Computational and Structural Biotechnology Journal 20 (2022) 4656–4666

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Univ-flu: A structure-based model of influenza A virus hemagglutinin for universal antigenic prediction

Jingxuan Qiu [a,1], Xinxin Tian [a,1], Yaxing Liu [a], Tianyu Lu [a], Hailong Wang [a], Zhuochen Shi [a], Sihao Lu [a], Dongpo Xu [a], Tianyi Qiu [b,*]

[a] School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
[b] Zhongshan Hospital, Fudan University, Shanghai 200032, China

ARTICLE INFO

Article history:
Received 24 May 2022
Received in revised form 23 August 2022
Accepted 24 August 2022
Available online 28 August 2022

Keywords:
Influenza virus
Hemagglutinin
Antigenic prediction
In-silico model

ABSTRACT

The rapid mutations on hemagglutinin (HA) of influenza A virus (IAV) can lead to significant antigenic variance and consequent immune mismatch of vaccine strains. Thus, rapid antigenicity evaluation is highly desired. The subtype-specific antigenicity models have been widely used for common subtypes such as H1 and H3. However, the continuous emerging of new IAV subtypes requires the construction of universal antigenic prediction model which could be applied on multiple IAV subtypes, including the emerging or re-emerging ones. In this study, we presented Univ-Flu, series structure-based universal models for HA antigenicity prediction. Initially, the universal antigenic regions were derived on multiple subtypes. Then, a radial shell structure combined with amino acid indexes were introduced to generate the new three-dimensional structure based descriptors, which could characterize the comprehensive physical–chemical property changes between two HA variants within or across different subtypes. Further, by combining with Random Forest classifier and different training datasets, Univ-Flu could achieve high prediction performances on intra-subtype (average AUC of 0.939), inter-subtype (average AUC of 0.771), and universal-subtype (AUC of 0.978) prediction, through independent test. Results illustrated that the designed descriptor could provide accurate universal antigenic description. Finally, the application on high-throughput antigenic coverage prediction for circulating strains showed that the Univ-Flu could screen out virus strains with high cross-protective spectrum, which could provide *in-silico* reference for vaccine recommendation.

1. Introduction

Influenza virus is one of the most serious threats to human public health, which causes about 3 million to 5 million cases of disease and 291,000 to 646,000 deaths globally each year [1]. Among all the human susceptible influenza types, influenza A virus (IAV) was the major infectious type that is highly contagious to human [2]. The subtypes of IAV are classified according to the two surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA) proteins [3]. The IAV was currently divided into two main groups, group 1 involved HA subtypes such as H1, H2, H5 H6 and H9, while group 2 involved HA subtypes such as H3 and H7[4]. Due to the rapid mutation of HA, antigenic variance was frequently occurred, resulting in the failure of vaccines effectiveness [5]. Therefore, accurate and rapid evaluation of antigenicity can not only help identifying antigenic drift strains, but also be used for vaccine recommendation [6].

Traditional experiment measurement of influenza antigenicity was mostly performed by the hemagglutinin inhibition test (HI) [7], which is the current golden standard for flu antigenicity criterion. Nevertheless, the time-consuming and labor-intensive nature, as well as the critical experimental condition, makes it unavailable for large-scale screening of IAV strains. More importantly, the sudden outbreak of new emerging IAVs such as H9 [8] makes it difficult for quick response of laboratory-based vaccine recommendation. Therefore, accurate and high-throughput approaches for antigenic measurement of IVA are highly desired.

In 2004, Smith et al. provided a comprehensive study to indicate the antigenic mapping of the historical influenza A/H3N2 viruses circulating from year 1968 to 2003 [9]. This work tested the pairwise antigenic relationship through anti-serum experiments and identified 11 antigenic clusters based on the two-dimensional

* Corresponding author.
E-mail address: ty_qiu@126.com (T. Qiu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2022.08.052

antigenic mapping of multidimensional scaling (MDS) [9]. This work was later defined as the standard dataset of H3 virus for multiple *in-silico* models. Since then, machine learning approaches were applied on antigenicity prediction for specific influenza subtypes. Most of them encoded the virus sequences or mutations through different descriptors. Typically, Liao's work generated a series of scoring scale to describe residue mutations, which involved amino acid groups with different polarity, charge, and aliphatic properties [10]. Combined with regression and classification models, those scoring scale could be used to predict antigenic variation of influenza A /H3N2 viruses [10]. Later, it is realized that structure information was critical for antigenicity measurement. Representative work such as PREDAC proposed a Naive Bayes prediction model, using a 12-bit binary vector to calculate structural and physicochemical characteristic differences for each HA sequence pairs [6]. To improve the prediction accuracy for fuzzy regions, Qiu et al. added the structural context description of influenza A/H3N2 HA protein into prediction model. Results showed that the structure model could accurately detect antigenic-escape strains, especially for those strain pairs around antigenic border regions, the model could be used in identifying those to-be-failed vaccine strains [5].

Besides H3 virus, the models for other subtypes were also constructed. For example, Peng's work integrated the HA antigen sites of human influenza H3 virus, the surface sites reported to be associated with immune escape of H5 virus, and the mutation sites with large historical information entropy to construct a comprehensive mutation sites [11]. Based on above sites, the computational model was constructed for the rapid and effective investigation of the highly pathogenic avian influenza H5N1 [11]. Similarly, based on the previous PREDAC model for H3N2, PREDAC-H1 was constructed to illustrate the antigenic evolution of H1N1 virus, focusing on the antigenic patterns and mutations of human influenza A/H1N1 virus from 1918 to 2014 [12].

Above subtype-specific models provide accurate prediction on antigenicity of influenza virus, while the application scope was limited on individual subtype. Currently, since H1 and H3 are the main IAV circulating in the community, their sequence data are very abundant. However, for some rare but epidemic-causing subtypes such as H5, H7, H9, or recombination cases such as the 2009 swine flu pandemic, the insufficient historical data makes it difficult for model construction. Thus, the need of universal computational model for influenza virus is ever-growing.

To solve this issue, some works have tried to reveal the common features among different IVA subtypes. For example, PREDAV-FluA analyzed the sequence mutation patterns of nine HA subtypes and found that different subtypes shared similar mutation patterns on HA protein, which provided the basis for the establishment of a universal computational model [13]. Then regional bands on the HA structure were divided based on the spatial distance of residues to the top of HA1, the mutation number in each regional band was calculated as descriptor. The universal model could achieve an accuracy rate of 77 % [13]. Meanwhile, a subtype-independent model, CFreeEnS was proposed using a context-free encoding scheme for protein sequences, which encodes protein sequence data sets into a numeric matrix and provide antigenic prediction for different influenza subtypes [14].

These universal models provide great instruction on subtype-free antigenic prediction. Nevertheless, the proposed descriptor was based on the sequence features on the full length of HA protein, rather than the antigenic-related sites. It is well known that the antigenicity changes were more related with the mutations on specific region which caused structure deviation or physical–chemical property changes [15], and it is urgently desired to design a structure based descriptor for antigenic sites, which could be suitable for different subtypes.

