Probing Harm-Related Signals in Pretrained Protein Language Models

Edric Castel C. Hao
Analog Devices, Inc.

Katrina Compendio
Independent Researcher

Rowell Herrera
De La Salle University

Ciaran O'Connell
Independent Researcher

Alessandro Lucatelli
ETH Zurich

With
Apart Research

Abstract

Protein safety classifiers, used to flag toxic, virulent, or otherwise hazardous sequences, are increasingly built on top of pretrained protein language models (PLMs), yet little is known about how these models represent harm-related properties or what this implies for their reliability. We probe ESM-2 (esm2_t6_8M_UR50D) on three binary classification tasks of biosecurity relevance: peptide toxicity, pore-forming toxin (PFT) identity, and virulence. Using mass-mean and logistic-regression linear probes applied to CLS-token activations at every transformer layer, we report three findings. First, the pretrained backbone, which has never seen harm-related labels, already encodes these tasks in a form that is largely linearly recoverable, with probe accuracies typically reaching ~75–90% in intermediate and deeper layers. Second, fine-tuning yields its largest gains on the mass-mean probe rather than the logistic-regression probe, suggesting that it primarily improves alignment of existing task-relevant structure with class-mean directions, rather than substantially increasing linear separability. Third, zero-shot cross-task evaluation reveals partial but non-trivial transfer among the three tasks, consistent with shared underlying structure, with virulence-trained models generalizing most broadly and PFT-trained models producing an inversely correlated signal on general toxicity. These results suggest that current PLM-based safety classifiers may rely heavily on

pre-existing, linearly accessible representations, potentially limiting robustness to distribution shift or adversarially constructed sequences. While linear probes demonstrate that harm-related information is present in model representations, they do not establish that deployed classifiers causally depend on the same features. Taken together, our findings highlight both the promise and limitations of PLM-based safety screening and motivate further work on robustness and failure modes.

1. Introduction

AI-assisted protein design has crossed a practical threshold. Tools like RFdiffusion [1] and ProteinMPNN [2] can generate functional proteins de novo in days, capabilities once confined to specialized structural biology labs and now accessible to anyone with a GPU. While this democratization accelerates drug discovery and enzyme engineering, it also lowers the barrier for misuse. The same pipeline used to develop therapeutics can, in principle, be redirected toward harm.

The biosecurity concern is concrete. Current toxicity screening relies largely on sequence homology to known toxins [3], a defense designed in the pre-GenAI landscape that is increasingly brittle with generative models. Fabio Urbina et al. [4] showed that repurposed generative models can rapidly produce harmful small-molecule candidates, and the same dual-use pattern plausibly extends to proteins. Here, the mitigation challenge is harder. Protein toxins are encoded in DNA and can be synthesized through commercial services. The limitation is structural rather than incidental. Classifiers such as ToxinPred2 [5] assume that harmful protein sequences occupy a learnable region of sequence space. This assumption holds for natural toxins but may fail once sequences are optimized to preserve function while diverging from known examples. What is missing is a detection approach grounded in protein activity rather than superficial similarity.

Protein language models provide a plausible substrate for this. Transformers trained on sequence alone recover contacts, secondary structure, and functional annotations from statistical patterns in sequence data [6]. If toxicity, constrained by underlying biophysics, is geometrically encoded in these representations, it could support a screening layer that is not easily circumvented by sequence novelty. To test this hypothesis, we probe frozen ESM-2 representations with linear classifiers across three biosecurity-relevant tasks: pore-forming toxin identification, virulence prediction, and peptide toxicity evaluated on the CAPTP adversarial dataset [7]. CAPTP directly instantiates the adversarial threat model motivating our work, providing sequences engineered to retain predicted toxicity while evading homology-based detection. Performance on this dataset is therefore the most practically relevant measure of whether PLM representations encode a detection signal robust to sequence novelty.

2. Related Work

Protein safety classifiers and screening.

Sequence-based toxicity prediction has been studied using both classical machine learning and deep learning models. CAPTP [7] is a recent benchmark for evaluating peptide toxicity under adversarial conditions. Among all the models tested, the best-performing model achieved 92.01% accuracy on the first test set and 90.55% accuracy on the second imbalanced test set. In parallel, gene synthesis providers deploy biosecurity screening systems designed to flag sequences associated with regulated pathogens or toxins. Recent evaluations show that these systems can be sensitive to distribution shifts introduced by AI-designed variants that retain function while evading detection. Subsequent updates have improved robustness to close synthetic homologs, but the broader question of how these models represent harmfulness rather than similarity remains open.

Embedding-based classifiers.
Beyond homology, several recent approaches leverage protein embeddings from pretrained language models to classify toxicity or pathogenicity. These embedding-based classifiers demonstrate that PLM representations contain safety-relevant information, but they typically train end-to-end predictors and report task-specific accuracy [8]. What remains underexplored is how harm-related properties are geometrically represented within PLMs, and whether linear probes can recover these signals without fine-tuning.

Protein language models.

ESM-2 [7] is a family of transformer encoders trained on UniRef sequences using masked amino acid prediction. Larger ESM-2 variants achieve state-of-the-art performance on structure-related downstream tasks and are widely used as feature extractors for sequence classification. We use the smallest publicly released variant (esm2_t6_8M_UR50D, 8M parameters, 6 transformer layers) for computational cost reasons. This design choice is a deliberate scope limitation discussed in §6.

Linear probing and representation analysis.

Linear probes are simple classifiers trained on frozen model activations. Their accuracy reflects the extent to which a target concept is linearly decodable from those representations [9]. In language model interpretability, probes have been used to study properties such as truthfulness, sentiment, refusal behavior, and entity attributes. Samuel Marks and Max Tegmark [10] formalize two probe variants, mass-mean and logistic regression, which we adopt here. The mass-mean probe is less expressive than logistic regression because it succeeds only when the relevant direction aligns with the difference in class means. The relative performance of the two probes therefore provides insight into whether a concept is encoded along interpretable axes in representation space, or requires a more complex linear boundary to be recovered.

Interpretability of biological models.

Prior interpretability work on protein language models has focused primarily on structural correlates of attention and residue-level features relevant to contact prediction. The application of probing methods to safety-relevant properties such as toxicity, virulence, and pore-forming toxin identity remains limited. This gap motivates the present study.

3. Methods

We use three datasets that together cover complementary failure modes of sequence-based screening, progressing from a controlled structural benchmark to adversarially designed evasion sequences.

3.1 TCDB Pore-Forming Toxins

Positive examples were drawn from the Transporter Classification Database (TCDB) [11], retrieving all entries under the 1.C superclass (pore-forming toxins). Two subfamilies were excluded after manual inspection: 1.C.39 (apextrin, poorly characterised experimental support) and 1.C.137 (hypothetical soil bacteria proteins, computationally predicted only). The final positive set comprises 728 sequences across 100 subfamilies, spanning bacteria, fungi, cnidarians, and other organisms, with lengths ranging from 100 to 5,206 amino acids (median 343 aa). Negative examples were drawn from Swiss-Prot reviewed entries lacking toxin (KW-0800) or transporter (KW-0813) annotations, length-matched and subsampled to 728. Final dataset: 1,456 sequences, balanced.

Pore‑forming toxins constitute a principled benchmark for this investigation. Despite sharing biophysical motifs such as hydrophobic segments and transmembrane propensity, alpha‑PFTs and beta‑PFTs achieve identical membrane‑disrupting function via completely different secondary structures, with no sequence similarity between families [12]. A probe that generalizes across this structural diversity cannot rely solely on motif detection, but must capture functional organization in representation space.

