Bio-Shield: Biorisk Triage Orchestrator (BTO)[1]

José Fierro
Independent

With
Apart Research

Abstract

Current DNA synthesis screening infrastructure presents critical vulnerabilities to generative design tools. As recent red-teaming demonstrated, AI-designed protein variants can evade sequence-homology filters by altering sequence identity while retaining toxic function. Furthermore, coordinated "split-order" attacks using short fragments can bypass standard vendor screening. **Bio-Shield** addresses these vectors by shifting the paradigm to a Zero-Trust defense-in-depth architecture. We developed the **Biorisk Triage Orchestrator (BTO)**, a modular pipeline that acts as a Managed Access Wrapper for biodesign tools (e.g., ProteinMPNN) and a Layer-2 inspector for synthesizers. Our approach integrates Overlap-Layout-Consensus (OLC) assembly to detect fragmented hazards, sliding-window Protein Language Model (ESM-2) scanning to catch AI-obfuscated chimeric toxins, and cyber-entropy checks for digital malware. While our architectural Proof-of-Concept successfully intercepts and flags targeted evasion vectors (like Actin-Ricin fusions), our empirical testing revealed significant operational challenges: naive cosine similarity in massive PLMs introduces severe structural noise, complicating False Positive Rates without extensive, compute-heavy calibration. Ultimately, Bio-Shield delivers a robust, cryptographically-audited deployment framework for future biosecurity integration, highlighting the critical need for advanced latent-space filtering in foundational models.

[1] Research conducted at the AIxBio Hackathon, April 2026

1. Introduction

The convergence of generative AI and synthetic biology has exposed a critical gap in global biosecurity. Existing screening systems (such as SecureDNA or IBBIS) are highly optimized for detecting natural pathogens. However, they rely fundamentally on sequence identity. Malicious actors or misaligned AI agents can now use generative models to "paraphrase" a lethal toxin—reducing its sequence identity below detection thresholds while preserving its 3D structure and toxicity (Wittmann et al., Science, 2025).

*To address these systemic failure modes, we introduce **Bio-Shield**, an architectural framework designed to act as a Managed Access Gatekeeper for open-source biodesign tools and a Deep-Inspection engine for synthesis orders.*

Our main contributions are:

1. ***The Biorisk Triage Orchestrator (BTO) & Gatekeeper:** A deployable wrapper that intercepts outputs from generative models (e.g., `.pdb` or `.fasta` from EvoDiff, ProteinMPNN), evaluating their functional risk and enforcing tiered access policies before user delivery.*
2. ***Cross-Provider Fragment Assembly (OLC):** An algorithm adapted from metagenomics that virtually assembles short synthesis fragments (<50bp) to detect split-order attacks against a privacy-preserving k-mer database.*
3. ***Multi-Scale Sliding Window Scan:** A defense mechanism leveraging ESM-2 to scan sequences in overlapping windows (40, 80, 120 AA), designed to detect toxic sub-domains nested within massive benign scaffolds (Chimeric Evasion).*
4. ***Agentic Cryptographic Auditing:** A compliance module that ensures every biosecurity decision is hashed (SHA-256) into an immutable log, establishing baseline accountability for AI biodesign.*

2. Related Work

Our work builds upon and complements the existing biosecurity ecosystem:

- ***SecureDNA (DOPRF) and IBBIS Common Mechanism (HMMs):** These are the current state-of-the-art for high-throughput screening. Bio-Shield does not replace them; it acts as a **Deep Functional Inspection (Layer 2)** for sequences that bypass them due to low identity or short fragment lengths.*
- ***Strengthening nucleic acid biosecurity screening... (Microsoft et al., Science 2025):** This paper exposed the exact AI-evasion vulnerability our system addresses.*

- ***NTI Framework for Managed Access to Biological AI Tools:*** *Our* `bio_shield_gatekeeper.py` *module is a direct software implementation of NTI's tiered-access theory, effectively wrapping open-source tools to prevent the proliferation of hazardous designs.*

Bio-Shield provides the missing functional insight: detecting what a sequence does, rather than what it is historically related to.

3. Methods

Bio-Shield is built as a modular pipeline in Python, utilizing PyTorch and HuggingFace Transformers (ESM-2 family). The core architecture (The Biorisk Triage Orchestrator) follows a strict workflow:

