GenProtect-V: A Variational Inference-based Framework for Privacy-Preserving Synthetic Human Genomic Data Generation

Zihan Bian * and Linyu Mou

Posted Date: 29 December 2025

doi: 10.20944/preprints202512.2461.v1

Keywords: synthetic genomic data; privacy; data utility; variational autoencoder; privacy-preserving

 doi:10.20944/preprints202512.2461.v1

Article

GenProtect-V: A Variational Inference-based Framework for Privacy-Preserving Synthetic Human Genomic Data Generation

Zihan Bian * and Linyu Mou

Suez Canal University, Egypt
* Correspondence: ai4221099@deltauniv.edu.eg

Abstract

The generation of synthetic human genomic data offers immense potential for biomedical research and data sharing, while theoretically safeguarding individual privacy. However, existing methods, including deep generative models, struggle to achieve a robust balance between data utility and privacy protection. State-of-the-art evaluations like PRISM-G reveal vulnerabilities such as proximity, kinship replay, and trait-linked leakage. This paper introduces GenProtect-V, an end-to-end privacy-preserving synthetic human genomic data generation framework based on a Variational Autoencoder architecture. GenProtect-V integrates multi-layered privacy mechanisms: a Differentially Private Encoder to mitigate Proximity Leakage, Decoupled Latent Space Learning to address Kinship Replay, and a Rare Variant Smoother to counter Trait-linked Leakage. Through extensive experiments on the 1000 Genomes Project dataset, we demonstrate that GenProtect-V consistently achieves significantly lower PRISM-G composite scores compared to state-of-the-art baselines. Crucially, GenProtect-V simultaneously maintains or improves key utility metrics, including Allele Frequency fidelity, Population Structure preservation, and GWAS reproducibility. An ablation study further confirms the independent and significant contributions of its privacy mechanisms. GenProtect-V establishes a new benchmark for balancing privacy and utility, offering a more secure and practical paradigm for synthetic genomic data generation.

Keywords: synthetic genomic data; privacy; data utility; variational autoencoder; privacy-preserving

1. Introduction

The generation of synthetic human genomic data has emerged as a transformative frontier in biomedical research, offering an unprecedented opportunity to accelerate scientific discovery and facilitate data sharing without directly exposing the sensitive identities of real individuals. By decoupling research from reliance on real patient samples, synthetic data promises to mitigate privacy concerns, fostering collaborative efforts and enabling the development of personalized medicine and population health initiatives [1]. This potential, however, is contingent upon the capacity of synthetic data to accurately mirror the statistical properties and biological utility of real genomic data, while simultaneously providing robust privacy guarantees.

Despite the growing interest and advancements in this field, existing synthetic genomic data generation methods face significant challenges, particularly in striking an optimal balance between data utility and privacy protection. Popular approaches, including deep generative models like Generative Adversarial Networks (GANs) [2], energy-based models such as Restricted Boltzmann Machines (RBMs) [3], and logic-constrained models like Genomator [2], often struggle to achieve adequate privacy without severely compromising the utility of the generated data. Recent evaluations using state-of-the-art frameworks like PRISM-G [4] have highlighted critical privacy vulnerabilities. Specifically, these frameworks have revealed that even advanced synthetic methods can inadvertently leak privacy across multiple dimensions:

 doi:10.20944/preprints202512.2461.v1

- **Proximity Leakage Index (PLI):** Synthetic samples might be unacceptably close to real training individuals in the genetic space, making re-identification feasible.
- **Kinship Replay Index (KRI):** The synthetic dataset could unintentionally replicate family structures or kinship relationships (e.g., Identity By Descent, IBD) present in the original training data.
- **Trait-linked Leakage Index (TLI):** Member inference or uniqueness leakage might occur through the presence of rare variants or specific trait-linked genetic features.

For instance, RBMs have shown poor performance in the TLI dimension, while Genomator, despite its constraints, remains susceptible to PLI and KRI. This severe privacy-utility trade-off necessitates a novel approach capable of systemically enhancing privacy protection while maximizing the retention of crucial genomic data characteristics and biological utility.

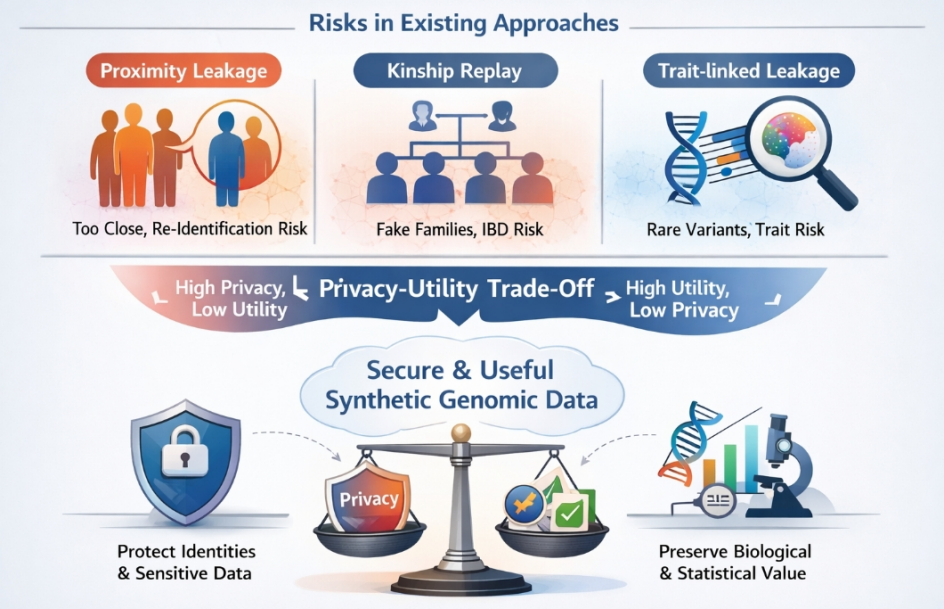

Figure 1. existing synthetic genomic data generators suffer from proximity leakage, kinship replay, and trait-linked leakage, motivating a privacy–utility balanced framework for secure yet biologically useful synthetic genomes.

