Exploring the antigenic space of Influenza A subtype H5 using Recurrent Neural Networks (RNN) and Transformers.

Ricardo Rivero, Washington State University.

December 12, 2024

Introduction

Influenza A is one of the most important human pathogens, having caused five pandemics in the last century [1]. The main key to its success is its high evolutionary rate, with an estimate of 3 to 4 amino acid mutations emerging per year in its hemagglutinin (HA) protein [7]. HA is the main antigen and cell entry protein for influenza A, and evolutionary successful clades are characterized by greater antigenic changes that confer these viruses with immune escape, increased ability to infect distant animal species, and increased transmissibility through mechanisms such as enhanced environmental stability and changes in α-2,3/2,6-linked sialic acid usage [2;3]. Experimental efforts have been made to characterize the effect of mutations in HA's ability to escape the immunity elicited by vaccination and prior infection, resulting in deep-mutational scan datasets that establish tradeoffs between escape-conferring mutations and viral fitness. In this context, fitness can be defined as the ability of a genotype to be efficiently transmitted at a population level, given environmental constraints such as existing immunity and the presence or absence of competing genotypes [3]. It has been previously demonstrated that these mutations are constrained by Influenza A's sequence space, which consists of the substitutions that the virus can accumulate through evolutionary processes without losing the protein's functionality, which yields that the number of possible sequences is lower than the number of possible proteins [10].

With the re-emergence of highly pathogenic avian influenza (HPAI) strains such as H5N1 2.3.4.4b, understanding the ability of a novel genotype to escape the existing vaccine and infection-elicited immunity is crucial for pandemic preparedness. In that sense, we propose to leverage the existing genetic diversity of influenza A, subtype H5Nx to train a Recurrent Neural Network model that takes protein sequences, classified in terms of antigenic cartographies to predict the antigenic distance of an unobserved sequence and its potential for immune escape. In this study, we also tested the applicability of state-of-the-art (SOTA) pretrained protein language models (PLMs) as a method for representing the protein sequences as structure-free embeddings that can be used as an input for transformed-based regression of the escape values of H5 genotypes. Our results show that LSTM-based architectures are suitable for the generation of novel Influenza H5 sequences that could be used for drug-screening. Furthermore, we employed a transformer-based architecture to predict antigenic escape values from naturally-occurring sequences, obtaining a Spearman correlation of 0.6299, demonstrating that our implementation is able to perform at the level more sophisticated PLM-based variant effect predictors (Spearman correlation = 0.69).

Methods.

Data acquisition and curation.

A total of 31,662 HA sequences from the H5 subtype of Influenza A were downloaded from the Global Initiative for Sharing All Influenza Data (GISAID) Epiflu platform [4]. This dataset, consisting of protein sequences, was processed to avoid overrepresentation of 2.3.4.4b genotypes derived from the current outbreak by performing a genetic similarity-based clustering in CD-HIT [8], where several similarity thresholds were tested to determine the effect of genetic diversity on the performance of the model. All sequences with a completeness lower than 70% were excluded from the analysis.

Data preprocessing.

The resulting dataset was aligned using MAFFT and transformed into a 20-dimensional space, where each possible aminoacidic value was encoded as a number from 0 to 19 before performing a non-metric multidimensional scaling to represent the sequences in a 2D space. Then, an antigenic cartography was built using the genetic and serological data from Dadonaite et al.[3] (available at `https://github.com/dms-vep/Flu_H5_American-Wigeon_South-Carolina_2021-H5N1_DMS/tree/main`). The resulting immune escape value was used as the model output.

Embedding representation of the protein sequences.

To obtain embedding representations of the protein sequences that could be used to train a transformer, we used two pre-trained protein language models based on Meta's ESM[9], namely, esm2_t6_8M_UR50D, and esm2_t30_150M_UR50D

As an alternative method, we used the escape values experimentally determined by Dadonaite et al[3], which are based on the evaluation of the effect of point mutations in pseudovirus neutralization using ferret sera. We further divided this into two methods:

1. Synthetic library of single-mutation Influenza Sequences: Based on the A/American Wigeon/South Carolina/USDA-000345-001/2021 strain of H5N1 Influenza, we generated one sequence per mutation per site, resulting in a total of 10,811 sequences. To each of these sequences, the antibody escape value associated with the specific point-mutation was assigned.

