Function Over Sequence: Empirical Evaluation of Protein Language Models for Biosecurity Screening

Saahir Dhanani
Independent Researcher

With
Apart Research

Abstract

DNA synthesis screening is a critical biosecurity chokepoint, but current tools detect dangerous sequences by similarity to known threats, a paradigm that collapses against AI-assisted protein design. Using ProteinMPNN and ESMFold, we generated 12,000 evasion variants of 50 known toxin proteins across 12 sampling temperatures, producing sequences with as little as 11% mean identity to known toxins. We evaluated ESM-C 600M protein language model embeddings as a function-aware alternative, comparing four classifier architectures against BLASTp and commec.

ESM-C classifiers consistently outperform both baselines. At T=1.5, BLASTp detects 3.5% of variants and commec detects 2.9%, while ESM-C kNN maintains 79.5%. Among variants structurally predicted to retain wild-type function at T=1.5, commec detects 0% and BLASTp detects 3.8%, while ESM-C kNN detects 100%. These results provide the first empirical temperature-sweep evaluation of protein language model embeddings for biosecurity screening, directly responding to the open problem identified by Wittmann et al. (2025) and Abel et al. (2026), and establish organism-matched negative construction as a necessary methodological requirement for honest evaluation in this space.

1. Introduction

DNA synthesis companies are a critical chokepoint in the production of dangerous biological agents. Before fulfilling orders, these companies screen submitted sequences against databases of known dangerous sequences using biosecurity screening software (BSS). The dominant approach is best-match screening: flag a sequence if it shares sufficient similarity to a known sequence of concern. This paradigm has been refined over time to resist naive attacks like codon optimization, but it has a structural weakness in that it can only detect what it has seen before.

AI-assisted protein design has made this weakness acute. Wittmann et al. (2025) demonstrated that freely available tools, including ProteinMPNN, can generate synthetic variants of known toxic proteins that preserve structural and functional properties while evading detection by commercial BSS tools. Across 76,080 synthetic variants of 72 proteins of concern, all four tested BSS tools showed meaningful miss rates. After patching, average detection improved to approximately 97% on probably-functional variants, but residual misses remained. The authors called explicitly for "homology detection using high-dimensional learned embedding spaces" as a necessary next direction. Abel et al. (2026) followed with a theoretical argument for function-based screening, identifying protein cytotoxins as the clearest starting point and calling for empirical evaluation. No such evaluation existed at the time of writing.

We present one. We evaluate ESM-C 600M protein language model embeddings as a function-aware biosecurity screener, testing whether representations learned from evolutionary data maintain detection signal on AI-generated evasion variants that fool sequence-similarity screening. We also identified and corrected a taxonomic bias in negative set construction that inflates detection metrics in naive evaluations of this kind.

Our main contributions are:

1. The first empirical temperature-sweep evaluation of protein language model embeddings for biosecurity screening, showing that ESM-C kNN maintains 79-84% detection across all tested temperatures while BLASTp and commec collapse to near-zero at high sequence diversity.
2. A structural validation of evasion variants using TM-Score, showing that among probably-functional T=1.5 variants (TM-Score > 0.5), commec detects 0% and BLASTp detects 3.8%, while ESM-C kNN detects 100%.
3. A taxonomic bias analysis showing that naive negative set construction inflates classifier AUC from 0.983 to 0.999 by allowing classifiers to learn organism identity rather than toxic function, along with a correction methodology using organism-matched negatives.

2. Related Work

Wittmann et al. (Science, 2025) is the foundational paper motivating this work. They used ProteinMPNN, EvoDiff-MSA, and EvoDiff-Seq to generate 76,080 synthetic variants of 72 wild-type proteins of concern and found that all four tested BSS tools showed meaningful miss rates before patching. After patching, residual misses concentrated in specific protein families. They concluded that sequence-based screening alone is unlikely to remain sufficient and called for learned embedding spaces as a complement. This paper is a direct empirical response to that call.

Abel et al. (SSRN, 2026) argued theoretically for function-based screening, contending that biologically functional proteins are constrained by biophysical requirements learnable from evolutionary data. They identified protein cytotoxins as the clearest starting point for empirical evaluation. We provide that evaluation.