3.2 UniProt Virulence Dataset

Virulence-annotated sequences were retrieved from UniProtKB/Swiss-Prot [13] using keyword KW-0843. The initial collection comprised 4,859 virulent sequences across 826 taxonomic groups. Nonvirulent sequences were drawn from the same taxa to control for organism-specific biases, yielding 136,352 candidate nonvirulent sequences. After sequence-level deduplication and removal of conflicting labels, where identical sequences were annotated as both virulent and nonvirulent likely due to context-dependent function, the dataset was reduced to 3,877 virulent and 97,974 nonvirulent sequences.

The resulting dataset is highly imbalanced, with nonvirulent sequences greatly outnumbering virulent ones. To control for this imbalance in the current study, we construct a balanced subset by randomly sampling nonvirulent sequences to match the number of virulent examples. This subsampling is intended to isolate representation quality from class imbalance effects and will be relaxed in future experiments that explicitly study performance under skewed distributions.

This dataset is designed to test whether toxicity-related signals in protein language model representations extend beyond a single mechanistic class to pathogenicity more broadly, spanning adhesins, secreted effectors, enzymatic toxins, and other virulence mechanisms.

3.3 CAPTP Adversarial Dataset

The CAPTP dataset [7] provides sequences engineered to retain predicted toxicity while evading BLAST-based homology detection. This directly instantiates the threat model: an adversary producing

functional toxic proteins that current screening pipelines cannot flag. Performance on this dataset is the most practically relevant measure of whether PLM representations encode a detection signal robust to sequence novelty and serves as a proxy for the AI-designed variant evasion scenario motivating this work.

---

4. Results

4.1 Model

We use the ESM-2 model esm2_t6_8M_UR50D: 8M parameters, 6 transformer layers, embedding dimension 320. All experiments use this same backbone.

For fine-tuning, we wrap the backbone in a feedforward classification head:

```
ESM-2 backbone
  └─ CLS token (position 0) at last transformer layer (layer 6)
       └─ Linear(320 → 128) → ReLU → Linear(128 → 1)
```

The scalar output is trained with `BCEWithLogitsLoss` and the CLS token serves as the sequence-level representation throughout the paper.

We consider three binary classification tasks:

Task	Source	Train	Val	Test
Peptide toxicity	Curated FASTA, labels in headers	6,387 (90% of train_data.txt)	10% split	`test1.txt` (1,126), `test2.txt` (582)
Pore-forming toxin (PFT)	TCDB family 1.C positives + non-PFT negatives	70% of 1,456	15%	15%
Virulence	UniProtKB keyword KW-0843	70%, downsampled 50/50	15%	15%

For toxicity, two test sets are provided. The file *test1.txt* has a class balance of approximately 72/28, similar to the training distribution. The file *test2.txt* has a class balance of about 92/8. We report accuracy on both, but treat *test1.txt* as the primary evaluation. For virulence, the natural class distribution is highly imbalanced. To address this, we downsample the majority class to obtain a balanced training set, consistent with the dominant convention for this benchmark.

Sequences are truncated to ESM-2's maximum usable length of 1,022 amino acids. Non-standard IUPAC characters absent from the ESM-2 alphabet (e.g., `J`) are remapped to canonical residues prior to tokenization.

4.3 Fine-tuning

All three tasks share the same hyperparameters. We use the Adam optimizer with a learning rate of 1e-5, train for ten epochs, and set the batch size to 4. The default split is 70/15/15 for train, validation, and test, with random_state=42. For toxicity, the benchmark specifies a held-out test set but not a validation set, so we create a 90/10 train/validation split to enable model selection.

Validation accuracy is computed after each epoch, and the best checkpoint is retained, so the saved model corresponds to the best-generalizing configuration rather than the final-epoch one. The full backbone is unfrozen during fine-tuning, and we do not use frozen-feature classifiers. Validation accuracy is computed after each epoch, and the best checkpoint is retained. Thus, the saved model corresponds to the configuration that generalizes best, rather than the final epoch.

