The Protein ID Card: A Semantically Aware Screening Framework for Pathogenic Sequence Detection Using ESM-2 and FAISS

Guillaume Zahnd

Apart Research, AIxBio Hackathon, April 2026

Abstract

The emergence of AI-powered biological design tools has necessitated a shift in biosecurity from sequence-alignment methods to function-prediction-based analysis. Current DNA screening protocols relying on BLAST are increasingly vulnerable to *de novo* designed sequences that evade similarity thresholds while retaining pathogenic functionality. We propose a scalable biosecurity screening pipeline that utilizes the ESM-2 transformer architecture to extract deep biological features from viral sequences, followed by similarity retrieval using FAISS. We implement a multi-class classification scheme to generate a functional "ID card" for proteins across five axes: Baltimore classification, Molecular function, Host category, Cellular tropism, and Zoonotic potential. Evaluating 15,154 viral samples from UniProt, our approach achieves a superior aggregate F_1-score of 0.89 compared to 0.77 for BLAST. The embedding-based pipeline demonstrates significant performance gains in complex domains such as Host category (0.94) and Cellular Tropism (0.98), where sequence identity often fails to reflect biological roles. These results indicate that high-dimensional embeddings successfully capture the structural and functional constraints of viral evolution, providing a robust, semantically aware guardrail for modern biosecurity.

Introduction

The rapid proliferation of high-throughput DNA synthesis technologies and the emergence of AI-powered biological design tools have fundamentally shifted the landscape of biosafety and biosecurity [Biden, 2023, HM Government, 2023, Wang et al., 2026]. Current DNA screening protocols, which serve as the primary defense against the illicit synthesis of pathogenic sequences, rely heavily on the Basic Local Alignment Search Tool (BLAST) to detect homologies with known regulated agents [Altschul et al., 1997]. However, sequence-alignment methods are increasingly ill-equipped to handle the rise of *de novo* designed proteins and highly divergent functional variants. Because these tools prioritize string matching over biological context, they frequently fail to detect dangerous sequences that have been modified to evade traditional sequence similarity thresholds while retaining their original pathogenic functionality [Sathyamoorthy and Puri, 2026].

To address these vulnerabilities, there is a critical need for screening frameworks that move beyond sequence identity toward function-prediction-based analysis [Radivojac et al., 2013]. By leveraging the semantic understanding of "protein language", it is possible to flag sequences based on their predicted biological roles, even when the underlying primary structure deviates significantly from known templates. This study specifically adapts the architecture of Large Language Models (LLMs) to the biological domain, treating amino acid sequences as a structured language to uncover latent functional patterns that are invisible to traditional alignment algorithms. By focusing on viral proteins gathered from the UniProt database, we aim to provide commercial synthesis providers with more robust guardrails to detect AI-generated sequences that might otherwise bypass "Best Match" screening criteria.

In this work, we propose a pipeline that utilizes Evolutionary Scale Modeling (ESM) to extract deep biological features from viral sequences. The implementation of the ESM-2 architecture allows the transformation of primary sequences into embeddings that capture complex biological patterns, providing a more robust basis for functional classification than sequence identity alone [Lin et al., 2023]. We implement a comprehensive multi-class classification scheme to establish a functional "ID card" for each protein across several analytical angles, including Baltimore classification, Host category, Molecular function, Cellular tropism, and Zoonotic potential. Our results demonstrate that this embedding-based approach significantly outperforms traditional BLAST-based methods in predicting biological context, providing a scalable and semantically aware solution for modern biosecurity screening.

Methods

Figure 1 provides an overview of the methodology: The screening pipeline follows a two-stage architecture: semantic encoding using ESM-2 followed by similarity retrieval using FAISS. The process begins with an input amino acid sequence, the final output is the resolution of the unknown sequence into specific sub-classes across five primary classification axes. The implementation details are structured hereafter.