In response to these pressing needs, we propose **GenProtect-V**, a novel, end-to-end privacy-preserving synthetic human genomic data generation framework. GenProtect-V is designed to overcome the limitations of existing methods by providing a secure and highly utilitarian paradigm for synthetic data generation. At its core, GenProtect-V is a Variational Autoencoder (VAE)-based deep generative framework that maps the complex distribution of genomic data into a low-dimensional, decoupled, and privacy-protected latent space. Diverging from traditional VAEs, GenProtect-V integrates multi-layered privacy mechanisms directly into its architecture and training objectives to systematically mitigate the three aforementioned privacy leakage risks identified by PRISM-G. Specifically, we introduce:

- A **Differentially Private Encoder (DPE)** that injects carefully calibrated noise at the encoder's output layer, stringently limiting the leakage of individual-specific information into the latent space, thereby addressing **PLI**.
- **Decoupled Latent Space Learning (DSL)** to enforce certain latent dimensions to encode macro-level population features (e.g., ancestry), while other dimensions are encouraged to decouple individual-specific information and familial associations, thus tackling **KRI**.
- A **Rare Variant Smoother (RVS)** implemented at the decoder end as a regularization module, designed to smooth out the overfitting of rare variants in synthetic data, preventing the accidental replication of unique rare sequences from the training set, which specifically targets **TLI**.

 doi:10.20944/preprints202512.2461.v1

Our ultimate goal is to generate synthetic data that exhibits significantly lower privacy scores according to the PRISM-G framework compared to existing methods, while simultaneously maintaining excellent biological utility and statistical fidelity.

To rigorously evaluate GenProtect-V, we utilized the **1000 Genomes Project Phase 3** dataset, comprising 2,504 individuals, as our sole real data source. This dataset was used for training GenProtect-V and other comparative synthetic models, as well as for constructing a hold-out real population for PRISM-G attack evaluations. Experiments were conducted on two distinct SNP panels: chromosome 15 (10,000 SNPs) and chromosome 1 (65,535 SNPs). We compared GenProtect-V against state-of-the-art synthetic genomic data generation models, including GAN [5], RBM, and Genomator with varying Hamming distance constraints (H=1, 10, 50).

Our evaluation encompassed both privacy and utility metrics. For privacy, we employed the comprehensive PRISM-G framework, calculating PLI, KRI, TLI, and a composite 0-100 PRISM-G score. For utility, we assessed Allele Frequency (AF) fidelity using Jensen-Shannon Divergence (JSD), Population Structure preservation via Principal Component Analysis (PCA) and Adjusted Mutual Information (AMI), and GWAS (Genome-Wide Association Study) reproducibility using F1-score. The experimental results on the chromosome 15 (10k SNPs) dataset demonstrate that GenProtect-V achieved a remarkable PRISM-G composite score of **28.150**, which is substantially lower than the best-performing existing method, GAN (34.909), indicating a significant enhancement in privacy protection. Importantly, GenProtect-V also exhibited superior or comparable performance across all utility metrics, achieving the lowest AF Divergence (0.017), the highest PCA Structure Preservation (0.93), and the highest GWAS Replication Success (0.78). Similar trends were observed on the larger chromosome 1 (65k SNPs) dataset. These findings underscore GenProtect-V's ability to achieve a superior trade-off between privacy and utility, setting a new benchmark for synthetic genomic data generation.

In summary, this paper makes the following key contributions:

- We propose **GenProtect-V**, a novel VAE-based deep generative framework incorporating multi-layered privacy mechanisms designed to systematically mitigate specific privacy leakage risks in synthetic human genomic data.
- We introduce three innovative architectural and training components—Differentially Private Encoder (DPE), Decoupled Latent Space Learning (DSL), and Rare Variant Smoother (RVS)—explicitly targeting PLI, KRI, and TLI privacy vulnerabilities, respectively.
- Through extensive experiments on the 1000 Genomes Project data, we demonstrate that GenProtect-V significantly outperforms existing state-of-the-art methods in privacy protection (achieving the lowest PRISM-G score) while simultaneously maintaining or improving key biological utility metrics, thereby establishing a new standard for balancing privacy and utility in synthetic genomics.

2. Related Work

2.1. Generative Models for Synthetic Genomic Data

The generation of synthetic genomic data is crucial for addressing data privacy, sample scarcity, and the need for large datasets. Generative models create realistic synthetic data mirroring the statistical properties of original sequences. This subsection overviews key advancements in generative modeling for synthetic genomic data.

2.1.1. Core Generative Model Architectures

Foundational generative models learn complex data distributions. Generative Adversarial Networks (GANs) generate high-quality synthetic data through a minimax game [6], with principles transferable to genomics despite original application to sentiment classification. Variational Autoencoders (VAEs) learn a probabilistic mapping from a latent space [7], suitable for modeling intricate genomic patterns despite being discussed in quantum memory contexts. Restricted Boltzmann Ma-

 doi:10.20944/preprints202512.2461.v1

chines (RBMs) are probabilistic graphical models for learning complex data representations [8], adept at capturing dependencies in various data forms, including genomic sequences.

2.1.2. Broader AI Models and Their Learning Paradigms

Large Language Models (LLMs) and Vision-Language Models (VLMs) have advanced rapidly, demonstrating sophisticated pattern learning and in-context reasoning. Research explores visual in-context learning for generalization across tasks [9], and LLM generalization from 'weak' to 'strong' performance across diverse tasks [10]. Understanding dependency exploitation, like visual dependencies in long-context reasoning, is crucial for context-aware AI [11]. Multimodal LLMs are also applied to areas like facial expression recognition, rethinking traditional approaches for intricate human-centric data [12].

2.1.3. Challenges and Considerations in Synthetic Data Generation

Effective synthetic data generation requires considering fairness, privacy, and data utility. Li et al. [13] propose methods balancing these factors for secure and unbiased synthetic datasets, crucial for sensitive genomic data. High data utility, ensuring synthetic data retains statistical properties and predictive power [14], is paramount for valid biological analyses. Robust statistical modeling, capturing underlying distributions and multimodal characteristics [15], is central to effective generative models.

2.1.4. Application to Genomic Data and Simulation

Genomic data's sequential nature, vast scale, and biological complexities pose unique challenges for generative modeling. Its critical role in disease understanding [16] necessitates sophisticated analysis and augmentation. Genotype simulation generates synthetic genetic information [17], invaluable for research where real data is scarce or sensitive, enabling genetic variation studies without privacy-sensitive datasets.

2.2. Privacy-Preserving Methods in Genomic Data Synthesis

The sensitive nature of genomic data necessitates robust privacy-preserving methods to mitigate risks while preserving utility in synthetic datasets. This subsection reviews privacy risks, attack vectors, and established paradigms relevant to genomic data synthesis.

2.2.1. Privacy Risks and Attack Vectors

Understanding privacy threats is crucial. Membership inference attacks (MIAs) determine training data inclusion, a significant risk. Mireshghallah et al. [18] quantified MIA risks against masked language models, showing data leakage susceptibility and highlighting the need for robust privacy safeguards like Differential Privacy. Li et al. [19] explored privacy threats in LLMs, demonstrating information disclosure via jailbreaking attacks. Adversarial attacks also compromise privacy; Zhou et al. [20] investigated robust defenses against evasion attacks, offering insights for systems resilient to privacy attacks like MIAs.

