


RESEARCH

Open Access



# InflANNet: a neural network predictor for Influenza A CTL and HTL epitopes to aid robust vaccine design

R. Karthika<sup>1</sup>, Sathya Muthusamy<sup>1</sup> and Prince R. Prabhu<sup>1,2,3\*</sup> 

## Abstract

**Background** An efficient and reliable data-driven method is essential to aid robust vaccine design, particularly in the case of an epidemic like Influenza A. Although various prediction tools are existing, most of them focus on the MHC-peptide binding affinity predictions. A tool which can incorporate more features other than binding affinity which characterizes the T-cell epitopes as vaccine candidates would be of much value in this scenario. The objective of this study is to develop two separate neural network models for the predictions of CTLs (cytotoxic T lymphocyte) and HTLs (helper T lymphocyte) with the manually curated datasets as a part of this study from the raw viral sequences of Influenza A.

**Results** The epitope datasets curated from the raw sequences of the broadly protective Neuraminidase protein were utilized for building and training the models for CTLs and HTLs. Each set consisted of nearly a balanced mix of vaccine candidates and non-vaccine candidates for both CTLs and HTLs. These were fed to neural networks as they are proven to be powerful for the predictions when compared with the other machine/deep learning algorithms. A set of epitopes experimentally proved were chosen to validate the model which was also tested through mutational analysis and cross-reactivity. The prepared dataset gave some valuable insights into the epitope distribution statistics and their conservancy in various outbreaks. An idea about the most probable range of peptide-MHC binding affinities was also obtained. Both the models performed well giving high accuracies when validated. These epitopes were checked for cross-reactivity with other antigens upon which it proved to be highly conservative and ideal for vaccine formulation.

**Conclusions** The combination of various features and the resulting model efficiencies in turn proved that the collected features are valuable in the easy identification of the vaccine candidates. This suggests that our proposed models have more potential for conserved epitope prediction compared to other existing models trained on similar data and features. The possibility of refining the model with more set threshold values based on more parameters is an added feature that makes it more user driven. Furthermore, the uniqueness of the model due to exclusive set of Neuraminidase epitopes paves a robust way for rapid vaccine design.

## Highlights

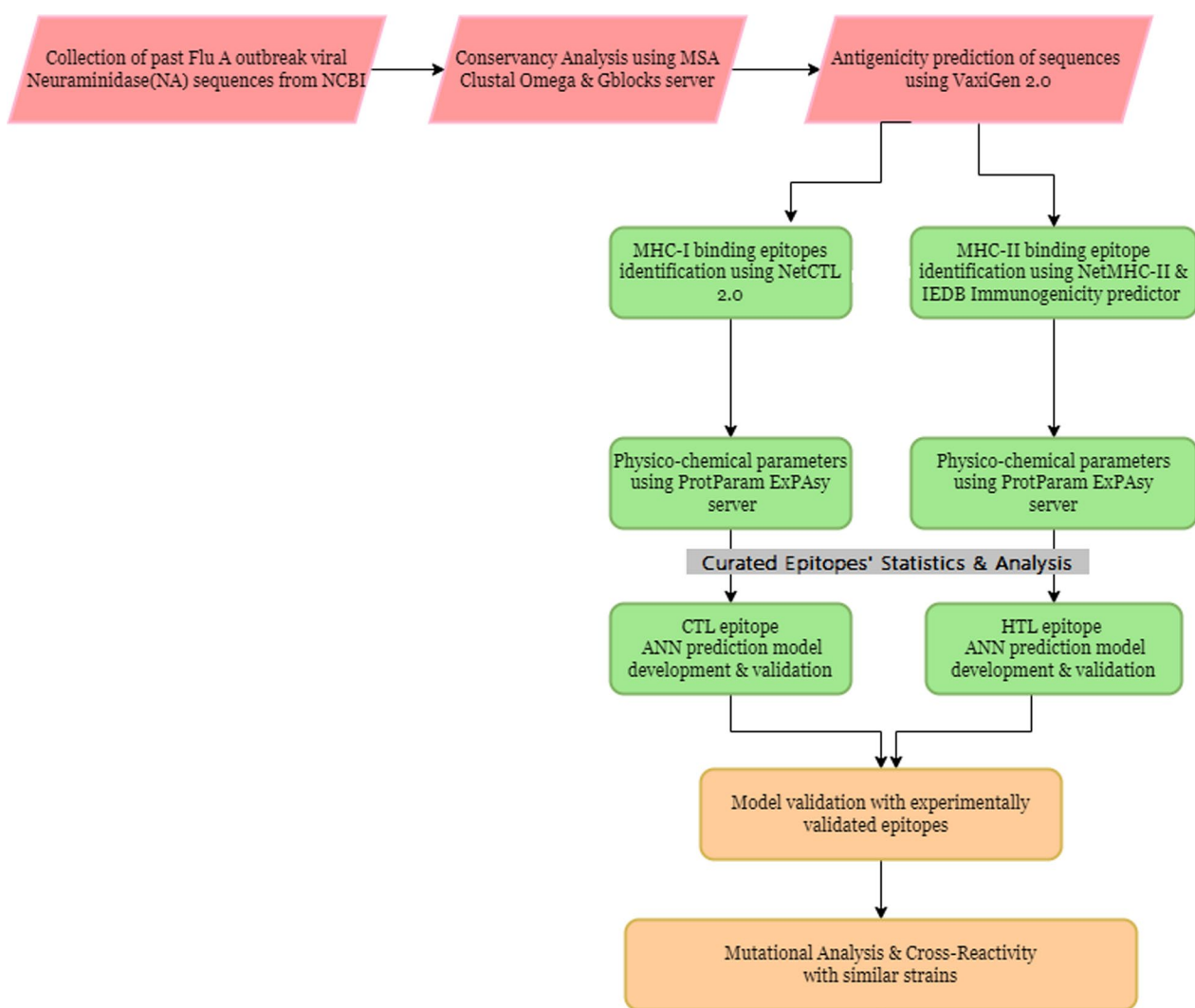
- 1 The analysis of the past viral strains of Neuraminidase, delivered an interesting list of possible epitopes for vaccine design.

\*Correspondence:  
Prince R. Prabhu  
princeprp@gmail.com  
Full list of author information is available at the end of the article

- 2 The class-I & class-II MHC binding reports of the test set of epitopes, strongly validate the high performance (~ or above 90%) of the ML models.
- 3 The homology studies with nematode antigens suggest that our predicted epitopes are free from cross-reactivity with parasitic epitopes.
- 4 The very few mutations reported for Neuraminidase (being the conserved one among the other proteins) from the mutational analysis prove that the vaccine formulations with our predicted epitopes are capable of overcoming any antigenic drifts in the future.

**Keywords** Vaccine design, Influenza A, T-cell epitopes, Neuraminidase, Machine learning, Deep learning, Immunology

**Graphical abstract**



### Background

Influenza A continues to be a global illness that causes huge morbidity and mortality in humans. Its ability to mutate, even now, termed as ‘antigenic shift’ and ‘drift’, makes the circulating strain prediction difficult and antigenic mismatch likely (Kim et al. 2022). And this poses a serious challenge for the scientists in each pandemic season for having an optimized vaccine candidate discovery due to the ineffectiveness of the already existing vaccines (Sheikh et al. 2016; Zeller et al. 2021). Thus, a good vaccination strategy along with the increasing need for easier prediction methods is highly in demand in this era.

