FuncScreen: Contrastive PLM Embeddings for Evasion-Resistant Biosecurity Screening*

Aheli Poddar
Institute of Engineering & Management, Kolkata
ahelipoddar2003@gmail.com

With
Apart Research

Abstract

Current DNA synthesis screening relies on sequence homology, which AI protein design tools like ProteinMPNN evade by generating functional threat variants with as low as 7% sequence identity to known threats. We introduce FuncScreen, a contrastive learning framework over frozen ESM-2 embeddings that screens by predicted biological function rather than sequence similarity. Trained with supervised contrastive loss, hard-negative mining, and embedding-space Mixup augmentation on 985 curated pore-forming toxin and benign homolog sequences, FuncScreen achieves 1.000 AUROC [1.000, 1.000] on standard and hard-negative splits. On 4,100 ProteinMPNN-designed adversarial variants, FuncScreen maintains 0.991 AUROC [0.988, 0.993] where homology drops to 0.952 [0.944, 0.959]. We provide a preliminary certified robustness analysis under biologically structured mutations (1,000 Monte Carlo samples, 100 sequences), finding an empirical-certified gap of at most 1%. We validate generalization on a second threat family (ribosome-inactivating proteins, AUROC 0.962) and report out-of-distribution false positive rates.

1 Introduction

DNA synthesis screening is a critical chokepoint for preventing misuse of synthetic biology. Current screening infrastructure including SecureDNA [1] and the IBBIS Common Mechanism [2] fundamentally relies on sequence homology: a query sequence is flagged if it is sufficiently similar to a known threat in a curated database.

However, AI protein design tools such as ProteinMPNN [3] can perform *inverse folding*: given a protein's 3D backbone structure, they generate entirely new amino acid sequences predicted to fold into the same structure and perform the same biological function. These designed sequences can share less than 30% sequence identity with any naturally occurring protein, falling below the detection threshold of homology-based screening. The cross-sector study by Horvitz et al. [4] demonstrated that AI-designed protein variants already evade existing screening pipelines, and recent work including GeneBreaker and SafeProtein (both ICLR 2026) has further characterized this attack surface.

The defensive response to this threat has been limited. While the attack side is well-characterized, screening systems that operate in *function space* rather than sequence space remain underexplored. FuncScreen addresses this gap. The individual techniques we employ—supervised contrastive learning [6], protein language model embeddings [5], Mixup augmentation [7]—are established; our contribution is their combination and systematic evaluation in the biosecurity screening domain, together with a novel adversarial benchmark and robustness analysis.

*Research conducted at the AIxBio Hackathon, April 2026. Code: https://github.com/XAheli/AiXBio

Our main contributions are:

1. **A ProteinMPNN adversarial evaluation benchmark** of 4,100 inverse-folded variants across 164 threat proteins at 5 sampling temperatures, providing a systematic stress test of screening against AI-designed evasion.

2. **FuncScreen**, a contrastive screening framework combining ESM-2 embeddings, hard-negative mining, Mixup augmentation, and adversarial training, evaluated with bootstrap confidence intervals and paired significance tests across 5 evaluation splits.

3. **A preliminary certified robustness analysis** adapting randomized smoothing to biologically structured mutation spaces, with 1,000 Monte Carlo samples per sequence.

4. **Ablation studies and generalization experiments** including projection dimension, temperature, hard-negative ratio, multi-scale classification, LOSO cross-validation, out-of-distribution false positive rates, and a second threat family (ribosome-inactivating proteins).

2 Related Work

DNA synthesis screening. SecureDNA [1] uses cryptographic distributed oblivious PRF to screen sequences as short as 30bp against a curated hazard database. The IBBIS Common Mechanism [2] employs HMM-based biorisk screening with best performance above 150bp. Both rely fundamentally on sequence similarity to known threats. The OSTP Framework [13] established federal guidance requiring compliant providers, and the Biosecurity Modernization and Innovation Act of 2026 (S.3741) mandates screening by gene synthesis providers. However, as Kim [12] notes, the legislation mandates homology-based screening without addressing AI-designed functional variants.

AI-assisted evasion of screening. Horvitz et al. [4] demonstrated that AI-designed synthetic protein variants can evade sequence-based screening. Edison et al. [11] showed that unregulated DNA fragments from dozens of providers are sufficient to assemble dangerous pathogens.

