# *In Silico* Vaccine Strain Prediction for Human Influenza Viruses

Thorsten R. Klingen[1*], Susanne Reimering[1*], Carlos A. Guzmán[2,3] and Alice C. McHardy[1,3,†]

[1]Department for Computational Biology of Infection Research, Helmholtz Centre for Infection Research, Braunschweig, Germany
[2]Department of Vaccinology and Applied Microbiology, Helmholtz Centre for Infection Research, Braunschweig, Germany
[3]German Centre for Infection Research (DZIF)

[†]Correspondence to: Alice.McHardy@helmholtz-hzi.de
*Co-first authors.

Keywords: influenza viruses, vaccine, computational predictions, GISRS, viral evolution

## Abstract

Vaccines preventing seasonal influenza infections save many lives every year; however, due to rapid viral evolution, they have to be updated frequently to remain effective. To identify appropriate vaccine strains, the World Health Organization (WHO) operates a global program that continuously generates and interprets surveillance data. Over the past decade, sophisticated computational techniques drawing from multiple theoretical disciplines have been developed that predict viral lineages rising to predominance, assess their suitability as vaccine strains, link genetic to antigenic alterations, as well as integrate and visualize genetic, epidemiological, structural and antigenic data. These could form the basis of an objective and reproducible vaccine strain selection procedure utilizing the complex, large scale data types from surveillance. Towards this end, computational techniques should already be incorporated into the vaccine selection process in an independent, parallel track and their performance continuously evaluated.

## Epidemiology and evolution of influenza viruses

Based on their internal proteins, influenza viruses circulating among humans are categorized into three types: A, B, and C. Of these, A and B are primarily responsible for the yearly epidemics, whereas type C is less prevalent and causes only mild infections. The influenza A viruses, which are historically also responsible for pandemics, are further divided into subtypes, based on the combination of their hemagglutinin (H or HA) and neuraminidase (N or NA) surface glycoproteins. There are 18 different hemagglutinin (H1-H18) and 11 neuraminidase (N1-N11) subtypes [1, 2]. The predominant subtypes circulating in humans are influenza A/H1N1 (pdm09), descending from the 2009 pandemic strain, and influenza A/H3N2, circulating since 1968. For H3N2, China, Southeast Asia and possibly India serve as a reservoir, from which viral lineages spread around the globe to cause seasonal epidemics, and air travel plays an important role in its spread [3-6]. Typically one antigenically distinct lineage (or variant) of influenza A subtypes predominates in seasonal epidemics, but currently the B/Yamagata and B/Victoria lineages are co-circulating and vary in local predominance.

The influenza A viral genome consists of 8 negative-sense, single-stranded RNA segments encoding more than 12 proteins, depending on the strain (Fig. 1, [1]). HA enables the virus to bind to its "receptor" (specific sugar structures) on the host cell surface, and initiates release of the virion content, once the particle has been endocytosed. NA facilitates the release of viral particles. The ribonucleoprotein (RNP) complex replicates and transcribes viral RNA segments. It includes the segments bound to nucleoprotein (NP), which is involved in nuclear import of the RNA, and a heterotrimeric polymerase complex of polymerase acidic protein (PA), the RNA-directed RNA polymerase catalytic subunit (PB1), and polymerase basic protein 2 (PB2). M2 is a transmembrane ion channel within the lipid envelope involved in virus uncoating during cell entry and virion assembly before release. The multifunctional M1 matrix protein, which surrounds the virion core, interacts with RNP, regulating RNA nuclear export and viral envelope formation (i.e. budding) [7]. Other proteins are the non-structural protein NS1, which acts as interferon antagonist regulating host gene expression and the nuclear protein (NEP or NS2), which mediates contact between the M1/RNP complex and the cellular exportin. The splice variant PB1-F2 exhibits pro-apoptotic activity, and PA-X, an RNA-endonuclease, also modulates host response [8, 9]. However, the proteome of the influenza A virus is far more complex than initially envisioned. Recently, several accessory proteins were identified (*e.g.* PB1-N40, PA-N155, PA-N182, M42, NS3) that are translated from alternative open reading frames and play a role in the viral life cycle [10].

One of the primary defense mechanisms in combating influenza infections are antibodies targeting HA and NA, which are produced by B cells. While antibodies against HA prevent infection of host cells, those binding NA interfere with viral replication and spread [11]. Antibodies can also contribute to the lysis of infected cells via complement activation, antibody dependent cell cytotoxicity or activation of natural killer cells [12-17]. The activation of these specific effector mechanisms can lead to improved viral clearance but also to immune pathology, such as more severe forms of disease as a result of immune complexes [18]. The internal proteins are the main target for cellular immune responses such as cytotoxic T cells, which lyse infected host cells.

RNA viruses have a high mutation rates [19]. The resulting amino acid changes in HA or NA can weaken viral particle binding by antibodies from prior infections or vaccination, a phenomenon called antigenic drift (Fig. 2) [1]. Since influenza viruses possess a segmented genome, reassortment can also occur, where viruses co-infecting the same cell inherit their segments from different parental strains. Reassortment of strains from the same or different human subtypes [20] can increase antigenic drift or alter overall viral fitness. Through antigenic variation created by antigenic drift [1], the viruses stay ahead of the host's production of specific antibodies. This leads to a succession of alternating states, in which the host's immune system recognizes the new viral particles and produces antibodies that are effective until a virus with altered surface proteins appears (Fig. 2). For H3N2 and H1N1 viruses, antigenically similar strains, a so called antigenic variant, predominate in seasonal epidemics, and circulate for several years, before being replaced by a novel one [21]. If reassortment generates human transmissible viruses with an HA segment of non-human lineages, such as from birds or pigs, this results in antigenic shift; a large change of antigenicity, correlating with low immune defenses in the population and initiating a pandemic [1].

