Probing Risk Representations in Protein Language Models[1]

Vishesh Gupta
Independent

With
Apart Research

Abstract

DNA synthesis screening currently relies on matching sequences against known-threat databases (BLAST-style). This breaks when AI-generated sequences retain dangerous function while differing enough in primary sequence to evade detection. We tested whether ESM-2, a widely-used protein language model, internally encodes a pathogenicity direction that could serve as a supplementary screening signal. Using 1,264 labelled sequences across 9 pathogen families and their benign relatives, we trained linear classifiers on ESM-2 activations and evaluated them on three completely withheld families. The central finding is negative: the global classifier fails to generalise across taxonomic families (test AUROC 0.694 for 650M, 0.622 for 3B), with no statistically significant advantage over BLAST (McNemar's $p = 0.128$). We also identify two failure modes relevant to AI-screening contexts: the classifier

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

flags scrambled nonsense sequences as dangerous (FPR jumps from 17.7% to 66.1%), and performance degrades substantially on short 60-residue fragments (TPR drops from 48.6% to 37.6%). Within individual families, however, family-specific classifiers work well (AUROC 0.828-0.992) and complement BLAST's annotation coverage gaps in Clostridiaceae. We propose a two-stage screening architecture: taxonomy routing via homology tools, followed by family-specific ESM-2 probes to address annotation gaps.

1. Introduction

Current DNA synthesis screening flags a sequence if it closely resembles something already in a threat database. This works for known variants of known pathogens. It fails when sequences are designed to retain dangerous function while looking novel enough to evade homology detection-exactly the scenario that AI protein design tools like RFdiffusion and Evo2 make increasingly plausible.

Protein language models like ESM-2 are trained on hundreds of millions of sequences and learn rich internal representations of protein biology. We asked: do those representations encode something about biosecurity risk? If ESM-2 has learned a detectable separation between dangerous and safe proteins in its activation space, probing that direction could catch sequences that evade BLAST. If not, that is itself a useful finding about the limits of using these models for screening.

Our main contributions:

- *A negative result with statistical support: global ESM-2 linear probes fail to generalise pathogenicity across unseen taxonomic families (AUROC approx 0.69), performing no better than BLAST keyword matching (McNemar's $p = 0.128$).*
- *Two failure modes specific to AI-screening contexts: OOD artefact sensitivity (scrambled sequences flagged as dangerous) and short-fragment degradation.*

- *A positive finding with caveats: within-family probes achieve AUROC 0.828-0.992 and complement BLAST's annotation gaps for Clostridiaceae.*

A proposed hierarchical architecture motivated by the complementary failure modes of BLAST and local ESM-2 probes.

2. Related Work

SecureDNA [1] screens synthesis orders via cryptographic k-mer matching, effective down to 30bp. The IBBIS Common Mechanism [2] uses profile HMMs, best above 150bp. Both are homology-based: they flag sequences that resemble known threats. AI-designed sequences can evade both by differing sufficiently in primary sequence while retaining function.

Wittmann et al. [3] demonstrated this vulnerability directly: using open-source AI protein design tools, they generated over 75,000 variants of proteins of concern and found that existing screening tools missed a substantial fraction. Their work motivates asking whether representation-based approaches could close the gap.

Lin et al. [4] showed that ESM-2 representations encode structural and evolutionary information at atomic resolution. Prior probing work has decoded secondary structure and contact maps from ESM-2 activations, but no prior work has specifically probed for biosecurity-relevant properties.

Liu et al. (ABC-Bench, NeurIPS 2025) [5] showed that LLM-based agents can perform biosecurity-relevant tasks end-to-end. Our work is complementary: we examine sequence model internals rather than chat model or agent outputs.

3. Methods

3.1 Dataset

We fetched labelled protein sequences from NCBI Protein (Entrez API, taxid-based queries) and UniProt SwissProt. Positives came from CDC/USDA Select Agent organisms; negatives came from taxonomically matched non-pathogenic relatives within the same family. After length filtering (80-2000 AA), deduplication, and

per-family balancing (max 1.4:1 negative-to-positive), the final dataset had 1,264 sequences across 9 families. Thin families (orthomyxovirus: 4 sequences, paramyxovirus: 9, plant_toxin: 2) were excluded from all probing analysis.

