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

Julia Doelger, Mehran Kardar, Arup K. Chakraborty

April 15, 2021

We developed a method for inferring the intrinsic fitness costs due to single and pairwise mutations of influenza-like antigens. Such inference can be useful for the design of novel long-term protective immunization strategies against seasonal influenza. Our inference method is based on immune-driven evolution and a strongly purifying selection regime. Inference performance and applicability range was tested with artificial sequence data that are created by a stochastic simulation of influenza-like sequence evolution.

1 Introduction

Influenza virus causes seasonal epidemics, which lead to the deaths of several hundred thousand people worldwide every year.

Due to their high mutation rate, seasonal influenza viruses like the subtype A(H3N2) continuously evolve and are thereby able to escape the population-wide immune memory responses, which are accumulated during every flu season.

There exist seasonally updated vaccines against influenza, which vary in effectivity, but there is until now no universal and long-term protective immunization method available that can protect against current and future viral strains.

The influenza spike protein hemagglutinin (HA) and in particular the five major epitope regions, labeled with letters A-E, form the dominant antigens, which elicit antibody responses in humans. Those regions on the head of HA are easily accessible but are also highly mutable. Antibodies targeting a specific epitope can usually be easily escaped from, often by single point mutations. However, there must be differences how easily each distinct epitope sequence can find mutational escape routes without costing the virus its functionality, and there further might be certain pairs of antibody targets that might have no viable escape path when both are targeted at the same time.

This vulnerability can be described with the concept of an intrinsic fitness landscape, where each viral unit is expressed with a fitness value, which has

contributions from each single and double mutation within its protein sequence and which describes the viability of the specific sequence.

In order to find the vulnerable sites and pairs of sites in the spike protein, we therefore would like to infer the intrinsic fitness coefficients, which describe the fitness costs due to single and double mutations.

The prevalence of different HA sequences has been recorded each year since 1968 and these sequences encode a snap shot information about the fitness of different strains at different time points. However this total fitness describes in general the ability of a virus with a specific sequence to spread in the human population at a given time, and this total fitness not only includes intrinsic viability of the viral sequence, but also depends on the human immune memory in the current population. A virus that for example could spread very easily with high intrinsic fitness in a completely susceptible population might not be able to proliferate in the actual population, which can block it from spreading with a strong immune memory response. Therefore we need to disentangle the intrinsic fitness effects from those of immune memory, when using the raw sequence data for inference.

We have some information about the accumulation of immune memory in the human population. (from the recorded past infections with different strains and from cross-immunity experiments)

Since we generally cannot measure the intrinsic ability of a single viral sequence to spread in a completely susceptible human population and therefore it is difficult to directly test our inference on real sequence data, we developed a simulation of a flu-like system of antigenic evolution, with which we can create flu-like artificial sequence data based on a known intrinsic fitness landscape. We use those simulated data to test the performance and applicability range of our inference method by comparing the inferred with the simulated fitness coefficients.

2 Model of influenza evolution based on intrinsic and immune-mediated fitness

We assume a model of influenza sequence evolution that is based on mutation and subsequent selection according to a fitness function with an intrinsic and an immune-mediated contribution.

At each coarse-grained time step (many viral generations) we assume that each sequence can mutate at each site with a specific mutational probability, which is summarized in the transition matrix $\mathbf{M}$ with entries M_{ji}, which describe the probability of mutating from one unique sequence $\mathbf{S}_i$ to another $\mathbf{S}_j$. The probability of finding a strain (unique sequence) $\mathbf{S}_j$ at time step t after mutation is then given as

$$x_{\mathrm{m}}(\mathbf{S}_j, t) = \sum_i M_{ji} x(\mathbf{S}_i, t), \tag{1}$$

as the sum of probabilities of descending from any sequence that was observed before mutation. Here $x(\mathbf{S}_i, t)$ is the probability of observing strain $\mathbf{S}_i$ at the same time step, but before mutation.

