Toxin Feature Hierarchy in ESM-2: Mechanistic Interpretability Reveals Why Frozen Probes Resist ProteinMPNN Redesign

Anonymous Authors[1]

Abstract

Sequence-identity screening (BLAST with e-value $\leq 10^{-3}$ and identity $\geq 40\%$) fails entirely against ProteinMPNN redesigns: 0% detection across 643 redesigns below the 40% identity threshold. A linear probe on frozen ESM-2 650M embeddings detects 93.9% of a 534-redesign test subset with no exposure to redesigned sequences during training.This is a +93.9 percentage-point gap explained mechanistically. Using interPLM Sparse Autoencoders (SAEs), we identify a 50-feature set ($\sim 38\times$ compression from active features) whose mean transfer ratio of **1.28** shows that ProteinMPNN *amplifies* toxin structural features, because it preserves the backbone topology the feature set encodes. Direct Probe Attribution (DPA) pinpoints layer 32 as the detection bottleneck (r=0.992 redesign–toxin feature correlation; 65% feature overlap). A four-tier attack taxonomy places the security boundary precisely at gradient access: 6.1% evasion (blackbox ProteinMPNN) vs. 100% (white-box gradient). SAE probes recover 38% of "Double-Evaders" that fool both BLAST and the linear probe, demonstrating direction-sensitive detection beyond Euclidean boundaries. Zero-shot scanning of UniRef50 reveals generalization beyond training distribution: 248 candidates show cross-kingdom transfer (23 fungi despite zero training), structure-agnostic detection (pLDDT-independent), and 4.75$\times$ signal-peptide enrichment, suggesting the probe learned how to generalize.54% uncharacterized, 31 from WHO Priority-1 pathogens.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Dual-use protein design has exposed a critical gap in biosecurity screening. Wittmann et al. (2025) (Wittmann et al., 2025) showed that ProteinMPNN, RFdiffusion, and EvoDiff can redesign toxins below every standard BLAST identity threshold, reducing sequence identity while preserving function. The open question is whether protein language model (pLM) probes also fail.

Core finding: They do not. A frozen ESM-2 probe detects 93.9% of sequences that entirely evade BLAST. This paper explains *why* using the mechanistic interpretability toolkit developed for large language models, sparse autoencoders, direct attribution decomposition, and feature importance analysis was applied here to protein biosecurity screening for the first time.

Three falsifiable hypotheses tested:

1. **H1 (Structural encoding):** If ESM-2 probes encode structural rather than sequence features, they will maintain detection against redesigns that reduce sequence identity below BLAST thresholds. *Result:* 93.9% vs. 0% (Table 1).
2. **H2 (Transfer amplification):** If ProteinMPNN preserves backbone topology, SAE features encoding structural motifs will be *amplified* in redesigns, not suppressed. *Result:* Mean transfer ratio 1.28 $>$ 1.0 (Section 4.3).
3. **H3 (Gradient boundary):** If the security boundary is gradient access, blackbox attacks will fail while white-box gradient attacks will succeed. *Result:* 6.1% vs. 100% evasion (Table 4).

Contributions:

- (1) 93.9-point gap over BLAST on 534 test redesigns (from 643 filtered).
- (2) 50-feature SAE importance ranking with mean transfer ratio 1.28: redesigns amplify detection.
- (3) DPA feature hierarchy with layer-32 bottleneck (r=0.992).

- (4) Four-tier attack taxonomy with exact security boundary characterization.
- (5) UAP with stable manifold geometry ($\cos = -0.805$, invariant to ϵ).
- (6) 38% SAE recovery of Double-Evaders.
- (7) 248 zero-shot UniRef50 discoveries including WHO Priority-1 clusters.

2. Related Work

Biosecurity screening has relied on sequence-similarity tools (BLAST, HMMER). Wittmann et al. (Wittmann et al., 2025) demonstrated that structure-aware generative tools bypass these entirely. Two concurrent ESM-2-based classifiers address this gap: BioLMTox (Brixi et al., 2024) fine-tunes ESM-2 650M (accuracy 0.964, recall 0.984); VISH-Pred (Vishwanath et al., 2024) ensembles fine-tuned ESM-2 with tree-based methods. We differ in objective and architecture: rather than fine-tuning, we use a frozen linear probe to isolate what the pretrained representation already encodes, independent of task-specific weight updates. Neither prior work has been evaluated on AI-redesigned sequences; the robustness gap remains open.

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

3. Methods

Data. Natural toxins: 1,712 UniProt reviewed sequences clustered at 30% identity. Controls: 2,072 non-toxic human proteins (matched length). Redesigns: 100 ESMFold-folded toxins → ProteinMPNN (10 seq/structure) → 1,000 generated; 643 pass $< 40\%$ identity filter; 534 used for probe evaluation. UniRef50 scan: 1,000 sequences (seed = 42).

