# Improving in silico prediction of influenza vaccine effectiveness by genome analysis incorporating epitope information

**Authors:**

Lirong Cao[1,2], Jingzhi Lou[1], Qi Li[1,2], Hong Zheng[1,2], Chris Ka Pun Mok[1,3], Zigui Chen[4], Renee Wan Yi Chan[5,6], Peter Pak Hang Cheung[7], Marc Ka Chun Chong[1,2], Eng Kiong Yeoh[1,8], William Ka Kei Wu[9,10], Jun Yu[10,11], Paul Kay Sheung Chan[5,12], Benny Chung-Ying Zee[1,2], and Maggie Haitian Wang[1,2*]

**Affiliations:**

[1] JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong SAR, China

[2] CUHK Shenzhen Research Institute, Shenzhen, China

[3] Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China.

[4] Department of Microbiology, The Chinese University of Hong Kong, Hong Kong SAR, China

[5] Department of Paediatrics, The Chinese University of Hong Kong, Hong Kong SAR, China

[6] Hong Kong Hub of Paediatric Excellence, The Chinese University of Hong Kong, Hong Kong SAR, China

[7] Department of Chemical Pathology, CUHK; Hong Kong SAR, China.

[8] Centre for Health Systems and Policy Research, CUHK; Hong Kong SAR, China.

[9]Department of Anaesthesia and Intensive Care, The Chinese University of Hong Kong, Hong Kong Special Administrative Region of China.

[10]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong Special Administrative Region of China.

[11]State Key Laboratory of Digestive Diseases, Institute of Digestive Diseases and Department of Medicine and Therapeutics, CUHK-Shenzhen Research Institute, Shenzhen. The Chinese University of Hong Kong, Hong Kong Special Administrative Region of China.[12] Stanley Ho Centre for Emerging Infectious Diseases, CUHK; Hong Kong SAR, China.

**Correspondence:** Maggie Haitian Wang, email: maggiew@cuhk.edu.hk, JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong SAR, China.

## Abstract

Fast evaluation of vaccine effectiveness (VE) is valuable for facilitating vaccine development and making vaccination strategy. In previous studies, we developed the computational model linking molecular variations and VE for the influenza and COVID-19, through which VE prediction prior to mass vaccination and infection is possible. In this study, we perform a complete survey of the predictive effect of major functional regions of the influenza virus for VE. Interestingly, we found that the genetic distance measured on the antigenic sites being also the effective mutations for epidemics is a strong predictor for influenza VE. Based on the identified optimal predictor codon set, we develop the improved VE-Genetic Distance model for influenza (VE-GD flu). The prediction accuracy of the new model is R-square 87.1% for H3N2 ($p$-value < 0.001) on VE data of the United States. Leave-one-out cross validation shows that the concordance correlation coefficient of the predicted and observed VE is 90.6% (95% CI: 73.8-96.9). Significant prediction improvement is also found for pH1N1. Accurate prediction of influenza VE before vaccine deployment may facilitate reverse vaccinology to optimize vaccine antigen design and government preparedness of influenza epidemics.

***Keywords***: Influenza; vaccine effectiveness; prediction; genetic distance; genetic mismatch; sequence analysis

## Introduction

Influenza causes numerous morbidity and mortality annually worldwide[1]. Many countries have implemented influenza vaccination programmes to reduce disease burden due to influenza epidemics. However, the performance of influenza vaccines varies by vaccine match to circulating viruses each year. While the protective effect of vaccine can be assessed by vaccine efficacy through randomized control trials (RCTs) or vaccine effectiveness (VE) in observational studies[2-4], rapid evaluation of VE before mass vaccination can provide timely information for vaccine design and policy-makers to evaluate epidemic risk and design optimal public health strategies.

Previously, we developed a computational framework to predict VE for the influenza[5,6] and COVID-19[7] by analysis of virus sequencing data. We found that genetic distance (GD) between vaccine strains and circulating viruses is significantly correlated with vaccine effectiveness (VE). The key to model the relationship between VE and GD lies in the appropriate quantification of GD. We found that GD measured on the effective mutations (EMs) showed good prediction performance of VE, which is composed of mutations significantly associated with epidemic level[8]. Evident linear relationships were identified between VE and GD at the EM sites for influenza A pH1N1, H3N2, and influenza B and were validated in independent data[5]. Given that mutations at antigenic sites or receptor binding domains were biologically responsible for major antigenic drift of the influenza virus[9-11], we hypothesize that appropriately incorporating their information may further improve the VE prediction on top of the current EM-based model. Therefore, in this study, we conducted a complete survey of genetic mismatch evaluated on the epitope regions on the hemagglutinin (HA) and neuraminidase (NA) proteins and EM sites. We also disentangle the effects of genetic escape at HA and NA on VE through mediation analysis.

The outcome of this study is quantified measures of genetic escape influencing vaccine protection and improved VE prediction model, ready-to-use for early evaluation of influenza VE of a given vaccine strain against any composition of circulating virus population.

## Methods

### Vaccine effectiveness data and viral sequences

The VE data were extracted from published epidemiological studies[12,13]. The VE reports included in the analysis satisfy the following inclusion criteria: (1) the study applied a test-negative case control design or cohort to evaluate VE; (2) used RT-PCR test to confirm influenza infections; (3) the target population is all-age group; (4) the primary outcome is influenza-like illness; (5) the study included vaccine effectiveness against influenza A. Since surveillance of influenza VE in the North America is systematic and continuous, we included all VE observations surveyed in the US and Canada meeting

the inclusion criteria. A complete summary of the VE data was described in the supplementary materials **Tables S1-2.**

Virus genetic sequencing data were downloaded from the NCBI[14] and GISAID EpiFlu databases[15]. All available sequences of influenza A H1N1pdm09 and H3N2 that matched to the flu seasons and locations of the observational studies were extracted, consisted of a total number of 3,974 hemagglutinin (HA) and neuraminidase (NA) protein sequences between 2009/10 and 2018/19 flu seasons. Multiple sequence alignment was performed by MEGA (7.0.26). The influenza vaccine strains recommended by the WHO for use in the Northern Hemisphere were summarized in the Supplement Materials **Table S3**.

**Antigenic sites (AS)**

Antigenic sites or epitopes of HA and NA for H1N1 and H3N2 viruses were collected from literature[16-22] , which are identified through experimental approach and resolving virus crystal structure (**Table S4)**.

**Effective mutations (EMs)**

The effective mutations (EMs) are a set of mutations that is significantly associated with influenza epidemic levels in a given geographical region. The EM-test was introduced in previous studies with applications on the H3N2 virus[8,23]. Herein, we give a brief description of the method. The EM framework aims to account for the time-dependency of mutation effect for evading population immunity when evaluating their statistical importance. In contrast to the traditional association tests that regard all mutations to have constant effect through time, the EM test simultaneously estimates the relevancy of a genomic position and the effective time period of a mutation. Location-wise, an EM needs to meet the criteria that its mutation prevalence has reached a "dominance threshold" (e.g. 0.5) in population; while temporally, the dominant mutations are also expected to act within its effective mutation period, such that the mutation will not always be accounted as a contributor for escaping population immunity which has already adapted over time.

Let $x_{jk}(t)$ denote the nucleotide or amino acid residue at genome position *j* and time *t*. $j \in J$ , where *J* is full sequence length. Let $p_j(t)$ denote mutation prevalence of $x_j(t)$. Let $t_j^0$ denote the time when $p_j(t^0)$=0 and $p_j(t^0+1) > 0$. Let θ denote the threshold of mutation prevalence, and $t_j^\theta$ is the earliest time point when $p_j(t) = \theta$. The EMs are defined by the following equation,

$$m_j(\theta, h, t) \triangleq I\{t_j^0 \leq t \leq t_j^\theta + h, j \in J\}. \qquad \textbf{Eq. 1}$$

Eq. 1 identifies the EMs with a period, which begins with the first non-zero prevalence of the mutation and ends after $h$ intervals following the mutation reaching a prevalence threshold $\theta$. The $(\theta, h)$ can be estimated by fitting a linear model between the epidemic level y(t) and an aggregated

mutation activity metric, the g-measure $(\theta, h, t) = \sum_{j,k} m_{jk}(\theta, h, t) p_{jk}(t)$, over a range of $(\theta, h)$ values, controlling for the covariates including temperate, absolute humidity and season. y(t) in this study is sero-positivity rate of the influenza subtypes by epidemic season, that is, from 1-Oct to 30-Apr. $\theta \in (0,1)$ and $\mathrm{h} = \{0,1,2, \ldots\}$. **Table S5** gives the complete set of fitted values.

The list of EMs identified in the United States can be found in **Table S5**. The AS and EM locations on stereo view of HA and NA protein structures were plotted in **Figures S1-4**, obtained from the RSCB Protein Data Bank (PDB) (https://www.rcsb.org). The summary of AS and the EM sites used in this study is provided in **Figure 1**. These sites serve as candidate codons for exploring their predictive effect for influenza VE. *P*-value <0.01 is considered as statistically significant.

**The VE-GD model for influenza virus (VE-GD flu)**

Let $W(\cdot)$ denote a set of predictor codon sites, and *s* denote the protein of concern. $s = \{HA, NA\}$. For a vaccine study *l*, the average genetic mismatch from vaccine strain $\{v_j^l\}$ to the circulating viruses on gene *s* corresponding to the study period can be obtained by,

$$d^l(s) = \sum_{i=1}^{n_l} \sum_{j \in W \cap s} I\left(v_j^l \neq x_j^l\right) / n_l, \qquad \textbf{Eq. 2}$$

where $n_l$ is the number of circulating sequences matched to study *l*. To integrate the genetic mismatch from multiple proteins, we utilize the polygenic risk score (PRS) method[24], which is a widely applied approach in human genetics for disease risk prediction[25,26]. First, the effect of protein s is evaluated by the linear regression,

$$VE^l = \beta_0 + \beta_s d^l(s) + \varepsilon, \qquad \textbf{Eq. 3}$$

obtaining the estimated coefficient $\hat{\beta}_s$. Next, the integrated genetic distance (GD) of HA and NA is obtained by,

$$GD^l = \sum_s \hat{\beta}_s d^l(s) / \sum_s \hat{\beta}_s. \qquad \textbf{Eq. 4}$$

Thus, the VE can be estimated by fitting the linear regression model,

$$VE^l = \beta_0 + \beta_1 GD^l + \varepsilon, \qquad \textbf{Eq. 5}$$

where $\varepsilon$ is the random error; $l \in L$, and $L$ is the total number of VE values. **Eq. 5** captures the relationship between VE and GD for a specific influenza virus subtype. Related VE-GD framework for predicting overall VE involving multiple subtypes is described in Cao et al[5], and for predicting VE from multiple technology platforms is described in Cao et al[7].

**Model evaluation**

Prediction accuracy is measured by the mean absolute error (MAE) between the predicted and observed VE values, as well as by the concordance correlation coefficient (CCC)[27]. Leave-one-out

cross-validation (LOOCV) setting is adopted in which the testing data is independent from the model training data.