In this study, a structure based universal model was constructed for multiple subtypes from both group 1 and group 2 of IAV. By introducing the shell structure model on HA protein, the structure information can be divided around the receptor binding sites. Then, antigenic descriptors encoding property features were generated to describe the property variance between two compared HA proteins. Finally, by combining the descriptors with machine learning approaches, a series of models were proposed to predict the antigenic variance for different IAV subtypes. Results showed that the Random Forest classifier with structure-based descriptors could reach the good performance for intra-subtype, inter-subtype, inter-group and universal prediction based on independent testing dataset. Moreover, we tested the application ability of this model to calculate the antigenic coverage for circulating strains of different subtypes. Several strains with high protective spectrum were proved to be potential vaccine strains through experimental validation [16,17].

2. Methods

2.1. Dataset

For model construction, three types of data are required: 1) amino acid sequences of HA protein, 2) three-dimensional structure of HA protein, and 3) HI assays between multiple influenza strain pairs. In this study, the HI assay for H1 and H3 were collected from related papers [5,13], HI assay for H5, H7 and H9 were collected from World Health Organization (WHO) Weekly Epidemiological Record (Supplementary Table S1).

Based on the HI test, the antigenic distance (D_{ab}) for strain a and strain b could be calculated as formula (1).

$$D_{ab} = \frac{1}{2}\log\left(\frac{H_{aa}H_{bb}}{H_{ab}H_{ba}}\right) \quad (1)$$

where the parameter of H_{ab} represents the maximum dilution value of antisera against strain a, which was effective to prevent the cell agglutination caused by strain b. Considering the influence of different experimental conditions, for the strain pair which obtained different D_{ab} calculated from different HI tables, the outliers were removed and the remained D_{ab} were normalized to calculate the final average $\bar{D}_{ab}$ [5,18]. Strain a and strain b were defined as antigenic escape (negative sample) pair, if $log^{-1}D_{ab}$ was greater than or equal to 4, otherwise, this pair was defined as the antigenic similar (positive sample) [10].

Sequence data of HA1 were collected from published papers [5,13] and Influenza Virus Resource [19]. All sequences were separately aligned according to subtype by Clustal X [20] through multiple sequence alignment (MSA) to reach the type-specific consensus length of HA1 sequences. After MSA, sequences with inserted gaps over 10 % of the aligned sequence were removed. Finally, a total number of 1,422 HA1 sequences were retained in our dataset including 68 for H1, 725 for H3, 162 for H5, 437 for H7 and 30 for H9, respectively. The consensus sequence length for five subtypes were 327 (H1), 330 (H3), 320 (H5), 317 (H7) and 317 (H9) amino acids, respectively. In MSA, the distribution frequency of 20 amino acids on each position was calculated. For each position or sites, if none of residues obtained distribution frequency more than 0.8, it is illustrated that none of dominant residues occurred on this position. Then the position was defined as frequently mutated sites.

For HI assays, a standard pair should involve two HA sequences of compared strains and their corresponding antigenic distance from HI test. If one strain included several different HA sequences, due to variation in different viral quasispecies or sequencing error, all of the sequences were included. All the possible sequences were

considered for one strain pair, which meant multiple HA pairs were derived for one strain pair. This may happen in the most abundant dataset of H3, in which 3,867 HA pairs were defined from 3,539 strain pairs for H3. In total, 355 pairs for H1, 3,867 pairs for H3, 317 pairs for H5, 6 pairs for H7 and 89 pairs for H9 were collected for further analysis.

For three abundant datasets of H1, H3 and H5, 80 % of the pairs for each subtype were randomly divided to construct the training dataset, without changing the proportion of positive samples. The remained 20 % of the dataset was defined as the independent testing dataset. The small dataset of H7 and H9 were used for independent testing set and application cases.

Protein structure of HA-antibody complexes were collected from Protein Data Bank [21], the dataset included 8 PDB structures for H1, 9 PDB structures for H3 and 3 PDB structures for H5, respectively (Supplementary Table S2). In each complex structure, the epitope regions were defined as those residues on HA with nearest atom distance towards the corresponding antibody less than 4 Å [15]. Template PDB structures for each subtype were also collected from Protein Data Bank [21] (pdb_id: 3LZG for H1, pdb_id 6AOU for H3, pdb_id: 2IBX for H5, pdb_id: 4KOL for H7, pdb_id: 1JSD for H9) [22–26]. The epitopes derived from different complexes were mapped on the template structures according to each subtype. Finally, the non-redundant epitope dataset can be derived for each subtypes.

2.2. Antigenic descriptor

The antigenic descriptors were designed to describe the influence of residue mutation on antigenicity, the generation of antigenic descriptors contains the following steps:

(1) Defining the center of antigenic sites on HA1 protein.

In order to construct an antigenic prediction model for all subtype influenza virus, a common region on HA1 protein should be selected as the target sites. Thus the receptor binding sites on HA1 protein, which was recognized by human receptor, were selected as the core region. According to Gamblin and colleagues [27], the receptor binding sites for the HA1 protein is formed by three secondary structural elements: the 130 loop, the 190 helix, and the 220 loop. The amino acids for three regions were mapped on the template structure of each subtype and the geometric center for three sub-regions was calculated through the Euclidean distance to determine the antigenic center.

(2) Dividing the spatial layout of residues by shell structure.

For each subtype, a sphere structure was generated by taking the antigenic center as the core with radius of 40 Å. By taking the step size as 4 Å, 10 shells can be divided for each sphere structure. Then, based on the Euclidean distance towards the antigenic center, each residue r can be divided into the certain layers of the shell structure, as formula (2) illustrated.

$$\text{if } (\mathrm{i}-1)*4\text{Å} \le ED(r,c) < \mathrm{i}*4\text{Å},\ r \in \text{layer}\,(\mathrm{i}-1) \tag{2}$$

where $ED(r,c)$ represented the Euclidean distance between residue r and antigenic center c, $r \in \text{layer}\,(\mathrm{i}-1)$ means residue r located in layer $(\mathrm{i}-1)$.

(3) Antigenic descriptor generation

There are many factors associated with HA protein that can influence the antigenicity of influenza viruses, including physicochemical characteristics [28], such as hydrophobicity, and spatial conformation characteristics [5]. At the same time, Yao et al. proved that these influencing factors do not exist independently, but work as a combination of multiple influencing factors [29]. By targeting the residues located in each layer of constructed shell structure, quantitative antigenic descriptors, included property descriptor and glycosylation descriptor, were calculated.

For property descriptor, the physical–chemical properties which affect the antigenic change were quantified. Properties which affect the protein interaction and immune-recognition such as isoelectric point score [15], van der Waals volume scores [15], and hydrophobicity-related indices [30] were introduced to quantify the property score of each residue. Here, four indices describing the hydrophobic property (AAindex ID: EISD840101, ARGP820101, PONP800101, GOLD730101) of each amino acid from AAindex [31] were used in this study. The hydrophobic score, along with the isoelectric point score (AAindex ID: ZIMJ680104), van der Waals volume score (AAindex ID: FAUJ880103, LEVM760106) were introduced to quantify the property score of each residue. For a specific HA, the score for each layer in shell structure were calculated as the total AAindex scores of residues included in the layer. For two compared virus pair, the score for each layer were calculated as the absolute difference between two HA proteins.

Glycosylation, one of the important post-transcriptional modifications, could provide great impact on antigenic recognition between antigen and antibody. The appearance of glycosylation shields underlying residues from the contact of immune-antibodies [32]. The *N*-glycosylation occurred in the sequons of Asn-X-Ser/Thr, where X represents any amino acid other than proline, and the glycan chain is modified on the sequencer's asparagine (Asn) [33]. In this study, the number of glycosylation sites was counted. For a virus pair, score for each layer was calculated as the absolute difference of glycosylation sites included in corresponding layer.