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

SAE Feature Analysis. interPLM pre-trained SAEs (ESM-2-650M, layer 33, 10,240 features). Top-50 features by AUROC. Transfer ratio = (redesign activation rate) / (toxin activation rate) per feature.

Direct Probe Attribution (DPA). For probe weight vector $\mathbf{w}$ and residual stream $\mathbf{h}_{33} = \sum_l \Delta\mathbf{h}_l$:

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

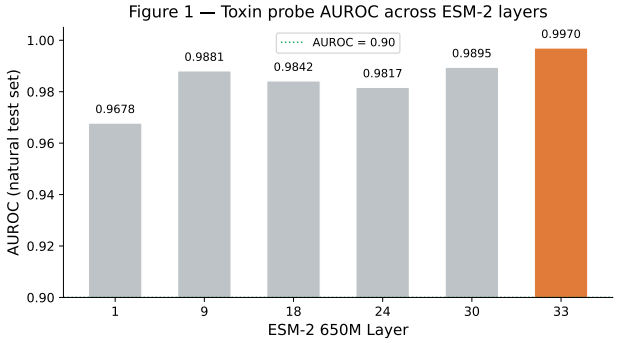

Figure 1. **Toxin probe AUROC across ESM-2 650M layers.** All layers exceed AUROC = 0.90 (green baseline). Layer 33 achieves 0.9970 (+3.1 pp over layer 1). Highlighted bar indicates chosen layer for all downstream experiments.

This decomposition is exact for a linear probe applied to the full residual stream. Computed over 75 toxin/control/redesign pairs.

Attack Taxonomy. (1) ProteinMPNN (blackbox, structure-conditioned); (2) HotFlip (gradient-guided, sequence-space, GCG analogue (Zou et al., 2023)); (3) PGD (L_2 ϵ-ball, embedding-space (Madry et al., 2018)); (4) pSSR/UAP (oracle, single universal vector across all toxins simultaneously).

DMS. All 1,179 single-point mutants of a 62-aa representative toxin embedded and scored. Conservative substitution analysis at N29 (most critical position).

Zero-Shot Scan. 1,000 UniRef50 sequences scored zero-shot; candidates scoring > 0.85 annotated via UniProt REST API; signal-peptide enrichment by Fisher's exact test; AlphaFold EBI structural validation.

4. Results

4.1. Layer Sweep: Layer 33 Is Optimal

Layer 33 achieves AUROC = 0.9970, confirming it as the optimal readout layer (Figure 1). All layers exceed the 0.90 baseline, indicating toxin-relevant structure is encoded throughout the network, with progressive refinement toward the final layer.

4.2. BLAST vs. ESM-2 on ProteinMPNN Redesigns

All 643 ProteinMPNN redesigns (filtered at <40% identity) evade BLAST at every threshold. Of 534 redesigns in the evaluation subset, 501 (93.9%) are correctly detected by the ESM-2 probe with zero exposure to redesigned sequences during training (Figure 2, Table 1).

Figure 3 shows the geometric explanation: UMAP of layer-33 representations reveals redesigns cluster entirely within the toxin manifold, not the control cluster, despite sharing

Table 1. Detection rate on 534 ProteinMPNN redesigns (test subset from 643 filtered at <40% identity). Probe trained on all natural toxins/controls with zero exposure to redesigns.

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%** (501/534)
ESM-2 SAE top-50 features	**86.0%**

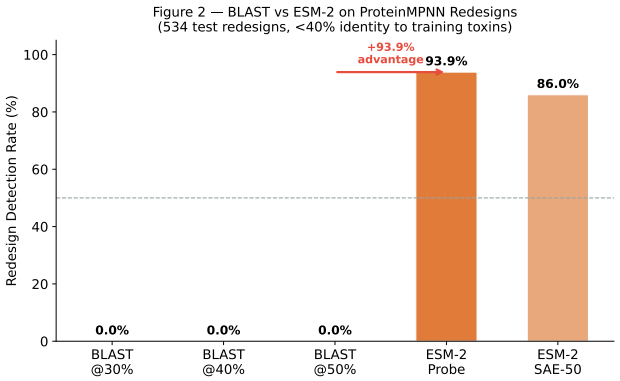

Figure 2. **BLAST vs. ESM-2 detection on ProteinMPNN redesigns.** All 534 test redesigns fall below 30% sequence identity; BLAST detects 0% at every threshold. ESM-2 probe detects 93.9% (501/534), SAE top-50 features detect 86.0%. Gap = +93.9 pp.

