BioChain: Cross-Vendor Threat Detection via Function-Aware DNA Fragment Screening

Arka Dash[1] Asutosh Rath[2] Caio Timm[3]

Igor Pereverzev[4] Yatharth Maheshwari[5]

[1-5]Independent

With
Apart Research

Abstract

Contemporary DNA synthesis screening systems - including SecureDNA and the IBBIS Common Mechanism - operate on single sequences submitted to individual vendors. This architecture is structurally blind to distributed synthesis attacks, in which a threat sequence is fragmented across multiple vendors such that no individual fragment triggers an alert. We present BioChain, a two-layer architecture for detecting distributed DNA synthesis attacks by linking cross-vendor fragment orders and scoring reassembled sets for biological threat potential. The first layer constructs a cryptographic audit trail across vendors using Order Commitment Records and locality-sensitive hashing, enabling cross-vendor fragment linking without disclosing raw sequences. The second layer scores candidate assembly sets using a permutation-invariant Set Transformer operating over ESM-3 (esm3_sm_open_v1) fragment embeddings - a multi-track protein language model that encodes sequence, structure, and functional priors jointly. On 5-fold cross-validation holding out entire toxin families, BioChain achieves AUC $= 0.907 \pm 0.032$ and AP $= 0.835 \pm 0.088$. Fragment-count robustness sweeps show AUC ≥ 0.90 across k $= 2$ to 15 fragments per assembly. We report two structural failure modes: poor probability calibration (ECE $= 0.1296$, where values ≥ 0.10 are generally considered poorly calibrated for production classifiers) and inability to distinguish neutralised toxin mutants from functional wild-types, and identify contrastive training on hard-negative pairs as the primary remediation. We position this work as an existence proof for the tractability of set-level distributed-attack screening, not a deployment-ready system; calibration and adversarial robustness require further work before operational use.

1. Introduction

DNA synthesis screening occupies a critical chokepoint in the biosecurity stack. As benchtop synthesizers approach the capability to produce virus-length sequences and open-weight biological foundation models lower the expertise barrier for biological design, the ability to detect dangerous synthesis orders before physical material is produced becomes an increasingly important layer of defence (Wittmann et al., 2025; NTI, 2023). Current commercial screening - using platforms such as SecureDNA and the IBBIS Common Mechanism - has demonstrated real-world value for catching naive orders of known select agents and toxins.

A documented gap, however, sits at the intersection of synthesis logistics and adversarial strategy: the *distributed synthesis attack*. A sufficiently motivated actor can fragment a target toxin or pathogen coding sequence into 10-20 innocuous-looking sub-sequences, order each fragment from a different vendor on different days, and assemble them downstream via Gibson assembly or ligation. Because no individual vendor sees more than one fragment, and because each fragment is too short and too dissimilar to any threat sequence in isolation, current per-sequence screening misses the threat entirely. This is not a hypothetical: Edison et al. demonstrated that MIT researchers were able to acquire unregulated DNA fragments sufficient for 1918 influenza reconstruction from dozens of providers, explicitly because no provider had visibility across the full order set (Edison, Toner, and Esvelt, 2026). The Biosecurity Modernization and Innovation Act of 2026 (S.3741) mandates homology-based screening but does not address this fragment-level distributed attack vector.

BioChain addresses this gap through a two-layer architecture. The cryptographic audit layer creates a tamper-evident, privacy-preserving record across vendors, enabling a central auditor to link fragments from the same customer without seeing raw sequences. The ML scoring layer takes the linked fragment set and applies a permutation-invariant classifier built on ESM-3 embeddings to produce a single threat probability for the assembly.

The choice to ground the ML layer in ESM-3 rather than a sequence-only protein language model is architecturally deliberate. Distinguishing a ricin A-chain fragment from a benign 50-residue peptide of similar length is not a sequence problem - at the fragment level, surface sequence similarity to known threats may be minimal. It is a structural and functional problem: does this fragment plausibly belong to a folded toxic domain? ESM-3's multi-track pre-training, which jointly conditions on sequence, structure, and function tokens, provides the inductive bias that sequence-only models cannot (Hayes et al., 2025).

Our main contributions are:

1. **We propose** a practical architecture for cross-vendor distributed synthesis screening, combining locality-sensitive hashing, blind-signature customer linking, and a Merkle-auditable log of Order Commitment Records (OCRs). The cryptographic layer is

presented as a system design contribution; empirical evaluation of Layer 1 (linking accuracy, false-link rate, scalability) is deferred to future work.

2. **We demonstrate** that ESM-3 embeddings combined with a Set Transformer achieve AUC 0.907 ± 0.032 on cross-family generalisation, with fragment-importance attribution that supports human biosafety review workflows.
3. We provide an honest characterisation of two structural failure modes - miscalibration and neutralised-mutant blindness - with a concrete remediation roadmap for each.

2. Related Work

Single-sequence screening systems. SecureDNA uses a cryptographic Distributed Oblivious Pseudorandom Function (DOPRF) to screen sequences as short as 30 bp without exposing the hazard database to customers or the submitted sequence to the screening authority (Sherman et al., 2026). The IBBIS Common Mechanism (commec) employs profile Hidden Markov Models (HMMs) calibrated against curated toxin families, with best performance above 150 bp (IBBIS, 2024). Both systems operate in isolation per submitted sequence. Neither has architectural support for reasoning over sets of related fragments across vendors.

