Inferring the intrinsic mutational fitness landscape of influenza-like evolving antigens from temporally ordered sequence data

Julia Doelger, Mehran Kardar, Arup K. Chakraborty

June 9, 2021

There still is no effective long-term protective vaccine against seasonal influenza, which continually evolves and regularly causes devastating epidemics in the human population. For finding such a broadly protective immunization strategy it is useful to know how easily the virus can escape via mutation from specific antibody responses. This information is encoded in the sequence fitness landscape of the viral antigens. Here we present a computational method to infer the intrinsic mutational fitness landscape of influenza-like evolving antigens from yearly sequence data. We test inference performance, which relies on stringent selection bottlenecks between epidemic seasons, with computer-generated sequence data that are based on stochastic simulations mimicking basic features of seasonal influenza evolution. Although the numerically simulated model does create a phylogeny based on the allowed mutations, the stringency-based inference scheme does not use this information. This provides a contrast to other methods that rely on reconstruction of phylogenetic trees. Our method just needs sufficiently many samples over multiple years. With our method we are able to infer single- as well as pairwise mutational fitness effects from the simulated sequence time series for short influenza-like antigens. The here presented fitness inference approach may have potential future use for immunization design by identifying intrinsically vulnerable immune target combinations on influenza antigenic proteins.

1 Introduction

Global seasonal influenza epidemics are caused by influenza A and B viruses that, although being effectively targeted by human immune responses and long-term immune memory, are able to persistently escape the population-wide immune memory via sequence mutations [Petrova and Russell, 2018]. The dominantly targeted antigen of influenza virus is the glycoprotein HA that is located on the viral surface together with the other surface glycoprotein NA, which also acts as antigen. HA is responsible for binding to sialic acid on human cell surfaces and it thereby enables viral cell entry. The human immune system

produces antibodies, which primarily bind to different regions (epitopes) on HA thereby blocking the virus from cell attachment and entry. There are 5 dominant and easily accessible epitope regions on the head of HA that have been identified in the circulating subtype H3, which are labeled with the letters A-E [Wiley et al., 1981, Skehel et al., 1984, Shih et al., 2007]. These represent the parts of the protein sequence, where the virus predominantly produces amino acid substitutions that abrogate antibody binding and thus lead to immune escape [Gerhard et al., 1981].

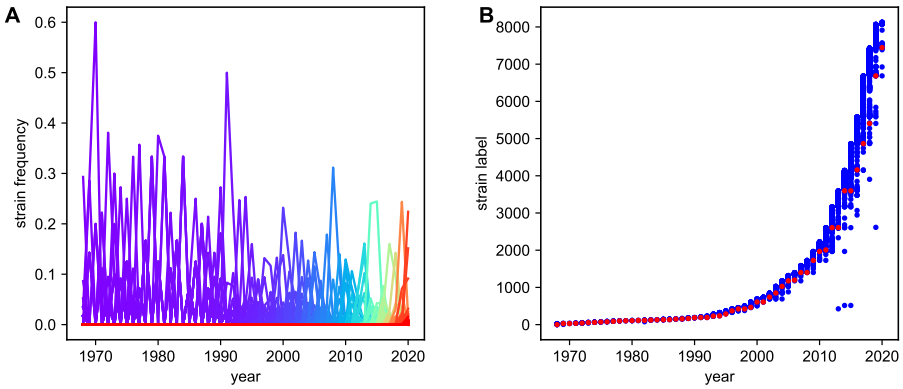

Figure 1: Strain succession for the evolution of HA (H3N2) sequences between 1968 and 2020. (A) Each unique HA amino acid sequence (strain) is shown with its observed frequency in each year as a solid line, with line colors ranging from purple (old strains) to red (new strains). (B) Strains are labeled with increasing numbers from old strains (low labels) to new strains (high labels). The respective strain, which is the most prevalent in each year is marked as red circle. Blue circles indicate strains that were observed with some non-zero frequency.

These interlinked dynamics of the mutating virus and responding human immunity cause a gradual evolution of the viral antigen that is known as antigenic drift [Smith et al., 2004], which leads to characteristic strain succession patterns in seasonal influenza (Fig. 1). Antigenic drift is also responsible for the fact that there is currently no long-term protective vaccine against seasonal influenza and why still around half a million people die globally from influenza infection [Carrat and Flahault, 2007]. Therefore it is important to create more effective vaccines and other immunization strategies, which target the virus where it is most vulnerable.

Even for the currently widely used seasonally updated influenza vaccines, the choice of vaccine strains is not trivial. For best efficacy one needs to make accurate predictions of the viral strains that will be prevalent in the future, based on past and current sequence information. Every year the WHO uses detailed information from international laboratories and worldwide experts to create rec-

ommendations on the composition of the influenza virus vaccine [WHO, 2021], but many seasonal vaccines still have a low effectivity compared to other viral vaccines. Thus many computational and experimental efforts are undertaken, which exclusively work on the task of analyzing and predicting the evolution of influenza antigenic sequences, with the goal of making seasonal vaccines more effective [Smith et al., 2004, Koel et al., 2013, Łuksza and Lässig, 2014, Neher et al., 2014, Bedford et al., 2014, Li et al., 2016, Hadfield et al., 2018]. But, although periodically updated vaccinations are continually improved and are currently the most effective method for preventive control of seasonal influenza epidemics, such relatively short-term predictions do not generally lead to long-term effective treatment plans [Paules et al., 2017].

Other approaches aim for cross-protective influenza treatments that are effective against a wide range of strains. Such approaches typically consider strongly conserved epitopes like the receptor binding site (RBS) or the stem of HA [Rajão and Pérez, 2018, Throsby et al., 2008, Ekiert et al., 2011, Corti et al., 2011, Dreyfus et al., 2012, Yamayoshi et al., 2017, Brandenburg et al., 2013, Ekiert et al., 2012, Schmidt et al., 2015, Whittle et al., 2011]. Methods targeting those regions require specialized methods for sophisticated vaccine protocols and drug designs [Steel et al., 2010, Yassine et al., 2015, Lu et al., 2014, Impagliazzo et al., 2015, Krammer et al., 2013, Hai et al., 2012, Nachbagauer et al., 2014, Eggink et al., 2014, Strauch et al., 2017, Kadam and Wilson, 2018, Amitai et al., 2020]. Additionally, for mutationally conserved sites it is generally not known if they are functionally conserved such that escape mutations are unviable or if they so far exhibited less amino acid substitutions, mainly because they are typically under a lower immune pressure than the exposed HA head epitopes [Amitai, 2020, Amitai et al., 2020, Park et al., 2020]. If the latter is true, those regions might not stay conserved for long if they are heavily targeted by new vaccines and if the virus can easily produce escape mutations at those previously less targeted regions.

Although the easily accessible sites on the head of HA are found to generally quickly escape human immune memory via amino acid substitution, mutations at some of those strongly targeted sites will be functionally more costly to the virus. For a long-term protective immunization approach it therefore would be useful to find and target primarily those sites on the HA head that are most vulnerable, i.e. that have difficulty finding viable mutational escape routes. We can further imagine targeting several sites simultaneously by specifically designed multi-clonal immune responses. In this case it would be useful to choose such combinations of sites as targets, which together are most vulnerable, and do not easily allow the combinations of mutations that lead to escape from the simultaneous responses. The information about the cost of such single and combined mutations at different protein sites is encoded in the intrinsic mutational fitness landscape of the viral sequence.

Previous studies were able to use approaches based on maximum entropy considerations and a method called Adaptive Cluster Expansion (ACE) to computationally infer intrinsic mutational fitness landscapes for other highly mutable viruses, HIV as well as polio, from sequence prevalence data [Dahirel et al., 2011,

Ferguson et al., 2013, Shekhar et al., 2013, Mann et al., 2014, Barton et al., 2016b, Butler et al., 2016, Chakraborty and Barton, 2017, Louie et al., 2018, Barton et al., 2019, Quadeer et al., 2020, Cocco and Monasson, 2011, Barton et al., 2016a]. The result of such fitness inference was used to propose a novel cross-protective immunization method against HIV using multidimensionally conserved parts of the proteome, which has been shown to be immunogenic in rhesus macaques [Murakowski et al., 2021]. Seasonal influenza evolves very differently in the human population than HIV. Since it is targeted by a population-wide immune memory that continually catches up with the viral evolution, it is permanently driven away from past strains as opposed to HIV, which evolves much more freely in its sequence landscape and is able to periodically revisit old strains [Pompei et al., 2012, Shekhar et al., 2013, Neher et al., 2014]. This immune-driven, non-equilibrium nature of influenza evolution requires a different method for the inference of the intrinsic mutational fitness landscape than the maximum entropy-based methods that were successful for HIV.

