Inferring the intrinsic mutational fitness landscape of influenza-like evolving antigens from protein sequences

Julia Doelger, Mehran Kardar, Arup K. Chakraborty

April 28, 2021

1 Introduction

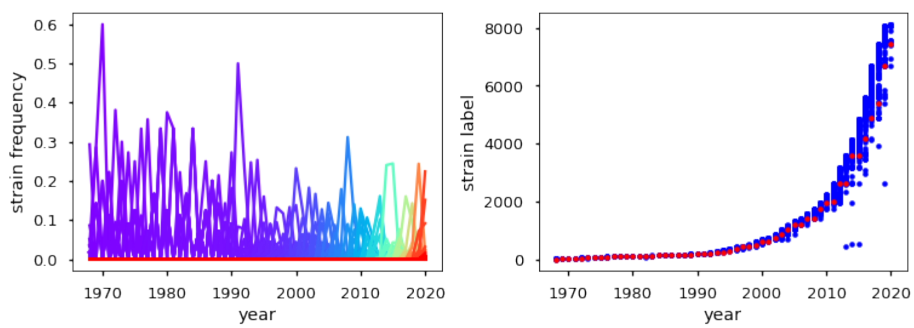

Figure 1: Strain succession for the evolution of HA (H3N2) sequences between 1968 and 2020. (left) Each unique HA sequence (strain) is shown with its observed frequency in each year as a solid line, with line colors ranging from purple (old strains) to red (new strains). (right) Strain labels here are counted 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.

Global seasonal influenza epidemics are caused by influenza viruses that, although being effectively targeted by human immune responses and long-term immune memory, are able to persistently escape this 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 and more prevalently represented than the other surface glycoprotein NA. 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]. These also 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].

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 universal and 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 recommendations 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 there are many computational and experimental efforts, which exclusively work on the task of analyzing and predicting the evolution of influenza antigenic sequences, in order to make 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 short-term predictions do generally not lead to long-term effective treatment plans [Paules et al., 2017].

Other approaches, aiming for cross-protective influenza treatments, consider strongly conserved epitopes like the receptor binding site (RBS) or the stem of HA which is shielded by the more easily targetable head of the protein [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]. These methods are complicated by the lower accessibility of the targeted regions to antibodies [Amitai, 2020], and specialized methods for sophisticated vaccine protocols and drug designs are needed to target these hidden protein regions [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, it is for such mutationally conserved sites often 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 targeted to a lesser extent than the classical HA head epitopes.

Although the easily accessible sites on the head of HA are found to generally quickly escape human immune memory via mutation, substitutions at some of the targeted sites will be 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. Additionally we can 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 each single and combined protein mutations at different sites is encoded in the intrinsic mutational fitness landscape of the viral sequence.

Previous studies were able to use equilibrium thermodynamics methods and an approach called Adaptive Cluster Expansion (ACE) to computationally infer the intrinsic mutational fitness landscape for other highly mutable viruses, HIV and 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 with multidimensionally conserved parts of the proteome, which is currently in clinical development [Murakowski et al., 2021]. However, the non-equilibrium nature of influenza evolution, which continuously evolves away from past strains, requires a different method for the inference of the intrinsic mutational fitness landscape.

Recently another fitness inference method, the so-called 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 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, since the population-wide immune-memory against each emerging mutant accumulates with every season.

Here we present a new 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 simulations and propose its application to investigate yearly HA sequence time series data from influenza A/H3N2 and to obtain vulnerable antibody targets on the HA head.

2 Model of influenza antigen evolution

In our model of influenza evolution, we assume that at the beginning of a flu season a population of viral units enters a human population. The distribution of unique antigenic sequences (strains) in the viral population is given by the observed frequency of sequences in the last season. For simplicity we treat mutation and fitness-based selection as separate steps in each season. Therefore each viral unit, before spreading in the new season, is allowed to mutate into a different sequence with a certain probability, which depends on the rate of mutation (as per nucleotide residue per cell cycle) and the time between flu seasons (as number of cell cycles per season).

