# Response Letter for "Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution" (15-06-2020-RA-eLife-60067)

Dear eLife editorial board and reviewers,

Thank you for the enthusiastic support for this manuscript and the detailed revisions and minor points. We have responded to these issues inline below, providing specific new text from the manuscript or additional figures to clarify our responses.

Sincerely,

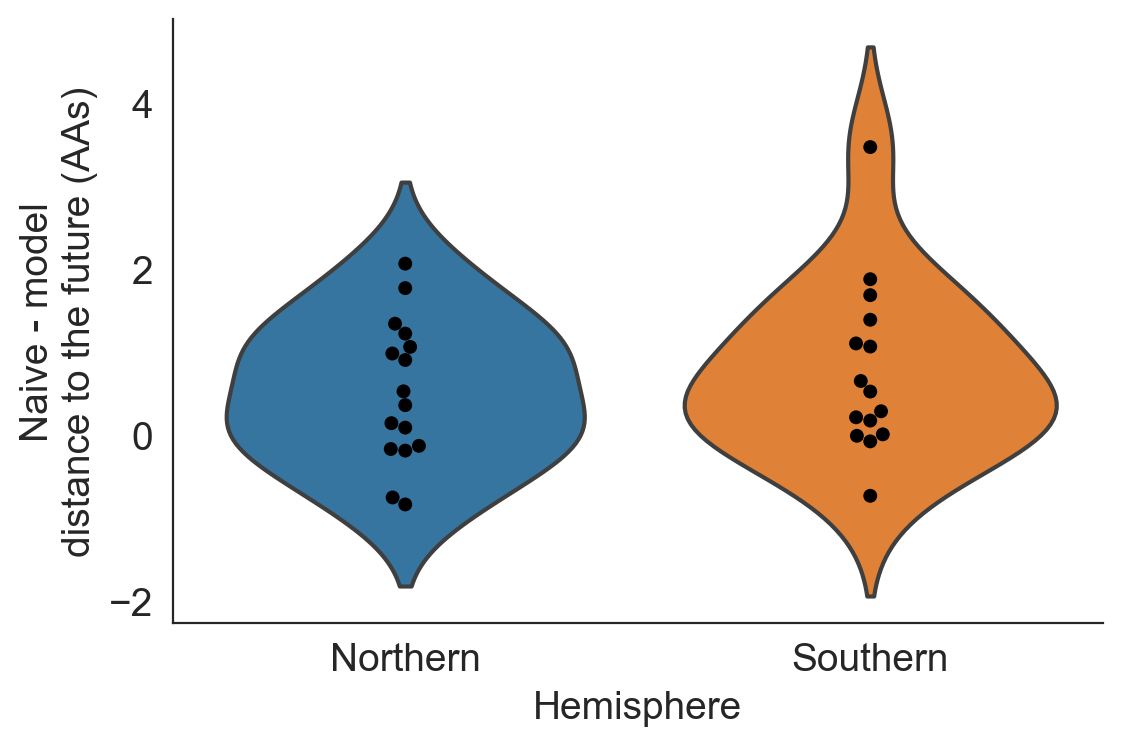
John Huddleston

## REVISIONS

**1. It would be helpful to further explore model performance in different contexts.**

**(a) Overall, do the models perform any better or worse at predicting Northern v Southern hemisphere viruses populations?**

We investigated whether our best model, HI antigenic novelty and mutational load, projected better estimates of the future for the Northern Hemisphere (forecasts made in October) or the Southern Hemisphere (forecasts made in April). To account for natural variation in distances between timepoints, we calculated the difference between the naive model’s distance to the future and our best model’s and plotted these values by Hemisphere (Response Figure 1). The median adjusted distance to the future for the Northern and Southern Hemispheres were 0.45 and 0.53 AAs, respectively. These results suggest that there is not a substantial difference in our ability to forecast based on the Hemisphere.

Response Figure 1. Distribution of distances to the future by Hemisphere for the best natural model (HI antigenic novelty and mutational load) subtracted from the corresponding distance to the future for the naive model at the same timepoint.

**(b) I’m surprised the authors don’t discuss more the recent problem of multiple H3 cocirculating clades (observed in Figure 9). Predicting H3 evolution has always been difficult, but at least there was generally a linear tree and a single dominant H3 clade at any given time. Did certain models better predict the emergence and persistence of this tree pattern? Do they offer insights?**

We agree that this was an oversight on our part to not mention the unusual diversity of H3N2 clades in the last decade. We have modified the following paragraph in the Discussion (lines 487-499) to mention this pattern and its potential role in the performance of our models (new text in bold):

