TRACE: Threat Recognition via Attention, Context, and Embedding Assembly for Context-Aware Biosecurity Intelligence

Samuel Kong[1,2] Mathias Ramm Haugland[1,3]
[1]Aix-Marseille Université, IBDM [2]Aix-Marseille Université, I2M [3]Aix-Marseille Université, LIS
tat-ching.kong@univ-amu.fr
mathias-ramm.haugland@univ-amu.fr
Research conducted at the AIxBio Hackathon, April 2026

Abstract

Generative protein design tools enable therapeutic innovation but introduce a critical dual-use failure mode: AI can paraphrase known toxins into synthetic homologs that preserve hazardous folds while evading homology-based DNA synthesis screening. Concurrently, regulatory frameworks mandate screening down to 50 bp and detection of cross-fragment assembly, yet current neural classifiers suffer from high false-positive rates, poor calibration, and uninterpretable outputs. We present TRACE, a context-aware escalation layer that bridges high-throughput first-line screening and human review. TRACE combines a deterministic short-window prefilter, a threat-pruned De Bruijn graph for cart-level assembly reconstruction, and a LoRA-fine-tuned ESM-2 protein risk scorer with temperature-scaled calibration and SHAP-based explainability. Evaluated on a family-held-out dataset of 32,526 sequences, TRACE achieves 95.1% recall at $\leq$2% false-positive rate, 0.987 PR-AUC, and 0.024 expected calibration error. Deployed as a lightweight CPU-optimized ONNX service with an interactive triage dashboard, TRACE provides regulator-ready evidence and plug-in guardrails for AI biological design tools, directly addressing OSTP 2024 and IGSC v3.0 mandates.

1 Introduction

The convergence of generative AI and synthetic biology has shifted biological design from empirical experimentation to computational prediction. Open-source tools such as ProteinMPNN, RFdiffusion, and ESM-3 can now design novel proteins with near-experimental accuracy. However, these same capabilities enable malicious actors to computationally "paraphrase" known toxins, generating synthetic homologs that preserve hazardous 3D folds and biochemical function while dropping primary sequence identity below 30% [1]. Traditional DNA synthesis screening infrastructure, which relies heavily on sequence alignment or exact-match cryptographic hashing, is fundamentally blind to these homology-independent threats.

Regulatory frameworks are rapidly tightening in response. The U.S. OSTP Framework for Nucleic Acid Synthesis Screening (2024) and the IGSC Harmonized Protocol v3.0 mandate that providers screen sequences as short as 50 bp and explicitly detect cross-fragment assembly potential by October 2026 [2]. Concurrently, the Biosecurity Modernization and Innovation Act (S.3741) establishes compliance requirements and a NIST governance sandbox. Yet current screening tools face three operational vulnerabilities: (1) *homology blindness* to AI-paraphrased variants, (2) a *short-sequence blindspot* where per-fragment neural classifiers produce high-variance embeddings and uncalibrated probabilities, inflating manual review costs ($\sim$\$15/order), and (3) a *black-box interpretability gap* that fails emerging auditability standards.

TRACE is proposed not as a standalone classifier, but as a **context-aware escalation layer** for DNA synthesis workflows. It sits between high-throughput first-line screens (e.g., IBBIS Common Mechanism, SecureDNA) and expert human review, reasoning over short windows, cart-level assembly context, and protein-function risk while producing regulator-ready explanatory evidence. Our main contributions are:

1. A hybrid three-stage architecture combining a deterministic short-window prefilter, a threat-pruned De Bruijn graph for fragmented cart reconstruction, and a LoRA-fine-tuned ESM-2 protein risk scorer.
2. Operational calibration via temperature scaling and threshold optimization, achieving $\geq$95% recall at $\leq$2% false-positive rate on family-held-out validation.
3. A hackathon-feasible, CPU-optimized ONNX deployment with a Streamlit triage dashboard and FastAPI guardrail endpoints for AI design tool integration.
4. Policy-aligned evidence packaging that maps directly to OSTP 2024, IGSC v3.0, and NTI managed-access principles.

