# A short-query DNA threat-screening prescreen

Tyler Rector
*Old Dominion University*
With Apart Research

**Abstract**

Existing DNA biosecurity screeners such as SecureDNA and IBBIS Common Mechanism are designed for high-precision homology or exact-match detection on long queries. Both are unable to operate meaningfully below approximately 30 base pairs, leaving a regime relevant to short-read screening uncovered. A gradient-boosted classifier was trained on DNA five-mer and six-frame protein three-mer presence vectors, with parent-level held-out splits, multi-length window augmentation, reverse-screening as a post-filter, low-complexity masking, and per-length threshold calibration. Evaluation followed a three-way protocol comprising a training holdout, a same-distribution similar set, and a phylogenetically distant external set. On the training distribution the model reaches AUC 0.77 at 15 to 29 bp, 0.82 at 30 to 59 bp, and 0.86 at 200 bp and above. Within-distribution generalization holds with AUC dropping by only 0.02 to 0.03 on the similar set. Out-of-distribution generalization fails: external AUC is 0.60 to 0.63 across all length buckets. AT-rich benign sequences are over-flagged (mean score difference +0.16) consistently across all three holdout sets, indicating a real model property tied to compositional bias rather than a training-set artifact. The model is a suitable prescreen on its training distribution at lengths where homology methods do not work, but is not safe to deploy on phylogenetically novel organisms without retraining on broader benign coverage.

## 1. Introduction

DNA synthesis screening is a key biosecurity control point. Synthesis providers screen incoming orders for sequences that match known biothreats before fulfilling them. Two production-grade tools dominate this space. SecureDNA [1] uses cryptographic hash matching against a curated hazard database, with a 30 bp minimum query length by design and a distributed architecture that requires certificate-based access to its hazard set. IBBIS Common Mechanism (commec) [2] runs a homology pipeline combining HMM searches and BLAST against multi-hundred-gigabyte reference databases. Both produce high-precision, low-false-positive output on long queries where homology signals are statistically detectable.

The regime below 30 bp is uncovered by both. SecureDNA rejects queries shorter than its minimum. Commec relies on BLAST and HMM, which both lose statistical power below approximately 10 to 20 amino acid residues, equivalent to 30 to 60 bp. This regime matters in practice because short-read sequencing platforms emit fragments at this scale, and a screening prescreen that operates here would let downstream homology tools focus their compute on candidate hits rather than every read.

Statistical machine learning is a natural fit for this regime, but it raises the question of whether a learned classifier generalizes beyond its training distribution. Most published threat-screening evaluations report only within-distribution performance and stop there. This is insufficient for biosecurity tools, where the cost of failing on a novel pathogen is exactly the deployment scenario the system is built for.

This work makes the following contributions.

1. A gradient-boosted DNA and protein k-mer classifier reaching AUC 0.77 at 15 to 29 bp on its training distribution, with calibrated per-length thresholds and a portable bundle that runs anywhere Python and XGBoost are available.
2. A three-way evaluation protocol (training holdout, similar-distribution, external out-of-distribution) that exposes the difference between within-distribution overfitting and true generalization failure, recommended as standard practice for short-query screening models.
3. This work includes a negative finding with operational consequences. The classifier achieves only an AUC of 0.60 on phylogenetically distant organisms, even with broad training data. This is a useful prescreen on its training distribution but is not safe to deploy on novel organisms without retraining.

## 2. Related Work

**SecureDNA** [1] is a distributed cryptographic hash-matching system. Sequence orders are hashed and queried against a precomputed hazard set held by independent keyservers, with the query never revealing the underlying sequence. Its architecture is designed for high precision on novel synthesis orders and a 30 bp minimum query length is enforced. The hazard database is not publicly distributed and access requires certificate provisioning.