AI evasion of existing screening. Wittmann et al. demonstrated that protein variants designed using generative protein design tools - specifically sequences that preserved biological function while minimising sequence similarity to known threats - evaded both SecureDNA and IBBIS with high reliability (Wittmann et al., 2025). The cross-sector response involved patching homology databases, but this is inherently reactive: any future generative model can again produce novel variants outside current databases. BioChain's function-based approach, embedding via ESM-3's structural priors rather than sequence homology lookup, is motivated directly by this finding.

Fragment-level regulation. Edison, Toner, and Esvelt established empirically that fragment-level acquisition of dangerous sequences from multiple uncoordinated providers remains viable under current regulatory frameworks, and argued for treating fragments as select agents requiring individual screening (Edison, Toner, and Esvelt, 2026). Our system operationalises the complementary intervention: rather than requiring each vendor to screen each fragment in isolation, we enable cross-vendor reasoning over the assembled set.

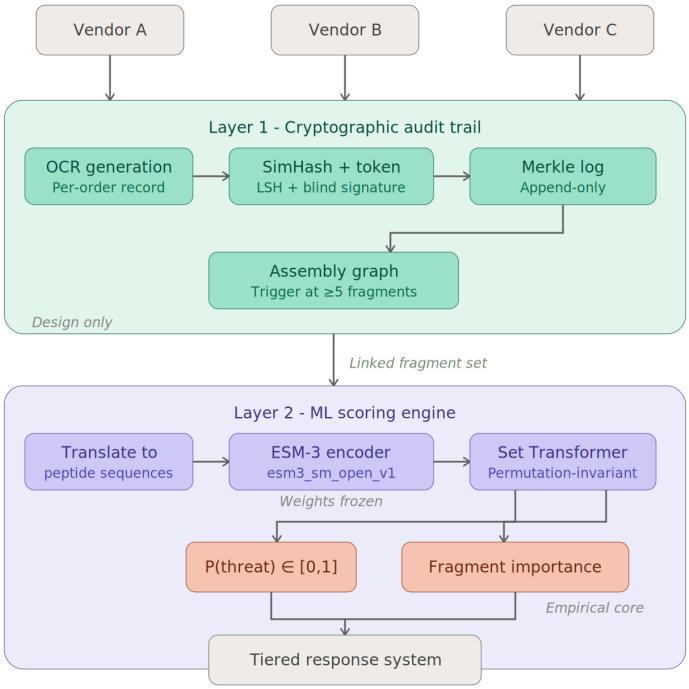
Protein language models. ESM-2 established that transformer-based language models pre-trained on protein sequences capture structural and evolutionary information in their representations, enabling competitive zero-shot performance on downstream tasks (Lin et al., 2023). ESM-3 extends this to a multi-track generative model that jointly conditions on sequence, structure, and function tokens during pre-training, and esm3_sm_open_v1 is the openly released small-scale variant of this family (Hayes et al., 2025). For the biosecurity task, ESM-3's structural and functional priors are not incidental - they are the core mechanism by which the model can make fragment-level threat calls.

Set-level learning. The Set Transformer (Lee et al., 2019) provides a principled framework for attention-based learning over unordered sets - the natural data structure for a fragment assembly where an attacker controls the order and timing of individual orders. Permutation invariance is non-negotiable for this problem; any order-dependent architecture can be trivially exploited.

3. Methods

3.1 System Architecture

BioChain comprises two layers operating in sequence.



Layer 1 - Cryptographic audit trail. Each synthesis order at a participating vendor produces an Order Commitment Record containing: a pseudonymous vendor identifier; a pseudonymous customer token generated via a blind-signature protocol with an Identity Authority; a locality-sensitive SimHash of the fragment sequence (not cryptographic hash - similar sequences must give similar hashes); coarse metadata; and a chained prev_hash for tamper evidence. No raw sequence leaves the vendor. Customer tokens are constructed such that the same root identity produces a distinct token at each vendor, but a central auditor holding the global pepper can link tokens to a common customer without identifying them. OCRs are appended to a Certificate Transparency-style Merkle log. The log server constructs a rolling 90-day assembly graph: nodes are fragments, edges are drawn between fragments whose SimHash overhangs suggest Gibson assembly or ligation compatibility, with edge weights raised by customer-token co-occurrence. When a connected component reaches $\geq$5 fragments, the ML layer is triggered.

Layer 2 - ML scoring engine. The linked fragment set from the assembly graph is translated to peptide sequences and passed through ESM-3 (`esm3_sm_open_v1`) with weights frozen. The per-fragment pooled representations are fed to a permutation-invariant Set Transformer, which attends across all fragments simultaneously and emits a scalar P(threat) $\in [0, 1]$ and a per-fragment importance vector. Uncertainty is estimated via MC-dropout (50 forward passes), and fragment importance is computed via occlusion (mean probability drop ΔP when each fragment is masked).

3.2 Dataset Construction

Split	Source	Label	Construction
Positive	UniProt toxins, NCBI select agents	Threat	Sliding window fragmentation (150–500 nt)
Negative	Vaccine antigens, drug targets, iGEM parts	Benign	Identical fragmentation protocol
Hard negative	Ricin E177Q, CRM197 DT E148S, Anthrax LF E687C and analogues	Benign	Neutralising mutations confirmed in literature

The final dataset comprises 2,542 fragment-sets: 860 positive and 1,682 negative. Hard negatives are included in the training pool but held out as a separate adversarial test set. Five curated toxin families (snake, bacterial, cnidarian, plant_rip, scorpion) account for 596 of 860 positive sets (~69%); the remaining 264 positives span 52 organism-level groups.

