Inferring the intrinsic mutational fitness landscape of influenzalike evolving antigens from temporally ordered sequence data

Julia Doelger

Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Mehran Kardar*

Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Arup K. Chakraborty†

Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;
Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;
Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;
Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;
and Ragon Institute of MGH, MIT, and Harvard, Cambridge, Massachusetts 02139, USA

(Received 28 July 2021; accepted 19 January 2022; published 2 February 2022)

There still are no effective long-term protective vaccines against viruses that continuously evolve under immune pressure such as seasonal influenza, which has caused, and can cause, devastating epidemics in the human population. To find 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 fitness landscape of the viral proteins (i.e., knowledge of the viral fitness as a function of sequence). Here we present a computational method to infer the intrinsic mutational fitness landscape of influenzalike evolving antigens from yearly sequence data. We test inference performance with computer-generated sequence data that are based on stochastic simulations mimicking basic features of immune-driven viral evolution. Although the numerically simulated model does create a phylogeny based on the allowed mutations, the 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 a sufficient number of 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 antigenic proteins. Our fitness inference approach may have potential future use for the design of immunization protocols by identifying intrinsically vulnerable immune target combinations on antigens that evolve under immune-driven selection. In the future, this approach may be applied to influenza and other novel viruses such as SARS-CoV-2, which evolves and, like influenza, might continue to escape the natural and vaccine-mediated immune pressures.

DOI: 10.1103/PhysRevE.105.024401

I. INTRODUCTION

Global seasonal influenza epidemics are caused by influenza A and B viruses that although being effectively targeted by natural immune responses, seasonal vaccination responses, as well as long-term immune memory, are able to persistently escape population-wide immunity via mutations [1]. The dominantly targeted antigen of the influenza virus is the glycoprotein HA that is located on the viral surface together with the other surface glycoprotein NA, which also acts as an 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 five 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 [2,3]. 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 [4].

These interlinked dynamics of the mutating virus and responding human immunity cause a gradual evolution of the viral antigens that is known as antigenic drift [5], which leads to characteristic strain succession patterns in seasonal influenza (Fig. 1). Each unique sequence arises at a specific time and persists for a small number of years in the population before being replaced by newer strains. Every year, a relatively small number of strains is observed, although the total number of strains grows rapidly. 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

[*] Corresponding author: kardar@mit.edu

[†] Corresponding author: arupc@mit.edu

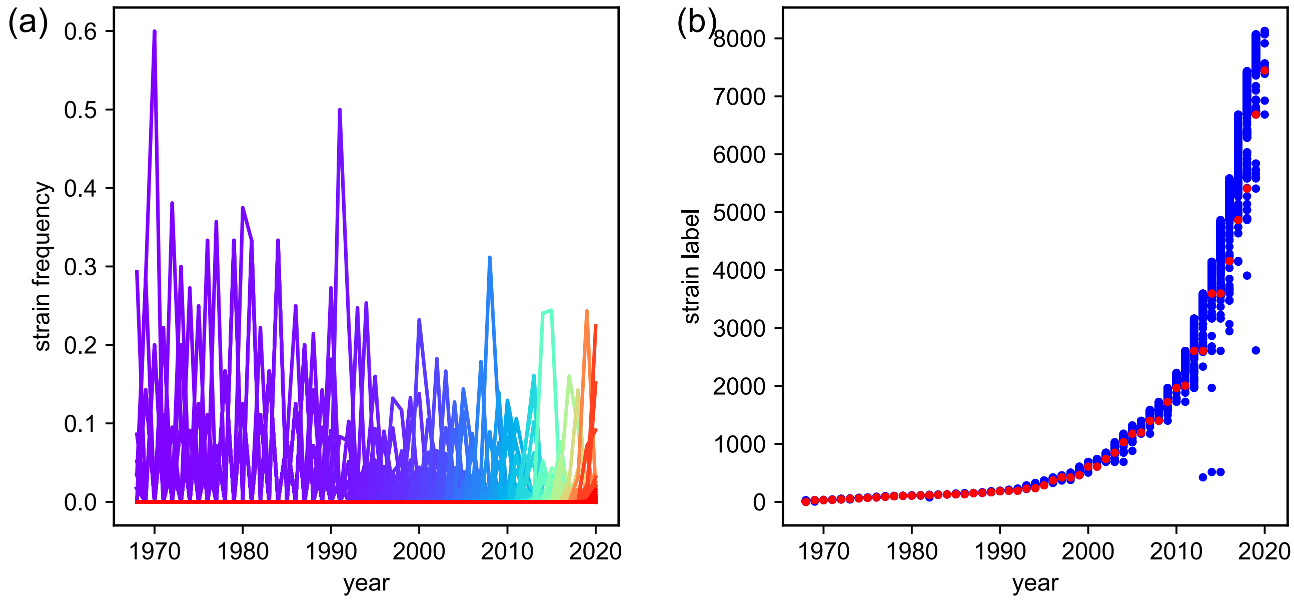

FIG. 1. Strain succession for the evolution of HA (H3N2) sequences between 1968 and 2020. (a) Each unique HA amino acid sequence (strain) is shown with its observed frequency in each year as a solid line, with line colors ranging from purple (old strains) to red (new strains). (b) Strains are labeled 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 a red circle. Blue circles indicate strains that were observed with some nonzero frequency. The apparent increase in the number of observed sequences at later years is likely an artifact of sampling more sequences, rather than indicative of increasing diversification of lineages.

infection [6]. 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 the 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 recommendations on the composition of the influenza virus vaccine [7], but many seasonal vaccines still have a low efficacy 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 [5,8–13]. 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 protection of the population [14].

Other approaches aim for cross-protective influenza treatments that are effective against a wide range of strains. Such approaches typically consider strongly conserved epitopes such as the receptor binding site (RBS) or the stem of HA [15–24]. Methods targeting those regions require specialized methods for sophisticated vaccine protocols and drug designs [25–35].

Easily accessible sites on highly mutable virus antigens, e.g., on the head of HA in the case of influenza, can generally quickly escape human immune memory via amino acid substitution. However, mutations at some of those strongly targeted sites will be functionally more costly to the virus than others. For a long-term protective immunization approach, it therefore would be useful to find and target primarily those easily accessible sites on viral antigens 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 multiclonal 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 [36–47]. 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 [48]. In fact, similar inverse statistical physics models have been extensively used in various contexts to learn from multiple sequence alignments about the structure and function of various pathogenic and human proteins [49].

Seasonal influenza, however, evolves very differently in the human population than viruses such as HIV, for which maximum entropy-based fitness inference methods have been successful. Since influenza is targeted by a population-wide immune memory, it is permanently driven away from past strains as opposed to HIV, which evolves much more freely within its fitness landscape and is able to periodically revisit old strains [10,38,50]. This immune-driven, nonequilibrium 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 other viruses.