**IBBIS Common Mechanism** [2] is an open-source homology-based pipeline. A query is searched against curated HMM profiles and BLAST databases of regulated pathogens. Decisions are made by aggregating hits with hand-tuned rules. The system is designed for orders longer than approximately 200 bp where homology signals are robust, and is computationally expensive (seconds to minutes per long query against multi-hundred-gigabyte databases).

Both systems share a design assumption. Queries must be long enough for biological signals to be detected through exact-match or alignment-based homology. Neither attempts to operate at the short-read scale (15 to 60 bp). The method described here targets that regime explicitly, accepting the tradeoff that statistical classification produces probabilistic scores rather than the deterministic hits these systems produce, and is therefore appropriate as a prescreen rather than a final decision tool. Table 1 summarizes the architectural differences.

| Property | SecureDNA | IBBIS commec | This work |
|---|---|---|---|
| Detection method | Cryptographic exact-match hash | BLAST + HMM homology | Gradient-boosted k-mer ML |

| Property | SecureDNA | IBBIS commec | This work |
|---|---|---|---|
| Min query length | 30 bp (by design) | ~30 bp BLAST, ~60 bp HMM | 15 bp (statistical, AUC 0.77) |
| Output | Match / no match | Hit list with E-values | Continuous score in [0,1] |
| Database size | Hazard set on keyservers | ~300+ GB BLAST + HMM | ~10 MB portable bundle |
| Per-query time | Sub-second hash lookup | Seconds to minutes | Sub-millisecond inference |
| Access requirements | Certificate from issuer | Open source + DB download | Open source + ~10 MB bundle |
| Failure mode | Misses non-listed hazards | Misses novel homologs | Mis-scores OOD organisms |

*Table 1: Architectural comparison of DNA threat-screening tools. SecureDNA and commec target the long-query, high-precision regime; this work targets the short-query, statistical-prescreen regime.*

# 3. Methods

## 3.1 Training data

Training data was imported from NCBI Nucleotide using two bacterial categories: threat (virulence factors, toxins, secretion systems, and named pathogens including Salmonella, Staphylococcus aureus, Bacillus anthracis, and Francisella tularensis) and benign (housekeeping genes, ribosomal proteins, and metabolic enzymes from non-pathogenic bacteria including Bacillus subtilis, Lactobacillus, and Streptomyces). A final training run used 5,000 threat and 10,000 benign parent sequences, deduplicated against a persistent accession log so repeated runs return novel sequences rather than re-fetching the same NCBI top hits.

All queries used the genomic-DNA filter (biomol_genomic) so the model trained exclusively on DNA sequences with valid A, C, G, T, N composition. A length filter of 200 to 5,000 bp was applied at fetch time to exclude both short fragments and full chromosomes.

## 3.2 Sequence preparation and feature extraction

Each input DNA sequence is processed through a fixed pipeline before scoring. The sequence is first uppercased and validated for A, C, G, T, N composition. When low-complexity masking is enabled, the DUST algorithm replaces homopolymer runs and simple-repeat regions with N characters using a triplet-entropy threshold of 2.5 over a 32 bp sliding window.

Two complementary feature representations are then extracted. For DNA k-mer features, a sliding 5-mer window walks the masked sequence and increments a count for every 5-mer present in the model's discriminative dictionary; the resulting count vector is normalized by sequence length. For protein k-mer features, the sequence is translated in all six reading frames (three forward frames from the input and three frames from its reverse complement, using the standard genetic code with stop codons mapped to a separator). A sliding 3-mer window over each of the six translated

protein strings increments counts for every protein 3-mer in the model's dictionary, accumulated into a single unnormalized vector across all six frames.

The DNA presence vector (412 features after dictionary filtering) and the protein presence vector (713 features) are concatenated into a single 1,125-dimensional vector that XGBoost scores. The raw probability output is then passed through an optional reverse-screening post-filter: 10-mer presence is checked against a threat reference set and an innocuous reference set, and the raw score is geometrically combined with the threat-to-innocuous hit ratio.

