Concepts and methods for predicting viral evolution

Matthijs Meijers*[a], Denis Ruchnewitz*[a], Jan Eberhardt[a], Malancha Karmakar[a], Marta Łuksza†[b], and Michael Lässig†[a]

[a]Institute for Biological Physics, University of Cologne, Zülpicherstr. 77, 50937, Köln, Germany
[b]Tisch Cancer Institute, Departments of Oncological Sciences and Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

April 29, 2024

Summary

The seasonal human influenza virus undergoes rapid evolution, leading to significant changes in circulating viral strains from year to year. These changes are typically driven by adaptive mutations, particularly in the antigenic epitopes, the regions of the viral surface protein haemagglutinin targeted by human antibodies. Here we describe a consistent set of methods for data-driven predictive analysis of viral evolution. Our pipeline integrates four types of data: (1) sequence data of viral isolates collected on a worldwide scale, (2) epidemiological data on incidences, (3) antigenic characterization of circulating viruses, and (4) intrinsic viral phenotypes. From the combined analysis of these data, we obtain estimates of relative fitness for circulating strains and predictions of clade frequencies for periods of up to one year. Furthermore, we obtain comparative estimates of protection against future viral populations for candidate vaccine strains, providing a basis for pre-emptive vaccine strain selection. Continuously updated predictions obtained from the prediction pipeline for influenza and SARS-CoV-2 are available on the website previr.app.

Keywords: antigenic evolution, population immunity, influenza vaccines, fitness models

1 Introduction

Global populations of the influenza virus encompass multiple antigenically distinct groups. The strains descending from a given mutant strain define a clade of the evolving viral population. A part of the genetic diversity encodes antigenic diversity: new strains escape human immunity generated by infection or vaccination with previous strains – partially and into different directions. Antigenic diversity enables fast evolution driven by human immune pressure, leading to genetic and antigenic turnover on timescales of a few years. Thus, influenza evolution continuously decreases the protection of a given vaccine and requires regular vaccine updates. Given the time required for the development, production, and delivery of vaccines, current decisions to update the vaccine strain are made about

*Equal contributions
†To whom correspondence should be addressed. Email: marta.luksza@mssm.edu, mlaessig@uni-koeln.de

nine months in advance of the next influenza season. Hence, optimal vaccines should be chosen in a pre-emptive way, to provide the best response to the strains circulating in the next season. These decisions benefit from modeling in two ways: to predict the antigenic characteristics of next year's circulating strains and to estimate the protection profile of different vaccine candidates against these strains.

Two kinds of methods are currently used for influenza predictions in the context of vaccine strain selection. One strategy is to estimate growth differences between viral clades from a sequence- or phenotype-based fitness model. The original predictive fitness model for influenza uses amino acid mutations within antigenic sites of influenza hemagglutinin as markers of positive selection for antigenic escape, as well as mutations outside epitopes as a measure of negative selection (mutational load) [1]. Antigenic assays measure the binding or neutralization capacity of host antisera against panels of test strains. More recently, suitably curated data from such assays [2] have been integrated into antigenic fitness models for predictions [3, 4, 5]. A complementary, model-free approach is to harvest genealogical trees built from viral sequences for information on the recent growth of genetic clades; this information can be extrapolated to predict near-future clade frequencies [6].

Digesting viral data into successful predictions poses a fundamental problem: to assess the effects of genetic and phenotypic changes in viral strains facing the heterogenous population of all of our immune systems. Moreover, viruses and immune systems are coupled in a fast co-evolutionary process, so successful antigenic predictions have to track, and project into the future, the changes in viruses and in human population immunity [5].

The computational analysis of influenza evolution, including predictions of near-future viral populations, is made possible by a unique combination of available data. First, sequence evolution of human influenza lineages is documented by worldwide surveillance over several decades [7]. Currently, about 40.000 sequences of human influenza are sequenced each year. Second, influenza incidences, in part subtyped by lineage, is reported from multiple countries on a weekly basis. Third, the interactions of human and ferret antisera and viruses are extensively characterized by antigenic assays, specifically hemagglutination-inhibition (HI) and neutralization tests [8, 9, 10, 11, 12, 13].

In this chapter, we describe an integrative suite of methods that combines all of these data to predict the prevalence and characteristics of emerging variants for fast-evolving respiratory viruses (Fig. 1). We focus on the analysis of human influenza, and we compare with parallel methods and results for SARS-CoV-2 where appropriate. We first describe the computational processing of viral sequence data, including the construction of genealogical trees and the evolutionary tracking of viral clades. As an application, we discuss how sequence and tree data can be combined to infer selection on the evolution of influenza proteins. Next, we describe the integration of epidemiological and antigenic data, which provide information on human immune pressure acting on viral evolution. We assemble this input into a data-driven fitness model for viral evolution and develop model-based computational estimates of vaccine protection.

2 Materials

2.1 Genetic data

The repositories GenBank [14] and GISAID [15] contain large numbers of influenza genome sequences (currently > 400.000 for human seasonal lineages). We use a primary curation procedure to exclude low-quality or incomplete sequences. We perform the following steps: (1) Initial curation: raw, unaligned consensus nucleotide sequences with $> 1\%$ ambiguous characters or without a complete collection

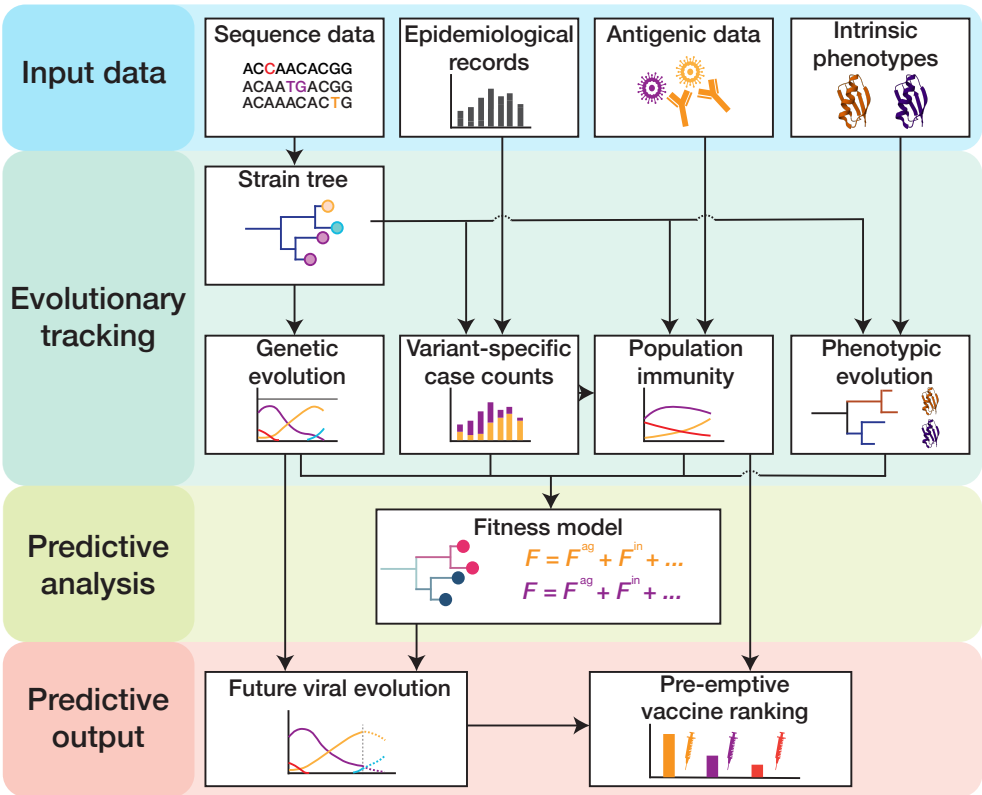

Figure 1. Pipeline for predictive analysis of viral evolution. (1) **Input data:** viral sequence data, epidemiological data, antigenic data, intrinsic viral phenotypes (protein stability, receptor binding, cell entry pathway, intra-cellular replication efficiency, ...). (2) **Evolutionary tracking:** viral evolution (clade frequency trajectories and empirical clade fitness trajectories), clade-specific incidence dynamics, population immunity, and phenotypic evolution are tracked using a timed strain tree. (3) **Predictive analysis:** fitness models are built from tracked immune data and tree-mapped sequence data and are calibrated by empirical fitness. (4) **Predictive output:** fitness models serve to predict clade trajectories and antigenic evolution and to rank candidate vaccines by their predicted protection against future viral strains.

date are excluded. (2) Alignment: the remaining sequences are aligned with MAFFT [16], using a reference isolate from GenBank (typically one of the oldest isolates available, e.g., A/HongKong/1-5-MA21-1/1968 for influenza A(H3N2)). (3) Secondary curation: From the aligned set, we discard sequences containing $> 1\%$ ambiguous characters. (4) Exclusion of outlier sequences (optional): We record pairwise sequence divergence and exclude isolates outside the 95th percentile within a given month. This step can filter out isolates from potential zoonotic events (transmissions from animals to human hosts). Together, these steps produce a set of aligned, high-quality sequences of viral isolates for all subsequent analyses (currently about 40.000 sequences/year for human seasonal influenza).

2.2 Epidemiological data

The size of influenza epidemics varies considerably between years, lineages, and geographical regions [17, 18, 19]. Factors determining the size of the influenza epidemic include the antigenic novelty of the predominant strains, as well as environmental, and socio-demographic factors [20]. We track the number of influenza cases using the epidemiological records provided by FluNet [21], a surveillance tool

of the Global Influenza Surveillance and Response System (GISRS) FluNet reports weekly numbers of influenza virus infections that are detected at the country-level, totalling about 370.000 influenza cases per year from 122 different countries, regions, or territories. Detected influenza cases are tested for influenza type (A or B); subtyping into lineages is provided for a large fraction of the cases. For cases that are not subtyped, we assign the case data to the different influenza subtypes, using the distribution of positive tests at that time in the corresponding region.

2.3 Antigenic data

Antigenic assays measure the neutralization capacity of antiserum from a human or animal host against a panel of test viruses. Common assays include the classical hemagglutination-inhibition (HI) assay [10] and various types of viral neutralization assays [11, 12]. In an HI assay, binding of viral particles to red blood cells prevents their hemagglutination; antibody-antigen binding can disrupt hemagglutination. The assay is performed with serial 2-fold dilutions of the antiserum. The maximum $\log_2$ dilution titer at which the red blood cells do not agglutinate, T, provides a readout of the net binding between antiserum and virions [22]. In neutralization assays, the antiserum is titrated in the presence of virions and target cells; the recorded titer measures the minimum serum concentration needed to inhibit cell infection by 50%. In both assays, higher titers correspond to better recognition of the test virus by the antiserum.

In the common animal model for influenza, naive ferrets are infected with an influenza strain which we call the reference strain. Antigenic assays record the titer T_i^j that quantifies the neutralization of virus i by antiserum raised against reference strain j. Absolute titer values are confounded by non-antigenic effects, such as the potency of the antiserum or the avidity of the red blood cells used in the HI assay [23] . Therefore, we preprocess titer data to remove spurious variation in the homologous titers T_j^j: we apply shifts $T_i^j \to T_i^j + (\langle T_j^j \rangle - T_j^j)$, where $\langle T_j^j \rangle$ is the average homologous titer in the raw data (for influenza ferret assays, $\langle T_j^j \rangle \approx 11$). These standardized titers are used in the remainder of the text. The transformation leaves titer differences $\Delta T_i^j = T_j^j - T_i^j$, which measure evolutionary differences between viral strains, unchanged. For the predictive data analysis reported in this chapter, we use a published set of primary antigenic data produced by the Worldwide Influenza Centre, London [24]. This dataset contains nearly 30.000 HI titers with test strains in the period 2010 – 2023. Reference and test strains are carefully chosen to cover the antigenic diversity of the evolving population of circulating strains.

Two further types of antigenic data have recently been included into evolutionary analysis: antigenic assays using polyclonal human antisera [25, 26] and deep mutational scanning data (DMS) [25, 26, 27]. In DMS experiments, a given backbone viral strain is used to construct a library of mutant strains, most of which contain a single point mutation away from the backbone strain. Then, in a high-throughput way, an escape score against a given antiserum is evaluated for each mutant [28, 29]. In sections 3.3 and 3.4, we discuss the extension of predictive methods to these data and highlight challenges specific to each data type.

3 Predictive analysis of viral evolution

3.1 Tracking viral evolution

The continuous, time-resolved surveillance of circulating influenza viruses is key to track evolutionary changes [30] and is the basis for predictive analysis. Here we discuss a pipeline for evolutionary

tracking that combines sequence data, epidemiological records, and phenotypic data. Based on a timed genealogical tree of the evolving viral population, we first construct trajectories of evolutionary change at the level of clade frequencies, clade fitness, and mutant allele frequencies. Second, integrating information from epidemiological records, we obtain trajectories of clade-specific case counts. Third, we discuss the mapping of molecular phenotypic data to the strain tree; the special case of antigenic phenotypes will be taken up in section 3.3.

