Gene-Guard: Real-Time Genomic Data Leak Prevention

Angana Mondal [1]
EPFL

Ahmed Rockey Saikia [2]
EPFL

Richard Guan [3]
Independent

Riccardo Campanella [4]
Utrecht University

Neha Suresh [5]
Independent Biosecurity Researcher

With
Apart Research

Abstract

Genomic data (such as DNA sequences) are both highly sensitive and uniquely vulnerable to accidental leakage in modern collaborative and AI-assisted workflows. This report introduces Gene-Guard, a two-layer detection and prevention system designed to safeguard genomic sequences from unauthorized exposure in real time. We articulate the growing risk through real-world scenarios: researchers inadvertently pasting DNA sequences into chat platforms or AI tools, and bio-lab staff emailing unencrypted gene data. To address this gap, Gene-Guard integrates an open DNA sequence dataset for training and benchmarking machine learning classifiers that flag genomic data with high recall and precision. Upon detection, an optional second layer provides an LLM-powered risk analysis, identifying the sequence and its potential biosecurity implications using tools like IBBIS's Common Mechanism and SeqScreen. Emphasis is placed on a lightweight, fast, platform-agnostic tool that can plug into any Data Loss Prevention pipeline without heavy resource demands or latency. Our results suggest Gene-Guard can significantly reduce the chance of genomic data leaks without impeding researchers' workflows, offering a novel safety net at the intersection of cyber and biosecurity.

Keywords: Genomic Data Leakage, Data Loss Prevention, Biosecurity, Insider Threat, Compliance, LLM Safety

1. Introduction

Genomic data (e.g. DNA sequences) are increasingly recognized as *sensitive information* that demands protection on par with personal identifiers or financial records. In fact, regulations like the EU's GDPR explicitly categorize genetic data as sensitive personal data requiring strict safeguards. Beyond privacy, the misuse of leaked DNA sequences poses serious **biosecurity risks** – for example, a leaked pathogenic genome could enable malicious synthesis of a virus. The 2023 breach of genetic testing firm 23andMe, which exposed raw genotype data of ~6.9 million people, underscores the high stakes and potential for abuse when genomic information is not secure.

Informed by discussions with genomics researchers and information security professionals, we identified a clear blind spot in current enterprise protection layers. Researchers admitted to casually pasting sequences into AI notebooks. Security engineers confirmed their DLP engines could not parse or prioritize such content. Gene-Guard is designed to close this gap.

Yet beyond external breaches, *insider accidents and negligence* are an overlooked threat vector for genomic data leakage. Researchers and lab personnel regularly handle DNA/RNA sequences in day-to-day work, sometimes in settings lacking adequate data confinement controls. Consider the following real-world scenario:

> Researchers increasingly use cloud-based AI tools (like ChatGPT or code assistants) to analyze or format genomic data. Corporate studies show a significant portion of employees paste confidential data into public LLMs, often unaware that these prompts are stored on external servers. A DNA sequence pasted into an AI chatbot may be retained and even used to train models, constituting an irreversible leak outside the organization's perimeter.

This scenario highlights a pressing need for proactive genomic Data Loss Prevention (DLP). Traditional DLP systems monitor common sensitive data (PII, PHI, credit card numbers, etc.) across endpoints, networks, and cloud apps, but *they are not tailored to recognize DNA sequences.* A string of A, C, G, T letters can easily bypass keyword-based filters, especially if the sequence is embedded in other text or deliberately obfuscated. Moreover, in AI-driven workflows, language-based transformations (summarizations, translations, or even DNA-to-protein conversions) could hide the content from simplistic rules. In short, current enterprise security tools often lack the "genomic awareness" to catch such leaks, creating a blind spot.

Threat Model**:** We consider three main vectors:

(a)AI or automated agents – systems like chatbots or office suite copilot features that might autonomously access or transmit genomic data (for instance, an LLM integrated with lab notebooks that might leak sequences if exploited).

(b) Malicious insiders – A rogue employee exfiltrating genomic IP (such as a vaccine DNA sequence) via copy-paste or by slowly leaking segments.

(c) Negligent or unaware insiders – well-intentioned staff whose normal use of cloud tools or messaging inadvertently exposes protected sequence data; and

(c) These threat vectors span both the insider threat domain and emerging *"shadow AI"* risks in which organizational data escapes via unsanctioned AI services.