Finally, for each pair of virus strains, 10 layers of shell structure was constructed for 7 amino acid indexes and 1 glycosylation descriptor, which lead to the total number of 80-bits quantitative descriptors for antigenic model construction.

2.3. Model construction

To construct the universal antigenic prediction model, 80-bit antigenic descriptor was used as the feature vector and the antigenic classification (antigenic similar or antigenic variant) was treated as label. Different machine learning approaches, including Logistic Regression, Bayes Net, Random Forest were introduced to build the *in-silico* model through Weka software [34]. The machine learning approach which could provide optimal prediction performance was chosen. The universal computational model to calculate antigenic relationship for different influenza virus subtypes was constructed according to the workflow shown in Fig. 1.

2.4. Model evaluation

The model was tested in three levels: 1) intra-subtype evaluation, 2) inter-subtype evaluation and 3) universal model evaluation, and the following parameters were introduced to evaluate model performance as formula (3) to (6) illustrated.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{4}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

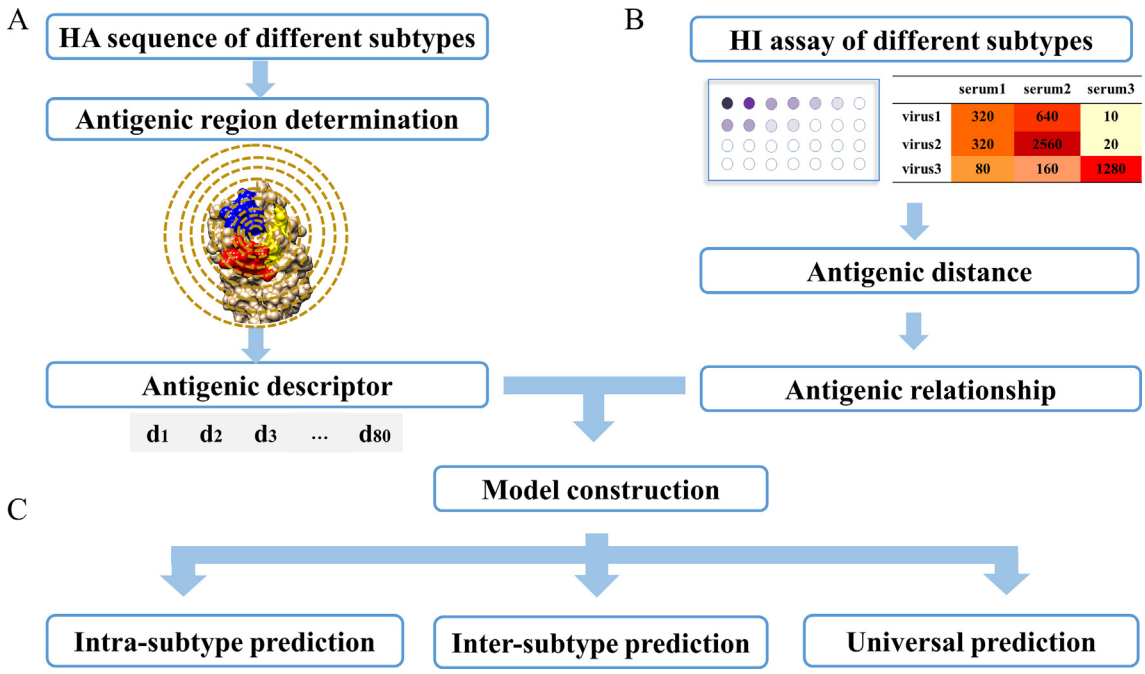


Fig. 1. The flowchart of model construction for antigenicity prediction. (A) HA sequences were collected from public resources. Antigenic center and shell structure were determined to describe the residue layout. Antigenic descriptor was designed to describe the property change of each HA pair. (B) Antigenic relationship was determined by HI assay. (C) Antigen prediction model was constructed.

$$\text{Balanced Accuracy} = \frac{Sensitivity + Specificity}{2} \quad (6)$$

Here, TP stands for true positive samples, TN stands for true negative samples, FP stands for false positive samples and FN stands for false negative samples.

2.5. Antigenic network construction for each virus subtype

For model application, all pairs of HA between any two influenza virus within each subtype were predicted by the proposed model in this study. And the antigenic network was constructed using Cytoscape [35], influenza strains were treated as nodes and the edge between two nodes referred to antigenic similar relationship. In antigenic network, the value of degree for each node was calculated [36,37] referring to the number of connected nodes. All the influenza nodes were ranked according to the degree value in descending order and top-ranked influenza strains refers to the strain with broad antigenic coverage.

3. Results

3.1. New determined antigenic regions could measure antigenic mutations on different subtypes of virus

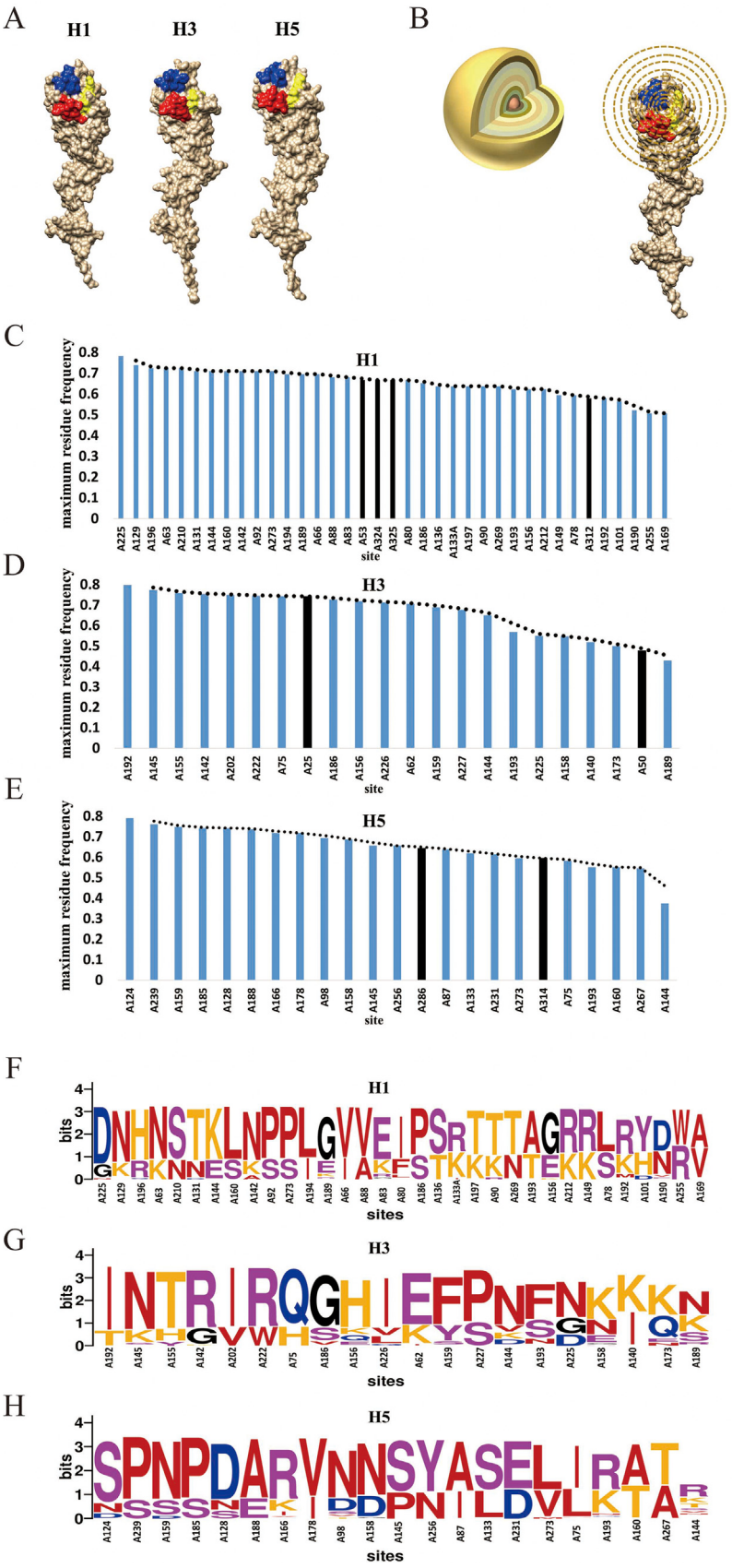
The foundation of constructing the universal model for all influenza subtypes was to determine the universal antigenic regions that could be applied on all HA proteins. It was noted that different influenza subtypes had different determined antigenic sites, such as the Sa, Sb, Ca, Cb, Pa, and Pb sites for H1 [38], the A, B, C, D, and E regions for H3 [39], these sites were overlapped but not completely identical. However, to make the universal antigenic regions applicable for all subtypes, the general antigenic sites should be derived.

