



OPEN ACCESS

EDITED BY
Xuyong Li,
Liaocheng University, China

REVIEWED BY
Wei Peng,
Kunming University of Science and
Technology, China
Jiahao Zhang,
Huazhong Agricultural University, China
Jihui Ping,
Nanjing Agricultural University, China

*CORRESPONDENCE
Huihui Kong
✉ konghuihui@caas.cn

RECEIVED 28 November 2023
ACCEPTED 08 January 2024
PUBLISHED 19 January 2024

CITATION
Li X, Li Y, Shang X and Kong H (2024) A
sequence-based machine learning model for
predicting antigenic distance for H3N2
influenza virus. *Front. Microbiol.* 15:1345794.
doi: 10.3389/fmicb.2024.1345794

COPYRIGHT
© 2024 Li, Li, Shang and Kong. This is an
open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

A sequence-based machine learning model for predicting antigenic distance for H3N2 influenza virus

Xingyi Li^{1,2}, Yanyan Li^{1,2}, Xuequn Shang^{1,2} and Huihui Kong^{3*}

¹School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi, China, ²Big Data Storage and Management MIIT Lab, Xi'an, Shaanxi, China, ³State Key Laboratory of Animal Disease Control and Prevention, Harbin Veterinary Research Institute, Chinese Academy of Agricultural Sciences (CAAS), Harbin, China

Introduction: Seasonal influenza A H3N2 viruses are constantly changing, reducing the effectiveness of existing vaccines. As a result, the World Health Organization (WHO) needs to frequently update the vaccine strains to match the antigenicity of emerged H3N2 variants. Traditional assessments of antigenicity rely on serological methods, which are both labor-intensive and time-consuming. Although numerous computational models aim to simplify antigenicity determination, they either lack a robust quantitative linkage between antigenicity and viral sequences or focus restrictively on selected features.

Methods: Here, we propose a novel computational method to predict antigenic distances using multiple features, including not only viral sequence attributes but also integrating four distinct categories of features that significantly affect viral antigenicity in sequences.

Results: This method exhibits low error in virus antigenicity prediction and achieves superior accuracy in discerning antigenic drift. Utilizing this method, we investigated the evolution process of the H3N2 influenza viruses and identified a total of 21 major antigenic clusters from 1968 to 2022.

Discussion: Interestingly, our predicted antigenic map aligns closely with the antigenic map generated with serological data. Thus, our method is a promising tool for detecting antigenic variants and guiding the selection of vaccine candidates.

KEYWORDS

influenza A H3N2 virus, antigenic distances, virus antigenicity prediction, antigenic drift, antigenic variants

1 Introduction

Each year, seasonal influenza results in an estimated 3–5 million cases of severe illness, culminating in roughly 290,000–650,000 respiratory-related deaths (Nelson and Holmes, 2007; Russell et al., 2008; Iuliano et al., 2018). The H3N2 influenza virus is one of the predominant subtypes responsible for these outbreaks. While vaccination remains the most effective measure to combat seasonal influenza, the perpetual evolution of influenza A viruses necessitates regular updates of the flu vaccines. Since the H3N2 viruses became prevalent in humans in 1968, they have spread worldwide and experienced significant antigenic evolution. This evolution is characterized by alternating phases: periods of relative stability are followed by phases of rapid phenotypic changes. Research conducted by Koel et al. (2013), as outlined in their antigenic cartography, reveals that from 1968 to 2003, ten distinct antigenic clusters of H3N2 viruses were identified. Moreover, data from the World Health Organization indicates that from 2003 to 2023,

multiple additional antigenic clusters of H3N2 viruses have emerged (<https://www.who.int/teams/global-influenza-programme/vaccines/who-recommendations/recommendations-for-influenza-vaccine-composition-archive>). The surface glycoprotein Hemagglutinin(HA), responsible for attachment and fusion to host cell membranes, is the primary target of neutralizing antibodies (Webster et al., 1992; Wille and Holmes, 2020). Consequently, mutation(s) in the HA protein may trigger antigenic drift which allows the virus to evade the host's immune defenses, facilitating its continued spread and infection (Caton et al., 1982; Wilson and Cox, 1990; Peng et al., 2022b). At present, researchers largely depend on serological assays, such as the hemagglutination inhibition (HI) assay, to determine viral antigenicity. However, these techniques are labor-intensive, time-consuming, and offer only medium throughput (Sun et al., 2013). This underscores the pressing need to devise fast and precise methods to identify antigenic variants. Such advancements will optimize the selection of vaccines that align with emerging antigenic strains.

With the advancements in sequencing technology and the concurrent accumulation of viral sequences and serological data in databases, efforts are being made to employ computational algorithms for the prediction of viral antigenicity (Kilbourne et al., 2002). Such methods have the potential to significantly enhance the detection of antigenic variants and optimize the selection of vaccine strains. For example, a statistical method to correlate the HI titer with the number of mutations in the HA sequence of viruses was developed by Lee and Chen (2004). Multiple and logistic regression were applied to assess the relationship between mutations in the HA sequence and HI data (Liao et al., 2008). Decision tree algorithms were applied to predict drift variants by extracting the association from HI data using information theory (Huang et al., 2009). While these methods lay the foundation for predicting the antigenic evolution of influenza viruses, they do not quantify the antigenic distance between the viruses. In recent years, there have been several attempts to establish a quantitative relationship between viral HA sequences and antigenic distance. For instance, Sparse learning methods have been proposed to identify key sites influencing antigenic changes and establish a quantitative relationship between key sites and antigenic distances (Sun et al., 2013; Yang et al., 2014). Previous research has already demonstrated that only a limited number of sites are active in the process of antigenic change (Smith et al., 2004). Regression models such as support vector regression and joint random forest regression were applied to establish a quantitative relationship between viral HA sequences and antigenic distances in order to identify drift variants (Ren et al., 2015; Yao et al., 2017). However, these methods only consider viral HA sequences, but biological experiments have unveiled many factors crucial to viral antigenicity, such as five primary antigenic regions, glycosylation of HA and so on. Some existing methods have already paid attention to the features related to antigenic change (Du et al., 2012; Han et al., 2019; Peng et al., 2022a), but the selected features either still are not associated with viral sequences or the number of selected features is limited and not comprehensive enough.

Here, we propose a novel computational method, named MFPAD, that establishes a quantitative relationship between viral

sequences and antigenic distances while integrating four categories of features influencing viral antigenicity. The overview of MFPAD is shown in Figure 1. MFPAD significantly improves the accuracy of identifying antigenic variants and reduces the prediction error of antigenic distance. We apply MFPAD to the H3N2 influenza A virus and successfully present its antigenic evolution patterns, and further confirm the positive impact of the four categories of features on prediction accuracy.

2 Materials and methods

2.1 HA sequence and serologic data

The HA sequences of H3N2 influenza viruses are collected from the Influenza Virus Database (<https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database>). The serological data consists of two parts: (1) the first part contains cross-reactive HI antibody titers of 79 antisera with 270 H3N2 viruses isolated from 1968 to 2003, which is generated by Koel et al. (2013), (2) the second part contains cross-reactive HI titers of 173 serum with 1493 virus sequences isolated between 2003 and 2022 (collected from annual and interim reports from Worldwide Influenza Centre Lab, <https://www.crick.ac.uk/research/platforms-and-facilities/worldwide-influenza-centre/annual-and-interim-reports>).

2.2 MFPAD model

Based on the integration of information from the literature, we select four categories of features based on HA sequence that are closely related to antigenic change. We then develop a novel computational framework integrating the HA sequence information with 12 features and propose a model to quantitatively correlate HA sequences with antigenic distance, thereby improving the accuracy of antigenic variant recognition and reducing the prediction error of antigenic distance. The computational framework consists of three steps: feature representation based on HA sequences, multi-task HI matrix completion for antigenic distances, training and prediction of the XGBoost model.

