BEYOND SEQUENCE SIMILARITY: PROTEIN AND DNA EMBEDDINGS FOR EVASION-RESILIENT SYNTHESIS SCREENING[1]

Robert Amanfu

With
Apart Research

Abstract

Biosecurity screening of synthesized or environmental DNA mostly asks: does this look like something on a known-toxin list? *That fails when an attacker (or evolution) changes enough letters to escape the lookup while keeping the protein's harmful function. We test whether biological sequence models trained on proteins and DNA can spot that function directly. On 4,060 toxin and benign coding sequences, evaluated with close homologs (evolutionarily related sequences) kept out of training, an ensemble of a DNA model (Evo2 7B) and a protein model (ESM-2) catches **85.8%** of toxins at a 1-in-100 false-alarm rate, versus **72%** for Evo2 alone, **71%** for ESM-2 alone, and **55%** for a simple 5-letter-pattern baseline. By comparison, the open-source COMMEC policy screen (biorisk-only mode) flags only **16.8%** of the same toxins, showing that learned models catch toxins the curated lookup databases miss. After mutating 60% of amino acids to disguise toxins, the protein screen still recovers **98%**; the baseline drops to **41%**. Reliable behavior requires DNA fragments ≥**1,500 bp**. Caveats: our mutations are random rather than structure-aware, and the length test was not run on independent, unseen samples.*

1. Introduction

The problem. When a customer orders synthetic DNA or a public health lab tests wastewater, the screening process checks if the sequence matches a list of known toxins. Comparing DNA letters (nucleotides) deters simple copy-paste attacks. However, this approach fails if someone alters the letters enough to avoid detection, even when the protein's 3D shape and harmful activity remain unchanged. Directly spotting the harmful biological function, rather than relying on surface resemblance, would close that gap.

Research question. *Can biological sequence models trained on proteins (ESM-2) and DNA (Evo2) spot dangerous sequences that standard letter-matching screens miss? How short can the input be before these models lose the ability to flag risks?* We split this into three sub-questions:

- **RQ1: Signal and combination.** Do these models pick up functional signals that simple letter-frequency baselines miss, even when we make the test harder by keeping close homologs out of training? Does combining their (standardized) decision values beat either model alone?
- **RQ2: Evasion.** When toxins are disguised, either by swapping in different DNA letters that code for the same protein or by mutating the protein itself, does the screen still catch them at a strict 1-in-100

[1]This work was produced at the AIxBio Hackathon (April 2026) under time constraints. Results have not been independently replicated. Findings should be treated as preliminary and interpreted with appropriate caution pending peer review.

false-alarm rate?

- **RQ3: Short reads.** How short can the DNA input get before the false-alarm rate balloons past its target?

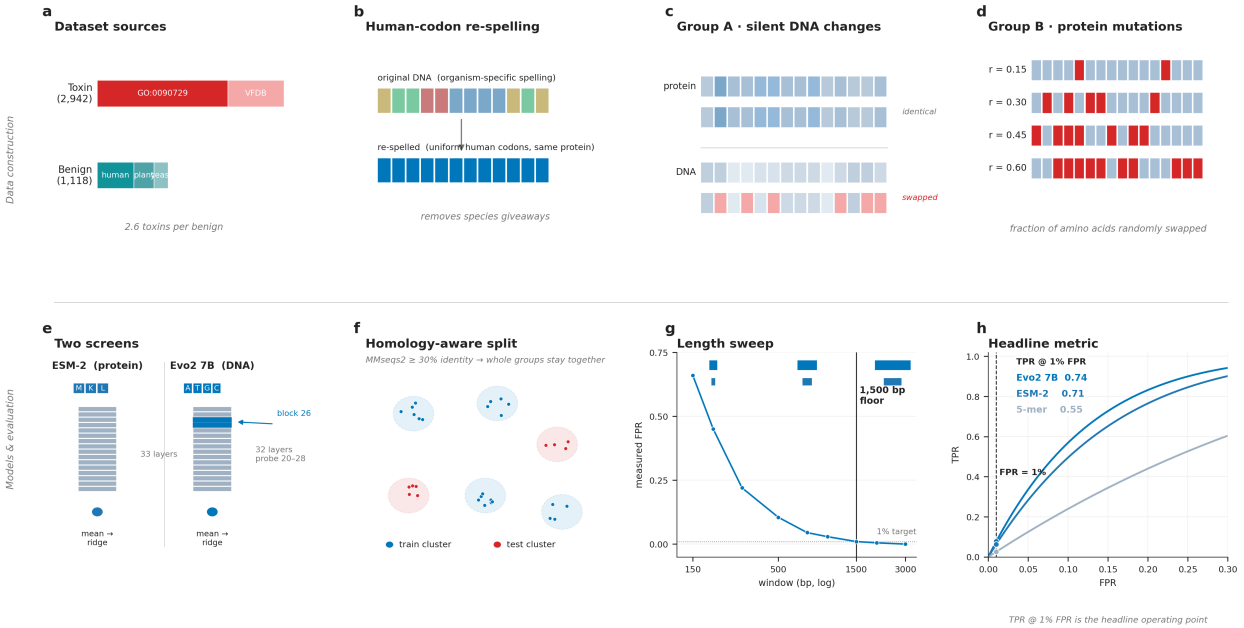

Figure 1. Experimental overview. ***Top, data construction.*** ***(a)*** *2,942 toxins (GO:0090729 + VFDB) and 1,118 benigns (human-secreted, plant, yeast).* ***(b)*** *Human-codon re-spelling: every coding sequence is rewritten with uniform human codons so any DNA-side signal must come from biology, not species-of-origin spelling.* ***(c)*** *Group A evasion: silent DNA changes leave the protein identical.* ***(d)*** *Group B evasion: random missense ladder at $r \in \{0.15, 0.30, 0.45, 0.60\}$.* ***Bottom, models and evaluation.*** ***(e)*** *ESM-2 (33 layers) and Evo2 7B (32 layers, probed at blocks 20–28, deployed at block 26), consumed as mean-pooled embeddings into a linear classifier.* ***(f)*** *Homology-aware split via MMseqs2 at ≥30% identity.* ***(g)*** *Length sweep locating the 1,500 bp working floor.* ***(h)*** *Headline ROC; the embedding screen is positioned as a complementary layer alongside alignment-based triage, not a replacement.*

