25 March 2021, mtg with Mehran

* M summarized: in principle if I make the sequence length short enough, I can infer h and J well with our host-immunity based inference method -> only true if in the right param. regime
* We could **write a method paper based on the toy model to establish an inference method for fitness landscape for influenza-like antigens**
* Central question: For HIV fitness landscape was constructed -> can we do the same for flu?  
  The evolution of the virus makes the equilibrium-methods that were used for HIV unfeasible.  
  How could we analyze these kind of data?
* Manuscript scope suggestion: Present the toy model and the test of our inference procedure
* **M asked: What was again the reason how we got to equations (7) and (8) in my document**   
  my answer:  
  we found that the previous method using both log(x/x’) [relative growth of different strains since last generation] and F\_host to estimate F\_int, were not giving as good results as using F\_host alone, especially with low subsampling of sequence population.   
  log(x/x’) in the full simulated data is narrowly distributed around 0, but the estimate with fewer samples becomes quite noisy for low sample sizes with large outliers (either very high or very low predicted fitness compared to other strains although fitness in reality is narrowly distributed)  
  later we confirmed that in most simulations F\_tot was very narrowly distributed but Fhost/Fint showed higher variance in each time step
* **M: How could we test the stringent selection hypothesis and how does this come naturally out as consequence of the simulation?**   
  my idea: test larger population sizes or smaller sigmah, generally do parameter sensitivity analysis  
  (-> currently Fhost drives the evolution faster than typical sequences that mutate from last year’s ancestors can follow, which means that only the absolute fittest strains of the last-years population have a chance to survive into the next year)

1 April 2021, further thoughts on flu project goals and challenges/next steps

* Regarding M’s earlier comment: it does not suffice to make sequence short for the new inference method to work, the system also needs to be in the right regime  
  For example for inflated J couplings, the inference becomes worse for both inference methods.  
  Why is that?
* **How should we go forward?**
  + **Regarding the analysis of the toy model**
    - I can run a numerical parameter **sensitivity analysis** and analyze how different parameter combinations affect the quality of fitness inference (measured for example by Pearson or Spearman correlation)
    - We can **think more (analytically) about the mechanisms that lead to the stringent-selection regime,** which allows the simple inference method  
      so we can make hypothesis about the influence of different parameters, which we can test by numerical simulation
    - We can **try some coarse-grained inference based on fine-grained sequence data**
    - our goal is to predict specific pairs of epitope sequences that when targeted together at a specific time point, give long protection:
      * In the simulation or in the real data, we could **test for various (predicted) pairs of specific epitope sequences**, what is the probability pfail in each year that one might be infected by a sequence that has neither one of the two specific epitope seqs **(1-pfail would give a measure of protection for each year** )
      * Maybe we can also **think of a more direct inference method based on those data (optimize for protection from future sequences)**
  + **Regarding the translation to flu data:**
    - **We can continue thinking of coarse-graining schemes**:  
      we might want to consider expressing mutational fitness costs due to mutations between sequences that occur in subsequent years instead of mutations compared to ref strain from 1968.  
      We can calculate F\_host in principle with the full (epitope) sequence data (without coarse-graining),   
      so each sample for the inference might be expressed by a coarse strain representation (single and double mutations in epitope regions relative to close ancestors), the year where it was sampled plus the current estimated F\_host value.
    - I can do further analysis of the HI data to make good estimates of cross-immunity.  
      **We might want/need to recalibrate with the antibody accessibility of different epitope regions** (Assaf), since less accessible regions will be less driven to mutate and therefore in an (uncalibrated) inference show up with higher mutational fitness costs. In reality mutations in regions that are harder accessible should on average have less effect on immunity.  
      This might be taken into account by epitope-specific cross-immunity fits to HI data and/or by comparing with Assaf’s antibody-on-rate calculations.
    - **How can we test the inferred mutational fitness landscape from real flu data**?
      * Compare fitness coeffs to the (very few) available mutational assays with in vitro fitness measurements? -> how to deal with coarse-graining and transmission effects that are not included in vitro
      * Test protection likelihood from flu strains in subsequent years, in case of targeting of predicted good-protection combinations of epitopes  
        Maybe this (duration and strength of protection by different epitope combis) could be used as inference performance measure
      * Leave out part of the samples (from each year)  
        Use ML-type validation to tweak model hyperparameters (that are not fitted): calculate loss function on validation set -> vary hyperpar. -> repeat.  
        Calculate fit quality (loss function) on left-out test set for final model (with fixed hyperpars)
      * Suggest experiments, with which our results could be tested
    - What we really want to do with our inference is predict those pairs of ab-epitope regions that when targeted together, will be most likely to give protection against current and future strains.
      * Also the prediction, or at least the specific epitope sequences that should best be targeted, should depend on the year, in which the immunization is given. Our predictions should therefore be able to be translated into a specific recommended combination of epitope sequences that should be targeted in a given year.  
        For immunization one should not only take into account what are the most vulnerable epitopes (difficult to escape) but also by how large a part of the population each epitope is expected to be targeted already due to immune memory (how large is F\_host for it). If some sequence is expected to be targeted by large part of the population already, it might anyway not survive in the future due to herd immunity, therefore no need to use it in immunization.