Even the most accurate models with few parameters will sometimes fail due to the probabilistic nature of evolution. For example, the model with the best performance across our validation data — mutational load and LBI — was also one of the worst models across our test data. **Although we cannot rule out the role of overfitting, this model’s poor performance coincided with unusual evolutionary circumstances. The diversity of H3N2 lineages during our test period was higher than the historical average (Koelle et al., 2006), with the most recent common ancestor of all circulating strains dating eight years back. This persistence of diversity may have reduced the effectiveness of the LBI metric that assumes relatively rapid population turnover. Additionally, this model’s poorest performance occurred in 2019 when it failed to predict the sudden decline of a dominant reassortant clade, A2/re. Only our models based on HI antigenic novelty and mutational load continued to perform as well or better than the naive model during the same time period. These results highlight the challenge of identifying models that remain robust to stochastic evolutionary events by avoiding overfitting to the past.**

As we note in this new text, we cannot disentangle model overfitting from changes in underlying evolutionary processes. This limitation unfortunately prevents us from identifying models that predicted the emergence or persistence of H3N2's current diversity. However, we can identify the models that are most robust to both overfitting and evolutionary change. We hope that this revised text clarifies these complexities and the reasons for selecting the HI antigenic novelty and mutational load model as our best model.

**2. Given that this study tackles a very real world problem (selecting strains for influenza vaccines), it would be helpful to have these results better translated for readers with public health backgrounds.**

**(a) A simple addition would be an opening table/chart that describes the different models and categorizes them (lab data, sequence-based, tree-based etc.)**

We have added the following table to the first section of the Results (Table 1) as a summary of all primary models used with the simulated and natural populations. We reference this new table from the initial list of models in the same section (Line 117).

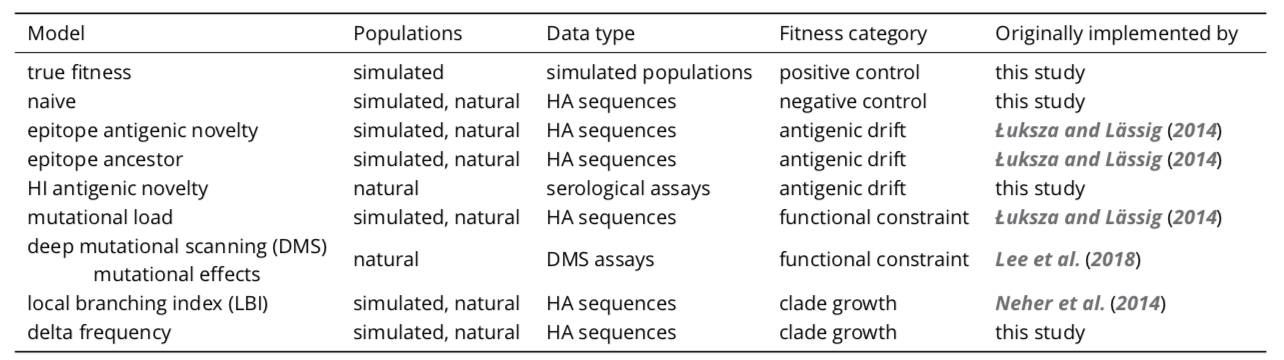
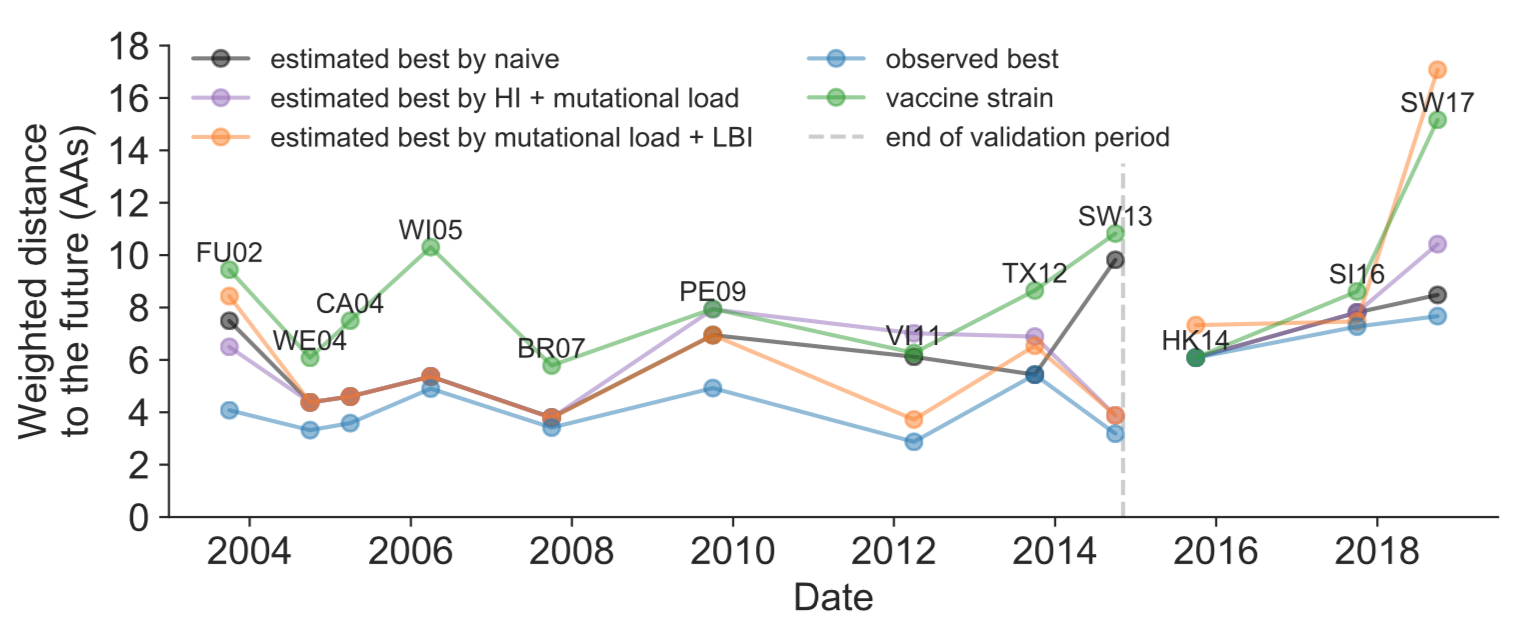