To derive the general antigenic regions which could be applied for all influenza subtypes, the well-studied receptors binding sites on HA including the 130-loop (6 amino acids), 190-helix (10 amino acids) and 220-loop (10 amino acids) (Supplementary Table S3–S5) were introduced to generate the universal antigenic sites, which were illustrated in Fig. 2A. The center of antigenic regions was set as the geometric center within the receptor binding sites, as illustrated in Fig. 2B. Then, residues located within 40 Å around the antigenic center were defined as antigenic region for model construction. The selected antigenic region was included in both head and stem [40] sub-domain of HA protein.

Even through, the length of HA protein for different virus types ranged from 317 to 330 amino acids, the receptor binding sites were located on the similar regions on the HA structure (Fig. 2A). The standard deviation of the geometric distances between the centers of three sub-regions and the defined antigenic centers in three influenza subtypes were only 0.09 Å to 0.54 Å (Supplementary Table S6), indicating similar structure shape of the antigenic regions in different HA subtypes. Thus, the antigenic site and shell structure division might be applied on different virus subtypes.

Further, the statistical analysis of mutation distribution was performed to evaluate whether the antigenic region determined by shell structure could include the majority of the frequently mutated sites on all HA subtypes. Here, the frequently mutated sites were defined as those sites with the maximum amino acid frequency less than 0.8 (see Methods), as illustrated in Fig. 2C–E. For those frequently mutated sites, the maximum residue frequency ranged from 0.507 to 0.783 for H1, 0.429 to 0.799 for H3, 0.374 to 0.791 for H5, indicating the variable mutation happened on each sites. Among these mutation sites, 89.19 % (33/37), 90.91 % (20/22) and 91.30 % (21/23) sites were included in the determined antigenic region for H1, H3, and H5, respectively. It is showed that the antigenic region divided by shell structure could incorporate approximately 90 % of the mutation sites. As been illustrated in Fig. 2F–H, the mutation sites included in antigenic region obtained more variable amino acid distribution and more possible residue types.

Previous studies have also shown that the antigenic drift of influenza virus is mainly mediated by mutations in HA epitopes [41,42]. By mapping the epitope regions on the HA protein of H1,


A
H1
H3
H5
B
C
H1
maximum residue frequency
site
D
H3
maximum residue frequency
site
E
H5
maximum residue frequency
site
F
H1
bits
sites
G
H3
bits
sites
H
H5
bits
sites

H3 and H5, the non-redundant epitope dataset of H1, H3 and H5 includes 61 residues, 58 residues and 31 residues respectively (Supplementary Table S7), these epitopes located on both the head and the stem region of HA. It could be found that 78.7 % (48/61), 75.4 % (43/57) and 90.3 % (28/31) of the epitopes were included in the antigenic region derived by shell structures, for H1, H3 and H5 virus respectively. Moreover, for epitope located on the head region of HA, 100 % (31/31), 93.5 % (43/46) and 100 % (28/28) of the epitopes were involved in antigenic region, for H1, H3 and H5 respectively. In general, the common antigenic region proposed in this study contained the majority of reported epitopes for different HA types including H1, H3 and H5.

3.2. Performance of intra-subtype, inter-subtype and universal model on antigenic prediction

By incorporating the antigenic descriptor and HI test between two compared strains, different machine learning approaches can be introduced to constructed classification model. Here, we have introduced three types of model including intra-subtype, inter-subtype and universal model to validate the applicability of our structure-based descriptors. The intra-subtype model was trained with specific HI-test of individual HA type and tested by the same HA type, for example, H1-intra-model used the HI-test of H1 for modeling and was also evaluated on the testing dataset of H1 variants. The inter-subtype model was trained with one specific type and tested by another type, for example, one inter-model used the HI-test of H1 for modeling and was evaluated on the testing dataset of H3. The universal model was training by mixed dataset of multiple HA types, and was also tested by any type of influenza virus.

3.2.1. Intra-subtype antigenic prediction performance

Initially, the intra-subtype models were constructed by Logistic regression, Bayes Net and Random Forest classifiers, and were evaluated by the specific subtype through 10-fold cross-validation. It can be found that Random Forest classifier could maintain the highest performance for all subtype by intra-subtype models of H1 (Fig. 3A), H3 (Fig. 3B), and H5 (Fig. 3C). Thus, Random Forest was used to construct antigenic prediction model in the following analysis. Results showed that, the H3 intra-subtype model could achieve the highest AUC value of 0.986 among all three subtypes, and followed by 0.928 for H5 intra-subtype model and 0.867 for H1 intra-subtype model. Moreover, for all three models, the accuracy, sensitivity and specificity could achieve a high level over 0.8 (Supplementary Table S8), indicating high prediction performance for 10 fold cross-validation. In addition, we performed an independent test (see Method) for each subtypes through Random Forest classifier (Fig. 3D). Similarly, the H3 intra-subtype model could achieve the highest AUC value of 0.988 on independent testing dataset. The similar good performance with AUC of 0.951 for H5 and AUC of 0.877 for H1 could also been observed (Supplementary Table S9).

3.2.2. Inter-subtype antigenic prediction performance

Further, the applicability of descriptors was evaluated by inter-subtype model. The training datasets including H1, H3 and H5, while the testing datasets including H1, H3, H5 and H9. As been illustrated in Fig. 3E, the inter-subtype model could achieved a AUC value ranged from 0.646 to 0.861, and the accuracy ranged from 0.676 to 0.797 (Supplementary Table S10). The performance was slightly lower than the results from intra-subtype model. Moreover, according to previous knowledge, the IAVs can be divided into two groups, group 1 and group 2 [4]. Thus, we constructed the inter-group model, which was trainined by one group and evaluated on the other one. In our dataset, the H1, H5 and H9 belonged to group 1, while H3 and H7 belonged to group 2. As been shown in Fig. 3F, the intra-group model illustrated better prediction performance (average AUC of 0.931) than the inter-group model. For inter-group prediction, the prediction AUC value and accuracy could still maintain the level over 0.7 (Supplementary Table S11). This results indicate that the inter-group prediction based on antigneic descriptors could also provide tolerable perdiction performance.

