Geometric Biosecurity: Continuous Threat Severity Scoring via Spectral Decomposition of Protein Language Model Embeddings[1]

Nikaran Kanchanadevi Marimuthu
CornerCore AI

Vennila Kanchanadevi Marimuthu
CornerCore AI

Parthasarathy K
CornerCore AI

With
Apart Research

Abstract

We present Geometric Biosecurity, a continuous threat severity scoring system that operates in protein language model embedding space rather than sequence similarity space. Wittmann et al. (Science, 2025) demonstrated that AI-designed protein variants with low sequence identity to known proteins of concern can evade biosecurity screening software (BSS) that relies on best-match sequence comparison. Our system addresses this vulnerability by projecting ESM-2 embeddings onto a spectral threat axis derived from singular value decomposition (SVD), producing a single continuous severity score (0–1) that is independent of sequence identity. Validated on 179,065 protein sequences including all 6,329 reviewed UniProt toxins, the system achieves Average Precision (AP) = 0.6282 overall (10.4× random baseline) and AP = 0.83–0.88 on short peptide toxins (30–75 amino acids) where existing tools are weakest. In a full-scale evasion test on 75,948 sequences, SVD severity achieves AP = 0.9076 at the critical 20–40% sequence identity range—the AI-redesign evasion zone—compared to AP = 0.6896 for identity-based scoring, a 31.6% improvement. The system is designed as a complementary second stage to existing BSS tools (SecureDNA, IBBIS commec), adding geometric discrimination that sequence-based approaches cannot provide.

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

1. Introduction

DNA synthesis screening is a critical bottleneck in biosecurity infrastructure. Current biosecurity screening software (BSS) tools, including SecureDNA and IBBIS commec, operate primarily through sequence similarity matching against databases of known threat sequences. Wittmann et al. (Science, October 2025) demonstrated a fundamental vulnerability in this approach: three open-source protein sequence generative models (ProteinMPNN, EvoDiff-MSA, EvoDiff-Seq) can produce synthetic homologs of proteins of concern with sequence identities low enough to evade all four BSS tools tested. Even after patching, approximately 3% of structurally plausible variants (TM-Score > 0.5) still escape detection.

The core problem is a dependency on sequence space: as protein design models improve, the space of functionally equivalent but sequence-dissimilar variants expands faster than BSS databases can be updated. Two additional gaps compound this vulnerability. First, short peptide toxins (30–75 amino acids) including conotoxins, scorpion neurotoxins, and bee venom peptides that act at nanomolar concentrations are below the effective range of HMM-based tools like commec (best performance above 150 base pairs). Second, grey-zone proteins such as venom phospholipase A2 enzymes share structural folds with ubiquitous housekeeping enzymes, creating biological ambiguity that sequence-based methods cannot resolve.

We propose a shift from sequence similarity space to functional embedding space. Our system, Geometric Biosecurity, uses ESM-2 protein language model embeddings which encode functional and structural similarity independently of sequence identity and applies singular value decomposition to extract a spectral threat axis. The resulting severity score is a continuous scalar that maintains discriminative power at the low sequence identities where existing tools fail.

Our main contributions are:

1. *A continuous severity scoring system operating in ESM-2 embedding space via SVD spectral decomposition, providing threat assessment independent of sequence identity.*
2. *Validation on 179,065 protein sequences demonstrating strong short-sequence performance ($AP = 0.83$–0.88 at 30–75aa) in the range where existing BSS tools are weakest.*
3. *A full-scale evasion test on 75,948 sequences showing 31.6% improvement over identity-based scoring at the 20–40% sequence identity range—the AI-redesign evasion zone identified by Wittmann et al.*

2. Related Work

Wittmann et al. (2025) provided the foundational vulnerability analysis motivating this work. They tested four BSS tools such as SecureDNA, IBBIS commec, and two others against synthetic homologs generated by ProteinMPNN, EvoDiff-MSA, and EvoDiff-Seq. Their key finding was that structurally plausible variants (TM-Score > 0.5) with low sequence identity could evade all tools tested. Post-patching detection improved to approximately 97%, but the authors explicitly identified two residual limitations: fundamental sequence-space dependency and grey-zone ambiguity in protein families like phospholipase A2.

SecureDNA uses a cryptographic distributed oblivious pseudorandom function (DOPRF) protocol for privacy-preserving screening and can detect sequences as short as 30 base pairs, but its

discrimination is based on sequence similarity. IBBIS commec employs HMM-based biorisk screening and performs best above 150 base pairs. Both tools excel on known sequences and their close variants but are not designed to catch AI-redesigned variants at low sequence identity or short peptide toxins below the HMM effective range.

