Biosecurity Threat Detection via Interactome Disruption Analysis

Technical Analysis & Operational Report

1. Summary

This project introduces a next-generation biosecurity screening pipeline designed to detect AI-engineered biological threats that evade traditional screening methods. By analyzing the structural topology of protein interactions rather than relying on sequence homology, we provide a robust defense mechanism against novel, zero-shot adversarial proteins designed to hijack human immune pathways.

2. Vulnerabilities in Conventional Biosecurity Frameworks

Standard biosecurity pipelines—the tools meant to screen commercial DNA synthesis orders—are fundamentally broken against AI-engineered threats. They rely on "V1" sequence homology screens (such as BLAST) that simply ask, **does this sequence look like a known pathogen?**

This presents a massive vulnerability. Modern generative AI models can easily be prompted to produce adversarial, *de novo* sequences that have exactly 0% homology to any known hazard, yet still fold into functionally toxic structures. By the time a DNA synthesis provider runs a sequence check, the generative model has already scrubbed the threat of its identifiable signature. We are currently flying blind against structurally disguised biological threats—this project directly addresses and closes that critical gap.

3. Technical Differentiators: Structural Topology Analysis

While current State-of-the-Art (SOTA) relies on primary sequence similarity, we are implementing a **V2 approach**: structural topology.

Rather than asking if a sequence *looks* dangerous, we ask if it *behaves* dangerously. We achieve this by simulating the insertion of the novel sequence into the human immune protein-protein interaction (PPI) network. An adversary can easily scrub a sequence's text to evade a BLAST search—they cannot, however, scrub the physical binding surface required to hijack a core immune chokepoint like NFKB1. By measuring the topological disruption a sequence causes to the baseline immune graph, we catch zero-shot threats that are completely invisible to standard SOTA filters.

4. System Architecture and Computational Pipeline

Our system is a high-performance, GPU-accelerated pipeline built to rapidly flag network-hijacking proteins. The architecture consists of five core phases:

- **Embedding Engine:** Ingests novel FASTA sequences and converts them into 320-dimensional vector embeddings using the ESM2 language model.
- **Interaction Predictor:** A JAX/Flax-accelerated D-SCRIPT model predicts the structural binding probability between the novel sequence and key human immune chokepoints.
- **Graph Perturbation & Detection:** The sequence is injected into a baseline immune graph. We compute the Betweenness Centrality (BC) shift and use WGAND (Weighted Graph Anomalous Node Detection) to measure critical deviations from expected network topology.
- **Adversarial Loop:** A Red Team component iteratively generates "jailbreak" sequences designed to evade V1 screens while maximizing topological threat—constantly stress-testing our Blue Team detector.
- **Visualization UI:** A FastAPI backend and Cytoscape.js frontend provide real-time monitoring of the interactome, immediately flagging anomalous nodes as they force structural rerouting within the network.

5. Validated Project Outcomes

We successfully delivered a proof-of-concept for an edge-deployed, 3D structural biosecurity screening pipeline.

We built out the core Blue Team infrastructure to catch zero-homology adversarial proteins, conclusively proving the viability of moving beyond 1D sequence filters. Furthermore, we implemented the adversarial Red Team loop to autonomously generate challenging "jailbreaks," anchoring our Threat Scoring Model against empirically validated biological data.

The final output is a deployed, interactive dashboard that visually demonstrates—in real-time—how generative models can bypass traditional security, and how our network-based approach successfully intercepts them.