Tree reconstruction and reassortment detection

Genealogical trees are graphical representations of inferred evolutionary relationships among collected samples, illustrating their common ancestry and divergence over time [31]. Two types of tree reconstruction methods are most commonly used for viral data analysis: maximum parsimony approaches analyze sequences directly to infer a tree topology with the least evolutionary changes. Maximum likelihood methods use statistical models of sequence evolution to estimate the probability of a particular tree structure given the observed data, together with the probability of the evolutionary steps along the tree. Maximum likelihood approaches are generally favored for their accuracy and flexibility, despite their higher computational demands [32].

We use a probabilistic, maximum-likelihood approach utilizing IQTree2 [33], which supports various evolutionary models. Following an inference of the tree topology, TreeTime [34] is employed to infer the time and sequences of internal nodes in the resulting trees. We perform the following steps: (1) Determine a maximum-likelihood sequence evolution model (if unknown). ModelFinder [35] identifies nucleotide substitution models that best fit the data based on likelihood metrics. For seasonal influenza, a suitable assumption is the general-time-reversible model (GTR) with invariant sites (I), nucleotide frequencies inferred from data (F), and site-specific rates of 8 categories (R8) (the corresponding model name for IQTree2 is GTR+F+I+R8). (2) Construct a maximum-likelihood tree of curated genetic data using IQTree2. The output from step 2 is a nexus file that stores the inferred topology. (3) Use TreeTime [34] to infer the ancestral sequences and the time of internal nodes of the genealogical tree.

The result of this procedure above is a timed, genealogical tree of available sequences (see Note 1 for further discussion). A strain tree for influenza A(H3N2) is shown in Figure 2A. The tree describes the tracked genetic evolution of the virus; individual nucleotide and amino acid changes are mapped on specific branches. Collected isolates are depicted on the tree as leaves (small circles). Internal nodes (positioned at junctions of branches) represent the inferred common ancestors of the isolates. The tree is colored by clades, which are genetically distinct groups of strains with a recent common ancestor. In defining a set of clades, the appropriate level of coarse-graining is an intricate problem and depends on the goals of the analysis. Criteria for delineating clades include epitope mutations or measured antigenic differences, changes in protein stability or other phenotypes, or significant growth increase compared to the ancestral clade. All of these changes are possible markers of fitness differences relevant for evolutionary predictions (section 3.4).

Influenza viruses have a segmented genome with multiple chromosome-like segments (8 segments for influenza A and B). These segments can mix during replication if a host is co-infected with two distinct strains. This process, called reassortment, significantly contributes to the genetic diversity of influenza, particularly for influenza type A viruses [36, 37]. Reassortment events break the assumption that each sequence originates from a single parent, which underlies tree reconstruction. Therefore, trees are commonly reconstructed for each segment independently; this step is followed by tree reconciliation analysis using visualization with tanglegrams, manual comparison [36, 38], or algorithmic approaches [39, 40, 41, 42]. Here, we implement a joint procedure of tree reconstruction and reassortment

detection for the genes encoding the surface proteins hemagglutinin (HA) and neuraminidase (NA), which accumulate more sequence changes than other genes. This procedure contains computationally efficient reassortment detection operating at the level of sequences: by comparison with earlier sequences, we identify isolates that have an exceptionally high number of changes in the NA segment, compared to a null model of sequence evolution by point mutations; a related approach has been proposed in ref. [43]. The algorithm partitions the sequence sample into bi-segment sequence clusters free of significant reassortments. For each cluster, we obtain a maximum-likelihood tree based on linked bi-segment sequences by the method described above, resulting in a subtree with a significantly reassorted sequence at its root. Finally, each cluster is mapped to a parent cluster, using the maximum-likelihood HA ancestor of its reassorted root sequence. The output is a bi-segment tree with point mutations and significant reassortments mapped on its branches.

Viral frequency trajectories

A partitioning of the strain tree into clades is defined by a set of internal nodes, each representing the last common ancestor of a given clade. Each strain is mapped to exactly one clade, given by the closest clade-defining node in its ancestral lineage. The evolutionary success of a given clade, α, is described by its time-dependent population frequency trajectory, $x^\alpha(t)$. We compute these frequency trajectories as follows. For each isolate i, we define a frequency $x_i(t) = w(t - t_i)/\sum_j w(t - t_j)$, where t_i is the reported collection date and $w(\tau) = \exp\left(-\tau^4/4\sigma^2\right)$ with $\sigma = 45$d. The weight function $w(\tau)$ is a smooth sliding window and a heuristic measure of the temporal distribution of cases related to each reported isolate. The frequency of each clade is then the sum of the frequencies of its strains,

$$x^\alpha(t) = \sum_{i \in \alpha} x_i(t). \tag{1}$$

Clade population frequencies, by definition, sum up to one, $\sum_\alpha x^\alpha(t) = 1$ for all t.

For each clade α, we can define the sublineage $S(\alpha)$ as the union of clade α and all its nested descendant clades. Sublineages are in one-to-one correspondence with subtrees of the strain tree. They define a second set of frequency trajectories,

$$X^\alpha(t) = \sum_{i \in S(\alpha)} x_i(t). \tag{2}$$

In a fast-evolving system like influenza, successful clades will rapidly produce new sequence variation on their genetic background, which defines new, nested descendant clades. Therefore, we rarely observe clade frequencies $x^\alpha(t)$ close to 1. By contrast, a sublineage frequency $X^\alpha(t)$ includes all nested subclades; it reaches 1 when the sublineage has displaced all competing clades. Fig. 2B shows clade frequency trajectories $x^\alpha(t)$ for the set of clades marked on the strain tree as a stacked plot; Fig. 2C shows the corresponding sublineage frequencies $X^\alpha(t)$.

A third set of trajectories tracks the time-dependent population frequency of a given (nucleotide or amino acid) allele a at a given genome position k,

$$x_{k,a}(t) = \sum_{i|k,a} x_i(t), \tag{3}$$

where the sum runs over all strains carrying allele a at position k. In contrast to clade and sublineage frequencies, allele frequencies count the occurrence of point mutations independently of the genetic background. Therefore, an observed increase of $x_{k,a}(t)$ may signal an allele-specific fitness advantage.

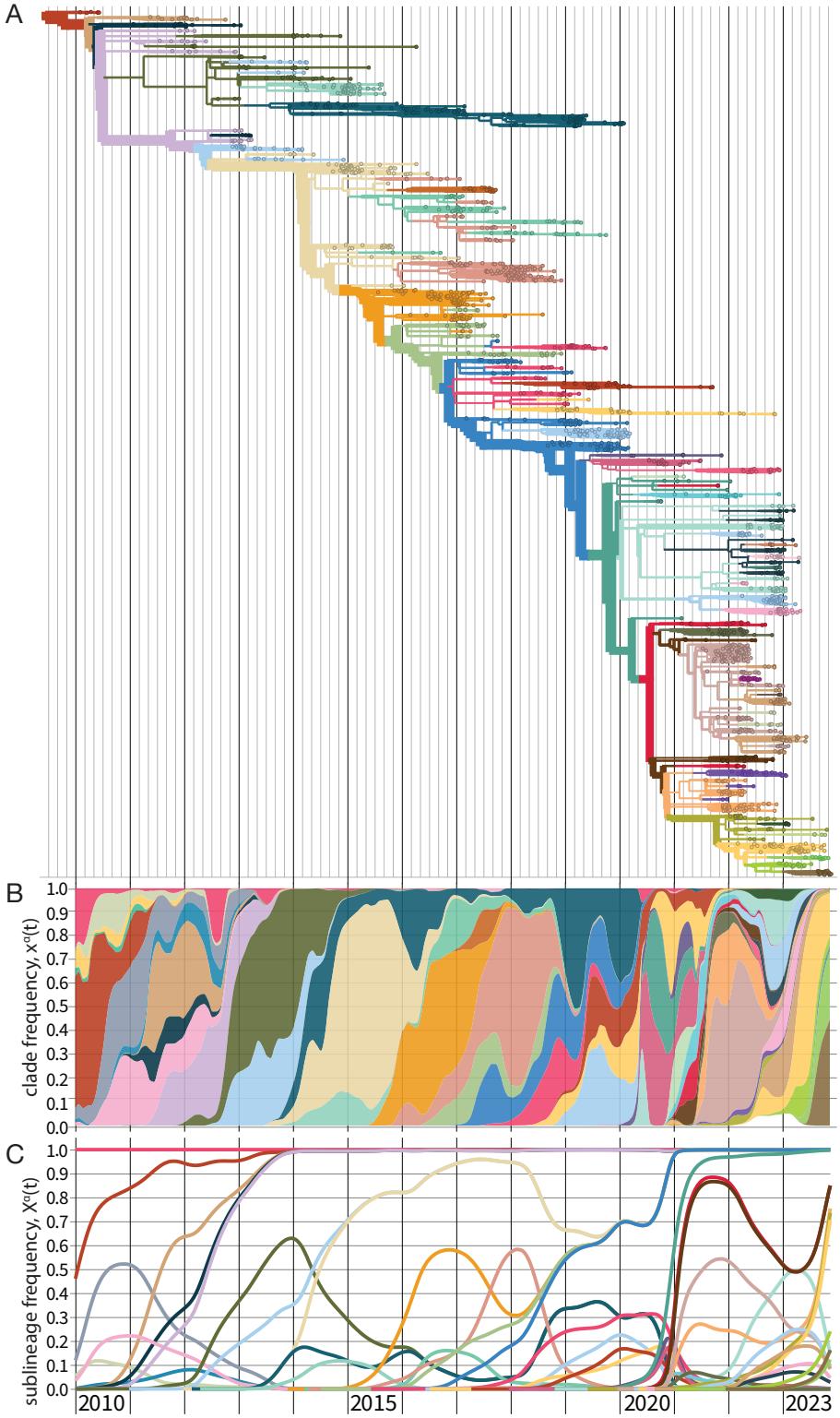

Figure 2. Evolutionary tracking of influenza A(H3N2). (A) Timed maximum-likelihood strain tree based on viral sequences from isolates with collection date from 2010 until end 2023. The tree is colored by clades α that reach a global frequency $X^\alpha(t) > 0.05$ and have potential antigenic differences to their parent clade (signalled by amino acid change(s) in an epitope or antigenic advance mapped from ferret antigenic data). (B) Stacked plot of global clade frequency trajectories, $x^\alpha(t)$, for this set of clades. (C) Global sublineage frequencies, $X^\alpha(t)$.

Empirical fitness trajectories

Given a set of clade frequency trajectories, $x^\alpha(t)$, we can also track the relative rates of change,

$$\hat{f}_\alpha(t) = \frac{\dot{x}^\alpha(t)}{x^\alpha(t)}, \tag{4}$$

where dots denote derivatives with respect to time. As discussed in section 3.4, the resulting trajectories $\hat{f}_\alpha(t)$ measure the relative fitness of each clade, relative to the mean population fitness, as a function of time. Clades with $\hat{f}_\alpha(t) > 0$ show a net increase in frequency, clades with $\hat{f}_\alpha(t) < 0$ a net decrease. A hat distinguishes these empirical fitness values derived from frequency tracking from the fitness models introduced below. The time-derivative is evaluated by numerical discretization around the time t or by Bayesian inference from the timed isolate counts. A related measure of relative clade growth that can be evaluated from tree topology is the Local Branching Index (LBI) introduced in ref. [6].

Regional frequency tracking

Influenza strain populations show significant variation across regions, generating different sets of major regionally competing strains. This variation is of particular importance for emerging clades that originate in a specific region and have to be gauged in that context. Moreover, a regionally distributed population of circulating strains generates region-specific exposure histories that are relevant for the antigenic fitness model described below.

Therefore, we track region-specific clade frequencies $x_r^\alpha(t)$, sublineage frequencies $X_r^\alpha(t)$, and allele frequencies $x_{r,k,a}(t)$; the frequencies in a given region r are obtained by summation over all strains collected in that region. An inference procedure for empirical fitness trajectories $\hat{f}_\alpha(t)$ from region-specific data is discussed in section 3.4. Major surveilled regions are roughly aligned with named WHO transmission zones; region-specific tracking can be refined to individual countries whenever this is relevant for a specific question.

Incidence tracking

By combining regional reported incidence numbers $I_r(t)$ obtained from FluNet [21] with regional clade frequencies $x_r^\alpha(t)$, we infer clade-specific incidence numbers,

$$I_r^\alpha(t) = I_r(t)\, x_r^\alpha(t), \tag{5}$$

which enter the tracking of population immunity discussed in section 3.3.

Regional incidence tracking can also serve to flag, and partially correct, differences in sequence sampling depth between regions. From sequence counts $N_r(t) = \sum_{i\in r} w(t-t_i)$ and incidence data $I_r(t)$ in a set of regions r, we record weight factors $m_r(t) = I_r(t)/N_r(t)$ measuring the incidence per sequence count in each region. Given region-specific clade frequency trajectories, $x_r^\alpha(t)$, we can then compute incidence-weighed global clade frequencies, $x^\alpha(t) = m^{-1}(t)\sum_r m_r(t)\, x_r^\alpha(t)$, where $m(t) = I(t)/N(t)$ is the ratio of global incidence and sequence counts. Incidence-weighed sublineage and allele frequencies are defined in an analogous way. Alternatively, differences in sampling depth can also be inferred from local sequence diversity [44, 45].