Challacombe & Haas (2024) demonstrated that ESM2 650M with a linear classification head achieves AUC approximately 0.99 on toxin versus non-toxin classification using Tox-Prot. This established that ESM2 embeddings carry toxin-relevant functional information. Their evaluation, however, used only known toxins and random benign proteins, with no adversarial evasion variants and no taxonomic bias analysis. We extend their finding to the adversarial setting and identify the negative set construction problem their evaluation did not surface.

ESM2 (Lin et al., Science, 2023) and ESM-C (Meta, 2024) are protein language models trained on hundreds of millions of sequences using masked language modeling. They learn to predict masked amino acids from evolutionary context, which forces internalization of co-occurrence patterns that proxy for structural and functional similarity. We use ESM-C 600M as our primary model and ESM2 650M for comparison.

ProteinMPNN (Dauparas et al., Science, 2022) generates amino acid sequences conditioned on a protein backbone structure. Its sampling temperature controls sequence diversity, making it a natural tool for generating evasion variants at controlled levels of adversarial difficulty.

commec is the open-source biosecurity screening tool maintained by IBBIS, using HMM-based search against curated profiles of dangerous protein families. We evaluate it as a production-BSS proxy alongside BLASTp.

3. Methods

3.1 Dataset Construction

We downloaded all reviewed toxin sequences from UniProt Swiss-Prot (keyword KW-0800), yielding 6,040 sequences after filtering to 50-1,022 amino acids. This dataset (Tox-Prot) includes animal venoms, bacterial toxins, and plant toxins with experimental evidence of toxicity, and is the standard reference used in prior biosecurity screening research including Challacombe & Haas (2024).

For the negative class, naive sampling from UniProt without taxonomic filtering produced a set dominated by human proteins, while Tox-Prot contains primarily proteins from venomous and pathogenic organisms. Classifiers trained on this imbalanced set exhibited 60.2% false positive rate on non-toxic proteins from venomous organisms — learning organism identity rather than toxic function. We corrected for this by building an organism-matched benign set: we extracted all 1,256 unique taxonomy IDs from Tox-Prot, queried UniProt per-organism for reviewed non-toxin proteins (up to 50 per organism), and subsampled to 8,000 sequences. All reported results use this debiased training set.

We selected 50 seed toxins from Tox-Prot (seed 42, filtered to 50-400 AA), predicted structures using the ESMFold API, and generated 20 variants per seed using ProteinMPNN (Dauparas et al., 2022) at 12 sampling temperatures (T = 0.1 to 1.5), producing 12,000 evasion variants total. ProteinMPNN was selected because it was one of the tools used in Wittmann et al. (2025) and because its temperature parameter provides a continuous axis of adversarial sequence diversity.

3.2 Structural Validation

To assess whether evasion variants plausibly retain toxic function, we followed Wittmann et al. (2025) in using TM-Score as a structural proxy. We folded a stratified sample of 200 variants per temperature using ESMFold and computed TM-Score against each variant's wild-type template PDB using TM-align. Variants with TM-Score $>$ 0.5 and delta-pLDDT $>$ -10 are classified as probably-functional, matching Wittmann et al.'s cutoffs exactly. Detection rates on this subset are our primary reported figures, as they approximate the realistic threat model: an adversary who pre-screens generated variants using structural metrics before ordering synthesis.

3.3 Baselines

We ran two sequence-similarity baselines. For BLASTp, we built a database from the full Tox-Prot FASTA and searched all evasion variants with e-value threshold 0.001, reporting detection at 40% identity. For commec, we installed the IBBIS open-source screener with its biorisk HMM database and ran it against all evasion sets. Our commec installation is lightweight, without the full taxonomy database, so results represent a lower bound on full-installation commec performance.

3.4 Embeddings and Classifiers

Our primary model is ESM-C 600M (Meta, 2024), which produces 1,152-dimensional representations across 36 layers. We also evaluated ESM-2 650M (Lin et al., 2023) for comparison with prior work. For each sequence we ran a single forward pass and mean-pooled the final layer across sequence positions to produce one vector per sequence. We trained four classifiers on the toxin and organism-matched benign embeddings using StandardScaler normalization inside sklearn Pipelines with 5-fold stratified cross-validation: Logistic Regression, Random Forest, MLP, and kNN cosine. kNN cosine is our primary reported classifier because it directly measures geometric proximity in embedding space without parametric assumptions, making it the most interpretable test of whether evasion variants occupy toxin-like regions of the representation.