2.2.1 Feature representation based on HA sequences

A pair of HA sequences (only HA1 sequences were used) is generally represented in two ways: one is binary representation, and the other is Pattern-Induced Multi-sequence Alignment (PIMA) scoring function (Smith and Smmith, 1992). Through comparison (see in Section 3.3), the binary method yields higher prediction accuracy, because of which we converted the HA sequence into the binary format. Based on features identified by various researchers that affect the antigenicity of influenza A viruses, a total of four categories of features are selected in this study. We calculate four categories of features between each pair of viruses based on the HA sequence. The primary category encompasses the number of substitutions in the viral HA sequences (Feature 1), since

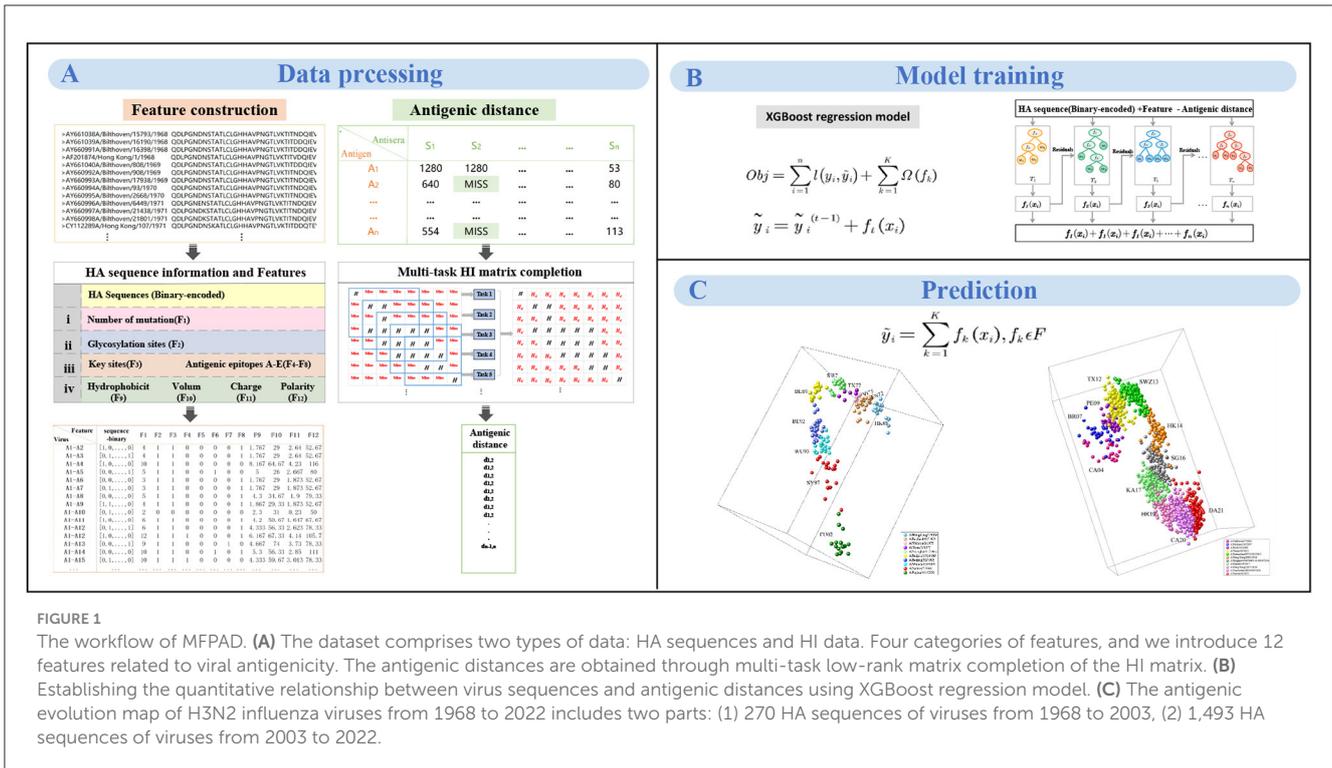


FIGURE 1 The workflow of MFPAD. (A) The dataset comprises two types of data: HA sequences and HI data. Four categories of features, and we introduce 12 features related to viral antigenicity. The antigenic distances are obtained through multi-task low-rank matrix completion of the HI matrix. (B) Establishing the quantitative relationship between virus sequences and antigenic distances using XGBoost regression model. (C) The antigenic evolution map of H3N2 influenza viruses from 1968 to 2022 includes two parts: (1) 270 HA sequences of viruses from 1968 to 2003, (2) 1,493 HA sequences of viruses from 2003 to 2022.

researchers have demonstrated that the number of mutations will substantially change the viral antigenicity (Lee and Chen, 2004). The second group pertains to the glycosylation sites (Feature 2), which was identified as the key factor affecting viral antigenicity (Wang et al., 2010; Tate et al., 2014; Hervé et al., 2015; Abdelwhab et al., 2016; Gu et al., 2019; York et al., 2019; Gao et al., 2021; Yin et al., 2021; Xu et al., 2022). The tertiary group integrates crucial antigenic positions (Feature 3) and substitutions within the five predominant antigenic regions designated as A, B, C, D, and E (Feature 4-8). Because many researchers have mapped the antigenic epitopes of H3N2 viruses and demonstrated that the viral antigenicity was mainly determined by five antigenic regions in the globular head of HA (Wiley et al., 1981; Tsuchiya et al., 2001; Hensley et al., 2009). More recently, Koel et al. and others further identified that key substitutions near the receptor-binding site in HA mainly determine the antigenic evolution of H3N2 viruses (Gerhard et al., 1981; Koel et al., 2013; Kong et al., 2021). The quaternary category embodies four intrinsic physicochemical properties of amino acids: hydrophobicity (Feature 9), volume (Feature 10), charge (Feature 11), and polarity (Feature 12), which significantly determine the antibody-protein interactions (Karadag et al., 2020). For a given pair of viruses i , feature j ($j = 1, \dots, N, N = 12$), N represents the number of features. f_{ij} represents feature j for a pair of virus i . For Feature 1, f_{ij} is calculated as the amino acid site mutations in the HA sequence of virus pair i . For Feature 2, we identify glycosylation sites in the HA sequence based on the NXT/S sequons (where X is any amino acid except proline), and then compare whether these sequons exist in the pair of virus i . If there is a glycosylation sequon, then $f_{ik} = 1$, otherwise $f_{ik} = 0$. For features 3-8, the calculation method for f_{ij} is similar to the calculation method for the Feature 2. If there is a mutation occurring at any site within the set, then $f_{ij} = 1$, otherwise $f_{ij} = 0$. For features 9-12, if

the number of amino acid mutations between virus pairs is less than 3, then f_{ij} is calculated as the mean difference in the quantitative values of the physicochemical properties of the mutation sites. If the number of mutations is greater than 3, only the 3 amino acid sites with the greatest differences in quantitative values are considered.

2.2.2 Multi-task HI matrix completion for antigen distances

The antigenic distances between viruses are used to measure the degree of antigenic difference between viruses. The antigenic difference is typically assessed and measured through HI assay. The HI matrix typically presents three different types of data: high-reacting values, low-reacting values, and missing values (Sun et al., 2013). After arranging the antigens and antisera in increasing order by year, the data within the HI matrix exhibits an overall banded distribution. The diagonal region primarily consists of high-reacting values and missing values, while others consist of low-reacting values and missing values. The completion of missing values in the HI matrix can be transformed into a low-rank matrix completion problem (Cai et al., 2010). To minimize the impact of low-reacting values, the completion task of the HI matrix is divided into multiple subtasks using a time sliding window (Cai et al., 2010; Sun et al., 2013). Due to the division of the task and the diversity of subsets, the amount of data and information involved varies from each subtask, resulting in different optimal ranks for each subtask. Nuclear norm regularization techniques can address the challenge of optimizing the general rank for all tasks (Jaggi et al., 2010; Han and Zhang, 2016).