2 Related Work

DNA synthesis screening has evolved through three generations. **Generation 1** tools (BLAST, HHsearch, SecureDNA) rely on local alignment or cryptographic exact-match hashing. SecureDNA screens down to 30 bp using DOPRF with reverse-screening to suppress false alarms [3], but cannot generalize to novel toxic folds outside precomputed hazard databases. **Generation 2** systems (BioLMTox-2, MultiTox) leverage fine-tuned protein language models (pLMs) like ESM-2 to detect latent structural signatures independent of alignment [4, 5]. While effective against paraphrased variants, they operate as black boxes, lack cart-level assembly context, and suffer from poor probability calibration on short fragments. **Generation 3**

approaches (ToxDL 2.0, IBBIS `commec`) integrate structural prediction or HMM-based biorisk scanning with taxonomy and low-concern filtering [6, 7]. ToxDL 2.0 achieves state-of-the-art detection by fusing ESM-2 embeddings with AlphaFold2 structures and GCNs, but requires heavy GPU compute and minutes per sequence, making it infeasible for real-time provider throughput.

A critical gap remains: no open, lightweight system combines short-window function screening, cart-level assembly context, calibrated triage, and defensible explainability. Furthermore, raw attention maps from transformers are mathematically unfaithful to model decisions [8], leaving analysts without biologically grounded evidence. TRACE addresses these gaps as a hybrid Gen-3 escalation layer optimized for provider throughput, hackathon feasibility, and regulatory auditability.

3 Methods

3.1 Architecture

TRACE employs a three-stage, latency-optimized pipeline:

1. **Short-Window Prefilter:** A 50-nt sliding window scans input DNA across six reading frames. Deterministic PROSITE/InterPro regex matching and candidate ORF extraction surface high-recall candidates in < 10 ms.
2. **Cart-Context Assembly Engine:** All oligos are translated to amino acid space to neutralize codon optimization. A minimizer-based De Bruijn graph ($k = 21$) is constructed with reverse complements. Threat-driven pruning expands only paths containing motif hits or high embedding scores, mitigating combinatorial explosion. Connected components yield virtual contigs for downstream scoring.
3. **Protein/Function Risk Scorer:** A LoRA-fine-tuned ESM-2 (650M parameters, $r = 8$, $\alpha = 16$ on Q/K/V projections) extracts task-specific embeddings. Residue-aware attention-weighted pooling preserves positional signals lost in naive mean-pooling. A calibrated logistic regression head fuses embeddings with motif hits to output a risk probability.

3.2 Data Pipeline

We constructed a leakage-controlled, operationally realistic dataset. *Positives* were curated from UniProt Tox-Prot and IBBIS hazard lists. *Hard negatives* included human/*E. coli* proteomes and realistic synthetic biology traffic (GFP, mCherry, luciferase, iGEM parts, Addgene plasmids) to prevent trivial taxonomic shortcuts and accurately model false-review burden. *Adversarial variants* were generated via ESM-2 iterative masked sampling, filtered to $< 30\%$ sequence identity while preserving length and physicochemical properties. *Fragmentation sets* simulated 50–100 bp oligos with 20 bp Gibson overlaps and codon variation. All splits were **family-held-out** to prevent homology memorization, yielding 26,246 training and 6,280 validation sequences.

3.3 Training & Calibration

Training was executed on a Colab T4 GPU using Hugging Face PEFT. Hyperparameters: batch size 2, gradient accumulation 4, fp16, cosine LR 2×10^{-4}, 2 epochs. Class imbalance was addressed via weighted cross-entropy loss. Post-training, we applied temperature scaling to minimize negative log-likelihood on validation logits [9], followed by precision-recall curve analysis to lock an operational threshold satisfying Recall ≥ 0.95 at Precision ≥ 0.98 (FPR ≤ 0.02). The merged model was exported to ONNX (opset 14) for CPU-optimized inference via `onnxruntime`.