Due to antigenic drift, vaccines against human influenza viruses have to be frequently updated. In the following we outline the vaccine selection process of the WHO and the associated challenges. We then describe computational techniques for predicting influenza evolution that might allow to further improve the process and the timely identification of even better matching strains to circulating viruses.

## Vaccines for seasonal influenza viruses

Influenza vaccines prevent infection by eliciting an antibody response against circulating viruses. Trivalent vaccines include a strain of influenza A/H1N1 and A/H3N2, as well as one strain of the B lineage that is currently predominant. A second B lineage is additionally represented in quadrivalent vaccines [22]. Two types of vaccines are available; one with inactivated viruses delivered via injection and a nasal spray of live attenuated viruses that has been less effective [23, 24]. Antigenic drift severely impacts vaccine efficacy: if circulating viruses are antigenically similar to the vaccine strains, the vaccine efficacy, measured by reduction of the infection risk, is 50-60% for the general population [25]. If the circulating viruses do not match the vaccine strains well, however, the efficacy is drastically reduced - such as to 19% in the 2014/2015 winter season [24]. Among the elderly, the efficacy in the best case only reaches around 19% [25]. Specific vaccines improve the efficacy for this high risk group, using adjuvants [26] or a higher antigen dosage [27].

To monitor the circulating viral populations for the emergence of new antigenic variants, the WHO has created the Global Influenza Surveillance and Response System (GISRS). This includes currently 143 National Influenza Centers (NICs) in 113 WHO member states and 6 WHO Collaborating Centers (CCs). The NICs collect and assess antigenic, genetic and epidemiological data of circulating viruses around the year: samples are obtained from patients with influenza symptoms and sent to the CCs, where viruses are isolated and analyzed to determine their type and subtype (Fig. 3). The CCs select representative viruses for a circulating clade and antigenically different ones for detailed genetic and antigenic analysis. Their antigenic relationship to other viruses is characterized with hemagglutination inhibition (HI) and virus microneutralisation (MN) assays using panels of post-infection ferret antisera and post vaccination human sera. For almost all viruses, genome sequencing is also performed.

A GISRS committee selects the vaccine strains for the next years' influenza season biannually. In February, vaccine strains are recommended for the next season in the northern hemisphere starting in October, whereas in September, vaccine strains are recommended for the next southern hemisphere season starting around April. If a new antigenic variant (defined by a more than fourfold reduced titre in HI or MN assays) is identified that is considered likely to become predominant, a new vaccine strain recommended. Subsequently, the CCs produce viruses of selected strains in hens' eggs and 'reassorting laboratories' create reassortants of their HA and NA segments with other segments from the A/Puerto Rico/8/34 strain. This improves growth in eggs, which are primarily used for vaccine production. The

reassortants are examined for antigenic and genetic changes and vaccine manufactures evaluate their growth properties. Based on these results, GISRS recommends high-growth vaccine strains that are antigenically similar to the selected viruses.

Creating an efficient influenza vaccine is challenging for several reasons: change of the receptor-binding properties of circulating H3N2 viruses reduced their binding avidity to avian red blood cells [28-30]. These were originally used to assess antigenicity in HI assays and were replaced with also problematic cells from guinea pigs or human [30, 31]. Measurements of antigenicity in HI assays are affected by changes in host cell receptor binding [32] that may also drive influenza evolution [33, 34]. The individual effects of antigenicity and receptor avidity on HI assays and their contributions as drivers of viral evolution are not fully understood. There are also differences between antigenicity measurements with human sera compared to more commonly used ferret sera [35]. Secondly, since evaluating the data and producing the vaccine requires up to 8 months, the vaccine strain recommendation is made almost two seasons before the start of the respective SH and NH seasons, from data available until mid-season (Fig. 3). As an antigenically novel strain subsequently rising to predominance may thus go undetected or antigenic drift can still occur, necessary updates are occasionally missed or an antigenically different strain is recommended. Both decreases vaccine effectiveness. False positive vaccine recommendations are rare and occurred only twice from 2003 to 2016/2017 for H3N2 viruses [36-38], namely when A/Wellington/1/2004 was recommended in 2004S [39] and A/Switzerland/9715293/2013 in 2014S [40]. But while the emerging strain was mostly correctly identified, it was recommended oftentimes less than two seasons before reaching predominance, which resulted in a mismatch due to the time-consuming vaccine production. This occurred for example for A/Fujian/411/2002 that was recommended in 2003N for the 2004N vaccine [41], but predominated already in 2003S [42], and A/Wisconsin/67/2005, included in 2006N for season 2007N [43] and predominant in 2006S [44]. Using this criterion, the vaccine composition exactly matched the predominant strain in only 13 of 26 seasons from 2003 to 2016/2017, with mismatches occurring mostly when a novel antigenic variant became predominant [38]. Third, recent H3N2 viruses grow poorly in eggs [45] and frequently acquire “egg adaptations” that alter antigenicity and lower vaccine efficacy, as for the 2012/2013 season [46]. Producing vaccines in mammalian cell culture is a still rarely used alternative with shorter production times and independence from an egg supply [47]. However, adaptive mutations also occur in cell culture [48], potentially leading to antigenic alterations. Overall, the lengthy vaccine production process and technical

issues with the utilized assays and production method contribute to reduced vaccine efficacies and make methodological innovations necessary.

## Computational prediction of viral evolution

Sophisticated computational techniques have been developed within the last decade to predict aspects of the genetic and antigenic evolution of seasonal influenza A viruses. We first discuss approaches that infer amino acid changes, protein sites or entire regions associated with antigenic change and then methods forecasting the predominant strain for the next season, and predicting vaccine strain updates (Table 1). A newly emerging lineage will only make a vaccine strain update necessary, if strains from this lineage are antigenically different from the current vaccine strains and dominant antigenic type.