2. Natural Influenza Genotypes: Using the Influenza sequences that remained after the similarity clustering, we calculated the total escape value by identifying the mutations with respect to the backbone sequence, and assuming an additive effect of presenting multiple mutations, we summed the individual escape value of the mutations to obtain the escape value of each H5 genotype.

Long Short-Term Memory (LSTM) Model for Protein Sequence Prediction.

Data Preparation

The dataset consisted of aligned H5 hemagglutinin protein sequences stored in FASTA format. The sequences were parsed, and each amino acid was encoded into a numeric format using a mapping of amino acids to integers (A=1, R=2, ..., V=20). Unknown amino acids were assigned a value of 0. The sequences were padded to ensure a uniform length corresponding to the maximum sequence length across the dataset. Each sequence was transformed into input-output pairs for training, where the input consisted of all subsequences up to a given position, and the output was the next amino acid in the sequence. This preprocessing resulted in a dataset of shape (num_samples, max_sequence_length), and one-hot encoding was applied to the output labels.

Model Architecture.

The LSTM model was implemented using TensorFlow/Keras. It included the following components:

1. Embedding Layer: Transformed input integers into dense vector representations of dimension 64. Padding was masked to prevent it from influencing training.

2. Bidirectional LSTM Layer: A bidirectional LSTM with 128 units processed the input sequence, allowing the model to capture both forward and backward dependencies.

3. Dense Output Layer: A fully connected dense layer with a softmax activation function was used to predict the probability distribution over 21 classes (20 amino acids and one padding/unknown class).

Training.

The model was compiled using the Adam optimizer and categorical cross-entropy loss. The training was performed for 10 epochs with a batch size of 256, and 10% of the dataset was used for validation. The training and validation accuracy were tracked to assess model performance. The training was done using two NVIDIA T4 GPUs in `https://www.kaggle.com/`

Sequence Prediction.

A function was developed to predict the next amino acid in a sequence by feeding a given sequence into the trained model. The predicted amino acid corresponded to the class with the highest probability in the model's output.

Sequence Generation.

A custom function was implemented to generate 20 full-length HA protein sequences using the trained model. Briefly, a seed sequence consisting in the first 20 amino acids of the reference protein were provided, then, we iteratively predicted the next amino acid. We used a temperature of 0.8 to control the randomness of predictions, allowing for the exploration of sequence diversity.

Sequence Identity Evaluation.

Generated sequences were compared to sequences from the training set to assess their similarity. Pairwise global alignments were performed using the pairwise2 module from Biopython, and sequence identity was calculated as the percentage of matching amino acids between aligned sequences.

Phylogenetic investigation of the generated sequences.

The generated sequences were aligned to all the sequences from the clustered dataset using MAFFT [6], and a maximum-likelihood phylogenetic tree was built using IQ-TREE2 [11] under a FLU+R5 protein substitution model. The resulting tree was rooted to the midpoint and visualized using Baltic `https://github.com/evogytis/baltic/tree/master`.

Transformer Architecture for Escape Probability Prediction.

To predict escape probabilities of Influenza H5 sequences, we implemented a transformer-based architecture trained on sequence embeddings derived from a pretrained model. Below is a detailed description of the model architecture and training procedure:

Input Representation.

The input data consisted of protein sequence embeddings with dimensions corresponding to (sequence_length, embedding_dimension). These embeddings were derived from a pretrained protein language model. The escape probabilities, provided as continuous values ranging from approximately -2.5 to 7.5, were normalized to the range [-1, 1] using MinMax scaling for numerical stability during training.

Positional Encoding.

A custom positional encoding layer was implemented to incorporate sequence order into the model. This layer adds sinusoidal positional information to the input embeddings, ensuring that the transformer could capture positional relationships between amino acids. The positional encoding was defined as follows:

$$\mathrm{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d}}\right), \quad \mathrm{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d}}\right),$$

where *pos* is the position, i is the dimension index, and d is the embedding dimension.

Transformer Encoder.

The core of the model was a stack of two transformer encoder blocks. Each block consisted of:

1. **Multi-Head Self-Attention:** This module allowed the model to attend to all positions in the input sequence simultaneously. Eight attention heads were used, with the embedding dimension as the key dimension. Regularization was applied through L2 penalties on the kernel and bias weights.

