CNNantigenic: predicting antigenic relationship of influenza A/H3N2 virus with convolutional neural network

Jing Meng[1,2#], Wenkai Song[3#], Jingze Liu[1,2#], Jiangyuan Wang [4], Le Zhang[3*], Taijiao Jiang[1,2,4*]

1 Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, China
2 Suzhou Institute of Systems Medicine, Suzhou 215123, China
3 College of Computer Science, Sichuan University, Chengdu 610065, China
4 Guangzhou Laboratory, Guangzhou 510005, China

#These authors contributed equally: Jing Meng, Wenkai Song, Jingze Liu.
*Correspondence and requests for materials should be addressed to Taijiao Jiang.

Introduction

The seasonal influenza virus causes millions of infections and hundreds of thousands of deaths (mainly the elderly and children) worldwide each year [1]. At present, the most effective and economical influenza prevention and control strategy is vaccination [2]. Influenza vaccines mainly compose of the surface antigen of hemagglutinin (HA) [3]. However, HA is susceptible to mutations, resulting in antigenic drift or shift [4]. If the antigenic distance between the circulating strain and the vaccine strain is distant, vaccine will not be able to play a good protective role [5]. Therefore, predicting the antigenic relationship of influenza viruses is a critical step in the vaccine strain recommendation process [6].

Experimental characterization of antigenic relationship relies mainly on hemagglutination inhibition (HI) assay [7]. The HI assay requires tissue culture for a specific virus and preparation of a specific serum, which is time-consuming and labor-intensive. Due to the development of high-throughput sequencing technologies, rapid determinations of whole genomes of influenza viruses are feasible [8]. Great efforts have been made towards constructing sequence-based computational models to predict antigenic relationship of influenza viruses [9].

Smith *et al*. employed a modification of metric multidimensional scaling (MDS) to determine the antigenic evolution and create an antigenic map of influenza A/H3N2 virus from 1968 to 2003 [10]. Since then, machine learning techniques have been used to quantify antigenic relationships. Lee *et al.* proposed that the extent of antigenic changes was associated with the number of mutated sites in the protein HA, and identified the best model that was based on the amino acid changes in the five epitopes [11]. Liao *et al.* measured the amino acid differences in terms of polarity, charge and structure, and evaluated the antigenic distances with a combination of scoring and regression methods [12]. Du *et al.* developed a Naïve Bayes classifier that integrated the differences of the structural and physicochemical features of the HA, including the number of mutations on each epitope, the five physicochemical properties of the amino acids on the surface residues, the changes of amino acids in the receptor binding region and at glycosylation sites [13]. Qiu *et al*. incorporated the structural information of HA into a linear model to predict the antigenic relationship of influenza A/H3N2 [14]. These methods account for only limited amino acids, and do not scale to weight the combinatorial effects of point mutations in the HA protein on the

antigenicity.

Given that deep neural networks can model the dependences in the context sequence, they have been becoming popular in bioinformatics [15-17]. Rui *et al.* designed IAV-CNN that was based on a 2D convolutional neural network to identify influenza antigenic variants [18]. IAV-CNN took as input ProtVec that was a distributed representation of amino acids, then the splittings and embeddings were employed to divide the HA1 sequence into shifted overlapping residue fragments by a window size of 3 and stride size of 1. The operation of splittings and embeddings broke the spatial structure of the HA1 sequence, which may lead the CNN to extracting misleading local features. Also, the distributed representations increased the model parameters, which may result in insufficient quantity of training samples.

To address the shortcomes of currently available approaches, we proposed CNNantigenic to predict the antigenic relationship of influenza A/H3N2 virus. CNNantigenic constructs a spatially oriented representation of the HA1 sequence adapted for the convolutional architecture, which enables it to explore the interactions of the amino acids in the context sequence. Moreover, instead of the redundant amino acid embeddings, CNNantigenic takes into consideration only physicochemical features determining antigenicity of influenza A/H3N2 virus. Together, CNNantigenic can effectively extract the features in the context sequence from local to global views, and investigate the combinatorial contributions of point mutations in the HA protein to the antigenicity. The 5-fold cross validations and retrospective tests show that CNNantigenic achieves significantly better performances than its competitors.