### Mediation analysis

With mediation analysis, we investigate the relationship between HA and NA mutations leading to decreased VE. We estimate the effect of HA mutations mediated by the co-mutations occurred on NA protein (referred to as an "indirect effect" in causal mediation terminology), and the drop in VE introduced via HA-only mutations through other immune mechanisms ("direct effect" not mediated by NA mismatch). All analyses was carried out in **R** statistical software[28] (version 4.0.3) and the R package mediation was used[29].

## Results

### Relevancy of mutation escape in alternative codon groups for influenza VE by gene

We first evaluate the relevancy of mutation escape in candidate codon sets for VE by gene, and in the following section evaluate the effect of the integrated genetic measure of HA and NA. We found that among all the combinations of epitope sites on the HA and NA and the EMs, the codon set composed of antigenic sites (AS) being also the effective mutations (EMs) is consistently the best estimator for VE of both H3N2 and pH1N1 ($p$-value < 0.01, **Table 1**). Among the five epitopes A-E of HA of H3N2, genetic distance on epitope A and B showed the strongest association among the epitope regions to the reduction of VE for H3N2 ($p$-value <0.01, **Table 1**). This finding is in agreement with the previous understanding of biological roles of these two epitopes towards antigenic drift[30-33]. The EMs are mutations that are found to be significantly associated with epidemic level in the United States during 2003/04-2018/19, which infers potential roles of these mutations in escaping population immunity. Interestingly, we found that when the EMs are not considered, genetic distance on the HA epitopes alone is not predictive to VE for H3N2 (R-sq = 0, $p$ = 0.95) and is only weakly associated with VE for pH1N1 (R-sq = 41%, $p$ =0.018); however, when genetic distance is evaluated on the AS being EMs, significant relationships are observed (**Fig 2a-b** for H1N1 and **Fig 2d-e** for H3N2). Scatter plots of the relationship of VE and all codon sets can be found in Supplementary Materials **Figures S5-12**.

### Relationship of influenza VE and the integrative genetic distance of HA and NA

Next, we estimate the relationship of VE and the integrated genetic distance (GD) of HA and NA using the predictor codon set composed of EM-AS. We found that the GD explains 87.1% ($p$ < 0.001, **Fig 2c**) variations of VE for the H3N2 and 55% ($p$ = 0.0037, **Fig 2f**) of the variations of VE for pH1N1. For each genetic mismatch occurred in the EM-AS loci, the expected VE reduction is 18.5% [95% CI: 14.4-22.5] for the H3N2 and 10.5% [95% CI:4.9-16.1] for pH1N1. Therefore, per mutation

impact on VE for the H3N2 is 1.8-fold more than that of pH1N1, indicating the particular challenge of vaccine strain selection for the H3N2 virus. In the ideal situation when a perfect vaccine match is achieved, the expected VE for the H3N2 is 55.7%, lower than the expected VE 68.3% for the pH1N1. The ceiling of VE of the two subtypes might be explained by the higher genetic diversity of the H3N2 virus comparing to the H1N1[34,35], such that it is more difficult to identify a single presentative strain for the entire virus population of the H3N2.

**Model validation by independent data**

First, leave-one-out cross validation was conducted for the VE-GD flu model. The average distance between the observed values and the predicted VE was 7.1% for H1N1 and 6.3% for H3N2. The calibration plot of **Figure 3** depicts that the validation data points are surrounding the 45° line, meaning that the predictions are nearly identical to the observed VE. The test statistic CCC was 90.6% (95% CI: 73.8-96.9) for the H3N2 model and 60.2% (95% CI: 13.3-85.1) for the H1N1 model. In **Figure 2**, the independent data from Canada also show an obvious negative relationship between GD and VE against pH1N1 and H3N2, closely clustering around the fitted line of US models for both single protein and two protein models. In addition, we conducted retroactive predictions for 2019/20 to 2022/23 flu seasons (**Figure 4**, supplementary materials **Table S6)** and found that the predicted H3N2 VE for US was 33.5% (95% prediction interval: 19.2-47.8), close to the observed value of 36% (95% CI: 20-49) in 2021/22 flu seasons[36]. In the next flu season, the estimated H3N2 VE was 49.0% (95% prediction interval: 34.0-64.1), which is in close agreement with the reported effectiveness of 45% (95% CI: 29-58) by US CDC[37], indicating good predictive power of the model.

**Mediation analysis of genetic factors on vaccine effectiveness**

Using the mediation analysis, we estimated that the average direct effect (ADE) of HA excluding the effect of NA is 4.7% (95% CI: 1.5 – 8.2, $p$ =0.01) decrease in VE for H3N2 per mismatch; and the indirect effect of HA mediated through the NA is 6.3% (95% CI: 1.5 – 10.6, $p$ =0.02) drop in VE per mismatch (**Figure 5**). The indirect effect due to HA-NA mismatch pathway accounts for 57% of the total effect of HA on VE. This result indicates the importance of considering HA and NA jointly in the optimization of influenza vaccine antigen for the H3N2. Currently, no significant indirect effect has been detected for pH1N1 virus.

