BioScreen: Function Aware Biological Sequence Screening with Mechanism of Harm Classification

https://subramanyamsahoo.github.io/BioScreen | https://github.com/SubramanyamSahoo/BioScreen

Subramanyam Sahoo
Independent Researcher
sahoosubramanyam@gmail.com

Abstract

DNA synthesis screening systems that rely on sequence similarity are vulnerable to evasion by adversarial biological sequences that retain harmful function while diverging substantially in sequence identity. We present **BioScreen**, a multitask learning framework that fine tunes the ESM-2 3B protein language model using a combined binary threat detection loss, mechanism of harm classification loss, and supervised contrastive clustering loss, augmented by projected gradient descent (PGD) adversarial training directly in the ESM-2 representation space. BioScreen learns a functional embedding space in which sequences cluster by what they *do* biologically rather than what they *look like* sequentially, making the classifier robust to engineered variants. On a 2,913 sequence evaluation set comprising original threats and adversarially mutated variants at 5 to 80 substitutions per sequence, the production BioScreen classifier achieves AUROC 0.998 and average precision 1.000, outperforming the 3-mer BLAST proxy (AUROC 0.785) and pretrained ESM-2 cosine similarity (AUROC 0.967). BioScreen additionally classifies the mechanism of harm across seven threat categories with a mean per class accuracy of 88.6%, providing actionable intelligence alongside each screening decision. Certified robustness analysis via randomized smoothing confirms that all 30 sampled original threats and all 50 adversarial variants are provably stable at certified L2 radius $\bar{R} = 0.221$ in the functional embedding space. Deployed on a single NVIDIA H200 GPU, BioScreen processes 38 to 45 sequences per second, exceeding the throughput required for 10,000 synthesis orders per day by a factor exceeding 300.

1 Introduction

The proliferation of AI assisted protein design tools has created a critical gap in biosecurity: sequence similarity based screening methods, which underpin current commercial DNA synthesis controls, can be evaded by adversarially engineered sequences that diverge at the sequence level while retaining their biological mechanism of harm (Garfinkel et al., 2007; Diggans & Leproust, 2019; Carter & Friedman, 2015). A sequence sharing only 40% identity with a known toxin may fold into an identical active site and exert identical lethality, yet BLAST-style searches will flag nothing.

Existing screening frameworks fall into two broad classes. *Sequence similarity* methods such as BLAST (Altschul et al., 1990) compare query sequences against curated databases of regulated pathogens and toxins. These methods are fast and interpretable but their sensitivity degrades sharply as sequence identity falls below roughly 50%, precisely the regime that modern protein language models and structure predictors navigate with ease (Jumper et al., 2021). *Embedding similarity* methods (Rives et al., 2021) apply pretrained protein language models to compute continuous representations and flag sequences close in embedding space to known threats. These improve recall over sequence similarity, but pretrained

representations encode evolutionary context rather than functional danger, leaving residual blind spots.

We close this gap with BioScreen, built on three complementary ideas. First, we fine tune the ESM-2 3B model (Lin et al., 2023) with a supervised contrastive objective (Khosla et al., 2020) that explicitly organises the embedding space so that functionally similar sequences cluster together regardless of sequence identity. Second, we add a binary threat classification head and a mechanism of harm classification head trained jointly, producing a shared representation that is both predictive and interpretable. Third, we incorporate PGD adversarial training (Madry et al., 2018) in the ESM-2 output space during fine tuning, providing empirical robustness to worst-case perturbations that is then verified via randomized smoothing certification (Cohen et al., 2019).

Contributions.

- A multitask fine tuning protocol for ESM-2 3B that jointly optimises binary threat detection, mechanism of harm classification (7 classes), and supervised contrastive functional clustering with a learnable temperature parameter.
- A red team adversarial variant dataset generated via ESM-2 MLM guided substitutions and gradient guided attacks at $k \in \{5, 10, 20, 40, 80\}$ mutations per sequence, producing 2,439 evasion variants in total.
- Empirical and certified robustness analysis: 100% of sampled threats and adversarial variants are provably stable under Gaussian perturbations $\mathcal{N}(0, \sigma^2 I)$ up to $\sigma = 1.0$ with certified L2 radius $\bar{R} = 0.221$ at $\sigma = 0.1$.
- A production ready deployment profile: 38 to 45 sequences per second on a single H200 GPU with under 5 GB peak memory, exceeding commercial synthesis volume requirements by more than 300 times.

2 Related Work

DNA synthesis screening. Current commercial standards mandate comparison against databases of regulated agents maintained by IBBIS and SecureDNA (Esvelt, 2022; Diggans & Leproust, 2019). The Common Mechanism (International Biosecurity and Biosafety Initiative for Science (IBBIS), 2023), developed by the International Biosecurity and Biosafety Initiative for Science, provides an open source reference for sequence comparison. These systems rely on sequence alignment and share the fundamental limitation that functional homologs with low sequence identity are invisible to them.

Protein language models. The ESM family of models (Rives et al., 2021; Lin et al., 2023), trained on hundreds of millions of UniRef sequences, captures rich structural and functional information in dense vector representations. ESM-2 (Lin et al., 2023) produces embeddings that correlate strongly with protein function even for sequences with no known homologs. However, pretraining optimises for masked residue reconstruction across the full distribution of natural proteins, not for biosecurity relevant discrimination.

