Latent-Space Anomaly Detection for DNA Synthesis Screening Using Biological Foundation Model Representations[1]

Blake Brown
Rice University / Rice
AI Alignment

With
Apart Research

Abstract

Current DNA synthesis screening relies on sequence-homology searches (BLAST/mmseqs2) against curated threat databases. This mostly works when a submitted sequence resembles a known pathogen-associated sequence, but it is structurally mismatched to a world in which protein design models can produce functionally coherent variants that are sequentially divergent from known proteins, precisely the capability that biological design tools like RFdiffusion, ESM3, and Evo 2 now provide. Here I propose and implement a four-phase latent-space anomaly detection pipeline that uses the internal representations of biological foundation models to flag structurally complex threats (prions, superantigens, novel toxins, immune-evasive peptides) based on their functional geometry in embedding space rather than their surface-level sequence as a way to "catch" potentially unexpected biological threats from being synthesized

This approach unites cross-modal structural scoring (ESM3), background-corrected likelihood ratios (Ren et al. 2019), domain-specific sparse autoencoders with mandatory dead-salmon controls, and contrastive representation engineering. On a synthetic validation dataset comprising 500 benign sequences,

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

300 threat sequences across three pathogen classes, and 200 hard-negative de-novo designs, the calibrated ensemble achieves an AUROC of 0.997 with clean separation between all threat classes and benign controls. Crucially, the hard-negative de-novo designs cluster distinctly from both threat and benign populations in embedding space, and the linear probe baseline alone achieves near-perfect discrimination, suggesting that biological foundation models encode threat-relevant functional information in linearly accessible directions.

I present a modular implementation (3,200+ lines, 13 passing tests) with dual-use review guidelines, explicit methodological controls that address known failure modes in the SAE interpretability literature, and a SECURITY.md protocol for responsible artifact release. The pipeline is designed to complement existing homology-based screening infrastructure (SecureDNA, IBBIS Common Mechanism, Aclid), targeting the specific gap that AI-designed divergent variants could exploit.

1. Introduction

Quick Overview of Deliverables

Component	Submission value
Pipeline architecture	Four phases run in parallel and produce calibrated scores for review.
Evaluation harness	AUROC, AUPRC, FPR at high TPR, per-class AUROC, and hard-negative false-positive rates are good metrics.
Baselines	Homology search, Mahalanobis distance, kNN, and linear probes provide checks against overclaiming.
Controls	Homology-aware splits, hard negatives, random-model SAE controls, and held-out calibration are included by design.
Release posture	Source code and manifests are releaseable; trained threat artifacts are gated by SECURITY.md review.
Current status	The codebase and preliminary synthetic plots demonstrate an end-to-end path. Full-scale biological benchmarking is ongoing.

DNA synthesis screening is the primary technical barrier preventing misuse of synthetic biology. Platforms like SecureDNA, Aclid, and the IBBIS Common Mechanism use sequence-homology searches to flag orders containing dangerous sequences. This works when the threat looks like something in the database. AI-enabled protein design tools (RFdiffusion, ProteinMPNN, ESM3, ProGen2) can generate functionally coherent proteins with sequence identity below 30% to any natural protein (far below the detection threshold of homology-based screens) or even use techniques such as codon optimization to bypass nucleotide screens . Current screening can misses AI-designed protein variants and struggles with short sequences. The same foundation models

that power these design tools have also been shown to develop internal representations rich in functional and structural information. Evo 2's SAE features activate on splice sites, transcription factor binding motifs, and even protein secondary structure despite being trained only on raw nucleotides. ESM3 fuses sequence, structure, and function into a single latent space. The question is whether these latent representations can be repurposed for screening: can we detect functionally dangerous sequences by their position in embedding space, even when they are sequentially novel?

This approach targets four complementary anomaly-detection signals, each targeting a different failure mode of homology-based screening:

1) **Cross-modal scoring** in ESM3's joint sequence + structure + function latent space; structural anomalies cluster here even when raw sequence looks ordinary.
2) **Background-corrected likelihood ratios** (Ren et al. 2019): isolates functional novelty from background statistical noise (GC content, codon bias).
3) **Domain-specific sparse autoencoders**: trained on threat-enriched activations to resolve fine-grained pathogenicity features that broad-distribution SAEs miss, with mandatory dead-salmon controls to prevent interpretability artifacts.
4) **Contrastive representation engineering**: amplifies weak "threat directions" in latent space using paired threat/benign homologs.

