Improved forecasts and visualization of seasonal influenza evolution

John Huddleston

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Trevor Bedford, Chair

Jesse Bloom

Harmit Malik

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Improved forecasts and visualization of seasonal influenza evolution

John Huddleston

Chair of the Supervisory Committee:
Associate Professor Trevor Bedford
Vaccine and Infectious Disease Division

The rapid evolution of seasonal influenza requires the development of new vaccines every one to two years. This evolution occurs through a process of antigenic drift where amino acid mutations in the hemagglutinin surface protein allow currently circulating viruses to evade adaptive immunity against previous viruses. Vaccine composition decisions are guided by predictions made from serological assays of antigenic drift and sequence-based forecasting models. These predictions do not account for functional effects of mutations measured by deep mutational scanning experiments or attempt to integrate fitness effects measured by experimental and sequence data. In this dissertation, I attempted to understand whether experimental measurements of antigenic drift and functional constraint could be used to improve forecasts of seasonal influenza evolution. I found that most estimates of seasonal influenza fitness could not robustly forecast future populations. Models that integrated serological measurements of antigenic drift with sequence-based estimates of functional constraint provided the most robust forecasts. I concluded that successful seasonal influenza predictions depend on the choice of prediction targets and fitness metrics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Thank you to Dr. Trevor Bedford for giving me the opportunity to work on challenging projects and the time and support to find my way through them. Thank you to my committee for guiding my research and career in the right direction when I was stuck in the weeds. Thank you to my excellent lab mates and collaborators for your constant examples of how to be excellent scientists and humans with humor and humility. I could not have done this without you all.

DEDICATION

To my parents who set me on this path,
to my wife who led the way,
and to my daughter who gave me every reason to keep going.

Chapter 1

INTRODUCTION

1.1 Why do we try to predict seasonal influenza evolution?

Seasonal influenza infects 5–15% of the global population yearly causing 100,000s of deaths and influenza subtype A/H3N2 is responsible for the bulk of human mortality and morbidity [World Health Organization, 2009]. Influenza virus naturally evolves to escape acquired immunity in the human population and this evolution results in loss of vaccine efficacy over time as the virus evolves away from the chosen vaccine strain. This process of *antigenic drift* necessitates yearly selection of vaccine strains by the World Health Organization (WHO) [Smith et al., 2004]. Manufacture and distribution of the influenza vaccine takes almost one year and the vaccine can contain only one representative strain per seasonal influenza subtype (A/H3N2, A/H1N1pdm, B/Victoria, and B/Yamagata) [Buckland, 2015]. As a result, WHO officials must predict which currently circulating A/H3N2 strain will be more representative of the influenza population one year in the future. The predictions required for vaccine development have been historically challenging, but recent improvements in sequencing throughput and developments in computational models have made the prediction of influenza virus evolution more tractable [Lässig et al., 2017, Morris et al., 2017]. The better these predictions are, the more likely the vaccine will prevent illness and death from infection.

1.2 How do we think seasonal influenza evolves?

Globally successful seasonal influenza viruses are usually antigenically distinct from previous lineages [Smith et al., 2004]. Thus, antigenic drift is an important predictor for influenza

surveillance and vaccine recommendations [Morris et al., 2017]. Antigenic drift occurs through mutations to the hemagglutinin (HA) surface protein that abrogate binding of preexisting human antibodies. Antigenic phenotype is most commonly measured through the hemagglutination inhibition (HI) assay, which quantifies the extent to which antisera blocks attachment of viruses to red blood cells [Hirst, 1943]. HI assays are the gold standard for measuring antigenic drift phenotypes, but these experiments are typically low-throughput and laborious compared to modern genome sequencing [Wood et al., 2012]. Thus, researchers have attempted to predict viral success by estimating antigenic drift from HA genome sequences alone [Łuksza and Lässig, 2014, Steinbrück et al., 2014, Neher et al., 2014].

Seasonal influenza viruses rapidly accumulate mutations during replication, due to their error-prone RNA polymerase [Petrova and Russell, 2018]. For most genes, most new amino acid mutations will weaken the functionality of their corresponding proteins and reduce viral fitness. An exception to this rule are amino acid mutations in HA or the other primary surface protein, neuraminidase (NA), that modify binding sites of host antibodies from previous viral exposure. These mutations contribute to antigenic drift and increase viral fitness by allowing viruses to escape existing antibodies (Figure 1.1). Thus, mutations in HA and NA create fitness trade-offs, where beneficial mutations facilitate antigenic drift against a background of deleterious mutations [Koelle and Rasmussen, 2015].

Viruses carrying beneficial mutations should grow exponentially relative to viruses lacking those mutations (Figure 1.2A) [Neher, 2013]. Beneficial mutations on different genetic backgrounds will compete with each other in a process known as clonal interference (Figure 1.2B). If beneficial mutations have large effects on fitness, the fitness of the genetic background where the beneficial mutations occur is less important for the success of the virus than the fitness effect of the beneficial mutations themselves (Figure 1.3). If beneficial mutations have similar, smaller effects on fitness, a virus's overall fitness depends on the effect of the beneficial mutations and the relative fitness of its genetic background. In this case, the ultimate success and fixation of these beneficial mutations depends, in part, on the number of deleterious

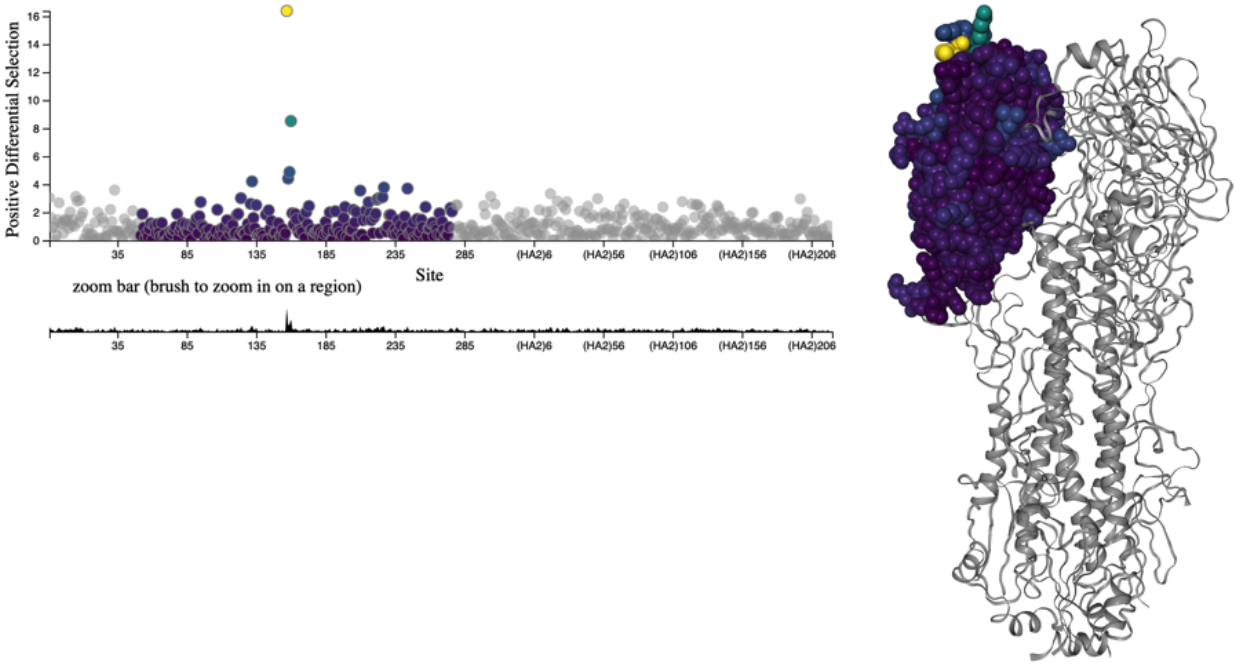

Figure 1.1: **HA accumulates beneficial mutations in its head domain (sites with color) that enable escape from antibody binding and deleterious mutations in its stalk domain (sites in gray) that reduce its ability to infect new host cells.** The linear genome view on the left shows how sites from HA's head domain map to the three-dimensional structure of an HA trimer. The site highlighted in yellow reveals where different amino acid mutations allowed a seasonal influenza virus to escape binding from existing antibodies in a human's polyclonal sera [Lee et al., 2019]. Explore this figure interactively with dms-view.

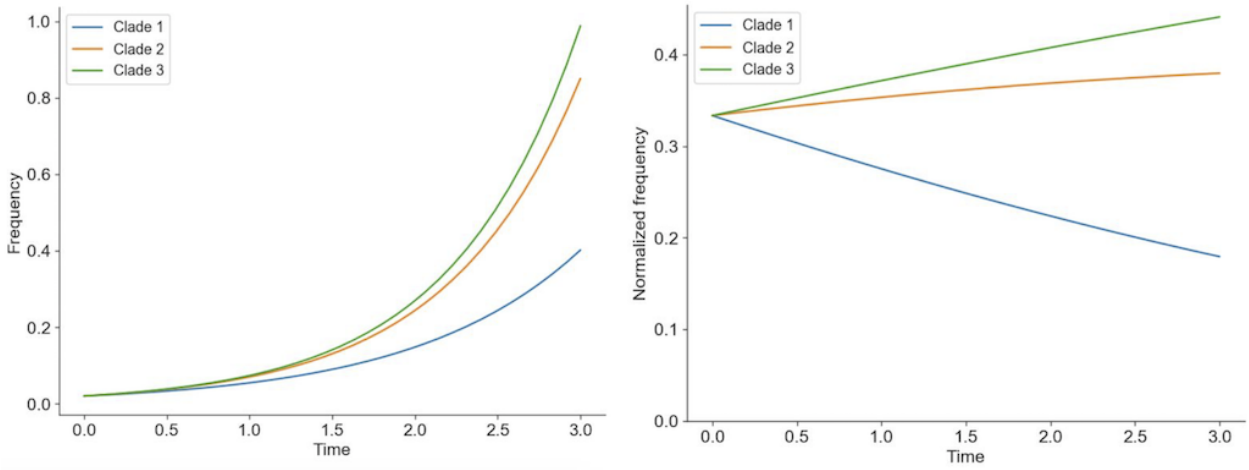

Figure 1.2: **Clonal interference constrains exponential growth of asexually reproducing populations.** Individuals in asexually reproducing populations tend to grow exponentially relative to their fitness (left). Normalization of frequencies to sum to 100% represents competition between viruses for hosts through clonal interference and reveals how exponentially growing viruses can decrease in frequency when their relative fitness is low (right).

mutations that already exist in the same genome (Figure 1.4).

1.3 What is predictable about seasonal influenza evolution?

The expectations from population genetic theory described above and previous experimental work suggest that aspects of seasonal influenza's evolution might be predictable. Mutations in HA and NA that alter host antibody binding sites and enable viruses to reinfect hosts should be under strong positive selection. We expect these strongly beneficial mutations to sweep through the global seasonal influenza population at a rate that depends on the importance of their genetic background. We also do not expect that every site in HA or NA will acquire beneficial mutations. For example, fewer than a quarter of HA's 566 amino acid sites are under positive selection [Bush et al., 1999], have undergone rapid sweeps [Shih et al.,

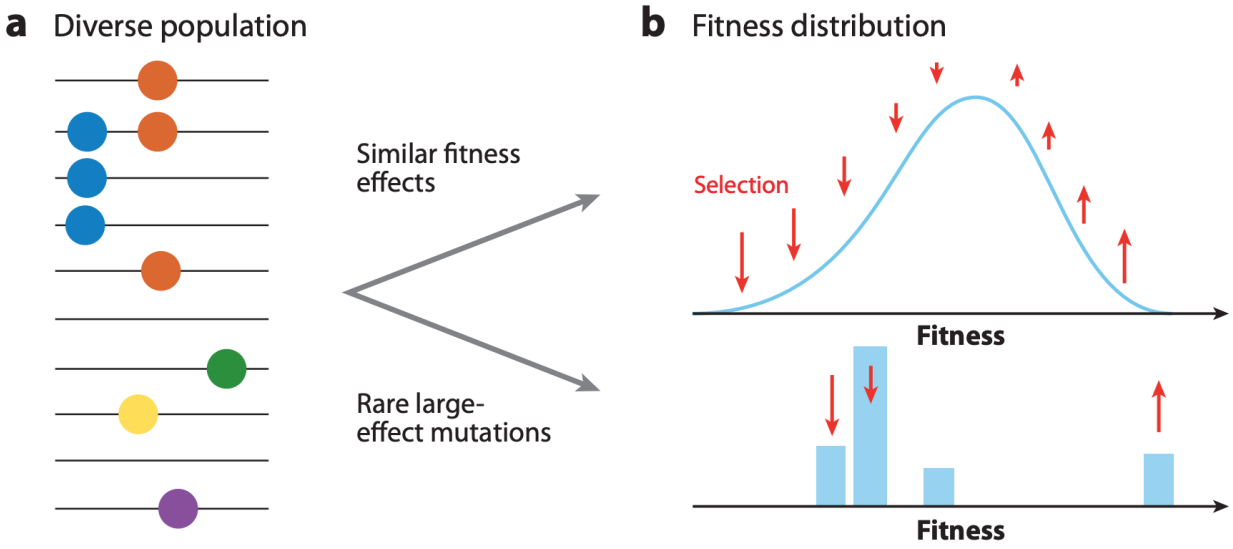

Figure 1.3: **The shape of fitness landscapes depends, in part, on mutation effect sizes.** Mutations with similar, smaller effects (blue and orange circles) produce a smooth Gaussian fitness distribution while mutations with large effect sizes (green, yellow, and purple circles) produce a more discrete fitness distribution. From Figure 1A and B of Neher [2013].

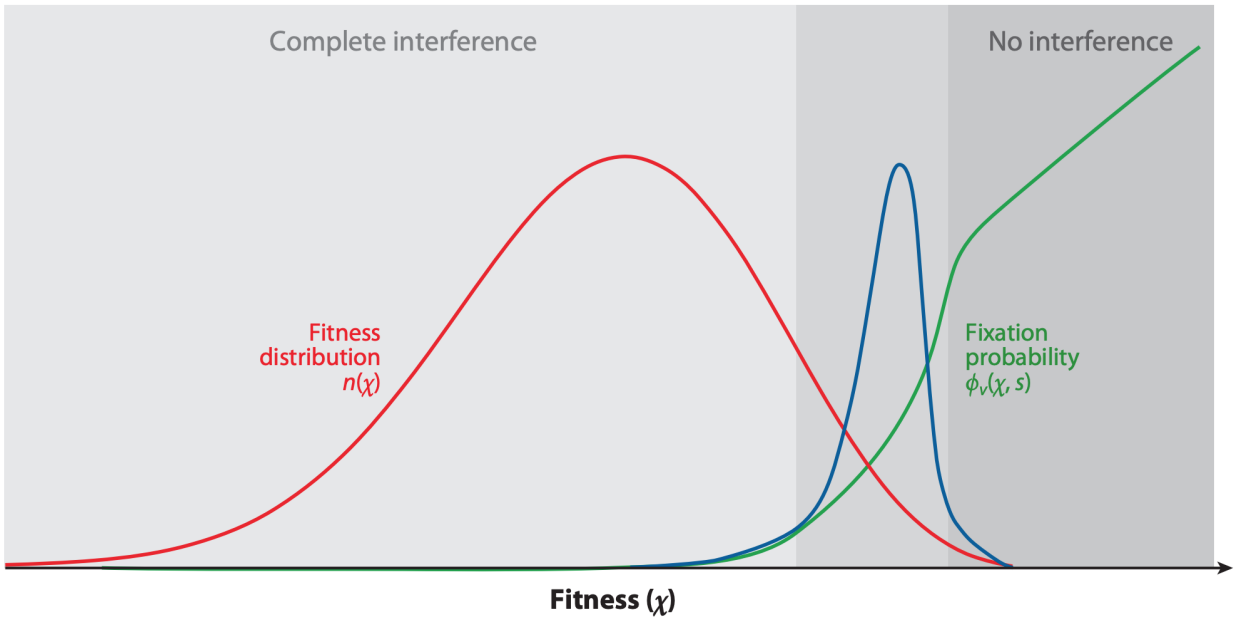

Figure 1.4: **The fixation probability of a beneficial mutation is a function of the mutation's genetic background.** When mutations have similar, smaller effects, the fitness of a beneficial mutation's genetic background (red) contributes to the mutation's fixation probability (green). Mutations that ultimately fix originate from distribution given by the product of the background fitness and the fixation probability (blue). From Figure 2C of Neher [2013].

2007], or contributed to antigenic drift [Wolf et al., 2006]. Importantly, not all of these sites contribute equally to antigenic drift [Koel et al., 2013]. Additionally, the complex and strong pressures of existing human immunity appear to constrain the space of antigenic phenotypes that viruses can explore at any given time [Smith et al., 2004, Bedford et al., 2012].

Recently, researchers have built on this evidence to create formal predictive models of seasonal influenza evolution. Neher et al. [2014] used expectations from traveling wave models to define the "local branching index" (LBI), an estimate of viral fitness. LBI assumes that most extant viruses descend from a highly fit ancestor in the recent past and uses patterns of rapid branching in phylogenies to identify putative fit ancestors (Figure 1.5). Neher et al. [2014] showed that LBI could successfully identify individual ancestral nodes that were highly representative of the seasonal influenza population one year in the future.

Łuksza and Lässig [2014] developed a mechanistic model to forecast seasonal influenza evolution based on population genetic theory and previous experimental work. This model assumed that seasonal influenza viruses grow exponentially as a function of their fitness, compete with each other for hosts through clonal interference, and balance positive effects of mutations at sites previously associated with antigenic drift and deleterious effects of all other mutations. Instead of predicting the most representative virus of the future population, Łuksza and Lässig [2014] explicitly predicted the future frequencies of entire clades.

Despite the success of these predictive models, other aspects of seasonal influenza evolution complicate predictions. When multiple beneficial mutations with large effects emerge in a population, the clonal interference between viruses reduces the probability of fixation for all mutations involved [Strelkowa and Lässig, 2012]. Seasonal influenza populations also experience multiple bottlenecks in space and time including transmission between hosts, global circulation, and seasonality [Xue et al., 2018, Petrova and Russell, 2018]. These bottlenecks reduce seasonal influenza's effective population size and reduce the probability that beneficial mutations will sweep globally. Finally, antigenic escape assays with polyclonal human sera suggest that successful viruses must accumulate multiple beneficial mutations of large effect

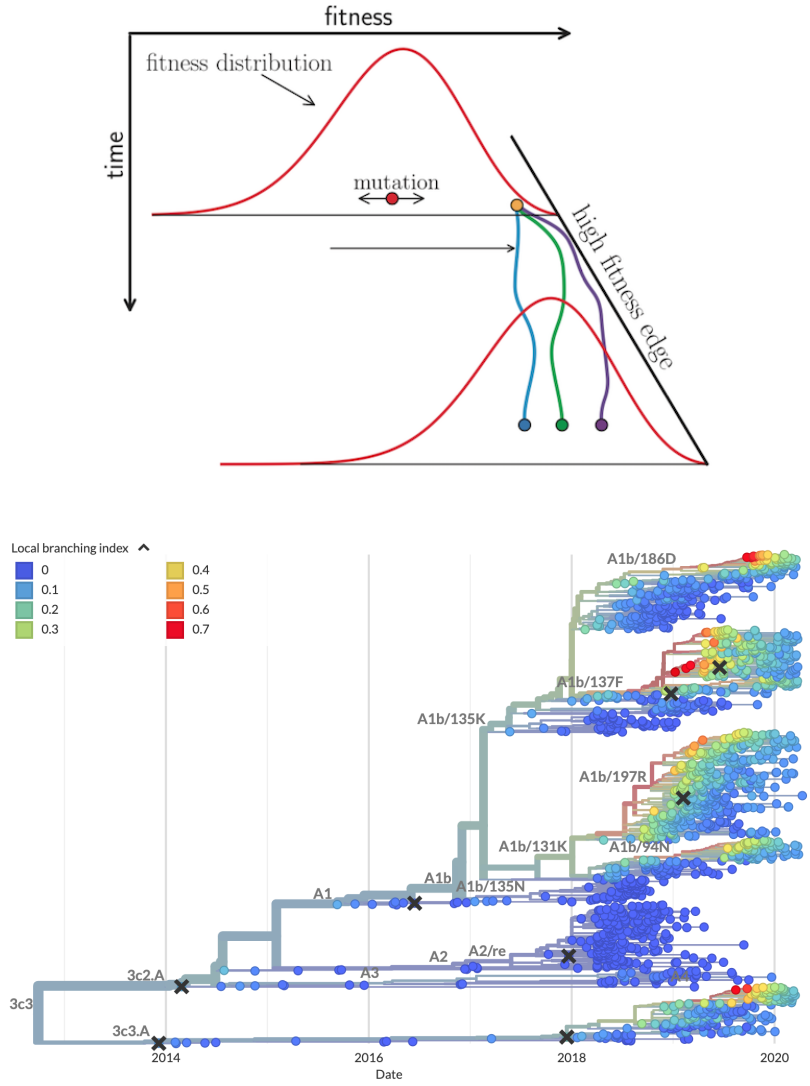

Figure 1.5: **Local branching index (LBI) estimates the fitness of viruses in a phylogeny.** A) LBI assumes that mutations at the high fitness edge of a current population will seed future populations. From Figure 5D of Neher [2013]. In practice, LBI tends to identify clusters of recently expanding populations, as shown in this seasonal influenza A/H3N2 phylogeny from Nextstrain. Explore LBI values in the current Nextstrain phylogeny for A/H3N2.

to successfully evade the diversity of global host immunity [Lee et al., 2019].

1.4 How has the field changed since the publication of the first predictive models?

In the years since the publication of these initial models, significant advances in influenza virology and computational modeling have paved the way for more biologically-informed predictive models. Recent computational methods can map HI measurements to phylogenetic trees of HA and accurately infer missing measurements between pairs of viruses using their shared ancestry [Neher et al., 2016]. By accounting for variation in viral avidity and serum potency, these methods provide a standardized measurement of antigenic phenotype that can inform existing predictive models. The application of deep mutational scanning to influenza proteins has enabled high-throughput quantification of functional constraints to protein evolution [Thyagarajan and Bloom, 2014, Wu et al., 2014, Doud and Bloom, 2016]. In addition to these improved measures of HA, recent studies highlight the importance of antigenic effects in neuraminidase (NA) [Chen et al., 2018] and fitness effects associated with the reassortment of HA with other proteins [Villa and Lässig, 2017]. The inclusion of evolutionary metrics for the entire influenza genome should therefore improve the predictive power of existing HA-only models. Finally, a detailed study of influenza phylodynamics has confirmed the importance of global circulation patterns to the success of influenza populations and revealed the variability of these patterns among influenza A and B subtypes [Bedford et al., 2015]. The resulting subtype-specific estimates of migration rate could readily benefit existing predictive models, which currently assume all viruses are panmictic. Despite the importance of these complementary characteristics of influenza fitness, no current predictive model of influenza evolution integrates all of these phenotypic, genomic, and geographic metrics with existing metrics of antigenic drift from HA sequences.

1.5 About this dissertation

In this dissertation, I describe my independent and collaborative efforts to improve our understanding of seasonal influenza A/H3N2 evolution through novel computational models and data visualizations. Chapter 2 describes a collaboration with Dr. Juhye Lee from the Bloom lab where Juhye performed deep mutational scanning experiments with an A/H3N2 virus and we attempted to understand the relationship between the resulting mutational preferences of A/H3N2 viruses in the lab and the success of mutations in natural populations. Chapter 3 describes a collaboration with Dr. Sarah Hilton from the Bloom lab where we developed a data visualization tool, *dms-view*, that allows virologists to rapidly explore their deep mutational scanning data in the linked contexts of the viral genome and protein structure. Chapter 3 also describes a preliminary tool for visualization of data from experiments that measure antigenic drift in A/H3N2. Chapter 4 describes my implementation of a long-term forecasting framework for A/H3N2 populations and how integration of genetic and phenotypic data in this framework produces the most accurate forecasts. Chapter 5 synthesizes the findings from the preceding chapters and a recent collaboration with Dr. Pierre Barrat-Charlaix from Dr. Richard Neher's lab. This conclusion places the results from this dissertation in the broader context of seasonal influenza evolutionary studies and provides recommendations for future research in the field. Finally, Appendix A describes Augur, a bioinformatics toolkit that I helped redesign and maintain during my doctoral work. This software was a critical component of the forecasting framework described in Chapter 4.

Chapter 2

DEEP MUTATIONAL SCANNING OF HEMAGGLUTININ HELPS PREDICT EVOLUTIONARY FATES OF HUMAN H3N2 INFLUENZA VARIANTS

With the exception of the subsection 2.4.1, this work was originally published in the *Proceedings of the National Academy of Sciences of the United States of America* at `https://doi.org/10.1073/pnas.1806133115`.

2.1 Introduction

Seasonal H3N2 influenza virus evolves rapidly, fixing 3 to 4 amino-acid mutations per year in its hemagglutinin (HA) surface protein [Fitch et al., 1997, Bhatt et al., 2011]. Many of these mutations contribute to the rapid antigenic drift that necessitates frequent updates to the annual influenza vaccine [Smith et al., 2004]. This evolution is further characterized by competition and turnover among groups of strains called clades bearing different complements of mutations [Bedford et al., 2011, Strelkowa and Lässig, 2012, Neher et al., 2014, Koelle and Rasmussen, 2015, Bedford et al., 2015]. Clades vary widely in their evolutionary success, with some dying out soon after emergence and others going on to take over the virus population. Several lines of evidence indicate that successful clades have higher fitness than clades that remain at low frequency [Bedford et al., 2011, Strelkowa and Lässig, 2012, Neher et al., 2014, Łuksza and Lässig, 2014]. A key goal in the study of H3N2 evolution is to identify the features that enable certain clades to succeed as others die out.

Two main characteristics distinguish evolutionarily successful clades from their competitors: greater antigenic change, and efficient viral growth and transmission. In principle, experiments

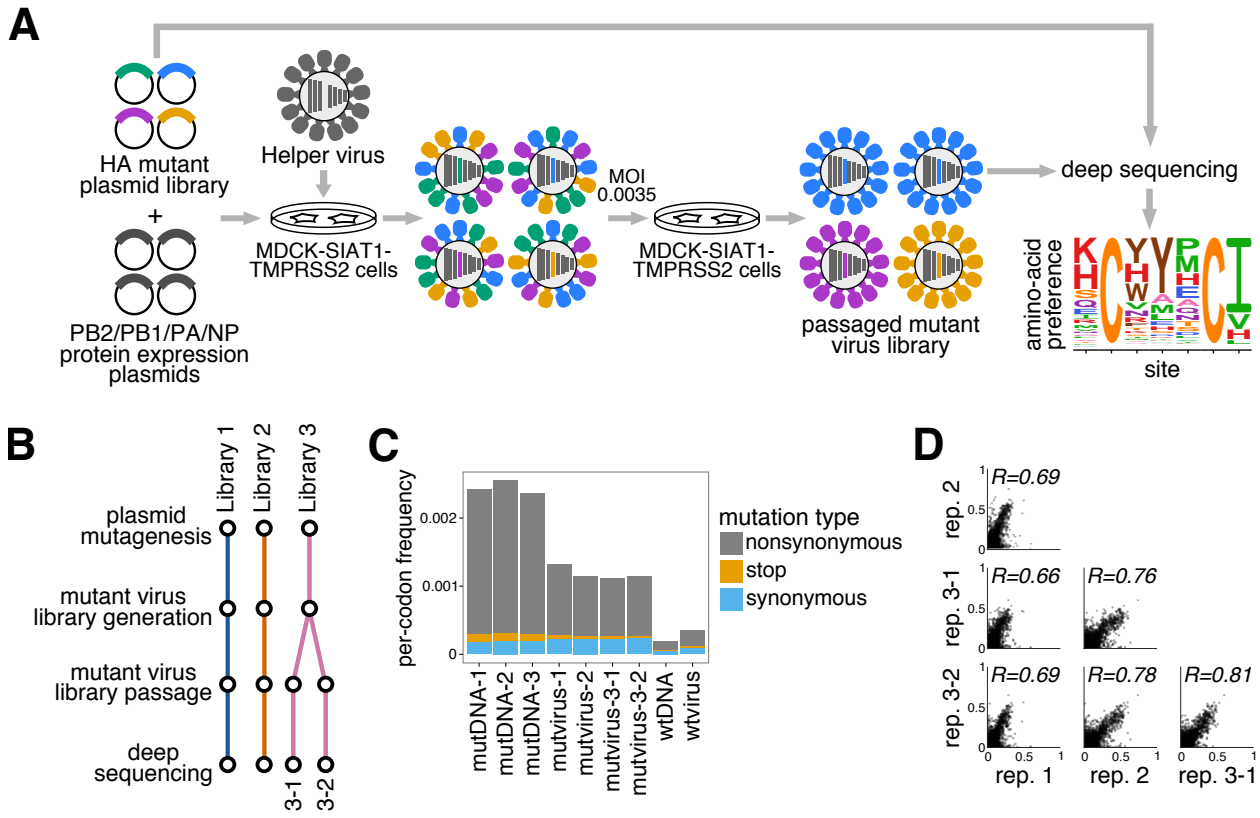

Figure 2.1: **Deep mutational scanning of the Perth/2009 H3 HA.** (A) We generated mutant virus libraries using a helper-virus approach [Doud and Bloom, 2016], and passaged the libraries at low MOI to establish a genotype-phenotype linkage and to select for functional HA variants. Deep sequencing of the variants before and after selection allowed us to estimate each site's amino-acid preferences. (B) The experiments were performed in full biological triplicate. We also passaged and deep sequenced library 3 in duplicate. (C) Frequencies of nonsynonymous, stop, and synonymous mutations in the mutant plasmid DNA, the passaged mutant viruses, and wildtype DNA and virus controls. (D) The Pearson correlations among the amino-acid preferences estimated in each replicate.

could be informative for identifying how mutations affect these features. Most work on influenza evolution to date has utilized experimental data to assess the antigenicity of circulating strains [Sun et al., 2013, Harvey et al., 2016, Neher et al., 2016, Koel et al., 2013, Chambers et al., 2015, Li et al., 2016]. However, the non-antigenic effects of mutations also play an important role [Pybus et al., 2007, Strelkowa and Lässig, 2012, Łuksza and Lässig, 2014, Koelle and Rasmussen, 2015]. Specifically, due to influenza virus's high mutation rate [Holland et al., 1982, Steinhauer and Holland, 1987, Lauring and Andino, 2010] and lack of intra-segment recombination [Boni et al., 2008], deleterious mutations become linked to beneficial ones. The resulting accumulation of deleterious mutations can affect non-antigenic properties central to viral fitness [Łuksza and Lässig, 2014]. However, there are no large-scale quantitative characterizations of how mutations to H3N2 HA affect viral growth.

It is now possible to use deep mutational scanning [Fowler and Fields, 2014] to measure the functional effects of all single amino-acid mutations to viral proteins [Thyagarajan and Bloom, 2014, Wu et al., 2014, Doud and Bloom, 2016, Haddox et al., 2016, Qi et al., 2015, Haddox et al., 2018]. However, the only HA for which such large-scale measurements have previously been made is from the highly lab-adapted A/WSN/1933 (H1N1) strain [Thyagarajan and Bloom, 2014, Wu et al., 2014, Doud and Bloom, 2016]. Here, we measure the effects on viral growth in cell culture of all mutations to the HA of a recent human H3N2 strain. We show that these experimental measurements can help discriminate evolutionarily successful mutations from those found in strains that quickly die out. However, the utility of the experiments for understanding natural evolution depends on the similarity between the experimental and natural strains: measurements made on an H1 HA are less informative for understanding the evolutionary fate of H3 viral strains.

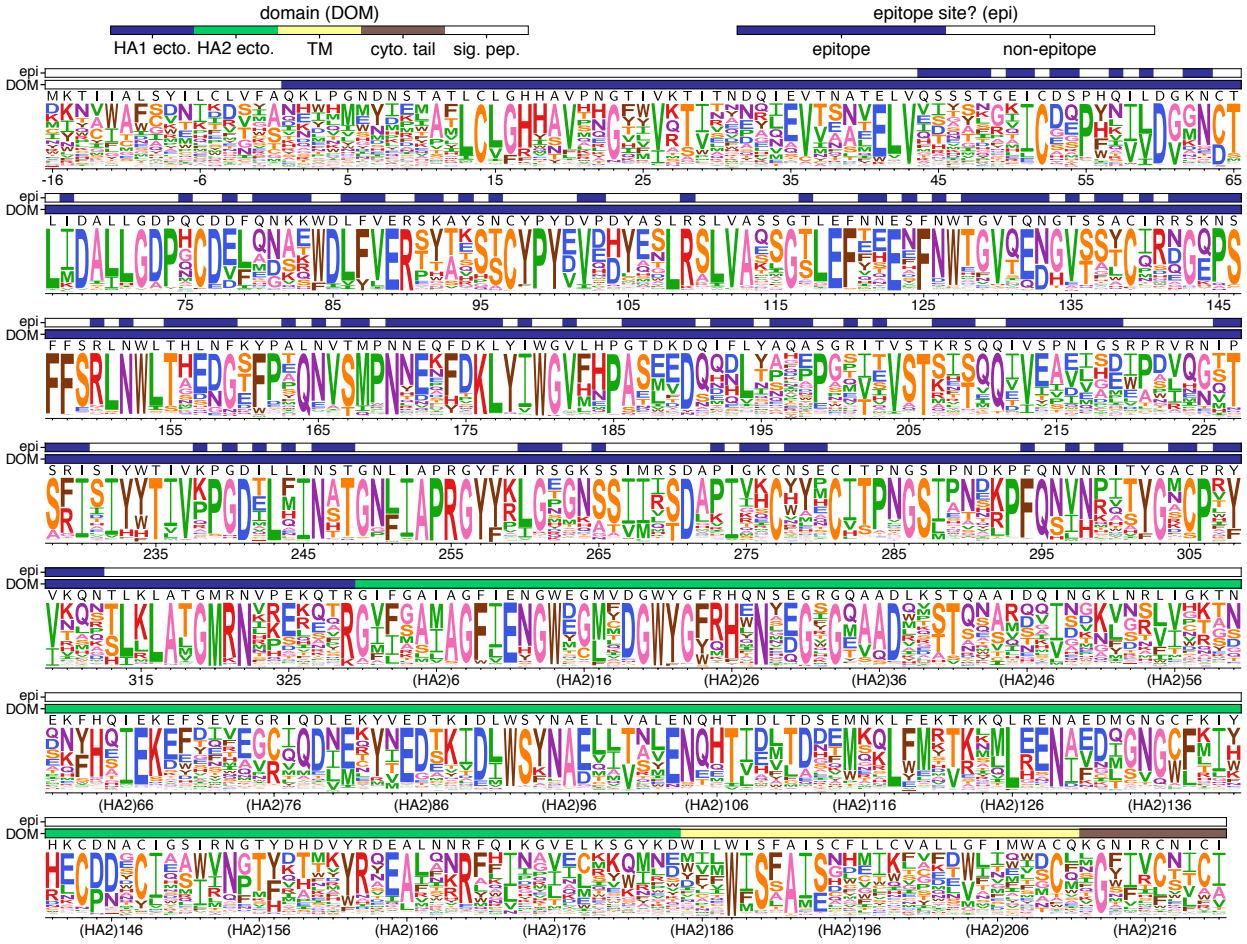

Figure 2.2: **The site-specific amino-acid preferences of the Perth/2009 HA measured in our experiments.** The height of each letter is the preference for that amino acid, after taking the average over experimental replicates and re-scaling [Hilton et al., 2017] by the stringency parameter in Table 2.1. The sites are in H3 numbering. The top overlay bar indicates whether or not a site is in the set of epitope residues delineated in Wolf et al. [2006]. The bottom overlay bar indicates the HA domain (sig. pep. = signal peptide, HA1 ecto. = HA1 ectodomain, HA2 ecto. = HA2 ectodomain, TM = transmembrane domain, cyto. tail. = cytoplasmic tail). The letters directly above each logo stack indicate the wildtype amino acid at that site.

2.2 Results

2.2.1 Deep mutational scanning of HA from a recent strain of human H3N2 influenza virus

We performed a deep mutational scan to measure the effects of all amino-acid mutations to HA from the A/Perth/16/2009 (H3N2) strain on viral growth in cell culture. This strain was the H3N2 component of the influenza vaccine from 2010-2012 [WHO, 2010, 2011]. Relative to the consensus sequence for this HA in Genbank, we used a variant with two mutations that enhanced viral growth in cell culture, G78D and T212I (see Lee et al. [2018] SI Appendix, Figure S1 and Dataset S1). The G78D mutation occurs at low frequency in natural H3N2 sequences, and T212 is a site where a mutation to Ala rose to fixation in human influenza in ∼2011.

We mutagenized the entire HA coding sequence at the codon level to create mutant plasmid libraries harboring an average of ∼1.4 codon mutations per clone (see Lee et al. [2018] SI Appendix, Figure S2). We then generated mutant virus libraries from the mutant plasmids using a helper-virus system that enables efficient generation of complex influenza virus libraries [Doud and Bloom, 2016] (Figure 2.1A). These mutant viruses derived all their non-HA genes from the lab-adapted A/WSN/1933 strain. Using WSN/1933 for the non-HA genes reduces biosafety concerns, and also helped increase viral titers. To further increase viral titers, we used MDCK-SIAT1 cells (Madin-Darby canine kidney cells overexpressing 2,6-sialyltransferase) [Matrosovich et al., 2003] that we engineered to constitutively express TMPRSS2 (Transmembrane Protease, Serine 2), which cleaves the HA precursor to activate it for membrane fusion [Böttcher et al., 2006, Böttcher-Friebertshäuser, et al., 2010].

After generating the mutant virus libraries, we passaged them at low multiplicity of infection (MOI) in cell culture to create a genotype-phenotype link and select for functional HA variants (Figure 2.1A). All experiments were completed in full biological triplicate (Figure 2.1B). We also passaged and deep sequenced library 3 in duplicate (library 3-1 and 3-2) to gauge experimental noise *within* a single biological replicate. As a control to measure sequencing

and mutational errors, we used the unmutated HA gene to generate and passage viruses carrying wildtype HA.

Deep sequencing of the initial plasmid mutant libraries and the passaged mutant viruses revealed selection for functional HA mutants. Specifically, stop codons were purged to 20-45% of their initial frequencies after correcting for error rates estimated by sequencing the wildtype controls (Figure 2.1C). The incomplete purging of stop codons is likely because genetic complementation due to co-infection [Marshall et al., 2013, Brooke et al., 2013] enabled the persistence of some virions with nonfunctional HAs. We also observed selection against many nonsynonymous mutations (Figure 2.1C), with their frequencies falling to 30-40% of their initial values after error correction.