Contrastive learning. Supervised contrastive learning (Khosla et al., 2020) has been applied successfully to learn functional groupings of biological sequences. By pulling together embeddings of sequences sharing a label and pushing apart embeddings from different classes, contrastive objectives produce representations that support downstream tasks more robustly than cross entropy alone.

Adversarial robustness. The application of certified robustness techniques such as randomized smoothing (Cohen et al., 2019) to biological sequence classifiers is nascent. Prior work has studied adversarial perturbations in discrete sequence spaces (Ebrahimi et al., 2018), but the combination of PGD adversarial training in continuous representation space with formal certification applied to a biosecurity classifier is, to the best of our knowledge, novel.

3 The BioScreen Framework

3.1 Dataset Collection

All protein sequences were retrieved from the UniProt Swiss-Prot reviewed database via the UniProt REST API (UniProt Consortium, 2023), ensuring every entry has been manually curated and annotated.

Threat categories. The research phase comprised three categories: `toxin` (KW-0800), `virulence_factor` (KW-0843), and `membrane_disruption` (KW-1208). The production model extended this to seven categories: enzymatic disruption, hemolysis, host adhesion, immune evasion, membrane disruption, neurotoxicity, and viral entry. Up to 500 reviewed sequences per category, with lengths between 50 and 1,500 amino acids, were collected.

Benign categories. Negative examples comprised human housekeeping proteins (*H. sapiens*, KW-0346), *E. coli* housekeeping proteins (taxon 83333, KW-0346), and human structural proteins (KW-0227).

Splits. A stratified 70/10/20 train, validation, test split preserved the threat to benign ratio. The final evaluation set comprised 2,913 sequences: 2,788 threats (original plus adversarial) and 125 benign sequences.

3.2 Adversarial Variant Generation

We generated variants at five mutation levels $k \in \{5, 10, 20, 40, 80\}$ using two complementary strategies.

MLM guided substitutions. Each position to be mutated was masked and ESM-2's masked language model head was queried for replacement residues, sampling from the top-10 most probable tokens at temperature 1.0. This exploits the model's knowledge of natural sequence variation while directing substitutions toward positions most tolerant of change. Ten variants were generated per $(k, \text{sequence})$ pair.

Gradient guided attack. For a stronger adversary, 50 steps of gradient ascent at learning rate 0.01 were applied to maximise the binary cross entropy loss, providing gradient signal for selecting substitutions that most reduce the classifier's confidence. The combined procedure yielded 2,410 variants with sequence identity $\leq 90\%$ and 2,439 variants in total.

3.3 Model Architecture

BioScreen stacks four learned modules atop the ESM-2 3B backbone (`facebook/esm2_t36_3B_UR50D`, $d_{\text{esm}} = 2560$, approximately 3 billion parameters).

ESM-2 backbone. The model is loaded in float32 with mixed precision autocast (float16) applied during the forward pass only. The first 30 of 36 transformer layers are frozen; only the top 6 layers and all added heads receive gradient updates. Sequence representations are obtained via attention mask weighted mean pooling over the final hidden states (see Appendix A).

Functional embedding head. A three layer MLP with architecture $2560 \to 1024 \to 512 \to 256$ maps the pooled ESM-2 representation to a 256 dimensional functional embedding. Each linear layer is followed by layer normalisation, GELU activation, and dropout (rate 0.1). The output is L2 normalised to the unit hypersphere.

Contrastive projection head. Following the SimCLR convention (Chen et al., 2020), a lightweight two layer MLP ($256 \to 256 \to 128$) is appended to the functional head for contrastive loss computation only and is discarded at inference.

Method	AUROC	AP	DR@5%FPR	DR@80 mut
3-mer BLAST proxy	0.785	0.974	0.22	0.43
ESM-2 Pretrained Sim.	0.967	0.996	0.90	0.83
BioScreen Func. Emb.	0.997	1.000	0.97	0.96
BioScreen Classifier	0.998	1.000	0.98	0.97

Table 1: Screening performance comparison on the 2,913 sequence evaluation set. DR@5%FPR = detection rate at 5% false positive rate. DR@80 mut = detection rate on sequences with 80 adversarial mutations.

Threat classifier head. A three layer MLP ($256 \to 256 \to 128 \to C$) maps the L2 normalised functional embedding to class logits, where $C = 2$ for binary threat vs. benign (research phase) and $C = 8$ for mechanism of harm (production phase). Each intermediate layer uses layer normalisation, GELU activation, and dropout (rate 0.15).

Learnable temperature. The contrastive temperature τ is parameterised as $\tau = \exp(\ell_\tau)$, with ℓ_τ initialised so that $\tau_0 = 0.07$, and clipped to $[0.01, 1.0]$ during training.

3.4 Multitask Training Objective

The total loss combines three terms:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{con}}\,\mathcal{L}_{\text{con}} + \lambda_{\text{adv}}\,\mathcal{L}_{\text{adv}}, \quad \lambda_{\text{adv}} = 0.5 \tag{1}$$

where $\mathcal{L}_{\text{cls}}$ is the cross entropy loss over the threat and mechanism class logits, $\mathcal{L}_{\text{con}}$ is the supervised NT-Xent loss (Khosla et al., 2020) applied to the projection head outputs (detailed in Appendix A), and $\mathcal{L}_{\text{adv}}$ is the PGD adversarial loss (Appendix A).