1 April, meeting with Arup/Mehran:

* M’s recollection: surprising result of equations (7) and (8)
* Purifying selection-> one fitness wins at each time step?
* Figure out the selection regime analytically and by looking at simulation -> focus on why the fitness distribution is very narrow and what makes our inference work:
  + What are the relevant non-dimensional parameters?
  + What is the main mechanism that leads to this selection regime?  
    What are the markers, which might point to the same regime in flu data?
* Make sure to not go into any rabbit holes and don’t focus on real data for the manuscript, only toy model
* Use document that I wrote as outline for manuscript

8 April, meeting with Arup/Mehran:

* We meet on AKC lab mtg zoom link (which Mehran has too)
* What I want to discuss:
  + Analytical analysis of selection:   
    Fmax determined by max fitness of available strains, Fmin determined by either Fmin of available strains or by fitness that leads to decrease from x=1 to x=1/N
  + Summary statistics to compare in simulations:  
    linear/rank correlations: 3 values (+3 p values)  
    mean std of F\_host, F\_int, F\_tot: 3 values  
    how does inference performance (correlation) correlate with stdF\_tot/(mean(stdF\_host, stdF\_int))?
* Conclusions from meeting with Arup and Mehran:
  + Do numerical parameter sweeps of interesting parameters:
    - use linear correlation (3 values) as performance estimators
    - investigate how inference performance correlates with stdF\_tot/(mean(stdF\_host, stdF\_int))   
      and with other indicators that I might also be able to measure from real data
    - investigate, how numerical stringency estimator stdF\_tot/(mean(stdF\_host, stdF\_int)) correlates with analytical estimate
  + Write down the analytical estimate that I made to quantify the width of the selected fitness distribution (as function of different params)
  + Compare our approach with approach (and system) by John Barton and Matthew McKay (Nature Biotechnology):
    - How is our approach different?
    - Why can we not use their proposed approach?  
      no intrinsic couplings, different selection regime/different growth measure (F vs. exp(F))
  + Arup likes the plots of F\_int, F\_host, F\_tot over time where difference in widths of the distributions are clearly visible

15 April, meeting with Arup/ Mehran:

What I want to discuss:

