An Integrated AI Approach to Predict Potential Influenza Variants

Meitner Cadena Cepeda and David Guevara-Barrientos

August 26, 2024

Abstract

A procedure involving several elements from biology, artificial intelligence, statistics, and mathematics to forecast influenza A variants to a very short term is proposed. This new mechanism tackles the huge complexity of this kind of forecast in a systematic way. Results show forecasts of both recurrent and new mutations. This procedure would also allow the identification of potential mutations of certain hemagglutinin and neuraminidase proteins, since its core is based on mining a number of protein relationships involving hemagglutinin and neuraminidase mutations. The findings of these mutations of the influenza A viruses that will circulate each season are essential to develop effective vaccines, since the rapid evolution of viruses is their main way of evading human immunity. The results show that the proposed procedure is promising to further improve this type of analysis strategy and, therefore, to discover true relationships giving true mutations.

Keywords: influenza A, hemagglutinin, neuraminidase, random forests, neural networks

1 Introduction

Influenza A virus is a recurring problem for human health due to its annual seasonal spread promoted by the ability of this virus to mutate rapidly and thereby infect humans with new variants (Short et al., 2015). These facts convert this virus into a world threat that eventually may become out of control producing pandemics. These viruses come from different reservoirs, some of them being non-human. This virus may provoke a wide range of symptoms and eventually compromise the health of individuals, leading to a risk of death. In fact, it is not well known how new influenza variants originate, since, until now, unexpected new variants can always emerge. Some of them became pandemics that have caused the death of a number of people, like the Spanish flu of 1918 and swine flu of 2009 (Borkenhagen et al., 2021). Despite these limitations, the accuracy of predictions of influenza variants is always crucial to developing highly effective vaccines, as it is one of the most important known ways to mitigate and prevent the spread of influenza (Lim et al., 2024).

One of the key difficulties to forecast variants is that the mechanisms to structure future variants are still elusive. This is one of the factors that have limited research in this field. Some key studies tackling this kind predictions have been one based on applied evolutionary biology to better understand influenza evolution (Morris et al., 2018), another one that considers laboratory research from biomolecular analysis and animal models to identify features of potential new strains related with mainly virulence (Harrington et al., 2021), and a third one that has addressed efforts to forecast risk levels associated with influenza (Turtle et al., 2021).

Given these constraints, an alternative analysis is to implement data mining on past experiences with influenza variants. In this sense, up to our knowledge, there is no study on combining several techniques from biology, artificial intelligence, statistics, and mathematics to discover future variants. The point would be to approach the problem in such a way that eventual patterns related to future variants can be found. This type of strategy would call upon several well-known procedures that should be properly integrated to maximize findings.

In this paper, we propose a focus on data concerning specific features of different types such as closeness among influenza variants, small periods of forecast, regression models and simulations of type bootstrap to cover eventual patterns that may be related to observed mutations. This analysis strategy focuses on hemagglutinin and neuraminidase, proteins found in the influenza virus, and their interactions. In this way, this approach will be a starting point for explorations that will strategically integrate more elements that contribute to discover mutations.

The rest of the paper describes the data studied, the methods to be used and how they are disposed to reach findings. The last sections show results for particular assumptions that will allow for standing out advantages of the present approach and delineate next steps for reaching next goals.

2 Materials and Methods

2.1 Data

Data of influenza A of type H1N1 were analyzed, focusing on its proteins hemagglutinin (HA) and neuraminidase (NA). They were obtained from Global Initiative on Sharing All Influenza Data (GISAID) (Shu & McCauley, 2017). Reported viruses by China, France, Germany, Hong Kong, Japan, Netherlands, Singapore, Sweden, UK, and USA between 2009 and 2023 were selected. Also, only records containing the same length of nucleotides for each of the proteins HA and NA were considered. Such lengths were fixed as the ones of the modes for observed HA and NA. This gave lengths of 566 for HA and 469 for NA. Moreover, viruses with ambiguous or missing nucleotides in their nucleotide sequences or showing a lack of one of their HA or NA proteins were excluded. This produced 14,020 variants observed throughout the analyzed period. Furthermore, the analyzed sequences were expressed in amino acids, which were obtained by translating their corresponding transcribed nucleotide sequences.

Furthermore, the viruses were organized in semesters according to their collection date. These semesters were established considering influenza seasonal periods, which gives semesters from mid October to mid April and from mid April to mid October (Zanobini et al., 2022). Therefore, the observed viruses were considered until mid April 2023. The identifiers of the final analyzed viruses are presented in the supplementary information of this paper.

The protein data described above must be prepared in numerical format in order to be used in operational models such as neural networks. This numerical requirement for amino acids given in 20 letters is usually accomplished using bidimensional matrix representations. For instance, each of those letters is represented by a vector of length 20 filled of 0s, excluding one exclusive position where the value 1 is assigned. Collecting each letter that appears in a protein, e.g. l, then that protein is represented by a numerical matrix $20 \times l$. Therefore, this 2-dimension design for proteins may be demanding if the available computational memory is limited as usually happens for many individuals because they work with their personal computers. To overcome this issue, we propose

the use of a 1-dimension design for proteins based on pre-trained models for proteins provided freely by Evolutionary Scale Modeling (ESM) from Facebook Research (Lin et al., 2022). This type of simplification and synthesis has meant a significant advancement in computational biology. Indeed, this was possible by configuring a profile that integrates characteristics such as complex biological, physical and chemical relationships that occur within protein structures in combination with new deep learning techniques like geometric networks (Wu et al., 2023). The set of pre-trained models chosen is called `esm2_t33_650M_UR50D`. This model has 650 million parameters and covers a wide variety of protein sequences. This model has been used in applications involving proteins, such as prediction of interactions between proteins (Sargsyan & Lim, 2024), and in the prediction of binding sites (Shenoy et al., 2024).

