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

26 April meeting

What I want to discuss:

* Which figures (parameter explorations) should we keep? -> 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 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?)

Conclusion from meeting: