Title: Predictive evolutionary modelling for influenza virus by site-based dynamics of mutations

Authors: Jingzhi Lou[1,2]†, Weiwen Liang[3]†, Lirong Cao[1,4]†, Shi Zhao[1,4], Zigui Chen[5], Renee Wan Yi Chan[6,7], Peter Pak Hang Cheung[8], Hong Zheng[1], Caiqi Liu[1], Qi Li[1], Marc Ka Chun Chong[1,4], Yexian Zhang[1,4], Eng-kiong Yeoh[1,9], Paul Kay-Sheung Chan[5,10], Benny Chung Ying Zee[1,4], Chris Ka Pun Mok[1,11]*, Maggie Haitian Wang[1,4]*

Affiliations:

[1] JC School of Public Health and Primary Care (JCSPHPC), the Chinese University of Hong Kong (CUHK); Hong Kong SAR, China.

[2] Beth Bioinformatics Co. Ltd; Hong Kong SAR, China.

[3] HKU-Pasteur Research Pole, School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong; Hong Kong SAR, China.

[4] CUHK Shenzhen Research Institute, Shenzhen, China.

[5] Department of Microbiology, CUHK; Hong Kong SAR, China.

[6] Department of Paediatrics, CUHK; Hong Kong SAR, China.

[7] Hong Kong Hub of Paediatric Excellence, CUHK; Hong Kong SAR, China.

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

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

[10] Stanley Ho Centre for Emerging Infectious Diseases, CUHK; Hong Kong SAR, China.

[11] Li Ka Shing Institute of Health Sciences, Faculty of Medicine, CUHK, Hong Kong SAR, China.

† These authors contributed equally to this work.

* Corresponding authors. Emails: kapunmok@cuhk.edu.hk, maggiew@cuhk.edu.hk

Abstract: A predictive evolutionary model was developed to forecast representative influenza viral strains and select vaccine strains for upcoming epidemic seasons. Influenza virus continuously evolves to escape human adaptive immunity and generates seasonal epidemics. A computational approach beth-1 that models site-wise mutation dynamics demonstrated remarkable matching of predicted strains to the circulating viruses in subsequent seasons for the influenza A(H1N1)pdm09 and A(H3N2) viruses in both retrospective and prospective validations. The method offers a promising and ready-to-use tool to facilitate vaccine strain selection for the influenza virus, achieved by capturing heterogeneous evolutionary dynamics over genome space-time and linking molecular variants to population immune response.

One-Sentence Summary: A computational model predicts virus evolution and facilitates vaccine strain selection for the influenza virus

Main Text:

The major driver of recurrent influenza epidemics is fast virus evolution that enables the influenza virus to escape from human immunity acquired from prior vaccination or infection[1]. In response, influenza vaccines need to be updated annually to match the circulating virus population[2]. Prediction of virus evolution has a critical role in ensuring the protective effect of influenza vaccines, which aids the selection of candidate vaccine strains nearly a year ahead before the arrival of next epidemic season[2,3]. We present a computational approach to predict influenza virus evolution through modelling the dynamic process of mutation adaptation at individual sites and locating the optimal wild-type strains by a combined evaluation of multiple gene segments for considerations of vaccine strains.

Virus evolution is shaped by a complex interplay of genetic mutations, host immune response, and epidemiology[1,4]. Thus, although mutation is stochastic[5,6], the evolutionary process could be traced with the observed viral genetic and antigenic profiling in the host population[6]. The influenza virus evolves through antigenic drift on the two surface proteins, hemagglutinin (HA) and neuraminidase (NA), the primary immuno-active components of influenza vaccines. Previous studies showed that mutations at epitope sites of the HA played a dominant role in characterizing virus antigenic change and were under higher immune selection pressure[5,7]. In addition to the major antigenic substitutions, the virus evolution is also critically influenced by epistatic mutations or mutation interference effects[8-10], making predictions challenging as the evolutionary dynamics are non-uniform across genomic regions and time.

Currently, influenza vaccine strain determination involves extensive surveillance and characterization of the virus in terms of genetic, antigenic evolution, and epidemiological profiles, a global effort coordinated by the World Health Organization (WHO)[11]. While antigenic evolution of influenza virus can be mapped by the antigenic cartography based on hemagglutinin inhibition (HAI) assay data[12], genetic evolution is mostly delineated with the phylogenetic trees[13]. Current prediction methods of virus evolution mainly focus on modelling the fitness of tree parts[14-16], using sequences of the HA or HA1 segment, with epitope or antigenic data incorporated as component information[14,15,17]. For instance, Luksza and Lassig calculated frequency of tree-clades and predicted the future predominant clade by an exponential function (the Malthusian model) with a certain growth rate[14]. Neher et al. used the local branching index (LBI) to rank tree-nodes in a given phylogeny to identify lineage with the highest fitness as a progenitor of strains in an upcoming influenza season, where fitness is

estimated by integrating exponentially discounted branch-length surrounding a node[16]. Steinbruck and McHardy used allele dynamics plots to identify the top three alleles characterizing antigenic novelty of tree-clades in a given season, pre-screened based on the epitope information, HI data, and phylogeny[17,18]. Also based on the Malthusian model, Huddleston et al regarded strain rather than clade as a primary unit of analysis, and investigated different combination of factors to calculate growth rate for strains[15]. These methods projected the most likely predominant future lineages by tracing the fitness of clusters of strains. However, as the virus evolution is driven by major antigenic substitutions, a substantial amount of information is contained in the dynamic process of the fitness of mutations, characterizing which might provide critical information for predicting genetic evolution.

A site-based dynamic model for mutation forecasting. Our prediction method is primarily based on modelling the time-resolved frequency pattern of mutations for individual sites across virus genome segments (**Fig. 1**). The selective advantage of a mutation could be reflected in their growing prevalence in the host population. To make a projection into the future, the velocity of mutation frequency growth can be estimated by solving the first-order derivative of a frequency function over a period of mutation adaption in the host population, which is captured by the mutation transition time (**Methods**). Since the transition time is specific to substituted residues, the dynamic model is site-based and time-dependent. Scanning over the genome restores a global picture of mutation-selection. Such design shares the advantage of the agent-based model that agglomerates individual agents carrying simple rules and avoids fitting a complicated system with copious parameters or unrealistic assumptions[19]. The potential relationship between the genome-wide dynamics of mutation prevalence and population epidemics is illustrated in **Extended Data Fig. 1**. The site-based model is also highly computationally efficient in analyzing large genomic datasets.

Specifically, transition time in this study is defined as the duration for a mutation to emerge until it reaches an influential frequency in the population. This quantity is related to but different from the conventional fixation time, which covers a period from the first observation time of a mutation to the point when it reaches 0.99 prevalence in the population. The fixation time was reported to span over a long period of 4-32 years for the influenza virus A(H3N2)[8]. In contrast, the transition time we identified for the same viral subtype had a median length of only ~17 months and ranged between 0-7 years (**Extended Data Fig. 2**). As the transition time calibrates the initial period of mutation adaptation, it may have the advantage

of informing the emerging genetic variants on a short-term time horizon. The transition time is determined with a frequency threshold (θ) indicating fitness strength, at which the overall mutation activities are detected to influence population epidemics significantly (**Extended Data Fig. 1**); and it can be estimated using the virus epidemic-genetic association model we previously developed[20] (**Methods**).

This site-based mutation dynamic model enables the prediction for fitness of competing residues at individual sites, thereby the construction of a genome-wide fitness landscape of the virus population in future time (**Fig. 1**). Thus, a consensus strain can be shaped containing all mutations showing selective advantage relative to their precedent or competing alleles in the upcoming epidemic season. It serves as a prototype to search for the optimal wild-type virus for consideration as vaccine strain.

Identification of optimal wild-type virus. The second part of our method is the selection of optimal wild-type strain based on the predicted virus population (**Fig. 1**). The optimal wild-type virus can be located by minimizing the weighted genetic distance between a candidate strain and the projected future virus population considering one or more proteins contained in vaccine antigen **(Methods)**. Although only HA concentration is standardized in the current vaccine production[21], both HA and NA genes are major components of the influenza vaccines; an integrative evaluation would provide a useful tool for strain selection and evaluation. This is realized by our recently developed model of predicting Vaccine Effectiveness (VE) through analysis of Genetic Distance (VE-GD)[22-24], in which effect size of genetic distance of HA and NA are estimated over the seasons and used for identifying an optimal wild-type; the model is also used to imply estimated VE given a vaccine strain based on historical data (**Methods)**.