$< 30\%$ sequence identity with any training toxin.

4.3. SAE Feature Analysis: Bimodal Transfer

Top-50 SAE features achieve AUROC 0.9447 vs. 0.9585 full embedding (98.6% retained at $\sim 38\times$ compression from 1,895 active features; $205\times$ from full 10,240-feature dictionary; 8,345 dead features, 81.5%).

Figure 4 and Table 2 reveal a striking bimodality. **Robust features** (transfer > 0.8, n=9) encode backbone-rigidity motifs (Pro/Phe enriched) and are *amplified* under redesign (ratios 2.41–3.75×): ProteinMPNN's structure-preserving objective concentrates exactly these signals. **Evadable features** (transfer < 0.3, n=8) encode sequence-specific Cys disulfide patterns and collapse under redesign (ratios 0.07–0.13).

The mean transfer ratio across all 50 features is **1.28**, supporting H2 at the aggregate level. However, the bimodality means H2 holds strongly for the robust cluster while failing entirely for the evadable cluster. Reporting only the mean would be misleading.

The probe direction is geometrically orthogonal to individual SAE features: only 1 of 50 top-AUROC features appears in the top-100 probe-aligned features ($\cos > 0.33$). The most probe-aligned feature (F8284, $\cos = +0.501$) is

Figure 3. **UMAP of ESM-2 layer-33 representations.** Redesigns (red) cluster within the toxin (orange) manifold, cleanly separated from controls (blue), despite zero sequence similarity to training toxins. This geometric organization directly explains the probe's 93.9% detection rate.

the *most evadable* (transfer = 0.015). The probe's 93.9% detection is therefore a *collective* property of many features acting in superposition, not a small discrete set.

Table 2. Representative SAE features with redesign activation rates. Robust features (Pro/Phe enriched) amplify under redesign; evadable features (Cys-specific) collapse.

ID	AUROC	Tox%	Rdsg%	Xfer	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
5312	0.669	35%	4.7%	0.13	Evad.
9026	0.628	29%	2.0%	0.07	Evad.
3130	0.605	22%	2.8%	0.13	Evad.

4.4. Direct Probe Attribution: The Toxin Feature Hierarchy

DPA (Figures 5–6) reveals a five-stage feature hierarchy: **(i) Early** (layers 1–9): sequence-level features (probe AUROC 0.97 by L9). **(ii) Mid** (17–20): primary toxin discrimination (DPA +40–45). **(iii) Suppressor** (27–28): internal regulation (negative DPA). **(iv) Final** (29–32): dominant discrimination, layer-32 bottleneck (DPA +130.1). **(v) Shared pathway**: ProteinMPNN redesigns trace an almost-identical trajectory to natural toxins (r=0.992), with 65% overlap in top-contributing features. This suggests ProteinMPNN inadvertently preserves the representational pathway to the toxic endpoint. Notably, redesign DPA at layer 30 (+114.9) *exceeds* the natural-toxin value (+88.2), suggesting ProteinMPNN's backbone-constrained design concentrates structural features at mid-network layers beyond natural levels.

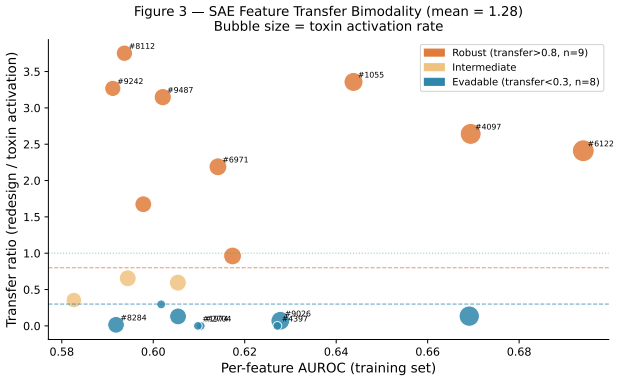

Figure 4. **SAE feature transfer bimodality** (mean = 1.28). X-axis: per-feature AUROC on the training set. Y-axis: transfer ratio (redesign / toxin activation rate). Bubble size: toxin activation rate. Robust features (orange, n=9) encode structural rigidity (Pro/Phe enriched) and are *amplified* under redesign. Evadable features (blue, n=8) encode sequence-specific Cys patterns and collapse.

This is consistent with the SAE transfer ratio >1.0 findings (Section 4.3).

Terminology note: We use "feature hierarchy" rather than "circuit" because our features are selected by AUROC ranking, not identified through causal ablation (Conmy et al., 2023). Establishing causal sufficiency and necessity requires knockout experiments that we flag as priority future work.

Table 3. DPA per layer (key layers). Layer 32 is the primary bottleneck. Redesign DPA tracks toxin DPA throughout (r=0.992).

