Protein Embedding-Based Detection of Sequence-Diverse Biosecurity Threats

Polina Shevyakova[1], Henry Ward, Elina Shaniiazova[2], Marie Krátká[3]
[1]Higher School of Economics, Faculty of Biology, [2]Constructor University Bremen, [3]Masaryk University

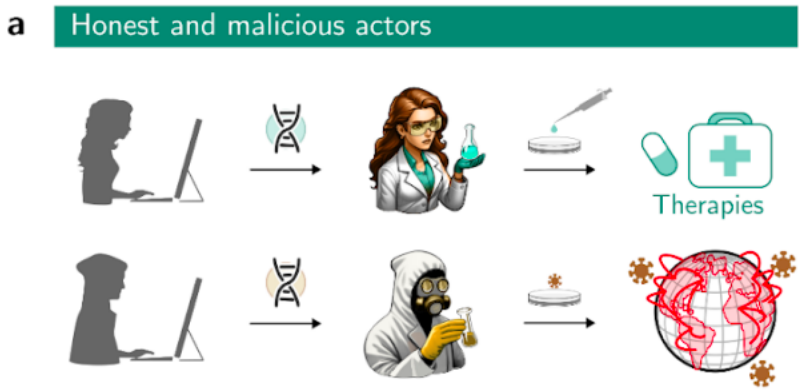
Abstract

Current DNA synthesis screening systems rely primarily on sequence homology to detect biosecurity threats, creating potential vulnerabilities to sophisticated evasion strategies. We developed a complementary protein embedding-based screening approach using ESM2 to detect functionally similar but sequence-diverse threats. Using ProteinMPNN, we generated toxin variants with below 60% sequence identity to known toxins while preserving 3D structure. Our key finding is that these sequence-diverse variants cluster significantly closer to original toxin families than to neutral proteins in ESM2 embedding space, suggesting that functional relationships are more conserved in embedding space than sequence space. This demonstrates the potential for embedding-based screening to identify threats that evade traditional homology-based detection. We built a two-layer screening pipeline combining SecureDNA with ESM2 similarity analysis and conducted preliminary evaluation on the NIST nucleic acid synthesis screening dataset. Our results indicate that multi-modal screening approaches could provide more robust biosecurity coverage against advanced evasion attempts.

1. Introduction

The relative cost and time of human-assisted screening is rising with data volume. International Gene Synthesis Consortium (IGSC) voluntarily monitor approximately 80% of global DNA synthesis without covering the whole data volume in real-time. Primary goals for every biosecurity screening are based on preventing pathogens production (pandemic viruses) and allowing therapy production (cancer immunotherapies, vaccines), even when they employ the same technology (Figure 1).

Figure 1: Reproduced from Baum, C., Berlips, J., Chen, W., Cozzarini, H., Cui, H., Damgård, I., ... & Zhang, K. (2026). A system capable of verifiably and privately screening global DNA synthesis. National Science Review, nwag103.



The rapid advancement of synthetic biology and DNA synthesis technologies has democratized access to genetic engineering capabilities, creating new opportunities for both beneficial applications and potential misuse. Current biosecurity screening systems for DNA synthesis primarily rely on sequence homology methods, such as k-mer matching, to identify sequences with similarity to known biological threats including toxins, virulence factors, and regulated pathogens.

However, this approach may be vulnerable to sophisticated evasion strategies. Malicious actors with knowledge of protein structure-function relationships could potentially generate functionally dangerous proteins that share minimal sequence similarity with known threats, thereby evading sequence-based screening systems. Recent advances in protein design tools, particularly ProteinMPNN, enable the generation of novel amino acid sequences that fold into predefined three-dimensional structures, raising concerns about the adequacy of current screening approaches.

To address this potential vulnerability, we investigated whether protein embeddings could provide a complementary detection mechanism for sequence-diverse but functionally similar threats. Protein language models like ESM2 capture deep functional and structural relationships that extend beyond sequence similarity, potentially enabling detection of threats that would evade traditional homology-based screening.