2. Related Work

Policy baselines and alignment-based screens. The U.S. government's screening guidance (HHS *Common Mechanism*, 2024) directs DNA synthesis providers to flag orders by comparing them with known-hazard databases, mainly through letter-by-letter alignment and short-pattern lookups. Wittmann & Olsen et al. (2024) showed these screens fail when a toxin's protein sequence is altered enough to disrupt the lookup but not the function. Tayouri et al. (2025) found that splitting a synthesis order across multiple providers also evades alignment-based screens. This motivates function-aware detection resilient to fragmentation and sequence disguise. To benchmark our learned screens against a real-world tool, we compare with IBBIS's Common Mechanism package (`commec`) in biorisk-only mode (Section 3.5). Our evasion test, deliberately simpler than structure-guided protocols, uses random mutations as a stress test rather than a full adversarial benchmark.

Biological sequence models. Two families of large models now read biological sequences end-to-end: protein language models such as ESM-2 (Lin et al., 2023) and DNA foundation models such as Evo2 (Arc/Stanford/Berkeley, 2024–2025). We also tested a much smaller DNA model (HyenaDNA-1M) early on, which lagged ESM-2 by a wide margin (Appendix A.6).

Simple baselines. Counting how often each short DNA pattern appears (the 5-letter-pattern baseline used here) gives us a floor to measure the learned screens against.

3. Methods

3.1 Dataset

We built a labeled dataset of toxin (positive) and benign (negative) protein-coding sequences. **Toxins** come from a Gene Ontology category for "toxin activity" (GO:0090729) and the Virulence Factor Database (VFDB). **Benigns** are from human-secreted, plant, and yeast proteins; none appear in the toxin catalogs. After filtering DNA and protein records for mismatches, we have **2,942 toxins and 1,118 benigns** (about 2.6 toxins per benign). We set aside a fixed 15% slice (610 sequences) for tuning the alarm threshold and use 3,450 for training and evaluation.

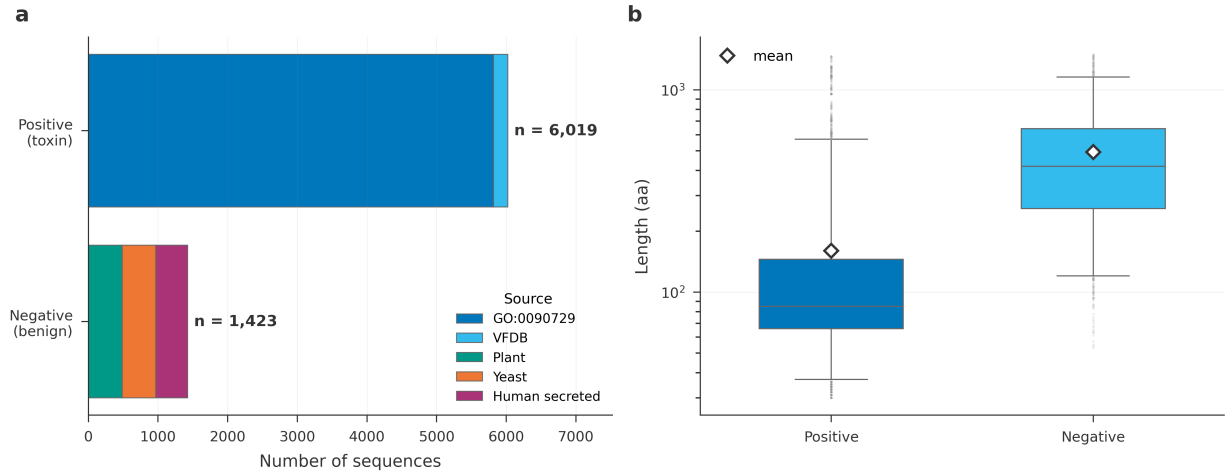

Figure 2. ***(a)*** *Counts of toxins and benigns, stacked by source catalog.* ***(b)*** *Protein length distribution; box covers the middle 50% of sequences, whiskers cover the 5th to 95th percentiles, diamond marks the mean.*

3.2 Homology-aware split

To prevent screens from getting credit for memorizing close homologs of training toxins, we group homologous sequences using **MMseqs2 at 30% protein identity**. We keep entire groups together, assigning each either to training or testing — never both. Without this guard, homologous toxins leak between folds and inflate metrics. Appendix A.6's HyenaDNA leaderboard for the older, non-grouped split illustrates this. All Section 4 numbers use the homology-aware split.

3.3 Models

ESM-2 (Lin et al., 2023) is a 33-layer protein language model that reads amino-acid sequences and learns to predict hidden amino acids, picking up patterns in protein shape and function. We extract a single numeric summary (mean-pooled embedding) per protein and feed it to a linear classifier.

Evo2 7B (Arc/Stanford/Berkeley, 2024–2025) is a 32-layer, 7-billion-parameter DNA foundation model. It reads raw nucleotide sequences and learns to predict the next letter across long genomic regions. We extract one numeric summary per internal layer (keeping layers 20–28 of the 32-layer stack) and train a linear classifier on each. Block 26 is used as the primary DNA screen, based on its ranking score.

3.4 Headline metric

The headline number is the **catch rate at a 1-in-100 false-alarm rate**, abbreviated **TPR@1%**. This is the true-positive rate when the false-positive rate is 1%. We also report AUROC (a 0-to-1 ranking score where 1.0 is perfect and 0.5 is chance) and TPR at the stricter 0.1% and looser 5% settings. The 15% held-out slice is used only to set the alarm threshold. We never use it to choose between models.

3.5 COMMEC evaluation

We evaluate COMMEC v1.0.0 in biorisk-only mode on our evaluation set. This mode uses only the curated biorisk database and the low-concern clearance databases. It skips a large-scale taxonomy search against NCBI reference databases (which would require 300–650 GB of reference data). COMMEC is a threshold-free flag system, not a continuous scorer, so AUROC and TPR@1% cannot be computed directly; we report its native Flag operating point as a conservative lower bound.