Formally, given a $m \times n$ matrix A , The matrix completion

problem infers missing values and replaces low-reacting values with more confident values by solving the following optimization problem:

$$\min \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^n (H_{ij}^{\Omega} - A_{ij}^{\Omega})(H_{ij}^{\Omega} > \theta) + \lambda \|H\|_* \quad (1)$$

$$\|H\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i \quad (2)$$

where the matrix H is the estimate of A , Ω is a set of low-reacting values and determined values. In equation (1), θ is denoted as the threshold for low-reacting values. In equation (2), $\|H\|_*$ is the nuclear norm, which is the sum of all singular values of the matrix H . And λ is a regularization parameter that strikes a compromise between data fitting and matrix rank regularization.

The antigenic distances between viruses are derived from HI matrix (Cai et al., 2010). Each unit of antigenic distance is equivalent to $2\log_2(HI)$. In the antigenic map, 2 units of the antigenic distance represent 4-fold change in HI titer. This threshold is also used as a criterion for assessing virus variants (Smith et al., 1999). If the antigenic distances between viruses are greater than 2 units, they are considered antigenically different and belong to different antigenic cluster. Conversely, they are deemed antigenically similar and belong to the same antigenic cluster.

2.2.3 Training and prediction of the XGBoost model

The overall goal of this study is to develop an HA sequence based method to predict viral antigenicity. We employ the XGBoost regression model to establish a quantitative relationship between viral sequences and antigenic distances. XGBoost is a machine learning model based on ensemble principles and its core idea root in gradient boosting and decision trees (Chen and Guestrin, 2016).

For a given virus pair i , the feature vector is represented as x_i , and y_i represents its antigenic distance. The objective function of the XGBoost model is defined as follows:

$$Obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \Omega(f_t) + C \quad (3)$$

where i represents a pair of virus, n denotes the total number of virus pairs, $l(y_i, \hat{y}_i^{(t)})$ is the loss function, $\hat{y}_i^{(t)}$ represents the current prediction, and y_i represents the true value. In equation (3), $\Omega(f_t)$ represents the complexity of the t -th tree, and C represents a constant term. The prediction result for the t -th round is obtained by summing the prediction results of the previous subtrees as illustrated in equation (4):

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (4)$$

$\hat{y}_i^{(t-1)}$ represents the model predictions from the previous $t - 1$ rounds, and $f_t(x_i)$ represents the tree. The loss function use mean squared error:

$$l(y_i, \hat{y}_i^{(t)}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i^{(t)})^2 \quad (5)$$

$\Omega(f_t)$ represents the complexity of the t -th tree in equation (5), where complexity is defined as the sum of the number of leaf nodes and the square sum of the weights of all leaf nodes:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (6)$$

T represents the number of leaf nodes, γ denotes the difficulty of node splitting, λ is used to indicate the sparsity of L2 regularization, and w_j represents the weight of the leaf node j in equation (6).

3 Results

3.1 MFPAD predicts the antigenic evolution of H3N2 influenza viruses

By combining HA sequences of the influenza viruses with 12 relevant features affecting viral antigenicity, we establish a quantitative relationship with viral HA sequences and antigenic distances, thereby improving the accuracy of assessing viral antigenicity. The MFPAD is applied to gain insight into the antigenic evolution of H3N2 viruses. After collecting available cross-reactive HI data between H3N2 viruses and sera, we process the missing values in HI data by using a multi-task low-rank matrix completion method. In addition to using the binary representation of the viral HA sequence, we additionally select four categories of features, including a total of 12 features closely related to viral antigenicity. These features encompass the number of mutations, glycosylation sites, key antigenic positions and mutations in five antigenic regions, and four kinds of amino acid physicochemical properties (hydrophobicity, polarity, charge, volume). We utilize the XGBoost regression model to establish a quantitative relationship between viral sequences and antigenic distances. Ultimately, we apply MFPAD to H3N2 virus sequences to predict the antigenic evolution process of H3N2 influenza viruses.

In order to obtain the completion matrix with the smallest error, we choose a parameter $\lambda = 0.3$, and set the size of the time sliding window to 12 during the multi-task low-rank matrix completion process (Cai et al., 2010; Sun et al., 2013). Error calculation is carried out by randomly selecting 10% of the high-response values in the matrix for testing.

During the building of the XGBoost regression model, we employ a stratified sampling method to select 90 virus sequences from virus isolated from 1968 to 2003 and 502 viruses sequences from viruses isolated from 2003 to 2022 as training data. The antigenic evolution map includes all the viruses in the dataset. The model performance evaluation is carried out using ten-fold cross-validation, and model hyperparameters are optimized using grid search. The selected parameters are as follows: booster type is based on tree models, maximum tree depth is set to 7, learning rate is 0.1, the number of decision trees is 200, the minimum loss reduction for tree growth is 0.1, and the remaining parameters are default values. The model achieves a Root Mean Square Error (RMSE) of 0.314 in the virus dataset from 1968 to 2003 and 0.326 in the virus dataset from 2003 to 2022. When the antigenic distances between viruses exceeded 2, it is considered that the antigenic drift occurred. Based on this, 2 units distance are set as the threshold for antigenic

variants. MFPAD yields a high prediction accuracy of 0.948 in the virus dataset from 1968 to 2003 and 0.942 in the virus dataset from 2003 to 2022 for identifying virus variants. Furthermore, we utilize historical training data from 1968 to the target prediction year to assess our model's predictive accuracy for upcoming seasons. The average accuracy in predicting antigenic variants emerging in the coming year reaches 92.3%. Detailed information can be found in [Supplementary Tables S2, S3](#).

To evaluate the accuracy of MFPAD, the MFPAD is compared with other models, including two single-task models [Lasso (Cai et al., 2012) and Antigen-Bridges (Sun et al., 2013)], two multi-task models [GG-MTSL (Han et al., 2019) and MTL (Liu et al., 2012)], and the PREDAC model based on network built on virus similarity (Du et al., 2012). Among the compared models, MFPAD exhibit the smallest error and the highest accuracy in predicting antigenic variants, as shown in [Table 1](#) and [Figure 2](#). These results indicate that the quantitative relationship established in this study is significantly effective and accurate in predicting antigenic distance between viruses and identifying antigenic variants.

3.2 Antigenic evolution of H3N2 influenza viruses

We apply MFPAD to H3N2 influenza viruses, which includes (1) 270 HA sequences of virus from 1968 to 2003 and (2) 1,493 HA sequences of virus from 2003 to 2022. We predict their antigenic distances and create antigenic evolution maps through Multidimensional Scaling(MDS). Through the antigenic evolution maps, we are able to identify major antigenic clusters and pathways of antigenic evolution, gaining insights into the antigenic distances among different virus strains. We identify a total of 10 virus antigenic clusters for the period from 1968 to 2003 (HK68, EN72, VI75, TX77, SI87, BE89, BE92, WU95, SY97, FU02), as well as 11 virus antigenic clusters for the period from 2003 to 2022 (CA04, BR07, PE09, TX12, SWZ13, HK14, SG16, KA17, HK19, CA20, DA21), as shown in [Figures 3A, D](#). To validate the accuracy and reliability of the antigenic evolution maps generated based on the computational method proposed in this study, we compare them with antigenic evolution maps constructed by using HI data measured with various antisera, as shown in [Figures 3B, E](#). The results show that both maps exhibit similar evolutionary patterns, with each major predicted antigenic cluster matching that generated with serological data. Furthermore, We use RAXML to construct the phylogenetic tree of H3N2 viruses from 1968 to 2022, which is based on the Maximum Likelihood estimation method, as shown in [Figures 3C, F](#). During the tree construction process, certain parameter choices are necessary. Specifically, we choose the General Time Reversible (GTR) substitution model and the Subtree Pruning and Regrafting (SPR) tree topology search algorithm. Additionally, RAXML supports Bootstrap analysis, and we performed 1,000 replicates for the analysis. Our results exhibit consistency and similarity when compared to the evolutionary patterns presented in phylogenetic trees, thus validating the accuracy and reliability of our predicted results.