2.2.2 Our measurements can help distinguish between mutations that reach low and high frequencies in nature

Mutations occurring in the H3N2 virus population experience widely varying evolutionary fates (Figure 2.3). Some mutations appear, spread and fix in the population, while others briefly circulate before disappearing. We take the maximum frequency reached by a mutation as a coarse indicator of its effect on fitness, since favorable mutations generally reach higher frequencies than unfavorable ones [Ewens, 2012]. Here, we follow the population genetic definition of *mutation* and track the outcome of each individual mutation event, e.g. although R142G occurs multiple times on the phylogeny we track each of these mutations occurring on different backgrounds separately. As such, each mutation is shown as a separate circle on a separate branch in Figure 2.3. However, because multiple mutations on the same phylogeny branch cannot be disentangled, when multiple mutations occurred on a single branch, we assigned a single mutational effect based on the sum of effects of each mutation.

After annotating mutations and their frequencies on the phylogeny in this way (Figure 2.3), it is visually obvious that there are relatively few circulating mutations that we measure to be strongly deleterious—and that such deleterious mutations rarely reach high frequency

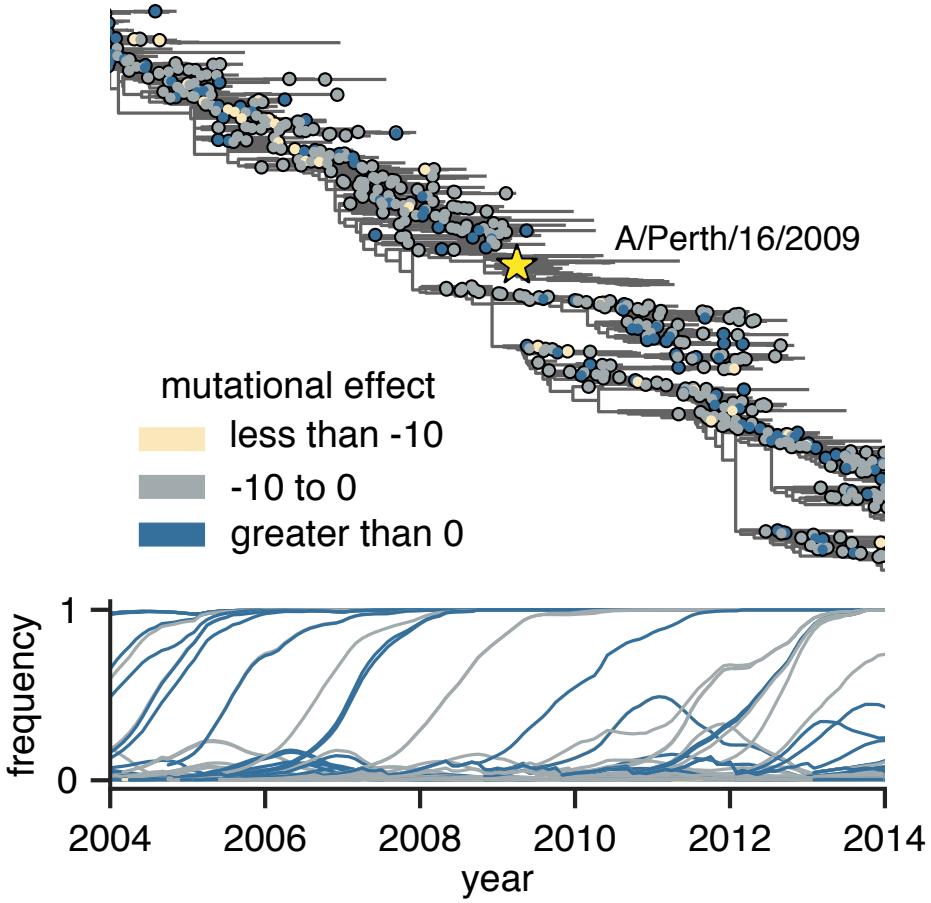

Figure 2.3: **Frequency trajectories of individual mutations and their relation to the experimentally measured effects of these mutations.** The top panel shows the subset of the full H3N2 HA tree (Figure 2.8) from 2004 to 2014. Circles indicate individual amino-acid mutations, and are colored according to the mutational effect measured in our deep mutational scanning (negative values indicate mutations measured to be deleterious to viral growth). The Perth/2009 strain is labeled with a star, and nodes in the clade containing the Perth/2009 strain were excluded from our analyses. The bottom panel shows the frequency trajectory of each mutation, with trajectories colored according to the mutational effects as in the top panel. It is clear that most mutations that reach high frequency are measured to be relatively favorable in our experiments. Figure 2.4 shows a similar layout but colors mutations by whether they are in HA's head or stalk domain.

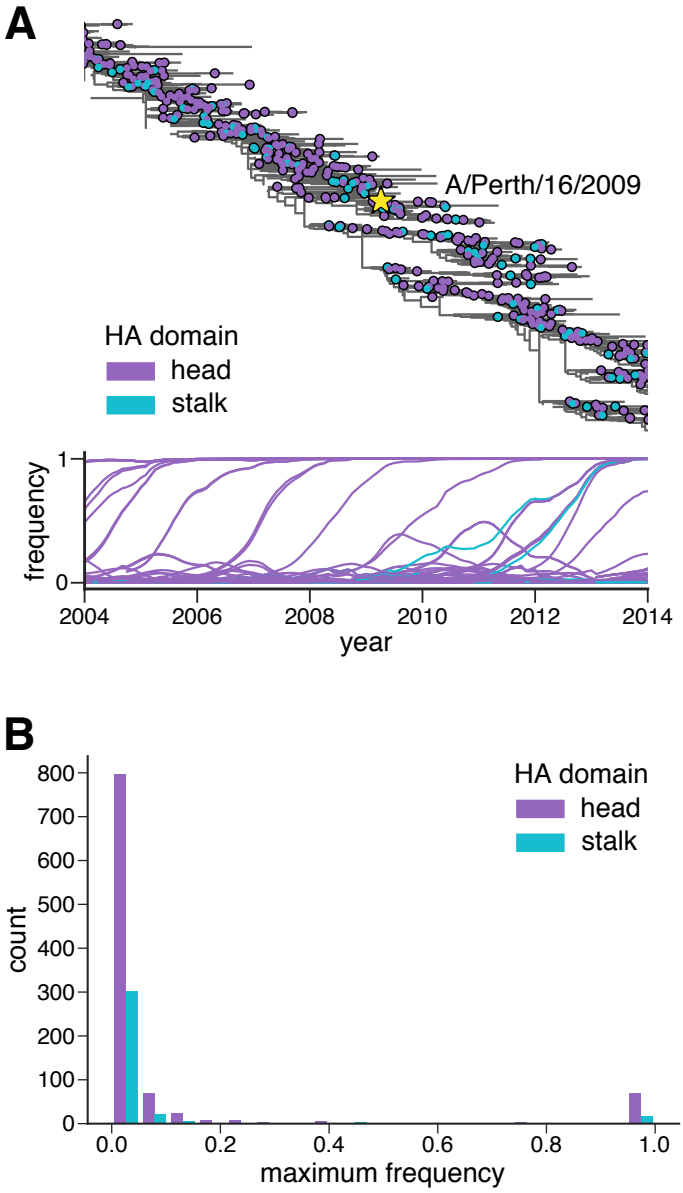

Figure 2.4: **Frequency trajectories of head and stalk domain mutations.** (A) This figure repeats the analysis of the H3N2 mutation frequencies in Figure 2.3, but colors amino-acid mutations by whether they occur in the head (purple) or the stalk (blue) domain. (B) Histogram of mutation maximum frequencies by the number of mutations in the head and stalk domains. It is clear that mutations in the head domain are more numerous than those in the stalk, particularly among mutations that reach high frequencies.

when they do occur.

We next sought to quantify the correlation between a mutation's experimentally measured effect and the maximum frequency it attained during natural evolution. To calculate a given mutation's effect, we simply took the logarithm of the ratio of the preferences for the mutant and wildtype amino acids at that site. To minimize effects related to the genetic background of the strain used in the experiment, we excluded mutations closely related to the experimental strain itself and partitioned the remaining mutations into 1,022 mutations pre-dating and 299 mutations post-dating the Perth/2009 strain (Figure 2.8). We additionally excluded mutations from the post-Perth partition that were sampled in 2014 or after, since these mutations have not had enough time for their evolutionary fates to be fully resolved. We used these pre-Perth and post-Perth partitions to test the utility of our measurements for both post-hoc and prospective analyses, respectively. We quantified the relationship between mutational effects and maximum mutation frequencies in the H3N2 phylogeny via Spearman rank correlation (Figure 2.5A). In both pre-Perth and post-Perth time periods, we found a modest, but statistically significant relationship between mutational effect and maximum mutation frequency (pre-Perth $\rho = 0.17$, post-Perth $\rho = 0.15$). The similar effect sizes for both the pre- and post-Perth partitions shows that our experimental measurements can help explain the evolutionary fates of mutations in strains that post-date the experimentally studied strain, as well as to retrospectively analyze mutations that precede the experimental strain.

Many of the HAs in sequence databases are from viral isolates that were passaged in cell-culture or eggs, which can cause lab-adaptation mutations that confound evolutionary analyses [McWhite et al., 2016]. To check that our results were robust to such lab-adaptation mutations, we repeated our analysis using only HA sequences derived from viruses that had not been passaged in the lab. Because sequencing of unpassaged primary isolates has only recently become commonplace, we could only perform this analysis for the post-Perth partition of the phylogenetic tree. Figure 2.5A shows that the correlation between our

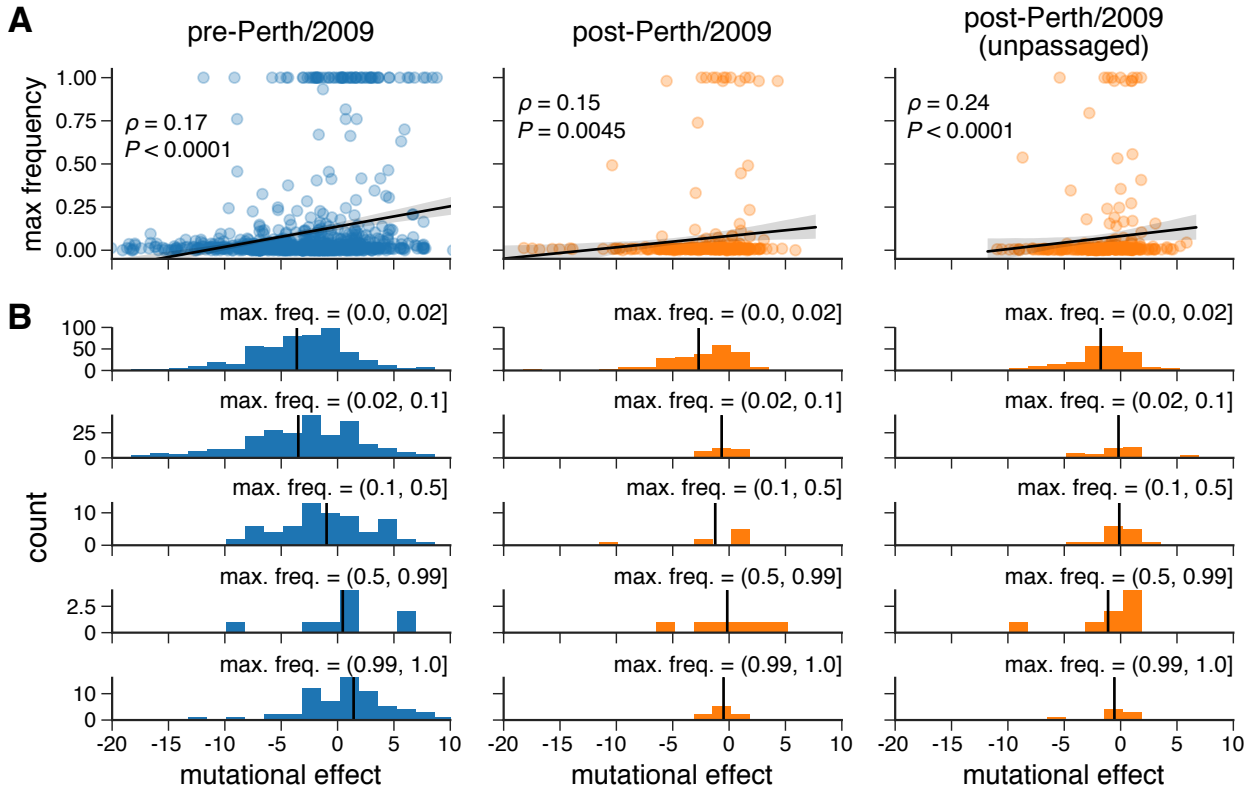

Figure 2.5: **Experimental measurements are informative about the evolutionary fate of viral mutations.** (A) Correlation between the effects of mutations as measured in our deep mutational scanning of the Perth/2009 HA and the maximum frequency reached by these mutations in nature. The plots show Spearman ρ and an empirical P-value representing the proportion of 10,000 permutations of the experimental measurements for which the permuted ρ was greater than or equal to the observed ρ. (B) The distribution of mutational effects partitioned by maximum mutation frequency. The vertical black line shows the mean mutation effect for each category. The analysis is performed separately for pre-Perth/2009, post-Perth/2009, and unpassaged isolates from the post-Perth/2009 partitions of the tree (Figure 2.8).

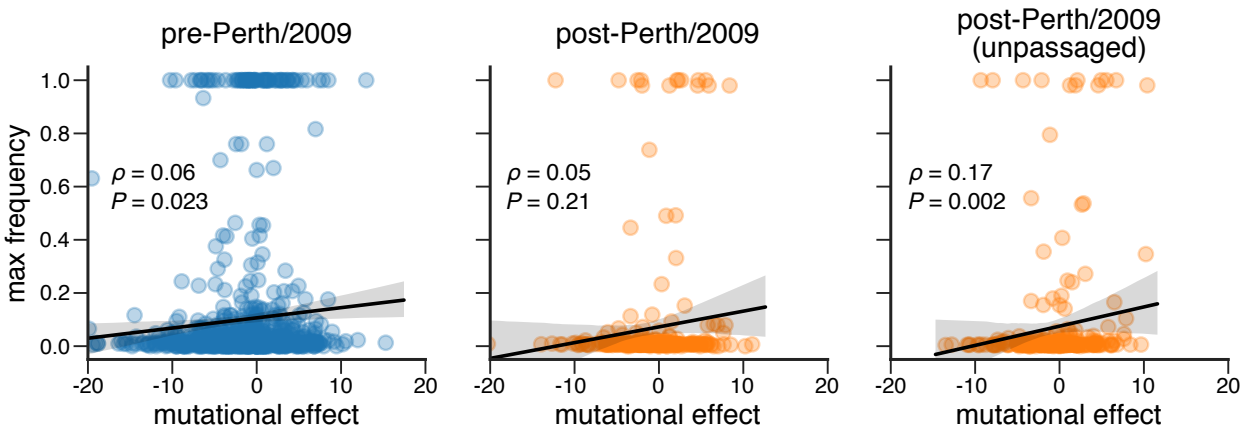

Figure 2.6: **Experimental measurements on an H1 HA are less informative about the evolutionary fate of H3N2 mutations.** This figure repeats the analysis of the H3N2 mutation frequencies in Figure 2.5A, but uses the deep mutational scanning data for an H1 HA as measured in [Doud and Bloom, 2016]. Figure 2.7 shows the histograms comparable to those in Figure 2.5B. The empirical P-value represents the result of 1,000 permutations.

measured mutational effects and the maximum frequency was even stronger for mutations from unpassaged viral isolates ($\rho = 0.24$).

The trends in Figure 2.5A are most strongly driven by the behavior of substantially deleterious mutations. We investigated this further by partitioning mutations into those that reach low, medium and high frequencies, and those that fix in the population (Figure 2.5B). The mutations that reach higher frequencies have a more favorable mean effect. Mutations measured to be substantially deleterious almost never reach high frequency. Overall, these results demonstrate that measurements of how mutations affect viral growth in cell culture are informative for understanding the fates of these mutations in nature: in particular, if a mutation is measurably deleterious to viral growth, that mutation is unlikely to prosper in nature.

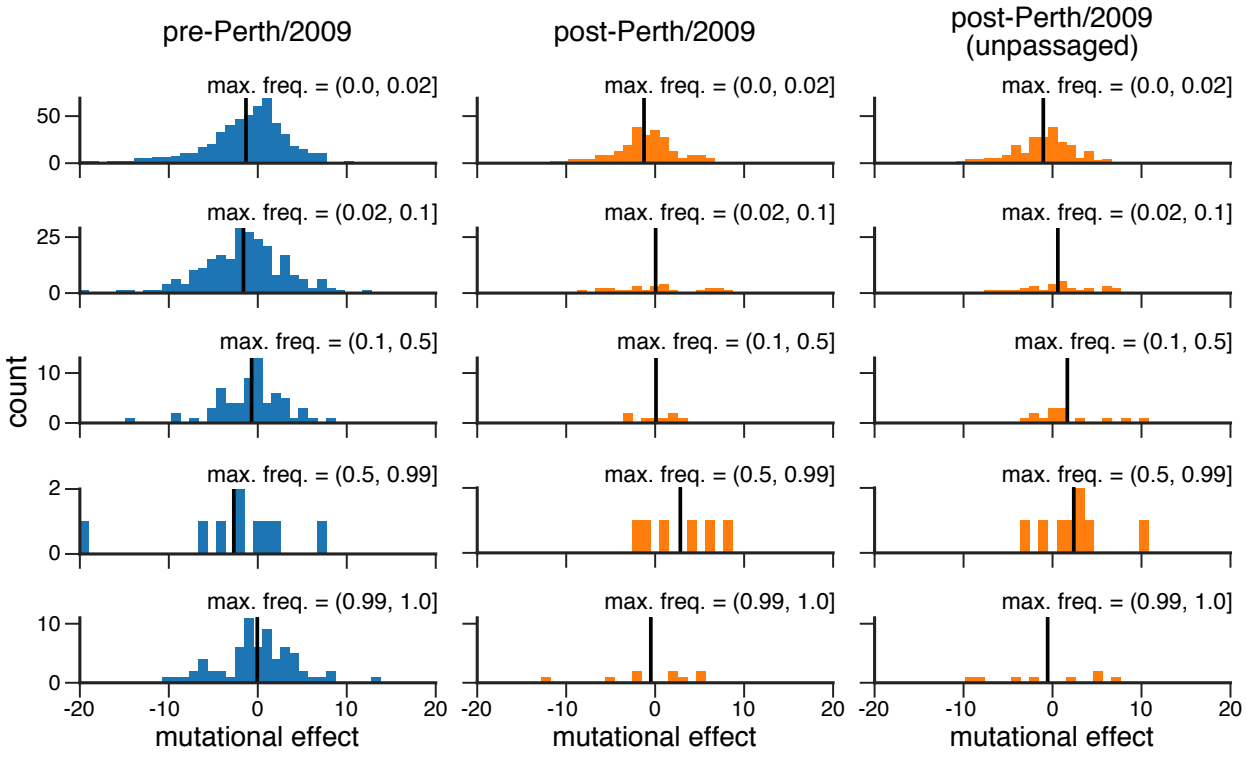

Figure 2.7: **The distribution of mutational effects measured in H1 HA among H3N2 mutations binned by the maximum frequency that they reach.** This figure repeats the analysis of the H3N2 mutation frequencies in Figure 2.5B, but uses the deep mutational scanning data for an H1 HA as measured in Doud and Bloom [2016].

2.2.3 Measurements made on an H1 HA are less informative for understanding the evolution of H3 influenza

To determine how broadly experimental measurements can be generalized across HAs, we repeated the foregoing analysis of H3N2 mutation frequencies, but using mutational effects measured in our prior deep mutational scanning of the WSN/1933 H1 HA [Doud and Bloom, 2016] (see Lee et al. [2018] SI Appendix, Figure S9), which is highly diverged from the Perth/2009 H3 HA (the two HAs only have 42% protein sequence identity). Figure 2.6 shows that the correlations between the H1 experimental measurements and the maximum frequency that mutations reach during H3N2 viral evolution are consistently weaker than those using H3 experimental measurements (compare Figure 2.6 to Figure 2.5A). Therefore, the utility of an experiment for understanding natural evolution degrades as the experimental sequence becomes more diverged from the natural sequences that are being studied.

2.3 Discussion

We have measured the effects of all possible single amino-acid mutations to the Perth/2009 H3 HA on viral growth in cell culture and demonstrated that these measurements have some value for understanding the evolutionary fate of these mutations in nature. Specifically, mutations measured to be more beneficial for viral growth tend to reach higher frequencies in nature than mutations measured to be more deleterious for viral growth. The fact that our experiments can help identify evolutionary successful mutations suggests that they might inform evolutionary forecasting. In their landmark paper introducing predictive viral fitness models that accounted for both antigenic and non-antigenic mutations, Łuksza and Lässig [2014] noted that the models could in principle be improved by integrating "diverse genotypic and phenotypic data" that more realistically represented the effects of specific mutations. Our work suggests that deep mutational scanning may be able to provide such data.

It is important to emphasize that measurements of viral growth in cell culture do *not* represent true fitness in nature. Indeed, a vast amount of work in virology has chronicled the ways in

which experiments can select for lab artifacts or fail to capture important pressures that are relevant in nature [Daniels et al., 1985, Sun et al., 2010, Lee et al., 2013, Wu et al., 2017]. As an example, although we identified G78D as favorable for viral growth in cell culture, this mutation never fixes in nature. Mutations in viral genes other than HA are also important in determining strain success [Memoli et al., 2009, Raghwani et al., 2017]. Given these caveats, it might seem surprising that measuring viral growth in cell culture can be informative about the success of viral strains in nature. Yet, prior to our work, there were no comprehensive studies of the functional effects of mutations to H3 HA on any property that even resembled viral fitness in nature, and modeling work has either omitted the non-antigenic effects of mutations [Sun et al., 2013, Harvey et al., 2016, Neher et al., 2016] or assumed that all non-epitope mutations had equivalent deleterious effects [Łuksza and Lässig, 2014]. The strength of our measurements are not that they perfectly capture fitness in nature, but that they are systematic and quantitative—and so represent an improvement over no information at all. We suspect that performing similar experiments using more realistic and complex selections (e.g., ferrets or primary human airway cultures) might further improve their utility and possibly their generalizability to more divergent strains.

We measured the effects of all single amino-acid mutations to a specific HA, and then generalized these measurements to other H3N2 HAs from a 50-year timespan. These generalizations will only be valid to the extent that the effects of mutations are conserved during HA's evolution. Extensive prior work has shown that epistasis can shift the effects of mutations as proteins evolve [Pollock et al., 2012, Shah et al., 2015, Gong et al., 2013, Natarajan et al., 2013, Harms and Thornton, 2014, Starr and Thornton, 2016, Starr et al., 2017]. Our work suggests that measurements on a HA from single human H3N2 viral strain can be usefully generalized to at least some extent across the entire evolutionary history of human H3N2 HA. On the other hand, when we compared our measurements for an H3 HA to prior measurements on H1 HA, we found substantial shifts at many sites—much greater than those observed in prior protein-wide comparisons of more closely related homologs [Doud et al., 2015, Haddox

et al., 2018]. Further investigation of how mutational effects shift as proteins diverge will be important for determining how broadly any given experiment can be generalized when attempting to make evolutionary forecasts.

Our work did not characterize the antigenic effects of mutations, which also play an important role in determining strain success in nature [Koel et al., 2013, Neher et al., 2016]. However, our basic selection and deep-sequencing approach can be harnessed to completely map how mutations affect antibody recognition [Doud et al., 2017, 2018]. But so far, experiments using this approach have not examined antibodies or sera that are relevant to driving the evolution of H3N2 influenza [Doud et al., 2017, 2018], or have used relevant sera but examined a non-comprehensive set of mutations [Li et al., 2016]. Future experiments that completely map how HA mutations affect recognition by human sera seem likely to be especially fruitful for informing viral forecasting.

2.4 Materials and Methods

Data and computer code

Deep sequencing data are available from the Sequence Read Archive under BioSample accessions SAMN08102609 and SAMN08102610. Computer code used to analyze the data is at `https://github.com/jbloomlab/Perth2009-DMS-Manuscript`.

HA numbering

Sites are in H3 numbering, with the signal peptide in negative numbers, HA1 in plain numbers, and HA2 denoted with "(HA2)". Sequential 1, 2, ... numbering of the Perth/2009 HA can be converted to H3 numbering by subtracting 16 for the HA1 subunit, and subtracting 345 for the HA2 subunit.

Table 2.1: **Substitution models informed by the experiments describe HA's evolution better than traditional models.** Maximum likelihood phylogenetic fit to an alignment of human H3N2 HAs using ExpCM [Hilton et al., 2017], ExpCM in which the experimental measurements are averaged across sites (site avg.), and M0 and M5 versions of the Goldman-Yang (GY94) model [Yang et al., 2000]. Models are compared by Akaike information criterion (AIC) [Posada and Buckley, 2004] computed from the log likelihood (LnL) and number of model parameters. The ω parameter is dN/dS for the Goldman-Yang models, and the relative dN/dS after accounting for the measurements for the ExpCM. For the M5 model, we give the mean followed by the shape and rate parameters of the gamma distribution over ω.

Model	**ΔAIC**	**LnL**	**Stringency**	ω
ExpCM	0.0	-8441	2.47	0.91
GY94 M5	2094	-9482	–	0.36 (0.30, 0.84)
ExpCM, site avg.	2501	-9692	0.67	0.32
GY94 M0	2536	-9704	–	0.31

Quantification of mutational effects

The effect of mutating site r from amino acid a_1 to a_2 was quantified as

$$\log_2 \frac{\pi_{r,a_2}}{\pi_{r,a_1}} \tag{2.1}$$

where π_{r,a_1} and π_{r,a_2} are the re-scaled preferences for amino acids a_1 or a_2 at site r as shown in Figure 2.2. The WSN/1933 H1 HA amino-acid preferences are the replicate-average values reported in [Doud and Bloom, 2016], re-scaled by a stringency parameter of 2.05 (see `https://github.com/jbloomlab/dms_tools2/blob/master/examples/Doud2016/analysis_notebook.ipynb`).

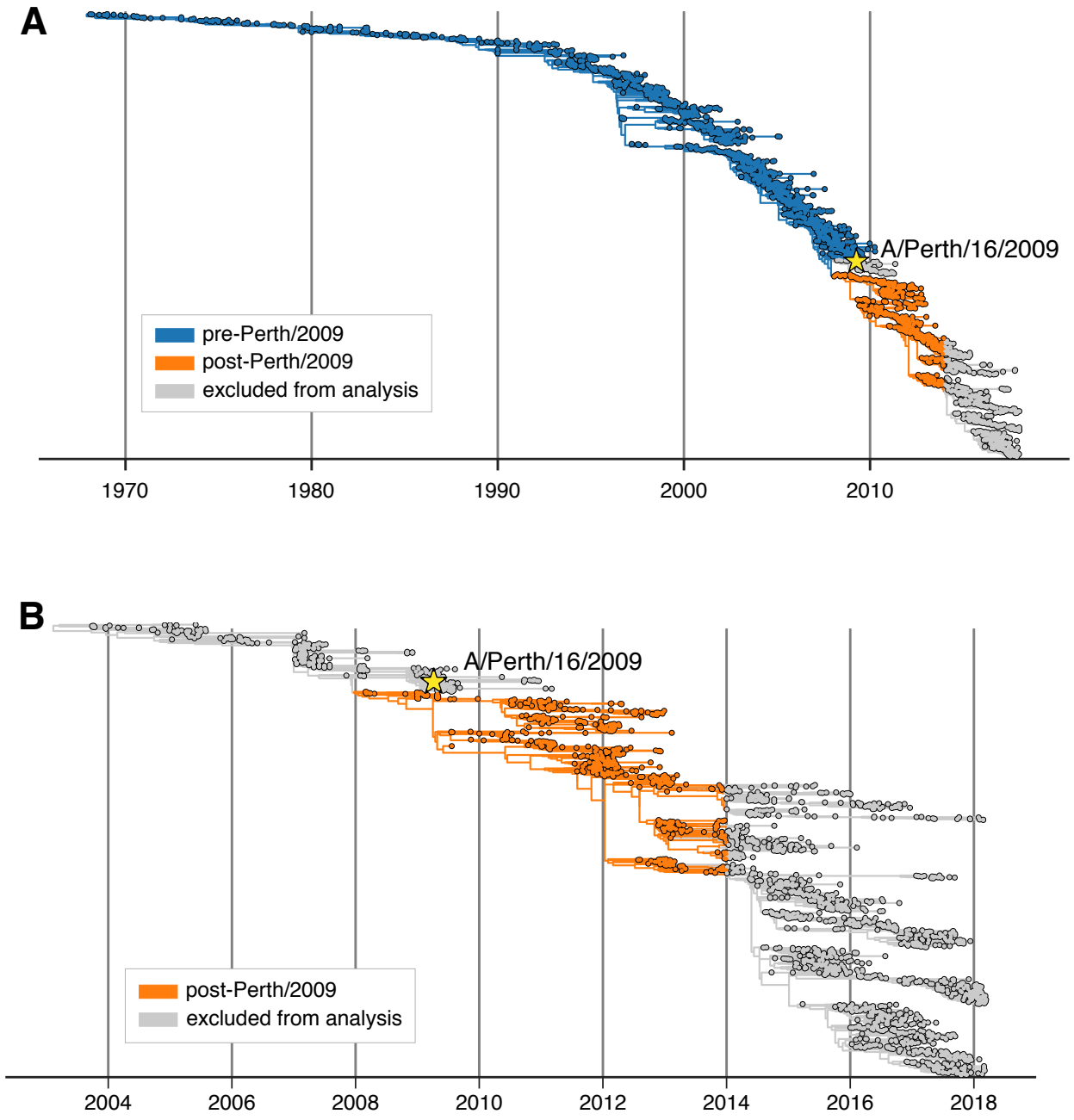

Figure 2.8: **A phylogenetic tree of all HA sequences used in our analysis of mutation frequencies.** (A) HA sequences were sampled at a rate of six viruses per month from January 1, 1968 through February 1, 2018. The Perth/2009 strain used in our experiments is indicated.

Figure 2.8: (continued) The rest of the tree is partitioned into nodes that preceded the split of the Perth/2009 strain from the trunk of the tree (blue) and nodes that branched off the trunk after the clade containing Perth/2009 (orange). In Figure 2.5, these two partitions of the tree are analyzed separately. Nodes in the clade containing the Perth/2009 strain and nodes sampled in 2014 or after were excluded from our analyses. The Perth/2009 strain was excluded to avoid artifacts related to mutations that occurred on the branches leading to the HA sequence used in the experiment. The post-2014 nodes were excluded because the evolutionary fates of many sequences after this date are not yet fully resolved. (B) The post-Perth/2009 partition of the tree containing only sequences from unpassaged isolates.

Inference of human H3N2 phylogenetic tree and calculation of maximum mutation frequencies

To generate the tree (Figure 2.8), we applied Nextstrain's augur pipeline [Hadfield et al., 2018] (`https://github.com/nextstrain/augur`; commit `006896d`) to publicly available H3N2 HA sequences from GISAID [Shu and McCauley, 2017] (see Lee et al. [2018] SI Appendix, Dataset S4), sampling six viruses per month over the time interval of January 1, 1968 to February 1, 2018. We aligned the resulting 2,189 HA sequences with MAFFT v7.310 [Katoh and Standley, 2013] and constructed a maximum likelihood phylogeny from this alignment with RAxML 8.2.10 [Stamatakis, 2006]. Ancestral state reconstruction and branch length timing were performed with TreeTime [Sagulenko et al., 2018]. The phylogenetic tree is available as a JSON file on GitHub at `https://github.com/jbloomlab/Perth2009-DMS-Manuscript/blob/master/analysis_code/data/flu_h3n2_ha_1968_2018_6v_tree.json.gz`. The tree was visualized using BALTIC (`https://github.com/blab/baltic`).

The frequency trajectory of each individual mutation on the phylogeny is estimated following Nextstrain's augur pipeline and as first implemented in Nextflu [Neher and Bedford, 2015]. Herein, mutation frequency dynamics are modeled according to a Brownian motion diffusion process discretized to one-month intervals. The number of viruses sampled in each interval

determines the denominator of the mutation frequency calculations. Relative to a simple Brownian motion, the expectation includes an "inertia" term ϵ that adds velocity to the diffusion and the variance includes a term $x(1-x)$ to scale variance according to frequency following a Wright-Fisher population genetic process. This results in the following diffusion process

$$x(t+\mathrm{d}t) = \mathcal{N}(x(t) + \epsilon\,\mathrm{d}x,\ \mathrm{d}t\,\sigma^2\,x(t)\,(1-x(t))), \tag{2.2}$$

with 'volatility' parameter σ^2. The term $\mathrm{d}x$ is the increment in the previous timestep, so that $\mathrm{d}x = x(t) - x(t-\mathrm{d}t)$. We used $\epsilon = 0.7$ and $\sigma^2 = 0.05$ to maximize fit to empirical trajectory behavior.

We also include an Bernoulli observation model for mutation presence / absence among sampled viruses at timestep t. This observation model follows

$$f(x,t) = \prod_{v \in V} x(t) \prod_{v \notin V} (1 - x(t)), \tag{2.3}$$

where $v \in V$ represents the set of viruses that have the mutation and $v \notin V$ represents the set of viruses that do not have the mutation. Each frequency trajectory is estimated by simultaneously maximizing the likelihood of the process model and the likelihood of the observation model via adjusting frequency trajectory $\mathbf{x} = (x_1, \ldots, x_n)$.

We also repeated the above analyses using only viruses that were sequenced directly without passaging. Routine direct sequencing did not begin until the early 2000s [McWhite et al., 2016]. To construct a tree with a similar number of viruses as the original analysis, we sampled 30 viruses per month between January 1, 2000 and April 1, 2018, producing a tree with 2,374 unpassaged viruses with augur (commit: `6d9f708`). We included the passaged DMS strain, A/Perth/16/2009, in the resulting tree to enable comparison between pre-Perth and post-Perth clades.

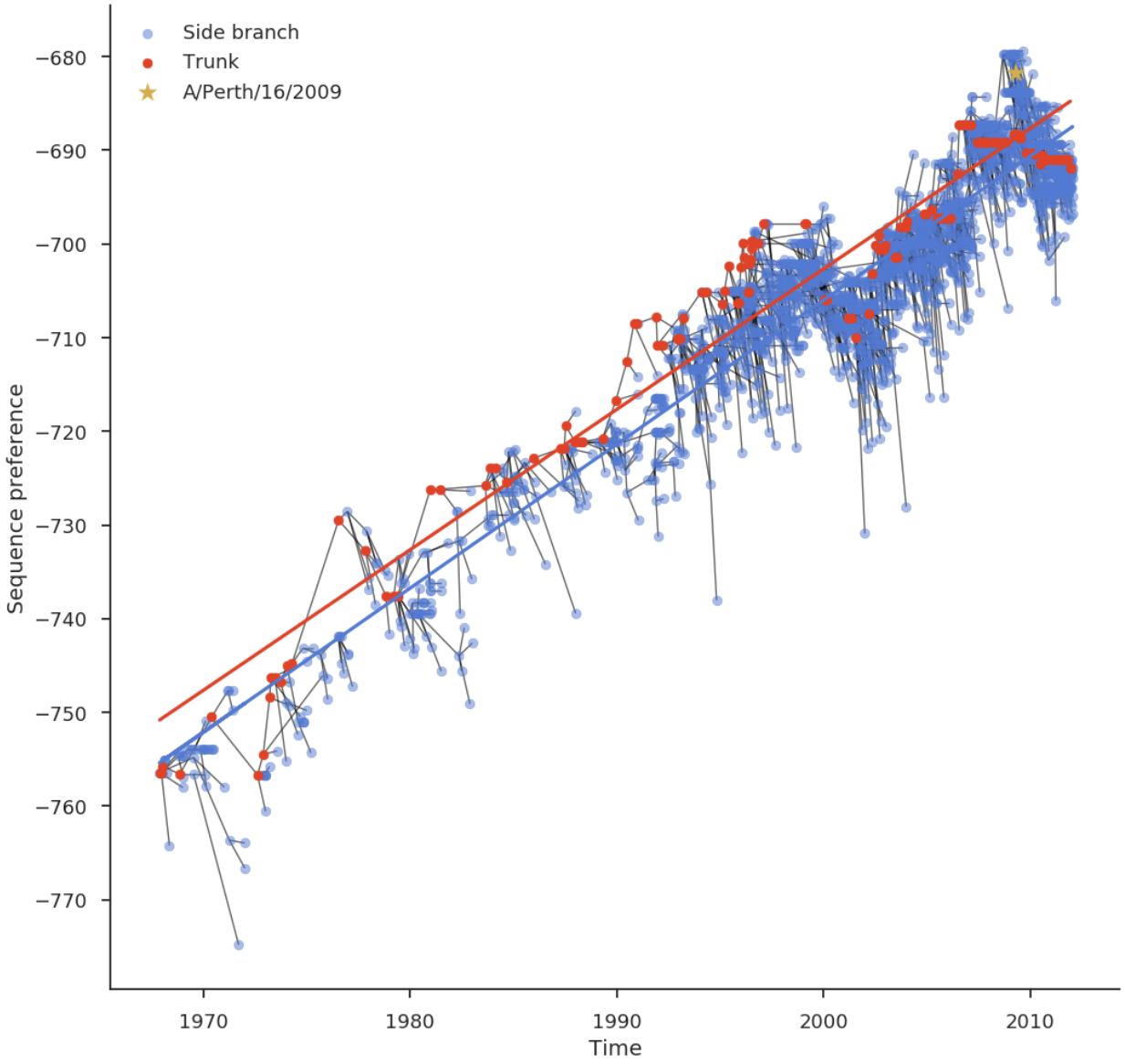

Figure 2.9: **Sequence preference by time for sequences from 1968–2012.** Least squares regression lines are shown for trunk (red) and side branch (blue) nodes. The trunk line corresponds to the equation $y = 1.54x - 3779.83$ with an adjusted R-squared value of 0.87 ($p << 0.01$). The side branch line corresponds to the equation $y = 1.50x - 3700$ with an adjusted R-squared value of 0.88 ($p << 0.01$). The differences between these two equations is effectively eliminated when the same analysis is performed without terminal nodes.

