Poster: Genomic Data Generation via Correlation-Guided and Privacy-Preserving Diffusion

You Li, Shuainan Liu, Shaojie Zhan, Zhongshuo Fang, and Tianxi Ji
Department of Computer Science, Texas Tech University
{You.Li, Shuainan.Liu, shzhan, zhonfang, tiji}@ttu.edu

***Abstract*—Access to human genomic data is critically limited by privacy regulations. In this work, we propose a privacy-preserving framework based on One-Dimensional Diffusion Transformer in the latent space. Our design incorporates a correlation-guided decoding strategy to enforce biological fidelity and a timestep-aware two-stage differential privacy mechanism to bound privacy risks without significantly compromising data utility. Preliminary evaluations support the feasibility of our proposed framework, demonstrating that the generated data can be effectively utilized for downstream tasks related to Genome-Wide Association Studies (GWAS).**

I. Introduction

Access to human genomic data is governed by stringent legal and ethical restrictions. Consequently, these factors pose significant challenges to the processing, analysis, and sharing of human genomic information. Researchers have employed generative artificial intelligence to produce synthetic genomic data [1]. By learning the underlying statistical distributions of real human genomes, this approach generates synthetic genomic data, thereby facilitating broad research applications while strictly safeguarding patient privacy. Among various forms of genomic variation, Single Nucleotide Polymorphism (SNP) is the most fundamental type. These variants serve as critical genetic markers for understanding the hereditary basis of phenotypic differences and disease susceptibility.

Researchers have leveraged machine learning models to effectively produce genomic data [2]. These synthetic SNP datasets have proven capable of performing downstream tasks, including population structure inference and genotype imputation. However, complex downstream tasks, such as Genome-Wide Association Studies (GWAS), have yet to be effectively addressed. This is primarily because accurately preserving the subtle yet critical statistical association signals between specific variant loci and traits during the generation process remains a significant challenge. Furthermore, Diffusion Models pose critical privacy risks when handling high-dimensional SNP data characterized by extreme uniqueness, as they are vulnerable to membership inference attacks and inverse reconstruction risks. Therefore, generating high-fidelity and privacy-preserving synthetic SNP datasets is of paramount importance.

We propose a privacy-preserving framework based on diffusion models. The architecture incorporates a design for a correlation-guided decoding strategy to guide SNP reconstruction. Furthermore, we outline the employment of a two-stage differential privacy mechanism to optimize the privacy-utility trade-off. This approach is conceptualized to enable the generation of high-fidelity genomic data suitable for complex tasks like population structure inference and GWAS under strict privacy guarantees.

II. Method

As illustrated in Fig. 1, this study adopts a 1D-DiT [3] architecture using a VAE to map discrete SNPs into a continuous latent space. This encoding reduces computational costs and aligns discrete genomic features with the Gaussian noise priors of diffusion models. To precisely replicate the complex biological correlations between SNP loci, we propose a correlation-guided decoding strategy. Specifically, the correlation matrix of the original data is calculated and subsequently employed as an explicit constraint during the decoding phase to guide SNP reconstruction. This mechanism ensures that the generated genomic data strictly adheres to authentic genetic structures, thereby significantly enhancing its utility for downstream tasks.

To address the privacy concerns associated with diffusion models, this paper leverages a timestep-aware two-stage differentially private training strategy [4]. In the first stage, the inherent noise injection in the diffusion process serves as a natural privacy barrier. In the second stage, DP-SGD is applied within the continuous latent space. This compressed representation enhances gradient norm controllability compared to sparse raw data, enabling optimized noise injection and a superior privacy-utility trade-off. This approach effectively safeguards data privacy under rigorous mathematical privacy budget constraints while mitigating the over-smoothing issue caused by global noise injection. Consequently, it maximizes the preservation of subtle signals essential for population structure inference and GWAS.

III. Evaluation

We present the preliminary evaluations on the 1000 Genomes Project dataset [5], including population structure inference and GWAS.

For population structure inference, PCA was applied for dimensionality reduction, followed by the deployment of three distinct machine learning models to classify the samples. Table I presents the inference results using original and

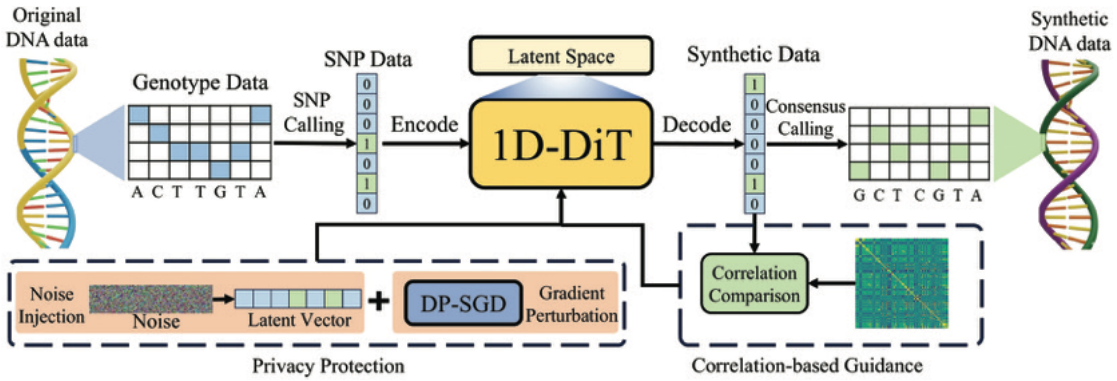


Fig. 1. Methodological Overview

TABLE I
POPULATION STRUCTURE INFERENCE

Model	Original data	Synthetic data
Logistic Regression	0.97	0.92
Support Vector Machine	0.96	0.92
Random Forest	0.92	0.84
Average	0.95	0.89