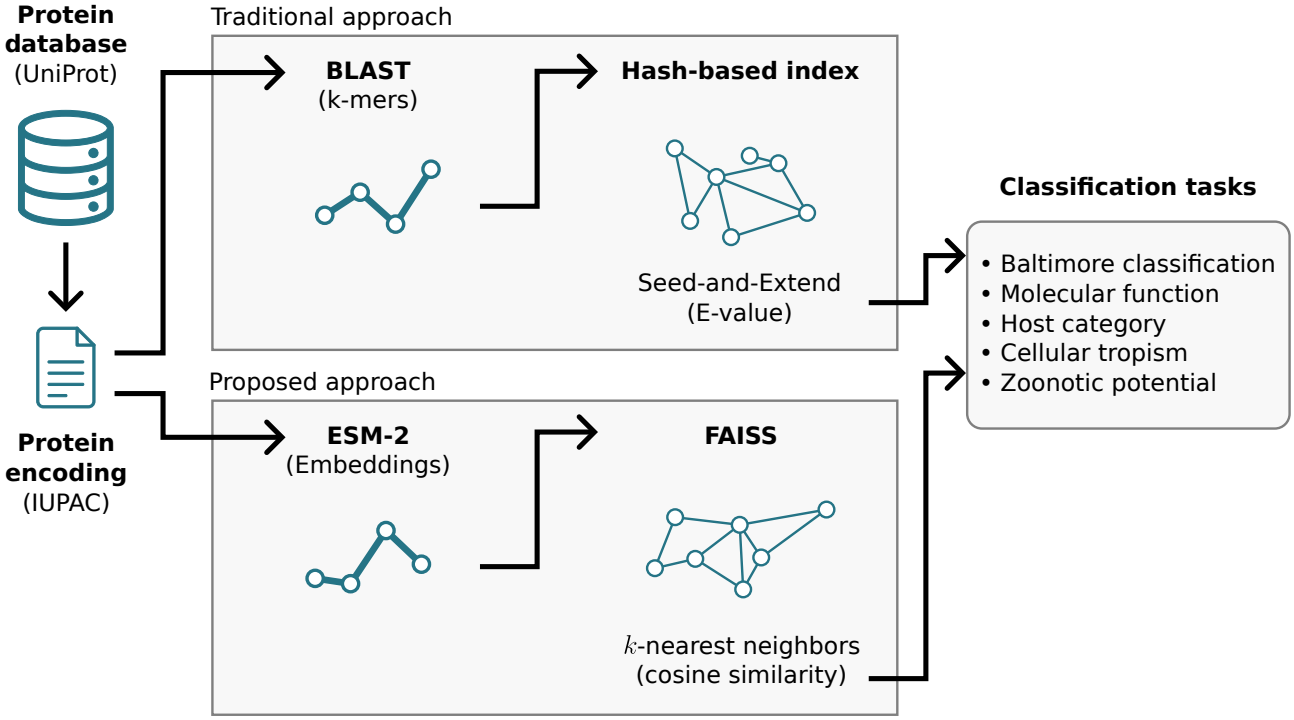

Figure 1: Comparison of the traditional and proposed approaches for protein classification.

Data

Data was gathered from the UniProt database [UniProt, 2025] using the following criteria: taxonomy ID = 10239 (to match viruses); minimum length = 64 (to avoid short peptides); maximum length = 1024 (to fit within the model's context window); reviewed = True (to avoid unverified or automated annotations); fragments = False (to avoid incomplete sequences or truncated proteins). A total of 15,154 samples were thusly gathered.

Five classification tasks were defined: Baltimore classification (identifying the genome type and replication strategy), Molecular function (specifying the biological role of the protein), Host category (defining the broad organismal group infected), Cellular tropism (identifying the specific cell or tissue types targeted), and Zoonotic potential (determining the capability for interspecies transmission between animals and humans).

For each task, samples were assigned to sub-classes via a hierarchical rule-based mapping of UniProt metadata fields, including taxonomic lineage, organismal host, and gene ontology annotations. Assignments were derived by parsing the controlled vocabularies within these fields and applying deterministic rules to resolve each sample to a single sub-class. Samples falling into uninformative categories (e.g., "General", "Unknown", or "Other") were excluded from further analysis to preserve the specificity and discriminative power of the resulting classification tasks. Figure 2 illustrates the distribution of each classification task into discrete sub-classes.

To ensure robust model evaluation, the dataset was partitioned into training, validation, and test sets using an 80/10/10 ratio. This partitioning process was conducted independently for each of the five classification tasks. For each task, a stratified sampling strategy was employed based on the specific sub-class labels to ensure that class distributions (see Figure 2) remained consistent across all three subsets despite class imbalance, thereby preventing label-specific bias during the learning and evaluation phases.