4. Results

4.1 Cross-Validation Performance

Table 1 reports cross-validation performance for ESM-C classifiers trained on the organism-matched benign set. All classifiers achieve AUC 0.981-0.989. CV false positive rate is 4.9-6.2% on organism-matched negatives, compared to 1.0-1.5% on random human-dominated benign sequences. The gap reflects the harder task: distinguishing toxic from non-toxic proteins within the same organisms rather than human proteins from venom proteins. For reference, classifiers trained on the naive human-dominated benign set achieved AUC 0.997-0.999, which is an inflation driven by taxonomic shortcutting, not functional discrimination.

Table 1: 5-fold cross-validation, ESM-C 600M with organism-matched benign (n=14,040)

Classifier	AUC (±std)	F1	CV FPR
kNN cosine	0.983 ±0.005	0.943	6.20%
Logistic Regression	0.981 ±0.003	0.933	5.90%
Random Forest	0.989 ±0.002	0.946	5.80%
MLP	0.986 ±0.004	0.948	4.90%

4.2 Detection at T=0.5

Table 2 shows detection rates on all 1,000 T=0.5 evasion variants. ESM-C kNN detects 79.2% versus BLASTp's 57.8% and commec's 19.9%. The commec result is lower than BLASTp across all temperatures on our evasion set, which we attribute partly to our lightweight installation and discuss in Section 5.

Table 3 shows detection restricted to the 160 probably-functional variants (TM-Score $>$ 0.5, delta-pLDDT $>$ -10, 80.4% of sampled variants at T=0.5). On this subset ESM-C kNN detects 86.2% versus BLASTp's 69.4% and commec's 24.4%. The larger gap on probably-functional variants is consistent with structurally-constrained sequences clustering more tightly around known toxins in embedding space.

Figure 1 shows the embedding geometry underlying these detection rates: evasion variants cluster predominantly in the toxin region of ESM-C embedding space despite low sequence identity to known toxins.

Figure 1: UMAP projection of ESM-C 600M embeddings. 1,000 organism-matched benign proteins (blue), 1,000 known toxins (red), and 1,000 T=0.5 evasion variants (orange), with probably-functional variants marked as stars. Evasion variants cluster predominantly in the toxin region.

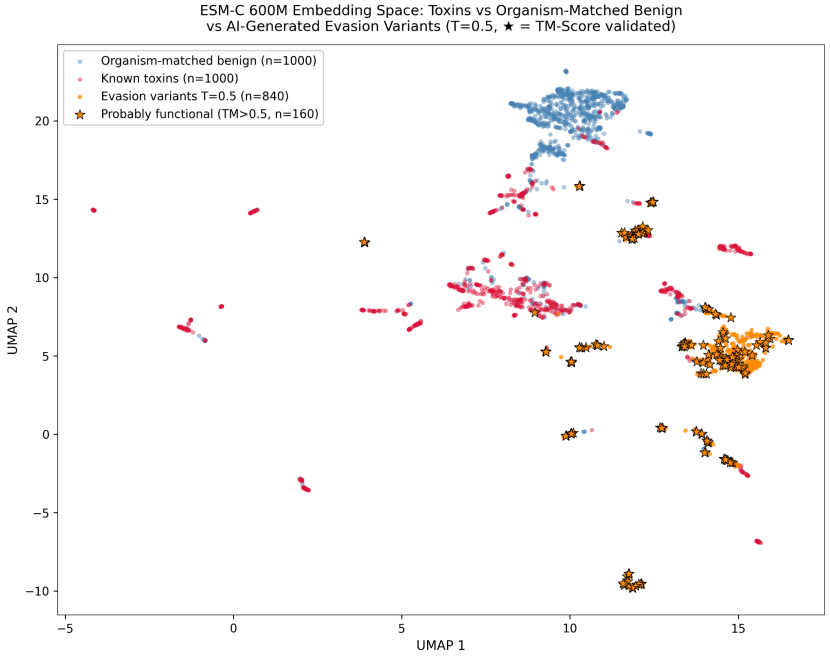


Table 2: Detection on all 1,000 T=0.5 variants

Method	Rate
commec biorisk HMMs	19.90%
BLASTp (>=40% identity)	57.80%
ESM-2 kNN	77.90%
ESM-C kNN	79.20%
ESM-C LR	61.40%
ESM-C RF	67.60%