This two-step evolution prediction and wild-type virus selection method are referred to as the "beth-1" for easy reference.

Genetic matching of retroactively predicted vaccine strains. We applied the beth-1 to predict vaccine strains of the influenza A (H1N1)pdm09 (pH1N1) and A (H3N2) viruses. Data was collected from the Global Initiative on Sharing All Influenza Data (GISAID)[25] between 1999/2000 and 2022/23, involving a total number of 13,192 HA and 11,260 NA sequences of pH1N1, and 37,093 HA and 34,037 NA sequences of H3N2 from ten geographical regions in the Northern Hemisphere, covering North America (New York State,

California State, Canada), Europe (United Kingdom, Germany, France), and Asia (Hong Kong SAR, South China provinces, Japan, Singapore) (**Extended Data Table 2**). Three-year data was burnt-in for model building. The prediction was performed using data up to February in season *T*, targeting the subsequent epidemic season *T*+1 (October-April next year). We calculate the average amino acid (AA) mismatch between the predicted strain and sequences of circulating viruses in the target season to determine prediction accuracy. To fully understand the performance of beth-1 in the context of existing methods, we calculated the following comparison groups: (1) WHO-recommended vaccine strains (the "current-system") for season *T*+1; (2) The local branching index (LBI) method[16], as the representative approach based on phylogenetic trees; (3) beth-1 (single protein): the beth-1 predicted strains for season *T*+1 by a single protein; (4) beth-1 (two-protein): predicted strains for season *T*+1 integrating two proteins; (5) the "answer": the observed representative strains in season *T*+1 (**Methods**, **Extended Data Table 3-4**).

In the retroactive data, prediction was conducted for seven seasons from 2012/13 to 2018/19 for pH1N1 and 17 seasons from 2002/03 to 2018/19 for the H3N2. beth-1 demonstrated significantly improved genetic matching to the future virus population compared to the LBI and the current-system on full-length HA and NA gene, their epitopes, and for both pH1N1 and H3N2 subtypes (**Fig. 2, Extended Data Table 5**). For example, the beth-1 (HA) model results in 7.5 AAs (SD 2.2) mismatch on the full-length HA protein of H3N2, while mismatch by the LBI and current-system are 9.5 AAs (SD 4.7) and 11.7 AAs (SD 5.1), respectively (pair-wise *t*-test p-value < 0.001) (**Extended Data Table 5**). The beth-1 (NA) gives 3.9 AAs (SD 1.5) mismatch on full-length NA protein of pH1N1, significantly lower than the 6.4 AAs (SD 2.1) by the LBI and 11.6 AAs (SD 4.4) by the current-system. Using the beth-1 (two-protein) model, the mismatch on the HA epitopes is 1.2 AAs (0.6) for pH1N1 and 5.1 AAs (SD 1.7) for H3N2. Particularly, the mismatch of beth-1 (two-protein) on the NA epitopes are 0.5 AA (SD 0.4) for pH1N1 and 0.6 AAs (SD 0.5) for the H3N2, close to the best possible outcome by the answer strains. In all these results, beth-1 delivers prevailingly smaller uncertainties (standard deviation) in prediction accuracy compared to the current-system and the LBI (**Extended Data Table 5**). We anticipate the improvement in VE corresponding to the increased genetic matching using the VE-GD model for influenza virus[23,24] (**Methods)**. Taking the United States (US) as an example, the predicted strains are expected to increase the VE of existing influenza vaccines by an absolute 13.1% (95% Prediction Interval - PI: 4.5-21.7) for

the pH1N1 in the six seasons when there is a difference in VE, and 11.4% (95% PI: 3.9-18.9) for the H3N2 in the 12 seasons when there is a difference in VE. Among the three seasons when predicted strains of beth-1 differ between the two-protein and the HA-only model, the two-protein model is expected to enhance VE by absolute 9.0% (95% PI: 1.5 – 16.4).

Seasonal and geographical breakdown of prediction accuracies were also analyzed. Year-by-year comparison with the LBI and current-system showed that the beth-1 gave vastly lower genetic mismatch throughout the epidemic seasons (**Fig. 3, Extended Data Figure 3, Extended Data Table 6-9**). A separate analysis by geographical regions showed no systematic difference in vaccine mismatch across continents by all prediction methods and for the two influenza A subtypes (**Extended Data Fig. 4, Extended Data Table 10-13**).

We performed two predictions for the 2019/20 epidemic season: a prospective prediction conducted in October 2019 using training data up to June 2019; and a retroactive prediction conducted in March 2022 using data up to March 2019 to better match the current-system's timeline **(Extended Data Table 14)**. Genetic mismatch was calculated for all prediction experiments. Between the 2019-03 and 2019-06 predictions by beth-1, the latter one recorded 0.1 less HA epitope mismatches for pH1N1 and 1.1 less HA epitope mismatch for H3N2. Both predictions showed significantly better genetic matching compared to the current-system on all protein segments of pH1N1, the NA of H3N2, and non-inferior matching in other segments. We further evaluated the immunogenicity of predicted strains for this particular season using clinical samples collected in Hong Kong (**Methods, Extended Data Fig. 5**). Against the viral isolates of H1 and H3 positive samples, beth-1 strains induced significantly higher neutralizing antibodies in $PRNT_{50}$ for pH1N1 and non-inferior antibodies for H3N2, compared to the current vaccines (**Extended Data Fig. 6**).

Prospective predictions and validations in 2020/21 - 2022/23. To further validate our prediction method, we conducted prospective predictions from 2020/21 to 2022/23. The predictions were sent to the WHO before the vaccine composition meetings for the Northern Hemisphere each year (**Extended Data Table 3**). To fully understand the relationship between the beth-1 strains and the current vaccine viruses, we mapped them on phylogenetic trees (**Fig. 4**). Generally, the beth-1 predicted strains extended more into the future clusters for both influenza subtypes in all seasons of prospective validations compared to the current vaccine strains. On the epitopes of HA and NA of pH1N1 and H3N2, beth-1 demonstrated non-inferior or significantly increased genetic matching to the future circulating viruses (**Extended Data**

Table 15). For instance, in 2020/21, beth-1 gave 2.7 AAs (SD 1.1) mismatch on the HA epitopes of pH1N1 and the current-system gave 3.8 AAs (SD 1.2) (**Extended Data Table 15**). In the 2021/22 season, beth-1 resulted in 4.2 AAs (SD 1.6) mismatches on the HA epitopes of H3N2, 6.6 AAs more accurate than the current-system's 10.8 AAs (SD 1.9) mismatches (*p*-value < 0.001), which implies an expected enhancement in VE by 19.3% (95% PI: 11.7 – 26.9) to 52.8% (PI: 37.4 – 68.2) against H3N2. Note that the observed and estimated VE of the current-system for H3N2 in the US in 2021/22 are 36% (95% CI: 20 – 49)[26] and 33.5% (95% PI: 19.2 – 47.8), respectively. In the 2022/23 season, beth-1 prediction resulted in 1.0 AA (SD 0.7) mismatch on the HA epitopes of pH1N1 and the current-system gave 3.0 AAs (SD 0.7) mismatch (*p*-value <0.001).

How predictable are the pH1N1 and H3N2 viruses? Predictability for the two influenza subtypes pH1N1 and H3N2 can be compared on genetic mismatch of HA epitopes in the retrospective data. The ceiling of prediction can be approximated by the answer strain, that is, the observed representative strain in respective seasons. For the pH1N1 virus, the lowest possible mismatch of HA epitope was 0.7 AAs (SD 0.4) achieved by the answer strain, and the beth-1's mismatch was 1.2 AAs (SD 0.6) **(Extended Data Table 5)**. This suggests that a highly precise hit by predictive modelling is achievable for the pH1N1, such that the resulting vaccine strain could provide an excellent match to the circulating viruses in regions of the Northern Hemisphere. However, for the H3N2 virus, the lowest possible HA epitope mismatch was as large as 3.5 AAs (SD 1.1), that is 2.8 more epitope mismatches compared to the pH1N1. Although the 5.1 AAs (SD 1.7) mismatch by beth-1 was only 1.6 residue away from the answer, the prediction for H3N2 would be ultimately bounded by the ceiling of genetic matching. The large genetic mismatch of the H3N2 answer strain might be attributed to the high genetic diversity of this virus[27,28], making the selection of a single representative wild-type strain challenging. New vaccinology strategies, such as developing broadly reactive vaccines, designing antigens containing multiple H3N2 strains, or preparing region-specific vaccines, may provide solutions from other dimensions to enhance vaccine protection against this subtype.

