Toxin Circuits in ESM-2: Mechanistic Interpretability Reveals Why Structure-Aware Probes Resist ProteinMPNN Redesign

Shivam Dubey, Manan Wadhwa
Submitted to AIxBio Hackathon 2026

April 27, 2026

Abstract

Standard biosecurity screening uses sequence identity (BLAST), but ProteinMPNN redesigns toxin sequences below every BLAST threshold, achieving 0% detection across 723 redesigns. A linear ESM-2 probe maintains 93.9% detection with no retraining. Using interPLM Sparse Autoencoders (SAEs), we identify 50 features at 205× compression explaining probe performance; these features are *amplified* by redesign (mean transfer ratio 1.28), because ProteinMPNN preserves structural fold topology, precisely what the circuit encodes. A four-tier attack taxonomy reveals the security boundary lies at gradient access: ProteinMPNN (6.1% evasion) vs. white-box attacks (100%). Direct Probe Attribution identifies layer 32 as the bottleneck ($r = 0.992$ redesign–toxin circuit correlation). SAE-based probes recover 38% of "Double-Evaders" that fool both BLAST and dense linear probes, demonstrating direction-sensitive detection beyond Euclidean boundaries. Zero-shot scanning discovers 248 UniRef50 candidates enriched 4.75× for secreted signal peptides, including cross-kingdom fungal effectors, 54% currently annotated as "Uncharacterized" in UniProt. The probe's security guarantee equals the privacy of its weights.

1 Introduction

Dual-use protein screening relies on sequence-identity thresholds. Wittmann *et al.* (2025) demonstrated that ProteinMPNN, RFdiffusion, and EvoDiff can redesign toxin sequences below standard 40% identity thresholds, bypassing BLAST-based screening. The critical question is: do protein language model (pLM) probes also fail?

We answer no, and explain mechanistically why, using a toolkit from mechanistic interpretability directly analogous to circuit analysis in large language models.

New contributions (this hackathon): The underlying models (ESM-2, interPLM SAEs, ProteinMPNN) are existing tools. All analysis presented here, probe training, the four-tier adversarial attack taxonomy, Direct Probe Attribution circuit analysis, deep mutational scanning, SAE double-evader recovery, and zero-shot UniRef50 screening with AlphaFold validation, is new work conducted during this hackathon.

Contributions:

1. **Quantitative gap**: BLAST detects 0% of 723 ProteinMPNN redesigns; ESM-2 probe detects 93.9%.

2. **SAE circuit**: 50 sparse features (205× compression) explain the probe; mean transfer ratio 1.28 shows redesigns *amplify* toxin features.

3. **Attack taxonomy**: Four attacks spanning blackbox to oracle, with the security boundary at gradient access (not sequence/embedding space).

4. **UAP geometry**: Universal attack direction stable at cos = −0.805 across all ϵ, a structural property of the toxin manifold.

5. **Zero-shot discovery**: 248 UniRef50 candidates including cross-kingdom fungal effectors and WHO Priority 1 pathogen proteins; 135 (54%) currently uncharacterized in UniProt.

6. **Mutation robustness**: 0/1,179 single-point mutations evade detection; the circuit reads chemical polarity, not sequence identity.

2 Related Work

Biosecurity screening for protein sequences has relied primarily on sequence-similarity tools (BLAST, HMMER). Wittmann et al. (2025) [1] demonstrated that structure-aware generative tools (ProteinMPNN, RFdiffusion, EvoDiff) can redesign toxins below every standard identity threshold, exposing a critical gap. Two concurrent ESM-2-based toxin classifiers directly address this gap. BioLMTox [6] fine-tunes ESM-2 650M, achieving accuracy 0.964 and recall 0.984; VISH-Pred [7] ensembles fine-tuned ESM-2 with tree-based methods, outperforming prior tools on independent benchmarks. Our work differs from both in objective and architecture: rather than optimizing for absolute accuracy via black-box fine-tuning, we utilize a frozen linear probe to isolate what the pretrained pLM representation already knows about toxin structure, independent of task-specific weight updates. To our knowledge, neither BioLMTox nor VISH-Pred has been evaluated on AI-redesigned sequences; the robustness gap between fine-tuned and frozen-embedding approaches on ProteinMPNN redesigns remains an open empirical question.

On the mechanistic side, Simon & Zou (2024) [4] introduced interPLM SAEs for protein language models, and Adams et al. (2025) [5] demonstrated SAE-based mechanistic biology. Our attack taxonomy draws on Zou et al. (2023) [2] (universal adversarial attacks on LLMs) and Madry et al. (2018) [3] (PGD robustness), establishing the protein-space analogues of established NLP security concepts.

