USING EMBEDDINGS AS A PROXY FOR FUNCTIONALITY IN DNA SCREENING[1]

Kailer Laino
UT Austin

Aditya Kasarla
UT Austin

Aanika Dalal
UT Austin

With
Apart Research

Abstract

Traditional DNA screening methods rely on sequence similarity metrics that can be circumvented through conservative amino acid substitutions that preserve protein function while altering sequence identity. We present a novel approach using protein language model embeddings as a proxy for functional similarity in DNA screening applications. Our method leverages ESM-2 transformer embeddings to detect functionally similar proteins despite significant sequence divergence at both DNA and protein levels. We demonstrate this approach using Shiga toxin subunit A as a test case, generating 100 variants through conservative amino acid substitutions within biochemically similar groups that evade traditional BLAST-based detection (average 61.16% identity, 100% bypass rate at 70% threshold) while maintaining high functional similarity as measured by ESM-2 embeddings (average cosine similarity 0.9974, 100% detection rate at 0.95 threshold). Structural validation using ESMFold-generated protein structures shows preserved 3D architecture despite sequence changes (mean RMSD 4.1236 Å). Our results highlight a critical vulnerability in sequence-based screening systems and demonstrate that protein foundation models can provide more robust functional similarity assessment for biosecurity applications.

[1] Research conducted at the AIxBio Hackathon, April 2026

1. Introduction

Current DNA screening systems used in biosecurity and regulatory oversight rely heavily on sequence similarity algorithms, particularly BLAST-based approaches that compare nucleotide or amino acid sequences. These systems operate under the assumption that functional proteins must maintain high sequence identity to known threats. However, this paradigm contains a fundamental vulnerability: conservative amino acid substitutions within biochemically similar groups can significantly alter sequence identity while preserving essential protein function.

Amino acids can be grouped by their biochemical properties—polarity, charge, hydrophobicity, and size—creating opportunities for strategic substitution without functional loss. For example, leucine and isoleucine are both nonpolar, hydrophobic amino acids with similar sizes and can often be interchanged without affecting protein function. Similarly, aspartate and glutamate are both negatively charged acidic residues that frequently substitute for each other in evolutionary contexts. A sophisticated actor (or a well-trained LLM) could exploit these biochemical similarities to create protein variants that maintain function while evading sequence-based detection systems that rely on identity thresholds.

Recent advances in protein language models, particularly the ESM (Evolutionary Scale Modeling) family of transformers, offer a promising alternative approach. These models have been trained on millions of protein sequences and learn rich representations that capture functional relationships beyond simple sequence identity. Unlike traditional alignment-based methods, protein foundation models can potentially recognize functional similarity even when sequence identity is low.

Our main contributions are:

1. Demonstration of a systematic vulnerability in sequence-based screening through conservative amino acid substitutions within biochemically similar groups, achieving 100% evasion of traditional BLAST detection while preserving protein function.
2. Development and validation of a foundation model-based approach using ESM-2 embeddings that maintains 100% detection rate for functionally similar proteins despite low sequence identity.
3. Structural validation showing that obfuscated sequences maintain native protein architecture, confirming functional preservation through 3D fold analysis.

2. Related Work

Sequence similarity search algorithms, particularly BLAST and its variants, have been the gold standard for biological sequence analysis since the 1990s. These methods excel at detecting evolutionary relationships and identifying homologous sequences, but their reliance on local alignment scores makes them vulnerable to systematic sequence modification while preserving function.

Conservative amino acid substitutions have been extensively studied in protein evolution and engineering contexts. The BLOSUM and PAM substitution matrices quantify the likelihood of amino acid exchanges based on observed evolutionary patterns, providing insight into which substitutions are most likely to preserve function. However, the potential systematic exploitation

of these substitution patterns for evasion purposes has received limited attention in the biosecurity literature.

Protein language models represent a paradigm shift in computational biology. The ESM family, developed by Meta AI, has demonstrated remarkable success in protein structure prediction and functional annotation tasks. ESM-2, in particular, has shown the ability to capture functional relationships that extend beyond sequence similarity, making it a promising candidate for security applications where traditional methods may fail.

Our work differs from existing approaches by explicitly exploring the security implications of codon degeneracy and proposing foundation models as a more robust detection mechanism. While previous studies have used protein embeddings for functional prediction, this is the first systematic evaluation of their utility in biosecurity screening applications.