Advantages of site-based dynamic model for evolution prediction. Strain-based dynamic model often requires the estimation of a single-valued growth rate for a given genome to project future fitness, whereas a site-based dynamic model is not constraint by a constant growth rate over genome space-time. This property leads to the following three advantages in

predicting virus evolution with the site-based angle. First, key mutations distributed over multiple clades could be captured as they arise, while the virus is trialing various epistatic combinations before shaping a stable lineage. Second, a site-based model avoids making assumption for directional mutation effects on fitness, which is self-evident in mutation frequency. In contrast, strain-based models often need to assume negative effect for the non-epitope mutations, to offset the genetic distance obtained from gene-based analysis in modelling evolutionary pathway[14-16], although studies suggested that the alternative might be true[29,30]. Third, the site-based dynamic model is adaptive to the altered residue fitness from epistasis and environmental factors[10,31], by sampling and re-estimating the dynamic function at each time period. When such adaptive framework is not in place, study showed that the fitness advantage estimated in the initial stage of mutation emergence cannot predict their ultimate fixation[32]. However, the proposed site-based dynamic model avoids making assumptions on the constant effect of mutations. Through adaptive estimation of fitness by genomic site and time, it projects a probable fitness landscape including all the trackable advantageous mutations to the near future.

"Representativeness" of viral strains in the perspective of strain-based and site-based analysis. Under the perspective of strain-based analysis, a representative virus is naturally indicated by the majority vote of strain or clades. While under the perspective of site-based analysis, a genome-wide fitness landscape can be constructed for a virus population, without missing a single mutation showing selective advantage. Based on this fitness landscape, a mode estimator can be operated at individual sites and generate a consensus sequence. Therefore, the consensus sequence is a natural representative strain for a virus population through the site-wise analysis of viral fitness. Beth-1 in its primary objective established the theoretical framework of modelling mutation dynamics site-wise and enabled forecasting for fitness landscape into future time.

Dissecting prediction accuracy. Since the consensus sequence serves as a good representative strain of a given virus population, by comparing the year-to-year consensus strains, we may depict the stepwise evolutionary change of virus population. For example, on the full-length HA protein of the H3N2, the average number of dominant mutations arise each year is 5.3 AA (SD 4.2), estimated from the 17 seasons. Therefore, regarding a given season's consensus strain (current-consensus) as the baseline, the LBI on average captures 1.5 AAs (28.3%) of the new dominant mutations in the upcoming season, while the beth-1 captures 2.6

AAs (49.1%). Interestingly, although the current-consensus as a predictor for future virus populations outperforms the LBI in terms of genetic mismatch[32], it whatsoever forecasts no evolutionary advancement in T+1. Rather, the prediction accuracy achieved by the current-consensus is solely contributed by capturing the center of viral cluster, which results in a smaller spread of genetic distance from a single strain to the circulating viruses. In sight of this, the genetic mismatch as a measure of prediction power is contributed from two aspects: the accuracy in forecasting evolutionary advancement of virus, and the accuracy in locating the center of the mass of viruses. The beth-1 deals with both aspects in a simple and elegant way through the site-based fitness modelling.

Limitation of the study. One major limitation in our analysis is the sparse time interval, constrained by the sequence sample size in the earlier years, which prevented us from using more complex trajectory functions for prediction. In the future, this sampling gap may be gradually closed with increasing surveillance strength and global collaborations, and non-linear functions can be explored. Second, the epitope mismatch could have been underestimated as it was subjected to the known epitope sites, especially in the immune-subdominant protein segments. Nevertheless, the comparison across prediction methods was fair with the same evaluation criteria.

Current WHO vaccine virus selection considerations. Currently, WHO selects vaccine strains by considering the emergence of virus with distinct genetic and antigenic characteristics, their geographical spread, and the potential loss of effective binding of antibodies from antisera of previously vaccinated subjects against the current circulating viruses[33]. These factors depict a picture of the current global virus population and their antigenic relationship to the previous representative viruses and vaccine strains, based on which recommendations of vaccine strains are made. One major consideration of vaccine strain selection is the HAI titre for antigenic characterization[34]. Nevertheless, the HAI only evaluates one specific type of reaction, that is the prevention of HA binding to sialic acid receptor on host cells inhibited by the anti-HA head antibodies[35], while other types of immune responses are undetected, such as the antibodies against HA-stem and T-cell responses that also play important roles in protection[36,37]. Developing improved assays with broader range or finer specificity for vaccine protection may facilitate the assessment and selection of vaccine strain[33]. Another major constraint of the current vaccine strain selection process is the limited availability of high-yield virus in embryonated eggs, with which more than 95% of the current

influenza vaccines are produced[38]. These candidate vaccine viruses (CVVs) are prepared from representative strains by the WHO Collaborating Centres (CC) and Essential Regulatory Labs (ERLs) before the annual WHO consultation meetings[33]. Thus, earlier prediction, even year-round projection of future representative strains may facilitate the preparation of CVVs and the subsequent vaccine strain selection, while the development of alternative vaccine technologies for influenza virus would sidestep the constraints of egg-based platform. The potential of computational optimized vaccine strains may be fully explored coupling with the availability of new production platforms, as "new wine in new wineskins".

In summary, we have introduced a new computational method, the beth-1, for predicting influenza evolution through the site-based fitness dynamic modelling and enables strain selection considering multiple proteins. The model demonstrated promising prediction performances in both retrospective and prospective real data applications. The framework has potentially wide applications by virology labs, vaccine manufacturers, health authorities and the WHO for indicating virus evolution and preparing vaccine virus, to facilitate influenza vaccine strain selection towards more effective vaccines.

References

1 Petrova, V. N. & Russell, C. A. The evolution of seasonal influenza viruses. Nat Rev Microbiol 16, 47-60, doi:10.1038/nrmicro.2017.118 (2018).

2 Ampofo, W. K. et al. Improving influenza vaccine virus selection: report of a WHO informal consultation held at WHO headquarters, Geneva, Switzerland, 14-16 June 2010. Influenza and other respiratory viruses 7 Suppl 2, 52-53, doi:10.1111/irv.12081 (2013).

3 Morris, D. H. et al. Predictive Modeling of Influenza Shows the Promise of Applied Evolutionary Biology. Trends in microbiology 26, 102-118, doi:10.1016/j.tim.2017.09.004 (2018).

4 Nelson, M. I. & Holmes, E. C. The evolution of epidemic influenza. Nat Rev Genet 8, 196-205, doi:10.1038/nrg2053 (2007).

5 Bush, R. M., Fitch, W. M., Bender, C. A. & Cox, N. J. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Molecular biology and evolution 16, 1457-1465, doi:10.1093/oxfordjournals.molbev.a026057 (1999).

6 Lässig, M., Mustonen, V. & Walczak, A. M. Predicting evolution. Nature ecology & evolution 1, 77, doi:10.1038/s41559-017-0077 (2017).

7 Koel, B. F. et al. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science 342, 976-979, doi:10.1126/science.1244730 (2013).

8 Shih, A. C., Hsiao, T. C., Ho, M. S. & Li, W. H. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. Proc Natl Acad Sci U S A 104, 6283-6288, doi:10.1073/pnas.0701396104 (2007).

9 Kryazhimskiy, S., Dushoff, J., Bazykin, G. A. & Plotkin, J. B. Prevalence of epistasis in the evolution of influenza A surface proteins. PLoS Genet 7, e1001301, doi:10.1371/journal.pgen.1001301 (2011).

10 Illingworth, C. J. & Mustonen, V. Components of selection in the evolution of the influenza virus: linkage effects beat inherent selection. PLoS Pathog 8, e1003091, doi:10.1371/journal.ppat.1003091 (2012).

11 Ampofo, W. K. et al. Improving influenza vaccine virus selection: report of a WHO informal consultation held at WHO headquarters, Geneva, Switzerland, 14-16 June 2010. Influenza and other respiratory viruses 6, 142-152, e141-145, doi:10.1111/j.1750-2659.2011.00277.x (2012).