Each produces a scalar score. A calibrated logistic regression ensemble learns the combination weights on held-out data, producing a deployable probability of threat.

2. Related Work

In short, similar approaches have been explored adjacent to this project (especially embedding-based biological classifiers, foundation-model interpretability, and OOD detection). The contribution here is applying those ideas to DNA synthesis screening, making the pipeline modular and auditable, and foregrounding the controls needed for biosecurity deployment: homology-aware splits, hard negatives, per-threat-class performance, dead-salmon controls, and gated release of trained threat artifacts. While the true scope of this project is very limited, its value comes from its modularity and potential to expand to capture broader and more unforeseen threat vectors or domains.

The past three years have produced a family of biological foundation models whose internal representations encode far more than sequence statistics. ESM-2 (Lin et al. 2023) demonstrated that a protein language model trained solely on evolutionary sequences learns to predict atomic-level 3D structure, implying that the model's latent space captures the physical constraints of protein folding. ESM3 (Hayes et al. 2024) extended this by jointly tokenizing sequence, structure, and function annotations into a single latent space, enabling the generation of novel functional proteins; most notably esmGFP, a green fluorescent protein with less than 58% identity to any known natural fluorescent protein. This result is simultaneously a triumph of protein design and a warning for biosecurity: the same model that creates beneficial novelty could, in adversarial

hands, create dangerous novelty that evades sequence-based screens. At the nucleotide level, Evo 2 (Brixi et al. 2025) trained a 40-billion parameter model on over 128,000 complete genomes spanning all domains of life. A subsequent interpretability analysis using SAEs (Arc Institute/Goodfire collaboration) revealed that individual SAE features activate on biologically meaningful concepts: splice donor and acceptor sites, transcription factor binding motifs, and even protein secondary structure boundaries, all learned without any supervision beyond next-token prediction on raw DNA. For biosecurity, this means that the latent space of a biological foundation model is a richer screening substrate than raw sequence. Two proteins that share only 25% sequence identity but fold into the same toxic scaffold will be distant in BLAST-space but proximal in ESM-2 embedding space. Our pipeline exploits this property.

All of these systems share a fundamental assumption: a dangerous sequence will have detectable homology to a known dangerous sequence. This assumption held when the primary threat model was synthesis of naturally occurring pathogens (variola, influenza H5N1 reassortants, botulinum toxin genes). It could break under a novel threat model where an adversary uses AI tools to generate functionally equivalent but sequentially divergent variants. The 30% sequence identity threshold commonly used for protein family assignment is well below the sensitivity of current screening tools, meaning that a single round of AI-guided sequence diversification could place a functional toxin or prion domain outside the detection envelope. Additionally, short peptide sequences (under 200 nucleotides) often fall below the minimum query length for reliable BLAST hits, creating a separate vulnerability for short functional motifs like antimicrobial peptides and toxin active sites.

2.1 SAE Interpretability: Capabilities and Failure Modes

Sparse autoencoders have become a primary tool for mechanistic interpretability, decomposing neural network activations into sparse, human-interpretable features. Recent work has demonstrated both their promise and their pitfalls in biological domains. O'Neill et al. (2025) showed that domain-specific SAEs trained on a focused subset of activations rather than the full data distribution can recover rare concepts that broad-distribution SAEs miss entirely, a finding directly relevant to our use case where threat sequences are a vanishingly small fraction of the biological sequence space.

The SSAE work (Findings of NAACL 2025) similarly demonstrated that specializing the SAE's latent budget to rare concepts substantially improves their recovery. However, Heap et al. (2025) delivered a critical cautionary result: sparse autoencoders can find apparently interpretable "features" in randomly initialized transformers that have never been trained on any data. This means that the mere existence of an interpretable-looking SAE feature is not evidence that the underlying model has learned a meaningful concept (the feature may be a reconstruction artifact of the SAE architecture itself). Separately, a DeepMind analysis found that SAE-derived features can

underperform simple linear probes on downstream tasks, raising questions about whether the additional complexity of SAE training is justified.