Tracking phenotypic evolution

Several independently measurable molecular phenotypes of influenza proteins show evolutionary variation and may contribute fitness effects. Examples are protein stability, conformation of binding interfaces, as well as local solvent accessibility and glycosylation status of antigenic epitopes. Using input from specific sequence data, recent computational methods serve to predict protein structures [46, 47], the impact of mutations on protein stability and binding of protein complexes [48], glycan shielding on the protein surface [49], cross-neutralization [50], variation of innate immune response [51, 52], as well as combinations of phenotypes relevant for growth [53]. Deep mutational scanning [29, 27] and high-throughput in-vitro evolution [54, 55] provide experimental genotype-phenotype maps.

Tracking phenotypic evolution from these data requires mapping significant changes onto branches of the strain tree. Here we use a tree-guided interpolation procedure to obtain clade-specific phenotype values E_α from data E_i measured for a limited set of strains. This procedure serves several purposes: (1) to average between measurements E_i within clades with measured data, (2) to infer maximum-likelihood values E_α for clades without data, (3) to detect measurement outliers, and (4) to identify branches with significant phenotypic changes. A specific tracking algorithm is described in the context of antigenic phenotypes (section 3.3).

3.2 Inference of selection on influenza proteins

Sequence data organized in a genealogical tree and partitioned into clades with time-dependent sublineage frequencies contain a wealth of evolutionary information. In particular, they can serve to infer the speed of molecular evolution and the underlying selective forces, in different influenza genes and gene segments with specific functions. Second, they harbour information on the type of selection acting on viral clades that can guide the predictive fitness models to be developed below.

As an example of the inference method, Table 1 records the number of synonymous and non-synonymous nucleotide changes in the HA protein, mapped on a strain tree constructed from > 100.000 isolates for influenza A(H3N2) in the period 1968-2020. The HA protein contains two domains, HA1 and HA2. The antigenic epitopes, which are the primary loci of interactions with the human immune system are found on the HA1 domain of the protein. These epitopes, designated A to E, contain a combined total of 62 amino acids [56, 57, 58, 59]. Figure 3A shows the location of the five epitopes (A-E) on the monomeric protein structure of HA. Accordingly, we partition the HA point mutations into four sequence classes: antigenic epitopes A, B, D (located on the HA1 head), epitopes C, E (located closer to the HA1 stem), the receptor binding domain (RBD, which overlaps with epitopes A and B), and the remaining non-epitope sequence. Each mutation originates on a given branch of the tree, i.e., on a given genetic background, and defines a sublineage of descendant strains. We record the time-dependent frequency, $X(t)$, and record the number of all mutations that reach a given frequency X, denoted by $m(X)$. A mutation reaching frequency $X = 1$ (in practice, we use a threshold $X = 0.99$) signals fixation of the descendant sublineage; we denote the number of such fixations by d. In most cases, sublineage fixation also implies fixation of the defining allele change in the viral genome. Table 1 shows the numbers $m(0)$, $m(0.05)$, and d for synonymous and non-synonymous changes in each sequence class. Here, the threshold frequency $X = 0.05$ serves to remove biases due to sampling depth (sampling in later years detects mutations at much lower initial frequency). Next, we compute the fraction of all mutations that reach a given frequency, $G(X) = m(X)/m(0)$. Throughout the influenza genome, most mutations remain at small frequency (i.e., close to the tips of the tree), only a small fraction reaches high frequency and eventual sublineage fixation (these mutations appear on the trunk of the tree).

Table 1. Selection inference for the HA protein of influenza A(H3N2). From left to right: (1) $m_{\rm syn}(0)$, number of all synonymous mutations; (2) $m_{\rm syn}(0.05)$, number of synonymous mutations reaching frequency $X > 0.05$; (3) $d_{\rm syn}$, number of synonymous mutations reaching sublineage fixation; (4) $m_{\rm nsyn}(0)$, number of all non-synonymous mutations; (5) $m_{\rm nsyn}(0.05)$, all non-synonymous mutations reaching sublineage frequency 0.05; (6) $d_{\rm nsyn}$, number of non-synonymous mutations reaching sublineage fixation; (7) propagator ratio, g; (8) mutation ratio, q. Data from influenza A(H3N2) strain tree 1968-2020, 120K isolates.

sequence class	$m_{\rm syn}$		$d_{\rm syn}$	$m_{\rm nsyn}$		$d_{\rm nsyn}$	g	q
	0.0	0.05		0.0	0.05			0.05
epitope A, B, D	4048	146	17	6160	483	120	3.06	1.58
epitope C, E	1669	68	8	3195	184	36	1.77	1.29
receptor binding domain	1998	103	15	1766	140	27	2.40	0.65
non-epitope	23499	1066	161	13982	542	40	0.45	0.24

To quantify selection, we record the ratio

$$g(X) = \frac{G_{\rm nsyn}(X)}{G_{\rm syn}(X)} \tag{6}$$

as a function of the frequency X for the different sequence classes. This measure, referred to as propagator ratio [60], gives the probability of non-synonymous mutations to reach sublineage frequency X, weighed by the corresponding probability for synonymous mutations. More generally, the ratio $g(X_2)/g(X_1)$ gives the conditional probability to reach frequency X_2, given an initial frequency X_1 already reached, again weighed by the corresponding probability for synonymous mutations. Hence, we obtain a frequency-sensitive measure of selection: the direction and amplitude of changes in $g(X)$ between frequencies X_1 and X_2 measure the amount of directional selection in the frequency interval $[X_1, X_2]$. The strongest signal of selection is contained in the propagator ratio evaluated over the full frequency interval $g = (d_{\rm nsyn}/m_{\rm nsyn}(0))/(d_{\rm syn}/m_{\rm syn}(0))$ (Table 1). We obtain a clear grading of selection in different HA gene segments of influenza A(H3N2): positive selection ($g > 1$) in antigenic epitopes (strongest in the head epitopes A, B, D, including the RBD sites overlapping with epitopes A and B), and negative selection ($g < 1$) in the remainder of the HA protein [60]. Thus, selection analysis is able to identify the sequence loci subject to positive selection for immune escape evolution. Numerous other studies have identified positive selection in the antigenic epitopes of influenza [61, 62, 63, 64, 65, 66, 67, 68] and in the RBD and nucleocapsid protein of SARS-CoV-2 [69]. Similarly, purifying selection on viral protein evolution was mapped in a number of studies for influenza [68, 70, 71] and SARS-CoV-2 [71, 72].

To obtain a more detailed picture of selection, we plot the propagator ratio $g(X)$ as a function of the sublineage frequency X. We observe that most of the selection signal is in the frequency range $X \lesssim 0.3$. For higher frequencies, the pattern of $g(X)$ flattens, indicating that initially successful escape variants can be driven to loss by the subsequent dynamics [60]. Two factors contribute to this effect, the relative strength of which delineates modes of evolution under directional vs. diversifying selection [73]. First, clonal interference – the selective competition between co-existing clades – introduces strong collective effects at high sublineage frequencies that override the selection coefficients of individual mutations and introduce so-called hitchhiking effects: neutral and even moderately deleterious mutations can reach high frequencies in a successful sublineage, while mutations under positively selected can be outcompeted by another clade [60]. Second, antigenic selection introduces time-dependent, non-linear effects [1, 5]: immune waning and previous infections deplete the pool of susceptible human hosts available to each of the competing clades. In particular, the self-coupling of an initially successful clade

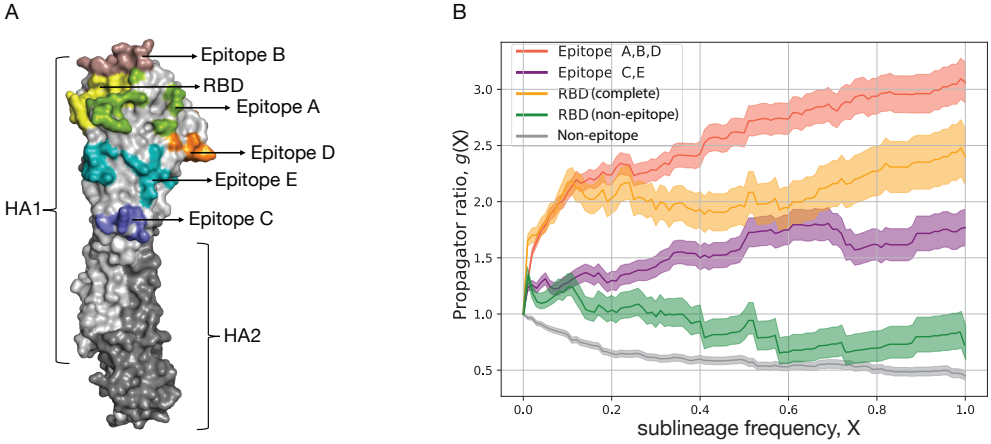

Figure 3. Selection on protein segments of influenza HA. (A) Protein structure of HA with marked regions: subdomains HA1 (light grey) and HA2 (dark grey), antigenic epitopes A – E, receptor binding domain (RBD). (B) Frequency propagator ratio, $g(X)$, as a function of the sublineage frequency threshold X, for different HA sequence classes of influenza A(H3N2). Predominantly positive selection ($g > 1$) is inferred for the epitopes including overlapping RBD sites, overall negative selection ($g < 1$) for the non-epitope HA sequence. Shading indicates standard error margins.

(that is, the consumption of susceptibles by previous infections from that clade and closely related clades) generates a systematic fitness decline at later times [5, 74].

The propagator ratio g is conceptually and computationally related to the McDonald - Kreitman test of selection [75]. We note that probability ratios of the form (6) are insensitive to uncertainties in entry frequency and timing of polymorphism histories, as well as to frequency-dependent bias in polymorphism numbers, as long as this bias does not depend on mutation class.

Alternatively, we can directly compare the numbers of synonymous and non-synonymous changes above a given sublineage frequency threshold. We compute the ratio

$$q(X) = \frac{m_{\text{nsyn}}(X)}{q_0\, m_{\text{syn}}(X)}, \tag{7}$$

where q_0 is the expected value under neutral genome evolution. We obtain $q_0 \approx 2.1$ by numerical evolution of influenza HA, using the reference sequence A/Hong Kong/1-5-MA21-1/1968 (see 2.1) as starting point, transition and transversion rates estimated by IQTree2, and discarding mutations that produce a stop codon. Other computational implementations of neutral evolution are discussed in refs. [76, 77]. Evaluating $q(0.05)$ in different sequence classes yields again a signal of positive selection in the antigenic epitopes, albeit weaker than by the propagator ratio method, and of negative selection in the remainder of the HA sequence (Table 1). The q measure is related to the classical dN/dS method [78, 79, 80]. Previous theoretical work has highlighted potential biases and inaccuracies in dN/dS estimates [81, 82, 83, 84, 85]. In a genealogical tree, the q ratio measures originations, rather than substitutions as in a strain tree across different species. With a given threshold frequency X, it captures only the part of selection acting below X.

The inference of selection can be refined to individual amino acid changes, integrating computational predictions of phenotypic effects [48, 50, 53] and experimental genotype-phenotype maps [29, 27, 54, 55]; see the discussion in section 3.1. A common challenge for experimental and computational analysis is to capture not only the effects of single mutations, but of combinations of multiple mutations that arise in circulating strains.

3.3 Tracking population immunity

The antigenic interaction between pathogen and host is a complex phenotype. The neutralization of a virus during an infection depends on the viral strain and on the individual immune system. The amount of neutralization shows considerable variation across the viral and the immune population. Moreover, both populations are highly dynamic: different viral clades and different classes of immune systems become prevalent over time. Here we develop a method to track these co-evolutionary dynamics, using a combination of molecular antigenic data, viral frequency data, and epidemiogical records.

The tracking of antigenic evolution at the level of molecular interactions, here cross-neutralization titers, is jointly organized along a strain tree of primary infections, generating immune classes [5], and another strain tree of secondary infections. Next, we define two antigenic interaction measures at the population level: population immunity trajectories, again organized by immune class and viral clade, and population immunity profiles that summarize the net immune pressure on a given viral clade. We first develop the tracking method for antigenic data from ferrets; the application to human antigenic data is discussed at the end of the section. A glossary of all antigenicity measures used in this Chapter is given in Note 2.

Tracking antigenic evolution

Antigenic titers T_i^j from ferrets are obtained from immune-naive animals infected with a reference strain j, whose antisera are subsequently assayed against a panel of test strains i representing potential secondary infections (see section 2.3). Titers from binding assays measure the reduced free energy of binding between virus and polyclonal antiserum; titers from neutralization assays remain strongly correlated with binding [86]. Reference strains and test strains are judiciously chosen from the evolving viral population but cover only a small fraction of the circulating strains. Here we describe a tree-guided interpolation method that produces a complete antigenic interaction matrix at the level of clades, T_α^κ (Fig. 4). Here κ is the clade of the primary infecting strain j, in the following also called immune class, and α denotes the clade partitioning of test strains i. The completed cross-neutralization matrix also quantifies local antigenic evolution: the antigenic advance from clade α to clade β,

$$\Delta T_{\alpha\beta}^\kappa = T_\alpha^\kappa - T_\beta^\kappa, \tag{8}$$

measures the immune escape effect of the mutation(s) separating these clades on the strain tree, as seen from immune class κ. Thus, antigenic advance is the change in the free energy of functional antibody-antigen binding induced by viral sequence evolution.