2.4.1 Analysis of full-length sequence preferences

We attempted to quantify the fitness of each H3N2 strain's complete HA sequence by summing the mutational preferences defined in Equation 2.1 across all positions in HA. We anticipated that in the absence of epistasis the resulting sums would provide a metric of each strain's fitness with respect to its functional constraints. Importantly, the sequence preference metric should also provide a fitness estimate that could be applied to any sequence without any additional phylogenetic information (i.e., the amino acid mutations present on branches leading to the strain in a phylogeny).

We calculated the sequence preference for all observed strains in the full phylogeny (Figure 2.8) and their inferred ancestral sequences corresponding to internal nodes of the tree. We additionally annotated each node of the tree by its status as belonging to the "trunk" or "side branch" of the phylogeny, using a previously published definition for these categories [Bedford et al., 2015]. Nodes that belonged to the trunk category represent sequences that seeded future H3N2 populations while nodes on side branches failed to propagate. We plotted the sequence preference of each node by its observed or inferred date and trunk status.

Surprisingly, we observed a positive linear trend for all sequences collected before the Perth/2009 strain (Figure 2.9). This pattern reversed for all sequences collected after the Perth/2009 strain, showing a negative linear trend. To better visualize and interpret these results, we fit linear regression models to both trunk and side-branch nodes and plotted the residuals for each sequence from the resulting models. The residuals recapitulated the trends we observed in the sequence preferences by time and highlighted the greater divergence of trunk nodes from the positive linear trend after 2009 (Figure 2.10). The same patterns appeared when we fit local weighted regression (LOWESS) lines to the residuals (Figure 2.11)

From these results, we concluded that the sequence preference metric was strongly influenced by the genetic background of the DMS preferences themselves (i.e., the Perth/2009 strain). We interpreted the positive linear trend toward the Perth/2009 era as a representation of

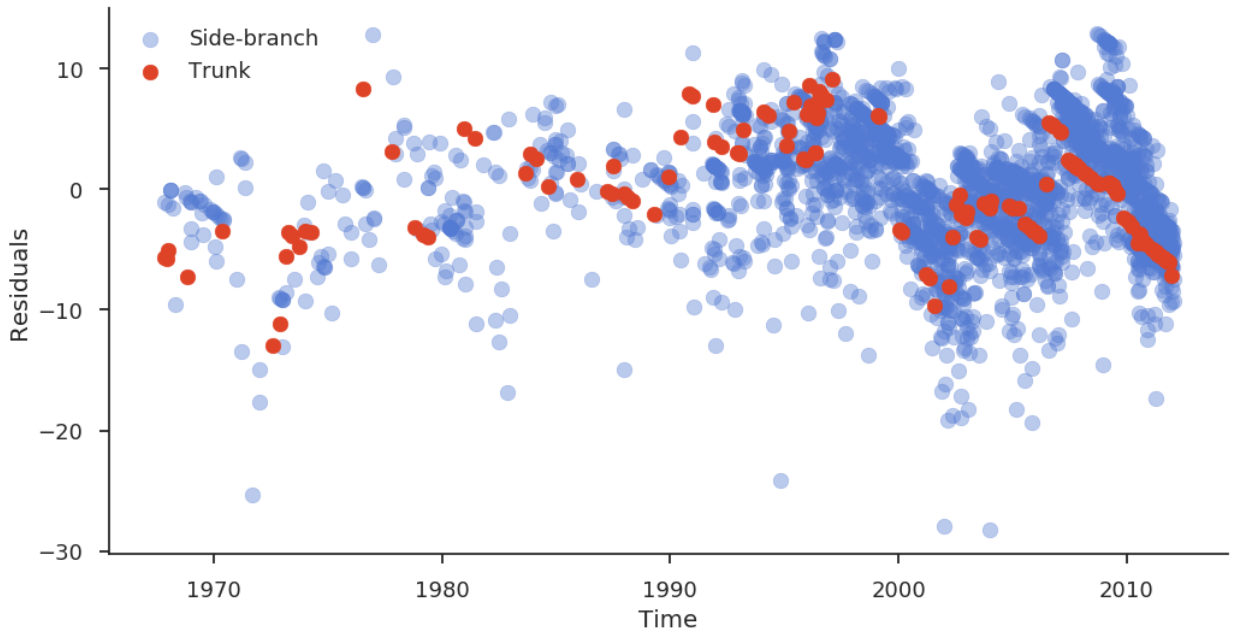

Figure 2.10: **Residuals by time based on linear regression equations in Figure 2.9.** The null expectation for the residuals is a horizontal line for both trunk and side-branch nodes with a trunk line slightly above the side-branch line.

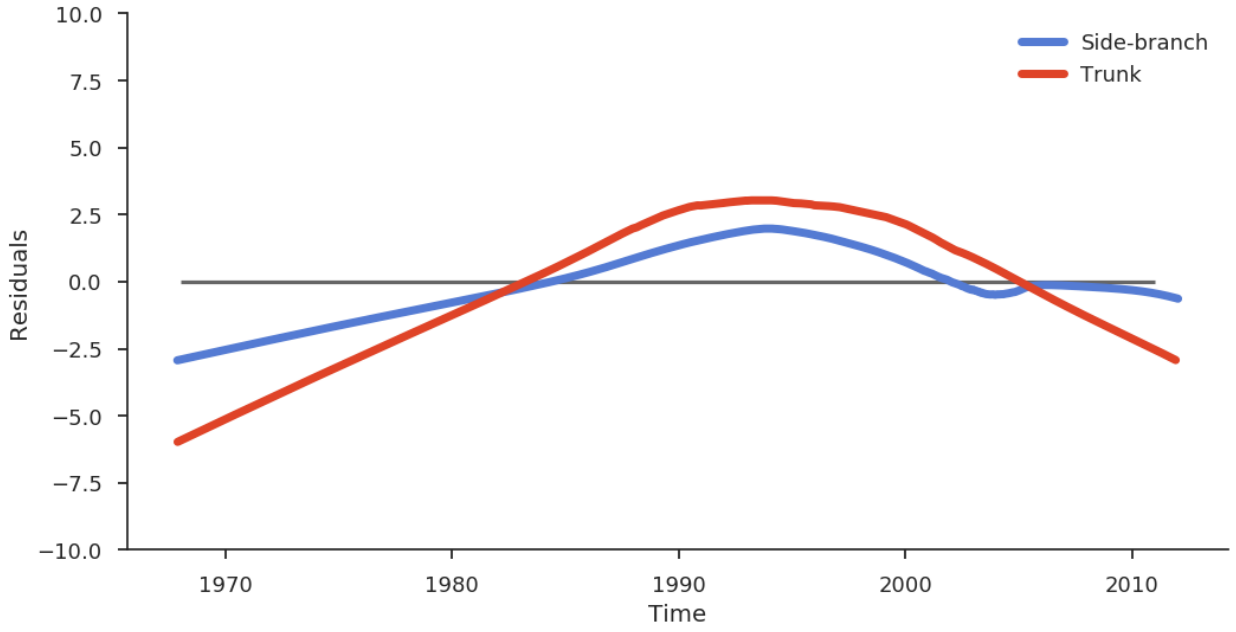

Figure 2.11: **Local weighted regression (LOWESS) lines for trunk and side-branch residuals in Figure 2.10.**

H3 sequences that gradually accumulated mutations that made them appear more like the Perth/2009 strain. This interpretation was supported by the fact that the Perth/2009 strain had one of the highest possible sequence preferences and therefore represented a kind of "fitness peak" for the sequence preference metric. From the decline of sequence preferences after 2009 even for trunk nodes, we concluded that the sequence preference was not a reliable fitness metric for long-term forecasts of H3N2 evolution. We resolved to use mutation-specific DMS preferences for the remainder of our analyses, to minimize the effects of epistasis on our results.

Chapter 3

VISUALIZATION OF SEASONAL INFLUENZA A/H3N2 EXPERIMENTAL PHENOTYPES

3.1 dms-view: Interactive visualization tool for deep mutational scanning data

With the exception of subsection 3.1.4, this work was originally published in *The Journal of Open Source Software* at `https://doi.org/10.21105/joss.02353`.

3.1.1 Summary and Purpose

The high-throughput technique of deep mutational scanning (DMS) [Fowler and Fields, 2014] has recently made it possible to experimentally measure the effects of all amino-acid mutations to a protein (Figure 3.1). Over the past five years, this technique has been used to study dozens of different proteins [Esposito et al., 2019] and answer a variety of research questions. For example, DMS has been used for protein engineering [Wrenbeck et al., 2017], understanding the human immune response to viruses [Lee et al., 2019], and interpreting human variation in a clinical setting [Starita et al., 2017, Gelman et al., 2019]. Accompanying this proliferation of DMS studies has been the development of software tools [Bloom, 2015, Rubin et al., 2017] and databases [Esposito et al., 2019] for data analysis and sharing. However, for many purposes it is important to integrate and visualize the DMS data in the context of other information, such as the 3-D protein structure or natural sequence-variation data. Currently, this visualization requires the use of multiple different tools including custom scripts, static visualization tools like MaveVis [Esposito et al., 2019], or protein structure software such as PyMol [Schrödinger, LLC, 2015]. No existing tools

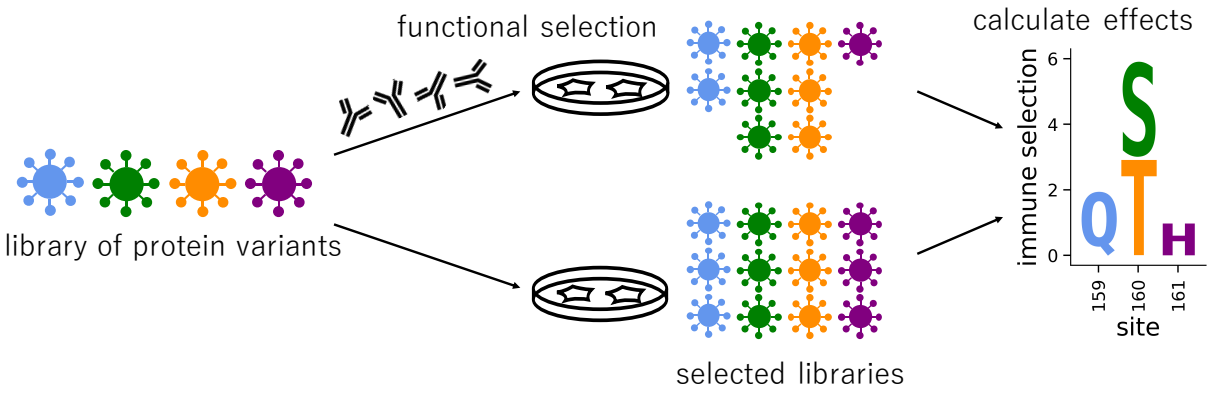

Figure 3.1: **Example deep mutational scanning workflow, modified from Lee et al. [2019].** The goal of this experiment is to quantify the how mutations affect a virus's ability to escape an antibody. The viral variant library contains all single amino-acid changes away from wildtype. The viral library is passaged in cell culture, with and without antibodies, to select for functional variants. Mutational effects are calculated based on deep sequencing of the pre-selected and post-selected libraries.

provide linked views of the protein structure and DMS data in a single interface to facilitate dynamic data exploration and sharing.

Here we describe *dms-view* (https://dms-view.github.io/), a flexible, web-based, interactive visualization tool for DMS data. *dms-view* is written in JavaScript and D3, and links site-level and mutation-level DMS data to a 3-D protein structure. The user can interactively select sites of interest to examine the DMS measurements in the context of the protein structure. *dms-view* tracks the input data and user selections in the URL, making it possible to save specific views of interactively generated visualizations to share with collaborators or to support a published study. Importantly, *dms-view* takes a flexible input data file so users can easily visualize their own DMS data in the context of protein structures of their choosing, and also incorporate additional information such amino-acid frequencies in natural alignments.

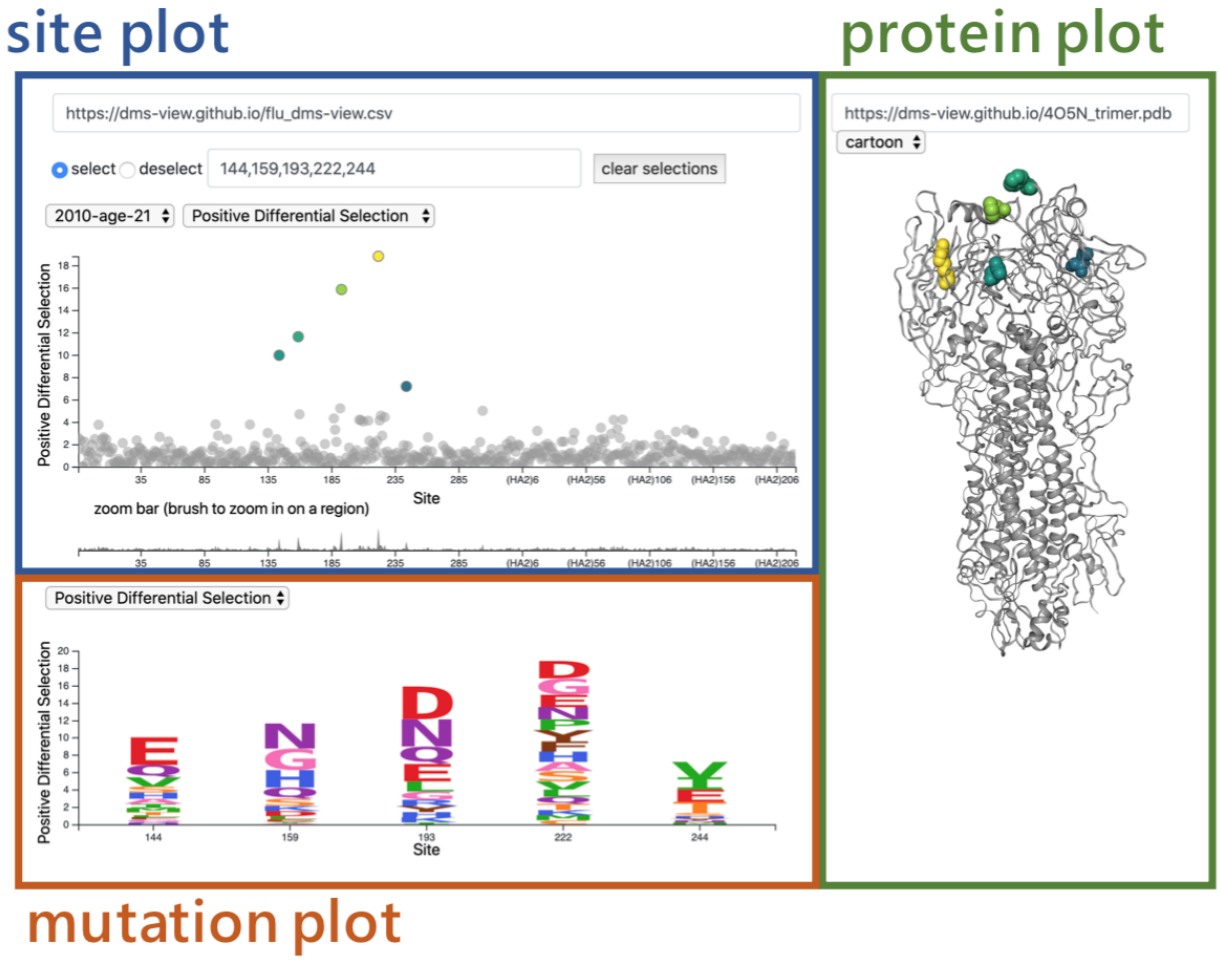

Figure 3.2: **Using *dms-view* to analyze DMS data.** For further exploration, please visit https://dms-view.github.io. The *dms-view* data section has three panels: the site plot, the mutation plot, and the protein structure plot. The interactive features for selecting sites and navigating are in the site plot panel. Here we show the five sites most highly targeted by human serum "2010-Age-21" from the study by Lee et al. [2019]. All five sites fall in the "globular head" of influenza virus HA.

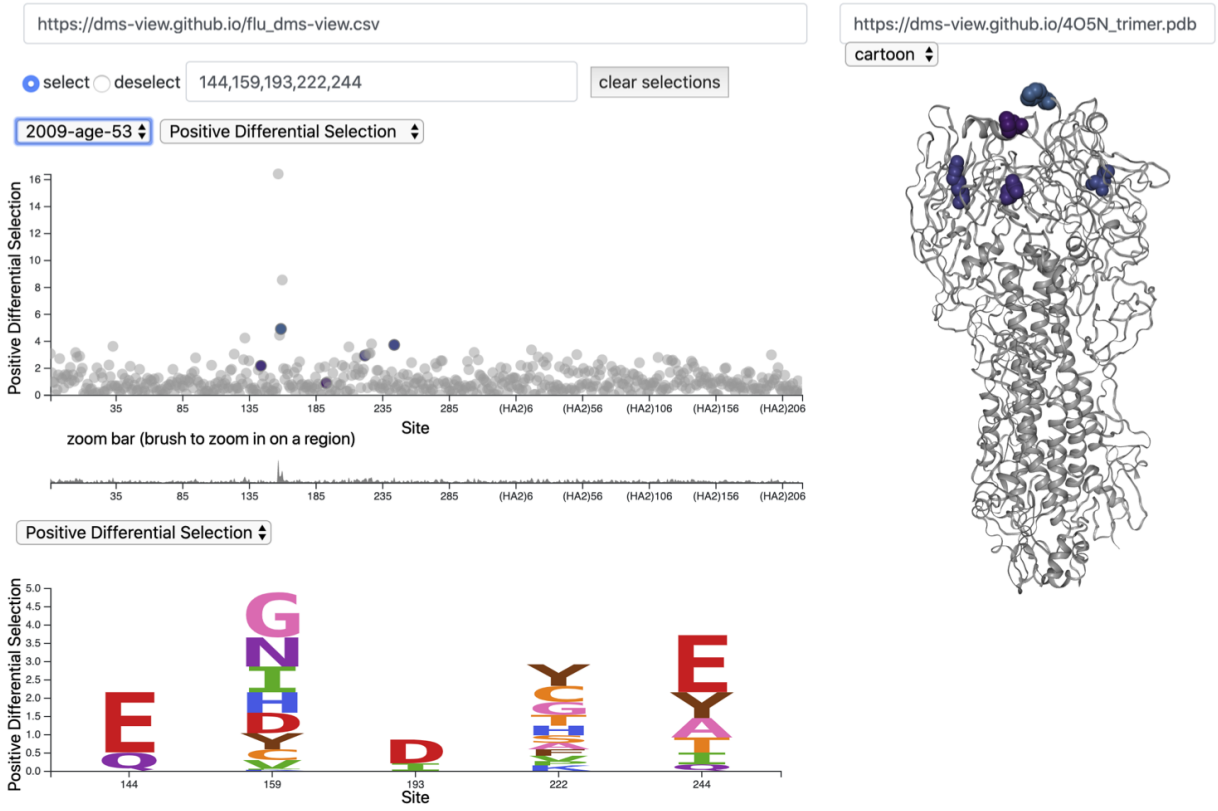

Figure 3.3: **The same five sites as in Figure 3.2 but now plotted with the data from a different human serum, "2009-age-53".** Using *dms-view* to compare, we see that different sites on HA are targeted by different sera.

Users can access *dms-view* at https://dms-view.github.io. The tool consists of a data section at the top and a description section at the bottom. The data section displays the user-specified data in three panels: the site-plot panel, the mutation-plot panel, and the protein-structure panel (Figure 3.2). When sites are selected in the site-plot panel, the individual mutation values are shown in the mutation-plot panel and highlighted on the protein structure. The user can toggle between different conditions, site- and mutation-level metrics, all of which are defined in the user-generated input file. The description section is at the bottom of the page, and allows the user to add arbitrary notes that explain the experimental setup, acknowledge data sources, or provide other relevant information. Note that dms-view is designed to visualize the effects of single mutations, not combinations of mutations.

Please visit the documentation at https://dms-view.github.io/docs to learn more about how to use the tool, how to upload a new dataset, or view case studies.

3.1.2 Case study: mapping influenza A virus escape from human sera

Using a DMS approach, Lee et al. [2019] measured how all amino-acid mutations to the influenza virus surface-protein hemagglutinin (HA) affected viral neutralization by human sera. For more information on the experimental setup, see the paper [Lee et al., 2019] or the GitHub repository (`https://github.com/jbloomlab/map_flu_serum_Perth2009_H3_HA`).

We visualized the Lee et al. [2019] serum mapping data using *dms-view*. To explore this dataset, please visit https://dms-view.github.io. In the *dms-view* visualization of these data, the conditions are the different human sera used for the selections. The site- and mutation-level metrics are different summary statistics measuring the extent that mutations escape from immune pressure.

Lee and colleagues asked two questions in their paper which can be easily explored using *dms-view*. First, are the same sites selected by sera from different people? To explore this question, we compared the site-level and mutation-level metric values for a specific set of sites between different conditions. Second, where on the protein structure are the highly selected

sites located? To explore this question, we selected specific sites of interest to be visualized on the 3-D protein structure.

Comparing site-level and mutation-level metric values for specific sites between conditions

To address whether or not the same sites are selected by different human sera using *dms-view*, we highlighted the most highly targeted sites for the human sera condition "Age 21 2010" in Figure 3.2 (144, 159, 193, 222, and 244). We then used the condition dropdown menu to toggle to the other sera. The highlighted sites remain highlighted after the condition is changed so we can easily see if the same sites are targeted in other conditions.

In Figure 3.3, we can see that there is no overlap of the sites selected by the human sera "2010-age-21" and the human sera "2009-age-53". These data are the default data for *dms-view*, so to explore this question in more detail please see https://dms-view.github.io.

View sites on the protein structure

To address where on the protein structure the targeted sites are located, we selected the most highly targeted sites (144, 159, 193, and 222) for the human sera condition "Age 21 2010" to highlight them on the protein structure.

In Figure 3.2, we can see that these sites cluster on the "head" of HA, which is known to be a common target of the human immune system [Chambers et al., 2015].

3.1.3 Code availability

dms-view is available at `https://dms-view.github.io`. Source code is available at `https://github.com/dms-view/dms-view.github.io`. Documentation is available at `https://dms-view.github.io/docs` and case studies are available at `https://dms-view.github.io/docs/casestudies/`.

3.1.4 Visual design decisions behind dms-view

With one notable exception, we designed *dms-view* to use the most expressive and effective representations of the underlying data. Expressiveness describes the ability of a visualization to communicate the most relevant information in a dataset, while effectiveness describes how rapidly a viewer can perceive the relevant information [Mackinlay, 1986]. Expressive and effective visualizations require the designer to select appropriate representations of the data based on the designer's understanding of human perception, available modes of representation, and the data being visualized.

To these ends, we selected encodings of DMS data that were most appropriate for these data. In the "site plot", we represented quantitative values as circles on a discrete genomic coordinate scale of the x-axis and a continuous scale on the y-axis. We chose a redundant color encoding of the site metric values to both emphasize the differences between multiple selected values and to link selected sites in the "site plot" with their placement on the protein structure. We colored unselected points gray to communicate their position in the site plot and distinguish them from selected points with color. We used reduced opacity for unselected points, to reduce occlusion caused by overlapping circles and maintain a reasonable circle size for user interactions. We increased the expressiveness of site plot data by providing details on demand that users can access by hovering over specific points. We increased the effectiveness of site plot data by adding filter, zoom, and select interactions that allow the user to reduce the number of data points to inspect at any given time.

In the "mutation plot", we used the same approach of representing quantitative data on a discrete genomic coordinate scale and a continuous "mutation metric" scale. In contrast to the site plot, we chose to only show mutation-level data for sites that users selected in the site plot. We also chose to represent these quantitative data by the height of the alphabetical letter associated with the amino acid of the mutation at each position. This encoding, also known as a "logo plot", allowed us to communicate additional information about the relative contribution of each amino acid to an overall site metric. Although logo plots are technically a

less effective representation of these quantitative data, due to their use of arbitrary differences in area associated with each letter, they are a standard representation of these data in the field of DMS research. We applied a color encoding to each mutation metric value such that amino acids were colored according to their biophysical group (hydrophobic, positively charged, negatively charged, etc.). As with the site plot, we provided details on demand when the user hovers over specific mutations at each genomic position. A major limitation of our logo plot implementation is its fixed width. As the user selects more sites in the site plot, we scale down the width of each logo plot column. When more than a couple dozen sites have been selected, the amino acid letters in the logo plot become illegible for all except the higher values.

Finally, we encoded the 3D structure of a given protein using a reduced space representation known as the "cartoon" view. In this view, selected sites appear more clearly as large bubbles against a ribbon of the unselected sites. We enabled the protein viewer's rotation functionality that allows users to spin the protein around in three dimensions. As with the site plot, we encoded the site metric of each position by color on a viridis scale. We made this choice primarily because the most effective representations of these quantitative data (the spatial x, y, and z axes) were already in use by the protein structure itself. We also provided details on demand for each protein site.

3.2 Visualization of antigenic phenotypes

3.2.1 Introduction

A primary component of seasonal influenza A/H3N2 evolution is the ability of viruses to acquire mutations that allow them to escape antibodies from previous infections. This process, known as antigenic drift, changes the appearance of viral surface proteins hemagglutinin (HA) and neuraminidase (NA). Viruses balance escape from antibodies with the maintenance of their protein functions. The HA surface protein allows viruses to bind to the surface of new host cells and initiate infection. When antibodies bind to HA, they can prevent viruses from

binding and infecting cells. Viruses that acquire mutations to HA that prevent antibodies from binding but do not disrupt the ability of HA to bind to host cells should be able to infect hosts.

Experimental measurements of antigenic drift allow researchers to quantify how well viruses with different HA mutations can escape detection by antibodies. Until recently, the most reliable of these experimental measurements were hemagglutination inhibition (HI) assays [Hirst, 1943]. In these assays, researchers place red blood cells into a multi-well plate and add one test virus per well. When a well contains only blood cells and virus, the virus binds to the blood cells causing them to agglutinate into a wide, red dot that fills the well. Next, researchers add two-fold dilutions of antisera from naive ferrets that were infected by a single reference virus. When the antisera contains enough antibodies to effectively bind the virus, the blood cells do not agglutinate and instead sink to the bottom of the well in a small red dot. The highest dilution of antisera required to inhibit agglutination provides the "titer" measurement between the reference and test viruses. When the test virus is the same as the reference virus, the assay provides an autologous titer measurement. When the test and reference viruses differ, the assay provides a heterologous titer measurement.

Researchers report titer measurements as both raw and normalized values. Raw titer measurements represent the denominator associated with the minimum dilution required to inhibit agglutination. These two-fold dilution series have raw measurements like 80, 160, 320, etc. such that lower numbers represent the presence of more antisera in the dilution. These raw measurements can also be represented more conveniently on a $\log_2$ scale [Smith et al., 2004, Bedford et al., 2014]. Antisera vary in their potency such that some sera always require higher or lower dilutions to inhibit agglutination regardless of the test virus. For example, a low potency antiserum requires a lower dilution to inhibit agglutination by the same reference virus resulting in a low autologous titer. To account for this variable potency, we normalize titer measurements by subtracting the $\log_2$ titer between a test and reference virus (the heterologous titer) from the $\log_2$ titer between the reference virus and its own

antisera (the autologous titer). The resulting normalized titers enable comparisons of antigenic distances across reference viruses. Importantly, viruses with a $\log_2$ distance greater than 2 are considered antigenically distinct for the purposes of deciding updates to vaccine composition [Katz et al., 2011].

3.2.2 Previous visual representations of antigenic distances

To understand broad patterns of antigenic drift beyond simple pairwise relationships, antigenic distance must be summarized by statistical or visual representations. For example, the method of antigenic cartography maps multidimensional antigenic distances to a two-dimensional space using dimensionality reduction methods [Smith et al., 2004, Bedford et al., 2014]. These map-like visualizations reveal long-term patterns and trends in antigenic drift within seasonal influenza lineages like the punctuated emergence of new antigenic clusters every few years (Figure 3.4). However, antigenic cartography is less suited to representations of short-term antigenic drift on the same time scale as annual vaccine updates in each hemisphere.

Alternate visualizations created in the nextflu [Neher and Bedford, 2015] and Nextstrain [Hadfield et al., 2018] frameworks address the needs of influenza researchers who make decisions about vaccine composition. Within nextflu, all available pairwise antigenic distances between a selected reference virus's antisera and test viruses are represented by colored tips on a phylogenetic tree constructed from HA sequences (Figure 3.5). This interactive visualization allows users to select a specific reference virus by clicking a "gear" icon drawn on the phylogeny where the reference virus occurs. All tips in the tree that have titer measurements to the selected reference appear as circles colored by the quantitative value of their antigenic distance (e.g., smaller distances are blue, larger distances are red). This representation of the pairwise data allows users to identify phylogenetic clades that are antigenically distance from a given antiserum or that are missing measurements from that antiserum. Influenza virologists use this information to select potential vaccine candidates that "cover" the most extant clades and prioritize which HI assays to perform next. The benefits of visualizing the phylogenetic

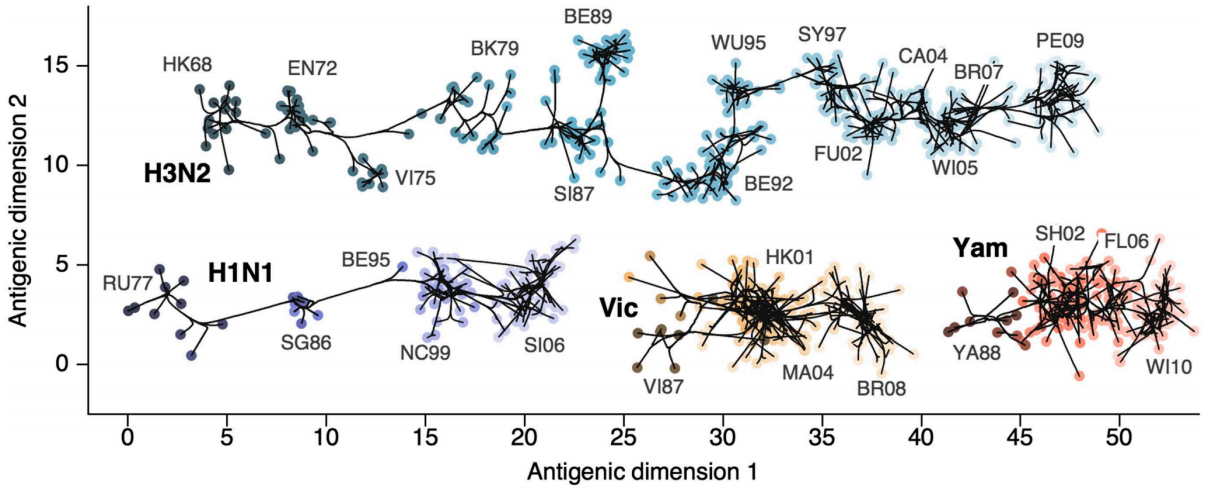

Figure 3.4: **Antigenic cartography of titer measurements from seasonal influenza lineages from Figure 1 of Bedford et al. [2014].** Raw titer measurements in a multidimensional matrix are transformed through Bayesian multidimensional scaling (BMDS) to a two-dimensional representation. These antigenic maps reveal long-term patterns of antigenic drift.

context of antigenic distances for a single antiserum are balanced by visual design costs of not showing antigenic distances for all antisera in a single view.

Another recent representation of pairwise antigenic distances attempts to address the limitations of the phylogenetic visualization. This representation summarizes the mean antigenic distances between a subset of relevant reference viruses (i.e., likely vaccine candidates) and all tests viruses within each extant clade. The resulting heatmaps use the x-axis to encode phylogenetic clades, the y-axis to encode reference viruses, and color to encode the mean $\log_2$ distance between a given reference virus and corresponding test viruses (Figure 3.6). These static heatmaps express more data than the phylogenetic representation, allowing decision-makers to identify qualitative trends across antisera and clades in biannual reports to the World Health Organization [Bedford and Neher, 2018, Bedford et al., 2019]. As with all heatmap, these titer matrices suffer reduced expressiveness by encoding the most valuable quantitative data with color instead of a spatial scale.

3.2.3 Improved visual representations of antigenic distances

Given the benefits and costs of existing visualizations of antigenic distances, we wondered whether we could apply user-driven design and established visual design principles to produce a more expressive and effective visualization for influenza virologists. To this end, we established the goals of users who would interact with our visualizations, identified the most effective encodings for the data users needed to explore, and composed an interactive visualization from these encodings to address the desired goals.

Based on informal user interviews with collaborators at the Influenza Division of the Centers for Disease Control and Prevention, we identified a list of primary goals for the phylogenetic and heatmap visualizations. Users wanted to know which currently circulating clades of influenza have measurements against a given serum, to prioritize which clades to select test viruses from in future HI experiments. Additionally, users wanted to know which available antiserum has the lowest antigenic distance across all circulating clades, to identify potential

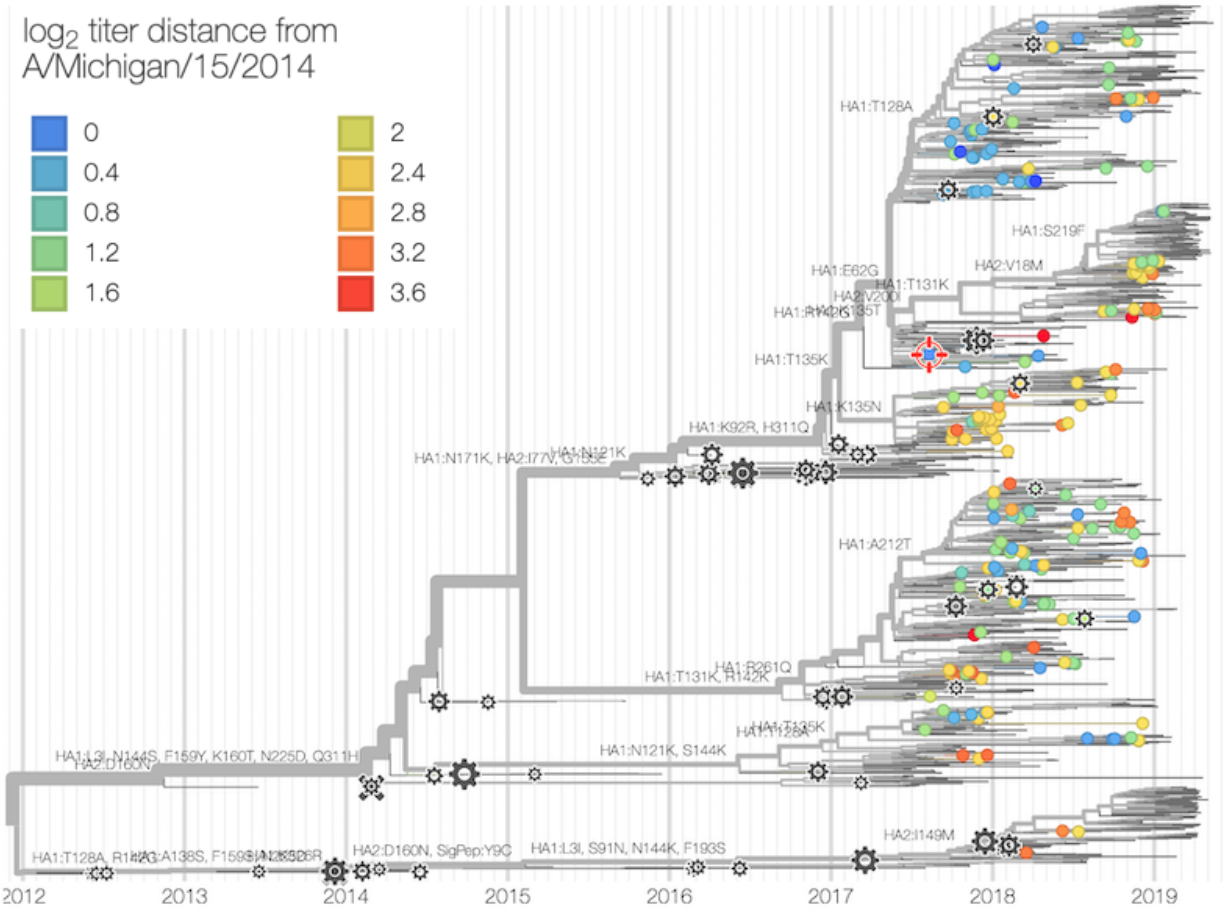

Figure 3.5: **Antigenic distance (mean $\log_2$ titer drop) between test strains (colored tips) and a selected reference virus's antiserum (red crossmark) in the context of a H3 phylogeny of recently circulating strains from nextflu [Neher and Bedford, 2015].**

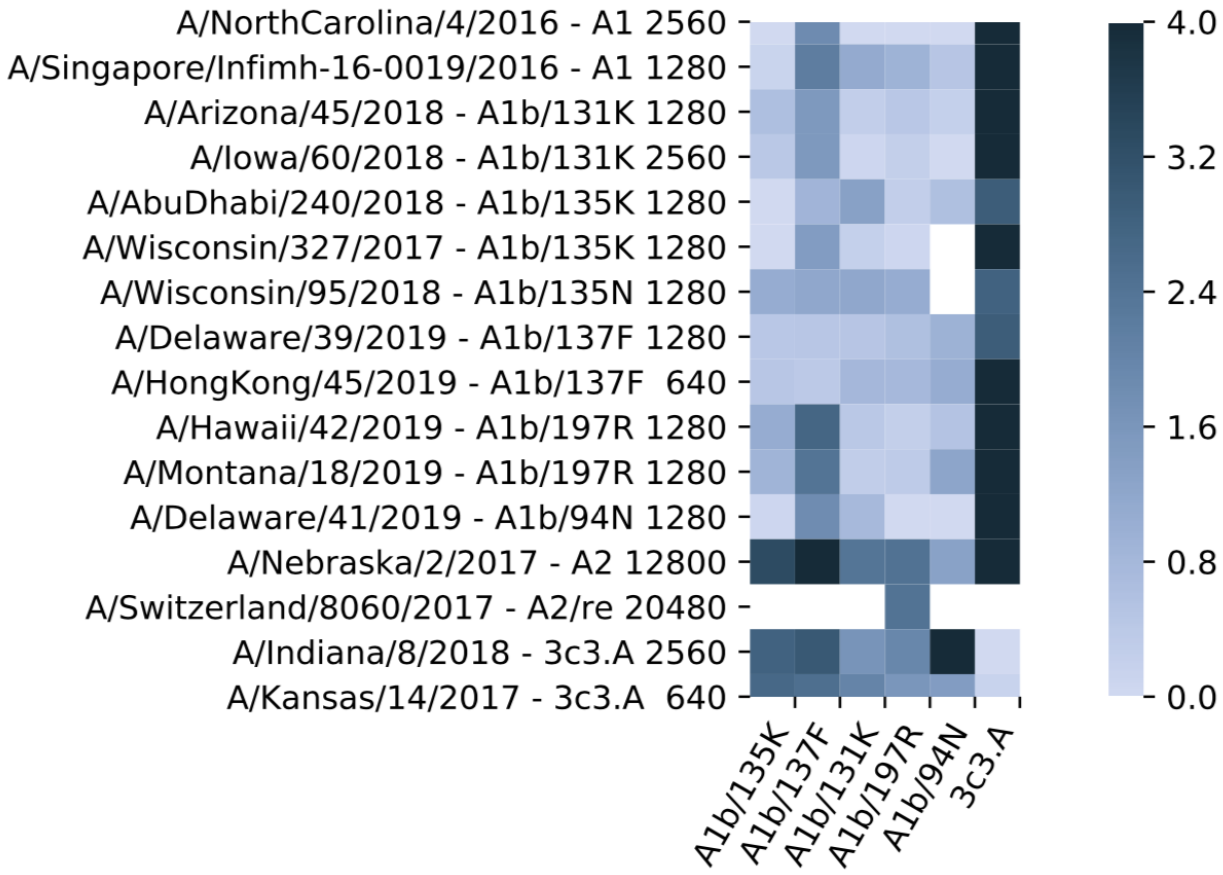

Figure 3.6: **Antigenic distance (mean $\log_2$ titer drop) between test strains sampled from recently circulating clades (columns) and antisera for representative viruses from these clades (rows).** The median autologous raw titer of antisera for representative viruses is shown after the corresponding virus's name and clade membership. White squares represent missing data.

vaccine candidates. Finally, users wanted to compare the antigenic diversity of extant clades. From these user goals, we decided that an optimal visualization would summarize the distribution of antigenic distances by antiserum and clade.