### *Inferring links between genotype and antigenicity*

For recommending vaccine strains based on surveillance data, knowing the most relevant amino acid substitutions, sites and regions of the HA for antigenic drift is essential. An extensive experimental study characterized the antigenic properties of single and double mutants from representative sequences of past antigenic variants of H3N2 [49]. Changes at seven sites close to the receptor binding site of HA were responsible for most change between consecutive variants in HI assays. Recent computational techniques may provide a rapid and cost-effective alternative to delineate such relationships. These assess the antigenic effects of amino acid changes from large numbers of sequences, including consideration of their evolutionary histories or of higher-order interactions, without requiring mutant strains. We classify the techniques as phylogenetic and population genetics-based or as statistical, including multivariate statistical learning methods (Table 1). The power of identified antigenicity-altering sites to predict future viral evolution is best evaluated on data from other time periods than the one used for inferring these relationships or optimizing model parameters, as successive selective sweeps of antigenically altered lineages (or variants) result in lower genetic and antigenic variation within a particular time period than across time periods, and different amino acid changes predominate (Fig. 4).

Methods that link antigenic and genetic variation rely primarily on statistical techniques or on information theory [50-56]. Suzuki [50] estimated antigenic distances between strains of H3N2 based on a model including physicochemical differences between amino acids, the distance between the site and receptor binding site, or to N-linked glycosylation sites, as well

as solvent accessibility. Though antigenic distances for the analyzed time period were predicted with lower error, the antigenic evolution of H3N2 was predicted with low accuracy. Cui et al. [51] inferred antigenic distances between H3N2 strains and antigenic variants using multivariate regression on multiple physicochemical properties of informative amino acid positions. Ren et al. [52] used multivariate regression and feature selection techniques for the HA of H1N1 viruses circulating until 2008 to identify combinations of protein sites that predict antigenic distances between strains. They thus identified the most relevant candidate sites for the antigenic evolution of the virus for the analyzed time period.

The key argument for methods from phylogenetics and population genetics is that genetic sequences of influenza viruses are closely related to each other, instead of being independent observations, as required for statistical analyses. Many differences that circulating viruses display relative to a previously circulating strain were acquired only once in their shared evolutionary histories. Thus, methods using features derived from sequence data directly count ancestral mutations multiple times, once for each descendant isolate, leading to overestimated significance and a strong effect of sampling. For instance, having data from one lineage overrepresented would likely lead to amino acid changes acquired once in its history to strongly affect the identified sites or changes. One can circumvent this problem by reconstructing a genealogy of the evolutionary relationships from these sequences, and inferring the history of evolutionary events, such as mutations leading to amino acid changes, for the branches of this tree, either with discrete or probabilistic approaches. The evolutionary events are independent from another and can be used for further analysis, including advanced statistics [57].

We developed a method for mapping antigenic distances onto a tree using least squares optimization, resulting in an "antigenic tree" [58]. For HI data mapped onto a HA genealogy of H3N2, this gave a comparatively good solution to representing antigenic distances in a two-dimensional map. The tree has antigenic weights for individual branches, identifying key branches and associated amino acid changes that altered the antigenicity in the evolution of successively circulating antigenic variants of H3N2. Neher et al. [59] described a related model for antigenic evolution on a tree, which also considers avidity and serum potency changes, and demonstrate its application to all circulating subtypes. In another study, antigenicity-altering sites determined from branches with high antigenic weight in the antigenic tree formed distinct patches on the HA structure [60]. Changes in two patches close

to the viral receptor binding site were primarily informative for detecting new antigenic variants.

Visualizing the antigenic relationships between viruses and antisera is commonly done with Antigenic Cartography [21], which uses non-metric multidimensional scaling (NMDS). NMDS places viruses and antisera on a low (usually 2) dimensional map, such that experimentally measured distances derived from HI assay data are best preserved. Future predominant lineages were predicted using a model derived from NDMS applied to protein sequences viral isolates for H3N2 and H1N1 viruses [61]. Bedford et al. [62] described a Bayesian version of antigenic cartography that also models avidity and potency changes and uses phylogenetic information to resolve uncertainty in placement of antigens and antisera.

*In silico prediction of evolution*

Computational methods for predicting the genetic and antigenic evolution of influenza viruses oftentimes identify a particular lineage with associated amino acid changes as becoming predominant in future seasons [38, 63-65]. Predicting vaccine strains can be evaluated with retrospective testing [38] (Fig. 4): a method is applied to data being collected until the meeting of the GISRS committee in season *t* and recommends to leave the vaccine unchanged or an update with a specific strain for season $t+2$, where the vaccine would be available using current production techniques. This recommendation is compared to the truly predominant circulating lineage in $t+2$, from which a matching vaccine strain would originate, based on its typical amino acid changes (found on their initiating branch in the HA genealogy) and whether it was antigenically novel. Predictive success is quantified either using a binary measure per season that compares the predicted lineage to the truly predominant one [38] or with a measure calculating performance per analyzed sequence and season by comparing predicted to observed lineage frequencies [64]. The latter depends on sampling and the relative proportions of different lineages in data and thus is less comparable across seasons.

One of the earliest studies predicted future predominant lineages of H3N2 based on the number of changes in a set of positive selected codons from a genealogy of HA [66]. With retrospective testing, Bush et al. analyzed whether the 'predictive isolate' was closest to the trunk node in a reference tree, from which the future dominant lineage descended. A caveat in this evaluation is the set of codons though, which was identified from data spanning the entire study period [63]; thus a proof of concept on data fully from future seasons is missing.

