Predicting Host Types of H5N1 flu with ESM Language Models

Rika Chan, Amabelle dela Pena Alcala, Pilar Isabel Maldonado

November 07, 2025

1 Project description

Language models have demonstrated significant potential for biological applications, particularly in viral evolution studies, where protein sequences serve as rich sources of evolutionary information. In this project, we leverage transformer-based protein language models developed by Facebook's FAIR team to predict viral host types directly from protein sequences of the H5N1 flu subtype (Github repo). We also analyze the attention mechanisms of the transformer model to identify regions of the sequences that could serve as transmission pathways between hosts. The H5N1 subtype is considered a highly pathogenic avian influenza, a severe strain of flu that primarily affects birds but can also infect cattle and, rarely, humans. There have been 70 reported cases of H5N1 flu in humans and 1 death (CDC, 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 animals to humans.

2 Previous works

Recent studies have applied protein language models to predict viral hosts directly from sequence data. For example, EvoMIL and HostNet use transformer-based embeddings to achieve high accuracy in virus–host classification (Liu et al. 2024, Ming et al. 2023). In influenza research, PB2 mutations such as E627K and D701N have been identified as key determinants of host adaptation. These findings highlight that pretrained sequence models can reveal biologically meaningful host-specific patterns and improve predictive performance.

3 Plans of accomplishment

At minimum, we aim to develop a robust and accurate model for classifying viral host types. If successful, we will further analyze the model's attention mechanisms to uncover sequence regions or variants most associated with each host. With additional time, we plan to validate these findings against known host-adaptive mutations to strengthen biological interpretability.

4 Measuring success

Success will be measured by how accurately the model predicts viral host types, evaluated using metrics such as AUROC, precision, F1, recall, and accuracy on test splits to ensure generalization beyond sequence similarity. The model should also highlight sequence regions consistent with known host-adaptive mutations, demonstrating biological interpretability. Ultimately, success means achieving strong predictive performance while uncovering biologically meaningful insights into host specificity mechanisms in H5N1.

5 Expected obstacles

1. Class imbalance: there are far less cases of H5N1 in humans than other mammals and birds. Hence, fine-tuning large models like ESM with limited H5N1-specific data risks overfitting unless techniques like low-rank adaptation (LoRA) or contrastive pretraining are applied.
2. Attention weights in transformers are not guaranteed to correspond to biological relevance. Inferring "causal" host-determining regions from attention maps requires careful validation.

6 Citations

1. CDC. (2025, September 16). H5 bird flu: Current situation. Avian Influenza (Bird Flu). https://www.cdc.gov/bird-flu/situation-summary/index.html

2. Facebookresearch/esm. (2025). [Python]. Meta Research. https://github.com/facebookresearch/esm (Original work published 2020)

3. Liu, D., Young, F., Lamb, K. D., Robertson, D. L., & Yuan, K. (2024). Prediction of virus-host associations using protein language models and multiple instance learning. PLOS Computational Biology, 20(11), e1012597. https://doi.org/10.1371/journal.pcbi.1012597

4. Ming, Z., Chen, X., Wang, S. et al. HostNet: improved sequence representation in deep neural networks for virus-host prediction. BMC Bioinformatics 24, 455 (2023). https://doi.org/10.1186/s12859-023-05582-9