Predicting Host Types of H5N1 flu with ESM-2 Protein Language Model

Rika Chan rc3517 **Amabelle dela Pena Alcala** ada2198 **Pilar Isabel Maldonado** pim2109

1 Introduction

Understanding how influenza viruses jump between species remains a critical challenge in viral epidemiology. Recently, large language models have demonstrated significant potential for biological applications, particularly in viral evolution studies, where protein sequences serve as rich sources of evolutionary information. Evolutionary Scale Modeling (ESM) is a transformer-based protein language model trained on millions of protein sequences that learns evolutionary relationships and can predict protein structure and function from sequence alone (Lin et al., 2023). In influenza research, the H5N1 subtype of influenza A is considered a highly pathogenic avian influenza, a severe strain of flu that primarily affects birds but can also infect cattle and, occasionally, humans. There have been 70 reported cases of H5N1 flu in humans and 1 death in the United States (Centers for Disease Control and Prevention, 2025). Currently, no person-to-person transmission has been documented. Understanding which parts of the protein sequence could act as determinants of host specificity may provide insights into the transmission mechanisms that enable H5N1 to jump from other animals to humans. Here, we aim to use ESM-2 to extract embeddings from the PB2 segment, the most important segment of a viral sequence, of the H5N1 influenza subtype in order to train a baseline classifier and fine-tune the model to predict viral host types and identify sequence determinants that could be a proxy for cross-species transmission.

2 Dataset

Data was collected from GISAID (Global Initiative on Sharing All Influenza) from three hosts: human, avian, and other mammals. Because there were far less H5N1 sequences in humans, all sequences in avian and mammal hosts have been randomly downsampled to 534 sequences, resulting in 1602 sequences across all three hosts. The mean sequence length is 739.8, 753.9, and 757.5 for human, avian, mammal hosts, respectively. The data was split into 60/20/20 for train, validation, and test sets, respectively. An example of what a sequence looks like is included in the Appendix.

3 Model/Approach

ESM-2 was used to extract the embeddings for each sequence. For each protein sequence, each letter, or more professionally called residue, gets a 1280-dimensional vector that represents various properties of that residue such as its chemical and structural properties. The sequence-level embedding was then obtained by averaging all residue embeddings in that sequence.

Baseline classification: Logistic regression and random forest classifiers were trained on the frozen ESM-2 embeddings to establish baseline performance for host species classification of the three hosts.

Fine-tune approach: A classification head will be added to the pre-trained ESM-2 model, and the entire network will be fine-tuned end-to-end for host type prediction to assess whether task-specific adaptation of the model improves upon the baseline performance.

Attention: For both the baseline and fine-tune approaches, contact maps representing attention patterns between residue pairs were also extracted from ESM-2 to identify which residue positions the model considers most important for sequence classification. These attention weights are used to understand sequence specificity for each host.

4 Detailed Plan of Work

Phase I: Data pre-processing & baseline classifiers (Completed)

- H5N1 PB2 sequences were obtained from GISAID. Pre-processing included concatenating all sequences from all hosts into one FASTA file with host labels. Avian and mammals sequences were randomly

downsampled to 534 sequences to match available H5N1 sequences found under humans.

- ESM-2 were used to extract the embeddings from the input data. The embeddings were then mapped onto a 2D space using the UMAP algorithm.
- Logistic regression and random forest were used as baseline classifiers with the ESM-2 embeddings as frozen features.

Phase II: Fine-tuning of ESM-2 for the classification task

- A classification head will be added to the pre-trained ESM-2 model to test if task-specific fine-tuning improves upon the baseline.
- Implement end-to-end fine-tuning with low learning rates to preserve the pre-trained knowledge while adapting to host specificity task.
- Evaluation will be done on learning rates, batch sizes, sequence length handling, and regularization techniques. Use validation set for early stopping and model selection to prevent overfitting on the limited dataset.

Phase III: Attention weights analysis

- Extract and visualize attention weights from ESM-2's final layer to identify residue-residue interactions deemed important by the model. This is done using the ESM-2's built-in contact maps function.
- Compare attention patterns between host types to identify regions of sequences of host specificity, and quantify the differences using statistical measures.
- If the attention weights are not interpretable, SHAP will be use to determine the most important regions of the sequences per each host. Preliminary biological validation will also be done.

Phase IV: Document results

- Gather findings to assess ESM-2's ability to capture host-specific evolutionary signatures in H5N1 PB2 sequences.
- Document performance improvements over the baseline and biological insights gained from attention analysis.