Discriminative dictionary filtering retains a k-mer if its log-odds score (log of threat-parent presence rate over benign-parent presence rate, with Laplace smoothing) exceeds 0.3 and the k-mer appears in at least five threat parents (DNA) or ten threat parents (protein).

## 3.3 Window fragmentation for short-query training and evaluation

Parent sequences from NCBI are typically several hundred to several thousand bp long, but deployment queries are often much shorter (15 to 200 bp). To bridge this gap, parents are fragmented into shorter windows during both training and evaluation. For training, each parent is sampled at five target lengths (15, 20, 25, 35, and 50 bp) with multiple random offsets per length, producing thousands of short labeled windows that the model sees alongside the full parent sequences. This window augmentation gives the model exposure to short-query distributions during training rather than only at inference time.

For evaluation, the same fragmentation procedure is applied to held-out parents at five target lengths (20, 30, 50, 100, and 200 bp) with three random offsets per parent per length. Each fragment becomes a separate scored query. Length-bucketed metrics are then reported by binning fragments into [15, 30), [30, 60), [60, 100), [100, 200), and [200, inf) bp groups. This produces apples-to-apples length-bucketed comparisons that reflect realistic short-read deployment performance rather than the inflated metrics that would come from scoring full parent sequences.

## 3.4 Honest evaluation protocol

All splits are made at the parent level (the original full-length sequence) rather than the fragment level, with stratification by source so that fragments derived from a single parent never appear in both the train and test sets. Standard splits use 65% train, 15% validation, and 20% test.

The pipeline tests four conditions head to head. These include a baseline trained on parents alone, a reverse-screening variant that scales the score by the ratio of long (10-mer) hits against threat versus innocuous reference sets, a synonymous-variant augmentation variant that adds five synonymous codon variants per train threat parent, and the combination. The winning condition is selected by held-out test AUC, then retrained on the multi-length windows described in Section 3.3. A second comparison evaluates GC-content matching of the benign training set and DUST low-complexity masking on top of the multi-length winner.

## 3.5 Three-way evaluation

Beyond the standard training holdout, the model is evaluated on two additional sets.

- Similar distribution refers to fresh accessions drawn from the same diverse query bank used in training, in a separate deduplication cache. Accessions in this set are guaranteed to be absent from training. Tests within-distribution generalization to unseen sequences from organisms the model has seen.
- External (out-of-distribution) includes phylogenetically distant organisms such as Borrelia, Treponema, Leptospira, Chlamydia, Mycoplasma, and Aeromonas, as well as novel benign sources including Acidobacterium, Verrucomicrobia, Aquifex, Nitrospira, Halobacterium, Sulfolobus, archaea, and environmental metagenomes. Tests true out-of-distribution generalization.

All three sets are fragmented and scored through the identical pipeline described in Section 3.3, producing matched length-bucketed metrics across all sets.

The trained model, dictionaries, long k-mer reference sets, calibrated thresholds, and a self-contained Python loader are packaged into a single tar.gz bundle (approximately 10 MB) so that evaluation does not require access to the training notebook. A standalone evaluation notebook loads the bundle, reads sequences from FASTA or fetches them from NCBI, and produces the three-way comparison automatically.

# 4. Results

## 4.1 Performance on the training distribution

Figure 1 shows length-dependent performance on the held-out test set. AUC rises monotonically with query length, from 0.77 at 15 to 29 bp to 0.86 at 200 bp and above. Sensitivity at 5% false-positive rate climbs from 26% at 15 to 29 bp to 49% at 200 bp and above. The model exceeds AUC 0.80 at all length buckets above 30 bp.