Only Influenza A has both seasonal, epidemic and pandemic capability and is classified into subtypes according to its hemagglutinin (HA) and neuraminidase (NA) surface glycoprotein antigens. A good vaccine recipe should target potential circulating strains of the virus in humans that could provide population immunity in the new flu season. Hence, targeting the right protein could help us rationalize with the right solution. Even though HA antigenic regions had been a long-standing interest, the more conserved the NA regions, the greater the potential for protective immunity (Almalki et al. 2022). Therefore, an effective vaccine recipe could be developed by targeting potentially conserved, cross-reactive T- and B-cell epitopes of this strain. Such ‘universal’ vaccine design can potentially be addressed by a T-cell epitope ensemble vaccine comprising short, highly conserved, immunogenic peptides from influenza to activate T-cells. All these shows the pre-dominant importance of the T-cell epitope candidates. The role of CTL in cellular immunity includes the direct clearance of virally infected cells and the indirect recruitment of other immune cells via chemokine and cytokine secretion (McGee and Huang 2022). CD4+ T cells’ primary roles include B cell stimulation leading to specific antigen antibody production as well as stimulating CD8+ proliferation and memory responses. CD4+ T-cells also mediate direct and indirect viral clearance, and symptom severity reduction in secondary infection. These T-cell epitope candidates are usually identified using computational prediction tools to

facilitate the efficient vaccine design (Sanchez-Trincado et al. 2017).

These can reduce the time and resources needed for epitope identification projects by narrowing down the peptide repertoire that needs to be experimentally tested. Most of these prediction tools are developed using various statistical and machine learning algorithms trained on mainly two types of data: binding affinities of peptides to specific MHC molecules generated using binding assays or sets of naturally processed MHC ligands found by eluting peptides from MHC molecules on cell surface and identifying them by mass spectrometry (Barra et al. 2018). So, there is a need to include more deciding features which can decide the immunogenic and thus vaccine candidate potential. But, if a new model could be developed by training the data on some more efficient features which decides its epitope’s vaccine candidate potential, it can be more rewarding to the vaccine research.

In the present scenario where many epitope prediction servers are available, it is necessary to keep comparing the performance of the different methods against each other, to rationally decide which methods to choose, and to allow developers to understand what changes can truly improve the prediction performance. Through this study, we also evaluated the existing servers for the Flu A epitope prediction by means of efficiency, accuracy classification parameters and ROC curves (Suri and Dakshanamurthy 2022; Xia et al. 2021). One issue with the past evaluations has been that they are commonly evaluated using the same set of data on which they were trained, which can impact the performance results. Another problem that arises when we employ this model is that the data used for training might exclusively use theoretical prediction results without relating to experimental results which offers no reliability or translation advantage for developing vaccine. Therefore, it makes necessary to validate the model with prior curated experimentally published data for qualitatively robust prediction (Ramírez-Salinas et al. 2020).

**Table 1** No. of viral strains of Neuraminidase protein retrieved

Period	Name of the flu	No. of sequences
1918–20	Spanish flu	1
1957–59	Asian flu	44
1968–69	Hong Kong flu	25
2009–10	Swine flu	1569

**Table 2** Search set parameters for CTL and HTL peptides from ProtParam server

Theoretical pl	Putative zwitterion status
Instability index	< 40 (peptide is stable)
Aliphatic index	(+) values indicates the increase in thermostability
GRAVY (Grand average of hydropathy)	(+) values indicate hydrophobicity (-) values indicate hydrophilicity

## Methods

### Dataset curation

#### Collection of raw sequences of viral strains and conservancy analysis

The raw sequences of Influenza A Neuraminidase strain virus of all the past four outbreaks were collected from NCBI Influenza Virus Resource Database (Table 1). All the sequences from these outbreak periods were collected according to the corresponding subtypes expressed with the other parameters.

These protein sequences were subjected to Multiple Sequence Alignment by means of Clustal Omega server. The aligned regions were curated from the results. The output aligned sequences from the MSA were further analysed for conserved regions by means of Castresana G blocks server. This server gives the blocks of conserved sequences with relevant nucleotide-base representation by removing least aligned regions and divergent sequences from the input of multiple sequence alignments. The conserved blocks were displayed within a Blue underlining at the bottom. Stringent selection was applied for not allowing many contiguous non-conserved positions.

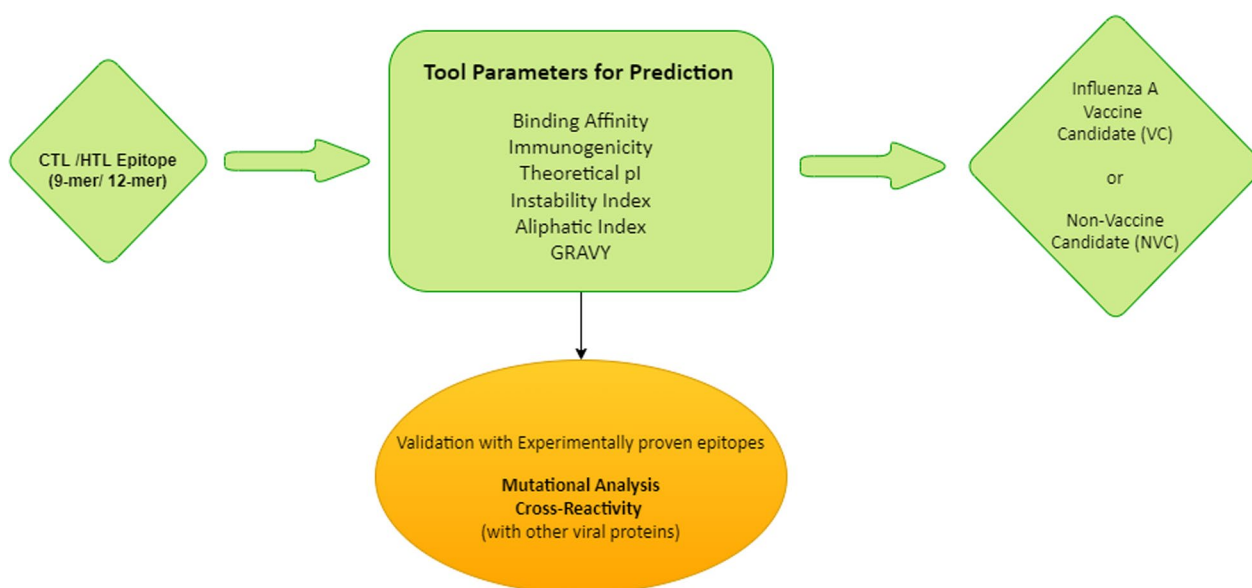
**Antigenicity determination** The strains collected were checked for its antigenic potential by means of Vaxijen 2.0 server. The optimum threshold was set as 0.5. The antigen classification is solely based on the physicochemical properties of proteins without recourse to sequence alignment. It can give the output list of the sequences with the Antigenic and Non-antigenic tags based on the threshold provided. The average range of antigenicity

score of antigenic strains was identified as: 0.5- 0.75. The dataset from flu 2009 contained some non-antigenic strains, and they were filtered out.

**CTL epitope prediction** NetC TL 1.2 server predicted the CTL (CD8+) epitopes in protein sequences for all the 12 MHC supertypes: A1, A2, A3, A24, A26, B7, B8, B39, B44, B58, B62 with the set threshold parameters. The probable epitopes for each supertype were screened out from all peptides based on the combined score and they were classified year-wise after removing the redundant ones. The immunogenicity scores were obtained for all the CD8+ epitopes by (Immune Epitope Database) IEDB T cell CD8+ Immunogenicity prediction tool.

**HTL epitope prediction** This process involved multiple screening steps considering the limitation of *in-silico* methods to predict only peptides that bind to MHC II. For the peptides to be epitopes, it is important to understand the activation of T helper cells. For this, the pMH-CII complexes should be immunogenic enough to induce the production of Tc-cells.