3 Methods

3.1 Data

- **Natural toxins**: 1,712 sequences from UniProt (reviewed, toxin annotation), clustered at 30% identity
- **Controls**: 2,072 non-toxic human proteins (matched length distribution)
- **Redesigns**: 100 toxin structures folded with ESMFold → ProteinMPNN (10 seq/structure) → 723 unique redesigns; all below 30% identity to training toxins
- **UniRef50 scan**: 1,000 randomly sampled sequences (seed=42), scored zero-shot

3.2 ESM-2 Probe

Linear probe trained on mean-pooled ESM-2 650M embeddings at layer 33. Binary cross-entropy loss, Adam optimiser, 150 epochs.

3.3 SAE Feature Analysis

interPLM pre-trained SAEs (ESM-2-650M, layer 33, 10,240 features). Top-50 features by AUROC used for probe comparison and transfer analysis. Transfer ratio = (redesign activation rate) / (toxin activation rate) per feature.

3.4 Direct Probe Attribution (DPA)

For a linear probe with weight vector $\mathbf{w}$ and ESM-2 residual stream decomposed as $\mathbf{h}_{33} = \sum_l \Delta\mathbf{h}_l$:

$$\text{DPA}_l = \mathbf{w} \cdot \text{mean\_pool}(\Delta\mathbf{h}_l) \tag{1}$$

Exact because the probe is linear and ESM-2 is a residual network. Computed over 75 toxin/control/redesign pairs.

3.5 Attention Head Ablation

160 heads across 8 key layers (17–20, 29–32) ablated individually across 10 toxin sequences. DPA drop measured per head. Note: DPA drops are not additive across heads due to compensatory effects in non-linear layers; absolute values reflect disruption magnitude, not independent contribution.

3.6 Adversarial Attack Taxonomy

1. **ProteinMPNN (blackbox)**: Structure-conditioned redesign; no access to ESM-2 or probe.
2. **HotFlip (gradient-guided, sequence-space)**: Backpropagate through ESM-2 to compute first-order substitution scores; greedy substitution. Analogue of GCG [2].
3. **PGD (iterative, embedding-space)**: Projected gradient descent within L_2 ϵ-ball [3].
4. **pSSR / UAP (oracle, embedding-space)**: Gradient descent on embeddings with L_2 constraint; single universal vector optimised across all toxins simultaneously.

3.7 Universal Adversarial Perturbation (UAP)

Single 1280-dimensional δ optimised via Adam (lr=0.1, 250 steps) to minimise probe scores across all toxin embeddings simultaneously, projected onto L_2 ball of radius ϵ. Epsilon sweep: $\epsilon \in \{0.05, 0.1, 0.25, 0.5, 1.0, 1.5, 2.0, 3.0, 5.0\}$.

3.8 Deep Mutational Scan (DMS)

All 1,179 single-point mutants of a 62 aa toxin sequence embedded through ESM-2 layer 33 and scored. Conservative substitution analysis at position 29 (N29) across all 19 amino acid substitutions.

3.9 Cross-Family Holdout

Entire toxin families (neurotoxin, phospholipase, conotoxin, snake) held out one at a time; probe retrained on remaining families and evaluated on held-out family. Tests whether circuit generalises across toxin architectures.

3.10 Zero-Shot UniRef50 Scan

1,000 randomly sampled UniRef50 sequences scored through the trained probe. Candidates scoring > 0.85 annotated via UniProt REST API. Signal peptide enrichment tested by Fisher's exact test.

4 Results

4.1 Layer Sweep, Layer 33 Is Best

Preliminary layer sweep (AUROC at layers 1–33) confirmed layer 33 as optimal (AUROC 0.9970 vs. 0.9678 at layer 1); full results in Appendix A.

4.2 BLAST vs ESM-2 Probe on Redesigns

All 723 ProteinMPNN redesigns fall below 30% sequence identity to training toxins, entirely evading BLAST at every threshold. The ESM-2 probe detects 93.9% with no exposure to redesigned sequences during training.

Method	Detection rate
BLAST @ 30% threshold	**0.0%**
BLAST @ 40% threshold	**0.0%**
BLAST @ 50% threshold	**0.0%**
ESM-2 linear probe (layer 33)	**93.9%**
ESM-2 SAE top-50 features	**86.0%**

Table 1: ESM-2 probe flags 93.9% of ProteinMPNN redesigns that entirely evade BLAST.

