Activation Probes for Synthetic Toxin Variant Detection[1]

Maxwell DeFanti
UC Berkeley

Ishaan Panigrahi
UC Berkeley

Kevin Zhang
UC San Diego

With
Apart Research

Abstract

DNA synthesis screening prevents bad actors from obtaining the physical sequences needed to produce dangerous toxins and pathogens. However, current screening tools like BLAST and SecureDNA rely on sequence similarity to known threats, and recent work has shown that AI protein design tools such as ProteinMPNN can generate functional toxin variants that evade these screens at rates approaching 100%. We introduce a screening approach that trains an activation probe on ESM-2 embeddings to recognize toxic function across diverged sequences; on held-out synthetic variants that are ~40% identity to their parents, our classifier maintains 86.7% recall while BLAST recall collapses to 46.7%. This provides initial evidence that protein language model embeddings can be a robust second layer of defense for DNA synthesis screening, complementing current similarity-based methods.

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

1. Introduction

Advanced AI tools may help both experts and novices design dangerous biological sequences, including viruses and toxins. While AI may lower barriers in designing dangerous sequences, bad actors still need to actually obtain the physical DNA from synthesis providers.

DNA synthesis screening currently relies on tools like SecureDNA, BLAST, and Common Mechanism, which compare orders to databases of known harmful sequences. These tools have high accuracy on known harmful sequences, but AI tools like ProteinMPNN [4] can create synthetic variants that are functionally similar yet evade detection. The Microsoft study "Strengthening nucleic acid biosecurity screening against generative protein design tools" demonstrated this vulnerability, with synthetic variants achieving 100% evasion rates against multiple screening tools [1].

The main issue is that screening tools are primarily pattern-matching algorithms, without deep understanding of biological function. We develop a classifier using ESM-2, a transformer-based protein language model, to detect harmful synthetic variants through functional similarity [1] rather than sequence similarity. To our knowledge, this is the first application of protein language models to DNA synthesis biosecurity screening.

Our main contributions are:

1. We introduce a protein language model-based approach for DNA synthesis biosecurity screening, using activation probes on ESM-2 embeddings
2. We show that an ESM-2 probe trained only on natural toxins already generalizes to ProteinMPNN-generated synthetic variants, and that adding synthetic toxins to training provides modest additional gains, primarily on natural sequences.
3. We demonstrate superior performance on detecting synthetic toxin variants compared to current sequence-based methods

2. Related Work

DNA Synthesis Screening. Current DNA synthesis screening relies on sequence similarity tools like BLAST, SecureDNA, and Common Mechanism, which compare orders to databases of known harmful sequences [3]. The Microsoft study exposed this vulnerability, showing that AI-designed synthetic variants evaded detection across multiple screening systems [1]. Recent work has focused on patching these sequence-based approaches [1], but all remain fundamentally limited by their reliance on sequence similarity rather than functional understanding.

Protein Language Models. Transformer-based models like ESM-2 [2] have shown strong performance on protein function prediction and variant effect prediction through self-supervised training on evolutionary sequences. However, to our knowledge, no prior work has applied these

models to biosecurity screening in the context of synthetic proteins. Our work is the first to leverage protein language model embeddings for detecting functionally dangerous proteins in DNA synthesis screening.

3. Methods

Figure 1. Overview of the toxin classifier. Real toxin sequences from ToxProt and DBETH are augmented with ProteinMPNN-generated variants and combined with harmless SwissProt proteins to train a lightweight probe on frozen ESM-2 embeddings, producing a binary toxin/safe prediction.

We frame toxicity detection as binary classification of protein sequences, using ESM-2 embeddings as the input representation. This is motivated by the fact that BLAST-style similarity fails on ProteinMPNN-generated variants that preserve function but diverge in sequence, while ESM-2 embeddings encode functional and structural signal directly.

We drew positives from ToxProt and DBETH (curated, experimentally validated toxins and venom components) and negatives from SwissProt (manually reviewed, high-confidence non-toxic annotations). To prevent homology leakage, we deduplicated positives at 90% identity with CD-HIT, filtered to lengths 50–1000, then clustered the full dataset at 50% identity and assigned whole clusters to train/val/test splits at 70/15/15. The final test set contains 1,254 sequences with 379 positives (~30% toxin prevalence).