During the flu season, each strain S_j is assumed to grow exponentially with a fitness (growth rate) $F(\mathbf{S}_j, t)$, i.e. if the frequency of strain $\mathbf{S}_j$ after mutation is given as $x_{\mathrm{m}}(\mathbf{S}_j, t)$, its frequency after growth and selection is calculated as

$$x(\mathbf{S}_i, t+1) = \frac{\exp\left(F(\mathbf{S}_j, t)\right) x_{\mathrm{m}}(\mathbf{S}_j, t)}{\sum_i \exp\left(F(\mathbf{S}_i, t)\right) x_{\mathrm{m}}(\mathbf{S}_i, t)}. \tag{1}$$

Here we assume that a fixed number N_{pop} of sequences survive into the next year and we can calculate the number of viral units $N(\mathbf{S}_j, t) = x(\mathbf{S}_j) N_{\mathrm{pop}}$ for each strain from the normalized strain frequency.

The fitness $F(\mathbf{S}_j, \mathbf{x}(t' < t)) = F_{\mathrm{int}}(\mathbf{S}_j) + F_{\mathrm{host}}(\mathbf{S}_j, \mathbf{x}(t' < t))$ is composed of two components, the intrinsic fitness $F_{\mathrm{int}}(\mathbf{S}_j)$ that signifies the intrinsic ability of a virus with that strain identity to spread in a completely susceptible human population, and a fitness cost $F_{\mathrm{host}}(\mathbf{S}_j, \mathbf{x}(t' < t)) < 0$ that depends on the accumulated amount of immune memory in the human host population against that specific strain.

The first, intrinsic fitness component, which is time-invariant is modeled as

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

Here F_0 represents the intrinsic fitness of a reference strain, the second term represents the fitness change due to independent mutations at each site α compared to the reference strain ($s_j^\alpha = 0$ if unmutated, 0 otherwise), and the last term represents the additional fitness change due to double mutations at pairs of sites α and β. The single-mutation coefficients h_α and the mutational coupling coefficients $J_{\alpha\beta}$ describe the intrinsic mutational fitness landscape, which we would like to infer from the sequence data. This fitness landscape determines, how easy or difficult it is for the virus to create escape mutations if specific sites or pairs of sites are targeted by a host response.

The host-dependent fitness component describes the decrease in the rate of spread of infections in a less susceptible population due to immune memory accumulated against each respective strain from previous infections. This component depends on the evolutionary history of the viral population and in our model it is calculated with a functional form similar to that previously used in other studies [Łuksza and Lässig, 2014], i.e.

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

This immune-mediated fitness decreases the strain fitness over time and is proportional to the prevalence $x(\mathbf{S}_i, t')$ of the same strain or of antigenically similar strains in previous years. 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 (Hamming) 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 each other's strains. Data from hemagglutination inhibition assays, which measure the cross-immunity between strains, suggest a typical cross-immunity distance of 5 amino acid residues for seasonal influenza A (H3N2) strains [].

The main motivation for the development of our model is to use the publicly available global, yearly sequence data of the spike protein HA, the major antigen of influenza, to infer the intrinsic mutational fitness landscape of influenza A (H3N2), i.e., the goal is to use our model together with time series data of influenza protein sequences to infer the intrinsic fitness coefficients $\{h, J\}$, in order to learn about the vulnerability of different pairs of regions to immune targeting.

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

Under a range of parameter choices, our simulations produce influenza-like immune-driven strain succession patterns, which are qualitatively similar to those observed for the evolution of HA (H3N2) in the human population (Figs. 1 and 2) and this similarity indicates that our model is able to capture the essential dynamics of antigenic evolution in seasonal influenza.