12 Smith, D. J. et al. Mapping the antigenic and genetic evolution of influenza virus. Science 305, 371-376, doi:10.1126/science.1097211 (2004).

13 Hadfield, J. et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics (Oxford, England) 34, 4121-4123, doi:10.1093/bioinformatics/bty407 (2018).

14 Luksza, M. & Lassig, M. A predictive fitness model for influenza. Nature 507, 57-61, doi:10.1038/nature13087 (2014).

15 Huddleston, J. et al. Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. eLife 9, doi:10.7554/eLife.60067 (2020).

16 Neher, R. A., Russell, C. A. & Shraiman, B. I. Predicting evolution from the shape of genealogical trees. eLife 3, doi:10.7554/eLife.03568 (2014).

17 Steinbruk, L., Klingen, T. R. & McHardy, A. C. Computational Prediction of Vaccine Strains for Human Influenza A (H3N2) Viruses. Journal of virology 88, 12123-12132, doi:10.1128/Jvi.01861-14 (2014).

18 Steinbruck, L. & McHardy, A. C. Allele dynamics plots for the study of evolutionary dynamics in viral populations. Nucleic Acids Res 39, doi:ARTN e410.1093/nar/gkq909 (2011).

19 Kennedy, J. & Eberhart, R. in Proceedings of ICNN'95-international conference on neural networks. 1942-1948 (IEEE).

20 Wang, M. H. et al. Characterization of key amino acid substitutions and dynamics of the influenza virus H3N2 hemagglutinin. The Journal of infection 83, 671-677, doi:10.1016/j.jinf.2021.09.026 (2021).

21 Krammer, F. et al. How Can Neuraminidase-Based Immunity Contribute to Better Influenza Virus Vaccines? mBio 9, doi:10.1128/mBio.02332-17 (2018).

22 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, doi:10.1038/s41591-022-01877-1 (2022).

23 Cao, L. et al. In silico prediction of influenza vaccine effectiveness by sequence analysis. Vaccine 39, 1030-1034, doi:10.1016/j.vaccine.2021.01.006 (2021).

24 Cao, L. et al. Improving in silico prediction of influenza vaccine effectiveness by genome analysis incorporating epitope information. medRxiv, 2023.2002. 2014.23285900 (2023).

25 Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin 22, doi:10.2807/1560-7917.Es.2017.22.13.30494 (2017).

26 Price A. M., Flannery B, Talbot H.K., et al. Influenza Vaccine Effectiveness Against Influenza A(H3N2)-Related Illness in the United States During the 2021-2022 Influenza Season. Clin Infect Dis. 2023;76(8):1358-1363. doi:10.1093/cid/ciac941

27 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, doi:10.1038/ncomms8952 (2015).

28 Ma, Y. et al. The Phylodynamics of Seasonal Influenza A/H1N1pdm Virus in China Between 2009 and 2019. Front Microbiol 11, 735, doi:10.3389/fmicb.2020.00735 (2020).

29 Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. Nature 444, 929-932, doi:10.1038/nature05385 (2006).

30 Li, C., Qian, W. F., Maclean, C. J. & Zhang, J. Z. The fitness landscape of a tRNA gene. Science 352, 837-840, doi:10.1126/science.aae0568 (2016).

31 Strelkowa, N. & Lassig, M. Clonal interference in the evolution of influenza. Genetics 192, 671-682, doi:10.1534/genetics.112.143396 (2012).

32 Barrat-Charlaix, P., Huddleston, J., Bedford, T. & Neher, R. A. Limited Predictability of Amino Acid Substitutions in Seasonal Influenza Viruses. Molecular biology and evolution 38, 2767-2777, doi:10.1093/molbev/msab065 (2021).

33 Group, W. H. O. W. et al. Improving influenza vaccine virus selection: report of a WHO informal consultation held at WHO headquarters, Geneva, Switzerland, 14-16 June 2010. Influenza Other Respir Viruses 6, 142-152, e141-145, doi:10.1111/j.1750-2659.2011.00277.x (2012).

34 Houser, K. & Subbarao, K. Influenza vaccines: challenges and solutions. Cell host & microbe 17, 295-300, doi:10.1016/j.chom.2015.02.012 (2015).

35 Stephenson, I., Wood, J. M., Nicholson, K. G. & Zambon, M. C. Sialic acid receptor specificity on erythrocytes affects detection of antibody to avian influenza haemagglutinin. J Med Virol 70, 391-398, doi:10.1002/jmv.10408 (2003).

36 Reber, A. & Katz, J. Immunological assessment of influenza vaccines and immune correlates of protection. Expert Rev Vaccines 12, 519-536, doi:10.1586/erv.13.35 (2013).

37 Doherty, P. C. et al. Effector CD4+ and CD8+ T-cell mechanisms in the control of respiratory virus infections. Immunol Rev 159, 105-117, doi:10.1111/j.1600-065x.1997.tb01010.x (1997).

38 Perdue, M. L. et al. The future of cell culture-based influenza vaccine production. Expert Rev Vaccines 10, 1183-1194, doi:10.1586/erv.11.82 (2011).

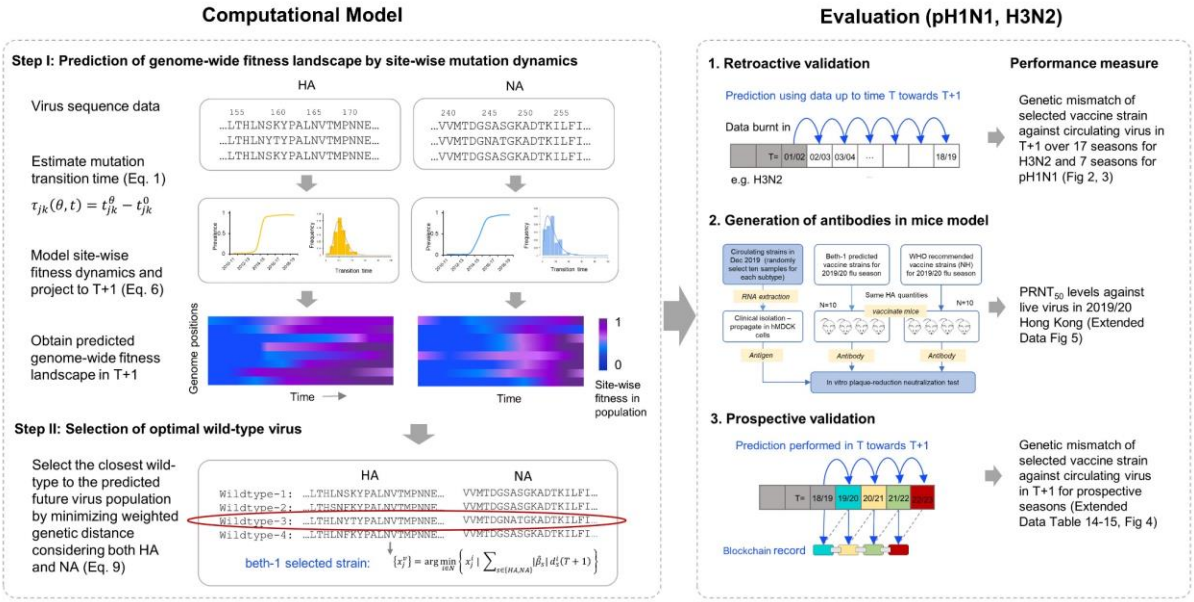

Fig. 1 | Overview of the computational model and study design.

Site-wise fitness dynamics are modelled and projected to the next season (*T*+1). Based on the predicted genome-wide fitness landscape of future virus population, an optimal wild-type virus can be selected integrating evolutionary information of both hemagglutinin and neuraminidase genes. The prediction is validated by genetic mismatch against observed circulating viruses in the Northern Hemisphere retroactively and prospectively for pH1N1 and H3N2. In addition, mouse model is used to evaluate the antibodies elicited by predicted vaccine strain in neutralizing the clinical isolates in 2019/20 season.