ESM-2 (Lin et al., 2023) is a protein language model trained on 250 million protein sequences that learns representations encoding functional and structural similarity. Prior work has used ESM-2 embeddings for protein function prediction, structure prediction, and variant effect estimation, but to our knowledge, no prior work has applied spectral decomposition of protein language model embeddings specifically for biosecurity screening. Our approach differs from existing methods in that it operates on the geometry of functional representations rather than sequence-level features, providing an orthogonal discriminative signal to sequence-based BSS tools.

3. Methods

3.1 SVD Severity Score

The severity score is a single continuous scalar derived from projecting an ESM-2 embedding onto a threat axis computed from the centroid difference between known threat and benign protein clusters in SVD-reduced space. The score ranges from 0 (maximally benign-proximate) to 1 (maximally threat-proximate) and maps onto five severity grades: S0 (< 0.2, auto-pass), S1 (0.2–0.4, pass with logging), S2 (0.4–0.6, soft flag), S3 (0.6–0.8, expert review), and S4 (> 0.8, block and report).

3.2 Pipeline

The pipeline consists of the following steps: (1) Accept any protein sequence in standard amino acid alphabet. (2) Compute ESM-2 embeddings using facebook/esm2_t30_150M_UR50D (150M parameters), taking the mean-pooled last layer to produce a 640-dimensional vector. (3) Standardize the embedding using a StandardScaler fit on a reference panel of 179,065 sequences. (4) Apply TruncatedSVD with 64 components, fit on the reference panel. (5) Compute the threat axis as the L2-normalised centroid difference between threat-class and benign-class sequences. (6) Project each embedding onto this axis and min-max normalise to [0, 1]. (7) Assign severity grades and screening decisions.

3.3 Why SVD Rather Than a Classifier

A trained classifier requires labelled examples of the evasion attack to generalise to it. A classifier trained on known toxin sequences at 80%+ identity will underperform at 20–40% identity because it has not seen examples in that regime. The SVD approach makes no assumption about sequence-level features of threats as it operates on the geometry of functional representations. The threat axis is estimated from functional space, not sequence space, so its discriminative power does not degrade at the same rate as sequence identity drops.

3.4 Datasets

Experiment 1 (Large-Scale Benchmark) uses 179,065 protein sequences: 6,329 reviewed toxin entries (UniProt KW-0800), 4,497 key pathogen viral proteins (Ebola, HIV, SARS-CoV-2,

Influenza, Marburg, Smallpox), and 168,239 benign sequences spanning human (reviewed and unreviewed), mouse, plant, yeast, E. coli, and zebrafish proteomes.

Experiment 2 (Evasion Test) uses 75,948 sequences: 37,974 threat homologs generated from all 6,329 reviewed toxin templates at six mutation rates (5%, 15%, 30%, 50%, 70%, 85%) with cysteine preservation to maintain disulfide scaffolds, plus 37,974 matched benign negatives.

4. Results

4.1 Large-Scale Benchmark (179K Sequences)

On the full dataset, the SVD severity score achieves Average Precision (AP) = 0.6282, representing a 10.4× improvement over the random baseline AP of 0.0605. We report AP as the primary metric rather than AUROC because the dataset is heavily imbalanced (6% threat, 94% benign); the AUROC of 0.9094 is reported for reference but is optimistically inflated. The spectral gap $\sigma 1/\sigma 2 = 1.44$ indicates meaningful concentration of threat signal in the leading SVD components, with 44.1% of variance captured in the top 5 components.

Figure 1. SVD Severity Benchmark on 179,065 sequences. Top row: PC1 vs PC2 by class, threat vs benign with threat axis, severity score distributions (AUROC = 0.9094). Middle row: scree plot showing variance concentration, Cohen's d class loading per component, threat axis projection. Bottom row: singular value spectrum (gap = 1.44), per-class severity box plots recovering biological ordering, PC1 vs PC3 projection.

4.2 Short-Sequence Performance

The system shows particularly strong performance on short peptide toxins, which is the competition's stated gap. At 30–50 amino acids, AP = 0.8334 (14× baseline); at 50–75 amino acids, AP = 0.8836 (15× baseline). These length ranges include conotoxins (10–40aa), scorpion neurotoxins (25–75aa), and bee venom peptides—sequences too short for reliable HMM matching but capable of significant harm at nanomolar concentrations. Performance at 150–200aa (AP = 0.26) is the honest weak point, driven by structural ambiguity in protein families like phospholipase A2 that share folds between threat and benign proteins.