Research Questions- *Can we build a real-time, low-overhead detection mechanism to identify genomic data in unstructured content and prevent its unauthorized sharing at an enterprise level?*

Our work contributes an affirmative answer by developing a two-layer detection pipeline for genomic data leak prevention. Gene-Guard is designed as a drop-in library that can augment existing DLP solutions. Crucially, it emphasizes speed, accuracy, and domain/context-awareness: the system should detect DNA sequences (even if hidden in noise), do so with minimal latency and resource use, and provide context on why a detected sequence is risky. By focusing on genomic data, Gene-Guard fills a critical gap in security tooling for biotech and healthcare organizations, aligning with calls for sector-specific DLP guidance. Ultimately, this approach aims to reduce accidental genomic leaks and to raise the hurdle for malicious actors trying to smuggle out dangerous genetic information.

2. Methods

To detect and assess the biosecurity risk of leaked genomic data, we developed Gene-Guard, a two-layer genomic data loss prevention (DLP) system that combines rapid sequence detection with contextual biological risk analysis.

Two-Layer Architecture

The first layer performs real-time screening on a document to identify actual or obfuscated genomic content. An expansive literature review suggested that although there is ample research in classifying DNA sequences, there is very little work focusing on detection/extraction of DNA sequences from unstructured data. First, we design a rule-based heuristic filter which can detect long ATCG-only substrings (e.g., $>$50 bp), nucleotide frequency imbalance characteristic of genomic sequences, and common formatting patterns such as FASTA headers. However, simple rule-based classifiers often miss the detection of obfuscated sequences (e.g., A→@, T→7 substitutions, or breaking the sequence with random characters). To address this, we present a natural language processing (NLP) model combined with logistic regression, using k-mer frequency distributions, entropy, and Markov transition matrices as feature representations (building on the works of Wren et al. (2005), Strzoda et al. (2023) and Wren et al. (2005)). The pipeline is further compared against a 1D convolutional neural network (1D-CNN) which directly learns sequence patterns from raw input. For training and evaluating the models, we curate an extensive open-source dataset incorporating more than ten "adversarial variations" engineered from raw public DNA data.

The second layer provides contextual risk analysis for sequences flagged as DNA-positive. The system first performs deterministic sequence extraction,

isolating the longest genomic substring and cleaning formatting artifacts. Sequence identity screening is then conducted using BLAST against curated biosecurity databases including NCBI pathogenic repositories and the U.S. Select Agents list, following the Common Mechanism biosecurity screening framework. Sequences exhibiting $\geq$90% identity over $\geq$300 bp to known harmful sequences are categorized as high risk, in alignment with established screening thresholds (zaixizhang.github.io).

For functional risk assessment, we integrate SeqScreen, an open-source functional screening platform that applies ensemble machine learning to assign taxonomic labels and Functions-of-Concern (FunSoC), identifying virulence factors, toxin domains, and antibiotic resistance genes even in novel sequences (Genome Biology, SeqScreen) (genomebiology.biomedcentral.com). This also enables detection of human genomic DNA, supporting privacy classifications relevant to regulatory compliance (e.g., HIPAA).

Large language models (LLMs) are used to generate concise risk summaries based on BLAST and SeqScreen outputs, aiding analyst interpretation without exposing full sequence content externally. In particular, LLMs can be employed to detect genomic sequences inside textual descriptions. In particular, a short dataset can be used to fine-tune a model such as GPT-2 to autoregressively predict the probability that a token belongs to a genomic sequence. The attention analysis not only serves as a biomarker, allowing to highlight the input flagged as DNA positive, but also to reveal how Transformers' attention heads specialize in detecting different types of DNA masking based on their difficulty.

Deployment and Evaluation

Gene-Guard was trained and validated using a curated dataset comprising authentic genomic sequences and adversarial negative samples, split into training and held-out test sets to measure generalizability. This layer processes kilobyte-scale documents in under one second, enabling integration into real-time DLP pipelines.

The system outputs a composite risk classification (Low–Critical), descriptive tags, and incident metadata for SIEM integration, supporting auditability and regulatory reporting. All screening can be performed locally to ensure data privacy and prevent secondary leakage.