3.6 Evasion design

We selected 1,000 toxins at random and created two disguised versions each. **Group A (silent DNA changes)** substitutes DNA letters with alternatives that encode the *same* protein, thereby disrupting DNA matching but preserving the protein. **Group B (protein mutations)** randomly alters amino acids at rates of 15%, 30%, 45%, and 60%, which changes the protein. We assess recovery as the fraction of disguised toxins still detected at TPR@1%, keeping the threshold fixed on held-out benigns. For scoring, mutated proteins are reprocessed through ESM-2 to obtain new embeddings.

Full details on the 5-letter-pattern baseline, embedding extraction (how we obtain numeric summaries from each model), layer-wise probing (testing each internal layer separately), and the length test are in Appendix A.

4. RQ1 — Signal and Combination

4.1 Model signal comparison

Both ESM-2 and Evo2 catch toxins far more reliably than the 5-letter baseline.

Table 1. Screen leaderboard. Close homologs kept apart between train and test (Section 3.2), sorted by TPR@1%.

Method	AUROC	PR-AUC	TPR@0.1%	TPR@1%	TPR@5%
ESM-2 + Evo2 block-26 (averaged scores)	**0.993**	**0.997**	**0.446**	**0.858**	0.967
Evo2 block-27 + Linear	0.990	0.995	0.132	0.737	**0.971**
Evo2 block-26 + Linear	0.990	0.996	0.366	0.722	0.965
ESM-2 + Linear	0.985	0.993	0.175	0.705	0.934
5-mer + Logistic	0.869	0.952	0.154	0.548	0.636

Method	AUROC	PR-AUC	TPR@0.1%	TPR@1%	TPR@5%
COMMEC biorisk-only (Flag)†	–	–	–	0.168‡	–

†COMMEC results use biorisk-only mode for comparable evaluation; a threshold-free flag system, not a continuous scorer, so AUROC and PR-AUC are not computable. ‡TPR at COMMEC's native Flag threshold (FPR = 0.3%, stricter than the 1% target), a conservative lower bound on TPR@1%.

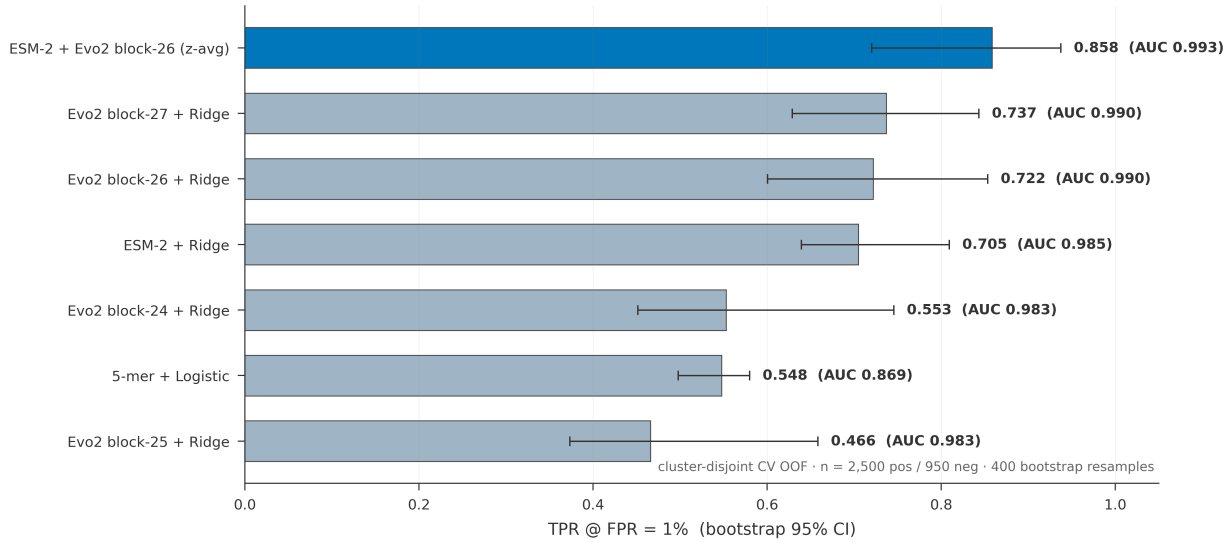

Figure 3. TPR@1% per screen; the ESM-2 + Evo2 block-26 ensemble (top bar) lifts catch rate to 0.858, beating either single-model screen by 12–15 percentage points without retraining.

Interpretation. Per-layer probes (Appendix A.2) place the DNA-side functional signal in the upper-middle layers (blocks 26–27 of the 32-layer stack), consistent with the idea that deeper layers encode what the protein does rather than its raw letter frequencies.

4.2 Ensemble combination

The simplest possible combination beats either model alone and outperforms COMMEC by a wide margin when paired with letter-matching. We standardize the ESM-2 and Evo2 block-26 decision scores independently: subtract each model's test-set mean, divide by its test-set standard deviation, then average the two normalized streams. On the test set, the ensemble reaches **TPR@1% = 0.858**. For comparison: ESM-2 alone scores 0.705, Evo2 block 26 alone 0.722, and the COMMEC policy baseline only 0.168 (Table 1). AUROC rises to 0.993 and TPR@0.1% to **0.446**, up from 0.175 and 0.366, respectively. The largest gain occurs at the strict 0.1% operating point, where missed toxins matter most. The two screens often disagree on hard cases. Averaging their scores recovers roughly an additional eighth of the toxins, without extra labeled data, additional training, or changes to the underlying models. The ensemble requires both ESM-2 and Evo2 7B to score new sequences, doubling the compute compared with single-screen baselines. Still, this may be worth it for the highest-stakes synthesis-order review tier, while single-model screens remain reasonable for bulk pre-filtering.

Holdout-threshold transfer caveat. The Table 1 number (0.858) is the test-set TPR when the threshold is set so that 1% of test benigns trigger an alarm; transferring instead the threshold calibrated on the 168

held-out benigns yields a test-set FPR of **4.5%** at which the ensemble catches **96.6%** of toxins, because 1% of 168 is fewer than two samples and the 99th-percentile cutoff lands more permissive than the true 1% point. A larger benign holdout would close this transfer gap directly.