Recently a marginal path likelihood (MPL) method has been proposed for the inference of mutational fitness effects from sequence time series [Sohail et al., 2020]. However, this method considers selection due to a fitness landscape that is assumed to be time-invariant, and it does not try to disentangle intrinsic from immune-mediated fitness effects. The assumption of time-invariance of total fitness is not true for seasonal influenza evolution in the human population, since the population-wide immune-memory against each emerging mutant accumulates with every season and thus creates an ever changing fitness landscape that depends on the viral evolutionary history. Such a changing fitness landscape has been referred to as a "seascape" [Mustonen and Lässig, 2009, Mustonen and Lässig, 2010].

Here we present a method, with which we can infer the single and pairwise mutational intrinsic fitness costs from population-level sequence time series of an influenza-like evolving antigen. We test our inference approach on computer simulations and propose its potential application in the future to investigate yearly protein sequence time series data from influenza A/H3N2, in order to obtain combinations of vulnerable antibody targets.

2 Model of influenza antigen evolution

In our model for influenza evolution, we consider each influenza epidemic season as an evolutionary step, in which different viral strains, represented as unique protein sequences, evolve and compete with each other according to intrinsic and host immunity-mediated driving forces. In the following we will describe the components of our model, which we use both to create computer-generated sequence data and to motivate the inference method, which we will describe in section 3. Our influenza evolution model is motivated and inspired by several previous modeling studies, which describe essential properties of the evolution of influenza-like pathogen populations that lead to the characteristic spindle-like phylogeny and strain succession pattern of seasonal

influenza [Gog and Grenfell, 2002, Łuksza and Lässig, 2014, Neher et al., 2014, Rouzine and Rozhnova, 2018, Yan et al., 2019]. Those pathogen models also relate to more general models of rapid adaptation in asexual populations that evolve towards increasing fitness in a traveling-wave type manner [Tsimring et al., 1996, Rouzine et al., 2003, Desai and Fisher, 2007].

2.1 Sequence representation

For the representation of viral strains we here use a binary sequence representation, in which a strain $\mathbf{S}_j = (s_j^1, s_j^2, ..., s_j^L)$, i.e. a unique sequence, is represented as a string of L ones and zeros. This is a coarse-grained representation of a real protein, wherein, in principle, there could be 20 possible amino acids at each residue. For proteins that do not mutate too much (like the p24 structural protein of HIV), a binary Ising like representation, instead of a Potts model, is reasonable [Mann et al., 2014]. Also, our approach could be generalized to Potts models. Here, we also consider sequences of $L < 100$, which is much shorter than real protein sequences.

In principle the full protein information could be contained in such binary sequences. However, we here mainly consider sequences of lengths $L < 100$ shorter than typical antigenic proteins, which typically contain > 100 amino acid residues and 20 possible amino acids per residue. Therefore our model sequences are not designed to contain full information but rather a coarse-grained representation of the viral proteins.

2.2 Fitness model

The time-dependent fitness landscape in our model, which defines the fitness of different strains, is composed of two components. The intrinsic fitness represents the intrinsic abilities of viruses belonging to a given strain to infect, reproduce and transmit in a susceptible human population. The host immunity-mediated fitness cost, on the other hand, represents the accumulated immunity against viruses belonging to a given strain in the host population, which reduces the number of susceptible hosts and therefore reduces the fitness of the respective strain. The total fitness of a strain $\mathbf{S}_j$ at time t, given the evolutionary history $\mathbf{x}(t' < t)$ of the whole virus population in humans, is modeled as:

$$F_{\text{total}}(\mathbf{S}_j, \mathbf{x}(t' < t)) = F_{\text{int}}(\mathbf{S}_j) + F_{\text{host}}(\mathbf{S}_j, \mathbf{x}(t' < t)) \tag{1}$$

with intrinsic and immunity-mediated fitness components F_{int} and F_{host}.

Intrinsic fitness model

The intrinsic fitness of a strain in our model is represented with a 2-point approximation as

$$F_{\text{int}}(\mathbf{S}_j) = F_0 + \sum_\alpha h_\alpha s_j^\alpha + \sum_{\alpha<\beta} J_{\alpha\beta} s_j^\alpha s_j^\beta. \tag{2}$$

Here F_0 represents the intrinsic fitness of a reference strain which is represented as a string of zeros, the second term represents the fitness change due to independent mutations at each sequence site α compared to the reference strain ($s_j^\alpha = 0$ if unmutated, 1 otherwise), and the last term represents the additional fitness change due to coupled mutations at pairs of sites α and β. The single-mutational fitness coefficients $\{h\}$ and the mutational coupling coefficients $\{J\}$ describe the intrinsic mutational fitness landscape, which we ultimately want to infer from the observed sequences. The intrinsic fitness coefficients describe how easy or difficult it is for the virus to create escape mutations if specific sites or pairs of sites are targeted by the host. Note, that by using this Ising-type approximation of the intrinsic fitness landscape, we reduce the number of fitness parameters for binary sequences with e.g. length $L = 20$ from $2^L = 1048576$ unique strains to $L*(L+1)/2 = 210$ fitness parameters $\{h, J\}$. The fitness model used in Eq. (2) is different compared to maximum entropy models wherein the fitness is the exponential of an expression like Eq. (2). We use this formulation for convenience and for demonstrating our method.

Representation of host immunity-mediated fitness costs

The host-dependent immunity-mediated fitness component depends on the evolutionary history of the viral population and in our model is calculated with a functional form similar to that previously used in other influenza fitness models [Gog and Grenfell, 2002, Łuksza and Lässig, 2014, Yan et al., 2019], i.e.

$$F_{\text{host}}(\mathbf{S}_j, \mathbf{x}(t' < t)) = -\sigma_{\text{h}} \sum_{t'<t} \sum_i x(\mathbf{S}_i, t') \exp\left(-|\mathbf{S}_j^{\text{ep}} - \mathbf{S}_i^{\text{ep}}|/D_0\right). \quad (3)$$

The immunity-mediated fitness component decreases the fitness of each emerging strain over time and is proportional to the prevalence $x(\mathbf{S}_i, t')$ of antigenically similar strains $\mathbf{S}_i$ in previous years t'. This accumulating fitness cost forces the virus to continuously evolve away from previously prevalent sequences. Here $|\mathbf{S}_j^{\text{ep}} - \mathbf{S}_i^{\text{ep}}|$ describes the mutational distance between strain $\mathbf{S}_i$ and $\mathbf{S}_j$ within their immune-targeted epitope regions and D_0 is the cross-immunity distance, i.e., the typical mutational distance within epitope regions, beyond which two strains are dissimilar enough to not be targeted by immune responses that were raised against the respective other. In the following we will assume for simplicity that all modeled sequence sites are equally immune-targeted and therefore $\mathbf{S}_j^{\text{ep}} = \mathbf{S}_j$, but the model can, in principle, be extended to incorporate less or untargeted sites in the model sequence $\mathbf{S}_j$.