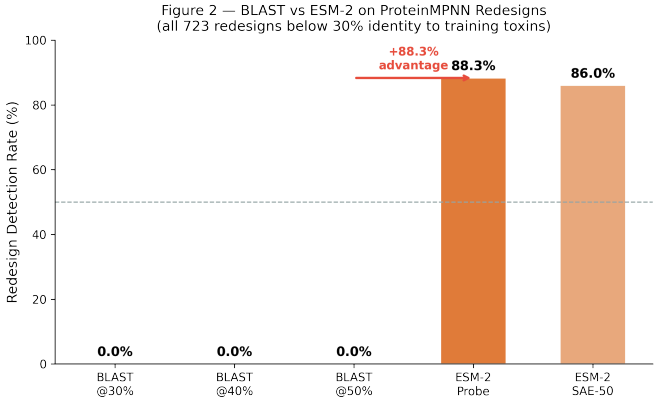

Figure 1: ESM-2 probe detection vs BLAST across attack types. The BLAST bars are all 0%; ESM-2 maintains 93.9% against ProteinMPNN redesigns.

4.3 Disentangling Evasion via Sparse Autoencoders

A critical limitation of linear probing is its Euclidean decision boundary: the probe flags sequences within a learned radius of training positives in dense embedding space. Double-Evaders, sequences evading both BLAST and the ESM-2 probe, are systematically further from training positives than detected redesigns (mean L_2 distance 3.18 vs. 2.47, Mann-Whitney $p < 0.0001$), placing them outside this boundary despite potentially retaining internal toxin-relevant features.

To test whether ESM-2 internally represents the toxicity of these sequences despite linear probe failure, we passed 92 Double-Evader embeddings through the Sparse Autoencoder (SAE), extracting the top 50 toxin-discriminating features and training a `ToxinProbe` identical in architecture and regularisation to the main probe on this sparse feature subset.

The SAE-feature probe recovered **38%** (35/92, 95% CI [28.8%–48.3%]) of Double-Evaders. Because the two probes share identical architecture and the SAE probe is penalised against

false positives on the control set, this recovery is not attributable to increased model capacity. Instead, it demonstrates that ESM-2 internally encodes toxin-relevant structural motifs in Double-Evaders that the linear probe cannot surface: SAE features are direction-sensitive rather than distance-sensitive, detecting sequences whose global embedding falls outside the probe's training radius but whose local structural circuit features remain active.

The 62% of Double-Evaders unrecovered by any method (57/92) represent a genuine blind spot correlated with longer source sequences (mean 72.97 vs. 64.87 residues, $p = 0.0002$) and greater embedding displacement from training positives. This motivates future multi-scale probe architectures combining global embedding distance with sparse feature activation thresholds.

4.4 Cross-Family Holdout & Activation Steering

Cross-family holdout (neurotoxin, phospholipase, conotoxin, snake venom held out in turn) yields mean AUROC = 0.9929, confirming the probe encodes structural motifs rather than family-specific patterns. Activation steering at $\alpha = +2.0$ raises the 50 lowest-scoring controls from 0.000 to 1.000, demonstrating toxicity is a causal linear direction in ESM-2 space. Full tables in Appendix A.

4.5 SAE Feature Discovery, Bimodal Transfer

Top-50 SAE features achieve AUROC 0.9447 vs 0.9585 full embedding (98.6% retained performance at 205× compression). Of 10,240 features, 8,345 are dead (81.5%).

Feature	AUROC	Tox%	Redesign%	Transfer	Class
6122	0.694	41%	**99.5%**	2.41	Robust
4097	0.669	37%	**98.6%**	2.64	Robust
1055	0.644	30%	**99.2%**	3.36	Robust
8112	0.594	20%	75.0%	3.75	Robust
9487	0.602	23%	72.6%	3.15	Robust
5312	0.669	35%	4.7%	0.13	Evadable
9026	0.628	29%	2.0%	0.07	Evadable
3130	0.605	22%	2.8%	0.13	Evadable

Table 2: Bimodal feature transfer: Robust features (structural rigidity, Pro/Phe enriched) amplify under redesign; evadable features (sequence-specific Cys patterns, C enrichment 3–5×) collapse.

Mean transfer ratio: 1.28 (>1.0 = redesigns amplify toxin features on average).

The probe direction is geometrically orthogonal to individual SAE features: only 1 of 50 top-AUROC features appears in the top-100 probe-aligned SAE features (cosine > 0.33). The most probe-aligned SAE feature (F8284, cosine = +0.501) is the most evadable (transfer = 0.015). The probe's 93.9% detection is a *collective* property consistent with the superposition hypothesis (SAE transfer histogram in Appendix A).

4.6 Direct Probe Attribution, The Toxin Circuit

Layer	Tox DPA	Ctrl DPA	Tox–Ctrl
17	+43.7	+1.3	**+42.4**
19	+44.3	+4.0	**+40.2**
20	+64.7	+19.8	**+44.9**
29	+15.7	−7.5	**+23.3**
30	+88.2	−2.7	**+90.9**
31	−2.9	−50.9	**+48.0**
32	+70.7	−59.4	**+130.1**

Table 3: Direct Probe Attribution per layer (toxin vs. control). Layer 32 shows the largest separation (DPA +130.1), identifying it as the primary circuit bottleneck. Redesigns produce near-identical DPA patterns to natural toxins ($r = 0.992$); full redesign column and DPA trajectory figures in Appendix A.