Materials and Methods

Datasets

The HA1 sequences of influenza A/H3N2 were downloaded from Influenza Research Database [19], Global Initiative on Sharing All Influenza Data [20] and NCBI Influenza Virus Database [21]. The HI assay values were collected from international organizations and publications, including World Health Organization, European Centre for Disease Prevention and Control, U.S. Food and Drug Administration (Data Collection of Supplementary Material). We did sequence deduplication, and kept only pairs of strains with no more than 18 amino acid substitutions. As a result, a total of 3671 strain pairs from 1968 to 2020 were obtained, which constituted the complete antigenic relations in this study.

Characterization of antigenic relationships of strain pairs

The antigenic distance D_{ij} between paired strains is represented by Archetti-Horsfall distance [22] as follows:

$$D_{ij} = \sqrt{\frac{H_{ii} \times H_{jj}}{H_{ij} \times H_{ji}}} \qquad (1),$$

where H_{ij} is the HI titer of strain i necessary to inhibit cell agglutination caused by strain j. Two strains are considered as antigenic variants when the D_{ij} is equal or greater than 4. Otherwise, the pair is treated as antigenic similar [12]. As a result, among the 3671 strain pairs, 525 and 3146 were defined as antigenic distant and antigenic similar, respectively.

Workflow of CNNantigenic

The workflow of CNNantigenic is shown in Figure 1. CNNantigenic takes as input the paired HA1 sequences of influenza A/H3N2. For training, the antigenic relations are constructed by incorporating the paired HA1 sequences and their corresponding antigenic relationship. Then CNNantigenic performs the mapping by encoding the chosen physicochemical features according to the feature dictionary. Next, CNNantigenic generates the input matrix corresponding to the paired HA1sequences (sequence 1 and sequence 2). The input matrix is a spatial representation of the HA1 sequences. The mapping of the physicochemical features of sequence 2 is appended to that of sequence 1. To overcome the drawback of the small size of the antigenic relations, CNNantigenic doubles the input matrix for each HA1 sequence pair by data amplification. The data amplification is implemented by exchanging the mappings of the physicochemical features of the paired HA1 sequences. Finally, CNNantigenic evaluates the information of the input matrix to infer the antigenic relationship of the paired strains of influenza A/H3N2 by the CNN model. The following provides the details of how CNNantigenic works.