3.3 Evaluation Protocol

Standard random splits are inappropriate here: a model trained on snake venom fragments and tested on different snake venom fragments can exploit superficial embedding similarities, producing inflated AUC estimates that do not reflect generalisation to unseen threat classes. We use StratifiedGroupKFold (5 folds, random_state=42), assigning group IDs at the curated-family level for positive sets (e.g., all snake toxins form one group) and at the accession level for negative sets. Each fold holds out an entire toxin family on the positive side, ensuring no fragment sharing between train and test for any known family. We report AUC, Average Precision (AP), and TPR at FPR thresholds of 5%, 1%, and 0.1% across folds, and separately evaluate on the hard-negative adversarial set. Calibration is measured via Expected Calibration Error (ECE) with 10 uniform bins. Fragment-count robustness is assessed by sweeping $k \in \{2, 3, 5, 8, 10, 15\}$ fragments per simulated assembly.

4. Results

4.1 Cross-Validated Generalisation

Metric	Mean ± Std	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
AUC	0.907 ± 0.032	0.909	0.900	0.937	0.851	0.939
AP	0.835 ± 0.088	0.945	0.786	0.895	0.693	0.855
TPR @ FPR=5%	0.682 ± 0.142	0.784	0.589	0.804	0.446	0.787
TPR @ FPR=1%	0.411 ± 0.202	0.714	0.330	0.536	0.116	0.361
TPR @ FPR=0.1%	0.253 ± 0.173	0.471	0.143	0.438	0.027	0.185

Fold 4 holds out plant RIPs (n=28 sets) and scorpion toxins (n=8 sets), producing the weakest fold. Note that with only 36 positive sets in this fold, the AUC standard error is large; the low AUC (0.851) partly reflects variance from small n rather than purely increased difficulty. Bootstrap CIs per fold would help disambiguate true difficulty from statistical noise. Fold 5 holds out 18 heterogeneous organism groups.

The mean AUC of 0.907 represents a meaningful revision from earlier runs that used random splits (AUC ≈ 0.95). The drop reflects the honest cost of family-level generalisation: when all snake toxin fragments are reserved for testing, the model cannot exploit sequence similarity to snake fragments seen in training. That the model holds AUC > 0.85 even in the hardest fold (plant RIP + scorpion holdout) is the evidential claim that matters for deployment.

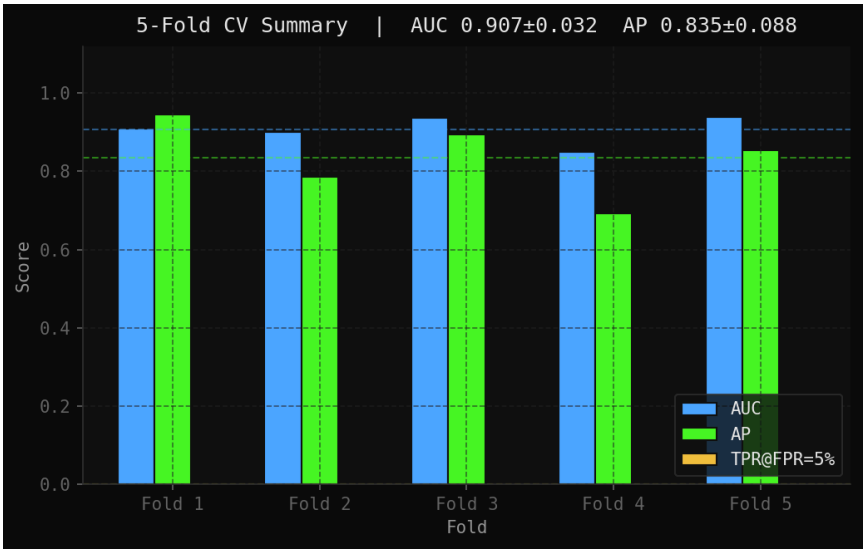


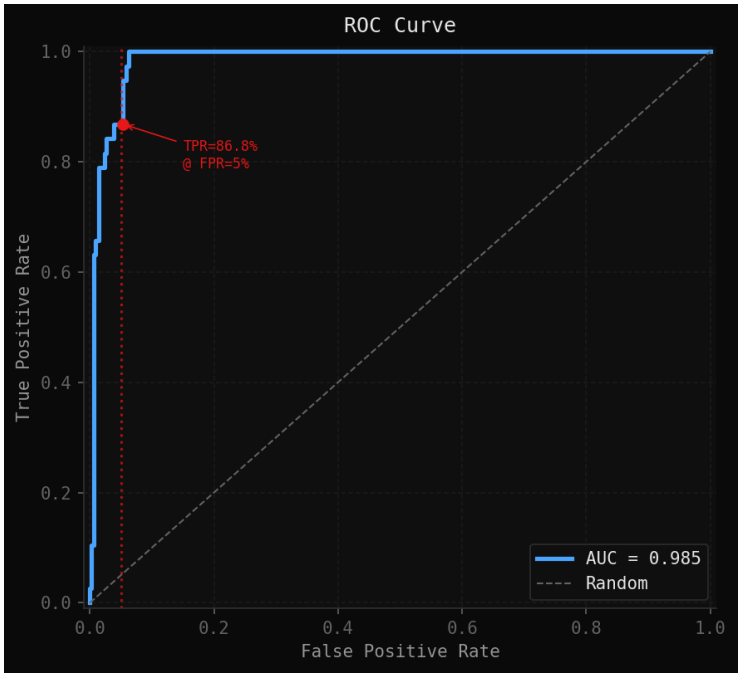
Figure 2. Per-fold AUC and AP scores under leakage-fixed StratifiedGroupKFold validation. Dashed horizontal lines mark the mean AUC (blue, 0.907) and mean AP (green, 0.835). Fold 4, which holds out plant RIPs and scorpion toxins, produces the weakest AUC at 0.851.