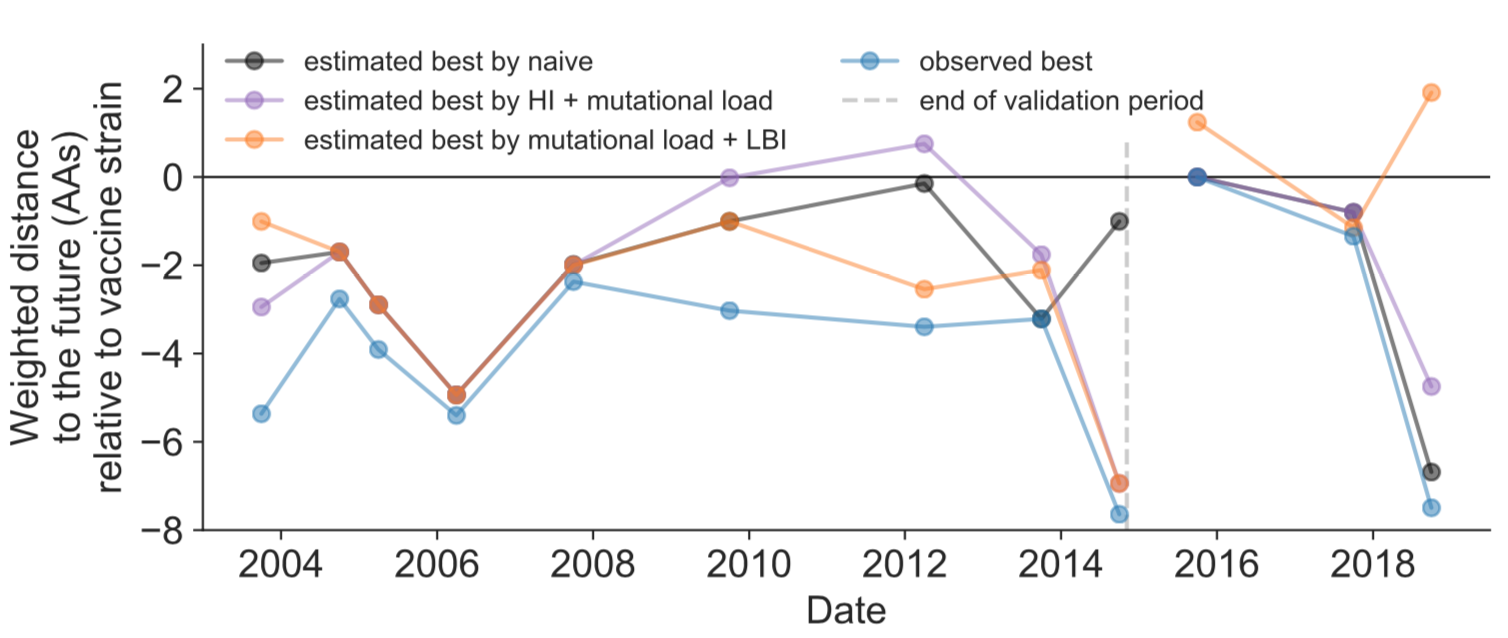
Table 1. Summary of models used with simulated and natural populations. Models are labeled by the type of population they were applied to, the type of data they were based on, and the component of influenza fitness they represent.