For influenzalike evolving viruses, in which the population-wide immune memory against each emerging mutant accumulates with every season, the effective fitness landscape depends on the viral evolutionary history and therefore changes in time. Such a changing fitness landscape has also been referred to as a "seascape" [51,52]. This time variance of the effective fitness landscape makes approaches that rely on conserved fitness landscapes, such as the recently proposed marginal path likelihood (MPL) method [53], generally unsuitable for fitness inference from sequence time series of influenzalike evolving antigens.

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 influenzalike evolving pathogen. We test our inference approach on sequence data generated by computer simulations and propose its potential application in the future to investigate yearly protein sequence time series data from influenzalike evolving viruses, in order to obtain combinations of vulnerable antibody targets.

II. MODEL OF INFLUENZA ANTIGEN EVOLUTION

In our model for influenzalike evolution, we consider each 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 Sec. III. Our influenza evolution model is motivated and inspired by several previous modeling studies, which describe the essential properties of the evolution of influenzalike pathogen populations that lead to the characteristic spindlelike phylogeny and strain succession pattern of seasonal influenza [9,10,54–56]. 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 [57–59].

A. Sequence representation

For the representation of viral strains, we use a binary sequence representation, in which a strain $\mathbf{S}_j = (s_j^1, s_j^2, \ldots, s_j^L)$, 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 (such as the p24 structural protein of HIV), a binary Ising-like representation, instead of a Potts model, is reasonable [39]. Also, our approach could be generalized to Potts models. Here, we consider sequences of $L < 100$, which are much shorter than real protein sequences.

B. 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 ability of a particular virus strain (with a specific sequence) to infect, reproduce, and transmit in a susceptible human population. We assume that this intrinsic landscape does not change as the host immunity evolves over time. This is tantamount to assuming that the basic functions of the viral proteins remain the same over the timescales of interest. 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

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

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

1. Intrinsic fitness model

The intrinsic fitness of a strain in our model is represented by a two-point approximation as

$$F_{\text{int}}(\mathbf{S}_j) = F_0 + \sum_{\alpha=1}^{L} h_\alpha s_j^\alpha + \sum_{\alpha<\beta} J_{\alpha\beta} s_j^\alpha s_j^\beta. \qquad (2)$$

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 $2^L = 104\,857\,6$ 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 such as Eq. (2). We use this formulation for convenience and for demonstrating our method.

2. 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

[9,54,56], i.e.,

$$F_{\mathrm{host}}[\mathbf{S}_j, \mathbf{x}(t' < t)] = -\sigma_{\mathrm{h}} \sum_{t'<t} \sum_{i=1}^{M=2^L} x(\mathbf{S}_i, t') \exp\left(-|\mathbf{S}_j^{\mathrm{ep}} - \mathbf{S}_i^{\mathrm{ep}}|/D_0\right). \quad (3)$$

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, $|\mathbf{S}_j^{\mathrm{ep}} - \mathbf{S}_i^{\mathrm{ep}}|$ describes the mutational distance between strains $\mathbf{S}_i$ 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 other). In the following, we will assume for simplicity that all modeled sequence sites are equally immune targeted and therefore $\mathbf{S}_j^{\mathrm{ep}} = \mathbf{S}_j$, but the model can, in principle, be extended to account for less or untargeted sites in the model sequence $\mathbf{S}_j$.

In this model, for fitness cost accumulation, it is assumed that immunity against each particular strain lasts forever. In fact, human immune memory against influenza strains seems to be able to persist for many decades [60]. However, if there is evidence for a particular memory duration or decay function, it is straightforward to include this in our model.

C. 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.,

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

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

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

where $x_{\mathrm{m}}(\mathbf{S}_j, t)$ 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.

D. 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 $\mathbf{x}_{\mathrm{m}}(t)$. We assume one symmetric mutation rate μ, per season, between the two different states at each site, such that the mutation probability $\mu_{ij} = \mu_{ji}$ for mutation between strains $\mathbf{S}_i$ and $\mathbf{S}_j$ is given as

$$\mu_{ij} = \mu^{|\mathbf{S}_i - \mathbf{S}_j|}(1-\mu)^{L-|\mathbf{S}_i - \mathbf{S}_j|}. \quad (6)$$

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 model is to generate a controlled dataset that can be used to develop and test a method for inferring the intrinsic mutational fitness landscape of influenzalike 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.

III. ANALYSIS AND INFERENCE BASED ON SIMULATED SEQUENCE DATA

A. Simulation produces influenzalike antigen evolution

Based on the presented model, we ran stochastic simulations to compare the computer-generated sequence evolution 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. Table I). In the beginning of each simulation, the population is initialized with the unmutated strain $\mathbf{S}_0 = (0, 0, \ldots, 0)$. 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 [39]. In each time step representing one epidemic season, sequences are first mutated according to rate μ. After mutation, the 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 immune-driven evolutionary 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 implies that our model is able to capture the essential dynamics of antigenic evolution for pathogens such as seasonal influenza. This also indicates that the inference approach, which we develop with the help of simulated data, can in principle be applied to influenza sequence data. One difference in the shown figures [Figs. 1(b)

TABLE I. Parameters for simulation of influenzalike 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
$N_{\rm pop}$	population size	10^5
$\sigma_{\rm h}$	host-fitness coefficient	1
$\{\lambda_h, \lambda_J, \lambda_{F^*}\}$	regularization coefficients for inference	$\{10^{-4}, 1, 10^{-4}\}$
$n_{\rm seasons}$	number of seasons used for inference	100
B	number of sampled sequences per season	10^3

and 2(b)] 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 the observed influenza sequences may be due to the rapid increase in the amount of yearly acquired sequencing data in the past years.

B. Observation of stringent selection regime

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 narrower 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

$$F_{\rm total}\{\mathbf{S}_j[t], \mathbf{x}(t' < t)\} \approx F[t, \mathbf{x}(t' < t)], \qquad (7)$$

with $F[t, \mathbf{x}(t' < t)]$ being dependent on time t but independent of sequence identity $\mathbf{S}$, given the specific evolutionary history $\mathbf{x}(t' < t)$. Here, $\mathbf{S}_j[t]$ denotes a sequence of identity $\mathbf{S}_j$ that is actually selected at time t.