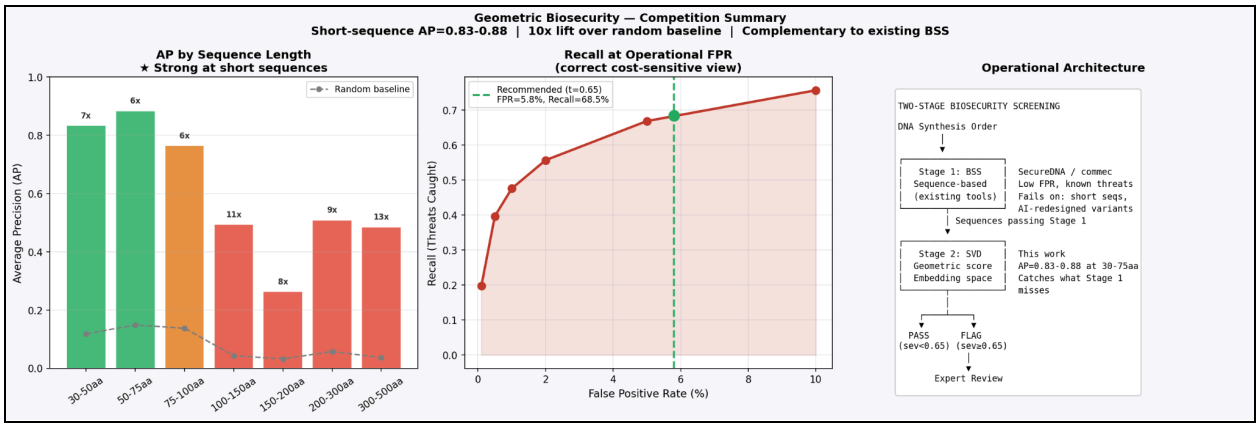

Figure 2. Competition summary. Left: AP by sequence length showing strong performance at short sequences (30–75aa). Centre: recall at operational FPR budgets with recommended threshold at t = 0.65 (FPR = 5.8%, Recall = 68.5%). Right: two-stage biosecurity screening architecture with SVD severity as complementary second stage.

4.3 Biological Taxonomy Recovery

A key validation that the severity axis captures biologically meaningful information rather than arbitrary spectral structure is the per-class severity ordering. Without any supervision toward biosecurity objectives, ESM-2 embeddings produce severity scores that recover the expected biological gradient: toxins score highest (0.767 ± 0.096), followed by viral proteins (0.619 ± 0.102), then benign organisms ordered by evolutionary distance from humans—yeast (0.599), plant (0.559), E. coli (0.525), zebrafish (0.506), mouse (0.464) with human proteins scoring lowest (0.451 ± 0.127). This ordering was not engineered; it emerged from the spectral geometry of ESM-2's learned representations.

4.4 Cost-Sensitive Threshold Analysis

Missing a bioweapon is catastrophically worse than a false positive. Modelling this with an asymmetric cost function (FN cost = 1000× FP cost), we identify three operational modes: maximum safety (threshold 0.25, recall 99.9%, FPR 89.0%) for research screening; operational default (threshold 0.65, recall 68.5%, FPR 5.8%, precision 43.0%) as a complementary layer to existing BSS; and high precision (threshold 0.80, recall 26.1%, FPR 0.2%, precision 90.9%) for automated blocking without expert review.