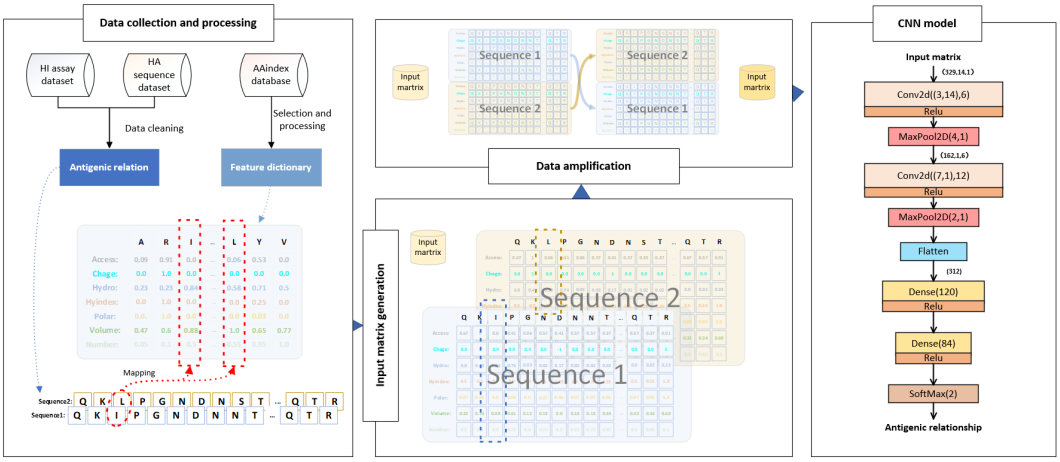

Figure1. Overview of CNNantigenic. CNNantigenic first takes as input the paired HA1 sequences of influenza A/H3N2. Next, CNNantigenic generates the input matrix corresponding to the paired HA1sequences (sequence 1 and sequence 2) by encoding the chosen physicochemical features according to the feature dictionary. Then, CNNantigenic doubles the input matrix for each HA1 sequence pair by data amplification to overcome the small sized antigenic relations. Finally, CNNantigenic evaluates the information of the input matrix to infer the antigenic relationship of the paired strains of influenza A/H3N2 by the CNN model.

Input matrix generation

The antigenicity depends on the interactions of the amino acids. Here we chose six antigenicity determining physicochemical properties of amino acids, including accessible surface, charge, hydrophobicity, hyindex (the number of hydrogen bond donors), polarity and volume[13]. There are multiple entries corresponding to each of the six physicochemical properties in the AAindex1 database [23, 24]. To scale the values to be in the interval [-1, 1], we processed the values in each entry for the six properties with the following formula

$$X_{norm} = \begin{cases} \dfrac{X_i}{\max\left(|X_{min}|, |X_{max}|\right)} & if(X_{min} < 0) \\ \dfrac{X_i - X_{min}}{X_{max} - X_{min}} & if(X_{min} \geq 0) \end{cases} \quad (3).$$

Then, to choose the optimal entry, we constructed a matrix of size 329*4 for each entry of each property, and chose the entry with the highest F1-score on the validations (Supplementary Figure S1). For example, for the property of volume with 12 entries, volume_1 was selected as the optimal entry. The feature dictionary contains the six physicochemical features with the optimal entry for 20 amino acids (Supplementary Table S1).

Next, CNNantigenic incorporates and encodes the six physicochemical features for the HA1 sequence into input matrix to feed into the CNN model. In order to fully explore the interactions of the amino acid sites in the context sequence, the input matrix for the paired HA1 sequences (sequence 1 and sequence 2) is constructed according to the feature dictionary. It is a spatially oriented representation of the HA1 sequence. In total, there are 14 rows (7 rows for each sequence) and 329 columns in the input matrix. One column represents one amino acid site in the HA1 sequence. In each column, the first and last 7 rows display the amino acid covering the site and its six physicochemical features of sequence 1 and sequence 2, respectively. For each HA1 sequence pair, CNNantigenic creates two input matrices by data amplification for training. The second input matrix is generated by appending the mappings of the physicochemical features of sequence 1 to that of sequence 2.