3.2 Extraction

We extracted mean pooled residue representations from all 34 hidden layers of ESM-2 650M and 7 checkpoint layers of ESM-2 3B. Sequences were truncated to 1,022 residues.

3.3 Holdout Split Design

A random split would let the probe memorise family signatures and pass in-distribution-not the threat model we care about. We used a zero-shot cross-family holdout instead: three complete families were withheld from training (poxvirus, enterobacteriaceae, coronavirus). The probe never sees any sequence from these families during training. Remaining data was split 90/10 for train/validation.

> *Caveat: the test set is 64% positive (109/62) because the holdout selected more positive than negative organisms. AUROC is unaffected; TPR and FPR are reported separately rather than relying on accuracy.*

3.4 Classifiers and Baselines

A logistic regression probe was trained at each layer (L2 regularisation, C=1.0, balanced class weights). Best layer selected by validation AUROC. A shallow MLP was trained at the best layer as a secondary comparison. For BLAST: BLASTp against a local SwissProt database (574,627 sequences), flagging sequences with hits matching 22 select-agent keywords at e-value $<$ 1e-5. McNemar's test (exact, two-sided) compared classifier accuracy on the same 171 test sequences.

3.5 Robustness Tests

Four additional experiments:

(1) intra-family probing via 5-fold CV within each family; (2) evasion test-random substitution of 0-30% of residues, re-extracting embeddings, comparing local probe vs BLAST detection; (3) fragment test-truncating positives to 60 AA; (4) OOD artefact test-randomly scrambling negative sequences and checking the global probe's false positive rate.

4. Results

4.1 Global Probe Fails on Novel Families

The probe achieves strong in-distribution validation performance (val AUROC 0.969 at layer 21). On three completely held-out families it collapses:

Method	*Test AUROC*	*FNR*	*F1*
ESM-2 650M LR (layer 21)	*0.694 (95% CI: 0.595-0.756)*	*0.495*	*0.625*
ESM-2 650M MLP (layer 21)	*0.610*	*0.404*	*0.657*
ESM-2 3B LR (layer 30)	*0.622*	*0.532*	*0.560*
BLAST (SwissProt keyword)	*-*	*0.376*	*0.680*

Table 1. Test-set results (N=171). FNR = fraction of dangerous sequences missed; lower is better for screening. BLAST has no continuous score so AUROC is not reported.

Validation AUROC of 0.969 collapsing to 0.694 on held-out families is the central finding: ESM-2 representations organise sequences by evolutionary family, not by biosecurity risk. Scaling to 3B parameters makes this worse, not better (test AUROC 0.622), ruling out model capacity as the explanation.

4.2 No Statistical Advantage Over BLAST

McNemar's test on paired predictions: ESM-2 correct and BLAST wrong on 35 instances; BLAST correct and ESM-2 wrong on 50 instances. $p = 0.128$. No statistically significant difference. A 650M parameter transformer provides no measurable advantage over keyword-based homology matching on this test set.

4.3 Two AI-Specific Failure Modes

OOD artefact: We scrambled the residues of 62 benign test sequences to create biophysically impossible sequences and checked how often the probe flagged them. False positive rate rises from 17.7% to 66.1%. The probe partly functions as an

out-of-distribution detector rather than a risk detector-a problem for screening AI-designed sequences that may have unusual patterns.

Fragment degradation: Truncating 109 positive test sequences to their first 60 residues (simulating short synthesis orders) drops TPR from 48.6% to 37.6%. Mean pooling over a short fragment destroys the risk signal.