5. RQ2 — Evasion Resilience

The strongest evidence for functional evasion resilience comes from toxins that actually escape BLAST's protein-level lookup. At the 45% and 60% mutation levels, only 6.5% and 23.9% of toxins evade BLAST's protein search respectively, but among the 304 toxins that do beat BLAST at these levels, the ESM-2 screen still catches **84.5%**, well above a coin flip (full BLAST hit-rate breakdown in Appendix A.4). Structure-guided evasion typically requires fewer mutations; 60% random swaps represents a conservative bound.

The full recovery ladder across all disguised toxins — not just the BLAST-evading subset — confirms the pattern:

Table 2. Recovery under disguise. Recovery is TPR@1% with the alarm threshold set on the held-out 15% slice and close homologs kept apart. Group A: silent DNA changes (same protein). Group B columns: fraction of amino acids randomly swapped.

Screen	Original	Group A: silent DNA	15% swapped	30%	45%	60%
ESM-2 (protein)	1.000	1.000	0.993	0.984	0.979	**0.983**
Evo2 7B (DNA, block 26)	0.974	**0.956**	–	–	–	–
5-letter-pattern baseline	0.653	0.659	0.593	0.538	0.475	0.407

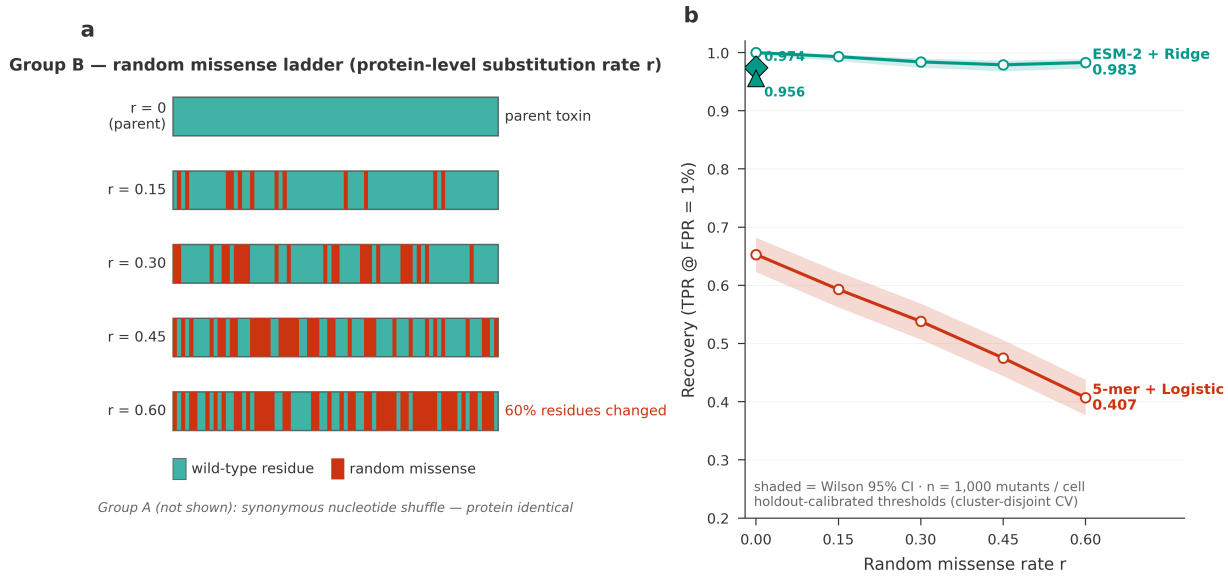

Figure 4. Recovery (TPR@1%) vs. fraction of amino acids randomly swapped, with the threshold set on

the held-out slice and close homologs kept apart between train and test. The protein screen retains 98.3% recovery at 60% mutation; the 5-letter baseline drops to 40.7%.

Interpretation. The two screens are reading different things: the protein screen tracks what the protein does and survives small changes in its sequence, while the letter-frequency baseline tracks DNA spelling and breaks down when those letters are disturbed. Evo2's modest 1.8-point drop on Group A — despite an average letter-by-letter match of 76.2% between original and codon-scrambled inputs — indicates the DNA model reads past raw spelling to functional signal. Group A (silent DNA changes) genuinely evades the DNA-level BLAST lookup: 93.3% of disguised toxins are no longer recognized by BLAST's DNA search, even though they spell the same protein.

6. RQ3 — Short Reads

Reliable screening requires DNA fragments ≥**1,500 bp**; full-length test results are in Appendix A.3.

7. Discussion and Limitations

What works. Both ESM-2 (protein) and Evo2 7B (DNA) provide strong functional signals with close homologs kept apart between train and test, and their combined score pushes TPR@1% from ~0.71 to **0.858** with no retraining. The evasion test confirms that the ESM-2 screen still recovers **98.3%** of toxins after 60% of amino acids are randomly swapped (versus 40.7% for the 5-letter baseline), and the Evo2 DNA screen retains **95.6%** recovery on codon-scrambled toxins despite an average letter-by-letter match of 76.2%. The length test sets a working window of **1,500 bp** with the alarm threshold set on the held-out slice. Crucially, the embedding screens are *complementary* to the COMMEC policy baseline rather than competing with it: COMMEC's curated biorisk database catches known regulated toxin families with high precision (16.8% catch rate at 0.3% FPR on our set), while the embedding ensemble generalizes to toxin families outside that curated list, suggesting a layered architecture in which alignment-based triage runs first and the embedding ensemble flags function-bearing sequences the policy tool misses.

Practitioner checklist.

1. Pair a letter-matching filter with a protein-level screen whenever an attacker could mutate the toxin while keeping its 3D shape.
2. Re-tune the alarm threshold for each deployment setting; do not reuse a paper's false-alarm rate on a 150-bp short-read pipeline.
3. For regulator-facing claims, re-run the evasion test (Section 3.6) with a structure-guided mutation protocol and an independent held-out set.
4. For highest-stakes review tiers, deploy the standardized ESM-2 + Evo2 combination; recovers ~15 percentage points more toxins at 1% FPR than single models.