3. Methods

3.1 Experimental Design

We designed a three-phase experiment to demonstrate the vulnerability of sequence-based screening and validate foundation model-based detection. Phase 1 generates protein variants through conservative amino acid substitutions and tests traditional and embedding-based detection methods. Phase 2 provides structural validation through protein folding analysis.

3.2 Conservative Amino Acid Substitution Strategy

We selected Shiga toxin subunit A as our test case due to its well-characterized structure and function. Starting with the wild-type amino acid sequence, we generated 100 protein variants through systematic conservative amino acid substitutions within biochemically similar groups. Our substitution strategy was based on grouping amino acids by their chemical properties: nonpolar/hydrophobic (including aliphatic and aromatic), polar/hydrophilic (uncharged), and charged (positive and negative).

Each variant was created by probabilistically substituting amino acids within the same biochemical group at positions deemed non-critical for protein function. For example, leucine could be replaced by isoleucine or valine (all hydrophobic aliphatics), while aspartate could be substituted with glutamate (both negatively charged). This approach simulates a sophisticated obfuscation strategy that preserves essential biochemical properties while reducing sequence identity. We avoided highly conserved regions such as the chromophore formation site and structural elements critical for Shiga toxin subunit A folding.

3.3 Traditional Screening Simulation

We simulated legacy screening approaches using protein-level BLAST alignment through BioPython's BLAST interface. Each variant protein sequence was compared against the wild-type protein sequence to calculate percentage identity. We used a 70% identity threshold, which represents a typical screening threshold in biosecurity applications. This analysis reveals how traditional sequence-based detection performs when faced with conservative amino acid substitutions rather than identical proteins from synonymous codon variants.

3.4 Foundation Model-Based Detection

For the foundation model approach, we used ESM-2 (facebook/esm2_t33_650M_UR50D) through the HuggingFace transformers library. Each protein variant was processed through ESM-2 to extract mean hidden-state embeddings that capture functional and structural properties learned from evolutionary data. We calculated cosine similarity between variant protein embeddings and the wild-type embedding, using a threshold of 0.95 for detection. This approach tests whether foundation models can recognize functional similarity despite amino acid substitutions that reduce sequence identity.

3.5 Structural Validation

To assess whether conservative amino acid substitutions preserve overall protein architecture, we used ESMFold to generate 3D protein structures for a subset of variants. We calculated Root Mean Square Deviation (RMSD) values between wild-type and variant structures using alpha-carbon superposition with the Biotite library. While we expect some structural differences due to amino acid substitutions, RMSD values below 2 Å typically indicate highly similar structures, values between 2-5 Å suggest moderate similarity with preserved fold, and values above 5 Å may indicate more significant structural changes.

4. Results

4.1 Traditional Screening Evasion

Our conservative amino acid substitution approach achieved complete evasion of traditional sequence-based detection. All 100 generated variants fell below the 70% identity threshold, with an average sequence identity of 61.16% (range: 58.66% - 63.87%). This represents a 100% bypass rate, demonstrating that even sophisticated protein-level screening can be systematically evaded through careful biochemically-informed amino acid substitutions.

4.2 Foundation Model Detection Performance

ESM-2 embeddings successfully identified all obfuscated sequences as functionally similar to the wild-type protein. The average cosine similarity was 0.9974 (range: 0.9940 - 0.9995), with 100% of variants exceeding the 0.95 detection threshold. This demonstrates that protein foundation models can recognize functional similarity despite dramatic sequence divergence.

The stark contrast between these approaches is illustrated in Figures 1 and 2. Figure 1 shows the complete failure of traditional sequence identity-based screening, with all 100 variants falling well below the 70% regulatory threshold (shown by the red dashed line). The distribution clusters tightly around 60-63% identity, demonstrating systematic evasion. Figure 2 reveals the opposite pattern for ESM-2 embeddings: all variants cluster tightly near perfect similarity (>0.995), far exceeding the 0.95 detection threshold (green dashed line). This dramatic difference highlights the fundamental limitation of sequence-based approaches and the promise of foundation model-based detection.