synthetic data. The generated SNP data effectively performs the population structure inference task, achieving accuracy comparable to the original data. In terms of GWAS analysis, we selected two distinct populations as the case and control groups, respectively. A dominant inheritance model and Z-test were employed to identify SNP loci significantly associated with the target phenotype, thereby evaluating whether the synthetic datasets preserve critical biological signals. We focused on SNPs with the highest statistical significance, specifically those within the top 10%, 20%, and 30% of the lowest p-values. We then calculated the overlap of significant SNPs between the original training and test sets, as well as the concordance between the original training set and both the vanilla synthetic data and the correlation-guided synthetic data. These results are summarized in Fig. 2. As illustrated, the synthetic data generated with correlation-prior guidance yields GWAS results that more closely align with the original datasets, demonstrating superior biological fidelity.

Initial results demonstrate that our proposed method successfully synthesizes biologically functional SNP data suitable for downstream tasks. These findings establish a solid foundation for the forthcoming integration of robust privacy-preserving mechanisms.

IV. CONCLUSION

This paper proposes a privacy-preserving architectural design for generating synthetic genomic data, integrating a 1D-DiT with a VAE. The framework is designed to incorporate a correlation-guided decoding strategy and a two-stage differentially private mechanism to theoretically ensure the rigorous protection of sensitive genetic information. While

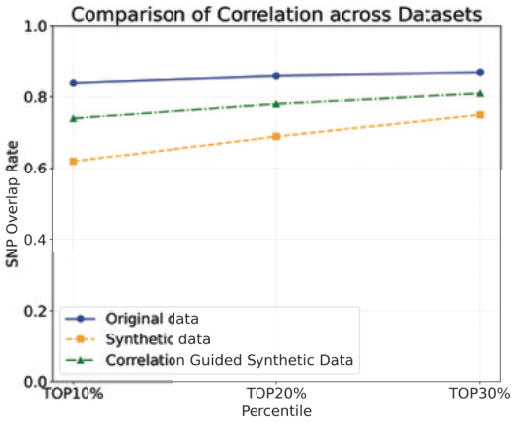


Fig. 2. Impact of Correlation Guidance on Synthetic Data Utility

preliminary results demonstrate the efficacy of the generative backbone, future work will focus on the full implementation and empirical validation of these privacy mechanisms. This framework provides a conceptual blueprint for overcoming privacy bottlenecks and paves the way for supporting critical downstream tasks such as population structure inference and GWAS.

REFERENCES

[1] S. Senan, A. J. Reddy, Z. Nussbaum, A. Wenteler, M. Bejan, M. I. Love, W. Meuleman, and L. Pinello, "Dna-diffusion: Leveraging generative models for controlling chromatin accessibility and gene expression via synthetic regulatory elements," in *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

[2] B. Yelmen, A. Decelle, L. L. Boulos, A. Szatkownik, C. Furtlehner, G. Charpiat, and F. Jay, "Deep convolutional and conditional neural networks for large-scale genomic data generation," *PLOS Computational Biology*, vol. 19, no. 10, p. e1011584, 2023.

[3] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[4] H. Wang, S. Pang, Z. Lu, Y. Rao, Y. Zhou, and M. Xue, "dp-promise: Differentially private diffusion probabilistic models for image synthesis," in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 1063–1080.

[5] L. Clarke, X. Zheng-Bradley, R. Smith, E. Kulesha, C. Xiao, I. Toneva, B. Vaughan, D. Preuss, R. Leinonen, M. Shumway *et al.*, "The 1000 genomes project: data management and community access," *Nature methods*, vol. 9, no. 5, pp. 459–462, 2012.

Genomic Data Generation via Correlation-Guided and Privacy-Preserving Diffusion

You Li†, Shuainan Liu†, Shaojie Zhan†, Zhongshuo Fang†, Tianxi Ji†

†Department of Computer Science, Texas Tech University

Motivation

Data Access Barriers: The acquisition of authentic genomic data is subject to stringent legal and ethical restrictions.

Utility Bottlenecks: Existing generative models struggle to preserve subtle yet critical statistical association signals.

Privacy Risks: Diffusion models pose severe privacy vulnerabilities regarding sensitive genetic information.

