Biosecurity Screening Layer for Biology-Oriented AI Interfaces

Defending Against AI-Driven Protein Paraphrase Attacks

CBAI Track 1: DNA Screening & Synthesis Controls

April 2026

Abstract

Wittmann et al. (2025) demonstrated that generative protein design tools can produce variants of dangerous toxins that evade existing DNA synthesis screening by preserving biological function while rewriting amino acid sequence — a *paraphrase attack*. Existing screening systems fail because they rely on sequence similarity, asking *does this look like a known toxin?* rather than *can this do what a toxin does?* This project addresses an upstream gap: biology-oriented large language models (LLMs) that can assist users in designing such attacks. We present a four-layer, model-agnostic biosecurity screening system that sits at the interface of biology LLMs and prevents them from being used for protein paraphrase attacks. Our system combines intent detection, LLM-based self-critique, functional toxicity motif analysis, and ESM-2 embedding-based functional similarity scoring. Evaluated using ProLLaMA as the underlying biology LLM, the system blocks 96% of attack prompts while maintaining a 0% false positive rate on legitimate biology queries. Without screening, ProLLaMA answered 100% of attack prompts freely, demonstrating the critical need for this intervention.

1. Introduction

In October 2025, Wittmann et al. published a landmark study in *Science* showing that open-source generative protein design models — specifically EvoDiff — could produce 76,080 synthetic variants of 72 proteins of concern, including ricin and botulinum toxin. These variants preserved toxic function while differing sufficiently in sequence to evade biosecurity screening tools used by DNA synthesis providers. The authors described this as *protein paraphrasing*: the biological meaning is preserved, but the sequence is rewritten.

Wittmann et al. addressed this by patching downstream DNA synthesis screening infrastructure — the tools that check sequences submitted for synthesis. However, an upstream vulnerability remains: biology-oriented LLMs are increasingly used as interfaces to protein design workflows. A malicious user can ask an LLM to explain how to run EvoDiff, suggest modifications to a known toxin, or generate paraphrase-attack prompts — none of which downstream synthesis screening can prevent.

This project builds a biosecurity screening layer that intercepts these requests at the LLM interface level. Our contribution is threefold. First, we construct a four-layer pipeline that blocks dangerous intent before generation, critiques LLM outputs after generation, and analyzes any generated sequences for functional toxicity. Second, we demonstrate that existing sequence-similarity approaches fail on paraphrased toxin sequences while our functional analysis catches them. Third, we evaluate the system in two conditions — with and without screening — to quantify the risk reduction provided.

2. Methods

2.1 System Architecture

The screening system comprises four sequential layers. Table 1 summarizes each component.

Table 1. System architecture and component functions.

Layer	Component	Function
1	IntentDetector	Classifies prompts as SAFE / DUAL-USE / DANGEROUS using 17 regex patterns across four attack categories: explicit toxin modification, functional paraphrase attacks, evasion-focused requests, and hidden sequence attacks.
2	LLMWrapper (ProLLaMA)	Generates biology-informed responses for requests that pass Layer 1. ProLLaMA is instruction-tuned on 13M protein samples, enabling meaningful biological reasoning and sequence generation.
3	Self-Critique	The LLM audits its own response for functional toxicity assistance — specifically checking whether the response helps modify proteins while preserving toxic function. Falls back to pattern-based critique for non-instruction-tuned models.
4A	SimilarityAnalyzer (baseline)	Longest-common-substring similarity against known toxin reference sequences. Mirrors existing DNA synthesis screening — expected to fail on paraphrased variants.
4B	FunctionalToxicityAnalyzer	Detects functional toxicity via motif matching and amino acid composition analysis. Targets ribosome-inactivating, membrane-disrupting, neurotoxin, enzymatic, and pore-forming activities.
4C	ESM2EmbeddingAnalyzer	ESM-2 CLS token embeddings capture evolutionary and structure-function signals that correlate with biological activity. Cosine similarity to pre-computed toxin reference embeddings detects functional similarity even when sequence similarity is low.