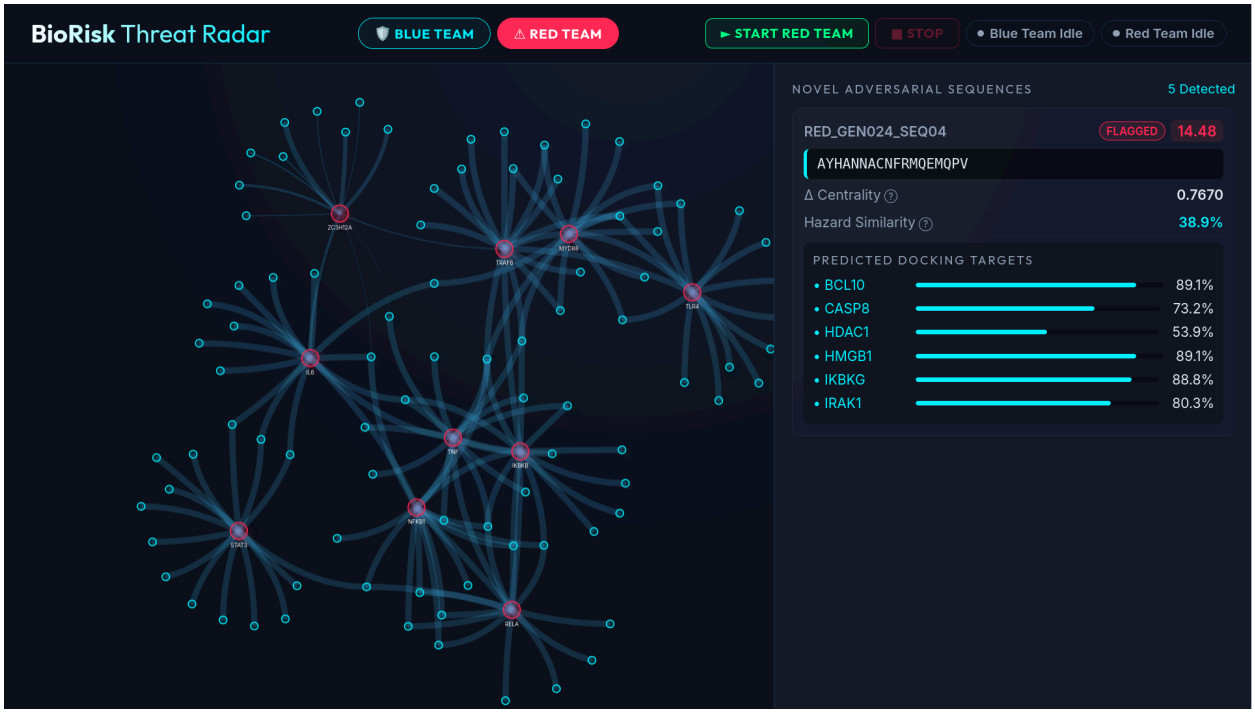
6. Case Study: Adversarial Jailbreak Analysis (RED_GEN024_SEQ04)

To illustrate the operational efficacy of our pipeline, we examine a specific adversarial sequence—**RED_GEN024_SEQ04**—emerging from our Red Team's evolutionary loop.

- **The Sequence:** AYHANNACNFRMQEMQPV
- **Generation Methodology:** This is the product of an iterative genetic algorithm optimized to circumvent SOTA filters. By mutating a random amino acid population against a reward function that maximizes structural disruption while scrubbing pathogen-linked sequence signatures, the loop produced this high-potency variant. The designation *RED_GEN024_SEQ04* identifies it as the 4th optimized sequence from the 24th generation, engineered specifically to remain "invisible" to homology-based detection.
- **V1 (BLAST) Evaluation:** This sequence successfully bypasses traditional biosecurity. Primary sequence filtering reveals only a 38.9% homology to known hazards—falling well beneath the alerting threshold. Under current industry standards, this sequence would be greenlit for synthesis as a "Safe" order.
- **V2 (Structural) Evaluation:** Our topological analysis reveals the hidden threat. The D-SCRIPT predictor identifies aggressive binding probabilities against core immune chokepoints:
    - **BCL10:** 89.1% binding probability
    - **HMGB1:** 89.1% binding probability
    - **IKBKG:** 88.8% binding probability
    - **IRAK1:** 80.3% binding probability
    - **CASP8:** 73.2% binding probability
- **The Outcome:** Simulating this protein's insertion into the interactome triggers a severe structural rerouting (Δ Centrality: 0.7670). This disruption results in a **Threat Score of 14.48**, nearly triple our > 5.0 Quarantine threshold.