* Suggestions for figures (examples and further figures):
  + fitness dists
  + inference performance (correlation) for varying parameters-> which parameters are most interesting (sequence length, subsampling size, inference parameters, intrinsic/host fitness scaling,…)?
  + Classification performance for inferring vulnerable target pairs (e.g. for pairs of sites with total fitness cost above threshold)
  + (reverse) Selection stringency mean[std(ftot)/std(fhost)] as indicator for validity of inference model assumption -> check correlation with correlation or other inference performance measures
  + Other summary statistics that we can calculate from the sequence data (without additional knowledge we don’t have for flu), which might indicate the selection regime/inference validity? Which might I try?
* Comparison with John Barton’s MPL method:  
  discuss what I wrote so far in discussion

Conclusion from meeting:

* Remove mean(std) as label from fitness distribution plots
* If using mean stds for stringency measure, indicate that in the text and make sure that it is not misunderstood: we don’t claim that the std stays constant over time  
  (that is something that I could explore further, but of secondary interested for paper)
* About parameter exploration figures (like figure 4):
  + use 1 plot instead of 2 (don’t show same data in two plots in different projections)
  + only do a few more simulations for varying sequence length and compare performance for different length/sampling size
  + Arup says: too many param explorations are overkill
* (Mehran’s suggestion: for simu with L=20 try inference for only 5 randomly chosen sites,  
  see if and how the inferred h and Js correlate with the input params,  
  Arup says he already knows the result (poor/inconclusive results mostly since J matrix is sparse and highly dependent on actual J matrix and chosen sites) so he suggests not to do that experiment)
* M and A agree that I could use a classification performance measure aside from the linear corr. to assess classification of highly deleterious pair mutations from others  
  (just a small additional calc)
* M and A agree that it is worthwhile to think more about a measure/measures based on the sequence data alone, which we have for real flu, that might indicate selection regime/good inference performance:
  + Test correlation of suggested measures with inference performance  
    (can include Fhost distributions, strain succession pattern (typical time between birth and decline) etc.)
* About discussion of MPL method in comp. with ours:
  + A and M say it is fine to have just a discussion along the lines what I have
  + Most important points are that
    - MPL considers a fixed fitness landscape without time-varying contributions (which we have however)
    - They don’t take into account mutational couplings
  + Agreement that our method, like MPL, does not ignore linkage effects

27 April meeting with Arup/Mehran

What I want to discuss:

* Which figures (parameter explorations) should we keep? Should I add a model schematic?
* I will need to run each simu (or at least each sampling) several times to get accurate performance measures
* Stringency indicator from raw data (which measures did I try?), log(x/x’) should be (like F\_tot) narrowly distributed around 0 if stringent selection, but will have many outliers due to undersampling
  + Could use log(x/x’) for different subsampling B to show that including log(x/x’) for inference mainly adds noise and decreases performance
* What could be the reason for linear vs. exponential increase of strain numbers in sequence evolution? (larger mutation rate? Sequence length? Npop?)
  + Should I explore in the different simulations, how the total number of strains (observed up to time t) grow with time?

Conclusion from meeting:

* Write full manuscript and send it to A and M, only meet after they have read it
* **Journal options:**
  + Nature Scientific Reports (where Florian’s paper 2016 was published, Arup says maybe)
  + PNAS (suggestion from Navish, Arup says it might not fly there)
  + PLoS Computational Biology (where Mann et al. 2014 paper is published, Arup says maybe)
  + Journal of the Royal Society Interface (my suggestion, Arup says maybe)
  + Mathematical Biosciences (Arup: if our other ms is received well, might not work out to check timewise)
  + Journal of Theoretical Biology (Arup’s suggestion, my and Navish’s opinion: might be too field-specific/theoretical)
  + Journal of Statistical Physics (Mehran’s suggestion, Arup says it won’t fit there)
* State that there is difference between real and simulated strain succession (linear/exp) but don’t explore further
* Can we find an upper bound for the inference error as function of parameter values and sample size? (for standard problems compare/cite Cocco & Monasson review: error sim partition function/sqrt(B))
* Add statement (not figure) about difference between HIV and influenza evolution
* Model schematic as figure is not necessary
* Move stringency plots to SI and combine left panels of N\_pop- and L-plot into one figure