4.4 Within-Family Probes and Complementary Failure Modes

When restricted to within a single family, ESM-2 performs substantially better:

Family	*N (pos/neg)*	*Local AUROC*	*Std Dev*
clostridiaceae	*105 / 119*	*0.992*	*0.009*
burkholderia	*141 / 158*	*0.956*	*0.027*
coronavirus	*44 / 32*	*0.951*	*0.062*
filovirus	*37 / 27*	*0.937*	*0.085*
coxiellaceae	*46 / 56*	*0.928*	*0.049*
enterobacteriaceae	*79 / 110*	*0.927*	*0.034*
francisellaceae	*36 / 26*	*0.896*	*0.072*
bacillus	*72 / 100*	*0.863*	*0.047*
poxvirus	*32 / 44*	*0.828*	*0.138*

Table 2. Intra-family probing (5-fold CV, in-distribution). Not all families exceed 0.90 AUROC: bacillus (0.863) and poxvirus (0.828) fall below. Poxvirus variance is high (small N).

The evasion test reveals that BLAST and local probes fail in different places:

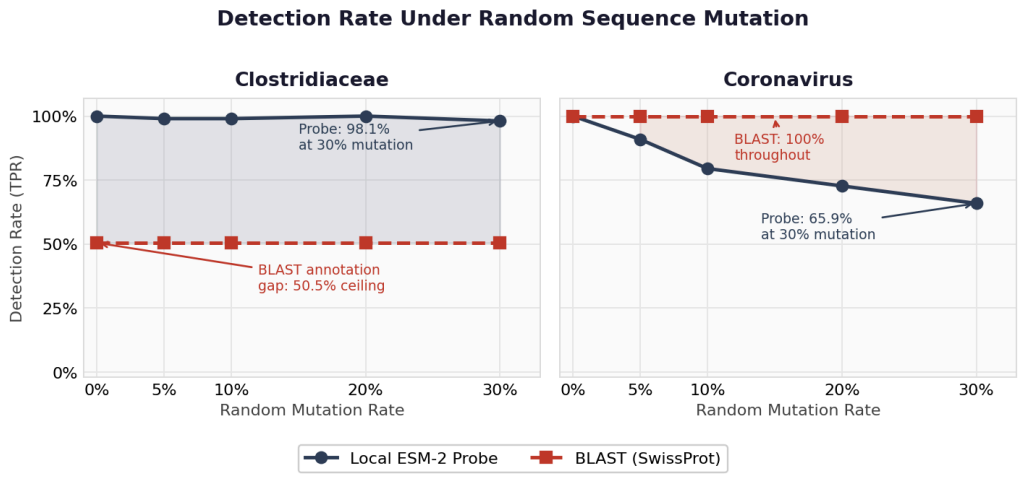

Figure 1. Detection rate vs random mutation rate for local ESM-2 probe vs BLAST. Left: Clostridiaceae-BLAST stagnates at 50.5% regardless of mutation (annotation gap); probe maintains ~98% through 30% mutation. Right: Coronavirus-BLAST maintains 100% (dense SwissProt coverage); probe degrades to 65.9% at 30% mutation. See Limitations for caveats on both results.

5. Discussion and Limitations

Discussion

The main finding is that ESM-2 is useful in a different way than hoped. A universal danger signal based on global linear probing does not exist in ESM-2's representation space. The model was trained to predict masked amino acids given evolutionary context, not to separate dangerous from safe. Within a family the relevant distinction is smaller-dangerous vs benign members of the same evolutionary group-and the representation geometry is more favourable for that task.

The complementary failure modes point toward a practical architecture: homology tools for fast taxonomy identification, then family-specific ESM-2 probes to address annotation gaps and evasive variants. This is a proposal, not a validated system.

