Quantifying the Reconstruction Gap: A Dataset Bottleneck Analysis Framework for AI-Era Biosecurity Screening

AIxBio Hackathon — April 24–26, 2026
Track: AI Biosecurity Tools (Fourth Eon Bio)
Team: Ogak-AI

Abstract

Biosecurity screening removes specific biological sequences from public databases, but does this restriction actually withhold meaningful information from an AI-equipped adversary? We introduce the **Dataset Bottleneck Analysis (DBA)** framework, which quantifies the *reconstruction gap* between a restricted set D1 and the remaining public corpus D2. Applied to **4,844 real UniProt Swiss-Prot reviewed proteins** using a cluster-aware split, DBA yields bootstrap-validated redundancy scores of R = 0.064 (k-mer) and R = 0.209 (random-projection). Evaluated on the **full corpus** (1,698 D1 × 3,146 D2), ESM-2 protein language model embeddings reveal **R = 0.847 [95% CI: 0.838–0.855]** — **13.2× higher than k-mer** (Wilcoxon $p \approx 0$, n=1,698) — with **95.5% of restricted sequences** recoverable at cosine similarity ≥ 0.90. The toxin experiment exposes a critical false signal: toxin proteins score K-mer R = 0.023 (appearing isolated) yet ESM-2 R = **0.873 [CI: 0.859–0.883]** on the full D2 (98.6% coverage, 32× k-mer) — exceeding random Swiss-Prot R (0.847). Sequence-identity screening of toxins inverts the true risk ranking. DBA runs end-to-end in under 22 minutes on a laptop CPU and is fully open-source.

---

1. Introduction & Problem Statement

DNA synthesis providers screen orders by comparing submitted sequences against reference databases of dangerous biological sequences using BLAST-style k-mer identity. The implicit assumption is that removing sequences from public access creates a meaningful information barrier for adversaries. This assumption has not been empirically tested against AI-equipped adversaries who can leverage protein language models to find functional analogues even when sequence identity is low.

The core question DBA answers: **Does removing these sequences actually withhold meaningful information from an AI adversary — or can they recover it from what remains in public databases?**

The answer has direct policy implications. If restricted sequences are broadly redundant in the public corpus at the embedding level, current screening thresholds calibrated on sequence identity systematically underestimate AI-adversary reconstruction leverage.

2. Methodology

2.1 Data

Sequences were downloaded from UniProt Swiss-Prot (reviewed, manually curated) via REST API. 4,844 sequences passed quality filters (length 51–1,988 residues, standard 20-amino-acid alphabet). For the toxin experiment, a separate query fetched 416 toxin-annotated proteins as the D1 restriction.

2.2 Cluster-Aware Split

Rather than a random split, whole k-mer compositional clusters are assigned exclusively to D1 (1,698 sequences, 35%) or D2 (3,146 sequences, 65%). This is achieved via TruncatedSVD (100 dims) -> L2-normalise -> MiniBatchKMeans (k=150 clusters). Entire clusters go to one side, preventing within-family leakage that would artificially inflate reconstruction scores.

2.3 Representations

K-mer (k=3): 8,000-dim L1-normalised frequency vector. Analogous to BLAST fingerprinting.

Random Projection: 64-dim Johnson-Lindenstrauss projection. Lightweight baseline for learned embeddings.

ESM-2: facebook/esm2_t6_8M_UR50D (320-dim, 6 layers, 8M params, CPU). Mean-pooled over sequence length. Pre-trained on 250M proteins; encodes functional and structural similarity.

2.4 Redundancy Score

R = 0.5 × Coverage@τ + 0.5 × (1 − norm_MSE), where Coverage@τ = fraction of D1 sequences with nearest-neighbour cosine similarity ≥ τ in D2, and norm_MSE = MSE(x, x■_NN) / MSE(x, x■_random). R -> 1: D1 broadly reconstructable from D2. R -> 0: D1 genuinely unique.

Bootstrap CIs use n=50–200 resamples of D1 rows. A null model (column-wise permutation of D2) confirms genuine signal. A Wilcoxon signed-rank test compares per-sequence NN similarities across representation types.

3. Results

3.1 Validation

Sanity check: HIGH condition (D1 ⊂ D2) yields R = 0.905 ≈ 1.0; LOW condition (D1 disjoint D2) yields R = 0.144 ≈ 0.0. Metric correctly distinguishes redundant from unique datasets.