2. **Feed-Forward Network:** Following the attention mechanism, a position-wise feed-forward network was implemented, consisting of two dense layers. The first layer expanded the embedding dimension to a feed-forward dimension of 256, followed by a ReLU activation, while the second layer projected the output back to the original embedding dimension.

3. **Residual Connections and Layer Normalization:** Residual connections and layer normalization ensured stable training and preserved gradient flow. Dropout with a rate of 0.1 was applied after the attention and feed-forward layers for additional regularization.

Output Layer.

After the encoder blocks, the sequence outputs were pooled using global average pooling to reduce the sequence dimension. A dense layer with a linear activation function was used to predict the escape probability for each sequence.

Training Procedure.

The model was trained using the GPU of an Apple M2 Pro with the Keras implementation of the Metal Performance Shaders (MPS) using the following procedure:

- Loss Function: The mean squared error (MSE) was used as the primary loss function.
- Metrics: The model was evaluated using the mean absolute error (MAE) and root mean squared error (RMSE).
- Optimizer: The Adam optimizer with a learning rate of 10^{-4} and gradient clipping at 1.0 was employed to stabilize training.
- Hyperparameters: The model was trained for 50 epochs with a batch size of 32. Early stopping was not applied in this training setup.

Evaluation.

The model's performance was evaluated on a validation set split from the dataset. After training, the predicted escape probabilities were denormalized back to their original range for downstream analysis. Spearman's rank correlation coefficient was computed to assess the relationship between true and predicted escape probabilities.

This architecture allowed the model to effectively learn from sequence embeddings while capturing sequence order and positional dependencies crucial for accurate escape probability predictions.

Results and discussion.

LSTM can generate novel Influenza H5 sequences.

A total of 20 novel sequences were generated using the trained LSTM, Protein-BLAST remote homology search showed that all of these sequences belonged to the H5 subtype but were significantly different to all of the sequences in the training data, with a mean pairwise similarity of 92.32%. Next, we set to explore the pairwise similarity of the generated sequences against the complete H5 HA dataset. We found that these sequences maintain the similarity/dissimilarity patterns found across the complete H5 subtype, with similarity percentages ranging from 58.21% to 98.48% as shown in Fig 1.