Limitations

- *Organism-level labels. All labels were assigned at organism level (Select Agent status), not protein level. The probe may be learning to distinguish organism identity rather than function.*
- *In-distribution local probing. Within-family probes in Table 2 and Figure 1 train on the same organisms they test on (via cross-validation). This is not zero-shot generalisation. True evaluation would require holding out organisms within a family.*
- *Evasion test is in-distribution and uses naive mutations. The probe trained on the exact sequences it then tested on (mutated). Random residue substitutions also destroy fold stability-a real adversary using RFdiffusion would design functionally stable variants, which is a harder problem. The ~100% TPR numbers are an upper bound.*
- *Narrow BLAST keyword list in evasion test. The Clostridiaceae BLAST stagnation at 50.5% is partly because the evasion-test script used only 6 keywords. The main test set evaluation used 22 keywords and reached 62.4% TPR.*
- *No comparison to production tools. The BLAST baseline is not SecureDNA or IBBIS Common Mechanism, which have different architectures and may perform differently.*
- *Probing is correlational, not causal. We show that pathogenicity labels are linearly decodable from ESM-2 activations. This does not mean the model uses this information for any biological task.*

Dual-Use and Ethical Considerations

This work has dual-use implications because it evaluates weaknesses in sequence-screening approaches. A malicious actor could misuse information about model failure modes, short-fragment degradation, or gaps between homology-based and representation-based screening to reason about screening bypasses. For this reason, we report only aggregate results and avoid releasing operational bypass instructions, deployment thresholds, or sequence-level evasion recipes.

The intended use of this work is defensive: to caution against naive deployment of global protein-language-model risk probes and to motivate more rigorous evaluation

of proposed AI-based screening tools. If future work identifies concrete vulnerabilities in deployed screening systems, those findings should be shared first with relevant screening providers or biosecurity organizations before broad public release.

Ethically, the main risks are false confidence and overblocking. A poorly validated model could miss genuinely concerning sequences while giving screening providers a misleading sense of safety. Conversely, high false-positive rates could burden legitimate research. Any future system based on this direction should include calibrated uncertainty, human review, and careful validation against production-grade screening tools.

Future Work

1. *Protein-level functional labels to separate organism identity from functional risk.*
2. *A within-family zero-shot holdout to get unbiased estimates of local probe generalisation.*
3. *Testing against genuinely AI-designed evasive sequences rather than random mutations.*

6. Conclusion

ESM-2 does not encode a universal, linearly separable pathogenicity signal. Global probes fail on novel taxonomic families (AUROC 0.694 for 650M, 0.622 for 3B) with no statistically significant advantage over BLAST ($p = 0.128$). Two additional failure modes make global probing unsuitable for AI-screening applications: OOD artefact sensitivity and short-fragment degradation.

Within families, local probes achieve AUROC 0.828-0.992 and complement BLAST's annotation coverage gaps for Clostridiaceae. The practical recommendation is a hierarchical design-homology tools for taxonomy routing, family-specific ESM-2 probes for annotation gaps-but deploying this properly requires protein-level labels, holdout evaluation of local probes, and testing against AI-designed rather than randomly-mutated sequences.

References

[1] SecureDNA Consortium. SecureDNA: Free, open-source cryptographic DNA synthesis screening. securedna.org.

[2] IBBIS. Common Mechanism (commec): HMM-based biorisk screening. github.com/ibbis-screening/common-mechanism.

[3] Wittmann, B.J., Alexanian, T., Bartling, C., et al. Strengthening nucleic acid biosecurity screening against generative protein design tools. Science 390, 82-87 (2025). DOI: 10.1126/science.adu8578.

[4] Lin, Z., Akin, H., Rao, R., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123-1130 (2023). DOI: 10.1126/science.ade2574.

[5] Liu, A.B., Nedungadi, S., Cai, B., Kleinman, A., et al. ABC-Bench: An Agentic Bio-Capabilities Benchmark for Biosecurity. NeurIPS 2025 Workshop on Biosecurity Safeguards for Generative AI.

[6] Biden White House OSTP. Framework for Nucleic Acid Synthesis Screening. April 2024.

[7] Biosecurity Modernization and Innovation Act of 2026 (S.3741). 119th Congress.

Appendix

Supplementary figures and detailed results. All numbers are from confirmed experimental outputs.