3.3 The impact of 12 features on model performance

In antigenic evolution studies of influenza viruses, viral sequences are commonly depicted using binary encoding or PIMA to quantify the genetic or antigenic distances between viruses. In this study, we investigate beyond solely considering virus sequences. Instead, we incorporate 12 features intrinsically linked to the antigenic evolution of influenza A viruses (Gerhard et al., 1981; Koel et al., 2013; York et al., 2019). Originating from viral HA sequences, these features are grouped into four categories. The first encompasses the number of amino acid mutations within the virus HA sequences. The second category of features include the glycosylation sites. The third category of features include the key antigenic positions, and the mutations in the five antigenic regions (A, B, C, D, and E). The fourth set of features focus on the physicochemical attributes of amino acids (hydrophobicity, volume, charge, and polarity). To substantiate the impact of these 12 features on enhancing the accuracy and reliability of antigenic evolution prediction, we undertake comparative tests. We conduct a comparative analysis between two methods: the first method involves training an XGBoost model utilizing solely the genetic distances represented by viral HA sequences, characterized through PIMA or binary encoding as features; the second method incorporates an additional 12 features into the model. Our experimental results demonstrate that the enriched model, supplemented with these 12 features, substantially outperform the model relying solely on PIMA or binary-encoded HA sequences. Notably, the inclusion of these 12 features markedly enhances the accuracy of predicting antigenic variants and also leads to a reduction in RMSE, as shown in [Figure 4](#).

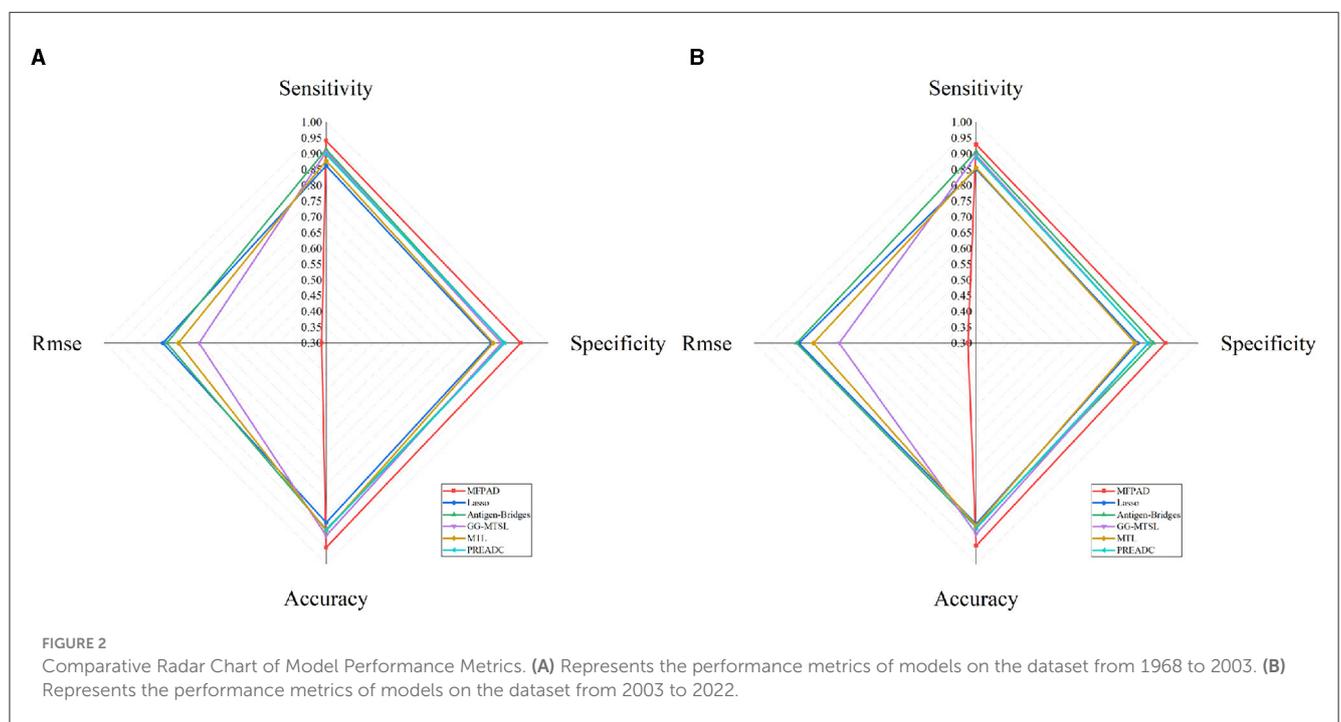
3.4 Changes in the 12 features during antigenic drift

To probe the significance of these supplementary features in the antigenic evolution of the H3N2 influenza viruses, we calculate the mean values of these 12 features between two adjacent antigenic clusters undergoing antigenic drift as well as within individual antigenic clusters. Compared to values derived from the individual antigenic clusters, the values of these features derived from two adjacent antigenic clusters display various levels of increase, as shown in [Figure 5](#). The accumulation of mutations leads to a higher probability of generating antigenic variants. In the virus dataset from 1968 to 2003, within one antigenic cluster, the value for virus mutations averages below 12, while in 66.7% of adjacent clusters, the value for virus mutations averages above 12. Glycosylation can either shield or reveal specific antigenic epitopes on HA proteins, thus plays an important role on viral antigenicity (York et al., 2019). A similar trend is observed for glycosylation sites, with 90% of individual antigenic clusters having mutation frequencies below 0.7, but in 66.7% of adjacent clusters, these frequencies are above 0.7. Studies have identified that there were five main antigenic regions (A-E) in HA (Wiley et al., 1981; Wilson et al., 1981), and the antigenicity of H3N2 viruses is mainly determined by key antigenic positions around the receptor binding site in HA

TABLE 1 Model performance comparison.

Model	Accuracy		RMSE		Sensitivity		Specificity	
	1968–2003	2003–2022	1968–2003	2003–2022	1968–2003	2003–2022	1968–2003	2003–2022
MFPAD	0.948	0.942	0.314	0.326	0.941	0.929	0.913	0.903
Lasso	0.869	0.873	0.815	0.859	0.862	0.851	0.822	0.809
Antigen-Bridges	0.891	0.883	0.803	0.866	0.914	0.907	0.861	0.859
GG-MTSL	0.911	0.905	0.701	0.732	0.906	0.895	0.853	0.842
MTL	0.895	0.879	0.766	0.813	0.878	0.855	0.827	0.801
PREADC	0.894	0.887	N/A	N/A	0.898	0.887	0.863	0.843

Performance of MFPAD, two single-task models (Lasso and Antigen-Bridges), two multi-task models (GG-MTSL and MTL), and the PREADC method. Root Mean Square Error (RMSE) represents the error in predicting antigenic distances, accuracy represents the performance of the model in identifying antigenic variants with a threshold of 2.



(Koel et al., 2013, 2014). In 80% of individual antigenic clusters, virus mutations at key antigenic positions remain below 0.8, while in 66.7% of neighboring clusters, they consistently exceed 0.8. Furthermore, mutations in the five major antigenic regions of the H3N2 viruses are closely associated with changes in antigenicity. In individual antigenic clusters, the virus mutation frequencies in antigenic region A are consistently below 0.8, while 66.7% of adjacent antigenic clusters have frequencies exceeding 0.8. For antigenic region B, 90% of viruses within antigenic clusters have mutation frequencies below 0.8, while in adjacent clusters, 88.9% are above 0.8, with 55.6% surpassing 0.9. Regarding antigenic region C, 60% of viruses within antigenic clusters have mutation frequencies below 0.3, while 44.4% of adjacent antigenic clusters have frequencies above 0.3. As for antigenic region D, in 90% of antigenic clusters, virus mutation frequencies are below 0.5, while 55.7% of adjacent antigenic clusters have frequencies above 0.5. Finally, for antigenic region E, in 70% of antigenic clusters,

virus mutation frequencies are below 0.2, while in 44.4% of adjacent antigenic clusters, they exceed 0.2. Overall, the mutation frequencies in antigenic regions C and E, as well as their variation within individual and adjacent antigenic clusters, are slightly lower than those in regions A, B, and D. In addition, we also examine four amino acid physicochemical properties (hydrophobicity, polarity, charge, and volume), which significantly determine the antibody-protein interactions (Karadag et al., 2020). In 80% of antigenic clusters, the differences in hydrophobicity between viruses are below 4, while in 77.8% of adjacent antigenic clusters, the differences in hydrophobicity are above 4. Regarding changes in amino acid volume, 80% of individual antigenic clusters remain below 55, while 77.8% of adjacent antigenic clusters exceed 55. Differences in amino acid charge are below 4 in 70% of individual antigenic clusters, but in 66.7% of adjacent antigenic clusters, they are above 4. Within individual antigenic clusters, differences in polarity between viruses remain below 75, while in 66.7% of