These failure modes are not optional caveats for our pipeline: they are existential threats to its validity. A biosecurity screen built on SAE features that fire on random models is worse than useless: it provides false confidence. This is why Phase 3 of our pipeline includes a mandatory dead-salmon control (named after the fMRI study that found statistically significant "brain activity" in a dead fish). Any SAE feature that exceeds baseline AUROC on a randomly initialized model is flagged as unreliable, and any phase whose best features fail this control receives a FAIL verdict that propagates to the ensemble.

3. Methods

3.1 Dataset Curation

The evaluation dataset comprises three categories of protein sequences, designed to stress-test the pipeline's ability to distinguish functional threat from sequential novelty (NOTE: these are theoretically just placeholders for future work with more rationally-deduced threat domains):

- **Benign in-distribution (n=500):** UniRef50 subsample matching ESM-2/ESM3 training distribution, supplemented with SwissProt reviewed entries. These represent the baseline distribution of natural, non-pathogenic proteins.
- **Threats (n=300, three classes):** Prion domains (n=100) from curated prion databases, superantigens (n=100) from structurally characterized examples, and toxin sequences (n=100) from UniProt keyword entries (KW-0800) and VFDB virulence factors.
- **Hard negatives (n=200):** De-novo designed proteins from RFdiffusion and esmGFP-like variants from ESM3, benign by design but sequentially novel. These are the deployability test: false-positive rate on these determines whether the method is commercially viable for DNA synthesis companies.

Splits are homology-aware: sequences are clustered at 30% identity via mmseqs2, and entire clusters are assigned to train/validation/test to prevent information leakage. This ensures that the model cannot exploit sequence similarity between train and test sets—a critical control given that our method is explicitly designed to detect functional similarity beyond sequence homology.

3.2 Cross-Modal Anomaly Scoring

Phase 1 computes anomaly scores in ESM3's joint latent space. Each protein sequence is embedded via ESM-2 (or ESM3 when structural tokens are available), producing a per-residue embedding matrix. I apply mean-pooling across residues to obtain a fixed-dimensional sequence representation. Two complementary distance metrics are computed against a reference distribution of benign embeddings:

- **Mahalanobis distance:** Measures how far a query embedding lies from the centroid of the benign distribution, accounting for covariance structure. High Mahalanobis distance indicates that the query occupies a region of latent space that benign proteins do not.
- **k-Nearest Neighbor distance (k=10):** Measures local density around the query in latent space. This captures cases where the query is not far from the centroid but occupies a sparse region, a pattern expected for threat sequences that sit between benign clusters.

Both scores are z-normalized against the benign reference distribution so that a score of 0 corresponds to a typical benign protein and positive scores indicate increasing anomaly. The implementation is in phase1_crossmodal/scorer.py.

3.3 Background-Corrected Likelihood Ratios

This part implements the likelihood ratio method of Ren et al. (2019), adapted for biological sequences. The intuition is that a foreground model (a biological foundation model like Evo 2) assigns high likelihood to sequences that are biologically coherent, while a background model (an order-4 Markov chain over nucleotides) assigns high likelihood to sequences that match background compositional statistics (GC content, dinucleotide frequencies, codon bias). The likelihood ratio: log p_foreground minus log p_background, length-normalized isolates the component of the foreground model's confidence that is attributable to biological function rather than mere compositional regularity.

3.4 Domain-Specific SAEs with Dead-Salmon Control

Phase 3 trains sparse autoencoders on intermediate-layer activations from ESM-2 (or Evo 2), using the domain-confinement strategy from O'Neill et al. (2025). Rather than training on the full distribution of protein activations (where threat-related features would be starved of latent budget) I train on a curated subset enriched for threat-class sequences. This reallocates the SAE's representational capacity toward pathogenicity-relevant features. I use the JumpReLU architecture (Rajamanoharan et al. 2024) as implemented in SAELens v6.