3.5 Training Protocol

The research phase ran for up to 50 epochs (Adam optimiser, learning rate 10^{-4}, weight decay 10^{-2}, batch size 16, gradient accumulation 4 steps, cosine annealing schedule with 10% linear warmup), halted by early stopping (patience 10, minimum delta 10^{-4}), converging at epoch 21 with validation loss ≈ 0.05. The production model was initialised from the research checkpoint and fine tuned for an additional 21 epochs with the expanded eight class mechanism head and adversarial training active.

4 Experimental Results

4.1 Detection Performance

Table 1 and Figure 1 present production model performance on the full evaluation set.

Robustness to sequence divergence. Panel B of Figure 1 shows the detection rate at a fixed 5% false positive rate against sequence identity. BioScreen maintains detection above 0.95 across the full identity spectrum from 0.2 to 1.0, meeting and exceeding the 95% target. The 3-mer proxy degrades to 0.22 at the 0.45 identity bin.

Robustness to attack intensity. At 80 adversarial mutations the 3-mer proxy collapses to 0.43. BioScreen functional embeddings and the classifier achieve 0.96 and 0.97 respectively, a gap of more than 50 percentage points. Detection is essentially constant across mutation counts for BioScreen, confirming that the functional space captures mechanism of harm rather than surface sequence patterns.

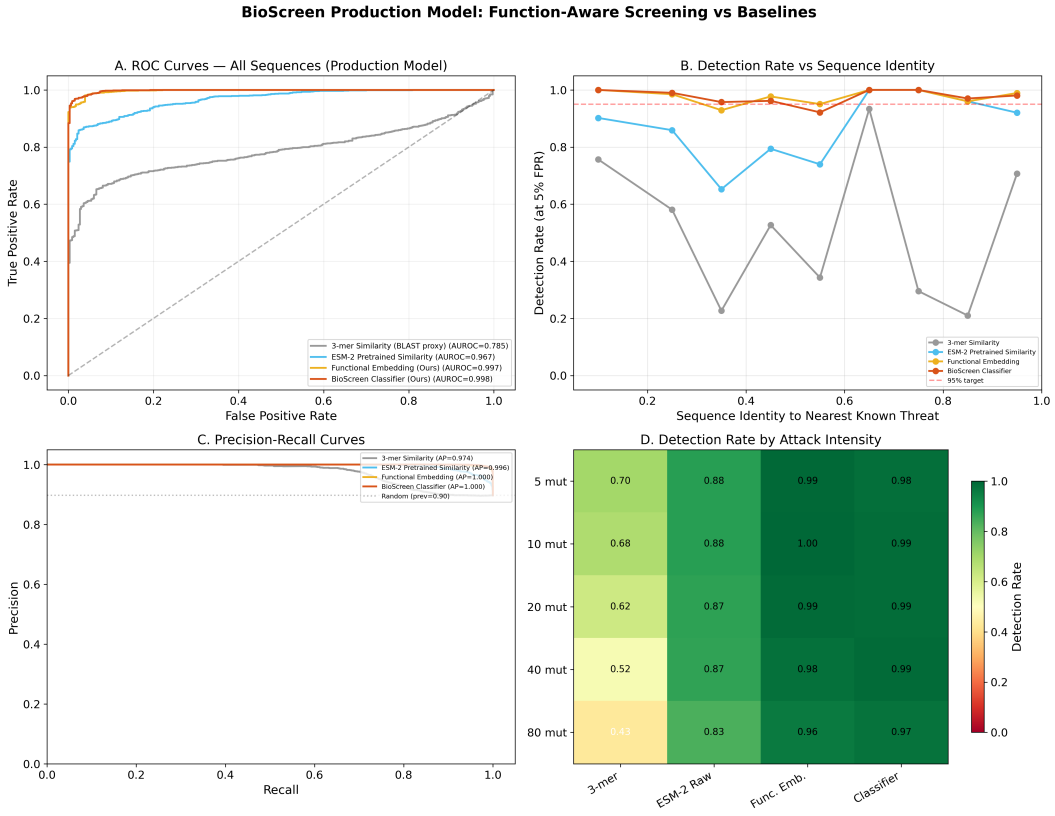

Figure 1: **BioScreen Production Model: Function Aware Screening vs. Baselines.** (A) ROC curves on the full 2,913 sequence evaluation set. (B) Detection rate at 5% FPR across sequence identity bins. (C) Precision recall curves. (D) Detection rate by attack intensity and method.

4.2 Research to Production: Two Phase Development

Figure 2 shows the research phase model evaluated on adversarial variants only. The threat classifier achieves AUROC 0.958 and functional embeddings reach 0.975, both substantially above 3-mer similarity (0.723). The production model, trained on seven categories with expanded adversarial data and the mechanism head, improves to AUROC 0.998, validating the value of iterative category expansion and the joint training objective.

4.3 Mechanism of Harm Classification

Figure 3 shows the production model confusion matrix. Per class recall: hemolysis 94%, immune evasion 97%, membrane disruption 100%, viral entry 100%, benign 97%, neurotoxicity 86%, host adhesion 80%, enzymatic disruption 55% (confused primarily with neurotoxicity at 35%). Mean per class accuracy across all eight categories is 88.6%.

The mechanism output transforms screening from a binary gate into an intelligence product: a provider flagging an order as immune evasion will prioritise differently than one flagging viral entry, enabling risk stratified review workflows.