In a first step, we infer matrix elements T_α^κ for clade pairs (α, κ) containing measurements T_i^j. We use a quadratic optimization procedure with a cost function

$$S_1 = \sum_{\alpha,\kappa} \sum_{(i,j)\in(\alpha,\kappa)} \left(T_i^j - \Delta_i - \Delta^j - T_\alpha^\kappa\right)^2 + \frac{1}{\sigma_{\text{av}}^2} \sum_i (\Delta_i)^2 + \frac{1}{\sigma_{\text{pot}}^2} \sum_j (\Delta^j)^2. \tag{9}$$

Here T_i^j are standardized titer data, as described in section 2.3. The correction terms Δ_i and Δ_j (with the constraint $\sum_i \Delta_i = \sum_j \Delta^j = 0$) account for strain-specific confounding factors caused by viral avidity and serum potency effects, respectively; the coefficients σ_{pot} and σ_{av} tune the amplitude of these corrections [23]. As shown in Fig. 4AB, this step achieves a partial data completion, setting titers T_α^κ close to the average of the measurements covering the corresponding clade pair (α, κ). In a second step, we infer the remaining T_α^κ, which belong to clade pairs (α, κ) without measurements. We

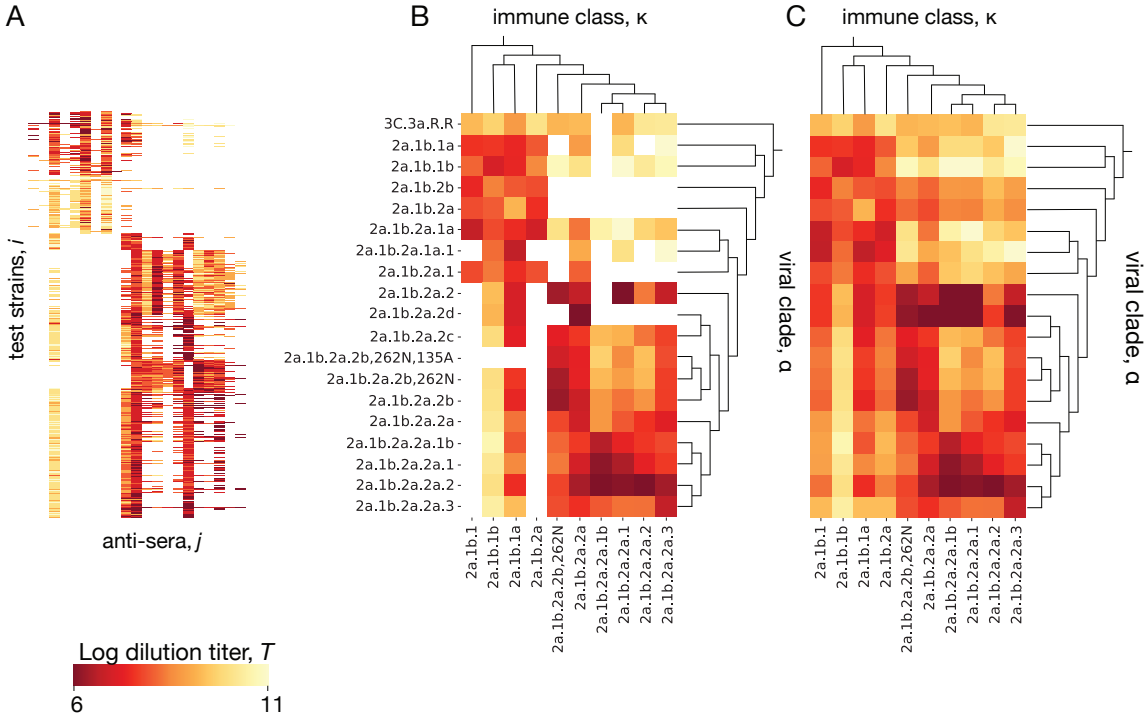

Figure 4. Inference of cross-neutralization. (A) Set of raw titers T_i^j between viral test strains i and reference strains j from antigenic assays. (B) Coarse-grained matrix elements T_α^κ for clade pairs (α, κ) containing measured strain pairs (i, j). (C) Full inferred cross-neutralization matrix T_α^κ covering major clades in the period 2020–2024; missing values are inferred using a tree-guided interpolation scheme. Ferret antigenic data for influenza A(H3N2) are from ref. [24], clade names follow WHO nomenclature.

use an interpolation procedure that minimizes the cost function

$$S_2 = \lambda_0 \sum_{\alpha,\kappa} \left(\Delta T^\kappa_{\mathcal{A}(\alpha)\alpha} \right)^2 + \lambda_1 \sum_{\alpha,\kappa} \left(\Delta T^{\mathcal{A}(\kappa)}_{\mathcal{A}(\alpha)\alpha} - \Delta T^\kappa_{\mathcal{A}(\alpha)\alpha} \right)^2 , \tag{10}$$

where $\mathcal{A}(\alpha)$ is the parent of clade α and $\mathcal{A}(\kappa)$ is the parent of clade κ. This step produces a complete interaction matrix T_α^κ (Fig. 4C). We test the accuracy of the inference scheme using a 90/10 training/test partitioning of the antigenic data and setting $\Delta_i = 0$ for strains that are not in the training set. We find a mean square error $\langle (T_i^j - \Delta_i - \Delta^j - T_\alpha^\kappa)^2 \rangle = 0.55 \pm 0.03$ (95% confidence interval) in the test set. This should be compared with the expected measurement error of ± 1 titer in an HI assay. We conclude that the curation and completion scheme reduces experimental noise; that is, the inferred titer values T_α^κ are likely to be more informative than individual antigenic measurements T_i^j.

The inference procedure presented here allows for antigenic advance between viral clades to depend on the immune class, $\Delta T^\kappa_{\alpha\beta} = T^\kappa_\alpha - T^\kappa_\beta$. It generalizes a previous method that infers antigenic advance with the constraint $\Delta T^\kappa_{\alpha\beta} = \Delta T_{\alpha\beta}$, where $\Delta T_{\alpha\beta}$ is a sum of uni-valued increments on the tree branches [2]. The increased method complexity is warranted by the data: from the ferret antigenic data for influenza, we find that some clade pairs (α, β) show a significant antigenic advance $\Delta T^\kappa_{\alpha\beta}$ in some immune classes κ but not in others; a similar effect for SARS-CoV-2 is discussed in ref. [5].

Inference of cross-immunity

How can the tracking of molecular antigenic evolution be used to build dynamical measures of antigenic interactions at the level of individuals and population segments? Here we relate clade-specific binding or neutralization titers, T_α^κ to the corresponding cross-immunity factors, c_α^κ, with measure the relative reduction of susceptibility to infection by viral strains from clade α, given a primary infection by a strain from clade κ [87, 88, 89]. We use a nonlinear map,

$$c_\alpha^\kappa = H(T_\alpha^\kappa) = \frac{1}{1 + \exp[(T_{50} - T_\alpha^\kappa)/\gamma]}, \tag{11}$$

with system-specific parameters T_{50}, γ (for standardized titers from influenza ferret data, $T_{50} = 8$ and $\gamma = 1.4$). This form follows the thermodynamic relation between binding probability and free energy, a rationale familiar from biophysical fitness models [90, 91, 92, 93, 94, 95, 96, 97]. However, equation (11) says more than molecular binding kinetics: it relates a molecular variable, T, to an organismic effect, c (the likelihood to get the flu). This relation is a heuristic; its validity has to be corroborated by data from both scales. For viral infections, a sigmoidal relationship between HI titers and measured protection from an infective challenge has first been reported in ref. [98]; later studies confirmed this form in several viral systems [99, 100, 101, 102, 5]. In using ferret antigenic data as an input for cross-immunity factors of human individuals, we assume that the ferret titers from immune class κ are a reasonable proxy for the neutralization capacity of humans who had their last infection by a strain from clade κ (Note 3).

Population immunity trajectories and profiles

To scale cross-immunity to the population level, we weigh the coefficients c_α^κ with epidemiological data of past infections. This defines a family of cross-immunity trajectories [5],

$$C_\alpha^\kappa(t) = c_\alpha^\kappa \, r_\kappa(t), \tag{12}$$

which describe the population immunity against viral clade α generated by the population segment in immune class κ. The time-dependent factors $r_\kappa(t)$ determine the weight of immune class κ in the human population at time t. We obtain these weights by summation over past infections, $r_\kappa(t) = \sum_r \int^t I_r^\kappa(t') \, K_\kappa(t - t') \, dt'$, using clade-resolved incidence numbers given by equation (5). The kernel K accounts for immunodominance-changing infections or vaccinations between t' and t; similar kernels describe waning of immune protection for SARS-CoV-2 [5].

By summation over immune classes, we obtain the population immunity profile,

$$\bar{C}_\alpha(t) = \sum_\kappa c_\alpha^\kappa \, r_\kappa(t), \tag{13}$$

which measures the total population immunity against clade α at time t. This profile describes the immune pressure on viral evolution and will be a key input to viral fitness models (section 3.4).

Tracking of human antigenic data

Antigenic data based on human antisera give a more direct representation of the human immune landscape, which is marked by complex histories of previous infections and vaccinations. Antigenic

assays performed on a sample of antisera from individuals ($n = 1, \ldots, N$) and a set of test viruses produce a set of titers T_α^n. We then compute a set of cross-immunity profiles specific to each individual,

$$c_\alpha^n = H(T_\alpha^n), \tag{14}$$

by applying a nonlinear map analogous to equation (11) with parameters T_{50}, γ appropriate for human data [98, 99, 102, 5]. Given a sufficiently large and representative population sample collected around a given time t, the sample-averaged immunity profile,

$$\bar{C}_\alpha(t) = \frac{1}{N} \sum_{n=1}^{N} c_\alpha^n, \tag{15}$$

may provide a good approximation of the underlying population immunity profile. In the context of predictive analysis, this type of inference has been applied to human antigenic data for SARS-CoV-2 [5]. In this case, individual samples may be assigned a likely recent infection or vaccination dominating cross-immunity; the sample average (15) can then be evaluated as an average over immune classes, equation (13), with weight functions informed by incidence tracking. More generally, the sample average (15) may still give a *bona fide* input for fitness models; however, detecting and correcting biases in the population sample requires at least partial knowledge of the underlying infection histories by available metadata or computational reconstruction. This practical problem indicates a fundamental challenge of human immune data: given the complexity of human immunization histories [103, 104, 105, 106], only the most common patterns (e.g., the last vaccination or infection) will be represented correctly in realistic population samples. This limits the complexity of immune imprinting and antigenic sin that can be included in data-driven predictive analysis. In the formalism developed here, immune classes define the appropriate coarse-graining of infection and vaccination histories [5]. In practice, we partition human antigenic data into immune classes labelled by one or few previous immunization events, each associated with a specific viral clade, depending on the availability of metadata (Note 3).

Recently, deep mutational scanning (DMS) experiments have been developed for the analysis of antigenic data from human population samples [25, 26]. These high-throughput experiments generate antigenic escape mutations away from a backbone genome under immune pressure from an individual's antiserum and measure the fraction of unbound viral particles. From this readout, free energy effects of individual mutations can be extracted [107], which correlate well with measured titer differences between the backbone strain and the single-mutant strain [25]. Assuming additivity of local free energy changes, DMS data can be extrapolated to map the antigenic advance of strains with multiple mutations away from the backbone strain [107]. Coarse-graining to the level of clades, we can track the advance $\Delta T_{\alpha\beta}^n$ from the backbone clade α to mutant clades β, as seen in a given individual n, similar to the antigenic advance (8) defined in the context of ferret data. With an additional measurement of the backbone titer T_α^n, the set of mutant titers $T_\beta^n = T_\alpha^n - \Delta T_{\alpha\beta}^n$ determines cross-immunities c_β^n and, by equation (15), a sample-averaged immunity profile $\bar{C}_\beta$. See also Note 4.

3.4 Evolutionary prediction and validation

Predictive analysis harvests information from a system's past evolution to determine its likely near-future trajectory. At the core of predictions is a fitness model: a computable relation between independently measurable genetic and molecular traits as input and fitness, quantifying the expected relative growth of competing viral clades, as output. In this section, we first define absolute and relative fitness. Next, we describe a method to build consistent fitness models from distinct components, representing different sources of input data. Fitness models, together with tracking of the past dynamics, determine likely

future frequency trajectories, as well as antigenic properties of emerging high-fitness variants. Finally, we discuss probabilistic measures to validate predictive methods and to quantify the time horizon of predictions.

Absolute and relative fitness

The absolute fitness of a viral clade is defined as the growth rate of the number of infections, $F_\alpha(t) = \dot{I}_\alpha(t)/I_\alpha(t)$. By defining fitness at the level of clades, we average over systematic growth differences between strains within the same clade. Absolute fitness depends on the effective reproductive number, $R_\alpha(t)$ (the average number of new infections generated by an infected individual), and on the distribution of generational intervals (the time between infection and transmission) [108, 109]. In the relevant parameter regime, this dependence is well approximated by $F_\alpha(t) = \log R_\alpha(t)/\tau$, where τ is the mean generational time interval [5].