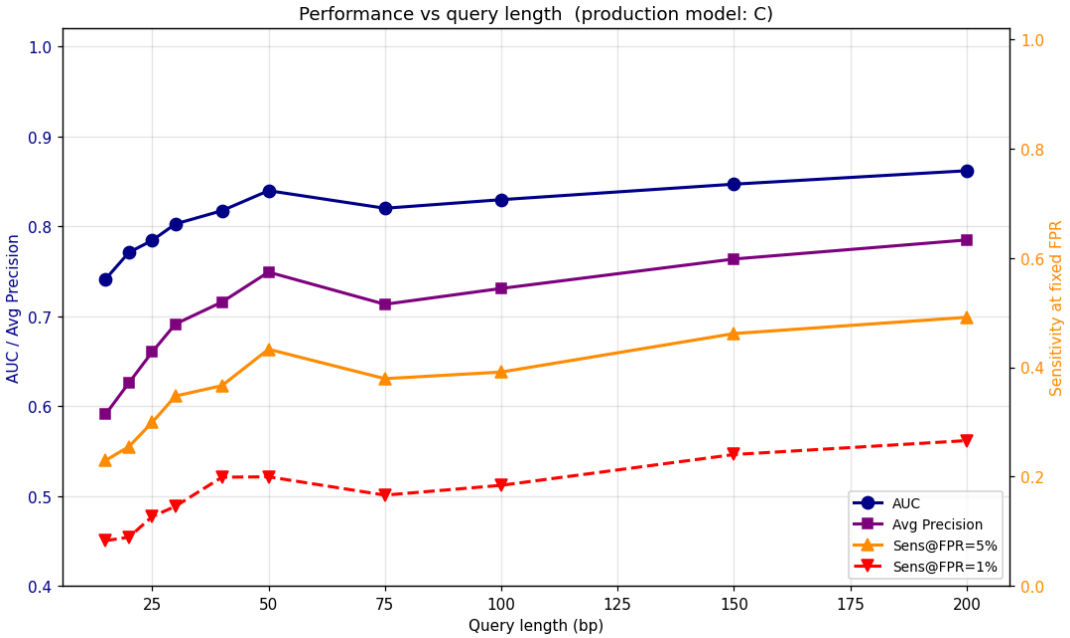

*Figure 1: Length-dependent performance of the production model on held-out test parents. AUC and Average Precision are read on the left axis; sensitivity at fixed false-positive rates of 1% and 5% is read on the right axis.*

Calibration is meaningful but imperfect. The Expected Calibration Error is 0.10, with the model under-predicting probabilities in the high-confidence range (sequences scoring around 0.5 are observed positives roughly 60% of the time). Score thresholds for given false-positive targets were calibrated separately for each length bucket from the validation split.