Protein language models. ESM-2 [5] is a 650M-parameter protein language model trained on 65 million sequences from UniRef50. ProteinMPNN [3] performs inverse folding, enabling generation of functionally equivalent but sequence-divergent proteins.

Contrastive learning and data augmentation. Supervised contrastive learning [6] has been applied to molecular property prediction and protein function classification. Mixup [7] provides regularization through linear interpolation of training examples. To our knowledge, this is the first application of contrastive learning with Mixup augmentation specifically to biosecurity screening, though the component techniques are individually well-established.

Certified robustness. Randomized smoothing [8] provides robustness guarantees for classifiers. RS-Del [9] extended this to edit-distance perturbations on discrete sequences. We adapt the framework to biologically structured mutation spaces.

3 Methods

3.1 Data Curation

We curated 985 protein sequences from UniProt Swiss-Prot (reviewed entries only) organized around **pore-forming toxins (PFTs)**.

- **Threat sequences** (335): pore-forming toxins (UniProt keyword KW-0800 + free-text "pore-forming"), aerolysin family (Pfam PF01338), cholesterol-dependent cytolysins (Pfam PF01289), hemolytic toxins (KW-0354 + KW-0800).
- **Hard negatives** (459): MACPF domain proteins (Pfam PF01823) that are *not* toxins, including complement components, perforin, gasdermin, and antimicrobial pore-forming peptides.
- **Easy negatives** (191): random human hydrolases with no toxin annotation.

3.2 Evaluation Splits

We constructed five evaluation splits:

1. **Standard** (186 sequences): held-out natural sequences.
2. **Hard negative** (299): threats paired exclusively with structurally similar benign proteins.
3. **Sequence divergent** (144): threat sequences with low k-mer similarity to training threats.
4. **BLOSUM62 adversarial** (2,242): computationally mutated variants using BLOSUM62-conservative substitutions at 5%–50% mutation rates.
5. **ProteinMPNN adversarial** (4,374): inverse-folded variants generated by ProteinMPNN at 5 sampling temperatures (T=0.1, 0.3, 0.5, 0.8, 1.0), producing variants with 7%–60% sequence identity. Structures from AlphaFold DB [10].

3.3 FuncScreen Architecture

FuncScreen consists of a **projection head** and a **classification head** trained jointly on frozen ESM-2 (650M) mean-pooled embeddings (1,280-dim).

The projection head maps embeddings to a 256-dim normalized space via three linear layers with batch normalization, ReLU, and dropout (0.1). The classification head takes the projection and produces a scalar threat probability.

Training objective.

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{SupCon}} + (1 - \alpha) \cdot \mathcal{L}_{\text{BCE}}, \quad \alpha = 0.5 \tag{1}$$

$\mathcal{L}_{\text{SupCon}}$ is the supervised contrastive loss [6] with temperature $\tau = 0.07$.

Hard-negative mining. For each threat protein, we identify the $k = 3$ closest benign proteins in embedding space (cosine similarity) and oversample them in training.

Mixup augmentation. We apply Mixup [7] in embedding space: for same-class pairs, $\mathbf{x}_{\text{mix}} = \lambda \mathbf{x}_a + (1 - \lambda)\mathbf{x}_b$ with $\lambda \sim \text{Beta}(0.2, 0.2)$, applied with probability 0.5 per batch.

Adversarial training. High-temperature ProteinMPNN variants (T=0.8, 1.0; 1,640 sequences) are added to training as additional threat examples, hardening the decision boundary against AI-designed evasion. Low-temperature variants (T=0.1, 0.3) are reserved for evaluation.

Training details. AdamW optimizer, learning rate 10^{-3}, weight decay 10^{-4}, cosine annealing schedule, batch size 64, 50 epochs. Best model selected by validation AUROC.

3.4 Baselines

1. **K-mer similarity** (k=5): max Jaccard similarity to known threats. Proxy for BLAST/homology screening.
2. **Cosine NN**: mean cosine similarity to 5 nearest threat embeddings in ESM-2 space.
3. **Linear classifier**: logistic regression on ESM-2 embeddings.
4. **KNN classifier**: 5-nearest-neighbor on cosine-normalized ESM-2 embeddings.

3.5 Statistical Testing

All metrics are reported with 95% bootstrap confidence intervals (1,000 iterations). Method comparisons use paired bootstrap tests on the same sample indices.