- (i) **MHC II-Peptide binding affinity prediction** NetMHCII 2.3 server predicted the binding of 15-mer peptides to HLA-DP, HLA-DQ and HLA-DR loci specific alleles using artificial neural network. The 9-mer binding cores were also obtained. Like MHC I prediction, output values were obtained in nM IC50 values, and the strong binders (SB) & weak binders (WB) were separately identified.
- (ii) **IEDB CD4+ T cell Immunogenicity predictor** Predicts the allele independent CD4 T cell immunogenicity at population level. The 1st, 2nd and



**Fig. 1** A schematic representation of the neural network-based prediction tool—InflANNet

C-terminus positions of the peptides were masked. RPCFWIELI epitope showed the highest score of 0.5193.

The final list of predicted 15-mer peptides were checked for the CD4+ immunogenicity by means of IEDB T-cell Immunogenicity prediction tool—CD4 episcore. This step is crucial in case of CD4+ epitope identification.

**Collection of other physico-chemical features of the epitopes**

Along with the binding affinity and immunogenicity values, various other physico-chemical properties of the epitopes based on their amino-acid sequence were collected from the ProtParam ExPASy server. This provided more reliable information about their antigenicity. These corresponding parameters were considered for each entry of CTL and HTL peptides and were screened based on thresholds (Table 2).

The potential features (Rostaminia et al. 2021) collected are as follows:

**Neural network development and training**

The dataset consisted of 1200 peptides for CTL model—Model I and 1400 for HTL- Model II comprising features with similar training criteria. The categorical encoding of the variables for the two output categories vaccine candidates (VC) & non-vaccine candidates (NVC) were also done. Similarly, as the mere alphabet representation of the amino-acids possess practical challenges in the

protein prediction problems, they were One-hot encoded appropriately. Then, the model was built and trained through various methods. For this problem of Binary Classification, train\_test\_split method from Sci-kit learns and some deep learning libraries—Keras, Tensorflow were also chosen. The data were split into train, test and validation sets in a ratio of 60:20:20 for both the datasets.

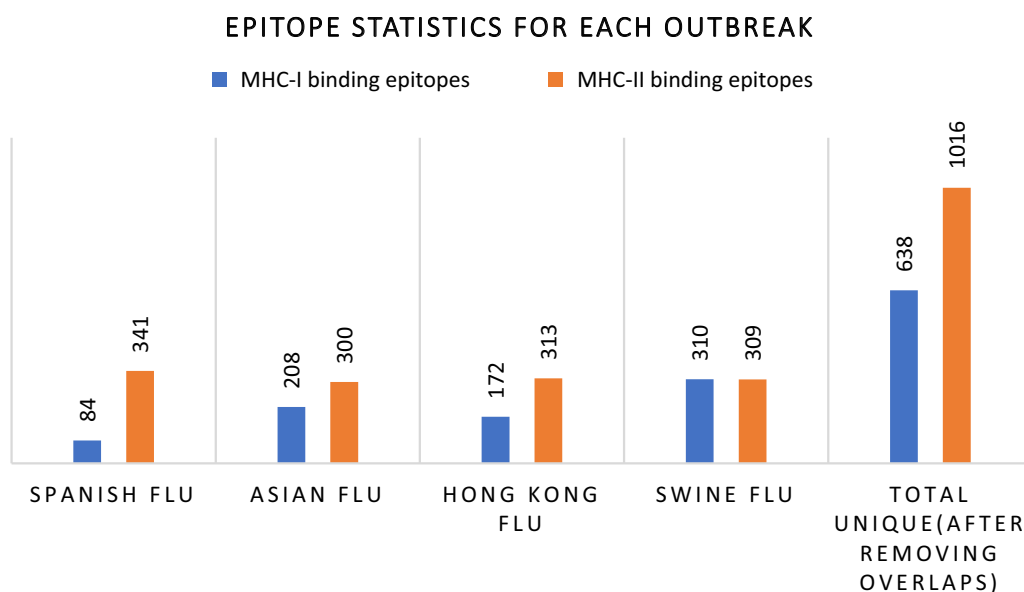
To develop a good performing model, hyperparameter tuning was done with appropriate number of network layers set. Different optimizers were tried such as Adam, RMSprop along with loss parametric—binary\_crossentropy and accuracy metric—binary\_accuracy. Different number of epochs 50, 80, 100, 150 were trained on. An overview of the constructed machine learning model is depicted in Fig. 1.

**Model validation**

The test set of epitopes contained a list of 9 peptide sequences for vaccine candidates (VC) and 9 sequences for non-vaccine candidates (NVC) as predicted by the ML model. All the peptides were checked for their effectiveness in binding with various alleles for MHC-I and MHC-II separately. The MHCI binding predictions were made for 27 different HLA alleles using the IEDB analysis resource NetMHCpan (ver. 4.1) tool (Reynisson et al. 2020). Similarly, the MHCII binding predictions were made using the IEDB analysis resource Consensus tool (Wang et al. 2008, 2010).

**Mutational analysis for neuraminidase protein**

Though Neuraminidase is considered as the most conserved protein of the Influenza virus, we suspected that



**Fig. 2** Epitope statistics

the few mutations that are reported till date, might overlap with our predicted vaccine candidates. Performing an in-depth mutational analysis and checking for mutations within the epitope region would render a better understanding while considering the epitopes for vaccine formulation. Hence the complete list of mutations, that have been reported till date for Neuraminidase was collected from various literature and a mutated model of Neuraminidase was generated using the SWISS-MODEL homology modelling server (Waterhouse et al. 2018). The overlap of sequences was also checked using the

ClustalW tool for multiple sequence analysis (Thompson et al. 1994).

**Investigation of the disease pathology and comorbidities of Flu A associated with other parasitic infections and antigens**

There have already been many clinical and immunological links found between parasitic infections and

**Table 3** Highly repeated epitopes from pMHC I binding

	1918 (4 times)	1957 (5 times)	1968 (5 times)	2009 (5 times)
1	ELNAPNYHY	LAATVTLHF	WTSNSIVVF	VSINQNLEY
2	VSFDQNLDY	WTSNSIVVF		
3	WTSGSSISF			

**Table 5** Highly repeated CD4+ epitope cores

Epitope core (9-mers)	Repeatability
FVIREPFIS	11
VWMGRTISK	10
IVHISPLSG	9
MIWDPNGWT	9
NWKGSNRPV	9