## 4.2 Three-way generalization

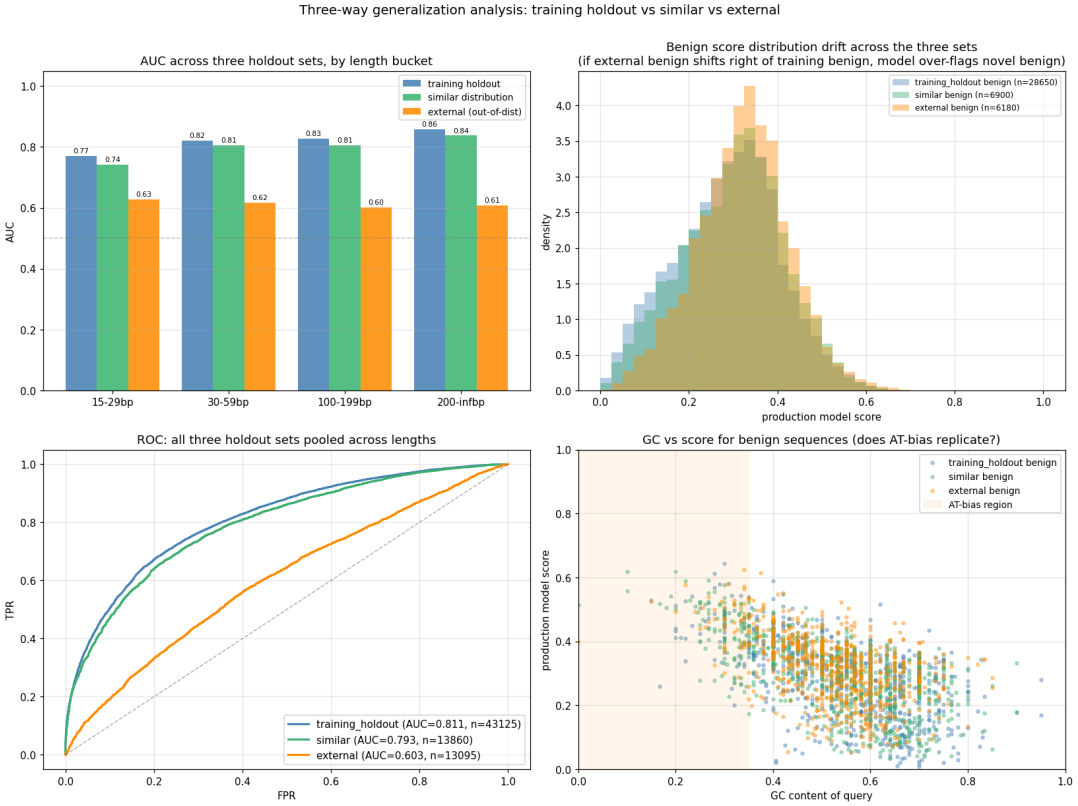


*Figure 2: Three-way evaluation on the production model. Top-left: AUC by length bucket across the three holdout sets. Top-right: benign score distributions overlaid; external benign do not shift right of training benign, indicating no over-flagging beyond what is already present. Bottom-left: pooled ROC curves; external curve sits well below the training and similar curves. Bottom-right: GC content versus model score for benign sequences from each set.*

The training holdout AUC of 0.77 to 0.86 (depending on length bucket) drops only slightly on the similar-distribution set (0.74 to 0.84), confirming that the model generalizes to novel sequences from organisms it was trained on. On the external out-of-distribution set, AUC collapses to 0.60 to 0.63 across all length buckets and stays near chance regardless of query length.

The benign score distribution comparison (Figure 2, top right) shows that external benign scores do not shift meaningfully right of training benign scores. The three benign-score distributions are nearly identical in shape and central tendency. This rules out an over-flagging mechanism in which

novel benign sequences are systematically called as threats. Instead, the failure on external data is in threat detection: external threats from phylogenetically distant pathogens score similarly to external benign, with mean separation collapsing from +0.150 (training) to +0.039 (external).