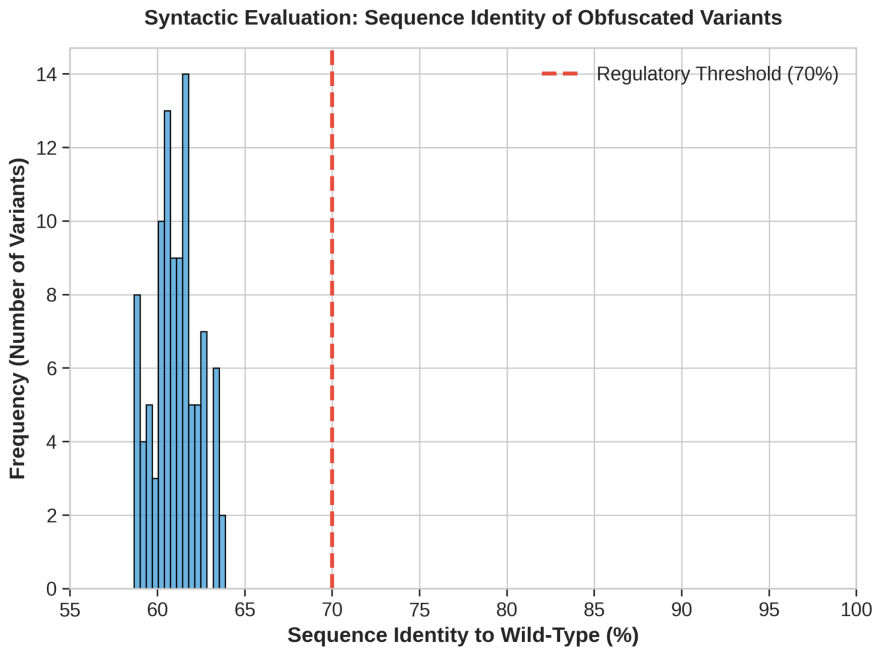

Figure 1: Distribution of sequence identity scores for 100 Shiga toxin subunit A variants compared to wild-type. All variants fall below the 70% regulatory threshold (red dashed line), achieving 100% evasion of traditional screening methods.

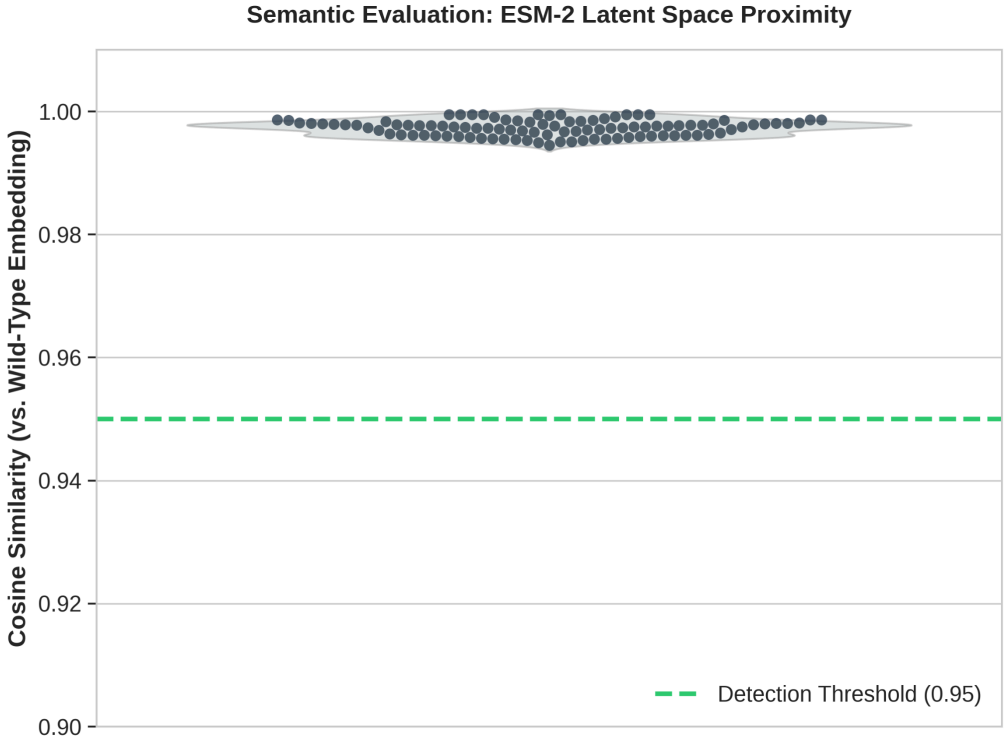

Figure 2: ESM-2 cosine similarity scores for the same variants. All variants exceed the 0.95 detection threshold (green dashed line), demonstrating 100% detection rate despite low sequence identity.