**(b) Figure 8 nicely visualizes model performance against actual vaccine strains. But can you quantify/summarize the results in this figure better (ie, exactly how much ‘closer’ to the future than the vaccine strain?). Perhaps also including additional models in those measures?**

To better represent and quantify how much closer to the future the observed and estimated closest strains are than the selected vaccine strains, we calculated the distance to the future of these strains relative to the vaccine strain at the corresponding timepoints and plotted these values in a new figure supplement (Figure 8-Figure supplement 1). Strains with relative distances greater than zero were farther from the future than the vaccine strain. We have also added the closest strain selected by the naive model to Figure 8 and its figure supplement 1, as an additional reference point for the strains selected by biologically-informed models. These two figures are included below as Response Figures 2 and 3, respectively.



Response Figure 2. Observed distance to natural H3N2 populations one year into the future for each vaccine strain (green) and the observed (blue) and estimated closest strains to the future by the mutational load and LBI model (orange), the HI antigenic novelty and mutational load model (purple), and the naive model (black). Vaccine strains were assigned to the validation or test timepoint closest to the date they were selected by the WHO. The weighted distance to the future for each strain was calculated from their amino acid sequences and the frequencies and sequences of the corresponding population one year in the future. Vaccine strain names are abbreviated from A/Fujian/411/2002, A/Wellington/1/2004, A/California/7/2004, A/Wisconsin/67/2005, A/Brisbane/10/2007, A/Perth/16/2009, A/Victoria/361/2011, A/Texas/50/2012, A/Switzerland/9715293/2013, A/HongKong/4801/2014, A/Singapore/Infimh-16-0019/2016, and A/Switzerland/8060/2017.



Response Figure 3. Relative distance to future H3N2 populations between vaccine strains and corresponding observed and estimated closest strains at each timepoint as in Figure 8. Strains with relative distances greater than zero were farther from the future than the selected vaccine strain, while strains below zero were closer to the future.

We have updated the corresponding section of the Results (Lines 388-411) to include a clearer description of how these closest strains are calculated, what they represent, and how much closer to the future each strain is than the vaccine strain (new text in bold):

For each season when the WHO selected a new vaccine strain and one year of future data existed in our validation or test periods, we measured the observed distance of that strain’s sequence to the future and the corresponding distances to the future for the observed closest strains **(Equation 3)**. We compared these distances to those of the closest strains to the future as estimated by our best models for the validation period (mutational load and LBI) and the test period (HI antigenic novelty and mutational load) **using Equation 4. The observed closest strain to the future represents the centroid of the observed future population, while the estimated closest strains are the models’ predictions of that future population’s centroid.** The mutational load and LBI model selected strains that were as close or closer to the future than the corresponding vaccine strain for 10 (83%) of the 12 seasons with vaccine updates (Figure 8). **On average, the strains selected by this model were closer to future than the vaccine strain by 1.93 AAs (Figure 8-Figure supplement 1).** For the two seasons that the model selected more distant strains than the vaccine strain, the mean distance relative to the vaccine strain was 1.58 AAs. The HI antigenic novelty and mutational load model performed similarly by identifying strains as close or closer to the future for 11 (92%) seasons **with an average improvement over the vaccine strains of 2.33 AAs**. For the one season that the model selected a more distant strain, that selected strain was 0.75 AAs farther from the future than the vaccine strain. **Interestingly, the strains selected by the naive model were always better than the selected vaccine strain. Since the naive model predicts that the future will be identical to the present, these strains represent the centroid of each current population. With an average improvement over the vaccine strains of 2.19 AAs, the naive model performed consistently better than the LBI-based model and nearly as well as the HI-based model. These results were consistent with our earlier observations that the naive model often performs as well as biologically-informed models when estimating a single closest strain to the future.**

**(c) In the Introduction can you provide more context for this study? What is the current range of influenza vaccine effectiveness? How frequent are H3 mismatches? And how well have we been trending at matching H3 vaccine strains to H3s in circulation? Are we making any discernible progress? Or have improvements in modeling been offset by the H3 cocirculating clades problem?**

We have modified the Introduction, splitting the first paragraph into two paragraphs (the first introducing seasonal influenza A/H3N2 and the second introducing predictive methods) and adding a new paragraph in between that summarizes vaccine effectiveness for H3N2 and factors affecting this effectiveness. This new paragraph is included below with the last sentence of the preceding paragraph and first sentence of the following paragraph included for context (new text in bold):

However, because the process of vaccine development and distribution requires several months to complete, optimal vaccine design requires an accurate prediction of which viruses will predominate approximately one year after vaccine viruses are selected.

**Historically, the vaccine effectiveness of the H3N2 vaccine component has been much lower than the other seasonal influenza subtypes. For example, H3N2’s mean vaccine effectiveness from 2004–2015 was 33% compared to 61% for H1N1pdm and 54% for influenza B viruses (Belongia et al., 2016). Multiple factors can reduce vaccine effectiveness including selection of a vaccine strain that is not antigenically representative of future populations (Belongia et al., 2016; Gouma et al., 2020) and adaptations of the selected strain to egg-passaging during vaccine production that alter the antigenicity of the resulting vaccine component (Zost et al., 2017b). Even when vaccine strains are well-matched antigenically, they may fail to induce a strong immune response due to previous infection history of vaccine recipients (Cobey et al., 2018). While all of these factors must be addressed to increase vaccine effectiveness, substantial effort has focused on the selection of the most representative strain for the next season’s vaccine.**

Current vaccine predictions focus on the hemagglutinin (HA) protein, which acts as the primary target of human immunity.