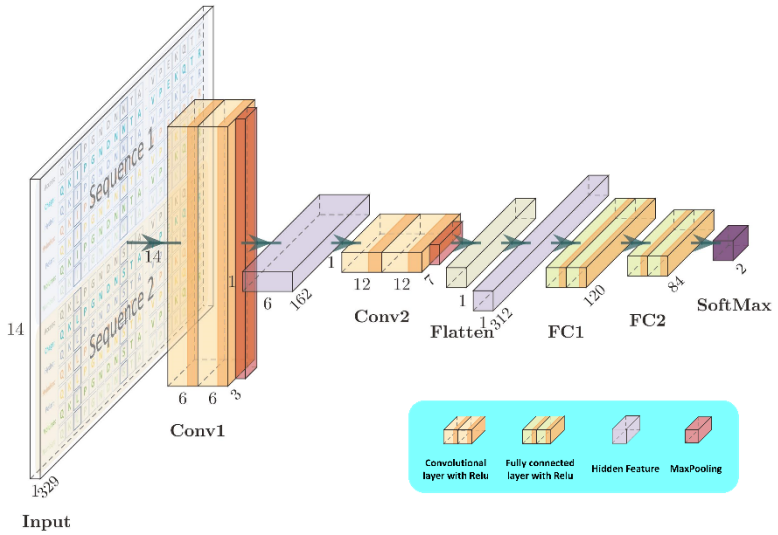

Figure2. The CNN model structure. The input matrix is of size 329*14*1 corresponding to a strain pair. The CNN model consists of two 2D convolution blocks (Conv1 and Conv2), two full connection blocks (FC1 and FC2), and a classification layer (SoftMax). The first convolution kernel of Conv1 has a size of 3*14*6. The second convolution kernel of Conv2 is of size 7*1*12. The classification layer infers the antigenic relationship (antigenic distant or antigenic similar) of a strain pair.

The CNN model

The input matrix is fed into the CNN model to infer the antigenic relationship of a strain pair. The

model structure is shown in Figure 2. The CNN model consists of two 2D convolution blocks (Conv1 and Conv2), two full connection blocks (FC1 and FC2), and a classification layer (SoftMax). The convolution block includes the convolutional layer, the activation function (Relu) and maxpooling. The full connection block includes the fully connected layer and the activation function.

To apply to the convolution operation, the input matrix is converted from 329*14 to 329*14*1. The first convolution kernel of Conv1 has a size of 3*14*6. The first dimension with size 3 allows the convolution implementation to extract features from both upstream and downstream amino acid sites around a site in the context sequence. The second dimension with size 14 is chosen, as there is no spatial representation of the mappings of the physicochemical features of a HA1 sequence pair. The third dimension with size 6 enriches the feature representations of the mappings[25]. Maxpooling follows to extract the informative mappings of the physicochemical features in small spatial regions. As a result, the hidden feature matrix of size 162*1*6 is obtained. Next, the second convolution kernel of Conv2 of size 7*1*12 is employed to further extract features in the context sequence. Finally, according to the abstract feature representations from the previous layers, the classification is performed by the classification layer to infer the antigenic relationship of a strain pair.

Experimental design

Firstly, in order to verify the importance of data amplification, we trained a new model CNNantigentic-single. It has the same parameters as the CNNantigentic model, but its dataset is the original dataset without data amplification. Both models use Adam as the optimizer, the initial learning rate is 0.001, and two hundred rounds of training are performed. At the same time, when the loss is no longer reduced, the size of the learning rate is dynamically reduced by half.

To evaluate the performance of our CNNantigentic model, compare it with several state-of-the-art approaches. Liao et al. proposed a method for predicting antigenic variation by combining scoring and regression methods[12]. Du et al. developed a Naive Bayes classifier to integrate the structural and physicochemical features of HA[13]. Lees et al. provide prediction methods based on five antigenic sites and increase additional assignments to establish five canonical regions [26].

To evaluate the generalization and predictability of the model, we will conduct five-fold crossover validation and retrospective testing, respectively. The five-fold crossover validation only needs to randomly divide the data set into five folds, and select one part as the test set and the rest as the training set in the five training process. Retrospective testing requires dividing the data by different years, and each HA pairs contains two years, so we use formula (2) to determine the year of each HA pairs.

$$HA\ pair_{year} = \max\left(Sequence1_{year}, Sequence2_{year}\right) \qquad (2)$$

In each simulate year (X), the year corresponding to X is used as the testing data, and collected from 1968 to X-1 as the training data. In this study, we choose a total of fifteen years from 2006 to 2020(X=2006, 2007, …, 2020). In order to compare the differences between models, we divided the results of the fifteen retrospective testing into four groups for comparison ([2006, 2008], [2009, 2012], [2013, 2016], [2017, 2020]). In this study, all the results will be presented in these four groups.

