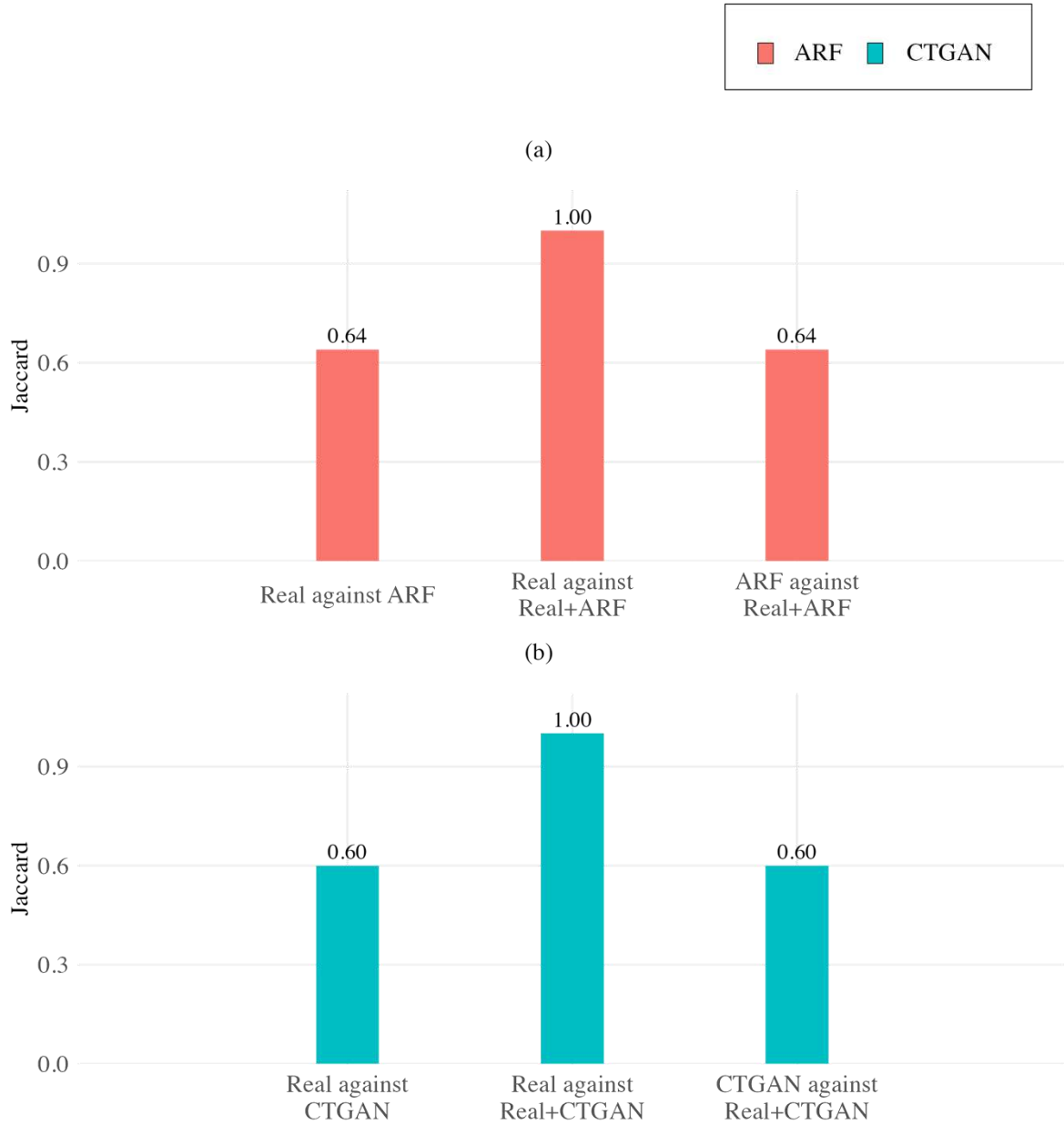


Fig.5. NPDR Top-K feature overlap ($K = 50$) across dataset combinations, summarized by Jaccard similarity. Panel (a) shows the ARF-based comparisons, and panel (b) shows the CTGAN-based comparisons. The bars represent real against synthetic, real against real+synthetic, and synthetic against real+synthetic comparisons. Higher Jaccard values indicate stronger agreement in the selected feature sets.