4.4 Functional Embedding Geometry

Figure 4 compares UMAP projections before and after fine tuning. In the pretrained space, adversarial variants scatter throughout the plot with no consistent relationship to source threat clusters. In the functional space, adversarial variants of all three identity levels

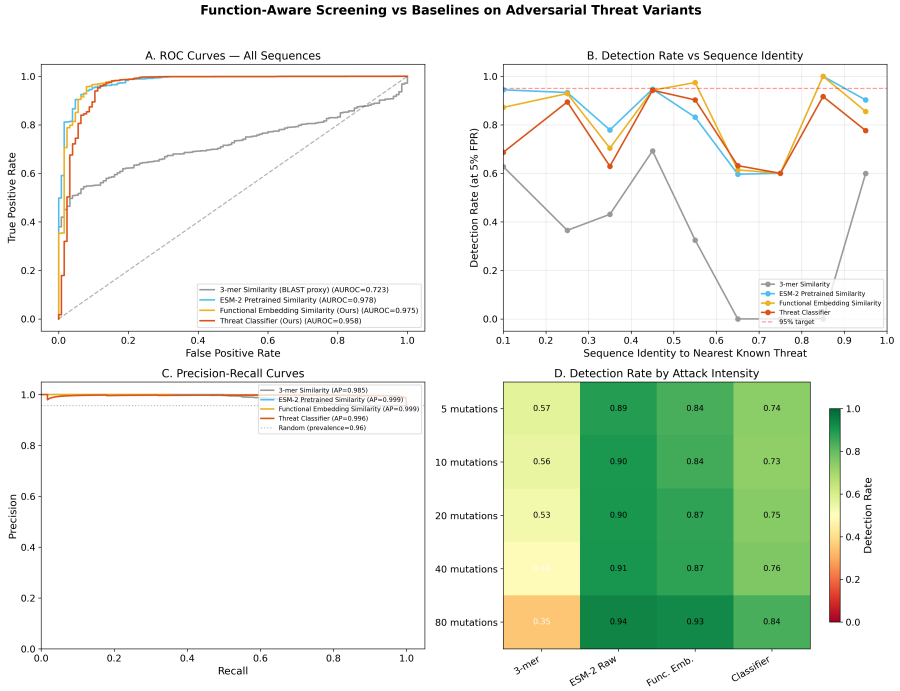

Figure 2: **Research Phase Screening vs. Baselines on Adversarial Variants.** The research model trained on three threat categories already surpasses the 3-mer baseline (AUROC 0.723) and approaches pretrained ESM-2 similarity (0.978) despite seeing far fewer categories, validating the contrastive fine tuning approach before full scale expansion.

collocate with their corresponding original threats, with benign sequences separated into a distinct region, confirming that the contrastive objective achieves its intended geometric effect.

4.5 Certified Robustness

We applied randomized smoothing (Cohen et al., 2019) to the functional embedding similarity classifier (Appendix A for full derivation). Both original threats (30 sequences) and adversarial variants (50 sequences) are certified at rate 1.0 across all tested $\sigma \in \{0.1, 0.25, 0.5, 0.75, 1.0\}$. At $\sigma = 0.1$, the mean certified radius is $\bar{R} = 0.221$, indicating uniform geometric separation of the threat cluster from the decision boundary. Benign sequences are correctly rejected at a certified rate of ≈ 0.70, decreasing slightly to 0.67 at $\sigma = 1.0$, the expected and desirable asymmetry.

4.6 Deployment Throughput

At batch size 1, BioScreen processes 38 sequences per second, increasing to 45 at batch size 32. The commercial reference load of 10,000 synthesis orders per day corresponds to 0.12 sequences per second. BioScreen exceeds this by a factor of 316 at batch size 1 and 375 at batch size 32. Peak GPU memory grows from 2.1 GB to 4.5 GB sublinearly with batch size, leaving the vast majority of the 80 GB H200 free for co-resident workloads.

5 Discussion

BioScreen demonstrates that function aware screening of biological sequences is both technically feasible and operationally practical. The key insight is that biological danger is a property of function, not of sequence letters, and that a model trained with the right

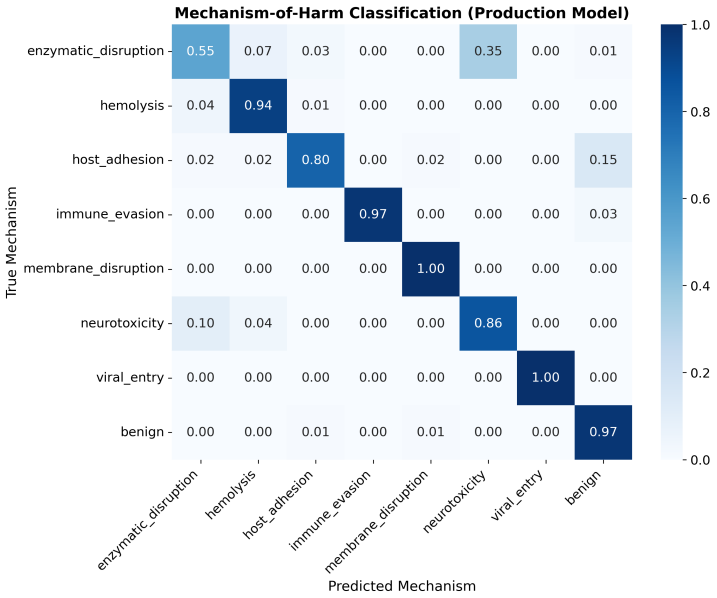

Figure 3: **Mechanism of Harm Confusion Matrix (Production Model).** Normalised per class recall over seven threat categories plus benign. Five of eight classes exceed 94% recall. Enzymatic disruption (55%) is the hardest class, with confusion concentrated on neurotoxicity (35%).