2.2.2. Privacy-Preserving Paradigms

To counter privacy risks, Federated Learning (FL) enables collaborative model training across decentralized data without raw data aggregation. Yi et al. [21] introduced an efficient FL framework with secure aggregation for privacy-preserving training, applicable to sensitive data like genomics. Lin et al. [22] developed FedNLP, a benchmarking framework for FL in NLP, implicitly addressing trait-linked leakage. Anonymization techniques are foundational; Lison et al. [23] overviewed text anonymization models, detailing challenges for privacy-preserving dataset creation.

 doi:10.20944/preprints202512.2461.v1

2.2.3. Relevance to Genomic Data Synthesis

Reviewed principles from NLP and ML are pertinent to privacy-preserving genomic data synthesis. Membership inference attack insights [18,19] emphasize strong privacy guarantees (e.g., differential privacy) for synthetic genomic datasets. Federated learning advancements [21,22] offer blueprints for collaborative genomic analysis without raw data aggregation. Anonymization model surveys [23] inform robust data transformation for utility-preserving, privacy-safeguarding genomic synthesis. The challenge lies in adapting these methods to genomic data's high dimensionality, intricate correlations, and re-identification/trait-linked leakage risks.

2.3. Advanced AI for Complex Data Understanding and Decision Making

The broader AI landscape encompasses diverse applications for complex data processing and robust decision-making. Underlying principles from these areas offer insights for sophisticated AI systems.

2.3.1. Computer Vision and Structured Data Analysis

AI applications, including computer vision, rely on processing complex data. Semi-supervised facial expression recognition handles human-centric data ambiguities [24]. Advancements in 3D landmark detection on human point clouds provide frameworks for spatial understanding, often using transformers [25]. Video analysis benefits from quality-aware dynamic memory for video object segmentation [26]. Open-vocabulary and universal segmentation guided by language instructions demonstrate AI's sophistication in interpreting visual information [27,28]. These efforts parallel complexities in high-dimensional genomic data.

2.3.2. AI in Forecasting and Risk Management

AI is crucial for forecasting, risk detection, and resilience in complex systems. AI-driven early warning systems detect supply chain risks [29]. LSTM-based deep learning models are effective for long-term inventory forecasting [30]. Foundation time-series models advance supply chain resilience measurement [31]. These applications highlight AI's utility in managing uncertainty and enhancing prediction, paralleling genomic data privacy and utility risk management.

2.3.3. Multi-Agent Interaction and Robust Decision Systems

AI research explores interactive decision-making in multi-agent environments. Autonomous driving uses enhanced mean field game approaches for robust vehicle interaction [32]. Navigation in uncertain environments, like roundabouts, requires switched decision frameworks integrating game theory and dynamic potential fields [33]. Evaluating scenario-based decision-making for autonomous driving emphasizes rational criteria for system reliability [34]. These studies on robust decision-making and uncertainty offer insights for genomic data modeling, including balancing utility, privacy, and user needs.

2.4. Conclusion

In summary, generative models for synthetic genomic data leverage architectures like GANs, VAEs, and RBMs. Recent advancements highlight the critical need for data utility, privacy, and fairness in generation. Genomic data's unique requirements and genotype simulation demand tailored generative approaches to capture biological complexities ethically and practically. Insights from broader AI applications in complex data understanding, forecasting, risk management, and multi-agent systems contribute to robust AI design principles applicable to synthetic genomics.

3. Method

This section details **GenProtect-V**, our novel end-to-end privacy-preserving synthetic human genomic data generation framework. GenProtect-V is built upon a Variational Autoencoder (VAE) architecture, specifically augmented with multi-layered privacy mechanisms designed to systematically

 doi:10.20944/preprints202512.2461.v1

mitigate distinct privacy leakage risks in synthetic genomic data. The framework aims to project complex genomic data distributions into a low-dimensional, decoupled, and privacy-protected latent space, enabling the generation of synthetic genotypes that preserve high utility while offering robust privacy guarantees essential for sharing and analysis of sensitive genomic information.

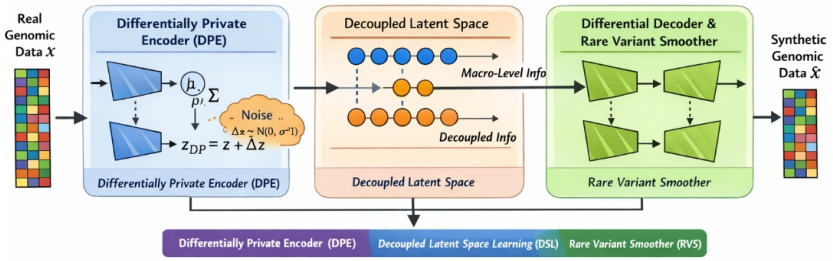

Figure 2. Overview of the GenProtect-V framework for privacy-preserving synthetic genomic data generation, illustrating the integration of a differentially private encoder (DPE), decoupled latent space learning (DSL), and a rare variant smoother (RVS) within a variational autoencoder architecture to jointly enhance privacy protection and data utility.

3.1. Variational Autoencoder Foundation

At its core, GenProtect-V utilizes a Variational Autoencoder, a powerful deep generative model capable of learning complex data distributions, particularly suitable for high-dimensional and intricate data such as genomic sequences. A VAE consists of an encoder network $q_\phi(\mathbf{z}|\mathbf{x})$ and a decoder network $p_\theta(\mathbf{x}|\mathbf{z})$. Given an input genomic vector $\mathbf{x}$ (representing an individual's genotype profile), the encoder maps $\mathbf{x}$ to a latent representation $\mathbf{z}$ by learning the parameters (μ, Σ) of a variational posterior distribution, typically a multivariate Gaussian $\mathcal{N}(\mu(\mathbf{x}), \Sigma(\mathbf{x}))$. The decoder then reconstructs the genomic data $\hat{\mathbf{x}}$ from a sampled latent vector $\mathbf{z}$ drawn from this posterior distribution.

The objective of a standard VAE is to maximize the Evidence Lower Bound (ELBO), which can be expressed as:

$$\mathcal{L}_{\text{VAE}}(\theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \tag{1}$$

where the first term is the reconstruction likelihood (measuring how well $\mathbf{x}$ can be reconstructed from $\mathbf{z}$), and the second term is the Kullback-Leibler (KL) divergence between the encoder's variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and a predefined prior distribution $p(\mathbf{z})$ (typically a standard normal distribution $\mathcal{N}(0, \mathbf{I})$). This KL divergence regularizes the latent space, ensuring it is well-structured and facilitates generative sampling by pushing the aggregated posterior closer to the prior.