We observed that the existing titer matrix heatmaps addressed most of the user goals except for communicating which extant clades were missing measurements. Both the phylogenetic and heatmap views use color to encode the most relevant quantitative data of antigenic distance. Previous visualization design research has shown that quantitative data are more effectively represented by positional encodings (e.g., x- or y-axis positions) whereas nominal data (e.g., phylogenetic clades) can be effectively encoded with color [Mackinlay, 1986]. In the phylogenetic view, the two available positional axes are used to represent time and the unitless phylogenetic position of nodes. Neither of these data are relevant to the user goals described above. In the titer matrix heatmaps, the two positional axes are used to encode two nominal data types (reference virus name and clade name).

We reasoned that we could make a more effective visualization that addressed most user goals by changing the encoding of data in the titer matrix heatmaps. Specifically, we chose to temporally omit clade names from the visualization and encode the antigenic distances on the positional x-axis. By encoding antigenic distance on a positional axis, we could annotate relevant thresholds for antigenic distances, show all available measurements for each reference virus at once, and display a summary statistic (mean antigenic distance) for each reference virus. We chose to maintain the encoding of nominal reference virus names on the y-axis, since most user goals require interrogation of specific antisera. With color available as an additional channel, we decided to encode the antigenic class of each antigenic distance with color. We implemented this design as an interactive visualization with the Altair visualization framework [VanderPlas et al., 2018] at `https://cse512-19s.github.io/A3-Influenza-vaccine/` (Figure 3.7).

This interactive visualization allows users to quickly compare which antisera have more measurements than others by the presence or absence of circles. By interactively selected

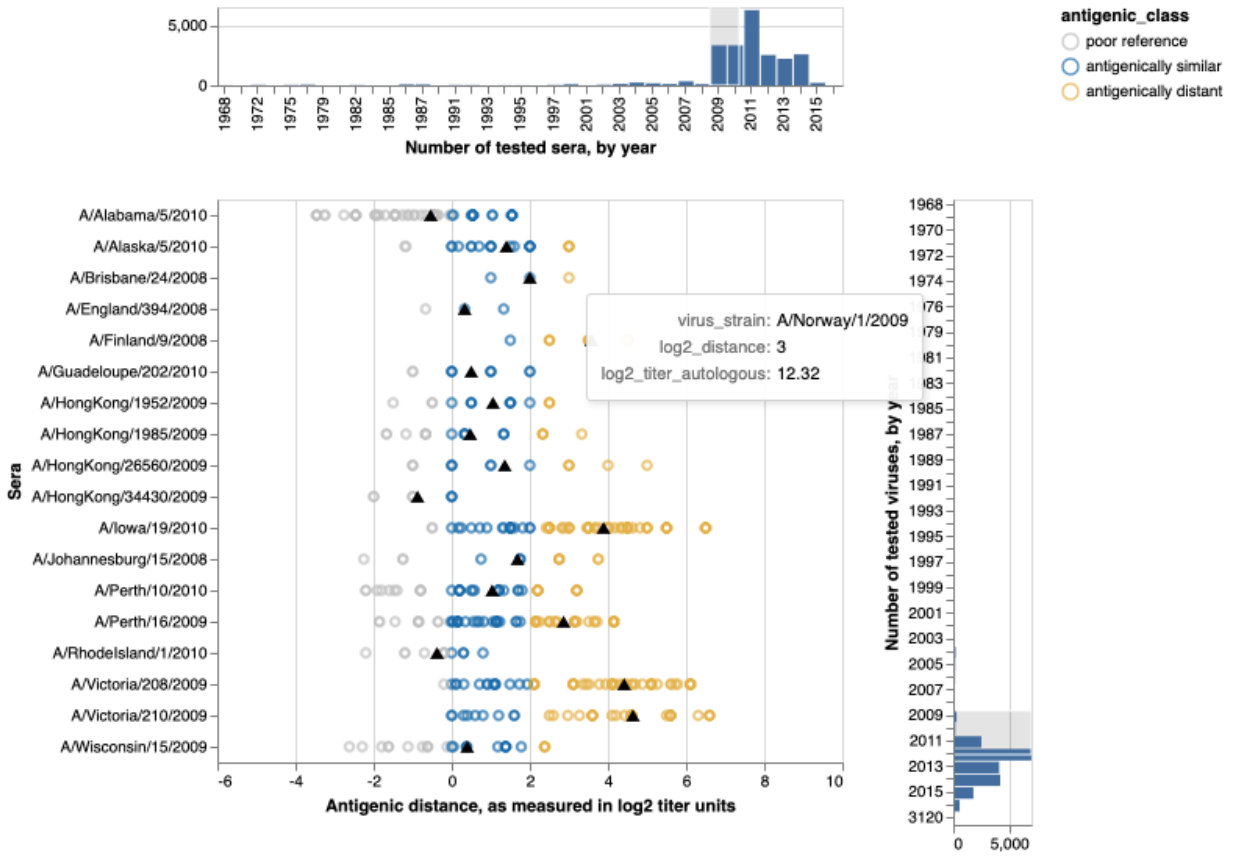

Figure 3.7: **Antigenic distance (mean $\log_2$ titer drop) between test viruses (circles) by serum reference virus.** Colors indicate distances at different actionable thresholds including measurements with a poor reference titer measurement (gray), measurements above two $\log_2$ units indicating substantial antigenic drift (yellow), and all other measurements (blue). Black triangles show the mean antigenic distance per serum. Histograms at the top and right represent the number of sera or test viruses per year, respectively. The interactive version of the visualization allows users to filter distances reported to a range of years for sera and test viruses by clicking and dragging across the corresponding histogram. The interactive visualization provides details on demand when users hover over specific circles including the name of the test virus, the $\log_2$ distance between test and reference viruses, and the $\log_2$ autologous distance of the corresponding serum.

years of test viruses to display, users can filter the display to see the distribution of antigenic distances (and mean distance) to recent viruses from all selected antisera. The distribution of antigenic distances identifies which antisera have perform poorly in HI assays (those with many negative normalized values, shown in gray) and which are antigenically distinct from most circulating viruses (i.e., poor vaccine candidates).

Despite its strengths, this visualization also has several weaknesses. Distances are not grouped anywhere by clade, making comparisons between antisera by clade impossible. This missing information also prevents users from identifying which clades need more titer measurements. Finally, the alphabetic order of reference viruses on the y-axis wastes an opportunity to communicate more important information to the user. For example, if users could choose to sort the reference viruses by their mean antigenic distance or by their number of measurements, this single visualization could help users more rapidly identify antisera to consider for vaccine candidates or for additional HI experiments. We anticipate that the inclusion of clade status in a future implementation (by color or in small multiples) and the default sorting of reference viruses by mean antigenic distance could address these remaining limitations (for example, Figure 3.8). The resulting visualization could provide a more effective real-time view of the same data currently displayed statically in titer matrix heatmaps.

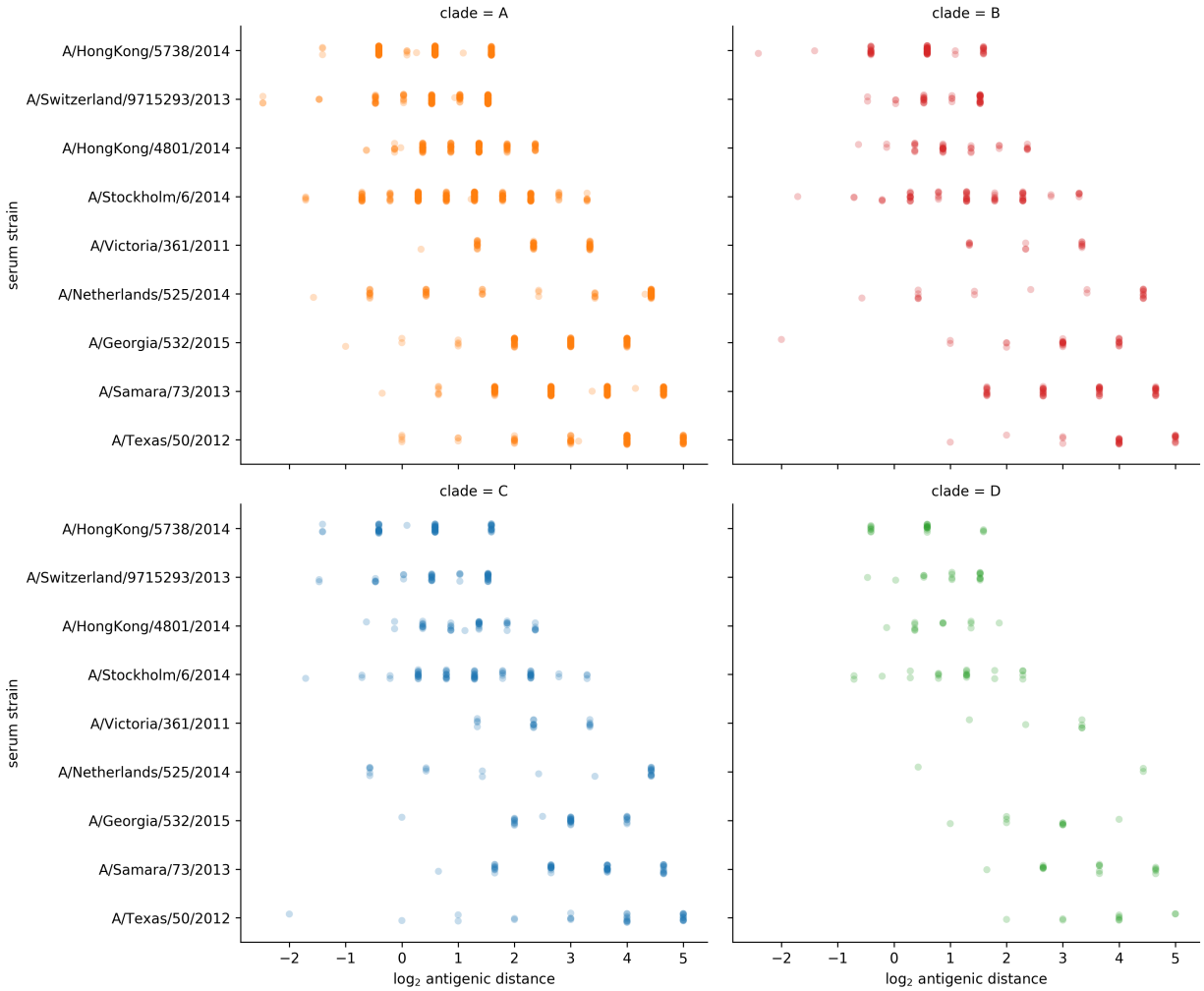

Figure 3.8: **Mockup of an alternative representation of antigenic distances by clade and antiserum.** This encoding of data would allow users to accomplish primary goals including identifying antisera without many measurements, distance of reference viruses across clades, and comparison of antigenic distances between clades.

Chapter 4

INTEGRATING GENOTYPES AND PHENOTYPES IMPROVES LONG-TERM FORECASTS OF SEASONAL INFLUENZA A/H3N2 EVOLUTION

This work was originally published in *eLife* at `https://doi.org/10.7554/eLife.60067`.

4.1 Introduction

Seasonal influenza virus infects 5–15% of the global population every year causing an estimated 250,000 to 500,000 deaths annually with the majority of infections caused by influenza A/H3N2 [World Health Organization, 2014]. Vaccination remains the most effective public health response available. However, frequent viral mutation results in viruses that escape previously acquired human immunity. 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. The WHO GISRS uses these data to select vaccine viruses that should best represent circulating viruses in the next influenza season. 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 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., 2017]. 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. Until recently, the hemagglutination inhibition (HI) assay has been the primary experimental measure of antigenic cross-reactivity between pairs of circulating viruses [Hirst, 1943]. Most modern H3N2 strains carry a glycosylation motif that reduces their binding efficiency in HI assays [Chambers et al., 2015, Zost et al., 2017], prompting the increased use of virus neutralization assays including the neutralization-based focus reduction assay (FRA) [Okuno et al., 1990]. Together, these two assays are the gold standard in virus antigenic characterizations for vaccine strain selection, but they are laborious and low-throughput compared to genome sequencing [Wood et al., 2012]. As a result, researchers have developed computational methods to predict influenza evolution from sequence data alone [Łuksza and Lässig, 2014, Steinbrück et al., 2014, Neher et al., 2014].

Despite the promise of these sequence-only models, they explicitly omit experimental measurements of antigenic or functional phenotypes. Recent developments in computational methods and influenza virology have made it feasible to integrate these important metrics of influenza fitness into a single predictive model. For example, phenotypic measurements of antigenic drift are now accessible through phylogenetic models [Neher et al., 2016] and functional phenotypes for HA are available from deep mutational scanning (DMS) experiments [Lee et al., 2018]. We describe an approach to integrate previously disparate sequence-only models of influenza evolution with high-quality experimental measurements of antigenic drift and

functional constraint.

The influenza community has long recognized the importance of incorporating HI phenotypes and other experimental measurements of viral phenotypes with existing forecasting methods to inform the vaccine design process [Gandon et al., 2016, Morris et al., 2017, Lässig et al., 2017]. Although several distinct efforts have made progress in using HI phenotypes to evaluate the evolution of seasonal influenza [Steinbrück et al., 2014, Neher et al., 2016], published methods stop short of developing a complete forecasting framework wherein the evolutionary contribution of HI phenotypes can be compared and contrasted with new and existing fitness metrics. However, unpublished work by Łuksza and Lässig submitted to the WHO GISRS network incorporates antigenic phenotypes into fitness-based predictions [Morris et al., 2017, Łuksza, 2020]. Here, we provide an open source framework for forecasting the genetic composition of future seasonal influenza populations using genotypic and phenotypic fitness estimates. We apply this framework to HA sequence data shared via the GISAID EpiFlu database [Shu and McCauley, 2017] and to HI and FRA titer data shared by WHO GISRS Collaborating Centers in London, Melbourne, Atlanta and Tokyo. We systematically compare potential predictors and show that HI phenotypes enable more accurate long-term forecasts of H3N2 populations compared to previous metrics based on epitope mutations alone. We also find that composite models based on phenotypic measures of antigenic drift and genotypic measures of functional constraint consistently outperform any fitness models based on individual genotypic or phenotypic metrics.

4.2 Results

4.2.1 A distance-based model of seasonal influenza evolution

We developed a framework to forecast seasonal influenza evolution inspired by the Malthusian growth fitness model of Łuksza and Lässig [2014]. As with this original model, we forecasted the frequencies of viral populations one year in advance by applying to each virus strain an exponential growth factor scaled by an estimate of the strain's fitness (Figure 5.2 and

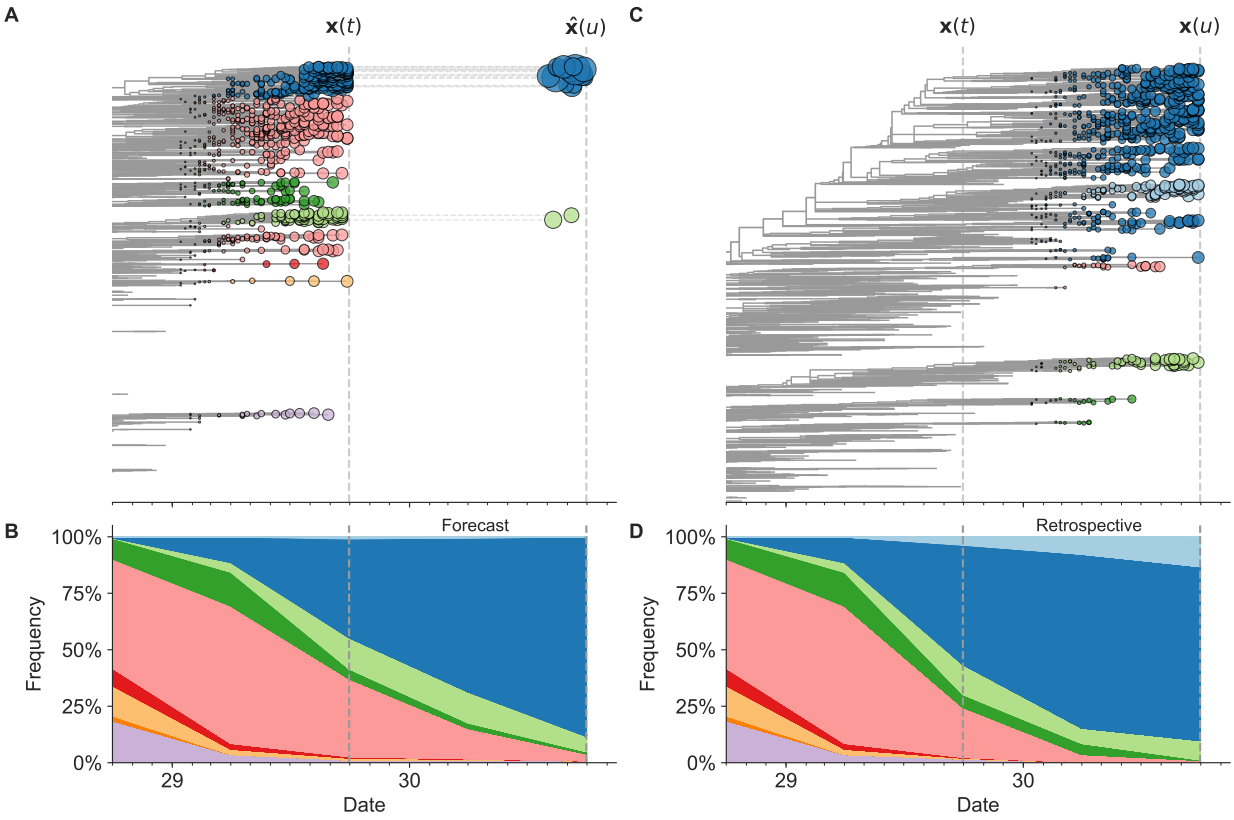

Figure 4.1: **Schematic representation of the fitness model for simulated H3N2-like populations.** The fitness of strains at timepoint t determines the estimated frequency of strains with similar sequences one year in the future at timepoint u. Genetically similar strains by amino acid sequence have similar colors (Methods). A) Strains at timepoint t, $\mathbf{x}(t)$, are shown in their phylogenetic context and sized by their frequency at that timepoint. The estimated future population at timepoint u, $\hat{\mathbf{x}}(u)$, is projected to the right with strains scaled in size by their projected frequency based on the known fitness of each simulated strain. B) The frequency trajectories of strains at timepoint t to u represent the predicted the growth of the dark blue strains to the detriment of the pink strains. C) Strains at timepoint u, $\mathbf{x}(u)$, are shown as in panel A. D) The observed frequency trajectories of strains at timepoint u broadly recapitulate the model's forecasts while also revealing increased diversity of sequences at the future timepoint that the model could not anticipate.

Model	Populations	Data type	Fitness category	Originally implemented by
true fitness	simulated	simulated populations	positive control	this study
naive	simulated, natural	HA sequences	negative control	this study
epitope antigenic novelty	simulated, natural	HA sequences	antigenic drift	Łuksza and Lässig [2014]
epitope ancestor	simulated, natural	HA sequences	antigenic drift	Łuksza and Lässig [2014]
HI antigenic novelty	natural	serological assays	antigenic drift	this study
mutational load	simulated, natural	HA sequences	functional constraint	Łuksza and Lässig [2014]
deep mutational scanning (DMS) mutational effects	natural	DMS assays	functional constraint	Lee et al. [2018]
local branching index (LBI)	simulated, natural	HA sequences	clade growth	Neher et al. [2014]
delta frequency	simulated, natural	HA sequences	clade growth	this study

Table 4.1: **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.

Equation 4.1). 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. 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 5.2).

We estimated viral fitness with biologically-informed metrics including those originally defined by Łuksza and Lässig [2014] of epitope antigenic novelty and mutational load (non-epitope mutations) as well as four more recent metrics including hemagglutination inhibition (HI) antigenic novelty [Neher et al., 2016], deep mutational scanning (DMS) mutational effects [Lee et al., 2018], local branching index (LBI) [Neher et al., 2014], and change in clade frequency over time (delta frequency) (Table 4.1). All of these metrics except for HI antigenic novelty and DMS mutational effects rely only on HA sequences. The antigenic novelty metrics estimate how antigenically distinct each strain at time t is from previously circulating strains based on either genetic distance at epitope sites or $\log_2$ titer distance from HI measurements. Increased antigenic drift relative to previously circulating strains is expected to correspond to increased viral fitness. Mutational load estimates functional constraint by measuring the number of putatively deleterious mutations that have accumulated in each strain since their ancestor in the previous season. DMS mutational effects provide a more comprehensive biophysical model of functional constraint by measuring the beneficial or deleterious effect of each possible single amino acid mutation in HA from the background of a previous vaccine strain, A/Perth/16/2009. The growth metrics estimate how successful populations of strains have been in the last six months based on either rapid branching in the phylogeny (LBI) or the change in clade frequencies over time (delta frequency).

We fit models for individual fitness metrics and combinations of metrics that we anticipated would be mutually beneficial. For each model, we learned coefficient(s) that minimized the earth mover's distance between HA amino acid sequences from the observed population one year in the future and the estimated population produced by the fitness model (Equation 4.2). We evaluated model performance with time-series cross-validation such that better models reduced the earth mover's distance to the future on validation or test data (Figures 4.21 and 4.23). The earth mover's distance to the future can never be zero, because each model makes predictions based on sequences available at the time of prediction and cannot account for new mutations that occur during the prediction interval. We calculated the lower bound for each

model's performance as the optimal distance to the future possible given the current sequences at each timepoint. As an additional reference, we evaluated the performance of a "naive" model that predicted the future population would be identical to the current population. We expected that the best models would consistently outperform the naive model and perform as close as possible to the lower bound.

4.2.2 Models accurately forecast evolution of simulated H3N2-like viruses

The long-term evolution of influenza H3N2 hemagglutinin has been previously described as a balance between positive selection for substitutions that enable escape from adaptive immunity by modifying existing epitopes and purifying selection on domains that are required to maintain the protein's primary functions of binding and membrane fusion [Bush et al., 1999, Neher, 2013, Łuksza and Lässig, 2014, Koelle and Rasmussen, 2015]. To test the ability of our models to accurately detect these evolutionary patterns under controlled conditions, we simulated the long-term evolution of H3N2-like viruses under positive and purifying selection for 40 years (Methods, Figure 4.21). These selective constraints produced phylogenetic structures and accumulation of epitope and non-epitope mutations that were consistent with phylogenies of natural H3N2 HA (Figure 4.22, Tables 4.4 and 4.5). We fit models to these simulated populations using all sequence-only fitness metrics. As a positive control for our model framework, we also fit a model based on the true fitness of each strain as measured by the simulator.

We hypothesized that fitness metrics associated with viral success such as true fitness, epitope antigenic novelty, LBI, and delta frequency would be assigned positive coefficients, while metrics associated with fitness penalties, like mutational load, would receive negative coefficients. We reasoned that both LBI and delta frequency would individually outperform the mechanistic metrics as both of these growth metrics estimate recent clade success regardless of the mechanistic basis for that success. Correspondingly, we expected that a composite model of epitope antigenic novelty and mutational load would perform as well as or better

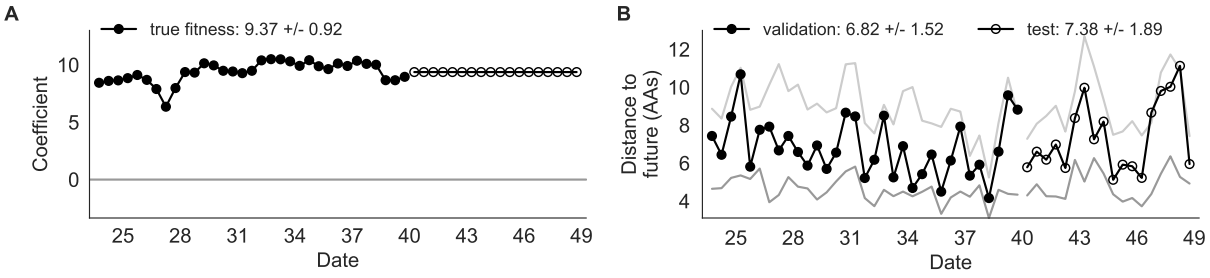

Figure 4.2: **Simulated population model coefficients and distances between projected and observed future populations as measured in amino acids (AAs).** A) Coefficients are shown per validation timepoint (solid circles, N=33) with the mean ± standard deviation in the top-left corner. For model testing, coefficients were fixed to their mean values from training/validation and applied to out-of-sample test data (open circles, N=18). B) Distances between projected and observed populations are shown per validation timepoint (solid black circles) or test timepoint (open black circles). The mean ± standard deviation of distances per validation timepoint are shown in the top-left of each panel. Corresponding values per test timepoint are in the top-right. The naive model's distances to the future for validation and test timepoints (light gray) were 8.97 ± 1.35 AAs and 9.07 ± 1.70 AAs, respectively. The corresponding lower bounds on the estimated distance to the future (dark gray) were 4.57 ± 0.61 AAs and 4.85 ± 0.82 AAs.

than the growth metrics, as this model would include both primary fitness constraints acting on our simulated populations.

As expected, the true fitness model outperformed all other models, estimating a future population within 6.82 ± 1.52 amino acids (AAs) of the observed future and surpassing the naive model in 32 (97%) of 33 timepoints (Figure 4.2, Table 4.2). Although the true fitness model performed better than the naive model's average distance of 8.97 ± 1.35 AAs, it did not reach the closest possible distance between populations of 4.57 ± 0.61 AAs. With the exception of epitope antigenic novelty, all biologically-informed models consistently outperformed the naive model (Figure 4.3, Table 4.2). LBI was the best of these models, with a distance to the future of 7.57 ± 1.85 AAs. This result is consistent with the fact that the LBI is a correlate of fitness in models of rapidly adapting populations [Neher et al., 2014]. Indeed, both growth-based models received positive coefficients and outperformed the mechanistic models. The mutational load metric received a consistently negative coefficient with an average distance of 8.27 ± 1.35 AAs.

Surprisingly, the composite model of epitope antigenic novelty and mutational load did not perform better than the individual mutational load model (Figure 4.4). The antigenic novelty fitness metric assumes that antigenic drift is driven by nonlinear effects of previous host exposure [Łuksza and Lässig, 2014] that are not explicitly present in our simulations. To understand whether positive selection at epitope sites might be better represented by a linear model, we fit an additional model based on an "epitope ancestor" metric that counted the number of epitope mutations since each strain's ancestor in the previous season. This linear fitness metric slightly outperformed the antigenic novelty metric (Table 4.2). Importantly, a composite model of the epitope ancestor and mutational load metrics outperformed all other epitope-based models and the individual mutational load model (Figure 4.4). From these results, we concluded that our method can accurately estimate the evolution of simulated populations, but that the fitness of simulated strains was dominated by purifying selection and only weakly affected by a linear effect of positive selection at epitope sites.

Model	Coefficients	Distance to future (AAs) Validation	Test	Model > naive Validation	Test
true fitness	9.37 +/- 0.92	6.82 +/- 1.52*	7.38 +/- 1.89*	32 (97%)	16 (89%)
LBI	1.31 +/- 0.33	7.24 +/- 1.66*	7.10 +/- 1.19*	32 (97%)	18 (100%)
+ mutational load	-1.77 +/- 0.49				
LBI	2.26 +/- 1.06	7.57 +/- 1.85*	7.51 +/- 1.20*	29 (88%)	17 (94%)
delta frequency	1.46 +/- 0.44	8.13 +/- 1.44*	8.65 +/- 1.99*	26 (79%)	13 (72%)
epitope ancestor	0.35 +/- 0.07	8.20 +/- 1.39*	8.17 +/- 1.52*	29 (88%)	17 (94%)
+ mutational load	-1.57 +/- 0.13				
mutational load	-1.49 +/- 0.12	8.27 +/- 1.35*	8.20 +/- 1.50*	29 (88%)	17 (94%)
epitope antigenic novelty	0.03 +/- 0.19	8.33 +/- 1.35*	8.22 +/- 1.51*	28 (85%)	17 (94%)
+ mutational load	-1.38 +/- 0.39				
epitope ancestor	0.14 +/- 0.11	8.96 +/- 1.35	9.03 +/- 1.68*	20 (61%)	13 (72%)
naive	0.00 +/- 0.00	8.97 +/- 1.35	9.07 +/- 1.70	0 (0%)	0 (0%)
epitope antigenic novelty	-0.03 +/- 0.19	9.03 +/- 1.37	9.07 +/- 1.69	14 (42%)	7 (39%)

Table 4.2: **Simulated population model coefficients and performance on validation and test data ordered from best to worst by distance to the future in the validation analysis.** Coefficients are the mean ± standard deviation for each metric in a given model across 33 training windows. Distance to the future (mean ± standard deviation) measures the distance in amino acids between estimated and observed future populations. Distances annotated with asterisks (*) were significantly closer to the future than the naive model as measured by bootstrap tests (see Methods and Figure 4.24). The number of times (and percentage of total times) each model outperformed the naive model measures the benefit of each model over a model than estimates no change between current and future populations. Test results are based on 18 timepoints not observed during model training and validation. Source data are available at `https://github.com/blab/flu-forecasting/blob/published/manuscript/Table_2-source_data_1.csv` and `https://github.com/blab/flu-forecasting/blob/published/manuscript/Table_2-source_data_2.csv`.

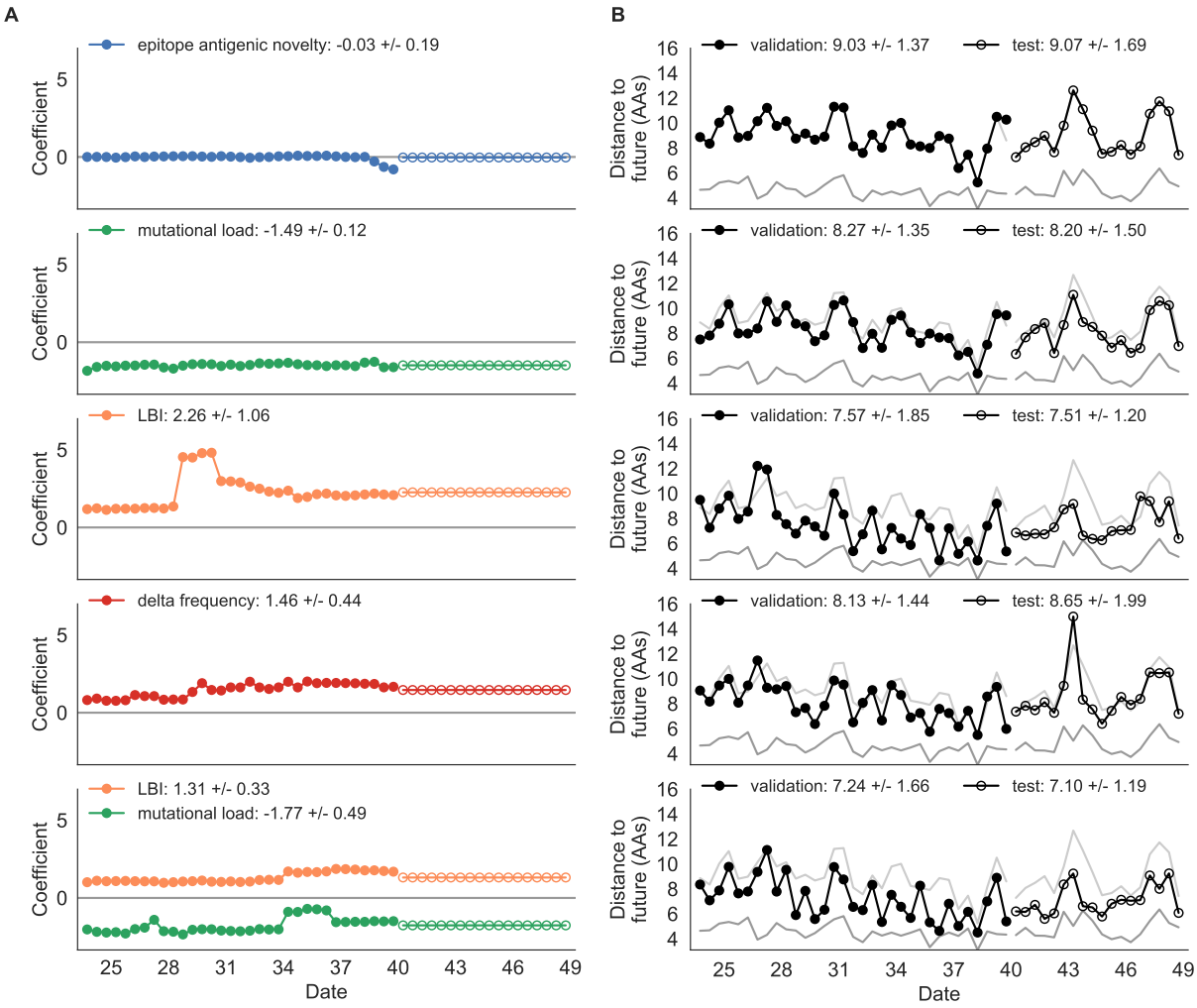

Figure 4.3: **Simulated population model coefficients and distances to the future for individual biologically-informed fitness metrics and the best composite model.** A) Coefficients and B) distances are shown per validation and test timepoint as in Figure 4.2.

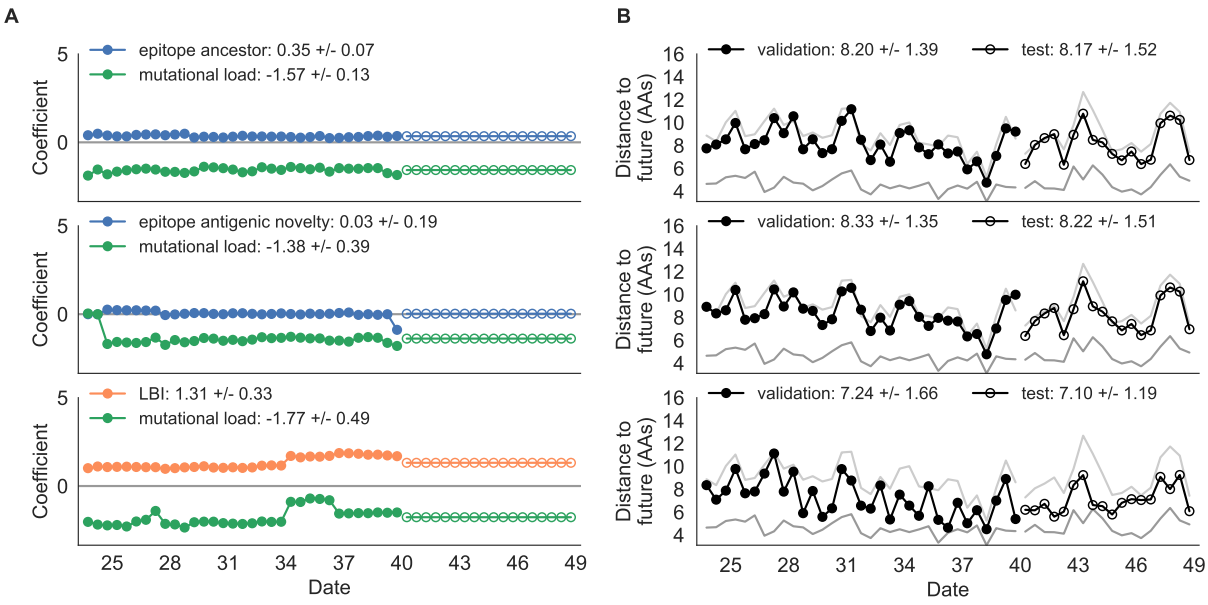

Figure 4.4: **Composite model coefficients and distances to the future for models fit to simulated populations.** A) Coefficients and B) distances are shown per validation timepoint and test timepoint as in Figure 4.2.

We hypothesized that a composite model of mutually beneficial metrics could better approximate the true fitness of simulated viruses than models based on individual metrics. To this end, we fit an additional model including the best metrics from the mechanistic and clade growth categories: mutational load and LBI. This composite model outperformed both of its corresponding individual metric models with an average distance to the future of 7.24 $\pm$ 1.66 AAs and outperformed the naive model as often as the true fitness metric (Figure 4.3, Table 4.2, Table 4.6). The coefficients for mutational load and LBI remained relatively consistent across all validation timepoints, indicating that these fitness metrics were stable approximations of the simulator's underlying evolutionary processes. This small gain supports our hypothesis that multiple complementary metrics can produce more accurate models.