3.6. Dilution of real feature ranks after synthetic augmentation

We assessed dilution by examining how the ranks of real Top-K features changed after synthetic augmentation and by measuring the proportion of real Top-K genes that were no longer retained in the augmented setting (Fig.6, Table 2). Mean rank shift and median rank shift summarize how strongly the original ranking was perturbed, whereas Top-K dropout (%)

indicates the percentage of real Top-K genes that fell below the Top-K threshold after synthetic data were added. Compared with ARF, CTGAN caused a larger disruption of the original feature hierarchy. ARF provided a mean rank shift of 417.7, a median rank shift of 110.5, and a Top-K dropout of 70%, whereas CTGAN provided corresponding values of 523.5, 375, and 82%. Taken together, these results suggest that CTGAN alters the ordering of discriminative genes more strongly than ARF after augmentation.

Table 2. Dilution summary for real Top-K ($K = 50$) features after synthetic augmentation.

Model	Mean rank shift	Median rank shift	Top-K dropout (%)
ARF	417.7	110.5	70
CTGAN	523.5	375	82

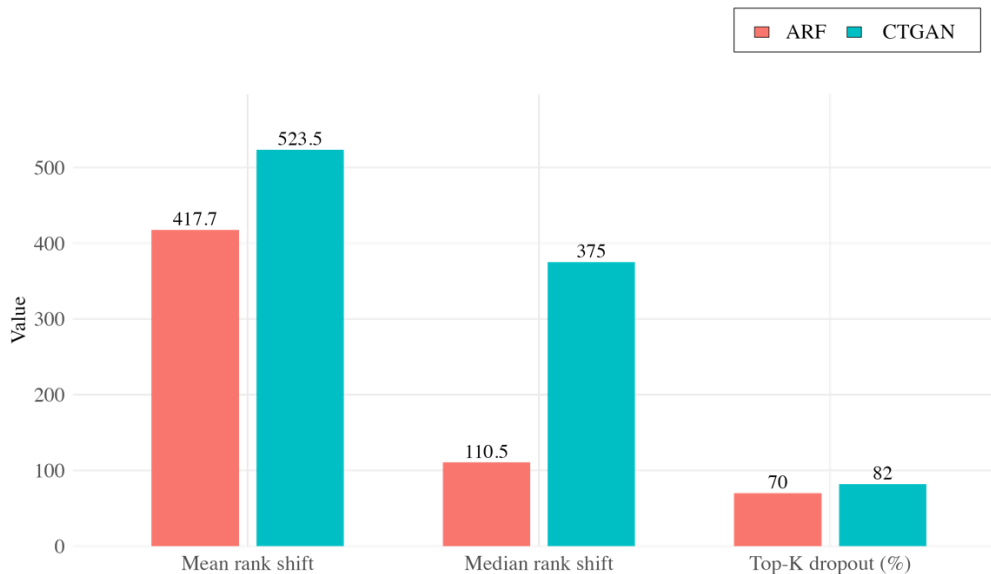


Fig.6. Dilution summary for real Top-K ($K = 50$) features after synthetic augmentation. Bars report mean rank shift, median rank shift, and Top-K dropout (%) for ARF and CTGAN. Higher values indicate greater disruption of the original real-data feature hierarchy.

3.7. Stability of feature selection

Feature-selection stability was evaluated by bootstrap resampling and summarized as the Jaccard overlap of Top-K features across resamples (Fig.7). Stability within each dataset was modest, with the real data showing slightly higher self-consistency than the corresponding synthetic datasets. By contrast, agreement between the real and synthetic settings was

substantially lower for both generators, indicating that synthetic data did not reproduce the same feature-selection priorities observed in the real data. ARF showed somewhat greater agreement with the real data than CTGAN, although both real-against-synthetic comparisons remained low overall. Taken together, these findings suggest that synthetic augmentation introduces additional variability in feature ranking, with ARF preserving the original feature-selection structure somewhat better than CTGAN.