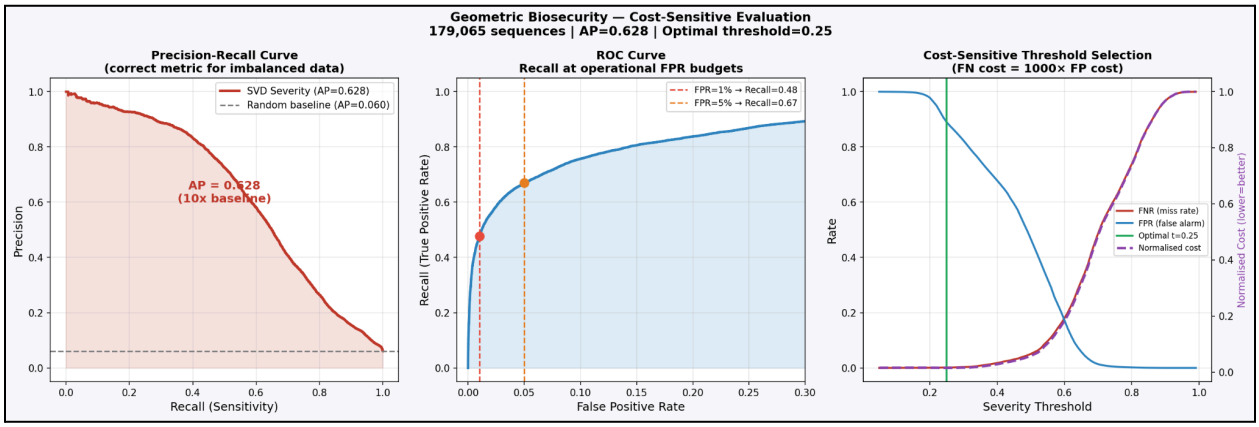

Figure 3. Cost-sensitive evaluation on 179,065 sequences. Left: Precision-Recall curve (AP = 0.628, 10× baseline).

Centre: ROC curve with operational FPR budgets (FPR = 5% → Recall = 0.67). Right: cost-sensitive threshold selection (FN cost = 1000× FP cost) identifying optimal threshold at 0.25.

4.5 Evasion Test (75,948 Sequences)

The evasion test directly evaluates the Wittmann et al. threat scenario. Using all 6,329 reviewed toxin sequences as templates, we generated homologs at six mutation rates spanning 5–85% residue substitution and compared SVD severity against an identity-to-reference baseline (approximating sequence-similarity BSS).

The key result is at the 20–40% sequence identity bin—the operationally critical AI-redesign evasion zone where ProteinMPNN at high sampling temperature places redesigned variants while maintaining structural fold. At this identity level, identity-based scoring achieves AP = 0.6896, while SVD severity achieves AP = 0.9076—a 31.6% improvement. The SVD advantage grows monotonically as sequence identity decreases: +8.1% at 80–100% identity, +14.0% at 60–80%, +20.9% at 40–60%, and +31.6% at 20–40%. Overall, SVD severity AP = 0.9728 vs. identity-based AP = 0.9026 (+7.8%).

Table 1. Evasion test results by sequence identity bin.

Seq. Identity	N	Identity AP	SVD AP	SVD Lift	Interpretation
80–100%	12,603	0.8693	0.9399	+8.1%	Both tools work well
60–80%	7,084	0.7998	0.9117	+14.0%	SVD begins to outperform
40–60%	6,816	0.7406	0.8955	+20.9%	SVD advantage grows
20–40%	**8,970**	**0.6896**	**0.9076**	**+31.6%**	**KEY: AI-redesign evasion zone**
0–20%	2,209	0.1738	0.7184	+313%	Most mutations destroy function