4.4 Linear Probes

We follow Marks and Tegmark [10] in using two probe types:

- **MMProbe (mass-mean):** the classification direction is the difference of per-class mean activations. Optionally scaled by Mahalanobis distance using the pooled within-class covariance matrix. The probe has no learned parameters beyond the threshold.
- **LRProbe (logistic regression):** sklearn.LogisticRegression fit to StandardScaler-normalized activations.

For each task, we extract CLS-token activations from every transformer layer (0–6, where Layer 0 is the input embedding before any transformer block) using both the pretrained backbone and its fine-tuned counterpart, evaluated on the same train/test splits. Probes are trained on training-set activations and evaluated on test-set activations. We report test accuracy as a function of layer depth for both backbones and both probe types.

The interpretive logic: the LRProbe upper-bounds the linearly decodable information about the task in a layer's representation. The MMProbe accuracy specifically measures whether that information aligns with the class-mean direction. A large MMProbe gain after fine-tuning, combined with a small LRProbe gain, suggests that fine-tuning primarily reorients existing representations rather than substantially increasing the amount of linearly decodable task-relevant information.

4.5 Cross-Task Evaluation

We additionally evaluate all three fine-tuned checkpoints on all three test sets in a zero-shot transfer setting. Each model's trained classifier head is applied directly to the test sets of the other tasks without further training or adaptation. This procedure yields a 3×3 accuracy matrix.

4.6 Reproducibility

All splits use random_state=42, and probe fitting is deterministic. The code, together with the three fine-tuned checkpoints, is provided alongside this submission. The pretrained backbone is the publicly released ESM-2 model esm2_t6_8M_UR50D.

5. Results

5.1 Fine-tuning achieves expected task performance

Task	Best validation accuracy	Test accuracy
Toxicity	0.9437	0.9352 (test1), 0.9399 (test2)
Pore-forming toxin	0.9220	0.8807
Virulence	0.8660	0.8514

Toxicity test2 accuracy should be interpreted relative to its 92% majority-class baseline. Test1, with a 72/28 balance, is the primary toxicity result. Virulence and PFT test sets are balanced 50/50, so accuracies there are directly interpretable.