Table 3: Detection on probably-functional T=0.5 variants (n=160)

Method	Rate
commec biorisk HMMs	24.40%
BLASTp (>=40% identity)	69.40%
ESM-2 kNN	85.00%
ESM-C kNN	86.20%

4.3 Temperature Sweep

Figure 2 shows detection rates across all 12 temperatures. As temperature increases and mean sequence identity to known toxins falls, BLASTp and commec collapse while ESM-C classifiers hold. At T=1.5, BLASTp detects 3.5% and commec detects 2.9%, while ESM-C kNN maintains 79.5% and ESM-C RF reaches 91.1%. ESM-C kNN variance across CV folds is narrow throughout (±0.6-2.1pp), and detection rate is deterministic across random seeds (79.2% ±0.0% across seeds 41, 42, 43), confirming results are stable to training set and initialization variation.

Table 4: Detection at selected temperatures

Temp	Mean ID%	BLASTp%	commec%	ESM-C kNN% (±std)	ESM-C RF%
0.1	37	61.4	21	80.2 ±2.1	42.5
0.5	34.2	57.8	19.9	79.2 ±1.2	67.6
1	23.5	25.9	14.1	83.4 ±0.8	89
1.5	11.1	3.5	2.9	79.5 ±1.4	91.1

Figure 2: Detection rate vs ProteinMPNN sampling temperature. ESM-C kNN, ESM-C RF, ESM-2 kNN, BLASTp, and commec across all 12 temperatures. BLASTp and commec collapse at high temperatures while ESM-C classifiers hold.

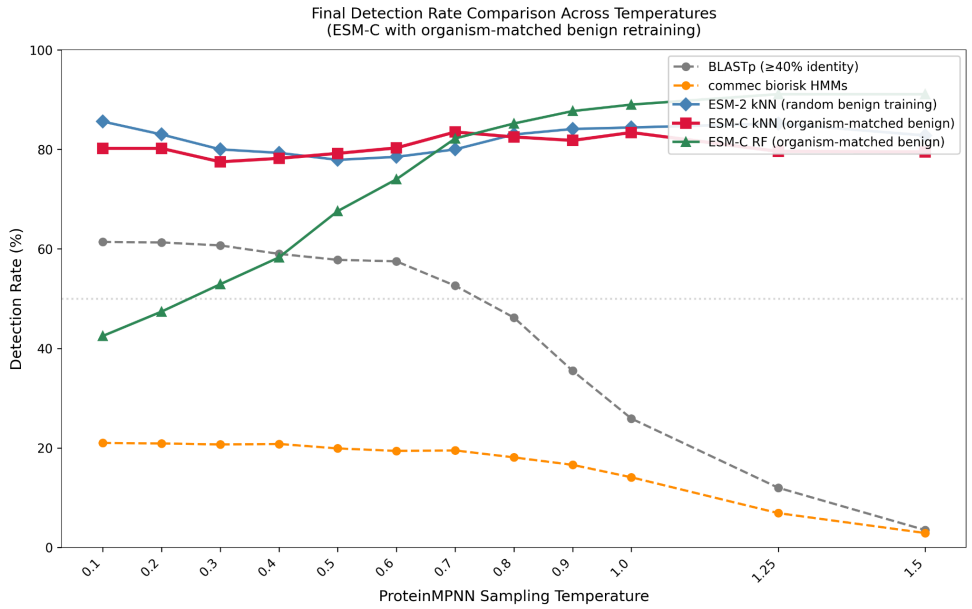
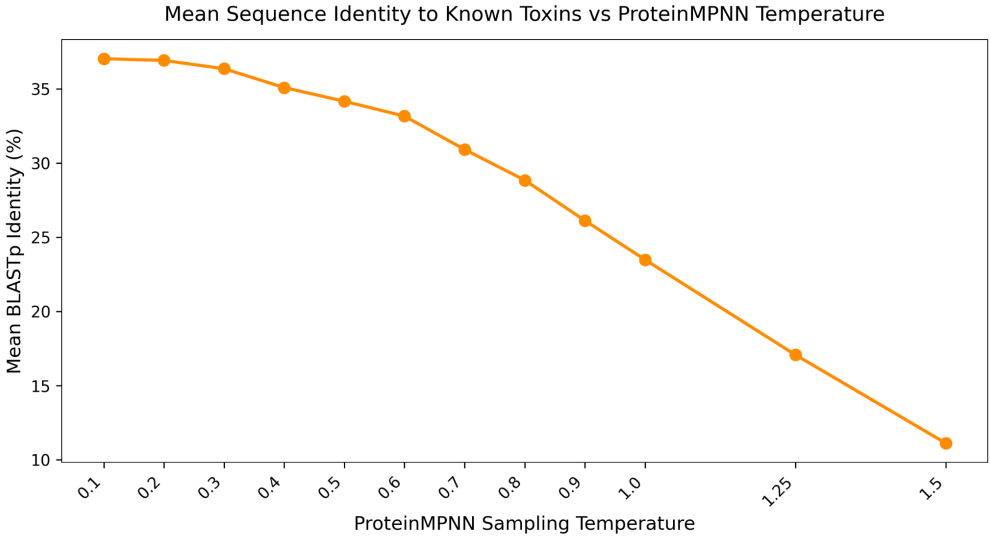