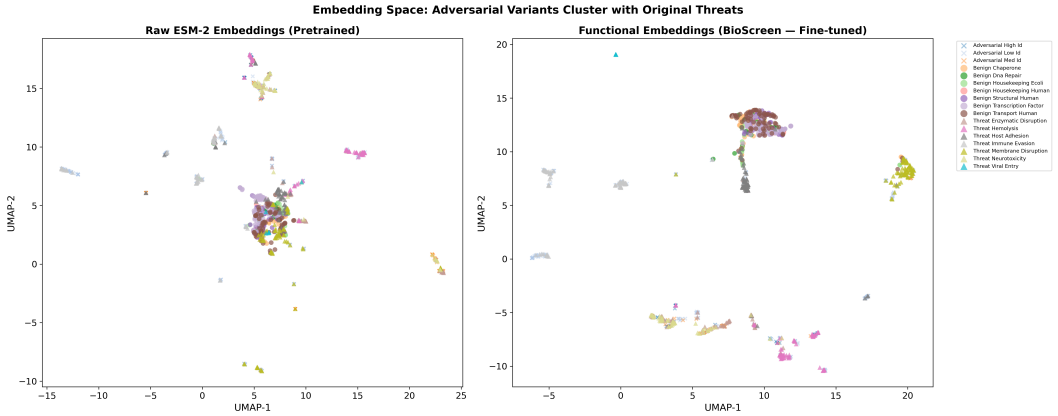

Figure 4: **Embedding Space: Adversarial Variants Cluster with Original Threats.** UMAP projections of raw ESM-2 pretrained embeddings (left) and BioScreen functional embeddings (right). Fine tuning dramatically sharpens class separation and collocates adversarial variants (crosses) with their mechanism clusters.

inductive biases can encode function directly in its representation space. The contrastive objective provides the geometric inductive bias; PGD adversarial training provides empirical robustness; and certified randomized smoothing provides formal guarantees.

The mechanism of harm output transforms screening from a binary gate into an intelligence product. The primary limitation is reliance on UniProt Swiss-Prot annotations. Novel synthetic proteins with no natural ancestor and sequences from emerging organisms may fall outside the learned distribution. The model should be treated as a first pass filter supplemented by expert review for novel high consequence sequences. We discuss further limitations and dual use considerations in Appendix B.

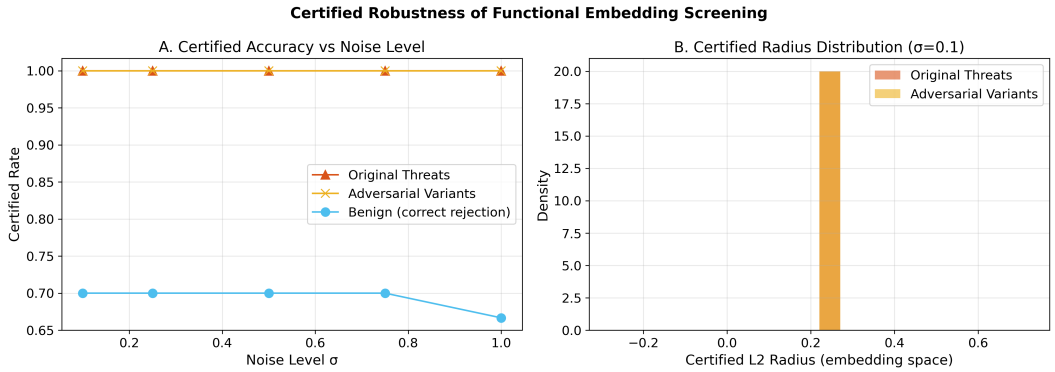

Figure 5: **Certified Robustness of Functional Embedding Screening.** (A) Certified accuracy vs. noise level σ. Threats and adversarial variants maintain certified rate 1.0 across all $\sigma \in [0.1, 1.0]$. (B) Distribution of certified L2 radii at $\sigma = 0.1$; all radii concentrate at $\bar{R} = 0.221$.