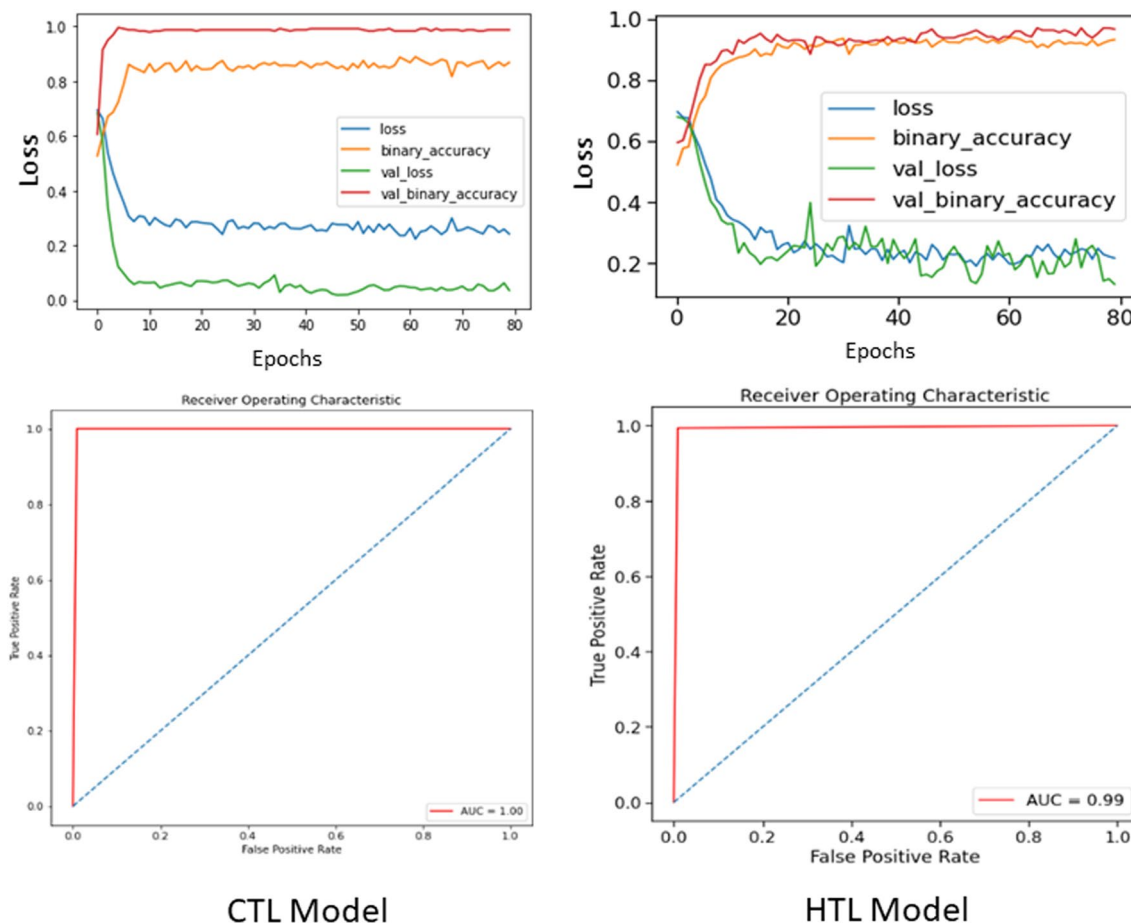
**Table 4** Highly repeated epitopes from pMHCII binding

SL.NO	1918 (3 times)	1957 (5 times)	1968 (4 times)	2009 (5 times)
1	CRTFFLTQGALLNDK	EGKIVHISPLSGSAQ	ADTRILFIEEGKIVH	RTFFLTQGALLNDKH
2	DSSFSVRQDIVAITD	FLMQIAILATTVTLH	ATASFYDGRVLDSI	SFKYGNVWIGRTKS
3	ECRTFFLTQGALLND	GKIVHISPLSGSAQH	EEGKIVHISPLSGSA	
4	FSFRYDNGVWIGRTK	GSVSLTIATVCFMLQ	EGKIVHISPLSGSAQ	
5	FVIREPFISCSHLEC	KEGKIVHISPLSGSA	ETRWWTSNSIVVFC	
6	GDVVFIREPFISCSH	KIVHISPLSGSAQHI	GKIVHISPLSGSAQH	
7	GFSFRYDNGVWIGRT	TRILFIEEGKIVHIS	GSVSLTIATVCFMLQ	
8	GIKGFSFRYDNGVWI		INRCFYVELIRGRKQ	
9	GQASYKILKIEGKGV		KIVHISPLSGSAQH	
10	IKGFSFRYDNGVWIG		KPQCQITGFAPFSK	
11	ISLILQIGNIISIVV		NRCFYVELIRGRKQE	
12	KGDVVFIREPFISCS		QETRVWTSNSIVVF	
13	LECRTFFLTQGALLN		QKIITIGSVSLTIAT	
14	NGIKGFSFRYDNGVW		RCFYVELIRGRKQET	
15	NRPWVSFDQNLDYQI		RILFIEEGKIVHISP	
16	NSRFESVAWSASACH		RVWTSNSIVVFCGT	
17	PWVSFDQNLDYQIGY		SKPQCQITGFAPFSK	
18	PYNSRFESVAWSASA		TIATVCFMLQIAILV	
19	QASYKILKIEGKVT		TRILFIEEGKIVHIS	
20	RTFFLTQGALLNDKH		TRWWTSNSIVVFCG	
21	SFRYDNGVWIGRTKS			
22	SKGDVVFIREPFISC			
23	SLILQIGNIISIVVS			
24	SRFESVAWSASACHD			
25	SYKILKIEGKVTKS			
26	YKILKIEGKVTCSI			
27	YNSRFESVAWSASAC			

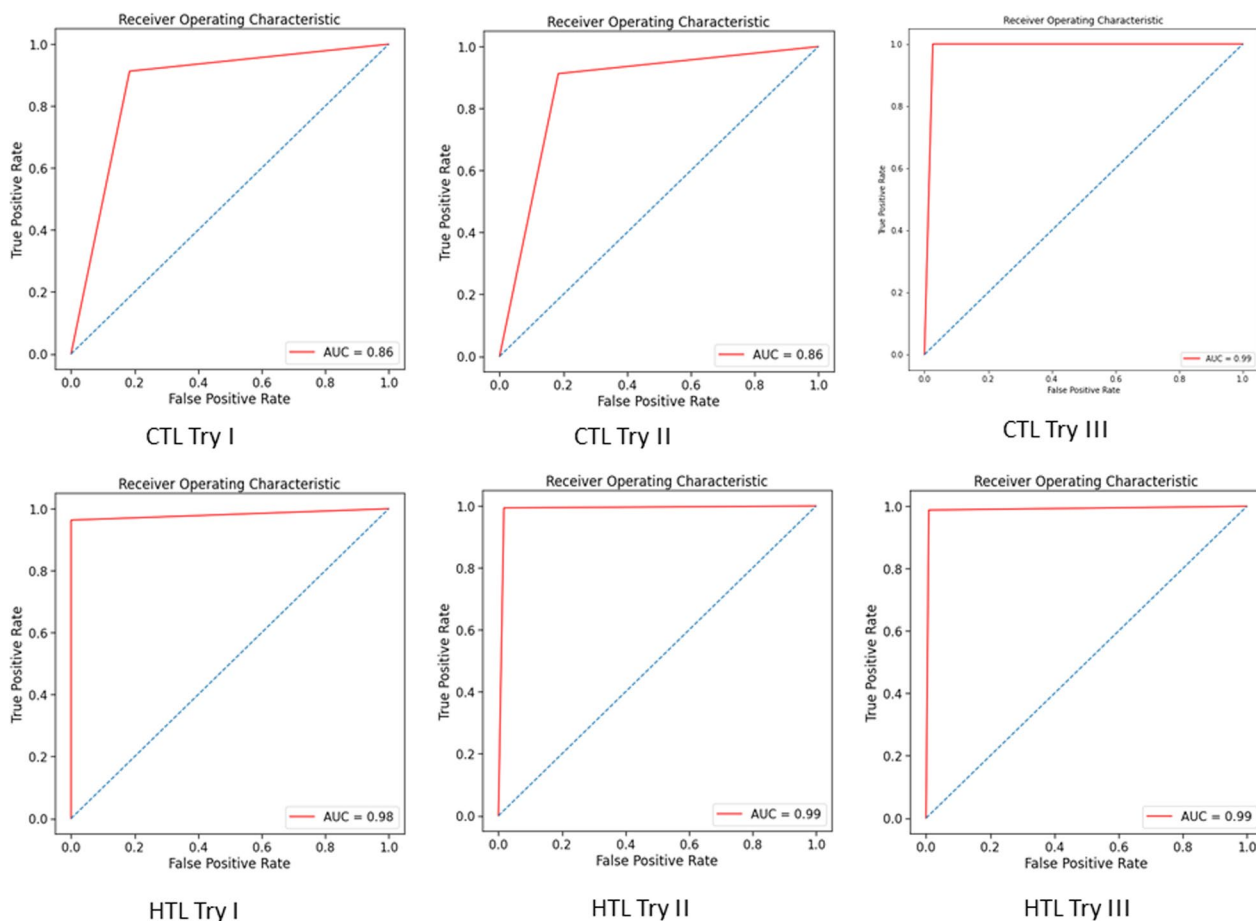
respiratory infections, particularly Influenza (Breloer and Hartmann 2023). It is through immune responses that these pathogens survive and reproduce, and these effects can be positive or negative. Comorbidities are complex in that they involve the immune system heavily. The human body responds better to Influenza A infection in humans with various helminthic infections. A good example is *Trichinella spiralis* (Furze et al. 2006). According to studies conducted on co-infections with *Trichinella pseudospiralis*, the bacterium suppresses inflammation and reduces cellular recruitment around implanted material (Stewart et al. 1985). A study by Kortzen et al. (2002) observed that the number of DX5+/CD3+ NK cells and DX5+/CD3+ T cells increased systemically during *Litomosoides sigmodontis* infection, so it is possible that the systemic release of parasite products and/or local cytokine and chemokine activity resulting from migration of new-born larvae through the capillary bed lining the lungs prompt the recruitment and activation of these cells.