5.2 Pretrained ESM-2 already encodes harm-related features linearly

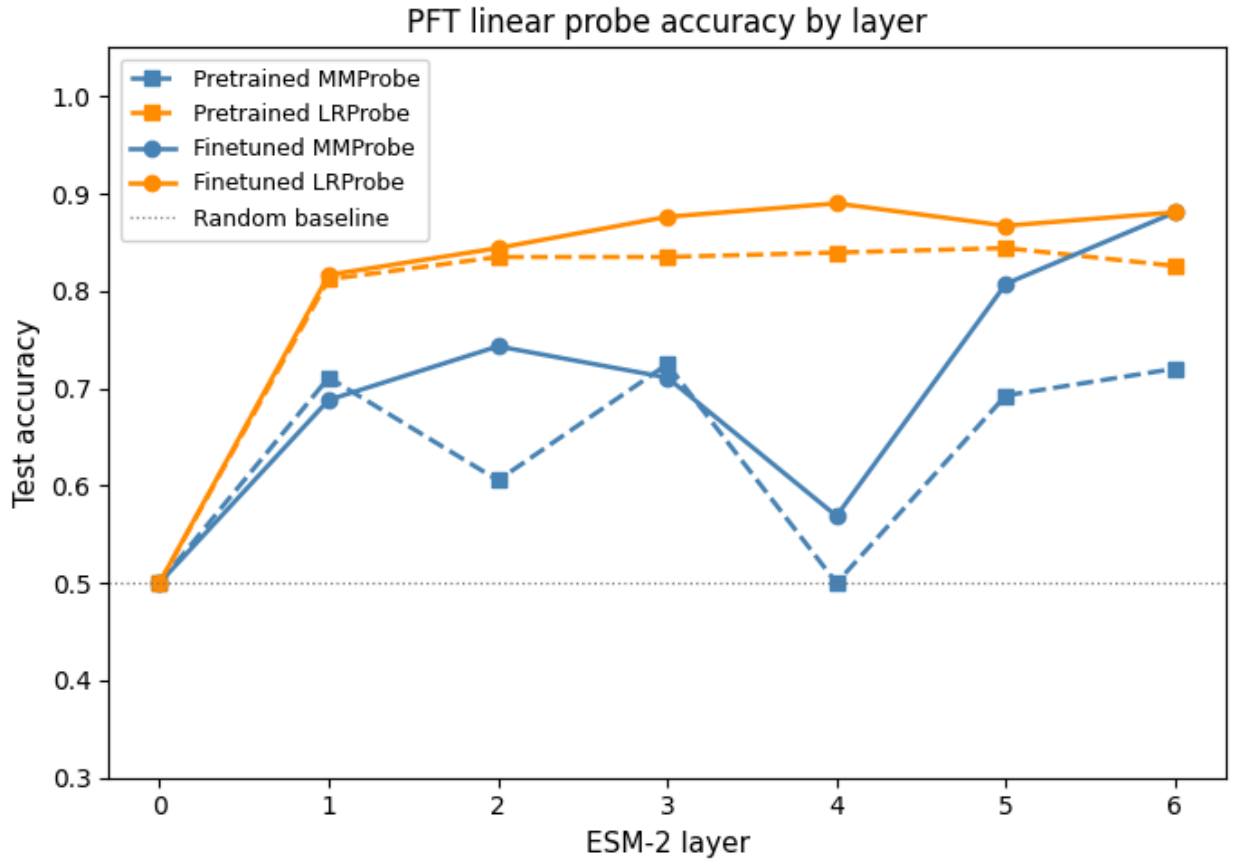

Figure 1. Performance of two linear probe types under pretrained and fine-tuned settings

Logistic regression probes on the **pretrained** (non-fine-tuned) backbone reach the following test accuracies at the most informative layer:

Task	Best pretrained LRProbe (test)	Layer	Accuracy at Layer 0	Δ(Best-Layer 0)
Toxicity (test1)	0.90	1	0.72	0.18
Toxicity (test2)	0.92	0	0.92	N/A
Pore-forming toxin	0.84	5	0.50	0.34
Virulence	0.80	5	0.50	0.30

For toxicity (test2), probe accuracy is already maximal at Layer 0, indicating that amino acid composition alone provides a strong signal. In contrast, pore-forming toxin and virulence tasks show substantial gains from Layer 0 to deeper layers, with improvements of 0.30–0.34. Figure 1 illustrates this pattern for pore-forming toxins, where accuracy rises sharply from Layer 0 to Layer 1, with slight fluctuations across subsequent layers, and the highest performance is observed at Layer 5. This trajectory highlights the contribution of contextualization introduced by the transformer blocks to classification performance on these tasks.

Across tasks, accuracy generally increases from Layer 0 to Layer 1, then shows minor fluctuations through Layers 1–6. This suggests that most task-relevant contextual structure is introduced in the first transformer block, while subsequent layers primarily refine representations within a relatively stable linearly decodable subspace.

5.3 Fine-tuning reorients existing geometry rather than creating new features

Comparing pretrained and fine-tuned probes at the deepest layer:

Toxicity (test1)	0.86 → 0.94 (+0.08)	0.87 → 0.94 (+0.07)
Toxicity (test2)	0.89 → 0.94 (+0.05)	0.85 → 0.94 (+0.09)
Pore-forming toxin	0.72 → 0.88 (+0.16)	0.83 → 0.88 (+0.05)
Virulence	0.68 → 0.85 (+0.17)	0.80 → 0.85 (+0.05)

The pattern across PFT and virulence is the clearest expression of the geometric reorientation hypothesis. The MMProbe gains 16–17 percentage points, while the LRProbe gains only 5. The LRProbe provides an upper bound on linearly decodable information in the representation. If fine-tuning had primarily added new features, both probes would have gained comparably. The fact that the MMProbe, which can only exploit the class-mean

axis, closes most of the gap to the LRProbe after fine-tuning indicates that the relevant subspace has been rotated into alignment with that axis. The information was largely present before, but fine-tuning made it accessible to a maximally simple decoder.