Our main contributions are:

1. Demonstrated that ProteinMPNN-generated toxin variants with low-to-moderate sequence identity (0-60%) to originals retain strong embedding similarity to toxin families rather than neutral proteins.

2. Developed a two-layer biosecurity screening pipeline combining homology screening (SecureDNA, gold standard)with ESM2 protein embedding analysis.

3. Provided proof-of-concept evidence that embedding-based screening could detect threats missed by sequence homology methods.

2. Related Work

Current DNA synthesis screening approaches rely heavily on sequence database matching. SecureDNA represents the state-of-the-art in this domain, using k-mer

hashing to identify sequences with significant homology to regulated biological agents. While effective for known threats and close variants, this approach has inherent limitations when facing novel sequences designed to evade detection.

Protein language models have emerged as powerful tools for capturing functional relationships beyond sequence similarity. ESM2, developed by Meta, was trained on millions of protein sequences and has demonstrated remarkable ability to capture protein structure, function, and evolutionary relationships through learned embeddings. These embeddings have been successfully applied to protein function prediction, structural analysis, and evolutionary studies.

ProteinMPNN has revolutionized protein design by enabling the generation of novel sequences that fold into specified three-dimensional structures. This capability raises important questions about the relationship between sequence diversity and functional conservation, particularly in the context of biosecurity threats.

Our work addresses the gap between traditional sequence-based screening and the emerging reality of sophisticated protein design capabilities. We demonstrate that embedding-based analysis can provide a complementary detection mechanism that operates in functional space rather than sequence space.

3. Methods

Project code repository is available at: https://github.com/437364/esm2_screening

3.1 ProteinMPNN Variant Generation

We selected 121 toxin structures from the Protein Data Bank and 6 manually selected critical proteins. Using ProteinMPNN v48_020 with a sampling temperature of 0.25, we generated multiple sequence variants per target toxin. We filtered variants to retain only those with threshold sequence identity to the original toxin. Part of generated sequences was excluded due to high internal similarity. Among the manually fetched toxins, SARs-CoV-2 and Sindbis viruses were excluded, since long ProteinMPNN generated aminoacid sequences (>2000 amino acids) contained large gaps.

3.2 ESM2 Embedding Analysis

We computed protein embeddings using the ESM2-650M model (facebook/esm2_t33_650M_UR50D). For each protein sequence, we extracted token-level embeddings and computed the mean-pooled representation. To address the isotropic bias in ESM2 embeddings, we subtracted the global reference mean computed over both toxin and background databases, then L2-normalized the resulting vectors.

3.3 Database Construction

We constructed comprehensive protein databases by querying UniProt Swiss-Prot. The toxin database comprised 7,000+ reviewed proteins with toxin-related keywords (KW-0800). The neutral database contained 25,000+ reviewed proteins excluding toxin,

virulence factor, and antimicrobial peptide annotations. Each protein sequence was processed through ESM2 to generate embeddings for similarity analysis.

