BioShield AI

Next-Generation AI-Powered Biosecurity Screening System

AI Bio Hackathon Submission

Domain	Type	Stack
AI Safety & Biosecurity	Red Teaming + Detection	ESM-2, ESMFold, XGBoost

1. The Problem

1.1 Current DNA Synthesis Screening Landscape

Current DNA synthesis screening tools — including SecureDNA and IBBIS — operate by comparing submitted sequences against databases of known dangerous sequences. If a submitted sequence looks similar to a known toxin or pathogen-related gene, it gets flagged and blocked.

This approach relies almost entirely on homology: if the letters of a new sequence closely match the letters of a known dangerous sequence, the system catches it.

1.2 The Critical Gap

> ⚠ **The Core Vulnerability**
> AI protein design tools — specifically ProteinMPNN and RFdiffusion — can now generate brand-new
> amino acid sequences that look completely different on paper but fold into the same 3D structure and perform the exact same dangerous biological function.
>
> These functionally equivalent variants slip through homology-based screening entirely.

This is not a theoretical risk. Researchers have already demonstrated that:

- A toxin's function is determined by its 3D folded structure, not its sequence letters
- Two proteins can share less than 20% sequence identity yet fold identically
- AI tools make it trivially easy to generate such evasion variants at scale

- No current commercial screening tool accounts for this attack surface

1.3 Why This Matters Now

Before AI Protein Design	After AI Protein Design
Dangerous sequences were mostly derivatives of known sequences. Homology screening was sufficient.	Novel sequences with no known homologs can now be rapidly generated to bypass all existing screening tools.

2. Our Solution — BioShield AI

2.1 Core Philosophy

Instead of comparing sequence letters (homology), BioShield AI compares biological meaning (function). We use protein language models to convert any sequence into a high-dimensional vector that encodes what the protein actually does — regardless of what the letters look like.

Two sequences that perform the same dangerous function will have similar biological fingerprints, even if their letter sequences are completely different.

2.2 The Full 5-Station Pipeline

Station	Name	What It Does	Technology
1	Functional Fingerprinting	Feed sequence into ESM-2 + ESMFold ensemble to get a 1280-dim biological embedding + 3D structure prediction	ESM-2 (650M), ESMFold, ProtTrans
2	Domain Risk Check	Cross-reference against Pfam domain DB + GO term mapping + NCBI Pathogen DB + UniProt reviewed toxins	Pfam, Gene Ontology, NCBI, UniProt
3	Pathway Assembly Check	Detect if multiple individually harmless sequences together reconstitute a dangerous biological pathway	Graph-based pathway analysis
4	Risk Scoring + Explainability	Trained XGBoost classifier outputs risk score + confidence interval + attention-based amino acid highlighting	XGBoost, ESM attention maps
5	Adversarial Red Team Loop	Auto-generate evasion attempts using ProteinMPNN/RFdiffusion and continuously harden the screener	ProteinMPNN, RFdiffusion, active learning

3. Technical Deep Dive

3.1 Station 1 — Functional Fingerprinting

ESM-2 Embeddings

ESM-2 (650M parameters, trained on 250M proteins) converts any amino acid sequence into a 1280-dimensional embedding vector. This vector captures biochemical properties, evolutionary information, and functional signatures of the protein.

Unlike raw sequence comparison, proteins with identical functions will cluster together in this embedding space — even if they share zero sequence identity.

ESMFold Structure Prediction

ESMFold predicts the full 3D atomic structure of the protein in seconds. Two proteins that fold into the same shape will have very similar structural fingerprints (TM-score > 0.7), even if their sequences are completely unrelated.

This is critical: a newly designed toxin variant will fold like the original toxin, so it will be caught at the structural level even if it evades the embedding check.

Ensemble Approach

Why an Ensemble?

- ESM-2 alone may miss proteins outside its training distribution
- ESMFold alone is computationally expensive at scale
- ProtTrans adds a complementary trained representation
- Combining all 3 signals dramatically reduces false negatives
- Any single model's blind spots are covered by the others

3.2 Station 2 — Domain Risk Check

Pfam Domain Analysis

The Pfam database contains profiles for over 19,000 protein families. Every submitted sequence is scanned for the presence of known dangerous functional domains — toxin domains, protease active sites, membrane-disrupting motifs, etc.