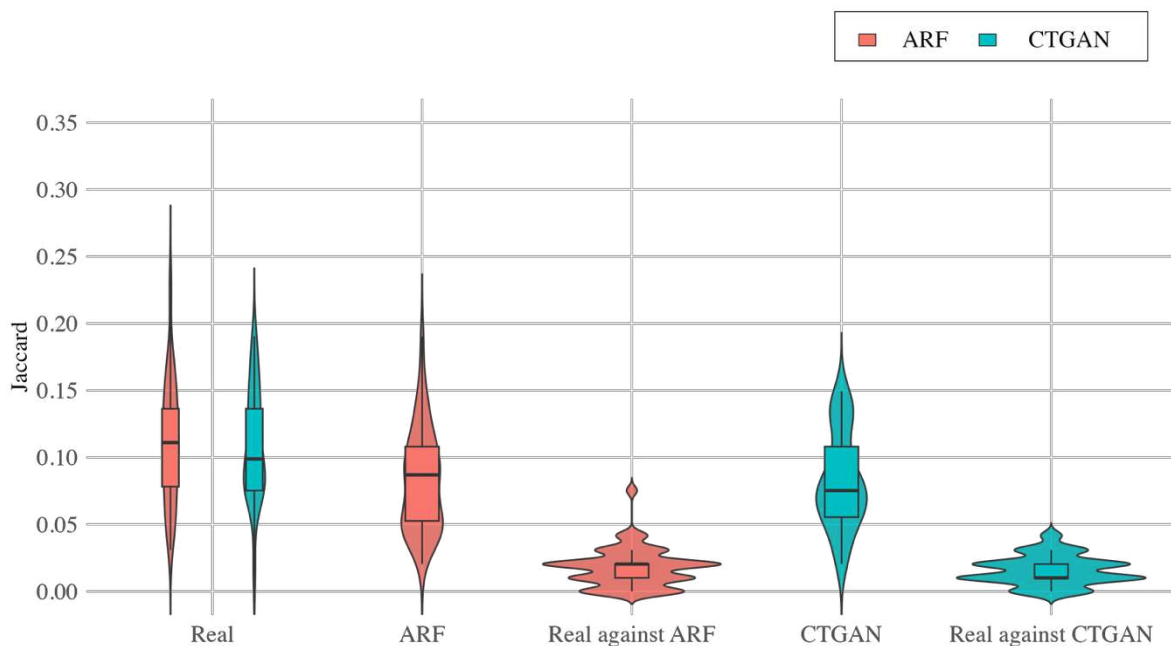


Fig.7. Bootstrap stability of Top-K ($K = 50$) feature selection, summarized by Jaccard overlap across resamples. Real, ARF, and CTGAN represent within-dataset stability, whereas Real against ARF and Real against CTGAN represent cross-dataset agreement. Higher Jaccard values indicate greater reproducibility of the selected feature sets.

3.8. Quantum-Adversarial Evaluation

We conducted a minimal quantum extension on the held-out real gene-expression test split to assess feasibility under the same standardized preprocessing used for the classical baselines. Here, *clean* refers to standardized test inputs with no adversarial perturbation applied ($\epsilon = 0$). On the clean test set, the RF and MLP models each achieved an accuracy of 0.575, whereas the 6-qubit PQC achieved 0.550, indicating broadly comparable performance despite the low-dimensional projection required for quantum embedding (Table 3).

Table 3. Test accuracies on the held-out real dataset under clean conditions ($\varepsilon = 0$). *Clean* denotes standardized test inputs with no adversarial perturbation applied.

RF	MLP	PQC
0.575	0.575	0.55

To examine robustness under transferred attacks, adversarial examples were generated using the FGSM on the MLP surrogate, and the same perturbed inputs were then evaluated on both the attacked MLP and the PQC. In this setting, ε denotes the perturbation magnitude applied to the standardized inputs. As ε increased, the accuracy of both models declined (Fig.8). The PQC retained slightly higher accuracy at $\varepsilon = 0.03$ (0.450 against 0.350 for the attacked MLP), whereas both models showed substantial degradation at larger perturbation magnitudes and approached very low accuracy by $\varepsilon = 0.20$. Overall, these results suggest modest transfer robustness of the PQC at small perturbation levels, although both models became highly vulnerable as attack strength increased.