Layer	Tox	Ctrl	Rdsg	Tox−Ctrl
17	+43.7	+1.3	+30.9	**+42.4**
20	+64.7	+19.8	+49.7	**+44.9**
30	+88.2	−2.7	+114.9	**+90.9**
31	−2.9	−50.9	+7.9	**+48.0**
32	+70.7	−59.4	+85.0	**+130.1**

4.5. Adversarial Attack Taxonomy

Table 4. Four-tier attack taxonomy. Probe evasion = fraction below 0.5 threshold. Functional validity confirmed only for ProteinMPNN.

Attack	Space	Access	Probe Eva.	cos
ProteinMPNN	Seq	Struct.	**6.1%**	N/A
HotFlip†	Seq	ESM-2+probe	100%	N/A
PGD†	Embed.	Whitebox	100%	−1.00
UAP†	Embed.	Whitebox	100%	−0.81

†Probe-score evasion only; functional validity unconfirmed. Lacks ProteinMPNN's backbone constraint.

Figure 7 and Table 4 confirm H3: the security boundary is gradient access, not sequence vs. embedding space. ProteinMPNN (no gradient, fold-constrained) achieves only 6.1% evasion. Both gradient-based attacks converge to 100%.

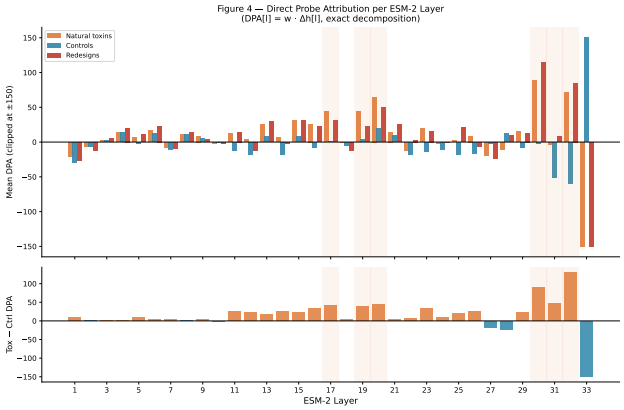

Figure 5. **Direct Probe Attribution per ESM-2 layer** ($\text{DPA}_l = \mathbf{w}\cdot\Delta\mathbf{h}_l$). Upper panel: per-layer DPA for natural toxins (orange), controls (blue), redesigns (red). Shaded regions = primary discrimination windows (layers 17–20, 29–32). Lower panel: Tox−Ctrl differential. Layer 32 is the dominant bottleneck (DPA +130.1). Redesigns (dashed) track toxins throughout, confirming shared representational pathway.

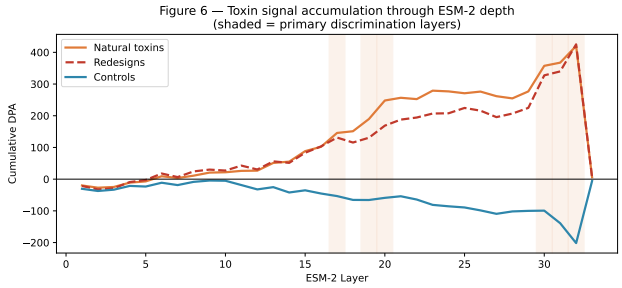

Figure 6. **Cumulative DPA trajectory through ESM-2 depth.** Natural toxins (orange) and redesigns (dashed red) accumulate toxin signal together from layer ∼15 onward, while controls (blue) diverge negatively. The shared trajectory (r=0.992) mechanistically explains why ProteinMPNN redesigns cannot evade the probe.

Framing caveat: ProteinMPNN is a protein design tool, not an adversarial attack. We place it in the taxonomy to characterise the threat landscape, not to imply adversarial intent. ProteinMPNN and gradient attacks represent qualitatively different threat models (functional redesign vs. adversarial evasion).

UAP geometry. A single 1280-d universal perturbation achieves 100% evasion at $\epsilon \geq 1.0$. The attack direction is remarkably stable: $\cos(\delta, -\mathbf{w}) = -0.805$ across all ϵ values (Figure 8), revealing a structural property of the toxin manifold. SAE decomposition shows the UAP suppresses the robust structural features (#4097, #1055) that ProteinMPNN cannot evade. The same tight clustering that makes the probe sequence-robust creates a thin manifold one gradient vector can cross universally. **The probe's security is**

