Title of Thesis

“Protein-based in silico models to predict infectious strains of Influenza A”

by

Trinita Roy

Under the Supervision of Prof. Dr. Gajendra P.S. Raghava

Indraprastha Institute of Information Technology Delhi

May 2022

Title of Thesis

“Protein-based in silico models to predict infectious strains of Influenza A”

by

Trinita Roy

Submitted

in partial fulfillment of the requirements for the degree of

Master of Technology

to

Indraprastha Institute of Information Technology Delhi

May 2022

Certificate

This is to certify that the thesis titled **"Protein-based in silico models to predict infectious strains of Influenza A"** being submitted by Trinita Roy to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

May 2022

Prof.Gajendra P.S. Raghava
Department of Computational
Biology Indraprastha Institute of Information
Technology Delhi
New Delhi 110 020

Acknowledgment

I would like to sincerely thank and acknowledge my MTech thesis supervisor Prof. Gajendra P Raghava from Indraprastha Institute of Information and Technology, Delhi, for allowing me to conduct research under his guidance, from introducing me to the topic to helping me mature as a researcher. His patience and diligence motivated me to persevere during my research. Besides my supervisor, I would like to express my sincere gratitude to the Ph.D. scholars Anjali Dhall and Sumeet Patiyal for their constant guidance, and motivation and for playing an instrumental role in the entire research process. I would like to express my gratitude to Khushal Sharma who collaborated with us to make the Genome module of the web server. Lastly, I would like to thank my family and friends for their constant encouragement and support, which enabled me to pursue research.

Trinita Roy
M.Tech CB

Contents

List of Figures

List of Tables

encoding feature for training and validation datasets.

Table 12: Comparison of FluSpred And BLAST Results.

Abstract

Influenza A, an infectious viral disease affecting the lungs, is a significant public health concern. It has already caused four pandemics in the past, and some strains are now seasonal. Being zoonotic, the virus is transmitted to humans from birds, which are usually aquatic, and swine and other mammals serve as intermediate hosts for its transmission. When present in aquatic birds, the virus is asymptomatic, predicting zoonotic strains that have the potential to cause an outbreak in humans. Gradually, this virus experiences host-adaptive mutations or reassortments in its genome, resulting in different variants which might trigger global health emergencies. Therefore, recognizing zoonotic strains that can cause an outbreak in humans and their origin is the need of the hour. In this analysis, we have devised a machine learning method to predict infectious strains of the Influenza A virus from avians/mammals to humans. The training and validation of the 15 protein sequence was conducted on data obtained from the Influenza Research Database. Random forest-based models using composition-based features attained maximum AUC for the 15 proteins ranging from 0.93 to 0.98 on the validation dataset. On training and validation datasets, the haemagglutinin (HA) protein has the highest AUC of 0.98. We have formulated an in-silico tool for the prediction of infectious strains from protein sequences as a service to the scientific community. The best models were incorporated in our web server named FluSPred which can be accessed freely at "https://webs.iiitd.edu.in/raghava/fluspred/". We expect that this research will assist in prioritizing high-risk viral strains hereafter and analyze the risk of a novel influenza virus emergence. This tool can be integrated with early warning systems and is beneficial for pandemic preparedness, disease surveillance, and determining the overall public-health impact.

Chapter 1: Introduction

Influenza A virus is responsible for causing a communicable viral disease, Influenza, which targets the respiratory system[1]. Influenza A virus, is the only species of the genus *Alphainfluenzavirus* of the virus family *Orthomyxoviridae*[2]. This virus has a zoonotic origin, occurring naturally among geese, swans, and waterfowl, which are generally known as wild aquatic birds, and it is accountable for causing avian influenza infection in domestic poultry [3]. Influenza is a viral disease characterized by sudden elevation of body temperature, dry cough, myalgia, lethargy, fatigue, and headache [4]. Flu-related complications such as pneumonia and encephalitis, as well as myocarditis [5], can be fatal. The prevalence of influenza infection is higher in individuals with chronic heart or lung diseases [6], immune disorders, diabetes [7], [8], [9].

The virus when it comes near to animals, and humans, in places such as forests, and live bird markets, genetic mixing occurs, and that enables the virus to jump the species-barrier to infect animals and humans[10]. This is a rare phenomenon although not unknown because there have been pandemics pertaining to this before. Mutations and reassortments originating from animal populations give rise to novel strains. These novel strains have the capacity to circumvent the host-species barrier and cause infections in humans[11]. As depicted in Figure 1, virus strains have the potential to infect humans and cause outbreaks as the strains may have the prospect to overcome species barriers and infect humans, get host-adaptive mutations and reassortments, thereby giving rise to human-to-human transmission[12]. This can lead to a potential epidemic or pandemic, which is a cause of great health concern.

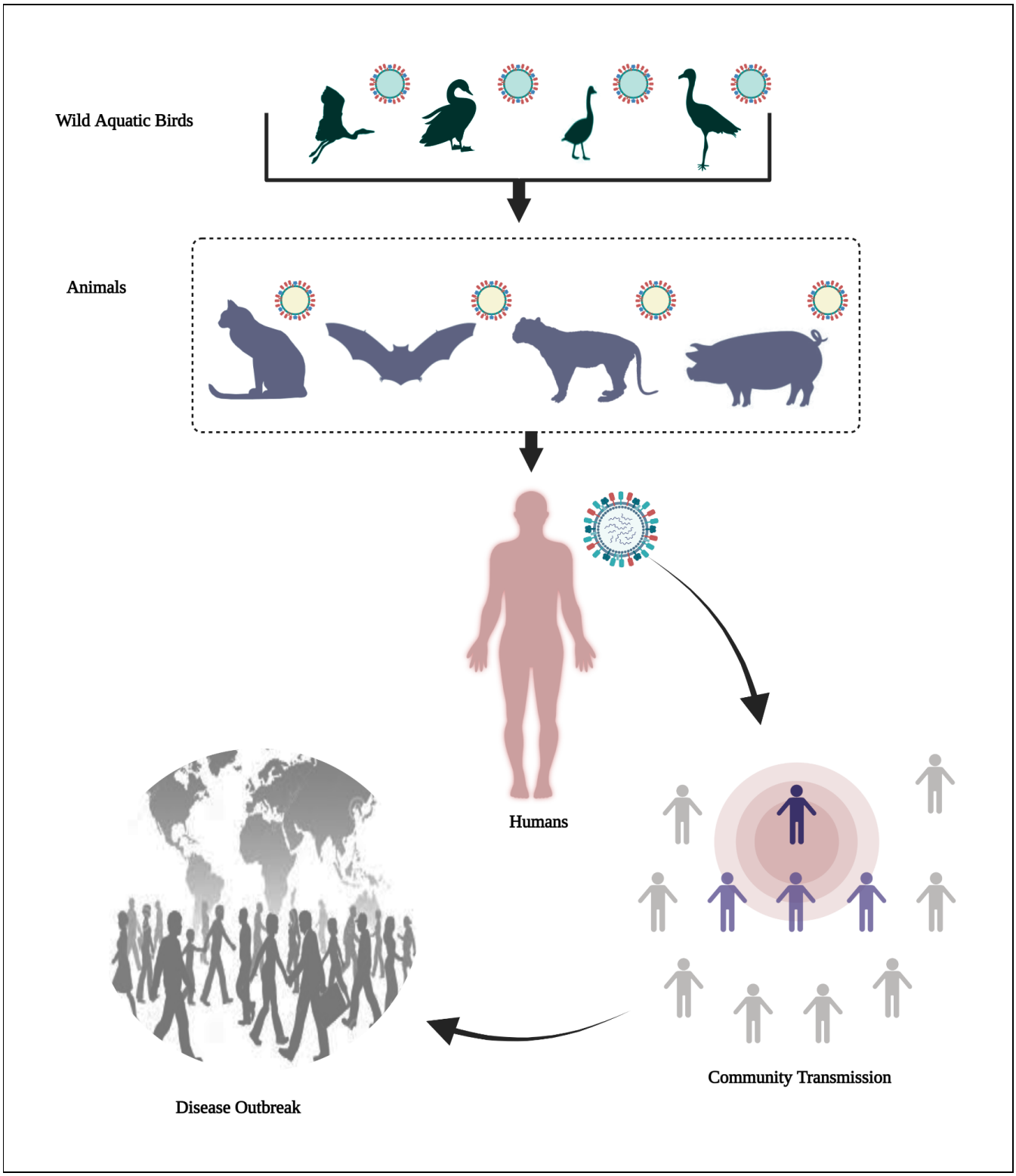


Figure 1: Illustration of zoonotic transmission of Influenza A virus from birds to animals. There is cross-transmission of the virus which may lead to the virus infecting humans with severity ranging from community transmission to disease outbreaks like epidemics and pandemics.

The seasonal flu is recurring in nature since the virus undergoes substitution point mutations causing their genomes to have gradual changes, a phenomenon known as antigenic drift [13]. The impact may be minor as some people have immunity from previous exposures and vaccines are readily available, but sometimes antibodies produced from the previous infection may not work therefore, vaccines against these seasonal strains need to be prepared and modified on an annual basis [14]. On the other hand, the pandemic flu is rare but has significant impacts, as people have little to no immunity from them, which increases the risk for vulnerable and healthy populations. Reassortment of genome segments are responsible for major antigenic changes. This phenomenon is known as antigenic shift, which gives rise to pandemic influeza strains[15]. Viral strains experience mutations and modifications that adapt to their host, which make them capable of infecting humans and spreading efficiently and sustainably [16]. The haemagglutinin protein(HA) can be altered by mutations in amino acid residues, which may enhance the affinity of the receptors for this protein[17].

Mutations or changes in the genome segments give rise to changes in the amino acid composition which can be attributed to the viral strains being capable of infecting human hosts from avian and mammalian reservoirs[18].

Pandemics in The Past

Influenza A is one of the most infectious sickness challenges around the world, which was responsible for four pandemics in contemporary history. In 1918, a subtype of the H1N1 influenza virus was driving one of the most deadly pandemics in the world with high mortality rates among children under 5 and people in their 20s to 40s. The infection had an avian origin and brought about 50 million assessed deaths around the world [18]. The H2N2 subtype happened in East Asia from an avian source in 1957, causing roughly 1.1 million casualties around the world [19]. In 1968, the H3N2 subtype arose in the United States of America and had north of 1 million deaths around the world. This infection later started spreading as the occasional influenza [20]. As of late, (H1N1)pdm09 was first noted in the United States of America in 2009. The infection is apparent from H1N1 and had caused 151,700-575,400 demises overall in the main year of disease. It then, from that point, began circling as seasonal influenza [21].

Structural Biology Perspective of Influenza A

As a single-stranded and segmented RNA virus, Influenza A is characterized by having a negative sense. Hemagglutinin(HA) and Neuraminidase(NA), which are essentially surface glycoproteins, are the subtypes of the virus. There are 18 HA subtypes and 11 NA subtypes associated with influenza viruses. For instance, the influenza virus HIN1 has both HA subtypes of type H1 and NA subtypes of type N1. HIN1, H2N2, and H3N2 are mainly found in humans, while H17N10 and H18N11 can be found in bats[22].

The influenza A subtypes that occasionally contaminate people during epidemics are H1N1, H2N2, and H3N2. The rest are known to contaminate birds, and a few mammals like H7N7 affects horses. Point mutations in a subtype may bring about different strains, being the essential driver of the infection's development, where new strains surpass the old ones through "genetic drift" [22]. Assume there is a virus having a new HA subtype, and they get reassorted to the genomic RNA fragments of the human and avian infectious strains. All things considered, it can possibly spread human-to-human and conceivably cause a pandemic.

The virus has a total of eight negative-stranded RNA genomic segments within its lipid bilayer envelope. Three of the most significant segments encode PA: Polymerase Acidic Protein, PB1: Polymerase Basic Protein 1, and PB2: Polymerase Basic Protein 2, which are the subunits of RNA-dependent RNA Polymerase proteins. PA is the protease part of RNA-dependent RNA Polymerase [23]. PB1 is the endonuclease part of RNA-dependent RNA Polymerase. It also plays a role in RNA [24]. PB2 protein recognizes the mRNA cap as well as plays an important role in host tropism as single glutamic acid residue at position 627 in avian strains gets replaced by lysine, which allows replication of the virus in humans [23]. The segment that encodes PB1 also encodes a non-structural protein PB1-F2. Three RNA segments encode for HA: Hemagglutinin, NA: Neuraminidase, and NP: Nucleoprotein. HA is a surface glycoprotein that plays an important role in host tropism as sialic acid receptors present in host cells bind with HA proteins at the surface of the viral envelope, which then mediates the internalization of the virus. α2,6-sialic acid linkages are recognized by human strains whereas avian strains recognize α2,3-sialic acid linkages [25], [26]. It is the primary antigen, having high plasticity and constantly changing due to a high error rate of the viral polymerase gives rise to antigenic drift [27]. NA is also a surface glycoprotein but it has a sialidase activity[28]. NP is an RNA binding protein that coats the polymerases and helps in the regulation of nuclear import [29]. Out of the

remaining two segments, the larger segment encodes M1: Matrix Protein 1, responsible for Nuclear import regulation [30], and, M2: Matrix Protein 2, which is an ion-channel protein, also present in the envelope as spikes. It helps in virus uncoating and assembly [30]. The smaller segment encodes NS1: Non-structural Protein 1, responsible for suppressing the production of host mRNA [31] and NS2: Non-structural Protein 2, helping in nuclear export [31]. Additionally, segment 2 encodes for PB1-F2 and PB1-N40. PB1-F2 is a small protein of a size of approximately 90 amino acid residues, which synthesizes capped RNA-primed mRNA[32]. PB1-N40 is the coatomer component of a length of 718 residues[33]. Likewise, segment 3 encodes PA-N155, PA-N182, and PAX. PA-N155 and PA-N182, 11th and 13th inframe AUG codon in PA mRNA respectively, which are N-terminally truncated forms of PA protein. They show no polymerase activity when expressed with PB1 and PB2[34]. PA-X is responsible for modulating host immune response [34].

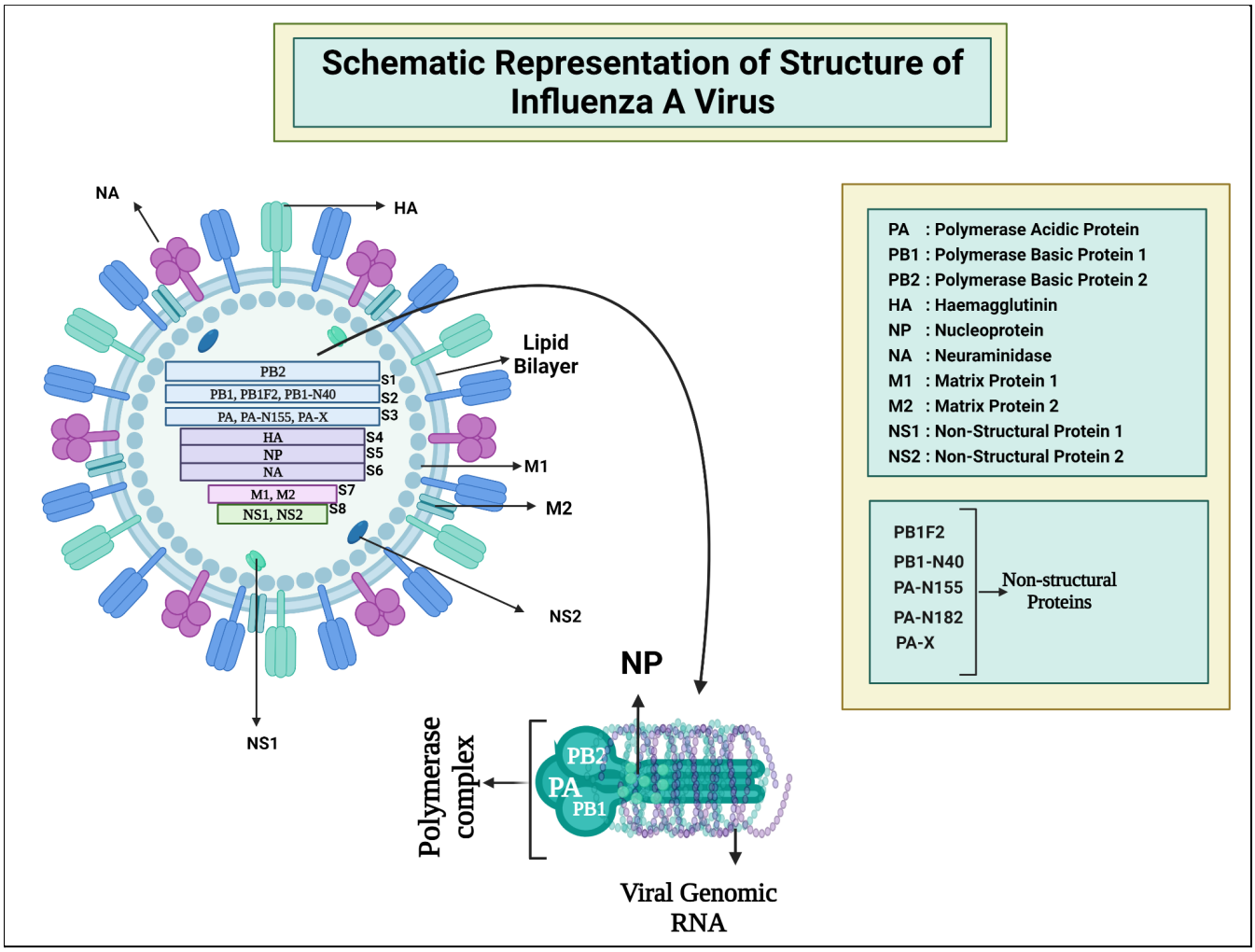


Figure 2: Schematic representation of Influenza A Virus with the segments and their respective proteins and description

Problem Statement

In this research, we have made a systematic endeavor to foster computational models to predict zoonotic host reservoirs of novel Influenza A viral strains, whether they will be infectious to humans or not. An open-source web server, FluSPred, alongside a standalone package, was made of the best models. The reason for the web-server is to support foreseeing irresistible Influenza A strains reservoir with the assistance of sequences of either of the 15 sorts of protein, to serve mainstream researchers in anticipating the zoonotic risk of the infection as well as act as an early warning system.

Literature Review

There have been studies where protein sequences were used to determine the host tropism of the Influenza A virus. Applying artificial neural networks to energy feature vectors from protein sequences, generated by wavelet packet decomposition method so as to distinguish between avian and human influenza virus[35].Another study was done to determine the avian-to-human transmission of influenza A virus using physicochemical properties for making the computational model[36]. The features were generated from the proteins PA, PB1, PB2, M1, NP, and NS1. Additionally, in another study, all the 11 proteins of Influenza A virus were used where amino acid composition along with physicochemical properties like charge, polarizability, solvent accessibility, hydrophobicity, etc was taken as the features for the machine learning-based model[37]. Identifying human adaption-associated genomic composition of Influenza A viruses using principal component analysis and hierarchical clustering have been developed[38]. The models were based on the composition of mono and dinucleotides, where 217549 full-length coding sequences of PA, PB1, PB2, HA, NP, and NA were used, taken from avian and human sources.

Feature extraction methods such as position-specific scoring matrix (PSSM), word embedding, and encoding were also used for host prediction of Influenza A virus using machine learning models[39]. Deep learning techniques like convolutional neural networks have also been used for predicting the virus phenotype using genome sequences[40]. Word vectors of nucleotide and amino acids of influenza A using word2vec for feature extraction were used for host prediction[41]. According to these investigations, protein sequences play a critical role in determining the host tropism of pathogens that cause zoonotic diseases, and modifications occur at the molecular level for the virus to be able to cross species barriers.