ESM-2 transformer model

Protein sequences encoded in the IUPAC single-letter alphabet were transformed into embeddings using the ESM-2 transformer model (checkpoint: `esm2_t48_15B_UR50D`) [Lin et al., 2023]. To manage the substantial

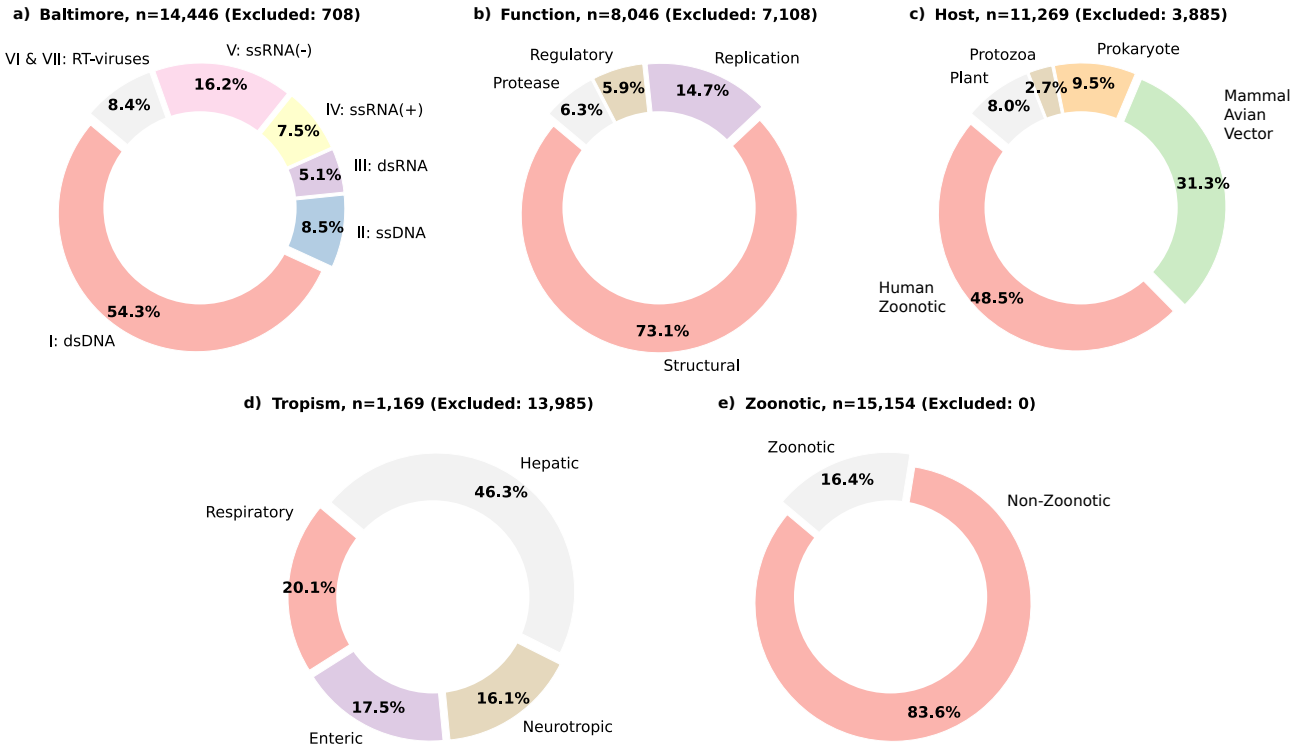

Figure 2: Distribution of protein samples across the five defined classification tasks. Each panel represents the relative frequency of sub-classes within the (a) Baltimore classification, (b) Molecular function, (c) Host category, (d) Cellular tropism, and (e) Zoonotic potential. For each task, a variable number of samples were excluded from the analysis due to assignment to an overly general and uninformative category.

memory requirements of the 15-billion parameter architecture, 4-bit NormalFloat (NF4) quantization with double quantization and a float16 compute dtype was implemented via the `bitsandbytes` library. For each protein, the last hidden state was extracted and global mean pooling was applied across the sequence length dimension to generate a fixed-length embedding vector.

Similarity search and classification

The FAISS (Facebook AI Similarity Search) library was used to perform efficient k-nearest neighbors (k-NN) retrieval within the embedding space [Johnson et al., 2019, Douze et al., 2025]. For cosine similarity tasks, vectors were L_2-normalized and indexed using an inner product flat index (`IndexFlatIP`), while Euclidean tasks utilized a squared L_2 distance index (`IndexFlatL2`). This framework enabled rapid comparison of query sequences against the training database to identify the most biologically relevant neighbors.

