# SynthGuard and BioLens: Discriminative Synthesis Screening and Intelligence-Aware Biosecurity Triage

A V S M Ashok Kumar
Independent Researcher
avsmashokkumar@gmail.com

Akhil C
Independent Researcher
akhilc0101@outlook.com

Submitted to AIxBio Hackathon 2026

**Focus:** DNA/protein synthesis screening; biosecurity operations tooling

**Abstract.** AI-enabled protein design and codon optimization stress synthesis-screening workflows that rely on sequence identity. We built **SynthGuard**, a lightweight DNA/protein screener using k-mer, codon-usage, codon-adaptation, amino-acid, physicochemical, and embedding features. On a 2,214-sequence held-out DNA test set, blastn at 70% identity achieved 99.4% recall but 94.3% false-positive rate, while **SynthGuard** achieved 88.5% recall, 8.0% false-positive rate, and 0.968 AUROC. On a verified 1,600-sequence out-of-distribution benchmark spanning seven unseen hazardous families, **SynthGuard** achieved 91.6% recall and 0.958 AUROC, while blastn again produced near-random discrimination. For protein screening, blastp at 50% identity achieved 0.0% recall on the translatable benchmark, while **SynthGuard** V4 achieved 86.0% recall and 0.944 AUROC. The hosted protein API uses adaptive model loading: it prefers the V4 ESM-2/k-mer model when available, can run ESM-2 on CPU when HuggingFace hosting does not provide a GPU, and falls back to the lighter V2 k-mer model only if the V4 artifact or embedding stack is unavailable. We integrated **SynthGuard** into **BioLens**, a nine-surface operator dashboard that converts screening outputs into auditable, intelligence-aware analyst workflows. The main takeaway is that AI-era biosecurity needs both stronger sequence discrimination and operational tooling: **SynthGuard** improves functional screening beyond identity search, while **BioLens** preserves raw model outputs, applies bounded intelligence-aware triage, and records analyst decisions for audit.

## 1 Introduction

Biological design tools are improving faster than the operational biosecurity systems used to screen DNA synthesis orders. Open biological foundation models, protein design tools, and codon-optimization pipelines make it easier to generate sequences that preserve function while changing surface sequence identity. At the same time, synthesis screening remains a critical intervention point: synthesis providers are one of the few places where potentially hazardous biological designs can be checked before physical construction [1, 4, 5].

Current screening workflows rely heavily on sequence identity. Nucleotide queries are aligned against known hazardous references, commonly using blastn-style thresholds, and protein sequences are screened using analogous amino-acid identity searches. This approach is useful for near-neighbor detection, but it is brittle when the threat model changes from exact sequence reuse to function-preserving redesign, codon optimization, or fragmented ordering [2, 3].

We address two linked problems. First, we ask whether a compact machine-learning screener can discriminate hazardous from benign sequences better than identity search under AI-era perturbations and unseen-family generalization. Second, we ask how such model outputs should be turned into analyst decisions. A hazard score alone is not an operational workflow: a practitioner still needs a queue, review state, intelligence context, audit history, escalation path, and report export.

Our submission combines two components:

1. **SynthGuard**, a sequence screener for DNA and protein synthesis workflows.
2. **BioLens**, an operator dashboard that wraps **SynthGuard** outputs in intelligence-aware triage, case management, audit logging, automation, analytics, and reporting.

The key scientific finding is that identity search fails differently across DNA and protein contexts. In our DNA

benchmarks, blastn did not primarily miss hazardous sequences; instead, it flagged almost everything, including benign sequences, producing an operationally unusable false-positive burden. In the protein benchmark, blastp at 50% identity had zero recall. **SynthGuard** improved discrimination in both settings and retained strong performance on seven unseen OOD hazardous families, while **BioLens** showed how those predictions can be converted into practical analyst workflow.

**Our main contributions are:**

1. A compact DNA screening model using 5,533 k-mer, codon-usage, codon-adaptation, and translated amino-acid features.
2. A protein screening model that combines amino-acid composition, dipeptides, physicochemical descriptors, and ESM-2 embeddings.
3. A hardware-aware protein API path that prefers the stronger V4 ESM-2/k-mer model, supports CPU execution when hosted GPU access is unavailable, and falls back to a lightweight V2 k-mer model only when required.
4. A real BLAST+ benchmark showing that blastn at 70% identity has high hazardous recall but poor discrimination on the DNA test set.
5. A verified out-of-distribution DNA benchmark across seven unseen hazardous families, showing that **SynthGuard** retains 91.6% recall and 0.958 AUROC without retraining.
6. A protein benchmark showing that blastp at 50% identity has zero recall on the evaluated translatable benchmark while **SynthGuard** V4 retains substantial recall.
7. **BioLens**, a nine-surface dashboard for intake, inbox, review, intelligence, automation, analytics, archive, reports, and audit.
8. An intelligence-aware triage mechanism that preserves raw model scores while computing bounded operational risk modifiers linked to active watchlists.