2.2 Strategy for Influenza Analysis

The study of viruses of influenza is challenging because several key features of their behavior and evolution are still not well understood. One of these aspects is how their diversity occurs. More precisely, how their mutations are developed. We propose the following approach to deal with this complexity. As a divide-and-conquer algorithm (Smith, 1985), the observed influenza viruses were organized in clusters that evolve over time. These clusters consisted in variants exhibiting close behaviors in their mutations. Next, for each of these clusters, simulated evolutions of variants belonging to the same cluster were developed. These simulations are in line with branches obtained from random forests, where simulations are based on bootstraps. Thus, the unknown knowledge on how viruses of influenza mutate is pretended to be covered through all those simulations. Next, for each of these clusters, a simple neural network to relate current with future variants through the application of autoregressive models was designed. In this way, the entire complexity was split in several problems that exhibited simpler complexity , since variants in each cluster show close variations from each other. Finally, because it is not possible to describe precise mutations for influenza virus, forecasts are modeled through probability density functions derived from all simulations generated.

Fig. 1 presents a flowchart of the main components of this analysis strategy. They are explained in what follows.

2.2.1 Clusters of Viruses

To form clusters of viruses over time, the last observed semester was considered at first. For this semester, clusters were formed by allowing variation for each position among viruses up to a fixed number of different amino acids from each other. For the construction of these clusters, those with the greatest number of viruses were prioritized. After several essays, up to 7 different amino acids for each protein of the couple (HA,NA) were allowed. This approach is in line with previous studies where viruses in each cluster had fewer nucleotide differences (Lavenu et al., 2006; Plotkin et al., 2002). Moreover, this rule required to belong to a cluster guaranteed to have members in a given cluster that show close behaviors among them in how many different nucleotides they can present with respect to other members. Further, the members of a given cluster for the last semester observed satisfy an equivalence relation through the application of this rule (Britz et al., 2001). This fact allows for all clusters to have no members in common. This procedure is represented in Algorithm 1, where all obtained clusters have at least three sequences.

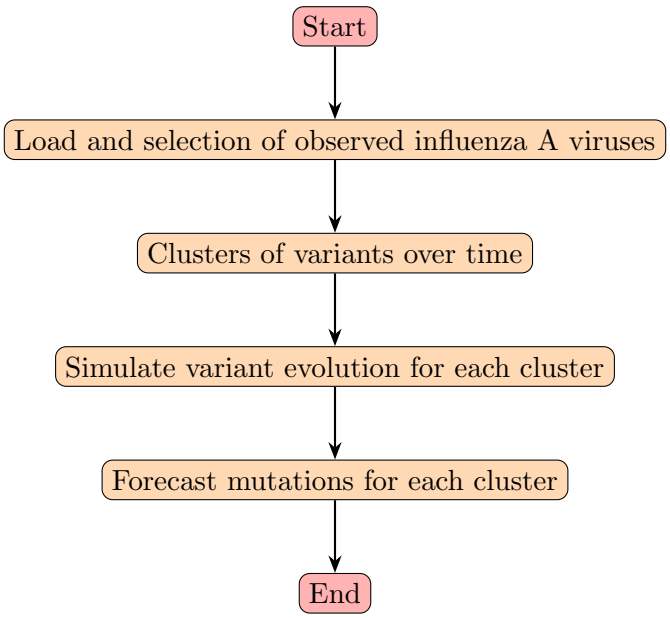

Figure 1: Flowchart of the Influenza A Virus mutation simulation process.

Algorithm 1 Clusters of Amino Acid Sequences of Viruses for the Last Observed Semester

Require: A set of amino acid sequences $S = \{\texttt{s1}, \texttt{s2}, \ldots, \texttt{sn}\}$
Require: Maximum allowed differences among sequences d
Ensure: A set of clusters C

Initialize an empty set of clusters C
for each sequence $\texttt{si} \in S$ **do**
    for each sequence $\texttt{sj} \in S$ such that $\texttt{si} \neq \texttt{sj}$ **do**
        Compute $d_{i,j}$ as the number of different amino acids between si and sj
    end for
end for
Let S_2 be the set of couples (si,sj) such that $d_{i,j} \leq d$
Let There.is.More.Clusters = 1
Let n = 2
while There.is.More.Clusters = 1 **do**
    n = n + 1
    Let S_n be the set of (si.1,...,si.n) such that $(\texttt{si.j}, \texttt{si.k}) \in S_2$, $\texttt{j}, \texttt{k} = 1, \ldots, n$
    if $S_n = \emptyset$ **then**
        Let There.is.More.Clusters = 0
    end if
end while
if $n = 3$ **then**
    Return C
else
    Include the cluster S_{n-1} into C
    $S = S \setminus S_{n-1}$
    Repeat the process from 2
end if

Fig. 2 gives an example of a set of amino acid sequences of length 50 with mutations to identify relationships among them by following Algorithm 1. For instance, if the maximum number of differences between two sequences is 2, we have that the cluster with the highest size satisfying this condition consists in `s7`, `s9`, `s10`, and `s15`. For that maximum number of differences, no other cluster can be formed. Increasing the maximum number of differences to 3, the cluster with the highest size satisfying this condition consists in `s7`, `s9`, `s10`, `s11`, `s13`, and `s15`, and the cluster with the second highest size consists in `s5` and `s6`, but only have two members. For such a condition, no other cluster can be formed.