## Discussion

The use of GD to predict vaccine protection can help to evaluate the impact of influenza immunization programme before mass vaccination and inform vaccination policy details, such as determining whether to introduce a particular vaccine in a specific population or identifying preferred vaccine products for tropical or subtropical regions. In this study, we optimized GD measure for HA

and NA by feature selection and found that GD on the intersection set of EM and AS is highly predictive of VE of influenza H1N1 and H3N2. However, the remaining AS but not EM sites showed few potentials in VE evaluation. The genetic mismatch on these AS but non-EM sites of HA and NA was consistently uncorrelated with VE of both influenza A subtypes. While literature reported amino acid substitutions emerged on HA or NA may weaken vaccine protection[38,39], few studies can quantify their effect on VE. This study bridges the molecular-level variations with population-level immunity and refines the quantification method of genetic distance in order to *in silico* predict the effectiveness of influenza vaccines in real time.

This study demonstrated that immunodominant antigenic epitopes are the optimal target to measure GD. Previous research reported that residues in epitopes A and B on HA of influenza H3N2 viruses are important in binding antibodies and genetic variants in these epitopes can help escape antibody response[30-33]. In line with previous findings, our results showed that genetic mismatch on epitopes A and B is associated with VE as amino acid changes on these epitopes could result in vaccine-mismatched influenza viruses predominant thereby causing a reduction in vaccine protection. Similar results are found for the H1N1pdm09 virus that the combined mismatch of sites Sa and Sb on HA can link to VE. The research on selective pressure also demonstrated that positively selected codons were associated with antibody combining epitopes A and B, while not all codons in these sites had predictive value for the evolution of human influenza A and only a few codons on epitopes C, D and E were under positive selection[40], which also support our findings.

Our previous research developed a novel bioinformatics approach to identify key residues relevant to escape from herd immunity, namely EM sites, and demonstrated the feasibility of using these sites to assess VE but failed for NA of H3N2. This work successfully determined genetic measure for H3N2 NA. While genetic mismatch on AS of NA showed the best performance, the superiority compared to using EM-AS codons is relatively small with a 9.1% difference in terms of the goodness of fit, which also proves the importance of EM-AS in VE evaluation. Previous research also tried to use the dominant epitope as that which shows the most antigenic drift to estimate VE[41-43]. However, a precise determination of which epitopes are dominant was not completed. This study provides useful evidence for identifying key epitopes and measuring vaccine matching levels.

This work can also inform ongoing research efforts to understand the role of NA in the effect of HA mutations on VE. The existing evidence showed that NA-induced antibodies could also provide protection against influenza viruses[44,45] and mutations on NA can also affect viral antigenicity[46]. Inclusion of HA and NA in influenza vaccines provides enhanced protection against influenza-associated illness compared to HA alone[47-49]. From the perspective of genetic epidemiology, we uncovered the potential mechanisms underlying the association of HA mismatch with VE that a significant mediating effect of NA mismatch existed for H3N2.

The limitations of this study include that only a few countries annually investigate the effectiveness of influenza VE and thus it is still a challenge to customize a prediction model for each region. Moreover, numerous studies reported that pre-existing immunity from previous infection and vaccinations can also affect vaccine protective effects[50-55]. The antigenic distance (AD) hypothesis proposed that when prior and current vaccines are antigenically distinct, the influence of the former vaccination should be minimal[56]. When their AD is smaller, positive interference to the current vaccine protection is expected when the AD between prior vaccines and epidemic strains is small, while the interference becomes negative when the AD between pre-existing immunity and circulating viruses is large. In future studies, we will attempt to quantify the impact of pre-existing immunity on VE as data on immune history increases. Despite the limitations, the relationship between GD and VE is deterministic and the key mutation positions involved in this study are also well demonstrated in plenty of biology experiments and bioinformatics analyses. Improved quantification of GD based on EM-AS codons could assist real-time prediction of VE and candidate vaccine strain selection. The predictive models for VE developed herein may facilitate future vaccine deployment and resource coordination.

## Acknowledgements

We thank the contributions of all the health care workers and scientists, the GISAID team, and the submitting and the originating laboratories. This work was partially supported by the Strategic Seed Funding for Collaborative Research Scheme [3136001], the Chinese University of Hong Kong Direct Grant [2022.02], and NSFC Excellent Young Scientists Fund [32322088]. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

## Author Contributions

M.H.W conceived the study. L.C processed data, carried out the analysis and wrote the manuscript. J.L., Q.L., H.Z., C.K.P.M, Z.C., R.W.Y.C., P.P.H.C., M.K.C.C., E.K.Y., P.K.S.C., and B.C.Y.Z. critically read and revised the manuscript and gave final approval for publication.

## Competing interests

M.H.W and B.C.Y.Z are shareholders of Beth Bioinformatics Co., Ltd. B.C.Y.Z is a shareholder of Health View Bioanalytics Ltd. All other authors declare no competing interests.