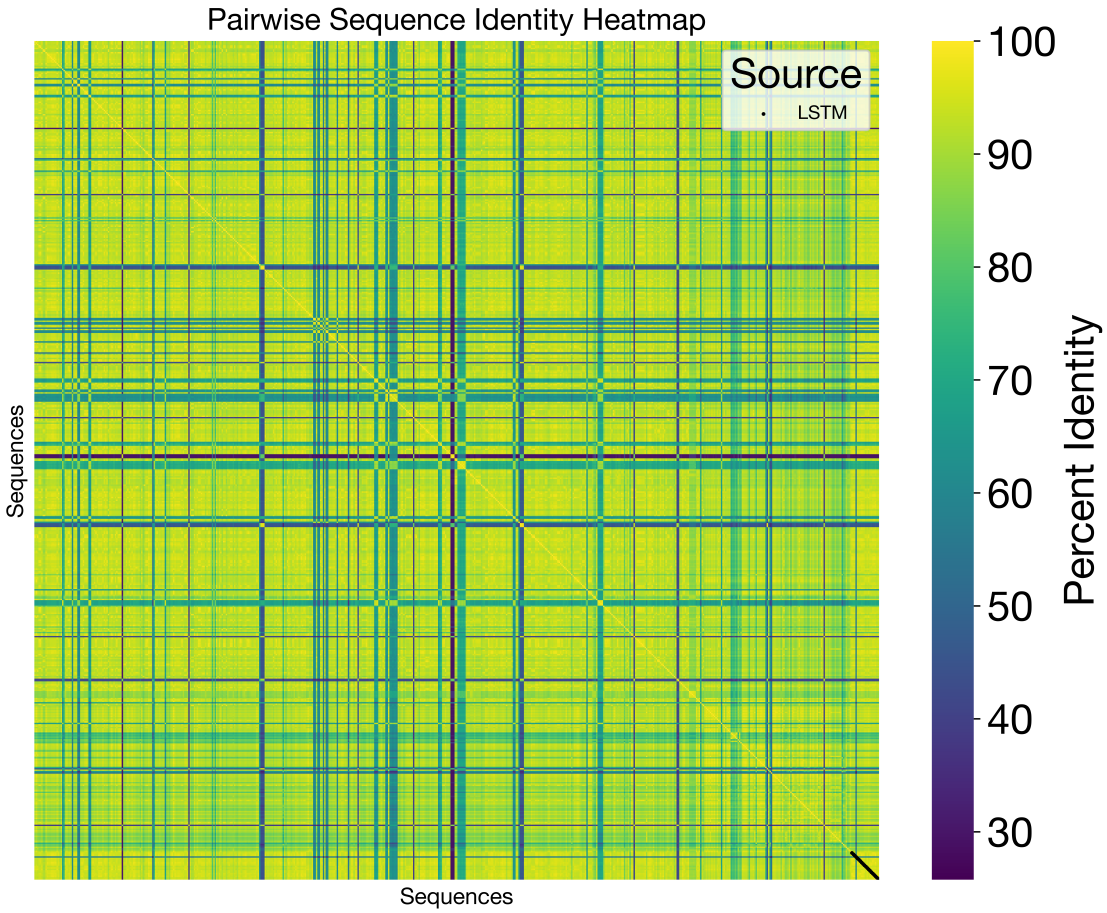

Figure 1: Pairwise sequence similarity matrix of the LSTM generated sequences and the natural H5 dataset.

LSTM generated Influenza sequence do not follow a clustering pattern.

Next, we sought to identify if the generated sequences were phylogenetically clustered. In this case, a single cluster would indicate that the model is not able to explore the sequence diversity in the generation process, which would suggest underfitting because it would fail to generate a new sequence based on the learned relationships between the amino acid positions.

From the phylogenetic tree we found that using a set temperature value of 0.8, the iterative generation process managed to produce distinct sequences that were scattered across the phylogenetic tree. Although no formal hypothesis testing was done, we expect a well-performing model to generate sequences that result in random placements in the phylogenetic tree as the number of generated sequences increases. In Fig 2, we can see that the 20 sequences generated by the LSTM are scattered across 6 distinct phylogenetic clusters.

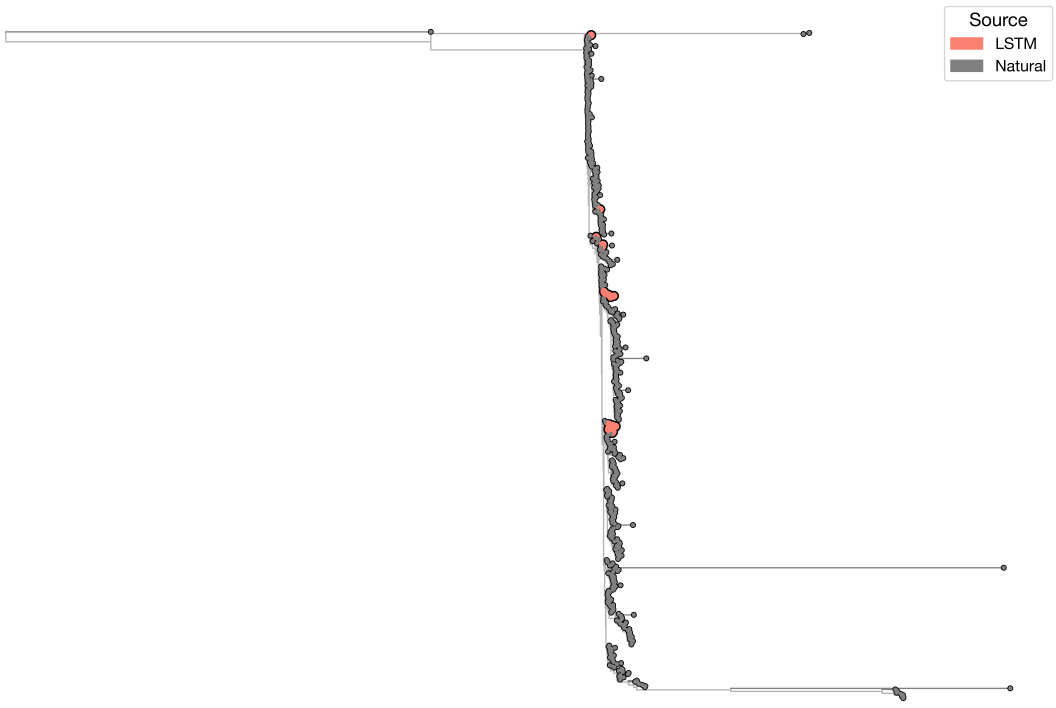

Figure 2: Maximum-likelihood phylogenetic tree of the LSTM-generated HA sequences in the context of the current Natural diversity of the H5 subtype. The generated sequences are represented in salmon color while the natural sequences are represented in gray.