All the approaches are implemented with Scikit-learn [27] and Tensorflow[28]. The antigenic distinct is labeled as '1' and antigenic similar is '0' for the relationship of two strains. In order to evaluate the effect difference between CNNantigentic and its peers, we select the evaluation indicators as accuracy and Matthews's Correlation Coefficient (MCC), intending to measure the overall accuracy of the model and solve the classification evaluation problem of two-type imbalanced datasets. Finally, we will perform five-fold cross-validation and retrospective testing to measure the robustness and generalization of the model, respectively.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \qquad (4)$$

$$MCC = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \qquad (5)$$

4 Result

4.1 Five-fold cross-validation results

Five-fold cross-validation is a process when all data is randomly split into 5 folds, and then the model is trained on the 4 folds, while one fold is left to test the model. It can test the performance of the model in order to show problems such as overfitting and selection bias. We calculated the training and testing for five times, and finally took the average of the accuracy and MCC of each round, and the results are shown in Figure 3. The accuracy and MCC values of CNNantigentic and CNNantigentic-single are higher than other models. It can be observed that the result of CNNantigentic-single is slightly lower than that of CNNantigentic, and the difference between the two models is only whether data amplification is performed. Among them, the MCC value of IAV-CNN is 0, mainly because the model cannot be fully trained in the case of a small amount of data.

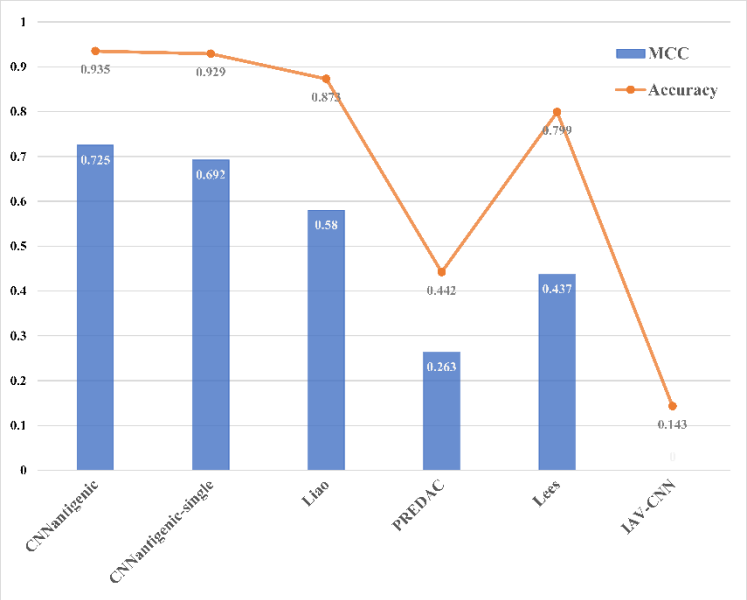

Figure3. Five-fold cross-validation results

4.2 Retrospective testing Results

Since the A/H3N2 virus evolves year by year, the site and antigenic epitope will change accordingly. In order to more effectively measure whether the model can capture the mutation relationship of the virus over time, we additionally use retrospective testing to evaluate the model (Method 2.3). We selected X year as testing data and 1968 to X-1 year as the training data, and conducted retrospective testing from 2006 to 2020. In order to show the results, we divided the results of fifteen years into four groups and results for each group were calculated separately for accuracy and MCC. The results shown in Figure 4.