**(d) A key message from this study is how challenging prediction H3 is, even with new analysis tools and new types of experimental data. We have so much further to go. It’s worth highlighting in the Discussion how CDC FluSight has collectively advanced epi flu forecasting by making weighted ensembles drawn from multiple modeling groups and how valuable a similar program would be for vaccine strain selection.**

We have updated the following sentence in the Discussion (Lines 505-507) to explicitly mention the CDC’s FluSight network (new text in bold):

The recent success of weighted ensembles for short-term influenza forecasting **through the CDC’s FluSight network** (Reich et al., 2019) suggests that long-term forecasting may benefit from a similar approach.

**3. Some epitopes are more important than others. Did you consider weighting epitopes differently? Or would that just exacerbate overfitting? And how do you handle glycosylations in the model?**

Prior to our discovery of overfitting related to previously defined epitope sites, we had planned to test models fit to weighted epitope sites (for example, assigning more weight to Koel et al.’s seven sites than others). In light of model performance based on unweighted sites, we reasoned that weighting sites would most likely increase overfitting.

We chose not to fit models based on glycosylation sites based on two lines of reasoning. As these sites accumulate slowly in H3N2 relative to epitope mutations (six sites between 1968 and 2012), we expected there would not be a strong signal in year-to-year forecasts despite the biological importance of these events when they do occur. We also considered the results of Łuksza and Lässig’s original analysis which found that adding glycosylation sites did not improve their full fitness model and that these sites appeared correlated with epitope mutations.

As we briefly describe in the Discussion (Lines 524-530), we anticipate that antigenic escape assays like those presented in Lee et al. 2019 (<https://elifesciences.org/articles/49324>) could allow us to identify and weigh contemporary epitope sites. How to effectively translate these antigenic escape data to fitness metrics is an active area of research in the Bedford lab.

## MINOR POINTS

**1. It would be helpful to clarify how “strain” and “clade” are defined and used throughout the manuscript.**

We have added the following sentence to the Introduction on lines 38-40 to define the meaning of the term “strain”:

The World Health Organization (WHO) Global Influenza Surveillance and Response System (GISRS) monitors influenza evolution by sampling currently circulating viruses, or strains, and analyzing these strains with genome sequencing and serological assays.

We have added the following sentence to the Results on lines 99-101 to define the meaning of the term “clade”:

Łuksza and Lässig (2014) measured model performance by identifying clades – groups of strains that all share a recent common ancestor – and comparing observed and estimated future clade frequencies.

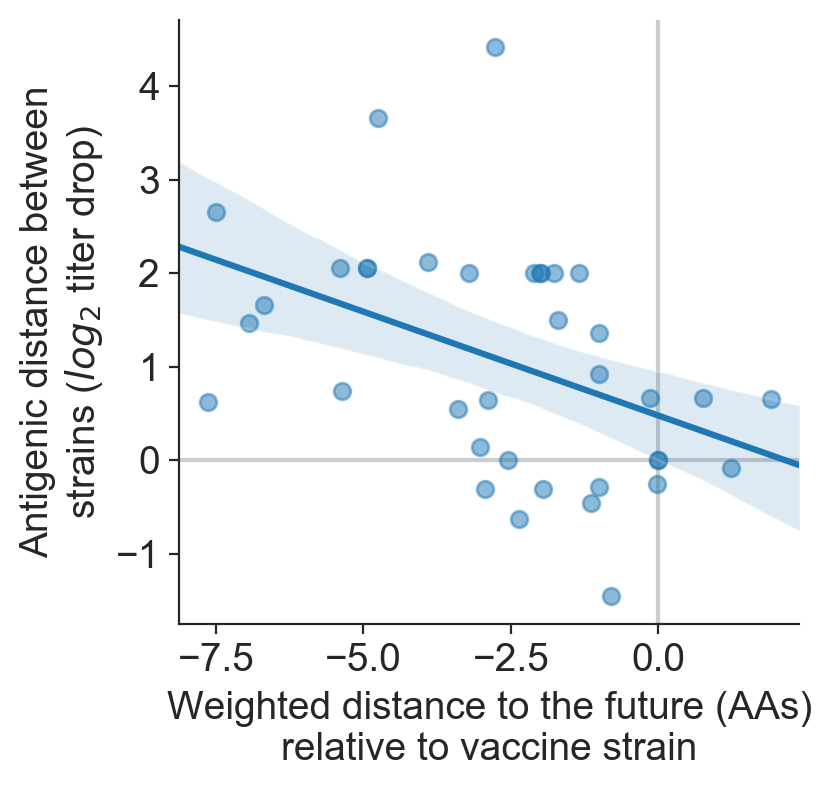
**2. A very brief description of earth mover’s distance and its application here would make the manuscript more accessible to a broader audience.**