After mutation we assume a growth and selection step, where we assume the total number of viral units that enter the next season form a constant-size bottleneck, i.e. we assume a Wright-Fisher-type system with constant population size. The probability of observing a strain $\mathbf{S}_i$ at time $t+1$ in the next season is calculated with

$$x(\mathbf{S}_i, t+1) = \frac{\exp\left(F(\mathbf{S}_j, t)\right) x_{\mathrm{m}}(\mathbf{S}_j, t)}{\sum_i \exp\left(F(\mathbf{S}_i, t)\right) x_{\mathrm{m}}(\mathbf{S}_i, t)}. \tag{2}$$

Each strain $\mathbf{S}_j$ has an assigned fitness (or effective growth rate) $F(\mathbf{S}_j, t)$ for each time step t, which determines how quickly each viral strain can spread. We assume independent growth of different strains until the transition to the next season, at which time a constant number of sequences is selected from the total sequence pool. It is therefore the relative fitness compared to other concurrent strains together with the previous strain frequency distribution, which determines if a strain will increase or decrease its proportion in the population. All of this is summarized in Eq. (2).

We assume the fitness function to be composed of two parts. The first contribution is the intrinsic fitness $F_{\mathrm{int}}(\mathbf{S}_j)$, which determines the growth rate of a strain $\mathbf{S}_j$ in a human population with no previously acquired immunity and which is purely based on the virus' ability to functionally infect, replicate and transmit in humans. This fitness component is assumed to not vary with time. We approximate the intrinsic fitness of a strain $\mathbf{S}_j$ as

$$F_{\mathrm{int}}(\mathbf{S}_j) = F_0 + \sum_{\alpha} h_\alpha s_j^\alpha + \sum_{\alpha<\beta} J_{\alpha\beta} s_j^\alpha s_j^\beta. \tag{3}$$

Here F_0 describes the fitness of a reference strain, whose sequence can be represented as a list of zeros, i.e., the mutational state at each site is labeled as *unmutated* for this reference strain. Each other strain is represented with its mutational state at each site compared to the reference sequence, with 0 being *unmutated* and 1 being *mutated*. Here we take into account the fitness effects of single-site mutations, with the coefficients $\{h\}$, as well as the deleterious or compensatory effects of pairwise-coupled mutations within the same sequence, with the coupling coefficients $\{J\}$. With this first-order coupling representation akin to an Ising model in statistical mechanics we reduce the number of parameters describing the fitness landscape from 2^L (the number of unique sequences with length L) to $L(L-1)/2 + L \sim L^2$ fitness coefficients.

The second part of the fitness function describes the immune-mediated fitness cost, which makes those strains less fit, which were prevalent in previous time steps and that led the human population to an acquire immune memory against them (and against their close relatives). We use a mean-field term for this immune-mediated fitness component, similar to one that has been

previously used in a similar context [Łuksza and Lässig, 2014]. The immune-mediated fitness component for strain $\mathbf{S}_j$ at time step t depends on the previous history of the evolving virus in the human population and is modeled as

$$F_{\text{host}}(\mathbf{S}_j, \mathbf{x}(t' < t)) = -\sigma_{\text{h}} \sum_{t'<t} \sum_i x(\mathbf{S}_i, t') \exp\left(-|\mathbf{S}_j^{\text{ep}} - \mathbf{S}_i^{\text{ep}}|/D_0\right). \quad (4)$$

The immune-mediated fitness cost accumulates over time with every occurrence $x(\mathbf{S}_i, t')$ of the same or a similar strain and this forces the virus to continuously evolve away from previously prevalent sequences. Here $|\mathbf{S}_j^{\text{ep}} - \mathbf{S}_i^{\text{ep}}|$ is the mutational distance between strain $\mathbf{S}_i$ and $\mathbf{S}_j$ within their immune-targeted epitope regions and D_0 is the cross-immunity distance, i.e., the typical mutational distance within the epitope regions between two strains, beyond which the strains are dissimilar enough to not be targeted by each other's immune response. Haemagglutinin inhibition data, which measure cross-immunity between strains, suggest a typical cross-immunity distance of 5 amino acids within the epitope regions of two strains. But there is likely a difference in cross-immunity if the mutations occur all within one epitope region or are spread across epitope regions, so for a more accurate representation of immune-mediated fitness against influenza it might be better to split the cross-immunity term into distinct terms taking into account the mutational distance within each epitope region separately.