In Steinbruck and McHardy [67], Allele Dynamics Plots were described, which utilize phylogenetics and sampling times to rank alleles (representing sets of amino acid changes from tree branches) by fitness. Fitness is estimated from their frequency change over consecutive seasons, which for constantly-sized populations identifies those most likely to be under positive selection. We demonstrate its value in determination of viral lineages rising to predominance and vaccine strain selection for H3N2 and H1N1pdm09. On nextflu.org, frequency changes of such alleles, or lineages, can be studied for all circulating influenza types in combination with additional information, such as their geographic origins or changes in epitope sites [68]. Sweep Dynamics (SD) Plots [69] assess the statistical significance for allele dynamics, and better resolve the dynamics of individual changes. SD Plots identified confirmed sites of functional relevance for host adaptation of H1N1pdm09 after 2009 and - combined with information on antigenic sites – outperformed GISRS in predicting vaccine strains for H3N2.

Luksza and Lassig [64] estimated the future fitness of lineages by combining an epidemiological SIR (susceptible-infectious-recovered) model with analysis of HA sequences for the circulating viruses. Strain fitness is determined by similarity to past and presently circulating strains in epitope and non-epitope sites. This method accurately predicted the evolutionary dynamics of H3N2, indicating its suitability for detecting newly emerging lineages. Neher et al. [70] described a lineage fitness model based on the local tree topology. They assumed that an exceptionally fit internal node in a gene tree will be the root of a rapidly branching, and hence expanding, lineage. For a given phylogenetic tree, the branch with the highest fitness predicts the progenitor sequences, which has predictive value regarding H3N2 evolution. Shaman and Karspeck [71] used techniques from numerical weather prediction to combine Google Flu Trends estimates of influenza-like illnesses with epidemiological models. Their approach predicts weekly influenza infection dynamics in real time from big data sets, which could inform vaccination and drug allocation strategies within a season.

Notably, allele and lineage characteristics could be influenced by viral passaging before sequencing [48, 72], geographically biased sampling or strongly reduced viral population sizes (bottlenecks) [73]. To reduce passaging effects, changes on terminal tree branches can be excluded [72], though this may not remove all adaptations [48]. Excluding all passaged isolates will remove most isolates from GISAID [48] and passaging information might be incomplete. Influenza A/H1N1 endemic until 2009 and the B/lineages circulate as distinct lineages over several seasons outside of a reservoir, which might give rise to bottlenecks and

require further consideration for predicting future predominant lineages [6]. For H3N2 viruses, however, relevant bottlenecks are unknown. A viral reservoir is maintained based on low-level year-round persistence and overlapping epidemics, from which it spreads across the globe [3]. Furthermore, changes at sites under positive selection and antigenicity-altering changes in HA are enriched in the surviving lineage, indicating that indeed fitness primarily drives viral evolution [58, 74].

Several methods jointly consider genetic and antigenic information in predicting influenza evolution. We combined the allele dynamic ranking with per-site antigenicity estimates from an antigenic tree [38], which improved H3N2 vaccine recommendations compared to GISRS in retrospective testing. Du et al. [75] inferred antigenic clusters of strains from an antigenic similarity network. The network edges were determined with a Naïve Bayes classifier from structural and physicochemical sequence properties and H3N2 vaccine strains were recommended based on increasing cluster prevalence within a season. This model was also applied to H1N1 viruses from before 2009, H5N1 viruses and all influenza subtypes [76-78]. Suzuki [79] predicted vaccine strains using a fitness model adapted from Luksza and Lassig [64] that considers cross-immunity to other strains and thermodynamic protein stability, optimized for predictive performance from antigenicity estimates.

The outlined computational methods combine techniques and concepts from multivariate statistics, population genetics, epidemiological modelling and phylogenetic theory. They have the potential to inform or even improve vaccine strain selection, due to their speed – for instance in determining antigenicity altering sites in comparison to extensive laboratory experiments of individual mutants - and ability to generate competitive predictive accuracies in vaccine strain prediction. They are well-equipped to handle large and high dimensional complex data types and make fully reproducible suggestions, given that codes and data are provided.

## Concluding Remarks

Over the last decade, in silico methods for predicting the (antigenic) evolution of seasonal influenza A viruses have made great strides. This has in part been made possible by the recent commitment of the GISRS and experimental labs to a timely release of sequences [80]. The importance of this practice for the computational and modelling fields to further mature cannot be emphasized enough. We urge all participating laboratories, agencies and research

teams to support this effort and apply this also for relevant experimental data, such as HI or neutralization assay information.

In comparison to GISRS, computational techniques identified suitable (strains of) vaccine lineages with improved performance in retrospective testing; however, this has limitations. Only GISRS currently solves the vaccine strain problem truly for the future. Furthermore, details of the performance evaluations differ: some studies predict strains for the immediately following season, while in practice vaccine strains are recommended two seasons before. For a realistic comparison, also only data available to GISRS at the vaccine strain meeting should be considered. Since a few years GISRS uses the sequences submitted to the GISAID EpiFlu database (http://platform.gisaid.org) until 10 to 12 days before the meeting and the number of deposited sequences has grown substantially. For instance, 13 sequences were isolated and submitted within the 2008S season, while it were 632 sequences in the 2016N season [69, 81]. Notably, though, another 2404 sequences isolated in the 2016N season were deposited after the meeting, and thus had no value for the prediction. Finally, GISRS sometimes could not select a well-matching strain, even though it attempted to, when the strain failed to grow in eggs [82]. However, the vaccine selection process itself is ideally suited for a realistic benchmarking of computational methods, which is usually not available in other disciplines: data generated by GISRS could be continuously publicized, and a public vaccine strain prediction track established in parallel to the GISRS procedure, with results and methods included in WHO reports. This could accelerate method development (see Outstanding Questions), and within a few years, ideally lead to improved and reproducible vaccine selection and realistic estimates of the inherent performance limits for the process