Figure 3: Mean sequence identity vs temperature. BLASTp percent identity to nearest Tox-Prot sequence falls from 37% at T=0.1 to 11% at T=1.5, validating temperature as a proxy for adversarial sequence diversity.



4.4 Structural Validation at T=1.5

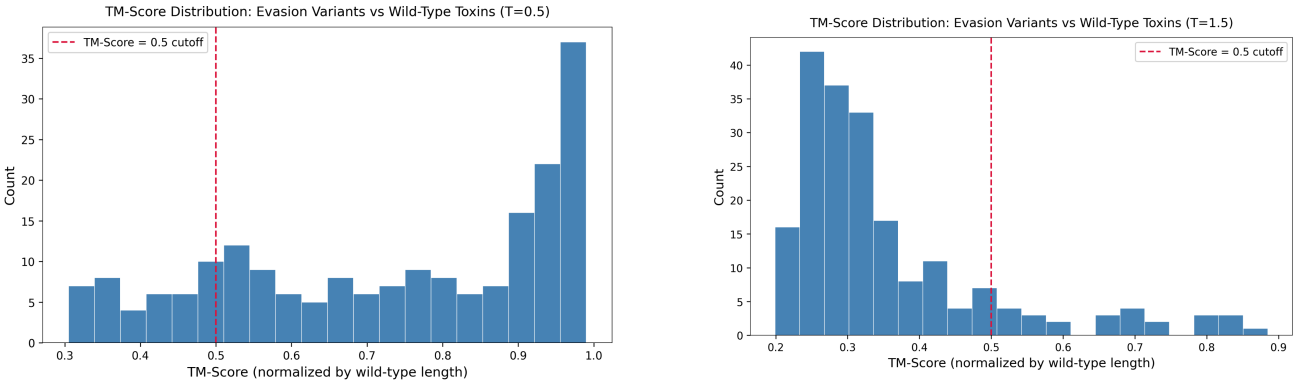
TM-Score analysis reveals a critical difference between the two temperature regimes. At T=0.5, 80.4% of sampled variants are probably-functional (mean TM-Score 0.730). At T=1.5, only 13.0% meet the threshold (mean TM-Score 0.355), meaning most T=1.5 variants are unlikely to retain toxic function despite evading screening. Table 5 reports detection on the 26 probably-functional T=1.5 variants. commec detects zero and BLASTp detects one. ESM-C kNN detects all 26.

Table 5: Detection on probably-functional T=1.5 variants (n=26)

Method	Detected	Rate
commec biorisk HMMs	0/26	0.00%
BLASTp (>=40% identity)	1/26	3.80%
ESM-2 kNN	23/26	88.50%
ESM-C kNN	26/26	100.00%

n=26 is a small sample due to ESMFold API rate limits. The directional result is clear but a larger sample would strengthen the claim.

Figure 4: TM-Score distributions at T=0.5 and T=1.5. Histograms of TM-Score for sampled variants at T=0.5 (left, n=199, 80.4% probably-functional) and T=1.5 (right, n=200, 13.0% probably-functional). The dashed vertical line marks TM-Score = 0.5.