## References


1. WHO. Influenza (Seasonal). Retrieve from https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal). (2023).
2. Smith, P.G., Rodrigues, L.C. & Fine, P.E. Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies. *Int J Epidemiol* **13**, 87-93 (1984).
3. Halloran, M.E., Struchiner, C.J. & Longini, I.M., Jr. Study designs for evaluating different efficacy and effectiveness aspects of vaccines. *Am J Epidemiol* **146**, 789-803 (1997).
4. De Serres, G., Skowronski, D.M., Wu, X.W. & Ambrose, C.S. The test-negative design: validity, accuracy and precision of vaccine efficacy estimates compared to the gold standard of randomised placebo-controlled clinical trials. *Euro Surveill* **18**(2013).
5. Cao, L., *et al.* In silico prediction of influenza vaccine effectiveness by sequence analysis. *Vaccine* **39**, 1030-1034 (2021).
6. Cao, L., *et al.* Differential Influence of Age on the Relationship between Genetic Mismatch and A(H1N1)pdm09 Vaccine Effectiveness. *Viruses* **13**(2021).
7. Cao, L., *et al.* Rapid evaluation of COVID-19 vaccine effectiveness against symptomatic infection with SARS-CoV-2 variants by analysis of genetic distance. *Nat Med* **28**, 1715-1722 (2022).
8. Wang, M.H., *et al.* Characterization of key amino acid substitutions and dynamics of the influenza virus H3N2 hemagglutinin. *J Infect* **83**, 671-677 (2021).
9. Lee, M.S. & Chen, J.S. Predicting antigenic variants of influenza A/H3N2 viruses. *Emerg Infect Dis* **10**, 1385-1390 (2004).
10. Smith, D.J., *et al.* Mapping the antigenic and genetic evolution of influenza virus. *Science* **305**, 371-376 (2004).
11. Koel, B.F., *et al.* Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science* **342**, 976-979 (2013).
12. US CDC. CDC Seasonal Flu Vaccine Effectiveness Studies. Retrieved from https://www.cdc.gov/flu/vaccines-work/effectiveness-studies.htm. (2023).
13. BC CDC. Canadian Sentinel Practitioner Surveillance Network (SPSN) influenza vaccine effectiveness estimates % (95% CI), 2004-05 to 2022-23 seasons. Retrieved from http://www.bccdc.ca/resource-gallery/Documents/Statistics%20and%20Research/Publications/Epid/Influenza%20and%20Respiratory/SPSN_VE_By_Year_Table.pdf. (2023).
14. Bao, Y., *et al.* The influenza virus resource at the National Center for Biotechnology Information. *J Virol* **82**, 596-601 (2008).
15. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**(2017).
16. Wiley, D.C., Wilson, I.A. & Skehel, J.J. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature* **289**, 373-378 (1981).

17. Gerhard, W., Yewdell, J., Frankel, M.E. & Webster, R. Antigenic structure of influenza virus haemagglutinin defined by hybridoma antibodies. *Nature* **290**, 713-717 (1981).

18. Wilson, I.A. & Cox, N.J. Structural basis of immune recognition of influenza virus hemagglutinin. *Annu Rev Immunol* **8**, 737-771 (1990).

19. Caton, A.J., Brownlee, G.G., Yewdell, J.W. & Gerhard, W. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). *Cell* **31**, 417-427 (1982).

20. Brownlee, G.G. & Fodor, E. The predicted antigenicity of the haemagglutinin of the 1918 Spanish influenza pandemic suggests an avian origin. *Philos Trans R Soc Lond B Biol Sci* **356**, 1871-1876 (2001).

21. Gulati, U., *et al.* Antibody epitopes on the neuraminidase of a recent H3N2 influenza virus (A/Memphis/31/98). *J Virol* **76**, 12274-12280 (2002).

22. Wan, H., *et al.* Structural characterization of a protective epitope spanning A(H1N1)pdm09 influenza virus neuraminidase monomers. *Nat Commun* **6**, 6114 (2015).

23. Zhao, S., *et al.* Quantifying the importance of the key sites on haemagglutinin in determining the selection advantage of influenza virus: Using A/H3N2 as an example. *J Infect* **81**, 452-482 (2020).

24. Purcell, S.M., *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).

25. Torkamani, A., Wineinger, N.E. & Topol, E.J. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* **19**, 581-590 (2018).

26. Lambert, S.A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum Mol Genet* **28**, R133-r142 (2019).

27. Lin, L.I. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255-268 (1989).

28. Team, R.C. R: A language and environment for statistical computing. (2013).

29. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. Mediation: R package for causal mediation analysis (2014).

30. Bush, R.M., Fitch, W.M., Bender, C.A. & Cox, N.J. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol* **16**, 1457-1465 (1999).

31. Cox, N.J. & Bender, C.A. The molecular epidemiology of influenza viruses. *Seminars in Virology* **6**, 359-370 (1995).

32. Beer, K., *et al.* Characterization of neutralizing epitopes in antigenic site B of recently circulating influenza A(H3N2) viruses. *J Gen Virol* **99**, 1001-1011 (2018).

33. Ndifon, W., Wingreen, N.S. & Levin, S.A. Differential neutralization efficiency of hemagglutinin epitopes, antibody interference, and the design of influenza vaccines. *Proc Natl Acad Sci U S A* **106**, 8701-8706 (2009).

34. Su, Y.C.F., *et al.* Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. *Nat Commun* **6**, 7952 (2015).

35. Ma, Y., *et al.* The Phylodynamics of Seasonal Influenza A/H1N1pdm Virus in China Between 2009 and 2019. *Front Microbiol* **11**, 735 (2020).

36. Price, A.M., *et al.* Influenza Vaccine Effectiveness Against Influenza A(H3N2)-Related Illness in the United States During the 2021-2022 Influenza Season. *Clin Infect Dis* **76**, 1358-1363 (2023).

37. US CDC. Preliminary Flu Vaccine Effectiveness (VE) Data for 2022-2023. Retrieved from https://www.cdc.gov/flu/vaccines-work/2022-2023.html. (2023).

38. Tenforde, M.W., *et al.* Effect of Antigenic Drift on Influenza Vaccine Effectiveness in the United States-2019-2020. *Clin Infect Dis* **73**, e4244-e4250 (2021).

39. Gao, J., *et al.* Antigenic Drift of the Influenza A(H1N1)pdm09 Virus Neuraminidase Results in Reduced Effectiveness of A/California/7/2009 (H1N1pdm09)-Specific Antibodies. *mBio* **10** (2019).