3.2.3. Universal antigenic prediction performances

Finally, the universal model of Univ-Flu was trained based on different subtypes and was expected to predict the antigenic relationship for all different IAV subtypes. Here, the most abundant dataset including subtype H1, H3 and H5 were used for model construction. By the 10-fold cross-validation based on the mixed dataset of H1, H3 and H5, the Random Forest classifier could achieve the best prediction performance with AUC of 0.977 and accuracy of 0.929 (Fig. 3G, Supplementary Table S12).

Moreover, the universal model Univ-Flu was evaluated through five independent testing dataset: 1) mixed dataset of H1, H3 and H5, 2) mono-type dataset of H1, 3) mono-type dataset of H3, 4) mono-type dataset of H5, and 5) mono-type dataset of H9. Results showed that the best prediction performance was achieved on H3 testing dataset with AUC value of 0.987 and accuracy of 0.937, followed by mixed dataset with AUC of 0.978 and accuracy of 0.915 (Fig. 3H, Supplementary Table S13). The performance on H1 and H5 testing set were higher than those of H9, which could achieve the AUC of 0.843 for H1 and AUC of 0.938 for H5. Even for H9, the performance could obtain with AUC value of 0.800 and accuracy of 0.730 (Fig. 3H, Supplementary Table S13).

It should be noted that, the dataset of H9 was not introduced to construct the universal model, thus the H9 testing set is totally independent. The above results indicated that the universal model trained by mixed subtypes could perform antigenicity prediction for different subtypes with good prediction performance. The antigenic descriptors of the universal model could integrate the HA and HI-test of different subtypes and be applied for other influenza virus.

3.3. Better antigenic escaping detection and application for vaccine recommendation

To explore the application scope of our method, the performance of Univ-Flu was evaluated and compared with available tools. Till now, only a few works [13,14] attempt to construct universal model for influenza virus and PREDAV-FluA [13] is the current cutting-edge algorithm with a user-friendly online server. The performances of two methods were evaluated through the testing

Fig. 2. Antigenic region for different virus subtypes. (A) Illustration of receptor binding sites on HA protein of different virus subtypes. Residues on 130-loop, 190-helix and 220 loop were labeled in yellow, blue and red for H1 (3LZG), H3 (6AOU) and H5 (2IBX). (B) Illustration of antigenic region determined by shell structure model. (C–E) Frequently mutated sites of H1, H3 and H5 subtypes. Each bar refers to one mutation site, the height of the bar refers to the maximum residue frequency on the sites. Blue bar represent the site located in antigenic region, black bar refers to the mutation sites located on the outside of antigenic region. (F–H) Residue distribution on mutation sites located with antigenic region. The horizontal axis labeled the mutation sites in antigenic region, and the vertical axis is sequence conservation for different amino acid at each site. The residues of K, T, H were labeled in orange, D and Q were labeled in blue, residues of A, V, L, I, P, W, F, M, N were labeled in red, residues of S, Y, R and E were labeled in purple. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

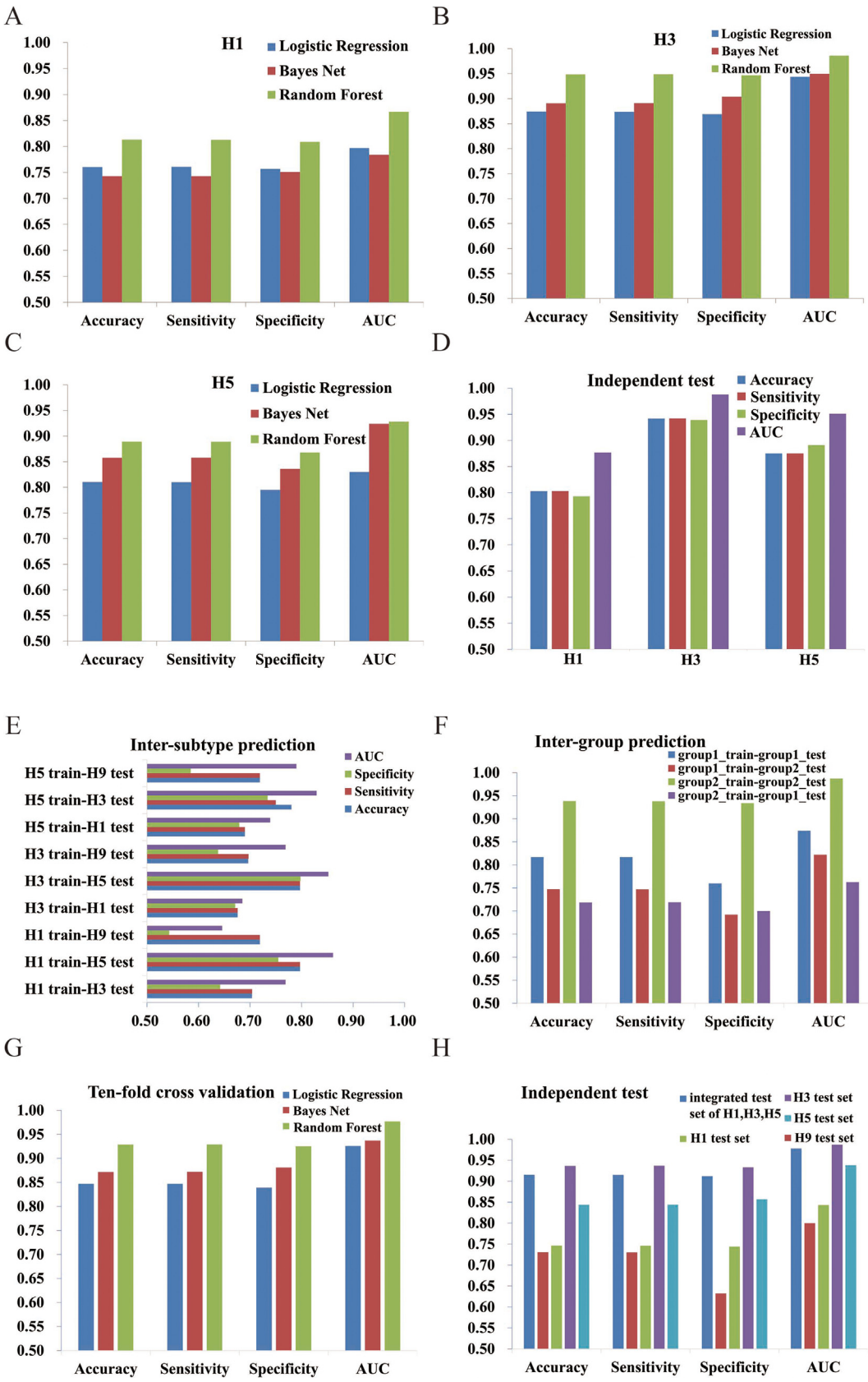


Fig. 3. Model performance on antigenic prediction. (A–C) Tenfold cross validation of intra-subtype classification based on different machine learning approaches. (D) Independent test performance of intra-subtype model. (E) Independent test performance of inter-subtype model. (F) Independent test performance of inter-group prediction. (G) Tenfold cross validation of universal antigenic prediction model based on the mixed training dataset of H1, H3 and H5. (H) Independent test of universal model. The integrated test set includes the test set of H1, H3 and H5.