2.3 Sequence selection

During the spread of viral infections in the course of a flu season, different strains are assumed to grow with a growth rate given by their respective fitness

(Eq. 1), i.e.,

$$F_{\text{total}}(\mathbf{S}_j, \mathbf{x}(t' < t)) = F_0 + \sum_{\alpha} h_\alpha s_j^\alpha + \sum_{\alpha<\beta} J_{\alpha\beta} s_j^\alpha s_j^\beta - \sigma_{\text{h}} \sum_{t'<t} \sum_i x(\mathbf{S}_i, t') \exp\left(-|\mathbf{S}_j^{\text{ep}} - \mathbf{S}_i^{\text{ep}}|/D_0\right). \quad (4)$$

At the end of a season a fixed number N_{pop} of sequences is assumed to survive into the next season. The expected frequency of a given strain $\mathbf{S}_j$ among the selected sequences in season $t+1$ is calculated as

$$p(\mathbf{S}_i, t+1) = \frac{\exp\left(F_{\text{total}}(\mathbf{S}_j, \mathbf{x}(t' < t))\right) x_{\text{m}}(\mathbf{S}_j, t)}{\sum_i \exp\left(F_{\text{total}}(\mathbf{S}_i, \mathbf{x}(t' < t))\right) x_{\text{m}}(\mathbf{S}_i, t)}, \quad (5)$$

where $x_{\text{m}}(\mathbf{S}_j, t)$ denotes the frequency of strain $\mathbf{S}_j$ in season t before growth and selection. The number of selected sequences $N(\mathbf{S}_j, t+1)$ belonging to strain $\mathbf{S}_j$ are drawn from a multinomial distribution with probabilities given by Eq. (5) and N_{pop} as the number of draws.

2.4 Sequence mutation

We assume that mutation is a separate step from selection in each flu season. Thus in every modeled season t before growth and selection, sequences are modeled to mutate and thereby create a new frequency distribution $\mathbf{x}_{\text{m}}(t)$. We assume one symmetric mutation rate μ, per season, between the two different states at each site. From the previously selected sequences given by the frequency distribution $\mathbf{x}(t)$, the probability, with which strain $\mathbf{S}_j$ is present after mutation, is given as

$$p_{\text{m}}(\mathbf{S}_j, t) = x(\mathbf{S}_j, t) + \sum_{i=1}^{M} \mu_{ij} \left(x(\mathbf{S}_i, t) - x(\mathbf{S}_j, t)\right) \quad (6)$$

with the mutation probability $\mu_{ij} = \mu_{ji}$ for mutation between strains $\mathbf{S}_i$ and $\mathbf{S}_j$ given as

$$\mu_{ij} = \mu^{|\mathbf{S}_i - \mathbf{S}_j|} (1-\mu)^{L - |\mathbf{S}_i - \mathbf{S}_j|}. \quad (7)$$

In Eq. (6) the sum iterates over all $M = 2^L$ possible strains, but in reality only a small fraction of possible strains is present in any season with $x(\mathbf{S}_i, t) \neq 0$. In a stochastic simulation procedure the mutated sequences can simply be created by randomly switching the state at each site in each selected sequence with probability μ.

As mentioned before, the main motivation for the above model is to develop a method for inferring the intrinsic mutational fitness landscape of influenza-like evolving antigenic sequences. The goal within our model framework is to infer the intrinsic fitness coefficients $\{h, J\}$ from yearly observations $\mathbf{x}(t < T)$ (until the most recent season T) of antigenic protein sequences, in order to learn

about the vulnerability and mutational escape likelihood at different single and combinations of sequence sites upon being targeted.

On this account we developed an inference approach, which we test on computer-generated data that we produced via simulation of our sequence evolution model with a known fitness landscape

3 Analysis and inference based on simulated sequence data

3.1 Simulation produces influenza-like antigen evolution

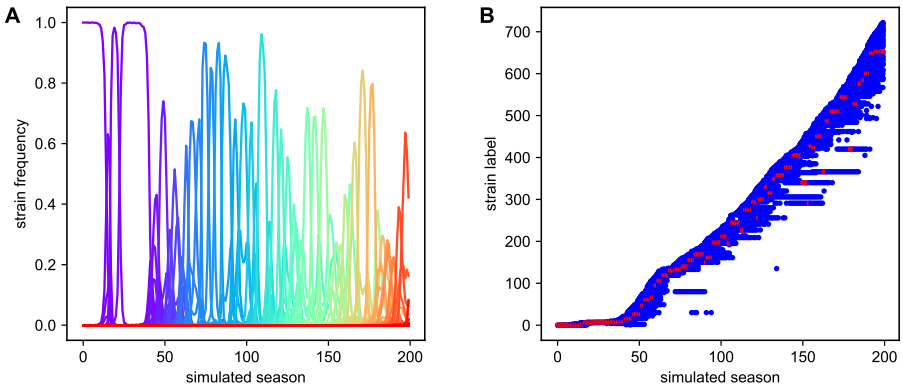

Figure 2: Strain succession for the evolution of simulated data over 200 time steps. (A) Each unique sequence (strain) is shown with its observed frequency in each simulated season as a solid line, with line colors ranging from purple (old strains) to red (new strains). (B) Strains are labeled with increasing numbers from old strains (low labels) to new strains (high labels). The respective strain, which is the most prevalent in each simulated season is marked as red circle. Blue circles indicate strains that were observed with some non-zero frequency. For the shown example the parameter values for simulation and analysis are: $N_{\text{pop}} = 10^5$, $L = 20$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D_0 = 5$, $N_{\text{simu}} = 200$, $B = 10^3$.

Based on the presented model we ran stochastic simulations to compare the computer-generated sequence succession to influenza sequence data and to test our fitness inference method. The simulation parameters are the sequence length L, population size N_{pop}, number of simulated seasons N_{simu}, mutation rate μ, cross-immunity distance D_0, host-immunity coefficient σ_{h} and intrinsic fitness coefficients $\{h, J\}$ (cf. Tab. 1). In the beginning of each simulation the population is initialized with the unmutated strain $\mathbf{S}_0 = (0, 0, ..., 0)$. Accordingly the initial strain frequency distribution is given by $x(\mathbf{S}_0, t = 0) = 1$. The intrinsic fitness landscape in our simulations is predetermined by the chosen intrinsic

fitness parameters $\{h, J\}$. As just an example, we sample a limited number of these parameters from Ising coefficients inferred for the HIV p24 protein using a maximum entropy model [Mann et al., 2014]. In each time step representing one epidemic season sequences are first mutated according to rate μ. After mutation, current fitness of each present strain is calculated with Eq. (4), based on which the selection probability (Eq. (5)) of each strain is determined. The sequence population for the next season is then sampled by N_{pop} random draws from a multinomial distribution with the individual probabilities given by the respective selection probabilities of each strain.

For a range of parameter choices our stochastic simulations produce influenza-like immune-driven strain succession patterns (Fig. 2), which are qualitatively similar to those observed for evolution of the influenza spike protein HA (H3N2) in the human population (Fig. 1). This similarity indicates that our model is able to capture the essential dynamics of antigenic evolution for pathogens like seasonal influenza. One difference in the shown figures (Figs. 1B and 2B) is the approximately exponential increase of total sequence diversity in strains based on full HA amino acid sequence data versus the more linear increase of total sequence diversity in a simulation of binary sequences of length 20. The dependence of this growth of strain diversity on various parameters and its underlying mechanisms should be further investigated when translating our procedures to infer the fitness landscape of influenza. We speculate that the exponential increase of sequence diversity in observed influenza sequences arises from the rapid increase of the amount of yearly acquired sequencing data in the past years.

3.2 Observation of stringent selection regime

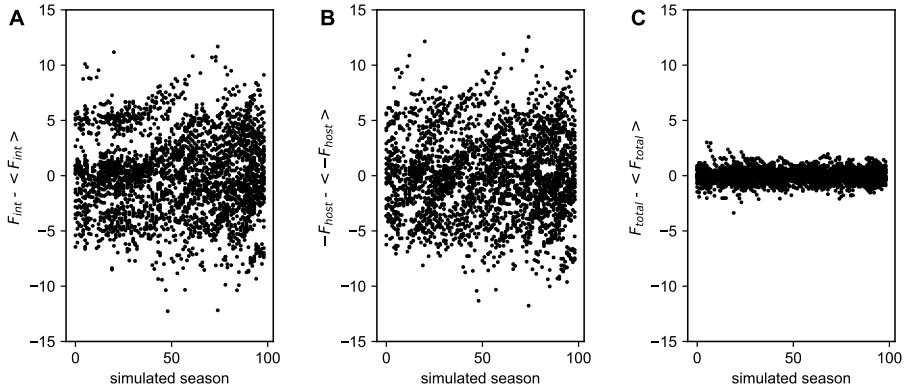

Figure 3: Fitness deviations from the mean of sampled strains for each simulated season between season 100 and 200. (A) intrinsic fitness component F_{int}, (B) immunity-dependent fitness component F_{host}, (C) total fitness $F_{\text{total}} = F_{\text{int}} + F_{\text{host}}$. For the shown example the parameter values for simulation and analysis are: $N_{\text{pop}} = 10^5$, $L = 20$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D_0 = 5$, $N_{\text{simu}} = 200$, $B = 10^3$.