Chapter 2: Prediction

Workflow

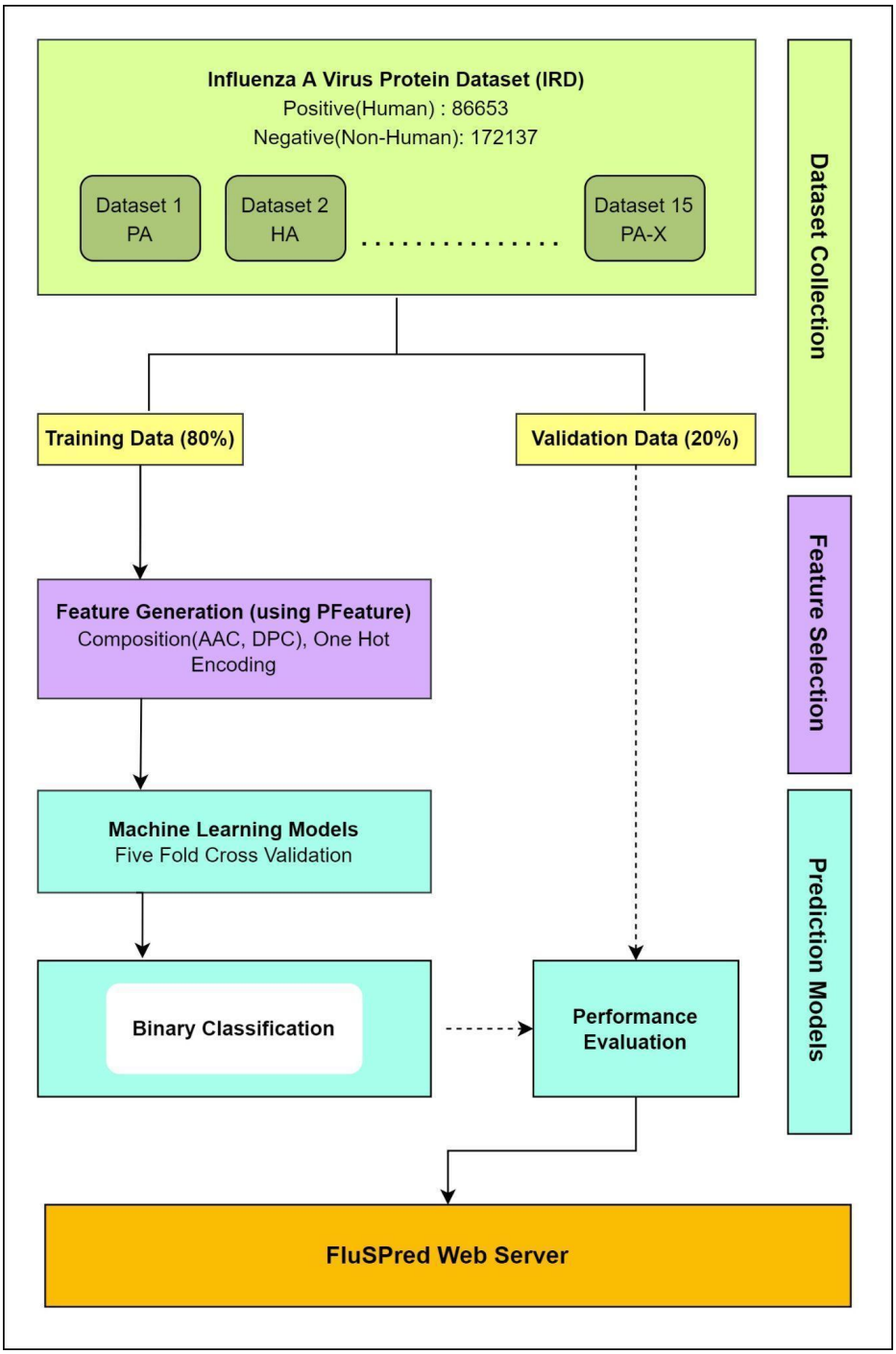


Figure 3: Workflow from Dataset Collection to Making of the FluSPred Web-server

Material and Methods

Dataset Preparation

The datasets were acquired from Influenza Research Database [42]. From this database, a sum of 985,720 protein sequences were extracted on 31.07.2021. These sequences were partitioned into 15 separate datasets relating to the 15 proteins that play a role in the Influenza A virus. These datasets contain sequences of viral proteins from humans as well as mammals and birds. We eliminated the repetitive and inadequate sequences to keep away from inclinations or biases. For each dataset, the incomplete and duplicate sequences were removed. Only sequences from the avian, mammalian, and human hosts were considered. The rest, including the ones from unknown hosts, were removed from the datasets.

Positive datasets were human contagious strains and related sequences, and negative datasets were non-human (avian and other mammalian) associated sequences. As a result of our initial cleaning, we obtained 258790 unique protein sequences, of which 86653 were positive (for example, viral protein sequences from humans) and 172137 were negative (for example, viral protein sequences from non-humans). The extensive distribution of the dataset for various Influenza A proteins is given in the table(Table 1) below.

Table 1: Distribution of positive and negative dataset for the Influenza A protein sequences

Protein Name	Positive Dataset	Negative Dataset	Total Dataset
Polymerase Acidic Protein(PA)	7722	14651	22373
Polymerase Basic Protein 1(PB1)	6011	12407	18418
Polymerase Basic Protein 2(PB2)	7961	14165	22126
Hemagglutinin(HA)	17999	27350	45349
Nucleoprotein(NP)	3716	9031	12747
Neuraminidase(NA)	13486	20315	33801
Matrix Protein 1(M1)	1309	2937	4246
Matrix Protein 2(M2)	1706	4533	6239
Non-Structural Protein 1(NS1)	5577	10414	15991
Non-Structural Protein 2(NS2)	1376	4158	5534
PB1-F2	2298	9441	11739
PB1-N40	5028	10893	15921
PA-N155	4687	10716	15403

PA-N182	4422	9739	14161
PA-X	3355	11387	14742

Feature Generation

Composition-based Features

We used Pfeature to calculate amino acid composition(AAC) and dipeptide composition(DPC) for the feature generation. Pfeature is a valuable tool for the annotation of structural and functional properties of protein sequences[43]. Composition-based features were computed for both the AAC and DPC models, with a feature vector length of 20 features for the former and 400 features for the latter.

AAC is the simplest feature extraction method used for the analysis of protein structures where the composition of each residue of the protein sequence is computed, using the formula:

$$AACi = Ri / L \quad \text{(i)}$$

where AACi is the amino acid composition of type i residue; Ri and L number of residues of type i and length of the sequence, respectively.

DPC considers the composition of amino acid residues and their local order, computed using the formula:

$$DPCij = Di\,j\,L - j \quad \text{(ii)}$$

where DPCij is the composition of the dipeptide of type i for jth order. Dij and L are the numbers of dipeptides of type i and the length of a protein sequence, respectively.

For higher-order dipeptide Dj, i is made of residue Ri and Ri+j where the value of j is 2 or more. In case j is equal to 1 then dipeptide is called a traditional dipeptide.

One-Hot Encoding based Features

Furthermore, one-hot encoding was another approach that we tried for feature extraction. For this process, there needs to be equal lengths of the sequences present in the dataset. However, the protein sequence data were of unequal lengths. As a result, we took the 0.95 quantiles of all the sequence lengths to equalize the lengths. By maintaining that threshold value, the sequences short of that were extended with normal repetitions until they reached the 0.95 quantiles threshold for all the sequences present in the dataset[40]. Sequences that exceeded threshold were truncated until they reached the 0.95 quantile of all sequences mark. In one hot encoding, the sequences are numerically encoded after their lengths are equalized. Throughout the sequence, the amino acid residues are converted into matrices of [20 x 1]. Specifically, 20 x 1 refers to the 20 canonical residue positions of a protein, where a particular residue is given 1 and the rest is given a value of 0. This generates a sparse matrix of

$$vector\ size\ =\ length\ of\ the\ sequence\ x\ 20 \quad \text{(iii)}$$

This renders a sparse matrix. Alanine, for example, is described as 10000000000000000000 in the 20-dimensional vector. Likewise, Cysteine(C) is represented as a 20-dimensional vector 00100000000000000000 in the sequences present in the dataset.

Feature Selection

A significant challenge in this study is to identify the set of critical features that will be important in in the machine learning models. Here, after the sequence data was converted into one-hot encoding,resultant matrix became sparse. In such cases. dimensionality reduction techniques are needed to reduce and make a choice of the number of components to use in the model. Truncated Singular Value Decomposition (tSVD) was used to reduce the sparse matrix into 100 principal components. This was done by computing the explained variance[44]. This dimensionality reduction helped improve the computation efficiency, aiding to executing the program faster as well as reduces the noise in the data.

Machine Learning Models

The model in this analysis is a binary classification model based on machine learning algorithms, which is a combination of K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF). An instance-based classification method, KNN relies on data points closest to an instance for classification [45]. In KNN, the classification of each data point is differentiated by its majority vote of the nearest neighbor based on its instances of the training variables. The RF classifier, being an ensemble-based approach for classification[46], fits a number of decision trees in parallel, to predict the response variable. Taking the mean of the DTs upgrades the prediction accuracy as well as its control on the overfitting of the models[47]. SVM or Support Vector Machine is a supervised machine learning algorithm that can be applied to both classification and regression[48]. Using the SVM algorithm, we aim to find a hyperplane in n-dimensional space that defines distinct classes for data points. The features of AAC, DPC, as well as One Hot Encoding, were taken into account for each protein using SVM, RF, and KNN. These classifiers were formulated using the sci-kit learn library of Python [49].

Cross-Validation

5-fold cross-validation was conducted on each model after splitting the data into training and validation datasets. The whole dataset for each protein was split into training and validation datasets in an 80:20 ratio, where 80% was kept for training and 20% was kept for validation. The validation dataset was kept independent so that it can be used for external validation. Additionally, during the five-fold cross-validation process, the training datasets were further divided into training and testing datasets. By doing this, the entire training dataset can be broken down into five equal folds. A training dataset consists of four folds, a testing dataset of one fold, and training itself is performed for each iteration. The whole process is iterated five times, where every fold is given the opportunity to be used as testing data. The mean of the results for each fold of the cross-fold validation was taken for the training dataset. This is a standard procedure that has been used in many analyses [50],[51],[52], [53].

Evaluation Parameters

Predictions were made using the saved model after cross-validation. This was a part of the external validation. As a standard measure, we computed Sensitivity, Specificity, Accuracy, Matthew's Correlation Coefficient (MCC), and Area Under the Curve (AUC) for performance evaluation of the trained models. From the above-mentoned metrics, threshold-dependent parameters are accuracy, sensitivity, and specificity. By plotting sensitivity vs 1-specificity, AUC is a threshold independent parameter. The parameters are calculated by:

$$Sensitivity = \frac{T_P}{T_P+F_N} \qquad (i)$$

$$Specificity = \frac{T_N}{T_N+F_P} \qquad (ii)$$

$$Accuracy = \frac{T_P+T_N}{T_P+T_N+F_P+F_N} \qquad (iii)$$

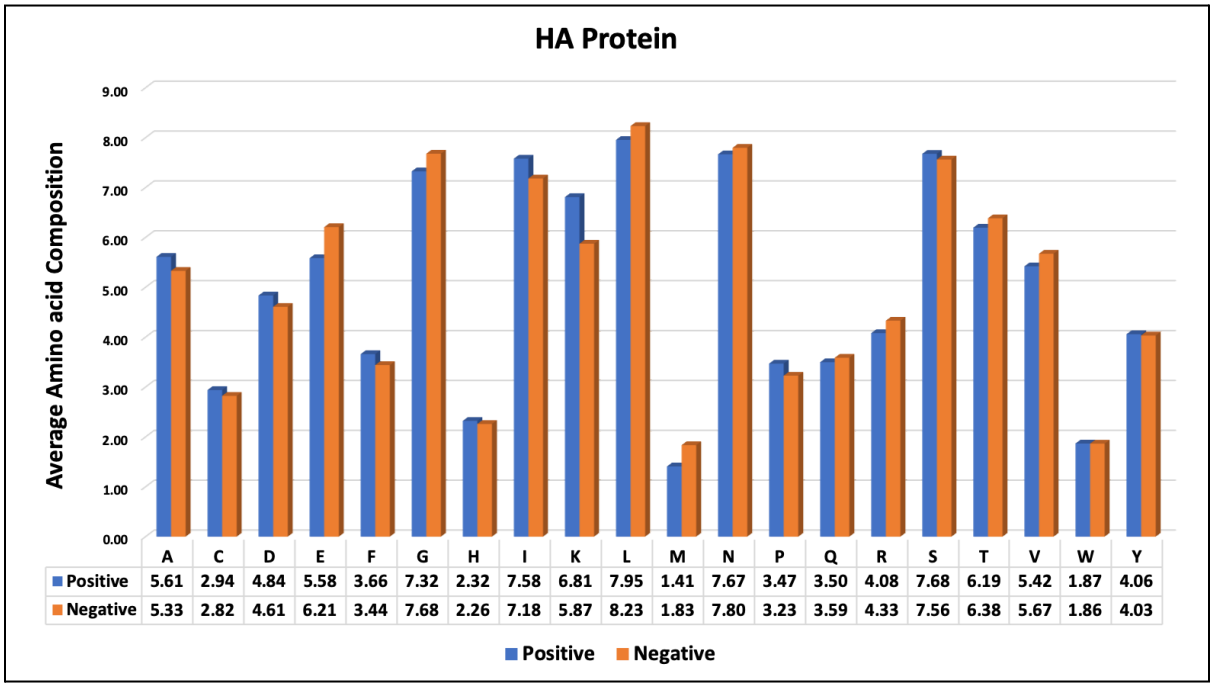
$$MCC = \frac{(T_P*T_N)-(F_P*F_N)}{\sqrt{(T_P+F_P)(T_P+F_N)(T_N+F_P)(T_N+F_N)}} \qquad (iv)$$

Where T_P is true positive, T_N is true negative, F_P is false positive and F_N is false negative.

Chapter 3: Results

Composition Based Analysis

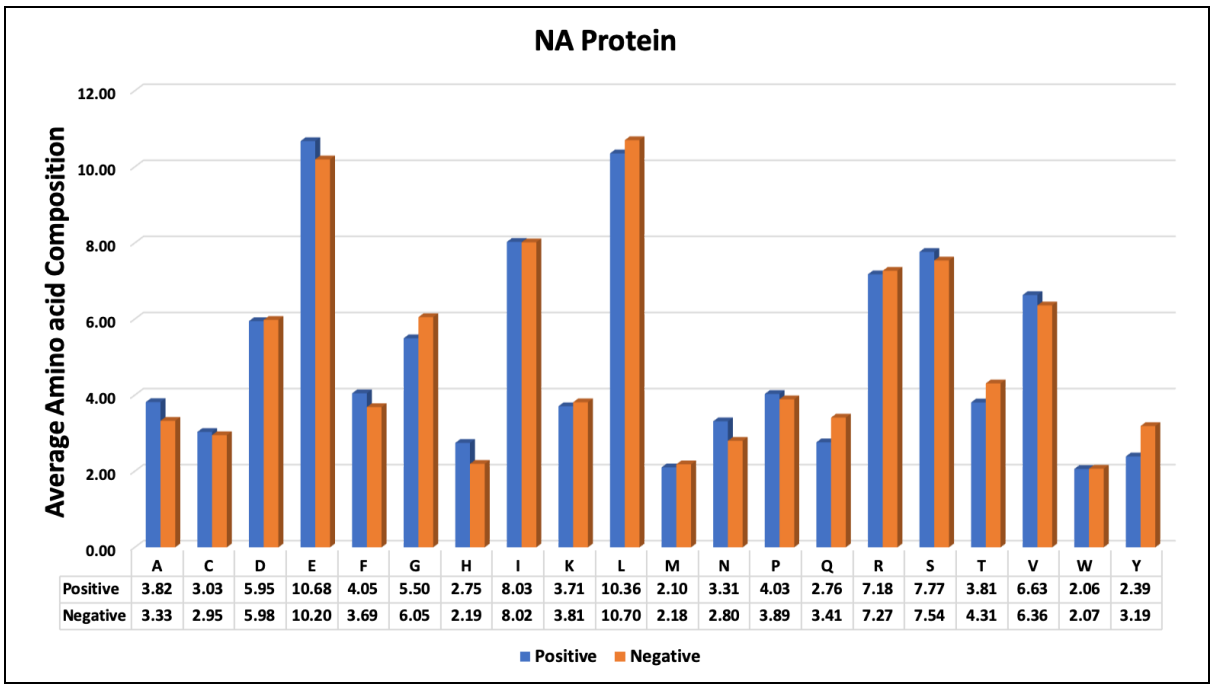
To determine the changes in average protein compositions, we calculated the average amino acid composition of 15 positive and negative sequences. Figure 4 portrays the compositional analysis for HA protein sequences based on its compositional analysis. This investigation aimed to sort out the compositional data of the proteins contrasting between the non-human and human arrangements. As referenced previously, HA and NA proteins assume a considerable part in host tropism variations in their genome structure is responsible for the different subtypes of Influenza A[54]. According to Figure 4, the mean composition of Lysine(K), Isoleucine (I), and Alanine (A) in the positive dataset is higher than that in the negative dataset for HA protein, where Glycine (G), Leucine (L), Glutamic Acid (E), and Methionine (M) are increased than those in the positive dataset.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Positive	5.61	2.94	4.84	5.58	3.66	7.32	2.32	7.58	6.81	7.95	1.41	7.67	3.47	3.50	4.08	7.68	6.19	5.42	1.87	4.06
Negative	5.33	2.82	4.61	6.21	3.44	7.68	2.26	7.18	5.87	8.23	1.83	7.80	3.23	3.59	4.33	7.56	6.38	5.67	1.86	4.03

Figure 4: Average Amino-Acid Compositional Analysis of HA Protein

Likewise, for NA protein (Figure 5), in the positive dataset, Histidine(H), Alanine(A), Glutamic Acid(E), and Asparagine(N) have higher average compositions than Glycine(G) and Tyrosine(Y)



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Positive	3.82	3.03	5.95	10.68	4.05	5.50	2.75	8.03	3.71	10.36	2.10	3.31	4.03	2.76	7.18	7.77	3.81	6.63	2.06	2.39
Negative	3.33	2.95	5.98	10.20	3.69	6.05	2.19	8.02	3.81	10.70	2.18	2.80	3.89	3.41	7.27	7.54	4.31	6.36	2.07	3.19

Figure 5: Average Amino-Acid Compositional Analysis of NA Protein

The compositional analysis has also been performed for the rest of the proteins, as illustrated in the figures below. Every protein shows a distinction in the average composition between the positive and negative datasets. The NS1 protein shown below has a higher composition of Glycine(G), Cysteine(C), Histidine(H), Methionine(M), and Glutamine(Q) with Lysine(K), Asparagine(N), and Valine(V) being significantly higher in the positive dataset as compared to Alanine(A), Aspartic Acid(D), Leucine(L), Isoleucine(I) and Arginine(R) being higher in the negative dataset.