Figure 2: Example for identifying relationships among sequences considering a fixed upper limit of differences between them. Mutations are shown without background color.

For data considered in this paper, Algorithm 1 produced 7 clusters with sizes 3, 5, 8, 11, 25, 35, and 80. The union of all these clusters did not contain all the variants observed in the last observed semester. Such cluster union represented 97.1 % of those observed variants.

Next, each cluster established in the last observed semester was projected onto the previous semester as follows. Each virus observed in the previous semester was related to a specific cluster if there was at least one virus from that cluster such that the number of differences between both viruses was at most d, this parameter as defined in Algorithm 1. This procedure was used to recursively continue the cluster projections in the rest of the first semesters. Furthermore, this maximum difference in amino acids generally allowed us to have non-empty protein sets for all semesters, which guaranteed generating simulations throughout the analysis period.

As a result, each group formed in the last observed semester is projected to the first semesters. This implies that the same number of clusters is maintained over time. However, due to this procedure itself, eventually some groups may merge in some of the first few semesters.

2.2.2 Simulations of Variant Evolutions

Since the evolution of influenza A variants is not well-understood, the possibility of different combinations between variants belonging to two contiguous semesters was adopted. This idea was implemented by randomly selecting one member from each of these two semesters, but choosing those members from a certain cluster. This means that each simulation is bootstrap-type, thus always guaranteeing that its members belong to said specific cluster.

For each cluster created over time, 500 bootstrap simulations were generated. This means that 500 variant evolutions were simulated, each of which covered the entire analyzed period. Note that each variant simulated over time involves its HA and NA proteins. This gives 500 evolutions of HA and NA separately, but they are related to each other as they are part of the same variant simulated in time. These simulations generate a random forest for modeling our specific need that is to represent evolutions of viruses. Such bootstrap simulations form branches of that random forest, where its relations between nodes are random but keeping some closeness. Also, these branches can be distinguished in ones for HA and others for NA. Other similar variants of random forests have been developed to address other specific types of problems (Zalavadia & Gildin, 2021; Zhang et al., 2022).

2.2.3 Neural Networks for Forecasting Proteins

Simple neural networks are proposed to forecast HA and NA in each cluster. However, because it is not well-understood how influenza A virus mutates, different combinations of evolutions are proposed to manage this issue. Such combinations are the simulated variant evolutions described above. This leads to consider a forecast as a set of possible scenarios. More precisely, describing a forecast using a probability density function (PDF) is more reasonable than using a single value. This type of outputs has also been used in other contexts such as biochemical models, operations research and SIR epidemic models (Calatayud et al., 2020; Jönsson et al., 2023; Wick et al., 2021). Notice that in our case, these PDFs are dynamic since HA and NA both are dynamic over time.

The neural network (NN) used was implemented in Keras (Chollet et al., 2015) to run in the Python environment (Foundation, 2024). Fig. 3 shows the design of this NN. It consisted in an input layer to enter all amino acids of both HA and NA, these expressions having been previously interpreted numerically as indicated in Subsection 2.1. Next, two hidden layers both of 1,000 nodes with activation function Rectified Linear Unit (ReLU) defined by $f(x) = \max(0, x)$ (Agarap, 2018) were included. The last layer produced the predicted protein while maintaining the same size as the input proteins. All models were trained with 50 epochs.

The above NN was designed to represent proteins that evolve over time (P_t) through autoregressive (AR) processes. These models are frequently used to represent a number of phenomena (Burant, 2022). Fig. 4 illustrates such a NN for that process of order p, AR-p. This model can be written as $P_t = f(P_{t-1}, P_{t-2}, \ldots, P_{t-p})$, where f is a nonlinear function estimated when the NN involved fits its output for a given set of inputs using an iterative training procedure (Tealab, 2018). Considering t as the reference time, P_{t-i} is a protein at time $t - i$.

Typically, NNs are trained with a subset of data and tested with another subset of those data. This evaluation seeks to ensure that the NNs are not over-fitted. Since 500 simulations were sampled to build random forests for each cluster, 400 of them were used to train AR models and the rest to test those models. This provides usual percentages used in random forests to train models, preventing them from overfitting (Harvey & McBean, 2014; Wang et al., 2018).

Finally, the NN produces a prediction of (HA,NA) in numerical values. A decoder linked to `esm2_t33_650M_UR50D` is then used to interpret these numerical results in terms of amino acids.