Indeed, we find with our simulation a clear 1:1 correspondence between the intrinsic fitness variation and immunity-

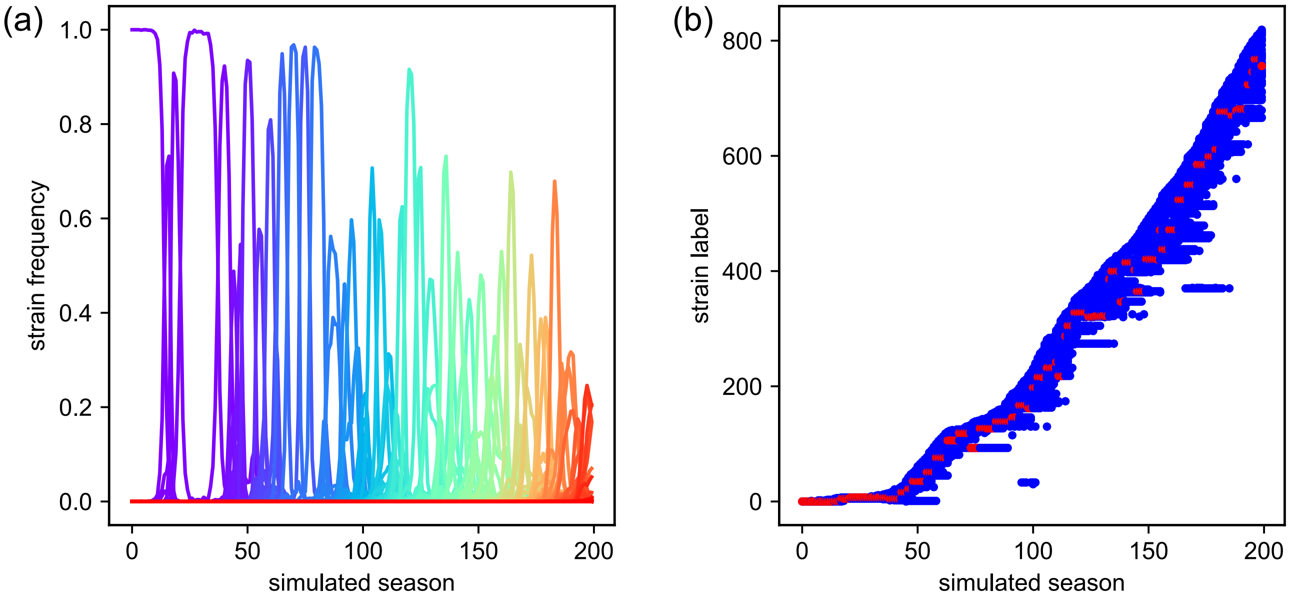

FIG. 2. Strain succession for the evolution of simulated data over 200 time steps. (a) Each unique sequence (strain) is shown with its observed frequency in each simulated season as a solid line, with line colors ranging from purple (old strains) to red (new strains). (b) Strains are labeled with increasing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

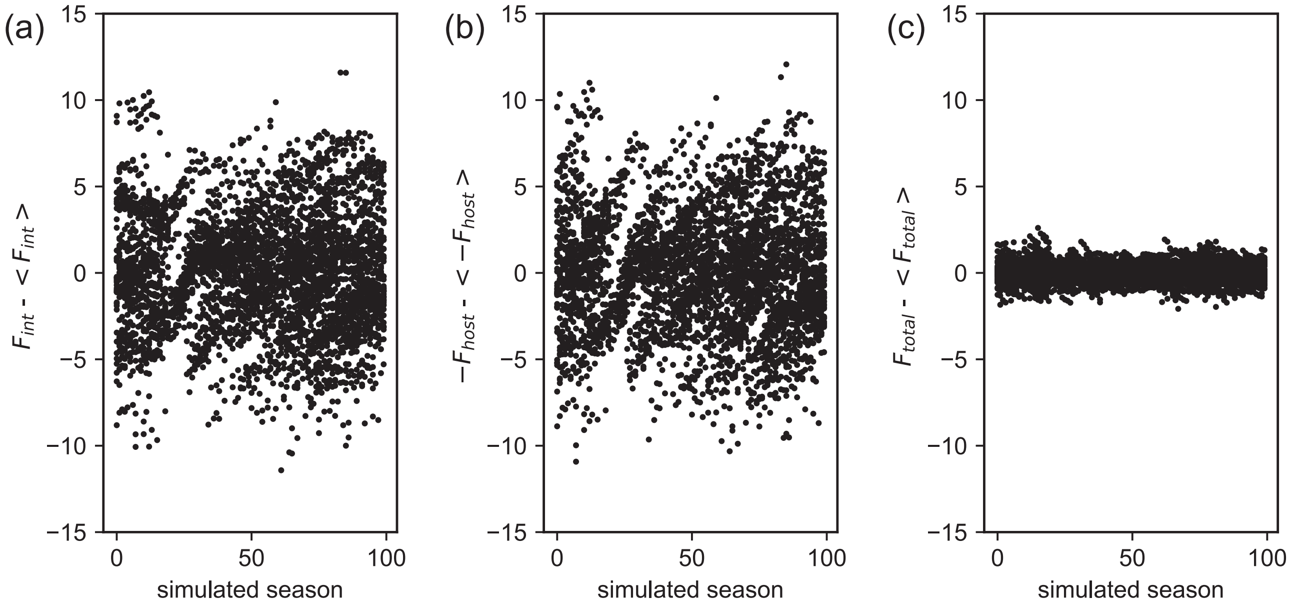

FIG. 3. Fitness deviations from the mean of sampled strains for each simulated season between seasons 100 and 200. (a) Intrinsic fitness component F_{int}, (b) immunity-dependent fitness component F_{host}, (c) total fitness $F_{\text{total}} = F_{\text{int}} + F_{\text{host}}$. For the shown example,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

dependent fitness variation in each season [Fig. 4(a)], which add up to a roughly constant total fitness in each season, as the stringency assumption [Eq. (7)] suggests. In Fig. 4(b), it can be observed that in our simulation, the absolute population fitness decreases with each year, due both to the emergence of less intrinsically fit strains and to the population-wide accumulation of immune pressure.

As for the evolution of influenza in the human population, its seasonal dynamics has been well described with traveling-wave models, which indicate a localized, narrow distribution of the viral population in fitness space at any given time point [57–59,61]. Such a narrow fitness distribution of concurrently selected viral antigenic sequences indicates that one necessary condition [Eq. (7)] for our stringency-based inference method may be fulfilled by seasonal influenza antigens.