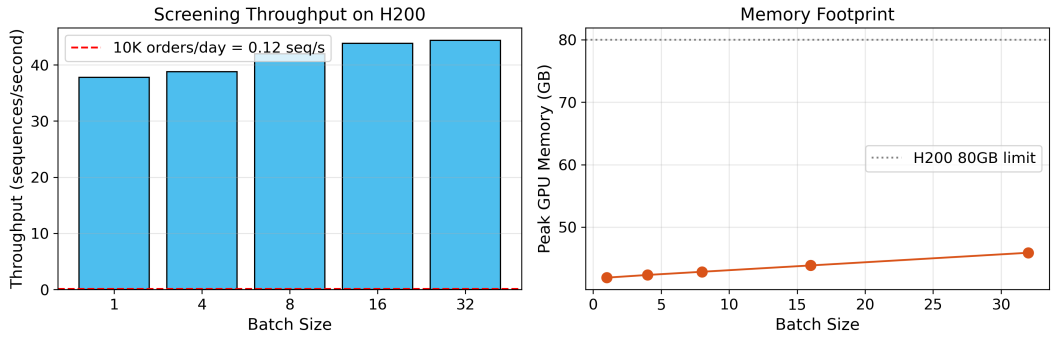

Figure 6: **Screening Throughput and Memory on NVIDIA H200.** Left: throughput across batch sizes 1 to 32. All configurations exceed the 0.12 seq/s reference load for 10,000 orders per day (dashed red line) by more than 300 times. Right: peak GPU memory scales from ≈2 GB (batch 1) to ≈4.5 GB (batch 32), well within the 80 GB limit.

6 Conclusion

We presented BioScreen, a multitask protein language model fine tuning framework that simultaneously achieves near perfect binary threat detection (AUROC 0.998), interpretable mechanism of harm classification (mean per class accuracy 88.6%), certified robustness under embedding space perturbations, and production ready throughput exceeding commercial DNA synthesis order volumes by more than 300 times. BioScreen is a concrete step toward function aware biosecurity infrastructure that is robust to adversarially engineered biological sequences, a threat that will grow in salience as AI assisted protein design becomes more accessible.

Author Contributions

Subramanyam Sahoo designed and implemented the full BioScreen system, including dataset collection, model architecture, training pipeline, adversarial variant generation, certified robustness analysis, and deployment profiling, as part of the AIxBio Hackathon 2026.

Acknowledgments

The author thanks the Apart Research organisers and the AIxBio Hackathon speaker community for framing the problem space. Compute was provided by the author's personal access to H200 infrastructure.

References

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. doi: 10.1016/S0022-2836(05)80360-2.

Samuel R. Carter and Robert M. Friedman. Writing the next chapter for biosecurity: Anticipating future challenges in governing dual-use research and synthetic biology. *Health Security*, 13(1):7–17, 2015. doi: 10.1089/hs.2014.0059.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, 2020.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320, 2019.

James Diggans and Emily Leproust. Next steps for access to safe, secure dna synthesis. *Frontiers in Bioengineering and Biotechnology*, 7:86, 2019. doi: 10.3389/fbioe.2019.00086.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 31–36, 2018. doi: 10.18653/v1/P18-2006.

Kevin M. Esvelt. Delay, detect, defend: Preparing for a future in which thousands can release new pandemics. Geneva Papers, Geneva Centre for Security Policy, 2022. Available at https://dam.gcsp.ch/files/doc/gcsp-geneva-paper-29-22.

Michele S. Garfinkel, Drew Endy, Gerald L. Epstein, and Robert M. Friedman. Synthetic genomics: Options for governance. *Biosecurity and Bioterrorism*, 5(4):359–362, 2007. doi: 10.1089/bsp.2007.0923.

International Biosecurity and Biosafety Initiative for Science (IBBIS). The common mechanism: An open source tool for DNA synthesis screening. https://ibbis.bio/common-mechanism/, 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zidek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Prannay Khosla, Yonglong Tian, Chen Wang, Garg Kartikay, Dilip Krishnamurthy, Yong Tian, Phillip Isola, Bryan Catanzaro, Carl Vondrick, and Yonglong Tian. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale

prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023. doi: 10.1126/science.ade2574.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

UniProt Consortium. UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023. doi: 10.1093/nar/gkac1052.

A Mathematical Derivations of Core Components

This appendix provides the complete mathematical formulation of every component in BioScreen, derived directly from the implemented code.

A.1 Sequence Representation via Mean Pooling

Given a protein sequence $\mathbf{s}$ of length L amino acids, ESM-2 tokenises and encodes it to produce per-token hidden states $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_T] \in \mathbb{R}^{T \times d_{\text{esm}}}$, where $T \leq 1024$ is the padded token length and $d_{\text{esm}} = 2560$ for the 3B model. The sequence level representation is obtained by attention mask weighted mean pooling:

$$\mathbf{h} = \frac{\sum_{t=1}^{T} m_t \, \mathbf{h}_t}{\sum_{t=1}^{T} m_t} \in \mathbb{R}^{d_{\text{esm}}}, \tag{2}$$

where $m_t \in \{0, 1\}$ is the attention mask (zero for padding tokens). This pools over real residues only, avoiding contamination from padding artefacts.

A.2 Functional Embedding Head

The functional embedding head $f_\theta : \mathbb{R}^{2560} \to \mathbb{S}^{255}$ (unit 256-sphere) is defined by the composition:

$$\mathbf{a}_1 = \text{Dropout}_{0.1}(\text{GELU}(\text{LN}(\mathbf{W}_1\mathbf{h} + \mathbf{b}_1))), \quad \mathbf{W}_1 \in \mathbb{R}^{1024 \times 2560} \tag{3}$$

$$\mathbf{a}_2 = \text{Dropout}_{0.1}(\text{GELU}(\text{LN}(\mathbf{W}_2\mathbf{a}_1 + \mathbf{b}_2))), \quad \mathbf{W}_2 \in \mathbb{R}^{512 \times 1024} \tag{4}$$

$$\tilde{\mathbf{e}} = \mathbf{W}_3\mathbf{a}_2 + \mathbf{b}_3, \quad \mathbf{W}_3 \in \mathbb{R}^{256 \times 512} \tag{5}$$

$$\mathbf{e} = \frac{\tilde{\mathbf{e}}}{\|\tilde{\mathbf{e}}\|_2} \in \mathbb{S}^{255}, \tag{6}$$

where LN denotes layer normalisation and $\text{GELU}(x) = x \cdot \Phi(x)$ with Φ the standard normal CDF. The L2 normalisation in Equation (6) ensures that cosine similarity in the functional space directly measures functional relatedness.