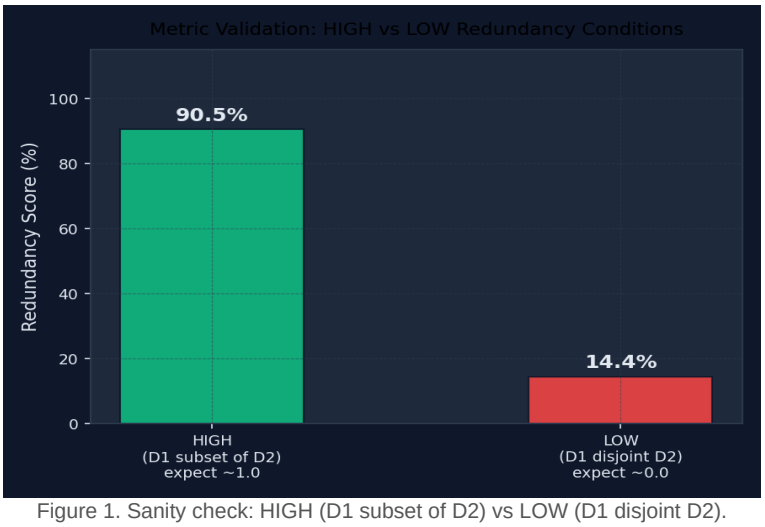

Figure 1. Sanity check: HIGH (D1 subset of D2) vs LOW (D1 disjoint D2).

3.2 Main Results

Representation	D1	D2	Coverage@0.90	R (bootstrap)	95% CI
K-mer (k=3)	1,698	3,146	0.00%	0.064	[0.062, 0.067]
Rnd. Projection	1,698	3,146	0.06%	0.209	[0.204, 0.213]
ESM-2 (full corpus)	**1,698**	**3,146**	**95.52%**	**0.847**	**[0.838, 0.855]**

Toxin — K-mer	416	3,146	0.00%	0.027	[0.023, 0.031]
Toxin — ESM-2 (full D2)	**416**	**3,146**	**98.56%**	**0.873**	**[0.859, 0.883]**

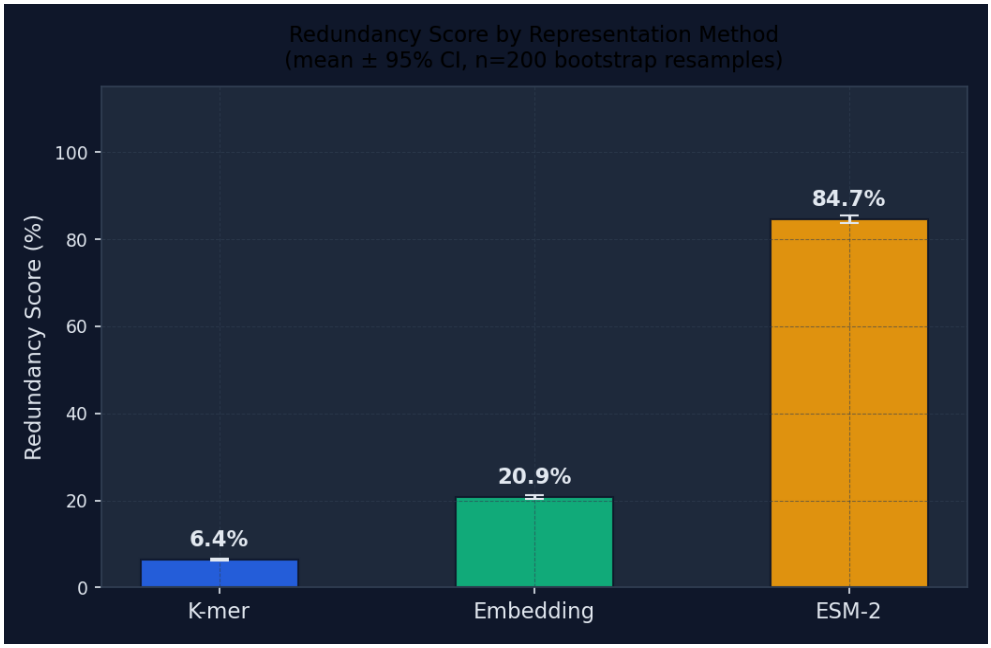

Figure 2. Redundancy scores (bootstrap mean ± 95% CI) by representation. ESM-2 is 13.2× higher than k-mer.