5 Preliminary Results

UMAP (Uniform Manifold Approximation and Projection) was used to project the sequence embeddings obtained from ESM-2 to a 2D space (McInnes et al., 2018).

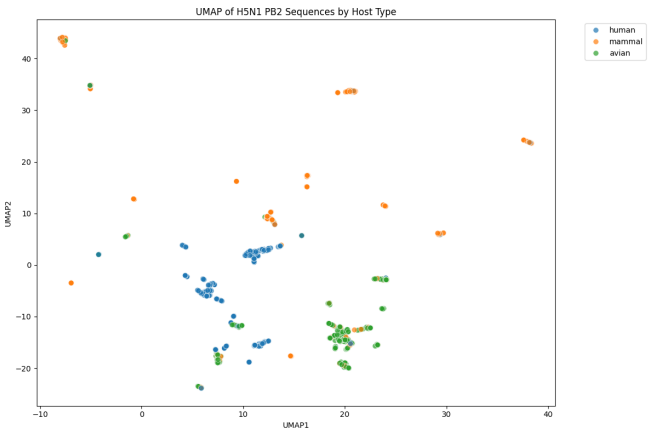

Figure 1: UMAP embedding of sequence embeddings obtained from ESM-2 colored by host types.

	precision	recall	f1-score
avian	0.73	0.93	0.81
human	0.96	0.73	0.83
mammal	0.88	0.86	0.87
accuracy			0.84
macro avg	0.86	0.84	0.84
weighted avg	0.86	0.84	0.84

Figure 2: Precision, recall, and F1 score for logistic regression classifier.

	precision	recall	f1-score
avian	0.90	0.91	0.90
human	0.94	0.93	0.93
mammal	0.89	0.90	0.89
accuracy			0.91
macro avg	0.91	0.91	0.91
weighted avg	0.91	0.91	0.91

Figure 3: Precision, recall, and F1 score for random forest classifier.

The UMAP projection of the sequence embedding showed somewhat a distinction between the different host types, suggesting that it could be linearly separable and simple models could distinguish between the three hosts (fig. 1). As such, the baseline results using just logistic regression and random forest classifiers showed a decent performance, with random forest being the better classifier, for distinguishing between different host types using the ESM-2 sequence embeddings (fig. 2 & 3).

6 Related Work

Several studies have leveraged machine learning approaches for influenza virus host classification, each deploying different strategies that present distinct limitations. Custom bidirectional long short-term memory (LSTM) and transformer models were trained from scratch on 848,630 influenza sequences including PB2 segments, treating viral sequences as natural language text by tokenizing nucleotides and residues for input to standard NLP architectures (Hatibi et al., 2023). They achieved good classification performance and discovered that mRNA sequences contained more predictive information than protein sequences. However, their method relied on training domain-specific models without leveraging evolutionary knowledge. It also required expensive computational resources and large datasets.

Alberts et al. (2024) applied a gradient boosting machine classifier to H3 influenza hemagglutinin (HA) segment sequences, achieving 98.0% accuracy for host prediction using ensemble methods. However, their methodology was limited to a single viral subtype and used traditional features extraction approaches, making it difficult to generalize across different influenza strains and capture complex evolutionary relationships.

Most recently, EvoMIL, which combined pre-trained ESM-1b protein language model embeddings with multiple instance learning for virus-host prediction, representing a significant advancement by leveraging evolutionary information (Liu et al., 2024). EvoMIL achieved substantial performance improvements, with median F1 score gains of 4.9% to 16.2% across different host categories compared to basic statistical representations of protein sequences. However, their multiple instance learning framework treated each virus as a "bag" of proteins from the entire genome, making it challenging to extract attribute host-specific regions. Additionally, this approach used frozen embeddings from ESM-1b rather than fine-tuning it, potentially limiting the model's ability to learn task-specific representations for host classification.

Our work addresses these specific limitations through several key methods specifically for the H5N1 subtype, a strain that recently has been infecting humans. First, we leverage the more advanced ESM-2 model with task-specific fine-tuning, allowing the model to adapt its pre-trained evolutionary knowledge, specifically for H5N1 host classification. Second, our method automatically learns relevant features through end-to-end training and focuses specifically on H5N1, rather having to do extensive feature engineering. Third, our method provides targeted analysis of the PB2 segment, which is crucial due to its primary role in viral replication, enabling more informative extraction of evolutionary signatures. Additionally, our attention weight analysis provides interpretability specifically for understanding the molecular basis of H5N1's cross-species transmission potential.