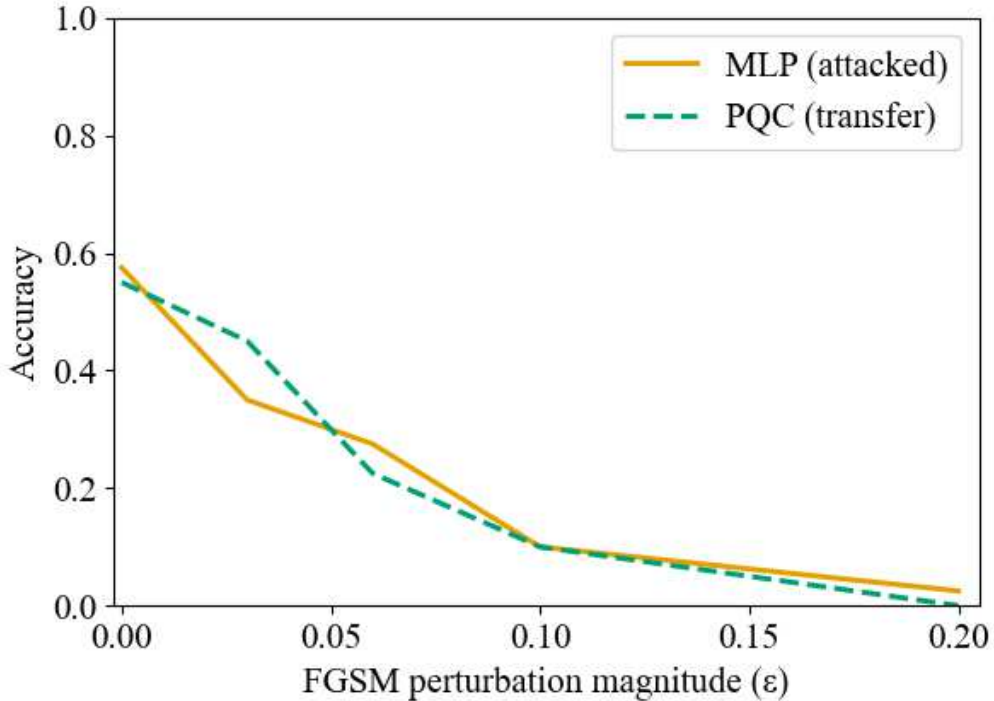


Fig.8. Adversarial accuracy on the held-out real dataset as a function of FGSM perturbation magnitude (ε). Adversarial examples were crafted on the MLP surrogate and evaluated on both the attacked MLP and the PQC under transfer. Accuracy declined for both models as perturbation magnitude increased, although the PQC retained slightly higher accuracy at $\varepsilon = 0.03$.

4. Conclusions

We presented a unified, reproducible, multi-metric framework for evaluating synthetic gene-expression generation in MDD. By jointly assessing downstream utility, statistical fidelity, privacy, feature-importance alignment, dilution, and bootstrap stability, this framework provides a transparent way to characterize privacy-utility-fidelity trade-offs in high-dimensional omics settings, where limited cohort size and data-sharing constraints often restrict robust model development and reproducibility.

Across the evaluation criteria considered here, ARF showed the most favorable overall balance relative to CTGAN. ARF better preserved downstream predictive performance under augmentation, remained more consistent with gene-wise distributional tests, and exhibited lower real-synthetic separability in the privacy analysis. In contrast, CTGAN introduced larger distributional shifts, reduced predictive utility, and more strongly disrupted feature-importance rankings.

A central contribution of this study is the dilution metric, which quantifies how synthetic augmentation alters the ranking of real discriminative features. In our experiments, CTGAN produced larger rank shifts and higher dropout among real Top-K features than ARF, indicating stronger distortion of the original feature hierarchy. The feature-alignment analyses supported the same pattern: although overlap between real and synthetic rankings remained limited for both generators, ARF showed more coherent agreement than CTGAN. The modest bootstrap stability observed across datasets was also consistent with the small-sample, high-dimensional setting, where multiple feature subsets may yield similar predictive performance.

From a practical perspective, these findings suggest that tree-based generative approaches such as ARF may be better suited to small-sample transcriptomic applications in which fidelity and

interpretability are important priorities. By contrast, GAN-based methods such as CTGAN may require larger cohorts or additional regularization to mitigate distributional drift. As a feasibility extension, we also included a minimal quantum-adversarial robustness experiment, showing that lightweight quantum models can be incorporated into the broader evaluation pipeline, while more comprehensive quantum analysis remains an important direction for future work.

Overall, feature alignment, dilution, and stability provide a complementary perspective on interpretability preservation. Whereas conventional metrics such as utility and fidelity assess predictive and statistical agreement, these additional metrics more directly evaluate whether synthetic data preserve the structure of biologically meaningful signals. In particular, dilution offers a practical and interpretable diagnostic of how strongly synthetic augmentation perturbs the ranking of discriminative genes. Taken together, this study provides a reproducible and extensible framework for evaluating synthetic biomedical data and highlights key considerations for balancing privacy, utility, biological validity, and interpretability in sensitive omics applications.

Data availability

All preprocessing, model training, and evaluation procedures were implemented in a fully reproducible pipeline. The code used to generate synthetic data, compute the evaluation metrics, including dilution, and reproduce all figures is publicly available at: <https://github.com/setarehakbari/synthetic-omics-arf-ctgan>. The real gene-expression dataset used in this study is not publicly available because of data-use restrictions. However, the repository includes scripts to reproduce all analyses for researchers with authorized access to the data.