In this study, we aim to investigate the effects of parasitic pathology on antigenic epitopes with flu's, and how the model predicts these effects. This requires a good understanding of the homology associated with the strains and the epitopes collected from various studies. Therefore, homology studies were performed with the other possible cross-reactive nematode antigens with Flu A using ClustalW for the following sequences retrieved from NCBI database.

1. Flu A—ADK33724.1 neuraminidase [Influenza A virus (A/Aarhus/INS242/2009(H1N1))]
2. Abundant Larval Transcript—CDW52393.1 abundant larval transcript 2 (alt2) protein [*Trichuris trichiura*]
3. Venom Allergen Homologue—XP\_042932260 [*Bru-gia malayi*]
4. 3.Filarial Thioredoxin—OZC12539.1 thioredoxin [*Onchocerca flexuosa*]



**Fig. 3** Accuracy–loss plots and ROC curves for original models

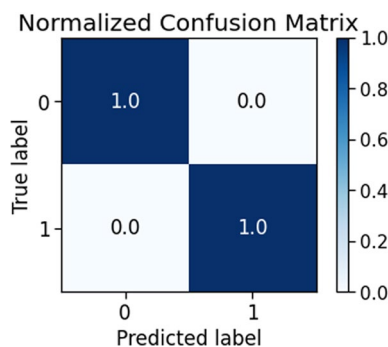


**Fig. 4** Comparison of various feature combination models

**Table 6** Analysis of strong features

	CTL model % accuracy	HTL model % accuracy
Try 1: Peptide only	86	98
Try 2: Peptide, GRAVY, pl, Aliphatic Index	85	99
Try 3: Binding affinity, Immunogenicity	99	99
With all features (Original model)	99.5	99

5. Filarial Glutathione S transferase—CCF78323.1 glutathione S-transferase [*Wolbachia endosymbiont of Onchocerca ochengi*]
6. GP29—CAA44965.1 gp29 [*Brugia pahangi*]
7. Thioredoxin Peroxidase—NP\_001285204.1 thioredoxin peroxidase 1, isoform E [*Drosophila melanogaster*]



**Fig. 5** Confusion matrix for validation set-CTL

**Results**

**Epitope statistics and analysis**

The initial stage results gave a good understanding of the epitope distribution statistics and the disease



	precision	recall	f1-score	support
0	1.00	1.00	1.00	5
1	1.00	1.00	1.00	9
micro avg	1.00	1.00	1.00	14
macro avg	1.00	1.00	1.00	14
weighted avg	1.00	1.00	1.00	14
samples avg	1.00	1.00	1.00	14

**Fig. 6** Classification report for validation set-CTL

prevalence (Fig. 2). The curated list from raw viral sequences of CTLs revealed NPNQKIITI as the highly conserved epitope throughout all four outbreaks. Swine flu (2009) showed the highest no. of around 300 unique CD8+ (CTL) epitopes while in the case of CD4+ T-cell (HTL) ones; Spanish flu (1918) showed the higher number of pMHC-II binding 15-mer peptides ranging around 340. The highly repeated 9-mer core which were found to be immunogenic was YKILKEIKG (1918) and YKIFREIKG (2009) from the unique 15-mer HTL epitopes obtained. Analysis also revealed that the high binding affinity was shown by the epitopes of Spanish and Asian flus.

The highly expressed epitopes among various years in MHC I binding are listed in Table 3.

The highly expressed epitopes among various years in MHC II binding are listed in Table 4.

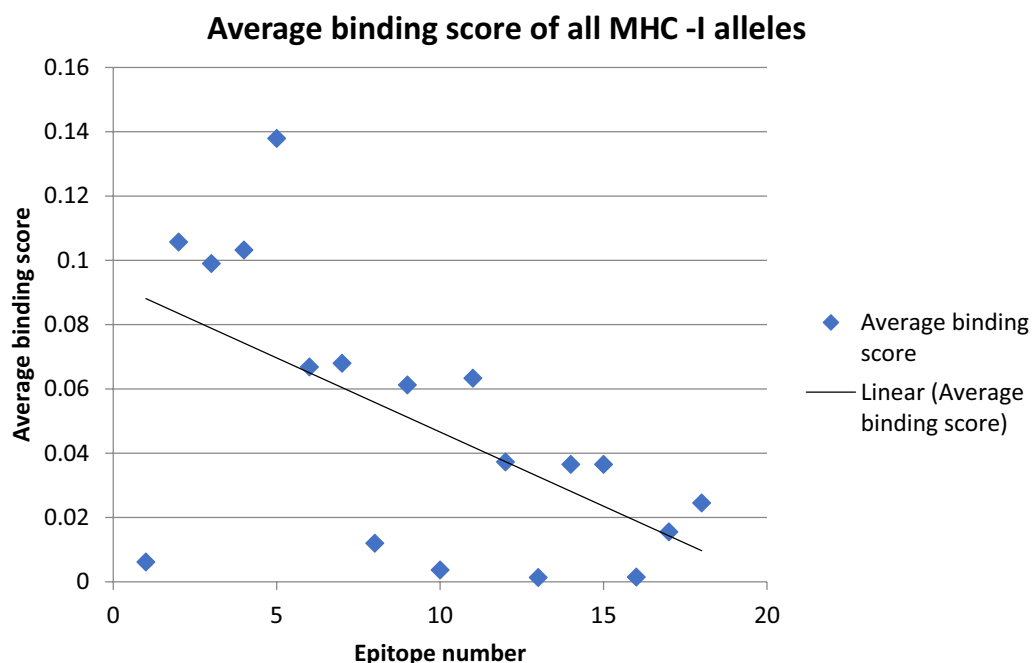
The highly repeated 9-mer binding cores are listed in Table 5.