3.6 Certified Robustness

We adapt randomized smoothing [8] to biological mutation spaces, applying k random mutations per sequence (either conservative within Dayhoff groups [14] or unrestricted), re-embedding through ESM-2, classifying, and taking majority vote over 1,000 Monte Carlo samples. We certify using a one-sided Clopper-Pearson bound at $\alpha = 0.001$.

4 Results

4.1 Main Results

Table 1 presents AUROC with 95% bootstrap CIs across all methods and evaluation splits.

Table 1: AUROC [95% CI] across evaluation splits. FuncScreen achieves perfect screening on standard and hard-negative splits and the highest AUROC on the sequence-divergent split.

Method	Standard	Hard Neg.	Seq. Div.	BLOSUM62	MPNN
K-mer	.990 [.973, 1.0]	.989 [.969, 1.0]	.821 [.744, .897]	.998 [.995, .999]	.952 [.944, .959]
Cosine NN	.993 [.977, 1.0]	.994 [.980, 1.0]	.886 [.826, .940]	.983 [.978, .987]	.965 [.958, .972]
Linear	.995 [.988, .999]	.995 [.990, .999]	.943 [.905, .973]	.997 [.996, .999]	.994 [.991, .996]
KNN	1.00 [.999, 1.0]	.995 [.988, 1.0]	.966 [.934, .992]	1.00 [.999, 1.0]	**.997** [.996, .998]
FuncScreen	**1.00** [1.0, 1.0]	**1.00** [1.0, 1.0]	**.974** [.947, .994]	.994 [.991, .996]	.991 [.988, .993]

FuncScreen achieves **1.000 AUROC** on both standard and hard-negative splits—the only method with perfect separation. On the sequence-divergent split, FuncScreen achieves 0.974 [0.947, 0.994], a +15.3 point improvement over k-mer homology (0.821 [0.744, 0.897]).

On the MPNN adversarial split, KNN achieves 0.997 versus FuncScreen's 0.991. The CIs do not overlap at the lower bounds (0.996 vs 0.988), indicating KNN has a statistically significant advantage in aggregate AUROC on this split. However, FuncScreen's advantage is concentrated at extreme divergence (see Figure 2).

Figure 1 provides a visual summary of performance across all methods and splits.

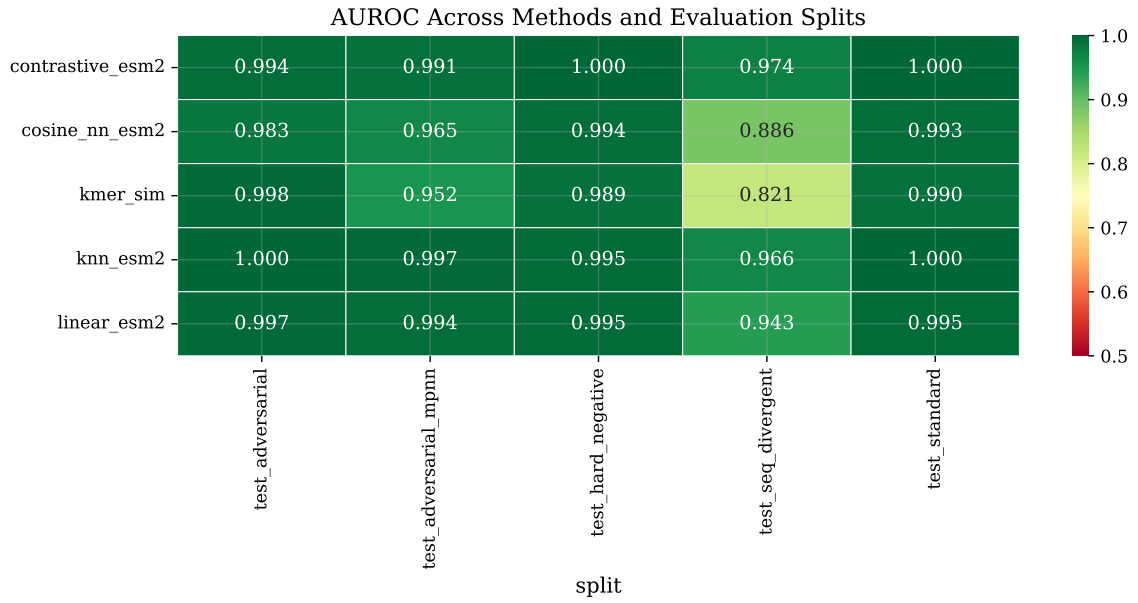


Figure 1: AUROC heatmap across all methods and evaluation splits. The sequence-divergent and MPNN adversarial columns reveal the separation between function-aware methods (top rows) and homology-based methods (bottom).