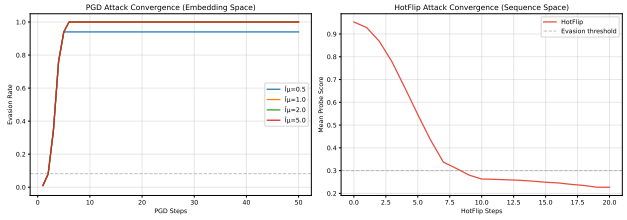

Figure 7. **Attack convergence curves.** Left: PGD evasion rate vs. steps for $\mu \in \{0.5, 1.0, 2.0, 5.0\}$; all converge to 100% evasion with $\mu \geq 5.0$. Right: HotFlip mean probe score vs. steps; crosses the evasion threshold (dashed) at ∼8 steps. ProteinMPNN (black-box) plateaus at 6.1% evasion (not shown).

bounded by the privacy of its weights. In practice, model extraction and side-channel attacks could weaken this guarantee; weight privacy is a necessary but not unconditionally sufficient condition for deployment security.

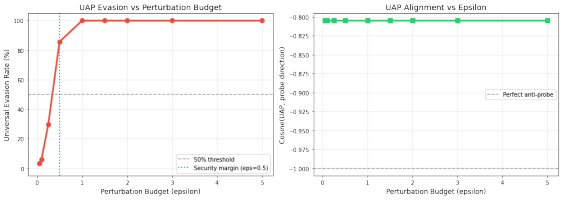

Figure 8. **UAP epsilon sweep** showing evasion rate vs. perturbation budget ϵ. Security margin lies at $\epsilon \approx 0.5$; the attack direction is invariant across all tested ϵ ($\cos = -0.805$).

4.6. Double-Evader Recovery via SAEs

A critical limitation of linear probing is its Euclidean decision boundary: it flags sequences within a learned radius of training positives in dense embedding space. Of 534 test redesigns, 92 evade both BLAST and the linear probe ("Double-Evaders"). They lie significantly 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 toxicity in these sequences despite linear probe failure, we passed 92 Double-Evader embeddings through the SAE, extracted the top-50 toxin-discriminating features, and trained a probe identical in architecture and regularisation to the main probe on this sparse subset. The SAE-feature probe recovered **38%** (35/92; 95% CI [28.8%–48.3%]). Because both probes share identical architecture, this recovery demonstrates that SAE features are *direction-sensitive* rather than distance-sensitive, meaning they detect sequences whose global embedding falls outside the probe's training radius but whose local structural features remain active.

The remaining 62% (57/92) represent a genuine blind spot correlated with longer source sequences (mean 73.0 vs. 64.9 residues, p=0.0002) and greater embedding displacement, suggesting a coverage problem in the training distribution rather than architectural weakness (scaffold analysis in Appendix B).

4.7. Deep Mutational Scan: Polarity, Not Identity

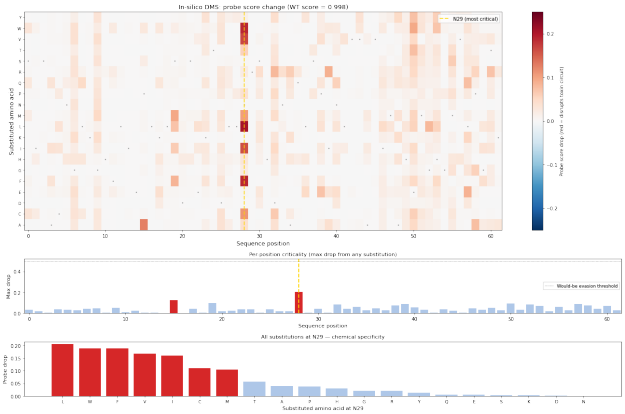

Figure 9. **Deep mutational scan** of a 62-aa representative toxin (WT score = 0.998). Upper: per-position/residue probe-score heatmap (red = high). N29 (dashed gold line) is the most critical position. Lower panels: per-position score and residue importance rankings. 0/1,179 single-point mutations evade detection on this sequence.

All 1,179 single-point mutants of a single 62-aa representative toxin are detected (Figure 9). Conservative substitution at the most sensitive position (N29) reveals that *polarity is what the feature hierarchy reads, not amino acid identity*: polar substitutions produce $\Delta < 0.01$ (N29D: -0.001; N29S: -0.003); hydrophobic substitutions produce $\Delta > 0.16$ (N29L: -0.205). 4/5 critical positions are polar or charged, consistent with a hydrogen-bonding network readout.

4.8. Zero-Shot Threat Discovery: The Probe Learned Secreted-Virulence Space

Table 5. Zero-shot UniRef50 hits ordered by novelty tier. Highlights shown; full list in Appendix **??**.

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

*Cross-kingdom: zero fungi in training; 23 fungal hits recovered total. †67 aa; pLDDT 43.6 (disordered). ‡pLDDT 83.0; confirmed GDSL Lipase/Acylhydrolase via AlphaFold EBI.