For the analysis of the simulated sequences we randomly sampled a number B of sequences per season to imitate the sampling properties of real observed protein data, which contain only subsets of the yearly circulating viruses. For an example set of sampled data from one simulation we see that the distribution of total fitness is more narrow than the distributions of the intrinsic and the immunity-dependent fitness components (Fig. 3). The narrow total fitness distribution in each season indicates a stringent selection regime, in which only those strains in a narrow fitness range around the currently fittest strain survive into the next season. In this observed regime we have

$$F_{\text{total}}(\mathbf{S}_j, \mathbf{x}(t' < t)) \approx F(t, \mathbf{x}(t' < t)) \tag{8}$$

with $F(t, \mathbf{x}(t' < t))$ being a constant for each season t, conditional on the specific evolutionary history $\mathbf{x}(t' < t)$.

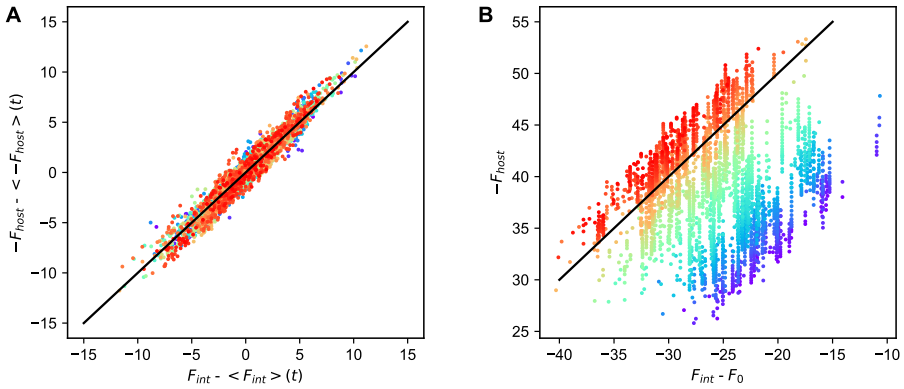

Figure 4: Negative immunity-dependent fitness cost F_{host} (y-axes) compared against intrinsic fitness F_{int} (x-axes) for sampled strains for each simulated season between season 100 and 200 (same data as in Fig. 3). (A) Fitness deviations from the mean as colored circles with a solid black line indicating 1:1 correspondence. (B) Absolute fitness components as colored circles with a solid black line indicating slope 1. Colors from purple to red in both panels indicate seasons from 100 to 200, in which the respective strains were sampled.

Indeed we find with our simulation a clear 1:1 correspondence between the intrinsic fitness variation and immunity-dependent fitness variation in each season (Fig. 4A), which add up to a roughly constant total fitness in each season, as the stringency assumption (Eq. (8)) suggests. In Fig. 4B it can be observed that in our simulation the absolute population fitness decreases with each year, both due to the emergence of less intrinsically fit strains and due to the population-wide accumulation of immune pressure.

3.3 Method for intrinsic fitness inference

From Eq. (1) together with Eq. (8) we obtain the following relation for the observed strains $\mathbf{S}_j$ in each given year in the case of stringent selection, i.e.,

$$-F_{\text{host}}(\mathbf{S}_j, \mathbf{x}(t' < t)) \approx \sum_{\alpha} h_\alpha s_j^\alpha + \sum_{\alpha<\beta} J_{\alpha\beta} s_j^\alpha s_j^\beta + F^*(t, \mathbf{x}(t' < t)), \quad (9)$$

where $F^*(t, \mathbf{x}(t' < t)) = F_0 - F(t, \mathbf{x}(t' < t))$ is another constant at time t, conditional on the evolutionary history until t. If we approximate the evolutionary history $\mathbf{x}(t' < t)$ with the observed strain frequencies starting from the first year of observation and assume the model parameters σ_{h} and D_0 to be known, e.g. as fit parameters to independent cross-immunity studies [Łuksza and Lässig, 2014], we can calculate $F_{\text{host}}(\mathbf{S}_j, \mathbf{x}(t' < t))$ for each observed strain in each season. We now use these host-dependent fitness values together with Eq. (9) to infer the intrinsic fitness coefficients $\{h, J\}$ as well as the additional parameters $\{F^*\}$ (one additional parameter per season). We here treat $\{F^*\}$ as independent parameters, although they generally depend on other model parameters and on the evolutionary history via the full path integral of the system. For the regression we minimize the sum of squared residuals between the data $Y_{\text{data}}(\mathbf{S}_j, t)$ given by the LHS of Eq. (9) and the model $Y_{\text{model}}(\mathbf{S}_j, t, \{h, J, F^*\})$ given by the RHS of Eq. (9), i.e.,

$$\{h, J, F^*\} = \arg \min_{\{h,J,F^*\}} \left[\frac{1}{2} \sum_j \left(Y_{\text{data}}(\mathbf{S}_j, t) - Y_{\text{model}}(\mathbf{S}_j, t, \{h, J, F^*\}) \right)^2 + \frac{\lambda_h}{2} \sum_\alpha h_\alpha^2 + \frac{\lambda_J}{2} \sum_{\alpha<\beta} J_{\alpha\beta}^2 + \frac{\lambda_{F^*}}{2} \sum_{t'} {F_{t'}^*}^2 \right], \quad (10)$$

where we also take into account regularization with coefficients $\lambda_h, \lambda_J, \lambda_{F^*}$ that in a Bayesian sense correspond to Gaussian prior distributions.

For inference we use the following equation, [Hastie et al., 2009, Eq. (3.44)],

$$\mathbf{M} = \left(\mathbf{X}^T \mathbf{X} + \mathbf{\Lambda} \right)^{-1} \mathbf{X}^T \mathbf{y} \quad (11)$$

to solve for the unique parameter values
$\mathbf{M} = (h_1, ..., h_L, J_1, ..., J_{L*(L-1)/2}, F_1^*, ..., F_{nseasons}^*)^T$, which minimize the sum of squared residuals subject to ridge-regularization (Eq. (10)). The feature vector for each sampled strain, which forms a row in the feature matrix $\mathbf{X}$, consists of binary features representing the single-mutational and double-mutational states of the respective sequence as well as its time of observation. $\mathbf{y}$ is a column vector, whose entries are given by the values $-F_{\text{host}}(\mathbf{S}_j)|\mathbf{x}(t' < t))$ (cf. Eq. (3)) for the respective sequence $\mathbf{S}_j$, sampled at time t. The non-zero regularization coefficients, $\{\lambda_h, \lambda_J, \lambda_{F*}\}$, are collected in the diagonal matrix $\mathbf{\Lambda}$, and regularization also ensures that no singularities are encountered at matrix inversion. The coefficients λ_h and λ_{F*} are set to very small values corresponding to a very wide, rather non-restrictive, prior distribution, while λ_J, corresponding to the assumed sparse mutational couplings, is set to 1.

3.4 Inferring the intrinsic mutational fitness landscape from simulated influenza-like sequence data

Table 1: Parameters for simulation of influenza-like sequence evolution and for intrinsic fitness inference

Parameter	Description	Default value
$\{h, J\}$	intrinsic fitness coefficients for single mutations and pairwise mutational couplings	sampled values from HIV protein p24
L	length of sequence representation	20
μ	mutation rate (per sequence site)	10^{-4}
D_0	cross-immunity distance	5
N_{pop}	population size	10^5
σ_{h}	host-fitness coefficient	1
$\{\lambda_h, \lambda_J, \lambda_{F^*}\}$	regularization coefficients for inference	$\{10^{-4}, 1, 10^{-4}\}$
n_{seasons}	number of years/seasons used for inference	100
B	number of sampled sequences per year	10^3

The parameters for simulation and inference with chosen default values are collected in Tab. (1).