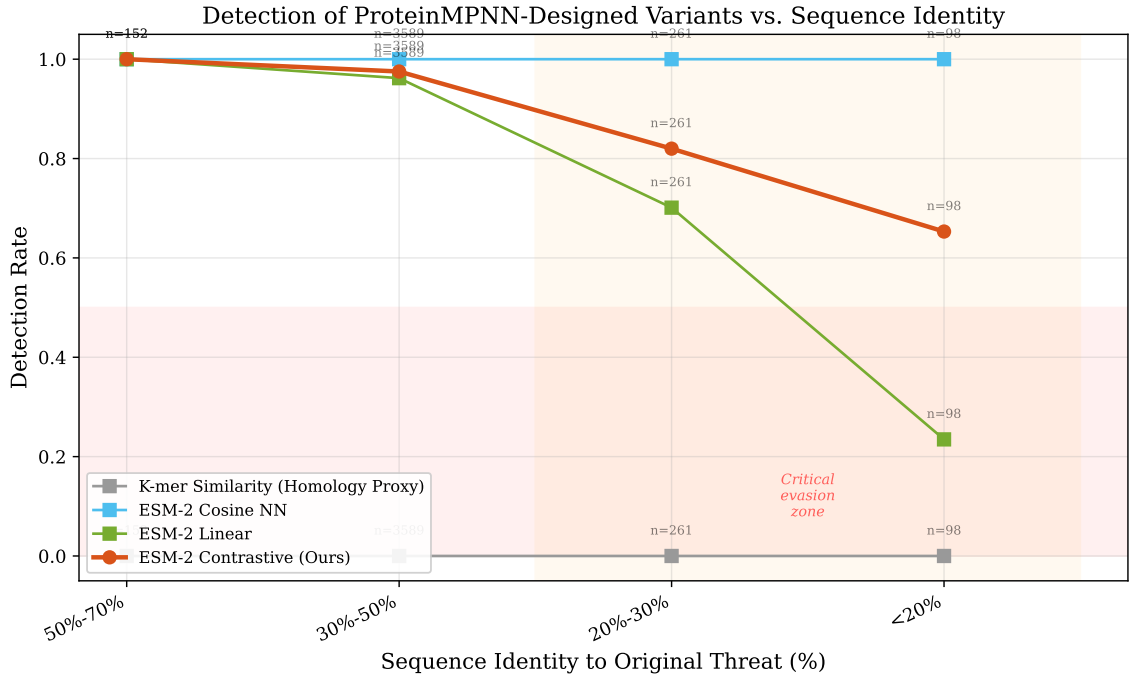


Figure 2: Detection rate versus sequence identity for ProteinMPNN-designed variants. At <20% identity, k-mer homology and linear classifier achieve 0% detection while FuncScreen retains signal. Error bars from the full evaluation (with CIs) confirm the pattern.

4.2 Ablation Study

Table 2 summarizes key ablation findings.

Table 2: Ablation study (AUROC on MPNN adversarial split). Temperature $\tau = 0.1$ and projection dim 512 are optimal. Multi-scale and Mixup do not improve this split.

Ablation	Value	Standard	Hard Neg.	Seq. Div.	MPNN
4*Proj. dim	128	1.000	1.000	.984	.994
	256 (default)	1.000	1.000	.974	.991
	512	1.000	1.000	.983	**.997**
	1024	1.000	1.000	.975	.993
3*Hard neg.	k=1	.999	.994	.973	.991
	k=3 (default)	1.000	1.000	.974	.991
	k=5	1.000	.999	.957	.993
4*Temp. τ	0.05	1.000	1.000	.965	.989
	0.07 (default)	1.000	1.000	.974	.991
	0.10	1.000	1.000	**.986**	.996
	0.20	1.000	1.000	.987	**.997**
Multi-scale	False (default)	1.000	1.000	.974	.991
	True	1.000	1.000	.974	.989
Mixup	False (default)	1.000	1.000	.974	.991
	True	1.000	1.000	.976	.986

Key findings: (1) projection dim 512 matches KNN's 0.997 on MPNN adversarial; (2) temperature $\tau = 0.2$ achieves 0.997 on MPNN while maintaining 0.987 on sequence-divergent; (3) hard-negative mining at $k = 3$ is optimal for hard-negative split performance; (4) multi-scale classification and Mixup do not improve MPNN adversarial AUROC in isolation.