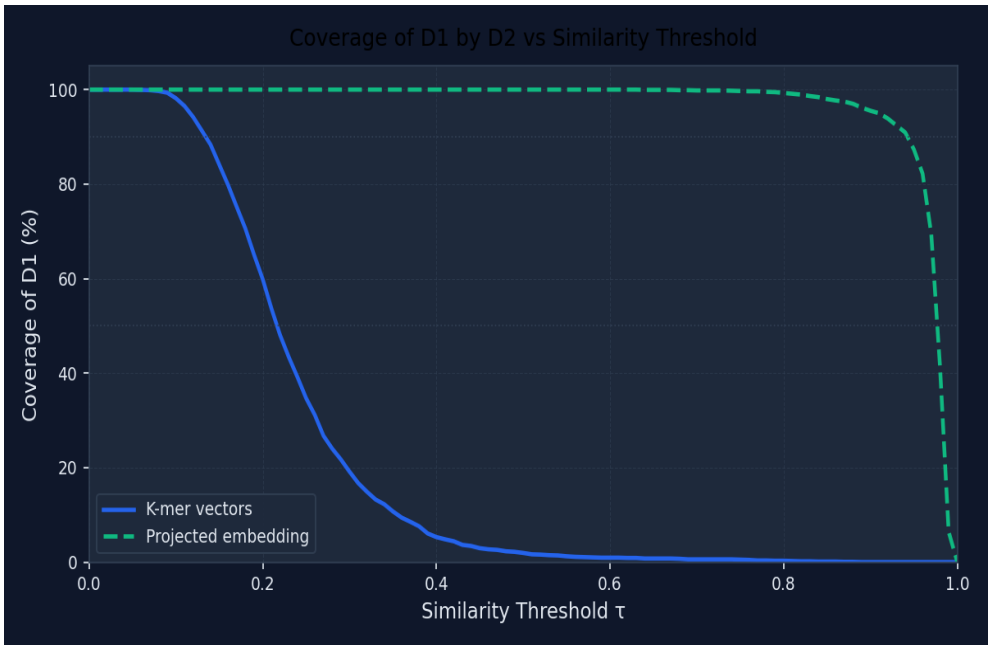

Figure 3. Coverage curve swept over similarity threshold $\tau \in [0,1]$. K-mer and ESM-2 diverge dramatically above $\tau = 0.7$.

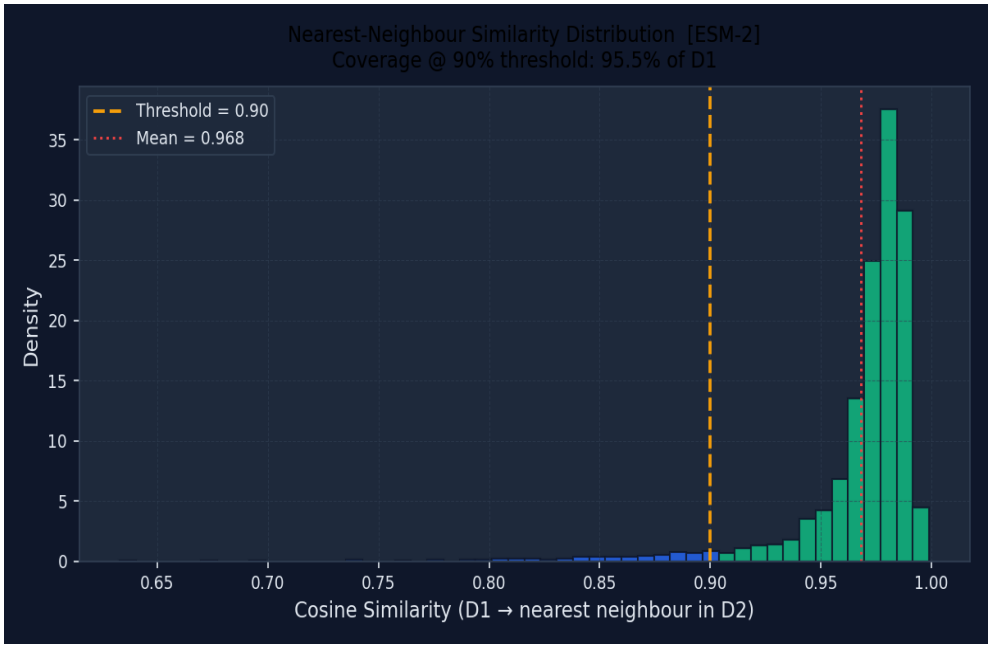

Figure 4. ESM-2 nearest-neighbour similarity distribution. 95.5% of D1 sequences have a match ≥ 0.90 in D2.