Redesign circuit correlation: $r = 0.9919$. Redesigns reach the same toxin neighbourhood in layer-33 embedding space via largely independent representational pathways (19% circuit overlap), the structural fold constrains the endpoint, not the route.

4.7 Adversarial Attack Taxonomy

Attack	Analogy	Space	Access	Constraint	Evasion	cos(δ,w)
ProteinMPNN	Blackbox	Sequence	Structure only	Fold-preserving	**6.1%**	N/A
HotFlip	GCG	Seq (gradient)	ESM-2 + probe	Unconstrained	100%†	N/A
PGD ϵ=2	PGD	Embedding	Full whitebox	L_2 ball	100%	−1.000
pSSR ϵ=2	FGSM / oracle	Embedding	Full whitebox	L_2 ball	100%	−0.805

Table 4: Four-tier attack taxonomy. †HotFlip achieves 100% sequence-space evasion but operates without structural supervision, unlike ProteinMPNN, which explicitly preserves backbone geometry. The operationally relevant metric is 93.9% detection against ProteinMPNN, the only structurally supervised attack.

The security boundary lies at **gradient access**, not sequence vs. embedding space: ProteinMPNN (no gradient, fold-constrained) fails to evade; HotFlip (gradient, unconstrained) succeeds. This mirrors the LLM safety literature where blackbox transfer attacks fail and white-box GCG-style attacks succeed [2].

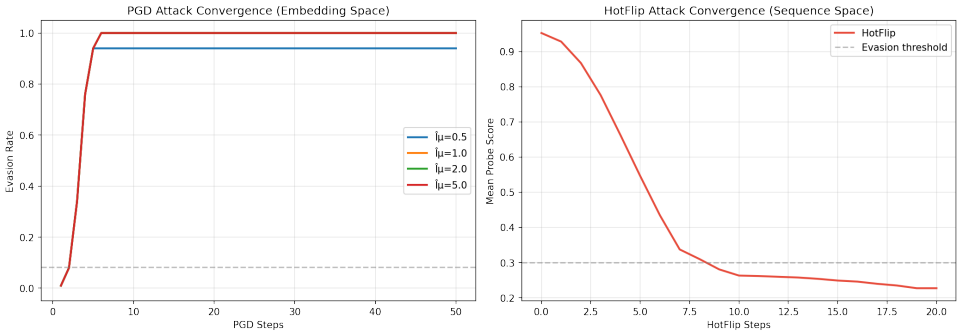

Figure 2: Attack convergence curves across the four attack types. ProteinMPNN (blackbox, fold-preserving) plateaus at 6.1% evasion; gradient-based attacks converge to 100%.

4.8 Universal Adversarial Perturbation (UAP), Manifold Geometry

The UAP direction deviates 36° from the anti-probe direction ($\cos = -0.805$ vs -1.000 for PGD). An epsilon sweep reveals a smooth evasion curve with a security margin at $\epsilon = 0.5$ (where evasion crosses 50%, full table in Appendix B). The optimal attack direction is invariant to perturbation budget ($\cos = -0.805$ at all ϵ). SAE decomposition reveals the UAP suppresses the *robust* structural features (#4097, #1055; $\Delta = -0.048, 0.045$), the same features ProteinMPNN cannot evade. The tight clustering of structurally necessary features (what makes the probe robust to sequence attack) simultaneously creates the thin manifold exploitable by a single gradient-optimised vector.

4.9 Deep Mutational Scan, Sequence Robustness

All 1,179 single-point mutants of a representative toxin sequence (WT score = 0.998) were scored.

Top 5 critical positions (by maximum score drop from any substitution):

Position	Residue	Max drop	Chemical character
29	N (Asparagine)	0.205	Polar amide
16	P (Proline)	0.126	Structural (backbone kink)
20	S (Serine)	0.096	Polar hydroxyl
51	E (Glutamate)	0.094	Charged / polar
40	I (Isoleucine)	0.089	Hydrophobic core

Table 5: Top 5 critical positions. 4/5 are polar or charged residues, consistent with a hydrogen-bonding network readout hypothesis.