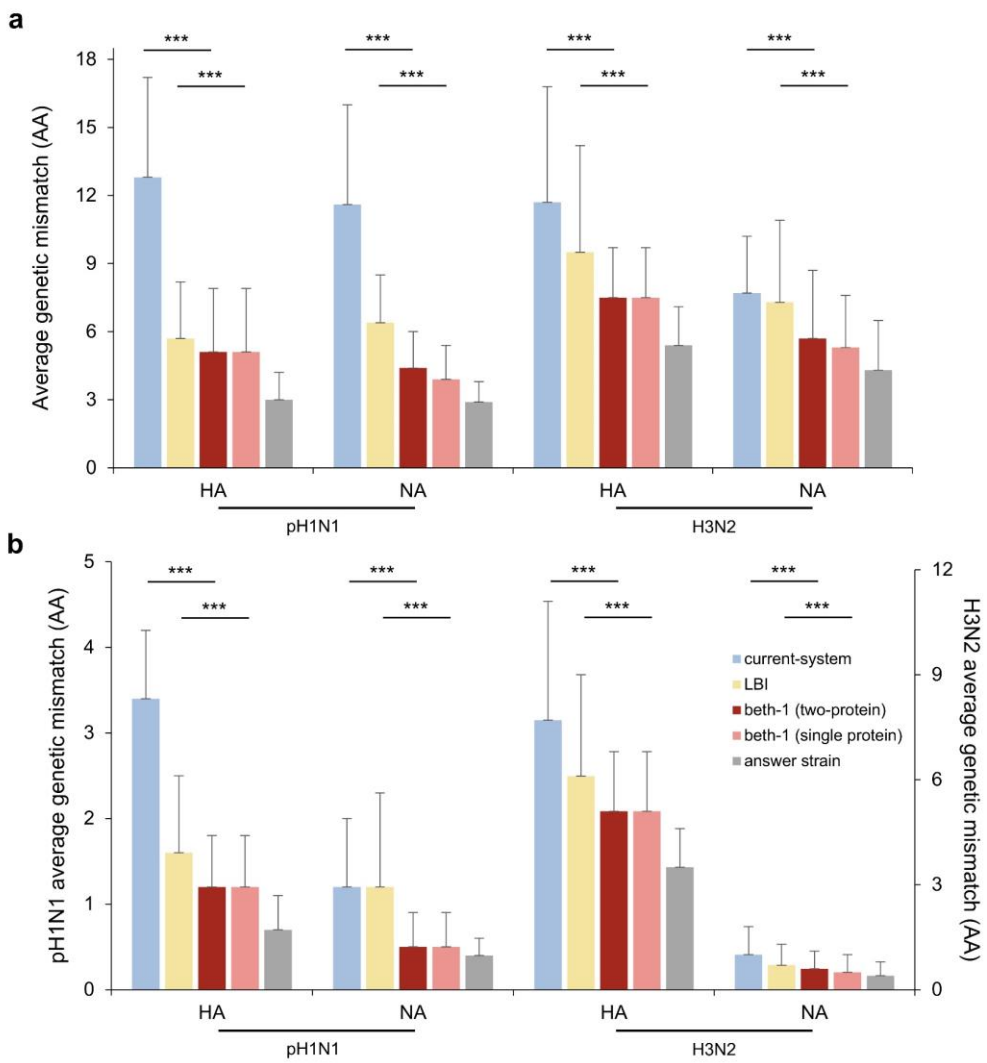

Fig. 2 | Prediction performance in the retrospective data for influenza pH1N1 and H3N2. **a**, Full-length protein. **b**, Epitope. The bar-plots display average genetic mismatch (amino acids, AA) over seven seasons in 2012/13-2018/19 for pH1N1 and 17 seasons in 2002/03-2018/19 for H3N2. Panel **b** left Y-axis: average genetic mismatch of pH1N1; right Y-axis: average genetic mismatch of H3N2. Prediction methods being compared include: the current-system, LBI, beth-1 (two-protein), beth-1 (single protein), and the actual representative strain. In the retrospective validations, beth-1 shows significantly lower genetic mismatch on all the protein segments evaluated for the two influenza subtypes, compared to the LBI and the current-system. Error bar: standard deviation of the average genetic mismatch by region and season. *p*-value is calculated using paired *t*-test between two methods matched by region and season. ***: *p*-value < 0.001.

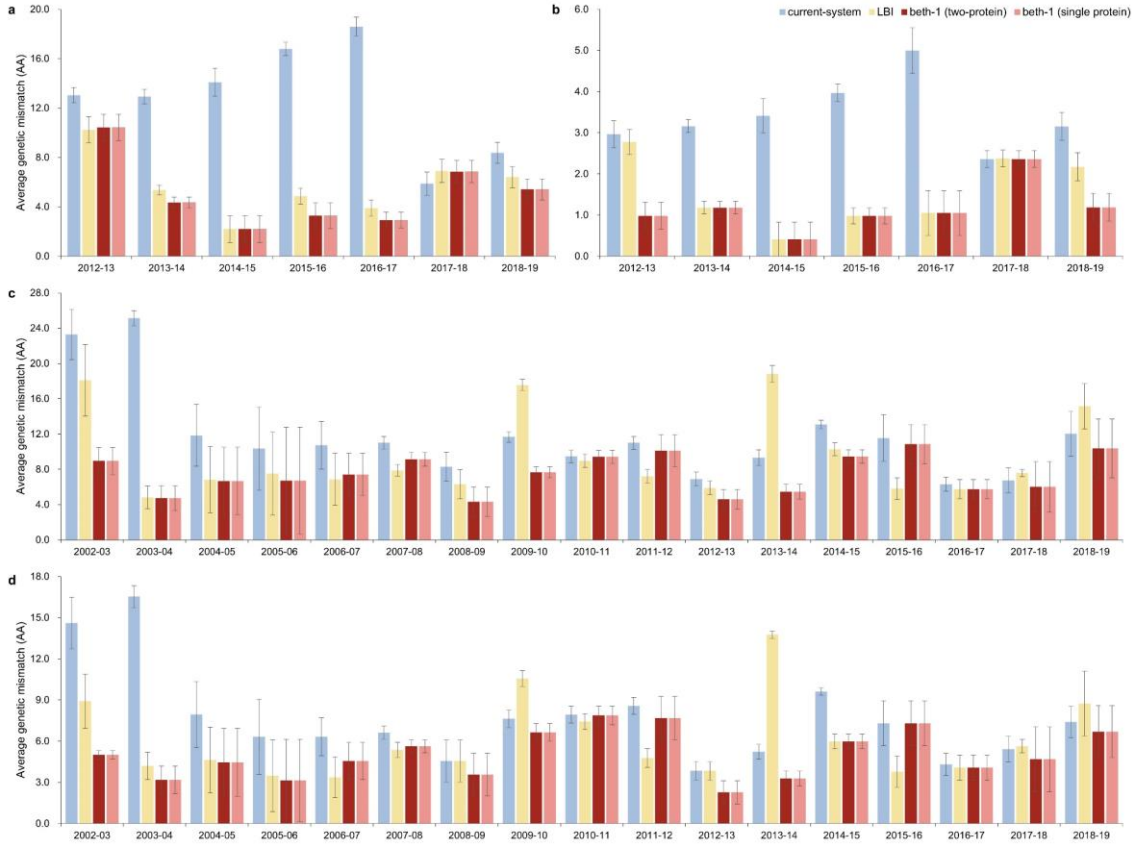

Fig. 3 | Prediction accuracy of alternative methods for HA evolution in retrospective data by season.

a, pH1N1, HA full sequence (566 codons). **b**, pH1N1, HA epitopes (50 sites). **c**, H3N2, HA full sequence (566 codons). **d**, H3N2, HA epitopes (131 sites). Prediction accuracy is assessed by the average amino acids mismatch (Y-axis) of the predicted strains against circulating viruses in ten geographical regions in the respective epidemic seasons (X-axis). Error bar: standard deviation of the average genetic mismatch by region in a given season. The average genetic mismatch of beth-1 is prevailingly lower compared to the LBI and the current-system.