The relative fitness of a clade is defined with respect to the mean fitness of the viral population,

$$f_\alpha(t) = F_\alpha(t) - \sum_\beta x^\beta(t) F_\beta(t). \tag{16}$$

Given that absolute fitness is proportional to the log of the reproductive numbers, relative fitness is invariant under a uniform rescaling $R_i \to a(t)R_i$, as expected for the temporal variation of reproductive numbers in the seasonal epidemics of influenza. Relative fitness governs the change of clade frequencies by selection,

$$\frac{\dot{x}^\alpha(t)}{x^\alpha(t)} = f_\alpha(t) + \ldots; \tag{17}$$

the omitted terms are stochastic changes by genetic drift. For sizeable clades of influenza, genetic drift can be neglected, and we obtain a deterministic relation between frequency trajectories and relative fitness that is the central equation of motion used for predictions. Given time-dependent frequency trajectories, we can also compute the change of the mean absolute fitness by evolution, $\Phi(t) = \sum_\alpha \dot{x}^\alpha(t) F_\alpha(t) = \sum_\alpha \dot{x}^\alpha(t) f_\alpha(t)$, the so-called fitness flux [110]. In the deterministic regime, the fitness flux equals the fitness variance in the population of circulating strains, $\Phi(t) = \mathrm{Var}\, F(t)$, which is Fisher's fundamental theorem.

Fitness models

In a fitness model, we write the absolute fitness of a clade as a sum of components that represent different molecular traits relevant for growth and can be quantified by independent input data. Here we use a simple two-component form

$$F_\alpha(t) = F_\alpha^{\mathrm{ag}}(t) + F_\alpha^{\mathrm{in}}. \tag{18}$$

The antigenic fitness component, $F_\alpha^{\mathrm{ag}}(t)$, is generated by the cross-immunity built up in the population by previous infections and vaccinations and is, therefore, explicitly time- and history-dependent. The intrinsic fitness component, F_α^{in}, collects the effects of other molecular traits that are relevant at specific points of the infection cycle. For example, changes in protein stability or host receptor binding affect the intra-host replication of the virus. Such changes are, in general, time-independent. In some of the literature, only the intrinsic fitness component is referred to as fitness, which is somewhat misleading. The total fitness, not any single of its components, enters Fisher's theorem and governs the dynamics of clade frequencies. Specifically, as detailed below, the relative fitness (16) computed from the fitness model (18) can be compared to the empirical relative fitness (4) inferred from tracked frequency trajectories.

Antigenic fitness

At the core of any predictive model is antigenic fitness, which has been a long-term driver of evolution in all human influenza lineages (and has recently become dominant in SARS-CoV-2 as well [5]). Here we use the explicit inference of population immunity described above to evaluate this fitness component: the clade-specific antigenic fitness cost of the viral population is proportional to the immunity profile $\bar{C}_\alpha(t)$ of the host population,

$$F_\alpha^{\rm ag}(t) = -\gamma_{\rm ag} \sum_\kappa c_\alpha^\kappa \, r_\kappa(t), \tag{19}$$

where $c_\alpha^\kappa = H(T_\alpha^\kappa)$ is given by equation (11). This form of the antigenic fitness contains the breakdown of $\bar{C}_\alpha(t)$ by immune classes, as given by equation (13). It reflects the clade-specific reduction of susceptible hosts and can be derived [1, 5] from an underlying multi-strain epidemiological model [88]. The antigenic fitness model is grounded upon the biophysics of molecular host-pathogen interactions, and it inherits the "thermodynamic" nonlinearity in the relation between titers and cross-immunity (section 3.3). The proportionality factor $\gamma_{\rm ag}$ sets the speed of evolution and is a free model parameter.

Given antigenic data from a human population sample, T_α^n $(n = 1, \ldots, N)$, we can also evaluate the biophysical model of antigenic fitness from individual-based cross-immunity profiles sampled around a given time t,

$$F_\alpha^{\rm ag}(t) = -\frac{\gamma_{\rm ag}}{N} \sum_{n=1}^{N} c_\alpha^n, \tag{20}$$

using the sigmoid map (14) and the approximation of the population immunity profile $\bar{C}_\alpha(t)$ as a sample average, equation (15). This approach has first been used to infer viral fitness for SARS-CoV-2, again breaking down the data into likely immune classes (see Note 3) [5]. For influenza, the correlation of $\bar{C}_\alpha(t)$ with subsequent clade growth has been shown in ref. [111].

Alternatively, antigenic fitness can be estimated from the HA protein sequence. Specifically, we can write the antigenic titer drop between a reference clade κ and a viral clade α as a sum of mutational effects in antigenic epitopes, $\Delta T_{\alpha\kappa} = \sum_{m \in M_{\alpha,\kappa}^{\rm ep}} \varepsilon_m$, where $M_{\alpha,\kappa}^{\rm ep}$ is the set of epitope mutations separating clades α and κ; this additive form has been corroborated in ref. [2]. Assuming uniform mutational effects, antigenic fitness retains the biophysical form (19) with a sequence-based approximation of the cross-immunity matrix, $c_\alpha^\kappa = H(D_{\alpha,\kappa}^{\rm ep})$, where $D_{\alpha,\kappa}^{\rm ep}$ is the amino acid distance in HA epitope sequence between clades α and κ. In this approximation, equation (19) reduces to the antigenic fitness model introduced in ref. [1].

Genome-based models with position-specific effects can be informed by DMS data. In recent work, this method has been used to map immune escape from human antisera [25, 26] and from monoclonal antibodies isolated from convalescent individuals [27]. Inferring neutralization titers T_α^n from the escape scores of individual mutations [107], DMS data can be fed into the biophysical model (20) of antigenic fitness [5, 112]. In other recent work, DMS escape scores have been directly related to empirical fitness differences between circulating clades [26]. This implies a linear sequence-fitness map, $F_\beta^{\rm ag} \sim \sum_k \varepsilon(a_{k,\alpha}, a_{k,\beta})$, where α is the clade of the backbone strain and β labels clades containing escape mutations. A common problem of additive sequence-based approximations is that epistatic effects between specific escape mutations present in a circulating strain are not systematically taken into account. See also Note 4.

Intrinsic fitness

Intrinsic fitness effects are a second component of fitness models. In sequence-based estimates, we can write the mutational load accumulated by HA protein changes again as a sum of mutational effects,

$$F_\alpha^{\text{in}}(t) = \sum_k s_k^{\text{in}}(a_{\alpha,k}). \tag{21}$$

In the simplest approximation, we count all recent amino acid changes outside antigenic epitopes with uniform deleterious effects, $F_\alpha^{\text{in}} = \gamma_{\text{in}} D_{\alpha,A(\alpha)}^{\text{ne}}$, where $A(\alpha)$ is a recent ancestor sequence of clade α [1]. Again, this estimate of mutational load can be improved by position- and amino-acid-specific effects measured in a DMS experiment [113]. In a recent deep learning approach, antigenic and intrinsic fitness effects have been jointly inferred solely from HA protein sequence data [53]. See also Note 5.

Model calibration

In a second step, we calibrate the model-based fitness by comparison with empirical fitness data. By equation (4), we obtain clade- and region-specific trajectories of empirical fitness, $\hat{f}_{\alpha,r}(t)$, in a representative set of regions r where clades are in direct competition (see section 3.1). We define a new, calibrated fitness model F^c by

$$F_{\alpha,r}^c(t) = F_{\alpha,r}(t) + \Delta F_\alpha. \tag{22}$$

The calibration combines clade- and region-specific trajectories of the primary model, here the two-component model $F_{\alpha,r}(t) = F_{\alpha,r}^{\text{ag}}(t) + F_\alpha^{\text{in}}$ computed with region-specific immune weights, and a clade-specific global correction term, ΔF_α. Relative fitness computed from the calibrated model, $f_{\alpha,r}^c(t)$, is then used as input for predictions.

The calibration algorithm operates by minimizing an error score over a suitable period up to the prediction baseline,

$$S = \sum_{\alpha,r} \int^{t_0} (\hat{f}_{\alpha,r}(t) - f_{\alpha,r}^c(t))^2 Y_r(t)\,\mathrm{d}t, \tag{23}$$

where $f_{\alpha,r}^c(t) = F_{\alpha,r}^c(t) - \sum_\beta F_{\beta,r}^c(t)\, x_r^\beta(t)$ and $Y_r(t) = N_r(t)/\sum_{r'} N_{r'}(t)$ is the fraction of sequences collected from region r at time t. We use a suitable prior to constrain non-zero corrections to few, significant values.

The calibration of model-based and empirical fitness serves three purposes: (1) The correction term systematically screens the set of trajectories for significant fitness effects that are not contained in the explicit fitness model. The specific calibration scheme given in equation (22) identifies global, time-independent mismatches of model-based and empirical fitness; examples are occasional positive fitness effects caused by intrinsic changes. Another example of potential mismatches are antigenic effects of neuraminidase changes, which are not mapped by HI assays. (2) The calibration yields optimal values of model parameters, here γ_{ag} and γ_{in}. (3) A variant of this method without primary model input and without prior can be used as a standalone inference method of empirical fitness from tracking data. This provides global, model-free fitness estimates $\hat{f}_\alpha(t)$ from regional frequency trajectories. The data partitioning by region is key for correct output, because the inference of relative fitness from global trajectories is confounded by differences in absolute growth between regions [5].

Evolutionary predictions

Clade frequency predictions require frequency tracking, $x^\alpha(t)$, and a fitness model $F_\alpha(t)$ informed by data up to a prediction baseline at time t_0 (we suppress the index c for calibrated models). The corresponding relative fitness serves to predict the most likely future frequency trajectories,

$$x^\alpha(t) = x^\alpha(t_0) \exp\left[\int_{t_0}^{t} f_\alpha(t')dt'\right], \tag{24}$$

over a limited time interval into the future of the prediction baseline. This relation follows by integration of equation (17) in the deterministic limit. Frequency tracking gives the initial condition $x^\alpha(t_0)$, the fitness model determines the future changes of the predicted trajectories.

The precise genetic changes that spawn new variants have a stochastic origin and may difficult to predict. The antigenic characteristics of successful emerging variants are more predictable, because the population immunity profile constrains of antigenic evolution. Specifically, as shown in ref. [5], temporal windows of strong antigenic selection in a given immune class κ is generated when high population immunity coincides with high expected loss of cross-immunity on the steep flank of the landscape $H(T)$ given by equation (11). Therefore, tracking of population immunity trajectories $C_\alpha^\kappa(t)$ can give important information on the direction of the next antigenic escape mutations.

A dedicated web platform, Previr (https://previr.app), reports continuously updated tracking and and fitness predictions obtained by the computational pipeline described here for human influenza A(H3N2), A(H1N1)pdm09, B(Vic), as well as SARS-CoV-2. See Note 6.

Model validation

Given a set of historical predictions, each based on data collected before a specific baseline t_0, we can compare the predicted trajectories with their posterior tracked counterparts. In this way, we can test prediction methods and quantify the so-called *prediction horizon* τ; that is, the characteristic temporal range of reliable predictions [1, 114, 115]. Care is to be taken that predictions are solely computed from data in the period up to t_0; for example, a strain tree computed from input and validation data would violate this requirement by propagating posterior information back into the training period.

We can quantify the quality of prediction by the *predictive information*, a probabilistic measure of how much a given model reduces the uncertainty about future trajectories [1, 114, 115]. Importantly, this measure separates the information obtained by tracking, which constrains the starting point of future evolutionary paths at time t_0, from the contribution of the fitness model, which explains a part of the evolutionary change from t_0 to $t_0 + \tau$. It also serves to define the prediction horizon as the characteristic time scale on which predictions show diminishing information return.

The prediction horizon of current predictions is limited due to the emergence of new variants. Historical predictions based on a two-component fitness model of the form (18) have a prediction time span of order 1 year [1, 114]. Alternatively, the empirical fitness inferred from recent growth data before the prediction baseline, $\hat{f}_\alpha(t_0)$, can be used as a standalone method for short-term predictions [6]. This method is independent of the specific assumptions of a fitness model, but it does not capture the non-linearities and time-dependence of antigenic fitness [74]. Recent method developments on predicting antigenic characteristics of new mutations can potentially increase the prediction horizon.

3.5 Vaccine protection

Evolutionary predictions as described above can be combined with antigenic data to select the vaccine composition for influenza, and in a similar way for SARS-CoV-2. Predictive modelling is essential for this decision, because the vaccine strain needs to pre-empt the most likely future viral population. Here we derive measures of vaccine protection against present and future viral populations that serve to rank candidate vaccines. These measures can be computed from tracked or predicted trajectories of viral evolution and population immunity.

Naive protection profiles

In analogy to the cross-protection by primary infections (section 3.3), we define the (cross-)protection of a vaccine as the relative reduction in the susceptibility to infections after vaccination. Resolving protection at the level of clades, we compute the protection profile of a given vaccine from antigenic data,

$$c_\alpha^{\text{vac}} = H(T_\alpha^{\text{vac}}), \tag{25}$$

using the nonlinear map (11). This profile applies to a population of otherwise naive hosts, because we single out vaccination as the only source of primary immunization. If the vaccine protection wanes over time (as comprehensively recorded in SARS-CoV-2 [116, 117]), the protection profile c_α^{vac} refers to infections shortly after vaccination, after full immune memory has built up but before substantial waning sets in. Given the protection profile for a given vaccine, we can also define the mean protection against circulating strains at a given time,

$$\bar{c}^{\text{vac}}(t) = \sum_\alpha x^\alpha(t) c_\alpha^{\text{vac}}. \tag{26}$$

Importantly, this measure is time-dependent because clade frequencies evolve over time.