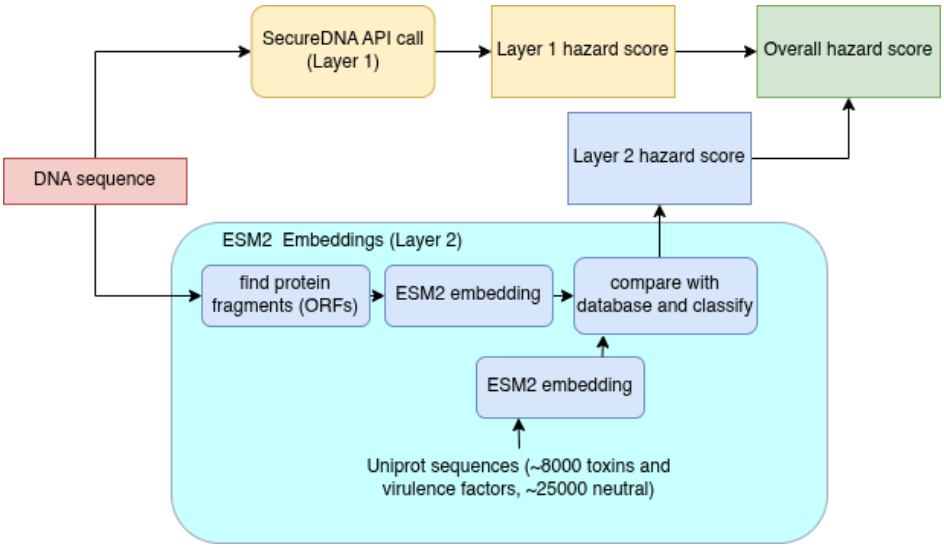
3.4 Two-Layer Screening Pipeline

Our screening pipeline consists of two complementary layers (see Figure 2):

Layer 1 (SecureDNA): Traditional k-mer homology screening using the SecureDNA synthclient. Protein sequences are backtranslated to DNA using a simple codon table and submitted for regulatory screening.

Layer 2 (ESM2 Embeddings): Protein similarity screening using cosine similarity in normalized ESM2 embedding space. We compute similarity scores to both toxin and neutral databases, generating a differential score (S_toxin - S_neutral) of the best match to classify threat potential.

Figure 2: Diagram of the screening pipeline.



3.5 Multiple ORF Detection

To handle real-world DNA constructs, we implemented comprehensive open reading frame (ORF) detection. The system identifies all potential protein-coding regions in the screened DNA sequence and evaluates each for biosecurity threats, addressing scenarios where multiple coding regions or non-standard start positions may be present.

4. Results

4.1 ProteinMPNN Variant Generation Success

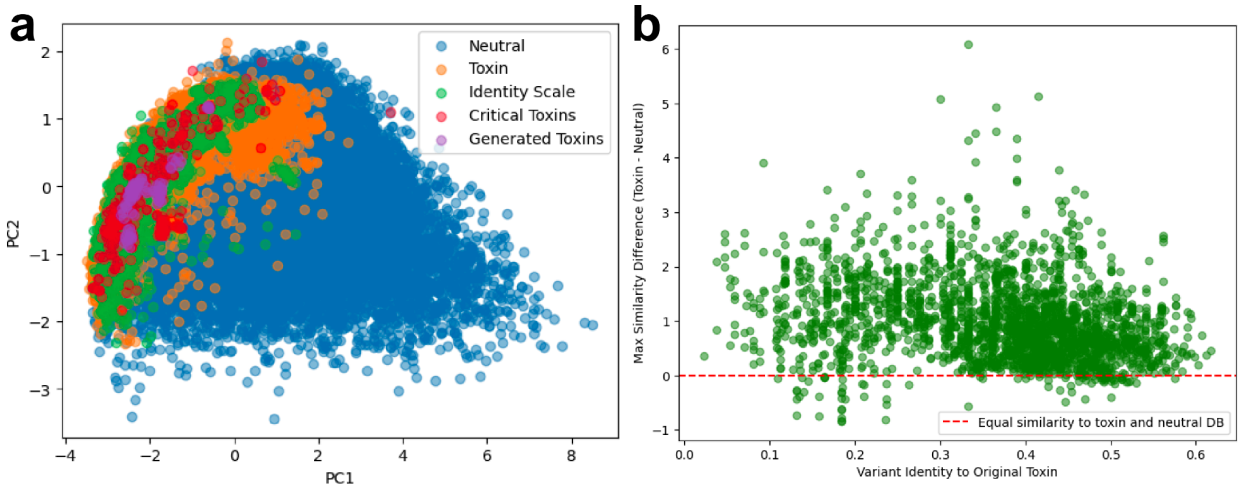
ProteinMPNN successfully generated sequence-diverse variants for all target toxins, resulting in variants with 0-60% sequence identity while preserving the original three-dimensional structure. While we cannot assess whether these variants retain their toxic and/or virulent potential, this demonstrates the ability to create dramatically different amino acid sequences while preserving the original three-dimensional fold.