**Model architecture and optimization**

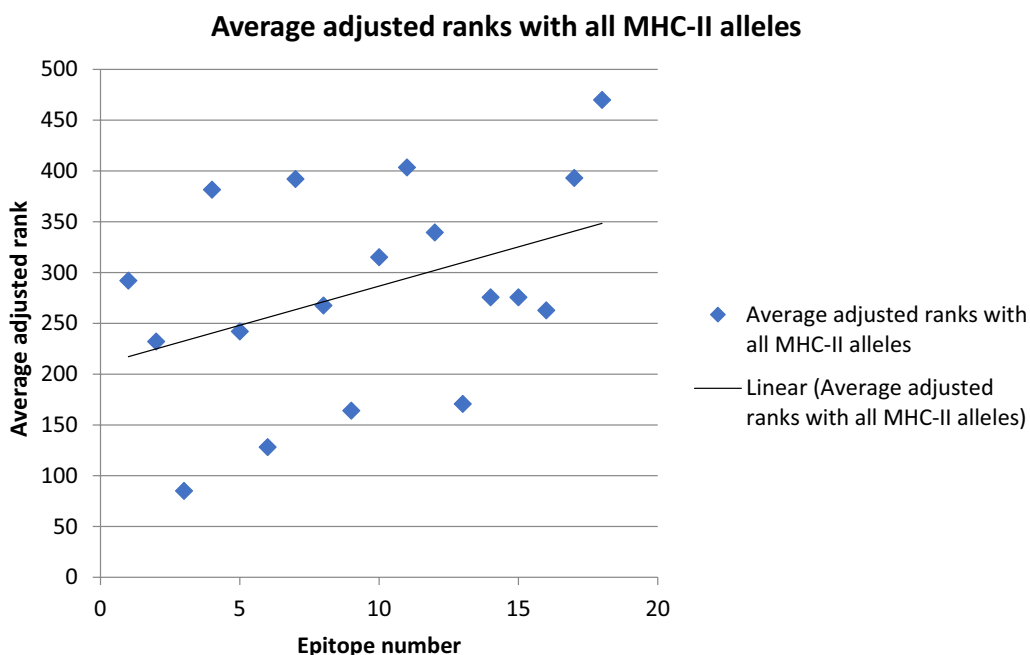
The neural network models were built for both 9-mer CTLs and 15-mer HTLs. A binary classifier defined with a train\_validation\_test split of 60:20:20 for both models with an optimized model architecture gave a good accuracy score of 99.5% for CTL model and 99% for HTL model with very minimum loss. The model architecture defined was of 3 layers with 10, 3, and 2 layers, respectively, with a dropout of 0.5 for each layer. Adam optimizer was employed, and the optimal number of training epochs was found to be 80.

**Model evaluation**

The model performances were evaluated using various metrics like accuracy, precision, F1- score and auc\_roc\_score. This being balanced datasets, accuracy was taken as the primary deciding measure. These prediction models were further validated with a set of experimentally validated CTLs and HTL epitopes collected from IEDB database which are experimentally curated by T-cell, B-cell, and MHC ligand assays. Some were also collected from cited literature findings (Mintaev et al. 2022). All



**Fig. 7** A scatter plot representing the range of binding scores for the various vaccine and non-vaccine candidates as predicted by the ML model. Epitope numbers 1–9 stand for vaccine candidates (VC) and epitope numbers 10 to 18 stand for non-vaccine candidates (NVC)

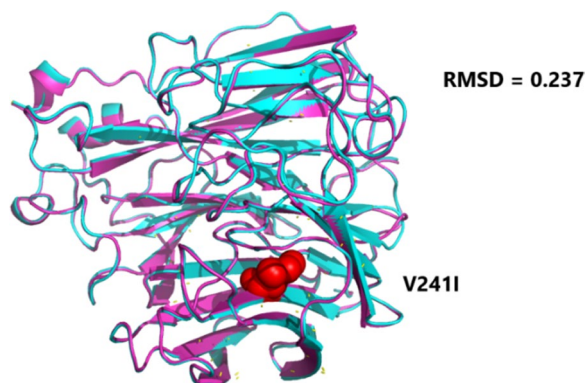


**Fig. 8** A scatter plot representing the range of average adjusted ranks for the various vaccine and non-vaccine candidates as predicted by the ML model. Epitope numbers 1–9 stand for vaccine candidates (VC) and epitope numbers 10–18 stand for non-vaccine candidates (NVC)

**Table 7** A tabular column representing the various epitope sequences with their binding activity for MHC-I and MHC-II alleles

Sequence Number	Peptide Sequence	Category as predicted by the ML model	Average Binding score with all MHC-I alleles	Average adjusted rank with all MHC-II alleles
1	CVNGSCFTV	Vaccine candidate	0.006149	292.017
2	ETFKVIGGW	Vaccine candidate	0.10571	232.0263
3	GLISLILQI	Vaccine candidate	0.09898	85.10259
4	NSDTV DWSW	Vaccine candidate	0.10318	381.5259
5	VSFDQNL DY	Vaccine candidate	0.137908	242.0352
6	QIAILVTTV	Vaccine candidate	0.066832	128.2119
7	DTVH D RTPY	Vaccine candidate	0.067977	392.1711
8	GRADTKILF	Vaccine candidate	0.011991	267.513
9	LMNELGV PF	Vaccine candidate	0.061249	164.0478
10	ASVLSGGEL	Non-Vaccine candidate	0.003676	315.0611
11	GELDRWEKI	Non-Vaccine candidate	0.063312	403.5344
12	GQLDRWEKI	Non-Vaccine candidate	0.03727	339.5222
13	RWEKIRLRP	Non-Vaccine candidate	0.001314	170.6922
14	KIRLRPGGK	Non-Vaccine candidate	0.036533	275.6174
15	KIRLRPGGK	Non-Vaccine candidate	0.036533	275.6174
16	IRLRPGGKK	Non-Vaccine candidate	0.001474	262.7644
17	RLRPGGKKQ	Non-Vaccine candidate	0.01552	393.1656
18	RPRPGGKKK	Non-Vaccine candidate	0.024531	469.9785

A high binding score corresponds to effective binding with MHC-I alleles and a low adjusted rank corresponds to effective binding with MHC-II alleles. For each epitope, the binding scores and the adjusted ranks were taken as an average of 27 different alleles for MHC-I and MHC-II separately



**Fig. 9** A cartoon representation of the superposition of mutated model with the native structure of Neuraminidase showing an RMSD score of 0.237, created using PyMol molecular visualizer (Yuan et al. 2017). The mutated model is represented in magenta and the native structure in cyan, while the mutated residues V241I are represented as spheres in red

these gave accurate predictions when tested on the developed models with the same parameters.

Different models were trained with various combinations of the features. The best model proved to be the ones with all features which denoted that all the features contributed well for the vaccine candidate design. Binding affinity and Immunogenicity were found to be the strong features of all with which the models gave an accuracy of 98% for both models. The ROC plots and accuracies of other combination results are as given below (Figs. 3 and 4).

The accuracy of the CTL model and HTL model for some of the strong features is listed in Table 6.

Figure 5 represents the normalized confusion matrix for the CTL validation set, and Fig. 6 represents the classification report for the CTL validation set.

#### Model validation

The binding activity checked with the epitope analysis resource provided by IEDB showed a strong correlation with the results obtained from our ML model with the test set of epitopes and thereby strengthening the performance of our ML model (Figs. 7, 8 and Table 7).

#### Mutational analysis for neuraminidase

The following mutations reported were collected from various literature sources: D151G, G147R, R292K, Q136K, E119K, V116A, I223R, S247N, H275Y, N295S, N200S, V241I (Hooper and Bloom 2013; Shao et al. 2017; Hossain et al. 2002; Eshaghi et al. 2014; Jain et al. 2018). The reference sequence for NA protein-Flu

A—ADK33724.1 neuraminidase [Influenza A virus (A/Aarhus/INS242/2009(H1N1))] was obtained from NCBI database and the novel mutations were introduced into the sequence. The multiple sequence analysis performed in ClustalW (Thompson et al. 1994) revealed that the mutation V241I was found to occur in the vaccine epitope region CVNGSCFTV as predicted by the ML model. However, the other mutations had no overlaps with any of the epitope regions (Fig. 9).

The observation leads to the inference that, most of the epitopes predicted as effective vaccine candidates from the ML model were highly conserved regions and hence considering them for vaccine formulation will prove to overcome any antigenic drifts in the future, which is also a major objective of our research.