The total fitness of a strain $\mathbf{S}_j$ at time t is the sum of intrinsic and immune-mediated fitness component, i.e.,

$$F(\mathbf{S}_j, \mathbf{x}(t' < t)) = F_{\text{int}}(\mathbf{S}_j) + F_{\text{host}}(\mathbf{S}_j, \mathbf{x}(t' < t)) \quad (5)$$

$$= F_0 + \sum_\alpha h_\alpha s_j^\alpha + \sum_{\alpha<\beta} J_{\alpha\beta} s_j^\alpha s_j^\beta - \sigma_{\text{h}} \sum_{t'<t} \sum_i x(\mathbf{S}_i, t') \exp\left(-|\mathbf{S}_j^{\text{ep}} - \mathbf{S}_i^{\text{ep}}|/D_0\right). \quad (6)$$

Our aim is to use the observed influenza spike protein sequences from the last 50 years together with our influenza evolution model and knowledge of parameter values such as mutation rate μ and cross-immunity distance D_0 to infer the intrinsic mutational fitness landscape of the protein, i.e., we want to infer the intrinsic fitness coefficients $\{h, J\}$.

On this account we developed an inference approach, which we will test on artificial data sets that are produced with a simulation according to our model for sequence evolution. The inference approach and its performance test on simulated data is described in the next section.

3 Inference of intrinsic fitness from flu-like antigenic sequence data

In this section I will describe an inference approach, which we developed for the inference of the intrinsic mutational fitness landscape of the influenza antigen.

This specific approach is based on the assumption that the selection of strains is very stringent at each time step, which is not unreasonable since observed influenza sequence diversity within each year is typically quite low.

Such a stringent selection in our case means that mainly sequences in a narrow fitness range around the maximum fitness survive into the next time step at each time. If this assumption is true, then those strains $\mathbf{S}_j^{\text{observed}}$ that are observed together at the same time step will have a similar total fitness, i.e.,

$$F(\mathbf{S}_j^{\text{observed}}, \mathbf{x}(t' < t)) \approx F(t, \mathbf{x}(t' < t)), \tag{7}$$

with $F(t, \mathbf{x}(t' < t))$ being a constant for the respective time step, given the evolutionary history. From this assumption follows

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

where $F^*(t, \mathbf{x}(t' < t)) = F_0 - F(t, \mathbf{x}(t' < t))$ is another constant at time t, given the evolutionary history. If we assume σ_{h} and D_0 to be known and if we have the whole sequence history, we can calculate $F_{\text{host}}(\mathbf{S}_j^{\text{observed}}, \mathbf{x}(t' < t))$ for each observed sequence at each time step. And if we naively ignore the dependence on the evolutionary history, i.e., assuming $F^*(t, \mathbf{x}(t' < t)) = F_t^*$ we can use Eq. (8) and $F_{\text{host}}(\mathbf{S}_j^{\text{observed}}, \mathbf{x}(t' < t))$ to try to infer $\{h, J\}$ as well as the additional parameters $\{F^*\}$ (one for each time step) via linear regression. For the regression we minimize the sum of squared residuals between the data given by the LHS of Eq. (8), which we call $Y_{\text{data}}(\mathbf{S}_j, t)$, and the corresponding model terms given by the RHS of Eq. (8), which we call $Y_{\text{model}}(\mathbf{S}_j, t, \{h, J, F^*\})$, i.e.,

$$\{h, J, F^*\} = \arg \min_{\{h,J,F^*\}} \frac{1}{2} \sum_j (Y_{\text{data}}(\mathbf{S}_j, t) - Y_{\text{model}}(\mathbf{S}_j, t, \{h, J, F^*\}))^2 + \frac{\gamma}{2} \sum_{\alpha<\beta} J_{\alpha\beta}^2, \tag{9}$$

where we also (in the last term) take into account a regularization of the coupling coefficients $\{J\}$, which is based on the assumption of a Gaussian prior distribution around 0 with e.g. $\gamma = 1$.

3.1 Test of the inference approach on computer-generated sequence data

[include plot of fitness coefficients from p24 that are used as input.]

In order to test, to what extent the above defined inference approach can recover the underlying intrinsic mutational fitness landscape from sequence time series, we simulated the sequence evolution according to our model, which was defined above. For this we used as input a known set of fitness coefficients {h,J}, which we chose from the previously inferred mutational fitness coefficients of HIV protein p24 [Mann et al., 2014], in order to get a realistic distribution of single- and pairwise mutational fitness effects for a viral protein (Fig. ...).