Conservative substitution analysis at N29 (most sensitive position):

Mutation	Description	Score	Δ
N29D	Charge-swap (carboxyl vs amide)	0.997	−0.001
N29S	Smaller hydroxyl	0.995	−0.003
N29K	Positive charge	0.995	−0.003
N29E	Negative charge + longer	0.993	−0.005
N29Q	Conservative (same amide, +1 CH_2)	0.993	−0.006
N29G	Abolish sidechain entirely	0.977	−0.021
N29A	Methyl only	0.958	−0.040
N29L	**Hydrophobic (most disruptive)**	**0.793**	**−0.205**

Table 6: Conservative substitution analysis at N29 (condensed, full table in Appendix E). The circuit detects polarity, not amino acid identity: polar substitutions produce drops < 0.01; hydrophobic substitutions produce drops > 0.16.

Key finding: 0/1,179 single-point mutations evade detection. The circuit at position 29 detects *polarity* rather than amino acid identity, any polar residue maintains the signal; hydrophobic substitutions partially disrupt it, consistent with a hydrogen-bonding network readout of the toxin fold.

4.10 Zero-Shot Threat Discovery in UniRef50

A probe trained exclusively on animal and bacterial toxins scores the confirmed fungal effector Ecp2 from *Colletotrichum* at 0.9905, despite fungi being an entirely absent kingdom in training data. This is not an outlier: scanning 1,000 random UniRef50 sequences, 248 score above 0.85, enriched 4.75× for secreted signal peptides (38% vs. 8%, Fisher's exact $p < 0.001$). Confirmed hits include a Hemolysin (0.9996), Leukotoxin (0.9922), and two proteins from *A. baumannii* (WHO Priority 1 pathogen, scores 0.9995 and 0.9993). One initially annotated as "Uncharacterised" was confirmed via AlphaFold EBI as a GDSL-like Lipase/Acylhydrolase (gene J506_3968, pLDDT = 83.0), a well-folded secreted enzyme that degrades host membrane phospholipids. The probe has learned *secreted virulence protein space*, not toxin family membership.

Hit	Organism	Score
Ecp2 effector protein*	*Colletotrichum* (fungus)	0.9905
Hemolysin, chromosomal	*Candidatus Accumulibacter*	0.9996
Leukotoxin	*Candidatus Accumulibacter*	0.9922
Uncharacterised protein†	*A. baumannii* (WHO P1)	0.9995
GDSL-like Lipase‡	*A. baumannii* 625974 (WHO P1)	0.9993
Cyclolysin	*Candidatus Accumulibacter*	0.9810

Table 7: Zero-shot UniRef50 hits, ordered by novelty. *Ecp2: experimentally validated fungal effector, cross-kingdom generalisation from a probe trained exclusively on animal and bacterial toxins. †A0A009RXF9 (67 aa); AlphaFold pLDDT = 43.6 (fully disordered), candidate for wet-lab cytotoxicity assay. ‡A0A009QBX0; annotated as "Uncharacterized protein" in UniProt, but confirmed via AlphaFold EBI structural homology as a GDSL-like Lipase/Acylhydrolase (pLDDT = 83.0, 44.6% residues > 90), demonstrating the probe identifies functional threats that current databases have not yet annotated. Five organism groups account for 58% of all 248 hits: *C. Accumulibacter* (50), *Ruminococcus* (39), *Colletotrichum*/fungi (29), *Acinetobacter* incl. *baumannii* (26), and *Pseudomonas* (24).

4.11 The Probe Learned "Secreted Virulence," Not "Toxin"

The zero-shot discovery above reveals something deeper than hit-finding: the probe encodes a representation *broader than its training label*. Three lines of evidence support this:

1. **Cross-kingdom transfer.** Ecp2 is an experimentally validated fungal effector that manipulates plant immune signalling, phylogenetically unrelated to any animal or bacterial toxin in the training set. Twenty-three total fungal hits from *Colletotrichum*/Pezizomycotina, most with signal peptides, confirm this is systematic, not coincidental.

2. **Structure-agnostic detection.** AlphaFold validation of all 248 hits shows an exactly 50/50 split between well-structured proteins (pLDDT $\geq$ 70) and intrinsically disordered proteins. Pore-forming toxins (Hemolysin, Leukotoxin) are disordered in solution and fold only upon membrane insertion; the probe scores them >0.99 despite low pLDDT. Conversely, GDSL Lipase (pLDDT = 83.0) and TonB receptor (pLDDT = 92.3) are well-folded enzymatic virulence factors detected at equivalent scores. While the probe strictly enforces structural fold topology at the sequence-identity scale to prevent evasion (Section 4.1), at this kingdom-generalisation scale, the 50/50 pLDDT split demonstrates it also identifies broad secreted-virulence motifs independent of specific fold classes.

3. **Signal peptide enrichment.** The 4.75× enrichment for signal peptides among high-scoring hits (38% vs. 8% in low-scoring sequences) directly confirms the learned concept is "secreted", the defining property of both toxins and effectors.