LSTM is not scalable for large-scale applications.

During the training, we noticed that the training was slow, even with the use of two NVIDIA T4 GPUs. With that in mind, we tested the time that the model takes to generate a protein sequence as a function of the expected sequence length. Sequences with length ranging from 50 to 568 AA in increments of 50. Our results showed that for a full-length sequence, the model takes 41.017 seconds. With these results (Fig 3.), we deemed the LSTM inapplicable for the prediction task, and switched to a transformer architecture.

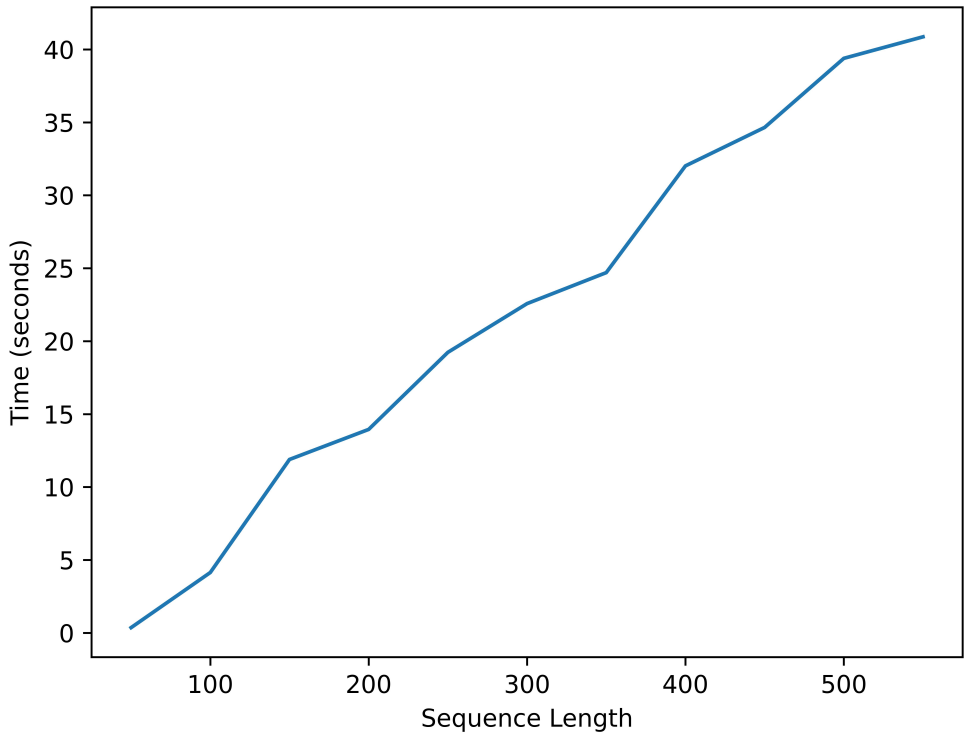

Figure 3: Runtime of the LSTM for protein sequence generation tasks as a function of the desired sequence length.

Transformer-based model successfully predicts antibody escape from ESM-encoded natural sequences.

We first tried to get embeddings from the synthetic dataset of point-mutated sequences using both 8M and 150M parameter versions of ESM2, however, neither of these implementations were suitable for fine tuning the transformer, with a Spearman's correlation in the order of 1e-2 (data not shown). We hypothesize that this poor performance is due to the nature of the sequence embeddings, which are encoded based on sequences that are only 1 mutation apart from each other, in this sense, the embeddings would not be different enough to be confidently learned by the model. To address this, we used the same transformer architecture, but this time, we calculated total escape values for the natural sequences collected from GISAID. For this, we assumed the effect of the mutations to be additive in terms of antibody escape. The total escape was then, the sum of all individual escape values associated with the detected mutations in respect with the reference sequence. For this implementation, we obtained a validation loss of 0.0487, and a Spearman's correlation of 0.6299 with a P-value of 1.6602e-14 between the predicted and true immune escape values (Fig 4).