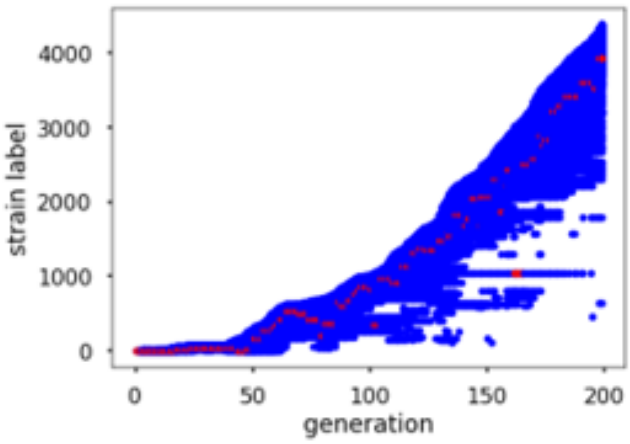

Figure 1: Evolution of strains in simulation, where strain labels indicate the time of first observation, older strains have lower labels. The strain, which is the most prevalent in each year is marked in red.

In an example simulation we used a population size of $N = 10^6$ sequences of length $L = 20$, starting at the unmutated reference sequence at time 0, and evolving for 200 time steps. For inference in this example we used the 100 latest time steps of the simulation and subsampled $B = 10^4$ sequences per time step. Model parameters were set to $\sigma_{\mathrm{h}} = 1$, $D_0 = 5$, and the probability to mutate at each site between two time steps was set to $\mu = 10^{-4}$ for (symmetric) mutation between the two mutational states. Fig. 1 shows the evolution of strains in that simulation with strain labels starting at 0 for the oldest, unmutated strain and increasing with each newly created strain.

Fig. 2 shows the distributions of intrinsic fitness, immun-mediated fitness and total fitness for the selected sequences at different time steps around their respective means. The distribution of total fitness is more narrow than its individual contributions, which is a first indication confirming the stringent selection regime.

Fig. 3 shows the performance of the proposed inference method, where the plots show the fitness coefficient that were inferred (y-axis) against the fitness coefficients (x-axis) that were used as simulation input. The correlations between input and inference in this example are high and therefore indicate high accuracy of our inference method to distinguish between single and pairs of sites, at which escape mutations imply low versus high fitness costs.

[include figures (like Fig. 4) comparing inference performance (correlation between fitness coefficients) for varying subsampling sizes, varying num. of inference time steps, varying sequence lengths, varying fitness coefficients, varying sigmah, varying mutation rate]

[include figures comparing the classification performance (e.g. for finding

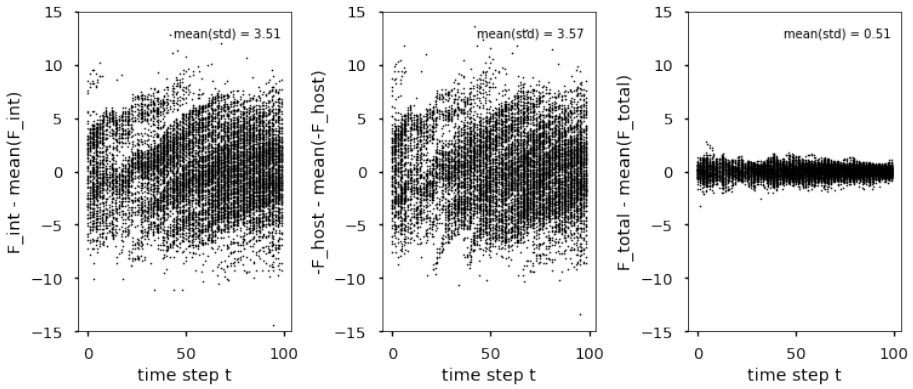

Figure 2: Fitness distributions of selected strains at different time points. In this example the simulation and sampling parameters are $L = 20$, $N_{\text{pop}} = 10^6$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D_0 = 5$, $B = 10^4$.

the pairs of sites, which have fitness below a threshold) for varying parameters]

[include figures (like Fig. 5) comparing relative widths of fitness distributions (stringency) across different parameter ranges; correlation with inference performance?]