3 Inference of intrinsic fitness coefficients for single and pairwise mutations from influenza-like sequence data

Our fitness inference approach is based on the assumption that the selection of strains that survive into the next year is very stringent in each season. Stringent selection in our case means that only (or mainly) sequences in a narrow fitness range around the currently fittest strain survive into the next season. With this assumption all strains $\mathbf{S}_j(t)$ that are observed, i.e. selected, in a given season t will have similar total fitness. Thus we assume

$$F(\mathbf{S}_j, \mathbf{x}(t' < t)) \approx F(t, \mathbf{x}(t' < t)) \tag{4}$$

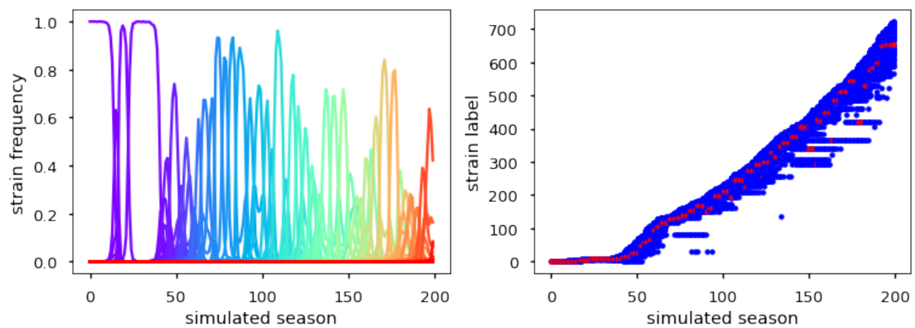

Figure 2: Strain succession for the evolution of simulated data over 200 time steps. (left) Each unique sequence (strain) is shown with its observed frequency in each simulated season as a solid line, with line colors ranging from purple (old strains) to red (new strains). (right) Strain labels here are counted 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: $N_{\mathrm{pop}} = 10^5$, $L = 20$, $\mu = 10^{-4}$, $\sigma_{\mathrm{h}} = 1$, $D_0 = 5$, $N_{\mathrm{simu}} = 200$, $B = 10^3$.

with $F(t, \mathbf{x}(t' < t))$ being a constant for each season t, conditional on the specific evolutionary history $\mathbf{x}(t' < t)$. From this assumption we obtain the following relation for the observed strains $\mathbf{S}_j$ in each given year, i.e.,

$$-F_{\mathrm{host}}(\mathbf{S}_j, \mathbf{x}(t' < t)) \approx \sum_{\alpha} h_\alpha s_j^\alpha + \sum_{\alpha<\beta} J_{\alpha\beta} s_j^\alpha s_j^\beta + F^*(t, \mathbf{x}(t' < t)), \qquad (5)$$

where $F^*(t, \mathbf{x}(t' < t)) = F_0 - F(t, \mathbf{x}(t' < t))$ is another constant at time t, conditional on the given evolutionary history. If we approximate the evolutionary history $\mathbf{x}(t' < t)$ with the observed sequences starting from the first year of observation and assume the model parameters $_{\mathrm{h}}$ and D_0 to be known, e.g. from independent cross-immunity studies, we can calculate $F_{\mathrm{host}}(\mathbf{S}_j, \mathbf{x}(t' < t))$ for each observed strain in each season. And if we ignore the dependence on the evolutionary history of $F^*(t, \mathbf{x}(t' < t)) = F_t^*$ we can use these host-dependent fitness values together with Eq. (5) to infer the intrinsic fitness coefficients $\{h, J\}$ as well as the additional parameters $\{F^*\}$ (one parameter per season). For this method we treat F_t^* as independent fit parameters, although they generally depend on other model parameters, but what we are mainly interested in are the coefficients $\{h, J\}$, which describe the intrinsic mutational fitness landscape of the virus. For the regression we minimize the sum of squared residuals between the data $Y_{\mathrm{data}}(\mathbf{S}_j, t)$ given by the LHS of Eq. (5) and the model