3. Results

Table 1 below summarises the overall performance of the different classifiers on our test data, which included multiple adversarial cases.

	Heuristic Classifier	**NLP Classifier**	**1-D CNN Classifier**
Test-set performance	**Accuracy : 0.8685** **Precision: 0.9776** **Recall : 0.7543** **F1 Score : 0.8515**	**Accuracy : 0.9629** **Precision: 0.9890** **Recall : 0.9362** **F1 Score : 0.9619**	**Accuracy : 0.8452** **Precision: 0.845** **Recall : 0.8448** **F1 Score : 0.8451**

Table 1 – Overall performance on test set

We observe that the NLP classifier performs best at detecting genome-like sequences, effectively closing many adversarial gaps that the heuristic filter alone cannot catch. To further assess the predictive capabilities of our technique, we curate a more 'advanced' test set with several adversarial cases which were not seen during training. We observe that the NLP and CNN models perform significantly better than the heuristic filter, signalling good transference to unseen adversarial variations. (For detailed results, please see Appendix, Tables 2 and 3).

GPT-2's attention analysis suggests that layer-specific attention heads specialize in recognizing specific genomic masking types based on the masking difficulty level (Figure 1, Figure 2.).

4. Discussion and Conclusion

Gene-Guard was conceived to fill a critical gap at the nexus of biosecurity and AI-driven data loss prevention. It shows that genomic data leakage is a solvable problem. Our solution is efficient, integrable, and urgently needed. As AI tools touch more of science and healthcare, the Gene-Guard approach can help secure the next frontier of digital biology.

Key Takeaways:

- Traditional DLP tools lack the domain intelligence to detect genomic sequences.
- Gene-Guard integrates detection using optimized lightweight machine learning, context-aware logging, and optional LLM risk previews.
- Its layered architecture balances speed, accuracy, and interpretability.

As genomic data increasingly circulates through AI tools, labs, and cloud platforms, conventional DLP systems fail to identify such sequences. Leaks of DNA sequences can violate privacy laws (when human-derived) and expose national security or intellectual property risks (when synthetic or pathogenic). Gene-Guard's real-time detection and incident response capabilities make it a critical component of forward-looking AI safety infrastructure.

5. References

- NIST IR 8432, *Cybersecurity of Genomic Data* – Technical report outlining challenges in genomic data protection (nvlpubs.nist.govnvlpubs.nist.gov.

- Suzanne Smalley, *NIST report identifies significant privacy gaps in genomic data handling*, Recorded Future News (2023) – News article summarizing NIST findings on genomic data sharing weaknesses (therecord.media).

- Cyberhaven Report (2023) – Study finding 11% of content employees paste into ChatGPT is confidential (cyberhaven.com).

- Ken Underhill, *77% of Employees Share Company Secrets on ChatGPT, Report Warns*, eSecurityPlanet (2025) (esecurityplanet.com.)

- PromptArmor, *Slack AI Bug could expose private data*, DarkReading (2024) – Describes Slack AI prompt injection risk and mentions sensitive data exposure via Slackdarkreading.comdarkreading.com.

- Lakera, *Data Loss Prevention: A Complete Guide for the GenAI Era* (2025) – Discusses how traditional DLP struggles with AI; notes employees pasting data into LLMslakera.ai and need for language-aware DLPlakera.ai.

- Microsoft TechCommunity Blog, *Purview tackles GenAI risks* (2025) – Highlights new Purview features for AI data protection: real-time blocking, audit trailstechcommunity.microsoft.com.

- Wiz.io, *Protect sensitive data with DSPM* – Cloud data security solution documentation; mentions scanning for PII, PHI, etc., and custom classifierswiz.io.

- Google Cloud, *Handling Genomic Data in the Cloud* – White paper noting Cloud DLP's ability to classify genomic datacloud.google.com.

- IBBIS, *Common Mechanism for DNA Screening* – Biosecurity initiative defining DNA sequence screening thresholds (90% identity to known pathogens)zaixizhang.github.io.