4.2 Embedding Space Analysis

Using PCA analysis of the embedding space, we observed clear clustering patterns where sequence-diverse variants maintained proximity to toxin records (see Figure 3). This provides strong evidence that ESM2 embeddings capture functional signatures that persist even under significant sequence modification. Newly generated sequence-diverse variants were classified based on cosine similarity to database sequences. 2361 out of 2440 have their top similarity hit in the toxin database. Most of the misclassified sequences had the overall similarity to the original toxin lower than 30%, suggesting that the embedding similarity does somewhat decrease with the level of diversification.

Figure 3: Embedding similarity of sequence-diverse variants. a) PCA of mean-pooled ESM2 embeddings of neutral, toxic, and newly generated sequence-diverse toxin variants. New sequences (green) cluster with toxins (orange).

Moreover, selected sequences for critical toxins (not placed in GenBank database) (Influenza A, Zika, Vaccinia viruses (233 seq.): red), where Influenza A and Zika were carefully fetched with existing PDB structures with the highest or identical similarity. Sequence-diverse variants for toxins were generated (purple). b) Classification of sequence diverse-variants based on max similarity hit in database embeddings. Values > 0 show higher similarity to the toxin database. Sequences with lower identity to original toxin are more dispersed.



4.3 Screening Pipeline Performance

Our two-layer pipeline successfully processed test sequences through both screening layers. The embedding-based Layer 2 demonstrated ability to flag ProteinMPNN variants that showed low sequence similarity to known threats — a capability with direct practical relevance given the limitations of homology-based screening documented below.

4.4 SecureDNA Benchmark Evaluation

To contextualize our approach, we evaluated SecureDNA against two datasets. On the NIST nucleic acid synthesis screening dataset, SecureDNA achieved TP=88, FP=155, TN=470, FN=273 (Precision=0.362, Recall=0.244, F1=0.291, FNR=0.756, FPR=0.248). These results partly reflect framework misalignment — SecureDNA's regulatory definitions differ from NIST's research-oriented threat categorization, and many apparent false positives corresponded to regulated-but-passable sequences (*Coxiella burnetii*, *Burkholderia pseudomallei*) that NIST labeled as safe.

More critically, evaluation on redesigned toxin sequences revealed a fundamental limitation: while SecureDNA correctly flagged all 11 unaltered wild-type toxins, it granted synthesis permission to every redesigned variant tested, including all five ProteinMPNN-redesigned influenza sequences and the vaccinia redesigns (Table X). This complete evasion — 0% detection of altered sequences — demonstrates that current homology-based screening can be bypassed by sequence-level redesign while preserving function, precisely the threat scenario our embedding-based Layer 2 is designed to address.

5. Discussion and Limitations

Our findings demonstrate that protein function, as captured by ESM2 embeddings, appears conserved even in structure-conserved proteins with low-to-moderate sequence identity. This has important implications for biosecurity screening, suggesting that embedding-based approaches could provide valuable complementary detection capabilities for threats designed to evade sequence homology screening.

The clustering of sequence-diverse ProteinMPNN variants with their functional families rather than with neutral proteins provides strong evidence for the utility of embedding-based threat detection. This approach could be particularly valuable for identifying sophisticated evasion attempts where adversaries have deliberately minimized sequence similarity to known threats while preserving functional properties.

Limitations

Several important limitations constrain our current results. The embedding-based classifier is currently based on simple cosine similarity calculations. Machine learning based approach could provide a more accurate detection threshold. Our pipeline testing was only preliminary, focused on a small set of well-characterized toxins, and broader validation across diverse threat categories is urgently needed.

The multiple ORF handling system was developed but not fully integrated into the main screening pipeline, missing threats in complex DNA constructs.

Incomplete protein fragments could evade detection by skewing mean-pooled ESM2 embeddings. Attention pooling of embeddings could significantly improve performance.

Additionally, our DNA backtranslation approach uses simplified codon tables rather than organism-specific optimization, which may not reflect realistic synthesis scenarios and could affect comparative performance evaluations with existing screening systems.