## 2 Related Work

**Synthesis screening.** The AIxBio synthesis-screening challenge asks for better DNA synthesis screening and synthesis-control tools, including systems that can detect dangerous sequences and address gaps caused by AI-designed proteins, short sequences, and current regulatory limitations [9]. Existing systems such as SecureDNA and the IBBIS Common Mechanism demonstrate important directions for open screening infrastructure. SecureDNA emphasizes privacy-preserving screening and detection of short sequences; IBBIS Common Mechanism provides an open HMM-based biorisk screener [4, 5].

**AI-era sequence redesign.** ProteinMPNN demonstrates that protein sequences can be redesigned while preserving structure and function [2]. Recent biosecurity work has shown that AI-designed variants can evade standard sequence-based screening, motivating defensive tools that detect functionally concerning sequences even when identity is low [3]. **SynthGuard** targets this gap by moving beyond identity-only matching toward function-aware and composition-aware screening.

**Operational biosecurity tooling.** Biosecurity practitioners need practical tools such as dashboards, rapid risk assessment systems, policy or threat-intelligence tools, and accessible workflows for under-resourced institutions [9]. **BioLens** fits this need as an operational layer rather than a new model. It turns a model output into a reviewable case with linked intelligence context, automation rules, and audit-ready reporting. Related practitioner needs are also reflected in broader biosecurity risk-assessment and policy resources [6, 7].

**Scope boundary.** This submission focuses on sequence screening and operator workflow. **SynthGuard** improves sequence-screening discrimination for DNA and protein synthesis workflows. **BioLens** turns model outputs into an operator-facing workflow for case review, intelligence-aware triage, automation, audit logging, analytics, and reports. **BioLens** includes intelligence context, but it is not a live pandemic early-warning system: the current implementation uses seeded demo alerts, manual alert creation, JSON import, and operator-curated watchlists rather than live ProMED, WHO DON, CDC/NWSS, wastewater, clinical, or news ingestion. We also do not claim a benchtop synthesizer security system; a split-order endpoint exists as a prototype, but a production split-order or device-security workflow was outside the completed submission scope.

## 3 Methods

### 3.1 System Overview

**SynthGuard** and **BioLens** are separated by a stable adapter contract. **SynthGuard** provides model-layer sequence screening. **BioLens** provides the practitioner-facing operational layer.

The adapter returns normalized fields such as `hazard_score`, `risk_level`, `confidence`, `category`, `explanation`, `baseline_result`, `model_name`,

DNA/protein sequence input → **SynthGuard** FastAPI or offline model route → stable adapter contract → dashboard heuristics → intelligence overlay → automation engine → SQLite store → Streamlit UI with Home, Screening, Inbox, Review, Analytics, Intelligence, Archive, Automation, and Reports

**Figure 1:** System architecture. **SynthGuard** provides model-backed screening endpoints and offline benchmarked models. **BioLens** provides the operator layer: adapter contract, dashboard-side heuristic overlays, intelligence-aware triage, automation, persistence, and analyst-facing workflow.

`threat_breakdown`, and optional attribution data. This boundary allowed backend changes during development without rewriting dashboard pages. Figure 1 summarizes the architecture.

### 3.2 Dataset Construction

The DNA dataset contains approximately 14,700 sequences across 37 organism families, with roughly balanced hazardous and benign labels. The split was 70% train, 15% validation, and 15% test, stratified by label and source family. The confirmed held-out DNA test set contains 2,214 sequences: 1,057 hazardous and 1,157 benign.

Hazardous examples cover three operational threat models:

- **Codon-shuffled variants:** synonymous codon substitutions applied to dangerous coding sequences at 25–45% codon substitution fraction.
- **Fragment-augmented sequences:** 50–300 bp windows simulating fragmented orders and short-sequence screening pressure.
- **Original sequences:** non-augmented hazardous sequences from public sequence resources.

Benign sequences were selected to reduce shortcut learning from GC content or organism background, including housekeeping proteins, *E. coli* metabolic genes, expression vectors, and sequences from diverse non-pathogenic organism families.

A separate targeted protein ablation dataset contains 255 ProteinMPNN structural redesigns across five toxin families: ricin, botulinum neurotoxin, anthrax lethal factor, abrin, and diphtheria toxin. This dataset was used for standalone protein ablation and is not part of the HuggingFace DNA/protein test split.

### 3.3 Out-of-Distribution Benchmark

To test generalization beyond the training families, we constructed a verified out-of-distribution DNA benchmark using seven hazardous families not present in training: tetanus toxin, *Francisella tularensis*, *Brucella abortus*, *Coxiella burnetii*, *C. difficile* toxin A, SARS-CoV-2 spike, and variola virus. The OOD evaluation set contained 1,600 sequences: 800 hazardous and 800 benign. Hazardous examples included 265 original and 535 codon-shuffled sequences; benign examples included 400 original and 400 codon-shuffled sequences. All OOD sequences were at least 150 bp, so the short-sequence specialist was not used.