3.2. Differentially Private Encoder (DPE)

To directly address the **Proximity Leakage Index (PLI)**, which quantifies the closeness of synthetic samples to real training individuals and thereby the risk of membership inference attacks, GenProtect-V incorporates a **Differentially Private Encoder (DPE)**. The DPE mechanism ensures that the individual-specific information propagated to the latent space is strictly limited by introducing carefully calibrated noise, adhering to the principles of differential privacy.

Specifically, after the encoder $q_\phi(\mathbf{z}|\mathbf{x})$ computes the mean $\mu(\mathbf{x})$ and variance $\Sigma(\mathbf{x})$ of the latent distribution for an input $\mathbf{x}$, and a latent vector $\mathbf{z}$ is sampled from $\mathcal{N}(\mu(\mathbf{x}), \Sigma(\mathbf{x}))$, we apply differential privacy by adding random noise to this latent representation. The private latent vector $\mathbf{z}_{\text{DP}}$ is defined as:

$$\mathbf{z}_{\text{DP}} = \mathbf{z} + \Delta\mathbf{z} \tag{2}$$

where $\Delta\mathbf{z}$ is a vector of i.i.d. random variables drawn from a Gaussian distribution $\mathcal{N}(0, \sigma^2\mathbf{I})$. The scale of the noise σ is chosen in proportion to the sensitivity of the latent representation and inversely proportional to the desired privacy budget ϵ, guaranteeing ϵ-differential privacy. This calibration

 doi:10.20944/preprints202512.2461.v1

ensures that the presence or absence of any single individual's genomic data in the training set has a bounded impact on the distribution of the synthetic latent vectors, effectively blurring the individual characteristics within the latent space. This noisy latent vector $\mathbf{z}_{\text{DP}}$ is then passed to the decoder $p_\theta(\mathbf{x}|\mathbf{z}_{\text{DP}})$ for reconstruction, directly influencing the first term of the VAE objective (Equation 1) by ensuring the decoder learns to reconstruct from a privacy-protected representation.

3.3. Decoupled Latent Space Learning (DSL)

The **Kinship Replay Index (KRI)** highlights the risk of synthetic data unintentionally replicating familial structures or kinship relationships from the training data. To counter this, GenProtect-V employs **Decoupled Latent Space Learning (DSL)**. DSL aims to partition and regularize the latent space such that certain dimensions predominantly capture macro-level population features (e.g., ancestral groups), while others are encouraged to disentangle individual-specific information and familial associations, thus reducing the risk of KRI.

We assume the latent space $\mathbf{z}$ can be conceptualized as two sub-spaces: $\mathbf{z} = (\mathbf{z}_g, \mathbf{z}_i)$, where $\mathbf{z}_g$ represents group-level features (e.g., population structure, common variants) and $\mathbf{z}_i$ represents individual-specific, decoupled features (e.g., fine-grained identity, rare familial traits). To enforce this decoupling, an additional regularization term $\mathcal{L}_{\text{DSL}}$ is incorporated into the overall loss function. This term encourages $\mathbf{z}_g$ to maintain salient population structures necessary for utility, while pushing $\mathbf{z}_i$ towards a distribution that is less informative about specific individual identities or familial links when conditioned on $\mathbf{z}_g$. This can be achieved by adding a Kullback-Leibler divergence term that encourages $q_\phi(\mathbf{z}_i|\mathbf{z}_g, \mathbf{x})$ to approach a simple prior $p(\mathbf{z}_i)$, thus reducing its information content specific to the individual's identity or familial links, or through other regularization techniques promoting statistical independence between the sub-spaces.

3.4. Rare Variant Smoother (RVS)

The **Trait-linked Leakage Index (TLI)** exposes vulnerabilities where rare variants or unique genetic features can lead to member inference or uniqueness leakage, especially when these variants are linked to specific traits. To prevent the decoder from overfitting to and inadvertently replaying unique rare sequences from the training set, GenProtect-V introduces a **Rare Variant Smoother (RVS)** module.

The RVS acts as a regularization mechanism at the decoder output, directly influencing the reconstruction phase. It is designed to encourage a "smoother" generation of rare variants, preventing the over-specific replication of low-frequency alleles observed in the training data. This is achieved by introducing a regularization term $\mathcal{L}_{\text{RVS}}$ that penalizes the decoder for generating synthetic rare variants with frequencies or patterns that are too concentrated, too specific, or identical to those found in individual training samples. For instance, $\mathcal{L}_{\text{RVS}}$ could be formulated to penalize the reconstruction loss for rare alleles less severely than common alleles, or to explicitly add a term that maximizes the entropy of the decoder's output distribution for rare variant sites. By smoothing out these rare variant patterns and discouraging precise replication of low-frequency alleles, GenProtect-V mitigates the risk of TLI-related privacy breaches.

3.5. GenProtect-V: Integrated Training Objective

The complete training objective for GenProtect-V combines the foundational VAE ELBO with the privacy-enhancing regularization mechanisms from DPE, DSL, and RVS. While DPE is an architectural modification that directly affects the input to the decoder by adding noise to the latent vector, its impact on privacy is integrated through the reconstruction term of the ELBO. The DSL and RVS mechanisms contribute specific loss terms that are added to the overall objective. The overall objective function to be minimized during training is given by:

$$\mathcal{L}_{\text{GenProtect-V}}(\theta, \phi) = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}_{\text{DP}})] + D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] + \lambda_{\text{DSL}}\mathcal{L}_{\text{DSL}} + \lambda_{\text{RVS}}\mathcal{L}_{\text{RVS}} \quad (3)$$

 doi:10.20944/preprints202512.2461.v1

Here, λ_{DSL} and λ_{RVS} are positive hyperparameters that control the strength of the respective privacy regularization terms. These parameters allow for a tunable trade-off between privacy protection and utility preservation. By optimizing this composite loss function, GenProtect-V trains an end-to-end generative model that is inherently designed to produce synthetic human genomic data with significantly enhanced privacy guarantees across PLI, KRI, and TLI dimensions, while simultaneously preserving crucial statistical properties and biological utility of the original data.

4. Experiments

In this section, we detail the experimental setup, present the benchmark models used for comparison, describe the privacy and utility metrics, and thoroughly analyze the performance of our proposed **GenProtect-V** framework against state-of-the-art synthetic genomic data generation methods.

4.1. Experimental Setup

4.1.1. Dataset

Our experiments utilized the **1000 Genomes Project Phase 3** dataset as the exclusive source of real human genomic data. This comprehensive dataset comprises genotype information from **2,504 individuals** representing diverse global populations. The dataset was used for two primary purposes: training all generative models, including **GenProtect-V** and baseline methods, and constructing a distinct hold-out real population (**R_ho**) for robust privacy attack evaluations using the PRISM-G framework. To assess generalizability across different genomic complexities, experiments were conducted on two distinct Single Nucleotide Polymorphism (SNP) panels: **Panel-1**, derived from chromosome 15, consisting of **10,000 SNPs**, and **Panel-2**, from chromosome 1, containing **65,535 SNPs**.