## 4.3 Compositional bias replicates

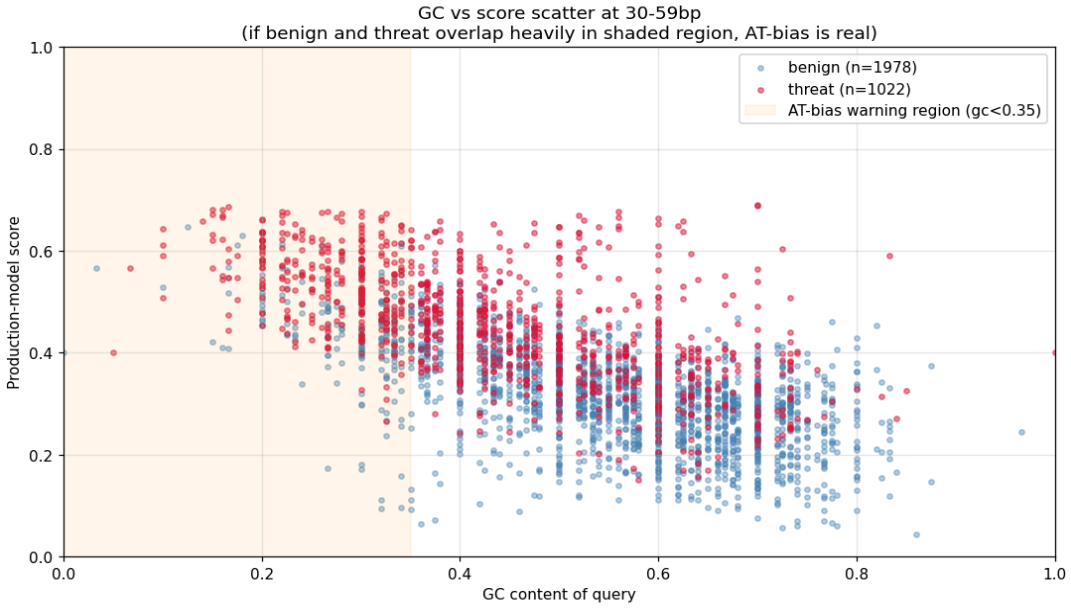


*Figure 3: GC content versus model score at 30 to 59 bp on training-holdout test queries. Threat sequences (red) cluster at higher scores than benign (blue), but score is also clearly correlated with low GC content: the upper-left region (low GC, high score) contains both threats and a population of false-positive benigns.*

The interpretability analysis reveals a real compositional bias in the learned model. The top DNA k-mers by XGBoost gain are dominated by AT-rich five-mers (ATATA with log-odds +0.58, TAATA with +0.55, TATAT with +0.53, TTATA with +0.51, TATAA with +0.50). On the held-out test set, mean score for AT-rich benign sequences (GC content below 0.35) is 0.43 compared to 0.28 for normal-GC benign sequences, a difference of +0.16. This bias replicates almost exactly on the similar set (+0.17) and the external set (+0.16), confirming it is a real model property and not an artifact of the training distribution.

The protein features tell a different and more biologically defensible story. Of total feature gain, 73% comes from protein k-mers and 27% from DNA k-mers. Top protein k-mers include hydrophobic and aromatic clusters consistent with transmembrane and aromatic-binding regions of bacterial proteins, suggesting that the model is learning real biology at the protein level even while relying on compositional bias at the DNA level.

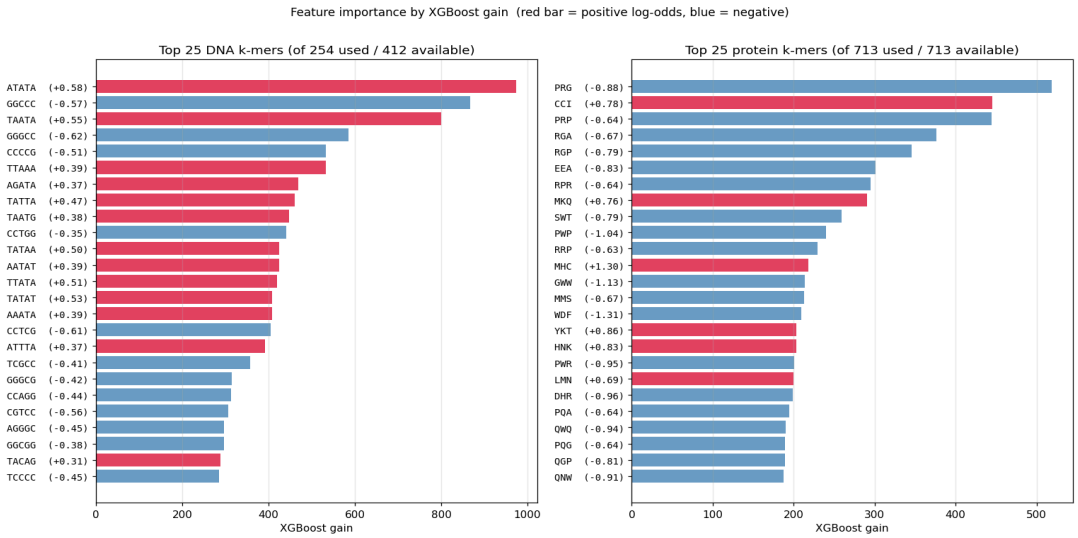


*Figure 4: Top DNA five-mers (left) and protein three-mers (right) by XGBoost gain. Red bars indicate positive log-odds (threat-indicative); blue bars indicate negative log-odds (benign-indicative). DNA features are dominated by AT-rich k-mers; protein features show more biologically interpretable composition.*