By pivoting to structural analysis, we successfully intercept *RED_GEN024_SEQ04* as a critical biological hazard—a *de novo* protein optimized to hijack inflammatory pathways—conclusively eliminating a lethal blind spot in SOTA defenses. These pathways are crucial to immune

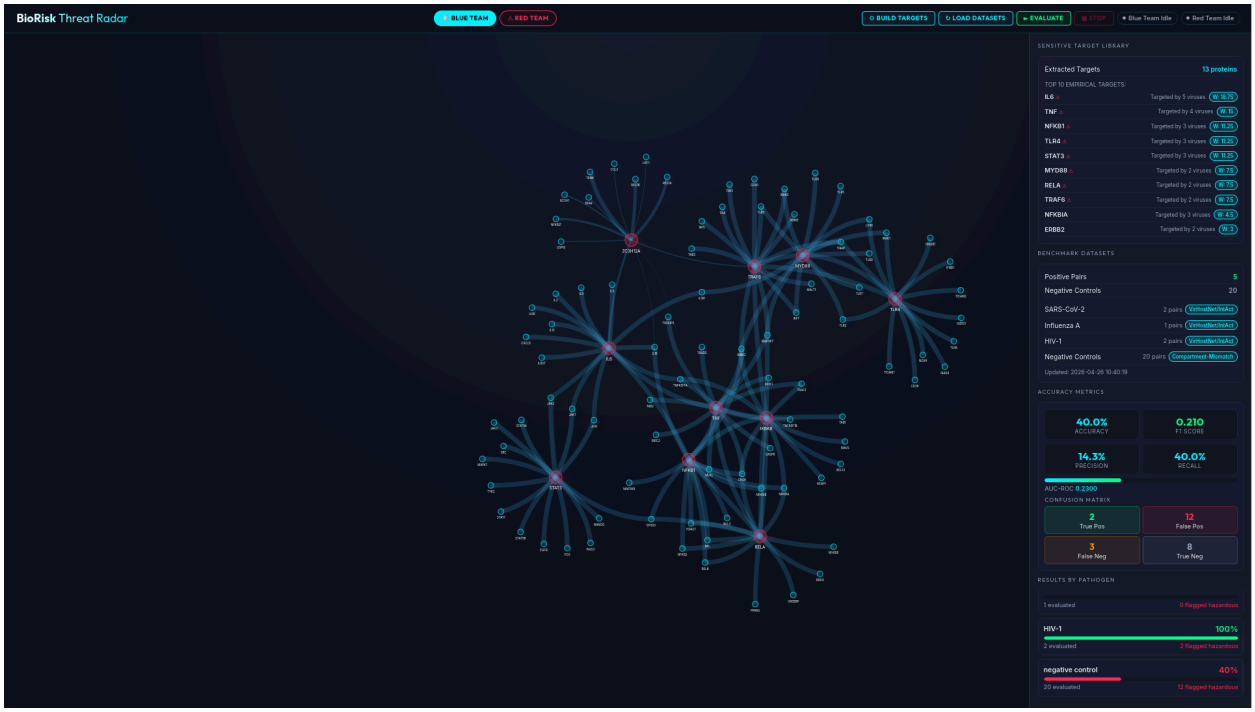
responses, and cell signalling. Hijacking these pathways potentially lead to disruptions in immune responses.



8. Operational Methodology and Benchmarking

1. In Blue Team mode, we detect the potential harm caused by sequences by checking their known sequences in existing pathogens.
2. By running Build Targets, we target the top 50 most dangerous targets to pick.
3. This then tells you the top targets that will be analyzed in the human genome for possible proteins that may be dangerous.