**Table 1:** SynthGuard DNA feature vector. RSCU = relative synonymous codon usage; CAI = codon adaptation index.

| Feature group | Description | Count |
|---|---|---|
| Global statistics | Length, GC%, AT%, N-fraction, entropy, sequence complexity | 6 |
| k-mer frequencies | All k-mers for $k \in \{3, 4, 5, 6\}$ | 5,440 |
| RSCU | Relative synonymous codon usage | 64 |
| CAI | CAI vs. *E. coli*, human, yeast | 3 |
| AA composition | Best-frame amino-acid composition | 20 |
| **Total** | | **5,533** |

### 3.4 DNA Feature Engineering

**SynthGuard** DNA uses a 5,533-dimensional feature vector, summarized in Table 1. The key design choice is to include codon-usage and codon-adaptation features. Codon optimization changes synonymous codon choices without changing the encoded protein. Identity search sees only nucleotide-level similarity; codon-usage features can detect host-optimization and organism-of-origin fingerprints that remain meaningful after synonymous substitution.

### 3.5 Model Training

**DNA. SynthGuard** uses two calibrated LightGBM classifiers. The general DNA triage model handles sequences at or above 150 bp; a short-sequence specialist handles sequences below 150 bp and is trained on fragment-augmented examples. The general model uses 500 estimators, maximum depth 7, 63 leaves, learning rate 0.05, subsample 0.8, column subsample 0.8, balanced class weights, and sigmoid calibration. The short specialist uses 400 estimators, maximum depth 5, and 31 leaves. At inference time, routing is based on sequence length. Benchmark metrics use a fixed 0.5 decision threshold, while the live API maps scores into operational tiers: `SAFE` below 0.30, `REVIEW` from 0.30 to 0.60, and `ESCALATE` at or above 0.60. The combined DNA model is approximately 5.4 MB and runs in about 2 ms per sequence on CPU.

**Protein.** The protein benchmark compares blastp at 50% amino-acid identity against **SynthGuard** protein models. The strongest benchmarked protein model, V4, adds 480-dimensional ESM-2 embeddings to the 426-feature baseline of amino-acid composition, dipeptide frequencies, and physicochemical descriptors, producing a 906-dimensional feature vector. V4 was evaluated on the 1,286-sequence translatable test set and in a separate 255-sequence ProteinMPNN ablation.

The hosted protein API uses adaptive runtime model selection. At startup, it first attempts to load the V4 ESM-2/k-mer model. If the V4 artifact is available, the API loads ESM-2 on CUDA when a GPU is available, or on CPU when hosted without GPU support. This means a HuggingFace CPU-only deployment can still serve the V4 path, although with higher latency. If the V4 artifact, PyTorch, Transformers, or ESM-2 loading path fails, the API falls back to the lighter V2 k-mer model. The external `/protein/screen` response contract remains the same across tiers; the health endpoint reports the active protein state as `v4-esm2kmer`, `v2-kmer`, or `none`.

### 3.6 BLAST Baselines

The BLAST database was constructed from 741 original, non-augmented hazardous training sequences. This design reflects real synthesis screening: providers generally maintain databases of known hazardous references rather than precomputed codon-shuffled or AI-redesigned variants. The nucleotide baseline used blastn 2.12.0+ at 70% identity. The protein baseline used blastp 2.12.0+ at 50% amino-acid identity [8]. For short nucleotide sequences, blastn-short settings were used. The OOD benchmark used the same BLAST database as the main benchmark.

### 3.7 Design Attempts and Negative Results

During development, we tested simpler identity-proxy baselines before running the final BLAST+ benchmark. Early experiments used a k-mer Jaccard proxy against a small hazardous reference set. This proxy substantially underestimated the sensitivity of real blastn and gave the misleading impression that blastn primarily failed by missing codon-shuffled hazardous sequences. The final benchmark replaced this proxy with real NCBI BLAST+ 2.12.0+ against 741 original hazardous training sequences, which changed the interpretation: blastn detects most hazardous DNA variants but cannot discriminate them from benign sequences at the tested threshold.

The same issue appeared in the early OOD benchmark. An older OOD script used a k-mer identity proxy and a feature-mismatched model path, underestimating **SynthGuard** performance. The final OOD results in this paper use real blastn 2.12.0+, the correct 5,533-feature model, and the same hazardous BLAST database as the main benchmark.

We also separated protein model capability from hosted deployment constraints. The strongest benchmarked protein model is V4 with ESM-2 embeddings. The public API attempts to serve V4 whenever the model artifact and dependency stack are available, including on CPU-only HuggingFace hosting. The lighter V2 model remains as a resilience fallback because it is CPU-only and fast, but it is less robust on adversarial protein redesigns. We report this distinction explicitly to avoid overstating either public-demo latency or fallback capability.

### 3.8 BioLens Operator Workflow

**BioLens** is a nine-surface Streamlit dashboard, summarized in Table 2. It is the practitioner-facing layer around **SynthGuard**: **SynthGuard** produces screening outputs, while **BioLens** turns those outputs into case records, analyst decisions, intelligence-aware triage states, automation events, audit logs, analytics, and compliance-style reports.