$Y_{\rm model}(\mathbf{S}_j, t, \{h, J, F^*\})$ given by the RHS of Eq. (5), i.e.,

$$\{h, J, F^*\} = \arg \min_{\{h,J,F^*\}} \left[\frac{1}{2} \sum_j \left(Y_{\rm data}(\mathbf{S}_j, t) - Y_{\rm model}(\mathbf{S}_j, t, \{h, J, F^*\}) \right)^2 \right.$$
$$\left. + \frac{\lambda_h}{2} \sum_\alpha h_\alpha^2 + \frac{\lambda_J}{2} \sum_{\alpha<\beta} J_{\alpha\beta}^2 + \frac{\lambda_{F^*}}{2} \sum_{t'} F_{t'}^{*2} \right], \quad (6)$$

where we also take into account regularization with coefficients $\lambda_h, \lambda_J, \lambda_F$ that are based on different Gaussian prior distributions for each type of coefficient.

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

Parameter	Description	Values
$\{h, J\}$	fitness coefficients for single mutations and pair-wise mutational couplings	values from HIV protein p24
L	length of sequence representation	$[5, 100]$
μ	mutation rate (per sequence site)	10^{-4}
D_0	cross-immunity distance	5
$N_{\rm pop}$	population size	$[10, 10^6]$
$\sigma_{\rm h}$	host-fitness coefficient	1
$(\lambda_h, \lambda_J, \lambda_{F^*})$	regularization coefficients for inference	$(0, 1, 0)$
$n_{\rm seasons}$	number of years/seasons used for inference	[10, 100]
B	number of sampled sequences per year	$[10, 10^6]$

The parameters for simulation and inference with explored values and ranges are collected in Tab. (1). For the simulations we used as input a set of fitness coefficients $\{h, J\}$, the values of which we chose from previously inferred mutational fitness coefficients of HIV protein p24 [Mann et al., 2014]. We used a fixed population size, to which the number of viral units was reduced at the beginning of each season. The evolution in the simulation was started at the unmutated reference sequence at season and we ran the simulation for 200 seasons. For inference we used data from a number $n_{\rm seasons}$ of seasons, without including the first 100 seasons, and for analysis we subsampled a number B of sequences per season. Mutation was assumed with a probability μ (per season) to mutate between the two mutational states per site.

For an example set of sampled data from one simulation we see that the distribution of total fitness is more narrow in each season than the distributions of the intrinsic and immune-dependent fitness components (Fig. 3), which indicates that the stringent selection assumption, which our inference approach depends on, is valid for these computer-generated data.

For each set of sampled sequences from each simulation, we can compare the inferred with the simulated intrinsic fitness coefficients (Fig. 4). The correlation

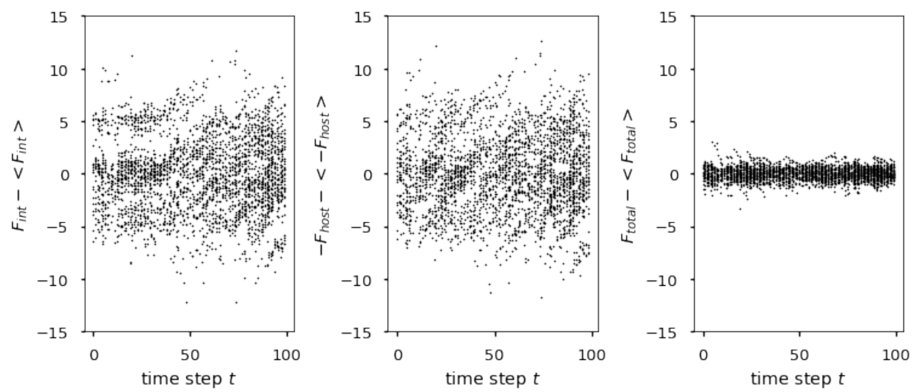

Figure 3: Fitness deviations from the mean of selected strains for each simulated season between season 100 and 200. (left) intrinsic fitness component F_{int}, (middle) immune-dependent fitness component F_{host}, (right) total fitness $F_{\mathrm{total}} = F_{\mathrm{int}} + F_{\mathrm{host}}$. For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

coefficients between simulated and inferred coefficients and in particular the 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, where escape mutations lead to low versus high (negative) fitness costs.