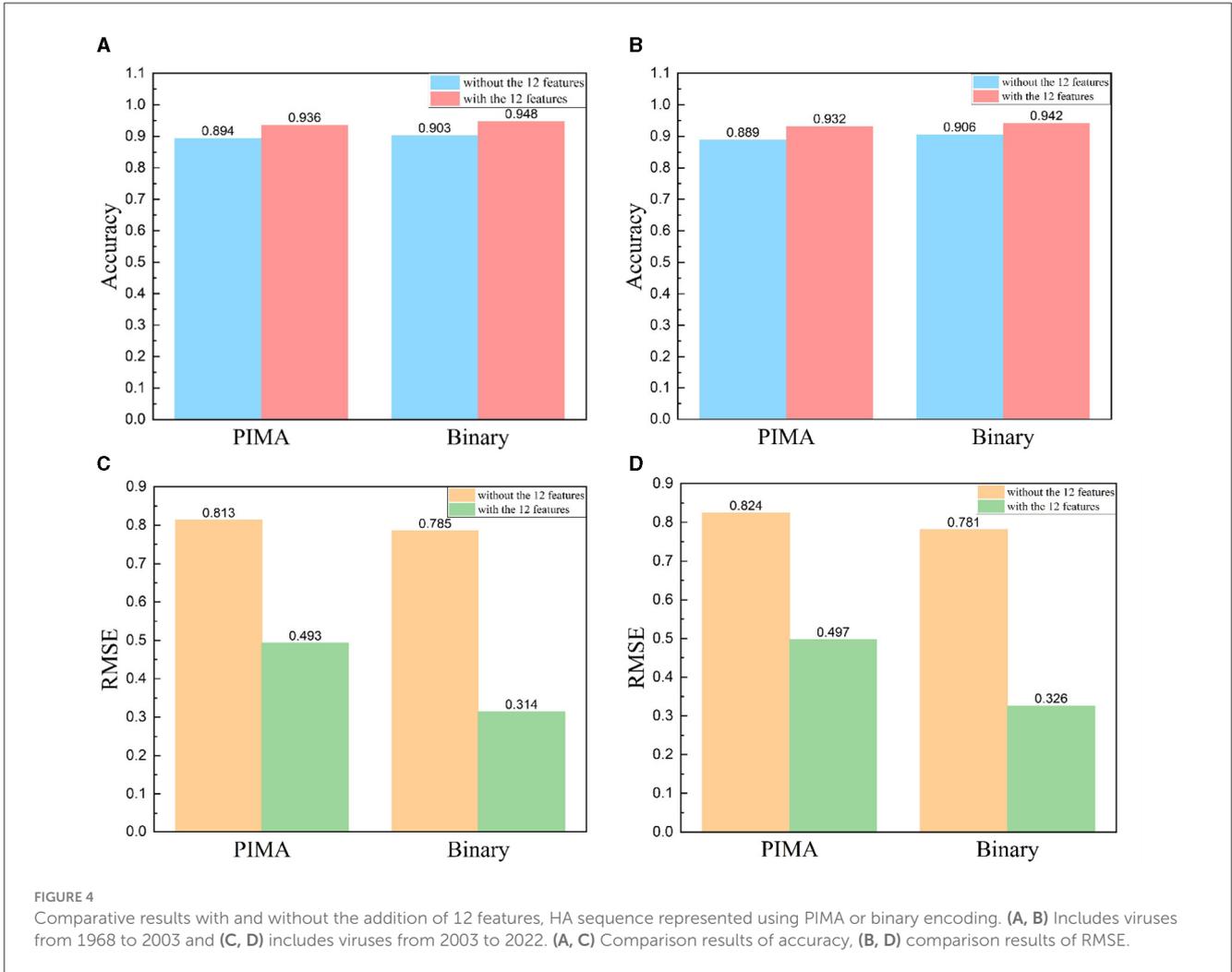


FIGURE 4 Comparative results with and without the addition of 12 features, HA sequence represented using PIMA or binary encoding. (A, B) Includes viruses from 1968 to 2003 and (C, D) includes viruses from 2003 to 2022. (A, C) Comparison results of accuracy, (B, D) comparison results of RMSE.

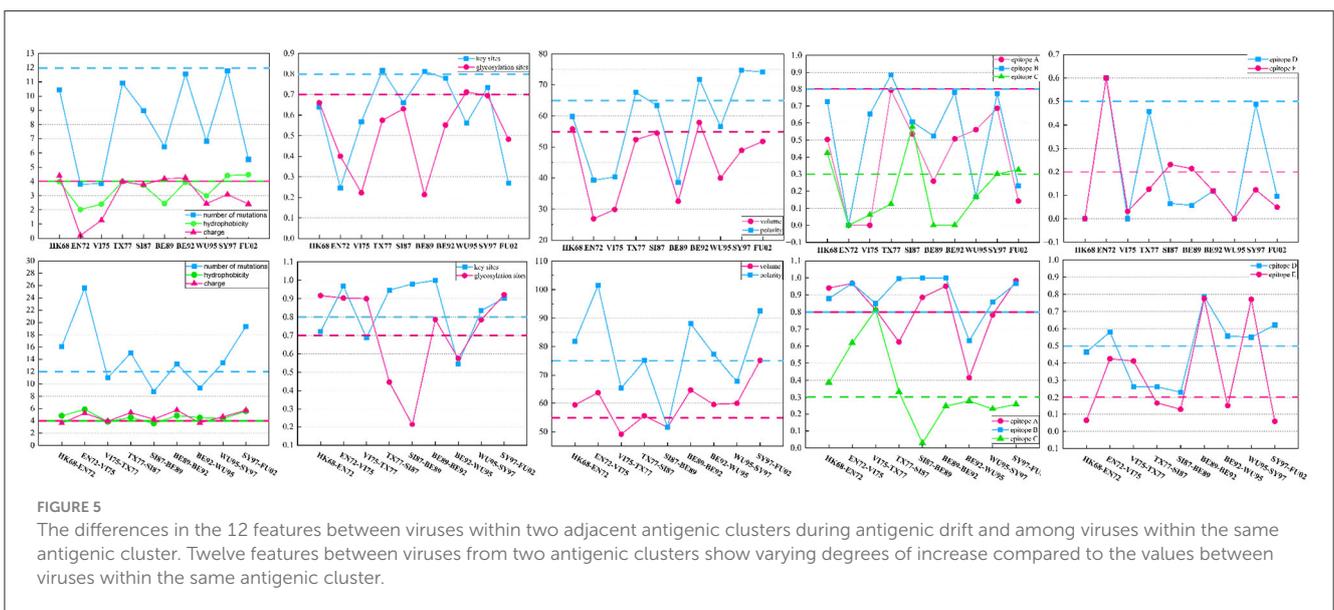
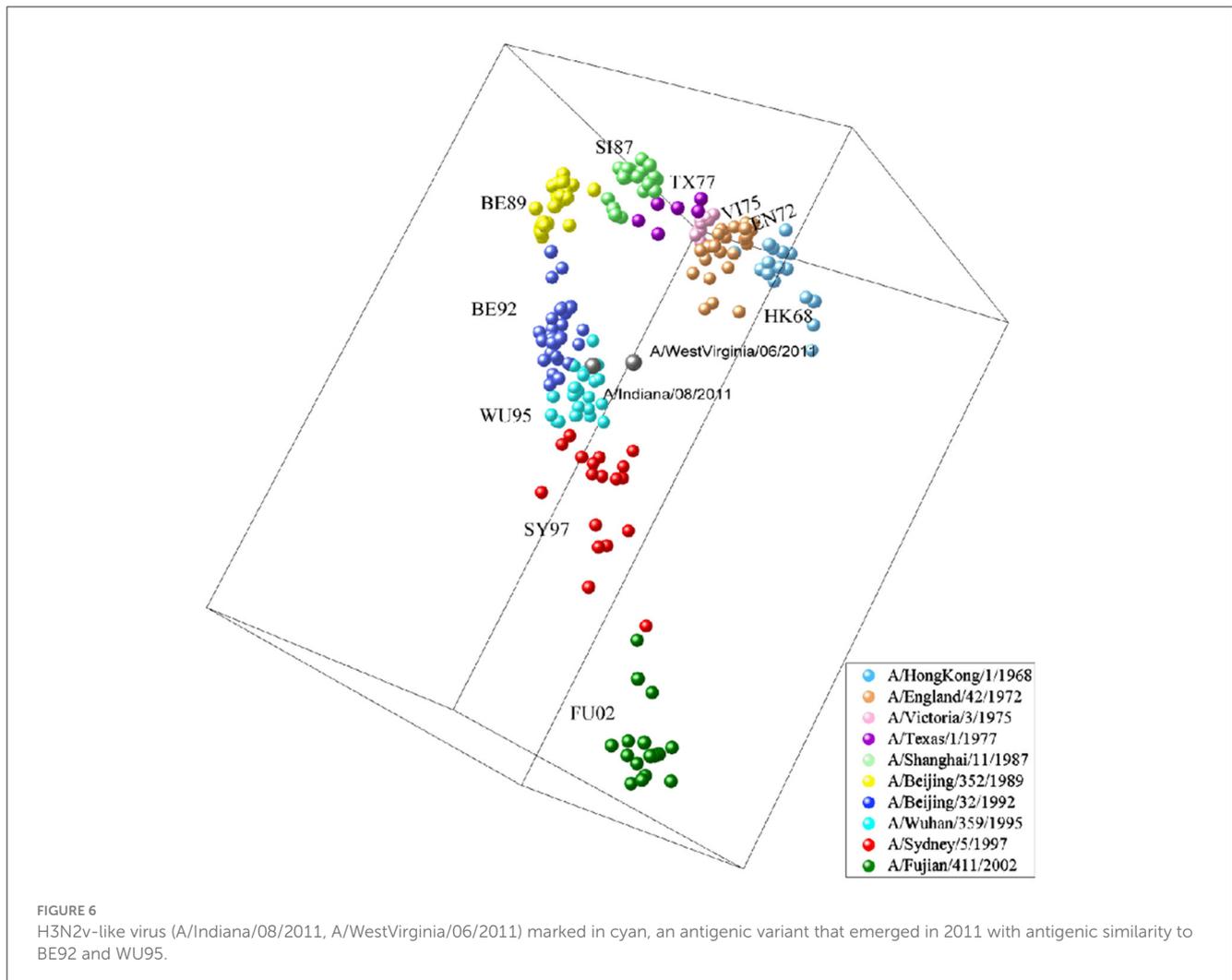