3.4 Explainability & Deployment

We replaced attention heatmaps with SHAP/Integrated Gradients attribution on the classifier head, overlaid with PROSITE motifs and nearest hazardous neighbors. The system is exposed via FastAPI (`/screen/sequence`, `/screen/cart`, `/guardrail`) and a Streamlit dashboard for interactive triage, returning `ALLOW/REVIEW/BLOCK` decisions with structured JSON evidence packages.

4 Results

4.1 Training Convergence & Core Metrics

TRACE converged stably across two epochs. Validation loss decreased from 0.0639 (Epoch 1) to 0.0546 (Epoch 2), indicating controlled generalization on the family-held-out split. Final validation metrics are summarized in Table 1.

4.2 Operational Calibration & Latency

Temperature scaling reduced model overconfidence by $\sim$60% (ECE dropped from 0.081 to 0.024). The optimized threshold $\theta = 0.47$ achieves the target operational regime: **Recall $\geq$0.95 at FPR $\leq$0.02**, directly aligning with IBBIS and OSTP provider standards. ONNX CPU inference averages < 85 ms per 1 kb sequence on local hardware, and cart assembly (50 oligos) completes in < 1.8 s, well within synthesis provider throughput bounds.

4.3 Baseline Comparison & Dashboard

TRACE outperforms a frozen ESM-2+SVM baseline (Recall 0.82, PR-AUC 0.91) and matches or exceeds `commec` HMM recall on AI-paraphrased variants while providing calibrated probabilities and assembly context. The deployed Streamlit

Table 1: Family-Held-Out Validation Metrics (Epoch 2)

Metric	Value
Accuracy	0.9892
F1 Score	0.9693
Precision	0.9881
Recall	0.9513
ROC-AUC	0.9980
PR-AUC	0.9870
Expected Calibration Error (ECE)	0.024
Operational Threshold (θ)	0.47
Temperature (T)	1.82

dashboard (`https://trace-dashboard.streamlit.app/`) demonstrates real-time triage with residue importance maps, De Bruijn graph visualizations, and guardrail API simulation.

5 Discussion

TRACE demonstrates that lightweight parameter-efficient fine-tuning, contextual assembly, and rigorous probability calibration can patch critical gaps in DNA synthesis screening without heavy structural compute. By positioning TRACE as an escalation layer rather than a replacement, we preserve provider economics while catching AI-paraphrased and fragmented threats that bypass first-line screens. The system's calibrated allow/review/block triage and structured evidence packaging directly satisfy OSTP 2024 audit requirements and NTI managed-access principles.

Limitations include the exclusion of full structure-aware reranking (ESMFold/Foldseek) for 48-hour feasibility, simulation of cross-provider cryptographic matching via customer hashes rather than production DOPRF, and Streamlit Cloud RAM constraints that restrict full ONNX loading in demo mode. Future work will integrate selective structural validation for top-5% ambiguous hits, implement privacy-preserving cross-order detection, expand hard negatives to full provider traffic distributions, and harden the API for production deployment.

6 Conclusion

TRACE provides a feasible, calibrated, and explainable escalation layer for DNA synthesis screening and AI biological design guardrails. Achieving 95.1% recall at $\leq$2% FPR with sub-100ms CPU latency, TRACE bridges the gap between rapid first-line screening and expert human review. The open pipeline, ONNX deployment, and interactive dashboard enable community iteration, provider adoption, and direct alignment with emerging biosecurity policy. By prioritizing operational calibration and defensible explainability over black-box accuracy, TRACE offers a practical pathway to safer AI-enabled biodesign.

Code and Data

- **Code repository:** `https://github.com/samueltckong/trace-dashboard`
- **Live Dashboard:** `https://trace-dashboard.streamlit.app/`

Author Contributions

The first author led project design, data pipeline, and LoRA training. Also implemented the De Bruijn assembly engine, ONNX export, calibration, and Streamlit dashboard. Both authors contributed to methodology, evaluation, and manuscript preparation.