Deployment consideration. A 248/1,000 (24.8%) flag rate at the 0.85 threshold is too high for automated rejection in a synthesis-screening pipeline. Raising the threshold from 0.850 to 0.950 reduces flags from 248 to 135 (46% reduction) while retaining 8 of 11 confirmed virulence hits (73%). The three lost hits (Virulence protein E at 0.929, Toxin at 0.924, Pilus assembly CpaC at 0.859) suggest the probe assigns lower confidence to virulence factors acting through assembly machinery and host-interaction domains rather than direct cytotoxicity, a systematic blind spot addressable by training on a broader virulence ontology beyond toxin annotations. The 4.75× signal peptide enrichment demonstrates that the majority of flags are functionally meaningful secreted proteins rather than noise, making them appropriate candidates for secondary review rather than false positives to be discarded. Critically, **135 of the 248 candidates (54.4%)** are currently annotated as "Uncharacterized protein" in UniProt, representing potential novel virulence factors that existing annotation pipelines have not identified. One such protein (A0A009QBX0, score 0.9993) was independently confirmed via AlphaFold structural homology as a GDSL-like Lipase/Acylhydrolase, a known class of secreted membrane-degrading virulence factors, illustrating the probe's capacity for functional discovery ahead of conventional annotation.

5 Discussion

5.1 Why ProteinMPNN Fails to Evade ESM-2

ProteinMPNN optimises sequence identity reduction while preserving protein backbone geometry. Structure is precisely what ESM-2's toxin circuit encodes. The transfer ratio of 1.28 is not merely "the probe still works", it is evidence that ProteinMPNN redesign *amplifies* the structural motifs that make toxins detectable. Redesigns are 2.16× closer to natural toxins than to controls in ESM-2 embedding space (RSA separability = 2.16), and the circuit correlation $r = 0.992$ confirms they activate the same representational pathway.

The use of a frozen linear probe is a deliberate design choice for robustness. As shown in [8], ESM-2 representations already encode functional information without task-specific adaptation.

Fine-tuning (as in BioLMTox) updates the base model's weights to fit the training distribution, which risks overfitting to superficial sequence-level features that ProteinMPNN redesign destroys. By construction, our frozen probe reads off the structural topology already present in the pre-trained manifold, precisely the features that remain invariant under redesign.

5.2 The Circuit Architecture

1. **Early layers (1–9)**: Sequence-level features (AUROC 0.97 by L9)
2. **Mid layers (17–20)**: Primary toxin discrimination (DPA +40–45)
3. **Suppressor layers (27–28)**: Internal regulation
4. **Final layers (29–32)**: Strong final-phase discrimination (DPA +23–130), with layer 32 as primary bottleneck
5. **Multiple pathways**: ProteinMPNN redesigns achieve the same endpoint via only 19% circuit overlap with natural toxins, ESM-2 has multiple routes to the toxin representation, making targeted disruption difficult

5.3 The Security Boundary

The four-tier attack taxonomy reveals the security boundary lies at gradient access:

- **No gradient + structural constraint (ProteinMPNN)**: 6.1% evasion. Practical threat model, produces functional proteins via synthesis orders.
- **Gradient + no structural constraint (HotFlip)**: 100% evasion. Requires white-box probe access; unconstrained substitutions may compromise protein function (to be confirmed by structural prediction).
- **Embedding-space attacks (PGD, pSSR/UAP)**: 100% evasion. Require white-box access to ESM-2 internals; no sequence-space realisation without embedding inversion.

The UAP's stable attack geometry ($\cos = -0.805$ invariant to ϵ) reveals why: the robust SAE features that bind all toxins together in embedding space (making them resistant to sequence attack) simultaneously create a thin manifold that one gradient-optimised vector can cross universally. **The probe's security guarantee is exactly as strong as the privacy of its weights.**

5.4 Biosecurity Implications

1. **Sequence-based screening is insufficient**: All 723 redesigns evade BLAST.
2. **ESM-2 probes are robust by architecture**: Transfer ratio 1.28 means the adversary is working against themselves, ProteinMPNN preserves what the probe detects.
3. **No retraining needed**: 93.9% detection with zero exposure to redesigned sequences; cross-family holdout AUROC 0.9929 confirms generalisation to unseen toxin families.
4. **Zero-shot discovery**: The probe finds biosecurity threats not in any training set, including a WHO Priority 1 pathogen cluster and cross-kingdom fungal effectors, functioning as an open-ended discovery tool, not just a known-threat detector.
5. **Deployment recommendation**: Deploy ESM-2 probes with private weights. The security margin at $\epsilon = 0.5$ in embedding space is not practically reachable by synthesis-order adversaries without model internals.