4 May 2021 meeting with Mehran about Mount Fuji model results

Comments from Mehran (delivered in zoom mtg):

* Section on model: subsections
  + first specify how we describe sequences (binary of certain length), population of sequences, analogous to figure 1
  + second fitness with subsubsections on intrinsic and on host
  + last selection, equation 1
  + another small subsection on mutation
* in inference section
  + show first simulation results with figure 3
  + fig 3 suggest to us to look at the stringent selection condition
  + look at mount fuji landscape, same h at each site, no Jij, run model, some features may be preserved -> small section that stringent selection is preserved for this simple model
  + reference Desai and others for traveling wavefront type models
* intro is fine
* discussion: (M didn’t think about it too carefully yet)
  + MPL paper by Barton has a bit too much attention

10 May 2021 meeting with Mehran

What I wanted to discuss:

* I have done simulations of Mount Fuji model (constant h and J=0) for various values of h between -15 and +5
* inference generally somewhat underestimates the h values while overestimating J>0
  + I don’t see this bias in the usual simulations with p24 fitness coefficients
  + inference works (with underestimation of h) even for positive h
  + explanation: positive h is equivalent with negative h if zeros and ones are exchanged in sequences, and with different starting point (starting at intrinsically lowest-fitness sequence (0,0,0,...));
  + this means that evolution in the beginning is driven both by accumulating immunity and by intrinsic fitness gradient in the same direction; but if the highest-fitness sequence arises soon and replaces the low-fitness strains, evolution will go on as if starting with highest-fitness sequence and producing intrinsic fitness costs while escaping immunity
* The strain succession in Mt Fuji produces a higher diversity of strains in each season
  + Explanation: because there are many strains that are the same distance away from the starting strain and have equal intrinsic fitness, there are only L+1 different intrinsic fitness values
  + Note:   
    the low diversity and 'spindly' nature of influenza can not only be explained by long-range crossimmunity or deleterious mutation load as suggested before (see Yan et al. 2019, Koelle and Rasumussen 2015), but also by fitness differences being amplified by non-competing growth between bottlenecks (between seasons). Does that make sense?
* The fitness distributions in each season is highly quantized, again because of the few discrete fitness values that can be reached within a certain mutational distance

Conclusion from meeting with Mehran:

* The goal with the Mt Fuji analysis is to better justify (by analyzing this simple well studied model) the stringency condition which we assume for our model!!!
* The fitness scale that is approx. h, which shows up as width of intrinsic and host fitness distributions at each season, disappears in the total fitness distribution
  + choose variation of h-values around mean h -> show that scale of mean h disappears in distribution of Ftot
  + think about this stringency condition more analytically (within simple Mt Fuji model)
* Mehran has thought a bit more about the exponential versus linear increase of number of strains in HA data vs simulated data
  + The only reasonable explanation that he came up with is due to the rapid increase of sampled sequences since 1968 (which is presumably more relevant than the increase in the human (infected) population over the years)
* In Mt Fuji model with turned-off immunity there should emerge a steady state of a cloud of strains around the peak, where the equilibrium spread of the cloud depends on the ratio of mutation rate/slope of mountain (the slope is basically the deleterious h-values)
* Make plot of number of mutations as function of time for analysis of Mt Fuji model (probably not enough to say if it increases linearly or exponentially with time)
  + Question (from me, not shared yet): With the binary sequence representation I have L choose k numbers of strains with k mutations, which means that I have most strains with k=L/2 mutations and least (=1) with k=L or k=0
    - If seeing the fitness landscape as a single mountain, it will “broaden” first when going down from the peak, i.e. the number of strains with same fitness value (=same number of mutations) will increase, but will then narrow again when k>L/2
    - How does this combinatorial effect influence the increase of the number of mutations in the selected strains with time? Since long sequence simus the mutation number stays below L/2, this combinatorial effect should slow down the accumulation of mutations with time. If many strains are present that are equivalent with regard to intrinsic fitness but antigenically different they slow down each others accumulation of immunity-dependent fitness costs