Table 1
Comparison of model performance.

Influenza subtype of test set	Prediction model	Accuracy	Sensitivity	Specificity	Balanced Accuracy
H1	Univ-Flu	0.747	0.746	0.744	0.745
	PREDAV-FluA [13]	0.662	0.632	0.697	0.665
H3	Univ-Flu	0.937	0.937	0.933	0.935
	PREDAV-FluA [13]	0.780	0.800	0.750	0.775
H5	Univ-Flu	0.844	0.844	0.857	0.851
	PREDAV-FluA [13]	0.813	0.923	0.640	0.782
H9	Univ-Flu	0.730	0.730	0.632	0.681
	PREDAV-FluA [13]	0.753	0.984	0.222	0.603

dataset of H1, H3, H5 and H9. Among them, H1, H3 and H5 training set were used for model construction in both Univ-Flu and PREDAV-FluA, while H9 was totally new for evaluation.

As been illustrated in Table 1, by setting the threshold, Univ-Flu could obtain the best prediction performance on H3 with accuracy of 0.937, and the accuracy reduced to 0.844 for H5, 0.747 for H1 and 0.730 for H9, respectively. For comparison, the classification results of Univ-Flu performed better than PREDAV-FluA in H1 and H3, and provided similar performance in H5. For the model-free subtype of H9, PREDAV-FluA performed slightly better than Univ-Flu. Above results showed that Univ-Flu could provide equally comparable classification than popular peer approach. On the other hand, the sensitivity and specificity between two approaches illustrated different application scope. It can be found that, the Balanced Accuracy (BA) of Univ-Flu (0.681–0.935) could outperform the BA of PREDAV-FluA (0.603–0.782) in all four independent testing dataset. Further, PREDAV-FluA provides high sensitivity rather than specificity on most of the subtype. This was more obvious in the prediction of H9, while sensitivity of 0.984 and specificity of 0.222 were obtained by PREDAV-FluA. The high sensitivity and relatively low specificity indicated the approach could provide high positive detection rate (antigenic similar events) and high false detection rate. In general, Univ-Flu could provide high prediction performance in H1, H3 and H5. For H9, Univ-Flu provides an improvement on specificity, which illustrated the advantages in detecting antigenic escape events (negative sample).

In the prevention and control of influenza virus, vaccine is one of the most effective measurements. For vaccine recommendation, the antigenic match between emerging strains and vaccine strain was of importance to guarantee the effectiveness of vaccine. Most importantly, the antigenic escape of vaccine strains should be detected, or considered for vaccine recommendation. Above results showed that our model could better perform on specificity, which indicated the better ability of antigenic escape detection. Thus, it could be used to predict the protective spectrum of potential vaccine strain with high-throughput antigenic screening.

To achieve that, the antigenic relationship of pair-wised strains, which including 2,278 for H1, 262,450 for H3, 13,041 for H5, 95,266 for H7 and 435 for H9, were predicted through the above universal antigenic model. Then, an antigenic network could be constructed to illustrate the antigenic relations between different strains. As been illustrated in Fig. 4A, each node refers to a strain and each line refers to antigenic similar relationship between two connected nodes. The size of nodes was proportioned to the degree, while higher degree refers to larger antigenic protection spectrum. By taking the smaller group of H9 as example for illustration, the results showed that A/swine/Hongkong/9/1998-like virus hold the highest degree, followed by A/swine/Shandong/Fjn/2003-like, A/chicken/Shenzhen/U/1999-like, and A/chicken/Hongkong/G9/1997-like. Among them, A/chicken/Hongkong/G9/1997 was selected as vaccine candidate for preventing highly pathogenic avian influenza in 2009 [16], indicating the successful detection of potential vaccine strain by our model. Moreover, the HA of this potential vaccine strain shared 97.8 % (310/317) sequence identity with the HA of top 1 ranked strain of A/swin/Hongkong/9/1998-like in our profile (Fig. S1).

Results for other subtypes are showed in Fig. 4B, we found that the top 1 ranked strains for H1, H3, H5 and H7 were A/Virginia/1/2006-like (H1N1), A/England/700/2009-like (H3N2), A/chicken/India/NIV-33487/2006-like (H5N2), and A/wild bird/Korea/5-77/2005-like (H7N9), respectively. Interestingly, by checking the top ranked strains in each subtype, A/turkey/Turkey/1/2005-like (H5N2) which ranked as top 3 in the protective spectrum list of H5, was selected as vaccine candidate for avian influenza in 2017 [17]. More importantly, the HA of recommended vaccine strain A/turkey/Turkey/1/2005-like (H5N2) and the HA of top 1 ranked vaccine candidate A/chicken/India/NIV-33487/2006-like (H5N2) shared 98.75 % (316/320) identity in amino acid sequence (Fig. S2). Above illustrated that Univ-Flu could be used to calculate antigenic coverage among the circulating strains. Further, it could provide guidance for vaccine recommendation. The top 10 ranking strains for four subtypes were listed in Supplementary Table S14–17 as reference.

4. Discussion

Due to the rapid evolution of virus and the constant antigenic drift events, high-throughput antigenicity measurement is a difficult but urgently desired task. By taking advantage of the accumulated HI assay and HA sequences, computational methods could be constructed for rapid and high-throughput antigenicity prediction. However, for new emerging subtypes, it is difficult to construct the subtype-specific model without enough historical data. Thus, it will be of great interests to construct a universal model which could be applied on different subtypes of influenza, especially for those new emerging or re-emerging influenza subtypes.

The key point to build a universal model is to derive the applicable features which could describe the characteristics of the mutants for each subtype. To achieve those, peers introduced the whole sequence features or whole structure features, such as whole sequence based mutation matrix [14] or regional bands which divided the HA spatial structure from top to the bottom to describe the amino acid mutations on different regional bands [13]. However, it is well known that antigenic drift events depend on the mutations located on the antigenic site, or epitopes [41,42]. Moreover, the mutations which caused the antigenicity variance may frequently involve large changes of structure or physical–chemical property, and whole HA-based descriptors might involve noises rather than information. For example, in the comparison of HA sequences between A/California/7/2009(H1N1) and A/Michigan/45/2015 (H1N1), there were 12 mutations on the whole HA1 sequence, which contains over 3.72 % (12/323) of mutations. For the whole sequence or whole structure based approaches [13,14], this pair was predicted as antigenic drift, due to the large number of mutations. However, by checking the mutations on the well-