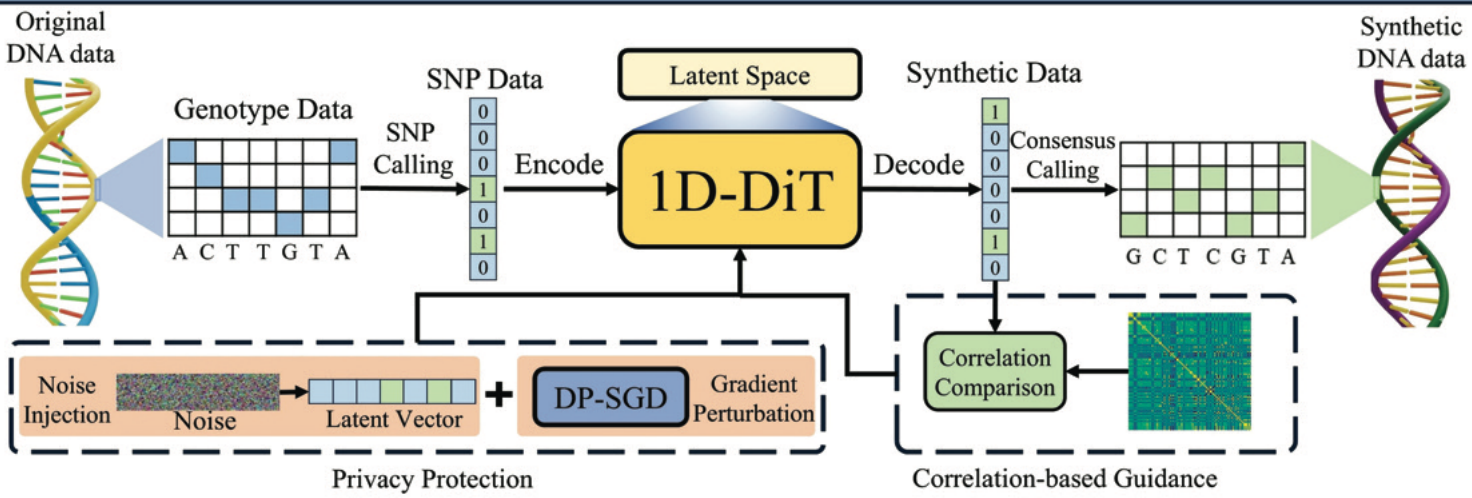
Goals

Architecture: Hybrid Generative Framework Combine 1D DiT with VAE to efficiently model high-dimensional SNP sequences.

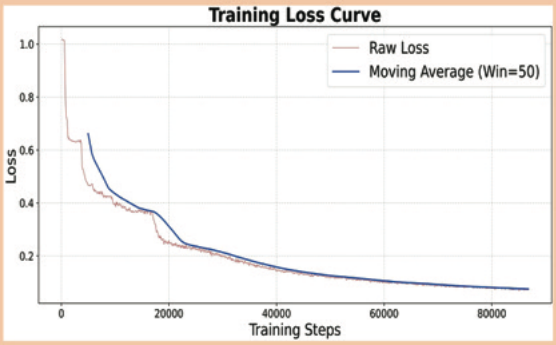
Fidelity: Utilize correlation-guided constraint during decoding to retain critical signals for GWAS.

Privacy: Proposes adopting a timestep-aware two-stage DP framework to bound privacy risks while optimizing the privacy-utility trade-off.

Design Overview



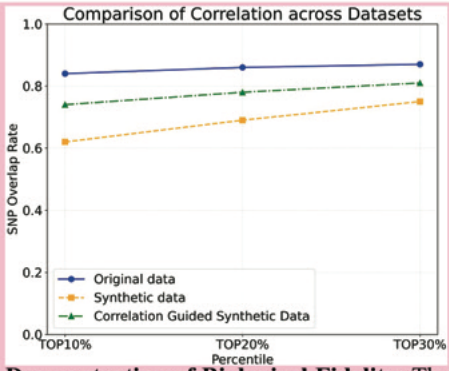
Evaluation



➢ **Training Dynamics**: The model showed robust convergence, stabilizing at 80k steps with a final loss of 0.0754. The smooth loss curve highlights high stability, indicating the model efficiently minimized the error in genomic signals.

Model	Original data	Synthetic data
Logistic Regression	0.97	0.92
Support Vector Machine	0.96	0.92
Random Forest	0.92	0.84
Average	0.95	0.89

➢ **Downstream Utility**: The data yielded an accuracy of 0.89 (near the 0.95 baseline), validating high fidelity in preserving population structures.



➢ **Demonstration of Biological Fidelity**: The significant SNP overlap observed at key p-value thresholds indicates the preservation of subtle signals. The correlation-guided synthetic data exhibits a superior alignment with real genomic data compared to the unguided baseline.

Key Results

- Privacy Architecture: Proposes adopting a timestep-aware two-stage DP strategy for strict latent-space protection.
- Model Performance: Achieved stable convergence to capture complex genomic signal distributions.
- High Utility & Biological Fidelity: Correlation-guided decoding significantly enhanced SNP overlap rates in GWAS analysis, ensuring the precise preservation of critical signals.