- Advait Balaji et al., *SeqScreen: accurate and sensitive functional screening of pathogenic sequences*, Genome Biology 23, 133 (2022) – Describes SeqScreen tool and its approach to labeling Functions-of-Concern in DNA sequences

- Strzoda, Tomasz, Lourdes Cruz-Garcia, Mustafa Najim, Christophe Badie, and Joanna Polanska., "A mapping-free natural language processing-based technique for sequence search in nanopore long-reads." Scientific Reports 13, no. 1 (2023): 1–12.

- Yang, Aimin, Wei Zhang, Jiahao Wang, Ke Yang, Yang Han, and Limin Zhang."Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA." Genes 12, no. 6 (2021): 920.

- Wren, Jonathan D., William H. Hildebrand, Sreedevi Chandrasekaran, and Ulrich Melcher., "Markov model recognition and classification of DNA/protein sequences within large text databases." Bioinformatics 21, no. 14 (2005): 3368–3376.

- Shtatland, Timur, Daniel Guettler, Misha Kossodo, Misha Pivovarov, and Ralph Weissleder., "PepBank: a database of peptides based on sequence text mining and public peptide data sources." BMC Bioinformatics 8, no. 1 (2007): 280.

6. Appendix

6.1. Tables and Figures

Table 2: Test Results of Heuristic Classifier and Machine Learning Classifiers - performance metrics for specific adversarial types within the test set.

	Heuristic Classifier	**NLP Classifier**	**1-D CNN Classifier**
Overall Test Performance	**Accuracy : 0.8685** **Precision: 0.9776** **Recall : 0.7543** **F1 Score : 0.8515**	**Accuracy : 0.9629** **Precision: 0.9890** **Recall : 0.9362** **F1 Score : 0.9619**	**Accuracy : 0.8452** **Precision: 0.845** **Recall : 0.8448** **F1 Score : 0.8451**
1. Add english words	Accuracy : 1.0 Precision: 1.0 Recall : 1.0 F1 Score : 1.0	Accuracy : 0.9833 Precision: 1.0 Recall : 0.9750 F1 Score : 0.9873	Accuracy : 0.775 Precision: 0.7676 Recall : 0.95 F1 Score : 0.8491
2. Add random text	Accuracy : 1.0 Precision: 1.0 Recall : 1.0 F1 Score : 1.0	Accuracy : 0.9425 Precision: 1.0 Recall : 0.9425 F1 Score : 0.9703	Accuracy : 0.5725 Precision: 1.0 Recall : 0.5725 F1 Score : 0.7281
3. Alternate random characters + random text	Accuracy : 0.6667 Precision: 1.0 Recall : 0.5	Accuracy : 1.0 Precision: 1.0 Recall : 1.0	Accuracy : 0.85 Precision: 0.8229 Recall : 0.9875

	F1 Score : 0.6667	F1 Score : 1.0	F1 Score : 0.8977
4. Compressed sequence + random text	Accuracy : 0.5062 Precision: 1.0 Recall : 0.3416 F1 Score : 0.5093	Accuracy : 1.0 Precision: 1.0 Recall : 1.0 F1 Score : 1.0	Accuracy : 0.8937 Precision: 0.8872 Recall : 0.9833 F1 Score : 0.9328
5. Compressed, replaced sequence + random text	Accuracy : 0.5125 Precision: 1.0 Recall : 0.35 F1 Score : 0.5185	Accuracy : 0.8437 Precision: 1.0 Recall : 0.7916 F1 Score : 0.8837	Accuracy : 0.85 Precision: 0.8380 Recall : 0.9916 F1 Score : 0.9083
6. Replaced sequence + random text	Accuracy : 0.675 Precision: 1.0 Recall : 0.5125 F1 Score : 0.6776	Accuracy : 0.9166 Precision: 1.0 Recall : 0.8750 F1 Score : 0.9333	Accuracy : 0.8 Precision: 0.7692 Recall : 1.0 F1 Score : 0.8695
7. Fake gene sequences (Negatives)	Accuracy : 0.0	Accuracy : 0.6	Accuracy : 0.05
8. Replacement with multi-character mapping	Accuracy : 0.6666 Precision: 1.0 Recall : 0.5 F1 Score : 0.6666	Accuracy : 0.925 Precision: 1.0 Recall : 0.8875 F1 Score : 0.9403	Accuracy : 0.85 Precision: 0.8229 Recall : 0.9875 F1 Score : 0.8977
9. Random spaces + Random text	Accuracy : 1.0 Precision: 1.0 Recall : 1.0 F1 Score : 1.0	Accuracy : 1.0 Precision: 1.0 Recall : 1.0 F1 Score : 1.0	Accuracy : 0.90625 Precision: 0.8889 Recall : 1.0 F1 Score : 0.9411
10. Random text breaks of variable length	Accuracy : 0.925 Precision: 1.0 Recall : 0.8875 F1 Score : 0.9403	Accuracy : 0.9583 Precision: 1.0 Recall : 0.9375 F1 Score : 0.9677	Accuracy : 0.7833 Precision: 0.7547 Recall : 1.0 F1 Score : 0.8602
11. Randomly generated text (Negatives)	Accuracy : 1.0	Accuracy : 0.9951	Accuracy : 1.0