The Intelligence surface is implemented as an operator-curated intelligence workspace. Analysts can use seeded demo alerts, manually create new alerts, import alerts from JSON, update alert status, promote alerts into active watchlists, and track watchlist effectiveness. The current system does not fetch live external intelligence feeds from ProMED, WHO DON, CDC/NWSS, wastewater surveillance, clinical feeds, or news APIs. This is why **BioLens** is framed as an operator workflow rather than an early-warning surveillance system.

**BioLens** persists cases, status changes, notes, alerts, watchlists, reports, and audit events in SQLite. SQLite is appropriate for a hackathon prototype and single-user demonstration, though not sufficient for production multi-analyst deployment.

### 3.9 Intelligence-Aware Operational Triage

**BioLens** separates raw model output from operational decision support. For each case, it stores `model_hazard_score`, `model_risk_level`, `intel_modifier`, `effective_hazard_score`, `effective_risk_level`, and the current dashboard score and tier. This design preserves the original **SynthGuard** prediction while allowing operator-curated intelligence context to influence workflow routing.

**Table 2:** BioLens dashboard surfaces. Together they implement screening workflow rather than merely displaying model predictions.

| Surface | Function |
|---|---|
| Home | System overview, KPIs, threat posture, recent activity |
| Screening | DNA/protein intake, FASTA upload, SynthGuard call |
| Inbox | Case queue, filtering, sorting, bulk actions |
| Review | Case detail, analyst notes, final action, audit trail |
| Analytics | Risk distribution, response time, alert statistics |
| Intelligence | Alerts, manual alert creation, JSON import, watchlists |
| Archive | Closed and cleared cases |
| Automation | Rules for auto-review or escalation |
| Reports | Markdown/JSON compliance-style exports |

**Table 3:** BioLens intelligence modifier policy. Modifiers are heuristic decision-support weights, not calibrated probability adjustments.

| Watchlist priority | Modifier |
|---|---|
| HIGH | +0.15 |
| MEDIUM | +0.08 |
| LOW | +0.03 |
| Total cap | +0.25 |

Operators can create or import intelligence alerts and promote relevant alerts to active watchlists. At screening time, **BioLens** matches cases against active watchlists using fields such as category, explanation, and sequence type. Watchlist keywords are expanded through a curated taxonomy map; for example, "hemorrhagic" expands to related terms such as filovirus, arenavirus, Ebola, Marburg, Lassa, and hemorrhage.

Matched watchlists apply bounded score modifiers, shown in Table 3. Automation rules can then act on watchlist priority or alert severity. For example, high-priority matches can move a case to `ESCALATE`, while medium-priority matches can route it to `IN_REVIEW`. Rule firings are logged as audit events.

The intelligence modifier is a deterministic workflow heuristic, not a calibrated epidemiological or probabilistic signal. It should be interpreted as operator-contextualized decision support: current threat context can change review priority, but the raw model score remains visible and auditable.

## 4 Results

### 4.1 DNA Screening: Identity Search Has High Recall but Poor Discrimination

On the full 2,214-sequence DNA test set, blastn at 70% identity achieved 99.4% recall but 94.3% false-positive rate, with AUROC 0.526. **SynthGuard** achieved 88.5% recall, 8.0% false-positive rate, and AUROC 0.968, as shown in Table 4.

This result changes the framing of the DNA screening problem. In our real blastn benchmark, identity search did not primarily fail by missing codon-shuffled hazardous sequences. At 35% codon shuffling, sequences still retain roughly 88% nucleotide identity and remain detectable at a 70% blastn threshold. The failure mode is non-discrimination: blastn also flags the overwhelming majority of benign sequences. At synthesis-provider scale, this would create an infeasible analyst review burden.

### 4.2 Out-of-Distribution Generalization

We next evaluated **SynthGuard** on seven hazardous families absent from training, using real blastn 2.12.0+ and the same 741-sequence hazardous BLAST database. The full OOD set contained 1,600 sequences: 800 hazardous and 800 benign. On this benchmark, blastn again achieved high recall but poor discrimination: 99.4% recall, 96.5% FPR, and 0.514 AUROC. **SynthGuard** achieved 91.6% recall, 12.2% FPR, and 0.958 AUROC (Table 5).

Per-family recall on original OOD hazardous sequences was 100% for tetanus toxin, *Francisella tularensis*, *Coxiella burnetii*, *C. difficile* toxin A, SARS-CoV-2 spike, and variola virus. The only observed gap was *Brucella abortus*, where **SynthGuard** achieved 85% recall on 20 original sequences. We interpret this as a data-distribution gap: *Brucella* has an unusual high-GC, intracellularly adapted codon-usage profile that diverges from the hazardous training distribution.