LLM Usage Statement

The authors used large language model–based tools during the research and writing process to assist with literature screening, code development and debugging, language refinement, and manuscript structuring. The outputs of these tools were treated as suggestions only. The authors independently verified the accuracy of all scientific content, references, code, analyses, and interpretations, and take full responsibility for the integrity and originality of the final manuscript.

References

[1] Wittmann, B. et al. Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science* (2025).

[2] White House OSTP. Framework for Nucleic Acid Synthesis Screening (2024).

[3] SecureDNA. Safeguarding DNA synthesis. `https://securedna.org/`.

[4] BioLMTox-2: Protein toxicity prediction via fine-tuned ESM-2. *Bioinformatics* (2024).

[5] MultiTox: Stacked ensemble for multiclass toxin prediction. *J. Chem. Inf. Model.* (2024).

[6] Shen, Z. et al. ToxDL 2.0: Protein toxicity prediction based on pretrained language model with graph neural networks. *GitHub* (2024).

[7] IBBIS. Common Mechanism for DNA Synthesis Screening. `https://github.com/ibbis-bio/common-mechanism`.

[8] Jain, S. & Wallace, B. Attention is not Explanation. *NAACL* (2019).

[9] Guo, C. et al. On Calibration of Modern Neural Networks. *ICML* (2017).

[10] Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS* (2021).

[11] Hu, E. et al. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* (2021).

[12] Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* (2015).

Appendix: Limitations and Dual-Use Considerations

Limitations

- **False Positives/Negatives:** While calibrated to $\leq 2\%$ FPR, extreme codon optimization or non-coding regulatory threats may evade detection. Recall drops on sequences < 30 aa due to transformer context limits.
- **Edge Cases:** The De Bruijn assembler may experience combinatorial explosion on carts > 200 oligos without aggressive pruning. Non-overlapping sticky ends with intronic spacers require enhanced graph heuristics.
- **Scalability Constraints:** Streamlit Community Cloud (1GB RAM) restricts full ONNX model loading; demo mode falls back to simulated inference. Production deployment requires $\geq$4GB RAM or edge GPU.
- **Structural Reranking:** ESMFold/Foldseek validation was scoped out for hackathon feasibility. Ambiguous high-risk hits currently rely on embedding + motif evidence alone.

Dual-Use Risks

TRACE is explicitly designed as a defensive escalation layer. The guardrail API returns binary refusal/escalation signals without exposing trigger logic, threshold values, or motif regex patterns, minimizing reverse-engineering risk. The system aligns with the Responsible AI x Biodesign commitments and NTI managed-access principles. While any screening tool could theoretically be probed for evasion boundaries, TRACE's calibrated thresholds, family-held-out validation, and closed-loop evidence packaging raise the technical barrier for adversarial optimization.

Responsible Disclosure Recommendations

Vulnerabilities identified in DNA synthesis screening pipelines should be reported through established channels: IBBIS Common Mechanism maintainers, SecureDNA cryptographic team, or OSTP/NIST biotechnology governance sandbox. TRACE's architecture intentionally avoids publishing adversarial benchmarks or evasion gradients to prevent misuse.

Ethical Considerations

The project adheres to open-science principles while prioritizing biosecurity. Model weights are shared for defensive research, but the guardrail API enforces inference-time circuit breaking for AI design tools. All training data is publicly sourced (UniProt, iGEM, Addgene), and no hazardous sequence generation or wet-lab validation was performed.

Suggestions for Future Improvements

- Integrate ESMFold + Foldseek for top-5% ambiguous hits to validate fold preservation.
- Implement SecureDNA-style DOPRF for privacy-preserving cross-provider order linkage.
- Expand hard negatives to full provider traffic distributions (viral vectors, CRISPR constructs, metabolic pathways).
- Deploy continuous adversarial red-teaming aligned with the Biosecurity Agent lifecycle framework.
- Harden FastAPI for production throughput with rate limiting, audit logging, and OSTP-compliant retention policies.