For toxicity, the picture is partially obscured by ceiling effects, as pretrained probes already reach 0.85+, leaving little headroom for either probe to gain. The PFT and virulence results are therefore the more diagnostic ones.

5.4 Cross-task transfer reveals limited overlap of harm-related subspaces

The zero-shot cross-evaluation matrix:

	Toxicity	PFT	Virulence
Toxicity-FT	**0.94**	0.53	0.50
PFT-FT	0.27	**0.94**	0.55
Virulence-FT	0.60	0.60	**0.88**

Three observations stand out:

The 0.27 cell (PFT-FT applied to toxicity) reflects systematic misclassification rather than random noise. This is expected, since pore-forming toxins are a narrower benchmark within the broader toxicity concept. Many PFT negatives overlap with toxicity positives, so the learned boundary is anti-aligned with toxicity labels.

Virulence-FT shows the broadest generalization, reaching 0.60 on both other tasks. The identical scores indicate that transfer is balanced across toxicity and PFT, but the level is modest. While above chance, the model still fails to identify 40 percent of threats in balanced test sets. Virulence is the broadest of the three concepts, so this partial transfer suggests the model has captured some general harm-related features, though not a fully shared subspace.

Toxicity-FT does not transfer (0.53 on PFT and 0.50 on virulence). The toxicity classifier appears to have learned distribution-specific features rather than a generalizable representation of harm.

---

6. Limitations

6.1 What we have shown

The strongest claim our results support is descriptive: **harm-related properties of protein sequences are linearly decodable from pretrained ESM-2 activations at accuracies well above chance, and fine-tuning predominantly reorganizes this existing structure rather than constructing new features.** This is a methodologically straightforward but empirically non-obvious finding. It places PLMs alongside language models as systems whose internal geometry exhibits emergent linear representation of properties they were not directly trained to predict.

The finding has a concrete deliverable: the probe weight vectors themselves are linear directions in ESM-2's representation space corresponding to "toxicity," "pore-forming activity," and "virulence." These directions are tractable objects for further mechanistic study.

6.2 What we have not shown

We are explicit about several gaps between our experiments and the broader biosecurity motivation in the introduction.

We have not tested AI-designed adversarial sequences. The held-out test sets in this paper are natural sequences drawn from the same distribution as those during training. The Science paper [2] reports failures of screening systems on AI-generated homologs, and addressing that finding directly would require evaluating probes (and downstream classifiers) on sequences generated by protein design models such as ProteinMPNN, ESM-IF, or RFdiffusion. We identify this as the primary follow-up experiment.

We have not evaluated production safety classifiers. We probe ESM-2 directly. Whether ToxDL 2.0 or commercial gene-synthesis screeners use representations similar to those we characterize is an empirical question we do not address. The relevance of our findings to deployed systems depends on the extent to which those systems share architectural and training similarities with ESM-2.

Linear probes, by construction, find linear structure. Non-linear features that contribute to classifier predictions are not captured by our methodology. Higher LRProbe accuracy than MMProbe accuracy is consistent with the relevant features being linearly separable. If LRProbe accuracy at the deepest layer were substantially lower than that of the fine-tuned classifier, it would suggest that non-linear features play an important role. In practice, LRProbe and the fine-tuned classifier achieve similar accuracy at the deepest layer. This similarity shows that a linear structure is sufficient for these tasks at this scale.

We have not performed causal interventions. Identifying a probe direction does not establish that the classifier uses that direction. Confirming causal use would require ablation experiments that project activations onto the orthogonal complement of the probe direction and measure classifier degradation, which we do not perform here.