404 **Declarations**

405 **Abbreviations**

406 Not applicable

407 **Ethics approval and consent to participate**

408 Not applicable

409 **Consent for publication**

410 Not applicable

411 **Availability of data and materials**

412 The datasets analyzed during the current study are openly available in the GitHub repository
413 at <https://github.com/insilico/GeneAgeMDD>. Further details regarding the original datasets and
414 their sources are described in [17].

415 **Competing interests**

416 The authors declare no competing interests.

417 **Funding**

418 Not applicable

419 **Author contributions**

420 S.A. conceptualized the study, developed the methodology, performed the software
421 implementation and validation, and wrote the original draft of the manuscript. Y.J.L.

contributed to the software, conducted the formal analysis, and reviewed and edited the manuscript. B.M. contributed to the methodology, curated the data, acquired funding, supervised the study, and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

References

1. Chen, Z.S., et al., *Modern views of machine learning for precision psychiatry*. Patterns, 2022. **3**(11).
2. Hicks, E.M., et al., *Integrating genetics and transcriptomics to study major depressive disorder: a conceptual framework, bioinformatic approaches, and recent findings*. Translational psychiatry, 2023. **13**(1): p. 129.
3. Erlich, Y. and A. Narayanan, *Routes for breaching and protecting genetic privacy*. Nature Reviews Genetics, 2014. **15**(6): p. 409-421.
4. Feldner-Busztin, D., et al., *Dealing with dimensionality: the application of machine learning to multi-omics data*. Bioinformatics, 2023. **39**(2): p. btad021.
5. Wang, Y., et al., *Generating bulk RNA-Seq gene expression data based on generative deep learning models and utilizing it for data augmentation*. Computers in Biology and Medicine, 2024. **169**: p. 107828.
6. Xu, L. and K. Veeramachaneni, *Synthesizing tabular data using generative adversarial networks*. arXiv preprint arXiv:1811.11264, 2018.
7. Breiman, L., *Random forests*. Machine learning, 2001. **45**(1): p. 5-32.

- 444 8. Adams, T., et al., *On the fidelity versus privacy and utility trade-off of synthetic patient*
445 *data*. Iscience, 2025. **28**(5).
- 446 9. Le, T.T., et al., *Differential privacy-based evaporative cooling feature selection and*
447 *classification with relief-F and random forests*. Bioinformatics, 2017. **33**(18): p. 2906-
448 2913.
- 449 10. Le, T.T., B.A. Dawkins, and B.A. McKinney, *Nearest-neighbor Projected-Distance*
450 *Regression (NPDR) for detecting network interactions with adjustments for multiple*
451 *tests and confounding*. Bioinformatics, 2020. **36**(9): p. 2770-2777.
- 452 11. Hernadez, M., et al., *Synthetic tabular data evaluation in the health domain covering*
453 *resemblance, utility, and privacy dimensions*. Methods of information in medicine,
454 2023. **62**(S 01): p. e19-e38.
- 455 12. Clough, L.A., et al., *Local-NPDR: a novel variable importance method for explainable*
456 *machine learning and false discovery diagnosis for ocean worlds biosignatures*.
457 *Frontiers in Astronomy and Space Sciences*, 2025. **12**: p. 1651953.
- 458 13. Wendlinger, M., K. Tschärke, and P. Debus. *A comparative analysis of adversarial*
459 *robustness for quantum and classical machine learning models*. in *2024 IEEE*
460 *International Conference on Quantum Computing and Engineering (QCE)*. 2024.
461 IEEE.
- 462 14. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and*
463 *powerful approach to multiple testing*. Journal of the Royal statistical society: series B
464 (Methodological), 1995. **57**(1): p. 289-300.
- 465 15. Meinshausen, N. and P. Bühlmann, *Stability selection*. Journal of the Royal Statistical
466 Society Series B: Statistical Methodology, 2010. **72**(4): p. 417-473.
- 467 16. Goodfellow, I.J., J. Shlens, and C. Szegedy, *Explaining and harnessing adversarial*
468 *examples*. arXiv preprint arXiv:1412.6572, 2014.

469 17. Li, Y.J., et al., *Differential expression of MDGA1 in major depressive disorder*. Brain,
470 Behavior, & Immunity-Health, 2022. **26**: p. 100534.
471