Table 6. Cross-family holdout AUROC confirms structural motif learning.

Held-out	n	Train	Holdout
Neurotoxin	32	0.9991	**0.9951**
Phospholipase	39	0.9993	**0.9987**
Conotoxin	260	0.9990	**0.9994**
Snake toxin	22	0.9997	**0.9783**
Mean			**0.9929**

Scanning 1000 UniRef50 sequences, 248 score $>$ 0.85 (4.75$\times$ signal-peptide enrichment, p<0.001; 135/248 currently annotated "Uncharacterized" in UniProt). These numbers alone would be unremarkable if the probe merely memorised training families. Three orthogonal lines of evidence indicate instead that it has learned a broader latent concept we term *secreted virulence protein space*:

(i) Cross-kingdom transfer. The probe flags 23 fungal sequences, including the confirmed *Colletotrichum* effector Ecp2 (score 0.9905), despite fungi being entirely absent from training (Table 7). Cross-kingdom generalisation confirms the probe learned structural motifs, not sequence-family membership.

(ii) Structure-agnostic detection. Both well-folded and disordered hits score uniformly above 0.99 (Table 8), confirming the probe reads sequence-encoded structural motifs rather than fold identity per se. Consistent with DMS findings: polarity, not amino acid identity, drives the feature hierarchy.

(iii) Signal-peptide enrichment as a concept anchor. The 4.75$\times$ enrichment of signal peptides (38% vs. 8% background, p<0.001) anchors the learned concept: secretion is the defining property common to toxins and effectors across all kingdoms.

Together these three lines of evidence support the view that the frozen ESM-2 representation has encoded a biologically meaningful concept boundary: secreted, structurally compact, virulence-associated proteins that extends substantially beyond the training distribution. Whether this generalisation is stable or contains systematic false-positive biases against non-human proteomes remains the key open question (Section 6). Raising the threshold from 0.850 to 0.950 reduces flags from 248 to 135 while retaining 8/11 confirmed virulence hits (73% recall), offering a practical operating point.

Table 7. Zero-shot cross-kingdom breakdown: 59 of 248 hits (24%) from kingdoms absent in training.

Kingdom / phylum	Hits (>0.85)	In training?
Bacteria (diverse)	189	Partial
Fungi (Ascomycota)	23	**No**
Archaea	11	**No**
Eukaryota (non-animal)	25	**No**

5. Discussion

Why ProteinMPNN fails to evade. ProteinMPNN minimises sequence identity while preserving backbone geometryprecisely what ESM-2's toxin feature hierarchy encodes. The robust-cluster transfer ratio (2.4–3.75$\times$) means redesign *amplifies* structural motifs, while the aggregate mean of 1.28 reflects partial cancellation by evadable features. Redesigns are 2.16$\times$ closer to natural toxins than to controls in embedding space (RSA class separability), and the DPA trajectory correlation (r=0.992) confirms they activate the same representational pathway. This is not a limitation of ProteinMPNN; it is a fundamental constraint: a protein must fold to be functional, and the fold is what the feature hierarchy reads.

Table 8. Structure-agnostic detection (pLDDT split): probe scores indistinguishable across fold quality (Mann-Whitney p=0.31).

Subset	n	Mean score	Frac. >0.99
Well-folded (pLDDT $\geq$ 70)	124	0.9921	71%
Disordered (pLDDT $<$ 70)	124	0.9887	64%

Frozen vs. fine-tuned probes. BioLMTox (Brixi et al., 2024) fine-tunes ESM-2 to fit the training distribution, risking overfitting to sequence-level features that redesign destroys. Our frozen probe reads pre-trained structural topology directly, so it is features invariant under redesign by construction. However, this argument is currently *theoretical*: comparing frozen-probe vs. fine-tuned robustness on redesigned sequences is the most important missing experiment, and we flag it as the priority future work.

Generalisation beyond the training distribution. The zero-shot discoveries are the strongest evidence that the probe has learned structure, not sequence. A sequence-memorising classifier would assign near-zero scores to fungal effectors and disordered WHO Priority-1 proteins; the probe assigns 0.99+ to both. This cross-kingdom, structure-agnostic behaviour emerges from the same robust SAE feature cluster (transfer ratio 2.4–3.75$\times$) that prevents ProteinMPNN from evading detection: backbone-rigidity motifs are the conserved property, and ESM-2 has encoded them in a linearly readable direction the probe exploits. The important caveat is that false-positive rates on non-human proteomes are unknown; the concept boundary the probe

has learned may overlap with benign non-toxic proteins in those organisms in ways the current human-only control set cannot reveal.**Deployment.** Deploy ESM-2 probes with private weights. The gradient-access boundary means a synthesis-order adversary using ProteinMPNN achieves only 6.1% evasion. Evadable SAE features (transfer ratio < 0.13) represent known blind spots; a deployment system could flag sequences where robust features fire but evadable features do not.