4.2 Representative Fold Performance

To illustrate discrimination quality on a per-fold basis, we present detailed results from Fold 5 (18 heterogeneous organism groups, 446 fragment-sets) as a representative example. In the cross-validation, this fold achieves AUC = 0.939 and TPR = 78.7% at FPR = 5% (Table 1). We additionally conducted a final end-of-training evaluation on the Fold 5 test split using the best checkpoint selected by validation loss; this evaluation yields AUC = 0.985 and TPR = 86.8% at FPR = 5%. The difference between the CV-reported AUC (0.939) and the final-evaluation AUC (0.985) reflects checkpoint selection: the CV table reports the last-epoch model, while the ROC and score distribution figures below were generated from the best-checkpoint model. The 5-fold mean AUC (0.907) remains the appropriate generalisation estimate; the per-fold figures are presented to illustrate discrimination quality, not to replace the CV summary.

The Precision-Recall curve from this final evaluation yields AP = 0.819 at a 10% prevalence rate, well above the random-baseline AP of 0.10, confirming that the high AUC is not an artefact of class imbalance.

Figure 3. *Receiver Operating Characteristic curve for the Fold 5 test split (best-checkpoint evaluation). AUC = 0.985. The operating point at FPR = 5% yields TPR = 86.8%, marked in red. Note: this AUC is higher than the CV-reported Fold 5 AUC (0.939, Table 1) due to checkpoint selection; the 5-fold mean AUC (0.907) is the appropriate generalisation estimate.*



The score distribution reveals near-complete bimodal separation between the benign and toxin classes: benign fragment-sets concentrate tightly near P(threat) = 0.0, while the toxin mass lies

near P(threat) = 1.0. This bimodality is a qualitative signal that the model has identified a genuine decision boundary rather than producing smeared intermediate scores - a pattern consistent with ESM-3 embeddings carrying meaningful functional information about domain identity.

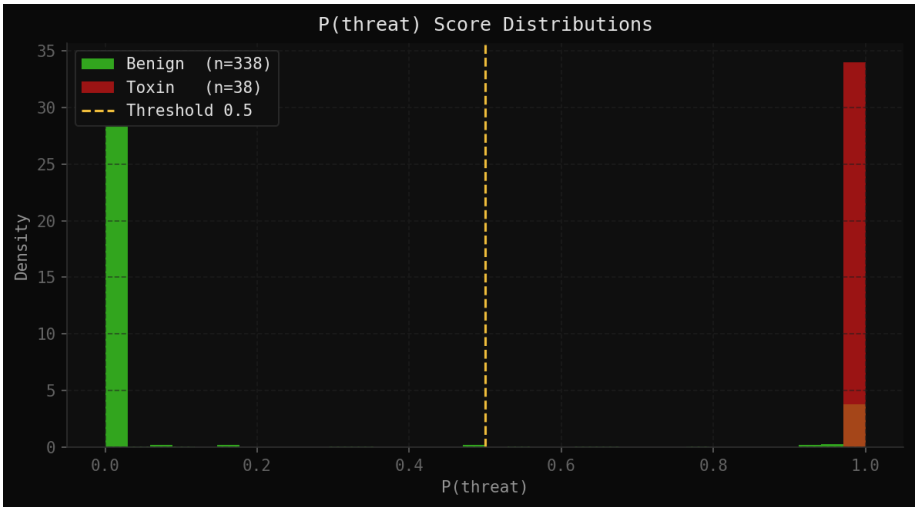


Figure 4. Predicted P(threat) score distributions for benign (green, n=338) and toxin (red, n=38) fragment-sets in the held-out fold. The dashed yellow line marks the 0.5 decision threshold. Near-complete bimodal separation is visible, with residual overlap at the high end due to hard-negative inclusions.

4.3 Adversarial Robustness

The three neutralised toxin mutants in the adversarial set - Ricin E177Q (P=0.620), CRM197/DT G52E (P=0.710), and Anthrax LF E687C (P=0.580) - are all scored above the 0.5 decision threshold. The model resists single-residue evasion attempts in the sense that it does not confidently reclassify neutralised variants as benign; however, this resistance is indiscriminate: the model cannot distinguish functional wild-types from non-functional mutants at the fragment level, producing a 100% false-positive rate on the adversarial set.