We validated the best performing model (true fitness) using two metrics that are relevant for practical influenza forecasting and vaccine design efforts. First, we measured the ability of the true fitness model to accurately estimate dynamics of large clades (initial frequency $> 15\%$) by comparing observed fold change in clade frequencies, $\log_{10} \frac{x(t+\Delta t)}{x(t)}$ and estimated fold change, $\log_{10} \frac{\hat{x}(t+\Delta t)}{x(t)}$. The model's estimated fold changes correlated well with observed fold changes (Pearson's $R^2 = 0.52$, Figure 4.5A). The model also accurately predicted the growth of 87% of growing clades and the decline of 58% of declining clades. Model forecasts were increasingly more accurate with increasing initial clade frequencies (Figure 4.5C). Next, we counted how often the estimated closest strain to the future population at any given timepoint ranked among the observed top closest strains to the future. We calculated the distance of each present strain to the future as the Hamming distance between the given strain's amino acid sequence and each future strain weighted by the future strain's observed or estimated frequency (Equations 4.3 and 4.4). The estimated closest strain was in the top first percentile of observed closest strains for half of the validation timepoints and in the top 20th percentile for 100% of timepoints (Figure 4.5B). Percentile ranks per strain based on their observed and estimated distances to the future correlated strongly across all strains and

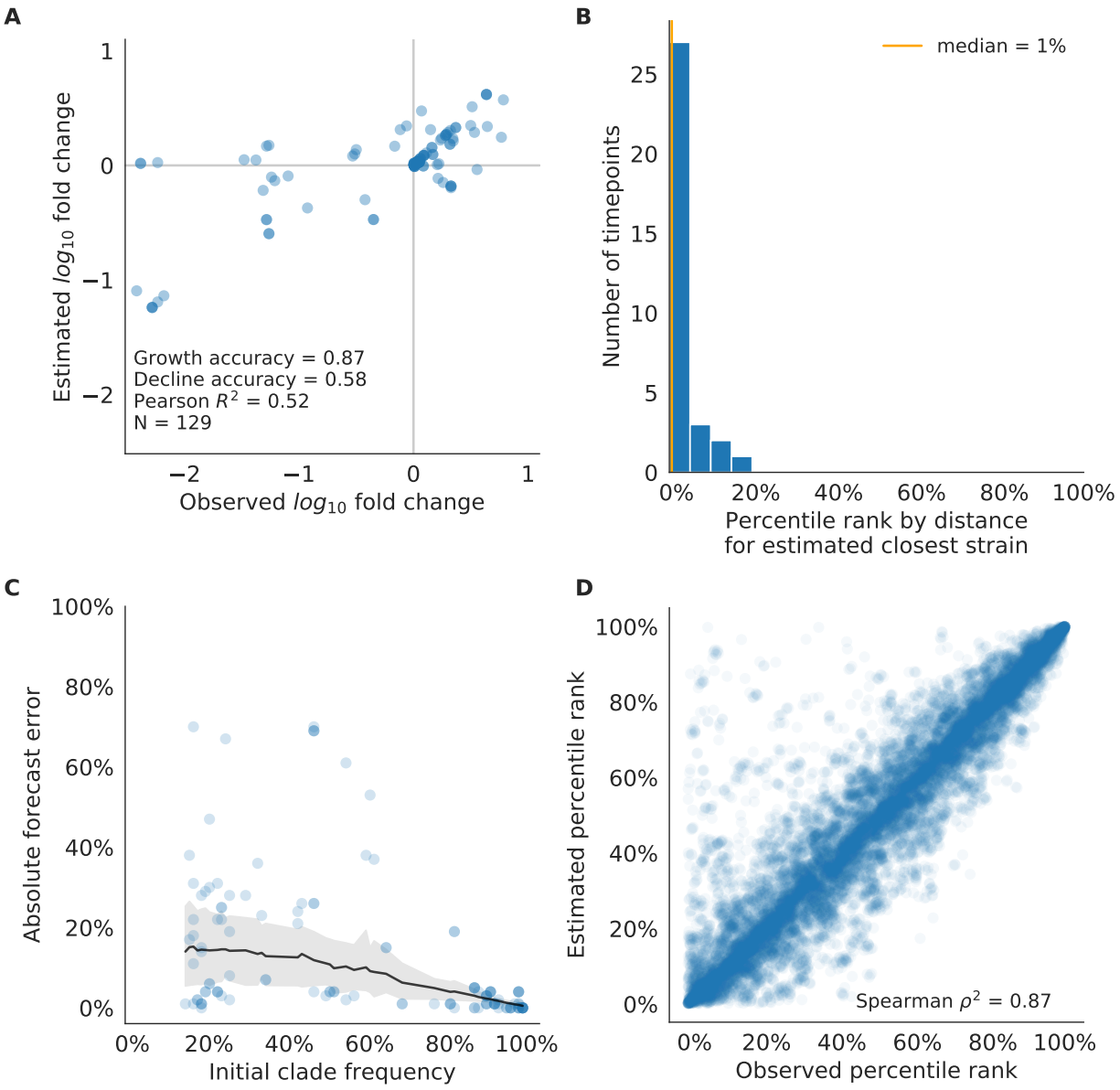

Figure 4.5: **Validation of best model for simulated populations of H3N2-like viruses for 33 timepoints (closed circles in Figure 4.21).** A) The correlation of estimated and observed clade frequency fold changes shows the model's ability to capture clade-level dynamics without explicitly optimizing for clade frequency targets. B) The rank of the estimated closest strain based on its distance to the future for 33 timepoints.

Figure 4.5: (continued) C) Absolute forecast error for clades shown in A by their initial frequency with a mean LOESS fit (solid black line) and 95% confidence intervals (gray shading) based on 100 bootstraps. D) The correlation of all strains at all timepoints by the percentile rank of their observed and estimated distances to the future. The corresponding results for the naive model are shown in Figure 4.6.

timepoints (Spearman's $\rho^2 = 0.87$, Figure 4.5D). In contrast, the naive model's forecasts of clade frequencies were considerably less accurate (Figure 4.6C). 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 $\rho^2 = 0.78$, Figure 4.6B 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.

Finally, we tested all of our models on out-of-sample data. Specifically, we fixed the coefficients of each model to the average values across the validation period and applied the resulting models to the next 9 years of previously unobserved simulated data. A standard expectation from machine learning is that models will perform worse on test data due to overfitting to training data. Despite this expectation, we found that all models except for the individual epitope mutation models consistently outperformed the naive model across the out-of-sample data (Figure 4.2, Figure 4.3, Figure 4.4, Table 4.2). The composite model of mutational load and LBI appeared to outperform the true fitness metric with average distance to the future of 7.10 $\pm$ 1.19 compared to 7.38 $\pm$ 1.89, respectively. However, we did not find a significant difference between these models by bootstrap testing (Table 4.6) and could not rule out fluctuations in model performance across a relatively small number of data points.

As with our validation dataset, we tested the true fitness model's ability to recapitulate clade dynamics and select optimal individual strains from the test data. While observed

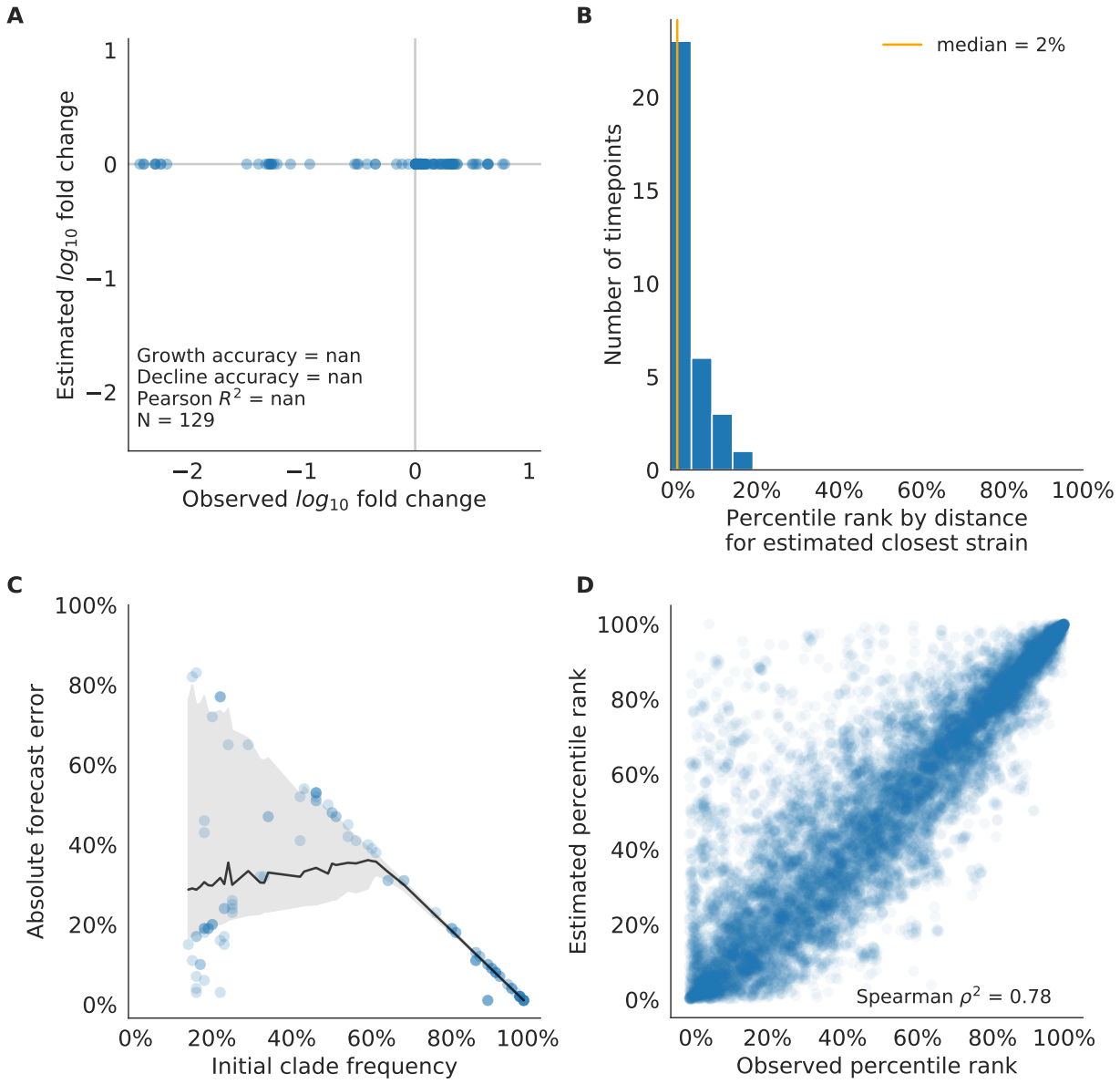

Figure 4.6: **Validation of naive model for simulated populations of H3N2-like viruses for 33 timepoints.** These timepoints correspond to the closed circles in Figure 4.21. Note that the naive model sets future frequencies to current frequencies such that there is no estimated fold change in frequencies for the first panel.

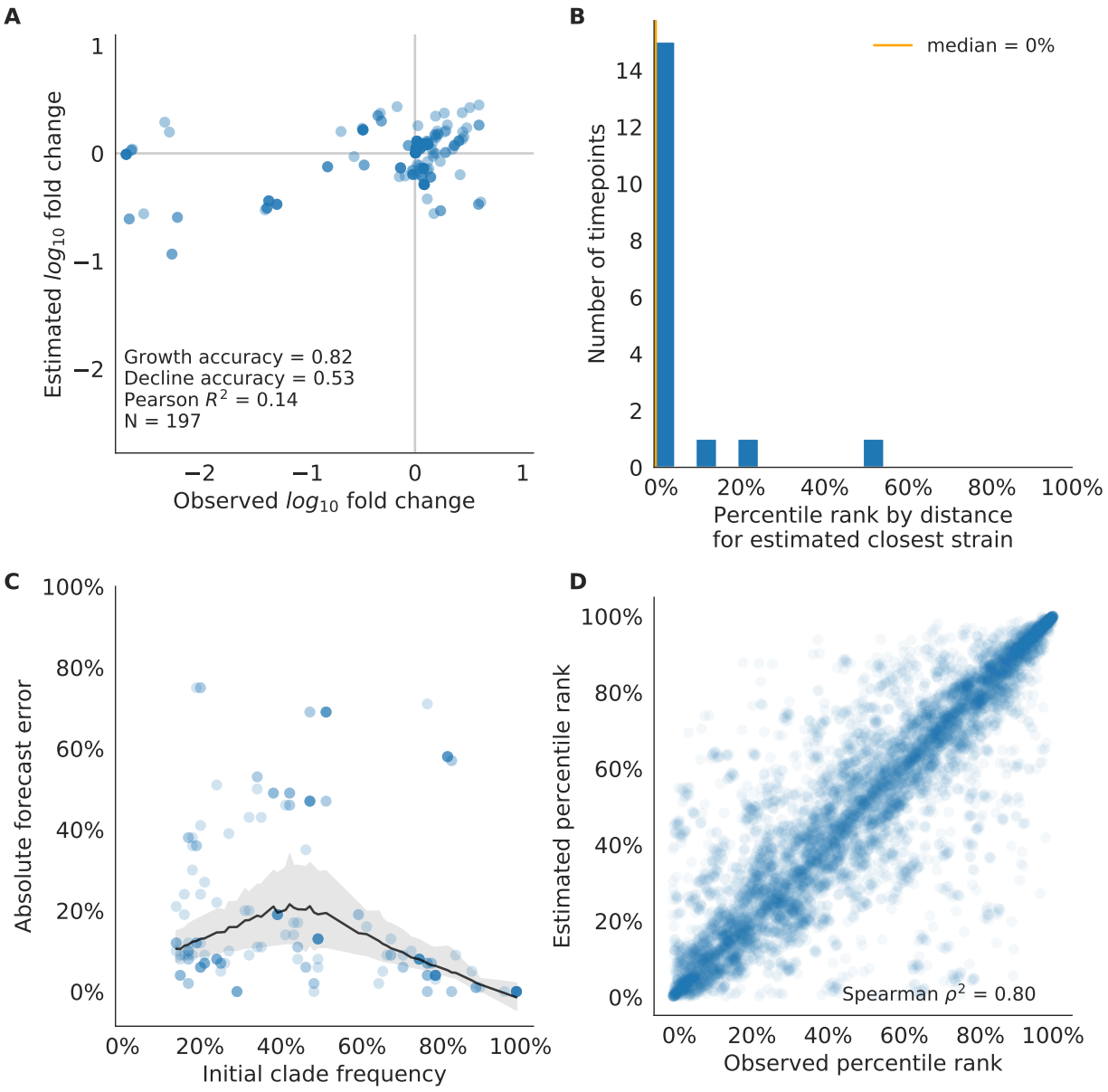

Figure 4.7: **Test of best model for simulated populations (true fitness) using 9 years (18 timepoints) of previously unobserved test data and fixed model coefficients (open circles in Figure 4.21).** A) The correlation of log estimated and observed clade frequency fold change. B) The rank of the estimated closest strain based on its distance to the future in the best model was in the top 20th percentile for 89% of 18 timepoints.

Figure 4.7: (continued) C) Absolute forecast error for clades shown in A by their initial frequency with a mean LOESS fit (solid black line) and 95% confidence intervals (gray shading) based on 100 bootstraps. D) The correlation of all strains at all timepoints by the percentile rank of their observed and estimated distances to the future. The corresponding results for the naive model are shown in Figure 4.8.

and estimated clade frequency fold changes correlated more weakly for test data (Pearson's $R^2 = 0.14$), the accuracies of clade growth and decline predictions remained similar at 82% and 53%, respectively (Figure 4.7A). We observed higher absolute forecast errors in the test data with higher errors for clades between 40% and 60% initial frequencies (Figure 4.7C). The estimated closest strain was higher than the top first percentile of observed closest strains for half of the test timepoints and in the top 20th percentile for 16 (89%) of 18 of timepoints (Figure 4.7B). Observed and estimated strain ranks remained strongly correlated across all strains and timepoints (Spearman's $\rho^2 = 0.80$, Figure 4.7D). 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 $\rho^2 = 0.82$, Figure 4.8). 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.

4.2.3 Models reflect historical patterns of H3N2 evolution

Next, we trained and validated models for individual fitness predictors using 25 years of natural H3N2 populations spanning from October 1, 1990 to October 1, 2015. We held

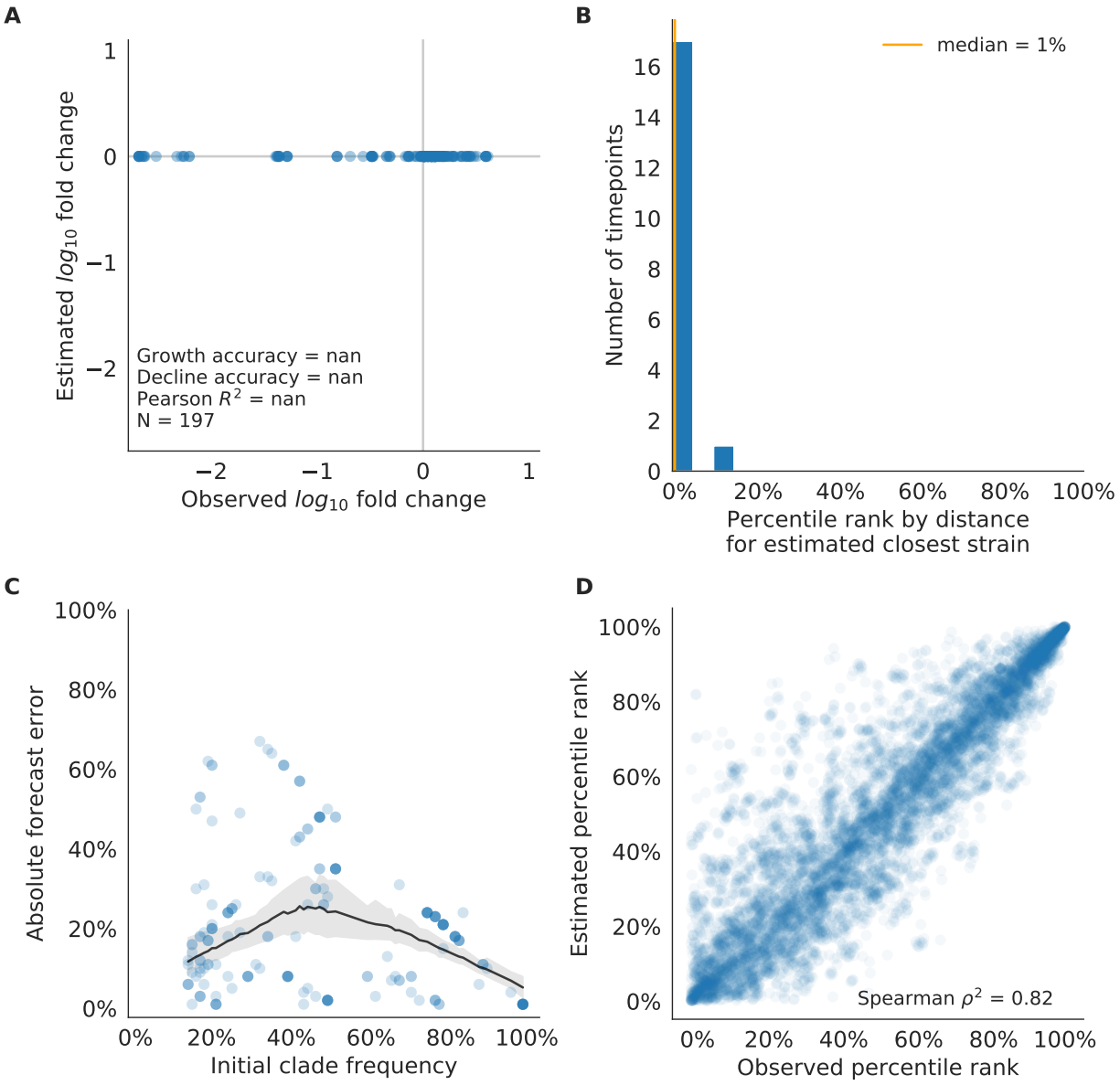

Figure 4.8: **Test of naive model for simulated populations of H3N2-like viruses for 18 timepoints.** These timepoints correspond to the open circles in Figure 4.21. Note that the naive model sets future frequencies to current frequencies such that there is no estimated fold change in frequencies for the first panel.

Model	Coefficients	Distance to future (AAs) Validation	Test	Model > naive Validation	Test
mutational load	-0.68 +/- 0.34	5.44 +/- 1.80*	7.70 +/- 3.53	18 (78%)	4 (50%)
+ LBI	1.03 +/- 0.40				
LBI	1.12 +/- 0.51	5.68 +/- 1.91*	8.40 +/- 3.97	17 (74%)	2 (25%)
HI antigenic novelty	0.89 +/- 0.23	5.82 +/- 1.50*	5.97 +/- 1.47*	17 (74%)	6 (75%)
+ mutational load	-1.01 +/- 0.42				
HI antigenic novelty	0.90 +/- 0.23	5.84 +/- 1.51*	5.99 +/- 1.46*	16 (70%)	6 (75%)
+ mutational load	-1.00 +/- 0.44				
+ LBI	-0.04 +/- 0.09				
HI antigenic novelty	0.83 +/- 0.20	6.01 +/- 1.50*	6.21 +/- 1.44*	16 (70%)	7 (88%)
delta frequency	0.79 +/- 0.47	6.13 +/- 1.71*	6.90 +/- 2.30	16 (70%)	5 (62%)
mutational load	-0.99 +/- 0.30	6.14 +/- 1.37*	6.53 +/- 1.39	17 (74%)	6 (75%)
naive	0.00 +/- 0.00	6.40 +/- 1.36	6.82 +/- 1.74	0 (0%)	0 (0%)
DMS mutational effects	1.25 +/- 0.84	6.75 +/- 1.95	7.80 +/- 2.97	11 (48%)	4 (50%)
epitope antigenic novelty	0.52 +/- 0.73	7.13 +/- 1.47	6.70 +/- 1.51	7 (30%)	5 (62%)

Table 4.3: **Natural population model coefficients and performance on validation and test data ordered from best to worst by distance to the future in the validation analysis, as in Table 4.2.** Distances annotated with asterisks (*) were significantly closer to the future than the naive model as measured by bootstrap tests (see Methods and Figure 4.25). Validation results are based on 23 timepoints. Test results are based on eight timepoints not observed during model training and validation. Source data are available at `https://github.com/blab/flu-forecasting/blob/published/manuscript/Table_7-source_data_1.csv` and `https://github.com/blab/flu-forecasting/blob/published/manuscript/Table_7-source_data_2.csv`.

out strains collected after October 1, 2015 up through October 1, 2019 for model testing (Figure 4.23). In addition to the sequence-only models we tested on simulated populations, we also fit models for our new fitness metrics based on experimental phenotypes including HI antigenic novelty and DMS mutational effects. We hypothesized that both HI and DMS metrics would be assigned positive coefficients, as they estimate increased antigenic drift and beneficial mutations, respectively. As antigenic drift is generally considered to be the primary evolutionary pressure on natural H3N2 populations [Smith et al., 2004, Bedford et al., 2014, Łuksza and Lässig, 2014], we expected that epitope and HI antigenic novelty would be individually more predictive than mutational load or DMS mutational effects. Previous research [Neher et al., 2014] and our simulation results also led us to expect that LBI and delta frequency would outperform other individual mechanistic metrics. As the earliest measurements from focus reduction assays (FRAs) date back to 2012, we could not train, validate, and test FRA antigenic novelty models in parallel with the HI antigenic novelty models.

Biologically-informed metrics generally performed better than the naive model with the exceptions of the epitope antigenic novelty and DMS mutational effects (Figure 4.9 and Table 4.3). The naive model estimated an average distance between natural H3N2 populations of 6.40 $\pm$ 1.36 AAs. The lower bound for how well any model could perform, 2.60 $\pm$ 0.89 AAs, was considerably lower than the corresponding bounds for simulated populations. The average improvement of the sequence-only models over the naive model was consistently lower than the same models in simulated populations. This reduced performance may have been caused by both the relatively reduced diversity between years in natural populations and the fact that our simple models do not capture all drivers of evolution in natural H3N2 populations.

Of the two metrics for antigenic drift, HI antigenic novelty consistently outperformed epitope antigenic novelty (Table 4.3). HI antigenic novelty estimated an average distance to the future of 6.01 $\pm$ 1.50 AAs and outperformed the naive model at 16 of 23 timepoints (70%).

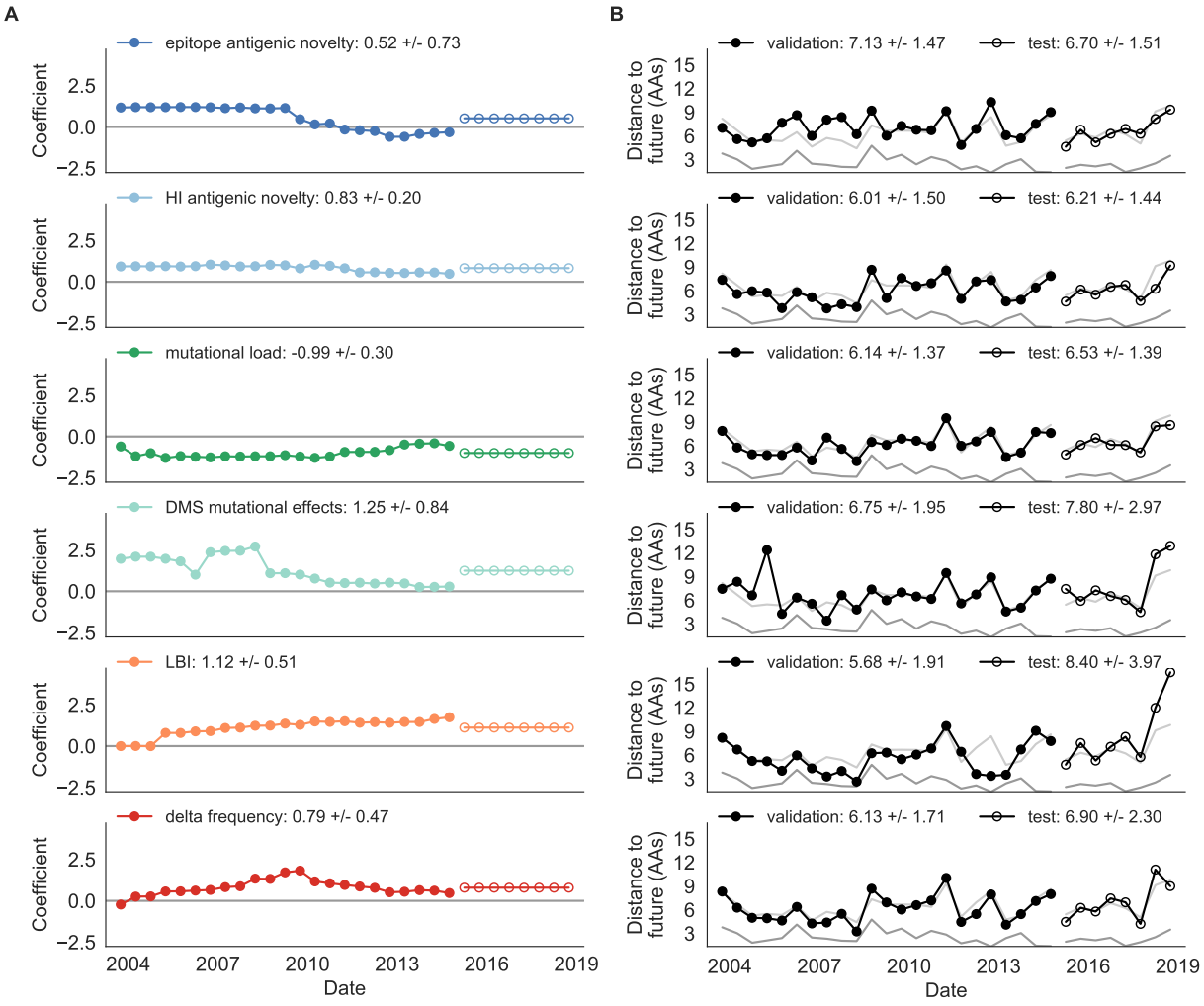

Figure 4.9: **Natural population model coefficients and distances to the future for individual biologically-informed fitness metrics.** A) Coefficients and B) distances are shown per validation timepoint (N=23) and test timepoint (N=8) as in Figure 4.2. The naive model's distance to the future (light gray) was 6.40 ± 1.36 AAs for validation timepoints and 6.82 ± 1.74 AAs for test timepoints. The corresponding lower bounds on the estimated distance to the future (dark gray) were 2.60 ± 0.89 AAs and 2.28 ± 0.61 AAs.

The coefficient for HI antigenic novelty remained stable across all timepoints (Figure 4.9). In contrast, epitope antigenic novelty estimated a distance of 7.13 ± 1.47 AAs and only outperformed the naive model at seven timepoints (30%). Epitope antigenic novelty was also the only metric whose coefficient started at a positive value (1.17 ± 0.03 on average prior to October 2009) and transitioned to a negative value through the validation period (-0.19 ± 0.34 on average for October 2009 and after). This strong coefficient for the first half of training windows indicated that, unlike the results for simulated populations, the nonlinear antigenic novelty metric was historically an effective measure of antigenic drift. The historical importance of the epitope sites used for this metric was further supported by the relative enrichment of mutations at these sites for the most successful "trunk" lineages of natural populations compared to side branch lineages (Table 4.5).

These results led us to hypothesize that the contribution of these specific epitope sites to antigenic drift has weakened over time. Importantly, these 49 epitope sites were originally selected by Łuksza and Lässig [2014] from a previous historical survey of sites with beneficial mutations between 1968–2005 [Shih et al., 2007]. If the beneficial effects of mutations at these sites were due to historical contingency rather than a constant contribution to antigenic drift, we would expect models based on these sites to perform well until 2005 and then overfit relative to future data. Indeed, the epitope antigenic novelty model outperforms the naive model for the first three validation timepoints until it has to predict to April 2006. To test this hypothesis, we identified a new set of beneficial sites across our entire validation period of October 1990 through October 2015. Inspired by the original approach of Shih et al. [2007], we identified 25 sites in HA1 where mutations rapidly swept through the global population, including 12 that were also present in the original set of 49 sites. We fit an antigenic novelty model to these 25 sites across the complete validation period and dubbed this the "oracle antigenic novelty" model, as it benefited from knowledge of the future in its forecasts. The oracle model produced a consistently positive coefficient across all training windows (0.80 ± 0.21) and consistently outperformed the original epitope model with an average distance to

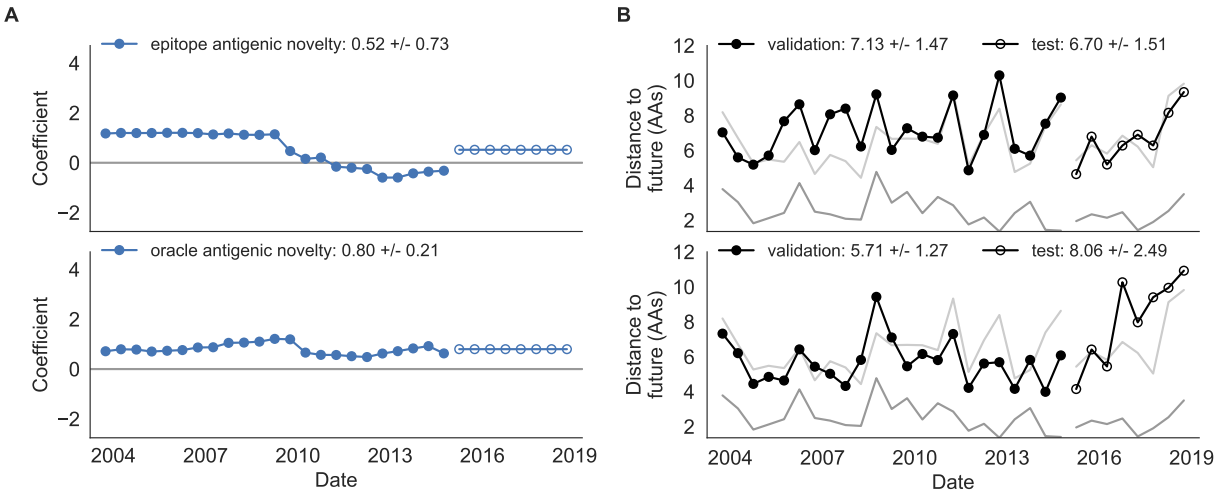

Figure 4.10: **Comparison of epitope-based models with knowledge of the future.** Model coefficients and distances to the future for antigenic novelty models fit to natural populations. A) Coefficients and B) distances are shown per validation timepoint and test timepoint as in Figure 4.2. The epitope antigenic novelty model relies on previously published epitope sites [Łuksza and Lässig, 2014]. The "oracle" antigenic novelty model relies on sites of beneficial mutations that were manually identified from the entire training and validation time period (Methods). The improved performance of the "oracle" model indicates that the sequence-based antigenic novelty metric can be effective when sites of beneficial mutations are known prior to forecasting.

the future of 5.71 ± 1.27 AAs (Figure 4.10). These results support our hypothesis that the fitness benefit of mutations at the original 49 sites was due to historical contingency and that the success of previous epitope models based on these sites was partly due to "borrowing from the future". We suspect that our HI antigenic novelty model benefits from its ability to constantly update its antigenic model at each timepoint with recent experimental phenotypes, while the epitope antigenic novelty metric is forced to give a constant weight to the same 49 sites throughout time.

Of the two metrics for functional constraint, mutational load outperformed DMS mutational effects, with an average distance to the future of 6.14 ± 1.37 AAs compared to 6.75 ± 1.95 AAs, respectively. In contrast to the original Łuksza and Lässig [2014] model, where the coefficient of the mutational load metric was fixed at -0.5, our model learned a consistently stronger coefficient of -0.99 ± 0.30. Notably, the best performance of the DMS mutational effects model was forecasting from April 2007 to April 2008 when the major clade containing A/Perth/16/2009 was first emerging. This result is consistent with the DMS model overfitting to the evolutionary history of the background strain used to perform the DMS experiments. Alternate implementations of less background-dependent DMS metrics never performed better than the mutational load metric (Table 4.7, Methods). Thus, we find that a simple model where any mutation at non-epitope sites is deleterious is more predictive of global viral success than a more comprehensive biophysical model based on measured mutational effects of a single strain.

LBI was the best individual metric by average distance to the future (Figure 4.9) and tied mutational load by outperforming the naive model at 17 (74%) timepoints (Table 4.3). Delta frequency performed worse than LBI and HI antigenic novelty and was comparable to mutational load. While delta frequency should, in principle, measure the same aspect of viral fitness as LBI, these results show that the current implementations of these metrics represent qualitatively different fitness components. The LBI and mutational load might also be predictive for reasons other than correlation with fitness, see Discussion.

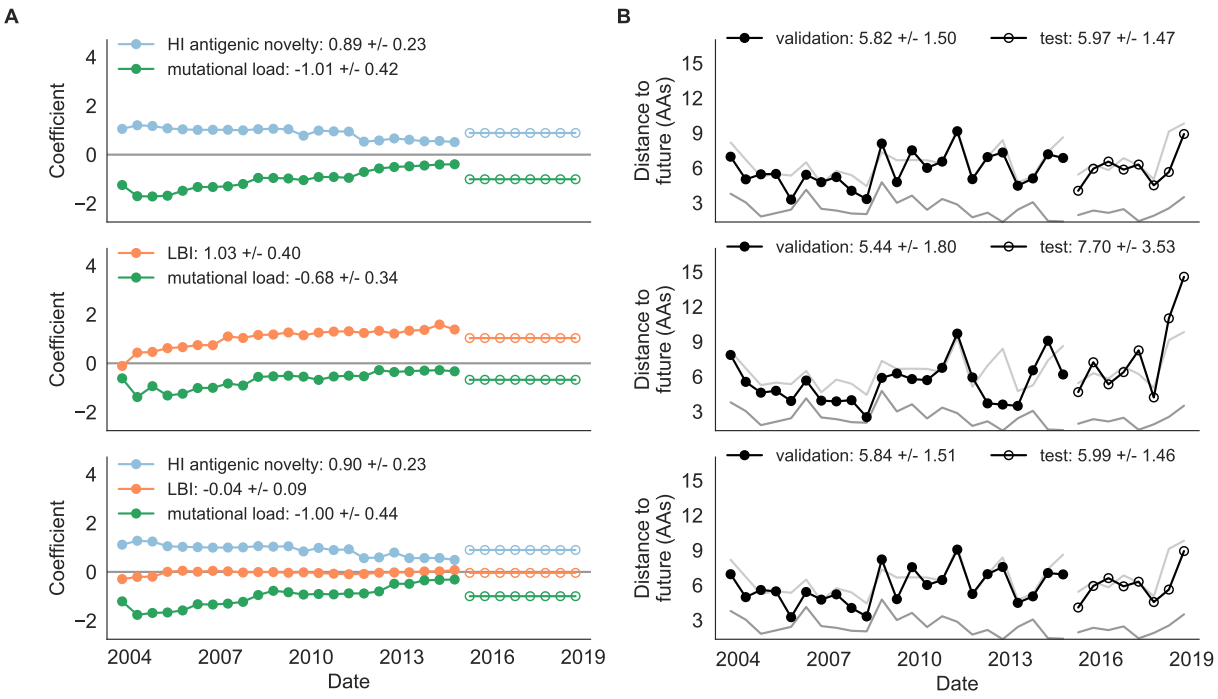

Figure 4.11: **Natural population model coefficients and distances to the future for composite fitness metrics.** A) Coefficients and B) distances are shown per validation timepoint (N=23) and test timepoint (N=8) as in Figure 4.2.

To test whether composite models could outperform individual fitness metrics for natural populations, we fit models based on combinations of best individual metrics representing antigenic drift, functional constraint, and clade growth. Specifically, we fit models based on HI antigenic novelty and mutational load, mutational load and LBI, and all three of these metrics together. We anticipated that if these metrics all represented distinct, mutually beneficial components of viral fitness, these composite models should perform better than individual models with consistent coefficients for each metric.

Both two-metric composite models modestly outperformed their corresponding individual models (Table 4.3, Figure 4.11, and Table 4.6). The composite of mutational load and LBI performed the best overall with an average distance to the future of 5.44 ± 1.80 AAs. The relative stability of the coefficients for the metrics in the two-metric models suggested that these metrics represented complementary components of viral fitness. In contrast, the three-metric model strongly preferred the HI antigenic novelty and mutational load metrics over LBI for the entire validation period, producing an average LBI coefficient of -0.04 ± 0.09. Overall, the gain by combining multiple predictors was limited and the sensitivity of coefficients to the set of metrics included in the model suggests that there is substantial overlap in predictive value of different metrics.