We have expanded the first paragraph of the Results section to include the following description of earth mover’s distance and justification of its use:

However, as clade definitions are inherently unstable between seasons, we evaluated our models by comparing the genetic composition of observed and estimated future populations with the earth mover’s distance metric. The earth mover’s distance calculates the minimum distance between two populations, given the frequency of each individual within a population and a pairwise “ground distance” between individuals (Rubner et al., 1998). We defined distinct amino acid haplotypes as individuals in our observed and estimated future populations. For frequencies of individuals, we used the observed frequencies of haplotypes in the future and our model’s estimated frequencies. We calculated the ground distance between individuals as the Hamming distance between haplotypes. With this implementation, more accurate projections of the future population’s composition produce smaller earth mover’s distances between the observed and estimated future (Figure 1).

**3. The model strains in Figure 7 are more similar to the future global population than those chosen for the vaccine. Do any of them differ antigenically from the vaccine strain?**

We determined whether any observed or estimated closest strains to the future differed antigenically from the corresponding vaccine strain using HI measurements that were normalized to relative log2 values as described in the Methods (Lines 614-616). For each closest strain to the future, we looked for its titer measurements against the corresponding vaccine strain. If no measurements existed between a given pair, we identified the first strain that had titer measurements against the vaccine strain and also had the most similar amino acid sequence to the closest strain. For 29 out of 37 non-redundant closest strains (78%), we found identical sequences to the closest strain. The remaining eight strains (22%) differed from the corresponding closest strains by one amino acid. We plotted the log2 distance between these strains and their vaccine strains against their relative distances to the future (as shown in Response Figure 3 above) and annotated a linear fit (Response Figure 4).



Response Figure 4. Antigenic distance between non-redundant observed or estimated closest strains to the future (N=37) and the corresponding vaccine strain (log2 titer drop) by relative distance to the future (AAs).

From these data, we find that strains that are closer to the future population than the corresponding vaccine strain (i.e., strains with more negative relative distances) also tend to be more antigenically distinct from the vaccine strain. It is important to note, though, that individual HI measurements are inherently noisy and the range of antigenic distances above represents this variance. For example, all antigenic distances less than zero indicate comparisons where serum against the reference virus inhibited the test virus better than the reference virus itself. Based on this reviewer point, we have realized that an important step in future influenza forecasting work could be the use of antigenic metrics as either a "ground distance" for our earth mover's distance calculations between populations, a per-strain measure of distance to the future for the selection of vaccine strains, or both.

**4. Many of the proposed models do better than the naive model in terms of minimizing earth mover’s distance from the observed future population composition. The authors also do a nice job showing that their approach predicts future strain frequencies reasonably well. However, it is unclear what effect a 0.85 lower EMD has on predicting the best strain. It would be useful to discuss how the naive model performs in predicting the frequency rank of future strains in the text (Figures S6, S7,S12 & S13). This comparison provides a baseline for figures 4 and 7 and indicates how much better the models are at predicting vaccine strains than the naive model. Which model best predicts the strains selected at WHO meetings? Antigenic novelty?**

To provide more context for the performance of our biologically-informed models, we have added descriptions of the naive model's performance for both simulated and natural populations. The specific additions to the text are included in detail below. Note that Supplemental Figures S6, S7, S12, and S13 are now labeled as Figure 4-Figure supplement 2, Figure 4-Figure supplement 3, Figure 7-Figure supplement 2, and Figure 7-Figure supplement 3.

In addition to the description of validation results for simulated populations, we added the following text (Lines 211-219) describing the naive model's performance on the validation data (new text in bold):

Percentile ranks per strain based on their observed and estimated distances to the future correlated strongly across all strains and timepoints (Spearman’s ⍴2 = 0.87, Figure 4-Figure supplement 1D). **In contrast, the naive model’s forecasts of clade frequencies were considerably less accurate (Figure 4-Figure supplement 2C). However, the naive model’s estimated closest strains to the future were consistently in the top fifth percentile of observed distances to the future and the correlation of its estimated percentile ranks and the observed ranks was strong (Spearman’s ⍴2 = 0.78, Figure 4-Figure supplement 2B and D). These results suggested that estimating a single closest strain to the future is a more tractable problem than estimating the future frequencies of clades.**

In addition to the description of test results for simulated populations, we added the following text (Lines 238-247) describing the naive model's performance on the test data (new text in bold):