We use the smallest ESM-2 variant. Larger ESM-2 models (650M to 15B parameters) may exhibit different probe profiles. In particular, the representation geometry may be cleaner or messier at scale, and the relative gains from fine-tuning may differ. The 8M variant was chosen for computational reasons, and our findings should be understood as applicable to that scale.

Test2 of the toxicity benchmark is class-imbalanced. With a 92 percent majority-class baseline, the headline accuracy of 0.94 on Test2 represents only a small absolute gain. For this reason, we treat Test1 as the primary toxicity evaluation throughout.

Three tasks are a small number with overlapping definitions of "harm." Our cross-task structure claim would be strengthened by tasks constructed to be more orthogonal (e.g., a non-harm-related protein property as a control) and weakened if the partial overlap we observe is partly an artifact of dataset construction. A control task is a clear methodological extension.

6.3 Interpreting the cross-task result carefully

The cross-task evidence for a "shared harm subspace" should be read as suggestive rather than definitive. Vir-FT shows partial transfer to both toxicity and PFT, which is consistent with virulence capturing some general harm-related features. In contrast, Tox-FT shows little transfer, and PFT-FT is anti-correlated with toxicity. This pattern suggests overlapping but distinct feature sets rather than a single unified subspace. Distinguishing these possibilities would require either more tasks, so that genuine sharing produces a low-rank structure across many transfer matrices, or direct geometric analysis of the probe direction vectors. We do not attempt either here.

The anti-correlated PFT-to-toxicity transfer is the most interpretable finding, since pore-forming toxins are a subtype of toxic peptides, and the biological relationship predicts the sign of the effect.

---

7. Conclusion

Protein language models that are never explicitly trained on harmful properties nonetheless develop internal representations in which harm-related features are approximately linearly separable. Fine-tuning on harm classification tasks largely reorganizes this pre-existing structure rather than creating it. The three harm-related tasks we study show partial representational overlap, with patterns that reflect their biological relationships.

These findings do not, by themselves, explain why deployed protein safety classifiers fail on AI-designed variants. They do, however, reveal the representational substrate on which a mechanistic explanation could be built. The probe directions we identify are concrete vectors in ESM-2's activation space that can be ablated, steered, and compared against features used by deployed classifiers. The natural next step is to evaluate these probes against AI-generated adversarial sequences and to perform causal interventions that test whether the directions we identify are the ones classifiers actually rely on.

Treating protein safety classifiers as objects of mechanistic study, rather than as black boxes whose failures are characterized only at the input/output level, offers a path toward understanding their failure modes that complements adversarial evaluation. This paper takes a first step in that direction.

Code and Data

- ***Code repository:*** *https://github.com/edric-hao/PLM-experiments*

References

1. *Watson et al. (2023). RFdiffusion. Nature. https://doi.org/10.1038/s41586-023-06415-8*
2. *Dauparas et al. (2022). ProteinMPNN. Science. https://doi.org/10.1126/science.add2187*

3. *Khan et al. (2025). SafeBench-Seq. arXiv arXiv:2512.17527. https://arxiv.org/abs/2512.17527*
4. *Urbina et al. (2022). Dual use of AI-powered drug discovery. Nature Machine Intelligence. https://doi.org/10.1038/s42256-022-00465-9*
5. *Sharma et al. (2022). ToxinPred2. Briefings in Bioinformatics. https://doi.org/10.1093/bib/bbac174*
6. *Lin et al. (2023). ESM-2. Science. https://doi.org/10.1126/science.ade2574*
7. *Jiao et al. (2024). Integrated convolution and self-attention for improving peptide toxicity prediction. Bioinformatics 40(5): btae297. https://doi.org/10.1093/bioinformatics/btae297*
8. *Zhu et al. (2025). ToxDL 2.0. Computational and Structural Biotechnology Journal. https://doi.org/10.1016/j.csbj.2025.04.002*
9. *Conneau et al. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. ACL 2018. arXiv:1805.01070. https://arxiv.org/abs/1805.01070*
10. *Marks and Tegmark (2023). Geometry of truth. arXiv arXiv:2310.06824. https://arxiv.org/abs/2310.06824*
11. *Saier et al. (2021). TCDB. Nucleic Acids Research. https://doi.org/10.1093/nar/gkaa1004*
12. *Dal Peraro and van der Goot (2016). Pore-forming toxins. Nature Reviews Microbiology. https://doi.org/10.1038/nrmicro.2015.3*
13. *UniProt Consortium (2025). UniProtKB 2025. Nucleic Acids Research. https://doi.org/10.1093/nar/gkae1010*