As with the simulated populations, we validated the performance of the best model for natural populations using estimated and observed clade frequency fold changes and the ranking of estimated closest strains compared to the observed closest strains to future populations. The composite model of mutational load and LBI effectively captured clade dynamics with a fold change correlation of $R^2 = 0.35$ and growth and decline accuracies of 87% and 89%, respectively (Figure 4.12A). Absolute forecasting error declined noticeably for clades with initial frequencies above 60%, but generally this error remained below 20% on average (Figure 4.12C). The estimated closest strain from this model was in the top first percentile of observed closest strains for half of the validation timepoints and in the top 20th percentile for 20 (87%) of 23 timepoints (Figure 4.12B). This pattern held across all

strains and timepoints with a strong correlation between observed and estimated strain ranks (Spearman's $\rho^2 = 0.66$, Figure 4.12D). 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 4.13B and C).

Finally, we tested the performance of all models on out-of-sample data collected from October 1, 2015 through October 1, 2019. We anticipated that most models would perform worse on truly out-of-sample data than on validation data. Correspondingly, only the three models with the HI antigenic novelty metric significantly outperformed the naive model on the test data (Table 4.3). The composite of HI antigenic novelty and mutational load performed modestly, although not significantly, better than the individual HI antigenic novelty model (Table 4.6). Surprisingly, the best model for the validation data – mutational load and LBI – was one of the worst models for the test data with an average distance to the future of 7.70 ± 3.53 AAs. The individual LBI model was the worst model, while mutational load continued to perform well with test data. LBI performed especially poorly in the last two test timepoints of April and October 2018 (Figure 4.9). These timepoints correspond to the dominance and sudden decline of a reassortant clade named A2/re [Potter et al., 2019]. By April 2018, the A2/re clade had risen to a global frequency over 50% from less than 15% the previous year, despite an absence of antigenic drift. By October 2018, this clade had declined in frequency to approximately 30% and, by October 2019, it had gone extinct. That LBI incorrectly predicted the success of this reassortant clade highlights a major limitation of growth-based fitness metrics and a corresponding benefit of more mechanistic metrics that explicitly measure antigenic drift and functional constraint. However, we cannot rule out the alternate possibility that the LBI model was overfit to the training data.

After identifying the composite HI antigenic novelty and mutational load model as the best model on out-of-sample data, we tested this model's ability to detect clade dynamics and select individual closest strains to the future for vaccine composition. The composite model

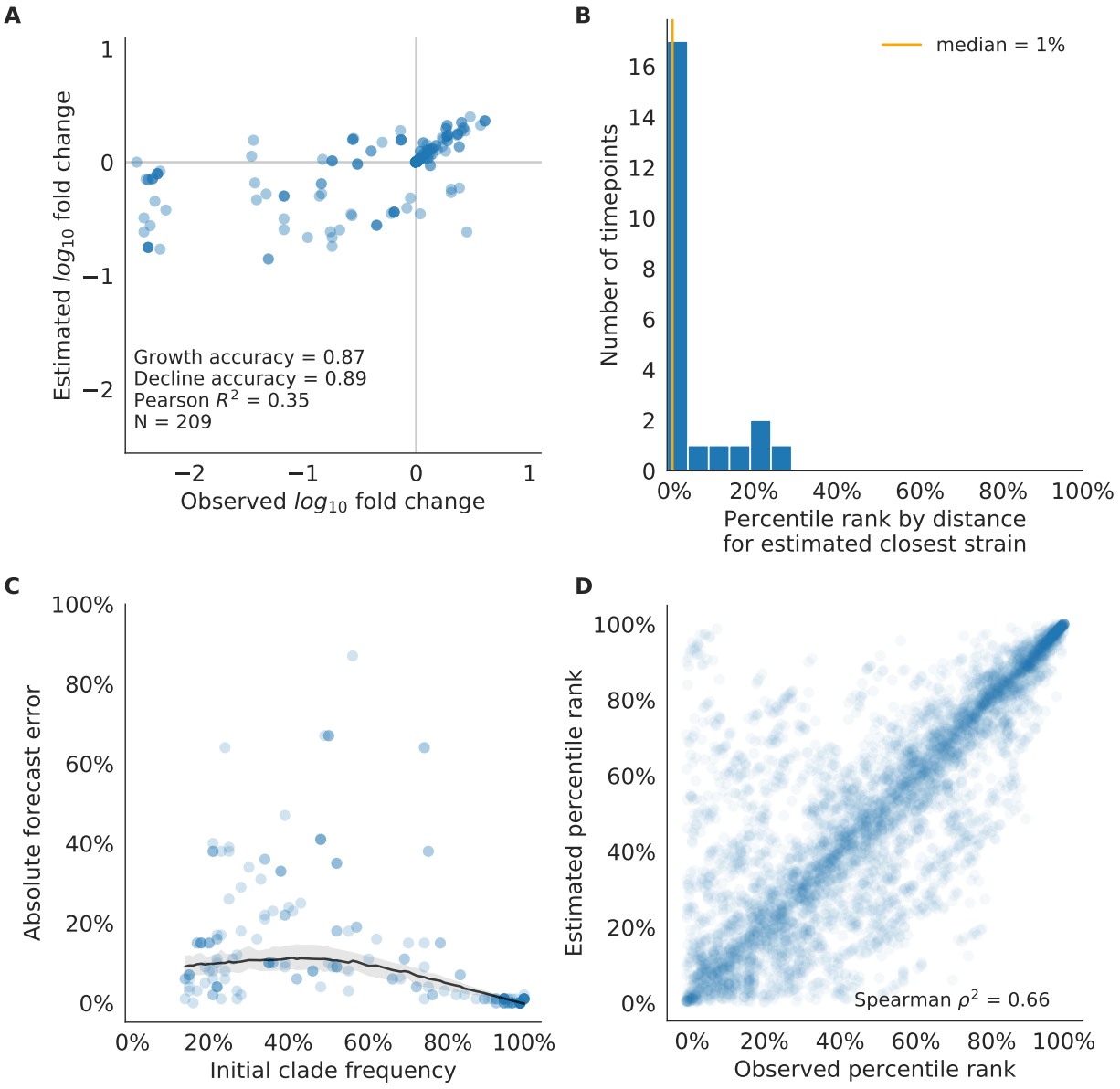

Figure 4.12: **Validation of best model for natural populations of H3N2 viruses for 23 timepoints (closed circles in Figure 4.23) using the composite model of mutational load and LBI.** A) The correlation of estimated and observed clade frequency fold changes shows the model's ability to capture clade-level dynamics without explicitly optimizing for clade frequency targets. B) The rank of the estimated closest strain based on its distance to the future for 23 timepoints.

Figure 4.12: (continued) C) Absolute forecast error for clades shown in A by their initial frequency with a mean LOESS fit (solid black line) and 95% confidence intervals (gray shading) based on 100 bootstraps. D) The correlation of all strains at all timepoints by the percentile rank of their observed and estimated distances to the future. The corresponding results for the naive model are shown in Figure 4.13.

partially captured clade dynamics with a Pearson's correlation of $R^2 = 0.46$ between observed and estimated growth ratios and growth and decline accuracies of 52% and 58%, respectively (Figure 4.14A). The mean absolute forecasting error with this model was consistently less than 20%, regardless of the initial clade frequency (Figure 4.14C). The estimated closest strain from this model was in the top first percentile of observed closest strains for half of the validation timepoints and in the top 20th percentile for 100% of timepoints (Figure 4.14B). Similarly, the observed and estimated strain ranks strongly correlated (Spearman's $\rho^2 = 0.72$) across all strains and test timepoints (Figure 4.14D). The estimated strain ranks of the naive model were not as well correlated (Spearman's $\rho^2 = 0.56$), 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 4.15B and D).

We further evaluated our models' ability to estimate the closest strain to the next season's H3N2 population by comparing our best models' selections to the WHO's vaccine strain selection. 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 4.3). 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 4.4. The observed closest strain to the future represents the centroid of the observed future population,

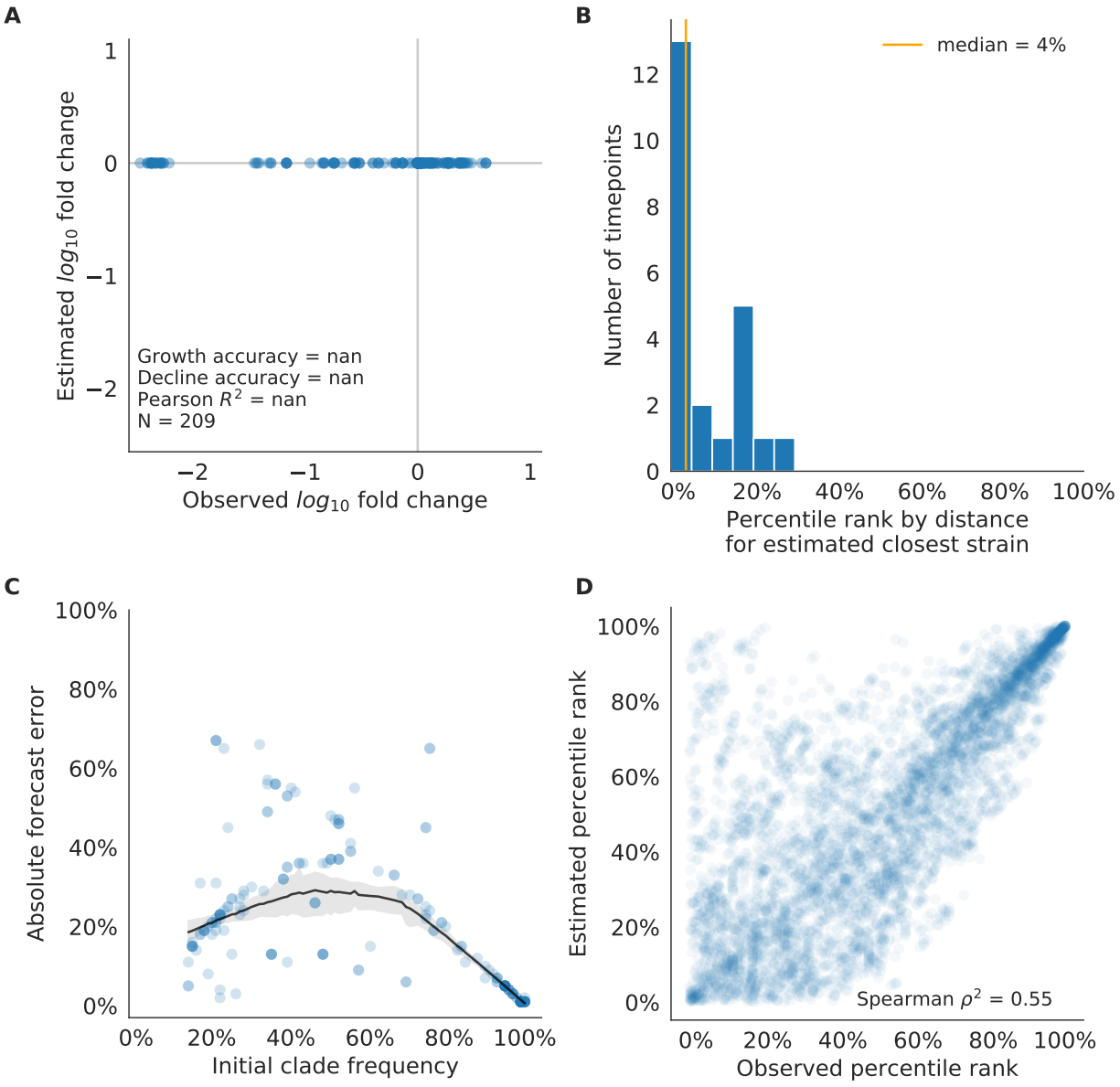

Figure 4.13: **Validation of naive model for natural populations of H3N2 viruses for 23 timepoints.** These timepoints correspond to the closed circles in Figure 4.23. Note that the naive model sets future frequencies to current frequencies such that there is no estimated fold change in frequencies for the first panel.

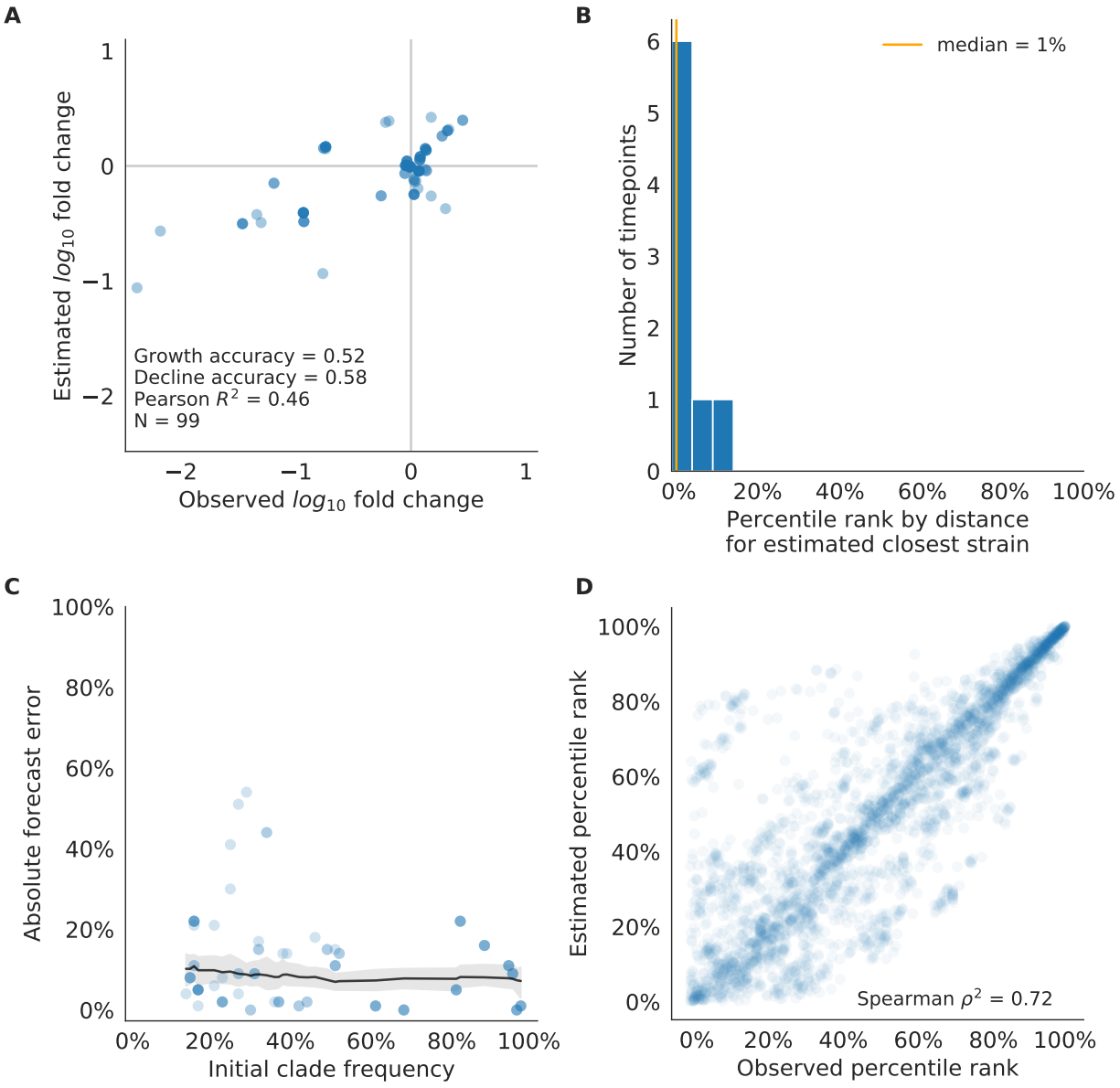

Figure 4.14: **Test of best model for natural populations of H3N2 viruses (HI antigenic novelty and mutational load) across eight timepoints (open circles in Figure 4.23).** A) The correlation of estimated and observed clade frequency fold changes shows the model's ability to capture clade-level dynamics without explicitly optimizing for clade frequency targets. B) The rank of the estimated closest strain based on its distance to the future for eight timepoints.

Figure 4.14: (continued) C) Absolute forecast error for clades shown in A by their initial frequency with a mean LOESS fit (solid black line) and 95% confidence intervals (gray shading) based on 100 bootstraps. D) The correlation of all strains at all timepoints by the percentile rank of their observed and estimated distances to the future. The corresponding results for the naive model are shown in Figure 4.15.

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 4.16). On average, the strains selected by this model were closer to future than the vaccine strain by 1.93 AAs (Figure 4.17). 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.

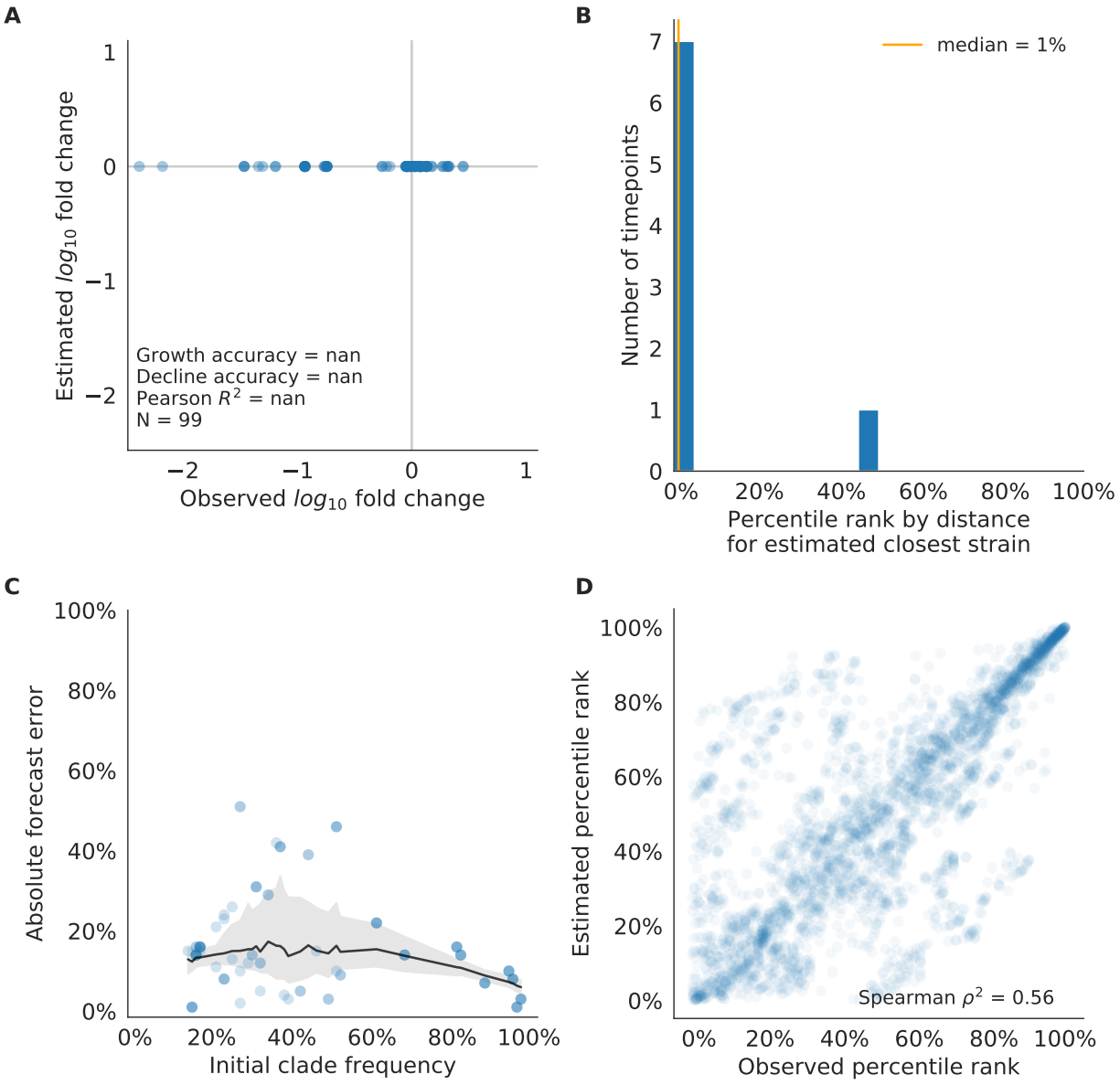

Figure 4.15: **Test of naive model for natural populations of H3N2 viruses for eight timepoints.** These timepoints correspond to the open circles in Figure 4.23. Note that the naive model sets future frequencies to current frequencies such that there is no estimated fold change in frequencies for the first panel.

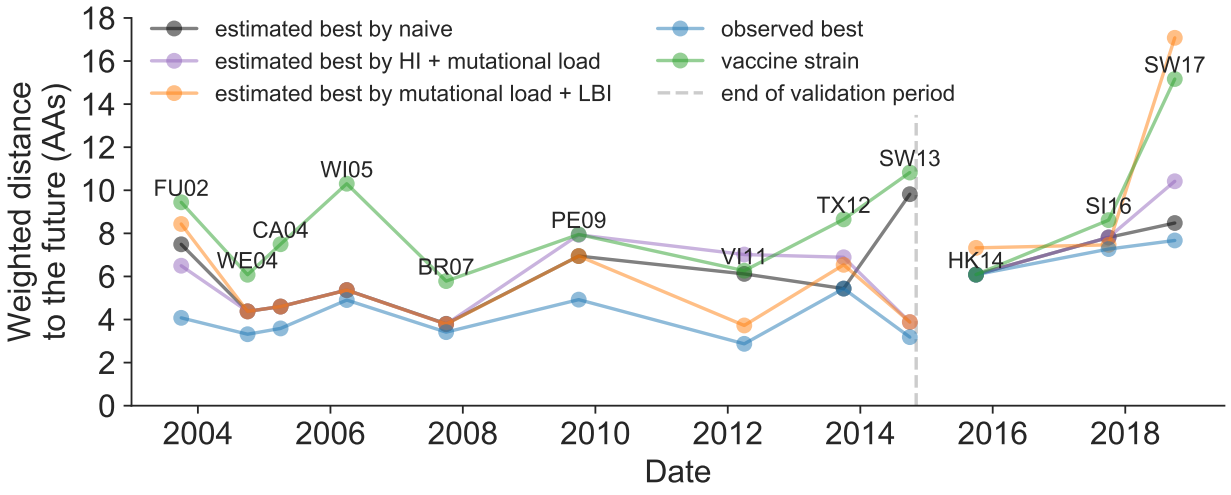

Figure 4.16: **Observed distance to natural H3N2 populations one year into the future for each vaccine strain (green) and the observed (blue) and estimated closest strains to the future by the mutational load and LBI model (orange), the HI antigenic novelty and mutational load model (purple), and the 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. Source data are available at `https://github.com/blab/flu-forecasting/blob/published/manuscript/Figure_8-source_data_1.csv`.

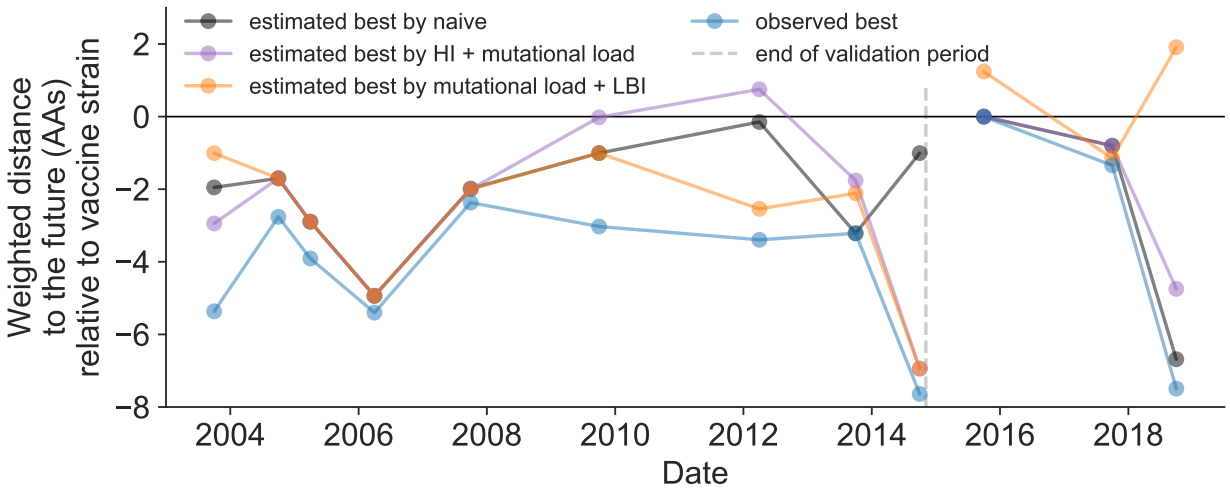

Figure 4.17: **Relative distance to future H3N2 populations between vaccine strains and corresponding observed and estimated closest strains at each timepoint as in Figure 4.16.** Strains with relative distances greater than zero were farther from the future than the selected vaccine strain, while strains below zero were closer to the future. Source data are available at `https://github.com/blab/flu-forecasting/blob/published/manuscript/Figure_8-source_data_1.csv`.

4.2.4 Historically-trained models enable real-time, actionable forecasts

To enable real-time forecasts, we integrated our forecasting framework into our existing open source pathogen surveillance application, Nextstrain [Hadfield et al., 2018]. Prior to finalizing our model coefficients for use in Nextstrain, we tested whether our three best composite models could be improved by learning new coefficients per timepoint from the test data. Additionally, we evaluated a composite of FRA antigenic novelty and mutational load. Since the earliest FRA data were from 2012, we anticipated that there were enough measurements to fit a model across the test data time interval. If modern H3N2 strains continue to perform poorly in HI assays, the FRA-based assay will be critical for future forecasting efforts.

Two of three models performed worse after refitting coefficients to the test data than their original fixed coefficient implementations (Figure 4.18). While, the mutational load and LBI model improved considerably over its original performance, it still performed worse than the naive model on average. These results confirmed that the coefficients for our selected best model would be most accurate for live forecasts. Interestingly, the FRA antigenic novelty metric received a consistently positive coefficient of 1.40 ± 0.24 in its composite with mutational load. Unfortunately, this model performed considerably worse than the corresponding HI-based model. These results suggest that we may need more FRA data across a longer historical timespan to train a model that could replace the HI-based model.

After confirming the coefficients for our best model of HI antigenic novelty and mutational load, we inspected forecasts of H3N2 clades using all data available up through June 6, 2020. Consistent with an average two-month lag between data collection and submission, the most recent data were collected up to April 1, 2020 and made our forecasts from this timepoint to April 1, 2021. Of the five major currently circulating clades, our model predicted growth of the clades 3c3.A and A1b/94N and decline of clades A1b/135K, A1b/137F, and A1b/197R (Figure 4.19). To aid with identification of potential vaccine candidates for the next season, we annotated strains in the phylogeny by their estimated distance to the future based on our best model (Figure 4.20).

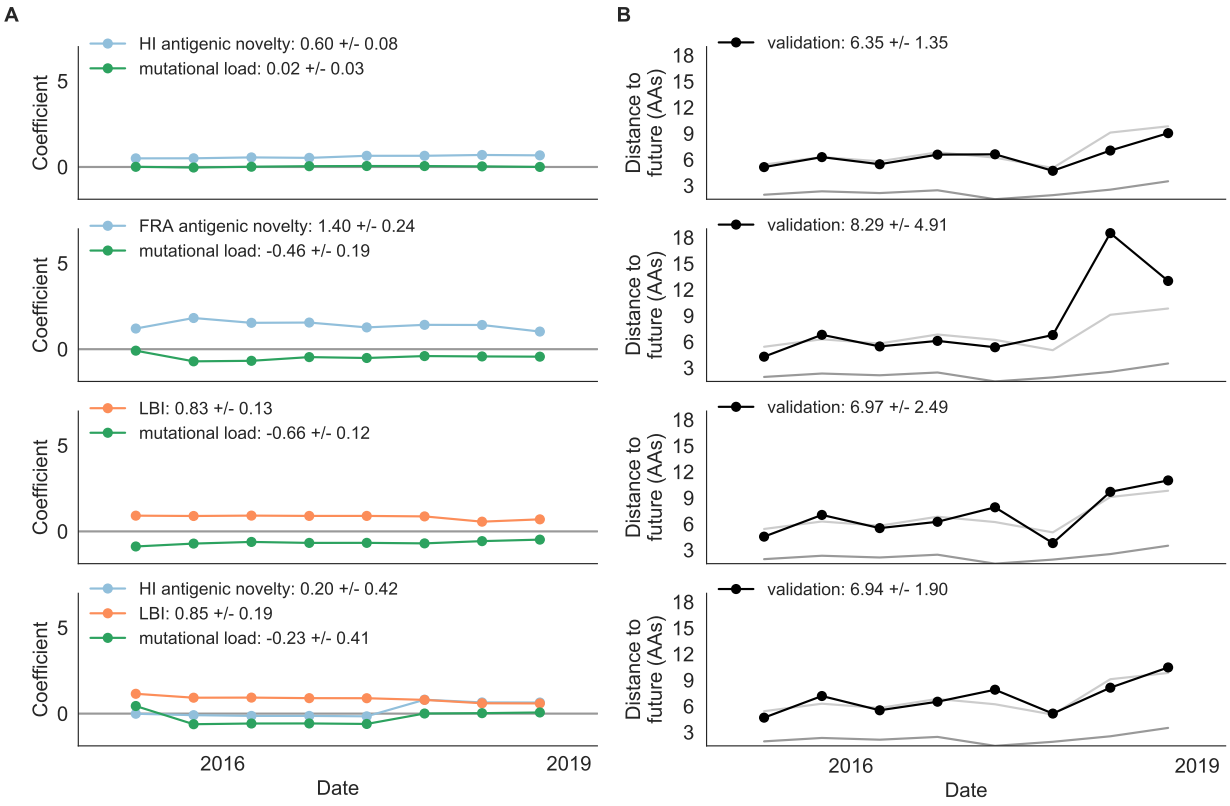

Figure 4.18: **Composite models fit to most recent data from natural populations.** Model coefficients and distances to the future for best composite models and a FRA-based composite fit to recent data from natural populations as in Figure 4.2. A) Coefficients and B) distances are shown per test timepoint (N=8). In contrast to the results for these models based on fixed coefficients from training/validation, these coefficients were learned for each six-year window prior to the corresponding test timepoint. The corresponding distances reflect the model's performance with updated coefficients on what is effectively new validation data. The naive model's distance to the future was 6.82 $\pm$ 1.74 AAs for these timepoints.

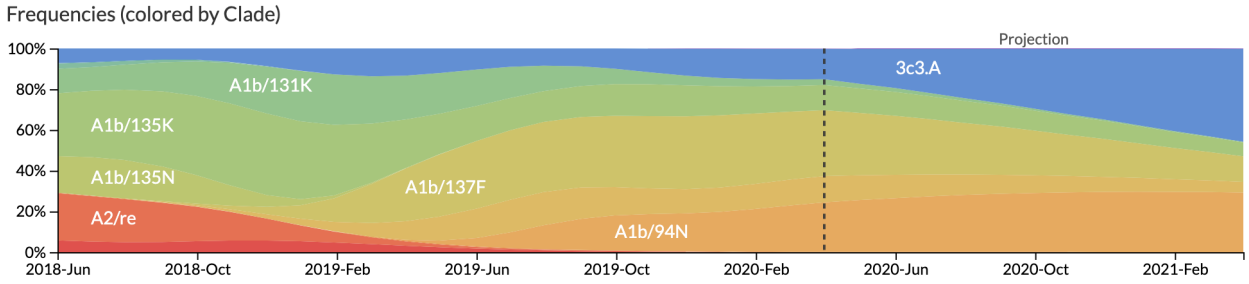

Figure 4.19: **Snapshot of live forecasts on nextstrain.org from our best model (HI antigenic novelty and mutational load) for April 1, 2021.** The observed frequency trajectories for currently circulating clades are shown up to April 1, 2020. Our model forecasts growth of the clades 3c3.A and A1b/94N and decline of all other major clades.

4.3 Discussion

We have developed and rigorously tested a novel, open source framework for forecasting the long-term evolution of seasonal influenza H3N2 by estimating the sequence composition of future populations. A key innovation of this framework is its ability to directly compare viral populations between seasons using the earth mover's distance metric [Rubner et al., 1998] and eliminate unavoidably stochastic clade definitions from phylogenies. The best models from this framework still effectively capture clade dynamics and accurately identify optimal vaccine candidates from simulated and natural H3N2 populations without relying on clades as model targets. We have further introduced novel fitness metrics based on experimental measurements of antigenic drift and functional constraint. 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. We have added this framework as a component of seasonal influenza analyses on nextstrain.org where it provides real-time forecasts for influenza researchers, decision makers, and the public.

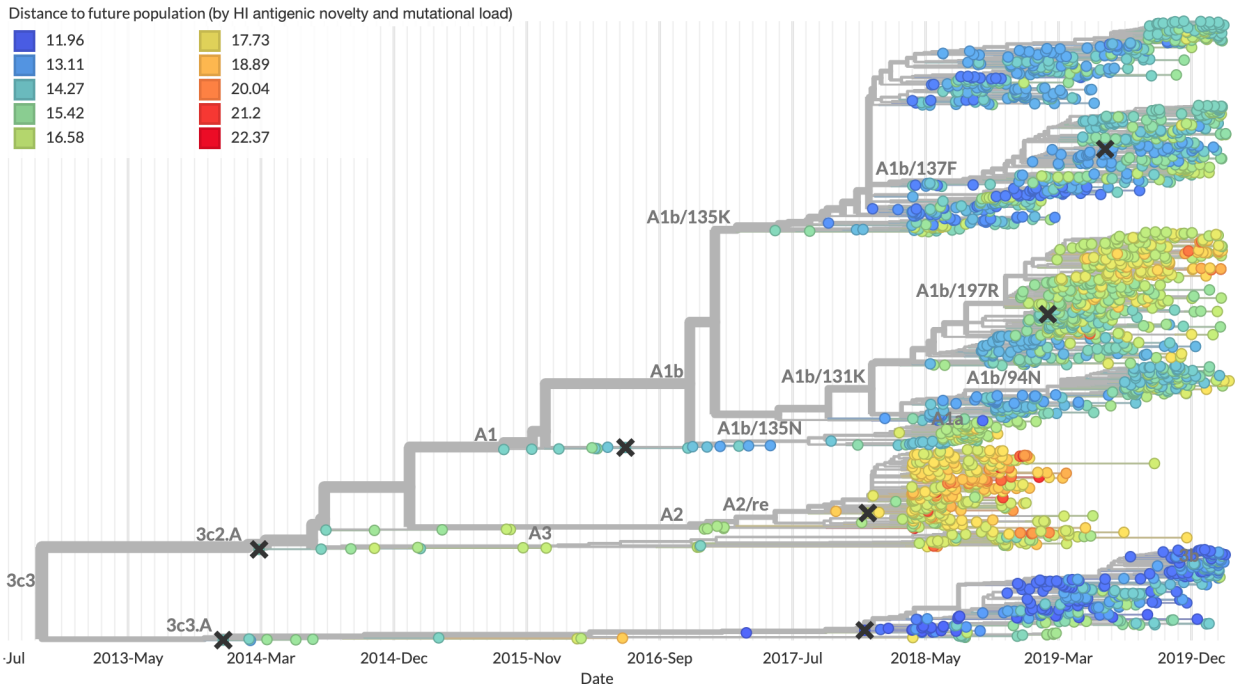

Figure 4.20: **Snapshot of the last two years of seasonal influenza H3N2 evolution on nextstrain.org showing the estimated distance per strain to the future population.** Distance to the future is calculated for each strain as the Hamming distance of HA amino acid sequences to all other circulating strains weighted by the other strain's projected frequencies under the best fitness model (HI antigenic novelty and mutational load).

4.3.1 Integration of genotypic and phenotypic metrics minimizes overfitting

Our evaluation of models by time-series cross-validation and true out-of-sample forecasts revealed substantial potential for model overfitting. We observed overfitting to both specific genetic backgrounds and general historical contexts. A clear example of the former was the poor performance of our DMS-based fitness metric compared to a simpler mutational load metric. Although the DMS experiments provided detailed estimates of which amino acids were preferred at which positions in HA, these measurements were specific to a single strain, A/Perth/16/2009 [Lee et al., 2018]. When we applied these measurements to predict the success of global populations, they were less informative on average than the naive model. To benefit from the more comprehensive fitness costs measured by DMS data, future models will need to synthesize DMS measurements across multiple H3N2 strains from distinct genetic contexts. We anticipate that these measurements could be used to define and continually update a modern set of sites contributing to mutational load in natural populations. This set of sites could replace the statically defined set of "non-epitope" sites we use to estimate mutational load here.

We observed overfitting to historical context in sequence-based models of antigenic drift. The fitness benefit of mutations that led to antigenic drift in H3N2 in the past is well-documented [Wiley et al., 1981, Smith et al., 2004, Wolf et al., 2006, Koel et al., 2013]. Although the antigenic importance of seven specific sites in HA were experimentally validated by Koel et al. [2013], these sites do not explain all antigenic drift observed in natural populations [Neher et al., 2016]. Other attempts to define these so-called "epitope sites" have relied on either aggregation of results from antigenic escape assays [Wolf et al., 2006] or retrospective computational analyses of sites with beneficial mutations [Shih et al., 2007, Łuksza and Lässig, 2014]. We found that models based on all of these definitions except for the seven Koel epitope sites overfit to the historical context from which they were identified (Table 4.7). These results suggest that the set of sites that contribute to antigenic drift at any given time may depend on both the fitness landscape of currently circulating strains and the immune

landscape of the hosts these strains need to infect. Recent experimental mapping of antigenic escape mutations in H3N2 HA with human sera show that the specific sites that confer antigenic escape can vary dramatically between individuals based on their exposure history [Lee et al., 2019]. In contrast to models based on predefined "epitope sites", our model based on experimental measurements of antigenic drift did not suffer from overfitting in the validation or test periods. We suspect that this model was able to minimize overfitting by continuously updating its antigenic model with recent experimental data and assigning antigenic weight to branches of a phylogeny rather than specific positions in HA.

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.

Correspondingly, we observed that composite models of multiple orthogonal fitness metrics often outperformed models based on their individual components. These results are consistent with previous work that found improved performance by integrating components of antigenic drift, functional constraint, and clade growth [Łuksza and Lässig, 2014]. However, the effective elimination of LBI from our three-metric model during the validation period (Figure 4.11)

reveals the limitations of our current additive approach to composite models. 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.