[include figures with summary statistics that are easily calculated from available raw sequence data, which might indicate selection regime/inference performance]

4 Discussion

Recently another inference method, the so-called marginal path likelihood (MPL) method has been proposed for the inference of mutational fitness costs from sequence time series [Sohail et al., 2020]. However, this method does not try to disentangle intrinsic from immune-mediated fitness effects and only considers selection due to total fitness. Secondly the MPL method considers a Wrightian fitness description, i.e., it considers linear growth between selection steps. For influenza we assume that viral populations grow exponentially in the human population between selection bottlenecks at the end of flu seasons and therefore we use a Malthusian growth description ($\sim \exp(F)$). Finally, we aim to not only infer single-mutation coefficients but also intrinsic fitness couplings $J_{\alpha\beta}$ between sites, which the MPL method does not try to infer. The MPL method in particular obtains improved inference performance compared to previous methods that do not take into account linkage effects like hitchhiking or clonal interference. Since we in our method use the whole-sequence information together with the time of observation to calculate the response function F_{host} we also take account for linkage effects and clonal interference or hitchhiking will likely not pollute our inference results [I will look more closely again at the MPL method

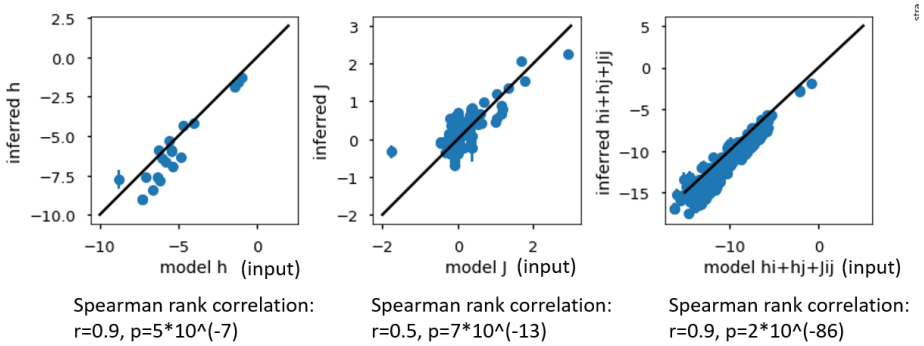

Figure 3: Inference test on computer-generated data. Inferred values of the fitness coefficients are shown against the fitness coefficients that were used as input values for the simulation. Left: single-site mutational fitness coefficients, middle: coupling coefficients for simultaneous mutation at two sites, right: total fitness changes due to simultaneous mutation at two sites i and j.

to make an accurate comparison].

In order to make meaningful predictions based on observed influenza protein sequence data, we need to translate our inference approach to this more complex system, which is constrained by small sampling sizes per year as well as by a high-dimensional sequence landscape. This means that we need to coarse-grain the protein sequence representation in a meaningful way to be able to reduce the number of inferred parameters but still be able to make predictions about the fitness effect of relevant coupled mutations, in particular between different immune epitope regions.

There are around hundred amino acids within the 5 major epitopes on the head region of HA. If we want to use the above inference approach we also need to make sure to confirm the functional form of the immune-mediated fitness component F_{host}, which drives the evolution. For that we can use haemagglutinin inhibition data, which measure the pairwise cross-reactivity between various strains. In order to test fitness inference on the real data, we might be able to compare to few available fitness measurements from mutational assays, although those are cumbersome and only estimate viral growth ability in cells, not rates of spreads across individuals in the human population. We might also use the inferred fitness to predict future strains, which might additionally provide some model validation, and if accurate might further help improve traditional seasonal vaccine formulations.