Figure A1. Within-Family (Local) Probe AUROC

5-fold stratified cross-validation AUROC for each family. Error bars show standard deviation across folds. Families below the 0.90 threshold (bacillus, poxvirus) are highlighted in red. These are in-distribution estimates; see Limitations on why they should not be read as zero-shot generalisation performance.

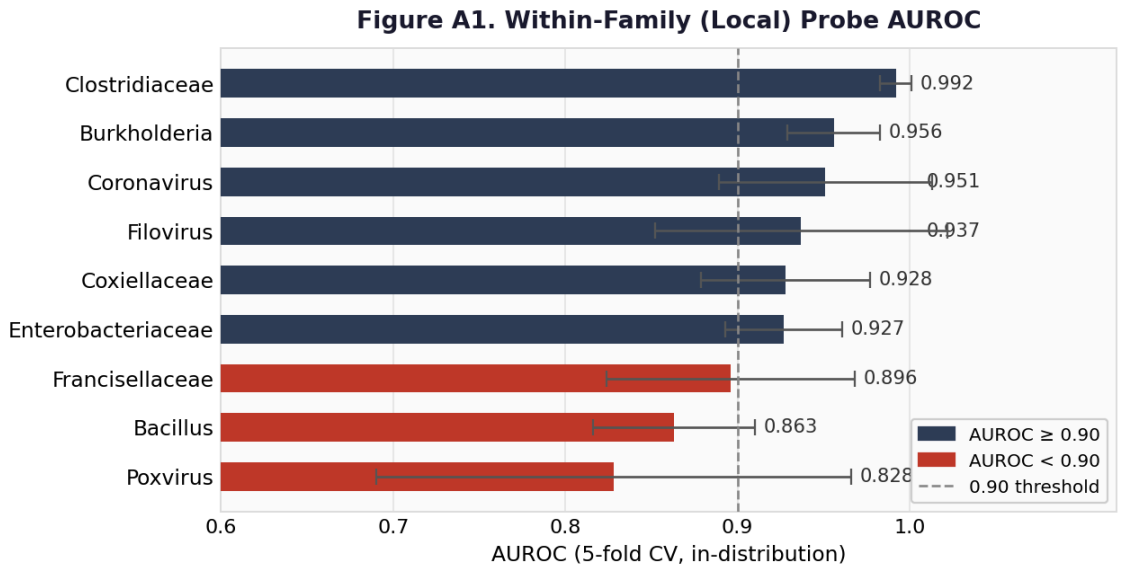

Figure A1. Local probe AUROC by family. Six of nine families exceed 0.90 AUROC within-family. Bacillus (0.863) and poxvirus (0.828) fall below. Poxvirus has high variance (SD = 0.138) reflecting its small positive sample size (N = 32).

Figure A2. Dataset Composition

Positive (pathogenic) and negative (benign) sequence counts per family after balancing. Numbers inside bars show per-class counts; numbers above bars show family totals. Families are sorted by total size.