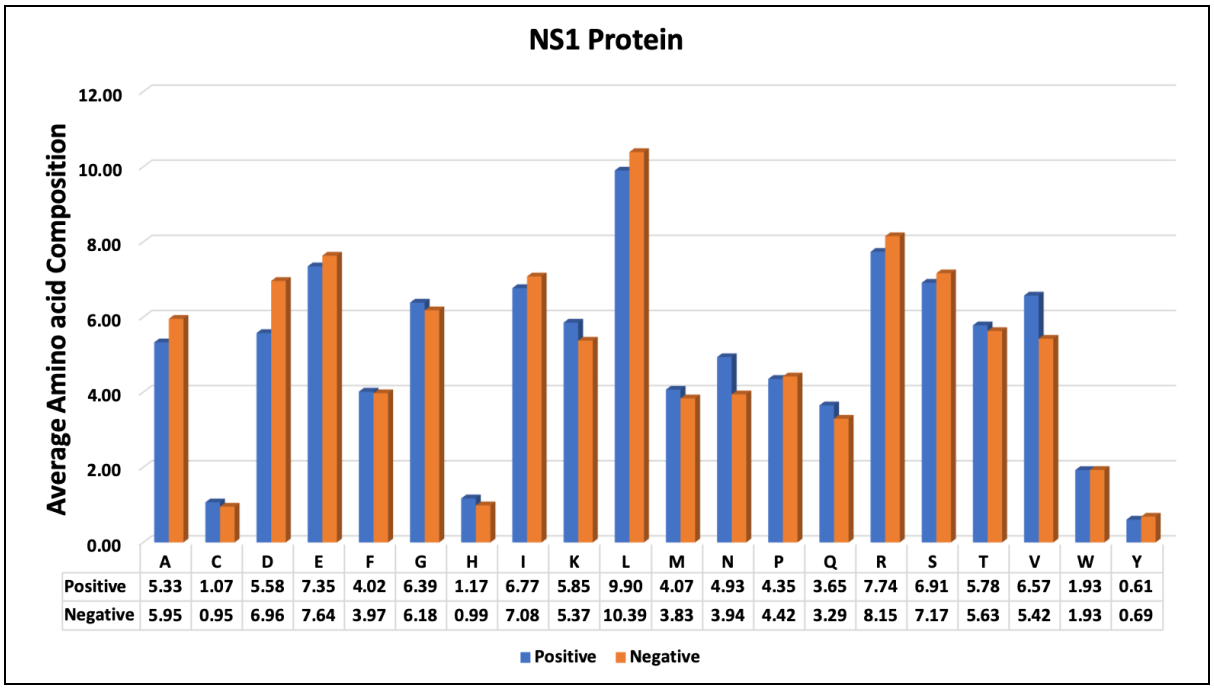


Figure 6: Average Amino-Acid Compositional Analysis of NS1 Protein

Figure 7 describes the compositional analysis of NS2 protein where the composition of Cysteine(C) and Proline(P) have a less impact compared to the other amino acids. Methionine(M) and Aspartic Acid(N) apart from Glutamic Acid(E), Glycine(G), and Histidine(H) have a significant contribution in the positive sequences compared to the negative ones including Serine(S), Threonine(T), Valine(V), Aspartic Acid(N), Isoleucine(I).

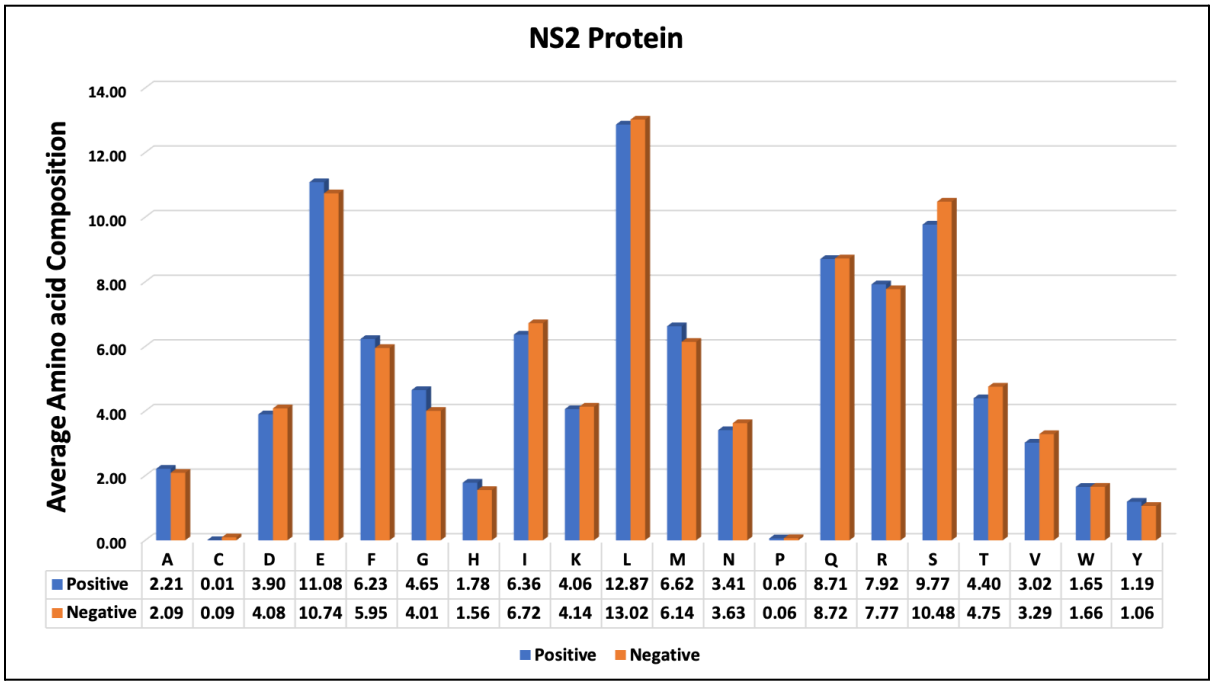
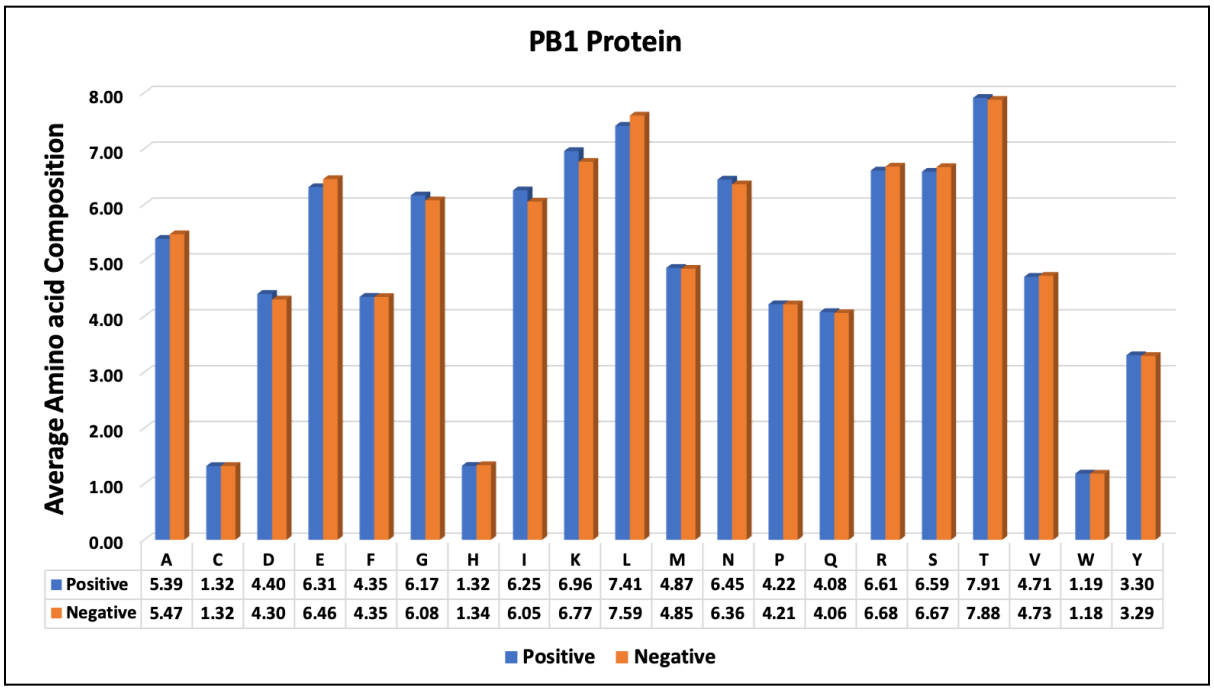


Figure 7: Average Amino-Acid Compositional Analysis of NS2 Protein

Both PB1 and PB2 are known to play a significant role in host tropism, as per the studies conducted before. The positive sequences have a higher composition of Aspartic Acid(D), Aspartic Acid(N), Isoleucine(I), and Lysine(K) whereas Leucine(L), Alanine(A), and Glutamic acid(E) are higher in the negative dataset.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Positive	5.39	1.32	4.40	6.31	4.35	6.17	1.32	6.25	6.96	7.41	4.87	6.45	4.22	4.08	6.61	6.59	7.91	4.71	1.19	3.30
Negative	5.47	1.32	4.30	6.46	4.35	6.08	1.34	6.05	6.77	7.59	4.85	6.36	4.21	4.06	6.68	6.67	7.88	4.73	1.18	3.29

Figure 8: Average Amino-Acid Compositional Analysis of PB1 Protein

In PB2, Serine(S) and Valine(V) are significantly higher in the positive dataset compared to the negative, where Leucine(L), Alanine(A), and Glutamic acid(E), and Glutamine(Q) are higher.

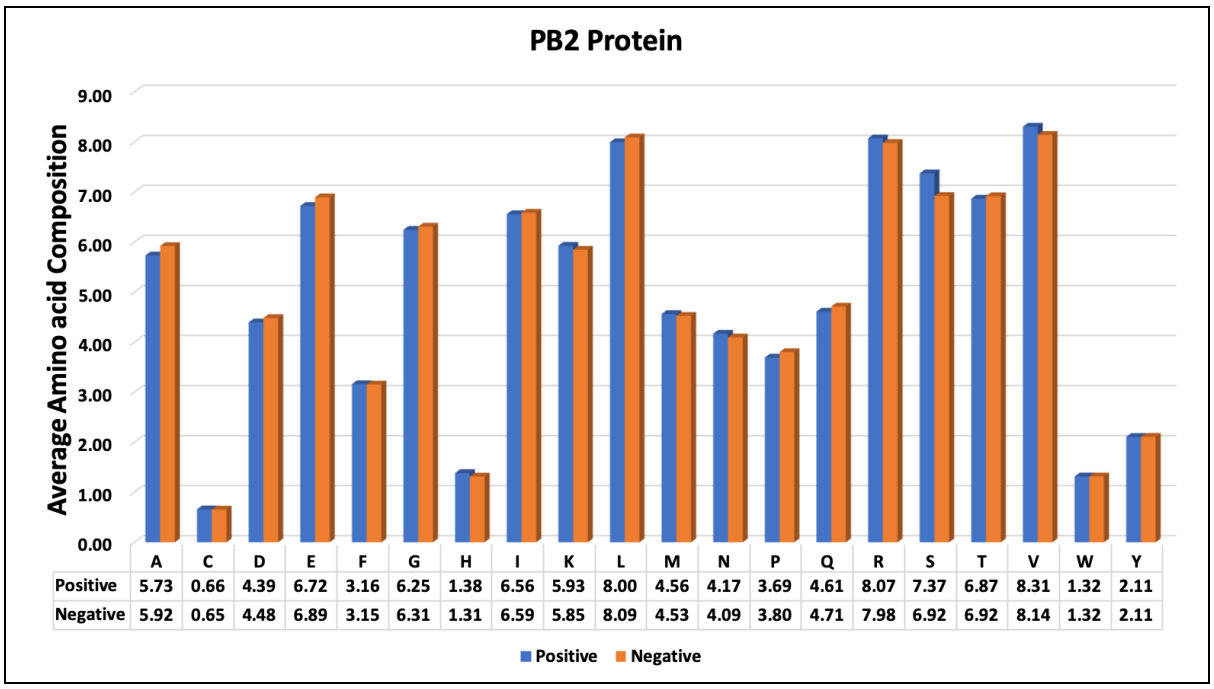


Figure 9: Average Amino-Acid Compositional Analysis of PB2 Protein

Aspartic Acid(N), Isoleucine(I), and Lysine(K) whereas Leucine(L), and Tyrosine(Y) have a higher composition in the positive data compared to Proline(P), Arginine(R), Serine(S), Threonine(T) in the negative dataset.

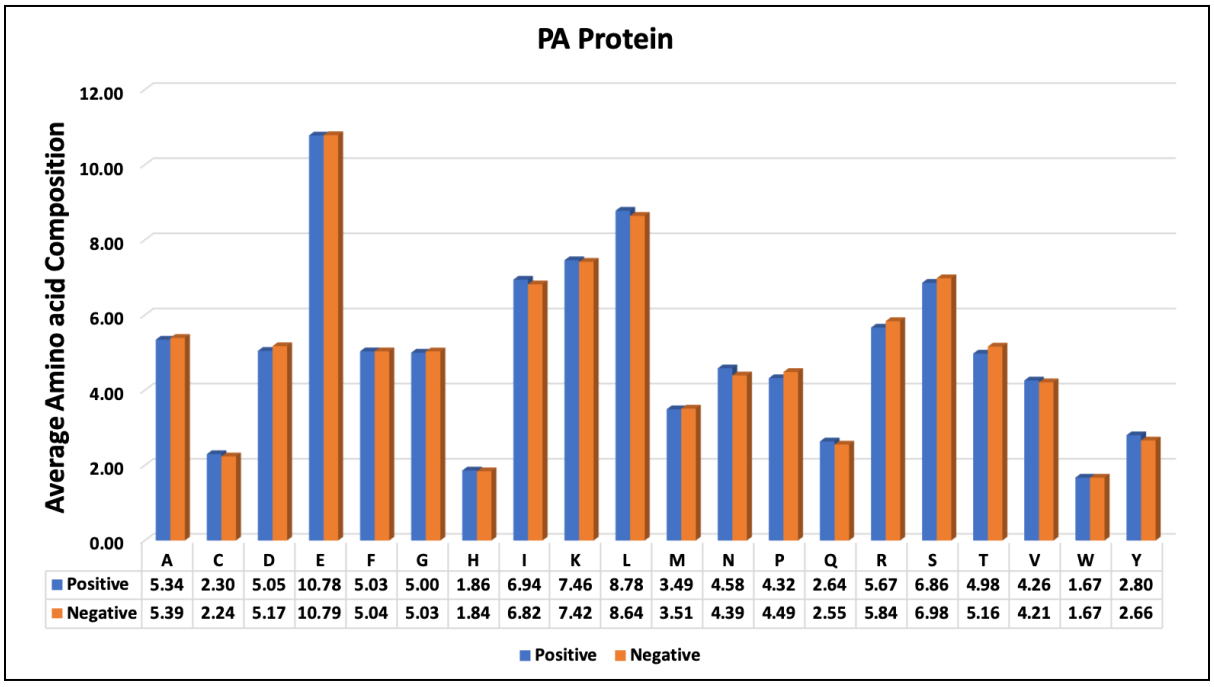
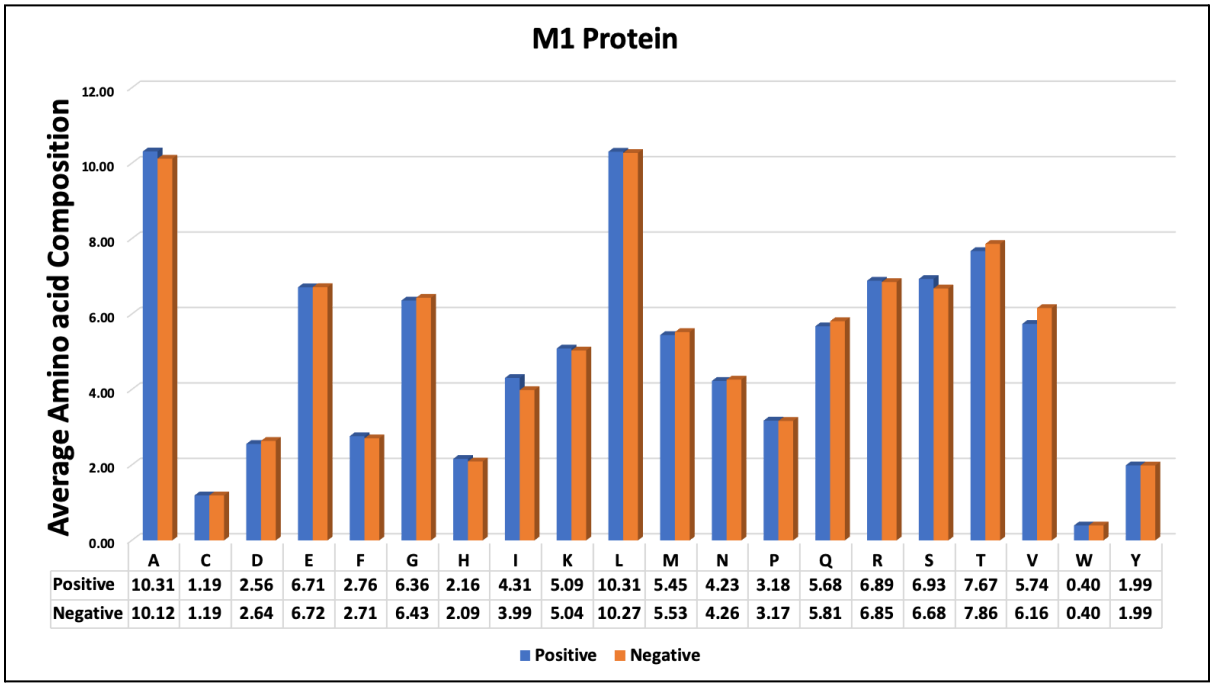


Figure 10: Average Amino-Acid Compositional Analysis of PA Protein

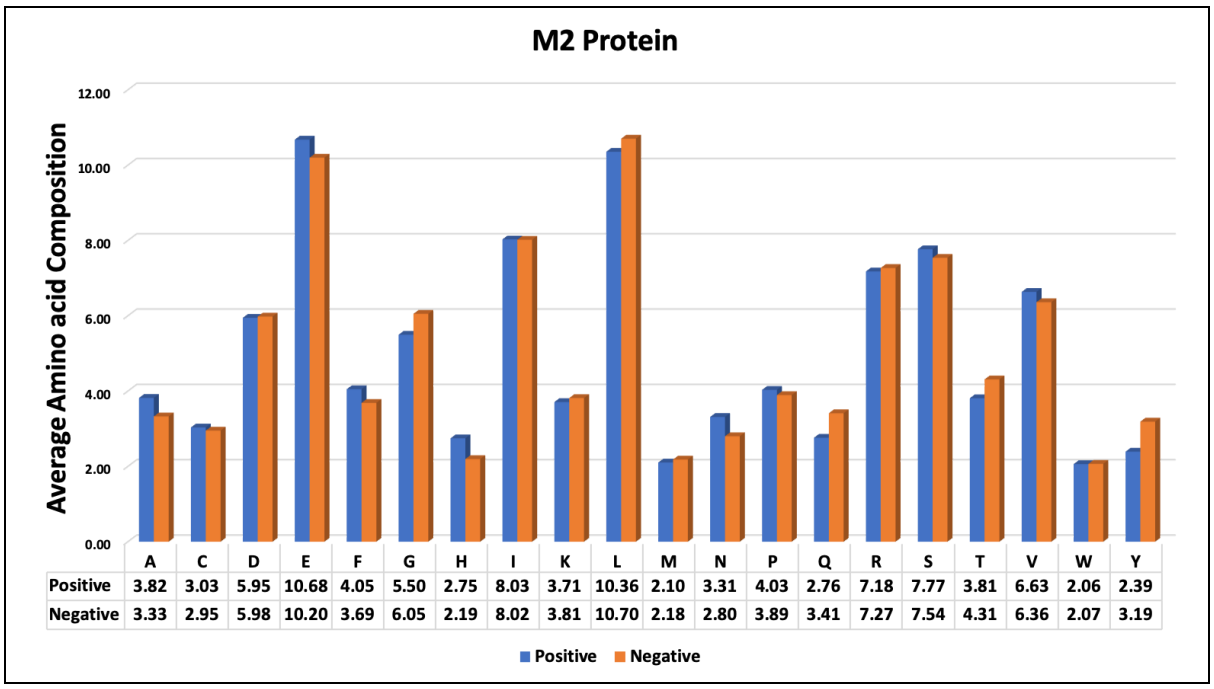
Serine(S) and Isoleucine(I), as seen in Figure 11, is much higher in the positive dataset. In the negative dataset, Threonine(T) and Valine(V) are much higher.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Positive	10.31	1.19	2.56	6.71	2.76	6.36	2.16	4.31	5.09	10.31	5.45	4.23	3.18	5.68	6.89	6.93	7.67	5.74	0.40	1.99
Negative	10.12	1.19	2.64	6.72	2.71	6.43	2.09	3.99	5.04	10.27	5.53	4.26	3.17	5.81	6.85	6.68	7.86	6.16	0.40	1.99

Figure 11: Average Amino-Acid Compositional Analysis of M1 Protein

M2 is a small protein but it is interesting to note that even that protein can be used to predict host tropism based on compositional features only. The difference in the average compositions of Alanine(A), Glutamic acid(E), Phenylalanine(F), Histidine(H), and Aspartic Acid(N) in the positive dataset and Glycine(G), Glutamine(Q), Threonine(T), Leucine(L), and Tyrosine(Y) in the negative dataset can be a contributing factor to for host tropism.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Positive	3.82	3.03	5.95	10.68	4.05	5.50	2.75	8.03	3.71	10.36	2.10	3.31	4.03	2.76	7.18	7.77	3.81	6.63	2.06	2.39
Negative	3.33	2.95	5.98	10.20	3.69	6.05	2.19	8.02	3.81	10.70	2.18	2.80	3.89	3.41	7.27	7.54	4.31	6.36	2.07	3.19

Figure 12: Average Amino-Acid Compositional Analysis of M2 Protein

In this analysis interestingly, non-structural proteins also show a difference in the average amino acid compositions between the positive and negative datasets. So in a way, they also are important for host tropism. As an example, PB1-F2 contains a remarkable amount of Glycine (G), Proline (P), and Glutamine (Q). This is evident in Figure 13 where the composition of Cysteine (C), Glutamic Acid (E), Lysine (K), and Leucine (L) is significantly higher in the negative dataset than it is in the positive dataset.