Figure 5A shows the protection profile c_α^{vac} computed by equation (25) for two successive vaccine components for influenza A(H3N2). The vaccine component based on the reference strain A/Cambodia/e0826360/2020 has been recommended for the Northern Hemisphere (NH) 21/22 season, the updated vaccine component based on the strain A/Darwin/6/2021 for the Southern Hemisphere (SH) 22 and 23 seasons and the NH 22/23 and 23/24 seasons [118] (these strains are marked by syringes). The protection profile of the A/Cambodia/e0826360/2020-like strain shows the partial escape of later circulating strains from vaccine protection, the updated A/Darwin/6/2021-like vaccine restores this protection. In Fig. 5B, we show another example, the protection profiles of the successive vaccine components A/Hawaii/70/2019 (NH 20/21 season) and A/Wisconsin/588/2019 (SH 21 and 22 seasons, NH 21/22 and 22/23 seasons) for influenza A(H1N1)pdm9. These vaccine strains have complementary protection profiles: the A/Hawaii/70/2019-like strain protects against the strains predominant before 2021, the A/Wisconsin/588/2019-like strain against later strains. In both examples, the corresponding mean protection trajectories, $\bar{c}^{\text{vac}}(t)$, of the successive vaccine components show an opposite time dependence (Fig. 5C,D). All of these protection estimates are computed retrospectively and are based on ferret data, which measure vaccine protection without pre-existing immunity.

Differential protection profiles

We can also consider the protection profile of a given vaccine, defined again as the relative reduction in the susceptibility to α-infections after vaccination, in a population with pre-existing immunity. This

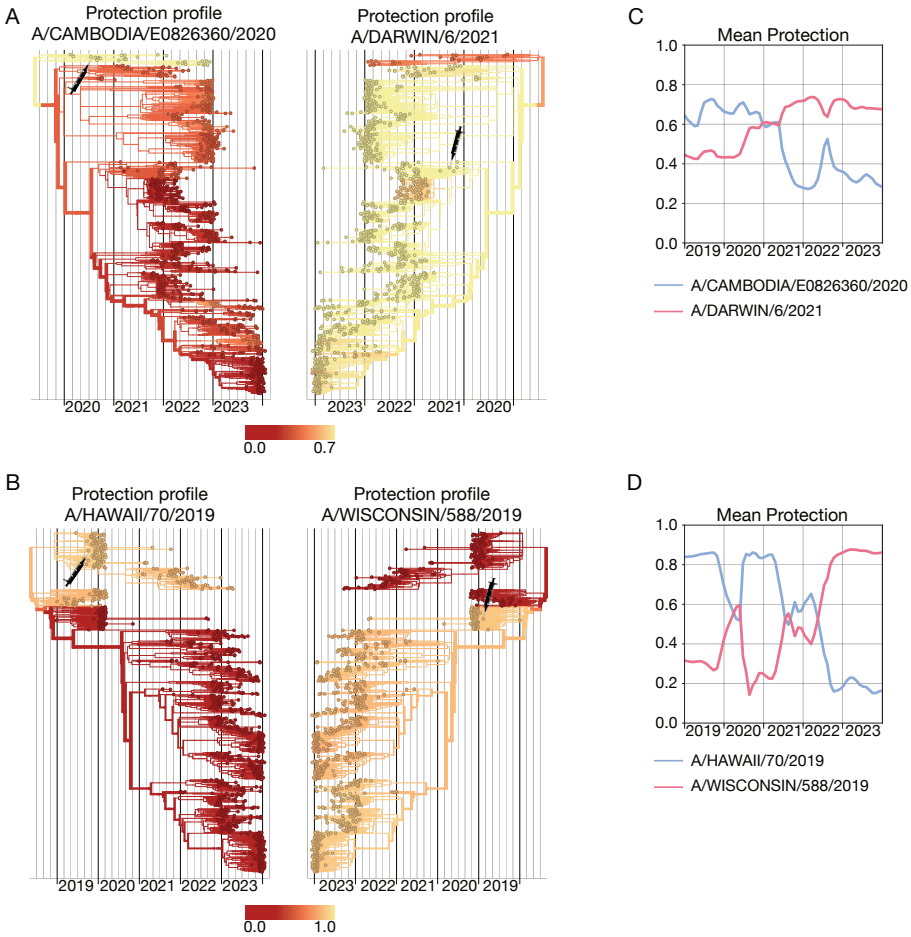

Figure 5. Cross-protection of successive vaccine strains. (A) Protection profiles, c_α^{vac}, of the influenza A(H3N2) A/Cambodia/e0826360/2020-like (left) and A/Darwin/6/2021-like (right) vaccine components, mapped onto the strain tree. Shading from yellow to red indicates decreasing cross-protection; vaccine strains are marked on the tree by syringe symbols. (B) A(H1N1)pdm09 A/Hawaii/70/2019-like (left) and A/Wisconsin/588/2019-like (right) vaccine components. (C,D) Corresponding trajectories of mean vaccine protection, $\bar{c}^{\text{vac}}(t)$. Based on ferret antigenic data for influenza A(H3N2) from ref. [24].

measure, the differential protection profile $\Delta c_\alpha^{\text{vac}}(t)$, is computed as

$$\Delta c_\alpha^{\text{vac}}(t) = \max\left[c_\alpha^{\text{vac}} - \bar{C}_\alpha(t), 0\right], \tag{27}$$

where c_α^{vac} is the naive vaccine protection given by equation (25) and $\bar{C}_\alpha(t)$ the pre-vaccination population immunity profile given by equation (13) or (15). Here we assume that the immuno-dominant repertoire component activated in response to an infecting strain α is induced by the vaccination or by previous immunizations, whichever component generates the stronger cross-immunity. The resulting differential vaccine protection of individuals, $\Delta c_\alpha^{\text{vac},n}(t) = \max[c_\alpha^{\text{vac}} - c_\alpha^n, 0]$, depends on their immunization history; the population average (27) then follows by equation (15). The underlying immuno-dominance assumption is consistent with a recent biophysical model of lineage activation in an acute infection that predicts a correlation between antigen binding and clone size for antigen-responding antibody lineages [119]. The specific form of complex immune responses remains to be tested in future

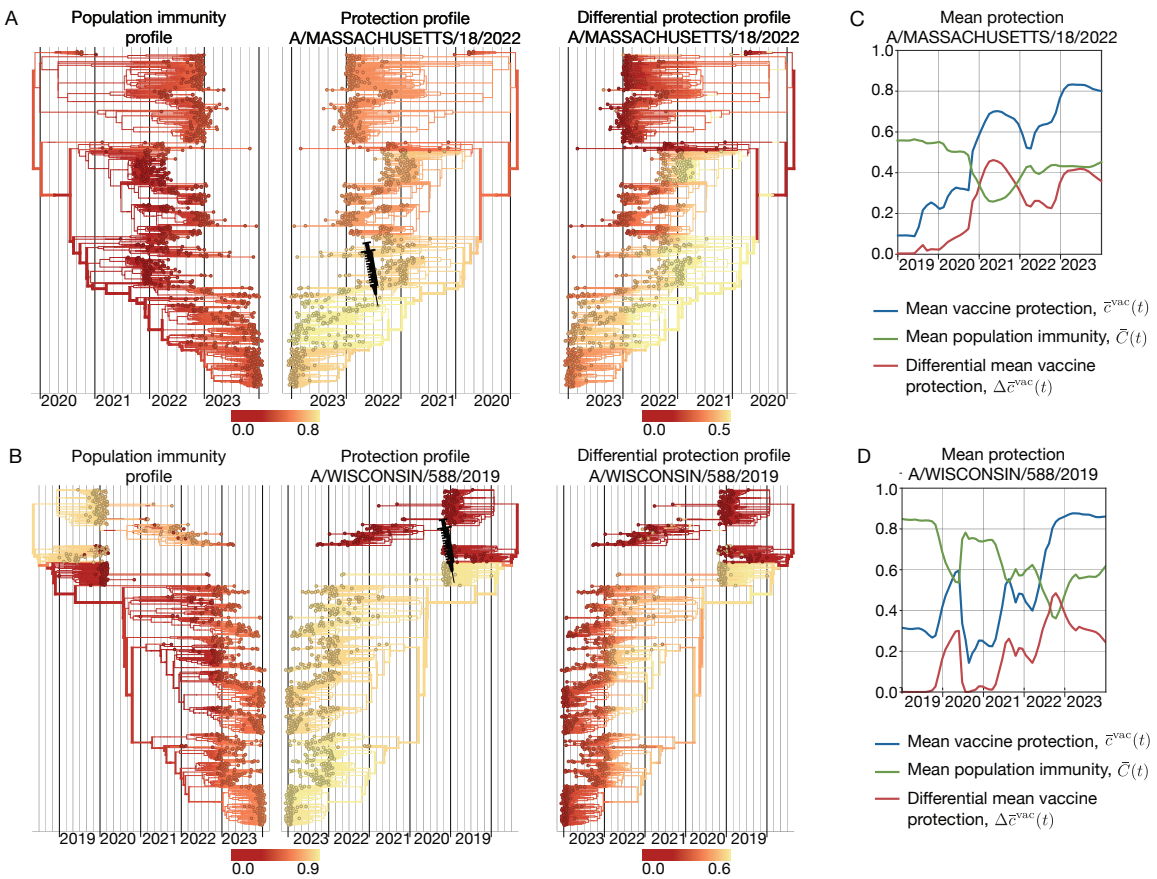

Figure 6. Differential protection of vaccines. (A) Influenza A(H3N2) population immunity profile, $\bar{C}_\alpha(t)$ (left); protection profile $c_\alpha^{\rm vac}$ of the A/Massachusetts/18/2022-like vaccine component (center); resulting differential vaccine protection profile $\Delta c_\alpha^{\rm vac}(t)$ (right). (B) Influenza A(H1N1)pdm09 population immunity profile, $\bar{C}_\alpha(t)$ (left); protection profile $c_\alpha^{\rm vac}$ of the A/Wisconsin/588/2019-like vaccine component (center); resulting differential vaccine protection profile $\Delta c_\alpha^{\rm vac}(t)$ (right). (C,D) Trajectories of mean population immunity, $\bar{C}(t)$ (green), mean vaccine protection, $\bar{c}^{\rm vac}(t)$ (blue) and differential mean vaccine protection, $\Delta\bar{c}^{\rm vac}(t)$ (red). Based on ferret antigenic data for influenza A(H3N2) from ref. [24].

studies. The resulting differential mean protection against circulating strains,

$$\Delta\bar{c}^{\rm vac}(t) = \sum_\alpha x^\alpha(t)\,\Delta c_\alpha^{\rm vac}(t), \tag{28}$$

quantifies the added value of vaccination over population immunity. This measure is time-dependent for two reasons: population immune dynamics affect $\bar{C}_\alpha(t)$ and viral evolution propagates the frequencies $x^\alpha(t)$. Notably, a good vaccine is expected to decrease in differential protection over time, as population immunity builds up against viral clades α covered by the vaccine.

By relating the differential mean protection to the total pre-vaccination population immunity $\bar{C}(t) = \sum_\alpha x^\alpha(t)\,\bar{C}_\alpha(t)$, we obtain a computational estimate of vaccine effectiveness (VE) [120, 121] in the population: the expected case numbers of vaccinated and unvaccinated individuals are proportional to $1-\bar{C}(t)-\Delta\bar{c}^{\rm vac}(t)$ and $1-\bar{C}(t)$, respectively, which gives

$$\mathrm{VE}\,(t) = \frac{\Delta\bar{c}^{\rm vac}(t)}{1-\bar{C}(t)}. \tag{29}$$

Importantly, vaccine effectiveness is seen to be a dynamical quantity: it can increase when the naive protection $\bar{c}^{\text{vac}}(t)$ increases and decrease when population immunity against clades covered by the vaccine builds up. In a similar way, we can compute estimates of vaccine efficacy as measured in clinical trials, by calibrating the mean protection $\bar{C}(t)$ to known characteristics of the trial cohort.

In Fig. 6A, we compare the protection profile c_α^{vac} of the A/Massachusetts/18/2022-like vaccine component (SH 24 season, NH 24/25 season) with the underlying population immunity profile $\bar{C}_\alpha(t)$ of influenza A(H3N2). These profiles result in the differential protection profile $\Delta c_\alpha^{\text{vac}}(t)$, as given by equation (27). In this case, the population immunity profile is broader but weaker than vaccine-induced protection, which generates added protection peaked around the vaccine clade (marked by a syringe). The corresponding analysis of the A/Wisconsin/588/2019-like vaccine component for influenza A(H1N1)pdm09 is shown in Fig. 6B. In the season 2021-22, population immunity and vaccine induced immunity cover complementary viral clades; the vaccine adds protection specifically against clades subject to weak population immunity. We can also track the corresponding trajectories of mean population immunity, $\bar{C}(t) = \sum_\alpha x^\alpha(t)\bar{C}_\alpha(t)$, and vaccine protection, $\bar{c}^{\text{vac}}(t)$ and $\Delta\bar{c}^{\text{vac}}(t)$ (Fig. 6C,D). In both examples, the differential vaccine protection shows a diminishing added value over time, in parallel to the buildup of population immunity against newer viral clades.