Table 2 – Performance on Test Data

	Heuristic Classifier	**NLP Classifier**	**1-D CNN Classifier**
1. Columnar Split	Accuracy: 1.0	Accuracy: 0.9	Accuracy: 1.0
2. Unicode Homoglyph Substitution	Accuracy: 1.0	Accuracy: 1.0	Accuracy: 1.0

3. Log/JSON Format	Accuracy: 0.0	Accuracy: 1.0	Accuracy: 1.0
4. Zero width character Injection	Accuracy: 1.0	Accuracy: 1.0	Accuracy: 1.0
5. Compressed Sequence	Accuracy: 0.0	Accuracy: 1.0	Accuracy: 1.0
6. Alternate Random Character	Accuracy: 0.4	Accuracy: 0.3	Accuracy: 0.8
7. Linguistic Camouflage	Accuracy: 0.1	Accuracy: 0.1	Accuracy: 1.0
8. Base64 Encoding	Accuracy: 0.0	Accuracy: 0.1	Accuracy: 1.0
9. Multicharacter Mapping	Accuracy: 0.0	Accuracy: 0.7	Accuracy: 0.8
11. Reverse Complement	Accuracy: 1.0	Accuracy: 1.0	Accuracy: 1.0

Table 3 – Performance on 'Advanced' test data

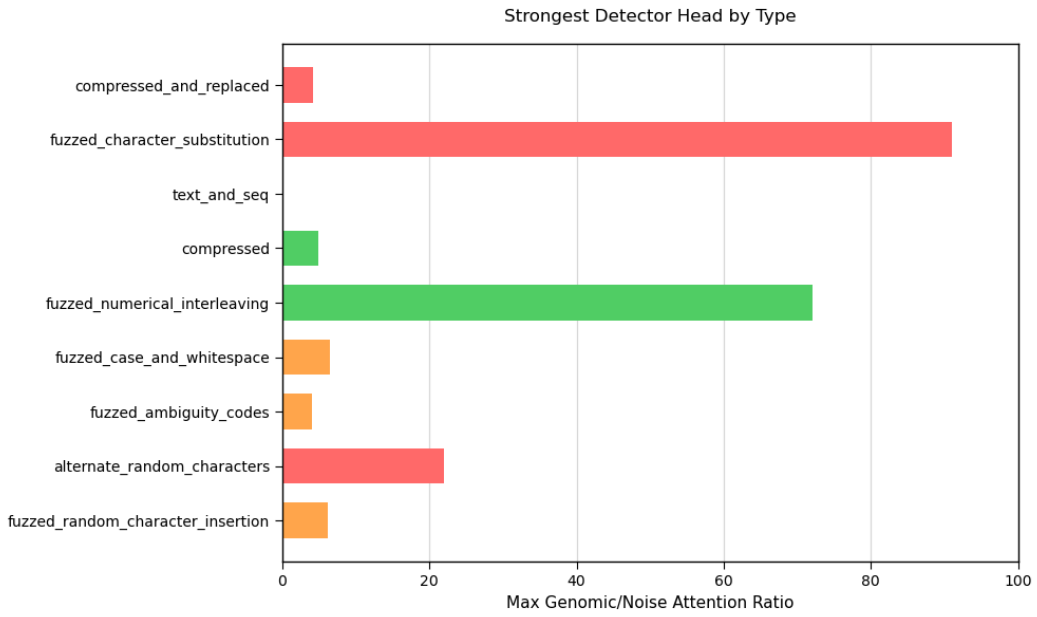

Figure 1 – Attention Ratio for different types of masking

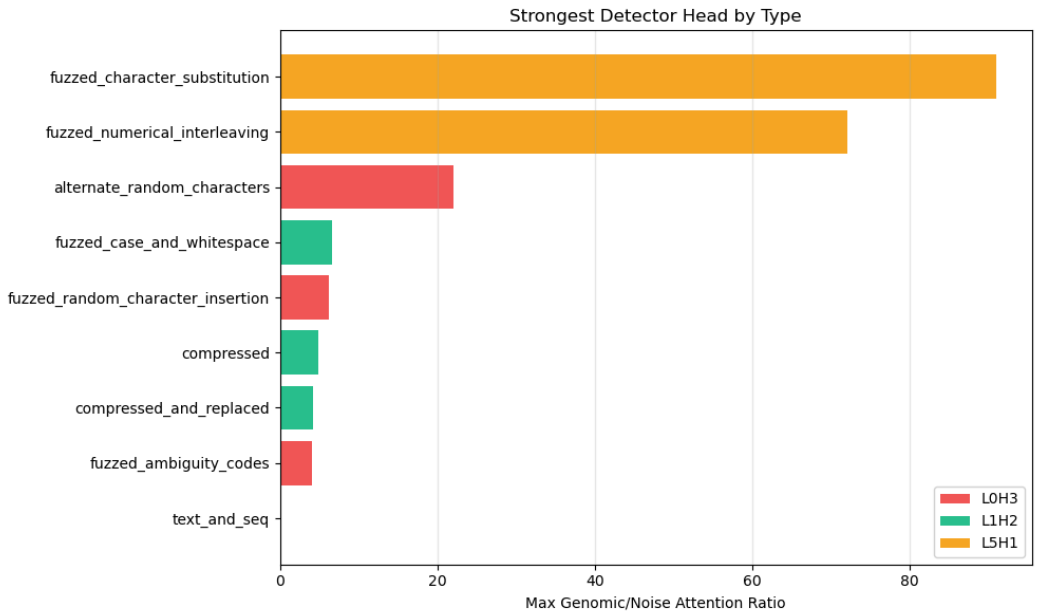

Figure 2 – Attention Ratio based for different attention heads and masking types

6.2. Data Acquisition

For detailed information regarding our data acquisition and adversarial sample generation pipeline, please see the documentation on our GitHub repository: https://github.com/rockysaikia730/GeneGuard/blob/main/data/DATASET.md.

6.3. Security Considerations

A. Evasion & Obfuscation
   - Adversaries could bypass detection by encoding DNA in alternate formats (e.g., amino acids, QR codes, or encryption).
   - Gene-Guard does not detect encrypted or heavily obfuscated content.