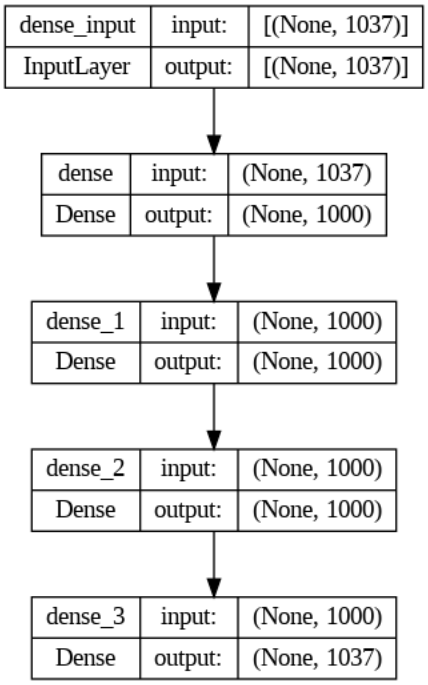

Figure 3: Neural network for forecasting proteins

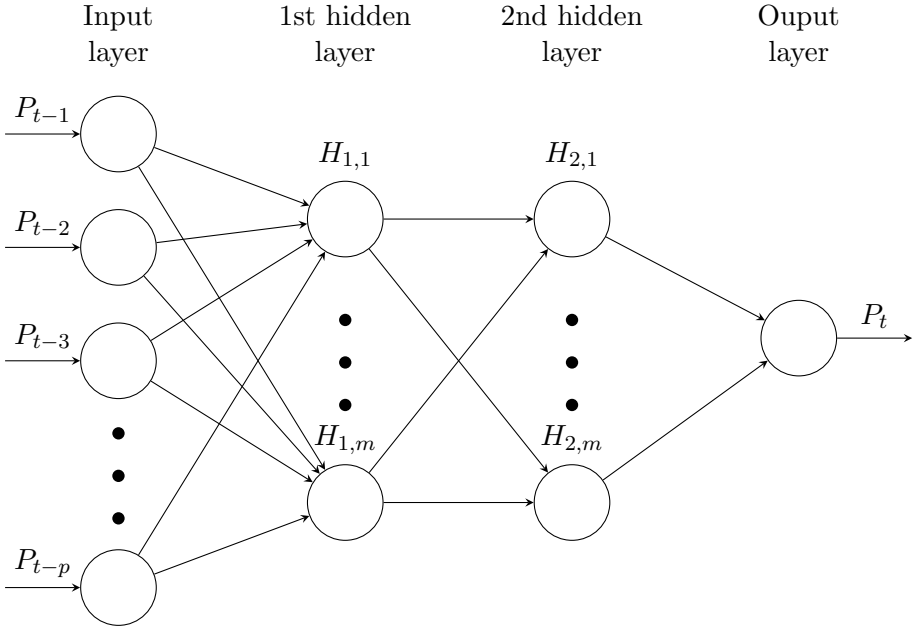

Figure 4: Neural network that incorporates autoregressive models for proteins varying over time (P_t)

2.3 Backtesting the Auto-Regressive Models

Since AR models are time series models, one way to validate that these models are appropriate is to run a backtest on the model (Bailey et al., 2015). It involves building another autoregressive model over a shorter time period, without using the last time period, and testing that model over that last

time period. Comparing the predicted and observed values during that last period would show how useful AR models are in obtaining forecasts. For this, a 1-semester backtesting was carried out.

2.3.1 Statistical Analysis of Predicted Proteins

The PDFs to consider are a bit different from typical PDFs derived from random variables. This is because of mutations. For a position in HA or NA where a mutation happens, the new amino acid may have up to 19 alternatives, since 20 amino acids are possible. This issue is driven using logos of proteins because these representations take into account all occurrences of amino acids observed for all positions in a protein. It means that protein logos can be seen as PDFs. These visual representations present heights that indicate the likelihood of several outcomes concerning a same amino acid (Schneider & Stephens, 1990). This leads to easily recognize the most frequent elements and also the ones with lower frequencies. In fact, some logos already address the presentation of such probabilities (Cao et al., 2023). Furthermore, these protein logos can also be used to statistically assess equality among variants of a protein as their image representations involve probabilities. For this purpose, a bootstrap procedure is proposed (Johnson, 2001). It consists in resampling the predicted proteins B times for a given position of a protein. The predicted amino acid frequencies of these proteins are then subtracted from the expected amino acid frequencies of the observed proteins. Next, averages of these frequencies for each amino acid for each resample are computed. This produces B values for each amino acid, which allow for building two-side confidence intervals for each of those positions. The null hypothesis saying that the frequency of an amino acid is statistically equal to the expected frequency of that amino acid is proven if 0 is contained in its corresponding confidence interval. Otherwise, it is concluded that the frequency for the concerned amino acid is not represented by the model that is tested.

3 Results

The AR-1 and AR-2 models with a 1-semester forecast were analyzed. Furthermore, the duration of the simulated protein time series was also evaluated considering all data (full data) and the last 8 semesters of data (partial data).

The backtesting analysis showed that all the analyzed models presented significant differences between the observed and predicted amino acid frequencies of the analyzed proteins.

Considering the clusters generated, there were 7, but the sixth could not be projected for all semesters since the proteins of those semesters were too far apart considering the rule imposed to form clusters. That is, the elements of the sixth group differed from the historical HA or NA proteins by more than 7 positions. Because the sixth cluster lasted only two semesters, this cluster was not taken into account for forecast.

Figure 5 shows the latest convergence behaviors for the NNs used to fit autoregressive models through all clusters 1, 2, 3, 4, 5, and 7. Each of them is noted as "n AR-p tD", where n is the cluster number indicated above, p is the order of the model AR, and t specifies the type of data as full (F) or partial (p). These outputs show similar NN convergence behaviors among all models used and a significant low variance through all those models. This is in line with the fact that most of the time the confidence intervals for the training means are included in the confidence intervals for the test means.