Ultimately, vaccines should stimulate broad long-term immunity across antigenically different strains, to render frequent vaccine updates unnecessary and maybe even protect against pandemic viruses [83]. The feasibility of this hypothesis was proven by the identification of broadly neutralizing antibodies against HA [84-86]. One method that predicts viral antigens for such a vaccine is COBRA [87]. It computationally identifies an epitope sequence representing a broad taxonomic range of isolates by multiple rounds of consensus generation from their sequences. Novel hemagglutinins generated that way for H5N1 and H1N1 elicited a broad antibody response in animal models [88, 89]. Nevertheless, universal long term protection will probably require both humoral and cell mediated immunity. This is supported by the correlation between pre-existing influenza specific T helper cells with protection against experimental influenza challenge in humans [90]. Other approaches target regions that

are highly conserved across different subtypes, such as the HA stem [91-93], the M2 ectodomain [94] or conserved T cell epitopes on internal proteins [95]. Most of these approaches are still being developed, with M2 and T cell based vaccines in clinical trials [47, 94, 95]. If successful, the number of vaccinations could be substantially reduced, with boosts of immunity required only every 5-15 years for protection against seasonal influenza viruses [96].

## Acknowledgements

We thank John McCauley for his comments and expert insights and Linus Roune for creating figures 2 and 3.

# References


1 Bouvier, N.M. and Palese, P. (2008) The biology of influenza viruses. *Vaccine* 26 Suppl 4, D49-53

2 Tong, S.*, et al.* (2013) New world bats harbor diverse influenza A viruses. *PLoS pathogens* 9, e1003657

3 Russell, C.A.*, et al.* (2008) The global circulation of seasonal influenza A (H3N2) viruses. *Science* 320, 340-346

4 Lemey, P.*, et al.* (2014) Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS pathogens* 10, e1003932

5 Bedford, T.*, et al.* (2010) Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). *PLoS pathogens* 6, e1000918

6 Bedford, T.*, et al.* (2015) Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* 523, 217-220

7 Gomez-Puertas, P.*, et al.* (2000) Influenza virus matrix protein is the major driving force in virus budding. *Journal of virology* 74, 11538-11547

8 Jagger, B.W.*, et al.* (2012) An overlapping protein-coding region in influenza A virus segment 3 modulates the host response. *Science* 337, 199-204

9 Bavagnoli, L.*, et al.* (2015) The novel influenza A virus protein PA-X and its naturally deleted variant show different enzymatic properties in comparison to the viral endonuclease PA. *Nucleic acids research* 43, 9405-9417

10 Vasin, A.V.*, et al.* (2014) Molecular mechanisms enhancing the proteome of influenza A viruses: an overview of recently discovered proteins. *Virus research* 185, 53-63

11 Eichelberger, M.C. and Wan, H. (2015) Influenza neuraminidase as a vaccine antigen. *Current topics in microbiology and immunology* 386, 275-299

12 O'Brien, K.B.*, et al.* (2011) A protective role for complement C3 protein during pandemic 2009 H1N1 and H5N1 influenza A virus infection. *PloS one* 6, e17377

13 Jegaskanda, S.*, et al.* (2017) Induction of H7N9-Cross-Reactive Antibody-Dependent Cellular Cytotoxicity Antibodies by Human Seasonal Influenza A Viruses that are Directed Toward the Nucleoprotein. *The Journal of infectious diseases* 215, 818-823

14 Terajima, M.*, et al.* (2015) High Antibody-Dependent Cellular Cytotoxicity Antibody Titers to H5N1 and H7N9 Avian Influenza A Viruses in Healthy US Adults and Older Children. *The Journal of infectious diseases* 212, 1052-1060

15 de Vries, R.D.*, et al.* (2017) Influenza virus-specific antibody dependent cellular cytoxicity induced by vaccination or natural infection. *Vaccine* 35, 238-247

16 Jegaskanda, S.*, et al.* (2013) Cross-reactive influenza-specific antibody-dependent cellular cytotoxicity antibodies in the absence of neutralizing antibodies. *Journal of immunology* 190, 1837-1848

17 Liu, Y.*, et al.* (2017) Uncompromised NK cell activation is essential for virus-specific CTL activity during acute influenza virus infection. *Cellular & molecular immunology* 10.1038/cmi.2017.10

18 Co, M.D.*, et al.* (2014) Relationship of preexisting influenza hemagglutination inhibition, complement-dependent lytic, and antibody-dependent cellular cytotoxicity antibodies to the development of clinical illness in a prospective study of A(H1N1)pdm09 Influenza in children. *Viral immunology* 27, 375-382

19 Sanjuan, R.*, et al.* (2010) Viral mutation rates. *Journal of virology* 84, 9733-9748

20 Holmes, E.C.*, et al.* (2005) Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS biology* 3, e300

21 Smith, D.J.*, et al.* (2004) Mapping the antigenic and genetic evolution of influenza virus. *Science* 305, 371-376

22 Tisa, V.*, et al.* (2016) Quadrivalent influenza vaccine: a new opportunity to reduce the influenza burden. *Journal of preventive medicine and hygiene* 57, E28-33