A.3 Supervised Contrastive Loss

The contrastive projection head $g_\phi : \mathbb{S}^{255} \to \mathbb{S}^{127}$ applies a two layer MLP followed by L2 normalisation to produce $\mathbf{z}_i = g_\phi(\mathbf{e}_i) \in \mathbb{S}^{127}$. Given a batch of N sequences with labels y_i, the supervised NT-Xent loss is:

$$\mathcal{L}_{\text{con}} = -\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\frac{1}{|P(i)|}\sum_{p\in P(i)}\left[\frac{\mathbf{z}_i^\top \mathbf{z}_p}{\tau} - \log\sum_{\substack{k=1\\k\neq i}}^{N}\exp\left(\frac{\mathbf{z}_i^\top \mathbf{z}_k}{\tau}\right)\right], \tag{7}$$

where $P(i) = \{p \neq i : y_p = y_i\}$ is the set of positives for anchor i, $\mathcal{V} = \{i : |P(i)| > 0\}$ is the set of valid anchors (those with at least one positive in the batch), and $\tau > 0$ is the learnable temperature. As implemented, the similarity matrix is $S_{ij} = \mathbf{z}_i^\top \mathbf{z}_j / \tau$; self-similarities are masked by subtracting 10^9 before computing the log-sum-exp denominator.

The temperature τ is trained as $\tau = \exp(\ell_\tau)$, initialised at $\ell_\tau^{(0)} = \log(0.07)$, and projected to $[0.01, 1.0]$ at each step. A low temperature sharpens the similarity distribution, encouraging tight clustering; the learned temperature adapts this trade off automatically.

A.4 PGD Adversarial Training

The adversarial training loss $\mathcal{L}_{\text{adv}}$ simulates an adversary that, given the ESM-2 output $\mathbf{h}$, finds the worst case perturbation within an ℓ_∞ ball of radius $\epsilon = 0.3$ in the ESM-2 representation space. Formally:

$$\mathcal{L}_{\text{adv}} = \mathcal{L}_{\text{CE}}(\phi(f_\theta(\mathbf{h} + \delta^*)), y), \tag{8}$$

where ϕ is the classifier head and δ^* is obtained by $K = 10$ steps of PGD starting from a uniform random initialisation:

$$\delta^{(0)} \sim \mathcal{U}(-\epsilon, \epsilon), \tag{9}$$

$$\delta^{(t+1)} = \Pi_\epsilon\left(\delta^{(t)} + \alpha \cdot \text{sign}\left(\nabla_{\delta^{(t)}}\mathcal{L}_{\text{CE}}\left(\phi\left(f_\theta\left(\mathbf{h} + \delta^{(t)}\right)\right), y\right)\right)\right), \tag{10}$$

where step size $\alpha = 0.05$, $\Pi_\epsilon(\cdot) = \text{clamp}(\cdot, -\epsilon, +\epsilon)$ is the ℓ_∞ projection, and $\text{sign}(\cdot)$ is the elementwise sign. Critically, perturbations are applied in the ESM-2 output space rather than in the discrete token space, simulating what an adversary who can perturb functional embeddings (e.g. via protein engineering) would achieve.

A.5 Full Training Objective

Combining Equations (1), (7), and (8), the complete training objective per batch is:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{CE}}(\phi(\mathbf{e}_i), y_i)}_{\text{classification}} + \underbrace{\mathcal{L}_{\text{con}}\left(\{g_\phi(\mathbf{e}_i), y_i\}_{i=1}^N; \tau\right)}_{\text{contrastive}} + \underbrace{0.5 \cdot \mathcal{L}_{\text{adv}}(\mathbf{h}_i, y_i; \epsilon, \alpha, K)}_{\text{adversarial}}. \tag{11}$$

The three terms are complementary by design: the classification term optimises decision boundaries; the contrastive term organises the embedding geometry; and the adversarial term hardens the classifier against worst-case perturbations.

A.6 Certified Robustness via Randomized Smoothing

Given the functional embedding similarity classifier $f : \mathbb{S}^{255} \to \{0, 1\}$, we construct a smoothed classifier $\tilde{f}$ by averaging predictions under Gaussian noise:

$$\tilde{f}(\mathbf{e}) = \underset{c\in\{0,1\}}{\arg\max}\, \mathbb{P}_{\boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}[f(\mathbf{e} + \boldsymbol{\varepsilon}) = c]. \tag{12}$$

By Theorem 1 of Cohen et al. (Cohen et al., 2019), if $\tilde{f}(\mathbf{e}) = c_A$ (the majority class) with probability lower bounded by $\underline{p}_A > \frac{1}{2}$, then $\tilde{f}(\mathbf{e}') = c_A$ for all $\mathbf{e}'$ with

$$\|\mathbf{e}' - \mathbf{e}\|_2 < R, \quad R = \sigma \cdot \Phi^{-1}\left(\underline{p}_A\right), \tag{13}$$

where Φ^{-1} is the inverse standard normal CDF and $\underline{p}_A$ is estimated via a one-sided binomial confidence interval at significance level $\alpha_{\text{stat}} = 0.001$ over $n_{\text{samples}} = 1000$ Monte Carlo draws. The certified radius R in Equation (13) gives a *provable* guarantee: no adversary can change the screening decision by perturbing the functional embedding by less than R in L2 norm. At $\sigma = 0.1$, every sampled threat achieves $R > 0$, with mean certified radius $\bar{R} = 0.221$.