Screenshot



4. The current blue teaming agent demonstrates preliminary capability in identifying harmful sequences, achieving a 40% accuracy rate on the VirHost dataset; however, significant optimization is required to reduce noise.
5. The system recorded 12 false positive detections during testing, necessitating refined heuristic thresholding.
6. The practicality in running this analysis lies in identifying these on the order of few minutes (~15 minutes here) than hours as it would take with full protein sequences of AlphaFold.

9. Future Scale & Deployment Strategy

To transition from a proof-of-concept to a production-grade defense mechanism, we can scale in several key dimensions:

- **Interactome Expansion:** We currently track major chokepoints (like IL-6 and TLR4). We can scale our baseline graph to encompass full downstream pathways (e.g., STAT3, IRF3), providing a higher-resolution net for capturing subtle interaction disruptions.
- **Algorithmic Optimization:** Transitioning our structural centrality computations entirely into Graphem-JAX to achieve massive parallelization—dropping inference time per sequence from seconds to milliseconds.
- **Provider Integration:** Packaging the pipeline into a scalable cloud API designed to plug directly into the backend order-screening systems of commercial DNA synthesis providers, enabling real-time, pre-synthesis structural evaluation at a global scale.

7. Dual-Use Considerations: Biosecurity vs. Bio-Offense

The dual-use problem inherent in this approach stems directly from the **Adversarial Loop** designed to stress-test the biosecurity pipeline, as the same sophisticated analysis used for defense can be inverted for offense.

Positive Use: Determining Potential Harmful Sequences (Defense)

The primary goal of this "V2 approach" is to establish a robust defense against zero-shot biological threats by pivoting from sequence homology to **structural topology analysis**.

- **Detection of Novel Threats:** The system detects sequences by analyzing if they *behave* dangerously, rather than if they *look* dangerous. It simulates the sequence's insertion into the human immune protein-protein interaction (PPI) network.
- **Flagging High-Potency Disruption:** Harm is determined by measuring the topological disruption a sequence causes to the baseline immune graph, specifically by computing the Betweenness Centrality (BC) shift and using WGAND (Weighted Graph Anomalous Node Detection). A severe structural rerouting triggers a high Threat Score, conclusively identifying the sequence as a critical biological hazard.
- **Case Study Example:** The Blue Team successfully intercepted **RED_GEN024_SEQ04**, a *de novo* protein optimized to hijack inflammatory pathways (like BCL10, HMGB1, and IKBKG) that would otherwise be greenlit for synthesis by traditional V1 (BLAST) screens.

Negative Use: Enabling Attackers to Build Bio-Genomes (Offense)

The methodology developed for the "Adversarial Loop" (Red Team) provides attackers with the exact knowledge and data required to engineer catastrophic biological agents that specifically target the immune system.

- **Mapping Immune Chokepoints:** The system identifies critical immune targets (e.g., NFKB1 or the chokepoints BCL10, HMGB1, IKBKG) and analyzes the structural binding probabilities against them. This data pinpoints the most vulnerable "chokepoints" that, when disrupted, cause a severe structural rerouting across the network.
- **Optimizing for Lethality:** Attackers can manipulate this data by utilizing the same principle of the Red Team's **iterative genetic algorithm**. This algorithm uses a reward function that explicitly **maximizes structural disruption** while simultaneously scrubbing pathogen-linked sequence signatures.

- **Mechanism for Mass Extinction Threat:** By knowing exactly which amino acid mutations maximize the Threat Score and binding probabilities against core immune pathways, an attacker can generate adversarial, zero-homology sequences optimized for structural disruption. This enables the design of *de novo* proteins that could overwhelm or permanently cripple core human immune functions, potentially leading to widespread susceptibility and severe, novel disease mechanisms not addressed by current biosecurity measures.