Observed and estimated strain ranks remained strongly correlated across all strains and timepoints (Spearman’s ⍴2 = 0.80, Figure 4D). **The naive model performed comparatively well on these test data with all its estimated closest strains to the future in the top 20th percentile and a slightly higher correlation between observed and estimated percentile ranks than the true fitness model (Spearman’s ⍴2 = 0.82, Figure 4-Figure supplement 3).** These results confirmed that our approach of minimizing the distance between yearly populations could simultaneously capture clade-level dynamics of simulated influenza populations and identify individual strains that are most representative of future populations. **However, they also supported the earlier finding that clade frequency forecasts may be inherently more challenging than identification of the closest strain to the future.**

In addition to the description of validation results for natural populations, we added the following text (Lines 350-355) describing the naive model's performance on the validation data (new text in bold):

This pattern held across all strains and timepoints with a strong correlation between observed and estimated strain ranks (Spearman’s ⍴2 = 0.66, Figure 7-Figure supplement 1D). **The naive model’s performance repeated the pattern we observed with simulated populations: it made poor forecasts of absolute clade frequencies, but its estimated closest strains to the future were consistently highly ranked among the observed closest strains (Figure 7-Figure supplement 2B and C).**

In addition to the description of test results for natural populations, we added the following text (Lines 382-386) describing the naive model's performance on the test data (new text in bold):

Similarly, the observed and estimated strain ranks strongly correlated (Spearman’s ⍴2 = 0.72) across all strains and test timepoints (Figure 7D). **The estimated strain ranks of the naive model were not as well correlated (Spearman’s ⍴2 = 0.56), but seven of its eight estimates for the closest strain to the future (88%) were in the top fifth percentile of observed closest strains (Figure 7-Figure supplement 3B and D).**

In addition to these descriptions, we have added the naive model's best picks to the vaccine comparison analysis (Response Figure 2) and referenced all of these results from the vaccine strain comparison section of the Results (Lines 398-409). This modified figure and the new Figure 8-Figure supplement 1 (described in response to Revision 2b above) explicitly quantifies how much better the biologically-informed models are at predicting vaccine strains than the naive model.

We have also added the following sentence to the Discussion (Lines 447-451), summarizing the general finding that the naive model performs well when selecting a single representative strain of the future (new text in bold):

We demonstrated that the integration of these phenotypic metrics with previously published sequence-only metrics produces more accurate forecasts than sequence-only models. **Interestingly, we found that a naive model that predicts no change over the course of one year can often identify a single representative strain of the future despite its inability to accurately forecast clade frequencies.**

Regarding the question of which model best predicts the strains selected at the WHO's vaccine composition meetings, we find that no single model consistently selects strains like those selected by the WHO. Although WHO strain selection has been historically informed by serological assay data like HI measurements, the use of formal models to synthesize and summarize these data is relatively recent and past decisions have relied more on the knowledge and discussions among experts. In this sense, the selections by the WHO represent predictions by a type of expert or judgmental forecasting model.

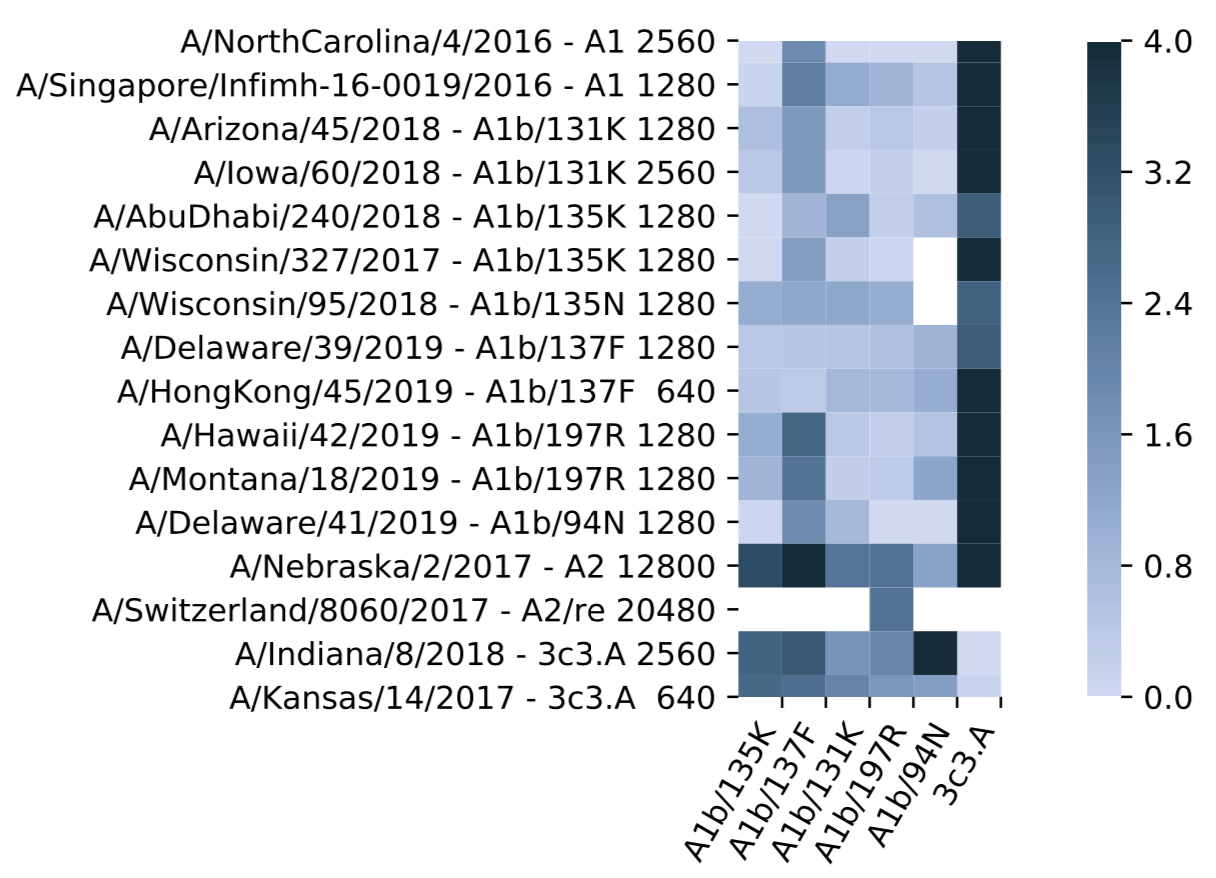
**5. Line 557: ‘We prioritized strains with more available HI titer measurements.’ Can you clarify whether the HI titers came from GISAID metadata?**