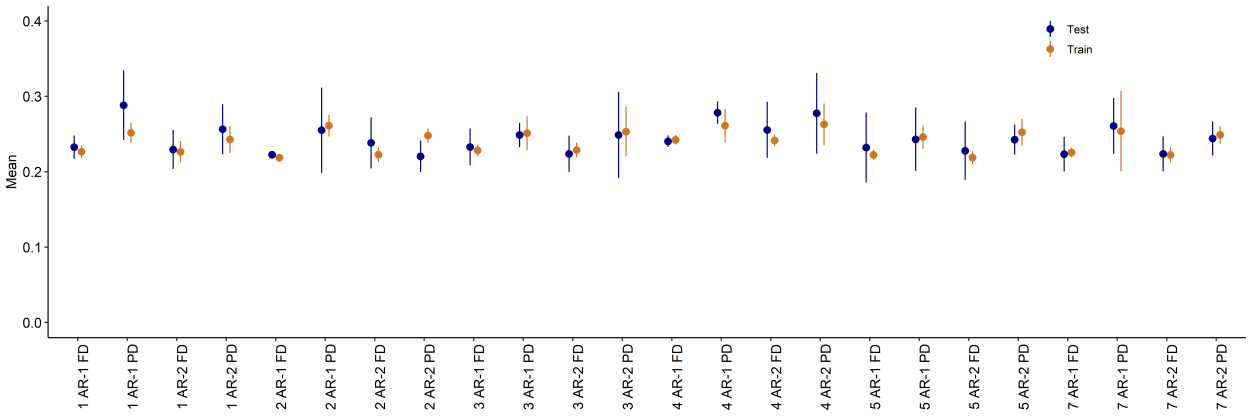

Figure 5: Behavior of NN convergence considering the 10 last epochs, by cluster

Table 1 presents the p value for each bootstrapped model. These values were calculated considering samples based on each position of the HA and NA proteins. In this regard, Table 2 shows that the majority of the positions conserve their amino acid, with the positions that report changes or mutations being marginal. A graphical illustration of this behavior can be seen in the examples provided in Fig. 7. These results show that all of these models are viable for forecasting future influenza A variants.

Cluster	Data type	Order of the AR model	
		$p = 1$	$p = 2$
1	Full	0.23	0.25
1	Partial	0.21	0.22
2	Full	0.22	0.26
2	Partial	0.18	0.19
3	Full	0.24	0.25
3	Partial	0.20	0.22
4	Full	0.24	0.26
4	Partial	0.17	0.19
5	Full	0.23	0.27
5	Partial	0.13	0.15
7	Full	0.24	0.26
7	Partial	0.16	0.17

Table 1: Backtesting of the models: p-value considering positions in both the HA and NA proteins

Below, Table 2 presents the mutations in the last observed semester and the predicted mutations in 1 semester for HA and NA by cluster, except for the sixth cluster. These results are the aggregation of the different simulations, both observed and predicted, since simulations of possible relationships are used as they do not have specific mutation mechanisms. More specifically, a mutation is identified as such if at the corresponding position two or more simulations provided different amino acids. In addition, the mutations of the last observed semester and the predicted semester are compared since it would be expected that these two consecutive semesters would have some type of relationship between them. Precisely, these types of unknown relationships are those that have been adopted to build the clusters that are currently analyzed. This table identifies two types of predicted mutations and for each HA and NA protein, recurrent and new. A recurrent mutation is a mutation in whose position it is identified that in the preceding (observed) semester a mutation already occurred.

That is, the interest in distinguishing this type of mutation is to identify possible positions that show susceptibility to mutations in consecutive semesters. On the other hand, a new mutation is a mutation in whose position no mutation is identified in the preceding (observed) semester. This would show the potential of a given position to generate mutations. Table 2 also shows the four types of models used for each cluster. The sixth cluster is excluded.

Model	HA Mutations			NA Mutations		
	Observed†	Recurrent	New	Observed†	Recurrent	New
Cluster 1						
AR-1 – Full data	54	6	8	25	3	3
AR-2 – Full data	54	4	4	25	4	3
AR-1 – Partial data	54	10	89	25	4	93
AR-2 – Partial data	54	6	57	25	3	39
Cluster 2						
AR-1 – Full data	54	6	2	25	3	4
AR-2 – Full data	54	5	2	25	4	3
AR-1 – Partial data	54	3	4	25	3	3
AR-2 – Partial data	54	0	5	25	0	3
Cluster 3						
AR-1 – Full data	54	6	4	25	3	3
AR-2 – Full data	54	2	2	25	4	1
AR-1 – Partial data	52	6	14	25	1	6
AR-2 – Partial data	54	5	4	25	2	3
Cluster 4						
AR-1 – Full data	54	9	4	25	4	2
AR-2 – Full data	54	8	86	25	7	80
AR-1 – Partial data	54	8	10	25	3	7
AR-2 – Partial data	54	4	3	25	1	9
Cluster 5						
AR-1 – Full data	54	5	1	25	1	1
AR-2 – Full data	X	5	3	43	5	3
AR-1 – Partial data	54	2	7	25	1	7
AR-2 – Partial data	54	4	6	25	1	3
Cluster 7						
AR-1 – Full data	53	3	3	25	2	0
AR-2 – Full data	53	10	8	25	3	6
AR-1 – Partial data	53	8	9	25	2	9
AR-2 – Partial data	53	5	12	25	2	15

† Observations from the precedent semester

Table 2: Mutations in the forecast semester by cluster

Fig. 6 illustrates examples of sequence logos of HA and NA for the predicted semester. These plots are for the cluster 1 and show the first 30 simulations predicted.