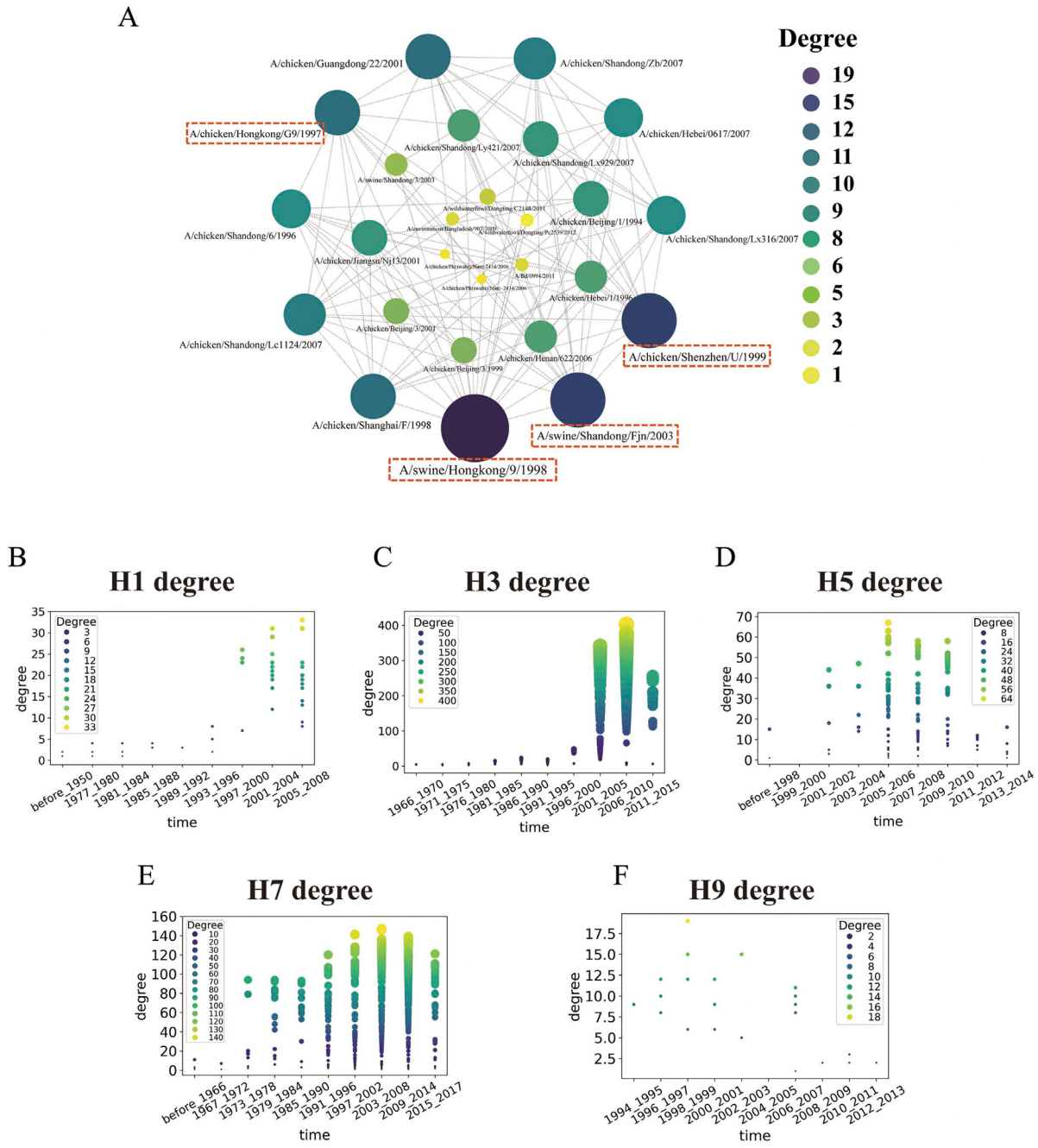


Fig. 4. Antigenic network illustration and antigenic degree distribution. (A) Antigenic network of H9 subtype, each strain node is arranged into circle from outside to inside according to degree value from high to low. (B-F) Degree distribution of virus strains for H1, H3, H5, H7 and H9 subtypes. X-axis refers to the emerging time for each strain. Y-axis refer to specific degree value ranked in descending order.

defined antigenic regions of Sa, Sb, Ca, Cb, Pa and Pb [38], 6 mutations were detected, including N84S, N97D, N162S, Q163K, T203S and T216I, among which only N97D, Q163K and T216I involved property change (Supplementary Figure S3). By checking the influenza surveillance report from WHO, those two strains were defined as antigenic similar through HI test [16], which was consisted with the prediction of Univ-Flu. The successful prediction of Univ-Flu may benefit from the physical chemical descriptors of the antigenic region, which could radially describe the local micro-environment around the antigenic regions. In this study, instead of the full-length mutation, the most common character of receptor binding sites shared by HA proteins from different influenza subtypes was considered for descriptor generation. In fact, the positions involved in our shell structure shared an overlap with 79.07 % (102/109) of the well-defined H1 antigenic sites and 83.21 % (109/131) of the well-defined H3 antigenic sites [38] (Supplementary Table S18–S19).

The prediction performance of Univ-Flu illustrated that the universal model could be applied on different influenza subtypes, even for those not involved in the training set. Due to the high similarity of structures in each subtype, it is reasonable that the prediction result of intra-subtype is better than those of inter-subtype or inter-group models. Nevertheless, in the inter-subtype prediction, the model constructed with H1 and H3 could provide accurate antigenic prediction for H5 and H9, which showed that the designed descriptor and model could make full use of data from

different subtypes and applied for other new emerging or re-emerging influenza viruses. More importantly, besides high AUC and BA performance, Univ-Flu seems to provide better specificity than sensitivity, which means it could accurately detect those antigenic drift events. For vaccine recommendation, the key issue is to calculate the protection spectrum of the immunogen. Although, relatively lower sensitivity indicated the potential miss of inspection on antigenic similar events and would lead to the decrease of the protective spectrum predicted in our model. The strains with high protective spectrum could still be considered as potential vaccine strains. As mentioned before, the strains of A/turkey/Turkey/1/2005-like (H5N2) and A/chicken/Hongkong/G9/1997-like (H9N2), which ranked as top 3 and top 4 in our list were validated to be potential vaccine. Those two strains hold high sequence identity with our top 1 strain in the prediction list.

Although, the Univ-Flu could provide good prediction performance for universal antigenic prediction of influenza A virus, there is still room for future improvement. For example, the accumulation of HI assays or new immune-recognition sites with experimental evidences will help improving the performance of our training-based model. Beside prediction performance, it also should be noticed that the application scope of our universal model is designed for fast response to new emerging influenza viruses. This could be used as a preliminary prediction tools for new occurred subtypes or subtypes without enough historical data. For the intensive prediction, such as predicting the antigenic relationship near the fuzzy region around the cutoff for specific subtypes [5] the subtype-specific model should be recommended.

On the other hand, it is also worth noting that this model provided is an "after mutation antigenic evaluation", which means the model could evaluate the consequences of mutation rather than "how it will mutate". In future, by incorporating with the mutation variants prediction, the universal model constructed in this study could provide comprehensive analysis of antigenic evolution for influenza virus.

CRediT authorship contribution statement

Jingxuan Qiu: Conceptualization, Methodology. **Xinxin Tian:** Formal analysis, Writing – original draft. **Yaxing Liu:** Data curation, Validation. **Tianyu Lu:** Data curation, Validation. **Hailong Wang:** Data curation, Validation. **Zhuochen Shi:** Data curation, Validation. **Sihao Lu:** Data curation, Validation. **Dongpo Xu:** Writing – review & editing. **Tianyi Qiu:** Writing – review & editing, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (32000470), the National Natural Science Foundation of China (31900483).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.08.052.