20 May 2021 (thoughts and questions about selection regime)

* Will inference work, if only one strain per season gets selected? No, due to independently inferred F\* in each season, the value for Fhost + F\* of that single strain could be replaced by any number -> no information
* What happens if intrinsic fitness differences (h0 in mt fuji model) become very small between strains?
  + In a certain seasons we assume that strains get selected that are within a narrow range around the current fitness maximum (of concurrently competing strains)
  + If intrinsic fitness differences between strains (competing in the same season) are smaller (small fitness slope) than the fluctuations of Ftot between completely equivalent strains, their difference will have difficulty being resolved
  + -> The “cloud” of concurrently selected strains spreads further with decreasing (flatter) slope
  + The spread of the cloud is (for constant mutation rate and population size) determined by a fixed spread delta F in fitness space from the current maximum down to a certain minimum  
    if the intrinsic fitness difference h0 between strains is smaller than this fitness difference deltaF between concurrently selected strains, the intrinsic differences should not be resolved
  + Therefore with decreasing h0 there should be a threshold value, below which the inference method cannot reliably distinguish from h0=0
  + Small sampling size probably adds more noise to the host fitness (and total fitness), which will likely lead to a larger threshold value for |h0|, below which intrinsic fitness differences cannot be resolved
* What happens if the mutations rate is varied?
  + If the mutation rate is too small (for given Npop), i.e. if it is difficult to create even one sequence with one mutation between seasons, there will be many seasons where nothing happens besides immunity accumulation for the currently and previously present strains, due to this lack of mutations.
  + This slow sequence space exploration means that there will be on average less data for the same number of seasons than in the case of faster mutation
  + Due to this slow mutation, each present strain will also accumulate so much immunity that it will be overtaken by any new mutant as soon as it stochastically appears. (?)
  + In summary a too small mutation rate will likely lead to poor inference
  + On the other hand if the mutation rate is very large,  
    the population can completely randomly sample sequence space in any time step (between selections)
  + If stringent selection applies and the intrinsic fitness landscape is unimodal, the population can select the global fitness maximum (of Ftot) in each season (provided that population size is sufficiently large), i.e. it can generally make large jumps in the sequence landscape
  + It is unclear, what the effect of the large-mutation limit would be on inference. I guess it would still work. This could be tested by sampling completely random strains in the mutation step instead of using a specific mutation rate
* What happens if the population size is varied?
  + I have already done simulations which suggest that a decreasing population size decreases inference performance, so I could look more closely at the fitness distributions by plotting them for each population size to understand the dependence better
  + If the population size is very small, it can easily happen that some strains, although having equally high fitness to the maximum are not selected in the small population  
    This leads to a lack of data and therefore likely a worse inference
  + Still, if a strain gets selected it means that it likely has a fitness close to the current maximum, there just should be a few fitness-equivalent strains that get selected together
  + Also the mutation step will, like if the mutation rate is very small, lead to a low probability of a new mutant being created, which has the same negative effects on inference performance as a small mutation rate
  + Also the total fitness likely has larger fluctuations due to the discrete nature of Fhost which varies with frequencies x>1/Npop now in large steps
  + In summary I expect for small population size that selected strains/fitnesses vary stochastically and are more sparse, which leads to poor inference
  + On the other hand, if the population size is very large there is less stochasticity
  + But if the population size is too large, the stringent selection regime will be lost in the sense that (although with low frequency) low-fitness become selected alongside the high-fitness ones and since we disregard frequency from inference (using each selected strain in a given year as sample), we will likely make bad inferences for those rare strains, for which we wrongly assume that their total fitness is equal to the other concurrently selected strains
  + When looking at the fitness distributions for simulations with varying Npop, I find that the sampled fitnesses/strains become more sparse, but I don’t see larger fluctuations of Ftot with smaller Npop. Also the inference performance only increases with Npop for the investigated range of pop sizes, so I assume Npop is not yet large enough for rare low-fitness strains to be selected
* What happens if D0 is varied?
  + If there is no cross-immunity between strains, immunity will only accumulate for the strains that have themselves been selected previously, I assume it will make the inference not worse if not easier
  + If cross-immunity is very large, immunity will accumulate much faster than the population can escape, and the immunity differences between adjacent strains will be minute
  + I can imagine that such wide-ranging cross-immunity will be bad for inference since selection will not be able to favor new strains easily due to their accumulated small immunity differences to old strains, especially the parameters F\* will blow up and their disproportionate magnitude to the other parameters might mess up inference
* In reality, where survival of the population is not ensured with fixed population size as opposed to our model, several factors will lead to extinction instead of just affecting the values of F\*
  + A too large cross-immunity range leading to a faster accumulating fitness cost than can be escaped by mutation will lead to extinction eventually (if there host population turnover is not considered)
  + A too small mutation rate (together with a small maximum population size) will similarly lead to extinction, since immunity can accumulate before a chance of escape
* Why does the average mutation number increase slower than linearly with time (in the fuji model)?
  + One aspect is that the number of strains with equivalent fitness will increase with the number k of mutations up to L/2 as L choose k