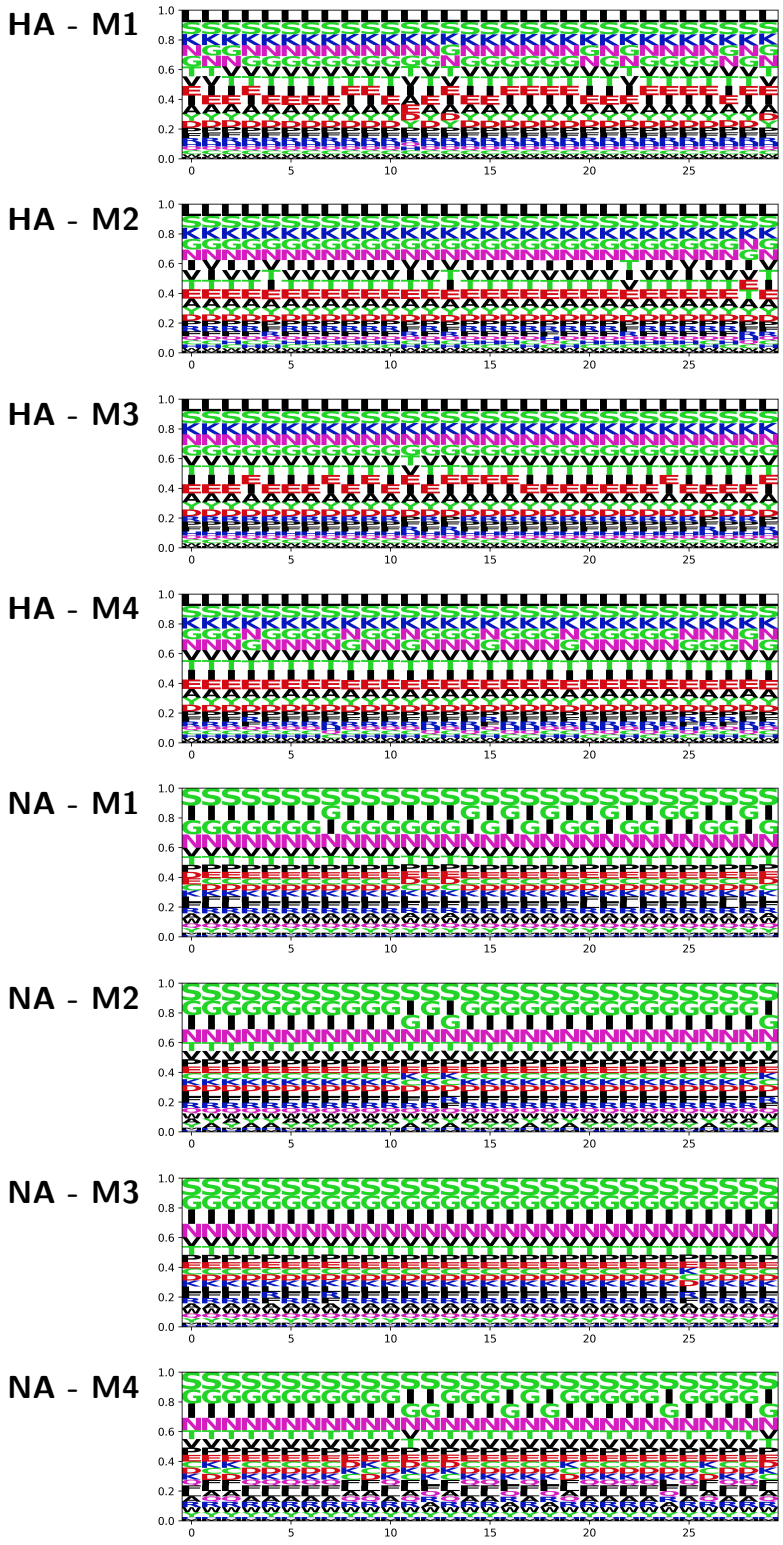

Figure 6: Cluster 1: sequence logos for the first 30 predicted HA and NA and each model applied. M1: AR-1 full data. M2: AR-1 partial data. M3: AR-2 full data. M4: AR-2 partial data

Fig. 7 presents examples of mutations of HA and NA for the predicted one. The proteins are scaled

to the perimeter of the circles, starting since the point blue and following in anticlockwise direction. As indicated above, mutations are any position where simulations present different amino acids. These mutations are shown as red radius.

4 Discussion

Our models have fulfilled the objective of making predictions of the HA and NA proteins even under serious difficulties as mainly the current situation of not having a good understanding of how these proteins mutate. Indeed, the backtesting analysis showed that the developed models are viable for forecasting the proteins HA and NA.

In fact, first of all, the applied procedure allowed us to detect a cluster for the two most recent semesters that would not be related to any pair (HA,NA) of the historical semesters. This suggests that new variants have appeared, but of unknown origin. Since the simulations for this cluster consist of only 2 semesters, no forecasts were made for this cluster. Next, our procedure was designed for dealing with problems exhibiting low complexity. This is precisely verified with the low variances observed in the convergences of the NNs. These results precisely suggest that the number of epochs necessary to achieve convergence of these NNs can be decreased. This is because the low variance would guarantee that the convergence that would be obtained with fewer epochs is very close to that achieved with 50 epochs. On the other hand, no single model is identified that predicts most mutations, but it seems to depend on the data to choose a convenient model. Further, some of those models generated a significant number of mutations, which are potential to find true mutations.