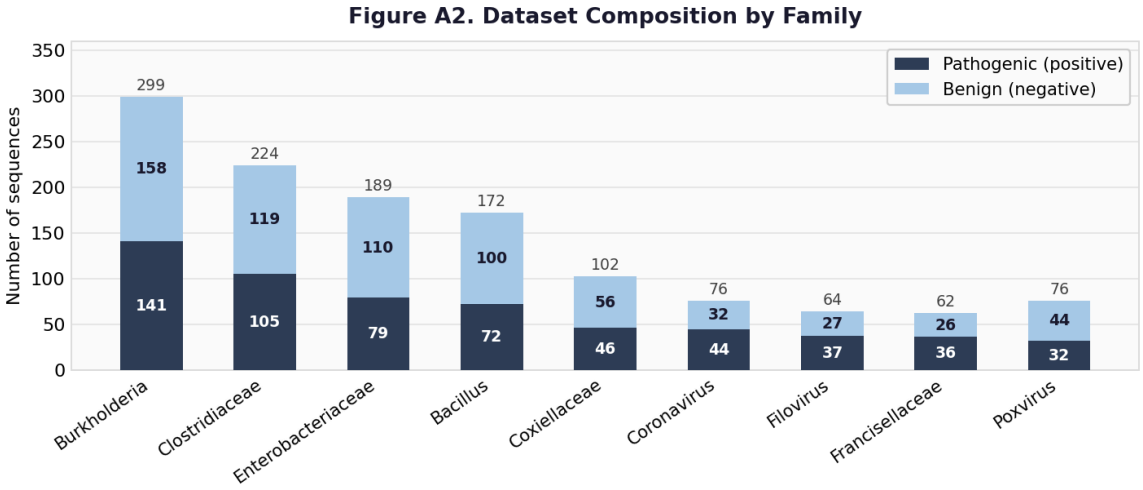

Figure A2. Dataset composition. Burkholderia and Clostridiaceae are the largest families. Thin families (orthomyxovirus, paramyxovirus, plant_toxin) are excluded from this chart as they were not used in probing analysis.

Table A1. Full Classification Report – ESM-2 650M Global Probe

Per-class precision, recall, and F1 for the LR and MLP classifiers on the 171-sequence test set. Note the test set is 64% positive, which affects precision and recall independently.

Classifier	Class	Precision	Recall	F1	Support
LR (layer 21)	Benign	0.48	0.81	0.60	62
LR (layer 21)	Pathogenic	0.82	0.50	0.62	109
LR (layer 21)	Macro avg	0.65	0.66	0.61	171
MLP (layer 21)	Benign	0.46	0.61	0.53	62
MLP (layer 21)	Pathogenic	0.73	0.60	0.66	109
MLP (layer 21)	Macro avg	0.60	0.60	0.59	171

Table A1. LR achieves higher AUROC (0.694) while MLP achieves lower FNR (0.404). The difference reflects different operating points rather than a meaningful superiority of either method.

Table A2. Train / Val / Test Split

Split	N	Positive	Negative	Positive Rate	Families
Train	984	435	549	0.44	8 (all excl. holdout)
Val	109	48	61	0.44	8 (random 10%)
Test	171	109	62	0.64	3 (poxvirus, entero., corona.)

Table A2. The test set positive rate of 0.64 arises because the three held-out families contributed more positive than negative organisms. AUROC is unaffected by this imbalance.

Table A3. Full Evasion Test Results

Family	Mutation	Probe TPR	BLAST TPR
Clostridiaceae	0%	1.000	0.505
Clostridiaceae	5%	0.990	0.505
Clostridiaceae	10%	0.990	0.505
Clostridiaceae	20%	1.000	0.505

Clostridiaceae	*30%*	*0.981*	*0.505*
Coronavirus	*0%*	*1.000*	*1.000*
Coronavirus	*5%*	*0.909*	*1.000*
Coronavirus	*10%*	*0.795*	*1.000*
Coronavirus	*20%*	*0.727*	*1.000*
Coronavirus	*30%*	*0.659*	*1.000*

Table A3. Complete evasion test results at all five mutation rates. BLAST Clostridiaceae rate is constant because the evasion-test keyword list (6 terms) misses most Clostridiaceae positives regardless of mutation. BLAST Coronavirus rate is constant because dense SwissProt coverage preserves hits even at 30% mutation.

Table A4. McNemar's Test Contingency Table

Paired comparison of ESM-2 LR (layer 21) vs BLAST binary predictions on the 171 test sequences.

	BLAST correct	*BLAST wrong*
ESM-2 correct	*84*	*35*
ESM-2 wrong	*50*	*2*

Table A4. Off-diagonal cells (35 and 50) are used in McNemar's test. $p = 0.128$ (exact, two-sided): no significant difference in error rates. BLAST was correct and ESM-2 wrong on 50 instances vs 35 in the reverse direction.

Code

- ***Code repository**: Github*

LLM Usage Statement

The text of this report was written by the author, with Claude utilized solely to correct grammatical errors and refine readability. Claude was also used to draft the initial iterations of the experimental code. Every line of generated code, along with

all experimental results and final claims, was manually reviewed, tested, and independently verified by the author.