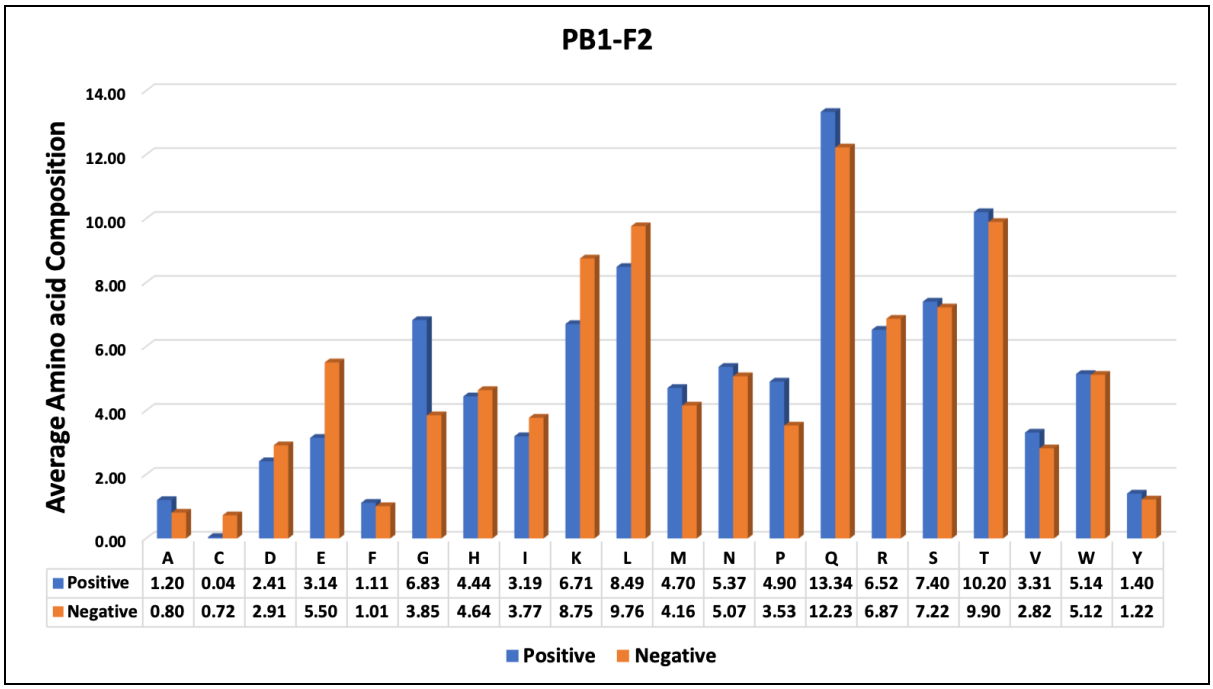


Figure 13: Average Amino-Acid Compositional Analysis of PB1F2 Protein

Similarly, Isoleucine(I), Lysine(K), and Glycine(G), are higher in positive and Leucine(L), Alanine(A), and Glutamic acid(E) are higher in the negative dataset for PB1-N40 protein.

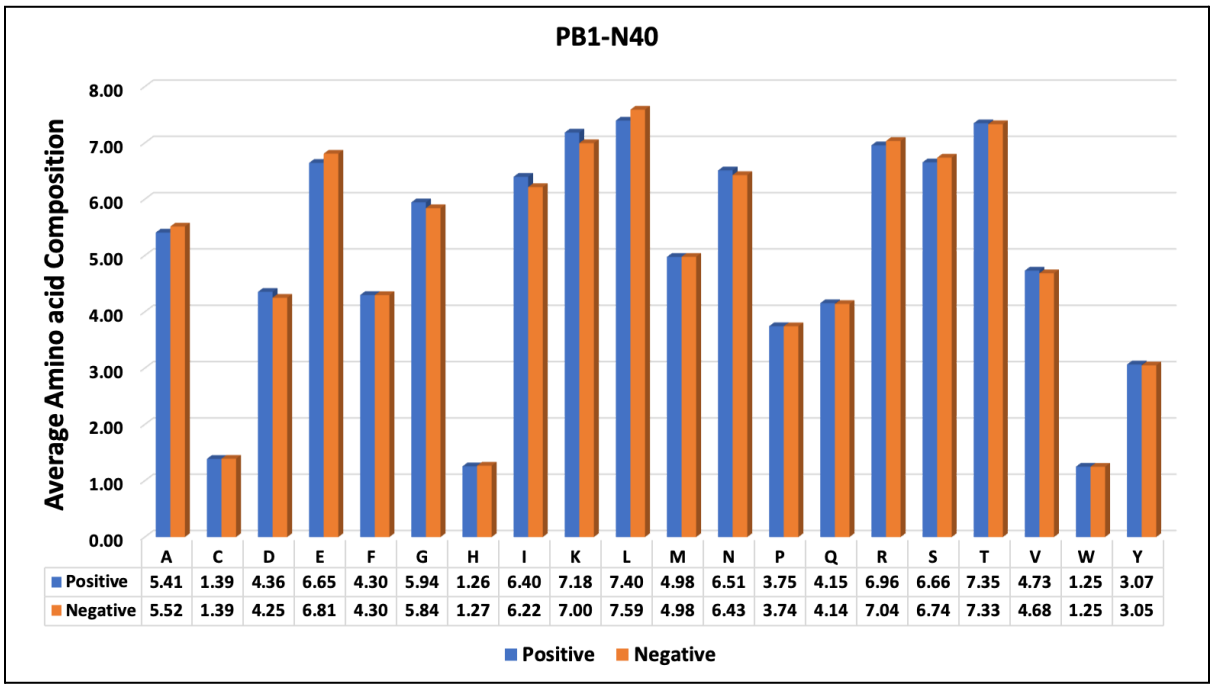
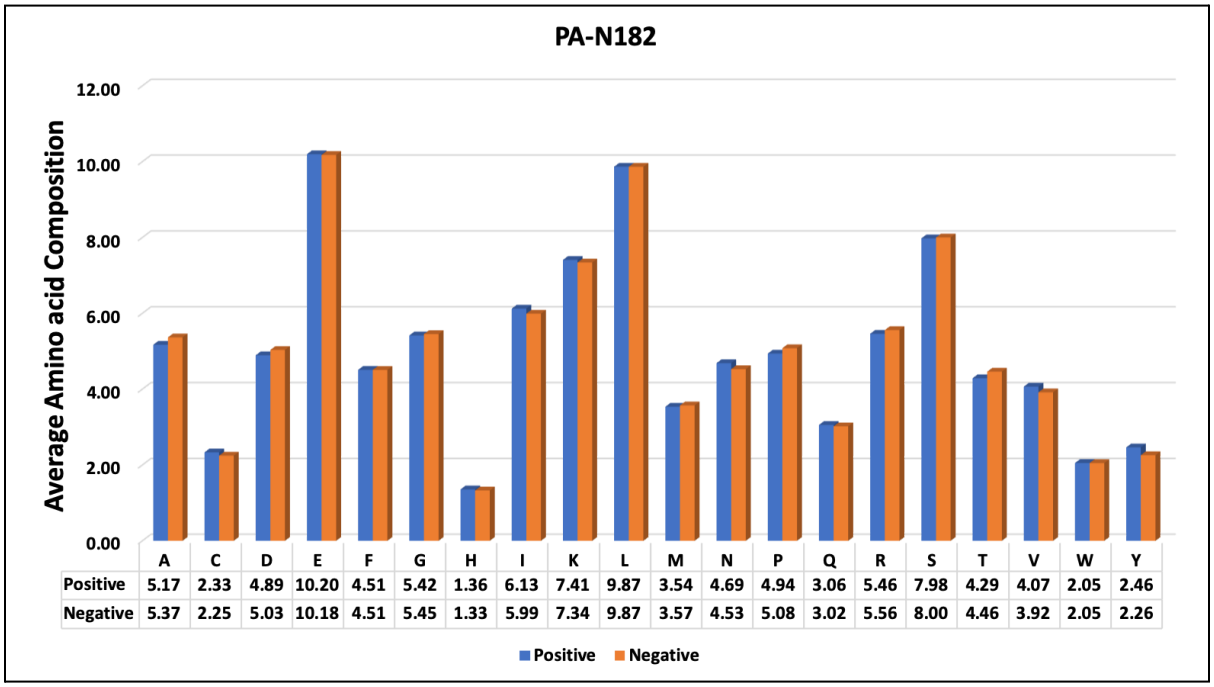


Figure 14: Average Amino-Acid Compositional Analysis of PB1-N40 Protein

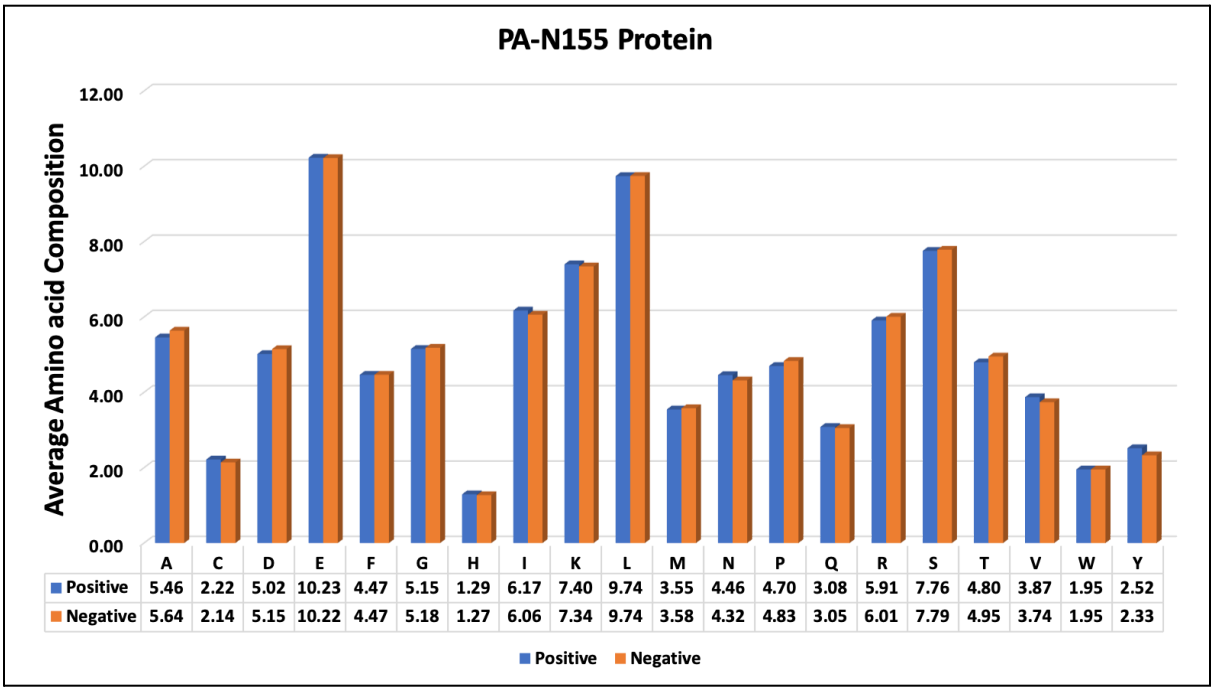
PA-N182 has a higher composition of Valine(V), Aspartic Acid(N), Isoleucine(I), Lysine(K) in positive and Alanine(A), Threonine(T), Proline (P) in negative data.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Positive	5.17	2.33	4.89	10.20	4.51	5.42	1.36	6.13	7.41	9.87	3.54	4.69	4.94	3.06	5.46	7.98	4.29	4.07	2.05	2.46
Negative	5.37	2.25	5.03	10.18	4.51	5.45	1.33	5.99	7.34	9.87	3.57	4.53	5.08	3.02	5.56	8.00	4.46	3.92	2.05	2.26

Figure 15: Average Amino-Acid Compositional Analysis of PA-N182 Protein

PA-N155 has a higher composition of Aspartic Acid(N), Tyrosine(Y) in positive and a higher composition of Proline (P), Arginine(R), and Alanine(A) in negative.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Positive	5.46	2.22	5.02	10.23	4.47	5.15	1.29	6.17	7.40	9.74	3.55	4.46	4.70	3.08	5.91	7.76	4.80	3.87	1.95	2.52
Negative	5.64	2.14	5.15	10.22	4.47	5.18	1.27	6.06	7.34	9.74	3.58	4.32	4.83	3.05	6.01	7.79	4.95	3.74	1.95	2.33

Figure 16: Average Amino-Acid Compositional Analysis of PA-N155 Protein

Lastly, PA-X protein shows a significant difference in Isoleucine(I), Aspartic Acid(N), Leucine(L) in positive and Asparrtic Acid(D),Glycine(G), Proline (P), Threonine(T), Arginine(R) in negative dataset.

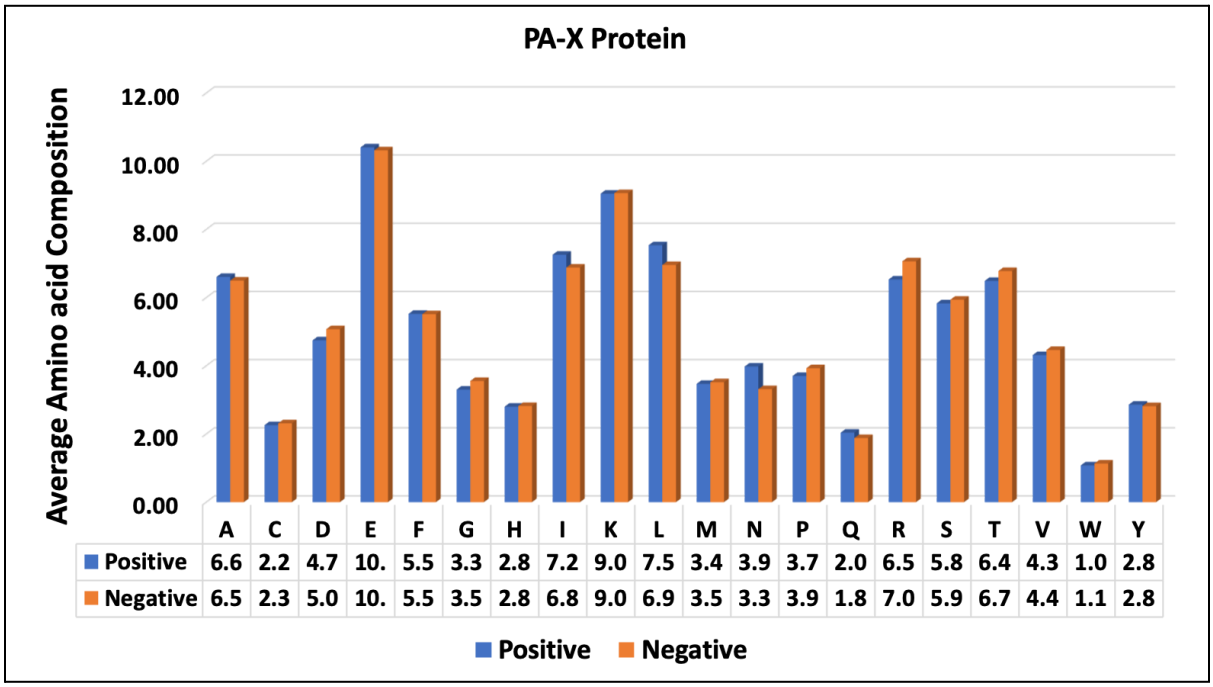


Figure 17: Average Amino-Acid Compositional Analysis of PA-X Protein

The compositional analysis also validates the study that composition-based features can be used to determine the host tropism. Mutations and reassortments in the segments can have the potential to change the underlying host-specificity of the virus. A difference in the compositions of the amino acids which play a role in the virus is capable of crossing the species barrier to infect humans.