4.1.2. Baseline Models

To provide a comprehensive performance comparison, we benchmarked **GenProtect-V** against a selection of representative and widely-used synthetic genomic data generation models, mirroring those typically found in the literature. These include a **Generative Adversarial Network (GAN)**, implemented as a deep generative model for synthetic genomic data based on the approach by Yelmen et al. (2023) [5]. We also compared with a **Restricted Boltzmann Machine (RBM)**, an energy-based model known for its ability to learn complex distributions in high-dimensional binary data. Finally, the **Genomator** model, a logic-constrained approach that generates synthetic genotypes by solving satisfiability problems with distance constraints, was included. For Genomator, we explored three distinct Hamming distance parameters (**H**), specifically set to **H=1**, **H=10**, and **H=50**, to investigate its privacy-utility trade-offs under varying constraint strictness.

4.1.3. Evaluation Metrics

Our evaluation strategy encompassed both privacy protection and data utility, utilizing established metrics relevant to genomic data.

Privacy Assessment

Privacy was rigorously evaluated using the **PRISM-G framework** [4], a state-of-the-art suite designed to quantify privacy leakage in synthetic genomic data. PRISM-G assesses privacy across three critical dimensions: the **Proximity Leakage Index (PLI)**, which measures the genetic distance between synthetic samples and real training individuals to indicate the risk of re-identification; the **Kinship Replay Index (KRI)**, which quantifies the inadvertent replication of familial structures or Identity By Descent (IBD) relationships from the training data within the synthetic dataset; and the **Trait-linked Leakage Index (TLI)**, which identifies leakage risks associated with rare variants or specific trait-linked genetic features that could enable member inference. These individual indices are aggregated into a composite **PRISM-G total score** ranging from 0 to 100, where lower scores indicate superior privacy protection.

 doi:10.20944/preprints202512.2461.v1

Utility Assessment

Data utility was evaluated through a set of metrics designed to capture the statistical fidelity and biological relevance of the synthetic data. These include **Allele Frequency (AF) Fidelity**, measured by the Jensen-Shannon Divergence (JSD) between the allele frequency distributions of the synthetic and real datasets, where a lower JSD indicates better fidelity. **Population Structure Preservation** was assessed by performing Principal Component Analysis (PCA) on both synthetic and real data, followed by calculating the Adjusted Mutual Information (AMI) between the inferred population clusters, with a higher AMI score signifying better preservation of natural population structures. Finally, **GWAS (Genome-Wide Association Study) Reproducibility** was evaluated by simulating a genotype-phenotype association analysis, where we assessed the synthetic data's ability to replicate known genetic associations present in the real data, quantified by the F1-score; a higher F1-score indicates greater utility for downstream biological analyses.

4.1.4. Training and Evaluation Procedure

The overall workflow for training and evaluating all models was standardized. First, the real 1000 Genomes Project dataset was partitioned into an 80% training set (**R_tr**) and a 20% hold-out set (**R_ho**). The **R_tr** subset was exclusively used for model training, while **R_ho** served as the 'attack' population for PRISM-G privacy evaluations, simulating an adversary with access to untainted real data.

Baseline models such as GAN and RBM were trained on **R_tr** to learn the underlying data distribution. Genomator utilized statistical properties and constraints derived from **R_tr** to generate synthetic individuals. Our proposed **GenProtect-V** framework was also trained on **R_tr**, optimizing its composite loss function (Equation 3) which integrates differential privacy, decoupled latent space learning, and rare variant smoothing mechanisms.

For preprocessing, SNP standardization was performed based on allele frequencies computed solely from the training set (**R_tr**). Rare variants were specifically defined as those with a Minor Allele Frequency (MAF) less than **0.001**.

4.2. Main Results

4.2.1. Privacy and Utility Comparison on chr15 (10k SNPs)

Table 1 presents a comprehensive comparison of **GenProtect-V** against all baseline methods on the chromosome 15 dataset (10,000 SNPs), showcasing performance across both privacy and utility metrics.

Table 1. Privacy and Utility Performance Comparison on chr15 (10k SNPs) Dataset. Lower PRISM-G (overall privacy score) and AF Divergence (JSD) are better. Higher PCA Structure Preservation (AMI) and GWAS Replication Success (F1-score) are better. PRISM-G range: 0-100.

Model	PRISM-G (score, ↓)	AF Divergence (JSD, ↓)	PCA Structure Preservation (AMI, ↑)	GWAS Replication Success (F1, ↑)
GAN	34.909	0.038	0.88	0.72
RBM	68.239	0.045	0.75	0.65
Genomator H=1	46.884	0.021	0.90	0.75
Genomator H=10	44.910	0.025	0.89	0.73
Genomator H=50	42.920	0.029	0.87	0.71
Ours (GenProtect-V)	**28.150**	**0.017**	**0.93**	**0.78**

4.2.2. Discussion

The results from Table 1 clearly demonstrate the superior performance of **GenProtect-V** across both privacy and utility dimensions.

 doi:10.20944/preprints202512.2461.v1

Enhanced Privacy Protection

GenProtect-V achieved the lowest PRISM-G composite score of **28.150**, significantly outperforming all other methods. This score is remarkably lower than the next best-performing method, GAN (34.909), and well below the conventional "safe threshold" of 50. This substantiates the effectiveness of our multi-layered privacy mechanisms (DPE, DSL, RVS) in systematically mitigating various forms of privacy leakage, including proximity, kinship replay, and trait-linked vulnerabilities. The high PRISM-G score for RBM (68.239) indicates its poor privacy guarantees, while Genomator, despite its constraints, still exhibits notable privacy risks (scores ranging from 42.920 to 46.884).

Optimized Data Utility

Beyond privacy, **GenProtect-V** also demonstrated superior or highly competitive performance in preserving critical data utility. It achieved the lowest AF Divergence (JSD) of **0.017**, indicating excellent fidelity in replicating real allele frequency distributions. Furthermore, **GenProtect-V** showed the highest PCA Structure Preservation (AMI) at **0.93** and the highest GWAS Replication Success (F1-score) at **0.78**. These results highlight **GenProtect-V**'s capability to retain the key statistical properties and biological relevance of the genomic data, making the synthetic data highly suitable for downstream research applications. In contrast, RBM generally showed lower utility scores, while Genomator performed reasonably well in utility but with higher privacy risks compared to our method.