Gene Ontology (GO) Term Mapping

GO terms describe what a protein does (molecular function), where it acts (cellular component), and what process it participates in (biological process). Even if a protein has no Pfam match, its predicted GO terms can flag dangerous biological functions.

External Database Cross-Reference

- NCBI Pathogen Reference Sequences: Flags sequences related to known pathogens
- UniProt Reviewed Toxins: Cross-references the curated Swiss-Prot toxin annotations
- CARD (Comprehensive Antibiotic Resistance DB): Detects antimicrobial resistance genes

3.3 Station 3 — Pathway Assembly Check

The Split Submission Attack

An attacker could split a dangerous protein into 3-4 harmless-looking fragments, submit them separately, and assemble them in the lab. Each fragment alone looks benign. Station 3 detects this by analyzing all sequences submitted in a batch together.

Station 3 builds a graph where nodes are submitted sequences and edges represent functional complementarity. It then checks whether any connected subgraph corresponds to a known dangerous biological pathway (toxin synthesis, viral replication machinery, etc.).

3.4 Station 4 — Risk Scoring & Explainability

Trained Classifier

An XGBoost model is trained on embeddings of known dangerous vs. benign proteins. It takes the combined signals from Stations 1-3 as features and outputs a risk score between 0 and 1, along with a calibrated confidence interval.

Explainability — Why Was This Flagged?

Biosecurity analysts need to understand why something was flagged, not just that it was. BioShield AI uses ESM-2 attention maps to highlight the specific amino acid regions most responsible for the risk score. This enables human experts to make informed final decisions.

Mutation Proximity Score

For any flagged sequence, BioShield AI also calculates how many point mutations away it is from the nearest known dangerous protein. A sequence 2 mutations from a known toxin is a very different risk than one that is 50 mutations away.

3.5 Station 5 — Adversarial Red Team Loop

The Killer Feature: Self-Hardening

BioShield AI doesn't just detect — it continuously attacks itself.

Using ProteinMPNN and RFdiffusion, the system generates evasion variants of known dangerous proteins and tests whether they slip through the screener.
Any that evade detection are added to the training set, making the system progressively harder to fool over time.

This is active adversarial robustness — a concept from AI red teaming applied directly to biosecurity.

This loop runs automatically on a schedule and can be triggered manually. It produces a continuously updated evasion resistance score that quantifies how hard the current screener is to bypass.

4. New Features Beyond Original Design

4.1 Mutation Trajectory Tracking

Given a flagged sequence, BioShield AI maps its evolutionary distance from all known dangerous proteins in our reference set. This produces a proximity landscape showing which dangerous proteins are nearby in sequence space, and what mutational paths could reach them.

This is actionable intelligence for biosecurity policy — it distinguishes between a sequence that accidentally resembles a toxin vs. one that is clearly converging toward it.

4.2 REST API + Audit Trail

BioShield AI ships as a fully-documented REST API, making it easy for DNA synthesis companies (Twist, IDT, Genscript) to integrate directly into their order processing pipelines.

Every query is logged with a cryptographic hash of the submitted sequence, timestamp, and risk score. This audit trail is essential for regulatory compliance and forensic analysis in the event of a biosecurity incident.

4.3 Batch & Real-Time Modes

Real-Time Mode	Batch Mode
Single sequence screened in < 5 seconds using pre-computed embedding indexes for Stations 1-2. Suitable for point-of-order screening.	Full pipeline including pathway assembly check and structure prediction. For order batches submitted together. Includes cross-sequence analysis.

4.4 Risk Dashboard

A web-based analyst dashboard visualizes the embedding space of submitted sequences in 2D (UMAP projection), showing dangerous clusters and the position of flagged sequences relative to known threats. Analysts can drill into any flagged sequence to see the full explainability report.

5. AI Safety Significance

5.1 Why This Is an AI Safety Problem

BioShield AI exists precisely because AI has created a new attack surface. Before ProteinMPNN and RFdiffusion, designing novel functional proteins required years of expertise. Now it takes minutes. The very AI tools that advance medicine have simultaneously lowered the barrier for misuse.