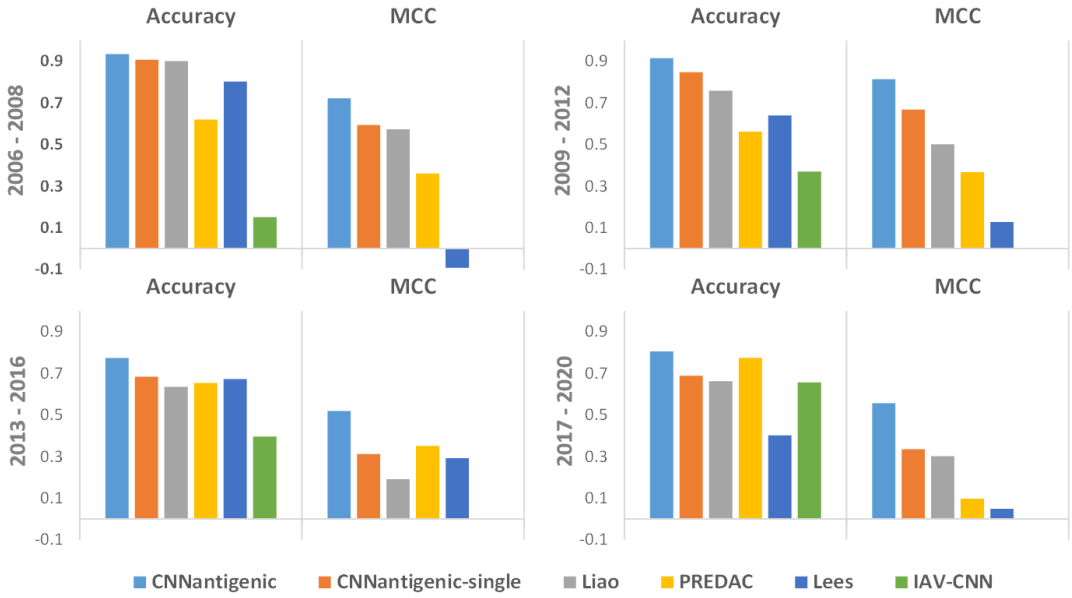

Figure4. Retrospective testing Results

Among the four groups, CNNantigentic scored the highest. CNNantigentic-single may be lower than some traditional methods without data amplification. For example, in group of 2017-2020, the accuracy of CNNantigentic-single is lower than that of PREDAC but MCC is higher than it. It shows that the training effect of the model without data expansion will indeed decline, but the stability of the MCC value also shows that the model itself can be effectively applied to the antigen prediction problem of H3N2.

4.3 ROC curve

All models will output the predicted probabilities for the two classes, and finally select the class corresponding to the largest predicted probability value as the result. Therefore, we also selected ROC curves and calculated AUC values to assess the specific performance of each model. An ROC curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters : TPR , FPR. An ROC curve plots TPR vs. FPR at different classification thresholds. The AUC stand for area under the ROC Curve, which provides a comprehensive performance measure for all possible classification thresholds. The advantage of AUC compared to MCC is that the classification threshold is unchanged. Regardless of the classification threshold chosen, it measures the quality of the model's predictions