Limitations.

- Class imbalance (2.6 toxins per benign) inflates PR-AUC; we rely on AUROC and FPR-thresholded metrics instead.
- Random amino-acid swaps stress the screen but are not structure-guided escape, so they do not target alignment blind spots specifically.
- Length-test fragments are drawn from the same pool as the training set; real wastewater or synthesis-order replication is required before any public-health claim.
- Setting the alarm threshold on the held-out slice rather than benign scores from held-out validation folds is conservative and likely contributes to the 1,500 bp working window.
- The ensemble threshold is calibrated on only 168 holdout benigns, so transferring it to the test set lands

at 4.5% FPR rather than the 1% target; a larger benign holdout would close this gap.

Future work. Run a structure-guided amino-acid mutation protocol and test on genuine environmental isolates and real DNA synthesis orders held out from training. Tune the ensemble weights and re-evaluate on a large benign holdout set to set a tight 1% threshold. Use the combined protein-plus-DNA score as a second stage after a Common-Mechanism-style first pass.

8. Conclusion

With close homologs kept apart between train and test and the alarm threshold set on the held-out 15% slice, an ensemble averaging Evo2 7B (block 26) and ESM-2 reaches an AUROC of **0.993** and **TPR@1% of 0.858** on a human-codon-matched benchmark of toxin and benign protein-coding sequences, substantially above either single model (0.722 and 0.705) and roughly five times the COMMEC biorisk-only policy baseline (16.8%). Under random amino-acid swaps, the protein screen still recovers **98.3%** of toxins at 60% of amino acids swapped, while the 5-letter screen drops to **40.7%**, a concrete gap between letter-frequency screens and protein-meaning screens. The length test puts the working window for the 1% target at **1,500 base pairs**. Embedding-based screening *complements* alignment-based policy screens, such as the Common Mechanism, but does not replace them.

Code and Data

- **Code repository:** Six end-to-end reproducibility notebooks: dataset assembly, codon-realization robustness, embedding extraction (ESM-2, Evo2 7B, HyenaDNA), classifier comparison, evasion evaluation, and length calibration.
- **Data/Datasets:** Frozen per-stage result tables and label tables are included with the submission artifacts.
- **Other artifacts:** All figures used in this paper are included as 300 dpi PNGs and vector PDFs, together with figure-rebuild scripts.

References

1. Wittmann, B. J. *et al.*, "Toward AI-Resilient Screening of Nucleic Acid Synthesis Orders: Process, Results, and Recommendations," *bioRxiv*, Dec 2024. DOI: 10.1101/2024.12.02.626439.
2. Wheeler, N. E. *et al.*, "Developing a Common Global Baseline for Nucleic Acid Synthesis Screening," *Applied Biosafety* **29**(2):71–78, 2024. DOI: 10.1089/apb.2023.0034.
3. Z. Lin *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science* **379**, 1123–1130 (2023) [ESM-2].
4. Evo2: Arc Institute / Stanford / UC Berkeley. Model card: `arcinstitute/evo2_7b`, 2025. https://huggingface.co/arcinstitute/evo2_7b.
5. Tang, Z. *et al.*, "Evaluating the representational power of pre-trained DNA language models for regulatory genomics," *Genome Biology* **26**:203, 2025. DOI: 10.1186/s13059-025-03674-8.
6. Nguyen, E. *et al.*, "HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution," NeurIPS (2023). https://proceedings.neurips.cc/paper_files/paper/2023/hash/86ab6927ee4ae9bde4247793c46797c7-Abstract-Conference.html.
7. S. F. Altschul *et al.*, "Basic local alignment search tool," *J. Mol. Biol.* **215**, 403–410 (1990); BLAST+ 2.17+ (NCBI).
8. Apart Research, AIxBio Hackathon internal notes (2026).
9. Tayouri, S. *et al.*, "Defending Synthetic DNA Orders Against Splitting-Based Obfuscation," *bioRxiv*, Mar 2025. DOI: 10.1101/2025.03.12.642526.

10. Chen, L. *et al.*, "VFDB: a reference database for bacterial virulence factors," *Nucleic Acids Research* **33**(Database issue):D325–D328, 2005. PMID: 15608208.
11. Steinegger, M. & Söding, J., "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," *Nature Biotechnology* **35**:1026–1028, 2017. DOI: 10.1038/nbt.3988.
12. McInnes, L. *et al.*, "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software* **3**(29):861, 2018. DOI: 10.21105/joss.00861.
13. Gene Ontology Consortium, "GO:0090729 — toxin activity." https://amigo.geneontology.org/amigo/term/GO:0090729.
14. IBBIS, `commec` v1.0.3 — Common Mechanism for DNA Synthesis Screening (2024–2026). https://github.com/ibbis-bio/common-mechanism.

Appendix

A.1 Methods detail

A.1.1 The 5-letter-pattern baseline. The reference baseline counts how often each of the 1,024 possible 5-letter DNA patterns appears in a sequence and feeds those counts to an L2-regularized logistic regression. We reran the human re-spelling step several times with different random choices to confirm that the baseline's ~0.87 AUROC reflects a real signal, not a quirk of a particular re-spelling.

A.1.2 Embeddings and quick separability checks. For each sequence, we extracted a numeric summary ("embedding") from each model: ESM-2 produces a single average vector per protein, while Evo2 7B produces one vector per internal layer (we kept layers 20–28 of its 32-layer stack). Before training a full classifier, simple linear probes on these vectors and UMAP plots (Figure A3) showed the protein-side embeddings cleanly separate toxins and benigns, while the DNA-side embeddings start mixed and separate only in deeper layers.

A.1.3 Comparing the screens. We compared three screens head-to-head: the 5-letter-pattern baseline; a linear classifier on top of ESM-2's protein embeddings; and a linear classifier on top of Evo2 7B's DNA embeddings, trained separately for each of layers 20–28. To prevent screens from getting credit for memorizing close homologs, we grouped homologous sequences using MMseqs2 at 30% protein identity and kept entire groups together in either training or testing — never split across both.