In Fig. 5 we compare the inferred and the simulated intrinsic fitness coefficients for one simulation. The correlation coefficients between simulated and inferred coefficients and in particular the Pearson correlation r_{hJ} between the total fitness effects of double mutations indicates if the specific fitness inference on the particular sequence data set can successfully distinguish between pairs of sites, at which escape mutations lead to either low or high (negative) fitness costs.

Besides the correlation coefficient r_{hJ} we use another measure for inference performance, which can be useful if we are mainly interested in identifying those pairs of sites that have the most deleterious fitness effect, i.e. those whose intrinsic fitness change compared to the reference sequence is below a certain negative threshold with

$$h_\alpha + h_\beta + J_{\alpha\beta} < F_{\text{threshold}} < 0. \tag{12}$$

In this case we can use typical classification performance measures to assess how well our inference method can distinguish between deleterious and more neutral or beneficial double mutations. We compare the classification of each pair (based on the inferred coefficients) with the classification of the simulation input values by calculating the precision-recall curve (PRC) as well as the receiver operating characteristic curve (ROC) and the respective areas under the curves (AUC) (Fig. 6), which approach 1 in the case of perfect classification skill.

When calculating the inference performance for one simulation with sequence length $L = 20$ in terms of correlation r_{hJ} and classification performance (AUC)

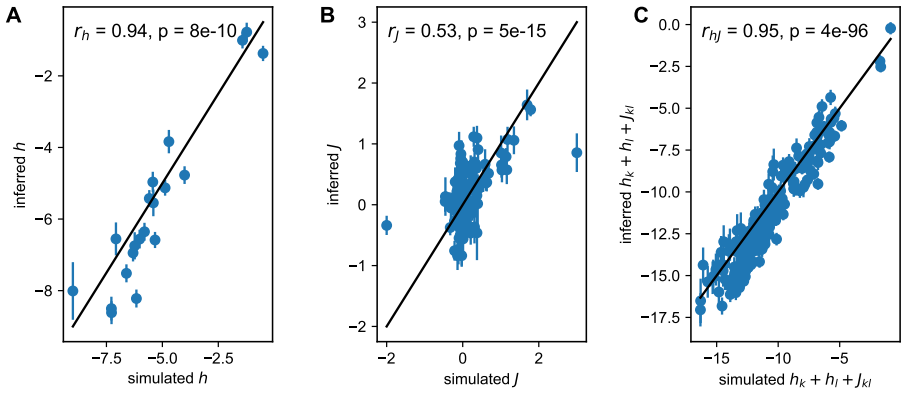

Figure 5: Parameter correlations for the inference on one simulated data set. Inferred values of the fitness coefficients are shown against the fitness coefficients that were used as input values for the simulation. (A) single-site mutational fitness coefficients h, (B) coupling coefficients J for simultaneous mutations at any two sites, (C) total fitness changes $h_k + h_l + J_{kl}$ due to simultaneous mutations at any two sites k and l. Pearson correlation coefficients r together with their respective p values are shown in each panel for the respective set of parameters. For the shown example the parameter values for simulation and analysis are: $N_{\text{pop}} = 10^5$, $L = 20$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D_0 = 5$, $N_{\text{simu}} = 200$, $B = 10^3$, $n_{\text{seasons}} = 100$, $\lambda_h = 10^{-4}$, $\lambda_J = 1$, $\lambda_{F^*} = 10^{-4}$.

for various sample sizes (Fig. 7), we find that a minimum total number of sampled strains $n_{\text{seasons}} * B$ is required for accurate inference. In the shown example a total sample size of $\geq 10^5$ strains is required for high inference performance (Fig. 7B). Since this is true for a sequence of length $L = 20$, a very large number of sequences would be needed for an inference based on a protein representation with all amino acid sites $L > 100$, which indicates that for real proteins sequence representations with strongly reduced dimensions are needed for inferences based on the available amount of observed data.

Regarding the amount of data needed for fitness inference of a given antigen, we can quantify our rough scaling expectations with the following argument. For a sequence of length L and n_{seasons} the number of observed epidemic seasons, we need to determine $m = n_{\text{seasons}} + L(L+1)/2 \approx n_{\text{seasons}} + L^2/2$ parameters. With the conservative assumption that we need m independent equations for this inference task, we can estimate that we approach optimal inference performance when $B\, n_{\text{seasons}}\, \mu \sim m$. Here the number of needed total samples $B * n_{\text{seasons}}$ is assumed to increase with decreasing mutation rate μ, since an independent set of samples is obtained only roughly every $1/\mu$ years.

The inference performance further strongly depends on other parameters such as the sequence length L (Fig. 8A) and on the population size N_{pop} (Fig. 8B). Inference performance in terms of the correlation r_{hJ} between inferred

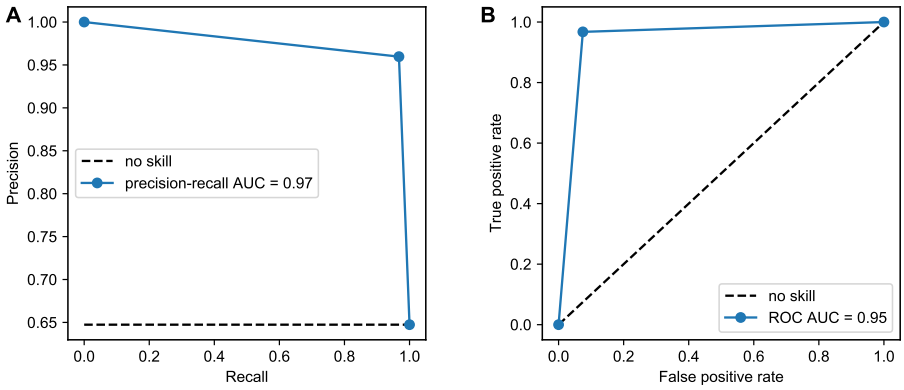

Figure 6: Classification performance for the inference on one simulated data set. Double mutations are classified as deleterious if their total fitness cost is lower than $F_{\text{threshold}} = -10$ (cf. Eq. (12)). (A) The precision-recall curve (PRC) and (B) the ROC curve for the classifier derived from inferred fitness coefficients. Blue, dashed lines show a no-skill classifier for comparison and the area under the classifier curve (AUC) is given in each panel, respectively. For the shown example the parameter values for simulation and analysis are: $N_{\text{pop}} = 10^5$, $L = 20$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D_0 = 5$, $N_{\text{simu}} = 200$, $B = 10^3$, $n_{\text{seasons}} = 100$, $\lambda_h = 10^{-4}$, $\lambda_J = 1$, $\lambda_{F^*} = 10^{-4}$, $F_{\text{threshold}} = -10$.

and simulated double-mutational fitness coefficients decreases with increasing sequence length and increases with increasing population size towards an upper limit ≤ 1. Thus, if the protein sequence representation is high-dimensional, a very large amount of data is needed for a high inference performance (indicating the need for dimensionality reduction) and secondly, if the effective population size that defines the selection bottleneck is small, inference can be poor. A large population size, however, will not help to have high inference performance if the sampling size B is low.

4 Discussion

We here presented a method for inferring the intrinsic mutational fitness landscape of influenza-like antigens from population-level protein sequence time series data. Our approach is able to infer single as well as pairwise mutational effects for binary sequences with several tens of sites. By simulating influenza-like evolutionary dynamics, we were able to analyze inference performance under different conditions such as for various sequence lengths and sample sizes. Our inference approach, in principle, only relies on the raw strain frequency data as function of time and does not depend on a separate inference of sequence phylogenies, as opposed to other analyses [Łuksza and Lässig, 2014, Neher et al., 2014].