Pre-emptive ranking of candidate vaccines

The mean naive and differential protection, $\bar{c}^{\text{vac}}(t)$ and $\Delta\bar{c}^{\text{vac}}(t)$, are directly related to viral fitness. By equations (19) and (26), the mean naive protection $\bar{c}^{\text{vac}}(t)$ is proportional to the vaccine-induced drop of the mean viral population fitness in a naive human population. The mean differential protection $\Delta\bar{c}^{\text{vac}}(t)$, equation (28), measures the analogous drop of viral fitness in a population with pre-existing immunity. Hence, $\Delta\bar{c}^{\text{vac}}(t)$ is an explicitly co-evolutionary measure, which depends on the distribution of circulating viral strains and on the distribution of human immunity. Here we have established a general method to compute estimates of vaccine protection for a rapidly changing population of viral strains and population immunity states. The specific evaluation of these measures is likely to undergo changes, as new types of input data become available.

The vaccine protection measures $\bar{c}^{\text{vac}}(t)$ and $\Delta\bar{c}^{\text{vac}}(t)$ can be used to rank available candidate vaccines for a given influenza season: higher naive and differential mean protection is predicted to induce lower viral growth in vaccinated population segments without and with previous immunity, respectively. Both measures change over time, underscoring the role of predictive analysis for vaccine strain selection. By using predicted clade frequencies, $x^\alpha(t)$, and predicted immune profiles, $\bar{C}_\alpha(t)$, we can optimize vaccines in a pre-emptive way, taking into account the likely viral evolution and population immune dynamics up to the season when the vaccine will be used. The protection profiles shown in Fig. 5 and 6 provide a more detailed picture than the mean protection, highlighting viral clades with maximal escape from vaccine-induced and pre-existing population immunity.

4 Notes

1. Due to intense sequencing efforts during the past decade, over 100,000 influenza and several million SARS-CoV-2 genomes have become available to the scientific community. Keeping track of the genetic variation observed in sequenced viral samples is a bioinformatic bottleneck [122]; the challenges for computational analysis will increase with time. Two recently published methods, MAPLE [123] and USHER [124], overcome the bioinformatic bottleneck using efficient algorithmic solutions within a maximum-parsimony framework. These approaches do not include the inference

of ancestral sequences and the timing of internal nodes, which represents a second computationally expensive process. To address the high computational demand of maximum-likelihood approaches, a common procedure is to subsample available data to a few thousand sequences. This method is adequate for displaying major clades of the evolving viral population, but subsampling may miss clades at small population frequency, including recently emerged, fast-growing clades. Here we implement a maximum-likelihood inference procedure for a subtree partitioning of the global tree, which avoids the need for subsampling.

2. Glossary of antigenicity measures. (1) Cross-neutralization matrix, T_α^κ: average neutralization titers of antisera from immune class κ against viral strains from clade α. (2) Cross-immunity matrix, c_α^κ: expected cross-protection, or reduction of susceptibility, of an individual from immune class κ against viral strains from clade α. (3) Population immunity trajectories, $C_\alpha^\kappa(t)$: population immunity, or reduction of viral growth, for strains of clade α induced by previous infections by strains from clade κ. (4) Population immunity profile, $\bar{C}_\alpha(t)$: total population immunity against strains from clade α. (5) Average population immunity, $\bar{C}(t)$: net population immunity against the population of circulating strains. (6) Vaccine protection profile, c_α^{vac}: cross-protection, or reduction of susceptibility, against viral strains from clade α induced by a given vaccine (vac) in a naive host. Results in a time-dependent mean protection, $\bar{c}^{\text{vac}}(t)$, against circulating strains. (7) Differential vaccine protection profile, $\Delta c_\alpha^{\text{vac}}(t)$: cross-protection, or reduction of susceptibility, against viral strains from clade α induced by a given vaccine (vac) in a population with pre-existing immunity profile $\bar{C}_\alpha(t)$. Results in a time-dependent differential mean protection, $\Delta\bar{c}^{\text{vac}}(t)$, against circulating strains.

3. The established animal model for influenza uses ferrets that are immune-naive against influenza prior to the infection with a reference strain. Neutralization tests of these ferret antisera capture vaccination-induced cross-protection against test strains without confounding effects of pre-existing immunity. The use of ferret data in fitness models for viral evolution in humans depends on a heuristic extrapolation: cross-immunity c_α^κ computed from ferrets infected with a strain from clade κ is a good proxi for humans who had their last immunization (infection or vaccination) by a strain from the same clade. An important role of the last immunization in the response to future exposures has been established [103], but older exposures may also be relevant in specific population segments [125, 25]. Another instance of more complex immunisation histories is combinations of a recent infection and vaccination; such cases can be modelled by composite immune classes, $\kappa = (\kappa_1, \kappa_2)$ [5]. The impact of immune complexity on viral dynamics has recently been studied in ref. [74].

4. Deep mutational scanning (DMS) [29, 27] and high-throughput in-vitro evolution [54, 55] are emerging as promising data sources for predictions of viral evolution. These data map the genomic distribution of high-fitness mutations and can inform antigenic and intrinsic components of genome-based fitness models. Using specific additivity assumptions, antigenic escape scores of individual point mutations can be processed to fitness estimates for strains with multiple mutations, including variants that have not yet been observed as circulating strains or have not yet been phenotypically characterized. However, this construction of fitness landscapes neglects epistasis between specific escape mutations. Other challenges in applying DMS data to predictive analysis of influenza evolution include the choice of the backbone strain from which the mutant library is constructed, as well as the choice of representative antisera driving the escape dynamics.

5. Key intrinsic fitness components, including protein stability and binding to human receptors, depend on protein structure. Computational models hold promise for a high-throughput evaluation of structure-based phenotypic and fitness effects [51, 52, 50, 48, 49]. A recent model combines

computational estimates of such effects with deep learning to predict the genetic evolution of influenza and SARS-CoV-2 [53]. Machine-learning tools to predict protein structures, like AlphaFold [46], can also inform predictions [47], but a successful application of these methods to evolutionary analysis requires sufficient accuracy in distinguishing wild-type and mutated protein structures [126, 127, 128].

6. The effective application of this prediction pipeline depends on comprehensive surveillance of influenza sequence evolution, antigenic evolution, and epidemiology. The collaborative efforts of researchers, national public health institutions, and the global WHO surveillance and response system are key to maintain global surveillance under comparable standards. The timely availability of sequence and antigenic data is key to flag new variants of concern early in their trajectory, or even before their appearance [5]. Surveillance is also necessary to support DMS and laboratory evolution experiments, to allow updating of backbone strains and of reference antisera generating immune pressure, in tune with the evolving viral population.

Acknowledgements

We thank all personnel at the Worldwide Influenza Centre, WHO Collaborating Centre for Reference and Research on Influenza, situated at The Francis Crick Institute, involved in the generation and sharing of antigenic data. Special thanks go to Nicola Lewis, Ruth Harvey, Monica Galiano, and John McCauley for discussions on multiple aspects of this work. We also thank all National Influenza Centers (NICs) and institutions who have contributed information, clinical specimens and viruses, epidemiological, and other associated data to the WHO Global Influenza Surveillance and Response System (GISRS), including FluNet. We are grateful to all institutions and individuals who have contributed clinical specimens, virus sequences, and associated data through the GISAID EpiFlu database.

This work was funded in part by Deutsche Forschungsgemeinschaft grant CRC 1310, Collaborative Research Center *Predictability in Evolution* (to ML), the Centers of Excellence for Influenza Research and Response (CEIRR) contract #75N93021C00014 (to MŁ and ML), the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIAID, NIH) grant R01AI165820 (to MŁ). MŁ was funded in part as a Pew Biomedical Scholar.

References

[1] Łuksza M and Lässig M (2014) A predictive fitness model for influenza. Nature 507(7490):57–61

[2] Neher RA, et al. (2016) Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proceedings of the National Academy of Sciences 113(12):E1701–E1709

[3] Morris DH, et al. (2018) Predictive modeling of influenza shows the promise of applied evolutionary biology. Trends in Microbiology 26(2):102–118

[4] Huddleston J, et al. (2020) Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. eLife 9:e60067

[5] Meijers M, et al. (2023) Population immunity predicts evolutionary trajectories of SARS-CoV-2. Cell 186(23):5151–5164 e13

[6] Neher RA, Russell CA, and Shraiman BI (2014) Predicting evolution from the shape of genealogical trees. eLife 3:e03568

[7] Bao Y, et al. (2008) The influenza virus resource at the National Center for Biotechnology Information. Journal of Virology 82(2):596–601

[8] World Health Organisation (2010) Serological diagnosis of influenza by microneutralization assay https://www.who.int/publications/i/item/serological-diagnosis-of-influenza-by-microneutralization-assay

[9] World Health Organisation (2011) Manual for the laboratory diagnosis and virological surveillance of influenza https://www.who.int/publications/i/item/manual-for-the-laboratory-diagnosis-and-virological-surveillance-of-influenza

[10] Pedersen JC (2014) Hemagglutination-inhibition assay for influenza virus subtype identification and the detection and quantitation of serum antibodies to influenza virus. Methods Molecular Biology 1161:11–25

[11] Jorquera PA, et al. (2019) Insights into the antigenic advancement of influenza A (H3N2) viruses, 2011–2018. Scientific Reports 9(1):2676

[12] Cuevas F, et al. (2022) An in vitro microneutralization assay for influenza virus serology. Current Protocols 2(7):e465

[13] Wang Y, Tang CY, Wan XF (2022) Antigenic characterization of influenza and SARS-CoV-2 viruses. Anal Bioanal Chem 414(9):2841–2881

[14] Benson D, et al. (2015) GenBank. Nucleic Acids Research 41:D36–D42

[15] Elbe S and Buckland-Merrett G (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges 1(1):33–46

[16] Katoh K and Standley D (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30(4):772–780

[17] Caini S, et al. (2016) Temporal patterns of influenza A and B in tropical and temperate countries: What are the lessons for influenza vaccination? PLoS One 11(3):1–15

[18] Dave K and Lee PC (2019) Global geographical and temporal patterns of seasonal influenza and associated climatic factors. Epidemiologic Reviews 41(1):51–68

[19] Zhou L, et al. (2019) Temporal patterns of influenza A subtypes and B lineages across age in a subtropical city, during pre-pandemic, pandemic, and post-pandemic seasons. BMC Infectious Diseases 19(1):1471–2334

[20] Coletti P, et al. (2018) Shifting patterns of seasonal influenza epidemics. Scientific Reports 8(1):12786

[21] World Health Organization, GISRS (2024) FluNet. https://www.who.int/tools/flunet.

[22] Spackman E and Sitaras I (2020) Hemagglutination inhibition assay. Methods Molecular Biology 2123:11–28

[23] Bedford T, et al. (2014) Integrating influenza antigenic dynamics with molecular evolution. eLife 3:e01914

[24] Worldwide Influenza Centre, CRICK Institute. Annual and interim reports. https://www.crick.ac.uk/research/platforms-and-facilities/worldwide-influenza-centre/annual-and-interim-reports

[25] Welsh FC, et al. (2023) Age-dependent heterogeneity in the antigenic effects of mutations to influenza hemagglutinin. bioRxiv DOI:10.1101/2023.12.12.571235

[26] Dadonaite B, et al. (2023) Full-spike deep mutational scanning helps predict the evolutionary success of SARS-CoV-2 clades. bioRxiv DOI:10.1101/2023.11.13.566961

[27] Cao Y, et al. (2023) Imprinted SARS-CoV-2 humoral immunity induces convergent omicron RBD evolution. Nature 614(7948):521–529

[28] Starr TN, et al. (2020) Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell 182(5):1295–1310

[29] Starr TN, et al. (2022) Deep mutational scans for ACE2 binding, RBD expression, and antibody escape in the SARS-CoV-2 Omicron BA. 1 and BA. 2 receptor-binding domains. PLoS Pathogens 18(11):e1010951

[30] Grubaugh ND, et al. (2019) Tracking virus outbreaks in the twenty-first century. Nature Microbiology 4(1):10–19

[31] Felsenstein J (2003) Inferring Phylogenies. Oxford University Press UK. ISBN 9780878931774

[32] Huelsenbeck JP and Crandall KA (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. Annual Review of Ecology and Systematics 28:437–466

[33] Minh BQ, et al. (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution 37(5):1530–1534

[34] Sagulenko P, Puller V, and Neher R (2018) TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evolution 1(4):vex042

[35] Kalyaanamoorthy S, et al. (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. Nature Methods 14:587–589

[36] Holmes EC, et al. (2005) Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. PLoS Biology 3(9):e300

[37] Nelson MI, et al. (2006) Stochastic processes are key determinants of short-term evolution in influenza a virus. PLoS Pathogens 2(12):e125

[38] Nelson MI, et al. (2008) Multiple reassortment events in the evolutionary history of H1N1 influenza a virus since 1918. PLoS Pathogens 4(2):e1000012

[39] Nagarajan N and Kingsford C (2011) GiRaF: robust, computational identification of influenza reassortments via graph mining. Nucleic Acids Res 39(6):e34