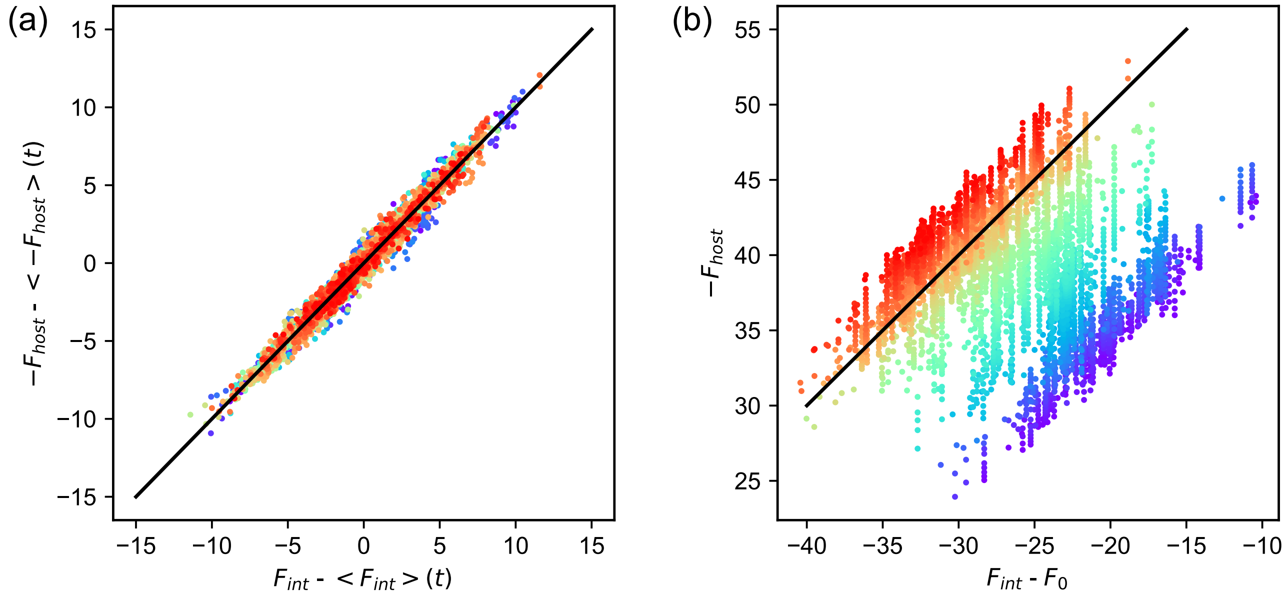

FIG. 4. Negative immunity-dependent fitness cost $-F_{\text{host}}$ (y axes) compared to intrinsic fitness F_{int} (x axes) for sampled strains for each simulated season between seasons 100 and 200 (same data as in Fig. 3). (a) Fitness deviations from the mean as colored circles with a solid black line 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.

C. Method for intrinsic fitness inference

From Eq. (1) together with Eq. (7), we obtain the following relation for the observed strains $\mathbf{S}_j[t]$ in each given year t in the case of stringent selection:

$$-F_{\text{host}}\{\mathbf{S}_j[t], \mathbf{x}(t' < t)\} \approx \sum_{\alpha} h_\alpha s_j^\alpha + \sum_{\alpha<\beta} J_{\alpha\beta} s_j^\alpha s_j^\beta + F^*[t, \mathbf{x}(t' < t)], \quad (8)$$

where $F^*[t, \mathbf{x}(t' < t)] = F_0 - F[t, \mathbf{x}(t' < t)]$ is a different constant at each 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 [9], we can calculate $F_{\text{host}}\{\mathbf{S}_j[t], \mathbf{x}(t' < t)\}$ for each observed strain in each season. We now use these host-dependent fitness values together with Eq. (8) to infer the intrinsic fitness coefficients $\{h, J\}$ as well as the additional parameters $\{F^*\}$ (one additional parameter per season). Here we treat $\{F^*\}$ as independent parameters, although they generally depend on other model parameters and on the history via the full evolutionary dynamics of the system. For the regression, we minimize the sum of squared residuals between the data,

$$Y_{\text{data}}(\mathbf{S}_j[t], t) = -F_{\text{host}}\{\mathbf{S}_j[t], \mathbf{x}(t' < t)\} = \sigma_{\text{h}} \sum_{t'<t} \sum_{i=1}^{M=2^L} x(\mathbf{S}_i, t') \exp\left(-|\mathbf{S}_j^{\text{ep}} - \mathbf{S}_i^{\text{ep}}|/D_0\right), \quad (9)$$

which are given by the left-hand side of Eq. (8), and the model function

$$Y_{\text{model}}(\mathbf{S}_j[t], t, \{h, J, F^*\}) = \sum_{\alpha} h_\alpha s_j^\alpha + \sum_{\alpha<\beta} J_{\alpha\beta} s_j^\alpha s_j^\beta + F_t^*, \quad (10)$$

which is given by the right-hand side of Eq. (8), i.e.,

$$\{h, J, F^*\} = \arg\min_{\{h,J,F^*\}} \left[\frac{1}{2}\sum_t \sum_j \{Y_{\text{data}}(\mathbf{S}_j[t], t) - Y_{\text{model}}(\mathbf{S}_j[t], t, \{h, J, F^*\})\}^2 + \frac{\lambda_h}{2}\sum_\alpha h_\alpha^2 + \frac{\lambda_J}{2}\sum_{\alpha<\beta} J_{\alpha\beta}^2 + \frac{\lambda_{F^*}}{2}\sum_{t'} F_{t'}^{*2}\right], \quad (11)$$

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 [62, Eq. (3.44)]:

$$\mathbf{M} = (\mathbf{X}^T\mathbf{X} + \mathbf{\Lambda})^{-1}\mathbf{X}^T\mathbf{y} \quad (12)$$

to solve for the unique parameter values $\mathbf{M} = (h_1, \ldots, h_L, J_1, \ldots, J_{L*(L-1)/2}, F_1^*, \ldots, F_{n\,\text{seasons}}^*)^T$, which minimize the sum of squared residuals subject to ridge regularization [Eq. (11)]. 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 $-F_{\text{host}}\{\mathbf{S}_j[t], \mathbf{x}(t' < t)\}$ [cf. Eq. (3)] for the respective sequence $\mathbf{S}_j[t]$, sampled at time t. The nonzero 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 nonrestrictive, prior distribution, while λ_J, corresponding to the assumed sparse mutational couplings, is set to 1.

D. Inferring the intrinsic mutational fitness landscape from simulated influenzalike sequence data

The parameters for simulation and inference with chosen default values are collected in Table I.

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

$$h_\alpha + h_\beta + J_{\alpha\beta} < F_{\text{threshold}} < 0. \quad (13)$$

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) 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 $\geqslant 10^5$ strains is required for high inference performance [Fig. 7(b)]. 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.