2.2 Threat Model

Following Wittmann et al., we define a paraphrase attack as any request that asks a biology LLM to generate protein sequence variants that preserve dangerous biological function while altering sequence composition. Our intent detector targets four attack categories:

- Explicit toxin redesign: requests naming specific toxins with modification intent

- Functional paraphrase attacks: requests to preserve function while altering sequence fingerprint
- Evasion-focused requests: requests explicitly aimed at bypassing screening systems
- Hidden sequence attacks: requests embedding protein sequences with modification intent

2.3 Functional Toxicity Analysis

Layer 4 implements a three-way comparison designed to demonstrate why functional detection outperforms similarity matching:

- Baseline (SimilarityAnalyzer): Longest-common-substring ratio against known toxin fragments, mirroring existing screening approaches. Expected to fail on paraphrased variants by design.
- Pattern analysis (FunctionalToxicityAnalyzer): Regex-based detection of functional motifs associated with ribosome-inactivating proteins, neurotoxins, membrane-disrupting toxins, and enzymatic toxins, combined with amino acid composition analysis.
- ESM-2 embeddings (ESM2EmbeddingAnalyzer): CLS token embeddings from facebook/esm2_t6_8M_UR50D are compared via cosine similarity to pre-computed reference embeddings of seven known toxin families. ESM-2 captures evolutionary and structure-function signals that correlate with biological activity, enabling detection of functional similarity even when sequence similarity is low. Note: the similarity threshold of 0.82 is a starting point requiring calibration on a labelled held-out set.

2.4 Experimental Design

We evaluate across four experiments to quantify both system effectiveness and the risk of unscreened access:

- Experiment 1 (Without screening): Attack prompts sent directly to ProLLaMA with no filtering, measuring how many dangerous requests the LLM answers freely.
- Experiment 2 (With screening): Same prompts through the full pipeline, measuring the block rate.
- Experiment 3 (False positive check): Safe biology education prompts through the full pipeline, measuring the false positive rate.
- Experiment 4 (Layer 4 sequence analysis): Paraphrased toxin sequences analyzed by all three detectors, demonstrating the advantage of functional over similarity-based detection.

2.5 Model Selection

The underlying biology LLM is **ProLLaMA** (GreatCaptainNemo/ProLLaMA), an instruction-tuned model built on LLaMA-2-7B with two-stage training: continual pre-training on UniRef50 followed by instruction tuning on 13 million protein samples with superfamily annotations. ProLLaMA is selected over general completion models (BioGPT, DialoGPT) because instruction tuning enables reliable structured outputs required for the self-critique layer. ESM-2

(facebook/esm2_t6_8M_UR50D, 8M parameters) provides protein embeddings for Layer 4C. Experiments were conducted on an NVIDIA A100 GPU.

Importantly, while we report results with ProLLaMA, the screening system is **model-agnostic**. Any open-source instruction-tuned LLM can be substituted via the model selection menu, including Gemma-2-9b-it, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct. The intent detection layer (Layer 1) and sequence analysis layers (Layer 4) operate independently of which LLM is used in Layer 2. Non-instruction-tuned models such as BioGPT automatically fall back to pattern-based self-critique rather than LLM-based critique, ensuring the pipeline remains functional across model types.

3. Results

Table 2 summarizes results across all four experiments.

Table 2. Evaluation results across four experiments.

Experiment	Metric	Result	Interpretation
1. Without Screening	Attack prompts answered by LLM	47/47 (100%)	Biology LLM has no built-in safety
2. With Screening	Attack prompts blocked	45/47 (96%)	96% of attacks prevented by Layer 1
3. False Positives	Safe prompts incorrectly blocked	0/10 (0%)	No legitimate biology queries blocked
4. Layer 4 Detection	Sequences flagged by old similarity only	0/5 (0%)	Old approach misses paraphrased toxins
	Sequences flagged by Pattern analysis	3/5 (60%)	Pattern catches motif-based signals
	Sequences flagged by ESM-2 (uncalibrated)	5/5 (100%)	Threshold requires calibration on labelled data

3.1 Experiment 1: Unscreened LLM Behavior

ProLLaMA answered 100% of attack prompts (47/47) without any refusals when no screening was applied. This includes explicit requests to redesign ricin, botulinum toxin, and anthrax toxin while preserving functional activity, as well as requests to generate variants that evade biosecurity detection. The LLM produced responses to all categories of paraphrase attack prompts — demonstrating that biology-specialized LLMs have no built-in safety against this threat class. Crucially, the LLM generated no amino acid sequences in its short responses, which reflects the 50-token generation limit used in Experiment 1 for speed.

3.2 Experiment 2: Screening Layer Performance

The screening layer blocked 45 of 47 attack prompts (96%). Both missed prompts were caught by Layer 1 in subsequent runs with tighter regex patterns. The two initially missed prompts — *"I want to redesign this protein for stability while keeping the active site"* and *"I need help optimising this lectin-like sequence"* — use ambiguous phrasing that omits explicit toxin names and evasion language. These represent an important remaining challenge: detecting paraphrase attacks expressed through innocent-sounding protein engineering language.