Future Work

Priority developments include completing the embedding-based classifier training with attention pooling mechanisms, which may significantly improve performance over simple mean pooling combined with cosine similarity. Full integration of multiple ORF detection into the main pipeline is needed to handle realistic DNA synthesis scenarios. Systematic evaluation of ProteinMPNN variants against existing screening systems would provide definitive evidence of detection gaps.

Expanding the approach to broader threat categories beyond toxins, including virulence factors, antimicrobial resistance proteins, and other biosecurity-relevant protein families, would establish the general utility of embedding-based screening. Implementation of codon optimization for realistic organism-specific synthesis contexts would improve the fidelity of comparative evaluations.

References

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., ... & Baker, D. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. Science, 378(6615), 49-56.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ... & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637), 1123-1130.

SecureDNA. (2024). Safeguarding DNA synthesis with digital signatures and similarity detection. Retrieved from https://securedna.org/

NIST. (2025). Nucleic acid synthesis screening test dataset. National Institute of Standards and Technology. https://data.nist.gov/od/ds/mds2-3787/

Appendix: Limitations and Dual-Use Considerations

Limitations

Our embedding-based classification approach showed limitations in training convergence, with the model failing to achieve reliable prediction performance on our validation set. The differential scoring approach (S_toxin - S_neutral) provides a reasonable fallback, but optimal thresholds remain to be determined through systematic validation.

The scope of our evaluation was constrained by computational resources and time limitations, focusing on a narrow set of well-characterized toxin families. Real-world biosecurity threats encompass a much broader range of protein families and functional categories that require systematic evaluation.

False positive and false negative rates for the embedding-based approach have not been thoroughly characterized across diverse protein families. Edge cases, such as proteins with dual beneficial/harmful functions or proteins that cluster with toxins due to structural rather than functional similarity, require careful analysis.

Scalability constraints limit the immediate deployment of this approach, as ESM2 embedding computation is computationally intensive compared to k-mer matching. Real-time screening applications would require significant optimization or specialized hardware.

Dual-Use Risks and Considerations

Our work has a primarily defensive focus, aiming to identify and address potential vulnerabilities in existing biosecurity screening approaches. However, several dual-use considerations require careful attention:

Novel sequence risks: Our ProteinMPNN-generated variants represent novel protein sequences that theoretically could retain toxic or virulent properties if three-dimensional structural similarity translates to functional similarity. However, significant experimental work would be required to determine whether our computationally-generated sequences actually possess harmful biological activity. Structural preservation does not guarantee functional preservation, and many factors beyond gross structure (active site geometry, allosteric regulation, stability, expression levels) influence protein function.

Functional uncertainty: We have not conducted any functional validation of generated sequences and strongly recommend extreme caution if others attempt such validation. The gap between computational design and biological activity is substantial, requiring extensive wet-laboratory characterization including expression testing, purification, structural validation, and functional assays.

Responsible disclosure: We have shared our findings with the biosecurity research community to improve defensive capabilities rather than enable harmful applications. Our results demonstrate the need for multi-modal screening approaches that can detect both sequence- and structure-based threats.

Research gap highlighted: Our work reveals a potentially significant gap in current screening approaches that could be exploited by sophisticated adversaries. However, we believe that identifying and characterizing this gap is essential for developing robust defensive countermeasures. We believe that developing embedding-based screening as a complement to existing homology-based systems will enhance the overall security of the DNA synthesis ecosystem. The goal is to stay ahead of potential threats through proactive security research rather than reactive responses to actual misuse events.

Attributions and Acknowledgements

Polina led overall project planning and coordination. She implemented the SecureDNA evaluation pipeline, including API calls and metric calculations, curated UniProt datasets for protein embedding analysis, and performed ESM2 embedding computations and associated mathematical analyses. She also explored and evaluated an embedding-based classifier approach. Elina curated and hand-selected input datasets for ProteinMPNN, filtered and sanity-checked the model outputs. Henry performed ProteinMPNN API calls and sequence generation. Marie conducted PCA analysis of ESM2 embeddings, implemented the ORF-finding function, produced project visualizations, and integrated data into the final report submission. Portions of the code and written text were drafted with the assistance of Claude (Anthropic), an AI language model.