References

[1] Iuliano AD et al. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. Lancet 2018;391:1285–300. https://doi.org/10.1016/S0140-6736(17)33293-2.
[2] Coates BM, Staricha KL, Wiese KM, Ridge KM. Influenza A virus infection, innate immunity, and childhood. JAMA Pediatr 2015;169:956–63. https://doi.org/10.1001/jamapediatrics.2015.1387.
[3] Tong S et al. New world bats harbor diverse influenza A viruses. PLoS Pathog 2013;9:e1003657.
[4] Trost JF et al. A conserved histidine in Group-1 influenza subtype hemagglutinin proteins is essential for membrane fusion activity. Virology 2019;536:78–90. https://doi.org/10.1016/j.virol.2019.08.005.
[5] Qiu J, Qiu T, Yang Y, Wu D, Cao Z. Incorporating structure context of HA protein to improve antigenicity calculation for influenza virus A/H3N2. Sci Rep 2016;6:31156. https://doi.org/10.1038/srep31156.
[6] Du X et al. Mapping of H3N2 influenza antigenic evolution in China reveals a strategy for vaccine strain recommendation. Nat Commun 2012;3:709. https://doi.org/10.1038/ncomms1710.
[7] Hirst GK. Studies of antigenic differences among strains of influenza a by means of red cell agglutination. J Exp Med 1943;78:407–23. https://doi.org/10.1084/jem.78.5.407.
[8] Song W, Qin K. Human-infecting influenza A (H9N2) virus: a forgotten potential pandemic strain? Zoonoses Public Health 2020;67:203–12. https://doi.org/10.1111/zph.12685.
[9] Smith DJ et al. Mapping the antigenic and genetic evolution of influenza virus. Science 2004;305:371–6. https://doi.org/10.1126/science.1097211.
[10] Liao YC, Lee MS, Ko CY, Hsiung CA. Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. Bioinformatics 2008;24:505–12. https://doi.org/10.1093/bioinformatics/btm638.
[11] Peng Y, Zou Y, Li H, Li K, Jiang T. Inferring the antigenic epitopes for highly pathogenic avian influenza H5N1 viruses. Vaccine 2014;32:671–6. https://doi.org/10.1016/j.vaccine.2013.12.005.
[12] Liu M et al. Antigenic patterns and evolution of the human influenza A (H1N1) virus. Sci Rep 2015;5:14171. https://doi.org/10.1038/srep14171.
[13] Peng Y et al. A universal computational model for predicting antigenic variants of influenza A virus based on conserved antigenic structures. Sci Rep 2017;7:42051. https://doi.org/10.1038/srep42051.
[14] Zhou X, Yin R, Kwoh CK, Zheng J. A context-free encoding scheme of protein sequences for predicting antigenicity of diverse influenza A viruses. BMC Genomics 2018;19:936. https://doi.org/10.1186/s12864-018-5282-9.
[15] Qiu T et al. CE-BLAST makes it possible to compute antigenic similarity for newly emerging pathogens. Nat Commun 2018;9:1772. https://doi.org/10.1038/s41467-018-04171-2.
[16] Karron RA et al. A live attenuated H9N2 influenza vaccine is well tolerated and immunogenic in healthy adults. J Infect Dis 2009;199:711–6. https://doi.org/10.1086/596558.
[17] Pitisuttithum P et al. Safety and immunogenicity of a live attenuated influenza H5 candidate vaccine strain A/17/turkey/Turkey/05/133 H5N2 and its priming effects for potential pre-pandemic use: a randomised, double-blind, placebo-controlled trial. Lancet Infect Dis 2017;17:833–42. https://doi.org/10.1016/S1473-3099(17)30240-2.
[18] Qiu T et al. A benchmark dataset of protein antigens for antigenicity measurement. Sci Data 2020;7:212. https://doi.org/10.1038/s41597-020-0555-y.
[19] Bao Y et al. The influenza virus resource at the National Center for Biotechnology Information. J Virol 2008;82:596–601. https://doi.org/10.1128/JVI.02005-07.
[20] Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, 10.1093/bioinformatics/btm404 (2007).
[21] Berman HM et al. The protein data bank. Nucleic Acids Res 2000;28:235–42. https://doi.org/10.1093/nar/28.1.235.
[22] Ha Y, Stevens DJ, Skehel JJ, Wiley DC. H5 avian and H9 swine influenza virus haemagglutinin structures: possible origin of influenza subtypes. EMBO J 2002;21:865–75. https://doi.org/10.1093/emboj/21.5.865.
[23] Shi Y et al. Structures and receptor binding of hemagglutinins from human-infecting H7N9 influenza viruses. Science 2013;342:243–7. https://doi.org/10.1126/science.1242917.
[24] Yamada S et al. Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors. Nature 2006;444:378–82. https://doi.org/10.1038/nature05264.
[25] Wu NC et al. A structural explanation for the low effectiveness of the seasonal influenza H3N2 vaccine. PLoS Pathog 2017;13:e1006682.
[26] Xu R et al. Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. Science 2010;328:357–60. https://doi.org/10.1126/science.1186430.
[27] Gamblin SJ et al. The structure and receptor binding properties of the 1918 influenza hemagglutinin. Science 2004;303:1838–42. https://doi.org/10.1126/science.1093155.
[28] Suzuki Y. Predictability of antigenic evolution for H3N2 human influenza A virus. Genes Genet Syst 2013;88:225–32. https://doi.org/10.1266/ggs.88.225.
[29] Yao Y et al. Predicting influenza antigenicity from Hemagglutintin sequence data based on a joint random forest method. Sci Rep 2017;7:1545. https://doi.org/10.1038/s41598-017-01699-z.

[30] Hajari T, Bandyopadhyay S. Water structure around hydrophobic amino acid side chain analogs using different water models. J Chem Phys 2017;146:. https://doi.org/10.1063/1.4985671225104.
[31] Kawashima S et al. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 2008;36:D202–5. https://doi.org/10.1093/nar/gkm998.
[32] Chang D, Zaia J. Why glycosylation matters in building a better flu vaccine. Mol Cell Proteomics 2019;18:2348–58. https://doi.org/10.1074/mcp.R119.001491.
[33] Lees WD, Moss DS, Shepherd AJ. A computational analysis of the antigenic properties of haemagglutinin in influenza A H3N2. Bioinformatics 2010;26:1403–8. https://doi.org/10.1093/bioinformatics/btq160.
[34] Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. Data Mining: Practical Machine Learning Tools and Techniques. *The WEKA Workbench* (2016).
[35] Shannon P et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–504. https://doi.org/10.1101/gr.1239303.
[36] Scardoni G, Petterlini M, Laudanna C. Analyzing biological network parameters with CentiScaPe. Bioinformatics 2009;25:2857–9. https://doi.org/10.1093/bioinformatics/btp517.
[37] Vilela M et al. Parameter optimization in S-system models. BMC Syst Biol 2008;2:35. https://doi.org/10.1186/1752-0509-2-35.
[38] Quan L et al. Cluster-transition determining sites underlying the antigenic evolution of seasonal influenza viruses. Mol Biol Evol 2019;36:1172–86. https://doi.org/10.1093/molbev/msz050.
[39] Nation, D. c. W. I. A. W. G. J. S. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature* (1981).
[40] Padilla-Quirarte HO, Lopez-Guerrero DV, Gutierrez-Xicotencatl L, Esquivel-Guadarrama F. Protective antibodies against influenza proteins. Front Immunol 2019;10:1677. https://doi.org/10.3389/fimmu.2019.01677.
[41] Couch RB, Kasel JA. Immunity to influenza in man. Annu Rev Microbiol 1983;37:529–49. https://doi.org/10.1146/annurev.mi.37.100183.002525.
[42] Bush RM, Fitch WM, Bender CA, Cox NJ. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol 1999;16:1457–65. https://doi.org/10.1093/oxfordjournals.molbev.a026057.