Motif-based Analysis

We employed MERCI software to figure out the different motifs surfacing solely in both the positive and negative datasets [55]. It identifies and finds top K motifs that are most common in positive sequences and are absent in negative sequences. By doing so, we added to our understanding of what motifs are likely to be present in the positive sequences, which might explain why they are infectious. The top 10 motifs of each of the proteins which are solely present in the positive and negative datasets are shown in the table below:-

Table 2: Top 10 Motifs of Each Protein Exclusive to Positive and Negative Datasets

Protein	Positive			Negative		
PA	Motifs	Motifs Length	Occurrences	Motifs	Motifs Length	Occurrences
	I R L P N G P P C F	10	259	A W K Q V L A E L Q D L	12	1734
	I R L P N G P P C F Q	11	259	A W K Q V L A E L Q D L E	13	1728
	I R L P N G P P C F Q R	12	259	L A W K Q V L A E L Q D L	13	1671
	I R L P N G P P C F Q R S	13	259	L A W K Q V L A E L Q D L E	14	1665
	I R L P N G P P C F Q R S K	14	259	L L A W K Q V L A E L Q D L	14	1663
	I R L P N G P P C F Q R S K F	15	259	Y L L A W K Q V L A E L Q D L	15	1663
	P I R L P N G P P C F	11	256	L L A W K Q V L A E L Q D L E	15	1657
	P I R L P N G P P C F Q	12	256	P E Q	3	1565
	P I R L P N G P P C F Q R	13	256	E P E Q	4	1563
	P I R L P N G P P C F Q R S	14	256	Q R S L	4	1553
	P I R L P N G P P C F Q R S K	15	256			
HA	Motifs	Motifs Length	Occurrences	Motifs	Motifs Length	Occurrences
	F L Y A Q S	6	4471	K A I D G I	6	2603
	F L Y A Q S S	7	4429	Q K A I D G I	7	2601
	F L Y A Q S S G	8	4429	T Q K A I D G I	8	2601
	I F L Y A Q S	7	4427	S T Q K A I D G I	9	2599
	Q I F L Y A Q S	8	4427	K A I D G I T	7	2594
	D Q I F L Y A Q S	9	4419	Q K A I D G I T	8	2592
	F L Y A Q S S G R	9	4416	T Q K A I D G I T	9	2592
	K D Q I F L Y A Q S	10	4410	S T Q K A I D G I T	10	2590
	F L Y A Q S S G R I	10	4396	K A I D G I T N	8	2553
	F L Y A Q S S G R I T	11	4388	Q K A I D G I T N	9	2551
				T Q K A I D G I T N	10	2551

M1	**Motifs**	**Motifs Length**	**Occurrences**	**Motifs**	**Motifs Length**	**Occurrences**
	G L K N D L L D	8	20	A S Q T R Q M V H	9	108
	G L K N D L L D N	9	20	A S Q T R Q M V H A	10	108
	G L K N D L L D N L	10	20	A S Q T R Q M V H A M	11	108
	G L K N D L L D N L Q	11	20	A S Q T R Q M V H A M R	12	108
	G L K N D L L D N L Q A	12	20	S Q T R Q M V H	8	108
	G L K N D L L D N L Q A Y	13	20	S Q T R Q M V H A	9	108
	G L K N D L L D N L Q A Y Q	14	20	S Q T R Q M V H A M	10	108
	G L K N D L L D N L Q A Y Q K	15	20	S Q T R Q M V H A M R	11	108
	H P S S S T G L K N D L L D	14	20	V A S Q T R Q M V H	10	107
	H P S S S T G L K N D L L D N	15	20	V A S Q T R Q M V H A	11	107
	K N D L L D	6	20	V A S Q T R Q M V H A M	12	107
				V A S Q T R Q M V H A M R	13	107
M2	**Motifs**	**Motifs Length**	**Occurrences**	**Motifs**	**Motifs Length**	**Occurrences**
	I F K	3	149	G W E C K C S	7	834
	R I F K	4	147	G W E C K C S D	8	779
	Y R I F K	5	143	G W E C K C S D S	9	778
	I F K H	4	140	G W E C K C S D S S	10	756
	I F K H G	5	140	G W E C K C S D S S D	11	749
	R I F K H	5	139	G W E C K C S D S S D P	12	737
	R I F K H G	6	139	P T R N G W E C K C	10	732
	I F K H G L	6	138	T P T R N G W E C K C	11	728
	I Y R I F K	6	137	E T P T R N G W E C K C	12	722
	R I F K H G L	7	137	G W E C K C S D S S D P L	13	720

NA	Motifs	Motifs Length	Occurrences	Motifs	Motifs Length	Occurrences
	V V S W S K E	7	1919	K Q N E C	5	2041
	D S V V S W S K E	9	1917	F K Q N E C	6	2039
	S V V S W S K E	8	1917	H F K Q N E C	7	1948
	V D S V V S W S K E	10	1913	L H F K Q N E C	8	1938
	L V D S V V S W S K E	11	1912	T L H F K Q N E C	9	1931
	V V S W S K E I	8	1878	D I N M A	5	1744
	D S V V S W S K E I	10	1876	D I N M A D	6	1719
	S V V S W S K E I	9	1876	D I N M A D Y	7	1706
	V V S W S K E I L	9	1875	D I N M A D Y S	8	1640
	V V S W S K E I L R	10	1875	I D I N M A	6	1628
NP	**Motifs**	**Motifs Length**	**Occurrences**	**Motifs**	**Motifs Length**	**Occurrences**
	F D K A	4	107	R R D G K W M	7	2579
	F D K A T	5	107	R R D G K W M R	8	2574
	L P F D K A	6	107	R R D G K W M R E	9	2571
	L P F D K A T	7	107	R R D G K W M R E L	10	2560
	P F D K A	5	107	R R D G K W M R E L I	11	2484
	P F D K A T	6	107	R R D G K W M R E L I L	12	2484
	D K A T I	5	106	R D G K W M R E L I L Y D K E	15	2475
	D K A T I M	6	106	R R R D G K W M	8	2454
	F D K A T I	6	106	R R D G K W M R E L I L Y	13	2451
	F D K A T I M	7	106	Y R R R D G K W M	9	2451
	F S V Q R N L P F D K A	12	106			
NS1	**Motifs**	**Motifs Length**	**Occurrences**	**Motifs**	**Motifs Length**	**Occurrences**
	N I M L K A N F N	9	1652	Y M A	3	1077

	K N I M L K A N F N	10	1650	M A R R	4	1053
	N I M L K A N F N V	10	1645	R Y M A	4	1046
	K N I M L K A N F N V	11	1643	K R Y M A	5	1035
	E K N I M L K A N F N	11	1627	Q K R Y M A	6	1034
	E K N I M L K A N F N V	12	1620	K Q K R Y M A	7	1029
	N I M L K A N F N V I	11	1599	P P K Q K R Y	7	1027
	K N I M L K A N F N V I	12	1597	G P P L P P K Q K R	10	1023
	N I M L K A N F N V I F	12	1595	Y M A R	4	1019
	K N I M L K A N F N V I F	13	1594	L P P K Q K R Y	8	1013
				P K Q K R Y M A	8	1013
				P P K Q K R Y M	8	1013
NS2	**Motifs**	**Motifs Length**	**Occurrences**	**Motifs**	**Motifs Length**	**Occurrences**
	V M R L G D L H S L Q H	12	72	I T F L	4	445
	V M R L G D L H S L Q H R	13	72	L Q A	3	445
	V M R L G D L H S L Q H R N	14	71	L Q A L	4	442
	A V M R L G D L H S L Q H	13	68	E Q I T F L	6	441
	A V M R L G D L H S L Q H R	14	68	F L Q A	4	441
	A V M R L G D L H S L Q H R N	15	67	Q I T F L	5	441
	V M R L G D L H S L Q H R N G	15	67	T F L Q	4	441
	E A V M R L G D L H S L Q H	14	66	T F L Q A	5	441
	E A V M R L G D L H S L Q H R	15	66	F E Q I T F L	7	440
	G E A V M R L G D L H S L Q H	15	66	F L Q A L	5	440
				I T F L Q	5	440
				I T F L Q A	6	440
				L Q A L Q	5	440
				T F L Q A L	6	440

PB1	**Motifs**	**Motifs Length**	**Occurrences**	**Motifs**	**Motifs Length**	**Occurrences**
	IKKLWDQTQSRT	12	1680	LNRRSYLIRA	10	413
	IKKLWDQTQSRTG	13	1680	LNRRSYLIRAL	11	412
	IKKLWDQTQSRTGL	14	1680	LNRRSYLIRALT	12	412
	IKKLWDQTQSRTGLL	15	1680	LNRRSYLIRALTL	13	411
	EIKKLWDQTQSRT	13	1679	LNRRSYLIRALTLN	14	411
	EIKKLWDQTQSRTG	14	1679	LNRRSYLIRALTLNT	15	411
	EIKKLWDQTQSRTGL	15	1679	RLNRRSYLIRA	11	381
	FEIKKLWDQTQSRT	14	1673	QRLNRRSYLIRA	12	380
	FEIKKLWDQTQSRTG	15	1673	RLNRRSYLIRAL	12	380
	SFEIKKLWDQTQSRT	15	1671	RLNRRSYLIRALT	13	380
				RLNRRSYLIRALTL	14	380
				RLNRRSYLIRALTLN	15	380
PB2	**Motifs**	**Motifs Length**	**Occurrences**	**Motifs**	**Motifs Length**	**Occurrences**
	KYPITADKRVTEM	13	581	AGALAEDPDEGTA	13	343
	MKYPITADKRVTEM	14	581	AGALAEDPDEGTAG	14	343
	AMKYPITADKRVTEM	15	578	ALAEDPDEGTA	11	343
	VTEMV	5	531	ALAEDPDEGTAG	12	343
	VTEMVP	6	531	GALAEDPDEGTA	12	343
	VTEMVPE	7	530	GALAEDPDEGTAG	13	343
	RVTEMV	6	527	AGALAEDPDEGTAGV	15	342
	RVTEMVP	7	527	ALAEDPDEGTAGV	13	342
	VTEMVPER	8	527	ALAEDPDEGTAGVE	14	342
	VTEMVPERN	9	527	GALAEDPDEGTAGV	14	342
	VTEMVPERNE	10	527	GALAEDPDEGTAGVE	15	342

PB1-F2	**Motifs**	**Motifs Length**	**Occurrences**	**Motifs**	**Motifs Length**	**Occurrences**
	Q I Q K L G R P S	9	694	M D H C L R T M	8	1033
	Q I Q K L G R P S S	10	686	W T R L	4	993
	Q I Q K L G R P S S T	11	665	R L T E	4	987
	Q I Q K L G R P S S T Q	12	661	W T R L T	5	985
	Q I Q K L G R P S S T Q L	13	658	R L T E H	5	972
	Q I Q K L G R P S S T Q L M	14	653	L M D H C L R T M	9	949
	Q I Q K L G R P S S T Q L M D	15	644	T R L T E	5	948
	R G G S G R Q	7	624	W T R L T E	6	946
	Q R G G S G R Q	8	622	R L T E H I	6	934
	R G G S G R Q I	8	589	T R L T E H	6	933
PB1-N40	**Motifs**	**Motifs Length**	**Occurrences**	**Motifs**	**Motifs Length**	**Occurrences**
	Q S R T	4	1310	L N R R S Y L I R A	10	352
	Q S R T G	5	1310	L N R R S Y L I R A L	11	351
	Q S R T G L	6	1309	L N R R S Y L I R A L T	12	350
	Q S R T G L L	7	1309	L N R R S Y L I R A L T L	13	349
	T Q S R T	5	1309	L N R R S Y L I R A L T L N	14	349
	T Q S R T G	6	1309	L N R R S Y L I R A L T L N T	15	349
	T Q S R T G L	7	1308	Q R L N R R S Y L I R A	12	322
	T Q S R T G L L	8	1308	R L N R R S Y L I R A	11	322
	Q T Q S R T	6	1306	Q R L N R R S Y L I R A L	13	321
	Q T Q S R T G	7	1306	R L N R R S Y L I R A L	12	321
PA-N155	**Motifs**	**Motifs Length**	**Occurrences**	**Motifs**	**Motifs Length**	**Occurrences**
	S E F N K A C E L T D S V	13	1227	A W K Q V L A E L Q D L	12	1245

	SEFNKACELTDSVW	14	1227	AWKQVLAELQDLE	13	1239
	QSEFNKACELTDSV	14	1226	LAWKQVLAELQDL	13	1189
	QSEFNKACELTDSVW	15	1226	LAWKQVLAELQDLE	14	1183
	SEFNKACELTDSVWI	15	1222	LLAWKQVLAELQDL	14	1181
	LRSLSSWIQS	10	1214	YLLAWKQVLAELQDL	15	1181
	LRSLSSWIQSE	11	1214	LLAWKQVLAELQDLE	15	1175
	LRSLSSWIQSEF	12	1214	PEQ	3	1066
	LRSLSSWIQSEFN	13	1213	EPEQ	4	1065
	LRSLSSWIQSEFNK	14	1212	AWKQVLAELQDLEN	14	1064
PA-N182	**Motifs**	**Motifs Length**	**Occurrences**	**Motifs**	**Motifs Length**	**Occurrences**
	SEFNKACELTDSV	13	1188	AWKQVLAELQDL	12	1213
	SEFNKACELTDSVW	14	1188	AWKQVLAELQDLE	13	1207
	QSEFNKACELTDSV	14	1187	LAWKQVLAELQDL	13	1158
	QSEFNKACELTDSVW	15	1187	LAWKQVLAELQDLE	14	1152
	SEFNKACELTDSVWI	15	1184	LLAWKQVLAELQDL	14	1150
	LRSLSSWIQS	10	1181	YLLAWKQVLAELQDL	15	1150
	LRSLSSWIQSE	11	1181	LLAWKQVLAELQDLE	15	1144
	LRSLSSWIQSEF	12	1181	AWKQVLAELQDLEN	14	1037
	LRSLSSWIQSEFN	13	1180	PEQ	3	1036
	LRSLSSWIQSEFNK	14	1179	EPEQ	4	1035
PA-X	**Motifs**	**Motifs Length**	**Occurrences**	**Motifs**	**Motifs Length**	**Occurrences**
	SPALRILEPMWMDL	14	166	SPTKVSHRTSPALKT	15	882
	TSPALRILEPMWMDL	15	164	LQEPCAGSPTKVS	13	698
	SPALRILEPMWMDLN	15	158	KLQEPCAGSPTKVS	14	686
	LNRTATLRASFLKCP	15	142	LKLQEPCAGSPTKVS	15	682

	AFRRTSPALR	10	136	LQEPCAGSPTKVSH	14	682
	AFRRTSPALRI	11	136	LQEPCAGSPTKVSHR	15	675
	AFRRTSPALRIL	12	136	KLQEPCAGSPTKVSH	15	670
	AFRRTSPALRILE	13	136	GLLTKVSHRTSPALK	15	360
	AFRRTSPALRILEP	14	136	PCAGLLTKVSHRT	13	268
	AFRRTSPALRILEPM	15	135	PCAGLLTKVSHRTS	14	268
	KAFRRTSPALR	11	135	PCAGLLTKVSHRTSP	15	268
	KAFRRTSPALRI	12	135			
	KAFRRTSPALRIL	13	135			
	KAFRRTSPALRILE	14	135			
	KAFRRTSPALRILEP	15	135			

Performance of Machine Learning Models

The host tropism of Influenza A virus, whether it is infectious to humans or not, has been predicted using machine learning models for all influenza A proteins using sequence-based features such as AAC, DPC, as well as encoding. For developing a prediction model based on protein sequence datasets, we used SVM, RF, and KNN classifiers.

Amino Acid Composition based Models

Based on validation data, the random forest models formulated using AAC features achieved an AUC of 0.973 with an accuracy of 97.5% for HA protein, as shown in Table 3. On both training and validation datasets, PB1-F2 prediction models deliver the highest accuracy of > 98.8%. The following table (Tables 3,4 and 5) below shows that the models performed similarly on the training and validation datasets.

Table 3: The performance of Support Vector Machine based models developed using AAC features for training and validation datasets.

	TRAINING					VALIDATION				
Proteins	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**MCC**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**MCC**
PA	0.87	0.84	0.89	0.87	0.72	0.88	0.77	0.93	0.85	0.72
HA	0.95	0.93	0.97	0.95	0.9	0.96	0.96	0.95	0.96	0.91
M1	0.82	0.81	0.83	0.82	0.56	0.82	0.52	0.95	0.73	0.55
M2	0.9	0.97	0.88	0.93	0.74	0.9	0.64	0.99	0.82	0.73
NA	0.94	0.91	0.96	0.93	0.87	0.93	0.93	0.94	0.93	0.86
NP	0.93	0.94	0.93	0.94	0.84	0.93	0.83	0.98	0.9	0.84
NS1	0.95	0.91	0.98	0.94	0.9	0.96	0.92	0.98	0.95	0.9
NS2	0.87	0.8	0.89	0.85	0.64	0.87	0.65	0.93	0.79	0.62
PB1	0.91	0.84	0.95	0.89	0.79	0.91	0.89	0.92	0.9	0.79
PB2	0.9	0.91	0.9	0.9	0.79	0.91	0.81	0.96	0.88	0.79
PB1-F2	0.99	0.98	0.98	0.98	0.92	0.89	0.98	0.98	0.94	0.92
PB1-N40	0.84	0.91	0.91	0.89	0.79	0.88	0.91	0.91	0.9	0.79
PA-N155	0.89	0.84	0.85	0.86	0.6	0.55	0.84	0.82	0.75	0.59
PA-N182	0.88	0.84	0.85	0.86	0.62	0.55	0.83	0.81	0.76	0.59
PA-X	0.87	0.93	0.93	0.91	0.8	0.84	0.93	0.93	0.9	0.8

Table 4: The performance of Random Forest-based models developed using AAC features for training and validation datasets.

	TRAINING					VALIDATION				
Proteins	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**	**MCC**	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**	**MCC**
PA	0.95	0.98	0.97	0.97	0.94	0.95	0.96	0.96	0.96	0.91
HA	0.96	0.99	0.98	0.97	0.95	0.98	0.97	0.97	0.97	0.94
M1	0.95	0.94	0.94	0.94	0.86	0.73	0.95	0.88	0.84	0.71
M2	0.94	0.97	0.97	0.96	0.91	0.83	0.96	0.93	0.9	0.81

NA	0.96	0.98	0.98	0.97	0.95	0.97	0.97	0.97	0.97	0.94
NP	0.96	0.98	0.98	0.97	0.94	0.95	0.98	0.97	0.97	0.94
NS1	0.95	0.98	0.97	0.97	0.94	0.96	0.97	0.96	0.96	0.92
NS2	0.93	0.98	0.97	0.96	0.93	0.89	0.96	0.94	0.93	0.85
PB1	0.94	0.98	0.96	0.96	0.92	0.92	0.96	0.94	0.94	0.88
PB2	0.95	0.98	0.97	0.96	0.93	0.95	0.96	0.96	0.96	0.91
PB1-F2	0.99	0.99	0.99	0.99	0.96	0.92	0.99	0.98	0.96	0.94
PB1-N40	0.94	0.97	0.96	0.96	0.91	0.91	0.97	0.95	0.94	0.88
PA-N155	0.94	0.98	0.97	0.96	0.92	0.92	0.97	0.95	0.94	0.89
PA-N182	0.95	0.98	0.97	0.97	0.93	0.94	0.96	0.95	0.95	0.9
PA-X	0.95	0.98	0.97	0.96	0.92	0.86	0.98	0.95	0.92	0.86

Table 5: The performance of K-Nearest Neighbour based models developed using AAC features for training and validation datasets.