7 Team Contribution Statement

Rika Chan is responsible for obtaining, pre-processing data, running the baseline classification, and analyzing the final attention weights of the fine-tune approach. She completed sections 1,2,3, and 5 of the mid-project review. Pilar Maldonado is fine-tuning the model and evaluating results of the fine-tune approach as well as the extracting the attention weights. She contributed to sections 4 and 7 of the review. Amabelle dela Pena Alcala will interpret the final results according to known literature and comparison to the baseline and completed section 6 of this paper.

References

Famke Alberts, Olaf Berke, Grazieli Maboni, Tatiana Petukhova, and Zvonimir Poljak. Utilizing machine learning and hemagglutinin sequences to identify likely hosts of influenza h3nx viruses. *Preventive Veterinary Medicine*, 233:106351, 2024. doi: 10.1016/j.prevetmed.2024.106351. URL `https://www.sciencedirect.com/science/article/pii/S016758772400237X`.

Centers for Disease Control and Prevention. H5 bird flu: Current situation, November 2025. URL `https://www.cdc.gov/bird-flu/situation-summary/index.html`. Accessed November 2025.

Nissrine Hatibi, Maude Dumont-Lagacé, Zakaria Alouani, Rachid El Fatimy, Mounia Abik, and Tariq Daouda. Misclassified: identification of zoonotic transition biomarker candidates for influenza a viruses using deep neural network. *Frontiers in Genetics*, 14:1145166, July 2023. doi: 10.3389/fgene.2023.1145166. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10415530/`.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023. doi: 10.1126/science.ade2574. URL `https://www.science.org/doi/abs/10.1126/science.ade2574`.

Dan Liu, Francesca Young, Kieran D. Lamb, David L. Robertson, and Ke Yuan. Prediction of virus-host associations using protein language models and multiple instance learning. *PLOS Computational Biology*, 20(11):e1012597, November 2024. doi: 10.1371/journal.pcbi.1012597. URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012597`.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, September 2018. doi: 10.21105/joss.00861. URL `https://joss.theoj.org/papers/10.21105/joss.00861`.

Appendix

```
>EPI4586312|PB2|A/dairy_cow/USA/019706-005/2025|EPI_ISL_20094669|A_/_H5N1|HOST_2
MERIKELRDLMSQSRTREILTKTTVDHMAIIKKYTSGRQEKNPALRMKWMMAMKYPITADKRIMEMIPERNEQGQTLWSK
TNDAGSDRVMVSPLAVTWWNRNGPTTSTIHYPKVYKTYFEKVERLKHGTFGPVHFRNQIKIRRRVDINPGHADLSAKEAQ
DVIMEVVFPNEVGARILTSESQLTITKEKKEELQDCKIAPLMVAYMLERELVRKTRFLPVAGGTSSVYIEVLHLTQGTCW
EQMYTPGGEVRNDDVDQSLIIAARNIVRRATVSADPLASLLEMCHSTQIGGIRMVDILRQNPTEEQAVDICKAAMGLRIS
SSFSFGGFTFKRTSGSSVKREEEVLTGNLQTLKIRVHEGYEGFTMVGRRATAILRKATRRLIQLIVSGRDEQSIAEAIIV
AMVFSQEDCMIKAVRGDLNFVNRANQRLNPMHQLLRHFQKNAKVLFQNWGIEPIDNVMGMIGILPDMTPSTEMSLRGIRV
SKMGVDEYSSTERVIVSIDRFLRVRDQRGNVLLSPEEVSETQGTEKLTITYSSSMMWEINGPESVLVNTYQWIIRNWETV
KIQWSQDPTMLYNKMEFEPFQSLVPKAARGQYSGFVRTLFQQMRDVLGTFDTVQIIKLLPFAAAPPEQSRLQFSSLTVNV
RGSGMRILIRGNSPVFNYNKATKRLTVLGRDAGALAEDPDEGTAGVESAVLRGFLILGKEDKRYGPALSINELSNLAKGE
KANVLIGQGDVVLVMKRKRDSSILTDSQTATKRIRMATN
```

Example of a sequence. The header starts with a > sign and denotes sequence ID, viral segment, species and collection date, database ID, flu type, and host type. HOST_0, HOST_1, HOST_2 denotes human, avian, and other mammals, respectively. Each letter in the sequence denotes a residue or also called an amino acid.