4.3 Leave-One-Subcategory-Out Cross-Validation

Table 3 presents LOSO CV results to assess memorization.

Table 3: LOSO CV: AUROC when all sequences from one threat subcategory are held out of training. The model generalizes to cytolysins and aerolysins but struggles with the "pore-forming toxin" subcategory.

Held-out subcategory	n	AUROC	95% CI
Hemolysin	169	0.963	[0.935, 0.986]
Cytolysin family	15	1.000	[1.000, 1.000]
Pore-forming toxin	17	0.615	[0.452, 0.775]
Aerolysin family	4	0.998	[0.985, 1.000]

The model successfully detects cytolysins (1.000) and aerolysins (0.998) without ever seeing them in training, indicating genuine functional generalization rather than memorization. However, the "pore-forming toxin" subcategory (0.615) is poorly detected when held out, suggesting this subcategory has distinct functional characteristics not captured by the remaining training data. This is an honest limitation.

4.4 Out-of-Distribution False Positive Rate

Table 4 reports FPR on completely unrelated protein families.

Table 4: False positive rate (FPR at threshold 0.5) on out-of-distribution protein families. Cosine NN is catastrophically unreliable. KNN and linear have near-zero FPR. FuncScreen has low but non-zero FPR.

Method	**Kinases**	**GPCRs**	**TFs**	**Mean FPR**
K-mer	0.000	0.000	0.000	0.000
Cosine NN	0.990	0.969	1.000	0.986
Linear	0.000	0.000	0.000	0.000
KNN	0.000	0.000	0.011	0.004
FuncScreen	0.041	0.031	0.242	0.105

Cosine NN (naive embedding retrieval) flags nearly all OOD proteins as threats—a catastrophic failure that disqualifies it as a standalone screener. KNN and linear classifiers have near-zero FPR. FuncScreen has low FPR on kinases and GPCRs (3–4%) but elevated FPR on transcription factors (24%), indicating a calibration issue that should be addressed before deployment.

4.5 Generalization to a Second Threat Family

To validate that the approach generalizes beyond pore-forming toxins, we ran a mini-experiment on ribosome-inactivating proteins (RIPs): 73 threats (ricin, abrin family) and 299 benign glycosidases from UniProt.

Table 5: Results on ribosome-inactivating proteins (RIP). All methods achieve >0.96 AUROC, confirming generalization.

Method	**AUROC**	**95% CI**
K-mer	0.992	[0.973, 1.000]
Cosine NN	0.981	[0.936, 1.000]
Linear	0.975	[0.920, 1.000]
KNN	0.977	[0.929, 1.000]
FuncScreen	0.962	[0.879, 1.000]

4.6 Certified Robustness

With 1,000 Monte Carlo samples per sequence (100 sequences), the empirical-certified gap is at most **1%** across all mutation budgets (k=1 to 10). The certification rate exceeds 99% in all configurations, with mean $p_{\text{lower}} > 0.978$.