	TRAINING					**VALIDATION**				
Proteins	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**MCC**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**MCC**
PA	0.98	0.99	0.97	0.98	0.95	0.95	0.91	0.97	0.94	0.88
HA	0.99	0.99	0.98	0.99	0.98	0.97	0.96	0.98	0.97	0.94
M1	0.93	0.98	0.92	0.95	0.84	0.87	0.71	0.94	0.82	0.68
M2	0.96	0.99	0.95	0.97	0.9	0.92	0.8	0.97	0.88	0.8
NA	0.98	0.99	0.98	0.99	0.97	0.96	0.94	0.97	0.96	0.92
NP	0.98	0.99	0.97	0.98	0.94	0.96	0.89	0.98	0.94	0.89
NS1	0.97	0.99	0.96	0.98	0.94	0.95	0.9	0.98	0.94	0.89
NS2	0.97	0.99	0.96	0.98	0.91	0.94	0.84	0.97	0.91	0.83
PB1	0.97	0.99	0.97	0.98	0.94	0.94	0.88	0.97	0.93	0.87
PB2	0.98	0.99	0.97	0.98	0.95	0.95	0.9	0.98	0.94	0.89

PB1-F2	0.99	0.99	0.99	0.99	0.97	0.99	0.94	0.99	0.97	0.96
PB1-N40	0.97	0.99	0.96	0.98	0.94	0.94	0.89	0.97	0.93	0.87
PA-N155	0.97	0.99	0.96	0.98	0.93	0.93	0.85	0.97	0.91	0.84
PA-N182	0.97	0.99	0.97	0.98	0.94	0.95	0.89	0.97	0.93	0.88
PA-X	0.97	0.99	0.97	0.98	0.92	0.95	0.86	0.98	0.92	0.86

Dipeptide Composition based Models

Among the composition based models, we developed several models derived from dipeptide composition of the peptides, where the random forest classifier performed better than the rest of the models using the same features. Both training and validation datasets gave the most accurate results to the HA protein random forest model, which achieved a 97.7% accuracy. On the validation dataset, the NP-based random forest model has an accuracy of 98.4% and an AUC of 0.99. The complete results are shown in Tables 6,7 and 8.

Table 6: The performance of Support Vector Machine based models developed using DPC features for training and validation datasets.

	TRAINING					**VALIDATION**				
Proteins	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**MCC**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**MCC**
PA	0.95	0.9	0.98	0.94	0.89	0.95	0.97	0.94	0.96	0.9
HA	0.97	0.95	0.99	0.97	0.94	0.97	0.98	0.97	0.97	0.94
M1	0.92	0.96	0.91	0.94	0.82	0.92	0.79	0.97	0.88	0.8
M2	0.94	0.95	0.94	*0.95*	*0.85*	0.94	0.82	0.99	0.9	0.85
NA	0.97	0.95	0.98	0.97	0.94	0.97	0.98	0.97	0.97	0.94
NP	0.97	0.94	0.98	0.96	0.92	0.97	0.96	0.97	0.97	0.93
NS1	0.96	0.92	0.98	0.95	0.91	0.95	0.97	0.95	0.96	0.9
NS2	0.92	0.98	0.91	0.95	0.77	0.92	0.75	0.98	0.87	0.79
PB1	0.96	0.92	0.98	0.95	0.9	0.95	0.92	0.96	0.94	0.88

PB2	0.96	0.92	0.98	0.95	0.91	0.96	0.97	0.95	0.96	0.91
PB1-F2	0.99	0.99	0.99	0.99	0.68	0.99	0.99	0.98	0.72	0.66
PB1-N40	0.92	0.95	0.95	0.95	0.9	0.93	0.95	0.95	0.95	0.89
PA-N155	0.86	0.94	0.94	0.92	0.86	0.95	0.94	0.94	0.94	0.87
PA-N182	0.88	0.94	0.94	0.93	0.87	0.96	0.94	0.94	0.95	0.88
PA-X	0.94	0.95	0.95	0.95	0.86	0.87	0.96	0.96	0.93	0.88

Table 7: The performance of Random Forest-based models developed using DPC features for training and validation datasets.

	TRAINING					**VALIDATION**				
Proteins	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**	**MCC**	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**	**MCC**
PA	0.96	0.99	0.97	0.97	0.94	0.97	0.97	0.97	0.97	0.93
HA	0.96	0.99	0.98	0.98	0.95	0.98	0.97	0.98	0.98	0.95
M1	0.96	0.97	0.96	0.96	0.91	0.82	0.97	0.92	0.89	0.81
M2	0.95	0.97	0.97	0.96	0.92	0.89	0.98	0.95	0.93	0.88
NA	0.96	0.99	0.98	0.98	0.96	0.98	0.97	0.97	0.97	0.94
NP	0.98	0.98	0.98	0.98	0.96	0.96	0.99	0.98	0.98	0.96
NS1	0.96	0.98	0.98	0.97	0.95	0.96	0.97	0.96	0.96	0.92
NS2	0.99	0.99	0.99	0.99	0.99	0.9	0.96	0.91	0.86	0.76
PB1	0.96	0.99	0.98	0.97	0.95	0.94	0.97	0.96	0.96	0.91
PB2	0.96	0.99	0.98	0.98	0.96	0.97	0.97	0.97	0.97	0.93
PB1-F2	0.99	0.99	0.99	0.99	0.99	0.87	0.99	0.99	0.83	0.75
PB1-N40	0.96	0.98	0.98	0.97	0.94	0.95	0.97	0.97	0.96	0.92
PA-N155	0.94	0.99	0.97	0.96	0.93	0.95	0.97	0.97	0.96	0.92
PA-N182	0.95	0.98	0.98	0.97	0.94	0.95	0.97	0.97	0.96	0.92
PA-X	0.95	0.98	0.98	0.97	0.93	0.94	0.98	0.97	0.96	0.92

Table 8: The performance of K-Nearest Neighbour based models developed using DPC features for training and validation datasets.

	TRAINING					VALIDATION				
Proteins	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**MCC**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**MCC**
PA	0.99	0.99	0.98	0.99	0.97	0.96	0.95	0.97	0.96	0.92
HA	0.99	0.98	0.99	0.99	0.98	0.97	0.96	0.98	0.97	0.94
M1	0.96	0.91	0.95	0.97	0.91	0.91	0.79	0.96	0.88	0.78
M2	0.97	0.92	0.95	0.98	0.91	0.93	0.8	0.98	0.89	0.82
NA	0.99	0.99	0.98	0.99	0.98	0.97	0.95	0.98	0.96	0.93
NP	0.99	0.96	0.98	0.99	0.97	0.98	0.94	0.99	0.97	0.94
NS1	0.98	0.96	0.97	0.99	0.96	0.95	0.91	0.97	0.94	0.88
NS2	0.98	0.91	0.97	0.99	0.94	0.95	0.85	0.98	0.92	0.86
PB1	0.99	0.99	0.99	0.99	0.98	0.96	0.92	0.98	0.95	0.91
PB2	0.99	0.99	0.98	0.99	0.98	0.96	0.95	0.97	0.96	0.92
PB1-F2	0.99	0.98	0.99	0.99	0.74	0.99	0.67	0.99	0.83	0.81
PB1-N40	0.99	0.99	0.98	0.99	0.97	0.96	0.93	0.98	0.95	0.92
PA-N155	0.99	0.99	0.98	0.99	0.97	0.96	0.91	0.98	0.94	0.9
PA-N182	0.99	0.99	0.98	0.99	0.97	0.97	0.94	0.98	0.96	0.92
PA-X	0.98	0.98	0.98	0.99	0.96	0.96	0.89	0.98	0.93	0.88

One Hot Encoding-based Models

We have also used binary or one-hot encoding features for the classification of human and non-human host sequences on being infectious to humans or not. We observed that the HA protein-based random forests model achieved 99.5 percent accuracy and 98.3 percent AUC, respectively, in the training and

validation data analyses. The data of validation support this by having a MCC of 0.9666 as shown in Table 9. The complete results are provided in the following tables (Tables 9, 10 and 11).

Table 9: The performance of Support Vector Machine-based models developed using one hot encoding feature for training and validation datasets.

	TRAINING					VALIDATION				
Proteins	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**MCC**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**MCC**
PA	0.95	0.9	0.98	0.96	0.89	0.95	0.97	0.94	0.95	0.89
HA	0.96	0.94	0.97	0.96	0.92	0.96	0.95	0.96	0.96	0.92
M1	0.92	0.95	0.91	0.93	0.8	0.92	0.79	0.98	0.88	0.81
M2	0.94	0.96	0.94	0.95	0.86	0.93	0.8	0.99	0.89	0.84
NA	0.97	0.94	0.98	0.96	0.93	0.97	0.97	0.96	0.97	0.93
NP	0.97	0.93	0.98	0.96	0.92	0.96	0.95	0.97	0.96	0.92
NS1	0.97	0.93	0.98	0.96	0.92	0.96	0.95	0.97	0.96	0.92
NS2	0.95	0.87	0.97	0.92	0.86	0.94	0.91	0.95	0.93	0.85
PB1	0.95	0.91	0.97	0.94	0.89	0.94	0.92	0.95	0.94	0.87
PB2	0.95	0.91	0.98	0.95	0.9	0.95	0.96	0.95	0.96	0.9
PB1-F2	0.99	0.98	0.99	0.99	0.95	0.92	0.98	0.98	0.96	0.95
PB1-N40	0.92	0.95	0.95	0.94	0.88	0.94	0.95	0.95	0.95	0.89
PA-N155	0.84	0.93	0.93	0.91	0.84	0.94	0.93	0.93	0.93	0.85
PA-N182	0.81	0.89	0.89	0.87	0.76	0.87	0.89	0.9	0.89	0.76
PA-X	0.94	0.93	0.94	0.94	0.81	0.77	0.94	0.93	0.88	0.81

Table 10: The performance of Random Forest-based models developed using one hot encoding feature for training and validation datasets.

	TRAINING					VALIDATION				
Proteins	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**	**MCC**	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**	**MCC**

PA	0.97	0.99	0.98	0.98	0.96	0.96	0.97	0.97	0.97	0.93
HA	0.99	0.99	0.99	0.99	0.98	0.98	0.99	0.98	0.98	0.97
M1	0.99	0.98	0.98	0.98	0.96	0.89	0.94	0.92	0.91	0.82
M2	0.99	0.99	0.99	0.99	0.97	0.85	0.98	0.94	0.91	0.86
NA	0.97	0.99	0.98	0.98	0.96	0.97	0.98	0.97	0.97	0.94
NP	0.99	0.98	0.99	0.99	0.97	0.93	0.98	0.97	0.96	0.92
NS1	0.99	0.99	0.99	0.99	0.97	0.93	0.98	0.97	0.96	0.92
NS2	0.98	0.98	0.98	0.98	0.95	0.88	0.97	0.94	0.92	0.85
PB1	0.97	0.99	0.98	0.98	0.96	0.94	0.97	0.96	0.96	0.91
PB2	0.98	0.99	0.98	0.98	0.96	0.95	0.98	0.97	0.97	0.93
PB1-F2	0.98	0.98	0.99	0.99	0.96	0.9	0.99	0.98	0.95	0.93
PB1-N40	0.97	0.98	0.98	0.97	0.95	0.95	0.97	0.96	0.96	0.91
PA-N155	0.97	0.99	0.98	0.98	0.96	0.93	0.98	0.96	0.95	0.91
PA-N182	0.82	0.94	0.9	0.88	0.77	0.85	0.91	0.89	0.88	0.75
PA-X	0.98	0.98	0.98	0.98	0.94	0.85	0.98	0.95	0.92	0.86

Table 11: The performance of K-Nearest Neighbour-based models developed using one hot encoding feature for training and validation datasets.

	TRAINING					**VALIDATION**				
Proteins	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**MCC**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**MCC**
PA	0.98	0.98	0.97	0.99	0.96	0.95	0.91	0.97	0.94	0.89
HA	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.98	0.97
M1	0.96	0.92	0.94	0.97	0.9	0.89	0.79	0.94	0.87	0.74
M2	0.97	0.9	0.96	0.98	0.92	0.92	0.79	0.97	0.88	0.8
NA	0.99	0.99	0.98	0.99	0.97	0.97	0.94	0.98	0.96	0.93
NP	0.98	0.98	0.98	0.99	0.96	0.96	0.92	0.98	0.95	0.91

NS1	0.98	0.98	0.98	0.99	0.96	0.96	0.92	0.98	0.95	0.91
NS2	0.97	0.91	0.97	0.98	0.93	0.94	0.85	0.98	0.91	0.84
PB1	0.98	0.97	0.97	0.99	0.95	0.96	0.91	0.98	0.94	0.9
PB2	0.98	0.93	0.97	0.99	0.96	0.96	0.92	0.98	0.95	0.9
PB1-F2	0.99	0.99	0.99	0.99	0.97	0.98	0.94	0.99	0.96	0.94
PB1-N40	0.98	0.97	0.97	0.98	0.95	0.95	0.9	0.97	0.94	0.88
PA-N155	0.98	0.96	0.97	0.98	0.94	0.95	0.88	0.98	0.93	0.87
PA-N182	0.86	0.87	0.87	0.87	0.68	0.89	0.84	0.91	0.88	0.75
PA-X	0.97	0.94	0.97	0.98	0.93	0.94	0.81	0.98	0.9	0.83

In our analysis, although HA and NA are often considered to be associated with host tropism by causing novel variants and subtypes, they are not the sole contributors to it. The other proteins also contribute to the host tropism, therefore, there is the scope for further research on sequence analysis or molecular studies on them. Furthermore, we see that proteins such as NS1 and NS2 or M1 and M2 also play a role. These proteins are encoded from the same segment, but their predictions differ. Despite sharing the same segments, they are functionally different even though they are encoded in the same way. As a result of this, each of the segments can be affected separately if a mutation occurs in one of them [56].

Best Model Selection

We chose not to use the random forest model using the encoding feature because the models were expensive to compute and time-consuming, even though following dimensionality reduction. The AAC and DPC models, on the other hand, permitted efficient execution of the models and as a result, we chose the DPC models for inclusion in the web-server since it achieved higher evaluation parameters than the AAC and generated features faster than the one-hot encoding of the sequences. As far as NS2 and PB1F2 are concerned, we chose the random forest model that uses AAC features because they showed better results than the DPC models.

Chapter 4: The FluSPred Web Server

Architecture of the Web-Server

For the benefit of the scientific research community, A webserver named "FluSPred" (https://webs.iiitd.edu.in/raghava/fluspred/) and a standalone package (https://github.com/raghavagps/FluSPred) has been made where, the best models for each of the 15 proteins was selected and incorporated. HTML, CSS were used for front-end structure of the web server and styling respectively. VanillaJS was used for client-side logic. PHP and Python were used for running the prediction scripts on the server and handling the backend tasks of the web-server. The webserver supports all kinds of device screen sizes such as mobile, tablets and laptops, i.e. the webserver is fully responsive.

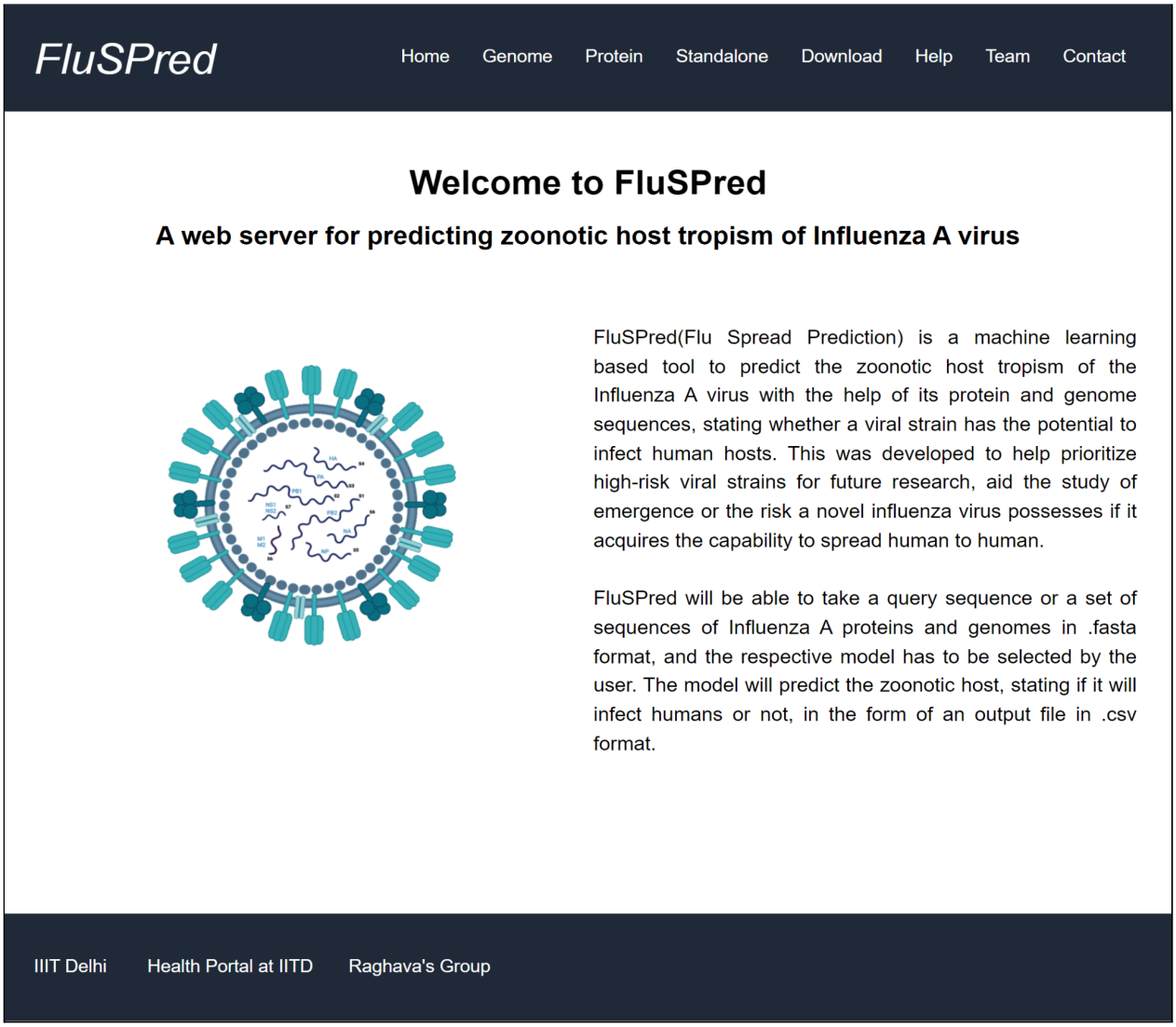


Figure 18: The FluSPred Web Server

Working of FluSpred

The FluSPred webserver has two modules, a) Genome and b) Protein. The module for Genome consists of one prediction model which predicts the host tropism of Influenza A using genome sequences. The Protein module of the web server encompasses prediction models of each of the 15 proteins. The best models for each protein selected were incorporated into this module.