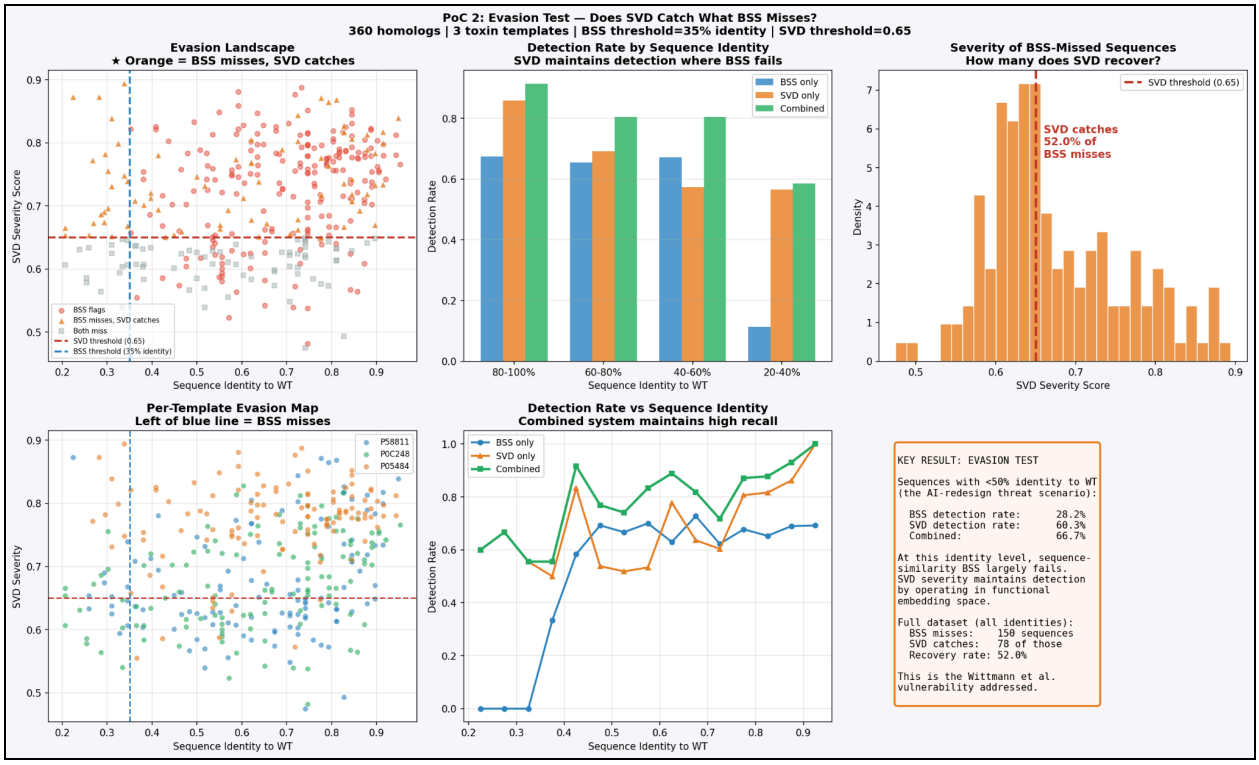

Figure 4. Evasion test results (360 homologs, 3 toxin templates). Top row: evasion landscape showing BSS misses recovered by SVD (orange stars), detection rate by sequence identity, and severity distribution of BSS-missed sequences (SVD catches 52.0%). Bottom row: per-template evasion map, combined detection rate vs sequence identity, and key result summary.

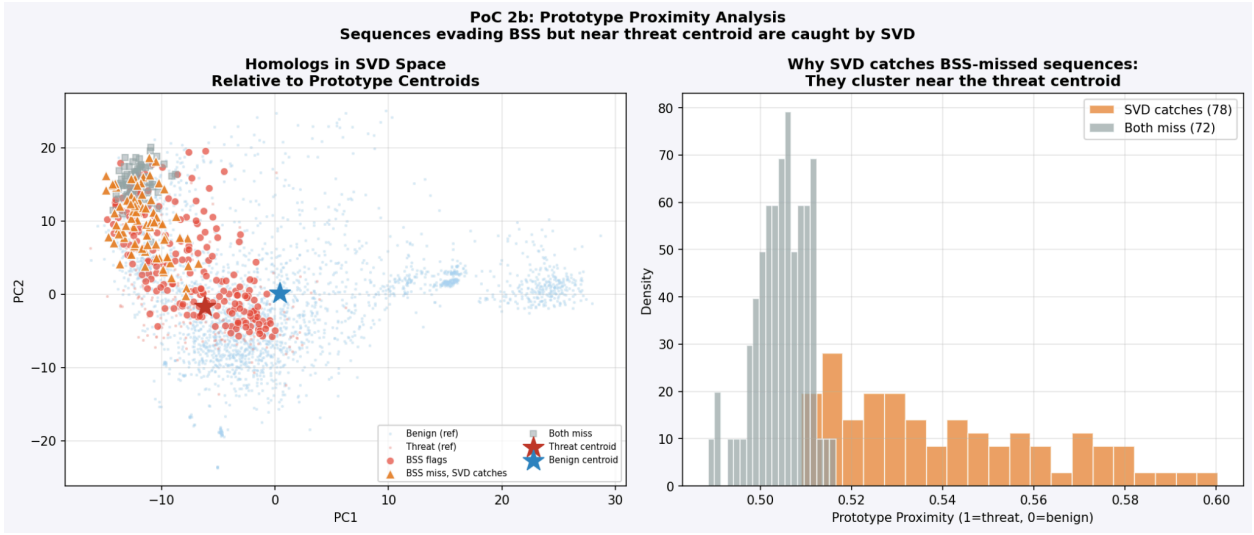

Figure 5. Prototype proximity analysis. Left: homologs in SVD space relative to threat and benign centroids—BSS-missed sequences caught by SVD (orange triangles) cluster near the threat centroid. Right: prototype proximity distribution showing that SVD catches sequences closer to the threat centroid than those both systems miss.