#### Homology studies with possible cross-reactive nematode antigens

For the homology studies performed with the various nematode antigens, no complete overlap was found for any of our predicted epitopes, hence suggesting that the reformulated vaccine with our predicted epitopes will have the least chances for cross-reaction with parasitic antigens (Breloer and Hartmann 2023).

#### Discussion

Influenza has been a globe-threatening infectious disease which is still a concern of the epidemiological, social, and biological scientists (Kim et al. 2022). Even though many tri-valent vaccines and anti-viral drugs exist, the Flu virus by virtue of antigenic drift and shift continues to circulate in humans and also in animal populations that results in seasonal evolution of influenza viruses (Kim et al. 2022; Petrova and Russell 2018). Predominantly, the surface proteins of the virus namely, Haemagglutinin (HA) and Neuraminidase (NA) that play a pivotal role in the pathogenesis is the focal point for vaccine design. In our study, we exploit the conserved regions of these proteins across the flu seasons to identify potential epitopes using neural network models incorporating more features for robust prediction. This approach could be a potential tool for proposing universal targets for vaccine development in Influenza from conserved regions. Discovering such a universal solution for these recurring episodes of flu via peptide vaccines could save the future generations from serious ailments.

Our study presented here provides such a possible solution through peptide vaccine design via an epitope prediction tool development. The current servers that exist for the epitope peptide prediction mostly focus on the binding affinity of the peptides to the MHC

complexes and is not customized for specific viral infections (Ras-Carmona et al. 2021; Desta et al. 2023). Our tool for epitope prediction is highly specific for Influenza A and carries additional feature of predicting the potential of a peptide's scope to be a putative vaccine candidate for any future outbreaks. These short-listed epitopes could be further validated by additional biological screening assays from clinical samples to develop a successful vaccine. Neuraminidase, being the choice for the protein vaccine design, has gotten to be the recent interest worldwide due to its highly conserved nature and vaccine effectiveness over the other proteins. We came up with this deep neural network model due to the high demand for the fast and accurate solutions that enables the effective Flu vaccine target epitope optimizations and predictions instead of the time-consuming simulation design and experimental methods. This tool was developed from trained B-cell and T-cell epitope sets that were curated independently incorporating validation features for vaccine effectiveness. The high accuracy of 99.5% for the CD8+CTL & 99% for the CD4+HTL models attributes to the high antigenicity shown by the Neuraminidase strains and the stability of the 9-mer, 15-mer peptides based on its strong physio-chemical parameters like binding affinity & immunogenicity. These peptides will also compliment effectively the existing resources for epitopes across populations due to its wide allelic coverage as observed from the IEDB resource results. The high repeated-ness of epitope scores among both sets also strengthens the fact that Neuraminidase could be a sole active target for Influenza vaccines in future. The Precision and Recall value of 1.0 of the models also validates the robustness of the model with no/less False Positives (FP) and True-Negatives (TN). The validation was done through a set developed from the experimentally proven vaccine epitopes from the literature (Kim et al. 2022; Lee et al. 2020). The strong correlation in the average binding score as demonstrated by the experimentally proven vaccine epitopes supports the robust design of the ML model. In addition, mutational analysis of the Neuraminidase protein showed that most of the predicted epitopes by the ML model lie in highly conserved regions of the protein. Therefore, the predicted epitopes by our ML model are expected to withstand seasonal epidemics and may provide long-term immune-resistance. In addition, the predicted epitopes were non-overlapping with the parasitic antigens, thereby to prevent cross-reactivity during vaccine administration. Interestingly, there are reports of diminished immune responses due to helminth infections for people living in infectious nematode endemic areas (Breloer and Hartmann 2023). Our homology studies on selected nematode antigens lacked

complete overlap with our predicted epitopes, suggesting that the epitopes predicted using our tool could provide putative epitopes for designing future Influenza vaccines more effective in helminth endemic regions. Overall, this ML model serves as a bridging gap in real world between vaccine development and its clinical application.

## Conclusions

This highly efficient model developed and presented here with a pan-data collected from across the globe would be the first-of-its-kind machine/deep learning algorithm that can predict the probable vaccine candidates even with minimal peptide data. This tool could independently compliment and mutually be cross-curated from majority of the other models that otherwise predicts only the MHC-binding affinity. Our prediction tool has a limitation for screening lengthy in-put protein or peptide sequences but works efficiently for independent, short peptide fragments. The potential putative epitopes predicted separately could be subsequently concatenated to enable faster vaccine design. Moreover, the recent research on Flu indicating the potency of T-cell epitope vaccines in providing a stable solution for the viral drift, makes this study and model even more worthwhile. Since, the cross-validation of the results obtained from the test set of peptides was in strong correlation with the IEDB Epitope analysis resource which is a renowned tool for epitope prediction suggesting the vigour our model. Additionally, our model stands unique from the other already existing models for being much efficient in overcoming antigenic drifts in the future, since the algorithm is built based on the most conserved Neuraminidase protein in influenza virus. The cross-reactivity analysis with the nematode antigens has provided a novel insight about our predicted epitopes when considered for vaccine reformulation in endemic areas with neglected nematode infectious diseases (Breloer and Hartmann 2023). We thereby suggest that the usefulness of InflAN-Net would prove to be beneficial to the overall scientific community and could be employed as a kick-start-tool in developing strategies for future clinical influenza vaccine experiments.

## Abbreviations

CTL	Cytotoxic T-Lymphocyte
HTL	Helper T-Lymphocyte
HA	Hemagglutinin
NA	Neuraminidase
NCBI	National Centre for Biotechnology Information
NK cells	Natural Killer cells
IEDB	Immune Epitope Database
ML	Machine Learning
RMSD	Root Mean Square Deviation

### Acknowledgements

We would like to thank Dr. Rahul Siddharthan (Professor, Computational Biology group, The Institute of Mathematical Sciences, Chennai); Dr. Suvro Chatterjee (Associate Professor, AU-KBC Research Centre, MIT, Anna University, Chennai); Chandrani Kumari (at The Institute of Mathematical Sciences, Chennai) and Yeshwanth Sripathy (The Institute of Mathematical Sciences, Chennai) for providing us ardent support towards executing this research project. We are also highly obliged in taking this opportunity to sincerely thank the authorities of the High Performance Computing(HPC) Centre, IMSc, Chennai and also the Biotechnology Information System (BTIS) facility of the Department of Biotechnology, Anna University, for providing us with the necessary computational facilities.

### Author contributions

KR collected literature data and designed the research protocol. KR and SM performed the in-silico work. PRP analysed and interpreted the results. KR and SM wrote the manuscript and created appropriate illustrations for the manuscript. PRP supervised the entire research work and edited the manuscript. All authors have read and approved the final manuscript.

### Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

### Availability of data and materials

All necessary data generated or analysed during this study are included in this article. The algorithm developed here, along with the library and versions, is available at github repository—<https://github.com/karthikavarmar/InfIANNNet>. Any additional data could be made available from the corresponding author upon request.

### Declarations

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

#### Author details

<sup>1</sup>Centre for Biotechnology, Anna University, Chennai, India. <sup>2</sup>The Hamburg Centre for Ultrafast Imaging (CUI), University of Hamburg, Hamburg, Germany. <sup>3</sup>Laboratory for Structural Biology of Infection and Inflammation, Institute for Biochemistry and Molecular Biology, University of Hamburg, c/o DESY, 22603 Hamburg, Germany.