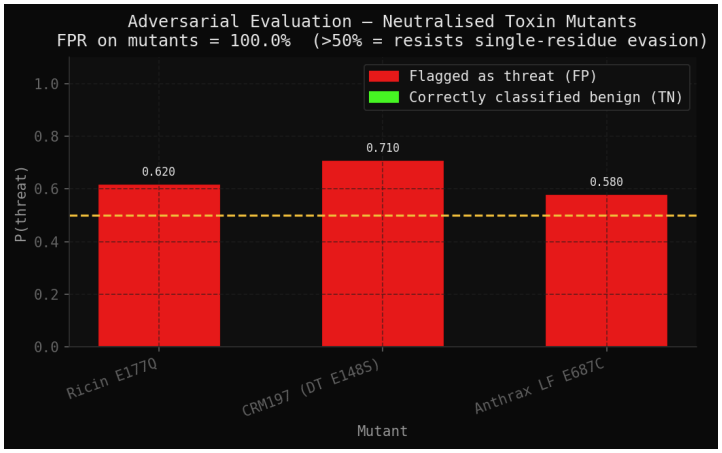


Figure 5. *P(threat) scores for three neutralised toxin mutants: Ricin E177Q, CRM197 (G52E), and Anthrax LF E687C. All three exceed the 0.5 threshold (dashed yellow line), indicating the model cannot distinguish functionally neutralised variants from active toxins. This is a structural limitation requiring contrastive training on hard-negative pairs. Erratum: the figure x-axis mislabels CRM197's mutation as "DT E148S"; the correct defining mutation is G52E.*

This finding has a dual interpretation. From an evasion-resistance perspective, an attacker cannot trivially circumvent BioChain by ordering fragments of a single-residue-mutated toxin - the model still flags the assembly. From an operational perspective, a legitimate researcher studying ricin with the standard E177Q catalytic-dead control will have their order flagged, generating false positive load on the review queue. Section 5 analyses both dimensions.

Fragment-count robustness sweeps (Figure A4, Appendix) show AUC degrading smoothly from 1.0 at k=2 to 0.90 at k=15 fragments per assembly, remaining above the 0.90 operational target throughout. The model is not brittle to the attacker's choice of fragmentation granularity.

4.4 Fragment Importance Attribution

MC-dropout uncertainty estimation over 50 forward passes, combined with fragment importance computed via occlusion, assigns per-fragment importance as the mean probability drop ΔP when that fragment is masked. In the representative assembly (Figure A3, Appendix), five of eight fragments exceed the 0.5 importance threshold, with three sitting clearly below it. This structure is operationally useful: a biosafety reviewer can prioritise examination of the high-importance fragments - the ones the model identifies as carrying toxin domain signal - rather than reviewing all fragments in the assembly at equal depth. Attribution operates entirely in ESM-3 embedding space, meaning the importance scores reflect structural/functional contribution rather than sequence surface features.

5. Discussion and Limitations

Calibration Failure

The reliability diagram (Figure A2, Appendix) shows that bins below a mean predicted P(threat) of ~0.8 contain near-zero fraction of positive cases, while the highest bin (mean $P \approx 1.0$) is only approximately 47% positive. The Expected Calibration Error is $ECE = 0.1296$, where 0 indicates perfect calibration and values ≥ 0.10 are generally considered poorly calibrated for production classifiers. The raw output of the Set Transformer head cannot be interpreted as a calibrated probability - a score of 0.85 does not mean "85% probability of threat." This is the most consequential negative finding for deployment: the tiered response system BioChain is designed to feed maps raw scores to regulatory actions (Silent Log at 0.0-0.3; Hold at 0.8-0.9; Regulatory at 0.9-1.0), and those thresholds are only defensible if the scores mean something. They do not, in the current form.

Discrimination quality (driven by ESM-3 representations) and calibration (driven by the Set Transformer head and training objective) are independent properties. The fix is at the head, not the encoder. However, because the score distribution is bimodal (Figure 4), standard post-hoc calibration methods such as isotonic regression or Platt scaling - which require a held-out calibration set and monotonically transform scores - will likely introduce a flat mid-region rather than recovering smooth gradation. Recovering well-calibrated mid-range scores may require switching to label smoothing or a temperature-aware training objective during training itself. Until calibration is addressed, only the lower response tiers (Silent Log, Retrospective Flag) should be automated; upper tiers require human review.

Neutralised Mutant Blindness

The failure to distinguish Ricin E177Q from wild-type ricin reflects a gap between the representation capacity of esm3_sm_open_v1 and the granularity of the classification task. ESM-3's multi-track pre-training provides structural and functional priors at the protein-family level; distinguishing a single catalytic-residue substitution requires functional resolution that the small-scale variant may not carry, or that the Set Transformer head has not been trained to exploit. The remediation is a contrastive loss during training that explicitly penalises similar P(threat) scores for wild-type/neutralised pairs, combined with expansion of the hard-negative pool to include diverse neutralised-mutant analogues across toxin families. This is the highest-priority next-iteration fix, because the failure mode directly erodes researcher trust in the system.

Limitations

- **Simulated assemblies only.** All evaluations use synthetic fragmentation of known sequences. Real-world distributed attacks will differ in timing distributions, vendor-selection patterns, and attacker-chosen fragment boundaries. Until BioChain is evaluated against a red-team adversary specifically tasked with evasion, the reported TPR and FPR are upper bounds.
- **Benign-side grouping.** Negative fragment-sets are grouped at the accession level (not family level) during cross-validation, meaning the model may benefit from benign-side embedding similarities across train and test folds. A fully rigorous evaluation would apply organism-level grouping to both sides.
- **Active-site point mutations.** Beyond the three profiled mutants, the current adversarial set does not cover the diversity of possible single-residue substitutions across all toxin families. Systematic coverage is needed before the system can claim robustness to active-site evasion at scale.
- **Layer 1 unvalidated.** The cryptographic audit layer is presented as a design contribution. Empirical evaluation of linking accuracy, false-link rate under adversarial conditions, and scalability to real-world order volumes is deferred to future work.