FIGURE 5 The differences in the 12 features between viruses within two adjacent antigenic clusters during antigenic drift and among viruses within the same antigenic cluster. Twelve features between viruses from two antigenic clusters show varying degrees of increase compared to the values between viruses within the same antigenic cluster.



4 Discussion

In this study, We propose a novel method for quantifying antigenic distances based on viral sequences. MFPAD not only takes into account the viral sequences but also integrates four categories of features related to antigenic change. Compared to previously established methods for predicting antigenic distances between viruses, MFPAD exhibits smaller errors in predicting antigenic distances between viruses and demonstrates higher accuracy in identifying antigenic variants.

With the development of high-throughput sequencing technologies, sequence data has become both cost-effective and rapid to obtain. Compared to hemagglutination inhibition (HI) data, sequence data is also more reliable and less susceptible to laboratory-specific variations. In this context, accurate computation of antigenic distances between viruses based on HA sequences has become critically important for virus classification, epidemiological investigation, and vaccine design. However, existing methods still have certain errors and limitations. Currently, research on the evolution of influenza viruses primarily focuses on viral sequence features, with a specific emphasis on amino acid mutation sites. For instance, applying regression models to construct HA sequences or identify key sites in

order to establish a quantitative relationship with antigenic distances (Sun et al., 2013; Ren et al., 2015; Yao et al., 2017; Han et al., 2019). Nevertheless, besides mutations in amino acid positions within viral sequences, there are multifaceted factors influencing antigenic evolution. Li et al. (2020) treat glycosylation sites as a separate category of features. Additionally, in the approach proposed by Du et al. (2012), features include various physicochemical properties of amino acids. These properties directly influence the structure, function, and stability of viral proteins. In this study, We integrate the viral sequence information along with categories of features related to antigenic change. These features include the number of sequence mutations, glycosylation sites, key antigenic positions, five major antigenic regions, and four physicochemical properties of amino acids, which encompass hydrophobicity, volume, charge, and polarity. By comprehensively considering these features, we can conduct a more comprehensive assessment of viral antigenic change, improve the accuracy of viral variant recognition, and reduce the error in predicting antigenic distances. When new antigenic variants emerge, MFPAD enables a rapid assessment of their antigenicity, determining their potential for epidemiological relevance and transmission risk. It can also provide essential guidance for vaccine preparation.

The 12 features in this study are closely associated with the antigenic evolution of H3N2 influenza virus. Among viruses from adjacent antigenic clusters, these 12 features exhibit varying degrees of increase in their values when compared to viruses within the same antigenic cluster. We observe that the frequency of mutations occurring in antigenic positions A, B, and D is higher in neighboring antigenic clusters compared to epitopes C and E. Particularly in antigenic epitope B, in nearly 90% of antigenic drift events, the mutation frequency between viruses exceeds 0.8, with over 50% of antigenic drift events even surpassing 0.9. Ndifon et al. (2009) also highlight that amino acid mutations occurring in antigenic epitopes with high antigenic efficiency (A, B, and D) exhibit a stronger correlation with viral antigenic drift when compared to epitopes with low to moderate efficiency (C and E). Notably, biological studies have demonstrated that antigenic region A and B play a major role to induce host immune response, and antigenic region B is immunodominant (Popova et al., 2012; Broecker et al., 2018; Wu et al., 2020). The consistency between the predicted result and biological findings further demonstrates that our model is feasible to predict the antigenicity of H3N2 viruses.

In this study, we apply MFPAD to predict the antigenic evolution of H3N2 influenza viruses based on HA sequences, including 270 sequences spanning from 1968 to 2003, and an additional 1493 sequences from 2003 to 2022. Our predictive analyses reveal the existence of 21 distinct major antigenic clusters, aligning closely with those identified through biological experiments. Although MFPAD is a robust system, our ability to represent the antigenic evolution of H3N2 viruses is somewhat limited. Rather than presenting it as a single integrated antigenic map, we are constrained to depict it in two separate maps. This limitation arises from the challenge of training the XGBoost model using HI data obtained from two independent sources, yielding a crowded and less reliable predicted map (data not shown). Several issues contribute to the challenges in processing and interpreting the HI data. Firstly, the HI data originates from different laboratories and are measured using varying protocols. Secondly, recent H3N2 viruses exhibit reduced binding to red blood cells (RBCs), necessitating a shift from turkey RBCs to guinea pig RBCs for HI assays (Lin et al., 2012). Thirdly, the surface protein NA of recent H3N2 viruses display RBC agglutination activity, which could introduce further variability into the HI data (Lin et al., 2010; Mögling et al., 2017). For assessing the antigenicity of influenza viruses, the gold standard is typically the Microneutralization assay (MN) or Focus Reduction Assay (FRA). Therefore, generating more reliable biological data, such as MN and FRA results, to train the prediction model could further enhance the accuracy of our predictive model.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

XL: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft, Writing – review & editing, Funding acquisition, Project administration, Resources, Supervision, Visualization. YL: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft, Writing – review & editing, Investigation. XS: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft. HK: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by the National Key Research and Development Program of China under Grant No. 2022YFD1801200, the National Natural Science Foundation of China under Grant No. 62202383, and the Fundamental Research Funds for the Central Universities under Grant No. D5000220129.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2024.1345794/full#supplementary-material>