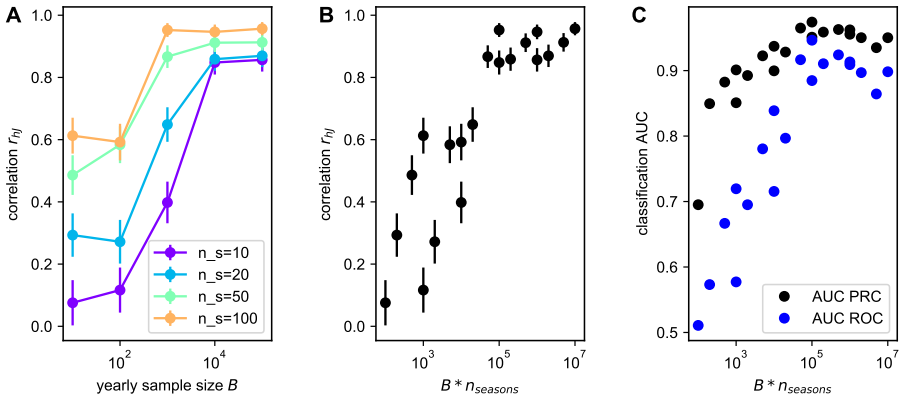

Figure 7: Inference performance for varying yearly sample size B per season and varying number n_{seasons} of seasons used for inference. (A) The correlation coefficient r_{hJ} between inferred and simulated double-mutational fitness costs as function of yearly sample size B for various n_{seasons}. (B) The correlation coefficient r_{hJ} as function of total sample size $B * n_{\text{seasons}}$. (C) The area (AUC) under the ROC curve and under the precision-recall curve (PRC) for classification of deleterious double mutations with classification threshold $F_{\text{threshold}} = -10$, shown as function of total sample size $B * n_{\text{seasons}}$. For the shown example the fixed parameter values for simulation and analysis are: $N_{\text{pop}} = 10^5$, $L = 20$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D_0 = 5$, $N_{\text{simu}} = 200$, $\lambda_h = 10^{-4}$, $\lambda_J = 1$, f$\lambda_{F^*} = 10^{-4}$.

In comparison to the recently proposed marginal path likelihood method (MPL) for sequence time series [Sohail et al., 2020], we were able to disentangle time-varying immunity-dependent fitness effects from intrinsic fitness, and we not only inferred the fitness effects of single mutations but also of double mutations at pairs of sites.

In order to make meaningful predictions based on observed influenza protein sequence data, our inference approach needs to be translated to this more complex system, which generally has a high-dimensional sequence landscape with around hundred residues in the head epitope regions of HA (A/H3N2) and 20 possible amino acids per residue. The inference performance will also be constrained by a relatively small number of samples, around $3 * 10^4$ HA sequences in total between 1968 and 2020 [Squires et al., 2012, Zhang et al., 2017] [put fasta file in SI and refer to it here].

For using our inference approach on the influenza protein data, one further needs to make sure that the cross-immunity function in $-F_{\text{host}}$ (Eq. 3), which we use as response variable, adequately captures the cross-immunity between different strains. The total mutational distance in the epitope regions, which we use in our model and which has been used in previous studies [Łuksza and Lässig, 2014] for estimating cross-immunity, only roughly captures the cross-immunity measurements from hemagglutination inhibition (HI) as-

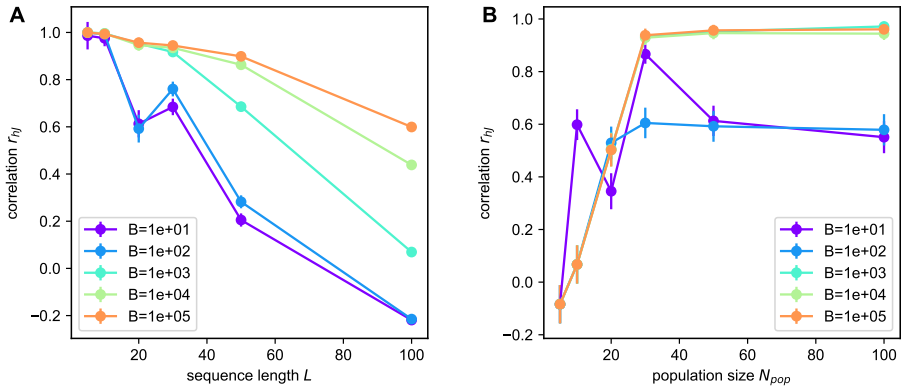

Figure 8: Inference performance in terms of the correlation coefficient r_{hJ} between inferred and simulated double-mutational fitness costs, for varying simulation and analysis parameters. (A) Inference performance as function of sequence length L, (B) inference performance as function of population size $N_{\rm pop}$. For both parameter explorations the yearly sample size B was varied between 10 and 10^5. For the shown simulation results the respective fixed parameter values for simulation and analysis are: $N_{\rm pop} = 10^5$, $L = 20$, $\mu = 10^{-4}$, $\sigma_{\rm h} = 1$, $D_0 = 5$, $N_{\rm simu} = 200$, $n_{\rm seasons} = 100$, $\lambda_h = 10^{-4}$, $\lambda_J = 1$, $\lambda_{F^*} = 10^{-4}$.

says. Analysis of such HI data, in which the proposed cross-immunity function is compared against measured cross-immunities, suggests a typical cross-immunity distance D_0 of 5 amino acids or 14 nucleotide residues for seasonal influenza A (H3N2) strains [Łuksza and Lässig, 2014, Ant, 2021], i.e., two strains that differ by more than 5 amino acid mutations within their epitope regions typically experience negligible cross-immunity to each other's immune responses. Better fitting cross-immunity functions could be constructed from strain antigenic distances in 2- to 5-dimensional antigenic maps, which can be inferred from the available HI data [Smith et al., 2004, Ant, 2021]. For this method, however, each viral strain needs to be measured against at least a few other viral strains to be accurately placed in the map. Another possibility might be to get a better cross-immunity prediction based on genetic data, if we used the separate mutational distances within each of the 5 head epitopes, instead of the total distance between epitope sequences. It is intuitive to assume that it does make a difference for the cross-immunity between two strains, whether the mutations between those strains occur only within one epitope region or whether the mutated sites are scattered across different epitopes thereby potentially inhibiting the binding of a wider range of antibodies that target different regions on the protein.

For testing fitness inference performance on real data, we generally do not have much direct information on the intrinsic effects of various mutations besides from some in-vitro mutational assays [Wu et al., 2017, Wu et al., 2018,

Wu et al., 2020], which are locally constrained to small parts of the sequence space and which only measure fitness in terms of functional replication in cells, not across the human population. The application of classical machine-learning methods of testing inference based on predictions on held-out data are also challenging due to the complex time-dependent nature and general sparsity and heterogeneity of available sequence data.

In conclusion we have proposed a method for inferring the intrinsic mutational fitness landscape of influenza-like viruses from time series of observed antigenic sequences, which can hopefully contribute to the development of new cross- and long-term protective immunization strategies against seasonal influenza.

References

[Ant, 2021] (2021). Antigenic cartography. www.antigenic-cartography.org. Accessed: 2021-04-29.

[WHO, 2021] (2021). WHO recommendations on the composition of influenza virus vaccines. https://www.who.int/influenza/vaccines/virus/recommendations/en/. Accessed: 2021-04-28.

[Amitai, 2020] Amitai, A. (2020). Viral surface geometry shapes influenza and coronavirus spike evolution. *bioRxiv*.

[Amitai et al., 2020] Amitai, A., Sangesland, M., Barnes, R. M., Rohrer, D., Lonberg, N., Lingwood, D., and Chakraborty, A. K. (2020). Defining and manipulating b cell immunodominance hierarchies to elicit broadly neutralizing antibody responses against influenza virus. *Cell Systems*, 11(6):573–588.

[Barton et al., 2016a] Barton, J. P., De Leonardis, E., Coucke, A., and Cocco, S. (2016a). Ace: adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics*, 32(20):3089–3097.

[Barton et al., 2016b] Barton, J. P., Goonetilleke, N., Butler, T. C., Walker, B. D., McMichael, A. J., and Chakraborty, A. K. (2016b). Relative rate and location of intra-host hiv evolution to evade cellular immunity are predictable. *Nature communications*, 7(1):1–10.

[Barton et al., 2019] Barton, J. P., Rajkoomar, E., Mann, J. K., Murakowski, D. K., Toyoda, M., Mahiti, M., Mwimanzi, P., Ueno, T., Chakraborty, A. K., and Ndung'u, T. (2019). Modelling and in vitro testing of the hiv-1 nef fitness landscape. *Virus evolution*, 5(2):vez029.