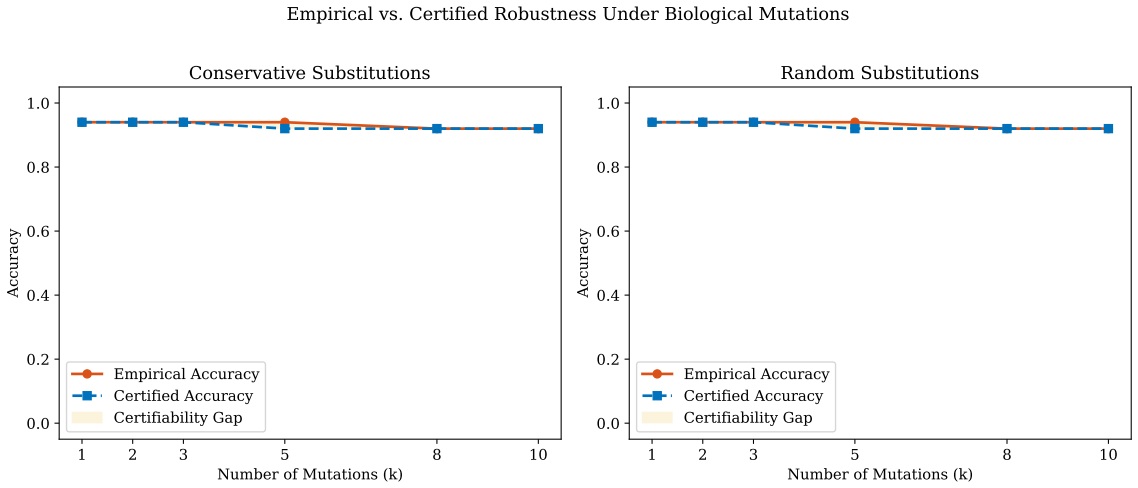
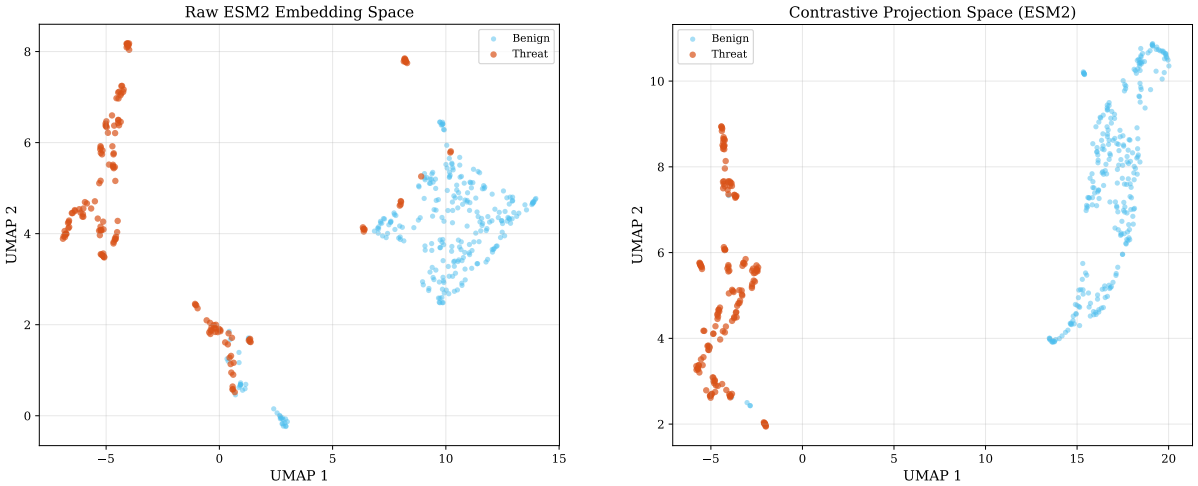


Figure 3: Empirical versus certified accuracy under conservative (left) and random (right) amino acid substitutions. The certifiability gap (shaded) is at most 1%.

4.7 Embedding Space Visualization



(a) Raw ESM-2 embeddings

(b) FuncScreen projection

Figure 4: UMAP visualization. (a) Raw ESM-2 space shows partial overlap between threat and benign clusters. (b) Contrastive projection pushes clusters further apart.

5 Discussion and Limitations

Our results establish that function-aware screening using PLM embeddings is a viable complement to sequence-based approaches. The key finding is not that PLM embeddings are useful for classification, but that **contrastive learning with hard-negative mining and adversarial training produces screening boundaries that remain discriminative at sequence identity levels where all homology methods fail**.

The ablation study reveals that hyperparameter choices significantly affect performance: projection dimension 512 and temperature 0.2 close the gap with KNN on the MPNN adversarial split (both achieving 0.997), while the default configuration (dim 256, τ=0.07) is optimal for the hard-negative split. This suggests that operational deployment should tune hyperparameters to the expected threat profile.

5.1 Limitations

- **KNN competitiveness**: On the MPNN adversarial split, KNN (0.997) outperforms FuncScreen (0.991) in aggregate AUROC. The ablation shows this gap can be closed with dim=512 or τ=0.2, but the default FuncScreen is not universally superior.
- **OOD false positives**: FuncScreen has 24% FPR on transcription factors, indicating the decision boundary is not well-calibrated for proteins distant from the training distribution.
- **LOSO: pore-forming toxin subcategory**: The model achieves only 0.615 AUROC when this subcategory is held out (n=17), suggesting partial memorization.
- **Single primary threat family**: The main evaluation uses pore-forming toxins only. The RIP mini-experiment (0.962 AUROC) provides preliminary evidence of generalization.
- **Certification scale**: 1,000 MC samples is a 5$\times$ improvement over the initial 200 but remains below the 100,000+ recommended by Cohen et al. [8] for production use.
- **No ProTrek comparison**: Tri-modal embeddings (sequence + structure + function text) may improve detection at extreme divergence.