Hyperparameter optimization

To determine the optimal configuration for classification tasks, a grid search was conducted using the validation set across two primary hyperparameters: the distance metric (including L2 and cosine similarity, both in uniform and distance-weighted variants) and the number of nearest neighbors ($k \in 1, \ldots, 5$). Performance was evaluated across the five distinct classification tasks (Baltimore classification, Molecular function, Host category, Cellular tropism, and Zoonotic potential) using the macro-averaged F_1-score as primary metric. Distance-weighted cosine similarity with $k = 2$ emerged as the top-performing configuration across all tasks and was subsequently utilized for evaluation on the held-out test set.

Results

Table 1 compares the macro-averaged F_1-scores, evaluated on the held-out test set, between the proposed approach (ESM-2 & FAISS) and the traditional approach (BLAST). Notably, the proposed pipeline achieves a superior aggregate F_1-score across all five tasks (0.89 vs. 0.77). The performance disparity between the two methods suggests that while BLAST relies on local sequence homology, the proposed pipeline effectively leverages high-dimensional embeddings to capture more complex, non-linear biological relationships. This is

particularly evident in tasks such as Host or Tropism classification, where BLAST's F_1-scores drop significantly (0.59 and 0.78, respectively) compared to the proposed approach (0.94 and 0.98), likely due to the model's ability to identify functional patterns even in the absence of high sequence identity. Conversely, the identical F_1-score of 0.91 for Baltimore classification indicates that genome replication strategies are often tied to highly conserved motifs that both local alignment and deep embeddings can successfully retrieve. The higher F_1-score for Zoonotic classification with BLAST (0.83 vs. 0.80) suggests that zoonotic potential may be strongly linked to highly specific, conserved sequence signatures or motifs that are more effectively captured by direct local alignment than by the broader semantic generalizations of protein language model embeddings.

Figure 3 illustrates the normalized confusion matrices obtained with our approach, for all classification tasks. High values along the diagonal confirm the proposed pipeline (ESM-2 & FAISS) has high classification accuracy across most protein characterization tasks, particularly for genome replication strategies where precision reaches 0.98 for ssRNA(-) and 0.94 for dsDNA viruses. Performance remains robust for molecular functions, with Structural (0.98) and Replication (0.97) proteins showing minimal cross-prediction, and within tropism categories where Neurotropic and Hepatic targets achieve perfect diagonal scores of 1.00. However, host classification reveals notable weaknesses, specifically for Protozoa (0.68) and Plant (0.69) hosts, with significant leakage observed between Protozoa and the Mammal/Avian/Vector category (0.26). Similarly, while non-zoonotic detection is strong at 0.92, zoonotic targets experience a 0.29 leakage rate into the non-zoonotic category, suggesting shared biological signatures that complicate discrimination in these specific domains.

Table 1: Macro-averaged F_1-scores for all classification tasks, evaluated on the held-out test set.

	Baltimore	**Host**	**Function**	**Tropism**	**Zoonotic**	**All**
ESM-2 & FAISS	0.91	0.94	0.80	0.98	0.80	0.89
BLAST	0.91	0.59	0.74	0.78	0.83	0.77