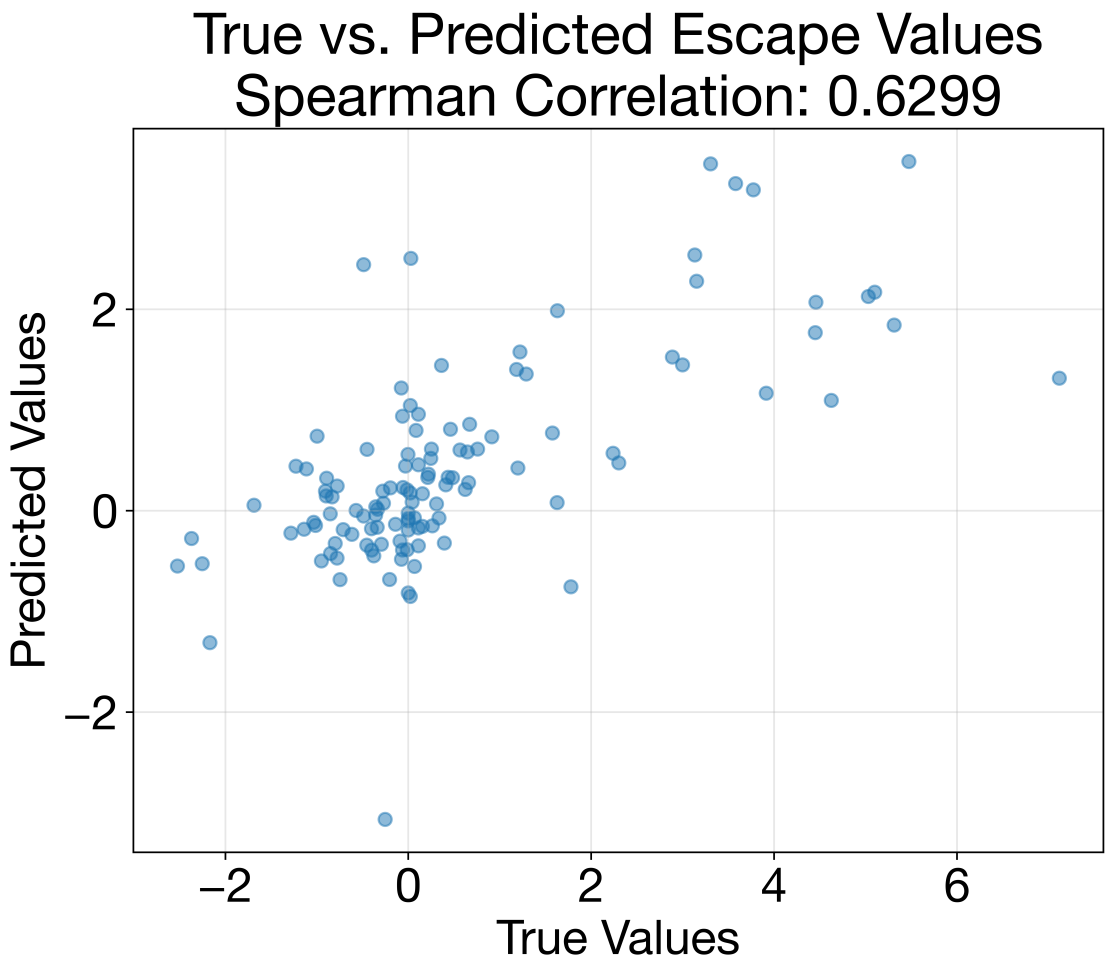

Figure 4: True versus predicted escape values using the transformer model trained on ESM2 150M embeddings. Spearman's rank correlation is shown at the top of the plot.