A.1.4 Length test. To see how short an input can be before the screens stop working, we cut benign sequences into DNA windows of 150, 200, 300, 500, 750, 1000, 1500, 2000, and 3000 base pairs. The protein-based screens (ESM-2 and the 5-letter baseline) first translate the longest open reading frame in the window; the Evo2 screen reads the DNA directly. We compare the false-alarm rate on real fragments against shuffled "garbage" controls (DNA scrambled to preserve only its short-range composition).

A.2 Per-layer Evo2 probes

Table A1. Per-layer probes of Evo2 (blocks 20–28). Ranking score climbs from 0.960 at the lowest probed layer to a plateau near 0.990 at layers 7–8, then dips at layer 9, consistent with deeper layers encoding what the protein does rather than its raw letter frequencies.

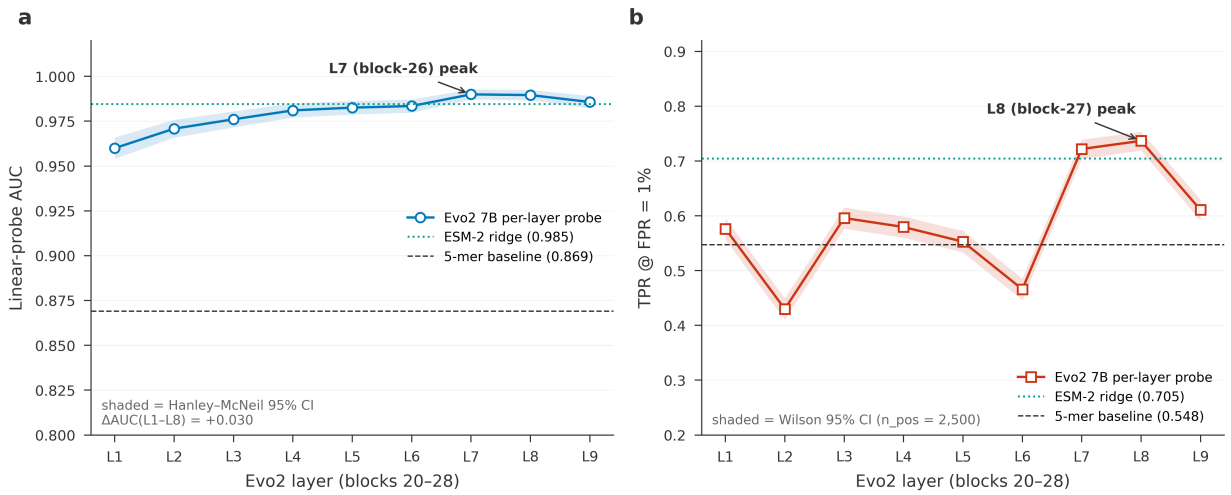

Figure A1. Per-layer linear probes on Evo2 7B (one average vector per layer), with close homologs kept apart between train and test. ***(a)*** *AUROC peaks at layer 7 / block 26 (0.990);* ***(b)*** *TPR@1% peaks at layer 8 / block 27 (0.737), with layer 7 1.5 pp behind. Grey dashed line: 5-letter baseline. Green dotted line: ESM-2.*

A.3 Length test (full results)

At 150–300 bp windows, the measured false-alarm rate at the 1% target rises to 18–49%; the curve only stays close to the 1% target from about 1,500 bp onward, with the alarm threshold set on the held-out 15% slice (Section 3.1).

Table A2. Measured false-alarm rate on benign DNA fragments when the screen is tuned to a 1% target. "Real" = actual benign fragments; "scrambled" = control DNA shuffled to preserve only short-range letter pairings.

Window (bp)	ESM-2 (real)	Evo2 layer 7 (real)	5-letter (real)	ESM-2 (scrambled)	Evo2 layer 7 (scrambled)
150	0.489	0.660	0.121	0.741	0.965
300	0.181	0.220	0.080	0.840	0.903
500	0.063	0.105	0.044	0.789	0.844
750	0.039	0.045	0.030	0.804	0.859
1000	0.028	0.029	0.024	0.783	0.851
1500	**0.010**	**0.010**	0.011	0.744	0.869
2000	0.002	0.005	0.012	0.763	0.891
3000	0.000	0.000	0.009	0.642	0.899

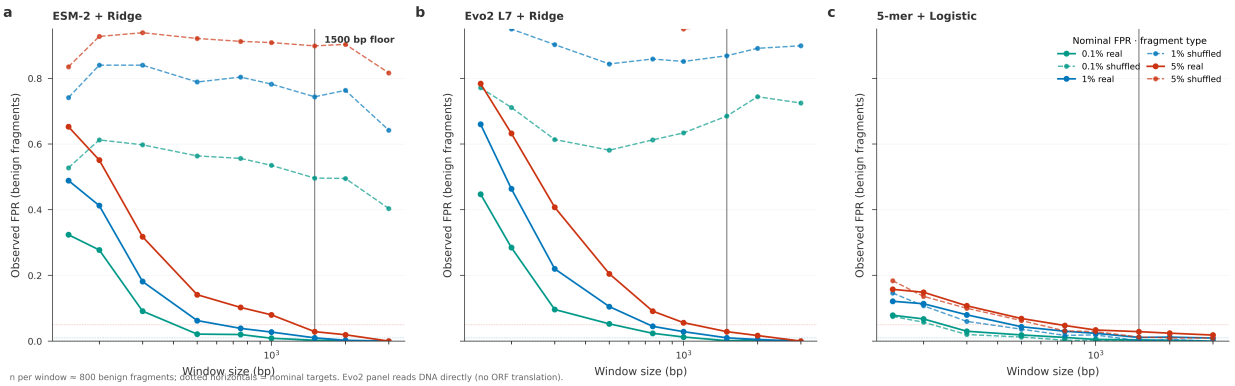

Figure A2. Measured false-alarm rate vs. DNA window size for ESM-2 (a), Evo2 layer 7 (b), and the 5-letter baseline (c) at three target rates (dotted lines: 0.1%, 1%, 5%); solid = real benign fragments, dashed = scrambled controls. The vertical line marks 1,500 bp, where the measured 1% rate first sits within 2× of the target across all three screens.