5. Discussion and Limitations

5.1 Discussion

Sequence-based biosecurity screening is structurally vulnerable to AI-assisted protein design. Our results quantify this vulnerability concretely: at ProteinMPNN T=1.5, BLASTp detects 3.5% of evasion variants and commec detects 2.9%, while ESM-C kNN maintains 79.5%. Among the

subset of T=1.5 variants that structurally resemble their wild-type templates, commec detects zero and BLASTp detects one. ESM-C detects all 26.

The implication for AI safety is that the gap between offensive and defensive capability is real and measurable today, using freely available tools. ProteinMPNN is open-source. ESMFold is a free API. The attack described in Wittmann et al. (2025) requires no specialized knowledge beyond what is publicly documented. Protein language model embeddings represent a practical defensive response that can be layered on top of existing screening infrastructure without replacing it.

The taxonomic bias finding carries its own implication. A classifier that achieves 0.999 AUC by learning organism identity rather than toxic function provides false confidence. Rigorous evaluation of biosecurity ML requires organism-matched negatives, not random protein databases. This applies to any future work in this space.

The commec result is harder to interpret. HMM-based search is theoretically more sensitive than BLASTp at low sequence identity, yet commec detection was lower than BLASTp across all temperatures on our evasion set. This likely reflects a combination of our lightweight installation, HMM profile coverage gaps for the specific toxin families in our seed set, and the fact that ProteinMPNN variants at moderate-to-high temperatures produce sequences that fall between known domain profiles. Whether this holds against full-database commec is an open question.

5.2 Limitations

We infer functional retention from TM-Score following Wittmann et al. (2025), but did not synthesize or experimentally characterize any variants. TM-Score is a noisy proxy: some variants meeting our probably-functional cutoff will not actually retain toxicity, and some below the cutoff may still be dangerous. Wet-lab validation using safe biological proxies, as in Ikonomova et al. (2025), is the appropriate next step.

Evasion variants were generated from seeds drawn from the same Tox-Prot dataset used to train classifiers, which means the classifier may be detecting family-level signals shared between a variant and its wild-type parent rather than generalizing to novel toxin families. The high-temperature results partially address this concern: at T=1.5, mean sequence identity is 11% and 640 variants have zero BLASTp hits. However, a held-out family evaluation is needed for a rigorous generalization claim.

Organism-matched retraining corrects the training distribution but we were unable to measure debiased ESM-C FPR on organism-matched negatives directly due to embedding infrastructure constraints. The pre-debiasing ESM-2 FPR of 60.2% on organism-matched negatives is the upper bound on remaining taxonomic bias. Whether debiased ESM-C FPR is acceptably low remains unmeasured.

Our commec results use a lightweight installation without the full taxonomy database, so full-database commec may perform differently. The directional finding that HMM-based screening collapses at high temperature is unlikely to reverse, but the magnitude may change.

Tox-Prot covers animal venoms, bacterial toxins, and plant toxins. Viral proteins and novel engineered sequences outside the evolutionary distribution of known toxins are not represented, and Abel et al. (2026) identified viral entry proteins as the natural next priority.

Finally, we measured false positive rate on held-out UniProt sequences rather than real synthesis order distributions. A production deployment would require FPR characterization on actual order data to avoid alert fatigue.

5.3 Future Work

The most important next steps are: a held-out toxin family evaluation where specific families are excluded from training entirely and their evasion variants used for testing; full-database commec comparison on a matched evasion set; FPR characterization on real synthesis order distributions; extension to viral entry proteins per Abel et al. (2026); and organism-matched FPR measurement for debiased ESM-C to confirm the taxonomic bias correction holds.

The deeper architectural question is whether protein language models trained with standard masked language modeling objectives can be made invariant to sequence diversity while maintaining functional discrimination. JEPA-DNA (Larey et al., 2026) showed that a joint embedding predictive objective improves functional generalization for genomic sequences. An analogous pretraining objective for protein language models of predicting the embedding of one functionally equivalent sequence from another, supervised by shared structural family membership would make functional invariance an explicit training objective rather than an emergent byproduct of evolutionary pretraining. This is the most direct path toward representations that generalize to sequences genuinely unlike anything in the training data.