References

- Abdelwhab, E.-S. M., Veits, J., Tauscher, K., Ziller, M., Grund, C., Hassan, M. K., et al. (2016). Progressive glycosylation of the haemagglutinin of avian influenza H5N1 modulates virus replication, virulence and chicken-to-chicken transmission without significant impact on antigenic drift. *J. General Virol.* 97, 3193–3204. doi: 10.1099/jgv.0.000648
- Biggerstaff, M., Reed, C., Epperson, S., Jhung, M. A., Gambhir, M., Bresee, J. S., et al. (2013). Estimates of the number of human infections with influenza a (H3N2) variant virus, United States, August 2011–April 2012. *Clin. Infect. Dis.* 57, S12–S15. doi: 10.1093/cid/cit273
- Broecker, F., Liu, S. T., Sun, W., Krammer, F., Simon, V., and Palese, P. (2018). Immunodominance of antigenic site b in the hemagglutinin of the current H3N2 influenza virus in humans and mice. *J. Virol.* 92, 10–1128. doi: 10.1128/JVI.01100-18
- Cai, Z., Ducatez, M. F., Yang, J., Zhang, T., Long, L.-P., Boon, A. C., et al. (2012). Identifying antigenicity-associated sites in highly pathogenic H5N1 influenza virus hemagglutinin by using sparse learning. *J. Mol. Biol.* 422, 145–155. doi: 10.1016/j.jmb.2012.05.011
- Cai, Z., Zhang, T., and Wan, X.-F. (2010). A computational framework for influenza antigenic cartography. *PLoS Comput. Biol.* 6, e1000949. doi: 10.1371/journal.pcbi.1000949
- Caton, A. J., Brownlee, G. G., Yewdell, J. W., and Gerhard, W. (1982). The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). *Cell* 31, 417–427. doi: 10.1016/0092-8674(82)90133-5
- Centers for Disease Control and Prevention (CDC) (2012). Evaluation of rapid influenza diagnostic tests for influenza a (H3N2) v virus and updated case count—United States, 2012. *MMWR Morb. Mortal. Wkly. Rep.* 61, 619–621.
- Chen, T., and Guestrin, C. (2016). “Xgboost: a scalable tree boosting system,” in *KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining San Francisco California USA August 13–17, 2016* (New York, NY: Association for Computing Machinery), 785–794.
- Du, X., Dong, L., Lan, Y., Peng, Y., Wu, A., Zhang, Y., et al. (2012). Mapping of H3N2 influenza antigenic evolution in china reveals a strategy for vaccine strain recommendation. *Nat. Commun.* 3, 709. doi: 10.1038/ncomms1710
- Feng, Z., Gomez, J., Bowman, A. S., Ye, J., Long, L.-P., Nelson, S. W., et al. (2013). Antigenic characterization of H3N2 influenza a viruses from Ohio agricultural fairs. *J. Virol.* 87, 7655–7667. doi: 10.1128/JVI.00804-13
- Gao, R., Gu, M., Shi, L., Liu, K., Li, X., Wang, X., et al. (2021). N-linked glycosylation at site 158 of the HA protein of H5N6 highly pathogenic avian influenza virus is important for viral biological properties and host immune responses. *Vet. Res.* 52, 1–14. doi: 10.1186/s13567-020-00879-6
- Gerhard, W., Yewdell, J., Frankel, M. E., and Webster, R. (1981). Antigenic structure of influenza virus haemagglutinin defined by hybridoma antibodies. *Nature* 290, 713–717. doi: 10.1038/290713a0
- Gu, C., Zeng, X., Song, Y., Li, Y., Liu, L., Kawaoka, Y., et al. (2019). Glycosylation and an amino acid insertion in the head of hemagglutinin independently affect the antigenic properties of H5N1 avian influenza viruses. *Sci. China Life Sci.* 62, 76–83. doi: 10.1007/s11427-018-9439-0
- Han, L., Li, L., Wen, F., Zhong, L., Zhang, T., and Wan, X.-F. (2019). Graph-guided multi-task sparse learning model: a method for identifying antigenic variants of influenza a (H3N2) virus. *Bioinformatics* 35, 77–87. doi: 10.1093/bioinformatics/bty457
- Han, L., and Zhang, Y. (2016). “Multi-stage multi-task learning with reduced rank,” in *30th Association-for-the-Advancement-of-Artificial-Intelligence (AAAI) Conference on Artificial Intelligence* (Phoenix, AZ: Assoc Advancement Artificial Intelligence), 30.
- Hensley, S. E., Das, S. R., Bailey, A. L., Schmidt, L. M., Hickman, H. D., Jayaraman, A., et al. (2009). Hemagglutinin receptor binding avidity drives influenza a virus antigenic drift. *Science* 326, 734–736. doi: 10.1126/science.1178258
- Hervé, P.-L., Lorin, V., Jouvion, G., Da Costa, B., and Escriou, N. (2015). Addition of n-glycosylation sites on the globular head of the H5 hemagglutinin induces the escape of highly pathogenic avian influenza a H5N1 viruses from vaccine-induced immunity. *Virology* 486, 134–145. doi: 10.1016/j.virol.2015.08.033
- Huang, J.-W., King, C.-C., and Yang, J.-M. (2009). Co-evolution positions and rules for antigenic variants of human influenza a/H3N2 viruses. *BMC Bioinform.* 10, 1–10. doi: 10.1186/1471-2105-10-S1-S41
- Iuliano, A. D., Roguski, K. M., Chang, H. H., Muscatello, D. J., Palekar, R., Tempia, S., et al. (2018). Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *Lancet* 391, 1285–1300. doi: 10.1016/S0140-6736(17)33293-2
- Jaggi, M., and Sulovsky, M. (2010). “A simple algorithm for nuclear norm regularized problems,” in *Proceedings of the 27th International Conference on Machine Learning (ICML) 2010* (Haifa), 471–478.
- Karadag, M., Arslan, M., Kaleli, N. E., and Kalyoncu, S. (2020). Physicochemical determinants of antibody-protein interactions. *Adv. Protein Chem. Struct. Biol.* 121, 85–114. doi: 10.1016/bs.apcsb.2019.08.011
- Kilbourne, E. D., Smith, C., Brett, I., Pokorny, B. A., Johansson, B., and Cox, N. (2002). The total influenza vaccine failure of 1947 revisited: major intrasubtypic antigenic change can explain failure of vaccine in a post-world war II epidemic. *Proc. Nat. Acad. Sci.* 99, 10748–10752. doi: 10.1073/pnas.162366899
- Koel, B. F., Burke, D. F., Bestebroer, T. M., Van Der Vliet, S., Zondag, G. C., Vervaeke, G., et al. (2013). Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science* 342, 976–979. doi: 10.1126/science.1244730
- Koel, B. F., van der Vliet, S., Burke, D. F., Bestebroer, T. M., Bharoto, E. E., Yasa, I. W. W., et al. (2014). Antigenic variation of clade 2.1 H5N1 virus is determined by a few amino acid substitutions immediately adjacent to the receptor binding site. *MBio* 5, 10–1128. doi: 10.1128/mBio.01645-14
- Kong, H., Burke, D. F., da Silva Lopes, T. J., Takada, K., Imai, M., Zhong, G., et al. (2021). Plasticity of the influenza virus H5 ha protein. *MBio* 12, 10–1128. doi: 10.1128/mBio.03324-20
- Lee, M.-S., and Chen, J. S.-E. (2004). Predicting antigenic variants of influenza a/H3N2 viruses. *Emerging Infect. Dis.* 10, 1385. doi: 10.3201/eid1008.040107
- Li, L., Chang, D., Han, L., Zhang, X., Zaia, J., and Wan, X.-F. (2020). Multi-task learning sparse group lasso: a method for quantifying antigenicity of influenza a (H1N1) virus using mutations and variations in glycosylation of hemagglutinin. *BMC Bioinform.* 21, 1–22. doi: 10.1186/s12859-020-3527-5
- Liao, Y.-C., Lee, M.-S., Ko, C.-Y., and Hsiung, C. A. (2008). Bioinformatics models for predicting antigenic variants of influenza a/H3N2 virus. *Bioinformatics* 24, 505–512. doi: 10.1093/bioinformatics/btm638
- Lin, Y. P., Gregory, V., Collins, P., Kloess, J., Wharton, S., Cattle, N., et al. (2010). Neuraminidase receptor binding variants of human influenza a (H3N2) viruses resulting from substitution of aspartic acid 151 in the catalytic site: a role in virus attachment? *J. Virol.* 84, 6769–6781. doi: 10.1128/JVI.00458-10
- Lin, Y. P., Xiong, X., Wharton, S. A., Martin, S. R., Coombs, P. J., Vachieri, S. G., et al. (2012). Evolution of the receptor binding properties of the influenza a (H3N2) hemagglutinin. *Proc. Nat. Acad. Sci.* 109, 21474–21479. doi: 10.1073/pnas.1218841110
- Liu, J., Ji, S., and Ye, J. (2012). Multi-task feature learning via efficient l2, 1-norm minimization. *arXiv [Preprint]*. arXiv: 1205.2631. doi: 10.48550/arXiv.1205.2631
- Mögling, R., Richard, M. J., Vliet, S., v. d., Beek, R., v., et al. (2017). Neuraminidase-mediated haemagglutination of recent human influenza a (H3N2) viruses is determined by arginine 150 flanking the neuraminidase catalytic site. *J. General Virol.* 98, 1274–1281. doi: 10.1099/jgv.0.000809
- Ndifon, W., Wingreen, N. S., and Levin, S. A. (2009). Differential neutralization efficiency of hemagglutinin epitopes, antibody interference, and the design of influenza vaccines. *Proc. Natl. Acad. Sci. U.S.A.* 106, 8701–8706. doi: 10.1073/pnas.0903427106
- Nelson, M. I., and Holmes, E. C. (2007). The evolution of epidemic influenza. *Nat. Rev. Genet.* 8, 196–205. doi: 10.1038/nrg2053
- Peng, W., Liu, H., Dai, W., Yu, N., and Wang, J. (2022a). Predicting cancer drug response using parallel heterogeneous graph convolutional networks with neighborhood interactions. *Bioinformatics* 38, 4546–4553. doi: 10.1093/bioinformatics/btac574
- Peng, W., Tang, Q., Dai, W., and Chen, T. (2022b). Improving cancer driver gene identification using multi-task learning on graph convolutional network. *Brief. Bioinform.* 23, bbab432. doi: 10.1093/bib/bbab432
- Popova, L., Smith, K., West, A. H., Wilson, P. C., James, J. A., Thompson, L. F., et al. (2012). Immunodominance of antigenic site B over site A of hemagglutinin of recent H3N2 influenza viruses. *PLoS ONE* 7, e41895. doi: 10.1371/journal.pone.0041895
- Ren, X., Li, Y., Liu, X., Shen, X., Gao, W., and Li, J. (2015). Computational identification of antigenicity-associated sites in the hemagglutinin protein of a/H1N1 seasonal influenza virus. *PLoS ONE* 10, e0126742. doi: 10.1371/journal.pone.0126742
- Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., et al. (2008). The global circulation of seasonal influenza a (H3N2) viruses. *Science* 320, 340–346. doi: 10.1126/science.1154137
- Smith, D. J., Forrest, S., Ackley, D. H., and Perelson, A. S. (1999). Variable efficacy of repeated annual influenza vaccination. *Proc. Nat. Acad. Sci.* 96, 14001–14006. doi: 10.1073/pnas.96.24.14001
- Smith, D. J., Lapedes, A. S., De Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D., et al. (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science* 305, 371–376. doi: 10.1126/science.1097211
- Smith, R. F., and Smmith, T. F. (1992). Pattern-induced multi-sequence alignment (PUMA) algorithm employing secondary structure-dependent gap penalties for