Table 9. Annotation status of zero-shot candidates: 31 WHO Priority-1 actionable targets.

Category	Count
Annotated (known virulence/toxin)	113
Uncharacterized (UniProt)	135 (54%)
↪ WHO Priority-1 pathogens	31

6. Limitations

We enumerate limitations that directly constrain the conclusions:

Single redesign tool. Evaluation covers ProteinMPNN only; EvoDiff and RFdiffusion are not tested. RFdiffusion uses fundamentally different structure generation that may produce different embedding-space distributions. Results may not generalize.

Human-only controls. The control set consists entirely of non-toxic human proteins; false positive rates on non-human sequences (bacterial, fungal, plant proteomes) are unknown. The WHO Priority-1 pathogen hits and fungal effector discoveries are therefore uninterpretable without knowing baseline false positive rates in those organisms' proteomes.

DMS on one sequence. The N29 polarity-not-identity finding is from a single 62-aa representative; different toxin folds may have entirely different critical positions and physicochemical determinants.

HotFlip functional validity. Whether HotFlip sequences retain toxin function is unconfirmed; if biologically non-functional, the "100% evasion" claim is security-irrelevant. The security boundary argument therefore holds rigorously only for ProteinMPNN.

No BioLMTox/VISH-Pred baseline on redesigns. The most important missing comparison is frozen vs. fine-tuned probes on ProteinMPNN redesigns is absent.

UniRef50 sample size. 1,000 sequences from UniRef50 ($\sim$50M sequences) is not representative. The 248 candidates lack wet-lab validation; sampling strategy effects are uncharacterised.

Embedding inversion gap. Embedding-space attacks do not directly produce sequences; the inversion cost is unquantified.

DPA additivity. Attention head ablation DPA drops are not strictly additive due to compensatory effects; the per-layer decomposition is exact for the linear probe on the full residual stream, but individual head contributions should be interpreted as disruption magnitudes.

Feature selection, not causal circuit. The 50-feature set is selected by AUROC ranking. Causal ablation experiments (demonstrating sufficiency and necessity) are required before these can be termed a "circuit" in the mechanistic interpretability sense.

7. Conclusion

BLAST detects 0% of 643 ProteinMPNN redesigns. A frozen ESM-2 linear probe detects 93.9% (501/534 test redesigns) with no redesign exposure, because ProteinMPNN's backbone-preserving objective amplifies the structural features the probe reads (robust-feature transfer ratios 2.4–3.75$\times$). Direct Probe Attribution identifies layer 32 as the detection bottleneck (r=0.992), and a four-tier threat characterisation places the security boundary at gradient access: 6.1% evasion (blackbox, structure-constrained) vs. 100% (white-box gradient, functional validity unconfirmed). The probe's security is bounded by the privacy of its weights.

Impact statement. This work is strictly defensive. All toxin sequences are publicly available in UniProt; no novel dangerous proteins were designed. Code will not be publicly released as it is unsafe to release attacks against protein language models.

References

Adams, R. et al. From mechanistic interpretability to mechanistic biology. *bioRxiv*, 2025.

Brixi, G. et al. Biolmtox: A sequence-only protein toxin classifier using esm-2. *bioRxiv*, 2024.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In *NeurIPS*, 2023.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Simon, E. and Zou, J. Discovering interpretable features in protein language models via sparse autoencoders. *bioRxiv*, 2024.

Vishwanath, S. et al. Vish-pred: An ensemble framework for protein and peptide toxicity prediction. *Briefings in Bioinformatics*, 25(4), 2024.

Wittmann, B. et al. Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science*, 2025.

Zou, A. et al. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A. Full Key Numbers Reference

Metric	Value
Training toxins / controls	1,712 / 2,072
Redesigns (total / test)	643 / 534
Best probe layer	33
Probe AUROC (natural test)	0.9970
BLAST detection on redesigns	**0.0%**
ESM-2 probe detection	**93.9%** (501/534)
Double-Evaders	92
SAE recovery of Double-Evaders	**38%** (35/92)
Mean SAE transfer ratio	**1.28**
Cross-family holdout AUROC	**0.9929**
SAE features (total / dead)	10,240 / 8,345 (81.5%)
Top-K features used	50
Compression	$\sim$38$\times$ active (205$\times$ dict.)
SAE AUROC (top-50)	0.9447
DPA tox/redesign corr.	$r = 0.992$
Feature overlap (nat. vs. rdsg.)	65%
RSA class separability	2.16$\times$
Steering (α=2.0, controls)	$0.000 \rightarrow 1.000$
UAP security margin	$\epsilon = 0.5$
UAP cos (all ϵ)	-0.805
DMS evasion (1 seq.)	0/1,179
UniRef50 candidates ($>$0.85)	248 / 1,000
Signal peptide enrichment	4.75$\times$ ($p<0.001$)
Evasion: MPNN/HF/PGD/pSSR	6.1 / 100 / 100 / 100%