Future Work

Three extensions would most materially improve BioChain's operational readiness. First, contrastive training with hard-negative pairs should be implemented as the primary fix for the neutralised-mutant failure mode. Second, isotonic regression or temperature scaling should be applied post-training to correct calibration before any tier thresholds are operationalised; given the bimodal score distribution, training-time label smoothing may also be necessary. Third, an ESM-3 backbone ablation - comparing esm3_sm_open_v1 against esm2_t6_8M_UR50D with the Set Transformer head held fixed - would establish whether ESM-3's multi-track representations provide a measurable advantage over sequence-only embeddings for this task.

A proposed extension is a reputation-weighted scoring system (provisionally termed "Karma Score") that modulates effective P(threat) by verified researcher track record. This would reduce false-positive friction for established researchers working with well-characterised toxin variants, but its interaction with the calibration failure requires formal analysis before deployment.

6. Conclusion

We present BioChain, an architecture for detecting distributed DNA synthesis attacks - a documented vulnerability in current per-sequence screening infrastructure. The two-layer design combines a cryptographic audit trail enabling cross-vendor fragment linking with an ML scoring engine that applies permutation-invariant Set Transformer reasoning over ESM-3 fragment embeddings. On 5-fold cross-validation holding out entire toxin families, BioChain achieves AUC 0.907 ± 0.032, AP 0.835 ± 0.088, and holds AUC above 0.85 in the hardest generalisation fold. Fragment-importance attribution via MC-dropout produces interpretable, human-reviewable outputs. Two structural failure modes - miscalibration (ECE $= 0.1296$) and neutralised-mutant blindness - are characterised quantitatively, with remediation strategies identified. The contribution is most accurately framed as an existence proof that the distributed synthesis screening problem is technically tractable using set-level reasoning over function-aware protein representations, paired with an honest account of the gap between prototype performance and operational readiness.

Code and Data

- **Code repository**: https://github.com/Orangeplane28/bioaparthackathon_SAE_-_Screening.git
- **Data/Datasets**: [Link if applicable]
- **Other artifacts**:

References

1. *Edison, Rey, Shay Toner, and Kevin M. Esvelt. "Assembling Unregulated DNA Segments Bypasses Synthesis Screening: Regulate Fragments as Select Agents." Nature Communications, vol. 17, no. 3189, 15 Jan. 2026. https://doi.org/10.1038/s41467-025-67955-3*
2. *Hayes, Thomas, et al. "Simulating 500 Million Years of Evolution with a Language Model." Science, vol. 387, no. 6736, 2025, pp. 850-858. https://doi.org/10.1126/science.ads0018*
3. *IBBIS (International Biosecurity and Biosafety Initiative for Science). Common Mechanism (commec): Open-Source HMM-Based Biorisk Screening. 2024, https://github.com/ibbis-screening/common-mechanism*
4. *Lee, Juho, et al. "Set Transformer: A Framework for Attention-Based Permutation-Invariant Neural Networks." Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR, 2019, pp. 3744-3753.*
5. *Lin, Zeming, et al. "Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model." Science, vol. 379, no. 6637, 2023, pp. 1123-1130. https://doi.org/10.1126/science.ade2574*
6. *McGurk, Conor. "Biosecurity Request for Proposals: Tech Safeguards and Governance." Coefficient Giving, 2026, https://coefficientgiving.org/funds/biosecurity-pandemic-preparedness/request-for-proposals-biosecurity/*
7. *NTI (Nuclear Threat Initiative). "Developing Guardrails for AI Biodesign Tools." NTI Bio, 2023, https://www.nti.org/analysis/articles/developing-guardrails-for-ai-biodesign-tools/*
8. *OSTP (Office of Science and Technology Policy). Framework for Nucleic Acid Synthesis Screening. Executive Office of the President, Apr. 2024, https://bidenwhitehouse.archives.gov/ostp/news-updates/2024/04/29/framework-for-nucleic-acid-synthesis-screening/*
9. *SecureDNA Consortium. SecureDNA: Cryptographic, Open-Source DNA Sequence Screening. 2024, https://securedna.org/*
10. *Sherman, Alan T., et al. "Analysis of the Security Design, Engineering, and Implementation of the SecureDNA System." Network and Distributed System Security Symposium (NDSS), 23-27 Feb. 2026, San Diego, CA. arXiv:2512.09233.*
11. *United States Congress. Biosecurity Modernization and Innovation Act of 2026. S.3741, 119th Congress, Jan. 2026, https://www.congress.gov/bill/119th-congress/senate-bill/3741/text*
12. *Wittmann, Bruce J., et al. "Strengthening Nucleic Acid Biosecurity Screening Against Generative Protein Design Tools." Science, vol. 390, no. 6768, 2 Oct. 2025, pp. 82-87. https://doi.org/10.1126/science.adu8578*