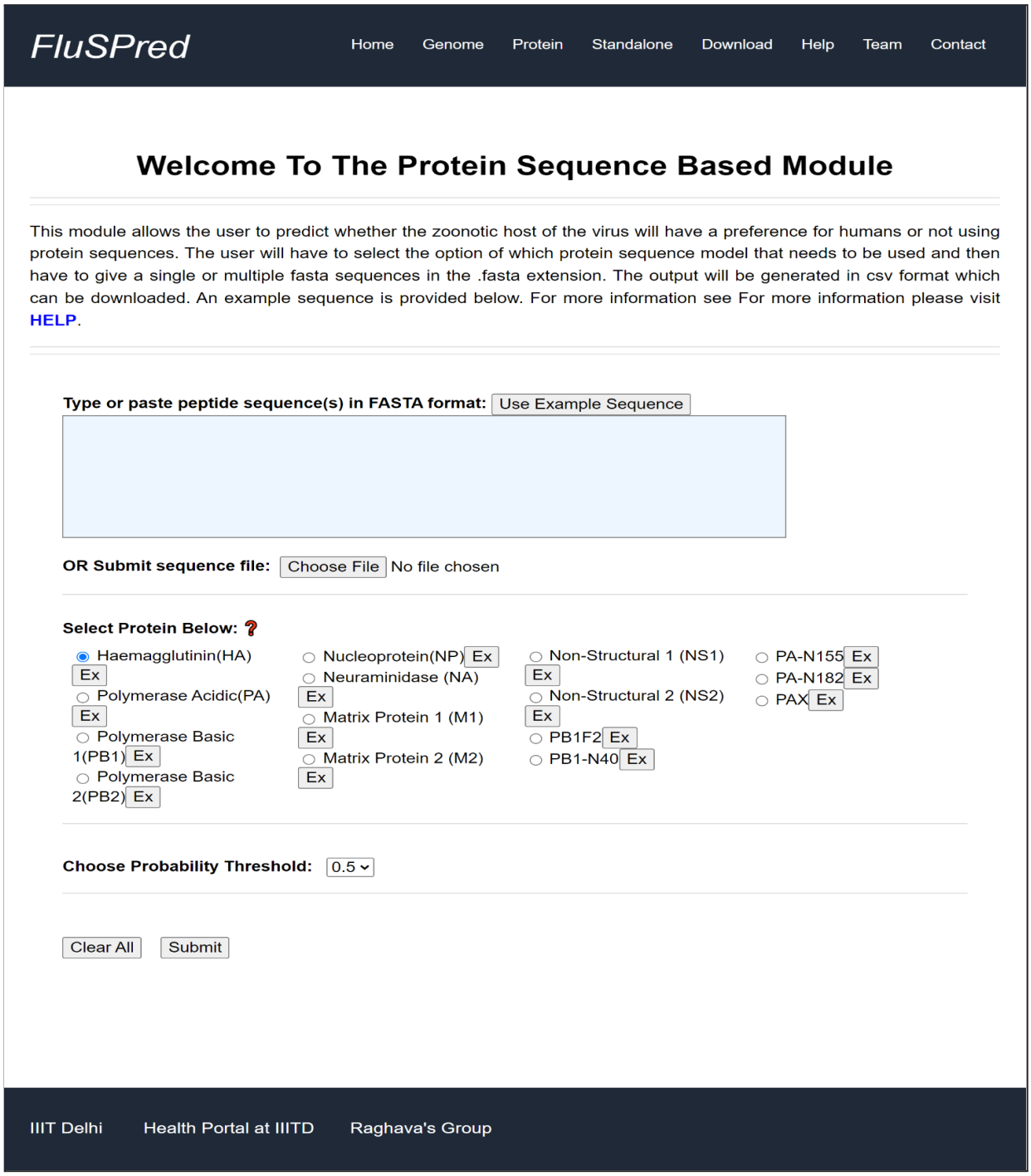


Figure 19: Protein Module of FluSPred

In the Protein module of the webserver, selecting either of the 15 models, a single or multiple sequences, can be used as an input in .fasta format. The user has to select the model that they want to run and their sequences needs to be provided in the input box. For more clarity, example sequences for each of the proteins have been added in the "Ex" buttons beside the radio buttons of their respective models.

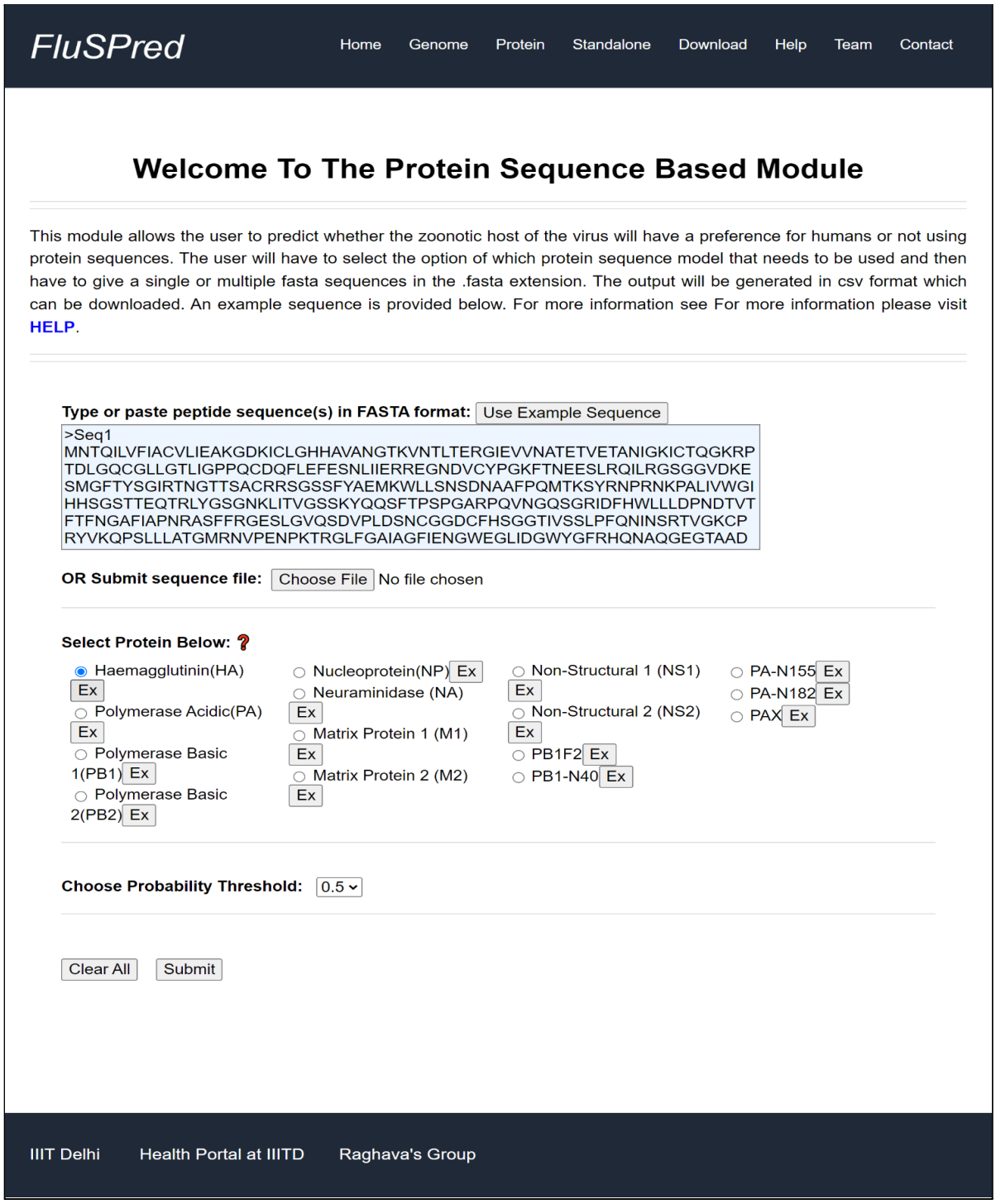
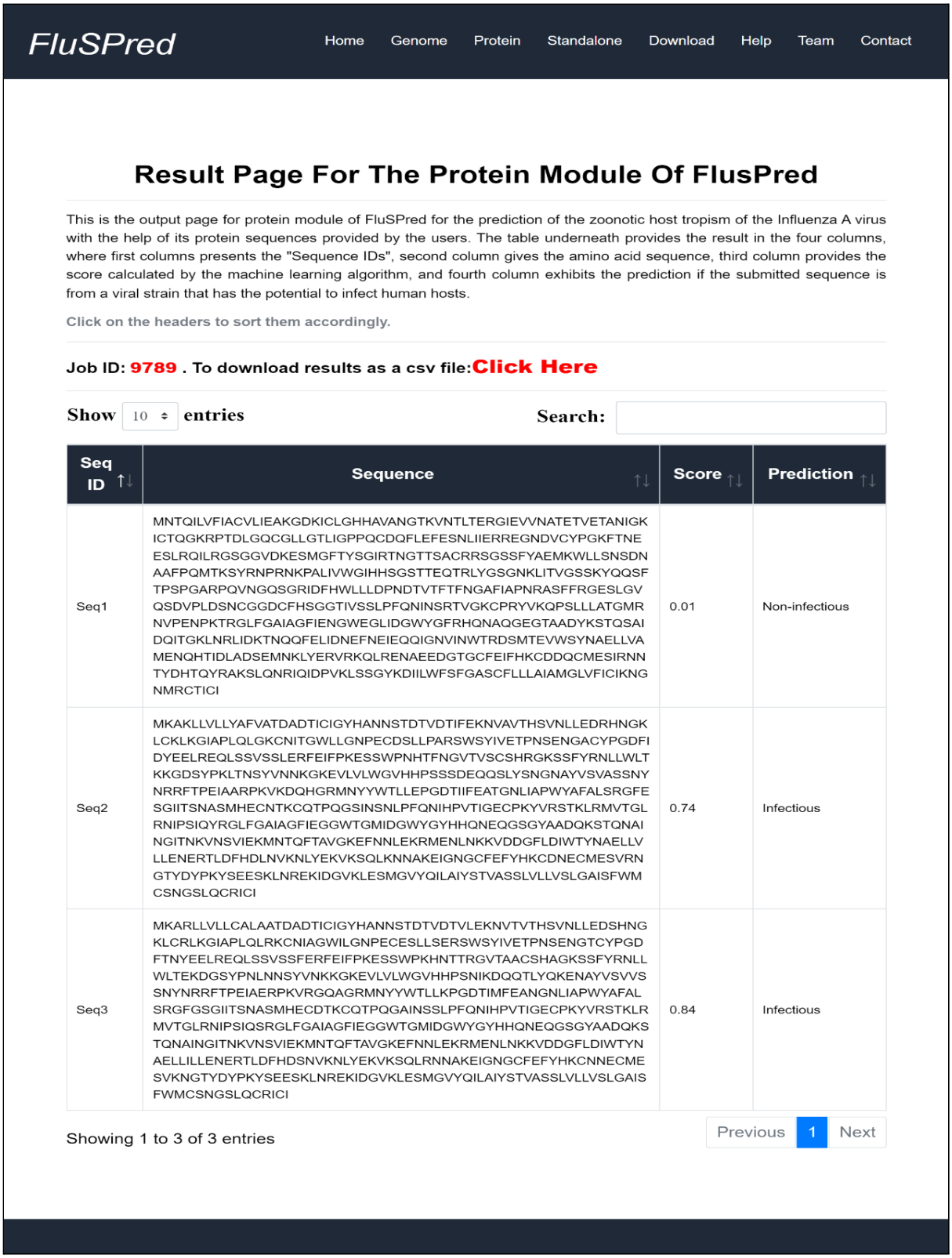


Figure 20: HA Model Selected and Example Sequence of the same Added in the Input Box

When a particular model is selected by the user and clicked on the “Submit” button, FluSPred then runs the respective model and provides the result in the form of a table which can be downloaded in .csv format. The result comprises the prediction of the human/non-human infectious strain.



FluSPred

Home Genome Protein Standalone Download Help Team Contact

Result Page For The Protein Module Of FlusPred

This is the output page for protein module of FluSPred for the prediction of the zoonotic host tropism of the Influenza A virus with the help of its protein sequences provided by the users. The table underneath provides the result in the four columns, where first columns presents the "Sequence IDs", second column gives the amino acid sequence, third column provides the score calculated by the machine learning algorithm, and fourth column exhibits the prediction if the submitted sequence is from a viral strain that has the potential to infect human hosts.

Click on the headers to sort them accordingly.

Job ID: 9789 . To download results as a csv file:Click Here

Show 10 entries Search:

Seq ID	Sequence	Score	Prediction
Seq1	MNTQILVFIACVLIEAKGDKICLGHHAVANGTKVNTLTERGIEVVNATETVETANIGK ICTQGKRPTDLGQCGLLGTLIGPPQCDQFLEFESNLIIERREGNDVCYPGKFTNE ESLRQILRGSGGVDKESMGFTYSGIRTNGTTSACRRSGSSFYAEMKWLLSNSDN AAFPQMTKSYRNPRNKPALIVWGIHHSGSTTEQTRLYGSGNKLITVGSSKYQQSF TPSPGARPQVNGQSGRIDFHWLLLDPNDTVTFTFNGAFIAPNRASFFRGESLGV QSDVPLDSNCGGDCFHSGGTIVSSLPFQNINSRTVGKCPRYVKQPSLLLATGMR NVPENPKTRGLFGAIAGFIENGWEGLIDGWYGFRHQNAQGEGTAADYKSTQSAI DQITGKLNRLIDKTNQQFELIDNEFNEIEQQIGNVINWTRDSMTEVWSYNAELLVA MENQHTIDLADSEMNKLYERVRKQLRENAEEDGTGCFEIFHKCDDQCMESIRNN TYDHTQYRAKSLQNRIQIDPVKLSSGYKDIILWFSFGASCFLLLAIAMGLVFICIKNG NMRCTICI	0.01	Non-infectious
Seq2	MKAKLLVLLYAFVATDADTICIGYHANNSTDTVDTIFEKNVAVTHSVNLLEDRHNGK LCKLKGIAPLQLGKCNITGWLLGNPECDSLLPARSWSYIVETPNSENGACYPGDFI DYEELREQLSSVSSLERFEIFPKESSWPNHTFNGVTVSCSHRGKSSFYRNLLWLT KKGDSYPKLTNSYVNNKGKEVLVLWGVHHPSSSDEQQSLYSNGNAYVSVASSNY NRRFTPEIAARPKVKDQHGRMNYYWTLLEPGDTIIFEATGNLIAPWYAFALSRGFE SGIITSNASMHECNTKCQTPQGSINSNLPFQNIHPVTIGECPKYVRSTKLRMVTGL RNIPSIQYRGLFGAIAGFIEGGWTGMIDGWYGYHHQNEQGSGYAADQKSTQNAI NGITNKVNSVIEKMNTQFTAVGKEFNNLEKRMENLNKKVDDGFLDIWTYNAELLV LLENERTLDFHDLNVKNLYEKVKSQLKNNAKEIGNGCFEFYHKCDNECMESVRN GTYDYPKYSEESKLNREKIDGVKLESMGVYQILAIYSTVASSLVLLVSLGAISFWM CSNGSLQCRICI	0.74	Infectious
Seq3	MKARLLVLLCALAATDADTICIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDSHNG KLCRLKGIAPLQLRKCNIAGWILGNPECESLLSERSWSYIVETPNSENGTCYPGD FTNYEELREQLSSVSSFERFEIFPKESSWPKHNTTRGVTAACSHAGKSSFYRNLL WLTEKDGSYPNLNNSYVNKKGKEVLVLWGVHHPSNIKDQQTLYQKENAYVSVVS SNYNRRFTPEIAERPKVRGQAGRMNYYWTLLKPGDTIMFEANGNLIAPWYAFAL SRGFGSGIITSNASMHECDTKCQTPQGAINSSLPFQNIHPVTIGECPKYVRSTKLR MVTGLRNIPSIQSRGLFGAIAGFIEGGWTGMIDGWYGYHHQNEQGSGYAADQKS TQNAINGITNKVNSVIEKMNTQFTAVGKEFNNLEKRMENLNKKVDDGFLDIWTYN AELLILLENERTLDFHDSNVKNLYEKVKSQLRNNAKEIGNGCFEFYHKCNNECME SVKNGTYDYPKYSEESKLNREKIDGVKLESMGVYQILAIYSTVASSLVLLVSLGAIS FWMCSNGSLQCRICI	0.84	Infectious

Showing 1 to 3 of 3 entries Previous 1 Next

Figure 21: Result Page of HA Model using the Example Sequences

The long-term objective of this server is to use it for research into Influenza A or public health and pandemic surveillance.

Standalone Package

For the benefit of the scientific community, FluSPred was made freely accessible or open-source. Along with the web-server, a standalone package was also developed. The standalone program needs three arguments to run for the Protein module. First, the user must provide an input file in .fasta format, which needs to be predicted. For proteins, the user must give the argument as 'P'. The second argument asks for the type of sequences the user provided in the input file. Third, the user must specify the protein name that the user's sequences belong to, which must be one of fifteen proteins listed below:

```
$ python main.py -h
loading...

usage: main.py [-h] -i INPUT -o OPTION [-pn PROTEINNAME]

Please provide following arguments to proceed

optional arguments:
  -h, --help            show this help message and exit
  -i INPUT, --input INPUT
                        Input File Name: protein or genome sequence in
FASTA format
  -o OPTION, --option OPTION
                        Select which kind of file you are giving,
Protein(P) or Genome(G)

                        P : Protein
                        G : Genome

  -pn PROTEINNAME, --proteinName PROTEINNAME
                        This argument is only required when choosing
OPTION as protein
                        enter the Protein name from 15 proteins listed
below
                        HA :  Haemagglutinin
                        PA :  Polymerase Acidic
                        PB1 :  Polymerase Basic 1
```

```
PB2 :  Polymerase Basic 2
NP :  Nucleoprotein
NA :  Neuraminidase
M1 :  Matrix Protein 1
M2 :  Matrix Protein 2
NS1 : Non-Structural 1
NS2 : Non-Structural 2
PB1F2 : PB1F2
PB1N40 : PB1-N40
PAN155 : PA-N155
PAN182 : PA-N182
PAX : PAX
```

Case Study: Evaluation of FluSPred

We conducted a study to further validate our web-server using newly retrieved sequence entries retrieved from IRD on 24.12.2021. Additional entries were added for HA, NP, and PA-X proteins. Redundancy was thoroughly examined for the new sequences, and those that were not in the original dataset that was used in the models were selected. The following strains were used, as demonstrated by the HA protein analysis: A/swine/Iowa/A02636065/2021, A/teal/Samara/Bolshechernigovsky/2021, A/Texas/01/2021. The HA prediction model was able to analyze these sequences and predict if the strains were infectious or not by using probability scores. For the non-infectious sequences, the scores were very low (0.02 for A/swine/Iowa/A02636065/2021 and 0.07 for A/teal/Samara/Bolshechernigovsky/2021 strains) compared with 0.8 for infectious sequences. NP prediction model was validated using the sequence of the NP protein from A/Texas/01/2021 strain with H3N2 subtype. The strain scored 0.57, indicating that it was infectious. Moreover, the strain carrying the PA-X protein sequence was also used to test the respective protein model. With a score of 0.98, the model also indicated that the strain was infectious.

BLAST Analysis

An analysis based on BLAST or Basic Local Alignment Search Tool[57] was also conducted for further evaluation of FluSPred. The dataset for few proteins (HA, NA, PB1, PB2) in .fasta format was taken and divided into train and test files. The command makeblastdb was run to make a local library for reference sequence. Then blastp was run where the input was given as test file and output with hits/nohits was executed. With the results received from BLAST, evaluation metrics such as accuracy, sensitivity, specificity, MCC and AUC were calculated. The comparison of the evaluation metrics of both FluSPred and Blast is given in the table(Table 12) below.

Table 12: Comparison of FluSpred And BLAST Results

Method	Protein	Sensitivity	Specificity	Accuracy	AUC	MCC
Blast	HA	0.98	0.66	0.89	0.82	0.73
FluSPred	HA	0.98	0.97	0.97	0.97	0.94
Blast	NA	0.96	0.91	0.93	0.93	0.87
FluSPred	NA	0.97	0.97	0.97	0.97	0.94
Blast	PB1	0.97	0.64	0.67	0.8	0.35
FluSPred	PB1	0.92	0.96	0.94	0.94	0.88
Blast	PB2	0.96	0.46	0.62	0.71	0.42
FluSPred	PB2	0.95	0.96	0.96	0.96	0.91

Chapter 5: Discussion

Approximately 2.4 billion cases illnesses pertaining to microorganisms and 2.7 million deaths are associated with zoonotic diseases, including infections from novel agents (such as bacteria, fungi, viruses, protozoa and pathogens) each year [58]. With increased human activity or interference, the frequency of zoonotic diseases in humans is increasing[59],[60]. It is imperative that we identify strains that have a high risk of causing diseases to emerge and that can motivate host acceptance of zoonotic infections. In analyzing disease outbreaks, forecasting tools and early warning systems can now be developed through advancements as well as improvements in technology and high-throughput sequencing[61]. It has been attempted by scientists worldwide in the last few years to develop computational methods to predict the progression of zoonotic diseases such as influenza, SARS, MERS, Ebola, and rabies [62],[63, 64],[65].In humans, the frequency of zoonotic illnesses is rising as human activity increases or human interference increases [59],[60]. The need of the hour is to identify strains with this capability of causing disease emergence and if this can promote host permissiveness for the spread of zoonotic infections [61].