A.7 3-mer Baseline Formulation

The 3-mer BLAST proxy represents each sequence $\mathbf{s}$ as a frequency vector over the $20^3 = 8{,}000$ possible amino acid trigrams:

$$v_k(\mathbf{s}) = \frac{1}{L-2} \sum_{t=1}^{L-2} \mathbf{1}[\mathbf{s}_{t:t+3} = k], \quad k \in \{A, C, \ldots, Y\}^3, \tag{14}$$

normalised to unit L2 norm. The screening score for a query sequence $\mathbf{q}$ is the maximum cosine similarity to any training set threat:

$$\text{score}_{3\text{mer}}(\mathbf{q}) = \max_{i \in \mathcal{T}} \frac{\mathbf{v}(\mathbf{q})^\top \mathbf{v}(\mathbf{s}_i)}{\|\mathbf{v}(\mathbf{q})\|_2 \, \|\mathbf{v}(\mathbf{s}_i)\|_2}, \tag{15}$$

where $\mathcal{T}$ is the set of training threats. This is the direct analogue of a BLAST hit score: high 3-mer overlap implies high sequence similarity. Its failure mode at low sequence identity is algebraically clear from Equation (15): 80 random substitutions in a 200 residue sequence change roughly 40% of all trigrams, reducing the numerator by approximately 40% and causing the score to fall below typical thresholds.

B Limitations and Dual Use Considerations

B.1 Limitations

Dataset imbalance. The evaluation set contains 2,788 threats and only 125 benign sequences (ratio 22:1). At the 5% false positive rate operating point, this ratio was taken into account; however, in a real deployment with a different ratio the operating threshold would need recalibration and benign rejection rates would require fresh empirical validation.

Enzymatic disruption recall. Enzymatic disruption sequences present a persistent false negative risk with only 55% recall, predominantly confused with neurotoxicity (35%). We hypothesise this confusion arises because both classes contain zinc binding motifs and beta sheet active site scaffolds that are difficult to disambiguate from sequence alone. Further curation of the training data for this mechanism class and integration of structure aware features are needed.

Coverage of novel threats. The training data was sourced entirely from UniProt Swiss-Prot, covering known, reviewed sequences. Novel synthetic proteins with no UniProt ancestor, emergent pathogen sequences, or sequences from poorly characterised organisms may fall outside the learned distribution. The model should be treated as a first pass filter supplemented by expert review for novel high consequence sequences.

Sequence length truncation. ESM-2 is truncated at 1,024 tokens in the current implementation, corresponding to approximately 750 amino acids. Longer sequences lose C terminal functional domains, which may matter for multi-domain virulence factors. Sliding window or hierarchical encoding strategies are natural extensions.

Sequence versus structure. BioScreen operates purely on sequence. Structure informed methods may improve accuracy further, particularly for the enzymatic disruption class where active site geometry is the primary determinant of function.

B.2 Dual Use Risks

The primary dual use risk is that a released model could be queried in a grey box fashion to guide design of sequences that maximise evasion. An adversary with query access could use gradient approximation or model extraction to learn the decision boundary and iteratively refine sequences away from the threat region.

Mitigations:

- Model weights should not be released publicly. Access should be mediated through a monitored API with rate limiting and anomaly detection on query patterns.
- Query logs should be monitored for patterns consistent with adversarial search (systematic exploration of sequence identity vs. score tradeoffs).
- The certified radius R from Equation (13) provides a lower bound on the perturbation budget required to flip a true threat across the decision boundary, enabling operators to quantify residual exposure.
- The mechanism of harm output should be withheld from API responses in production, with access restricted to cleared reviewers, as it could in principle guide optimisation of a specific biological mechanism.

B.3 Responsible Disclosure

No novel pathogen sequences, enhanced potential pandemic pathogen sequences, or regulated agent sequences were synthesised or tested during this project. All data was drawn from publicly available, reviewed UniProt annotations. Any deployment in a production DNA synthesis context should be coordinated with relevant national regulatory bodies and biosecurity organisations such as IBBIS and SecureBio.

B.4 Ethical Considerations

The training data was collected exclusively from public databases with no proprietary or personally identifiable information. The adversarial variant generation procedure was applied only to sequences already present in public databases and does not constitute the design or enhancement of biological agents. All experiments were conducted in silico.

B.5 Future Work

Key extensions include: (1) expansion of mechanism categories to cover chemical weapon relevant proteins and emerging pathogen families; (2) integration with AlphaFold predicted structures for structure aware screening; (3) a closed loop retraining pipeline that ingests newly flagged sequences to maintain coverage of the evolving threat landscape; and (4) rigorous wet lab validation of false negative sequences to quantify true biological risk in the enzymatic disruption class.