A limitation of these study is that the effect of multiple escape-conferring mutations is not additive at the phenotypic level. In reality, higher-order interactions such as epistatic interactions, and structural changes might arise, affecting the susceptibility to neutralization in a non-additive way. Furthermore, protein stability is a tradeoff that often limits the antigenic space of viral glycoproteins, and in that sense, not all of the mutations characterized by Dadonaite et al[3] may emerge in nature. In further studies, we expect to integrate global epistatic models and structural embeddings to obtain a better representation of the association between sequence, structure, and immune escape. Overall, the transformer had a good performance in predicting the escape values from the test sequences. Other authors have fine-tuned ESM2-150M for variant effect prediction in different proteins, obtaining Spearman correlation values from 0.69 to 0.88[12]. This suggests that our implementation of the transformer architecture achieved a comparable performance to that of more elaborate fine-tuned models.

The evasion landscape of H5 Hemagglutinin.

We sought to identify the sites in the protein sequence that had the highest effect on immune escape. The results showed that mutations in antigenic regions A and B of the HA1 subdomain of H5 Hemagglutinin had the largest effect in neutralization sensitivity, either to decrease it or to make the phenotype more susceptible to antibody neutralization. These results are in line with neutralization studies that have identified these regions as part of the receptor-binding motif of HA, which is the main target for neutralizing antibodies[5;13].

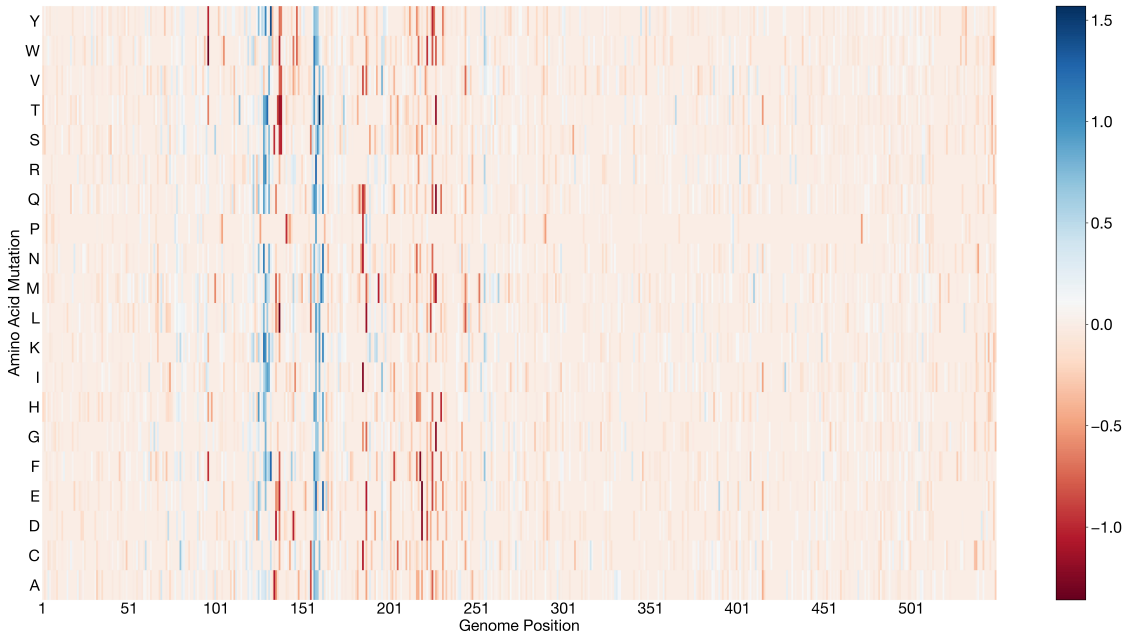

Figure 5: Heatmap of the effect of amino acid mutations in the immune escape of H5 HA.

Lessons learned.

During the execution of this project, I learned about embeddings, pre-trained protein language models, recurrent neural networks, and transformers. For the completion of this machine-learning endeavor, I had to test different sequence datasets, preprocessing (clustering thresholds, one-hot encodings, embeddings, paddings, etc.), and PLMs such as ESM2-8M, ESM-150M, and ESMC-600M. Through this journey, I learned that different PLM complexities are able to extract embeddings at different levels of details, with ESM-150M performing the best among the three PLMs tested.

Finally, the take-home message for me is that with domain expertise, we can integrate biological and machine-learning concepts to solve tasks that are of vital importance for public health, such as determining the potential that a novelly identified genotype has to escape from current vaccine formulations. In further iterations, I expect to enhance this model to account for the non-additive effects of mutation-saturated genotypes in immune escape.