Using ProteinMPNN with PDB parent structures drawn exclusively from the training split, we generated variants at temperatures 0.1, 0.5, and 1.0, yielding ~85%, 55–65%, and 35–50% identity to the parent. Variants inherited their parent's split assignment and were down-weighted to 0.7 in the loss to reflect lower label confidence. We confirmed no test redesign exceeded 50% identity to any natural training sequence, and validated structural plausibility on 15 random variants with ESMFold (all met TM-score $>$ 0.7 against the parent backbone, the same bar used in [1] as a proxy for functional similarity).

We embedded each sequence using frozen ESM-2 650M, mean-pooling the 33rd-layer representation over true positions to a 1280-d vector. On top we trained our activation probe, a two-layer MLP with weighted binary cross-entropy (positives upweighted $\sim 2\times$ to match the negative-to-positive ratio), AdamW, gradient clipping, and early stopping on validation AUC-ROC. ESM-2 was kept frozen because fine-tuning at this dataset size risks collapsing the general representations.
Evaluation. We compared against (i) a BLAST baseline flagging hits with e-value < 0.001 and identity $> 30\%$ against the training positives, and (ii) an ablation classifier identical in every respect but trained on natural sequences only. The primary metric is recall across natural sequences and ProteinMPNN variants at each diversity level.

4. Results

On the full test set of 1,254 sequences (379 positives, 875 negatives), the augmented classifier achieved an AUC-ROC of 0.983, recall of 93.1%, and F1 of 0.899, compared to the natural-only classifier (AUC-ROC 0.981, recall 89.7%, F1 0.903) and the BLAST baseline (AUC-ROC 0.928, recall 87.3%, F1 0.913). Both ESM-2 classifiers improved AUC-ROC and AUC-PR substantially over BLAST, but BLAST achieved higher precision (0.957) and a much lower false positive rate (1.7% vs 6.1% for the augmented classifier and 3.9% for the natural-only classifier). BLAST still missed 48 of 379 true toxins (12.7%), genuine toxins that fell below its detection threshold given the cluster-level split, which enforces out-of-distribution evaluation even for natural sequences. While results have varied across different experimental runs, the overall trend of superior performance compared to BLAST remains consistent across the board.

Metric	BLAST	Our Method (Natural only)	Our Method (Natural + Synthetic)
AUC-ROC	0.928	0.981	0.983
AUC-PR	0.874	0.955	0.957
Recall	0.873	0.897	0.931
Precision	0.957	0.909	0.869
F1	0.913	0.903	0.899
Accuracy	0.95	0.942	0.937
FPR	0.017	0.039	0.061
TP	331	340	353
FP	15	34	53

FN	48	39	26
TN	860	841	822

Table 1. Overall classification performance

The central result is shown in Table 2. We find that BLAST recall degrades monotonically as sequence divergence increases, from 89.5% on natural sequences to 86.7% at high similarity, 80.0% at medium similarity, and 46.7% at low similarity. At the lowest divergence level BLAST misses more than half of redesigned toxins. Both ESM-2 classifiers maintain substantially higher recall across all redesign levels, holding 86.7% recall at low similarity where BLAST drops to 46.7%, with the advantage over BLAST growing from 3.9 percentage points on natural sequences to 40.0 percentage points at low similarity.We find that on natural toxic sequences the augmented classifier modestly outperforms the natural-only classifier, 93.4% vs 89.5% recall, while on synthetic toxic sequences the two models are nearly indistinguishable: both achieve 91.1% recall, with mean predicted scores of 0.895 (augmented) and 0.907 (natural-only). The natural-only classifier is in fact slightly more confident on synthetic positives than the augmented one. This is the key finding of the ablation: ESM-2 embeddings already carry enough function-relevant signal that a probe trained only on natural toxins generalizes to ProteinMPNN redesigns at all three divergence levels, including ~40% identity, with no exposure to synthetic variants during training. The augmentation does buy a small overall improvement (AUC-ROC 0.983 vs 0.981; recall 93.1% vs 89.7%), but at the cost of a higher false positive rate (6.1% vs 3.9%), and the contribution comes from natural sequences rather than synthetic ones. The robustness to ProteinMPNN-induced sequence divergence appears to be a property of the underlying ESM-2 representations, not of the augmentation procedure.