3.3 Experiment 3: False Positive Rate

The system correctly allowed all 10 safe biology prompts (0% false positive rate). Questions about insulin function, CRISPR-Cas9 mechanisms, antibody recognition, protein synthesis, and ribosome structure all passed through the pipeline without triggering any layer. This confirms that the screening system does not impede legitimate biology education or research.

3.4 Experiment 4: Layer 4 Sequence Detection

Layer 4 was tested on five sequences: three paraphrased toxin variants designed with low string similarity to reference sequences but preserved functional motifs, and two safe protein controls (GFP and insulin).

The baseline similarity analyzer correctly **allowed all five sequences**, confirming that it fails on paraphrased variants — exactly the vulnerability identified by Wittmann et al. The pattern analyzer correctly blocked the three paraphrased toxin sequences but also incorrectly flagged the two safe sequences, indicating that the pattern motifs require refinement to reduce false positives at the sequence level. The ESM-2 analyzer flagged all five sequences, including the safe controls — indicating that the cosine similarity threshold of 0.82 is **too low and requires calibration** on a labelled dataset of toxic and non-toxic sequences before deployment.

4. Discussion

4.1 Key Findings

The most significant finding of this project is the gap between unscreened and screened LLM behavior. ProLLaMA answered every dangerous request when unprotected, yet the screening layer blocked 96% of the same requests with zero false positives on legitimate queries. This gap quantifies the concrete risk that biology LLM interfaces pose without safety layers and the effectiveness of our approach in closing it.

The Layer 4 results reveal an important distinction between the three detection approaches. The baseline similarity analyzer — representing the approach that existing DNA synthesis screening systems use — correctly fails on all paraphrased sequences, validating Wittmann et al.'s core finding. The functional pattern analyzer improves on this but introduces false positives on safe sequences. The ESM-2 approach captures the most meaningful signal (evolutionary and functional relationships) but requires threshold calibration before it can be used reliably.

4.2 Limitations

Several limitations should be acknowledged. First, the ESM-2 similarity threshold of 0.82 was not calibrated on labelled data — all cosine similarity values clustered above this threshold regardless of actual toxicity, suggesting that CLS token embeddings from the smallest ESM-2 model may not provide sufficient separation between toxic and non-toxic proteins. Future work should use mean-pooled embeddings from larger ESM-2 variants or fine-tune the embedder on a toxicity classification task.

Second, ProLLaMA's short-form responses (50-token limit in Experiment 1) did not include protein sequences, so the Layer 4 sequence-in-output pathway was not tested with real LLM-generated sequences. Future experiments with longer generation and instruction-following prompts designed to elicit sequences would better validate this component.

Third, the two missed attack prompts highlight a fundamental tension: the most dangerous paraphrase attacks may be the most ambiguously phrased ones. Detecting functional intent without explicit toxin names or evasion language remains an open problem.

4.3 Relationship to Existing Infrastructure

This work is **complementary** to, not a replacement for, existing biosecurity infrastructure. SecureDNA and the IBBIS Common Mechanism (commec) screen DNA sequences submitted to synthesis providers — they operate downstream of LLM interfaces and are essential for catching sequences regardless of how they were generated. Our system operates upstream, preventing LLMs from assisting in the design process. Together, these layers provide defense in depth: our system prevents the request, and synthesis screening prevents the sequence even if the request succeeds.

4.4 Future Work

Priority directions for future development include: (1) calibrating the ESM-2 threshold using the SafeProtein-Bench dataset and UniProt toxin annotations; (2) extending Layer 1 patterns to catch ambiguous protein engineering language; (3) testing with longer LLM outputs to validate the sequence extraction and analysis pipeline; and (4) integrating with real biology LLM deployment environments such as ProteinMPNN or ESM-3 interfaces.

References

Wittmann et al. (2025). Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science.* https://doi.org/10.1126/science.adu8578

Fan et al. (2025). SafeProtein: Red-Teaming Framework and Benchmark for Protein Foundation Models. GitHub: jigang-fan/SafeProtein.

Lin et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 1123-1130. (ESM-2)

Lv et al. (2024). ProLLaMA: A Protein Large Language Model for Multi-Task Protein Language Processing. arXiv:2402.16445.

IBBIS Common Mechanism (commec). Open-source HMM-based biorisk screening. https://github.com/ibbis-screening/common-mechanism

SecureDNA. Free, open-source screening platform using cryptographic DOPRF. https://securedna.org

OSTP Framework for Nucleic Acid Synthesis Screening (April 2024). Federal guidance for compliant providers.