References

[1] Orthomyxoviridae | ICTV. URL `https://ictv.global/report_9th/RNAneg/Orthomyxoviridae`.

[2] Leonardo C. Caserta, Elisha A. Frye, Salman L. Butt, Melissa Laverack, Mohammed Nooruzzaman, Lina M. Covaleda, Alexis C. Thompson, Melanie Prarat Koscielny, Brittany Cronk, Ashley Johnson, Katie Kleinhenz, Erin E. Edwards, Gabriel Gomez, Gavin Hitchener, Mathias Martins, Darrell R. Kapczynski, David L. Suarez, Ellen Ruth Alexander Morris, Terry Hensley, John S. Beeby, Manigandan Lejeune, Amy K. Swinford, François Elvinger, Kiril M. Dimitrov, and Diego G. Diel. Spillover of highly pathogenic avian influenza h5n1 virus to dairy cattle. pages 1–8. ISSN 1476-4687. doi: 10.1038/s41586-024-07849-4. URL `https://www.nature.com/articles/s41586-024-07849-4`. Publisher: Nature Publishing Group.

[3] Bernadeta Dadonaite, Jenny J. Ahn, Jordan T. Ort, Jin Yu, Colleen Furey, Annie Dosey, William W. Hannon, Amy L. Vincent Baker, Richard J. Webby, Neil P. King, Yan Liu, Scott E. Hensley, Thomas P. Peacock, Louise H. Moncla, and Jesse D. Bloom. Deep mutational scanning of h5 hemagglutinin to

inform influenza virus surveillance. URL https://www.biorxiv.org/content/10.1101/2024.05.23.595634v2. Pages: 2024.05.23.595634 Section: New Results.

[4] GISAID. GISAID initiative. 2008:1–7. URL https://www.epicov.org/epi3/frontend#41d4ce.

[5] Yanmei Hu, Hannah Sneyd, Raphael Dekant, and Jun Wang. Influenza a virus nucleoprotein: a highly conserved multi-functional viral protein as a hot antiviral drug target. 17(20):2271–2285. ISSN 1568-0266. doi: 10.2174/1568026617666170224122508. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5967877/.

[6] Kazutaka Katoh and Martin C. Frith. Adding unaligned sequences into an existing alignment using MAFFT and LAST. 28(23):3144–3146. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTS578. URL https://academic.oup.com/bioinformatics/article/28/23/3144/193620. Publisher: Oxford Academic.

[7] Juhye M. Lee, John Huddleston, Michael B. Doud, Kathryn A. Hooper, Nicholas C. Wu, Trevor Bedford, and Jesse D. Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human h3n2 influenza variants. 115(35):E8276–E8285. doi: 10.1073/pnas.1806133115. URL https://www.pnas.org/doi/10.1073/pnas.1806133115. Publisher: Proceedings of the National Academy of Sciences.

[8] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. 22(13):1658–1659. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl158. URL https://doi.org/10.1093/bioinformatics/btl158.

[9] 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. 379(6637):1123–1130. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574. Publisher: American Association for the Advancement of Science.

[10] John Maynard Smith. Natural selection and the concept of a protein space. 225(5232):563–564. ISSN 1476-4687. doi: 10.1038/225563a0. URL https://www.nature.com/articles/225563a0. Publisher: Nature Publishing Group.

[11] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt Von Haeseler, Robert Lanfear, and Emma Teeling. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. 37(5):1530–1534. ISSN 15371719. doi: 10.1093/molbev/msaa015. URL https://academic.oup.com/mbe/article/37/5/1530/5721363. Publisher: Oxford Academic.

[12] Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein language models boosts predictions across diverse tasks. 15(1):7407. ISSN 2041-1723. doi: 10.1038/s41467-024-51844-2. URL https://www.nature.com/articles/s41467-024-51844-2. Publisher: Nature Publishing Group.

[13] Nicholas C. Wu and Ian A. Wilson. Influenza hemagglutinin structures and antibody recognition. 10(8):a038778. ISSN , 2157-1422. doi: 10.1101/cshperspect.a038778. URL http://perspectivesinmedicine.cshlp.org/content/10/8/a038778. Publisher: Cold Spring Harbor Laboratory Press.