Using computational methods, we developed models of the proteins and genome in order to predict the zoonotic hosts of novel influenza A viruses. Sequence data was obtained from the Influenza Research Database. The feature extraction and encoding were performed using compositional based feature extraction. To train and validate the model, we selected the relevant features. Our compositional analysis indicates which residues were preferred over the others in sequences from human hosts, for each protein. The results show that simple composition based techniques can identify the important features and aid in prediction. Of the hot encoding features, the HA protein achieved the highest AUC of 0.991 in training datasets and 0.982 in validation datasets, using random forest models. Analyses of the sequence data using MERCI identified important motifs that were present on the human hosts but not on the non-human sequences. It gives us an overview of the motifs that can play a role in zoonotic transmission. As they were not as computationally expensive as one-hot encoding features, Random Forest models of AAC and DPC for the corresponding proteins were selected as the best models. FluSPred, an open-source web-server that incorporates the best models of each of the 15 proteins, was developed.

There have been several attempts by scientists all over the world to make computational models for the prediction of zoonotic events of different pathogens. Most of the available tools are complex and do not

have a web-server which can be used by the community. There was less focus on the role of proteins in zoonotic transmission of Influenza A, and where used, the number of proteins considered were limited, with a maximum consideration of 11 proteins [66]. The prediction models made by Wang *et. al*[36] used only 6 influenza A proteins, which did not include HA and NA. Although historically we have seen HA and NA being responsible for the Influenza A subtypes for the four pandemics. Also, both HA and NA showed high outcomes on the computational models, which plainly demonstrate their significance in zoonosis. Additionally, our models require protein or genome sequences of the virus from the respective hosts. It does not require the sequences of the hosts, which can be cumbersome since their limited availability[67, 68].[69]Zhang et al.[70] and Galiez et al[71] combined species and genera to higher taxonomic groups whereas our approach takes account of the host species for each strains of the virus. *Mock et al*[40]performed deep learning, which is computer expensive, on multiclass classification of genomic data that was imbalanced, having preference on certain species over the others. They have achieved an average accuracy of 97.46 and an AUC of 0.94 on the influenza A dataset. On the other hand, we have performed binary classification with a focus on human and non-human hosts, achieving a higher prediction accuracy and AUC of 98% by using simple composition-based features, which are not computer expensive and time consuming. Li and Sun[72]used SVM, alignment-based and without alignment based methods to predict the host of influenza A genomic data. Their dataset was small (1200 sequences and 6 hosts), and the average accuracy was 84%, 85.67%, and 87% for alignment based, SVM and alignment free methods respectively [72], [40]. Our models incorporated a wide scope, i.e. 308632 of sequence data, pertaining to 34 hosts as well as accomplished a lot higher accuracy.

With our methodical approach, we have developed a computational tool that predicts human infectious strains of the influenza A virus based on 15 proteins. From sequence datasets provided by IRD, we compute compositional-based descriptors/features and one hot encoding-based feature using Pfeature. These features were used to train and validate the model. Using compositional analysis, we identify which residues (A, I, K, E, H, G, P, Q) are most frequently observed in three major zoonotic proteins such as HA, NA and PB1-F2. It is demonstrated that simple composition-based techniques can be used to identify and predict important features. An analysis of the human sequences for motifs showed that there were several important motifs not present in the non-human sequences.

Our webserver, FluSPred, is user-friendly and the first web server to incorporate all the 15 influenza A protein prediction models all at one place. It is a machine-learning-based tool that has been trained by

composition-based features(AAC, DPC), having models on each of the 15 proteins. The composition-based feature extraction is simple and much faster compared to the rest of the methods provided by others/available, and yet highly accurate, as described in the results. The simplicity in the webserver lies with the fact that the user would only have to provide the sequence of either of the proteins, and the models will be able to predict whether the sequence pertaining to the virus is infectious to human hosts. The data on which our models were trained covers a huge span of time and a vast range of influenza causing viral proteins and genomes from diverse hosts. BLAST Analysis was also conducted on few proteins which in literature is know to be of great importance for host tropism. The results of our models surpass the BLAST results indicating that our models can predict the infectious strains at a higher accuracy as compared to BLAST. The purpose of this web server is to serve the scientific community for predicting the zoonotic risk of the virus as a part of the early warning system. To the best of our knowledge, this is the first attempt to develop a web server that has computational models for all the 15 Influenza A proteins and genome at one place.

Conclusion

With a history of four pandemics in the past as well as recurring seasonal influenza every year, the need for international measures depend heavily on the investigations, analyses and researches carried out by public health centres and other infection sites to trace the outbreak. One of the critical factors and primary research must involve around finding out the origin or the root cause of the infectious disease emergence. There are traces of evidence or proof that cross-species transmission or zoonotic spillovers from animals to humans take place which can eventually lead to outbreaks. Serological surveys of the animals prone to zoonosis living in proximity to humans and places where human and animals can come in close contact is essential to prevent likely spillovers.

A robust disease surveillance system is of utmost importance for prompt detection of zoonotic spillovers. This surveillance also includes detection of infectious agents while crossing the species barrier before it starts to circulate among human populations. This would help arrest a possible outbreak, be it epidemics or pandemics.

Here, "One Health" approach tends to play a very important role as this vision involves public health, veterinary health, epidimeological knowledge and medical science for investigating the risks, predisposition and mitigating any likely outbreaks. It can also help with better prevention and control strategies. Relying completely on public health measures is not enough. Hence there should be efforts to arrest emmerging zoonotic events at all levels. Our web server FluSPred, even if it plays a small part, would be of great help to the communities and people in curbing such events. However, there is still scope of more research in this area.

Future Objectives

There is scope of further research where insights can be driven on cross-species transmission which is still unknown. Figuring out which exact position mutations are taking place which enables cross species transmission would be a good way to start. Our models help distinguish between humans and non humans, and which reservoir the strains belong to. However, further research can be carried out which helps determine animal to animal transmissions, what triggers them and their effects. After all, adequate research, management and control of emerging zoonoses will offer a possibility for containing health risks of zoonotic infections that are of global health concern and make the world safer from emerging and re-emerging pathogens.

References

1. Taubenberger, J.K. and D.M. Morens, *The pathology of influenza virus infections.* Annu Rev Pathol, 2008. **3**: p. 499-522.
2. Bouvier, N.M. and P. Palese, *The biology of influenza viruses.* Vaccine, 2008. **26 Suppl 4**: p. D49-53.
3. Webster, R.G., *Influenza: an emerging disease.* Emerg Infect Dis, 1998. **4**(3): p. 436-41.
4. Gaitonde, D.Y., F.C. Moore, and M.K. Morgan, *Influenza: Diagnosis and Treatment.* Am Fam Physician, 2019. **100**(12): p. 751-758.
5. Sellers, S.A., et al., *The hidden burden of influenza: A review of the extra-pulmonary complications of influenza infection.* Influenza Other Respir Viruses, 2017. **11**(5): p. 372-393.
6. Fleming, D.M., *The contribution of influenza to combined acute respiratory infections, hospital admissions, and deaths in winter.* Commun Dis Public Health, 2000. **3**(1): p. 32-8.
7. Chow, E.J., J.D. Doyle, and T.M. Uyeki, *Influenza virus-related critical illness: prevention, diagnosis, treatment.* Crit Care, 2019. **23**(1): p. 214.
8. Lyytikainen, O., et al., *Influenza A outbreak among adolescents in a ski hostel.* Eur J Clin Microbiol Infect Dis, 1998. **17**(2): p. 128-30.
9. Thomas, P., et al., *Fatal influenza A virus infection in a child vaccinated against influenza.* Pediatr Infect Dis J, 2003. **22**(2): p. 201-2.
10. Webster, R.G., *The importance of animal influenza for human disease.* Vaccine, 2002. **20 Suppl 2**: p. S16-20.
11. Eng, C.L.P., J.C. Tong, and T.W. Tan, *Predicting Zoonotic Risk of Influenza A Viruses from Host Tropism Protein Signature Using Random Forest.* Int J Mol Sci, 2017. **18**(6).
12. Long, J.S., C.T. Benfield, and W.S. Barclay, *One-way trip: influenza virus' adaptation to gallinaceous poultry may limit its pandemic potential.* Bioessays, 2015. **37**(2): p. 204-12.
13. Naeem, A., et al., *Antigenic drift of hemagglutinin and neuraminidase in seasonal H1N1 influenza viruses from Saudi Arabia in 2014 to 2015.* J Med Virol, 2020.
14. Clem, A. and S. Galwankar, *Seasonal influenza: waiting for the next pandemic.* J Glob Infect Dis, 2009. **1**(1): p. 51-6.
15. Nachbagauer, R., et al., *A chimeric hemagglutinin-based universal influenza virus vaccine approach induces broad and long-lasting immunity in a randomized, placebo-controlled phase I trial.* Nat Med, 2021. **27**(1): p. 106-114.

16. Rothberg, M.B. and S.D. Haessler, *Complications of seasonal and pandemic influenza.* Crit Care Med, 2010. **38**(4 Suppl): p. e91-7.
17. Yamada, S., et al., *Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors.* Nature, 2006. **444**(7117): p. 378-82.
18. Slaine, P.D., et al., *Adaptive Mutations in Influenza A/California/07/2009 Enhance Polymerase Activity and Infectious Virion Production.* Viruses, 2018. **10**(5).
19. Viboud, C., et al., *Global Mortality Impact of the 1957-1959 Influenza Pandemic.* J Infect Dis, 2016. **213**(5): p. 738-45.
20. Esposito, S., et al., *Impact of pandemic A/H1N1/2009 influenza on children and their families: comparison with seasonal A/H1N1 and A/H3N2 influenza viruses.* J Infect, 2011. **63**(4): p. 300-7.
21. Rubinson, L., et al., *Impact of the fall 2009 influenza A(H1N1)pdm09 pandemic on US hospitals.* Med Care, 2013. **51**(3): p. 259-65.
22. Wang, Y., C.Y. Tang, and X.F. Wan, *Antigenic characterization of influenza and SARS-CoV-2 viruses.* Anal Bioanal Chem, 2022. **414**(9): p. 2841-2881.
23. Song, M.S., et al., *The polymerase acidic protein gene of influenza a virus contributes to pathogenicity in a mouse model.* J Virol, 2009. **83**(23): p. 12325-35.
24. Biswas, S.K. and D.P. Nayak, *Influenza virus polymerase basic protein 1 interacts with influenza virus polymerase basic protein 2 at multiple sites.* J Virol, 1996. **70**(10): p. 6716-22.
25. Matrosovich, M.N., et al., *Avian influenza A viruses differ from human viruses by recognition of sialyloligosaccharides and gangliosides and by a higher conservation of the HA receptor-binding site.* Virology, 1997. **233**(1): p. 224-34.
26. Rogers, G.N. and J.C. Paulson, *Receptor determinants of human and animal influenza virus isolates: differences in receptor specificity of the H3 hemagglutinin based on species of origin.* Virology, 1983. **127**(2): p. 361-73.
27. Nachbagauer, R. and F. Krammer, *Universal influenza virus vaccines and therapeutic antibodies.* Clin Microbiol Infect, 2017. **23**(4): p. 222-228.
28. McAuley, J.L., et al., *Influenza Virus Neuraminidase Structure and Functions.* Front Microbiol, 2019. **10**: p. 39.
29. Xu, J., et al., *Evolutionary dynamics of influenza A nucleoprotein (NP) lineages revealed by large-scale sequence analyses.* Infect Genet Evol, 2011. **11**(8): p. 2125-32.
30. Furuse, Y., et al., *Evolution of the M gene of the influenza A virus in different host species: large-scale sequence analysis.* Virol J, 2009. **6**: p. 67.

31. Hao, W., L. Wang, and S. Li, *Roles of the Non-Structural Proteins of Influenza A Virus.* Pathogens, 2020. **9**(10).
32. Varga, Z.T. and P. Palese, *The influenza A virus protein PB1-F2: killing two birds with one stone?* Virulence, 2011. **2**(6): p. 542-6.
33. Wang, Q., et al., *Host cell interactome of PB1 N40 protein of H5N1 influenza A virus in chicken cells.* J Proteomics, 2019. **197**: p. 34-41.
34. Muramoto, Y., et al., *Identification of novel influenza A virus proteins translated from PA mRNA.* J Virol, 2013. **87**(5): p. 2455-62.
35. Qiang, X. and Z. Kou, *Predicting interspecies transmission of avian influenza virus based on wavelet packet decomposition.* Comput Biol Chem, 2019. **78**: p. 455-459.
36. Wang, J., et al., *Predicting transmission of avian influenza A viruses from avian to human by using informative physicochemical properties.* Int J Data Min Bioinform, 2013. **7**(2): p. 166-79.
37. Eng, C.L., J.C. Tong, and T.W. Tan, *Predicting host tropism of influenza A virus proteins using random forest.* BMC Med Genomics, 2014. **7 Suppl 3**: p. S1.
38. Li, J., et al., *Machine Learning Methods for Predicting Human-Adaptive Influenza A Viruses Based on Viral Nucleotide Compositions.* Mol Biol Evol, 2020. **37**(4): p. 1224-1236.
39. Tsukiyama, S., et al., *LSTM-PHV: prediction of human-virus protein-protein interactions by LSTM with word2vec.* Brief Bioinform, 2021. **22**(6).
40. Mock, F., et al., *VIDHOP, viral host prediction with deep learning.* Bioinformatics, 2021. **37**(3): p. 318-325.
41. Xu, B., et al., *Predicting the host of influenza viruses based on the word vector.* PeerJ, 2017. **5**: p. e3579.
42. Zhang, Y., et al., *Influenza Research Database: An integrated bioinformatics resource for influenza virus research.* Nucleic Acids Res, 2017. **45**(D1): p. D466-D474.
43. Pande, A., et al., *Computing wide range of protein/peptide features from their sequence and structure.* bioRxiv, 2019: p. 599126.
44. Kuzmin, K., et al., *Machine learning methods accurately predict host specificity of coronaviruses based on spike sequences alone.* Biochem Biophys Res Commun, 2020. **533**(3): p. 553-558.
45. Zhang, Z., *Introduction to machine learning: k-nearest neighbors.* Ann Transl Med, 2016. **4**(11): p. 218.
46. Rigatti, S.J., *Random Forest.* J Insur Med, 2017. **47**(1): p. 31-39.

47. Gautam, A., et al., *In silico approaches for designing highly effective cell penetrating peptides.* J Transl Med, 2013. **11**: p. 74.
48. Ben-Hur, A. and J. Weston, *A user's guide to support vector machines.* Methods Mol Biol, 2010. **609**: p. 223-39.
49. Bac, J., et al., *Scikit-Dimension: A Python Package for Intrinsic Dimension Estimation.* Entropy (Basel), 2021. **23**(10).
50. Agrawal, P., et al., *In Silico Approach for Prediction of Antifungal Peptides.* Front Microbiol, 2018. **9**: p. 323.
51. Nagpal, G., et al., *Computer-aided prediction of antigen presenting cell modulators for designing peptide-based vaccine adjuvants.* J Transl Med, 2018. **16**(1): p. 181.
52. Qureshi, A., N. Thakur, and M. Kumar, *VIRsiRNApred: a web server for predicting inhibition efficacy of siRNAs targeting human viruses.* J Transl Med, 2013. **11**: p. 305.
53. Patiyal, S., et al., *NAGbinder: An approach for identifying N-acetylglucosamine interacting residues of a protein from its primary sequence.* Protein Sci, 2020. **29**(1): p. 201-210.
54. Pappas, C., et al., *Single gene reassortants identify a critical role for PB1, HA, and NA in the high virulence of the 1918 pandemic influenza virus.* Proc Natl Acad Sci U S A, 2008. **105**(8): p. 3064-9.
55. Vens, C., M.N. Rosso, and E.G. Danchin, *Identifying discriminative classification-based motifs in biological sequences.* Bioinformatics, 2011. **27**(9): p. 1231-8.
56. Ito, T., et al., *Evolutionary analysis of the influenza A virus M gene with comparison of the M1 and M2 proteins.* J Virol, 1991. **65**(10): p. 5491-8.
57. McGinnis, S. and T.L. Madden, *BLAST: at the core of a powerful and diverse set of sequence analysis tools.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W20-5.
58. Taylor, L.H., S.M. Latham, and M.E. Woolhouse, *Risk factors for human disease emergence.* Philos Trans R Soc Lond B Biol Sci, 2001. **356**(1411): p. 983-9.
59. McArthur, D.B., *Emerging Infectious Diseases.* Nurs Clin North Am, 2019. **54**(2): p. 297-311.
60. Jones, K.E., et al., *Global trends in emerging infectious diseases.* Nature, 2008. **451**(7181): p. 990-3.
61. Han, B.A., et al., *Rodent reservoirs of future zoonotic diseases.* Proc Natl Acad Sci U S A, 2015. **112**(22): p. 7039-44.
62. Banerjee, A., et al., *Bats and Coronaviruses.* Viruses, 2019. **11**(1).
63. Dhama, K., et al., *SARS-CoV-2 jumping the species barrier: Zoonotic lessons from SARS, MERS and recent advances to combat this pandemic virus.* Travel Med Infect Dis, 2020. **37**: p. 101830.

64. Fischhoff, I.R., et al., *Predicting the zoonotic capacity of mammals to transmit SARS-CoV-2.* Proc Biol Sci, 2021. **288**(1963): p. 20211651.
65. Cui, J., F. Li, and Z.L. Shi, *Origin and evolution of pathogenic coronaviruses.* Nat Rev Microbiol, 2019. **17**(3): p. 181-192.
66. Eng, C.L., J.C. Tong, and T.W. Tan, *Distinct Host Tropism Protein Signatures to Identify Possible Zoonotic Influenza A Viruses.* PLoS One, 2016. **11**(2): p. e0150173.
67. Dilcher, M., et al., *Genetic characterization of Tribec virus and Kemerovo virus, two tick-transmitted human-pathogenic Orbiviruses.* Virology, 2012. **423**(1): p. 68-76.
68. Teeling, E.C., et al., *Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species.* Annu Rev Anim Biosci, 2018. **6**: p. 23-46.
69. Pagel Van Zee, J., et al., *Tick genomics: the Ixodes genome project and beyond.* Int J Parasitol, 2007. **37**(12): p. 1297-305.
70. Zhang, M., et al., *Prediction of virus-host infectious association by supervised learning methods.* BMC Bioinformatics, 2017. **18**(Suppl 3): p. 60.
71. Galiez, C., et al., *WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs.* Bioinformatics, 2017. **33**(19): p. 3113-3114.
72. Li, H. and F. Sun, *Comparative studies of alignment, alignment-free and SVM based approaches for predicting the hosts of viruses based on viral sequences.* Sci Rep, 2018. **8**(1): p. 10032.