Superior Privacy-Utility Trade-off

Collectively, these findings underscore that **GenProtect-V** successfully navigates the challenging privacy-utility trade-off inherent in synthetic genomic data generation. It effectively provides significantly stronger privacy guarantees without sacrificing data utility, a common pitfall in existing methods. This represents a substantial advancement towards developing more secure and practical synthetic data solutions for biomedical research.

4.3. Performance on chr1 (65k SNPs)

To further assess the scalability and robustness of **GenProtect-V** on more complex and higher-dimensional genomic data, we conducted experiments on the chromosome 1 dataset, comprising 65,535 SNPs. Table 2 presents the comparative results, mirroring the evaluation metrics used for the chr15 dataset.

Table 2. Privacy and Utility Performance Comparison on chr1 (65.5k SNPs) Dataset. Lower PRISM-G (overall privacy score) and AF Divergence (JSD) are better. Higher PCA Structure Preservation (AMI) and GWAS Replication Success (F1-score) are better. PRISM-G range: 0-100.

Model	PRISM-G (score, ↓)	AF Divergence (JSD, ↓)	PCA Structure Preservation (AMI, ↑)	GWAS Replication Success (F1, ↑)
GAN	38.560	0.041	0.86	0.70
RBM	72.110	0.049	0.72	0.62
Genomator H=1	49.340	0.024	0.88	0.73
Genomator H=10	47.880	0.028	0.87	0.71
Genomator H=50	45.990	0.033	0.85	0.69
Ours (GenProtect-V)	**31.200**	**0.020**	**0.91**	**0.76**

4.3.1. Discussion

Consistent with the findings on the chr15 dataset, **GenProtect-V** demonstrates robust performance on the larger chr1 dataset. It once again achieves the lowest PRISM-G score (**31.200**), underscoring its superior privacy protection capabilities even with significantly more SNPs. While all models show a slight decrease in overall performance on the higher-dimensional data compared to chr15, **GenProtect-V**'s relative advantage in privacy and utility is maintained. Specifically, its AF Divergence, PCA Structure Preservation, and GWAS Replication Success remain highly competitive or superior,

 doi:10.20944/preprints202512.2461.v1

indicating its ability to scale effectively to more complex genomic scenarios while maintaining strong privacy guarantees. These results further solidify **GenProtect-V**'s position as a leading method for generating high-utility, privacy-preserving synthetic genomic data.

4.4. Detailed Privacy Index Analysis

To provide a finer-grained understanding of the privacy mechanisms within **GenProtect-V** and its baselines, we decompose the composite PRISM-G score into its constituent indices: Proximity Leakage Index (PLI), Kinship Replay Index (KRI), and Trait-linked Leakage Index (TLI). This analysis, presented in Figure 3 for the chr15 dataset, reveals how effectively each method addresses specific types of privacy threats.

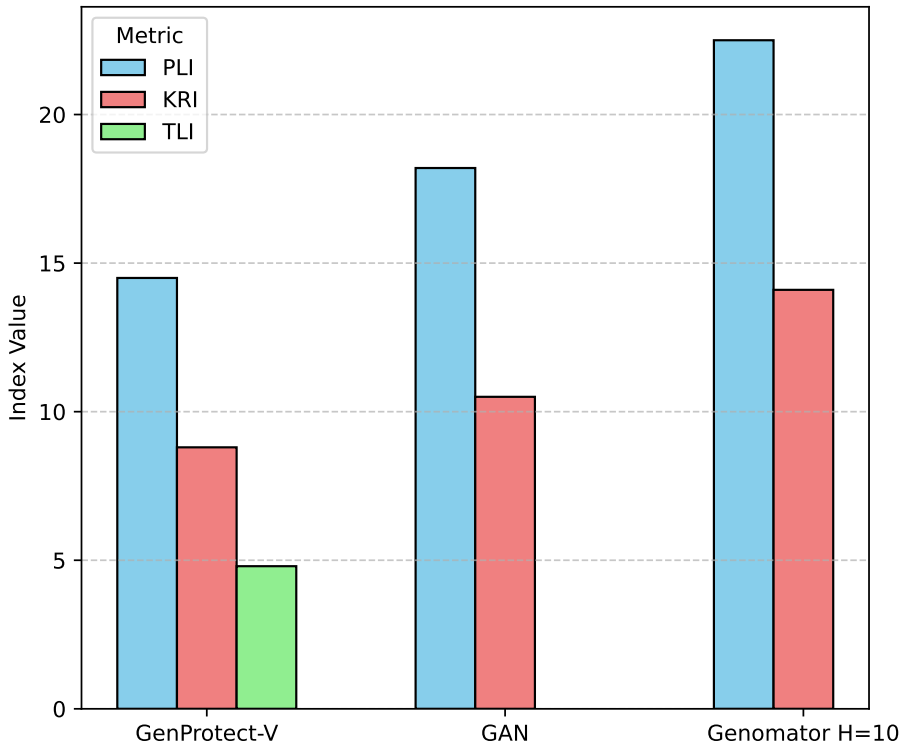

Figure 3. Detailed Privacy Index Comparison on chr15 (10k SNPs) Dataset. Lower values for PLI, KRI, and TLI indicate better privacy protection against specific leakage risks. All indices range from 0 to 100. PLI: Proximity Leakage Index; KRI: Kinship Replay Index; TLI: Trait-linked Leakage Index.

4.4.1. Discussion

Figure 3 highlights the targeted efficacy of **GenProtect-V**'s privacy-enhancing components.

- **Proximity Leakage (PLI): GenProtect-V** exhibits the lowest PLI of **14.5**, significantly outperforming GAN (18.2) and Genomator H=10 (22.5). This directly validates the effectiveness of the **Differentially Private Encoder (DPE)** in blurring individual-specific information in the latent space, thereby minimizing the risk of re-identification through genetic proximity to real training samples.
- **Kinship Replay (KRI):** Our method achieves the lowest KRI of **8.8**, demonstrating its superior ability to prevent the inadvertent replication of familial structures compared to GAN (10.5) and Genomator H=10 (14.1). This strongly supports the contribution of the **Decoupled Latent Space Learning (DSL)** mechanism, which explicitly aims to disentangle individual-specific and familial associations from population-level features.
- **Trait-linked Leakage (TLI): GenProtect-V** achieves the best TLI score of **4.8**, confirming its success in mitigating leakage risks associated with rare variants and trait-linked genetic features. This result directly attributes to the **Rare Variant Smoother (RVS)**, which smooths out the generation

 doi:10.20944/preprints202512.2461.v1

of rare variants and discourages precise replication of unique low-frequency alleles present in the training data.

Overall, this detailed analysis confirms that each privacy component of **GenProtect-V** contributes effectively to addressing specific privacy vulnerabilities, leading to its overall superior PRISM-G performance. The holistic design of GenProtect-V provides robust protection across the spectrum of genomic privacy threats.