### 4.3 Protein Screening: Identity Search Misses Low-Identity Redesigns

For protein screening, blastp at 50% identity achieved 0.0% recall and AUROC 0.500 on the translatable benchmark, while **SynthGuard** protein V4 achieved 86.0% recall, 13.2% false-positive rate, and AUROC 0.944, as shown in Table 6. This is the cleaner low-identity redesign finding: when amino-acid identity falls below the threshold used by the baseline, identity search becomes blind. **SynthGuard** recovers discriminatory signal from amino-acid composition, dipeptide patterns, physicochemical features, and ESM-2 embeddings.

### 4.4 ProteinMPNN Ablation

The separate 255-sequence ProteinMPNN ablation evaluated whether adding structural embedding features and targeted synthetic examples improves coverage across toxin folds. Recall rose from 34.5% in V1 to 100.0% in V4, as shown in Table 7. This ablation is a

**Table 4:** DNA benchmark using real blastn 2.12.0+ at 70% identity. FPR = false-positive rate.

| Slice | n | Haz | Ben | blastn Rec. | blastn FPR | blastn AUROC | SG Rec. | SG FPR | SG AUROC |
|---|---|---|---|---|---|---|---|---|---|
| Full test set | 2,214 | 1,057 | 1,157 | 99.4% | 94.3% | 0.526 | 88.5% | 8.0% | 0.968 |
| Original sequences | 532 | 152 | 380 | 98.7% | 97.4% | 0.507 | 96.1% | 1.1% | 0.989 |
| Fragment <150 bp | 555 | 236 | 319 | 99.6% | 98.7% | 0.504 | 76.3% | 16.6% | 0.878 |
| Any augmented variant | 1,682 | 905 | 777 | 99.6% | 92.8% | 0.534 | 87.2% | 11.3% | 0.954 |

**Table 5:** Verified OOD benchmark using real blastn 2.12.0+. The seven hazardous families were absent from training.

| Slice | n | blastn Rec. | blastn FPR | blastn AUROC | SG Rec. | SG FPR | SG AUROC |
|---|---|---|---|---|---|---|---|
| Full OOD set | 1,600 | 99.4% | 96.5% | 0.514 | 91.6% | 12.2% | 0.958 |
| Original sequences | 665 | 100.0% | 98.0% | 0.510 | 98.9% | 4.8% | 0.986 |
| Codon-shuffled variants | 935 | 99.1% | 95.0% | 0.520 | 88.0% | 19.8% | 0.924 |

**Table 6:** Protein benchmark using real blastp 2.12.0+ at 50% identity.

| Metric | blastp | SynthGuard V4 |
|---|---|---|
| Test set size | 1,286 | 1,286 |
| Recall | 0.0% | 86.0% |
| FPR | 0.0% | 13.2% |
| AUROC | 0.500 | 0.944 |

**Table 7:** Standalone ProteinMPNN ablation on 255 structural redesigns. This dataset is separate from the main HuggingFace test split.

| Model | Added signal | Recall |
|---|---|---|
| Identity proxy | k=7 Jaccard | 24.7% |
| V1 | 426 protein features | 34.5% |
| V2 | + ProteinMPNN data | 52.9% |
| V3 | + ESM-2 embeddings | 79.2% |
| V4 | + targeted 1BC7 data | 100.0% |

Identity-proxy recall is inflated by one coincidental family; excluding it, recall is approximately 6%.

targeted stress test, not the same distribution as the main 1,286-sequence translatable benchmark.

Two diagnostic trajectories are important. First, botulinum neurotoxin variants were never in training, yet recall rose from 2% in V1 to 96% in V3 and 100% in V4, suggesting that structural embeddings improved fold-level transfer. Second, diphtheria toxin recall was 0% for V1–V3 and rose to 100% after adding 50 ProteinMPNN variants of the 1BC7 fold, without architectural change. This suggests that targeted computational augmentation can close coverage gaps for underrepresented toxin folds.

### 4.5 BioLens Demonstration Results

**BioLens** demonstrates that screening results can be operationalized rather than merely displayed. The dashboard supports case creation, queueing, analyst review, operator-curated intelligence alerts, active watchlists, intelligence-adjusted scoring, automation rules, audit logging, and report export.

A representative triage path is:

> Raw **SynthGuard** result: score 0.25, tier `SAFE`
> Matched intelligence: high-priority toxin watchlist
> **BioLens** modifier: +0.15
> Effective operational result: score 0.40, tier `REVIEW`
> Audit trail: raw score, modifier, linked alert, status transition

The design explicitly preserves model transparency. Analysts can see the raw model prediction, the matched watchlist context, the bounded modifier, and the reason a case was operationally escalated. This matters because current threat context can affect triage priority without hiding the underlying model output.

The intelligence module should not be interpreted as live outbreak detection. In the current prototype, alerts are seeded, manually created, or imported from JSON by an operator. **BioLens** therefore demonstrates intelligence-aware triage inside an operator dashboard, not automated public-health surveillance ingestion.

### 4.6 Robustness and Interpretation of Claims

The strongest claims in this report are based on real BLAST+ benchmarks rather than earlier proxy baselines. The main DNA benchmark used a held-out test set of 2,214 sequences with 1,057 hazardous and 1,157 benign examples, and the BLAST database was built only from original hazardous training sequences rather

than augmented test variants. The OOD benchmark used seven hazardous families absent from training, 1,600 total sequences, and the same BLAST database. This setup avoids directly giving the identity baseline access to codon-shuffled or fragment-augmented test sequences.