[40] Yurovsky A and Moret BM (2011) Fluref, an automated flu virus reassortment finder based on phylogenetic trees. BMC Genomics 12:S3

[41] Svinti V, Cotton JA, and McInerney JO (2013) New approaches for unravelling reassortment pathways. BMC Evol Biol 13:1

[42] Barrat-Charlaix P, Vaughan TG, and Neher RA (2022) TreeKnit: Inferring ancestral reassortment graphs of influenza viruses. PLoS Comput Biol 18(8):e1010394

[43] Rabadan R, Levine AJ, and Krasnitz M (2008) Non-random reassortment in human influenza A viruses. Influenza Other Respir Viruses 2(1):9–22

[44] Khatri B and Burt A (2019) Robust estimation of recent effective population size from number of independent origins in soft sweeps. Molecular Biology and Evolution 36(9):2040–2052

[45] Smith M, Trofimova M, and Weber Aea (2021) Rapid incidence estimation from SARS-CoV-2 genomes reveals decreased case detection in europe during summer 2020. Nat Commun 12:6009

[46] Jumper J, Evans R, and Pritzel Aea (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583—589

[47] Akdel M, et al. (2022) A structural biology community assessment of AlphaFold2 applications. Nature Structural & Molecular Biology 29(11):1056–1067

[48] Pun MN, et al. (2024) Learning the shape of protein microenvironments with a holographic convolutional neural network. Proceedings of the National Academy of Sciences 121(6):e2300838121

[49] Tsai YX, et al. (2024) Rapid simulation of glycoprotein structures by grafting and steric exclusion of glycan conformer libraries. Cell 187(5):P1296–1311.E26

[50] Harvey WT, et al. (2023) A Bayesian approach to incorporate structural data into the mapping of genotype to antigenic phenotype of influenza A(H3N2) viruses. PLoS Comput Biol 19(3):e1010885

[51] Greenbaum BD, et al. (2008) Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLOS Pathogens 4:1–9

[52] Greenbaum BD, et al. (2014) Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses. Proceedings of the National Academy of Sciences 111(13):5054–9

[53] Thadani NN, et al. (2023) Learning from prepandemic data to forecast viral escape. Nature 622(7984):818–825

[54] Moulana A, et al. (2023) Genotype–phenotype landscapes for immune–pathogen coevolution. Trends in Immunology 44(5):384–396

[55] Zahradník J, et al. (2021) SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. Nature Microbiology 6(9):1188–1198

[56] Wiley D, Wilson I, and Skehel J (1981) Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature 289(5796):373–378

[57] Bush RM, et al. (1999) Predicting the evolution of human influenza A. Science 286(5446):1921–1925

[58] Munoz ET and Deem MW (2005) Epitope analysis for influenza vaccine design. Vaccine 23(9):1144–1148

[59] Macken C, et al. (2001) The value of a database in surveillance and vaccine selection. International Congress Series 1219:103–106

[60] Strelkowa N and Lässig M (2012) Clonal interference in the evolution of influenza. Genetics 192(2):671–682

[61] Fitch WM, et al. (1991) Positive darwinian evolution in human influenza A viruses. Proceedings of the National Academy of Sciences 88(10):4270–4274

[62] Bush RM, et al. (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A. Molecular Biology and Evolution 16(11):1457–1465

[63] Shih ACC, et al. (2007) Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. Proceedings of the National Academy of Sciences 104(15):6283–6288

[64] Bhatt S, Holmes EC, and Pybus OG (2011) The genomic rate of molecular adaptation of the human influenza A virus. Molecular Biology and Evolution 28(9):2443–2451

[65] Lin YP, et al. (2012) Evolution of the receptor binding properties of the influenza A(H3N2) hemagglutinin. Proceedings of the National Academy of Sciences 109(52):21474–21479

[66] Meyer AG, Dawson ET, and Wilke CO (2013) Cross-species comparison of site-specific evolutionary-rate variation in influenza haemagglutinin. Philosophical Transactions of the Royal Society B: Biological Sciences 368(1614):20120334

[67] Arunachalam R (2013) Detection of site-specific positive darwinian selection on pandemic influenza A/H1N1 virus genome: integrative approaches. Genetica 141:143–155

[68] Klingen TR, et al. (2019) Structures and functions linked to genome-wide adaptation of human influenza A viruses. Scientific reports 9(1):6267

[69] Rochman ND, et al. (2021) Ongoing global and regional adaptive evolution of SARS-CoV-2. Proceedings of the National Academy of Sciences 118(29):e2104241118

[70] Lin JJ, et al. (2019) Many human RNA viruses show extraordinarily stringent selective constraints on protein evolution. Proceedings of the National Academy of Sciences 116(38):19009–19018

[71] Ghafari M, et al. (2022) Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza. Molecular Biology and Evolution 39(2):msac009

[72] Neher RA (2022) Contributions of adaptation and purifying selection to SARS-CoV-2 evolution. Virus Evolution 8(2):veac113

[73] Frazão N., et al. (2022) Two modes of evolution shape bacterial strain diversity in the mammalian gut for thousands of generations Nature Communications (13):5604

[74] Barrat-Charlaix P and Neher RA (2024) Eco-evolutionary dynamics of adapting pathogens and host immunity. arXiv DOI:10.48550/arXiv.2402.07252

[75] McDonald JH and Kreitman M (1991) Adaptive protein evolution at the Adh locus in drosophila. Nature 351(6328):652–654

[76] Yang Z and Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular Biology and Evolution 17(1):32–43

[77] Goldman N and Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution 11(5):725–736

[78] Miyata T, Yasunaga T, and Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proceedings of the National Academy of Sciences 77(12):7328–7332

[79] Li WH, Wu CI, and Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Molecular Biology and Evolution 2(2):150–174

[80] Nei M and Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3(5):418–426

[81] Rocha EP, et al. (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. Journal of Theoretical Biology 239(2):226–235

[82] Kryazhimskiy S and Plotkin JB (2008) The population genetics of dN/dS. PLoS Genetics 4(12):e1000304

[83] Mugal CF, Wolf JB, and Kaj I (2014) Why time matters: codon evolution and the temporal dynamics of d N/d S. Molecular Biology and Evolution 31(1):212–231

[84] Meyer AG, et al. (2015) Time dependence of evolutionary metrics during the 2009 pandemic influenza virus outbreak. Virus Evolution 1(1):vev006

[85] Rahman S, et al. (2021) Weak selection on synonymous codons substantially inflates dn/ds estimates in bacteria. Proceedings of the National Academy of Sciences 118(20):e2023575118

[86] Veguilla V, et al. (2011) Sensitivity and specificity of serologic assays for detection of human infection with 2009 pandemic H1N1 virus in US populations. Journal of Clinical Microbiology 49(6):2210–2215

[87] Andreasen V, Lin J, and Levin SA (1997) The dynamics of cocirculating influenza strains conferring partial cross-immunity. Journal of Mathematical Biology 35:825–842

[88] Gog JR and Grenfell BT (2002) Dynamics and selection of many-strain pathogens. Proceedings of the National Academy of Sciences 99(26):17209–17214

[89] Gog J and Swinton J (2002) A status-based approach to multiple strain dynamics. Journal of Mathematical Biology 44(2):169–184

[90] Gerland U and Hwa T (2002) On the selection and evolution of regulatory DNA motifs. Journal of Molecular Evolution 55:386–400

[91] Berg J, Willmann S, and Lässig M (2004) Adaptive evolution of transcription factor binding sites. BMC Evolutionary Biology 4(1):1–12

[92] Zeldovich KB, Chen P, and Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. Proceedings of the National Academy of Sciences 104(41):16152–16157

[93] Mustonen V, et al. (2008) Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. Proceedings of the National Academy of Sciences 105(34):12376–12381

[94] Manhart M and Morozov AV (2015) Protein folding and binding can emerge as evolutionary spandrels through structural coupling. Proceedings of the National Academy of Sciences 112(6):1797–1802

[95] Chi PB and Liberles DA (2016) Selection on protein structure, interaction, and sequence. Protein Science 25(7):1168–1178

[96] Rodrigues JV, et al. (2016) Biophysical principles predict fitness landscapes of drug resistance. Proceedings of the National Academy of Sciences 113(11):E1470–E1478

[97] Held T, Klemmer D, and Lässig M (2019) Survival of the simplest in microbial evolution. Nature Communications 10(1):2472

[98] Hobson D, et al. (1972) The role of serum haemagglutination-inhibiting antibody in protection against challenge infection with influenza A2 and B viruses. Epidemiology & Infection 70(4):767–777

[99] Coudeville L, et al. (2010) Relationship between haemagglutination-inhibiting antibody titres and clinical protection against influenza: development and application of a bayesian random-effects model. BMC Medical Research Methodology 10(1):1–11

[100] Rotem A, et al. (2018) Evolution on the biophysical fitness landscape of an rna virus. Molecular Biology and Evolution 35(10):2390–2400

[101] Meijers M, et al. (2021) Predicting in vivo escape dynamics of HIV-1 from a broadly neutralizing antibody. Proceedings of the National Academy of Sciences 118(30):e2104651118

[102] Khoury DS, et al. (2021) Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection. Nature Medicine 27(7):1205–1211

[103] Fonville JM, et al. (2014) Antibody landscapes after influenza virus infection or vaccination. Science 346(6212):996–1000

[104] Lee JM, et al. (2019) Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. eLife 8:e49324

[105] Dugan HL, et al. (2020) Preexisting immunity shapes distinct antibody landscapes after influenza virus infection and vaccination in humans. Science Translational Medicine 12(573):eabd3601

[106] Auladell M, et al. (2022) Influenza virus infection history shapes antibody responses to influenza vaccination. Nature Medicine 28(2):363–372

[107] Yu TC, et al. (2022) A biophysical model of viral escape from polyclonal antibodies. Virus Evolution 8(2):veac110

[108] Park SW, et al. (2019) A practical generation-interval-based approach to inferring the strength of epidemics from their speed. Epidemics 27:12–18

[109] Park SW, et al. (2022) The importance of the generation interval in investigating dynamics and control of new SARS-CoV-2 variants. Journal of The Royal Society Interface 19(191):20220173

[110] Mustonen V and Lässig M (2010) Fitness flux and ubiquity of adaptive evolution. Proceedings of the National Academy of Sciences 107(9):4248–4253

[111] Kim K, et al. (2024) Measures of population immunity can predict the dominant clade of influenza A (H3N2) and reveal age-associated differences in susceptibility and specificity. medRxiv DOI:10.1101/2023.10.26.23297569

[112] Von Kleist M, et al. (2023) SARS-CoV-2 evolution on a dynamic immune landscape. ResearchSquare, DOI:10.21203/rs.3.rs-3366919/v1

[113] Lee JM, et al. (2018) Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human h3n2 influenza variants. Proceedings of the National Academy of Sciences 115(35):E8276–E8285

[114] Lässig M, Mustonen V, and Walczak AM (2017) Predicting evolution. Nature Ecology & Evolution 1(3):0077

[115] Lässig M, Mustonen V, and Nourmohammad A (2023) Steering and controlling evolution—from bioengineering to fighting pathogens. Nature Reviews Genetics 24(12):851–867

[116] Iyer AS, et al. (2020) Persistence and decay of human antibody responses to the receptor binding domain of SARS-CoV-2 spike protein in COVID-19 patients. Science Immunology 5(52):eabe0367

[117] Israel A, et al. (2021) Large-scale study of antibody titer decay following BNT162b2 mRNA vaccine or SARS-CoV-2 infection. Vaccines 10(1):64

[118] World Health Organisation. Recommendations for influenza vaccine composition. `https://www.who.int/teams/global-influenza-programme/vaccines/who-recommendations`

[119] Morán-Tovar R and Lässig M (2023) Immune recognition dynamics of a primary infection. bioRxiv DOI:10.1101/2023.04.05.535743

[120] Jackson ML and Nelson JC (2013) The test-negative design for estimating influenza vaccine effectiveness. Vaccine 31(17):2165-8

[121] Trombetta CM and Kistner O and Montomoli E and Viviani S and Marchi S (2022) Influenza Viruses and Vaccines: The Role of Vaccine Effectiveness Studies for Evaluation of the Benefits of Influenza Vaccines. Vaccines 10:714

[122] Hodcroft EB, et al. (2021) Want to track pandemic variants faster? Fix the bioinformatics bottleneck. Nature 591(7848):30–33

[123] De Maio N, et al. (2023) Maximum likelihood pandemic-scale phylogenetics. Nature Genetics 55(5):746–752

[124] Turakhia Y, et al. (2021) Ultrafast sample placement on existing trees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nature Genetics 53(6):809–816

[125] Zhang A, et al. (2019) Original antigenic sin: how first exposure shapes lifelong anti–influenza virus immune responses. The Journal of Immunology 202(2):335–340

[126] McBride JM, et al. (2023) AlphaFold2 can predict single-mutation effects. Physical Review Letters 131(21):218401

[127] Buel GR and Walters KJ (2022) Can AlphaFold2 predict the impact of missense mutations on structure? Nature Structural & Molecular Biology 29(1):1–2

[128] Pak MA, et al. (2023) Using AlphaFold to predict the impact of single mutations on protein stability and function. PLoS One 18(3):e0282689