Besides the correlation coefficient r_{hJ} we can use another measure for inference performance, if we are only interested in identifying those pairs of sites that have a high fitness cost as those, whose fitness cost is below a certain threshold, i.e. with

$$h_\alpha + h_\beta + J_{\alpha\beta} < F_{\mathrm{threshold}} < 0. \tag{7}$$

In this case we can use classification performance measures to assess how well our inference is performing on this specific classification task.

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. 5).

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. 6), we find that a minimum total number of sampled strains $n_{\mathrm{sample}} = n_{\mathrm{seasons}} * B$ is required for accurate inference. In the shown example a total sample size of $n_{\mathrm{sample}} \geq 10^5$ strains is required for high inference performance.

The inference performance further strongly depends on the sequence length L (Fig. 7) and on the population size N_{pop} (Fig. 8). Inference performance in

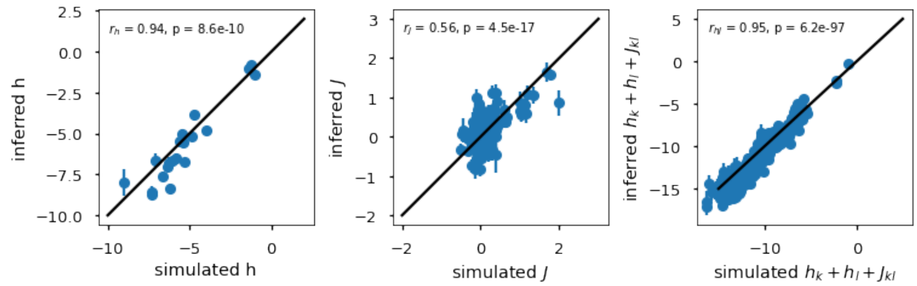

Figure 4: Parameter correlations for the inference on one simulated data set. Inferred values of the fitness coefficients are shown against the fitness coefficients that were used as input values for the simulation. (left) single-site mutational fitness coefficients h, (middle) coupling coefficients J for simultaneous mutations at any two sites, (right) 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: $N_{\text{pop}} = 10^5$, $L = 20$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D_0 = 5$, $N_{\text{simu}} = 200$, $B = 10^3$, $n_{\text{seasons}} = 100$, $\lambda_h = 0$, $\lambda_J = 1$, $\lambda_{F^*} = 0$.

terms of the correlation r_{hJ} between inferred and simulated double-mutational fitness coefficients decreases with increasing sequence dimension and increases with increasing population size towards its maximum 1.

4 Discussion

Appendix

Indicators for inference performance

We define a measure of reverse stringency as the ratio between the time-averaged yearly standard deviations of total and host-dependent fitness, i.e. as

$$\frac{\langle \text{std}(F_{\text{tot}}) \rangle}{\langle \text{std}(F_{\text{host}}) \rangle}. \tag{8}$$

This reverse stringency measure, which compares the widths of the distributions of the different fitness components (cf. Fig. 3), shows a strong negative rank correlation with the inference performance measure r_{hJ}, both for variation of population size and sequence length, as well as sample size (Fig. ...).