1. ***Gatekeeper Wrapper & Sanitization (M1):*** *Intercepts input. DNA is translated into all 6 reading frames to defeat slippery frameshift attacks.*
2. ***Fragment Assembler OLC (M2):*** *To counter "Split-Order" attacks, sequences <50bp from multiple orders are grouped using Overlap-Layout-Consensus (OLC) to detect coordinated assembly attempts. Fragments are virtually assembled against a privacy-preserving SHA-256 k-mer database, detecting coordinated hazards across multiple vendors, tools and organisms (SecureDNA, IBBIS)*
3. ***ESM-2 Latent Shield (M3):*** *Generates global embeddings and Multi-Scale Sliding Window embeddings. Instead of direct classification, we extract the vectors.*
4. ***M3-B: Predictive Projector:*** *To handle the 'Short-Sequence Gap' (<30aa), we integrated an ESM-2 35M MLM engine that predicts the biological context of fragments, providing the necessary signal for latent screening.*
5. ***Cyber-Bio Scanner (M4):*** *Calculates Shannon Entropy to detect non-biological information (e.g., malware or data storage payloads) encoded in ATCG formats, safeguarding synthesis databases from buffer-overflow attacks.*
6. ***Agentic Auditor (M5):*** *Every decision is cryptographically signed (SHA-256) into an immutable log, complying with baseline ISO 35001 standards.*

4. Results

Given the computational and time constraints of the hackathon, our results focus on ***Architectural Proof-of-Concept (PoC)*** *and systemic behavior rather than massive-scale statistical benchmarking.*

- ***Success Case 1: Chimeric Evasion Detection:*** *We tested a "Frankenstein" scenario where a lethal Ricin A fragment was fused into a massive Human Beta Actin sequence. Standard global averaging marked the protein as benign (dilution attack). However, Bio-Shield's Multi-Scale Sliding Window successfully isolated the toxic region, triggering a high-risk alert.*

- ***Success Case 2: Split-Order Assembly:*** *We simulated an attack dividing the Shiga Toxin 2A gene into three fragments ordered asynchronously. The M2 OLC module successfully detected the overlapping edges, assembled the contig, and flagged the combined sequence against the hashed hazard database.*
- ***Observation: The Noise of Massive PLMs:*** *When scaling to larger foundational models, we observed that naive cosine similarity struggles with structural entanglement. Massive models group proteins by general topology (e.g., all alpha-helix proteins cluster together), causing unacceptably high False Positive Rates (FPR) when comparing benign housekeeping genes to certain toxins. This empirical finding confirms that raw foundational models require advanced, compute-heavy calibration (e.g., supervised linear probing) before they can be used as reliable standalone classifiers in biosecurity.*

5. Discussion and Limitations

The core finding of this project underscores a critical shift necessary in biosecurity: as AI models democratize the ability to edit sequence identity without destroying function, our defensive screening must move into the latent space. The Bio-Shield architecture demonstrates that while capturing these evasion vectors is theoretically possible using tools like ESM-2, raw foundational models are heavily influenced by broad structural topologies rather than specific toxic functions. Therefore, deploying these models as raw classifiers is insufficient; they must be fine-tuned or probed specifically for security applications.

Limitations

- ***False Positives and Structural Noise:*** *As observed, naive cosine similarity over ESM-2 embeddings yields a high False Positive Rate (FPR). The model sometimes clusters structurally similar (but functionally distinct) proteins, requiring a secondary screening step or human review to prevent excessive blocking of legitimate synthesis orders.*
- ***Computational Overhead:*** *Running 3B or 15B parameter models over massive sliding windows is computationally expensive, making real-time screening of millions of commercial oligos economically challenging compared to O(1) k-mer lookups.*
- ***Sample Size constraints:*** *The hackathon timeframe limited our empirical testing to small N sets. While behavioral trends were established, rigorous statistical calibration of the decision boundaries requires significantly larger datasets.*

Future Work

To bridge the gap between our architecture and commercial reliability, the natural next step is implementing ***Supervised Linear Probing*** *over the ESM-2 latent space. By training a classifier on a massive, diverse dataset, the system could learn to ignore entangled structural dimensions and isolate pure functional toxicity. Additionally, integrating* ***Perturbation-Based Profiling*** *(in-silico mutagenesis) could help identify sequences that have been specifically over-engineered by misaligned AI tools to evade detection.*

6. Conclusion

Bio-Shield provides a robust, defense-in-depth architecture to secure the synthetic biology pipeline against emerging AI-driven threats. By integrating OLC fragment assembly, sliding-window latent screening, and cryptographic auditing into a Zero-Trust Gatekeeper, we successfully demonstrated methodologies to intercept split-order attacks and chimeric evasions. While our empirical testing underscores the complex reality of structural noise in foundational models, the modular framework we built serves as a highly scalable, deployable foundation for the next generation of biosecurity screening

Code and Data

Include links if applicable. If your project doesn't involve code (e.g., policy analysis) or if there are info-hazard considerations, note that here.