The inference performance further strongly depends on the sequence length L [Fig. 8(a)] as well as on the population size

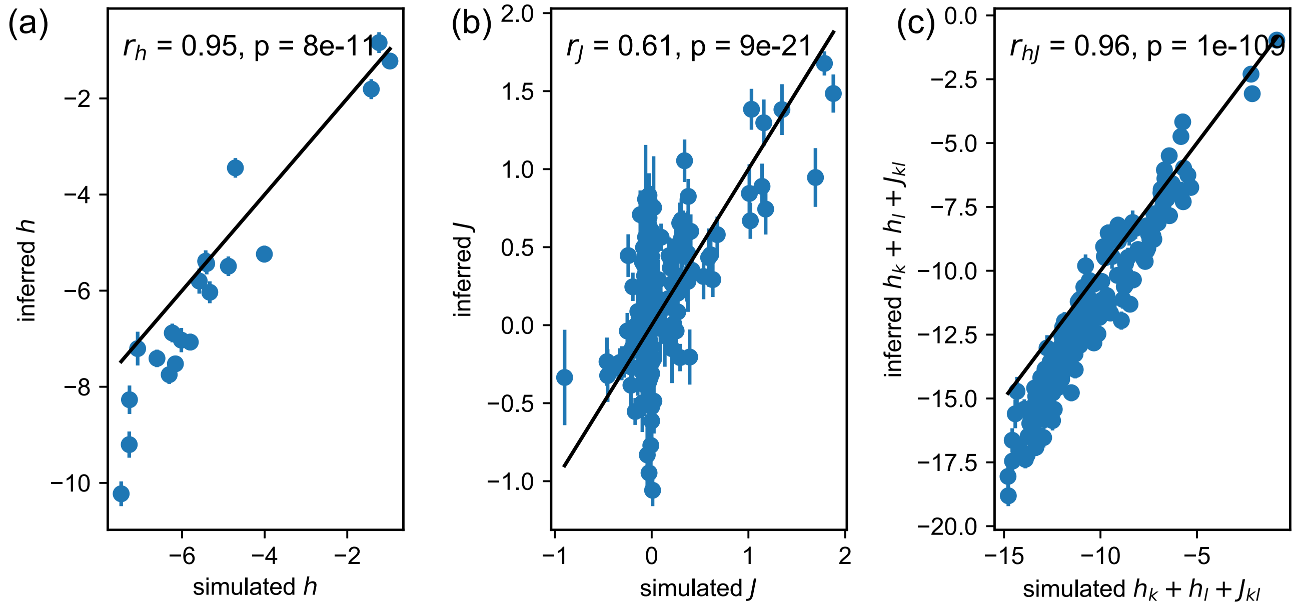

FIG. 5. Parameter correlations for the inference on one simulated data set. Inferred values of the fitness coefficients are shown against the fitness coefficients that were used as input values for the simulation. (a) Single-site mutational fitness coefficients h, (b) coupling coefficients J for simultaneous 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 $N_{\rm pop} = 10^5$, $L = 20$, $\mu = 10^{-4}$, $\sigma_{\rm h} = 1$, $D_0 = 5$, $N_{\rm simu} = 200$, $B = 10^3$, $n_{\rm seasons} = 100$, $\lambda_h = 10^{-4}$, $\lambda_J = 1$, and $\lambda_{F*} = 10^{-4}$.

$N_{\rm pop}$ [Fig. 8(b)]. Inference performance in terms of the correlation $r_{\rm hJ}$ between inferred and simulated double-mutational fitness coefficients decreases with increasing sequence length and increases with increasing population size towards an upper limit $\leqslant$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, second, if the effective population size that defines the selection bottleneck is small, inference can be poor. A large population size, however, will not contribute to high inference performance if the sample size B is low.

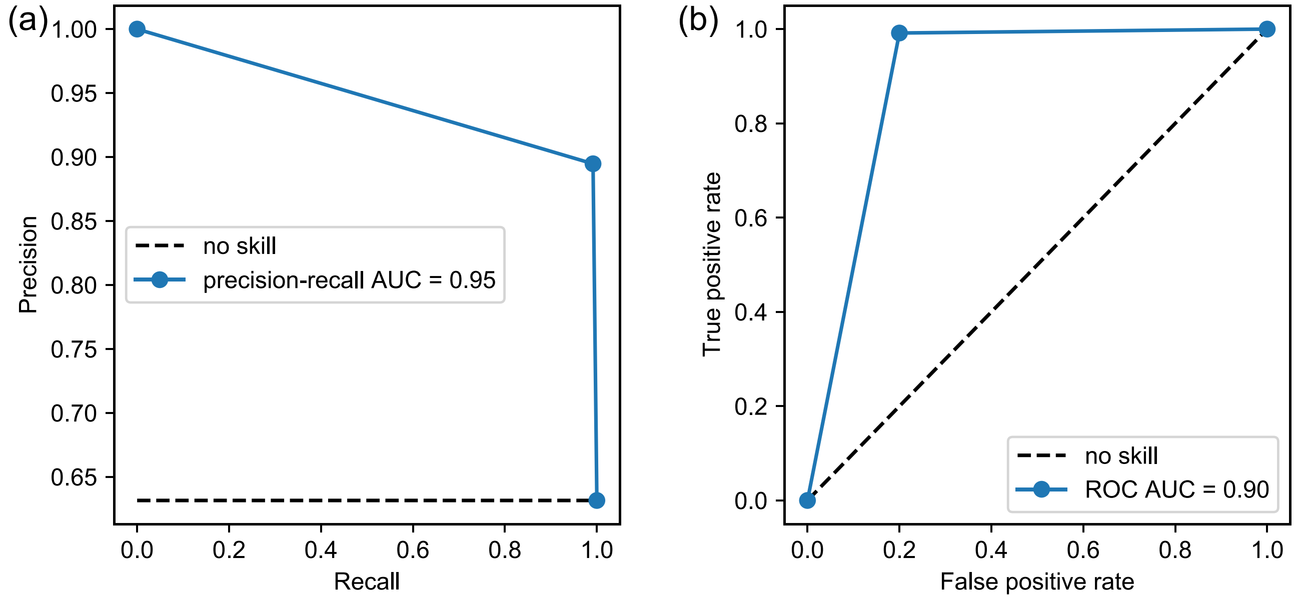

FIG. 6. Classification performance for the inference on one simulated data set. Double mutations are classified as deleterious if their total fitness cost is lower than $F_{\rm threshold} = -10$ [cf. Eq. (13)]. (a) The precision-recall curve (PRC) and (b) the ROC curve for the classifier derived from inferred fitness coefficients. Black 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 $N_{\rm pop} = 10^5$, $L = 20$, $\mu = 10^{-4}$, $\sigma_{\rm h} = 1$, $D_0 = 5$, $N_{\rm simu} = 200$, $B = 10^3$, $n_{\rm seasons} = 100$, $\lambda_h = 10^{-4}$, $\lambda_J = 1$, $\lambda_{F*} = 10^{-4}$, and $F_{\rm threshold} = -10$