The 5-letter baseline keeps a tighter false-alarm rate at short windows (0.121 at 150 bp vs. 0.489 for ESM-2 and 0.660 for Evo2), but this is not a sign that it screens better: its TPR@1% is 15–19 points below the embedding screens at full length (0.548 vs. 0.705–0.737), so it catches fewer real toxins regardless of window size. Short-read contexts such as metagenomics and wastewater monitoring require either retuned alarm thresholds or a less strict target false-alarm rate, paired with follow-up confirmation. For DNA synthesis orders of 1,500 bp or longer, all three screens converge and the embedding screens hold near the 1% target with high reliability.

A.4 BLAST evasion check

The headline BLAST-evasion result — ESM-2 catches 84.5% of the toxins that beat BLAST's protein search — is presented in Section 5. Here we report the detailed BLAST hit rates. **Group A (silent DNA changes):** 93.3% (933 of 1,000) of disguised toxins are no longer recognized by BLAST's DNA search, even though they spell the same protein. Swapping in different DNA letters across the whole sequence breaks the short stretches of identical letters that BLAST relies on. **Group B (random protein mutations):** only 6.5% of toxins escape BLAST's protein search after 45% of amino acids are swapped, and 23.9% escape after 60% are swapped.

A.5 Visualizing the embeddings

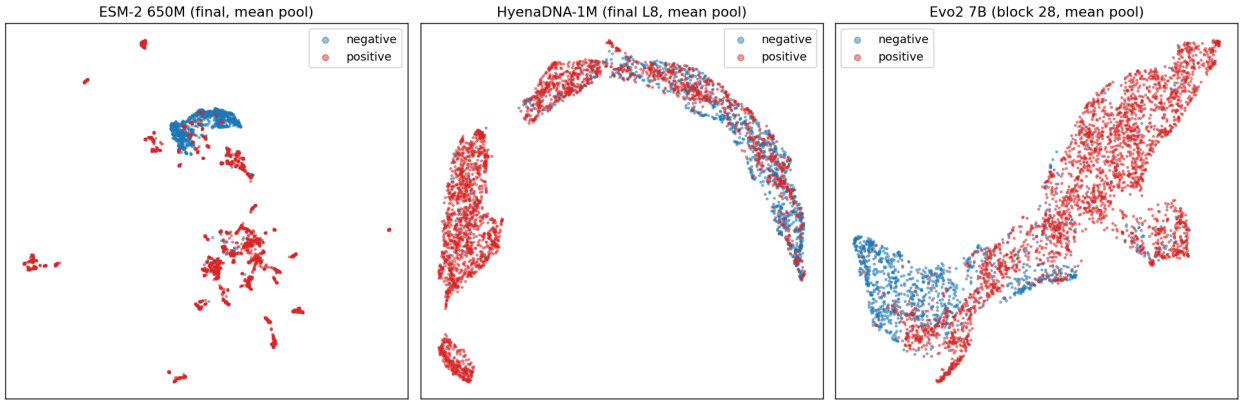

*Figure A3. UMAP projections of each model's embeddings (red = toxin, blue = benign). **(Left)** ESM-2*

(protein): clean split. ***(Middle)*** *HyenaDNA-1M (final layer): heavy overlap.* ***(Right)*** *Evo2 7B (final layer): two arc-shaped clusters with cleaner separation than HyenaDNA, though not fully split at the final layer; the deployed Evo2 screen uses block 26, not the final layer.*

A.6 Results from the smaller DNA model (HyenaDNA-1M)

We also ran a much smaller DNA model, **HyenaDNA-1M** (8 layers, ~1 million parameters), as a same-family comparison point for Evo2 7B. **Important caveat:** these numbers were collected before we switched to the homology-aware split (Section 3.2), so homologous sequences could leak between training and test folds, inflating the absolute values; they are not directly comparable to Table 1, but the gap between HyenaDNA and Evo2 is the point.

Table A3. Leaderboard from the older homology-leaking split, with HyenaDNA variants alongside ESM-2.

Method	AUROC	PR-AUC	TPR@0.1%	TPR@1%	TPR@5%
ESM-2 + Linear	**0.996**	**0.998**	**0.540**	**0.937**	**0.995**
One-hot CNN (Tang)	0.952	0.988	0.320	0.657	0.843
HyenaDNA dual-linear on covariance pool	0.963	0.991	0.248	0.622	0.863
HyenaDNA mean-pool (L8) + Linear	0.948	0.985	0.026	0.557	0.790
5-mer + Logistic	0.872	0.969	0.197	0.554	0.674
MLP on HyenaDNA covariance pool	0.917	0.977	0.036	0.406	0.736

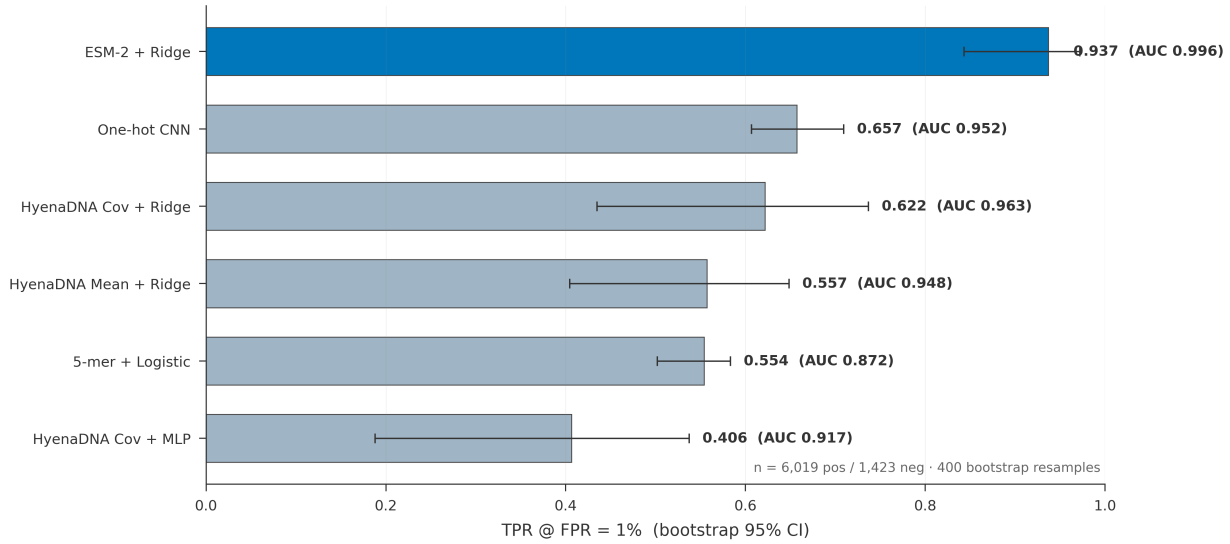

Figure A4. Older leaderboard with HyenaDNA-1M variants alongside ESM-2 and the 5-letter baseline; ordering matches Figure 3, but absolute numbers are inflated by homology leakage.