5 Conclusions

This article has analyzed HA and NA sequences maintaining the same lengths for each. The proposed procedure to predict these proteins over a period of 1 semester has been successful since mutation predictions could be obtained. This suggests that clusters may generate their own dynamics of evolution. This procedure is also found to be successful since it allowed the identification of a cluster that appears to have originated recently since it is not related to historical data.

Despite these advances, more research is required to identify properties that identify and differentiate the clusters found. Therefore, This article constitutes a promising first step towards a better understanding of the evolution of influenza A.

References

Agarap, A. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Bailey, D. H., Ger, S., de Prado, M. L., & Sim, A. (2015). Statistical overfitting and backtest performance. In *Risk-based and factor investing* (pp. 449–461). Elsevier.

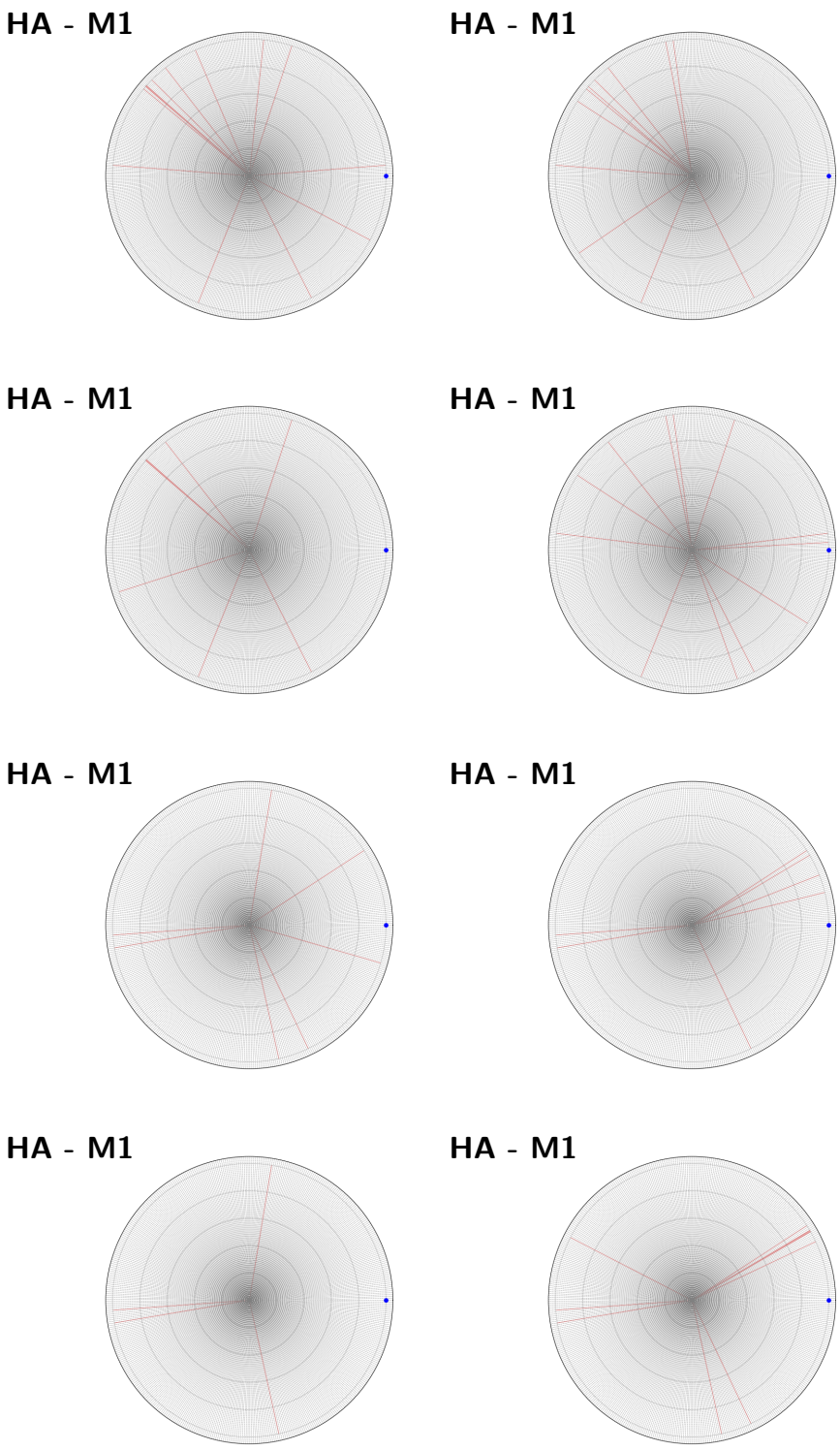

Figure 7: Cluster 1: mutations of predicted HA and NA and each model applied. M1: AR-1 full data. M2: AR-1 partial data. M3: AR-2 full data. M4: AR-2 partial data

Borkenhagen, L. K., Allen, M. W., & Runstadler, J. A. (2021). Influenza virus genotype to phenotype predictions through machine learning: a systematic review: computational prediction

of influenza phenotype. *Emerging microbes & infections*, *10*(1), 1896–1907.

Britz, T., Mainetti, M., & Pezzoli, L. (2001). Some operations on the family of equivalence relations. *Algebraic Combinatorics and Computer Science: A Tribute to Gian-Carlo Rota*, 445–459.