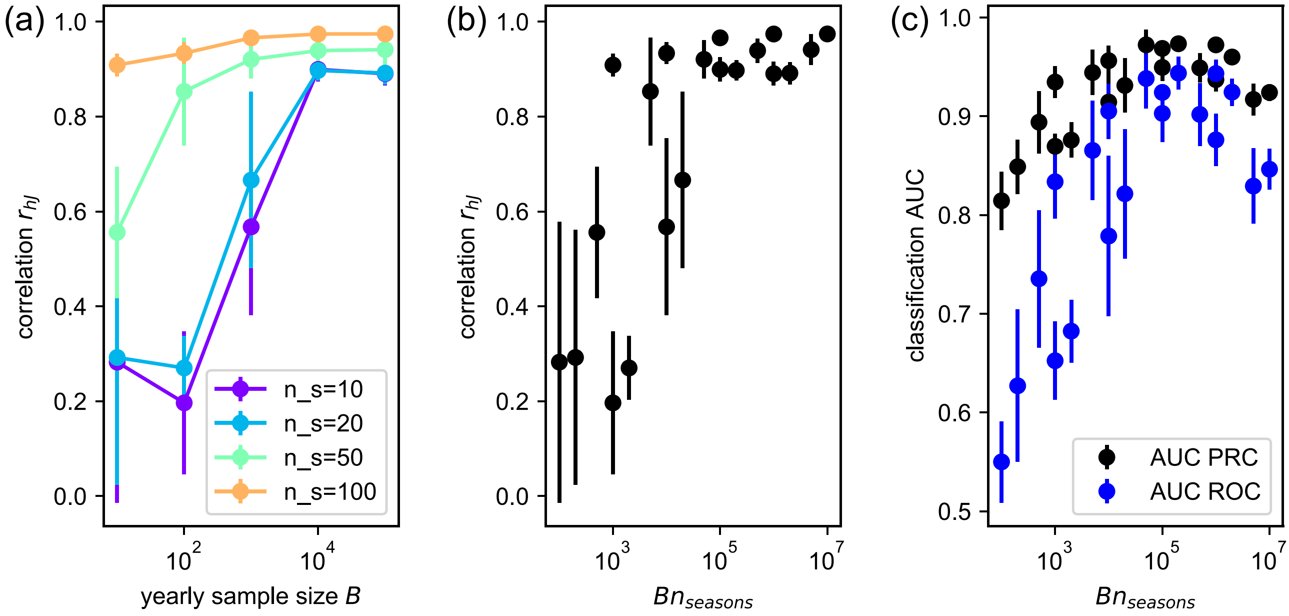

FIG. 7. Inference performance for varying yearly sample size B per season and varying number n_{seasons} of seasons used for inference. (a) The correlation coefficient r_{hJ} between inferred and simulated double-mutational fitness costs as a function of yearly sample size B for various n_{seasons}. (b) The correlation coefficient r_{hJ} as a function of total sample size, Bn_{seasons}. (b) The area (AUC) under the ROC curve and under the precision-recall curve (PRC) for the classification of deleterious double mutations with classification threshold $F_{\text{threshold}} = -10$, shown as a function of total sample size Bn_{seasons}. Each value is averaged over six simulations, respectively, and error bars show the respective sample standard deviations. For the shown example, the fixed parameter values for simulation and analysis are $N_{\text{pop}} = 10^5$, $L = 20$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D_0 = 5$, $N_{\text{simu}} = 200$, $\lambda_h = 10^{-4}$, $\lambda_J = 1$, and $\lambda_{F^*} = 10^{-4}$.

The five epitope regions of influenza HA contain 131 amino acid sites [2,63]. If we represent this antigenic region with a binary representation, i.e., 1 for being mutated, 0 for not being mutated compared to a reference sequence, we remain with a sequence length of $L = 131$. On the other hand, the total number of H3N2 sequences that have been collected

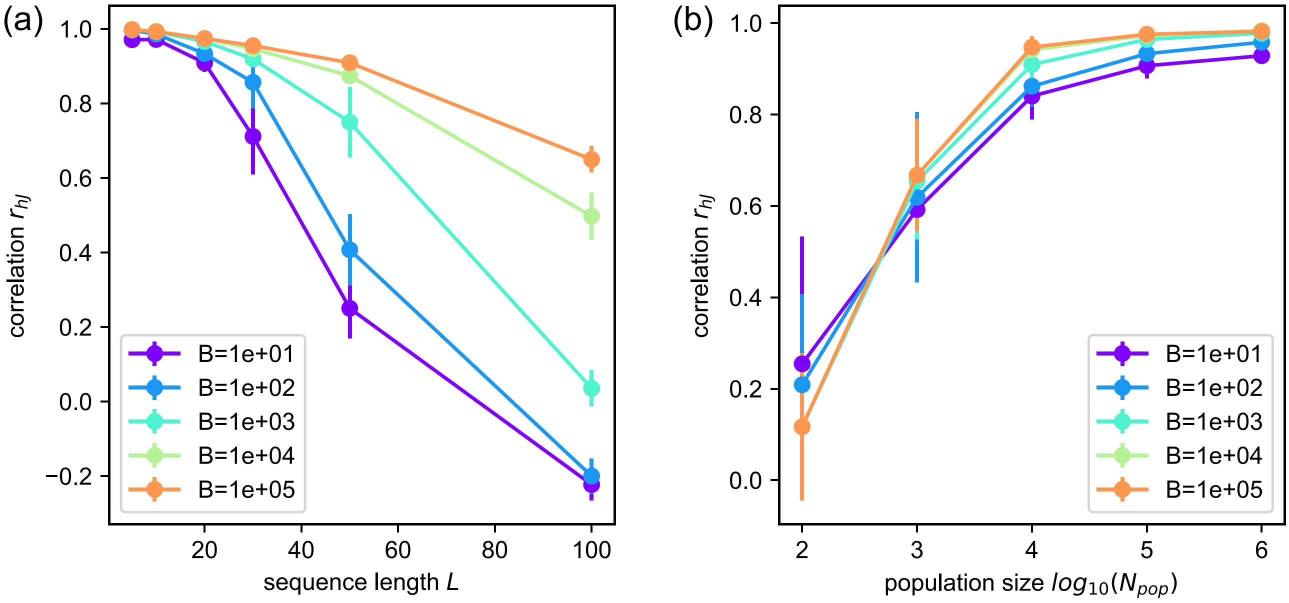

FIG. 8. Inference performance in terms of the correlation coefficient r_{hJ} between inferred and simulated double-mutational fitness costs, for varying simulation and analysis parameters. (a) Inference performance as a function of sequence length L, (b) inference performance as a function of population size N_{pop}. For both parameter explorations, the yearly sample size B was varied between 10 and 10^5. Each value is averaged over six simulations, respectively, and error bars show the respective sample standard deviations. For the shown simulation results, the respective fixed parameter values for simulation and analysis are $N_{\text{pop}} = 10^5$, $L = 20$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D_0 = 5$, $N_{\text{simu}} = 200$, $n_{\text{seasons}} = 100$, $\lambda_h = 10^{-4}$, $\lambda_J = 1$, and $\lambda_{F^*} = 10^{-4}$.

since 1968 is of the order of $Bn_{\mathrm{seasons}} \sim 10^4$, which is less data than minimally required for accurate inference on simulated sequences of shorter length [Figs. 7(b) and 8(a)]. Therefore, we do not expect a significant inference performance with our method applied to the currently available influenza sequence data. Nevertheless, we have conducted an inference trial on the HA epitope sequence data, the code and results of which can be found on GitHub [64]. This trial confirmed our expectations of insufficient inference performance by comparison with deep-mutational scanning measurements from Lee *et al.* (2018) [65].