Divergence level	N+	BLAST	Natural	Natural + Synthetic	Advantage over BLAST
Natural	334	0.895	0.895	0.934	0.039
High similarity (~85%)	15	0.867	0.933	0.933	0.067
Medium similarity (~60%)	15	0.8	0.933	0.933	0.133
Low similarity (~40%)	15	0.467	0.867	0.867	0.4

Table 2. Robustness to sequence redesign (recall by divergence level)

5. Discussion and Limitations

The primary constraints of our work were limited experimental validation, limited data scope, and limited toxicity verification. First, we were only able to perform a small handful of experiments and empirical validations due to computational and time constraints, which limit our ability to establish firm statistical evidence. Over these few runs, we observed a fair amount of variance in results, however, outperformance of BLAST in AUC and True Positive Rates for synthetic toxins were observed across the board. Second, due to the aforementioned constraints, we were only able to use a few thousand registered protein sequences and several hundred synthetic sequences. We would love to see future verification and improvements to the techniques presented in our work. Finally, without access to wet lab experimentation, it is impossible to verify whether or not our synthetic proteins would be functionally toxic, so we rely on the claim from [1] that structural similarity is a reasonable proxy for functional similarity.

Future Work

The most impactful next steps are:

1. Expanding synthetic detection to viruses, bacterias, and dual-use research proteins
2. Validating that the synthetics we generated are actually toxic in a wet-lab
3. Explore robustness of the method to more sophisticated synthetic generation methods than Protein MPNN and adversarial methods

6. Conclusion

We address a major biosecurity threat where AI tools like ProteinMPNN can create synthetic toxin variants that evade current DNA synthesis screening. Current screening relies on sequence similarity methods like BLAST, which fail when synthetic variants preserve toxic function while changing sequence patterns. We develop a protein language model-based screening approach, using ESM-2 embeddings to detect functional similarity rather than sequence similarity. Our classifier maintains 87% recall on synthetic variants at ~40% identity to their parents, while BLAST's performance collapses from 90% to 47% recall as sequence similarity decreases. This demonstrates that function-based screening can provide robust defense against AI-generated biological threats that would otherwise evade detection.

Code and Data

- **Code:** https://github.com/maxdefanti/ProteinToxicityClassifier
- **Data:** We will not be providing our synthetic dataset or data generation scripts out of concern for dual-use applications. Should you want access to our dataset or synthetic toxin generation code for validation or experimental purposes, please reach out to us via email.

References

1. Wittmann, B. J., Alexanian, T., Bartling, C., Beal, J., Clore, A., Diggans, J., Flyangolts, K., Gemler, B. T., Mitchell, T., Murphy, S. T., Wheeler, N. E., & Horvitz, E. (2025). Strengthening nucleic acid biosecurity screening against generative protein design tools. Science, 390(6617), 82-87. https://doi.org/10.1126/science.adu8578
2. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637), 1123-1130. https://doi.org/10.1126/science.ade2574
3. Alley, E. C., Turpin, M., Liu, A. B., Kulp-McDowall, T., Swett, J., Edison, R., Von Kleist, M. P., Kelley, J., Bernstein, A., Linfield, J., Siegel, M., Hobbs, N., Esvelt, K., & Church, G. (2024). A system capable of verifiably and privately screening global DNA synthesis. Science, 386(6723), 732-740. https://doi.org/10.1126/science.adl4635
4. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., & Baker, D. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. Science, 378(6615), 49-56. https://doi.org/10.1126/science.add2187

LLM Usage Statement

We used Claude to brainstorm approaches and help draft sections. All results and claims were independently verified.
We used Claude code to assist in development and data visualization