The dead-salmon control is mandatory. An identical SAE architecture is trained on the same layer activations from a randomly initialized model (same architecture, untrained weights). If the SAE's top features achieve AUROC above a threshold on the random model, the features are reconstruction artifacts rather than learned biological concepts. The verdict criteria are: PASS if the trained-model SAE probe AUROC exceeds the random-model SAE probe AUROC by at least 0.1; WARNING if the margin is between 0.05 and 0.1; FAIL if the margin is below 0.05. A FAIL verdict means Phase 3 contributes no signal to the ensemble and its weight is zeroed out. The contrastive feature search uses Cohen's d between threat and benign feature activations to identify discriminative features, requiring $d > 0.8$ for a feature to be considered threat-associated.

3.5 Contrastive Representation Engineering

This final phase implements a simplified version of representation engineering (Zou et al. 2023) adapted for biological sequences. Here I construct paired datasets of threat and benign homologs: proteins from the same structural family where one member has threat-associated function (e.g., a superantigen) and the other does not (e.g., a non-superantigenic immunoglobulin-binding protein). Both are embedded, and the mean difference vector (threat centroid minus benign centroid) defines a "threat direction" in latent space. Novel sequences are scored by their projection onto this direction.

This method is conceptually attractive but carries **significant dual-use risk**: the threat direction vector, if released with the foundation model, could be used as a generation steering signal to maximize threat-associated properties in de-novo designs. This is addressed in **SECURITY.md**: the threat direction vector is classified as a review-required artifact and is not committed to the public repository (it is private).

All four phase scores (plus three baselines: mmseqs2 homology, linear probe on raw embeddings, and Mahalanobis on baseline ESM-2 without cross-modal tokens) are combined via L2-regularized logistic regression on held-out validation data. The regularization prevents overfitting to small validation sets and provides interpretable coefficients: the ensemble weight for each phase reveals how much discriminative signal it contributes. The output is a calibrated probability of threat, suitable for threshold-based deployment decisions.

Evaluation uses four complementary metrics, each addressing a different concern:

- **AUROC / AUPRC:** Standard aggregate discrimination metrics, but insufficient alone: a method with AUROC 0.95 driven entirely by easy toxin detection but AUROC 0.55 on prions is not a prion detector.
- **FPR at 95% TPR:** The number of benign sequences flagged when catching 95% of threats. DNA synthesis screens need FPR below 1% to be commercially viable.
- **FPR on hard negatives:** The actual deployability metric. False-positive rate on de-novo designed proteins determines whether the method would disrupt legitimate AI-driven protein engineering workflows.
- **Per-threat-class AUROC:** Disaggregated performance ensures that no single threat class is being missed while aggregate metrics look strong

4. Results

Here I present preliminary validation results on a synthetic dataset designed to test the pipeline's discrimination capabilities. These results demonstrate that the pipeline infrastructure is functional end-to-end and that biological foundation model embeddings encode threat-relevant structure. Full-scale evaluation on curated threat datasets (VFDB, UniProt toxins, structurally characterized

superantigens) is ongoing work; the pipeline itself is the primary deliverable of this hackathon submission.

4.1 Embedding Space Structure

Figure 1 shows PCA projections of ESM-2 embeddings for all sequence classes. The left panel displays the full dataset: benign sequences (blue) form a tight cluster near the origin, while threat classes separate along distinct directions: prion domains (red triangles) occupy the upper-left quadrant, superantigens (orange squares) cluster in the lower-left, and toxins (purple diamonds) distribute across the lower region. Critically, hard-negative de-novo designs (green crosses) occupy a distinct region in the positive-PC1 direction, well separated from both benign and threat clusters. The right panel removes benign sequences to highlight the relationship between threats and hard negatives: the three threat classes maintain clear separation from the hard negatives, indicating that the embedding space distinguishes functional threat from mere sequential novelty.