23 Gaglani, M.*, et al.* (2016) Influenza Vaccine Effectiveness Against 2009 Pandemic Influenza A(H1N1) Virus Differed by Vaccine Type During 2013-2014 in the United States. *The Journal of infectious diseases* 213, 1546-1556
24 Zimmerman, R.K.*, et al.* (2016) 2014-2015 Influenza Vaccine Effectiveness in the United States by Vaccine Type. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 63, 1564-1573
25 Simpson, C.R.*, et al.* (2015) Trivalent inactivated seasonal influenza vaccine effectiveness for the prevention of laboratory-confirmed influenza in a Scottish population 2000 to 2009. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 10.2807/1560-7917.ES2015.20.8.21043
26 Van Buynder, P.G.*, et al.* (2013) The comparative effectiveness of adjuvanted and unadjuvanted trivalent inactivated influenza vaccine (TIV) in the elderly. *Vaccine* 31, 6122-6128
27 DiazGranados, C.A.*, et al.* (2014) Efficacy of high-dose versus standard-dose influenza vaccine in older adults. *The New England journal of medicine* 371, 635-645
28 Medeiros, R.*, et al.* (2001) Hemagglutinin residues of recent human A(H3N2) influenza viruses that contribute to the inability to agglutinate chicken erythrocytes. *Virology* 289, 74-85
29 Nobusawa, E.*, et al.* (2000) Change in receptor-binding specificity of recent human influenza A viruses (H3N2): a single amino acid change in hemagglutinin altered its recognition of sialyloligosaccharides. *Virology* 278, 587-596
30 Lin, Y.P.*, et al.* (2012) Evolution of the receptor binding properties of the influenza A(H3N2) hemagglutinin. *Proceedings of the National Academy of Sciences of the United States of America* 109, 21474-21479
31 Ampofo, W.K.*, et al.* (2013) Strengthening the influenza vaccine virus selection and development process: outcome of the 2nd WHO Informal Consultation for Improving Influenza Vaccine Virus Selection held at the Centre International de Conferences (CICG) Geneva, Switzerland, 7 to 9 December 2011. *Vaccine* 31, 3209-3221
32 Li, Y.*, et al.* (2013) Single hemagglutinin mutations that alter both antigenicity and receptor binding avidity influence influenza virus antigenic clustering. *Journal of virology* 87, 9904-9910
33 Meyer, A.G. and Wilke, C.O. (2015) Geometric Constraints Dominate the Antigenic Evolution of Influenza H3N2 Hemagglutinin. *PLoS pathogens* 11, e1004940
34 Hensley, S.E.*, et al.* (2009) Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. *Science* 326, 734-736
35 Linderman, S.L.*, et al.* (2014) Potential antigenic explanation for atypical H1N1 infections among middle-aged adults during the 2013-2014 influenza season. *Proceedings of the National Academy of Sciences of the United States of America* 111, 15798-15803
36 WHO (2005) Recommended composition of influenza virus vaccines for use in the 2005–2006 influenza season. *WHO Weeky Epidemiological Record* 80, 65-76
37 WHO (2015) Recommended composition of influenza virus vaccines for use in the 2015-2016 northern hemisphere influenza season. *WHO Weekly Epidemiological Record* 90, 97-108
38 Steinbruck, L.*, et al.* (2014) Computational prediction of vaccine strains for human influenza A (H3N2) viruses. *Journal of virology* 88, 12123-12132
39 WHO (2004) Recommended composition of influenza virus vaccines for use in the 2005 influenza season. *WHO Weeky Epidemiological Record* 79, 369-376
40 WHO (2014) Recommended composition of influenza virus vaccines for use in the 2015 southern hemisphere influenza season. *WHO Weeky Epidemiological Record* 89, 441-456
41 WHO (2003) Recommended composition of influenza virus vaccines for use in the 2003–2004 influenza season. *WHO Weeky Epidemiological Record* 78, 58-62
42 WHO (2003) Recommended composition of influenza virus vaccines for use in the 2004 influenza season. *WHO Weeky Epidemiological Record* 78, 375-379
43 WHO (2006) Recommended composition of influenza virus vaccines for use in the 2006–2007 influenza season. *WHO Weeky Epidemiological Record* 81, 81-88