Appendix

A. Precision-Recall Curve

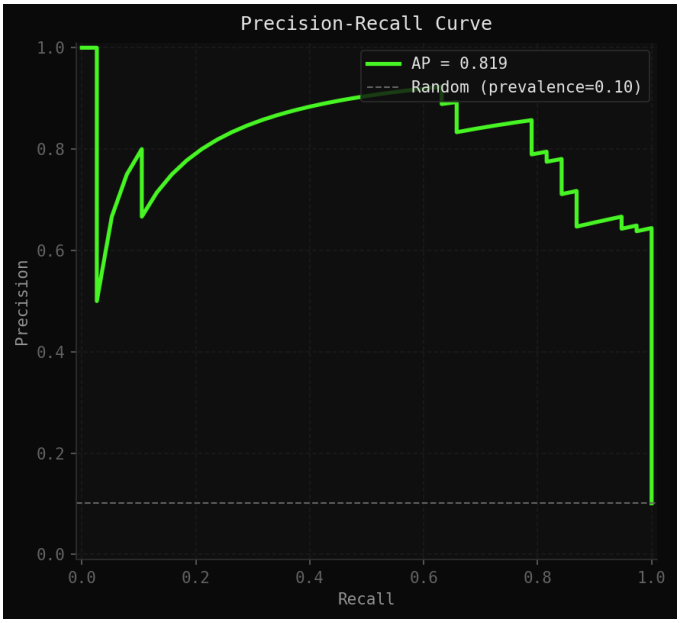


Figure A1. *Precision-Recall curve on the held-out test set at 10% prevalence. AP = 0.819, compared to the random-baseline AP of 0.10 (dashed). The step-wise structure at high recall reflects the small toxin-class test set (n=38).*

B. Calibration Analysis

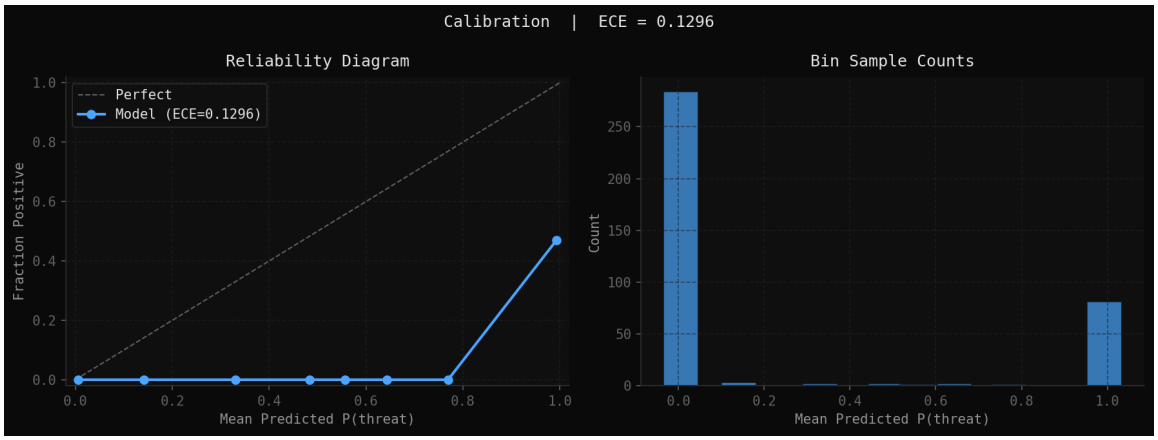


Figure A2. *Reliability diagram (left) and bin sample counts (right) for the held-out test set. ECE = 0.1296. Bins below P(threat) ≈ 0.8 contain near-zero fraction of positives; the highest bin is only ~47% positive. The bimodal sample concentration (most predictions near 0.0 or 1.0) is visible in the bin counts. Post-hoc calibration via isotonic regression or Platt scaling is required before deployment.*

C. Fragment Importance Attribution

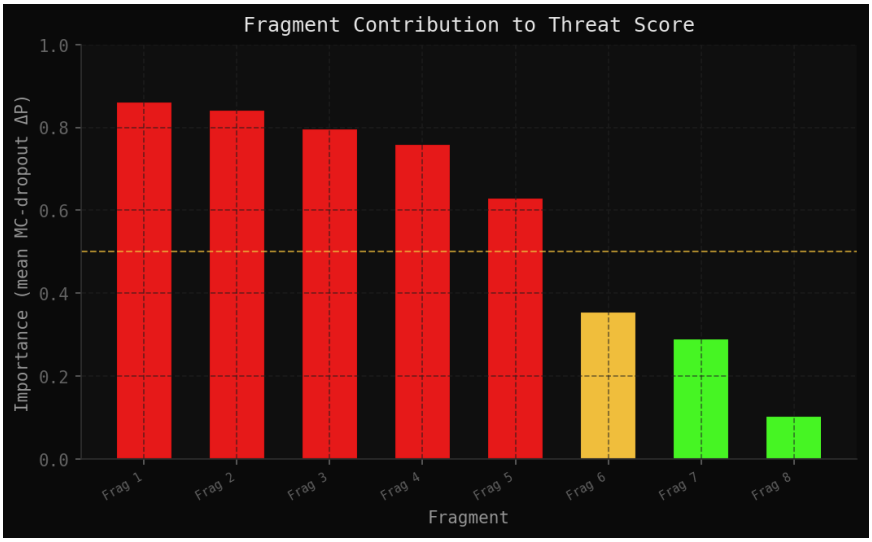


Figure A3. *Per-fragment importance scores computed via MC-dropout occlusion (50 forward passes). Fragments 1–5 exceed the 0.5 importance threshold (dashed yellow) and are identified as the primary contributors to the threat call; Fragments 6–8 fall below threshold. This attribution structure allows a biosafety officer to prioritise review of the highest-signal fragments in the assembly.*