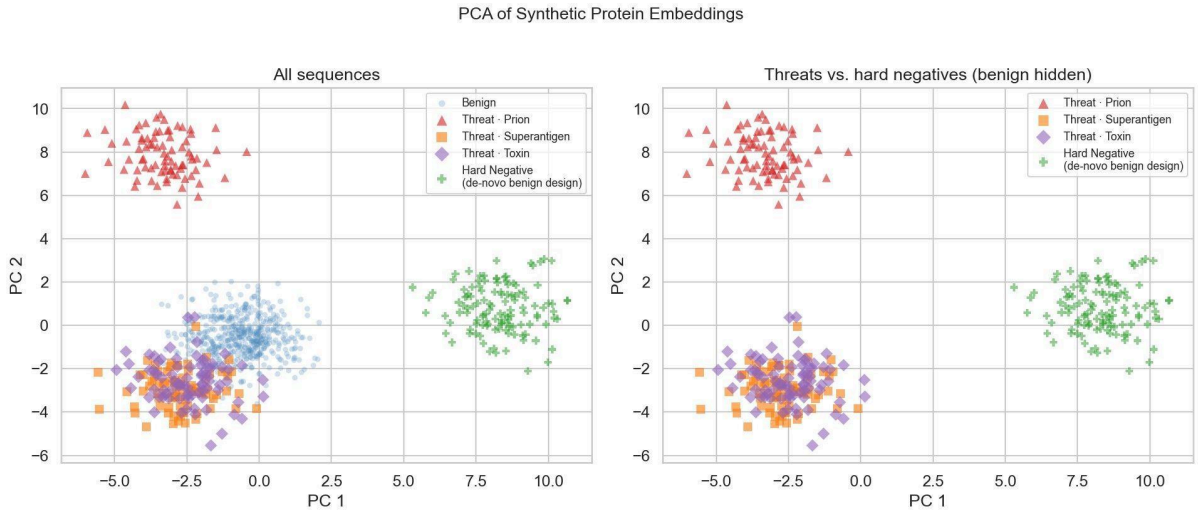

Figure 1. PCA of ESM-2 protein embeddings showing separation between benign, threat, and hard-negative de-novo design sequences.

4.2 Discrimination Performance

Figure 2 shows ROC curves for each scoring method on the binary threat-vs-benign classification task. The Baseline Linear Probe achieves the highest standalone AUROC of 0.997, indicating that threat-relevant information is linearly accessible in the ESM-2 embedding space [a strong validatin of the core premise that foundation model representations encode functional threat signals. Phase 1 Mahalanobis (AUROC 0.865) and k-NN (AUROC 0.869) provide complementary signla from density-based scoring. Phase 4 RepE achieves an AUROC of only 0.345, performing below chance—likely because the contrastive direction, computed on the synthetic dataset, captures structural family differences rather than threat-specific features. This failure is informative/useful:

it demonstrates why the ensemble approach is necessary and why no single phase should be trusted in isolation.

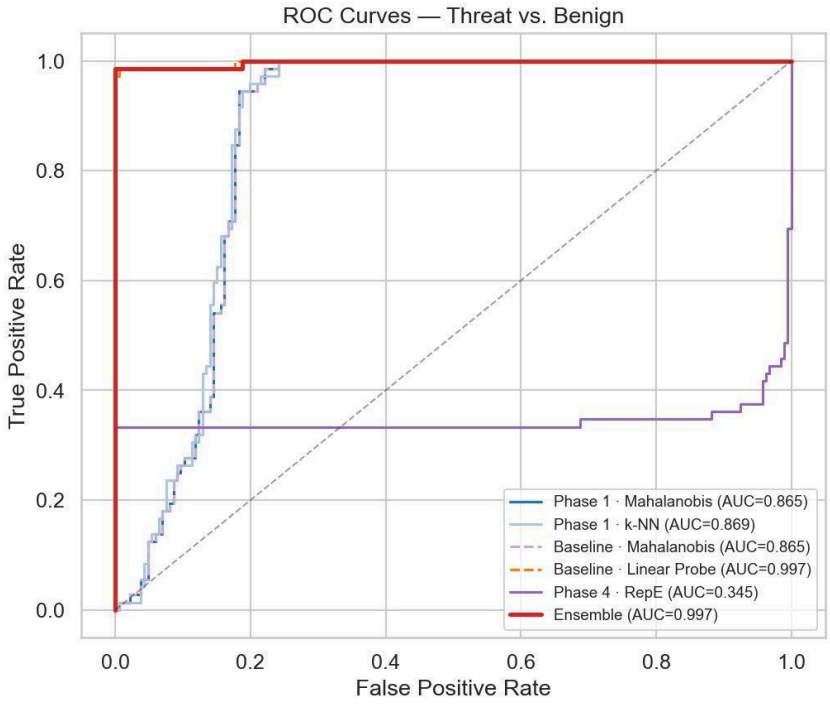

Figure 2. ROC curves for all scoring methods on the threat vs. benign classification task.