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 $m = n_{\mathrm{seasons}} + L(L+1)/2 \approx n_{\mathrm{seasons}} + L^2/2$ parameters. With the simplifying assumption that we need m independent equations for this inference task, we can estimate that we approach optimal inference performance when $B\,n_{\mathrm{seasons}}\,\mu \sim m$. Here the number of needed total samples Bn_{seasons} is assumed to increase with decreasing mutation rate μ, since an independent set of samples is obtained only roughly every $1/\mu$ years.

As more sequences become available in the future, we expect our method to become useful for real data for viruses under strong selection pressure due to human immunity, which could soon include SARS-CoV-2. Another interim approach that we are beginning to consider is to coarse grain these 131 residues into groups, and then infer the fields and couplings for the groups. Then, we successively add more residues to the groups between which large couplings exist, and carry out this procedure iteratively. But, such a study is beyond the scope of this paper.

IV. DISCUSSION

Here we presented a method for inferring the intrinsic mutational fitness landscape of influenzalike 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 influenzalike 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 a function of time and does not depend on a separate inference of sequence phylogenies, in contrast to other analyses [9,10].

In comparison to the recently proposed marginal path likelihood method (MPL) for sequence time series [53], 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. Although there have been previous studies inferring double-mutational fitness effects from influenza protein sequences [66,67], those approaches do not generally attempt to systematically decouple intrinsic fitness effects from time-varying, immunity-dependent effects. These studies further tend to focus on the detection of positive epistasis, i.e., more likely than neutral double mutations, that are generally more easily detected than negative epistatic fitness effects, which lead to more rarely observed mutation events. Detecting positive fitness effects is important for predicting sequence mutations that are most likely to evolve in the future. In this study, however, we are especially interested in pairwise mutations that incur a large negative mutational fitness effect since those might point towards effective immunization targets.

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 100 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 [68,69].

For using our inference approach on the influenza protein data, one further needs to make sure that the cross-immunity function in $-F_{\mathrm{host}}$ [Eq. (3)], which we use as the 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 [9] for estimating cross immunity, only roughly captures the cross-immunity measurements from hemagglutination inhibition (HI) assays [5,70]. Analysis of such HI data, in which the proposed cross-immunity function is compared to 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 [9,70], 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.

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, which are locally constrained to small parts of the sequence space or which only consider single-mutational fitness effects based on a given reference strain [65,71–73]. Furthermore, these empirical studies only measure fitness in terms of functional replication in cells, not in terms of spread 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.

Our computer simulations with a well-defined model of the evolution of a mutable virus subjected to human immune pressure over time have generated a data set of temporally ordered sequences. These sequences could be used in the future to test the veracity of different inference schemes against data that is the "ground truth." For example, do existing models developed for predicting the most likely influenza strains given data until the preceding year [9,10] give the right answers for the data set that we have generated?

In conclusion, we have proposed a method for inferring the intrinsic mutational fitness landscape of influenzalike viruses from time series of observed antigenic sequences. This approach can, with increasing availability of sequence data in the future, contribute to the development of new cross-

and long-term protective immunization strategies against viruses that evolve due to immune-driven selection. Like seasonal influenza, SARS-CoV-2 and other novel viruses might become endemic by evolving under vaccine and natural immune pressure, and our approach might provide valuable insight into their intrinsic fitness landscapes and reveal their vulnerabilities.

The computer code used for simulations and analyses as well as the data that were used to produce the figures in the paper are available on GitHub [64].

ACKNOWLEDGMENTS

This research was supported by the Ragon Institute of MGH, MIT, and Harvard and NIH Grant No. R01HL120724-01A1. M.K. acknowledges support by the NSF through Grants No. DMR-1708280 and No. PHY-2026995.

A.K.C. and M.K. initiated the project. J.D., M.K., and A.K.C. designed the research, analyzed the results, and wrote the paper. J.D. performed the calculations and generated all the figures.

The authors declare no competing interests.

[1] V. N. Petrova and C. A. Russell, The evolution of seasonal influenza viruses, Nat. Rev. Microbiol. **16**, 47 (2018).

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

[3] J. Skehel, D. Stevens, R. Daniels, A. Douglas, M. Knossow, I. Wilson, and D. Wiley, A carbohydrate side chain on hemagglutinins of Hong Kong influenza viruses inhibits recognition by a monoclonal antibody, Proc. Natl. Acad. Sci. USA **81**, 1779 (1984).

[4] W. Gerhard, J. Yewdell, M. E. Frankel, and R. Webster, Antigenic structure of influenza virus haemagglutinin defined by hybridoma antibodies, Nature (London) **290**, 713 (1981).

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

[6] F. Carrat and A. Flahault, Influenza vaccine: The challenge of antigenic drift, Vaccine **25**, 6852 (2007).

[7] WHO recommendations on the composition of influenza virus vaccines, https://www.who.int/influenza/vaccines/virus/recommendations/en/

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

[9] M. Łuksza and M. Lässig, A predictive fitness model for influenza, Nature (London) **507**, 57 (2014).

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

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

[12] C. Li, M. Hatta, D. F. Burke, J. Ping, Y. Zhang, M. Ozawa, A. S. Taft, S. C. Das, A. P. Hanson, J. Song *et al.*, Selection of antigenically advanced variants of seasonal influenza viruses, Nat. Microbiol. **1**, 16058 (2016).

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

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

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

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

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

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

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

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

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

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

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

[24] J. R. Whittle, R. Zhang, S. Khurana, L. R. King, J. Manischewitz, H. Golding, P. R. Dormitzer, B. F. Haynes, E. B. Walter, M. A. Moody *et al.*, Broadly neutralizing human antibody that recognizes the receptor-binding pocket of influenza

virus hemagglutinin, Proc. Natl. Acad. Sci. USA **108**, 14216 (2011).

[25] J. Steel, A. C. Lowen, T. T. Wang, M. Yondola, Q. Gao, K. Haye, A. García-Sastre, and P. Palese, Influenza virus vaccine based on the conserved hemagglutinin stalk domain, mBio **1**, e00018-10 (2010).

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

[27] Y. Lu, J. P. Welsh, and J. R. Swartz, Production and stabilization of the trimeric influenza hemagglutinin stem domain for potentially broadly protective influenza vaccines, Proc. Natl. Acad. Sci. USA **111**, 125 (2014).

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

[29] F. Krammer, N. Pica, R. Hai, I. Margine, and P. Palese, Chimeric hemagglutinin influenza virus vaccine constructs elicit broadly protective stalk-specific antibodies, J. Virol. **87**, 6542 (2013).