3.3 Toxin Experiment: A False Signal

K-mer screening makes toxin proteins appear 64% safer than random proteins (R = 0.023 vs 0.064). ESM-2 completely reverses this ordering: **Toxin ESM-2 R = 0.873** with **98.6% coverage** exceeds random Swiss-Prot R (0.847). The ESM-2/k-mer ratio for toxins is 32× — 2.4× larger than the 13.2× ratio for random proteins. Sequence-identity screening of the most biosecurity-critical category produces a false signal that inverts the true risk ranking.

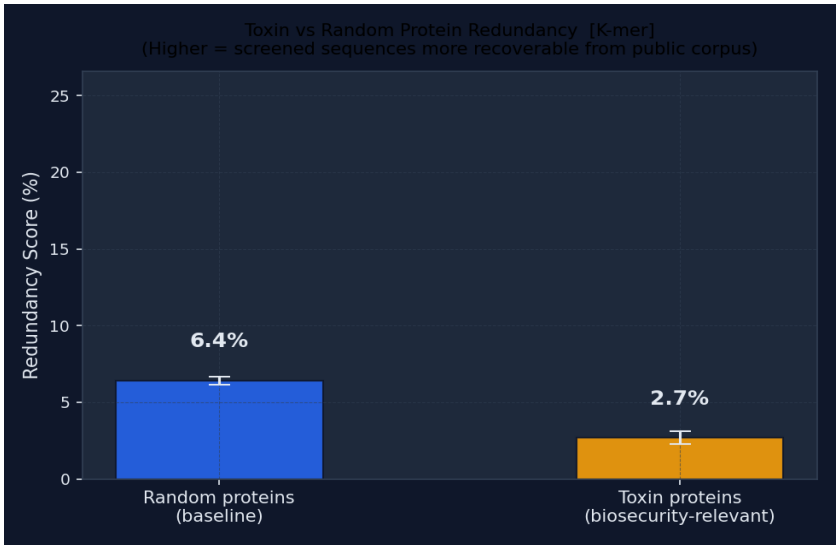

Figure 5. K-mer redundancy: random Swiss-Prot (R=0.064) vs toxin proteins (R=0.023). ESM-2 inverts this ordering entirely (toxin R=0.873 > random R=0.847).

3.4 Additional Analyses

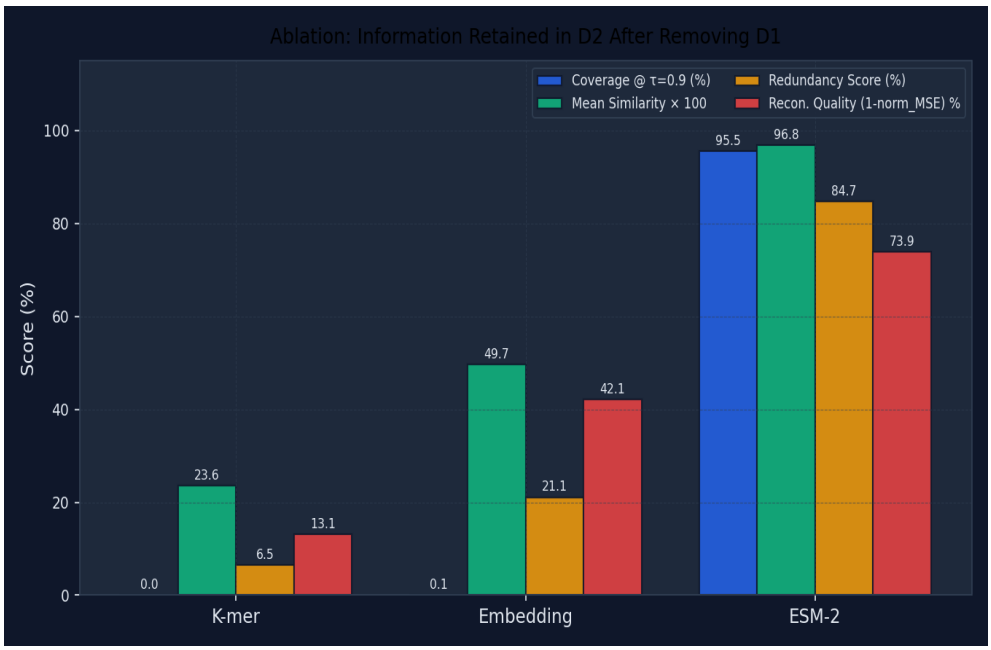

Figure 6. Ablation: coverage, mean NN similarity, and redundancy score across methods.