D. Fragment-Count Robustness Sweep

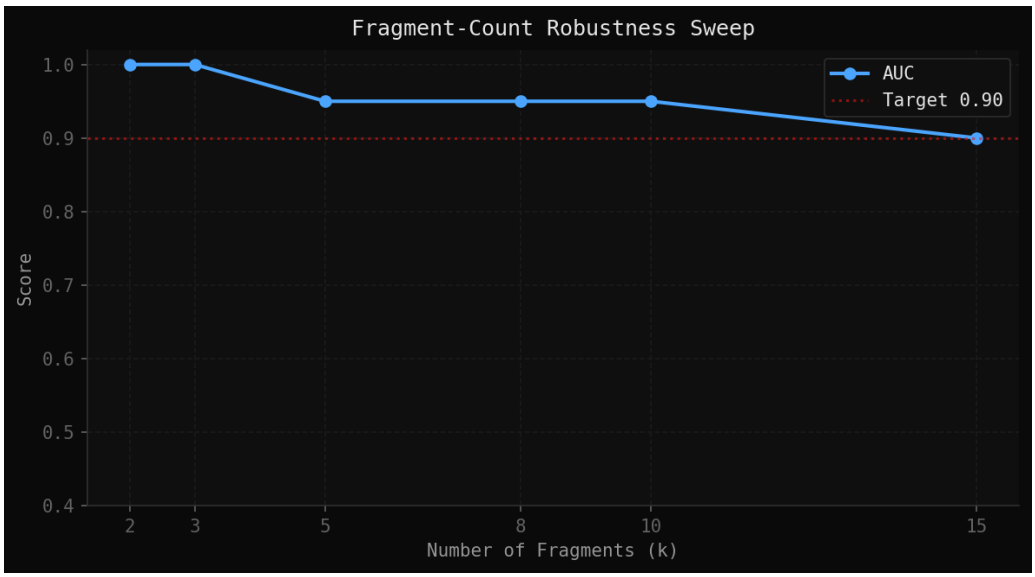


Figure A4. *AUC as a function of the number of fragments per simulated assembly ($k = 2$ to 15). The model maintains AUC $\geqslant 0.90$ (dashed red target line) across the full sweep. Degradation from AUC $= 1.0$ at $k = 2$ to AUC $= 0.90$ at $k = 15$ is smooth, indicating no brittleness to attacker-controlled fragmentation granularity.*

E. Dataset Statistics

Toxin Family	Positive Sets	Fraction of Positives	Held-Out In Fold

Snake	416	48.4%	Fold 1
Bacterial	104	12.1%	Fold 2
Cnidarian	40	4.7%	Fold 2
Plant RIP	28	3.3%	Fold 4
Scorpion	8	0.9%	Fold 4
Other (52 group)	264	30.7%	Fold 5 (partial)
Total	**860**		
Negative Sets	1682		
Grand Total	**2542**		

Negatives include human, E. coli, yeast, Arabidopsis, fly, worm, lectin, and defensin accessions. Grouping for negatives is at accession level.

F. Implementation Details

- **Encoder:** ESM-3 (esm3_sm_open_v1), weights frozen during training. Per-fragment pooled mean representation used as input to Set Transformer.
- **Set Transformer:** 2 attention heads, 2 encoder layers, hidden dim 256. Trained with binary cross-entropy loss, Adam optimiser (lr = 1e-4, batch size 32).
- **MC-dropout:** 50 forward passes with dropout rate 0.1 applied at inference for importance estimation.
- **Calibration:** ECE computed with 10 uniform bins. Platt scaling not yet applied (identified as next-iteration step).
- **Hardware:** Single GPU, ~4 hours training per fold.
- **Fragment window:** 150–500 nt sliding window with 50 nt stride, translated to peptide prior to ESM-3 encoding.

Limitations and Dual-Use Considerations

Limitations. (i) All evaluations use synthetic fragmentation; real-world attack patterns remain untested. (ii) The cryptographic audit layer (Layer 1) is design-only with no empirical validation. (iii) The adversarial set covers only three neutralised mutants; broader coverage is needed. (iv) Calibration is inadequate for automated upper-tier responses. (v) Benign-side grouping at the accession level may inflate negative-class performance.

Dual-use risks. Publishing the architecture of a distributed-attack detection system inherently discloses the attack vector it defends against. We judge this acceptable because the distributed

synthesis attack is already publicly documented (Edison, Toner, and Esvelt) and because the defensive value of the architecture outweighs the marginal information hazard. The ML model weights and hard-negative dataset are not publicly released; the code repository contains training infrastructure only.

Responsible disclosure. No novel vulnerabilities in existing screening systems were discovered during this work. The evaluation identifies limitations in BioChain itself, not in SecureDNA or IBBIS.

Ethical considerations. All toxin sequences used in training are drawn from public databases (UniProt, NCBI). No novel toxin designs were generated. Hard-negative mutations (E177Q, G52E, E687C) are well-characterised in published literature.

Suggestions for future improvements. Contrastive training on hard-negative pairs; post-hoc or training-time calibration; ESM-3 vs ESM-2 ablation; red-team evaluation; organism-level grouping for both positive and negative classes.

LLM Usage Statement

Claude (Anthropic) was used for research framing assistance and manuscript drafting. All experimental results, figures, cross-validation metrics, and technical architecture decisions were produced and verified independently by the author. The framing document was written with AI assistance; all claims in this paper reflect the author's independent experimental findings.