40. Bush, R.M., Bender, C.A., Subbarao, K., Cox, N.J. & Fitch, W.M. Predicting the evolution of human influenza A. *Science* **286**, 1921-1925 (1999).

41. Bonomo, M.E. & Deem, M.W. Predicting Influenza H3N2 Vaccine Efficacy From Evolution of the Dominant Epitope. *Clin Infect Dis* **67**, 1129-1131 (2018).

42. Xu, Y. & Ye, J. Predicting Influenza Vaccine Effectiveness With Hemagglutinin Sequences Is Not Enough. *Clin Infect Dis* **68**, 346 (2019).

43. Muñoz, E.T. & Deem, M.W. Epitope analysis for influenza vaccine design. *Vaccine* **23**, 1144-1148 (2005).

44. Giurgea, L.T., Morens, D.M., Taubenberger, J.K. & Memoli, M.J. Influenza Neuraminidase: A Neglected Protein and Its Potential for a Better Influenza Vaccine. *Vaccines (Basel)* **8** (2020).

45. Eichelberger, M.C. & Monto, A.S. Neuraminidase, the Forgotten Surface Antigen, Emerges as an Influenza Vaccine Target for Broadened Protection. *J Infect Dis* **219**, S75-S80 (2019).

46. Wan, H., *et al.* The neuraminidase of A(H3N2) influenza viruses circulating since 2016 is antigenically distinct from the A/Hong Kong/4801/2014 vaccine strain. *Nat Microbiol* **4**, 2216-2225 (2019).

47. Johansson, B.E., Matthews, J.T. & Kilbourne, E.D. Supplementation of conventional influenza A vaccine with purified viral neuraminidase results in a balanced and broadened immune response. *Vaccine* **16**, 1009-1015 (1998).

48. Bosch, B.J., *et al.* Recombinant soluble, multimeric HA and NA exhibit distinctive types of protection against pandemic swine-origin 2009 A(H1N1) influenza virus infection in ferrets. *J Virol* **84**, 10366-10374 (2010).

49. Johansson, B.E. Immunization with influenza A virus hemagglutinin and neuraminidase produced in recombinant baculovirus results in a balanced and broadened immune response superior to conventional vaccine. *Vaccine* **17**, 2073-2080 (1999).

50. Skowronski, D.M., *et al.* Serial Vaccination and the Antigenic Distance Hypothesis: Effects on Influenza Vaccine Effectiveness During A(H3N2) Epidemics in Canada, 2010-2011 to 2014-2015. *J Infect Dis* **215**, 1059-1099 (2017).

51. Keitel, W.A., Cate, T.R., Couch, R.B., Huggins, L.L. & Hess, K.R. Efficacy of repeated annual immunization with inactivated influenza virus vaccines over a five year period. *Vaccine* **15**, 1114-1122 (1997).

52. McLean, H.Q., *et al*. Impact of repeated vaccination on vaccine effectiveness against influenza A(H3N2) and B during 8 seasons. *Clin Infect Dis* **59**, 1375-1385 (2014).

53. Skowronski, D.M., *et al.* A Perfect Storm: Impact of Genomic Variation and Serial Vaccination on Low Influenza Vaccine Effectiveness During the 2014-2015 Season. *Clin Infect Dis* **63**, 21-32 (2016).

54. Ohmit, S.E., *et al*. Influenza vaccine effectiveness in households with children during the 2012-2013 season: assessments of prior vaccination and serologic susceptibility. *J Infect Dis* **211**, 1519-1528 (2015).

55. Kim, J.H., *et al*. Prior infection with influenza virus but not vaccination leaves a long-term immunological imprint that intensifies the protective efficacy of antigenically drifted vaccine strains. *Vaccine* **34**, 495-502 (2016).

56. Katz, J.M., Hancock, K. & Xu, X. Serologic assays for influenza surveillance, diagnosis and vaccine evaluation. *Expert Rev Anti Infect Ther* **9**, 669-683 (2011).

57. Skowronski, D.M., *et al.* Interim estimates of 2019/20 vaccine effectiveness during early-season co-circulation of influenza A and B viruses, Canada, February 2020. *Euro Surveill* **25**(2020).

## Table

| VE | Protein | Codon sets | $\widehat{\beta}_s$ (SE) | $R^2$ | *p*-value |
|---|---|---|---|---|---|
| pH1N1 | HA | All Antigenic Sites (AS) | -7 (2.5) | 41.0% | 0.018 |
| | | Site Sa | -11.2 (7.1) | 18.5% | 0.142 |
| | | Site Sb | 2.1 (8.4) | 0.6% | 0.808 |
| | | Site Ca1 | -3.9 (6.1) | 3.6% | 0.534 |
| | | Site Ca2 | -0.5 (28.5) | 0 | 0.986 |
| | | Site Cb | -4.3 (9.3) | 1.9% | 0.653 |
| | | AS excluding EM | -1.2 (6.3) | 0.4% | 0.847 |
| | | EM | -2.3 (0.8) | 41.7% | 0.017 |
| | | **EM-AS** | **-8.4 (2.7)** | **47.3%** | **0.009*** |
| | NA | AS | -9.9 (2.9) | 51.7% | 0.006* |
| | | AS excluding EM | -16.5 (22.5) | 4.6% | 0.479 |
| | | EM | -2.2 (0.7) | 51.2% | 0.006 |
| | | **EM-AS** | **-10.6 (3.1)** | **52.4%** | **0.005*** |
| H3N2 | HA | All Antigenic Sites (AS) | -0.1 (2.0) | 0 | 0.953 |
| | | Epitope A | -28.6 (8.5) | 48.5% | 0.006* |
| | | Epitope B | -9.1 (2.2) | 57.7% | 0.002* |
| | | Epitope C | 14.3 (6.3) | 30.1% | 0.042 |
| | | Epitope D | 7.3 (3.9) | 22.8% | 0.084 |
| | | Epitope E | 8.9 (5.2) | 19.4% | 0.116 |
| | | AS on epitope A and B | -7.7 (1.7) | 61.5% | < 0.001* |
| | | AS excluding EM | -11 (5.8) | 23.0% | 0.08 |
| | | EM | -6.7 (2.0) | 49.8% | 0.005* |
| | | **EM-AS** | **-11 (2.3)** | **65.1%** | **< 0.001*** |
| | NA | AS | -19.8 (2.7) | 81.6% | < 0.001* |
| | | AS excluding EM | 47.1 (75.3) | 3.1% | 0.543 |
| | | EM | 1.1 (2.5) | 1.6% | 0.667 |
| | | **EM-AS** | **-20.2 (2.7)** | **81.8%** | **< 0.001*** |