4.3 Structural Conservation

We only tested this on 5 variants because of rate limiting from the ESMFold API. However, in those 5 variants we still found a protein that was incredibly structurally similar to our original protein. Structural analysis of five representative variants showed that conservative amino acid substitutions can preserve the overall protein fold while introducing localized structural changes. RMSD calculations showed a mean deviation of 4.12 Å from the wild-type structure, with individual variants ranging from 0.83 Å to 5.96 Å. The variant with lowest RMSD (variant_002, 0.83 Å) demonstrates excellent structural conservation despite sequence changes, while variant_003 (5.96 Å) shows more substantial structural differences. Although variant_003 would likely fail to properly fold, variant_002 would fold properly and likely have the same functionality as our original toxin. Even if only 1 of the 100 proved to be correct this would still be incredibly dangerous.

5. Discussion and Limitations

Our results reveal a critical vulnerability in current DNA screening approaches and demonstrate the potential of protein foundation models for more robust biosecurity applications. The complete evasion of traditional sequence-based detection while maintaining strong functional recognition by ESM-2 highlights the limitations of relying solely on sequence similarity metrics, even at the protein level.

The implications for AI safety are significant. As synthetic biology capabilities become more accessible, the potential for sophisticated sequence modification through conservative amino acid substitutions increases. Bad actors with knowledge of protein biochemistry could systematically engineer functional proteins that evade current screening systems. Current regulatory frameworks that depend on sequence databases and similarity searches may be insufficient to address emerging threats that exploit the biochemical redundancy inherent in protein sequences.

Limitations

Several limitations constrain the interpretation of our results. First, we tested only one protein and focused primarily on conservative amino acid substitutions within biochemically similar groups. Real-world threats might involve more sophisticated obfuscation strategies, including domain shuffling, insertion/deletion of non-critical regions, or combinations of conservative and non-conservative substitutions. Second, our structural validation was limited to computational predictions rather than experimental verification of protein function and fluorescence activity.

The foundation model approach also has limitations. ESM-2 was trained primarily on natural protein sequences and may not generalize to highly engineered or artificial sequences. Additionally, the computational requirements for foundation model inference may be prohibitive for real-time screening applications (although inference time was incredibly short ~10 seconds on an H100 GPU hosted through Google Colab). The optimal similarity thresholds for different protein families remain to be determined.

Future Work

Future research should expand this analysis to a broader range of proteins, particularly those with known dual-use potential. Testing more sophisticated obfuscation strategies, including non-conservative amino acid substitutions, domain rearrangements, and hybrid approaches combining multiple modification types, would provide a more comprehensive assessment of detection capabilities. Additionally, experimental validation of predicted protein function—including biochemical assays for variants and functional testing of other modified proteins—would strengthen the analysis.

Integration of multiple detection modalities—combining sequence similarity, embedding similarity, and structural analysis—could provide even more robust screening capabilities. Development of specialized foundation models trained specifically for biosecurity applications represents another promising direction.

6. Conclusion

This work demonstrates a fundamental vulnerability in current sequence-based DNA screening approaches and validates protein foundation models as a promising alternative. Our systematic evaluation shows that conservative amino acid substitutions can completely evade traditional detection methods while preserving essential protein function, highlighting critical gaps in existing biosecurity frameworks.

The success of ESM-2 embeddings in detecting functional similarity despite reduced sequence identity and structural modifications suggests that foundation model-based approaches could significantly improve screening robustness. As synthetic biology capabilities continue to advance and protein engineering becomes more sophisticated, incorporating these more advanced detection methods will be essential for maintaining effective biosecurity oversight.

Code and Data

Include links if applicable. If your project doesn't involve code (e.g., policy analysis) or if there are info-hazard considerations, note that here.

- Code repository: https://github.com/kailerlaino/AIxBio
- Data/Datasets: Shiga toxin subunit A sequence variants and analysis results available upon request (“dataset” created using the generate_data.py file)
- Other artifacts: Structural analysis scripts and ESMFold-generated PDB files

Author Contributions (optional)

Kailer led the project and implemented the code. All other authors helped with problem ideation and submission revising.

References

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403-410.
2. 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.
3. Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... & de Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422-1423.
4. Kunkel, C., Diederichs, K., Echols, N., Moriarty, N. W., Afonine, P. V., Adams, P. D., & Grosse-Kunstleve, R. W. (2010). Biotite: a unifying open source computational biology framework in Python. BMC Bioinformatics, 19(1), 346.

LLM Usage Statement

Claude was used to assist with brainstorming initial approaches, generating Python code for sequence manipulation and analysis, and helping draft sections of this report. All experimental results were independently verified, and the core methodology and findings represent original work conducted during the hackathon.