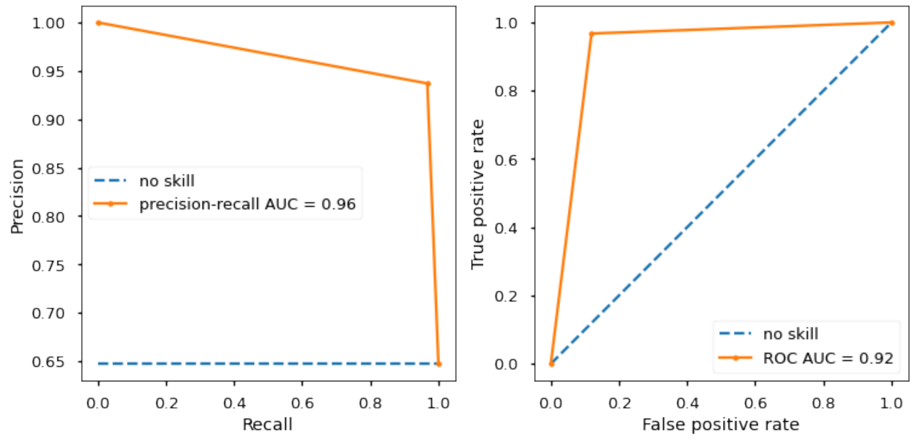

Figure 5: Classification performance for the inference on one simulated data set. Double mutations are classified as deleterious if their total fitness cost is lower than $F_{\text{threshold}} = -10$. (left) The precision-recall curve (PRC) and (right) the ROC curve for the deleterious-mutation classifier from inferred fitness coefficients. Blue, dashed lines show a no-skill classifier for comparison and the area under the classifier curve is given in each panel. For the shown example the parameter values for simulation and analysis are: $N_{\text{pop}} = 10^5$, $L = 20$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D_0 = 5$, $N_{\text{simu}} = 200$, $B = 10^3$, $n_{\text{seasons}} = 100$, $\lambda_h = 0$, $\lambda_J = 1$, $\lambda_{F^*} = 0$, $F_{\text{threshold}} = -10$.

References

[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

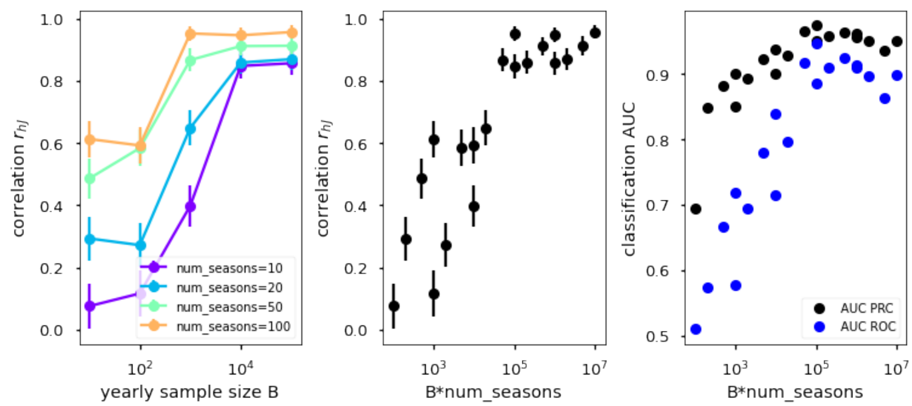

Figure 6: Inference performance for varying yearly sample size B per season and varying number n_{seasons} of seasons used for inference. (left) The correlation coefficient r_{hJ} between inferred and simulated double-mutational fitness costs as function of yearly sample size B for various n_{seasons}. (middle) The performance measure r_{hJ} as function of total sample size $B * n_{\text{seasons}}$. (right) The area (AUC) under the ROC curve and under the precision-recall curve (PRC) for classification of deleterious double mutations with $F_{\text{threshold}} = -10$, shown as function of total sample size $B * n_{\text{seasons}}$. For the shown example the fixed parameter values for simulation and analysis are: $N_{\text{pop}} = 10^5$, $L = 20$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D_0 = 5$, $N_{\text{simu}} = 200$, $\lambda_h = 10^{-4}$, $\lambda_J = 1$, $\lambda_{F^*} = 10^{-4}$.

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.