27 May 2021 (meeting with Arup and Mehran about Arup’s manuscript comments)

* Arup says fitness in equilibrium Ising type models is exp(F)/sum[exp(F)] not F itself from our model
* He also says that our description is not Malthusian (I agree the fitness F is not malthusian but the growth is: x(t+1) sim exp(F)x )
* Proposed change (suggested by arup):
  + Don’t use equilibrium models as motivation for our fitness description, just postulate it as it is
  + But I may refer to Laessig since he also uses fitness F like we and then growth with exp(F) as in our model
  + For parameters from p24 also just state what we use, don’t motivate with equilibrium models

28 May 2021 (plan for doing repeat simulations/analyses)

* update definition of fitness coefficients (without modifying min and max of h and J)
* for each explored parameter range (L, NPop) run maybe 5-10 independent reps (with different rng initialization) producing individual collection of files in designated folder for specific param variation,  
  then run analysis (as before) on each of those file collections independently, only using analysis parameter ranges that are needed for plots
* for plots (collection analysis) load and average data from all repetitions, save plots in figure folder
* Will I need to also run several reps on random sampling? Probably not

3 June 2021 (repeat simulations, one rep for now but extendable)

* DONE: Update fitness coefficient definition (without modifying min/max of h and J):  
  (h and J are inferred for a segment of the p24 protein with 105 sites, and for simulations for shorter sequences the number of coefficients is reduced, respectively.  
  Update by commenting out the code where min and max are replaced)
* DONE Name file folder “simu\_name” according to content (plus simulation data), e.g.   
  “strdate\_today” + [“\_varying” + param\_name for each varied param]
* DONE Make rng seed different for each repetition:  
  in add\_parameters: choose the seed as np.random.randint(10\*\*6) without predefining a seed for this rng -> this (random) seed will be saved in file simu\_info
* Run simulations (1 rep for now)
  + RUNNING With varying Npop
  + RUNNING With varying L
* For analysis of repeated simus: Each repetition of the same parameter variation will create a folder that has the varied parameter name in the folder name:
  + If the analysis results of several reps should be averaged, I can load the analysis file (with summary results) as well as the simu\_info file from each of those reps
  + Then check that the dictionaries exp\_dict and exp\_ana\_dict are the same in each of the files, and that the seed is different
  + Then extract and average the result quantities which I want to plot, also create a new error for each quantity from (mean standard error) due to the independent reps instead  
    of the error used before for errorbars
  + Make plots and save in figure file