**Table 1**. **Relevancy of genetic escape on alternative codon sets for VE**

Legend: *p*-value is significance of the coefficient. * $p < 0.01$. R-square is goodness-of-fit of Eq. 3 in the training data of the United States.

# Figures

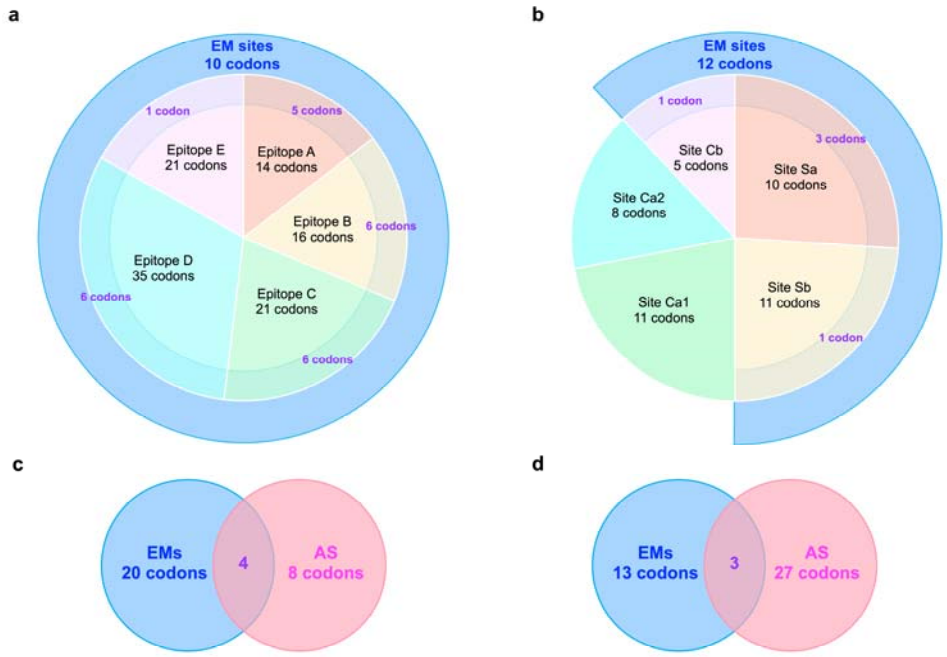


**Figure 1. Venn diagram for comparing antigenic sites (AS) and effective mutation (EM) sites.**

We compared EM sites and AS and identified the intersection codons of two genetic regions. The list of AS and EM sites are provided in supplementary materials Table S4 and Table S5, respectively.

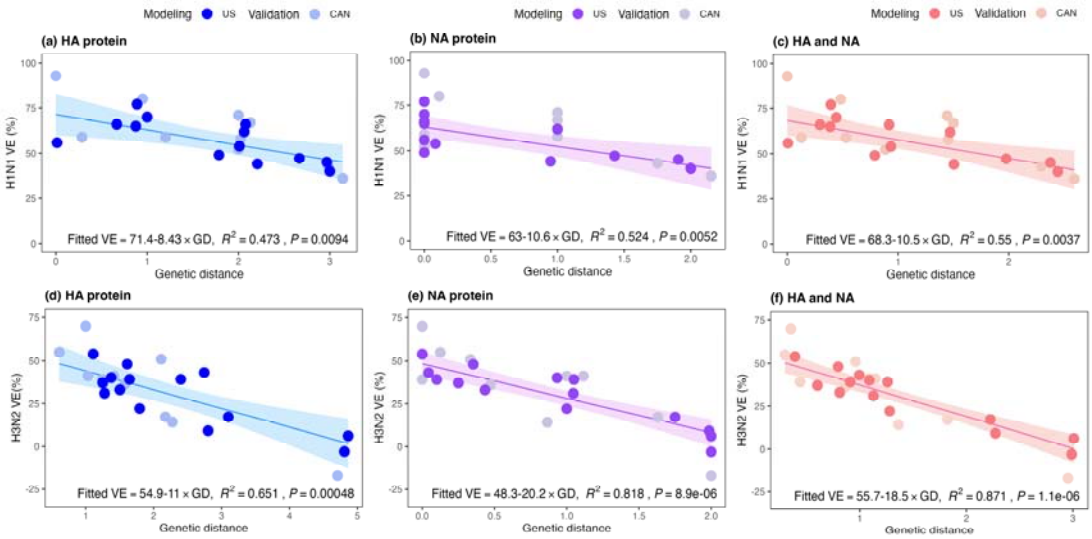


**Figure 2. Relationship between vaccine effectiveness (VE) and genetic distance (GD) on the antigenic sites (AS) being also effective mutations (EMs).**