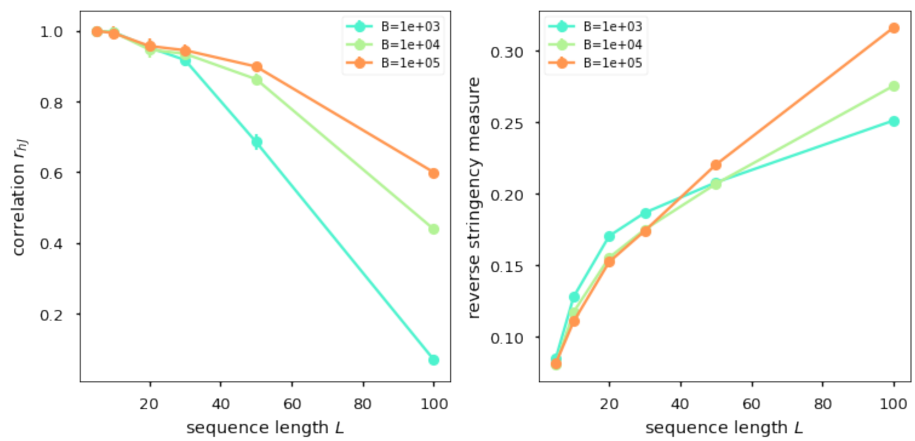

Figure 7: Inference performance and stringency measure for simulations with varying sequence lengths L and yearly sample size B. (left) Correlation coefficient r_{hJ} between inferred and simulated double-mutational fitness costs as function of sequence length L for various sample sizes B. (right) Reverse stringency measure for each simulated data set in the left panel. The Spearman rank correlation coefficient between the measures for inference performance r_{hJ} and reverse stringency is calculated as $\rho = -0.83$. For the shown example the fixed parameter values for simulation and analysis are: $N_{\text{pop}} = 10^5$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D_0 = 5$, $N_{\text{simu}} = 200$, $num_{\text{seasons}} = 100$, $\lambda_h = 10^{-4}$, $\lambda_J = 1$, $\lambda_{F^*} = 10^{-4}$.

[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.

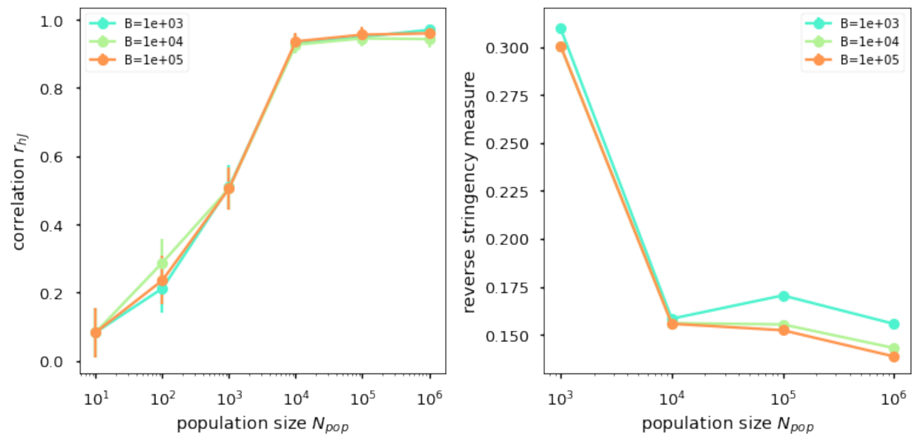

Figure 8: Inference performance and stringency measure for simulations with varying population size N_{pop} and yearly sample size B. (left) Correlation coefficient r_{hJ} between inferred and simulated double-mutational fitness costs as function of population size N_{pop} for various sample sizes B. (right) Reverse stringency measure for each simulated data set in the left panel. The Spearman rank correlation coefficient between the measures for inference performance r_{hJ} and reverse stringency is calculated as $\rho = -0.85$. For the shown example the fixed parameter values for simulation and analysis are: $L = 20$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D_0 = 5$, $N_{\text{simu}} = 200$, $n_{\text{seasons}} = 100$, $\lambda_h = 0$, $\lambda_J = 1$, $\lambda_{F^*} = 0$.

[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.

[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.

[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).

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.