44 WHO (2006) Recommended composition of influenza virus vaccines for use in the 2007 influenza season. *WHO Weekly Epidemiological Record* 81, 390-395
45 Lu, B*., et al.* (2005) Improvement of influenza A/Fujian/411/02 (H3N2) virus growth in embryonated chicken eggs by balancing the hemagglutinin and neuraminidase activities, using reverse genetics. *Journal of virology* 79, 6763-6771
46 Skowronski, D.M*., et al.* (2014) Low 2012-13 influenza vaccine effectiveness associated with mutation in the egg-adapted H3N2 vaccine strain not antigenic drift in circulating viruses. *PloS one* 9, e92153
47 Soema, P.C*., et al.* (2015) Current and next generation influenza vaccines: Formulation and production strategies. *European journal of pharmaceutics and biopharmaceutics : official journal of Arbeitsgemeinschaft fur Pharmazeutische Verfahrenstechnik e.V* 94, 251-263
48 McWhite, C.D*., et al.* (2016) Sequence amplification via cell passaging creates spurious signals of positive adaptation in influenza virus H3N2 hemagglutinin. *Virus evolution* 10.1093/ve/vew026
49 Koel, B.F*., et al.* (2013) Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science* 342, 976-979
50 Suzuki, Y. (2013) Predictability of antigenic evolution for H3N2 human influenza A virus. *Genes & genetic systems* 88, 225-232
51 Cui, H*., et al.* (2014) Using multiple linear regression and physicochemical changes of amino acid mutations to predict antigenic variants of influenza A/H3N2 viruses. *Bio-medical materials and engineering* 24, 3729-3735
52 Ren, X*., et al.* (2015) Computational Identification of Antigenicity-Associated Sites in the Hemagglutinin Protein of A/H1N1 Seasonal Influenza Virus. *PloS one* 10, e0126742
53 Xia, Z*., et al.* (2009) Using a mutual information-based site transition network to map the genetic evolution of influenza A/H3N2 virus. *Bioinformatics (Oxford, England)* 25, 2309-2317
54 Lees, W.D*., et al.* (2010) A computational analysis of the antigenic properties of haemagglutinin in influenza A H3N2. *Bioinformatics (Oxford, England)* 26, 1403-1408
55 Huang, J.W*., et al.* (2009) Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses. *BMC bioinformatics* 10 Suppl 1, S41
56 Liao, Y.C*., et al.* (2008) Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. *Bioinformatics (Oxford, England)* 24, 505-512
57 Felsenstein, J. (2003) *Inferring Phylogenies*. Sinauer
58 Steinbruck, L. and McHardy, A.C. (2012) Inference of genotype-phenotype relationships in the antigenic evolution of human influenza A (H3N2) viruses. *PLoS computational biology* 8, e1002492
59 Neher, R.A*., et al.* (2016) Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. 113, E1701-1709
60 Kratsch, C*., et al.* (2016) Determination of antigenicity-altering patches on the major surface protein of human influenza A/H3N2 viruses. *Virus evolution* 10.1093/ve/vev025
61 Ito, K*., et al.* (2011) Gnarled-trunk evolutionary model of influenza A virus hemagglutinin. *PloS one* 6, e25953
62 Bedford, T*., et al.* (2014) Integrating influenza antigenic dynamics with molecular evolution. *eLife* 3, e01914
63 Bush, R.M*., et al.* (1999) Predicting the evolution of human influenza A. *Science* 286, 1921-1925
64 Luksza, M. and Lassig, M. (2014) A predictive fitness model for influenza. *Nature* 507, 57-61
65 He, J. and Deem, M.W. (2010) Low-dimensional clustering detects incipient dominant influenza strain clusters. *Protein engineering, design & selection : PEDS* 23, 935-946
66 Bush, R.M*., et al.* (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Molecular biology and evolution* 16, 1457-1465
67 Steinbruck, L. and McHardy, A.C. (2011) Allele dynamics plots for the study of evolutionary dynamics in viral populations. *Nucleic acids research* 39, e4
68 Neher, R.A. and Bedford, T. (2015) nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics (Oxford, England)* 31, 3546-3548

69 Klingen, T.R.*, et al.* (2017) Sweep Dynamics (SD) plots: Computational identification of selective sweeps to monitor the adaptation of influenza A viruses. *bioRxiv* 10.1101/110528
70 Neher, R.A.*, et al.* (2014) Predicting evolution from the shape of genealogical trees. *eLife* 10.7554/eLife.03568
71 Shaman, J. and Karspeck, A. (2012) Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences of the United States of America* 109, 20425-20430
72 Chen, H.*, et al.* (2016) Dynamic Convergent Evolution Drives the Passage Adaptation across 48 Years' History of H3N2 Influenza Evolution. *Molecular biology and evolution* 33, 3133-3143
73 Kim, K. and Kim, Y. (2015) Episodic nucleotide substitutions in seasonal influenza virus H3N2 can be explained by stochastic genealogical process without positive selection. *Molecular biology and evolution* 32, 704-710
74 Fitch, W.M.*, et al.* (1991) Positive Darwinian evolution in human influenza A viruses. *Proceedings of the National Academy of Sciences of the United States of America* 88, 4270-4274
75 Du, X.*, et al.* (2012) Mapping of H3N2 influenza antigenic evolution in China reveals a strategy for vaccine strain recommendation. *Nature communications* 3, 709
76 Peng, Y.*, et al.* (2017) A universal computational model for predicting antigenic variants of influenza A virus based on conserved antigenic structures. *Scientific reports* 7, 42051
77 Peng, Y.*, et al.* (2017) Continual Antigenic Diversification in China Leads to Global Antigenic Complexity of Avian Influenza H5N1 Viruses. *Scientific reports* 7, 43566
78 Liu, M.*, et al.* (2015) Antigenic Patterns and Evolution of the Human Influenza A (H1N1) Virus. *Scientific reports* 5, 14171
79 Suzuki, Y. (2015) Selecting vaccine strains for H3N2 human influenza A virus. *Meta gene* 4, 64-72
80 Shu, Y. and McCauley, J. (2017) GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 10.2807/1560-7917.ES.2017.22.13.30494
81 Klingen, T.R.*, et al.* (2017) hzi-bifo/SDplots: First release of SD plots data. 10.5281/zenodo.831631
82 WHO (2003) Addendum to the recommended composition of influenza virus vaccines for use in the 2003-2004 influenza season. *WHO Weekly Epidemiological Record* 78, 77
83 Nabel, G.J. and Fauci, A.S. (2010) Induction of unnatural immunity: prospects for a broadly protective universal influenza vaccine. *Nature medicine* 16, 1389-1391
84 Corti, D.*, et al.* (2011) A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science* 333, 850-856
85 Ekiert, D.C.*, et al.* (2011) A highly conserved neutralizing epitope on group 2 influenza A viruses. *Science* 333, 843-850
86 Dreyfus, C.*, et al.* (2012) Highly conserved protective epitopes on influenza B viruses. *Science* 337, 1343-1348
87 Giles, B.M. and Ross, T.M. (2011) A computationally optimized broadly reactive antigen (COBRA) based H5N1 VLP vaccine elicits broadly reactive antibodies in mice and ferrets. *Vaccine* 29, 3043-3054
88 Carter, D.M.*, et al.* (2016) Design and Characterization of a Computationally Optimized Broadly Reactive Hemagglutinin Vaccine for H1N1 Influenza Viruses. *Journal of virology* 90, 4720-4734
89 Crevar, C.J.*, et al.* (2015) Cocktail of H5N1 COBRA HA vaccines elicit protective antibodies against H5N1 viruses from multiple clades. *Human vaccines & immunotherapeutics* 11, 572-583
90 Wilkinson, T.M.*, et al.* (2012) Preexisting influenza-specific CD4+ T cells correlate with disease protection against influenza challenge in humans. *Nature medicine* 18, 274-280
91 Impagliazzo, A.*, et al.* (2015) A stable trimeric influenza hemagglutinin stem as a broadly protective immunogen. *Science* 349, 1301-1306
92 Yassine, H.M.*, et al.* (2015) Hemagglutinin-stem nanoparticles generate heterosubtypic influenza protection. *Nature medicine* 21, 1065-1070
93 Valkenburg, S.A.*, et al.* (2016) Stalking influenza by vaccination with pre-fusion headless HA mini-stem. *Scientific reports* 6, 22666