- use in comparative protein modelling. *Protein Eng. Design Select.* 5, 35–41. doi: 10.1093/protein/5.1.35
- Sun, H., Yang, J., Zhang, T., Long, L.-P., Jia, K., Yang, G., et al. (2013). Using sequence data to infer the antigenicity of influenza virus. *MBio* 4, 10–1128. doi: 10.1128/mBio.00230-13
- Tate, M. D., Job, E. R., Deng, Y.-M., Gunalan, V., Maurer-Stroh, S., and Reading, P. C. (2014). Playing hide and seek: how glycosylation of the influenza virus hemagglutinin can modulate the immune response to infection. *Viruses* 6, 1294–1316. doi: 10.3390/v6031294
- Tsuchiya, E., Sugawara, K., Hongo, S., Matsuzaki, Y., Muraki, Y., Li, Z.-N., et al. (2001). Antigenic structure of the haemagglutinin of human influenza A/H2N2 virus. *J. General Virol.* 82, 2475–2484. doi: 10.1099/0022-1317-82-10-2475
- Wang, W., Lu, B., Zhou, H., Suguitan Jr, A. L., Cheng, X., Subbarao, K., et al. (2010). Glycosylation at 158N of the hemagglutinin protein and receptor binding specificity synergistically affect the antigenicity and immunogenicity of a live attenuated H5N1 A/Vietnam/1203/2004 vaccine virus in ferrets. *J. Virol.* 84, 6570–6577. doi: 10.1128/JVI.00221-10
- Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M., and Kawaoka, Y. (1992). Evolution and ecology of influenza A viruses. *Microbiol. Rev.* 56, 152–179. doi: 10.1128/mr.56.1.152-179.1992
- Wiley, D., Wilson, I., and Skehel, J. (1981). Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature* 289, 373–378. doi: 10.1038/289373a0
- Wille, M., and Holmes, E. C. (2020). The ecology and evolution of influenza viruses. *Cold Spring Harb. Perspect. Med.* 10, 7. doi: 10.1101/cshperspect.a038489
- Wilson, I. A., and Cox, N. J. (1990). Structural basis of immune recognition of influenza virus hemagglutinin. *Annu. Rev. Immunol.* 8, 737–787. doi: 10.1146/annurev.iy.08.040190.003513
- Wilson, I. A., Skehel, J. J., and Wiley, D. (1981). Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. *Nature* 289, 366–373. doi: 10.1038/289366a0
- Wu, N. C., Otwinowski, J., Thompson, A. J., Nycholat, C. M., Nourmohammad, A., and Wilson, I. A. (2020). Major antigenic site B of human influenza H3N2 viruses has an evolving local fitness landscape. *Nat. Commun.* 11, 1233. doi: 10.1038/s41467-020-15102-5
- Xu, N., Wu, Y., Chen, Y., Li, Y., Yin, Y., Chen, S., et al. (2022). Emerging of H5N6 subtype influenza virus with 129-glycosylation site on hemagglutinin in poultry in china acquires immune pressure adaption. *Microbiol. Spect.* 10, e02537–02521. doi: 10.1128/spectrum.02537-21
- Yang, J., Zhang, T., and Wan, X.-F. (2014). Sequence-based antigenic change prediction by a sparse learning method incorporating co-evolutionary information. *PLoS ONE* 9, e106660. doi: 10.1371/journal.pone.0106660
- Yao, Y., Li, X., Liao, B., Huang, L., He, P., Wang, F., et al. (2017). Predicting influenza antigenicity from hemagglutinin sequence data based on a joint random forest method. *Sci. Rep.* 7, 1545. doi: 10.1038/s41598-017-01699-z
- Yin, X., Deng, G., Zeng, X., Cui, P., Hou, Y., Liu, Y., et al. (2021). Genetic and biological properties of H7N9 avian influenza viruses detected after application of the H7N9 poultry vaccine in china. *PLoS Pathog.* 17, e1009561. doi: 10.1371/journal.ppat.1009561
- York, I. A., Stevens, J., and Alymova, I. V. (2019). Influenza virus N-linked glycosylation and innate immunity. *Biosci. Rep.* 39, BSR20171505. doi: 10.1042/BSR20171505