We interpret AUROC as the main threshold-independent discrimination metric. This is important because synthesis-screening operators may tune thresholds depending on institutional risk tolerance. Under this framing, the DNA result is not merely that **SynthGuard** improves one fixed operating point, but that blastn at the evaluated identity-search setup provides near-random discrimination, while **SynthGuard** provides substantially stronger separation between hazardous and benign examples.

The protein benchmark is robust in a different sense: blastp at 50% amino-acid identity produced no positive detections on the translatable benchmark, while **SynthGuard** recovered substantial recall. However, the targeted 255-sequence ProteinMPNN ablation should be interpreted as a separate stress test rather than as the same distribution as the main held-out benchmark. The adaptive protein deployment path should also be interpreted separately from the benchmark: it is an availability and latency-management feature for hosted inference, not a new biological validation result.

## 5 Discussion and Limitations

**SynthGuard** and **BioLens** address complementary failure modes. **SynthGuard** improves the technical screening layer by using features that are not reducible to local sequence identity. **BioLens** improves the operational layer by converting model outputs into auditable, intelligence-aware decisions.

The DNA benchmark highlights an underappreciated issue: identity search can appear safe if judged only by hazardous recall, but fail in deployment because it lacks specificity. A 94.3% false-positive rate on the main benchmark, and 96.5% on the OOD benchmark, would overwhelm manual review, especially for high-volume synthesis providers. **SynthGuard** reduces this burden while preserving substantial recall.

The protein benchmark shows the opposite identity-search failure mode: under low amino-acid identity, blastp misses the relevant redesigned sequences entirely. This supports the need for protein-level function-aware or structure-aware features in synthesis screening, especially as generative design systems become more accessible.

**BioLens** adds a practical workflow around these models. In real operations, screening is not a single prediction. It is an end-to-end process: intake, queueing, review, intelligence context, escalation, notes, audit, and reporting. **BioLens** demonstrates one way to build this workflow while preserving transparency between raw model output and contextual operational triage.

### 5.1 Main Limitations and Assumptions

**Short-sequence performance gap.** Performance degrades for sequences shorter than 150 bp: the short-sequence specialist achieved 76.3% recall and 16.6% FPR, below the full-test performance of 88.5% recall and 8.0% FPR. At this length, k-mer frequencies are less stable and codon-usage fingerprints are harder to estimate reliably.

**Chimeric constructs.** The model was not evaluated on chimeric sequences in which a hazardous functional domain is embedded within a benign coding scaffold. Such constructs could partially mask hazardous signal if the benign scaffold dominates codon-usage or k-mer features.

**De-novo backbone generation.** Evaluation was limited to ProteinMPNN redesigns of known toxin structures. De-novo backbone generation, including RFdiffusion-style workflows, represents a threat vector beyond this evaluation's scope because generated proteins may share little sequence or structural similarity with known toxins.

**OOD codon-bias gap.** Among the seven OOD hazardous families, *Brucella abortus* was the exception at 85% recall on original sequences; all other OOD families achieved 100% recall. This suggests that unusual intracellular codon-adaptation profiles can produce coverage gaps and should be addressed through targeted data augmentation.

**Computational-only validation.** All results are computational. No wet-lab validation of hazard function was performed, consistent with the role of synthesis screening as a risk-stratification gate rather than a biological confirmation assay.

**Adaptive protein deployment limitation.** The protein API includes runtime fallback logic to keep the public endpoint available across different hosted environments. The preferred path is V4, which combines 426 baseline protein features with 480 ESM-2 embedding features. On GPU-backed hosting, V4 can run ESM-2 on CUDA with substantially lower latency. On HuggingFace CPU-only hosting, V4 can still run by loading ESM-2 on CPU, but latency is higher. If the V4 model artifact is missing, or if PyTorch, Transformers, or ESM-2 fails to load, the API falls back to the lighter

V2 k-mer model. This improves reliability for a public demo, but fallback has a real capability cost: V2 is faster and CPU-only, but it is less robust on adversarial protein redesigns than V4.

**BioLens intelligence limitation.** **BioLens** currently uses demo/operator-curated intelligence rather than live external intelligence ingestion. It supports seeded alerts, manual alert creation, JSON import, watchlist promotion, and watchlist matching, but it does not fetch live ProMED, WHO DON, CDC/NWSS, wastewater, clinical, or news feeds. It also does not perform public-health anomaly detection or forecasting. Accordingly, **BioLens** should be interpreted as an operator workflow with early-warning-inspired context, not as a pandemic early-warning system.

**BioLens operational prototype limitations.** **BioLens** uses SQLite persistence and Streamlit session-state role controls, which are suitable for a hackathon demonstration but not production-grade multi-analyst operations. A production version would require PostgreSQL or another robust database, real authentication and authorization, rate limiting, monitoring, backup, stronger access control, and tamper-evident audit guarantees. Push notifications and supervisor notification are also better interpreted as workflow flags/logged events unless real email or messaging integration is added.