4.3.2 Forecasting framework aids practical forecasts

By forecasting the composition of future H3N2 populations with biologically-informed fitness metrics, our best models consistently outperformed a naive model (Table 4.3). While this performance confirms previously demonstrated potential for long-term influenza forecasting [Łuksza and Lässig, 2014], the average gain from these models over the naive model appears low at 0.96 AAs per year for validation data and 0.85 AAs per year for test data. However, these results are consistent with the observed dynamics of H3N2. First, the one-year forecast horizon is a fraction of the average coalescence time for H3N2 populations of about 3–8 years [Rambaut et al., 2008]. Hence, we expect the diversity of circulating strains to persist between seasons. Second, H3N2 hemagglutinin accumulates 3.6 amino acid changes per year [Smith et al., 2004]. This accumulation of amino acid substitutions contributes to the distance between annual populations observed by the naive model. In this context, our model gains of 0.96 and 0.85 AAs per year correspond to an explanation of 27% and 24% of the expected additional distance between annual populations, respectively.

Several clear opportunities to improve forecasts still remain. Integration of more recent experimental data may improve estimates of antigenic drift. Despite the weak performance of our FRA antigenic novelty model on recent data, continued accumulation of FRA measurements over time should eventually enable models as accurate as the current HI-based models. In addition to these FRA data based on ferret antisera, recent high-throughput antigenic escape assays with human sera promise to improve existing definitions of epitope sites [Lee et al., 2019]. These assays reveal the specific sites and residues that confer antigenic escape from polyclonal sera obtained from individual humans. A sufficiently broad geographic and

temporal sample of human sera with these assays could reveal consistent patterns of the immune landscape H3N2 strains must navigate to be globally successful. Models should also integrate information from multiple segments of the influenza genome and will need to balance the fitness benefits of evolution in genes such as neuraminidase [Chen et al., 2018] with the costs of reassortment [Villa and Lässig, 2017]. 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. Finally, forecasting models need to account for the geographic distribution of viruses and the vastly different sampling intensities across the globe. Most influenza sequence data come from highly developed countries that account for a small fraction of the global population, while globally successful clades of influenza H3N2 often emerge in less well-sampled regions [Russell et al., 2008, Rambaut et al., 2008, Bedford et al., 2015]. Explicitly accounting for these sampling biases and the associated migration dynamics would allow models to weight forecasts based on both viral fitness and transmission.

4.3.3 The nature of the predictive power of individual metrics remains unclear

Prediction of future influenza virus populations is intrinsically limited by the small number of data points available to train and test models. Increasingly more complex models are therefore prone to overfitting. Across the validation and test periods, we found that antigenic drift and mutational load were the most robust predictors of future success for seasonal influenza H3N2 populations.

Several metrics like the rate of frequency change or epitope mutations are naively expected to

have predictive power but do not. Others metrics like the mutational load are not expected to measure adaptation but are predictive. These results point to one aspect that often overlooked when comparing the genetic make-up of an asexual population at two time points: the future population is unlikely to descend from any of the sampled tips but ancestral lineages of the future population merge with those of the present population in the past. Optimal representatives of the future therefore tend to be tips in the present that tend to be basal and less evolved. The LBI and the mutational load metric have the tendency to assign low fitness to evolved tips. The LBI in particular assigns high fitness to the base of large clades. Much of the predictive power, in the sense of a reduced distance between the predicted and observed populations, might be due to putting more weight on less evolved strains rather than *bona fide* prediction of fitness. In a companion manuscript, Barrat-Charlaix et al. [2020] show that LBI has little predictive power for fixation probabilities of mutations in H3N2.

Our framework enables real-time practical forecasts of these populations by leveraging historical and modern experimental assays and gene sequences. By releasing our framework as an open source tool based on modern data science standards like tidy data frames, we hope to encourage continued development of this tool by the influenza research community. We additionally anticipate that the ability to forecast the sequence composition of populations with earth mover's distance will enable future forecasting research with pathogens whose genomes cannot be analyzed by traditional phylogenetic methods including recombinant viruses, bacteria, and fungi.

4.3.4 Model sharing and extensions

The entire workflow for our analyses was implemented with Snakemake [Köster and Rahmann, 2012]. We have provided all source code, configuration files, and datasets at https://github.com/blab/flu-forecasting.

4.4 Materials and methods

4.4.1 Simulation of influenza H3N2-like populations

We simulated the long-term evolution of H3N2-like viruses with SANTA-SIM [Jariani et al., 2019] for 10,000 generations or 50 years where 200 generations was equivalent to 1 year. We discarded the first 10 years as a burn-in period, selected the next 30 years for model fitting and validation, and held out the last 9 years as out-of-sample data for model testing (Figure 4.21). Each simulated population was seeded with the full length HA from A/Beijing/32/1992 (NCBI accession: U26830.1) such that all simulated sequences contained signal peptide, HA1, and HA2 domains. We defined purifying selection across all three domains, allowing the preferred amino acid at each site to change at a fixed rate over time. We additionally defined exposure-dependent selection for 49 putative epitope sites in HA1 [Łuksza and Lässig, 2014] to impose an effect of antigenic novelty that would allow mutations at those sites to increase viral fitness despite underlying purifying selection. We modified the SANTA-SIM source code to enable the inclusion of true fitness values for each strain in the FASTA header of the sampled sequences from each generation. This modified implementation has been integrated into the official SANTA-SIM code repository at `https://github.com/santa-dev/santa-sim` as of commit e2b3ea3. For our full analysis of model performance, we sampled 90 viruses per month to match the sampling density of natural populations. For tuning of hyperparameters, we sampled 10 viruses per month to enable rapid exploration of hyperparameter space.

4.4.2 Hyperparameter tuning with simulated populations

To avoid overfitting our models to the relatively limited data from natural populations, we used simulated H3N2-like populations to tune hyperparameters including the KDE bandwidth for frequency estimates and the L1 penalty for model coefficients. We simulated populations, as described above, and fit models for each parameter value using the true fitness of strains from the simulator.

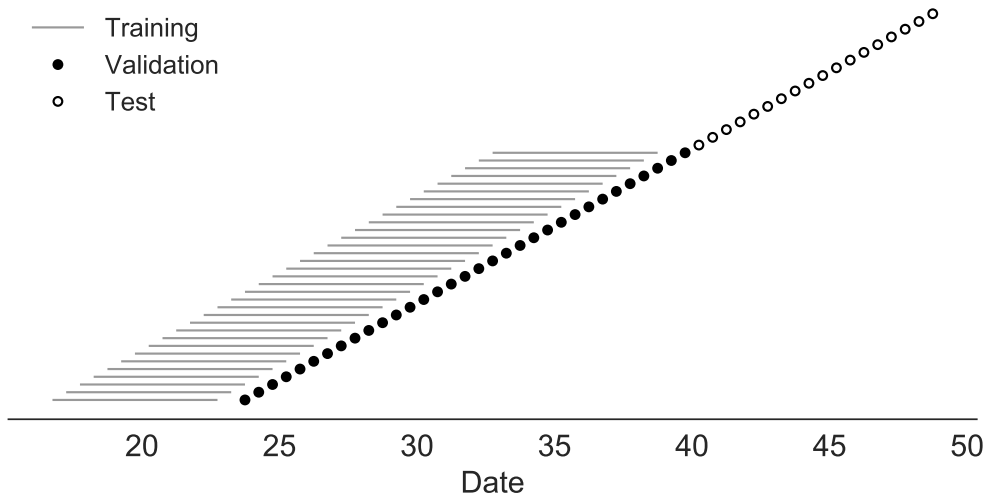

Figure 4.21: **Time-series cross-validation scheme for simulated populations.** Models were trained in six-year sliding windows (gray lines) and validated on out-of-sample data from validation timepoints (filled circles). Validation results from 30 years of data were used to iteratively tune model hyperparameters. After fixing hyperparameters, model coefficients were fixed at the mean values across all training windows. Fixed coefficients were applied to 9 years of new out-of-sample test data (open circles) to estimate true forecast errors.

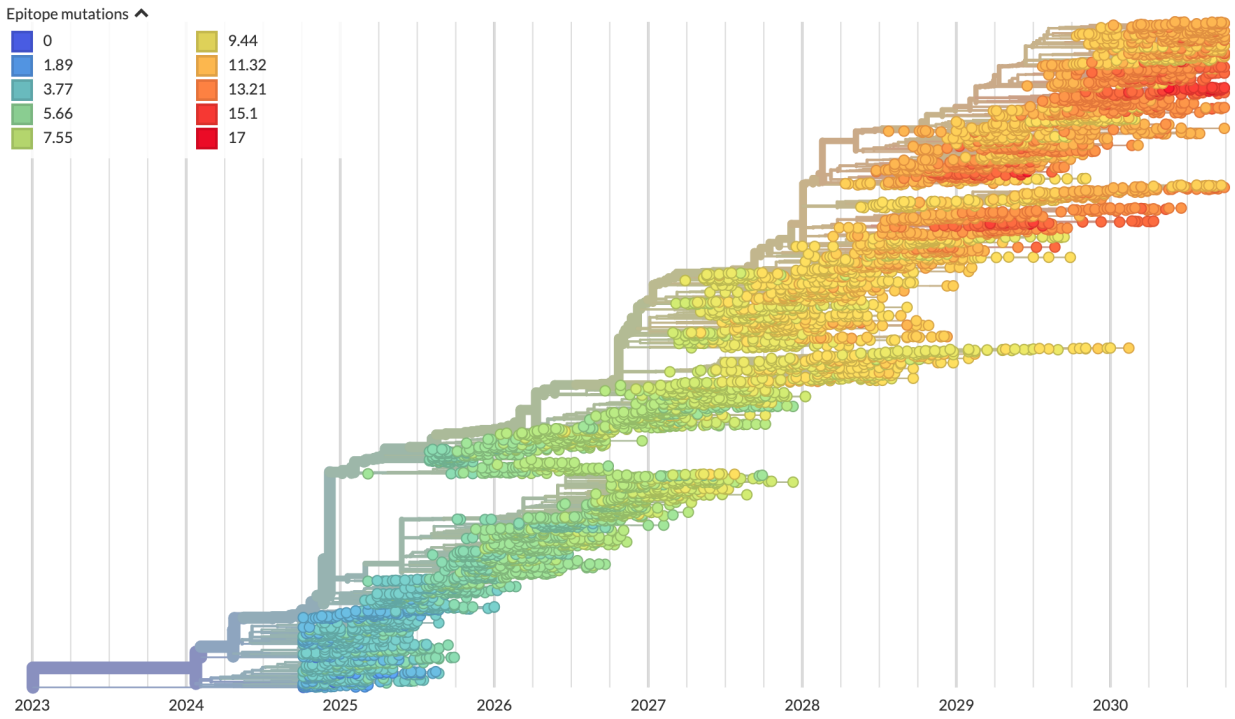

Figure 4.22: **Phylogeny of H3N2-like HA sequences sampled between the 24th and 30th years of simulated evolution.** The phylogenetic structure and rate of accumulated epitope and non-epitope mutations match patterns observed in phylogenies of natural sequences. Sample dates were annotated as the generation in the simulation divided by 200 and added to 2000, to acquire realistic date ranges that were compatible with our modeling machinery.

branch type	epitope mutations	non-epitope mutations	epitope-to-non-epitope ratio
side branch	590	1327	0.44
trunk	23	12	1.92

Table 4.4: **Number of epitope and non-epitope mutations per branch by trunk or side branch status for simulated populations.** Epitope sites were defined previously described [Łuksza and Lässig, 2014]. Annotation of trunk and side branch was performed as previously described [Bedford et al., 2015]. Mutations were calculated for the full validation tree for simulated sequences samples between October of years 10 and 40.

branch type	epitope mutations	non-epitope mutations	epitope-to-non-epitope ratio
side branch	485	1177	0.41
trunk	50	32	1.56

Table 4.5: **Number of epitope and non-epitope mutations per branch by trunk or side branch status for natural populations.** Epitope sites were defined previously described [Łuksza and Lässig, 2014]. Annotation of trunk and side branch was performed as previously described [Bedford et al., 2015]. Mutations were calculated for the full validation tree for natural sequences samples between 1990 and 2015.

We identified the optimal KDE bandwidth for frequencies as the value that minimized the difference between the mean distances to the future from the true fitness model and the naive model. We set the L1 lambda penalty to zero, to reduce variables in the analysis and avoid interactions between the coefficients and the KDE bandwidths. Higher bandwidths completely wash out dynamics of populations by making all strains appear to exist for long time periods. This flattening of frequency trajectories means that as bandwidths increase, the naive model gets more accurate and less informative. Given this behavior, we found the bandwidth that produced the minimum difference between distances to the future for the true fitness and naive models instead of the bandwidth that produced the minimum mean model distance. Based on this analysis, we identified an optimal bandwidth of $\frac{2}{12}$ or the equivalent of 2-months for floating point dates. Next, we identified an L1 penalty of 0.1 for model coefficients that minimized the mean distance to the future for the true fitness model.

4.4.3 Antigenic data

Hemagglutination inhibition (HI) and focus reduction assay (FRA) measurements were provided by WHO Global Influenza Surveillance and Response System (GISRS) Collaborating Centers in London, Melbourne, Atlanta and Tokyo. We converted these raw two-fold dilution measurements to log_2 titer drops normalized by the corresponding log_2 autologous measurements as previously described [Neher et al., 2016].

4.4.4 Strain selection for natural populations

Prior to our analyses, we downloaded all HA sequences and metadata from GISAID [Shu and McCauley, 2017]. For model training and validation, we selected 15,583 HA sequences $\geq$900 nucleotides that were sampled between October 1, 1990 and October 1, 2015. To account for known variation in sequence availability by region, we subsampled the selected sequences to a representative set of 90 viruses per month with even sampling across 10 global regions including Africa, Europe, North America, China, South Asia, Japan and Korea, Oceania, South America, Southeast Asia, and West Asia. We excluded all egg-passaged strains and all

strains with ambiguous year, month, and day annotations. We prioritized strains with more available HI titer measurements provided by the WHO GISRS Collaborating Centers. For model testing, we selected an additional 7,171 HA sequences corresponding to 90 viruses per month sampled between October 1, 2015 and October 1, 2019. We used these test sequences to evaluate the out-of-sample error of fixed model parameters learned during training and validation. Supplemental File S1 describes contributing laboratories for all 22,754 validation and test strains.

4.4.5 Phylogenetic inference

For each timepoint in model training, validation, and testing, we selected the subsampled HA sequences with collection dates up to that timepoint. We aligned sequences with the augur align command [Hadfield et al., 2018] and MAFFT v7.407 [Katoh et al., 2002]. We inferred initial phylogenies for HA sequences at each timepoint with IQ-TREE v1.6.10 [Nguyen et al., 2014]. To reconstruct time-resolved phylogenies, we applied TreeTime v0.5.6 [Sagulenko et al., 2018] with the augur refine command.

4.4.6 Frequency estimation

To account for uncertainty in collection date and sampling error, we applied a kernel density estimation (KDE) approach to calculate global strain frequencies. Specifically, we constructed a Gaussian kernel for each strain with the mean at the reported collection date and a variance (or KDE bandwidth) of two months. The bandwidth was identified by cross-validation, as described above. This bandwidth also roughly corresponds to the median lag time between strain collection and submission to the GISAID database. We estimated the frequency of each strain at each timepoint by calculating the probability density function of each KDE at that timepoint and normalizing the resulting values to sum to one. We implemented this frequency estimation logic in the augur frequencies command.

4.4.7 Model fitting and evaluation

Fitness model

We assumed that the evolution seasonal influenza H3N2 populations can be represented by a Malthusian growth fitness model, as previously described [Łuksza and Lässig, 2014]. Under this model, we estimated the future frequency, $\hat{x}_i(t + \Delta t)$, of each strain i from the strain's current frequency, $x_i(t)$, and fitness, $f_i(t)$, as follows where the resulting future frequencies were normalized to one by $\frac{1}{Z(t)}$.

$$\hat{x}_i(t + \Delta t) = \frac{1}{Z(t)} x_i(t) \exp(f_i(t)\Delta t) \tag{4.1}$$

We defined the fitness of each strain at time t as the additive combination of one or more fitness metrics, $f_{i,m}$, scaled by fitness coefficients, β_m. For example, Equation 4.2 estimates fitness per strain by mutational load (ml) and local branching index (lbi).

$$f_i(t) = \beta_{\text{ne}} f_{i,\text{ml}}(t) + \beta_{\text{lbi}} f_{i,\text{lbi}}(t) \tag{4.2}$$

Model target

For a model based on any given combination of fitness metrics, we found the fitness coefficients that minimized the earth mover's distance (EMD) [Rubner et al., 1998, Kusner et al., 2015] between amino acid sequences from the observed future population at time $u = t + \Delta t$ and the estimated future population created by projecting frequencies of strains at time t by their estimated fitnesses. Solving for EMD identifies the minimum about of "earth" that must be moved from a source population to a sink population to make those populations as similar as possible. This solution requires both a "ground distance" between pairs of strains from both populations and weights assigned to each strain that determine how much that strain contributes to the overall distance.

For each timepoint t and corresponding timepoint $u = t + 1$, we defined the ground distance as the Hamming distance between HA amino acid sequences for all pairs of strains between timepoints. For strains with less than full length nucleotide sequences, we inferred missing nucleotides through TreeTime's ancestral sequence reconstruction analysis. We defined weights for strains at timepoint t based on their projected future frequencies. We defined weights for strains at timepoint u based on their observed frequencies. We then identified the fitness coefficients that provided projected future frequencies that minimized the EMD between the estimated and observed future populations. With this metric, a perfect estimate of the future's strain sequence composition and frequencies would produce a distance of zero. However, the inevitable accumulation of substitutions between the two populations prevents this outcome. We calculated EMD with the Python bindings for the OpenCV 3.4.1 implementation [Bradski, 2000]. We applied the Nelder-Mead minimization algorithm as implemented in SciPy [Virtanen et al., 2020] to learn fitness coefficients that minimize the average of this distance metric over all timepoints in a given training window.

Lower bound on earth mover's distance

The minimum distance to the future between any two timepoints cannot be zero due to the accumulation of mutations between populations. We estimated the lower bound on earth mover's distance between timepoints using the following greedy solution to the optimal transport problem. For each timepoint t, we initialized the optimal frequency of each current strain to zero. For each strain in the future timepoint u, we identified the closest strain in the current timepoint by Hamming distance and added the frequency of the future strain to the optimal frequency of the corresponding current strain. This approach allows each strain from timepoint t to accumulate frequencies from multiple strains at timepoint u. We calculated the minimum distance between populations as the earth mover's distance between the resulting optimal frequencies for current strains, the observed frequencies of future strains, and the original distance matrix between those two populations.

Strain-specific distance to the future

We calculated the weighted Hamming distance to the future of each strain from the strain's HA amino acid sequence and the frequencies and sequences of the corresponding population one year in the future. Specifically, the distance between any strain i from timepoint t to the future timepoint u was the Hamming distance, h, between strain i's amino acid sequence, s_i, each future strain j's amino acid sequence, s_j, and the frequency of strain j in the future timepoint, $x_j(u)$.

$$d_i(u) = \sum_{j \in s(u)} x_j(u) h(s_i, s_j) \tag{4.3}$$

We calculated the estimated distance to the future for live forecasts with the same approach, replacing the observed future population frequencies and sequences with the estimated population based on our models.

$$d_i(\hat{u}) = \sum_{j \in s(\hat{u})} x_j(\hat{u}) h(s_i, s_j) \tag{4.4}$$

Time-series cross-validation

To obtain unbiased estimates for the out-of-sample errors of our models, we adopted the standard cross-validation strategy of training, validation, and testing. We divided our available data into an initial training and validation set spanning October 1990 to October 2015 and an additional testing set spanning October 2015 to October 2019 (Figure 4.23). We partitioned our training and validation data into six month seasons corresponding to winter in the Northern Hemisphere (October–April) and the Southern Hemisphere (April–October) and trained models to estimate frequencies of populations one year into the future from each season in six-year sliding windows. To calculate validation error for each training window, we applied the resulting model coefficients to estimate the future frequencies for the year after the last timepoint in the training window. These validation errors informed our tuning of

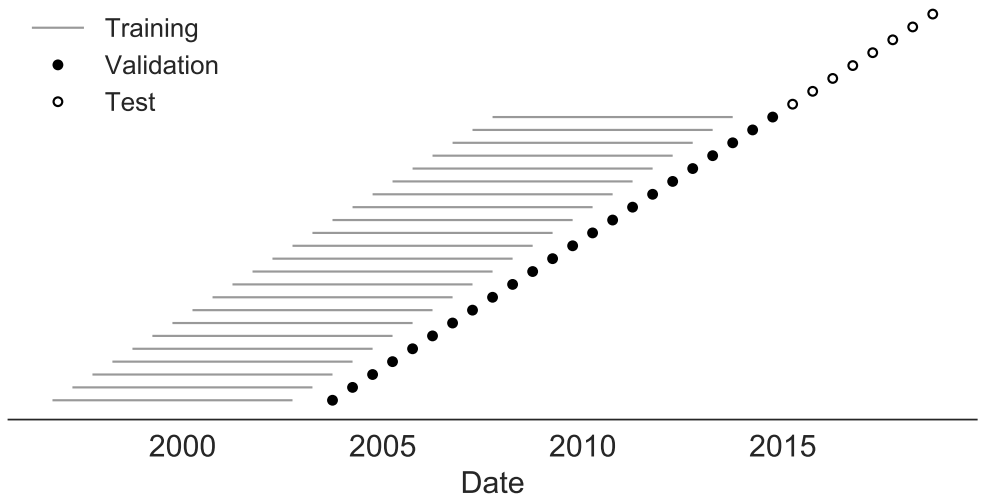

Figure 4.23: **Time-series cross-validation scheme for natural populations.** Models were trained in six-year sliding windows (gray lines) and validated on out-of-sample data from validation timepoints (filled circles). Validation results from 25 years of data were used to iteratively tune model hyperparameters. After fixing hyperparameters, model coefficients were fixed at the mean values across all training windows. Fixed coefficients were applied to four years of new out-of-sample test data (open circles) to estimate true forecast errors.

hyperparameters. Finally, we fixed the coefficients for each model at the mean values across all training windows and applied these fixed models to the test data to estimate the true forecasting accuracy of each model on previously unobserved data.

Model comparison by bootstrap tests

We compared the performance of different pairs of models using bootstrap tests. For each timepoint, we calculated the difference between one model's earth mover's distance to the future and the other model's distance. Values less than zero in the resulting empirical

sample	error_type	individual_model	composite_model	bootstrap_mean	bootstrap_std	p_value
simulated	validation	true fitness	mutational load + LBI	0.42	0.23	0.9644
simulated	validation	mutational load	mutational load + LBI	-1.03	0.21	<0.0001
simulated	validation	LBI	mutational load + LBI	-0.33	0.14	0.0091
simulated	test	true fitness	mutational load + LBI	-0.28	0.26	0.1392
simulated	test	mutational load	mutational load + LBI	-1.11	0.25	<0.0001
simulated	test	LBI	mutational load + LBI	-0.42	0.16	0.0001
natural	validation	mutational load	mutational load + LBI	-0.69	0.28	0.0036
natural	validation	LBI	mutational load + LBI	-0.23	0.09	0.0025
natural	validation	mutational load	mutational load + HI antigenic novelty	-0.31	0.18	0.0417
natural	validation	HI antigenic novelty	mutational load + HI antigenic novelty	-0.18	0.11	0.0513
natural	test	mutational load	mutational load + LBI	1.19	0.79	0.9432
natural	test	LBI	mutational load + LBI	-0.70	0.24	<0.0001
natural	test	mutational load	mutational load + HI antigenic novelty	-0.56	0.33	0.0133
natural	test	HI antigenic novelty	mutational load + HI antigenic novelty	-0.24	0.18	0.0999

Table 4.6: **Comparison of composite and individual model distances to the future by bootstrap test (see Methods).** The effect size of differences between models in amino acids is given by the mean and standard deviation of the bootstrap distributions. The p values represent the proportion of n=10,000 bootstrap samples where the mean difference was greater than or equal to zero.

distribution represent when the first model outperformed the second model. To determine whether the first model generally outperformed the second model, we bootstrapped the empirical difference distributions for n=10,000 samples and calculated the mean difference of each bootstrap sample. We calculated an empirical p value for the first model as the proportion of bootstrap samples with mean values greater than or equal to zero. This p value represents how likely the mean difference between the models' distances to the future is to be zero or greater. We measured the effect size of each comparison as the mean ± the standard deviation of the bootstrap distributions. We performed pairwise model comparisons for all biologically-informed models against the naive model (Figures 4.24 and 4.25). We also compared a subset of composite models to their respective individual models (Table 4.6).

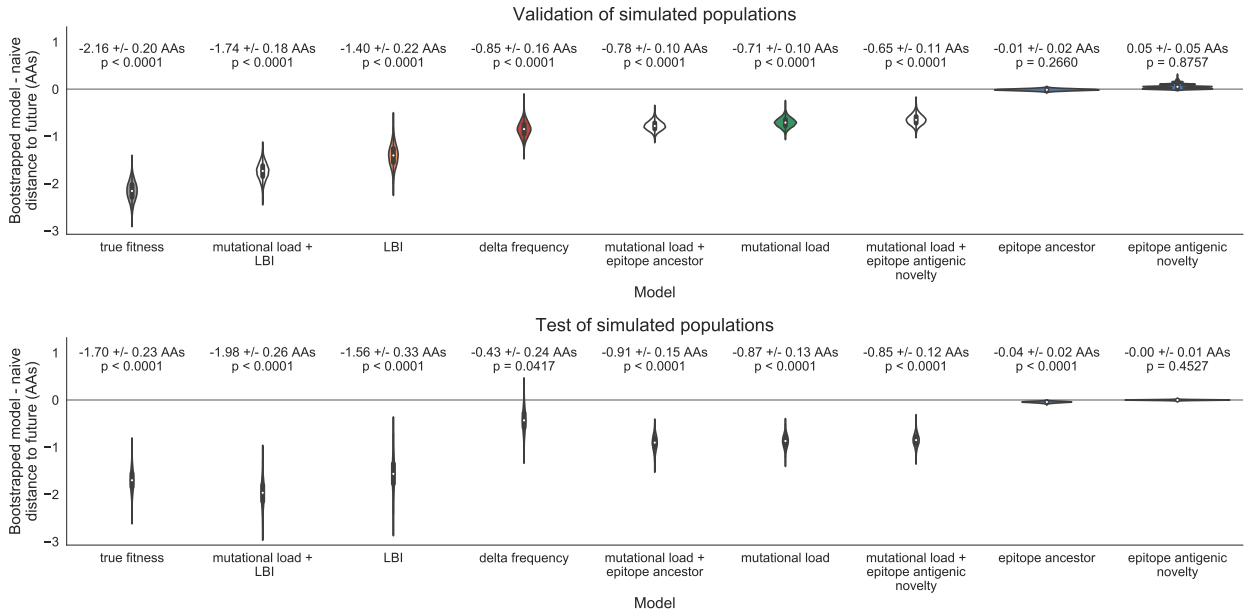

Figure 4.24: **Bootstrap distributions of the mean difference of distances to the future between biologically-informed and naive models for simulated populations.** Empirical differences in distances to the future were sampled with replacement and mean values for each bootstrap sample were calculated across n=10,000 bootstrap iterations. The horizontal gray line indicates a difference of zero between a given model and its corresponding naive model. Each model is annotated by the mean ± the standard deviation of the bootstrap distribution. Models are also annotated by the p-value representing the proportion of bootstrap samples with values less than zero (see Methods).

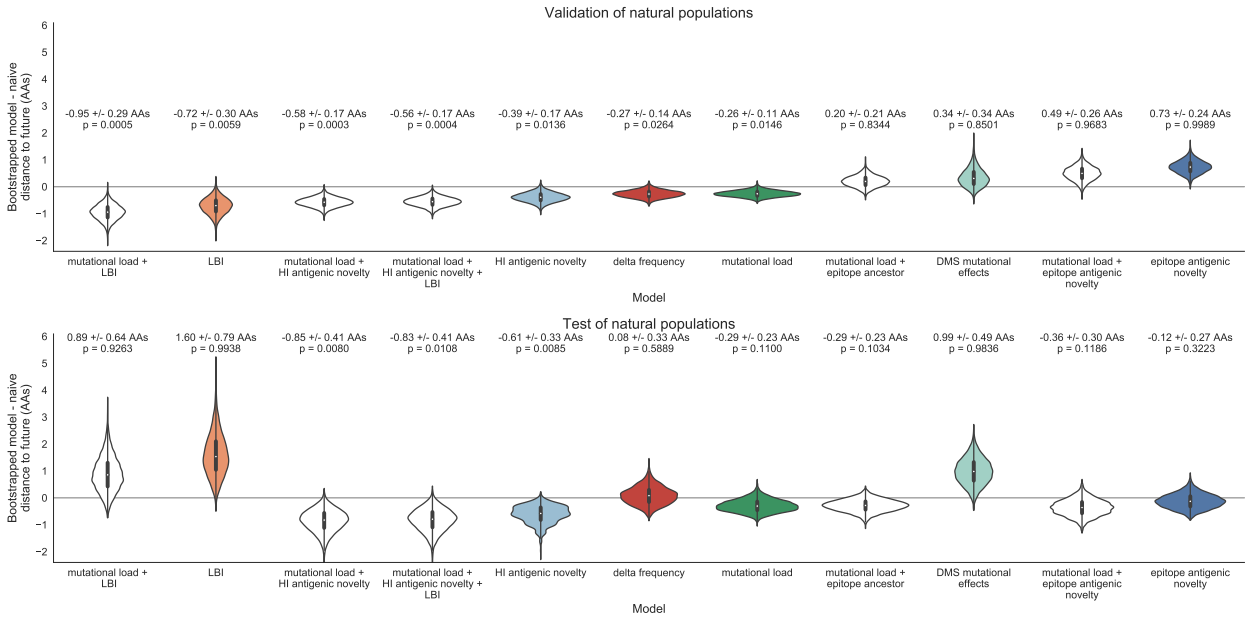

Figure 4.25: **Bootstrap distributions of the mean difference of distances to the future between biologically-informed and naive models for natural populations.** Empirical differences in distances to the future were sampled with replacement and mean values for each bootstrap sample were calculated across n=10,000 bootstrap iterations. The horizontal gray line indicates a difference of zero between a given model and its corresponding naive model. Each model is annotated by the mean ± the standard deviation of the bootstrap distribution. Models are also annotated by the p-value representing the proportion of bootstrap samples with values less than zero (see Methods).

4.4.8 Fitness metrics

We defined the following fitness metrics per strain and timepoint.

Antigenic drift

We estimated antigenic drift for each strain using either genetic or HI data. To estimate antigenic drift with genetic data, we implemented an antigenic novelty metric based on the "cross-immunity" metric originally defined by Łuksza and Lässig [2014]. Briefly, for each pair of strains in adjacent seasons, we counted the number of amino acid differences between the strains' HA sequences at 49 epitope sites. The one-based coordinates of these sites relative to the start of the HA1 segment were 50, 53, 54, 121, 122, 124, 126, 131, 133, 135, 137, 142, 143, 144, 145, 146, 155, 156, 157, 158, 159, 160, 163, 164, 172, 173, 174, 186, 188, 189, 190, 192, 193, 196, 197, 201, 207, 213, 217, 226, 227, 242, 244, 248, 275, 276, 278, 299, and 307. We limited pairwise comparisons to all strains sampled within the last five years from each timepoint. For each individual strain i at each timepoint t, we estimated that strain's ability to escape cross-immunity by summing the exponentially-scaled epitope distances between previously circulating strains and the given strain as in Equation 4.5. We defined the constant $D_0 = 14$, as in the original definition of cross-immunity [Łuksza and Lässig, 2014]. To compare these epitope sites with other previously published sites, we fit epitope antigenic novelty models based on sites defined by Wolf et al. [2006] and Koel et al. [2013].

$$f_{i,\mathrm{ep}}(t) = \sum_{j:t_j<t_i} -\max(x_j)\exp\left(-D_{\mathrm{ep}}(a_i,a_j)/D_0\right) \tag{4.5}$$

To test the historical contingency of the epitope sites defined above, we additionally identified a new set of sites with beneficial mutations across the training/validation period of October 1990 through October 2015. Following the general approach of Shih et al. [2007], we manually identified 25 sites in HA1 where mutations rapidly swept through the global population. We required mutations to emerge from below 5% global frequency and reach >90% frequency.

Although we did not require sweeps to complete within a fixed amount of time, we observed that they required no longer than one to three years to complete. To minimize false positives, we eliminated any sites where one or more mutations rose above 20% frequency and subsequently died out. If two or more sites had redundant sweep dynamics (mutations emerging and fixing at the same times), we retained the site with the most mutational sweeps. Based on this requirements, we defined our final collection of "oracle" sites in HA1 coordinates as 3, 45, 48, 50, 75, 140, 145, 156, 158, 159, 173, 186, 189, 193, 198, 202, 212, 222, 223, 225, 226, 227, 278, 311, and 312.

To estimate antigenic drift with HI data, we first applied the titer tree model to the phylogeny at a given timepoint and the corresponding HI data for its strains, as previously described by Neher et al. [2016]. This method effectively estimates the antigenic drift per branch in units of log_2 titer change. We selected all strains with nonzero frequencies in the last six months as "current strains" and all strains sampled five years prior to that threshold as "past strains". Next, we calculated the pairwise antigenic distance between all current and past strains as the sum of antigenic drift weights per branch on the phylogenetic path between each pair of strains. Finally, we calculated each strain's ability to escape cross-immunity using Equation 4.5 with the pairwise distances between epitope sequences replaced with pairwise antigenic distance from HI data. As with the original epitope antigenic novelty described above, this HI antigenic novelty metric produces higher values for strains that are more antigenically distinct from previously circulating strains.

Functional constraint

We estimated functional constraint for each strain using either genetic or deep mutational scanning (DMS) data. To estimate functional constraint with genetic data, we implemented the non-epitope mutation metric originally defined by Łuksza and Lässig [2014]. This metric counts the number of amino acid differences at 517 non-epitope sites in HA sequences between each strain i at timepoint t and that strain's most recent inferred ancestral sequence in the

previous season $(t-1)$.

We estimated functional constraint using mutational preferences from DMS data as previously defined [Lee et al., 2018]. Briefly, mutational effects were defined as the log ratio of DMS preferences, π, at site r for the derived amino acid, a_i, and the ancestral amino acid, a_j. As with the non-epitope mutation metric above, we considered only substitutions in HA between each strain i and that strain's most recent inferred ancestral sequence in the previous season. We calculated the total effect of these substitutions as the sum of the mutational preferences for each substitution, as in Equation 4.6.

$$f_{i,\mathrm{DMS}}(t) = \sum_{r \in r, a_i != r, a_j} \log_2 \frac{\pi_{r,a_i}}{\pi_{r,a_j}} \tag{4.6}$$

To determine whether DMS preferences could be used to define fitness metrics that were less dependent on the historical context of the background strain, we implemented two additional DMS-based metrics: "DMS entropy" and "DMS mutational load". For both metrics, we calculated the distance between HA amino acid sequences of each strain and its ancestral sequence in the previous season, to enable comparison of these metrics with the DMS mutational effects and mutational load metrics. For the "DMS entropy" metric, we calculated the distance between sequences such that each mismatch was weighted by the inverse entropy of DMS preferences at the site of the mismatch. We expected this metric to produce a negative coefficient similar to the mutational load metric, as higher values will result from mutations at sites with lower entropy and, thus, lower tolerance for mutations. For the "DMS mutational load" metric, we defined a novel set of non-epitope sites corresponding to each position in HA with a standardized entropy less than zero. With this metric, we sought to identify more highly conserved sites without weighting any one site differently from others. We anticipated that this lack of site-specific weighting would make the DMS mutational load metric even less background-dependent than the DMS entropy and DMS mutational effect metrics.

Clade growth

We estimated clade growth for each strain using local branching index (LBI) and the change in frequency over time (delta frequency). To calculate LBI for each strain at each timepoint, we applied the LBI heuristic algorithm as originally described [Neher et al., 2014] to the phylogenetic tree constructed at each timepoint. We set the neighborhood parameter, τ, to 0.3 and only considered viruses sampled in the last 6 months of each phylogeny as contributing to recent clade growth.

We estimated the change in frequency over time by calculating clade frequencies under a Brownian motion diffusion process as previously described [Lee et al., 2018]. These frequency calculations allowed us to assign a partial clade frequency to each strain within nested clades. We calculated the delta frequency as the change in frequency for each strain between the most recent timepoint in a given phylogeny and six months prior to that timepoint divided by 0.5 years.