This project applies AI safety methodology — adversarial red teaming, robustness evaluation, and continuous monitoring — to biosecurity. It is a direct translation of AI alignment work into a critical real-world domain.

5.2 Connection to AI Red Teaming

AI Red Teaming Concept	How BioShield AI Uses It
Adversarial examples	Evasion variants generated by ProteinMPNN
Model robustness evaluation	Continuous self-hardening red team loop
Out-of-distribution detection	Flagging sequences outside safe embedding clusters
Confidence calibration	Risk scores with calibrated confidence intervals
Explainability	Attention-based amino acid highlighting

6. Validation Strategy

6.1 Ground Truth Dataset

Training and evaluation uses a curated dataset of:

- ~2,400 known protein toxins from UniProt Swiss-Prot (reviewed, manually annotated)
- ~500 Select Agent proteins from the CDC/USDA Select Agent list
- ~10,000 benign proteins sampled uniformly from UniProt to establish the safe baseline
- ~300 synthetic evasion variants generated by our own red team loop (held-out test set)

6.2 Key Metrics

Metric	Target	Why It Matters
False Negative Rate	< 2%	Missing a true danger is catastrophic
False Positive Rate	< 15%	Too many false alarms = system ignored
Evasion Variant Detection	> 85%	Core novelty of BioShield AI
Real-time latency	< 5 sec	Practical for production use
Confidence calibration	ECE < 0.05	Risk scores must be trustworthy

7. Technology Stack

Component	Technology	Purpose
Protein Embeddings	ESM-2 (Meta), ProtTrans	Functional fingerprinting
Structure Prediction	ESMFold (Meta)	3D structure-based detection
Protein Design (Red Team)	ProteinMPNN, RFdiffusion	Generating evasion variants
Domain Scanning	Pfam, HMMER	Known domain detection
Functional Annotation	Gene Ontology (GO)	Function-level risk flags
Classifier	XGBoost + calibration	Final risk score output
Pathway Analysis	NetworkX graph lib	Multi-sequence pathway check
API	FastAPI (Python)	Production integration
Dashboard	React + Plotly UMAP	Analyst visualization
Databases	NCBI, UniProt, CARD	Reference threat data

8. BioShield AI vs. Existing Tools

Feature	SecureDNA	IBBIS	BioShield AI	
Detects known sequences	✅ Yes	✅ Yes	✅ Yes	
Detects functional analogs	❌ No	❌ No	✅ Yes	
3D structure-based detection	❌ No	❌ No	✅ Yes	
Multi-sequence pathway check	❌ No	❌ No	✅ Yes	
Explainability / Why flagged	❌ No	❌ No	✅ Yes	
Adversarial self-hardening	❌ No	❌ No	✅ Yes	
Confidence intervals	❌ No	❌ No	✅ Yes	
API for integration	✅ Yes	Partial	✅ Yes	

9. Roadmap

Phase 1 — Hackathon MVP

Stations 1-2 functional. ESM-2 embeddings + Pfam scanning. Basic risk scoring. Proof-of-

Phase 2 — Full Pipeline

All 5 stations active. Full validation on curated dataset. API deployment. Analyst dashboard.

concept evasion detection on synthetic test set.	Partnership outreach to DNA synthesis companies.

Phase 3 — Production & Policy

- Integration with major DNA synthesis order pipelines (Twist, IDT)
- Continuous adversarial red team loop running in production
- Publishing results as peer-reviewed biosecurity research
- Policy recommendations for AI-enabled biosecurity screening standards
- Open-source release of non-sensitive components

10. Why Our Team

Our team sits at the precise intersection of the two skills this problem demands: biological domain knowledge and AI safety methodology.

- Biotechnology/Biochemistry background — we understand protein structure, function, and the threat landscape from first principles
- AI red teaming expertise — we apply adversarial ML concepts that most pure biology teams would never consider
- This combination is exceptionally rare and is exactly what this problem requires

Our Unique Advantage

Most teams entering this space come from either pure biology or pure CS.
We are neither. We are AI safety researchers with biological intuition.
BioShield AI could only be designed by someone who understands
both how proteins work AND how AI models fail.

BioShield AI — Protecting the World from AI-Enabled Biological Threats