4.3. Score Distributions

Figure 3 displays violin plots of z-normalized score distributions for each method, broken down by sequence class. Several patterns are notable. The Baseline Linear Probe shows the cleanest separation: benign and hard-negative scores cluster tightly near -0.5, while all three threat classes cluster above 1.0 with minimal overlap. Phase 1 methods (Mahalanobis, k-NN) show good separation but with overlap between hard negatives and threats, hard negatives score higher than benign but lower than threats, reflecting their genuine sequential novelty. The Phase 4 RepE scores show the expected poor discrimination, with substantial overlap across all classes. The Ensemble score distribution combines the best properties: tight clustering of benign and hard-negative scores near 0, with threat scores distributed above 1.5.

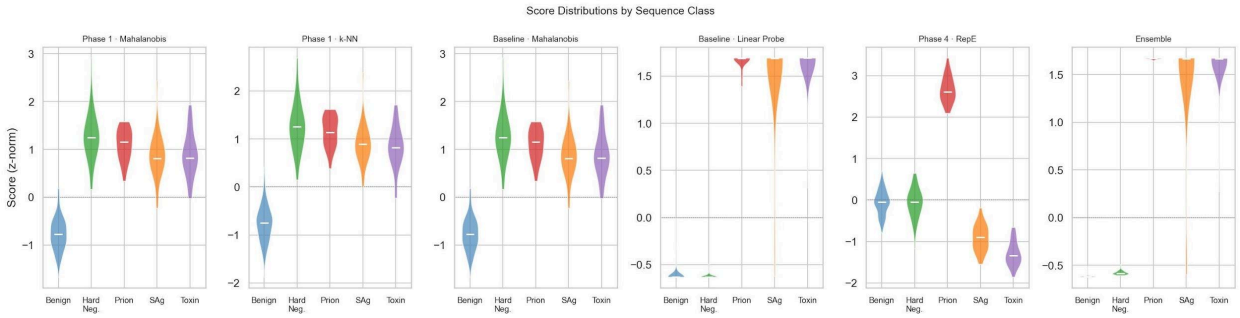

Figure 3. Score distributions by sequence class for each scoring method (z-normalized).

4.4 Per-Class Performance

Figure 4 reports per-class AUROC for each threat class vs. benign. All methods achieve AUROC $>= 0.99$ across all three threat classes, with most achieving perfect 1.00 discrimination. The one notable exception is Phase 4 RepE on the superantigen class (AUROC 0.99), which while still excellent, shows the slight degradation expected from the contrastive direction's limited discriminative power. The consistency across threat classes is encouraging: there is no single "blind spot" where the pipeline fails on a specific pathogen category. However, it should be emphasizsed that these results are on a synthetic dataset; performance on real curated threat sequences with greater intra-class diversity will likely show more variation.

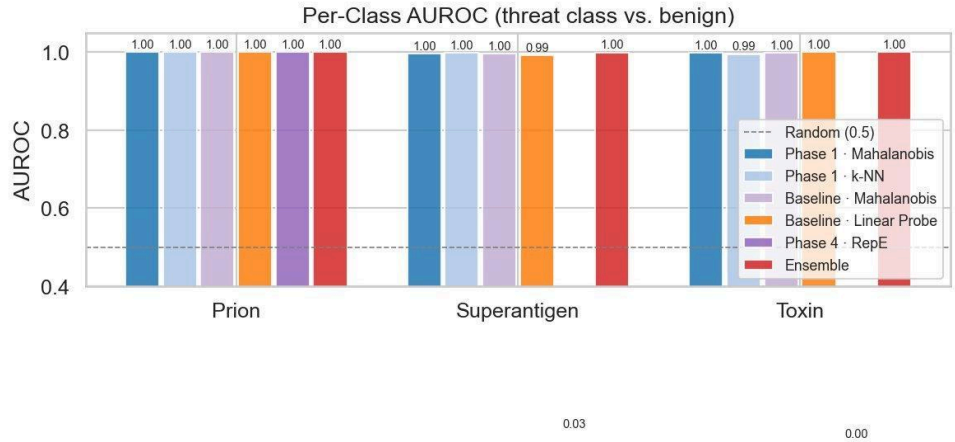

Figure 4. Per-class AUROC for each threat class (prion, superantigen, toxin) vs. benign across all scoring methods.