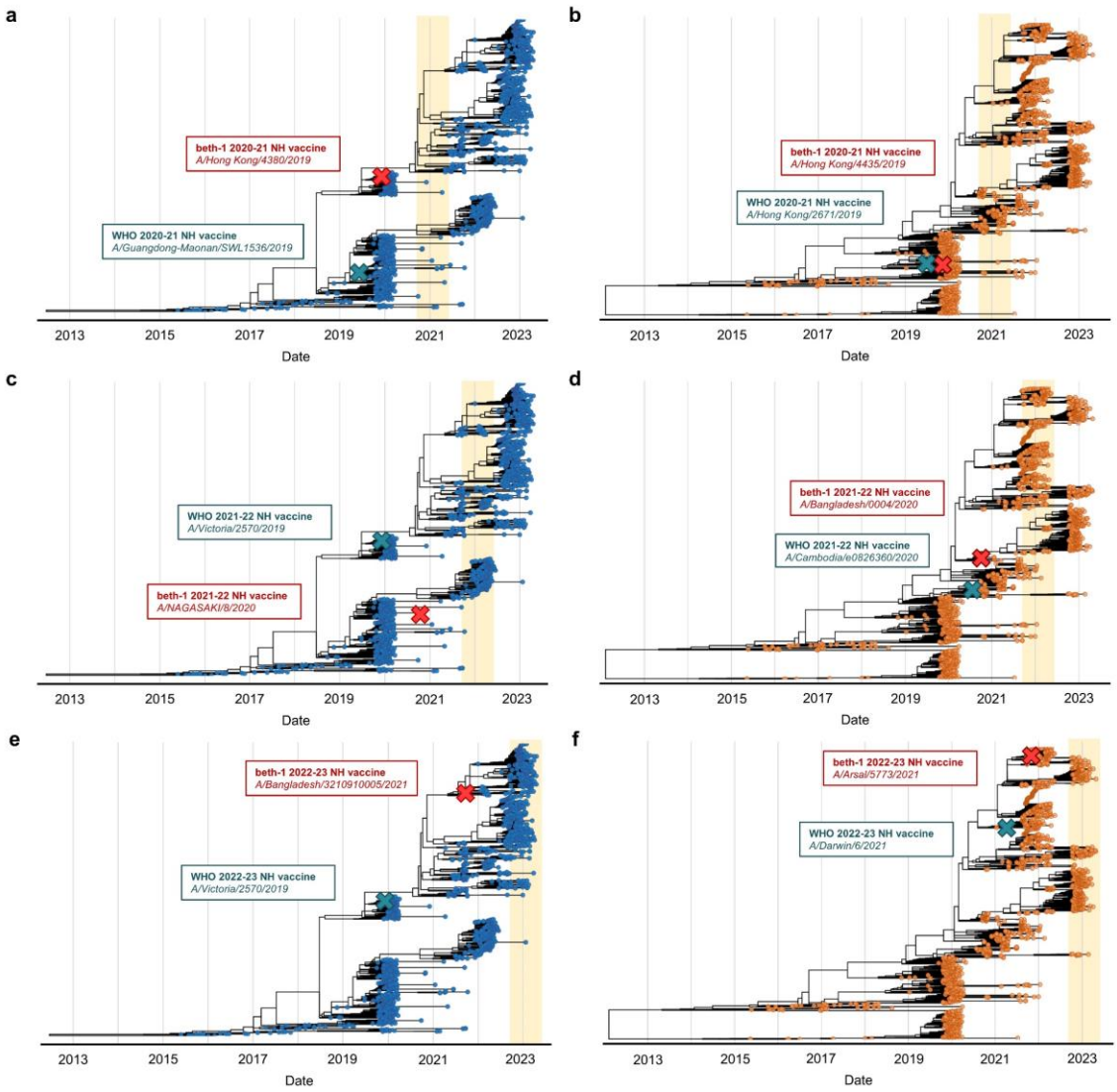

Fig. 4 | Prospectively predicted strains on phylogenetic tree from 2020/21 to 2022/23. **a**, 2020/21, pH1N1. **b**, 2020/21, H3N2. **c**, 2021/22, pH1N1. **d**, 2021/22, H3N2. **e**, 2022/23, pH1N1, **f**, 2023/23, H3N2. Predicted strains by beth-1 (two-protein) and the current-system are marked on phylogenetic trees. Predictions were made prospectively for the next epidemic season in the Northern Hemisphere (highlighted by yellow band in the background). The phylogeny relationship shows that beth-1's predictions are generally more advanced into the future.

Methods

Dataset We downloaded genetic sequences of the influenza virus from the Global Initiative on Sharing All Influenza Data (GISAID)[25]. All samples were retrieved and analyzed if the strain has complete sequence and were isolated from the target epidemic seasons and geographical regions (**Extended Data Table 2**). The data of genetic sequences for pH1N1 spanned from 2009/10 to 2022/23, and for the H3N2 from 1999/2000 to 2022/23. Ten geographical regions in the Northern Hemisphere (NH) were considered, including North America (New York State, California State, Canada), Europe (United Kingdom, Germany, France), and Asia (Hong Kong SAR, South China provinces, Japan, Singapore). In 2020/21 and 2021/22, due to sharp decline of available samples in these regions, all sequences in NH were considered. Overall, the total number of genetic sequences used in analysis was 50,285 for the HA and 45,297 for the NA. Sequence alignment was performed by MEGA (7.0.26)[39]. All statistical analysis were conducted in R version 4.1.3[40]. All *p*-values reported are two-sided.

The vaccine strain prediction and selection model: beth-1 In our prior work[20], we used the *threshold* θ to detect the change-point of mutation prevalence in relation to a significant rise in epidemic level, and the *extended effective mutation period* (h) to measure the duration of a mutation to remain in selective advantage in the host population after reaching θ. Let $x_{jk}(t)$ denote amino acid residue or nucleotide type at site j in time t. $j \in J$, where J is the sequence length. $k \in K_j$, and K_j is the total number of unique substitutions observed at site j. $K_j \leq 4$ for nucleotide sequence, and $K_j \leq 20$ for amino acid sequence. Amino acid sequences are used in the prediction modelling of this study. Denote prevalence of $x_{jk}(t)$ in host population as $p_{jk}(t)$. For a mutation $x_{jk}(t)$, let t_{jk}^0 denote the time point when $p_{jk}(t^0)$=0 and $p_{jk}(t^0+1) > 0$; and t_{jk}^θ denote the time when $p_{jk}(t^\theta) = \theta$. The *transition time* (τ) for a mutation $x_{jk}(t)$ is defined as,

$$\tau_{jk}(\theta, t) = t_{jk}^\theta - t_{jk}^0. \qquad \textbf{Eq. 1}$$

τ describes the duration that a mutation takes from its first emergence to reaching an influential prevalence in population; $\tau > 0$. The parameters (θ, h) are estimated from fitting the epidemic trend and a summary statistic of genetic activities, the g-measure[20], towards optimizing a goodness-of-fit statistic $S(\cdot)$. We have,

$$(\hat{\theta}, \hat{h}) = \arg \max_{\theta \in \Theta, h \in \mathrm{H}} S[y(t), g(\theta, h, t)], \qquad \textbf{Eq. 2}$$

where $y(t)$ is a measure of epidemic level through time. $\Theta \in (0,1)$, and $\mathrm{H} = \{0,1,2,\ldots\}$. The *g-measure* is written as,

$$g(\theta, h, t) = \mathbf{m}(\theta, h, \mathrm{t}) \cdot \mathbf{p}(\mathrm{t}) = \sum_{j,k} m_{jk}(\theta, h, t) p_{jk}(t), \qquad \textbf{Eq. 3}$$

in which $m_{jk}(\theta, h, t)$ is a time-dependent indicator function of influential mutations (or *effective mutations*, EM)[20], such that $m_{jk}(\theta, h, t) \triangleq I\{t_{jk}^{0} \leq t \leq t_{jk}^{\theta} + h, j \in J, k \in K_j\}$. Thus, the subset of influential mutations varies across epidemic seasons (**Extended Data Table 1**). We showed in other works that the *g*-measure is a good predictor for epidemic cycles of the A(H3N2)[20], A(H1N1)pdm09[41] and COVID-19[42,43]. In this study, $S(\cdot)$ is the R-square of the following linear model,

$$y(t) \sim g(\theta, h, t) + Z(t), \qquad \textbf{Eq. 4}$$

in which $y(t)$ is the annual sero-positivity rate, $g(\theta, h, t)$ is the g-measure, and covariates $Z(t)$ include mean temperature, absolute humidity and season, fitted separately by geographical regions in the training set (**Extended Data Table 16**). When sufficient sequences and epidemic data are available, the time-varying reproduction number R_t is a preferred quantity for $y(t)$[42,43].