## 6 Future Work

Future work should prioritize external evaluation against provider-scale order distributions and independent benign corpora; additional OOD evaluation across more Tier 1 Select Agent families; fragment-aware architectures for sequences below 150 bp; evaluation against chimeric constructs, domain swaps, RFdiffusion-generated de-novo backbones, and adversarially optimized sequences; calibration of **SynthGuard** thresholds against operational review costs and institutional risk tolerance; and production hardening of **BioLens**.

For the protein API, future work should harden adaptive model selection with clearer startup diagnostics, explicit active-tier reporting in model cards or deployment metadata, cached embedding paths where appropriate, and GPU-backed hosting for lower-latency V4 inference. The CPU V4 pathway should remain supported for accessibility, while V2 should remain a transparent fallback rather than being treated as equivalent to the strongest benchmarked model.

For **BioLens**, future work should add governed live intelligence ingestion from trusted public-health and policy sources, such as ProMED, WHO DON, CDC/NWSS, wastewater surveillance, and curated policy feeds. This should be implemented with source attribution, deduplication, analyst review, rate limiting, and human approval before external signals affect case routing. Additional production work should include real authentication, role-based authorization, PostgreSQL-backed persistence, tamper-evident audit logs, and optional notification integrations.

## 7 Conclusion

**SynthGuard** shows that lightweight ML features can substantially improve discrimination over identity-only screening for AI-era synthesis risks. The strongest DNA result is not that blastn misses all hazardous variants, but that it flags nearly all benign sequences as well, producing poor discrimination. This result persists on a verified OOD benchmark spanning seven unseen hazardous families. The strongest protein result is that blastp at 50% identity has zero recall under the evaluated low-identity setting, while **SynthGuard** recovers substantial signal.

The adaptive protein pathway improves deployment practicality. It allows the same public API endpoint to prefer the stronger V4 ESM-2/k-mer model, run that model on CPU-only HuggingFace hosting when GPU access is unavailable, and fall back to the lightweight V2 model only when the V4 artifact or embedding stack cannot load. This preserves demo availability without hiding the capability difference between the tiers.

**BioLens** extends the project from a model into a practitioner tool. It provides a dashboard, case queue, review workflow, operator-curated intelligence alerts, watchlists, intelligence-aware triage, automation, audit trail, analytics, and reports. Together, **SynthGuard** and **BioLens** form a defensive prototype for discriminative synthesis screening and auditable biosecurity triage.

## 8 Code and Data

The project includes public code, model artifacts, dataset links, and live demonstrations. Table 8 lists the public project artifacts.

**Table 8:** Public project artifacts.

| Artifact | Link |
|---|---|
| Code repository | `https://github.com/Ashok-kumar290/synthscreen` |
| SynthGuard API | `https://seyomi-synthguard-api.hf.space` |
| BioLens dashboard | `https://seyomi-biolens-dashboard.hf.space` |
| DNA/protein k-mer models | `https://huggingface.co/Seyomi/synthguard-kmer` |
| ESM-2 protein model | `https://huggingface.co/Seyomi/synthguard-esm2` |
| Dataset | `https://huggingface.co/datasets/Seyomi/synthscreen-dataset` |

The code repository contains the implementation, with the main submission branch named `synthguard`. The model repositories contain the DNA general model, DNA short-sequence specialist, protein V2 fallback model, benchmarked protein V4 ESM-2 model, and associated metadata. The protein API uses adaptive startup logic to choose the strongest available protein tier while preserving a stable `/protein/screen` endpoint. The dataset repository contains the curated sequence corpus used for training and evaluation. The main DNA, OOD DNA, and protein benchmark results reported here were generated with real NCBI BLAST+ 2.12.0+ on Google Colab A100 on April 26, 2026.

We intentionally do not include hazardous example sequences in the report text. Public artifacts should be handled with appropriate biosecurity caution, and any future release of higher-risk evaluation cases should use controlled access or responsible disclosure channels.

## 9 Author Contributions

Ashok worked on model training, benchmark evaluation, adaptive protein API integration. Akhil worked on the dashboard, Streamlit workflow, repository cleanup, and biosecurity operations integration. Both authors contributed to project design, testing, result verification, final drafting, and submission preparation.

## References

[1] Office of Science and Technology Policy. Framework for Nucleic Acid Synthesis Screening. 2024.

[2] J. Dauparas et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 2022.

[3] B. J. Wittmann, T. Alexanian, C. Bartling, J. Beal, A. Clore, J. Diggans, K. Flyangolts, B. T. Gemler, T. Mitchell, S. T. Murphy, N. E. Wheeler, and E. Horvitz. Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science*, 390(6768):82–87, 2025. doi:10.1126/science.adu8578.

[4] SecureDNA. Open-source privacy-preserving sequence screening documentation.