Received: 5 June 2023 Accepted: 9 August 2023

Published online: 22 August 2023

### References

- Almalki S, Beigh S, Akhter N, Alharbi RA (2022) In silico epitope-based vaccine design against Influenza A neuraminidase protein: Computational analysis established on B- and T-cell epitope predictions. *Saudi J Biol Sci* 9:103283. <https://doi.org/10.1016/j.sjbs.2022.103283>
- Barra C, Alvarez B, Paul S, Sette A, Peters B, Andreatta M, Buus S, Nielsen M (2018) Footprints of antigen processing boost MHC class II natural ligand predictions. *Genome Med* 10:84. <https://doi.org/10.1186/s13073-018-0594-6>
- Breloer M, Hartmann W (2023) Filial infections compromise influenza vaccination efficacy: Lessons from the mouse. *Immunol Lett* 255:62–66. <https://doi.org/10.1016/j.imlet.2023.03.001>
- Desta IT, Kotelnikov S, Jones G, Ghani U, Abyzov M, Kholodov Y, Standley DM, Beglov D, Vajda S, Kozakov D (2023) The ClusPro AbEMap web server for the prediction of antibody epitopes. *Nat Protoc* 18:1814–1840. <https://doi.org/10.1038/s41596-023-00826-7>
- Eshaghi A, Shalhoub S, Rosenfeld P, Li A, Higgins RR, Stogios PJ, Savchenko A, Bastien N, Li Y, Rotstein C, Gubbaya JB (2014) Multiple Influenza A (H<sub>2</sub>N<sub>2</sub>) mutations conferring resistance to neuraminidase inhibitors in a bone marrow transplant recipient. *Antimicrob Agents Chemother* 58:7188–7197. <https://doi.org/10.1128/AAC.03667-14>
- Furze RC, Hussell T, Selkirk ME (2006) Amelioration of influenza-induced pathology in mice by coinfection with *Trichinella spiralis*. *Infect Immun* 74:1924–1932. <https://doi.org/10.1128/IAI.74.3.1924-1932.2006>
- Hooper KA, Bloom JD (2013) A mutant influenza virus that uses an N1 neuraminidase as the receptor-binding protein. *J Virol* 87:12531–12540. <https://doi.org/10.1128/jvi.01889-13>
- Hossain MG, Akter S, Dhole P, Saha S, Kazi T, Majbauddin A, Islam MS (2021) Analysis of the genetic diversity associated with the drug resistance and pathogenicity of Influenza A virus isolated in Bangladesh from 2002 to 2019. *Front Microbiol* 12:1–11. <https://doi.org/10.3389/fmicb.2021.735305>
- Jain A, Prakash S, Jain B (2018) Changes in hemagglutinin and neuraminidase genes of pH1N1 Influenza virus strains collected from a North Indian tertiary care hospital during 2015. *Intervirology* 60:263–270. <https://doi.org/10.1159/000489632>
- Kim Y-H, Hong K-J, Kim H, Nam J-H (2022) Influenza vaccines: past, present, and future. *Rev Med Virol* 32:e2243. <https://doi.org/10.1002/rmv.2243>
- Korten S, Volkman L, Saefel M, Fischer K, Taniguchi M, Fleischer B, Hoerauf A (2002) Expansion of NK cells with reduction of their inhibitory Ly-49A, Ly-49C, and Ly-49G2 receptor-expressing subsets in a murine helminth infection: contribution to parasite control. *J Immunol* 168:5199–5206. <https://doi.org/10.4049/jimmunol.168.10.5199>
- Lee EK, Tian H, Nakaya HI (2020) Antigenicity prediction and vaccine recommendation of human influenza virus A (H<sub>2</sub>N<sub>2</sub>) using convolutional neural networks. *Hum Vaccin Immunother* 16:2690–2708. <https://doi.org/10.1080/21645515.2020.1734397>
- McGee MC, Huang W (2022) Evolutionary conservation and positive selection of Influenza A nucleoprotein CTL epitopes for universal vaccination. *J Med Virol* 94:2578–2587. <https://doi.org/10.1002/jmv.27662>
- Mintaev RR, Glazkova DV, Bogoslovskaya EV, Shipulin GA (2022) Immunogenic epitope prediction to create a universal influenza vaccine. *Heliyon* 8:e09364. <https://doi.org/10.1016/j.heliyon.2022.e09364>
- Petrova VN, Russell CA (2018) The evolution of seasonal influenza viruses. *Nat Rev Microbiol* 16:47–60. <https://doi.org/10.1038/nrmicro.2017.118>
- Ramírez-Salinas GL, García-Machorro J, Rojas-Hernández S, Campos-Rodríguez R, de Oca AC-M, Gomez MM, Luciano R, Zimic M, Correa-Basurto J (2020) Bioinformatics design and experimental validation of Influenza A virus multi-epitopes that induce neutralizing antibodies. *Arch Virol* 165:891–911. <https://doi.org/10.1007/s00705-020-04537-2>
- Ras-Carmona A, Pelaez-Prestel HF, Lafuente EM, Reche PA (2021) BCEPS: a web server to predict linear B cell epitopes with enhanced immunogenicity and cross-reactivity. *Cells*. <https://doi.org/10.3390/cells10102744>
- Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M (2020) NetMHCpan-4.1 and NetMHCIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* 48:W449–W454. <https://doi.org/10.1093/nar/gkaa379>
- Rostaminia S, Aghaei SS, Farahmand B, Nazari R, Ghaemi A (2021) Computational design and analysis of a multi-epitope against Influenza A virus. *Int J Pept Res Ther* 27:2625–2638. <https://doi.org/10.1007/s10989-021-10278-w>
- Sanchez-Trincado JL, Gomez-Perosanz M, Reche PA (2017) Fundamentals and methods for T- and B-cell epitope prediction. *J Immunol Res*. <https://doi.org/10.1155/2017/2680160>
- Shao W, Li X, Goraya MU, Wang S, Chen JL (2017) Evolution of Influenza A virus by mutation and re-assortment. *Int J Mol Sci*. <https://doi.org/10.3390/ijms18081650>
- Sheikh QM, Gatherer D, Reche PA, Flower DR (2016) Towards the knowledge-based design of universal influenza epitope ensemble vaccines. *Bioinformatics* 32:3233–3239. <https://doi.org/10.1093/bioinformatics/btw399>
- Stewart GL, Wood B, Boley RB (1985) Modulation of host response by *Trichinella pseudospiralis*. *Parasite Immunol* 7:223–233. <https://doi.org/10.1111/j.1365-3024.1985.tb00072.x>

- Suri S, Dakshnamurthy S (2022) IntegralVac: a machine learning-based comprehensive multivalent epitope vaccine design method. *Vaccines*. <https://doi.org/10.3390/vaccines10101678>
- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680. <https://doi.org/10.1093/nar/22.22.4673>
- Wang P, Sidney J, Dow C, Mothé B, Sette A, Peters B (2008) A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput Biol* 4:e1000048. <https://doi.org/10.1371/journal.pcbi.1000048>
- Wang P, Sidney J, Kim Y, Sette A, Lund O, Nielsen M, Peters B (2010) Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinform* 11:568. <https://doi.org/10.1186/1471-2105-11-568>
- Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 46:W296–W303. <https://doi.org/10.1093/nar/gky427>
- Xia YL, Li W, Li Y, Ji XL, Fu YX, Liu SQ (2021) A deep learning approach for predicting antigenic variation of Influenza A H3N2. *Comput Math Methods Med*. <https://doi.org/10.1155/2021/9997669>
- Yuan S, Chan HCS, Hu Z (2017) Using PyMOL as a platform for computational drug design. *Wiley Interdiscip Rev Comput Mol Sci*. <https://doi.org/10.1002/wcms.1298>
- Zeller MA, Gauger PC, Arendsee ZW, Souza CK, Vincent AL, Anderson TK (2021) Machine learning prediction and experimental validation of antigenic drift in H3 Influenza A viruses in swine. *Msphere*. <https://doi.org/10.1128/mSphere.00920-20>

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

---

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)

---