The offensive capability side of this problem is also accelerating. King et al. (2025) demonstrated that Evo 1 and Evo 2, genomic foundation models trained on billions of DNA sequences, can generate novel functional bacteriophage genomes with as little as 40% nucleotide identity to natural sequences. The same generative capability that makes these models valuable for therapeutic design makes them a potential tool for generating sequences that fall entirely outside the evolutionary distribution of known threats. Screening approaches trained only on ESM-C or ESM-2 representations may not generalize to sequences generated by genome-scale models operating at the nucleotide level rather than the amino acid level. Extending function-aware screening to cover Evo-generated sequences, and evaluating whether Evo 2 embeddings themselves could serve as a detection backbone for nucleotide-level threats is a natural and urgent next direction.

6. Conclusion

Current DNA synthesis screening rests on a sequence-similarity paradigm that is measurably inadequate against AI-assisted protein design. At ProteinMPNN T=1.5, where mean sequence identity to known toxins is 11%, BLASTp detects 3.5% of evasion variants and commec detects 2.9%. Among the variants structurally predicted to retain wild-type function, those numbers fall to 3.8% and 0.0% respectively. ESM-C kNN detects 79.5% of all T=1.5 variants and 100% of probably-functional ones.

Beyond the detection results, this work surfaces a methodological problem relevant to any future evaluation in this space: naive negative set construction produces classifiers that learn organism identity rather than toxic function, inflating AUC by up to 0.016 and FPR by 58 percentage points on organism-matched negatives. Organism-matched negative sets are a necessary condition for honest evaluation of toxin detection classifiers. The two findings together – that protein language model embeddings provide meaningful detection signal, and that evaluating them correctly requires careful dataset construction – define the practical agenda for function-aware biosecurity screening.

Code and Data

Code repository: https://github.com/saaahir/aixbiohackathon

Evasion variant sequences are not publicly released given their potential for misuse. They are available to verified biosecurity researchers upon request, subject to responsible disclosure review. Training data (Tox-Prot, organism-matched benign) is publicly available from UniProt.

References

1. Wittmann, B.J., et al. "Strengthening nucleic acid biosecurity screening against generative protein design tools." Science 390, 82-87 (2025). https://doi.org/10.1126/science.adu8578

2. Abel, G., et al. "Beyond Sequence Similarity: Toward Function-Based Screening of Nucleic Acid Synthesis." SSRN (2026). https://doi.org/10.2139/ssrn.6444478

3. Lin, Z., et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model." Science 379, 1123-1130 (2023). https://doi.org/10.1126/science.ade2574

4. Dauparas, J., et al. "Robust deep learning-based protein sequence design using ProteinMPNN." Science 378, 49-56 (2022). https://doi.org/10.1126/science.add2187

5. Larey, A., et al. "JEPA-DNA: Grounding Genomic Foundation Models through Joint-Embedding Predictive Architectures." arXiv:2602.17162 (2026). https://doi.org/10.48550/arXiv.2602.17162

6. Ikonomova, S.P., et al. "Experimental evaluation of AI-driven protein design risks using safe biological proxies." bioRxiv 2025.05.15.654077 (2025). https://doi.org/10.1101/2025.05.15.654077

7. The UniProt Consortium. "UniProt: the Universal Protein Knowledgebase in 2023." Nucleic Acids Research 51, D523-D531 (2023). https://doi.org/10.1093/nar/gkac1052

8. Challacombe, C.A., and Haas, N.S. "Towards a Dataset for State of the Art Protein Toxin Classification." bioRxiv 2024.04.14.589430 (2024). https://doi.org/10.1101/2024.04.14.589430

9. King, S.H., et al. "Generative design of novel bacteriophages with genome language models." bioRxiv 2025.09.12.675911 (2025). https://doi.org/10.1101/2025.09.12.675911

10. EvolutionaryScale. ESM Cambrian (ESM-C). (2024). https://github.com/evolutionaryscale/esm

LLM Usage Statement

Claude (Anthropic) was used throughout this project for brainstorming approaches, generating and debugging code, searching literature, and drafting sections of this report. All experimental results were independently verified by running code locally and inspecting outputs directly. All scientific claims reflect actual experimental findings. The authors take full responsibility for the accuracy of all results and interpretations.