[30] R. Hai, F. Krammer, G. S. Tan, N. Pica, D. Eggink, J. Maamary, I. Margine, R. A. Albrecht, and P. Palese, Influenza viruses expressing chimeric hemagglutinins: Globular head and stalk domains derived from different subtypes, J. Virol. **86**, 5774 (2012).

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

[32] D. Eggink, P. H. Goff, and P. Palese, Guiding the immune response against influenza virus hemagglutinin toward the conserved stalk domain by hyperglycosylation of the globular head domain, J. Virol. **88**, 699 (2014).

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

[34] R. U. Kadam and I. A. Wilson, A small-molecule fragment that emulates binding of receptor and broadly neutralizing antibodies to influenza a hemagglutinin, Proc. Natl. Acad. Sci. USA **115**, 4240 (2018).

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

[36] V. Dahirel, K. Shekhar, F. Pereyra, T. Miura, M. Artyomov, S. Talsania, T. M. Allen, M. Altfeld, M. Carrington, D. J. Irvine *et al.*, Coordinate linkage of HIV evolution reveals regions of immunological vulnerability, Proc. Natl. Acad. Sci. USA **108**, 11530 (2011).

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

[38] K. Shekhar, C. F. Ruberman, A. L. Ferguson, J. P. Barton, M. Kardar, and A. K. Chakraborty, Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes, Phys. Rev. E **88**, 062705 (2013).

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

[40] J. P. Barton, N. Goonetilleke, T. C. Butler, B. D. Walker, A. J. McMichael, and A. K. Chakraborty, Relative rate and location of intrahost HIV evolution to evade cellular immunity are predictable, Nat. Commun. **7**, 11660 (2016).

[41] T. C. Butler, J. P. Barton, M. Kardar, and A. K. Chakraborty, Identification of drug resistance mutations in hiv from constraints on natural evolution, Phys. Rev. E **93**, 022412 (2016).

[42] A. K. Chakraborty and J. P. Barton, Rational design of vaccine targets and strategies for hiv: A crossroad of statistical physics, biology, and medicine, Rep. Prog. Phys. **80**, 032601 (2017).

[43] R. H. Louie, K. J. Kaczorowski, J. P. Barton, A. K. Chakraborty, and M. R. McKay, Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies, Proc. Natl. Acad. Sci. USA **115**, E564 (2018).

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

[45] A. A. Quadeer, J. P. Barton, A. K. Chakraborty, and M. R. McKay, Deconvolving mutational patterns of poliovirus outbreaks reveals its intrinsic fitness landscape, Nat. Commun. **11**, 377 (2020).

[46] S. Cocco and R. Monasson, Adaptive cluster expansion for inferring Boltzmann machines with noisy data, Phys. Rev. Lett. **106**, 090601 (2011).

[47] J. P. Barton, E. De Leonardis, A. Coucke, and S. Cocco, ACE: Adaptive cluster expansion for maximum entropy graphical model inference, Bioinformatics **32**, 3089 (2016).

[48] D. K. Murakowski, J. P. Barton, L. Peter, A. Chandrashekar, E. Bondzie, A. Gao, D. H. Barouch, and A. K. Chakraborty, Adenovirus-vectored vaccine containing multidimensionally conserved parts of the HIV proteome is immunogenic in rhesus macaques, Proc. Natl. Acad. Sci. USA **118**, e2022496118(2021).

[49] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Inverse statistical physics of protein sequences: A key issues review, Rep. Prog. Phys. **81**, 032601 (2018).

[50] S. Pompei, V. Loreto, and F. Tria, Phylogenetic properties of RNA viruses, PLoS One **7**, e44849 (2012).

[51] V. Mustonen and M. Lässig, From fitness landscapes to seascapes: Nonequilibrium dynamics of selection and adaptation, Trends Genet. **25**, 111 (2009).

[52] V. Mustonen and M. Lässig, Fitness flux and ubiquity of adaptive evolution, Proc. Natl. Acad. Sci. USA **107**, 4248 (2010).

[53] M. S. Sohail, R. H. Louie, M. R. McKay, and J. P. Barton, MPL resolves genetic linkage in fitness inference from complex evolutionary histories, Nat. Biotechnol. **39**, 472 (2021).

[54] J. R. Gog and B. T. Grenfell, Dynamics and selection of many-strain pathogens, Proc. Natl. Acad. Sci. USA **99**, 17209 (2002).

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

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

[57] L. S. Tsimring, H. Levine, and D. A. Kessler, RNA virus evolution via a fitness-space model, Phys. Rev. Lett. **76**, 4440 (1996).

[58] I. M. Rouzine, J. Wakeley, and J. M. Coffin, The solitary wave of asexual evolution, Proc. Natl. Acad. Sci. USA **100**, 587 (2003).

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

[60] X. Yu, T. Tsibane, P. A. McGraw, F. S. House, C. J. Keefer, M. D. Hicar, T. M. Tumpey, C. Pappas, L. A. Perrone, O. Martinez *et al.*, Neutralizing antibodies derived from the b cells of 1918 influenza pandemic survivors, Nature (London) **455**, 532 (2008).

[61] J. Marchi, M. Lässig, A. M. Walczak, and T. Mora, Antigenic waves of virus-immune co-evolution, J. Proc. Natl. Acad. Sci. **118**, e2103398118 (2021).

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

[63] Y. Suzuki, Natural selection on the influenza virus genome, Mol. Biol. Evol. **23**, 1902 (2006).

[64] See https://github.com/JDoelger/InfluenzaFitnessInference.git.

[65] J. M. Lee, J. Huddleston, M. B. Doud, K. A. Hooper, N. C. Wu, T. Bedford, and J. D. Bloom, Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants, Proc. Natl. Acad. Sci. USA **115**, E8276 (2018).

[66] S. Kryazhimskiy, J. Dushoff, G. A. Bazykin, and J. B. Plotkin, Prevalence of epistasis in the evolution of influenza a surface proteins, PLoS Genet. **7**, e1001301 (2011).

[67] N. Strelkowa and M. Lässig, Clonal interference in the evolution of influenza, Genetics **192**, 671 (2012).

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

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

[70] Antigenic cartography, www.antigenic-cartography.org.

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

[72] N. C. Wu, A. J. Thompson, J. Xie, C.-W. Lin, C. M. Nycholat, X. Zhu, R. A. Lerner, J. C. Paulson, and I. A. Wilson, A complex epistatic network limits the mutational reversibility in the influenza hemagglutinin receptor-binding site, Nat. Commun. **9**, 1264 (2018).

[73] N. C. Wu, J. Otwinowski, A. J. Thompson, C. M. Nycholat, A. Nourmohammad, and I. A. Wilson, Major antigenic site b of human influenza h3n2 viruses has an evolving local fitness landscape, Nat. Commun. **11**, 1233 (2020).