[5] International Biosecurity and Biosafety Initiative for Science. Common Mechanism documentation: open-source synthesis-screening software.

[6] T. Webster, R. Moulange, B. Del Castello, J. Walker, S. Zakaria, and C. Nelson. Global Risk Index for AI-enabled Biological Tools: Summary Assessment and Methods Report. RAND Europe and Centre for Long-Term Resilience, 2025.

[7] J. S. Morrison and M. Simoneau. Eight Commonsense Actions on Biosafety and Biosecurity. Center for Strategic and International Studies, 2023.

[8] T. Madden. BLAST Command Line Applications User Manual. National Center for Biotechnology Information, 2008–.

[9] Apart Research. AIxBio Hackathon 2026 submission guidelines. 2026.

## A Limitations and Dual-Use Considerations

### A.1 Technical Limitations

**SynthGuard** is a research prototype. It has not undergone production validation, external red-teaming, or wet-lab validation. The dataset is synthetic and curated; performance on real provider order streams may differ. The short-sequence specialist still has a 16.6% false-positive rate on fragments below 150 bp, which may be too high for production-scale synthesis screening without additional triage layers.

The model was not evaluated on chimeric domain-fusion constructs, RFdiffusion-generated de-novo backbones, domain-swapped proteins, or fully adversarially optimized sequences. These cases are important because future design systems may intentionally optimize against screening features.

The OOD benchmark covers seven unseen hazardous families, not all Tier 1 Select Agent families. The *Brucella abortus* gap suggests that unusual organism-specific codon-usage profiles can reduce recall and should be addressed through targeted augmentation.

The protein API uses adaptive model selection rather than a single fixed hosted model. The preferred tier is V4, which combines 426 baseline protein features

with 480 ESM-2 embedding features and achieved the strongest reported protein-screening results. On GPU-backed hosting, V4 can use CUDA acceleration. On HuggingFace CPU-only hosting, V4 can still run by loading ESM-2 on CPU, but inference is slower. If the V4 artifact is missing or the ESM-2 dependency path fails, the API falls back to the lighter V2 k-mer model. This fallback improves robustness and public-demo availability, but it is not capability-equivalent to V4: the overall AUROC difference is small, but V2 is substantially weaker on the ProteinMPNN redesign stress test and should be treated as a fallback tier, not the main scientific result.

**BioLens** is an operator dashboard, not a live pandemic early-warning platform. Its intelligence module currently uses seeded demo alerts, manual alert creation, JSON import, and operator-curated watchlists. It does not fetch live ProMED, WHO DON, CDC/NWSS, wastewater, clinical, or news feeds; it does not perform automated external polling; and it does not implement anomaly detection or forecasting over surveillance feeds.

**BioLens** also has prototype infrastructure limitations. It uses SQLite and Streamlit session-state role controls, which are appropriate for a hackathon prototype but not production-grade multi-analyst operations. A production version would require PostgreSQL or another robust database, real authentication, authorization, rate limiting, monitoring, backup, push-notification or messaging integrations if desired, and tamper-evident audit guarantees.

### A.2 Dual-Use Risks

A public screening model can itself create dual-use risk. Adversaries could probe model boundaries, search for sequences that score below thresholds, or use explanations to optimize evasion. The same features that improve defensive discrimination could reveal which sequence properties are being monitored.

We reduce this risk in the report by presenting aggregate metrics and system architecture rather than hazardous sequences or evasion recipes. We do not publish dangerous example sequences in the paper. Production deployments should avoid unrestricted anonymous screening APIs, should monitor for probing behavior, and should use rate limits, access controls, and responsible logging.

**BioLens** also introduces operational risk if intelligence modifiers are treated as calibrated probabilities or automated public-health signals. They are not. The current modifiers are heuristic decision-support weights based on operator-curated watchlists, and they should be reviewed by human analysts. **BioLens** should support, not replace, analyst judgment.

### A.3 Responsible Disclosure and Governance

If screening weaknesses are discovered through future evaluation, they should be disclosed to relevant screening infrastructure maintainers, synthesis providers, and biosecurity governance organizations before public release. Vulnerability examples should be shared in controlled form, with enough detail for defensive patching but without enabling misuse.

Deployments should use a managed-access model for high-risk capabilities, maintain audit logs, and define escalation protocols for concerning submissions. Public demos should either use synthetic benign examples or restrict outputs to high-level risk tiers.

### A.4 LLM Usage Statement

We used Claude Code to assist with code development and debugging. We used ChatGPT to assist with literature review, wording and title refinement, drafting portions of the report, proofreading, and converting the final writeup into LaTeX format. Core research direction, experimental design decisions, benchmark execution, model outputs, and final claims were verified by the authors against the underlying evaluation artifacts.

### A.5 Ethical Considerations

The goal of this project is defensive: reducing the risk that AI-generated or modified biological sequences evade screening while preserving legitimate research access. False positives can delay beneficial science, while false negatives can create serious security risks. The appropriate deployment posture is therefore not full automation, but human-in-the-loop triage with transparent model outputs, contextual intelligence, and appealable review.