$$TPR = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad (6)$$

$$FPR = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \qquad (7)$$

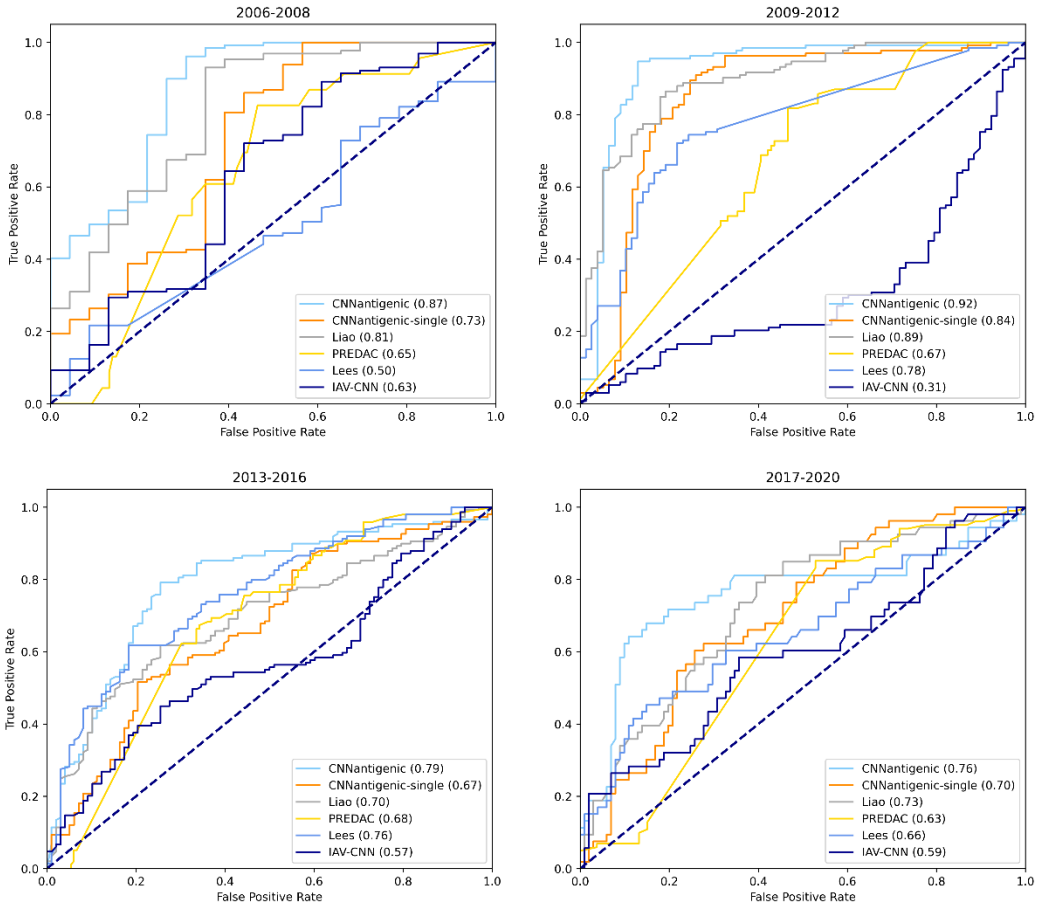

Figure5. ROC curve results The lower right corner indicates the color corresponding to each method, and the corresponding AUC values are indicated in parentheses.

Correspondingly, we still draw the ROC curves of the corresponding four sets of results from 2006 to 2020. As shown in Figure 5, the ROC results of the four groups of retrospective analysis experiments show that the AUC values of CNNantigentic are the highest, indicating that our model can achieve better results no matter what classification threshold is selected. For the ROC curve, our model curve is also the closest to the upper left corner and obtains the best performance evaluation results. Similarly, the effect of data augmentation on the model can be seen in the ROC. This shows that our model can clearly classify antigenic variation and antigenic similarity, and can capture the corresponding evolutionary laws over time.

4 Discution

In this paper, we propose a CNN model based on the physical and chemical properties of amino acids based on the idea of data augmentation commonly used in the image field to solve the shortcomings of traditional methods such as single matrix design and insufficient data volume, which lead to the ineffective training of the model. We refer to the research of related papers and

choose accesible, charge, hydrophobicity, hyindex(number of hydrogen bond donors), polarity and volume as the representative physicochemical properties of amino acids. After relevant processing, a Feature dictionary is constructed to expand the sequence pair into a 329*14 input matrix. Using the idea of data augmentation, the data volume is doubled by changing the positions of Sequence1 and Sequence2..This fully shows that in the direction of deep learning, the amount of data is still very important. CNNantigentic uses data amplification to effectively improve the performance of the model, effectively utilize limited data, and help the model to effectively extract features. Therefore, compared with the traditional model, we extend the sequence information, use the amino acid physicochemical properties to help the model predict antigenic relation, and we perform feature extraction for every three sites and scan the entire sequence instead of part of the sequence. Compared with other more complex encoding methods, our matrix can be as small as possible while ensuring that all sequences exist, to ensure that the parameters can be fully trained.

However, we can see that although the model obtained a good score in the five-fold cross-check, in the retrospective analysis, the accuracy of the results of the latter two groups (2013-2016, 2017-2020) was around 0.8, which was relatively high. Big room for improvement. It shows that the model still needs more improvement for some evolution problems of viruses.