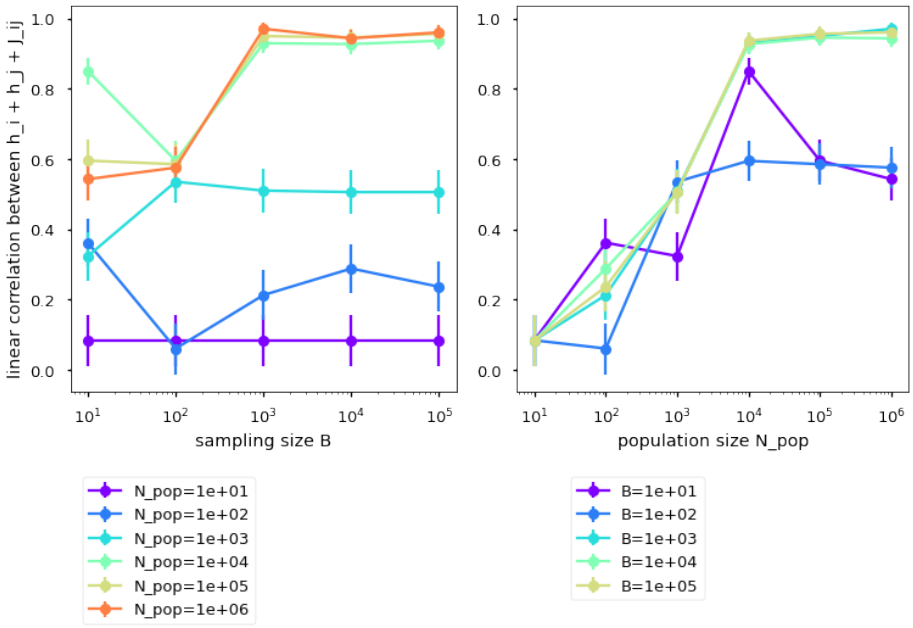

Figure 4: Linear correlation between simulated and inferred fitness coefficients for simulations with varying population size N_{pop} and for varying subsampling sizes B per time step. For these simulation with $L = 20$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D = 5$ and $N_{\text{inf}} = 100$ time steps used for inference, a sampling size of 1000 sequences per time step or higher results in a high correlation close to 1.

References

[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.

[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.

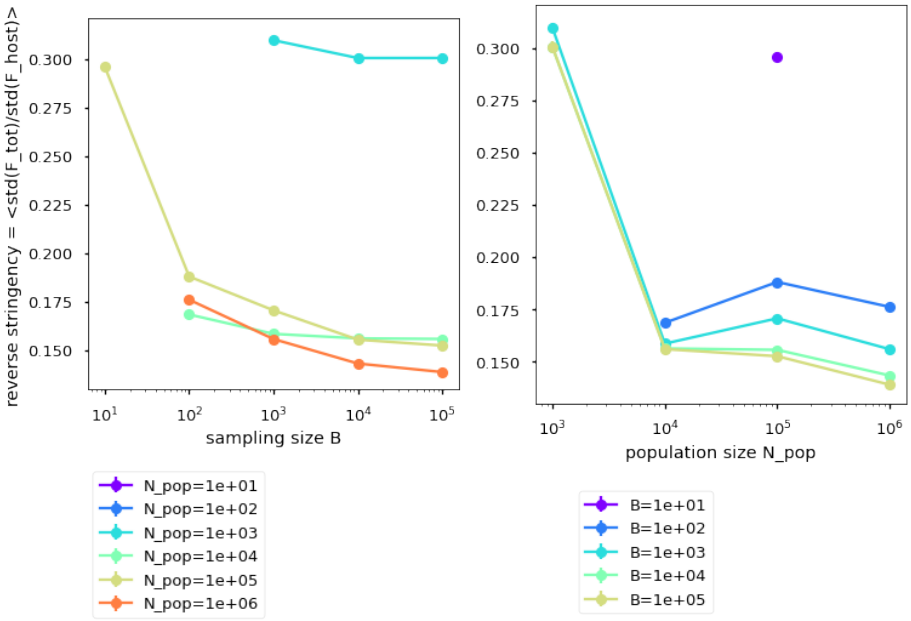

Figure 5: Reverse stringency estimate for simulations with varying population size N_{pop} and for varying subsampling sizes B per time step. For these simulation with $L = 20$, $\mu = 10^{-4}$, $\sigma_{\text{h}} = 1$, $D = 5$ and $N_{\text{inf}} = 100$ time steps used for inference, the reverse stringency estimate decreases with increasing sampling size and with increasing population size. A large stringency (low reverse stringency estimate) roughly correlates with a high inference performance measured with the linear correlation between simulated and inferred fitness costs (cf. Fig. 4).