94 Turley, C.B.*, et al.* (2011) Safety and immunogenicity of a recombinant M2e-flagellin influenza vaccine (STF2.4xM2e) in healthy adults. *Vaccine* 29, 5145-5152
95 Lillie, P.J.*, et al.* (2012) Preliminary assessment of the efficacy of a T-cell-based influenza vaccine, MVA-NP+M1, in humans. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 55, 19-25
96 Cimons, M. (2016) Universal Influenza Vaccine: Quest in Sight? *Microbe Magazine* 11, 433-437
97 McHardy, A.C. and Adams, B. (2009) The role of genomics in tracking the evolution of influenza A virus. *PLoS pathogens* 5, e1000566

## Figure Legends

**Figure 1: Schematic view of an influenza A virion and its RNA segments (from [97]).**

The genome of influenza A viruses consists of eight segments (bottom left corner) encoding 12 or more proteins, including splice variants. HA, NA, M1, M2, PA, PB1, PB2 and NP are shown in the figure.

**Figure 2: Evolutionary arms race between seasonal influenza viruses and their human host**

Influenza viruses are shown with their genomic segments inside the virion and the proteins hemagglutinin and neuraminidase on the surface. Different antigenic properties of the hemagglutinin are indicated by different colors (yellow, red, green). Antibodies produced by the host's immune system are shaped like Ts, with colors indicating which viral hemagglutinin the antigen-binding site is able to recognize. After the host acquires immunity against a specific influenza strain, either by previous infections or vaccination, neutralizing antibodies bind to the hemagglutinin on the virus surface, contributing to clearance of the infection. The viral hemagglutinin accumulates mutations that change its antigenic properties, allowing the virus to escape the host's immune response, until immunity is acquired again, as indicated by the dotted arrow.

**Figure 3: GISRS vaccine strain selection**

Timeline of GISRS surveillance, data analysis, vaccine strain selection and vaccine production for both the Northern (blue, inner circle) and Southern hemisphere (yellow, outer circle). Isolate collection is performed year round, with isolates from October to end of January considered for the vaccine in the Northern hemisphere and isolates from March to the end of August considered for the Southern hemisphere. Antigenic and genetic data is analyzed in January and August, to decide the vaccine composition at meetings in February and September. The production and analysis of candidate vaccine viruses starts in parallel to the isolate collection to ensure the availability of a candidate virus at the vaccine selection meeting. In the following eight months, the vaccine is produced and available at the beginning of the Northern or Southern hemisphere influenza season in the October or May, respectively.

**Figure 4: Genealogy of HA for H3N2 viruses from 1968 onwards, demonstrating the concept of retrospective testing.**

Leaf nodes are colored by year of isolation for viral isolates, indicating the temporal structure of the genealogy and the presence of one surviving lineage over time. After 2002, amino acid changes for the trunk and major branches are indicated. To evaluate the performance of vaccine strain predictions, data until the time of the GISRS meeting in season $t$ (red box) is analyzed. Based on these data, a vaccine lineage or strain is predicted for season $t + 2$, when the vaccine would become available. This prediction is compared to the predominant lineage for this season (black box). The figure was adapted from [60]. Green indicates changes in antigenicity-altering "patches" of residues on the protein structure.

# Tables

| Data/Information used | Method category | Predicted outputs | References |
|---|---|---|---|
| S | P, ST | C | Bush et al. [63] |
| S | P | C, AA | Steinbruck and McHardy [67] |
| S | P, ST, R | C | Luksza and Lassig [64] |
| S, H | P, ST | C, AA | Steinbruck et al. [38] |
| S, H, PR, O (Physicochemical Properties) | P, R | C | Suzuki [79] |
| S | P | C | Neher et al. [70] |
| S,(H), PR | P, ST | C, AA | Klingen et al. [69] |
| E (Infection Rates) | R | O (Influenza Peaks) | Shaman and Karspeck [71] |
| S | O (Information Theory) | AA | Xia et al. [53] |
| S,H | P, ST | AA | Steinbruck and McHardy [58] |
| S, H, PR, O (Physicochemical Properties) | ST | AC | Du et al. [75], Liu et al. [78], Peng et al. [77], Peng et al. [76] |
| PR, O (Physicochemical Information) | ST | AA | Suzuki [50] |
| S, H | ST | AA | Cui et al. [51] |
| S, H | P, ST | O (Antigenic Map) | Bedford et al. [62] |
| S, H | ST | AA | Ren et al. [52] |
| S, H, PR | P, O (Graph Theory) | AA | Kratsch et al. [60] |
| S, H | P, ST | AA | Neher et al. [59] |

**Table 1: Recent computational methods predicting antigenicity-altering sites, future predominant lineages or vaccine strains for human influenza A viruses.**

S: Viral Sequences; H: HI Assay Data; (H): Derivative of HI assay data (e.g. antigenic patches); PR: Protein Structure; E: Epidemiological Information; P: Phylogenetics and Population Genetics; ST: Statistical Methods; R: Epidemiological Models; C: Lineages/Clades; AA: Amino Acids; AC: Antigenic Cluster; O: Other.