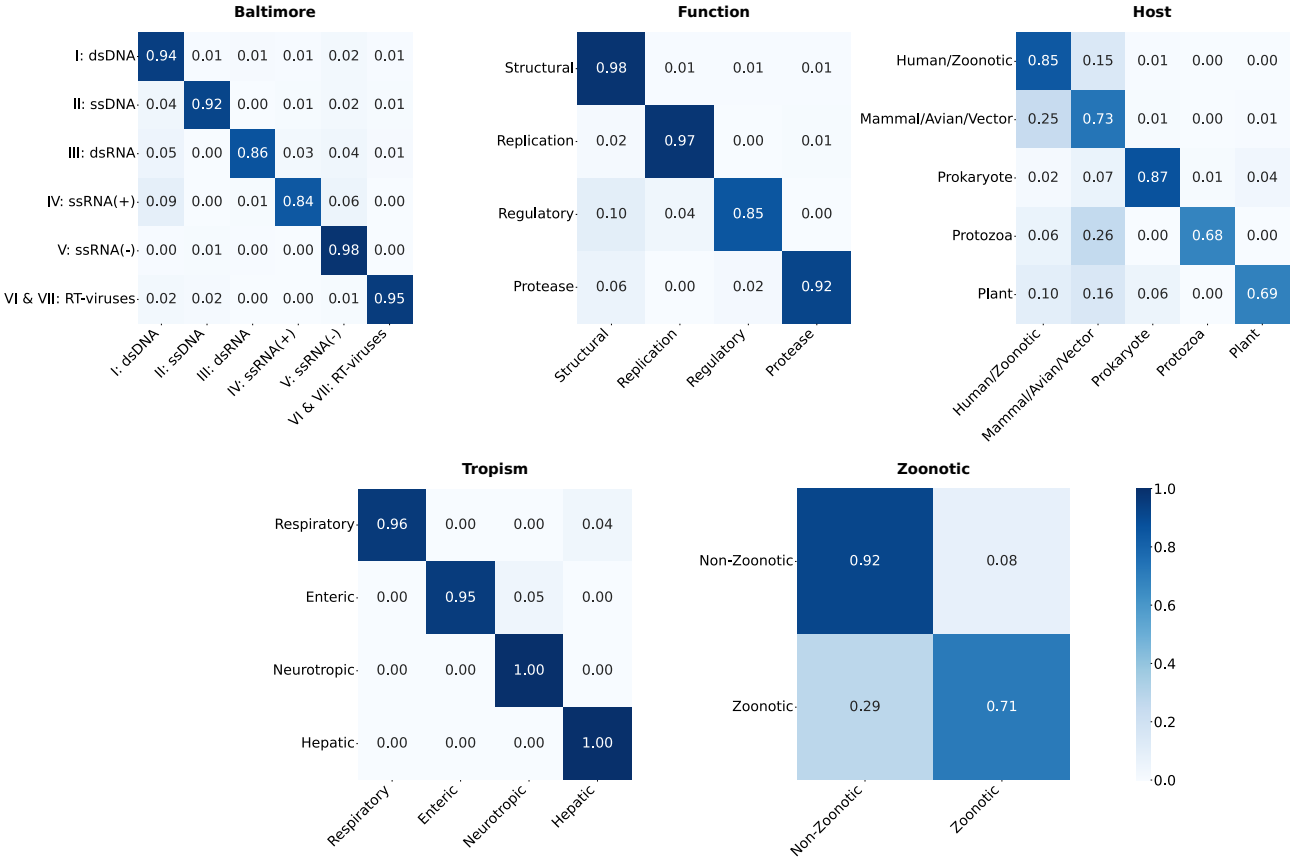

Figure 3: Normalized confusion matrices for all classification tasks, evaluated on the held-out test set.

Conclusion

The integration of ESM-2 embeddings with FAISS-based retrieval provides a significant advancement over traditional alignment methods for viral protein characterization, achieving a superior aggregate F_1-score (0.89 vs. 0.77). By capturing latent functional semantics rather than relying on exact sequence matches, the pipeline excels in complex domains such as Host (0.94) and Cellular Tropism (0.98), where sequence identity often fails to reflect biological roles. While BLAST remains effective for identifying highly conserved motifs (such as those defining Zoonotic potential), our embedding-based approach offers a more resilient defense against the evasion tactics possible with *de novo* designed sequences. By transforming amino acids into high-dimensional embeddings via a foundation transformer model, we capture the structural and functional constraints imposed by billions of years of evolution, allowing biological structure and function to emerge from scaling unsupervised learning to millions of protein sequences. This framework establishes a robust, semantically aware "ID card" for unknown proteins, offering commercial synthesis providers a scalable and context-driven guardrail essential for modern biosecurity.

Code and data

Our code is available at: https://github.com/GuillaumeZahnd/aixbio.

References

Joseph R Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. 2023.

HM Government. Uk biological security strategy. Technical report, Department for Science, Innovation and Technology, 2023. URL https://www.gov.uk/government/publications/uk-biological-security-strategy.

Dianzhuo Wang, Marian Huot, Zechen Zhang, Kaiyi Jiang, Eugene I Shakhnovich, and Kevin M Esvelt. Without safeguards, AI-biology integration risks accelerating future pandemics. *Frontiers in Microbiology*, 16:1734561, 2026.

Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

Rajesh Sathyamoorthy and Munish Puri. Protein language models outperform BLAST for evolutionarily distant enzymes: A systematic benchmark of EC number prediction. *bioRxiv*, pages 2026–03, 2026.

Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature methods*, 10(3):221–227, 2013.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

UniProt. UniProt: the universal protein knowledgebase in 2025. *Nucleic acids research*, 53(D1):D609–D617, 2025.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE transactions on big data*, 7(3):535–547, 2019.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library. *IEEE Transactions on Big Data*, 2025.