[Bedford et al., 2014] Bedford, T., Suchard, M. A., Lemey, P., Dudas, G., Gregory, V., Hay, A. J., McCauley, J. W., Russell, C. A., Smith, D. J., and Rambaut, A. (2014). Integrating influenza antigenic dynamics with molecular evolution. *elife*, 3:e01914.

[Brandenburg et al., 2013] Brandenburg, B., Koudstaal, W., Goudsmit, J., Klaren, V., Tang, C., Bujny, M. V., Korse, H. J., Kwaks, T., Otterstrom, J. J., Juraszek, J., et al. (2013). Mechanisms of hemagglutinin targeted influenza virus neutralization. *PloS one*, 8(12):e80034.

[Butler et al., 2016] Butler, T. C., Barton, J. P., Kardar, M., and Chakraborty, A. K. (2016). Identification of drug resistance mutations in hiv from constraints on natural evolution. *Physical Review E*, 93(2):022412.

[Carrat and Flahault, 2007] Carrat, F. and Flahault, A. (2007). Influenza vaccine: the challenge of antigenic drift. *Vaccine*, 25(39-40):6852–6862.

[Chakraborty and Barton, 2017] Chakraborty, A. K. and Barton, J. P. (2017). Rational design of vaccine targets and strategies for hiv: A crossroad of statistical physics, biology, and medicine. *Reports on Progress in Physics*, 80(3):032601.

[Cocco and Monasson, 2011] Cocco, S. and Monasson, R. (2011). Adaptive cluster expansion for inferring boltzmann machines with noisy data. *Physical Review Letters*, 106(9):090601.

[Corti et al., 2011] Corti, D., Voss, J., Gamblin, S. J., Codoni, G., Macagno, A., Jarrossay, D., Vachieri, S. G., Pinna, D., Minola, A., Vanzetta, F., et al. (2011). A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza a hemagglutinins. *Science*, 333(6044):850–856.

[Dahirel et al., 2011] Dahirel, V., Shekhar, K., Pereyra, F., Miura, T., Artyomov, M., Talsania, S., Allen, T. M., Altfeld, M., Carrington, M., Irvine, D. J., et al. (2011). Coordinate linkage of hiv evolution reveals regions of immunological vulnerability. *Proceedings of the National Academy of Sciences*, 108(28):11530–11535.

[Desai and Fisher, 2007] Desai, M. M. and Fisher, D. S. (2007). Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics*, 176(3):1759–1798.

[Dreyfus et al., 2012] Dreyfus, C., Laursen, N. S., Kwaks, T., Zuijdgeest, D., Khayat, R., Ekiert, D. C., Lee, J. H., Metlagel, Z., Bujny, M. V., Jongeneelen, M., et al. (2012). Highly conserved protective epitopes on influenza b viruses. *Science*, 337(6100):1343–1348.

[Eggink et al., 2014] Eggink, D., Goff, P. H., and Palese, P. (2014). Guiding the immune response against influenza virus hemagglutinin toward the conserved stalk domain by hyperglycosylation of the globular head domain. *Journal of virology*, 88(1):699–704.

[Ekiert et al., 2011] Ekiert, D. C., Friesen, R. H., Bhabha, G., Kwaks, T., Jongeneelen, M., Yu, W., Ophorst, C., Cox, F., Korse, H. J., Brandenburg, B., et al. (2011). A highly conserved neutralizing epitope on group 2 influenza a viruses. *Science*, 333(6044):843–850.

[Ekiert et al., 2012] Ekiert, D. C., Kashyap, A. K., Steel, J., Rubrum, A., Bhabha, G., Khayat, R., Lee, J. H., Dillon, M. A., O'Neil, R. E., Faynboym, A. M., et al. (2012). Cross-neutralization of influenza a viruses mediated by a single antibody loop. *Nature*, 489(7417):526–532.

[Ferguson et al., 2013] Ferguson, A. L., Mann, J. K., Omarjee, S., Ndung'u, T., Walker, B. D., and Chakraborty, A. K. (2013). Translating hiv sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity*, 38(3):606–617.

[Gerhard et al., 1981] Gerhard, W., Yewdell, J., Frankel, M. E., and Webster, R. (1981). Antigenic structure of influenza virus haemagglutinin defined by hybridoma antibodies. *Nature*, 290(5808):713–717.

[Gog and Grenfell, 2002] Gog, J. R. and Grenfell, B. T. (2002). Dynamics and selection of many-strain pathogens. *Proceedings of the National Academy of Sciences*, 99(26):17209–17214.

[Hadfield et al., 2018] Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123.

[Hai et al., 2012] Hai, R., Krammer, F., Tan, G. S., Pica, N., Eggink, D., Maamary, J., Margine, I., Albrecht, R. A., and Palese, P. (2012). Influenza viruses expressing chimeric hemagglutinins: globular head and stalk domains derived from different subtypes. *Journal of virology*, 86(10):5774–5781.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

[Impagliazzo et al., 2015] Impagliazzo, A., Milder, F., Kuipers, H., Wagner, M. V., Zhu, X., Hoffman, R. M., van Meersbergen, R., Huizingh, J., Wanningen, P., Verspuij, J., et al. (2015). A stable trimeric influenza hemagglutinin stem as a broadly protective immunogen. *Science*, 349(6254):1301–1306.

[Kadam and Wilson, 2018] Kadam, R. U. and Wilson, I. A. (2018). A small-molecule fragment that emulates binding of receptor and broadly neutralizing antibodies to influenza a hemagglutinin. *Proceedings of the National Academy of Sciences*, 115(16):4240–4245.

[Koel et al., 2013] Koel, B. F., Burke, D. F., Bestebroer, T. M., Van Der Vliet, S., Zondag, G. C., Vervaet, G., Skepner, E., Lewis, N. S., Spronken, M. I., Russell, C. A., et al. (2013). Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science*, 342(6161):976–979.

[Krammer et al., 2013] Krammer, F., Pica, N., Hai, R., Margine, I., and Palese, P. (2013). Chimeric hemagglutinin influenza virus vaccine constructs elicit broadly protective stalk-specific antibodies. *Journal of virology*, 87(12):6542–6550.

[Li et al., 2016] Li, C., Hatta, M., Burke, D. F., Ping, J., Zhang, Y., Ozawa, M., Taft, A. S., Das, S. C., Hanson, A. P., Song, J., et al. (2016). Selection of antigenically advanced variants of seasonal influenza viruses. *Nature Microbiology*, 1(6):1–10.

[Louie et al., 2018] Louie, R. H., Kaczorowski, K. J., Barton, J. P., Chakraborty, A. K., and McKay, M. R. (2018). Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proceedings of the National Academy of Sciences*, 115(4):E564–E573.

[Lu et al., 2014] Lu, Y., Welsh, J. P., and Swartz, J. R. (2014). Production and stabilization of the trimeric influenza hemagglutinin stem domain for potentially broadly protective influenza vaccines. *Proceedings of the National Academy of Sciences*, 111(1):125–130.

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

[Mann et al., 2014] Mann, J. K., Barton, J. P., Ferguson, A. L., Omarjee, S., Walker, B. D., Chakraborty, A., and Ndung'u, T. (2014). The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS Comput Biol*, 10(8):e1003776.

[Murakowski et al., 2021] Murakowski, D. K., Barton, J. P., Peter, L., Chandrashekar, A., Bondzie, E., Gao, A., Barouch, D. H., and Chakraborty, A. K. (2021). Adenovirus-vectored vaccine containing multidimensionally conserved parts of the hiv proteome is immunogenic in rhesus macaques. *Proceedings of the National Academy of Sciences*, 118(5).

[Mustonen and Lässig, 2009] Mustonen, V. and Lässig, M. (2009). From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends in Genetics*, 25(3):111–119.

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

[Nachbagauer et al., 2014] Nachbagauer, R., Wohlbold, T. J., Hirsh, A., Hai, R., Sjursen, H., Palese, P., Cox, R. J., and Krammer, F. (2014). Induction of broadly reactive anti-hemagglutinin stalk antibodies by an h5n1 vaccine in humans. *Journal of virology*, 88(22):13260–13268.

[Neher et al., 2014] Neher, R. A., Russell, C. A., and Shraiman, B. I. (2014). Predicting evolution from the shape of genealogical trees. *eLife*, 3:e03568.