Prediction of future mutation prevalence can be made by solving a classical initial value problem of the differential equation,

$$\begin{cases} f'(t) = F\left[t, p_{jk}(t), \tau_{jk}(\theta, t)\right], & t > T \\ f(T) = \eta \end{cases}, \qquad \textbf{Eq. 5}$$

where $f(t)$ is a function describing mutation prevalence through time, $F\left[t, p_{jk}(t), \tau_{jk}(\theta, t)\right]$ is the velocity of the prevalence change, η is the initial value, and T is the latest observation time. Using Euler's method of approximation, the projected mutation prevalence at time T+1 for substitution k at position j is,

$$\hat{p}_{jk}(T+1) = \{p_{jk}(T) + F\left[T, p_{jk}(T), \hat{\tau}_{jk}(\theta, T)\right]\} \cdot C, \qquad \textbf{Eq. 6}$$

in which $k \in K_j$, and $j \in J$. $C = 1/\sum_{k \in K_j}\{p_{jk}(T) + F\left[T, p_{jk}(T), \hat{\tau}_{jk}(\theta, T)\right]\}$ is a normalization constant. $\hat{\tau}_{jk}(\theta, T)$ is estimated with history transition event at site j by the time of making prediction. When no history of transition event is available for the site, its transition time is estimated by the mean transition time of mutations from the same protein in the training data (**Extended Data Table 16**).

Eq. 5 and **6** describe the prediction step. In practice, the model form of $F(t)$ shall be chosen such that the training data can support a robust estimation. In the earlier seasons, only a dozen influenza sequence samples were available each year in many regions. Thus, yearly interval is chosen and the simplest form of $F(t)$ as a rate is adopted. Specifically,

$$F\left[T, p_{jk}(T), \hat{\tau}_{jk}(\theta, T)\right] = \frac{\Delta p_{jk}(T)}{\Delta t} = \frac{p_{jk}(T) - p_{jk}[T - \hat{\tau}_{jk}(\theta, T)]}{\hat{\tau}_{jk}(\theta, T)} \quad . \qquad \textbf{Eq. 7}$$

For viruses with more abundant samples at denser time interval, other model form of $F(t)$ can be incorporated.

Based on the projected genome-wide fitness landscape given by **Eq. 6**, we can thus select the most advantageous mutations per site in T+1 using the mode estimator,

$$k_j^* = \arg\max_{k \in K_j} \hat{p}_{jk}(T+1).$$

This gives us a consensus strain to represent the projected future virus population,

$$\left\{x_j^c(T+1) \mid c = k_j^*\right\}. \qquad \textbf{Eq. 8}$$

Since a high mutation prevalence implies selective advantage of the mutation in population, the mode sequence carries all the advantageous mutations at time T+1. The prediction model is trained using sample sequences from the South-east Asia region that is the source of influenza A epidemics[44,45]; for the 2020/21 and 2021/22 seasons during COVID-19 pandemic, all Northern Hemisphere samples were used.

Next, we locate the optimal wild-type virus $\left\{x_j^v\right\}$ that is closest to the predicted HA or/and NA sequences by minimizing the weighted distance,

$$\left\{x_j^v\right\} = \arg\min_{i \in N}\left\{ x_j^i \mid \sum_{s \in S} |\hat{\beta}_s|\, d_s^i(T+1) \right\}. \qquad \textbf{Eq. 9}$$

in which,

$$d_s^i(T+1) = \sum_{j \in W(t), t \in [1, T+1]} I\{x_j^i \neq x_j^c(T+1)\}. \qquad \textbf{Eq. 10}$$

$d_s^i(T+1)$ is the hamming distance between a candidate wild-type virus $\{x_j^i\}$ and $\{x_j^c(T+1)\}$, measured over the predictor codon sites $W(t)$, which is a subset of EMs residing on antigenic sites (**Extended Data Table 1**). EMs in T+1 can be obtained using the projected $\hat{p}_j^k(T+1)$ in **Eq. 6** and fitted (θ, h) (**Extended Data Table 16**). The weights $|\hat{\beta}_s|$ are the relative importance of mutations of gene s, obtained by the VE-GD model described below. $s \in S = \{HA, NA\}$ for

beth-1(two-protein) model, and $S = \{HA\ or\ NA\}$ for beth-1 (single protein) model. N is the total number of available wild-type viruses on and before time T. **Eq. 9** gives the optimal wild-type virus identified by beth-1 for considerations of vaccine strains.

A brief account of the VE-GD model for influenza[23,24] is described here. The model captures the statistical relationship between VE and genetic mismatch of vaccine strains against circulating viruses. VE data was collected from published observational studies over multiple seasons. Let $\{x_j^l\}$ denote a vaccine sequence in study l. d_s^l measures genetic mismatch, the average hamming distance on the predictor codon set of gene s between a vaccine strain $\{x_j^l\}$ and the circulating viruses in the same period and geographical region of study l. $s \in S = \{HA, NA\}$. The effect of genetic mismatch was first estimated for individual genes by $VE \sim d_s$ in a linear model, resulting the estimated effect size $\hat{\beta}_s$. Since systematic surveillance of influenza VE is available for the United States, VE data and the corresponding viral sequences from this region are used for model fitting. The fitted coefficients of HA and NA are in ratio of 1: 1.3 for the pH1N1 and 1: 1.8 for the H3N2 (p-value < 0.001)[24]. Next, normalized effect of multiple genes is combined through the polygenic risk score (PRS)[46,47] to form an integrated measure of genetic mismatch, $GD = \sum_{s \in S} |\hat{\beta}_s|\, d_s$, which serves as the predictor for VE in a linear regression model.

Vaccine strains predicted by alternative models. The list of predicted strains by alternative methods can be found in **Extended Data Table 3-4**. The answer strain is set as the closest wildtype to the consensus sequence in season T+1 obtained using the observed data. The predicted strains of LBI[16] are obtained by running the model's source code in python (https://github.com/rneher/FitnessInference) on the same training data as the beth-1 by gene. In each season, a new tree is built for the LBI using default parameters (eps_branch =10^{-5}, tau = 0.0625, outgroup= "A/Puerto Rico/8/1934" for pH1N1, "A/Hong Kong/1/1968" for H3N2). When the highest LBI score corresponds to more than one strain, one of the strains is randomly picked. Two-sample *t*-test is used to compare log mismatch against circulating viruses between alternative prediction models in a given season. When comparing model performances over multiple seasons, paired *t*-test of two methods by season is used to control seasonal variation.

Phylogenetic tree To better visualize the predicted strains, we present them in the time-scaled maximum-likelihood phylogenies generated from TreeTime[48]. The input tree was built in MEGA (7.0.26) using the neighbor-joining (NJ) approach with Jukes-Cantor gamma distance

of shape parameter 4[39]. Genetic sequences for building the tree were sampled in proportion to continent distributions in the corresponding epidemic season, targeting a total number of 400 sequences per season for each of the pH1N1 and H3N2 subtype per season. In 2020/21, all available sequences of pH1N1 in the NH were used.

Clinical samples for animal experiment Between October 2019 to February 2020, a total number of 117 positive H1 and 43 H3 nasopharyngeal swab samples were collected in the Prince of Wales Hospital, Hong Kong SAR. Ten samples from each of the H1 and H3 subtypes were randomly selected for virus isolation in hMDCK cells. The sequences of the HA gene from each virus was then determined by sanger sequencing. Phylogenetic analysis showed that the viruses belonged to the same clusters of their respective subtype in these isolates, of which HA ectodomain shared >98% amino acids similarity for the pH1N1 and >96% for H3N2. Thus, one pH1N1 and one H3N2 virus from these isolates were randomly picked for animal experiments. Human study ethics approval has been obtained from the Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee.

Cell culture Humanized Madin-Darby canine kidney (hMDCK) cells[49] and human embryonic kidney (HEK) 293T cells were used in this study. Both cells were maintained in minimum essential medium (MEM) supplemented with 10% fetal bovine serum (FBS; Gibco), 25 mM HEPES, and 100 U/mL penicillin-streptomycin (PS; Gibco).