### 4.4 Intervention ablations did not help

To test whether the AT-content bias was reducible by simple training-time interventions, the multi-length baseline was compared head-to-head against three alternatives: GC-matching the benign training parents to the threat GC distribution (B), DUST masking of low-complexity regions (C), and both combined (BC). Mean AUC over the short length range (15 to 30 bp) was 0.7709 for the baseline, 0.7707 for B, 0.7718 for C, and 0.7713 for BC. None of these differences exceeds plausible run-to-run noise. The AT-rich content of the top false positives is unchanged across conditions. The AT-bias is therefore an irreducible property of the training data rather than a preprocessing artifact, consistent with the observation that the bias replicates on the external set.

## 5. Discussion and Limitations

### 5.1 Implications

The model is capable of discriminating threat versus benign in the presence of organisms whose composition matches its training set. The result says nothing positive about the model's behavior on novel organisms. The model is a defensible prescreen for sequences from training-adjacent organisms but should not be trusted on phylogenetically novel queries without manual review or paired homology screening.

### 5.2 Limitations

Several constraints limit the scope of the conclusions. First, calibration drifts from perfect (ECE = 0.10) and the per-length thresholds are fit on a finite validation split; deployment use should re-calibrate on the deployment distribution. Second, the external set is constructed from queries on phylogenetically distant organisms, but bacteria are still over-represented relative to viruses,

fungi, and eukaryotic threats; conclusions about generalization to non-bacterial threat agents are not supported by these results. Third, the feature space is fixed (DNA 5-mers + protein 3-mers); architectures using learned representations such as transformer-based DNA language models might transfer differently. Fourth, the AT-bias arose despite GC-matching and DUST masking; whether it is fundamentally reducible through a different feature representation, or whether it reflects a genuine compositional difference between bacterial threat and benign coding regions, is not resolved here.

### 5.3 Future work

The natural next step is broader benign coverage during training, specifically extending the negative class to include benign sequences from the same phylogenetic clades currently used as the external evaluation set. If retraining closes the external AUC gap to within 0.05 of the training holdout, the architecture is sound and the prior failure was a data coverage issue. If the gap remains, the limitation is in the feature representation and a learned-representation model would be the right next direction.

## 6. Conclusion

A gradient-boosted DNA and protein k-mer classifier achieves an AUC of 0.77 for 15 base pair queries on its training distribution. This statistical prescreen fills a functional gap left by exact-match systems like SecureDNA and homology pipelines like commec. These existing systems cannot operate effectively at the short-read scale. While the classifier currently struggles to generalize to phylogenetically distant organisms, it successfully functions as a probabilistic prescreen for organisms within its training distribution.

## Code and Data

Source code, training notebook, evaluation notebook, and a portable model bundle are included in the submission archive.

Datasets used include NCBI Nucleotide (genomic DNA filter, length 200 to 5,000 bp), accessed via Bio.Entrez. Specific accession lists are in the training_sequences_log.csv produced by the training notebook.

## References

[1] SecureDNA Foundation. SecureDNA: cryptographic biosecurity screening. Source repository: github.com/SecureDNA/SecureDNA. Accessed 2026.

[2] International Biosecurity and Biosafety Initiative for Science (IBBIS). Common Mechanism (commec): an open-source biosecurity screening pipeline. Source repository: github.com/ibbis-bio/common-mechanism. Accessed 2026.

[3] Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. KDD '16.

[4] Wootton, J. C., and Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. Methods in Enzymology, 266, 554-571.

[5] National Center for Biotechnology Information. Entrez Programming Utilities. www.ncbi.nlm.nih.gov/books/NBK25497/.

## LLM Usage Statement

Claude to aid in implementing the project. Claude was used in structuring and the drafting of this paper. All claims were verified independently.