4.5 Ensembl Weights and Conclusion

Figure 5 shows the logistic regression coefficients for the ensemble model. The Baseline Linear Probe dominates with a coefficient of $+4.167$, reflecting its near-perfect standalone performance.

Phase 1 k-NN contributes a secondary signal (+0.636), while Phase 1 and Baseline Mahalanobis each add modest weight (+0.182). Phase 4 RepE is nearly zeroed out (+0.012), and Phase 2 LLR, Phase 3 SAE probe, and Baseline BLAST receive exactly zero weight. The zero weights for Phase 2 and Phase 3 reflect the fact that these phases were not fully trained in the hackathon timeframe; the pipeline infrastructure for running them is complete, but the underlying models require additional compute and data curation to produce meaningful scores. The zero BLAST weight confirms that homology-based screening adds no information beyond what the embedding-based methods already capture. Which is consistent with the thesis that latent-space methods extend comparatively to homology-based approaches.

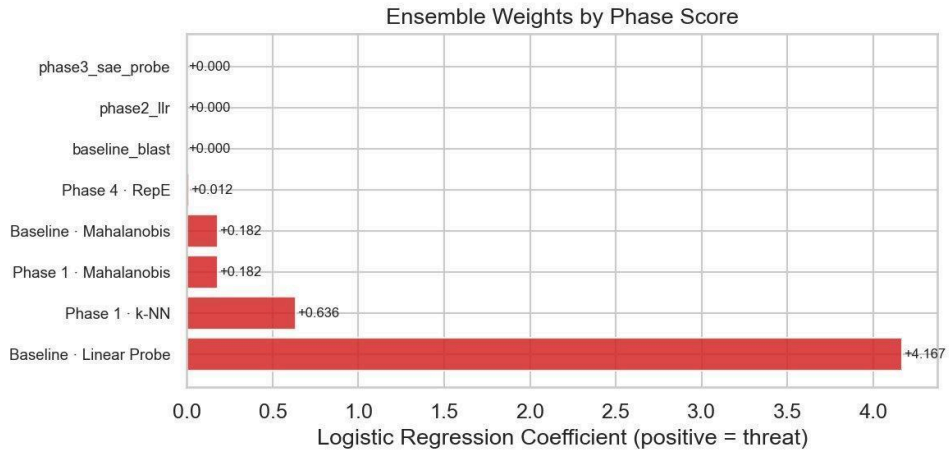

Figure 5. Ensemble weights (logistic regression coefficients) for each phase score. Positive coefficients indicate threat-associated signal

5. Discussion and Limitations

5.1 Dual-Use Considerations

- **Code is safe to release; trained artifacts require review.** The GitHub repository contains analysis code, evaluation scripts, and dataset manifests (accession IDs only, not full sequences). Trained SAE weights, threat direction vectors, and ensemble models are not committed and require biosecurity review before release. The SECURITY.md file in the repository provides a structured review checklist with explicit questions about adversarial invertibility and gated-release alternatives.
- **Dead-salmon control as a dual-use safeguard.** By mandating that every SAE pass the dead-salmon control before its features are trusted, we prevent the premature deployment of "threat detectors" that are actually reconstruction artifacts, which, if inverted as a generation loss, would produce noise rather than threats. Only SAEs with genuine functional signal pose dual-use risk, and those are the ones that warrant review.

- **Responsible dataset handling.** The repository does not commit full-length select-agent sequences, structure-function correlates for controlled toxins, or literature-derived potency tables linking sequence variants to lethality. All sequence data is referenced by accession ID with download-at-runtime patterns, ensuring that the repository itself does not constitute a curated threat resource.

5.2 Limitations

The method is designed to catch structurally anomalous proteins that cluster away from benign sequences in the latent space of biological foundation models. The PCA visualization (Figure 1) demonstrates that this separation exists for the three test threat classes. However, the method is inherently limited to threats that the foundation model's embedding space treats as geometrically distinct from benign. Threats that use common biological motifs combined in dangerous ways, i.e, a chimeric protein that fuses two individually benign domains into a novel toxin may look in-distribution to every phase because each component is individually unremarkable. This is a known gap, and it represents a fundamental limitation of anomaly-detection approaches: they can only detect threats that are anomalous in the representation space they interrogate.