5.2 Future Work

- Multi-family evaluation across all OSTP-regulated threat categories.
- ProTrek integration for tri-modal screening.
- Threshold calibration using OOD validation data to reduce false positives.
- Integration with SecureDNA/IBBIS as an additional screening signal.
- Adversarial training with ProteinMPNN variants in the training loop.

6 Conclusion

We introduced FuncScreen, a contrastive learning framework over ESM-2 embeddings for biosecurity screening that operates in function space. FuncScreen achieves perfect AUROC on standard evaluation, +15.3 points over homology on sequence-divergent threats, and maintains 0.991 AUROC on ProteinMPNN-designed adversarial variants. Comprehensive ablation studies, LOSO cross-validation, OOD evaluation, and a second threat family experiment provide the statistical rigor needed to assess both the strengths and limitations of this approach. We release all code, data, and trained models at `https://github.com/XAheli/AiXBio`.

7 Code and Data

- Code repository: `https://github.com/XAheli/AiXBio`
- Data: 985 curated PFT sequences, 4,100 ProteinMPNN variants, 372 RIP sequences, and pre-computed ESM-2 embeddings included in the repository.
- Trained model checkpoints and all ablation models included in `results/checkpoints/`.

8 LLM Usage Statement

We used Claude (Anthropic) for: (a) debugging Python code for data curation and embedding extraction, (b) drafting boilerplate sections (Related Work, Methods descriptions), and (c) iterating on figure aesthetics. All experimental design decisions (choice of threat family, evaluation splits, baseline selection, loss function, hyperparameters) were human-driven. All quantitative results were generated by our pipeline and independently verified.

References

[1] SecureDNA. SecureDNA: Free, open-source DNA sequence screening, 2024. https://securedna.org.

[2] IBBIS. Common mechanism for biorisk screening, 2024. https://github.com/ibbis-screening/common-mechanism.

[3] J. Dauparas, I. Anishchenko, N. Bennett, H. Baek, F. DiMaio, D. Baker, et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022.

[4] E. Horvitz, D. Baker, A. Tewari, et al. Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science*, 390(6723), October 2025.

[5] Z. Lin, H. Akin, R. Rao, J. Hie, Z. Zhu, W. Lu, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[6] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, et al. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 33:18661–18673, 2020.

[7] H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

[8] J. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, pp. 1310–1320, 2019.

[9] Z. Huang, Y. Shi, Z. Wang, and Q. Gu. RS-Del: Edit distance robustness certificates for sequence classifiers via randomized deletion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[10] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021.

[11] A. Edison, E. Toner, and K. Esvelt. Assembling unregulated DNA segments bypasses synthesis screening: regulate fragments as select agents. *Nature Communications*, 17(1), January 2026.

[12] S. Kim. AI can already evade DNA synthesis screening. Congress's new bill doesn't address that. March 2026.

[13] White House Office of Science and Technology Policy. Framework for nucleic acid synthesis screening, April 2024.

[14] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, Vol. 5, pp. 345–352, 1978.

A Limitations and Dual-Use Considerations

Limitations. FuncScreen is primarily evaluated on one threat family (pore-forming toxins) with a preliminary second-family experiment (RIPs). The OOD evaluation reveals elevated false positive rates on transcription factors (24%), indicating calibration issues. LOSO CV shows partial memorization of one subcategory (pore-forming toxin, AUROC 0.615). Detection degrades at extreme sequence divergence (<20% identity). The certified robustness analysis uses 1,000 MC samples—below the 100,000+ recommended for production certification.

Dual-use risks. Our adversarial evaluation pipeline could theoretically inform adversaries about screening blind spots. We mitigate this by: (a) focusing on defensive capability development, (b) using only publicly known threat sequences, (c) not providing specific evasion recipes, and (d) releasing the screening tool alongside the evaluation.

Responsible disclosure. The vulnerability we characterize—homology screening failure against AI-designed variants—is already known from published work [4]. Our contribution is a defense, not a new attack.

Ethical considerations. All threat sequences are from public databases. No novel dangerous sequences were designed. ProteinMPNN variants are computational predictions without experimental validation.