Legend: Evaluation the relationship between VE and GD calculated on the best predictor codon set, EM-AS. Panels show the analysis by gene: (a) HA of H1N1, (b) NA of H1N1, (c) HA and NA of H1N1; (d) HA of H3N2, (d) NA of H3N2, and (e) HA and NA of H3N2. Dark dot: training data of the United States; light dot: validation data of Canada

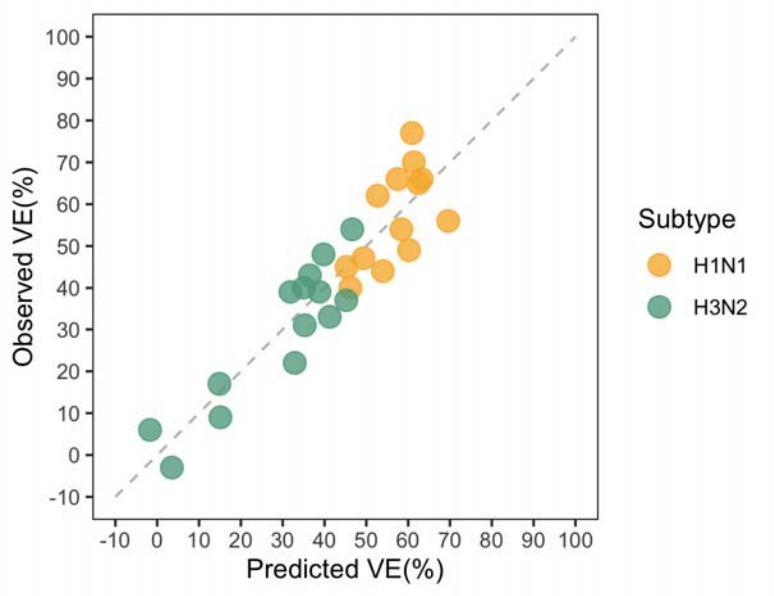


**Figure 3. Calibration plot for the prediction accuracy in leave-one-out cross validation.**

Data points near the dashed line mean that the predicted value is close to the observed VE.

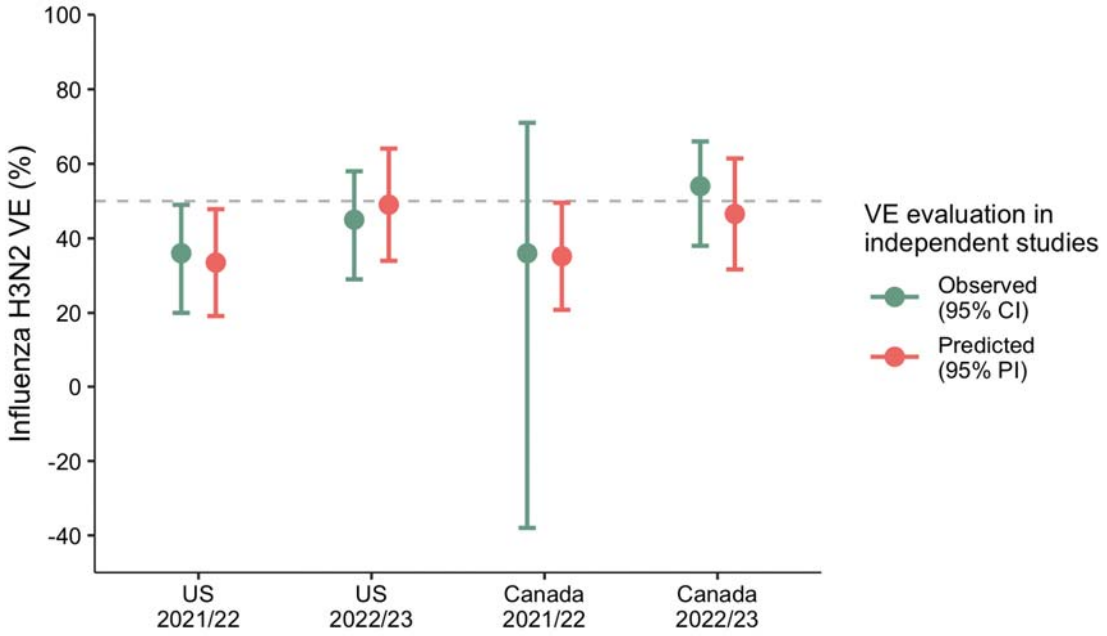


**Figure 4. Retroactive prediction for H3N2 VE from 2019/20 to 2022/23 flu seasons.**

VEs from 2019/20 to 2022/23 flu seasons were retroactively predicted.

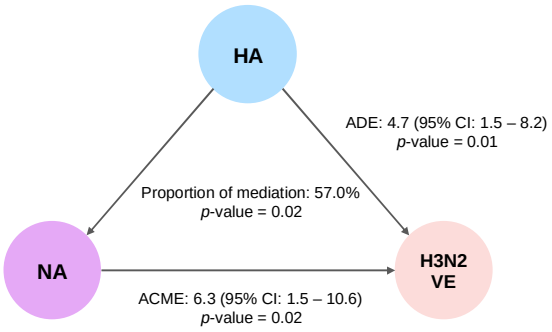


**Figure 5. Mediation analysis of the effect of HA and NA on the VE of influenza H3N2.**

ADE: average direct effect; ACME: average causal mediation effects.