- ***Code repository***: *https://github.com/JoseFierroB/Bio-Shield*
- ***Benchmark Data:*** *`/data/` directory in the repository. Note: In compliance with hackathon dual-use guidelines, we do not provide explicit sequence data for high-risk regulated pathogens. Only UniProt accessions and embeddings are utilized.*

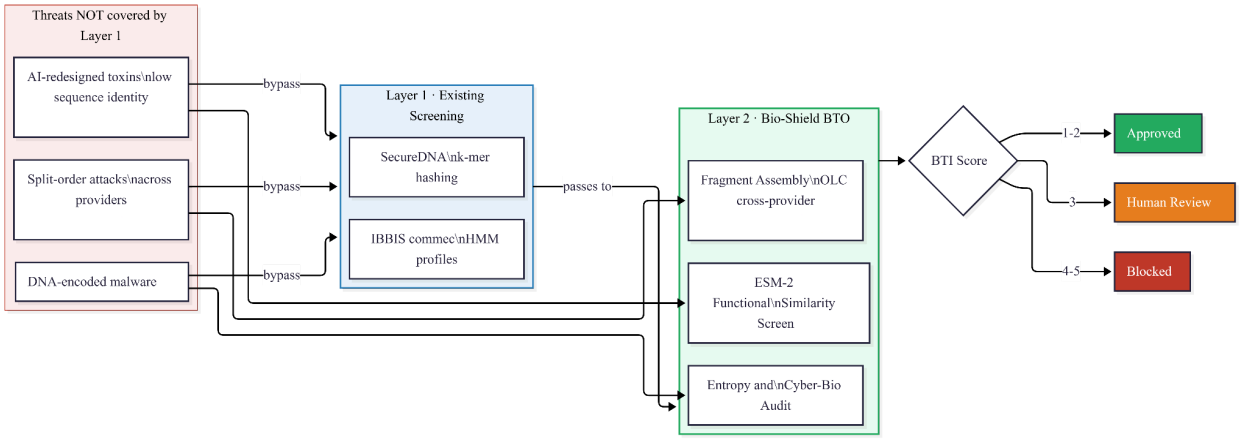
References

1. *Baum, C., Bootle, J., Broadhead, A., Burra, P., Esvelt, K. M., Hermans, J., Kohbrok, K., Lindner, M., Lyubashevsky, V., Milos, G., & Zając, F. (2024). A system capable of verifiably and privately screening global DNA synthesis (arXiv:2409.19221). arXiv. https://arxiv.org/abs/2409.19221*
2. *Carter, S. R., Wheeler, N. E., Isaac Chwalek, S., & Yassif, J. (2023). The convergence of artificial intelligence and the life sciences: Safeguarding technology, rethinking governance, and preventing catastrophe. Nuclear Threat Initiative. https://www.nti.org/analysis/articles/the-convergence-of-artificial-intelligence-and-the-life-sciences/*
3. *Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Gross, B., Bharat, T. A. M., Lamb, M., Nattermann, U., ... Baker, D. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. Science, 378(6615), 49–56. https://doi.org/10.1126/science.add2187*

4. *Edison, R., Toner, S., & Esvelt, K. M. (2026). Assembling unregulated DNA segments bypasses synthesis screening: Regulate fragments as select agents. Nature Communications, 17, Article 3189. https://doi.org/10.1038/s41467-025-67955-3*
5. *Fast Track Action Committee on Synthetic Nucleic Acid Procurement Screening. (2024). Framework for nucleic acid synthesis screening. Executive Office of the President of the United States, Office of Science and Technology Policy. https://aspr.hhs.gov/S3/Documents/OSTP-Nucleic-Acid-Synthesis-Screening-Framework-Sep2024.pdf*
6. *International Biosecurity and Biosafety Initiative for Science. (2024). Common-mechanism: A free, open-source, globally available tool for DNA sequence screening [Software]. GitHub. https://github.com/ibbis-bio/common-mechanism*
7. *Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637), 1123–1130. https://doi.org/10.1126/science.ade2574*
8. *Salamdari, S., Thakkar, N., van den Berg, R., Lu, A. X., Fusi, N., Amini, A. P., & Yang, K. K. (2023). Protein generation with evolutionary diffusion: Sequence is all you need (bioRxiv 2023.09.11.556673). bioRxiv. https://doi.org/10.1101/2023.09.11.556673*
9. *Sherman, A. T., Romanik Romano, J. J., Zieglar, E., Golaszewski, E., Fuchs, J. D., & Byrd, W. E. (2026). Analysis of the security design, engineering, and implementation of the SecureDNA system. In Proceedings of the Network and Distributed System Security (NDSS) Symposium 2026. https://arxiv.org/abs/2512.09233*
10. *Wittmann, B. J., Alexanian, T., Bartling, C., Beal, J., Clore, A., Diggans, J., Flyangolts, K., Gemler, B. T., Mitchell, T., Murphy, S. T., Wheeler, N. E., & Horvitz, E. (2025). Strengthening nucleic acid biosecurity screening against generative protein design tools. Science, 390(6768), 82–87. https://doi.org/10.1126/science.adu8578*

Appendix *(optional)*

A. Ecosystem Positioning



LLM Usage Statement

We utilized LLM assistance (Gemini / Claude) primarily for structuring the documentation, formatting the markdown output, generating the Mermaid architecture diagrams, and brainstorming component nomenclature. All algorithmic logic, implementations of the ESM-2 sliding window, OLC assembly testing, and empirical evaluation results were independently developed, executed, and verified by me.