   - Mitigation: Add OCR/image scanning, file signature detection, insider behavior monitoring.

B. False Positives & Negatives
   - Small chance of false positives (benign text flagged) or false negatives (short/obscured sequences missed).
   - Mitigation: Tuned for high precision, shows contextual feedback, continuous model updates using new samples.

C. Privacy & Data Handling
   - Logs contain sensitive data and must be access-controlled.
   - LLM-based analysis must avoid sending raw DNA to third-party APIs.
   - By default: local analysis only, logs mask sequence data.

D. Scope of Application
    - Specialized for genomic data only.
    - Not suitable for other sensitive assets (Future iterations can and should include e.g., protein structures, clinical trials).
    - Best deployed in genomics-rich environments (labs, biotech, hospitals).

6.4. Limitations and Future Work:

- Current detection does not cover images or encrypted payloads.
- The machine learning classifier trained on obfuscated genomic data fails on extreme cases of obfuscation. Addressing this might require a larger pool of synthetically generated adversarial sequences (potentially leveraging GAN models) to make the classifier more robust.
- Expanding to RNA/protein detection is a future goal.
- We also aim to improve LLM interpretability for novel sequences and offer Gene-Guard as a plugin for major cloud and productivity platforms.
- Attention-based genomic monitoring systems might be vulnerable to cyber attacks targeting their weights.