Table A4. Per-layer probes of HyenaDNA (older homology-leaking split). Catch rates barely change across HyenaDNA's 8 layers, consistent with a shallow architecture in which every layer already mixes context from across the whole sequence.

Layer	AUROC	PR-AUC	TPR@0.1%	TPR@1%	TPR@5%
1	0.953	0.988	0.112	0.588	0.805
2	**0.956**	**0.988**	0.077	**0.631**	**0.823**
3	0.951	0.986	0.061	0.544	0.801
4	0.945	0.984	0.049	0.476	0.779
5	0.944	0.984	0.030	0.506	0.781
6	0.948	0.985	0.020	0.538	0.786
7	0.948	0.985	0.027	0.554	0.786
8	0.948	0.985	0.026	0.557	0.790

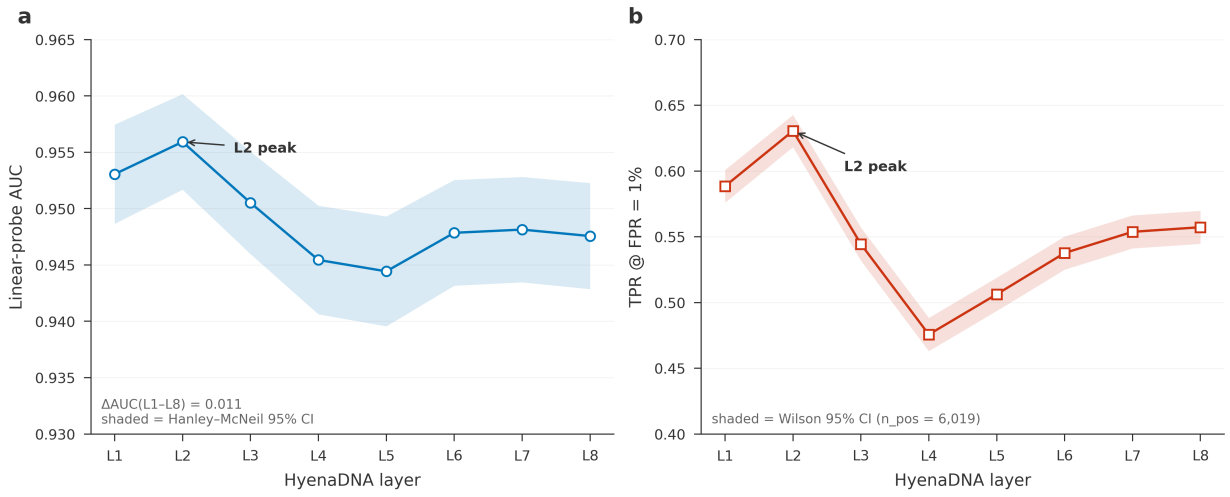

Figure A5. Layer-by-layer probes of HyenaDNA-1M on the older split: AUROC is nearly flat (0.945–0.956) and TPR@1% peaks at layer 2, very different from the steady climb across Evo2's layers in Figure A1.

Takeaway. Within the same model family, scale does the work: the 8-layer HyenaDNA-1M sits about 38 percentage points below ESM-2 on TPR@1%, while the 32-layer Evo2 7B (Table 1) catches up to and slightly beats ESM-2. This is what convinced us to move the DNA screen to Evo2 7B.

A.7 Limitations and Dual-Use Considerations

This work is intended to strengthen biosecurity screening, but the same tools and findings carry dual-use risks that warrant explicit discussion.

Adversarial optimization. A continuous scoring model, unlike a binary flag system, returns graded confidence values. An adversary with query access could iteratively mutate a toxin sequence, observe how the score changes, and hill-climb toward variants that evade detection while retaining function. Deploying the ensemble as a black-box API with rate limits and audit logging, rather than releasing model weights or per-query scores, mitigates this risk but does not eliminate it. Future deployments should evaluate adversarial robustness under query-budget constraints.

Dataset and classifier release. The toxin sequences used here are drawn from public databases (Gene Ontology, VFDB) and pose no incremental information hazard. However, releasing a trained classifier together with its decision boundary could lower the barrier to testing evasion strategies. We release code and evaluation artifacts but recommend that trained model weights be shared only with vetted screening providers under access controls.

Embedding models as dual-use tools. ESM-2 and Evo2 are general-purpose biological sequence models developed for broad scientific use. Our work shows that their internal representations encode functional toxin signatures, a property useful for defense but also informative to an adversary seeking to understand which sequence features trigger detection. This dual-use tension is inherent to any function-aware screening approach and is not unique to embedding models; alignment-based screens face analogous risks when their reference databases are public. The net benefit of transparent, reproducible screening research outweighs the marginal risk, provided deployment follows responsible-disclosure norms.

Scope of evaluation. Our evasion test uses random mutations, not structure-guided or fitness-aware protocols, and our dataset covers annotated toxins rather than novel or engineered threats. These limitations, discussed in Section 7, mean the screening performance reported here is an upper bound on real-world resilience. Any operational deployment should be validated against a broader and more adversarial threat model before public-health claims are made.

LLM Usage Statement

We used Claude Code for drafting and editing prose, brainstorming figure layouts, and refactoring notebook scaffolding. All numeric results, tables, and figures were generated from the code and data in this repository; every claim in the paper was checked against the artifacts before inclusion.