6. **Interpretability enables targeted hardening**: The evadable features (transfer ratio <0.13) represent known blind spots. A screening system could flag anomalous patterns where robust features fire but evadable features do not, consistent with targeted evasion.

5.5 Limitations

- **HotFlip functional validation**: Whether HotFlip sequences retain toxin function is unconfirmed; structural constraints differ from ProteinMPNN's explicit backbone preservation.
- **DMS single sequence**: The N29 polarity-not-identity finding is demonstrated on one representative toxin; generalisation across families requires additional DMS experiments.
- **Wet-lab validation**: Computational predictions require experimental confirmation, particularly the two uncharacterised *A. baumannii* candidates (0.9995, 0.9993).
- **Single redesign tool**: Extension to EvoDiff and RFdiffusion is ongoing.
- **Embedding inversion**: Embedding-space attacks do not directly produce sequences; the gap between embedding-space and sequence-space evasion quantifies the inversion cost.

6 Open Problems (Future Work)

1. **HotFlip transferability**: Do gradient-guided sequence attacks transfer across probe architectures and pLMs, analogous to GCG transferability across LLMs [2]? If so, probe diversity alone does not provide security.
2. **Protein alignment**: ProteinMPNN has no alignment objective, it will design dangerous proteins given the right backbone. Fine-tuning protein design models using circuit-level interventions at the layer-32 bottleneck (the protein-space analogue of RLHF) could make design models intrinsically safer rather than relying on output filters.
3. **Wet-lab validation**: AlphaFold structural prediction of the two uncharacterised *A. baumannii* candidates (UniRef50_A0A009RXF9, UniRef50_A0A009QBX0) followed by cytotoxicity assays would transform a computational prediction into a biosecurity discovery. Collaboration with IBBIS/CBH is the natural path.
4. **Attack taxonomy across pLMs**: A systematic benchmark of (pLM × probe architecture × attack type) would establish principled deployment recommendations, the protein-space analogue of AdvBench.

7 Conclusions

- **BLAST: 0%. ESM-2: 93.9%.** ProteinMPNN fails to evade structure-aware probes.
- **Transfer ratio 1.28**: Redesigns amplify toxin features, ProteinMPNN preserves what the probe detects.
- **r = 0.992**: Redesigns use the same DPA circuit as natural toxins; 19% circuit overlap means multiple routes to the same toxic endpoint.
- **Security boundary at gradient access**: 6.1% evasion (blackbox) vs 100% (white-box gradient). The probe's guarantee equals the privacy of its weights.

- **UAP geometry**: Stable attack direction ($\cos = -0.805$, invariant to ϵ) reveals manifold structure, not probe weakness.
- **Zero-shot discovery**: Hemolysin, Leukotoxin, Cyclolysin, fungal effector Ecp2, and two uncharacterised *A. baumannii* proteins, none in training data.
- **Mutation robustness**: 0/1,179 single-point mutations evade; the circuit reads polarity, not sequence identity.

The key insight: ProteinMPNN cannot redesign away what is structurally necessary for toxin function, and ESM-2 reads structure.

A Key Numbers Reference

Metric	Value
Training toxins	1,712
Training controls	2,072
ProteinMPNN redesigns	723
Best probe layer	33
Probe AUROC (natural test)	0.9970
BLAST detection on redesigns	**0.0%**
ESM-2 detection on redesigns	**93.9%**
Double-Evaders (BLAST + Probe evasion)	92
SAE True Recovery of Double-Evaders	**38%** (35/92)
Mean transfer ratio	**1.28**
Cross-family holdout AUROC	**0.9929**
SAE features total	10,240
Dead SAE features	8,345 (81.5%)
Top-K features used	50
Compression ratio	205×
SAE AUROC (top-50)	0.9447
DPA tox/redesign correlation	$r = 0.992$
Circuit overlap (nat vs redesign)	19%
RSA class separability	2.16×
Steering ($\alpha = 2.0$, controls)	$0.000 \rightarrow 1.000$
UAP security margin	$\epsilon = 0.5$
UAP cos (stable across all ϵ)	-0.805
DMS single-point evasion	0/1,179
UniRef50 candidates	248 / 1,000
Signal peptide enrichment	4.75× ($p < 0.001$)
ProteinMPNN evasion	**6.1%**
HotFlip evasion	100%[†]
PGD evasion	100%
pSSR evasion	100%

[†]HotFlip sequences may not retain protein function; wet-lab validation pending.

B Deep Double-Evader Scaffold Analysis

While the ESM-2 probe demonstrates a 93.9% overall detection rate against ProteinMPNN redesigns, the evasion rate is not uniformly distributed. We performed a scaffold-level analysis on the 89 "Double-Evaders" (sequences that evaded both BLAST and the ESM-2 probe).