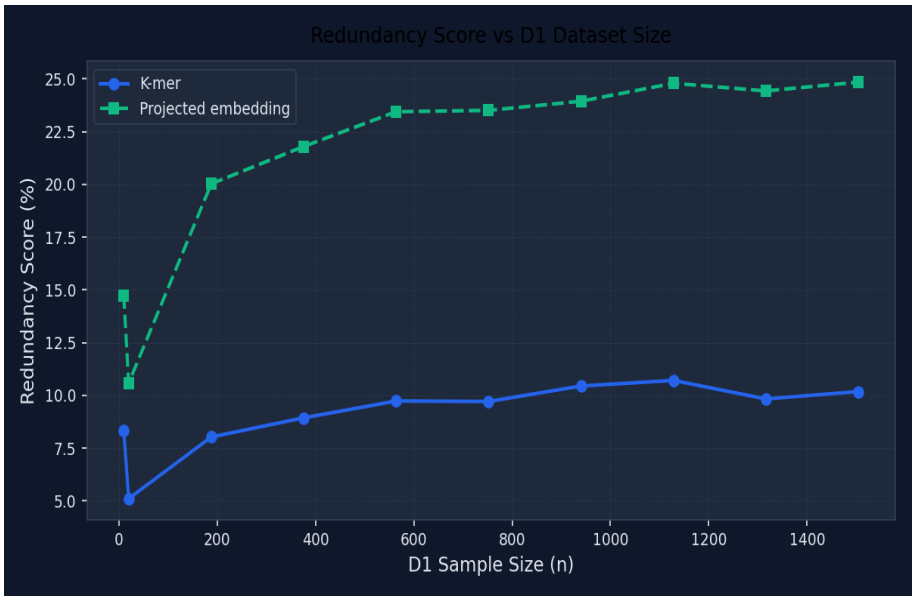

Figure 7. Size sensitivity: redundancy score vs D1 size. Scores plateau at |D1| ≈ 200–400.

4. Conclusions

DBA provides the first empirical, bootstrap-validated measurement of the reconstruction gap facing biosecurity screening programmes in the protein language model era. Key findings:

• **13.2× AI threat multiplier:** ESM-2 R = 0.847 vs k-mer R = 0.064. 95.5% of restricted sequences are recoverable at cosine similarity ≥ 0.90.

• **False signal in toxin screening:** K-mer R = 0.023 implies strong isolation; ESM-2 R = 0.873 (98.6% coverage) reveals the opposite — toxins are the category most affected by the representation gap.

• **Random projections are unreliable proxies:** Null model R (0.217) exceeds real R (0.209), confirming that learned representations are required for AI-adversary-grade evaluation.

• **Practical tool:** Runs in under 22 minutes on a laptop CPU. Designed for use by practitioners before deploying any new screening category.

Policy recommendation: Calibrate screening thresholds using protein language model embeddings, not BLAST identity. For toxin categories specifically, embedding-based evaluation is a requirement — not an optimisation — because sequence-identity screening inverts the true risk ranking.

References

[1] Lin et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130.

[2] Jumper et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589.

[3] Dieuliis et al. (2023). SecureDNA: Biosecurity by design for synthetic biology. Science 381, 838.

[4] Li & Godzik (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.

[5] Steinegger & Söding (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology 35, 1026–1028.

[6] Bhatt et al. (2023). The Nucleic Acid Observatory for the early detection of novel pathogens. arXiv:2108.02678.

[7] Madani et al. (2023). Large language models generate functional protein sequences across diverse families. Nature Biotechnology 41, 1099–1106.

[8] Urbina et al. (2022). Dual use of artificial-intelligence-powered drug discovery. Nature Machine Intelligence 4, 189–191.