Burant, C. J. (2022). A methodological note: an introduction to autoregressive models. *The International Journal of Aging and Human Development*, *95*(4), 516–522.

Calatayud, J., Cortés, J. C., & Jornet, M. (2020). Computing the density function of complex models with randomness by using polynomial expansions and the rvt technique. application to the sir epidemic model. *Chaos, Solitons & Fractals*, *133*, 109639.

Cao, T., Li, Q., Huang, Y., & Li, A. (2023). plotnineseqsuite: a python package for visualizing sequence data using ggplot2 style. *BMC genomics*, *24*(1), 585.

Chollet, F., et al. (2015). *Keras.* `https://github.com/keras-team/keras`. GitHub.

Foundation, P. S. (2024). Python language reference [Computer software manual]. Retrieved from `https://docs.python.org/3/reference/` (Accessed: [insert date accessed])

Harrington, W. N., Kackos, C. M., & Webby, R. J. (2021). The evolution and future of influenza pandemic preparedness. *Experimental & molecular medicine*, *53*(5), 737–749.

Harvey, R. R., & McBean, E. A. (2014). Predicting the structural condition of individual sanitary sewer pipes with random forests. *Canadian Journal of Civil Engineering*, *41*(4), 294–303.

Johnson, R. W. (2001). An introduction to the bootstrap. *Teaching statistics*, *23*(2), 49–54.

Jönsson, B. F., Follett, C. L., Bien, J., Dutkiewicz, S., Hyun, S., Kulk, G., ... others (2023). Using probability density functions to evaluate models (pdfem, v1. 0) to compare a biogeochemical model with satellite-derived chlorophyll. *Geoscientific Model Development*, *16*(16), 4639–4657.

Lavenu, A., Leruez-Ville, M., Chaix, M.-L., Boelle, P.-Y., Rogez, S., Freymuth, F., ... Carrat, F. (2006). Detailed analysis of the genetic evolution of influenza virus during the course of an epidemic. *Epidemiology & Infection*, *134*(3), 514–520.

Lim, C. M. L., Komarasamy, T. V., Adnan, N. A. A. B., Radhakrishnan, A. K., & Balasubramaniam, V. R. (2024). Recent advances, approaches and challenges in the development of universal influenza vaccines. *Influenza and Other Respiratory Viruses*, *18*(3), e13276.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ... others (2022). Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, *2022*, 500902.

Morris, D. H., Gostic, K. M., Pompei, S., Bedford, T., Łuksza, M., Neher, R. A., ... McCauley, J. W. (2018). Predictive modeling of influenza shows the promise of applied evolutionary biology. *Trends in microbiology*, *26*(2), 102–118.

Plotkin, J. B., Dushoff, J., & Levin, S. A. (2002). Hemagglutinin sequence clusters and the antigenic evolution of influenza a virus. *Proceedings of the National Academy of Sciences*, *99*(9), 6263–6268.

Sargsyan, K., & Lim, C. (2024). Using protein language models for protein interaction hot spot prediction with limited data. *BMC bioinformatics*, *25*(1), 115.

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, *18*(20), 6097–6100.

Shenoy, A., Kalakoti, Y., Sundar, D., & Elofsson, A. (2024). M-ionic: prediction of metal-ion-binding sites from sequence using residue embeddings. *Bioinformatics*, *40*(1), btad782.

Short, K. R., Richard, M., Verhagen, J. H., van Riel, D., Schrauwen, E. J., van den Brand, J. M., ... Herfst, S. (2015). One health, multiple challenges: The inter-species transmission of influenza a virus. *One health*, *1*, 1–13.

Shu, Y., & McCauley, J. (2017). Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, *22*(13), 30494.

Smith, D. R. (1985). The design of divide and conquer algorithms. *Science of Computer Programming*, *5*, 37–58.

Tealab, A. (2018). Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal*, *3*(2), 334–340.

Turtle, J., Riley, P., Ben-Nun, M., & Riley, S. (2021). Accurate influenza forecasts using type-specific incidence data for small geographic units. *PLoS Computational Biology*, *17*(7), e1009230.

Wang, Q., Nguyen, T.-T., Huang, J. Z., & Nguyen, T. T. (2018). An efficient random forests algorithm for high dimensional data classification. *Advances in Data Analysis and Classification*, *12*, 953–972.

Wick, F., Kerzel, U., Hahn, M., Wolf, M., Singhal, T., Stemmer, D., ... Feindt, M. (2021). Demand forecasting of individual probability density functions with machine learning. In *Operations research forum* (Vol. 2, pp. 1–39).

Wu, F., Wu, L., Radev, D., Xu, J., & Li, S. Z. (2023). Integration of pre-trained protein language models into geometric deep learning networks. *Communications Biology*, *6*(1), 876.

Zalavadia, H., & Gildin, E. (2021). Two-step predict and correct non-intrusive parametric model order reduction for changing well locations using a machine learning framework. *Energies*, *14*(6), 1765.

Zanobini, P., Bonaccorsi, G., Lorini, C., Haag, M., McGovern, I., Paget, J., & Caini, S. (2022). Global patterns of seasonal influenza activity, duration of activity and virus (sub) type circulation from 2010 to 2020. *Influenza and Other Respiratory Viruses*, *16*(4), 696–706.

Zhang, C., Wang, W., Liu, L., Ren, J., & Wang, L. (2022). Three-branch random forest intrusion detection model. *Mathematics*, *10*(23), 4460.