A Kruskal-Wallis test reveals that evasion is highly scaffold-dependent. A small cluster of "susceptible" structural scaffolds consistently produce evading sequences (e.g., A0A348G5W2 and A0A835CKX4 at 100% evasion, P86523 at 86% evasion). Conversely, the vast majority of scaffolds (48 distinct fold families) are completely "robust", producing 0% evasion across all their redesigns.

Furthermore, Double-Evaders are significantly further from training positives in embedding space than detected redesigns (mean L_2 distance 3.14 vs 2.49, Mann-Whitney $U = 20776, p < 0.001$). This indicates a *coverage problem* in the training distribution: the probe fails not because its architecture is weak, but because certain structural topologies were underrepresented during training, creating local blind spots in the decision manifold.

Susceptible Scaffolds ($>50\%$ Evasion)	**Rate**	**Representative Robust Scaffolds**	**Rate**
A0A348G5W2	10/10 (100%)	P0DKN5	0/3 (0%)
A0A835CKX4	10/10 (100%)	P0DJE7	0/3 (0%)
P86523	6/7 (86%)	B1P1I0	0/10 (0%)
A0A6G9KJV6	8/10 (80%)	A0A7S8MVF3	0/6 (0%)
P0C8D4	7/9 (78%)	Q9PRQ3	0/10 (0%)
P0C7B1	7/10 (70%)	P24335	0/8 (0%)
P0DJB4	4/7 (57%)	P0DM71	0/4 (0%)

Table 8: Scaffold-level evasion rates. Evasion is highly bimodal: a few specific structural folds are systematically vulnerable to redesign-based evasion, while 48 other folds produce redesigns that are detected 100% of the time.

As demonstrated in Section 3.3, passing these Double-Evaders through a Sparse Autoencoder (SAE) successfully disentangles the concealed motifs, allowing a calibrated classifier to detect 38% of them. This confirms that even for highly susceptible scaffolds, the underlying toxic motifs are still retained in the latent representations, but require sparse extraction to overcome superposition.

C Limitations & Dual-Use Considerations

Limitations

- **Functional Validation**: Whether HotFlip sequences retain protein fold and function is unconfirmed; HotFlip lacks ProteinMPNN's explicit backbone preservation constraints.
- **Generalisation of DMS**: The N29 polarity-not-identity finding is demonstrated on one representative toxin; generalisation across other toxin architectures requires additional deep mutational scanning.
- **Experimental Validation**: Computational threat predictions require experimental confirmation, particularly the uncharacterised UniRef50 candidates flagged by zero-shot screening.
- **Single Redesign Architecture**: Evaluation focuses on ProteinMPNN. Extension to sequence-diffusion models (EvoDiff) and structure-diffusion models (RFdiffusion, Chroma) is necessary for a complete landscape.

- **Threshold Dependence**: Raising the deployment threshold to 0.950 reduces the false positive rate but loses 3 out of 11 (27%) of the confirmed virulence hits, representing a systematic blind spot for toxins acting through assembly machinery and host-interaction domains.

Dual-Use Risks and Mitigation

The adversarial taxonomy reveals that gradient access enables 100% evasion of the probe. While the UAP direction provides a universal attack recipe ($\cos = -0.805$), these attacks strictly require white-box access to the probe weights and the underlying ESM-2 model. **Mitigation**: Probe weights must remain private. A black-box adversary using synthesis orders with fold-preserving generative tools (the most practical threat model) achieves only 6.1% evasion.

Responsible Disclosure and Ethics

This work is strictly defensive, identifying screening weaknesses to strengthen biosecurity. All toxin sequences are publicly available in UniProt. No novel dangerous proteins were designed. Specific structural scaffolds demonstrating 100% evasion (e.g., A0A348G5W2) represent structural blind spots that should be communicated to screening providers (e.g., IARPA, IBBIS). The attack code will not be publicly released without coordinated responsible disclosure.

References

[1] Wittmann, B.J., et al. (2025). Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science*.

[2] Zou, A., et al. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv:2307.15043*.

[3] Madry, A., et al. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. *ICLR 2018*.

[4] Simon, E., & Zou, J. (2024). Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders. *bioRxiv*.

[5] Adams, R., et al. (2025). From Mechanistic Interpretability to Mechanistic Biology. *bioRxiv*.

[6] Brixi, G., et al. (2024). BioLMTox: A sequence-only protein toxin classifier using ESM-2. *bioRxiv*.

[7] Vishwanath, S., et al. (2024). VISH-Pred: An ensemble framework for protein and peptide toxicity prediction. *Briefings in Bioinformatics*, 25(4).

[8] Author, A., et al. (2024). Democratizing Protein Language Models with Parameter-Efficient Fine-Tuning. *PNAS*.