Virus rescuing by reverse genetics The H1N1 A/Puerto Rico/8/1934 (PR8) eight-plasmid reverse genetic system was used to produce the recombinant influenza virus with the predicted HA and NA genes[50]. Chimeric 6:2 recombinant viruses with six internal genes from PR8 (PB2, PB1, PA, NP, M, and NS) and 2 genes (HA and NA) from strain of interest were synthesized by Sangon Technology (Guangzhou, China) and were cloned into the pHW2000 vector[51]. The HEK 293T cells and hMDCK cells were then mixed at a ratio of 2:1 one-day before transfection and co-cultured in a 6-well plate until they reached 70% confluence. For each recombinant virus, 16 μL TransIT®-LT1 (Mirus) and 1 μg for each of the 8 plasmids encoding the corresponding virus fragments were mixed thoroughly for transfection. The medium was replaced with 1mL of MEM supplemented with 25 mM HEPES and 100 U/mL PS at 6-hour post transfection. At 24-hour post-transfection, another 1mL of MEM was added, supplemented 25 mM HEPES, 100 U/mL PS, and 1 μg/mL tosylphenylalanyl chloride methyl ketone (TPCK) trypsin (Sigma). At 72-hour post-transfection, cell supernatants were inoculated into hMDCK cells maintained in MEM containing 25mM HEPES, 100 U/mL PS,

and 1 μg/mL TPCK trypsin. The viruses were harvested at 72 hours after infection and titrated by plaque assay.

Immunization of the influenza vaccine virus in a mouse model The research protocol of animal experiment was carried out in strict accordance with the recommendations and was approved by the Teaching and Research Committee on the Use of Live Animals (CULATR 5598-20) of the University of Hong Kong. Batches of ten 8-10 weeks old female BALB/c mice were inoculated intraperitoneally (i.p) with 10^5 PFU of Addavax-adjuvant recombinant virus in 200 μL of phosphate-buffered saline (PBS; Gibco). At twenty-one days post immunization, peripheral blood was drawn from the immunized mice by cardiac puncture, and the serum was collected by centrifugation. All sera were treated with receptor destroying enzyme (RDE; Sigma) for 16-18 hours to reduce non-specific antibodies, then heated at 56 degrees for 30 minutes to inactivate the RDE residue. The serum was stored at -80 degrees for further experiments.

Immunogenicity assessment based on plaque reduction neutralization test hMDCK cells were used for virus titration by plaque assay and plaque reduction neutralization test (PRNT). hMDCK cells were prepared in a 12-well plate at least 24 hours before PRNT and cultured until obtaining a 100% confluent monolayer in each well. Two-fold serial dilutions of antiserum from 1:10 to 1:1280 were prepared, where each 100 μL was mixed with 100 PFU of viruses and incubated at 37°C for 1 hour. Cells were washed once with PBS and inoculated with corresponding virus-antiserum mixture for further incubation at 37°C for one hour. After that, the supernatant in each well was replaced with 3 mL of MEM supplied with a final concentration of 25 mM HEPES, 100 U/mL PS, 1% agarose gel, and 1 μg/mL TPCK trypsin. Against the live viruses, the highest dilution of the antiserum that inhibited at least 50% of the viral plaques ($PRNT_{50}$ titer) was recorded. The plate was inverted and incubated at 37°C for 48 hours. Finally, the cells were fixed with 4% formaldehyde and stained with 0.5% crystal violet for virus plaque counting. We recorded the highest dilution of the antiserum that inhibited at least 50% of the viral plaques as the $PRNT_{50}$ titer of the antiserum. Detailed information of the reagents and cell lines used in this study were listed in **Extended Data Table 17**.

38 Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular biology and evolution* **33**, 1870-1874, doi:10.1093/molbev/msw054 (2016).
39 Team, R. C. R: A language and environment for statistical computing. (2013).

40 Wang, M. H. *et al.* Characterization of key amino acid substitutions and dynamics of the influenza virus H3N2 hemagglutinin. *The Journal of infection* **83**, 671-677, doi:10.1016/j.jinf.2021.09.026 (2021).
41 Lou, J. *et al.* Predicting the dominant influenza A serotype by quantifying mutation activities. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases* **100**, 255-257, doi:10.1016/j.ijid.2020.08.053 (2020).
42 Lou, J. *et al.* Quantifying the effect of government interventions and virus mutations on transmission advantage during COVID-19 pandemic. *Journal of infection and public health* **15**, 338-342, doi:10.1016/j.jiph.2022.01.020 (2022).
43 Zhao, S. *et al.* Modelling the association between COVID-19 transmissibility and D614G substitution in SARS-CoV-2 spike protein: using the surveillance data in California as an example. *Theoretical biology & medical modelling* **18**, 10, doi:10.1186/s12976-021-00140-3 (2021).
44 Rambaut, A. *et al.* The genomic and epidemiological dynamics of human influenza A virus. *Nature* **453**, 615-619, doi:10.1038/nature06945 (2008).
45 Bedford, T. *et al.* Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* **523**, 217-220, doi:10.1038/nature14460 (2015).
46 International Schizophrenia, C. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752, doi:10.1038/nature08185 (2009).
47 Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* **19**, 581-590, doi:10.1038/s41576-018-0018-x (2018).
48 Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus evolution* **4**, vex042, doi:10.1093/ve/vex042 (2018).
49 Takada, K. *et al.* A humanized MDCK cell line for the efficient isolation and propagation of human influenza viruses. *Nature microbiology* **4**, 1268-1273, doi:10.1038/s41564-019-0433-6 (2019).
50 Hoffmann, E., Neumann, G., Kawaoka, Y., Hobom, G. & Webster, R. G. A DNA transfection system for generation of influenza A virus from eight plasmids. *Proc Natl Acad Sci U S A* **97**, 6108-6113, doi:10.1073/pnas.100133697 (2000).
51 Liang, W. *et al.* Egg-adaptive mutations of human influenza H3N2 virus are contingent on natural evolution. *PLoS Pathog* **18**, e1010875, doi:10.1371/journal.ppat.1010875 (2022).

Data availability

All data used in this study is publicly available. Viral sequence data were downloaded from the global initiative on sharing all influenza data (GISAID) at *http://platform.gisaid.org/* and the accession numbers were provided in the online supplementary acknowledgment table.

Code availability

The code is available at *https://github.com/mwanglab/beth-1*. Access and use of the code is subject to a revocable, non-transferable, and limited right for the exclusive purpose of undertaking academic or not-for-profit research. Use of the Code or any part thereof for commercial purposes requires a Commercial License Agreement from Beth Bioinformatics (info@bethbio.com).

Acknowledgments

We thank the GISAID Initiative for critical surveillance efforts and open data sharing, and we greatly acknowledge the authors' contributions, submitting and originating laboratories.

Funding

Health and Medical Research Fund, the Food and Health Bureau, the Government of the Hong Kong Special Administrative Region: INF-CUHK-1 (E.K.Y.), COVID190103 (M.H.W.) and 19180932 (C.K.P.M.)

National Natural Science Foundation of China: 32322088 (M.H.W.) and 71974165 (M.K.C.C.) the Chinese University of Hong Kong Direct Grant: PIEF/Ph2/COVID/06 and 2022.02 (M.H.W.)

Guangdong-Hong Kong-Macau Joint Laboratory of Respiratory Infectious Disease: 20191205 (C.K.P.M.)

Visiting scientist scheme from Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore (C.K.P.M.)

Research Grants Committee, General Research Fund: 16301319 and C6036-21G (P.P.H.C.)

Author contributions

M.H.W. conceived the study. M.H.W. and J.L. developed the prediction method. J.L. analyzed the data and prepared the results. C.K.P.M. designed the animal experiment. P.K.C. and Z.C. provided clinical samples. W.L. and C.K.P.M. conducted the immunology experiments. L.C., S. Z., H.Z., Y.Z., C.L. and Q.L. contributed to data collection and interpretation. E.Y., M.H.W. and C.K.P.M. acquired funding. M.H.W. and J.L. wrote the original manuscript. L.C., C.K.P.M.

and W.L. revised the manuscript. Z.C., R.W.Y.C., P.P.H.C., and M.K.C.C. reviewed and edited the manuscript. B.C.Y.Z. and E.K.Y. supervised the project.

Competing interests

M.H.W., J.L., M.K.C.C. and B.C.Y.Z. filed a related patent PCT/CN2019/091652. 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. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper.

Correspondence and requests for materials should be addressed to M.H.W. and C.K.P.M.