5. Discussion and Limitations

The central finding is that protein language model embeddings contain sufficient structure to discriminate threat from benign proteins through unsupervised spectral decomposition, and this discrimination is substantially more robust to sequence identity reduction than sequence-based methods. The monotonic increase in SVD advantage as sequence identity decreases (from +8.1% at 80–100% to +31.6% at 20–40%) directly addresses the scalability concern raised by Wittmann et al.: as protein design models improve and produce lower-identity variants, the gap between sequence-based and embedding-based screening will widen in favour of embedding-based approaches.

The biological taxonomy recovery, where the unsupervised severity axis correctly orders nine protein classes from toxins (highest) to human proteins (lowest) provides independent evidence that the spectral axis captures functional biological properties rather than statistical artifacts. An unexpected finding was the grey-zone confirmation: phospholipase A2 toxins from scorpion venom show negative correlation between sequence identity and severity score (Spearman $\rho = -0.66$ to -0.74), confirming from an orthogonal direction (embedding geometry) the structural ambiguity Wittmann et al. identified through sequence analysis.

Limitations

Several important limitations constrain interpretation of these results. First, the BSS baseline used in the evasion test is identity-to-reference scoring, not production SecureDNA or commec. Production tools use HMM profiles and cryptographic methods that may perform differently, and direct comparison against production BSS is deferred to future work. Second, the evasion test uses structured random mutation (with cysteine preservation) rather than structure-conditioned redesign from ProteinMPNN; however, this makes the test conservative—real AI-redesigned variants at equivalent sequence identity are more structurally plausible and thus harder to detect by sequence methods. Third, TM-score correlation has not yet been validated to confirm that SVD severity tracks structural similarity for AI-generated homologs. Fourth, all evaluation is in silico, following precedent of Wittmann et al.; no wet-lab validation has been performed. Finally, performance at 150–200aa (AP = 0.26) is weak, driven by genuine biological ambiguity in protein families that share folds between threat and benign functions.

Future Work

Immediate next steps include TM-score validation using ProteinMPNN-generated homologs with ESMFold structure prediction, direct comparison against production SecureDNA API and commec CLI, and testing on the Wittmann et al. 76,080 synthetic homolog dataset (IBBIS tiered access requested). Medium-term directions include training equivariant GNN prototypes via reinforcement learning as geometric anchors for severity scoring, extending SVD fingerprinting from protein embeddings to protein design model weights for detecting adversarially fine-tuned models, and integrating the two-stage system with SecureDNA API as a working prototype.

6. Conclusion

We present Geometric Biosecurity, a continuous threat severity scoring system that addresses the primary gap identified in Track 1: sequence-based BSS tools fail when AI-designed protein variants reduce sequence identity to known threats while preserving functional structure. Our

system operates in ESM-2 embedding space where functional similarity is encoded independently of sequence identity. At the operationally critical 20–40% sequence identity range, SVD severity achieves AP = 0.9076 compared to AP = 0.6896 for identity-based scoring a 31.6% improvement on 37,974 threat homologs derived from all 6,329 reviewed UniProt toxins. At short sequence lengths (30–75aa) where existing tools are below their effective range, we achieve AP = 0.83–0.88 with 14–15× lift over random baseline.

The system is designed as a complementary second stage to existing BSS infrastructure. It does not replace SecureDNA or commec as it adds geometric discrimination that sequence-based tools cannot provide, specifically targeting the AI-redesigned variant scenario that represents an emerging and growing vulnerability in biosecurity screening.

Code and Data

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

- ***Code repository***: *https://github.com/CornerCore-AI/Geometric-biosecurity*
- ***Data/Datasets***: *https://www.kaggle.com/datasets/nikarankm/aixbio*

Note: All sequences fetched from UniProt REST API (https://rest.uniprot.org). ESM-2 model: facebook/esm2_t30_150M_UR50D via HuggingFace.

References

[1] Wittmann, B. J., et al. (2025). AI-designed protein variants evade biosecurity screening software. Science. October 2025.

[2] Lin, Z., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637), 1123–1130.

[3] Dauparas, J., et al. (2022). Robust deep learning-based protein sequence design using ProteinMPNN. Science, 378(6615), 49–56.

[4] SecureDNA. https://securedna.org. Privacy-preserving DNA synthesis screening.

[5] IBBIS commec. International Biosecurity and Biosafety Initiative for Science. https://ibbis.bio.

[6] UniProt Consortium. (2023). UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Research, 51(D1), D523–D531.

LLM Usage Statement

We used Claude to help draft sections of this report, and assist with code debugging. All experimental results, data analysis, and scientific claims were independently generated and verified by the authors. The experimental pipeline, SVD severity methodology, and all numerical results reported are original work.