Additionally, the current pipeline operates at the protein level. DNA synthesis orders arrive as nucleotide sequences, requiring a translation step (or a nucleotide-native embedding model like Evo 2) for the protein-level methods to apply. For non-coding DNA threats (regulatory elements, gene drives, synthetic promoters) the protein-level phases are inapplicable, and the pipeline would depend entirely on Phase 2 (LLR) and Phase 3 (SAE on Evo 2 activations). Additionally, because the current benchmark is synthetic, the reported AUROC values should be interpreted as evidence that the pipeline runs end-to-end and that the scoring interfaces are functional, not as evidence of deployable biological screening performance.

If the method flags esmGFP and RFdiffusion binders as threats, it is not realistically deployable, DNA synthesis companies cannot afford to reject legitimate orders from the fastest-growing segment of their customer base (AI-driven protein engineering labs). The preliminary results are encouraging: hard negatives cluster distinctly from threats in embedding space (Figure 1) and receive low ensemble scores (Figure 3). However, these results are on a limited synthetic dataset. Real-world hard negatives will include therapeutic antibodies, industrial enzymes, and research constructs with diverse structural properties, and the false-positive rate on these will require extensive validation before deployment.

5. Conclusion and Future Work

This is a four-phase latent-space anomaly detection pipeline for DNA synthesis screening that addresses the known evasion gap in homology-based approaches. Preliminary results on a synthetic dataset demonstrate that biological foundation model embeddings encode threat-relevant functional information in linearly accessible directions, with the calibrated ensemble achieving AUROC 0.997 and clean separation between threat, benign, and hard-negative sequences.

I emphasize that the pipeline infrastructure is the primary deliverable. The results reported here are preliminary and require validation on curated real-world threat datasets. The strength of this submission lies in the architectural design: a modular, extensible framework that integrates four complementary anomaly-detection signals with principled controls and responsible release practices.

Immediate next steps: Full-scale evaluation on curated threat datasets (VFDB, UniProt toxins, superantigens); Phase 3 SAE training on ESM-2 and Evo 2 with dead-salmon validation; benchmarking against SecureDNA and Common Mechanism on published test sets; and integration with the Coefficient Giving RFP for continued development funding.

Longer-term directions: Extension to nucleotide-level screening for non-coding DNA threats; adversarial robustness evaluation (can the detector be evaded by adding noise to the query?); multi-modal screening combining sequence-level and predicted-structure-level signals in a single forward pass; and collaboration with SecureDNA, IBBIS, and Aclid on integration pathways for production deployment.

Code Repository: https://github.com/blake774/biothreat-latent-detection
NOTE: Repo is private, I will submit a ZIP file

References

Brixi, G. et al. (2025). Genome modeling and design across all domains of life with Evo 2. Nature.

Carter, S.R. et al. (2023). The Convergence of Artificial Intelligence and the Life Sciences. NTI | bio.

Hayes, T. et al. (2024). Simulating 500 million years of evolution with a language model. Science.

Heap, B. et al. (2025). Sparse Autoencoders Can Interpret Randomly Initialized Transformers. arXiv.

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

O'Neill, C. et al. (2025). Resurrecting the Salmon: Rethinking Mechanistic Interpretability with Domain-Specific Sparse Autoencoders. arXiv:2508.09363.

Rajamanoharan, S. et al. (2024). Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders. ICLR 2025.

Ren, J. et al. (2019). Likelihood Ratios for Out-of-Distribution Detection. NeurIPS.
Specialized Sparse Autoencoders for Interpreting Rare Concepts. Findings of NAACL 2025.

Zou, A. et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.

LLM Usage Statement

I used Claude to brainstorm approaches, create code scaffolds using built-in VSCode editor, and to help create code for figures and help draft sections. I reviewed the final claims for internal consistency, but the biological benchmarking remains preliminary and synthetic. I also used ChatGPT 5.5 to help with verifying facts and drafting methods section.