Synthetic Data Generation Methods and Robustness in Tabular Foundation Models under Adversarial Perturbations

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the choice of synthetic data generation method (e.g., GANs, VAEs, diffusion models) influence the robustness of tabular foundation models on the TabMNAR benchmark when evaluated under. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LapDDPM: A Conditional Graph Diffusion Model for scRNA-seq Generation with Spectral Adversarial Perturbations. Research question: How does the choice of synthetic data generation method (e.g., GANs, VAEs, diffusion models) influence the robustness of tabular foundation models on the TabMNAR benchmark when evaluated under adversarial perturbations?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

14 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.8/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Early approaches for generating synthetic cellular profiles often adapted models from general machine learning, such as	×	0.07
VAE-based models learn a low-dimensional latent representation and reconstruct gene expression, often accounting for spa	×	0.03
GANs aim to learn a mapping from a simple prior distribution to the complex data distribution through an adversarial tra	×	0.11
Flow-based models have been explored for their exact likelihood estimation and invertible mappings.	×	0.04
GNNs have been applied to various tasks in single-cell biology, including cell type annotation, trajectory inference, an	×	0.06
Diffusion Probabilistic Models (DPMs) have emerged as a powerful class of generative models, demonstrating state-of-the-	×	0.07
The overall training procedure combines diffusion, reconstruction, and KL divergence losses, with the encoder being trai	×	0.08
Given a scRNA-seq dataset consisting of N cells and D genes, represented as a count matrix X $\in$RN$\times$D, we first preprocess	×	0.08
Prior to graph construction, genes expressed in fewer than a specified threshold of cells are filtered out to reduce spa	×	0.02
The raw count data is then normalized and log-transformed for stable numerical operations during feature extraction.	×	0.04
Laplacian Positional Encoding (LPE) is used to capture biologically meaningful relationships and reduce the dimensionali	×	0.08
A k-NN graph is constructed on the cells using Euclidean distance in the PCA-reduced space.	×	0.02
For each cell (node), its k nearest neighbors are identified, and edges are formed between them.	×	0.02

References

- http://arxiv.org/abs/2502.17119v2
- http://arxiv.org/abs/2512.03307v1
- http://arxiv.org/abs/2506.13344v1