[Park et al., 2020] Park, J.-K., Xiao, Y., Ramuta, M. D., Rosas, L. A., Fong, S., Matthews, A. M., Freeman, A. D., Gouzoulis, M. A., Batchenkova, N. A., Yang, X., et al. (2020). Pre-existing immunity to influenza virus hemagglutinin stalk might drive selection for antibody-escape mutant viruses in a human challenge model. *Nature medicine*, 26(8):1240–1246.

[Paules et al., 2017] Paules, C. I., Marston, H. D., Eisinger, R. W., Baltimore, D., and Fauci, A. S. (2017). The pathway to a universal influenza vaccine. *Immunity*, 47(4):599–603.

[Petrova and Russell, 2018] Petrova, V. N. and Russell, C. A. (2018). The evolution of seasonal influenza viruses. *Nature Reviews Microbiology*, 16(1):47–60.

[Pompei et al., 2012] Pompei, S., Loreto, V., and Tria, F. (2012). Phylogenetic properties of rna viruses. *PLoS One*, 7(9):e44849.

[Quadeer et al., 2020] Quadeer, A. A., Barton, J. P., Chakraborty, A. K., and McKay, M. R. (2020). Deconvolving mutational patterns of poliovirus outbreaks reveals its intrinsic fitness landscape. *Nature Communications*, 11(1):1–13.

[Rajão and Pérez, 2018] Rajão, D. S. and Pérez, D. R. (2018). Universal vaccines and vaccine platforms to protect against influenza viruses in humans and agriculture. *Frontiers in Microbiology*, 9:123.

[Rouzine and Rozhnova, 2018] Rouzine, I. M. and Rozhnova, G. (2018). Antigenic evolution of viruses in host populations. *PLoS pathogens*, 14(9):e1007291.

[Rouzine et al., 2003] Rouzine, I. M., Wakeley, J., and Coffin, J. M. (2003). The solitary wave of asexual evolution. *Proceedings of the National Academy of Sciences*, 100(2):587–592.

[Schmidt et al., 2015] Schmidt, A. G., Therkelsen, M. D., Stewart, S., Kepler, T. B., Liao, H.-X., Moody, M. A., Haynes, B. F., and Harrison, S. C. (2015). Viral receptor-binding site antibodies with diverse germline origins. *Cell*, 161(5):1026–1034.

[Shekhar et al., 2013] Shekhar, K., Ruberman, C. F., Ferguson, A. L., Barton, J. P., Kardar, M., and Chakraborty, A. K. (2013). Spin models inferred from patient-derived viral sequence data faithfully describe hiv fitness landscapes. *Physical Review E*, 88(6):062705.

[Shih et al., 2007] Shih, A. C.-C., Hsiao, T.-C., Ho, M.-S., and Li, W.-H. (2007). Simultaneous amino acid substitutions at antigenic sites drive influenza a hemagglutinin evolution. *Proceedings of the National Academy of Sciences*, 104(15):6283–6288.

[Skehel et al., 1984] Skehel, J., Stevens, D., Daniels, R., Douglas, A., Knossow, M., Wilson, I., and Wiley, D. (1984). A carbohydrate side chain on hemagglutinins of hong kong influenza viruses inhibits recognition by a monoclonal antibody. *Proceedings of the National Academy of Sciences*, 81(6):1779–1783.

[Smith et al., 2004] Smith, D. J., Lapedes, A. S., De Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D., and Fouchier, R. A. (2004). Mapping the antigenic and genetic evolution of influenza virus. *science*, 305(5682):371–376.

[Sohail et al., 2020] Sohail, M. S., Louie, R. H., McKay, M. R., and Barton, J. P. (2020). Mpl resolves genetic linkage in fitness inference from complex evolutionary histories. *Nature Biotechnology*, pages 1–8.

[Squires et al., 2012] Squires, R. B., Noronha, J., Hunt, V., García-Sastre, A., Macken, C., Baumgarth, N., Suarez, D., Pickett, B. E., Zhang, Y., Larsen, C. N., et al. (2012). Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza and other respiratory viruses*, 6(6):404–416.

[Steel et al., 2010] Steel, J., Lowen, A. C., Wang, T. T., Yondola, M., Gao, Q., Haye, K., García-Sastre, A., and Palese, P. (2010). Influenza virus vaccine based on the conserved hemagglutinin stalk domain. *MBio*, 1(1).

[Strauch et al., 2017] Strauch, E.-M., Bernard, S. M., La, D., Bohn, A. J., Lee, P. S., Anderson, C. E., Nieusma, T., Holstein, C. A., Garcia, N. K., Hooper, K. A., et al. (2017). Computational design of trimeric influenza-neutralizing proteins targeting the hemagglutinin receptor binding site. *Nature Biotechnology*, 35(7):667–671.

[Throsby et al., 2008] Throsby, M., van den Brink, E., Jongeneelen, M., Poon, L. L., Alard, P., Cornelissen, L., Bakker, A., Cox, F., van Deventer, E., Guan, Y., et al. (2008). Heterosubtypic neutralizing monoclonal antibodies cross-protective against h5n1 and h1n1 recovered from human igm+ memory b cells. *PloS one*, 3(12):e3942.

[Tsimring et al., 1996] Tsimring, L. S., Levine, H., and Kessler, D. A. (1996). Rna virus evolution via a fitness-space model. *Physical review letters*, 76(23):4440.

[Whittle et al., 2011] Whittle, J. R., Zhang, R., Khurana, S., King, L. R., Manischewitz, J., Golding, H., Dormitzer, P. R., Haynes, B. F., Walter, E. B., Moody, M. A., et al. (2011). Broadly neutralizing human antibody that recognizes the receptor-binding pocket of influenza virus hemagglutinin. *Proceedings of the National Academy of Sciences*, 108(34):14216–14221.

[Wiley et al., 1981] 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.

[Wu et al., 2020] Wu, N. C., Otwinowski, J., Thompson, A. J., Nycholat, C. M., Nourmohammad, A., and Wilson, I. A. (2020). Major antigenic site b of human influenza h3n2 viruses has an evolving local fitness landscape. *Nature communications*, 11(1):1–10.

[Wu et al., 2018] Wu, N. C., Thompson, A. J., Xie, J., Lin, C.-W., Nycholat, C. M., Zhu, X., Lerner, R. A., Paulson, J. C., and Wilson, I. A. (2018). A complex epistatic network limits the mutational reversibility in the influenza hemagglutinin receptor-binding site. *Nature communications*, 9(1):1–13.

[Wu et al., 2017] Wu, N. C., Xie, J., Zheng, T., Nycholat, C. M., Grande, G., Paulson, J. C., Lerner, R. A., and Wilson, I. A. (2017). Diversity of functionally permissive sequences in the receptor-binding site of influenza hemagglutinin. *Cell Host & Microbe*, 21(6):742–753.

[Yamayoshi et al., 2017] Yamayoshi, S., Uraki, R., Ito, M., Kiso, M., Nakatsu, S., Yasuhara, A., Oishi, K., Sasaki, T., Ikuta, K., and Kawaoka, Y. (2017). A broadly reactive human anti-hemagglutinin stem monoclonal antibody that inhibits influenza a virus particle release. *EBioMedicine*, 17:182–191.

[Yan et al., 2019] Yan, L., Neher, R. A., and Shraiman, B. I. (2019). Phylodynamic theory of persistence, extinction and speciation of rapidly adapting pathogens. *Elife*, 8:e44205.

[Yassine et al., 2015] Yassine, H. M., Boyington, J. C., McTamney, P. M., Wei, C.-J., Kanekiyo, M., Kong, W.-P., Gallagher, J. R., Wang, L., Zhang, Y., Joyce, M. G., et al. (2015). Hemagglutinin-stem nanoparticles generate heterosubtypic influenza protection. *Nature medicine*, 21(9):1065–1070.

[Zhang et al., 2017] Zhang, Y., Aevermann, B. D., Anderson, T. K., Burke, D. F., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C. N., Lee, A. J., et al. (2017). Influenza research database: An integrated bioinformatics resource for influenza virus research. *Nucleic acids research*, 45(D1):D466–D474.