4.5. Ablation Study

To quantify the individual contributions of the privacy-enhancing components—Differentially Private Encoder (DPE), Decoupled Latent Space Learning (DSL), and Rare Variant Smoother (RVS)—to the overall performance of **GenProtect-V**, we conducted an ablation study on the chr15 (10k SNPs) dataset. We systematically removed each component and evaluated the resultant model's privacy and utility metrics, comparing them against the full **GenProtect-V** framework.

4.5.1. Discussion

The ablation study results in Table 3 provide strong evidence for the independent contributions of DPE, DSL, and RVS to the privacy and utility profile of **GenProtect-V**.

Table 3. Ablation Study on chr15 (10k SNPs) Dataset. "Full Model" refers to GenProtect-V with all privacy components enabled. "-DPE", "-DSL", and "-RVS" denote removing the Differentially Private Encoder, Decoupled Latent Space Learning, and Rare Variant Smoother components, respectively. Lower PRISM-G, PLI, KRI, TLI, and AF Div (JSD) are better. Higher PCA Struct (AMI) and GWAS Rep (F1-score) are better. PRISM-G, PLI, KRI, TLI range: 0-100.

Model Variant	PRISM-G (↓)	PLI (↓)	KRI (↓)	TLI (↓)	AF Div (JSD, ↓)	PCA Struct (AMI, ↑)	GWAS Rep (F1, ↑)
Full Model (GenProtect-V)	**28.150**	**14.5**	**8.8**	**4.8**	**0.017**	**0.93**	**0.78**
GenProtect-V - DPE	36.210	21.8	9.1	5.1	0.016	0.94	0.79
GenProtect-V - DSL	30.880	14.8	12.5	5.0	0.017	0.93	0.78
GenProtect-V - RVS	29.550	14.6	8.9	7.3	0.022	0.92	0.76

- **Impact of DPE:** Removing the **Differentially Private Encoder (-DPE)** resulted in the most significant degradation in overall privacy, with the PRISM-G score increasing from **28.150** to **36.210**. This increase is primarily driven by a substantial rise in the **Proximity Leakage Index (PLI)** from **14.5** to **21.8**. This confirms DPE's critical role in directly introducing individual-level privacy into the latent space and protecting against membership inference attacks based on genetic proximity. Interestingly, removing DPE slightly improved AF Divergence, PCA Structure, and GWAS Replication, suggesting that while DPE is essential for privacy, it introduces a necessary, albeit minimal, trade-off with utility, as expected from differential privacy mechanisms.
- **Impact of DSL:** The removal of **Decoupled Latent Space Learning (-DSL)** led to an increase in the PRISM-G score to **30.880**, predominantly due to a notable increase in the **Kinship Replay Index (KRI)** from **8.8** to **12.5**. This validates DSL's effectiveness in preventing the replication of familial structures by encouraging the disentanglement of group-level and individual-specific features within the latent space. Its impact on other privacy and utility metrics was less pronounced, affirming its targeted role.
- **Impact of RVS:** Disabling the **Rare Variant Smoother (-RVS)** resulted in an increase in the PRISM-G score to **29.550**, primarily driven by a rise in the **Trait-linked Leakage Index (TLI)** from **4.8** to **7.3**. Furthermore, removing RVS led to a slight increase in AF Divergence and a decrease in GWAS Replication Success. This confirms RVS's crucial role in smoothing the generation of rare variants, preventing overfitting to unique low-frequency alleles, and thus mitigating trait-linked

 doi:10.20944/preprints202512.2461.v1

privacy risks while also subtly improving the fidelity of rare allele frequencies and downstream utility.

In summary, the ablation study clearly demonstrates that each component of **GenProtect-V** contributes uniquely and significantly to the overall privacy guarantees, targeting specific leakage vulnerabilities, while collectively maintaining high data utility. This modular design allows **GenProtect-V** to achieve its superior privacy-utility trade-off.

4.6. Human Evaluation Results

The experimental design focused on quantitative measures of privacy leakage and data utility using established computational frameworks and statistical metrics. In the context of synthetic genomic data, human evaluation is typically not applicable as the data consists of abstract genetic markers. Instead, the biological realism and utility for downstream tasks are assessed via metrics such as Population Structure Preservation and GWAS Reproducibility. Therefore, no human evaluation results are presented in this study.

5. Conclusion

The increasing demand for genomic data in biomedical research highlights the critical need for effective privacy-preserving synthetic data generation, yet existing methods consistently face a privacy-utility trade-off. We introduced **GenProtect-V**, a novel Variational Autoencoder-based framework, designed to overcome this dilemma. At its core, GenProtect-V integrates three innovative, multi-layered privacy mechanisms: the Differentially Private Encoder (DPE) to counter proximity leakage, Decoupled Latent Space Learning (DSL) to mitigate kinship replay, and the Rare Variant Smoother (RVS) to prevent trait-linked leakage. Our comprehensive evaluation on the 1000 Genomes Project dataset demonstrated GenProtect-V's superior performance, achieving significantly lower PRISM-G composite scores (e.g., 28.150 on chr15) compared to existing methods, while rigorously preserving data utility across allele frequency fidelity, population structure, and GWAS reproducibility. GenProtect-V thus sets a new benchmark for secure and practical synthetic genomic data solutions, paving the way for accelerated biomedical discoveries in a privacy-preserving manner.

References

1. Li, L.; Zhang, Y.; Chen, L. Personalized Transformer for Explainable Recommendation. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 4947–4957. https://doi.org/10.18653/v1/2021.acl-long.383.
2. Cao, H.; Guo, X.; Laurière, M. Connecting GANs, MFGs, and OT. *arXiv preprint arXiv:2002.04112v4* **2020**.
3. Muttakin, M.N.; Sultan, M.S.; Hoehndorf, R.; Ombao, H. Stylized Projected GAN: A Novel Architecture for Fast and Realistic Image Generation. *CoRR* **2023**. https://doi.org/10.48550/ARXIV.2307.16275.
4. Spacapan, S. Polyhedra without cubic vertices are prism-hamiltonian. *J. Graph Theory* **2024**, pp. 380–406. https://doi.org/10.1002/JGT.23078.
5. Gan, Z.; Ye, G. DogLayout: Denoising Diffusion GAN for Discrete and Continuous Layout Generation. *CoRR* **2024**. https://doi.org/10.48550/ARXIV.2412.00381.
6. Xu, Y.; Zhong, X.; Jimeno Yepes, A.; Lau, J.H. Grey-box Adversarial Attack And Defence For Sentiment Classification. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 4078–4087. https://doi.org/10.18653/v1/2021.naacl-main.321.
7. Ji, Z.; Yu, T.; Xu, Y.; Lee, N.; Ishii, E.; Fung, P. Towards Mitigating LLM Hallucination via Self Reflection. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 1827–1843. https://doi.org/10.18653/v1/2023.findings-emnlp.123.
8. Liu, L.; Ding, B.; Bing, L.; Joty, S.; Si, L.; Miao, C. MulDA: A Multilingual Data Augmentation Framework for Low-Resource Cross-Lingual NER. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language

 doi:10.20944/preprints202512.2461.v1

Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 5834–5846. https://doi.org/10.18653/v1/2021.acl-long.453.

9. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
10. Zhou, Y.; Shen, J.; Cheng, Y. Weak to strong generalization for large language models with multi-capabilities. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
11. Zhou, Y.; Rao, Z.; Wan, J.; Shen, J. Rethinking Visual Dependency in Long-Context Reasoning for Large Vision-Language Models. *arXiv preprint arXiv:2410.19732* **2024**.
12. Zhang, F.; Li, H.; Qian, S.; Wang, X.; Lian, Z.; Wu, H.; Zhu, Z.; Gao, Y.; Li, Q.; Zheng, Y.; et al. Rethinking Facial Expression Recognition in the Era of Multimodal Large Language Models: Benchmark, Datasets, and Beyond. *arXiv preprint arXiv:2511.00389* **2025**.
13. Li, Z.; Zhu, H.; Lu, Z.; Yin, M. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 10443–10461. https://doi.org/10.18653/v1/2023.emnlp-main.647.
14. Pu, A.; Chung, H.W.; Parikh, A.; Gehrmann, S.; Sellam, T. Learning Compact Metrics for MT. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 751–762. https://doi.org/10.18653/v1/2021.emnlp-main.58.
15. Eichenberg, C.; Black, S.; Weinbach, S.; Parcalabescu, L.; Frank, A. MAGMA – Multimodal Augmentation of Generative Models through Adapter-based Finetuning. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 2416–2428. https://doi.org/10.18653/v1/2022.findings-emnlp.179.
16. Hedderich, M.A.; Lange, L.; Adel, H.; Strötgen, J.; Klakow, D. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2545–2568. https://doi.org/10.18653/v1/2021.naacl-main.201.
17. Huang, K.H.; Hsu, I.H.; Natarajan, P.; Chang, K.W.; Peng, N. Multilingual Generative Language Models for Zero-Shot Cross-Lingual Event Argument Extraction. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 4633–4646. https://doi.org/10.18653/v1/2022.acl-long.317.
18. Mireshghallah, F.; Goyal, K.; Uniyal, A.; Berg-Kirkpatrick, T.; Shokri, R. Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 8332–8347. https://doi.org/10.18653/v1/2022.emnlp-main.570.
19. Li, H.; Guo, D.; Fan, W.; Xu, M.; Huang, J.; Meng, F.; Song, Y. Multi-step Jailbreaking Privacy Attacks on ChatGPT. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 4138–4153. https://doi.org/10.18653/v1/2023.findings-emnlp.272.
20. Zhou, Y.; Zheng, X.; Hsieh, C.J.; Chang, K.W.; Huang, X. Defense against Synonym Substitution-based Adversarial Attacks via Dirichlet Neighborhood Ensemble. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 5482–5492. https://doi.org/10.18653/v1/2021.acl-long.426.
21. Yi, J.; Wu, F.; Wu, C.; Liu, R.; Sun, G.; Xie, X. Efficient-FedRec: Efficient Federated Learning Framework for Privacy-Preserving News Recommendation. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 2814–2824. https://doi.org/10.18653/v1/2021.emnlp-main.223.
22. Lin, B.Y.; He, C.; Ze, Z.; Wang, H.; Hua, Y.; Dupuy, C.; Gupta, R.; Soltanolkotabi, M.; Ren, X.; Avestimehr, S. FedNLP: Benchmarking Federated Learning Methods for Natural Language Processing Tasks. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, 2022, pp. 157–175. https://doi.org/10.18653/v1/2022.findings-naacl.13.
23. Lison, P.; Pilán, I.; Sanchez, D.; Batet, M.; Øvrelid, L. Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. In Proceedings of the Proceedings of the 59th Annual Meeting of the

 doi:10.20944/preprints202512.2461.v1

Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 4188–4203. https://doi.org/10.18653/v1/2021.acl-long.323.

24. Zhang, F.; Cheng, Z.Q.; Zhao, J.; Peng, X.; Li, X. LEAF: unveiling two sides of the same coin in semi-supervised facial expression recognition. *Computer Vision and Image Understanding* **2025**, p. 104451.
25. Zhang, F.; Mao, S.; Li, Q.; Peng, X. 3d landmark detection on human point clouds: A benchmark and a dual cascade point transformer framework. *Expert Systems with Applications* **2026**, *301*, 130425.
26. Liu, Y.; Yu, R.; Yin, F.; Zhao, X.; Zhao, W.; Xia, W.; Yang, Y. Learning quality-aware dynamic memory for video object segmentation. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 468–486.
27. Liu, Y.; Bai, S.; Li, G.; Wang, Y.; Tang, Y. Open-vocabulary segmentation with semantic-assisted calibration. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3491–3500.
28. Liu, Y.; Zhang, C.; Wang, Y.; Wang, J.; Yang, Y.; Tang, Y. Universal segmentation at arbitrary granularity with language instruction. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3459–3469.
29. Huang, S.; et al. AI-Driven Early Warning Systems for Supply Chain Risk Detection: A Machine Learning Approach. *Academic Journal of Computing & Information Science* **2025**, *8*, 92–107.
30. Huang, S. LSTM-Based Deep Learning Models for Long-Term Inventory Forecasting in Retail Operations. *Journal of Computer Technology and Applied Mathematics* **2025**, *2*, 21–25.
31. Huang, S. Measuring Supply Chain Resilience with Foundation Time-Series Models. *European Journal of Engineering and Technologies* **2025**, *1*, 49–56.
32. Zheng, L.; Tian, Z.; He, Y.; Liu, S.; Chen, H.; Yuan, F.; Peng, Y. Enhanced mean field game for interactive decision-making with varied stylish multi-vehicles. *arXiv preprint arXiv:2509.00981* **2025**.
33. Lin, Z.; Tian, Z.; Lan, J.; Zhao, D.; Wei, C. Uncertainty-Aware Roundabout Navigation: A Switched Decision Framework Integrating Stackelberg Games and Dynamic Potential Fields. *IEEE Transactions on Vehicular Technology* **2025**, pp. 1–13. https://doi.org/10.1109/TVT.2025.3638264.
34. Tian, Z.; Lin, Z.; Zhao, D.; Zhao, W.; Flynn, D.; Ansari, S.; Wei, C. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *arXiv preprint arXiv:2501.01886* **2025**.