Appendix A: Limitations and Dual-Use Considerations

A.1 Scope limitations (technical)

The technical limitations of this work are listed in §6.2. We summarize them here for ease of reference:

- *Smallest ESM-2 variant only; no scaling study*
- *Three tasks with overlapping harm-related definitions; no orthogonal control task*
- *No evaluation on AI-designed or adversarial sequences*
- *No evaluation on production safety classifiers (e.g., ToxDL 2.0, commercial synthesis screeners)*
- *No causal intervention experiments confirming that probe directions are the directions classifiers actually use*
- *Linear probes only; non-linear feature use is not characterized.*
- *Class imbalance in toxicity test 2 partially obscures the true performance lift on that test set.*

A.2 Dual-use considerations

The general concern with publishing interpretability work on protein safety classifiers is that a sufficiently mechanistic understanding of those classifiers could, in principle, inform attempts to evade them. We have considered this carefully and judge the present work to be on the safe side of the relevant tradeoff for the following reasons.

The findings are at the level of "such directions exist," not "here is how to perturb a sequence to avoid detection." *We characterize that ESM-2 representations linearly separate harm-related classes. We do not provide gradient-based attack recipes, generate adversarial sequences, or train models intended to evade screeners. The information uplift to a hypothetical bad actor is small relative to the published work that already documents classifier failures on AI-designed variants [2].*

The probe directions we identify are not directly weaponizable. *A linear direction in ESM-2's 320-dimensional CLS-activation space is not a sequence. Mapping it back to an adversarial sequence would require substantial additional work (gradient inversion, generative modeling, or design-then-screen pipelines), which we neither perform nor enable beyond what is already in the public PLM literature.*

The constructive use case is clearer than the destructive one. *The natural follow-up is to test whether classifiers rely on the directions we identify and to use those directions to build more robust screeners, which is a squarely defensive application. Our recommended next experiments, including causal interventions, evaluation on AI-designed variants, and comparison to production classifiers, are diagnostic rather than generative.*

We use a small open-source model. *Our findings are about representations in esm2_t6_8M_UR50D, the smallest publicly available ESM-2 variant. We make no claims about whether the same structure exists in larger variants or proprietary models, and we do not provide tooling to generalize an attack across models.*

We support the development of biosecurity-relevant interpretability work being conducted openly within research communities that maintain dual-use review norms. We would not support publication of follow-up work that crosses the line from characterizing representations to providing operational attack recipes against deployed screeners. That determination should be made on a case-by-case basis with input from biosecurity practitioners.

A.3 What would change our risk assessment

If subsequent work were to (a) demonstrate that the probe directions we identify causally drive the predictions of production safety classifiers, and (b) provide a tractable method for generating sequences that perturb these directions while preserving function, the resulting bundle would warrant restricted disclosure. We flag this in advance because the natural research trajectory from this paper passes through (a) before reaching (b), and we believe the field benefits from authors stating in advance which results they would and would not publish openly.

A.4 Data and compute statement

All datasets used in this work are publicly available (UniRef50, UniProtKB, TCDB, public peptide toxicity benchmarks). No private or restricted-access data was used. Total compute is on the order of a few GPU-hours on consumer hardware. This work is reproducible by any team with access to a single modern GPU.

LLM Usage Statement

We used Claude to help write code snippets and draft sections. All results and claims were independently verified.