References

1. WHO. *Influenza fact sheet*. 2018; Available from: https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal).
2. Virelizier, J.-L.J.T.J.o.I., *Host defenses against influenza virus: the role of anti-hemagglutinin antibody.* 1975. **115**(2): p. 434-439.
3. Bouvier, N.M. and P.J.V. Palese, *The biology of influenza viruses.* 2008. **26**: p. D49-D53.
4. Nakaya, H.I., et al., *Systems biology of vaccination for seasonal influenza in humans.* 2011. **12**(8): p. 786-795.
5. Stöhr, K.J.T.L.i.d., *Influenza—WHO cares.* 2002. **2**(9): p. 517.
6. Houser, K., K.J.C.h. Subbarao, and microbe, *Influenza vaccines: challenges and solutions.* 2015. **17**(3): p. 295-300.
7. Pedersen, J.C., *Hemagglutination-inhibition assay for influenza virus subtype identification and the detection and quantitation of serum antibodies to influenza virus*, in *Animal influenza virus*. 2014, Springer. p. 11-25.
8. Reuter, J.A., D.V. Spacek, and M.P.J.M.c. Snyder, *High-throughput sequencing technologies.* 2015. **58**(4): p. 586-597.
9. Klingen, T.R., et al., *In silico vaccine strain prediction for human influenza viruses.* 2018. **26**(2): p. 119-131.
10. Smith, D.J., et al., *Mapping the antigenic and genetic evolution of influenza virus.* 2004. **305**(5682): p. 371-376.
11. Lee, M.-S. and J.S.-E.J.E.i.d. Chen, *Predicting antigenic variants of influenza A/H3N2 viruses.* 2004. **10**(8): p. 1385.
12. Liao, Y.-C., et al., *Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus.* 2008. **24**(4): p. 505-512.
13. Du, X., et al., *Mapping of H3N2 influenza antigenic evolution in China reveals a strategy for vaccine strain recommendation.* 2012. **3**(1): p. 1-9.

14. Qiu, J., et al., *Incorporating structure context of HA protein to improve antigenicity calculation for influenza virus A/H3N2.* 2016. **6**(1): p. 1-9.
15. Min, S., B. Lee, and S.J.B.i.b. Yoon, *Deep learning in bioinformatics*. 2017. **18**(5): p. 851-869.
16. Li, Y., et al., *Deep learning in bioinformatics: Introduction, application, and perspective in the big data era.* 2019. **166**: p. 4-21.
17. Meng, J., et al., *DeepSSV: detecting somatic small variants in paired tumor and normal sequencing data with convolutional neural network*. 2021. **22**(4): p. bbaa272.
18. Yin, R., et al., *IAV-CNN: a 2D convolutional neural network model to predict antigenic variants of influenza A virus.* 2021.
19. Squires, R.B., et al., *Influenza research database: an integrated bioinformatics resource for influenza research and surveillance.* 2012. **6**(6): p. 404-416.
20. Shu, Y. and J.J.E. McCauley, *GISAID: Global initiative on sharing all influenza data–from vision to reality.* 2017. **22**(13): p. 30494.
21. Bao, Y., et al., *The influenza virus resource at the National Center for Biotechnology Information.* 2008. **82**(2): p. 596-601.
22. Ndifon, W., J. Dushoff, and S.A.J.V. Levin, *On the use of hemagglutination-inhibition for influenza surveillance: surveillance data are predictive of influenza vaccine effectiveness.* 2009. **27**(18): p. 2447-2452.
23. Kawashima, S. and M.J.N.a.r. Kanehisa, *AAindex: amino acid index database*. 2000. **28**(1): p. 374-374.
24. Wu, A., et al., *Correlation of influenza virus excess mortality with antigenic variation: application to rapid estimation of influenza mortality burden.* 2010. **6**(8): p. e1000882.
25. Liu, Z., et al. *A convnet for the 2020s*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
26. Lees, W.D., D.S. Moss, and A.J.J.B. Shepherd, *A computational analysis of the antigenic properties of haemagglutinin in influenza A H3N2.* 2010. **26**(11): p. 1403-1408.
27. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* 2011. **12**: p. 2825-2830.
28. Abadi, M., et al. *{TensorFlow}: a system for {Large-Scale} machine learning*. in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*. 2016.