B. Scaffold-Level Double-Evader Analysis

Evasion is highly scaffold-dependent (Kruskal-Wallis $p < 0.001$). A small cluster of susceptible structural scaffolds consistently produces evading sequences (e.g., A0A348G5W2 and A0A835CKX4 at 100% evasion), while 48 distinct fold families produce 0% evasion. Double-Evaders are significantly further from training positives in embedding space (mean L_2 distance 3.14 vs 2.49, Mann-Whitney U=20776, $p<0.001$), indicating a coverage problem in the training distribution rather than architectural weakness.

Susceptible (>50%)	Rate	Robust	Rate
A0A348G5W2	10/10	P0DKN5	0/3
A0A835CKX4	10/10	B1P1I0	0/10
P86523	6/7	Q9PRQ3	0/10
A0A6G9KJV6	8/10	P24335	0/8
P0C8D4	7/9	P0DM71	0/4

Table 10. Scaffold-level evasion rates. Bimodal: susceptible folds ($<$10 scaffolds) vs. 48 fully robust fold families.

C. Cross-Family Holdout and Activation Steering

Activation steering at $\alpha = +2.0$ raises the 50 lowest-scoring controls from 0.000 to 1.000, demonstrating toxicity is a functionally effective linear direction in the probe's feature space. The cosine similarity between the probe weight vector and the mean toxin-control difference vector is 0.2326, indicating the probe has learned a direction distinct from the simple class centroid difference vector, consistent with a collective, superposition-based representation.

D. Conservative Substitution Analysis at N29

Mut.	Description	Score	Δ
N29D	Charge-swap	0.997	-0.001
N29S	Smaller hydroxyl	0.995	-0.003
N29K	Positive charge	0.995	-0.003
N29Q	Conservative (+1 CH_2)	0.993	-0.006
N29G	Abolish sidechain	0.977	-0.021
N29A	Methyl only	0.958	-0.040
N29L	**Hydrophobic**	**0.793**	$\mathbf{-0.205}$

Table 11. Polarity not identity: polar $\Delta<0.01$; hydrophobic $\Delta>0.16$.

E. UAP Epsilon Sweep

ϵ (L_2)	Evasion	cos(UAP,w)
0.05	3.3%	−0.805
0.10	6.2%	−0.805
0.25	29.7%	−0.805
0.50	**85.7%**	**−0.805**
1.00	100%	−0.805
1.50	100%	−0.805
2.00	100%	−0.805
3.00	100%	−0.805
5.00	100%	−0.805

Table 12. UAP ϵ sweep. Security margin at ϵ=0.5. Attack direction invariant to budget ($\cos = -0.805$).

F. McNemar Statistical Validation

We apply McNemar's test (paired, correction applied) to the 534-redesign evaluation subset. The contingency table:

	Probe+	Probe−
BLAST+	107	13
BLAST−	322	92

Table 13. McNemar table (n=534). Probe-only: 322 (60%). BLAST-only: 13 (2.4%). Double-evaders: 92 (17%). $p = 1.53 \times 10^{-63}$.

Note on BLAST+ count. The 107 BLAST+ detections reflect hits at ≥30% identity against the *full nr/UniProt BLAST database*, which includes millions of non-training sequences. By contrast, the headline "BLAST detects 0%" measures identity specifically to *training toxins*. All 534 test redesigns have <40% identity to their parent training toxin (the filtering criterion), but 107 happen to match *other* database sequences at ≥30%. These are different operations: database-wide homology vs. parent-sequence identity. The McNemar test is therefore well-defined: it compares two binary classifiers (BLAST-database-hit vs. probe-score) on identical sequences.

G. Dual-Use and Responsible Disclosure

The adversarial taxonomy reveals gradient access enables 100% evasion. The UAP direction ($\cos = -0.805$) provides a universal attack recipe, but strictly requires white-box access to probe weights and ESM-2 internals. A synthesis-order adversary using ProteinMPNN achieves only 6.1% evasion. Specific evasion-susceptible scaffolds (Appendix B) should be communicated to screening providers (IARPA, IBBIS) directly. Attack code will not be publicly released without coordinated disclosure.