We have revised the original sentence to clarify the source of the HI titer measurements. The sentence (Lines 625-626) now reads (new text in bold):

We prioritized strains with more available HI titer measurements **provided by the WHO GISRS Collaborating Centers**.

**6. In Figure 9, how much cross-reactivity is there between different clades?**

Most extant clades in the 3c2.A lineage cross-react in focus-reduction neutralization assays (FRAs) based on ferret antisera (light blue squares in Response Figure 5). The exception is the A1b/137F clade which is more antigenically distinct from its sibling clades (darker blue squares in the A1b/137F column of Response Figure 5). [The most recent H3 vaccine component was a virus from this clade](https://nextstrain.org/flu/seasonal/h3n2/ha/2y?s=A/HongKong/2671/2019). The 3c3.A clade is the least cross-reactive with other circulating clades, as seen by its consistent 4-fold mean log2 titer drop relative to antisera from viruses in the 3c2.A lineage.



Response Figure 5. Antigenic distance (mean log2 titer drop) between test strains sampled from recently circulating clades (columns) and antisera for representative viruses from these clades (rows). The median autologous raw titer of antisera for representative viruses is shown after the corresponding virus's name and clade membership. White squares represent missing data.

**7. Does anyone ever talk about including 2 H3 strains in the vaccine? Clearly there will be times when the model predicts two clades rising in frequency like in Fig 9?**

Multivalent H3 vaccines have been considered and discussed. Any vaccines with an additional H3 valency would need to show superior performance over the current vaccine in clinical trials, avoid immunodominance of one valency over the other, and provide a benefit that outweighs the increased cost of producing a new vaccine. If the current diversity of influenza A/H3N2 lineages persists and 3c3.A continues to circulate as an antigenically distinct lineage, there may be increasing benefits to a vaccine with multiple H3 components. The introduction of modern quadrivalent influenza vaccines with two influenza B representatives provides a precedent for this kind of change.

The complexity and difficulty of selecting a single representative H3 strain for the vaccine highlights the limitations of automated forecasts and the continued importance of expert knowledge in the final vaccine composition decisions.

**8. The authors include a line in the Discussion that other segments of the genome need to be considered in forecasts. The line is similar to ines that have been repeated at the end of HA-focused studies for over a decade with little practical effect. If mutational load in the HA is as important in viral forecasts as suggested in this study, it doesn’t take a lot of imagination to realize how much of the picture we’re missing by just focusing on HA antigenic evolution. Maybe worth elaborating a little on feasibility, including the availability of WGS data?**

We have added the following text to our discussion of models based on multiple gene segments (lines 533-540), to summarize technical and scientific feasibility:

Our forecasting framework makes the inclusion of fitness metrics based on additional gene segments technically straightforward. However, the definition of appropriate fitness metrics for neuraminidase and other genes remains an important scientific challenge. An additional challenge to model training is a relative lack of historical strains for which all genes have been sequenced. Of the 34,312 H3N2 strains in GISAID with all eight primary gene segments and collection dates between October 1, 1990 and 2019, the majority (24,466 or 71%) were collected after October 1, 2015. Data availability will therefore inform which gene segments are prioritized for inclusion in future models.

**9. The legend title in figure 10 is cut off.**

We have corrected the legend in Figure 10 to display the full title of “Distance to future population (by HI antigenic novelty and mutational load)“.