4.4.9 Clustering of amino acid sequences for visualization

For the purpose of visualizing related amino acid sequences in Figure 5.2, we applied dimensionality reduction to pairwise amino acid distances followed by hierarchical clustering. Specifically, we selected a representative tree from our simulated population of viruses at month 10 of year 30. From this tree, we selected all strains with a collection date in the previous two years. We calculated the pairwise Hamming distance between the full-length HA amino acid sequences for all selected strains and applied t-SNE dimensionality reduction [van der Maaten and Hinton, 2008] to the resulting distance matrix (n=2 components, perplexity=30.0, and learning rate=400). We assigned each strain to a cluster based on its two-dimensional t-SNE embedding using DBSCAN [Ester et al., 1996] with a maximum neighborhood distance of 10 AAs and a minimum of 20 strains per cluster. Despite known limitations of applying hierarchical clustering to manifold projections that do not preserve sample density, this approach allowed us to effectively assign strains to qualitative genetic

Model	Coefficients	Distance to future (AAs)		Model > naive	
		Validation	Test	Validation	Test
mutational load	-0.68 +/- 0.34	5.44 +/- 1.80*	7.70 +/- 3.53	18 (78%)	4 (50%)
+ LBI	1.03 +/- 0.40				
LBI	1.12 +/- 0.51	5.68 +/- 1.91*	8.40 +/- 3.97	17 (74%)	2 (25%)
oracle antigenic novelty	0.80 +/- 0.21	5.71 +/- 1.27^	8.06 +/- 2.49^	18 (78%)	2 (25%)
HI antigenic novelty	0.89 +/- 0.23	5.82 +/- 1.50*	5.97 +/- 1.47*	17 (74%)	6 (75%)
+ mutational load	-1.01 +/- 0.42				
HI antigenic novelty	0.90 +/- 0.23	5.84 +/- 1.51*	5.99 +/- 1.46*	16 (70%)	6 (75%)
+ mutational load	-1.00 +/- 0.44				
+ LBI	-0.04 +/- 0.09				
HI antigenic novelty	0.83 +/- 0.20	6.01 +/- 1.50*	6.21 +/- 1.44*	16 (70%)	7 (88%)
delta frequency	0.79 +/- 0.47	6.13 +/- 1.71*	6.90 +/- 2.30	16 (70%)	5 (62%)
mutational load	-0.99 +/- 0.30	6.14 +/- 1.37*	6.53 +/- 1.39	17 (74%)	6 (75%)
Koel epitope antigenic novelty	0.28 +/- 0.36	6.22 +/- 1.26^	6.72 +/- 1.51^	18 (78%)	4 (50%)
naive	0.00 +/- 0.00	6.40 +/- 1.36	6.82 +/- 1.74	0 (0%)	0 (0%)
DMS entropy	-0.03 +/- 0.10	6.40 +/- 1.36^	6.81 +/- 1.73^	9 (39%)	6 (75%)
DMS mutational load	-0.02 +/- 0.13	6.45 +/- 1.42^	6.82 +/- 1.73^	7 (30%)	5 (62%)
epitope ancestor	0.53 +/- 0.52	6.60 +/- 1.34	6.53 +/- 1.51	12 (52%)	4 (50%)
+ mutational load	-0.77 +/- 0.32				
DMS mutational effects	1.25 +/- 0.84	6.75 +/- 1.95	7.80 +/- 2.97	11 (48%)	4 (50%)
Wolf epitope antigenic novelty	0.31 +/- 0.51	6.83 +/- 1.30^	6.97 +/- 1.41^	4 (17%)	3 (38%)
epitope ancestor	0.23 +/- 0.51	6.89 +/- 1.39^	6.82 +/- 1.67^	8 (35%)	4 (50%)
epitope antigenic novelty	0.57 +/- 0.77	6.89 +/- 1.42	6.46 +/- 1.31	7 (30%)	4 (50%)
+ mutational load	-0.77 +/- 0.27				
epitope antigenic novelty	0.52 +/- 0.73	7.13 +/- 1.47	6.70 +/- 1.51	7 (30%)	5 (62%)

Table 4.7: **All model coefficients and performance on validation and test data for natural populations ordered from best to worst by distance to the future, as in Table 4.2.** Distances annotated with asterisks (*) were significantly closer to the future than the naive model as measured by bootstrap tests (see Methods and Figure 4.25).

Table 4.7: (continued) Distances annotated with carets (∧) were not tested for significance relative to the naive model. Validation results are based on 23 timepoints. Test results are based on eight timepoints not observed during model training and validation. Model results for additional variants of fitness metrics including those based on epitope mutations and DMS preferences are included for reference. Source data are available at `https://github.com/blab/flu-forecasting/blob/published/manuscript/Table_7-source_data_1.csv` and `https://github.com/blab/flu-forecasting/blob/published/manuscript/Table_7-source_data_2.csv`.

clusters for the purposes of visualization.

4.5 Supplemental Files

Supplemental File S1. GISAID accessions and metadata including originating and submitting labs for natural strains used across all timepoints.

Chapter 5

CONCLUSIONS

In the preceding chapters, I have shown how experimentally-informed computational models and interactive data visualizations can improve our understanding of seasonal influenza evolution. Clearly, many questions remain about seasonal influenza evolution and their answers are currently debated by experts in the field. One relevant example of these unanswered questions is how predictable the frequencies of individual mutations in seasonal influenza populations are. In the last years of this work, I collaborated with Dr. Pierre Barrat-Charlaix from Dr. Richard Neher's lab to help answer this question. The results from this collaboration provide a helpful counterpoint to the results I've presented in this dissertation. Below, I provide a summary of Dr. Barrat-Charlaix's findings, how they relate to this dissertation, and how these results have changed our understanding of influenza evolution.

5.1 Does seasonal influenza evolve like we think it does?

In Barrat-Charlaix et al. [2020], we investigated the predictability of seasonal influenza mutation frequencies. We explicitly avoided modeling seasonal influenza evolution and focused on an empirical account of long-term outcomes for mutation frequency trajectories. We selected all available HA and NA sequences for seasonal influenza lineages A/H3N2 and A/H1N1pdm, performed multiple sequence alignments per lineage and gene, binned sequences by month, and calculated the frequencies of mutations per site and month. From these data, we constructed frequency trajectories of individual mutations that were rising in frequency from zero. We expected these rising mutations to represent beneficial, large-effect mutations

that would sweep through the global population as predicted by the population genetic theory [Neher, 2013]. By considering individual mutations, we effectively averaged the outcomes of these mutations across all genetic backgrounds. We evaluated the outcomes of trajectories for mutations that had risen from 0% to approximately 30% global frequency and classified trajectories for mutations that fixed, died out, or persisted as polymorphisms.

The average trajectory of individual rising A/H3N2 mutations failed to rise toward fixation (Figure 5.1). Instead, the future frequency of these mutations was no higher on average than their initial frequency. We repeated this analysis for mutations with initial frequencies of 50% and 75% and for mutations in A/H1N1pdm and found nearly the same results. From these results, we concluded that it is not possible to predict the short-term dynamics of individual mutations based solely on their recent success.

Next, we calculated the fixation probability of each mutation trajectory based on its initial frequency. Surprisingly, we found that the fixation probabilities of A/H3N2 mutations were equal to their initial frequencies. This pattern corresponds to what we expect for mutations evolving neutrally, where population genetic theory predicts that fixation probability is equal to current mutation frequency. Generally, the pattern remained the same even when we binned mutations by high LBI, presence at epitope sites, multiple appearances of a mutation in a tree, geographic spread, or other potential metrics associated with high fitness. We concluded that the recent success of rising mutations provides no information about their eventual fixation.

We tested whether we could explain these results by genetic linkage or clonal interference by simulating seasonal influenza-like populations under these evolutionary constraints. Mutation trajectories from simulated populations were more predictable than those from natural populations. The closest our simulations came to matching the uncertainty of natural populations was when we dramatically increased the rate at which the fitness landscape of simulated populations changed. These results suggested that we cannot explain the unpredictable nature of seasonal influenza mutation trajectories by linkage or clonal interference alone.

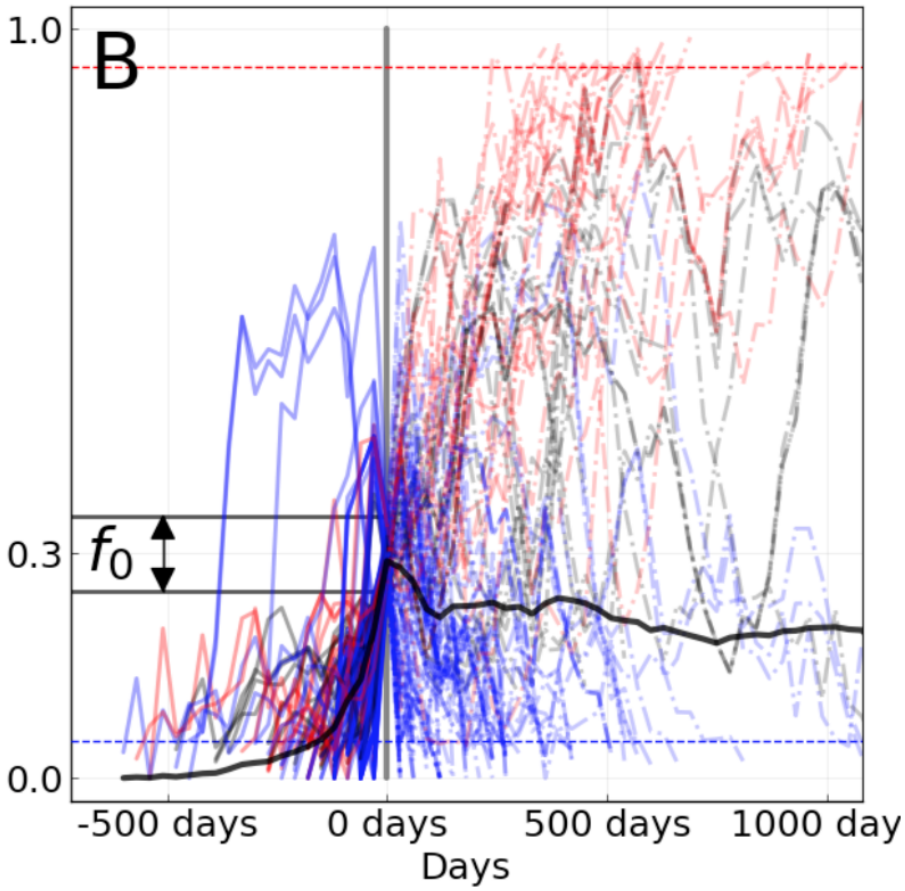

Figure 5.1: **Mutation trajectories for seasonal influenza A/H3N2 where mutations rose from a frequency of zero to approximately 30% frequency.** Dashed horizontal lines represent thresholds for fixation (red) and loss (blue). Trajectory colors also indicate eventual fixation (red), loss (blue), or persistence as a polymorphism (black). The thick black dashed line indicates the average frequency of all trajectories shown. For the interactive figure, hover over individual trajectories to highlight their full extent and details about the current frequency of a given mutation at each timepoint. Use the radio buttons to filter trajectories by segment and outcome. (After Figure 1B in Barrat-Charlaix et al. [2020].) Explore an interactive version of this figure.

Since seasonal influenza mutation trajectories lacked "momentum" and LBI did not provide information about eventual fixation of mutations, we wondered whether we could identify the most representative sequence of future populations with a different metric. The consensus sequence is provably the best predictor for a neutrally evolving population. We found that the consensus sequence is often closer to the future population than the virus sequence with the highest LBI. Indeed, we found that the top LBI virus was frequently similar to the consensus sequence and often identical.

Taken together, our results from this empirical analysis reveal that beneficial mutations of large effect do not predictably sweep through seasonal influenza populations and fix. Instead, the average outcome for any individual mutation resembles neutral evolution, despite the strong positive selection expected to act on these mutations. Although simulations rule out clonal interference between large effect mutations as an explanation for these results, we cannot discount the role of multiple mutations of similar, smaller effects in the overall fitness of seasonal influenza viruses and the fixation of multiple co-evolving mutations.

5.2 Can we forecast seasonal influenza evolution?

In Huddleston et al. [2020], we built a modeling framework based on the approach described in Łuksza and Lässig [2014] to forecast seasonal influenza A/H3N2 populations one year in advance. We used this framework to predict the sequence composition of the future population, the frequency dynamics of clades, and the virus in the current population that most represented the future population. As in Barrat-Charlaix et al. [2020] and Łuksza and Lässig [2014], we assumed that viruses grow exponentially as a function of their fitness and that viruses with similarly high fitness compete with each other under clonal interference. In contrast to Barrat-Charlaix et al. [2020], we considered the fitness of complete amino acid haplotypes instead of individual mutations.

We estimated fitness with metrics based on HA sequences and experimental measurements of antigenic drift and functional constraint. The sequence-based metrics included the epitope

cross-immunity and mutational load estimates defined by Łuksza and Lässig [2014], LBI from Neher et al. [2014], and "delta frequency", a measure of recent change in clade frequency analogous to Barrat-Charlaix's rising mutations. The experimental metrics included a cross-immunity measure based on hemagglutination inhibition (HI) assays [Neher et al., 2016] and an estimate of functional constraint based on mutational preferences from deep mutational scanning experiments [Lee et al., 2018].

We trained models based on each of these metrics independently and in relevant combinations of complementary metrics. For each model, we fit coefficients per fitness metric that minimized the distance between the estimated and observed amino acid haplotype composition of the future (Figure 5.2). These coefficients represent the effect of each metric on seasonal influenza fitness. As a control, we also calculated the distance to the future population for a "naive" model that assumed the future population is the same as the current population. To test our framework, we simulated 40 years of evolution for seasonal influenza-like populations with SANTA-SIM and fit models to these data. After verifying our framework with simulated populations, we trained models for natural A/H3N2 populations using 25 years of historical data. We tested the accuracy of each model by applying the coefficients from the training data to forecasts of new out-of-sample data from the last 5 years of A/H3N2 evolution.

We found that the most robust forecasts depended on a combined model of experimentally-informed antigenic drift and sequence-based mutational load. Importantly, this model explicitly accounts for the benefits of antigenic drift and the costs of deleterious mutations. This model also slightly outperformed the naive model in its estimation of future clade frequencies. However, we found that the naive model often selected individual strains that were as close to the future population as the best biologically-informed model. The naive model's estimated closest strain to the future is effectively the weighted average of the current population and conceptually similar to the consensus sequence of the population. From these results, we concluded that the predictive gains of fitness models depend on the prediction target.

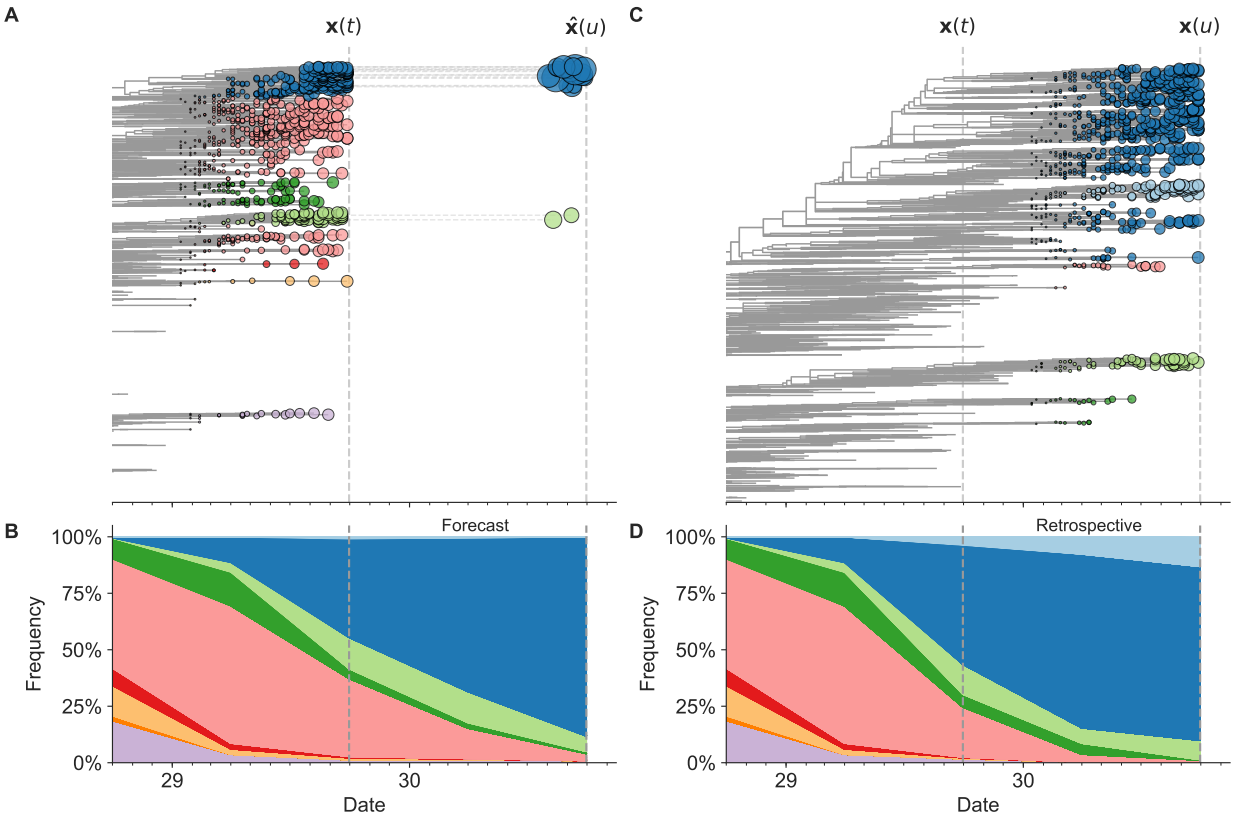

Figure 5.2: **Schematic representation of the fitness model for simulated H3N2-like populations.** The fitness of strains at timepoint t determines the estimated frequency of strains with similar sequences one year in the future at timepoint u. Genetically similar strains by amino acid sequence have similar colors. A) Strains at timepoint t, $\mathbf{x}(t)$, are shown in their phylogenetic context and sized by their frequency at that timepoint. The estimated future population at timepoint u, $\hat{\mathbf{x}}(u)$, is projected to the right with strains scaled in size by their projected frequency based on the known fitness of each simulated strain. B) The frequency trajectories of strains at timepoint t to u represent the predicted the growth of the dark blue strains to the detriment of the pink strains. C) Strains at timepoint u, $\mathbf{x}(u)$, are shown as in panel A. D) The observed frequency trajectories of strains at timepoint u broadly recapitulate the model's forecasts while also revealing increased diversity of sequences at the future timepoint that the model could not anticipate.

Surprisingly, the sequence-based metrics of epitope cross-immunity and delta frequency and the mutational preferences from DMS experiments had little predictive power. These metrics failed to make accurate forecasts because of their dependence on a specific historical context. For example, the original epitope cross-immunity metric [Łuksza and Lässig, 2014] depends on a predefined list of epitope sites that were originally identified in a retrospective study of seasonal influenza sequences up through 2005 [Shih et al., 2007]. This metric correspondingly failed to predict the future after 2005, suggesting that its previous success depended on inadvertently borrowing information from the future. Similarly, the mutational preferences from DMS experiments measure effects of all single amino acid mutations to the genetic background of the virus A/Perth/16/2009. The metric based on these preferences failed to predict the future after 2009, reflecting the strong dependence of these preferences on their original genetic background. Both delta frequency and LBI suffered from overfitting to the training data, in a more general form of historical dependence.

5.3 How do results from our two studies compare?

The two studies we have presented here use different approaches to analyze the same natural seasonal influenza populations. We were especially interested to understand how simulated populations from the two studies differed and whether the optimal predictor from Barrat-Charlaix et al. [2020] could also be an accurate fitness metric in the modeling framework from Huddleston et al. [2020].

Simulated populations play an important role in our two studies. We generated these simulated data as a source of truth where we understand the population dynamics because we defined them. In Barrat-Charlaix et al. [2020], the simulated binary populations from ffpopsim [Zanini and Neher, 2012] evolved under strong epistasis and immune escape pressure. These populations showed us that mutation trajectories could be predictable under these population genetic constraints. In Huddleston et al. [2020], the simulated nucleotide populations from SANTA-SIM [Jariani et al., 2019] also evolved under strong epistasis, purifying selection,

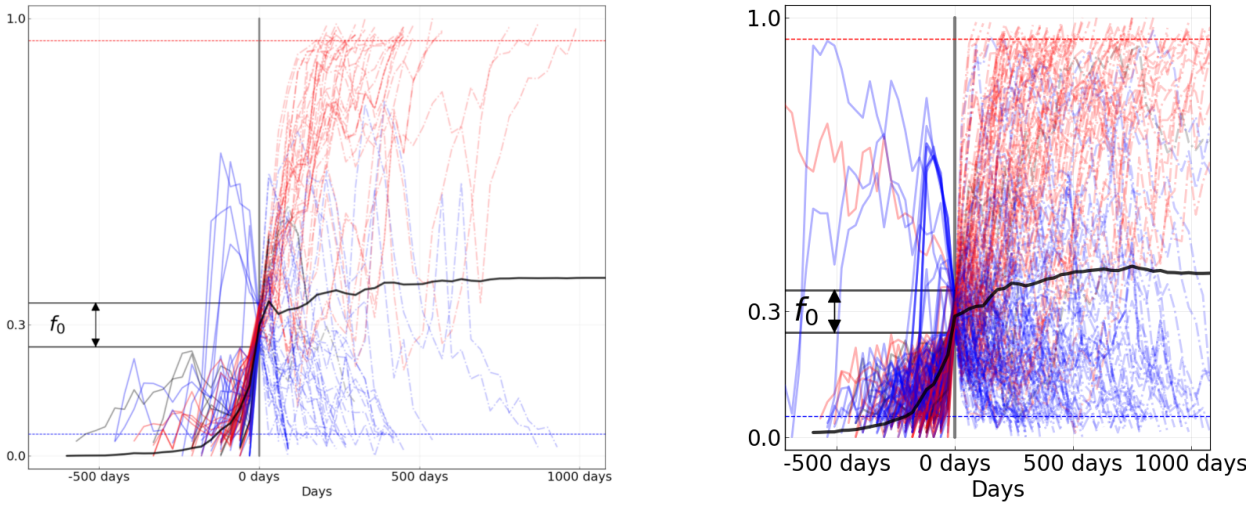

Figure 5.3: **Comparison of rising trajectories for natural H1N1pdm trajectories from Barrat-Charlaix et al. [2020] and simulated seasonal influenza-like populations from Huddleston et al. [2020].** A) Rising trajectories for H1N1pdm mutations as reported in Figure S9 of Barrat-Charlaix et al. [2020]. B) Rising trajectories for seasonal influenza-like populations simulated with SANTA-SIM in Huddleston et al. [2020]. Mutation trajectories from simulated populations resemble those of natural H1N1pdm mutations.

and an "exposure dependent" fitness function that mimics immune escape pressure. We used these populations to confirm that our forecasting framework could accurately predict the composition of future populations. Interestingly, when we inspected the predictability of the mutation trajectories for these simulated populations, we found that they resembled the weak predictability of natural H1N1pdm trajectories (Figure 5.3). Despite the weak predictability of mutation trajectories from these simulated populations, we were able to forecast the composition of their future populations. These results highlight the importance of using complete haplotypes to make predictions, as individual mutation trajectories remain difficult to predict.

We also wanted to know whether the optimal metric from Barrat-Charlaix et al. [2020] for selecting a representative of the future, the consensus sequence of the current population, could make accurate forecasts in the modeling framework from Huddleston et al. [2020]. We noted above that the closest strain to the future selected by the naive model from Huddleston et al. [2020] is analogous to the consensus sequence of the current population. One important difference is that the naive model has to select a previously sampled strain while the consensus sequence represents a hypothetical strain that may not exist in nature. To understand whether the consensus sequence could also improve forecasts of the future population's haplotype composition, we developed a new fitness metric called the "distance from consensus". For each timepoint in our forecasting analysis, we constructed the amino acid consensus sequence from all extant strains and calculated the pairwise distance between the consensus and each extant strain. If the consensus sequence is the best representation of the future population, we expected the corresponding model's coefficients to be consistently negative. This negative coefficient would have the effect of penalizing strains whose amino acid sequences diverged greatly from the consensus sequence.

We fit a model to this new metric using the same 25 years of historical A/H3N2 data described in Huddleston et al. [2020] and tested the robustness of the model on the last 5 years of A/H3N2 data. We compared the performance of this model to models for LBI and experimental measures of antigenic drift (HI antigenic novelty). For the first half of the training period, the distance to consensus metric received a coefficient of zero, meaning it did not improve forecasts over the naive model (Figure 5.4). In the second half of the training period, the metric received a strong negative coefficient, as we expected. When we applied the mean coefficient from the training period to out-of-sample data in the test period, we found that the distance from consensus metric outperformed LBI and performed only slightly worse than the antigenic drift metric. These results support findings from both of our studies. The consensus sequence is a more robust representative of the future than LBI, as shown in Barrat-Charlaix et al. [2020]. However, experimental measurements of antigenic drift still

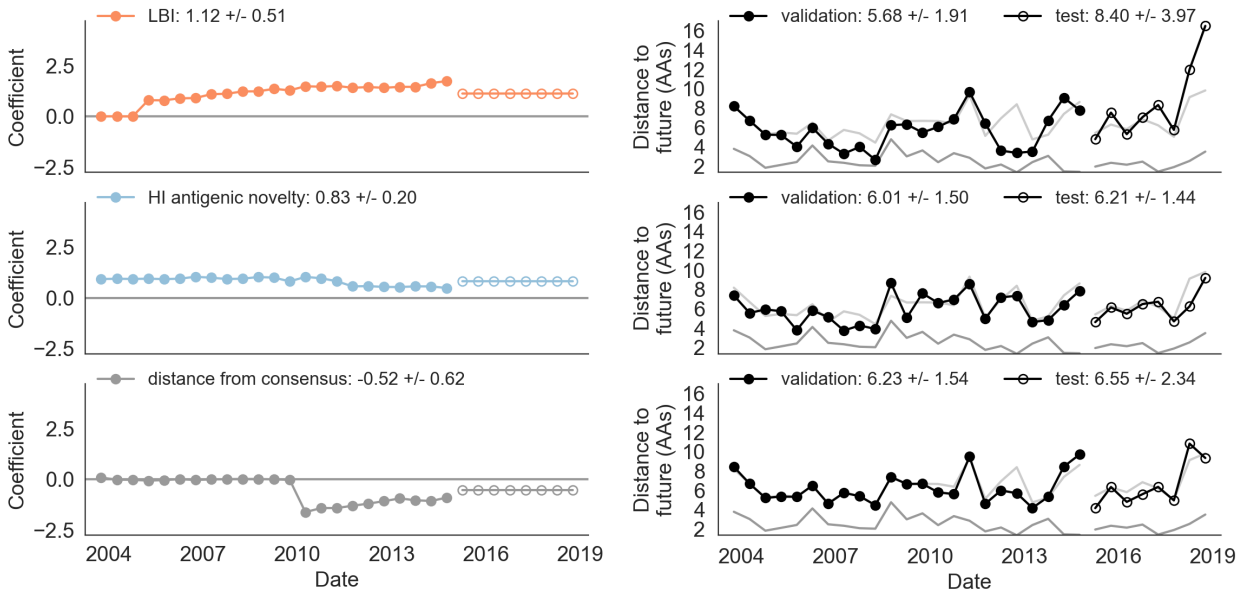

Figure 5.4: **Model coefficients and distance to the future for LBI, HI antigenic novelty, and distance from consensus metrics.** A) Coefficients are shown per validation timepoint (solid circles, N=23) with the mean +/- standard deviation in the top-left corner. For model testing, coefficients were fixed to their mean values from training/validation and applied to out-of-sample test data (open circles, N=8). B) Distances between projected and observed populations are shown per validation timepoint (solid black circles) or test timepoint (open black circles). The mean +/- standard deviation of distances per validation timepoint are shown in the top-left of each panel. Corresponding values per test timepoint are in the top-right. The naive model's distance to the future (light gray) was 6.40 +/- 1.36 AAs for validation timepoints and 6.82 +/- 1.74 AAs for test timepoints. The corresponding lower bounds on the estimated distance to the future (dark gray) were 2.60 +/- 0.89 AAs and 2.28 +/- 0.61 AAs.

provide more information about the future population than sequence-only metrics, as shown in Huddleston et al. [2020]. We anticipate that this new distance from consensus metric could eventually replace the existing mutational load metric in a combined model with HI antigenic novelty. This new combined model could potentially provide better estimates of functional constraint (by limiting changes from the consensus) and antigenic drift (by using experimental measures of antigenic drift phenotypes.)

5.4 How have these results changed how we think about seasonal influenza evolution?

In general, we found that the evolution of H3N2 seasonal influenza populations remains difficult to predict. The frequency dynamics and fixation probabilities of individual mutations resemble neutrally evolving alleles. We can weakly predict the frequency dynamics of seasonal influenza clades when we combine experimental and genetic data in models that account for antigenic drift and mutational load. In the best case, we can use these same biologically-informed models to predict the sequence composition of future seasonal influenza populations. However, these complex fitness models do not always outperform simpler models, when predicting which individual virus is the most representative of the future population. In Barrat-Charlaix et al. [2020], the consensus sequence of the current population was as close or closer to the future population than the sequence with the highest local branching index. In Huddleston et al. [2020], a naive model estimated the single closest strain to the future nearly as well as the best biologically-informed models.

Successful seasonal influenza predictions depend on the choice of prediction targets and fitness metrics. Future prediction efforts should attempt to estimate the composition of future populations instead of future clade frequencies. Fitness models should account for the genetic background of beneficial mutations and favor fitness metrics that are the least susceptible to model overfitting and historical contingency. The benefits of considering the genetic background of individual mutations in HA suggest that considering the context of all genes should yield gains, too. We need measures of antigenic drift from human antisera

to complement current measures based on ferret antisera. We may also improve forecast accuracy by accounting for seasonal influenza's global migration patterns. Most importantly, we should make the forecasting problem itself easier by embracing efforts to reduce the lag between vaccine composition decisions and distribution to the public.

With a validated forecasting framework in hand, we can readily produce real-time forecasts at nextstrain.org and as part of recommendations submitted to the World Health Organization prior to biannual vaccine composition meetings Bedford et al. [2019]. Other modeling groups produce similar reports with forecasts of even higher precision. However, there is no expectation or formal requirement for us to revisit and evaluate the accuracy of our predictions. The long-term historical accuracy of forecasting models does not matter, if these models fail to perform consistently in modern forecasts. For this reason, we recommend that modeling groups formally archive their forecasts for each composition meeting and periodically assess the performance of their models in real forecasting conditions. Modeling groups that participate in short-term or within-season seasonal influenza forecasts have already adopted this approach, as part of the CDC FluSight network. The Zoltar Forecast Archive [Reich et al., 2020] was designed with these forecasts in mind and could be modified to support storage of long-term forecasts. We expect that these formal archives of real forecasts will improve accountability of modeling groups, increase the accuracy of forecasts made by their models, and encourage community consensus on model targets.

BIBLIOGRAPHY

Pierre Barrat-Charlaix, John Huddleston, Trevor Bedford, and Richard A. Neher. Limited predictability of amino acid substitutions in seasonal influenza viruses. *bioRxiv*, 2020. doi: 10.1101/2020.07.31.231100. URL https://www.biorxiv.org/content/early/2020/07/31/2020.07.31.231100.

T. Bedford, A. L. Greninger, P. Roychoudhury, L. M. Starita, M. Famulare, M. Huang, A. Nalla, G. Pepper, A. Reinhardt, H. Xie, L. Shrestha, T. N. Nguyen, A. Adler, E. Brandstetter, S. Cho, D. Giroux, P. D. Han, K. Fay, C. D. Frazar, M. Ilcisin, K. Lacombe, J. Lee, A. Kiavand, M. Richardson, T. R. Sibley, M. Truong, C. R. Wolf, D. A. Nickerson, M. J. Rieder, J. A. Englund, The Seattle Flu Study Investigators, J. Hadfield, E. B. Hodcroft, J. Huddleston, L. H. Moncla, N. F. Müller, R. A. Neher, X. Deng, W. Gu, S. Federman, C. Chiu, J. S. Duchin, R. Gautom, G. Melly, B. Hiatt, P. Dykema, S. Lindquist, K. Queen, Y. Tao, A. Uehara, S. Tong, D. MacCannell, G. L. Armstrong, G. S. Baird, H. Y. Chu, J. Shendure, and K. R. Jerome. Cryptic transmission of SARS-CoV-2 in Washington state. *Science*, September 2020. ISSN 0036-8075. doi: 10.1126/science.abc0523. URL https://science.sciencemag.org/content/early/2020/09/09/science.abc0523.

Trevor Bedford and Richard Neher. Seasonal influenza circulation patterns and projections for feb 2018 to feb 2019. *bioRxiv*, 2018. doi: 10.1101/271114. URL https://www.biorxiv.org/content/early/2018/02/25/271114.

Trevor Bedford, Sarah Cobey, and Mercedes Pascual. Strength and tempo of selection revealed in viral gene genealogies. *BMC evolutionary biology*, 11(1):220, 2011.

Trevor Bedford, Andrew Rambaut, and Mercedes Pascual. Canalization of the evolutionary

trajectory of the human influenza virus. *BMC Biology*, 10(1):38, Apr 2012. ISSN 1741-7007. doi: 10.1186/1741-7007-10-38. URL https://doi.org/10.1186/1741-7007-10-38.

Trevor Bedford, Marc A Suchard, Philippe Lemey, Gytis Dudas, Victoria Gregory, Alan J Hay, John W McCauley, Colin A Russell, Derek J Smith, and Andrew Rambaut. Integrating influenza antigenic dynamics with molecular evolution. *Elife*, 3:e01914, 2014.

Trevor Bedford, Steven Riley, Ian G Barr, Shobha Broor, Mandeep Chadha, Nancy J Cox, Rodney S Daniels, C Palani Gunasekaran, Aeron C Hurt, Anne Kelso, Alexander Klimov, Nicola S Lewis, Xiyan Li, John W McCauley, Takato Odagiri, Varsha Potdar, Andrew Rambaut, Yuelong Shu, Eugene Skepner, Derek J Smith, Marc A Suchard, Masato Tashiro, Dayan Wang, Xiyan Xu, Philippe Lemey, and Colin A Russell. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523(7559):217–220, July 2015.

Trevor Bedford, John Huddleston, Barney Potter, and Richard A. Neher. Seasonal influenza circulation patterns and projections for september 2019 to september 2020. *bioRxiv*, 2019. doi: 10.1101/780627. URL https://www.biorxiv.org/content/early/2019/09/30/780627.

Edward A Belongia, Melissa D Simpson, Jennifer P King, Maria E Sundaram, Nicholas S Kelley, Michael T Osterholm, and Huong Q McLean. Variable influenza vaccine effectiveness by subtype: a systematic review and meta-analysis of test-negative design studies. *The Lancet Infectious Diseases*, 16(8):942–951, August 2016. ISSN 14733099. doi: 10.1016/S1473-3099(16)00129-8. URL https://linkinghub.elsevier.com/retrieve/pii/S1473309916001298.

Samir Bhatt, Edward C Holmes, and Oliver G Pybus. The genomic rate of molecular adaptation of the human influenza A virus. *Molecular Biology and Evolution*, 28(9):2443, 2011.

Allison Black, Duncan R. MacCannell, Thomas R. Sibley, and Trevor Bedford. Ten recommendations for supporting open pathogen genomic analysis in public health. *Nature Medicine*, 26(6):832–841, Jun 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0935-z. URL https://doi.org/10.1038/s41591-020-0935-z.

Jesse D Bloom. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics*, 16(1):1, 2015. doi: 10.1186/s12859-015-0590-4.

Maciej F Boni, Yang Zhou, Jeffery K Taubenberger, and Edward C Holmes. Homologous recombination is very rare or absent in human influenza A virus. *Journal of Virology*, 82 (10):4807–4811, 2008.

Eva Böttcher, Tatyana Matrosovich, Michaela Beyerle, Hans-Dieter Klenk, Wolfgang Garten, and Mikhail Matrosovich. Proteolytic activation of influenza viruses by serine proteases TMPRSS2 and HAT from human airway epithelium. *Journal of Virology*, 80:9896–9898, 2006.

E Böttcher-Friebertshäuser,, C Freuer, F Sielaff, S Schmidt, M Eickmann, J Uhlendorff, T Steinmetzer, H Klenk, and W Garten. Cleavage of influenza virus hemagglutinin by airway proteases TMPRSS2 and HAT differs in subcellular localization and susceptibility to protease inhibitors. *Journal of Virology*, 11:5605–5614, 2010.

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Christopher B Brooke, William L Ince, Jens Wrammert, Rafi Ahmed, Patrick C Wilson, Jack R Bennink, and Jonathan W Yewdell. Most influenza A virions fail to express at least one essential viral protein. *Journal of Virology*, 87(6):3155–3162, 2013.

Barry C Buckland. The development and manufacture of influenza vaccines. *Human Vaccines & Immunotherapeutics*, 11(6):1357–1360, 2015. doi: 10.1080/21645515.2015.1026497. URL https://doi.org/10.1080/21645515.2015.1026497. PMID: 25844949.

R M Bush, C A Bender, K Subbarao, N J Cox, and W M Fitch. Predicting the evolution of human influenza A. *Science*, 286(5446):1921–1925, December 1999.

Benjamin S Chambers, Kaela Parkhouse, Ted M Ross, Kevin Alby, and Scott E Hensley. Identification of hemagglutinin residues responsible for H3N2 antigenic drift during the 2014-2015 influenza season. *CellReports*, 12(1):1–6, July 2015.

Yao-Qing Chen, Teddy John Wohlbold, Nai-Ying Zheng, Min Huang, Yunping Huang, Karlynn E Neu, Jiwon Lee, Hongquan Wan, Karla Thatcher Rojas, Ericka Kirkpatrick, Carole Henry, Anna-Karin E Palm, Christopher T Stamper, Linda Yu-Ling Lan, David J Topham, John Treanor, Jens Wrammert, Rafi Ahmed, Maryna C Eichelberger, George Georgiou, Florian Krammer, and Patrick C Wilson. Influenza infection in humans induces broadly cross-reactive and protective neuraminidase-reactive antibodies. *Cell*, 173(2): 417–429.e10, April 2018.

Sarah Cobey, Sigrid Gouma, Kaela Parkhouse, Benjamin S Chambers, Hildegund C Ertl, Kenneth E Schmader, Rebecca A Halpin, Xudong Lin, Timothy B Stockwell, Suman R Das, Emily Landon, Vera Tesic, Ilan Youngster, Benjamin A Pinsky, David E Wentworth, Scott E Hensley, and Yonatan H Grad. Poor immunogenicity, not vaccine strain egg adaptation, may explain the low H3N2 influenza vaccine effectiveness in 2012–2013. *Clinical Infectious Diseases*, 67(3):327–333, 02 2018. ISSN 1058-4838. doi: 10.1093/cid/ciy097. URL `https://doi.org/10.1093/cid/ciy097`.

RS Daniels, JC Downie, AJ Hay, M Knossow, JJ Skehel, ML Wang, and DC Wiley. Fusion mutants of the influenza virus hemagglutinin glycoprotein. *Cell*, 40(2):431–439, 1985.

Michael B Doud and Jesse D Bloom. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses*, 8(6), June 2016.

Michael B Doud, Orr Ashenberg, and Jesse D Bloom. Site-specific amino acid preferences are

mostly conserved in two closely related protein homologs. *Mol. Biol. Evol.*, 32:2944–2960, 2015.

Michael B Doud, Scott E Hensley, and Jesse D Bloom. Complete mapping of viral escape from neutralizing antibodies. *PLoS Pathog.*, 13(3):e1006271, March 2017.

Michael B Doud, Juhye M Lee, and Jesse D Bloom. How single mutations affect viral escape from broad and narrow antibodies to H1 influenza hemagglutinin. *Nature Communications*, DOI 10.1038/s41467-018-03665-3, 2018.

Daniel Esposito, Jochen Weile, Jay Shendure, Lea M Starita, Anthony T Papenfuss, Frederick P Roth, Douglas M Fowler, and Alan F Rubin. Mavedb: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology*, 20(1):1–11, 2019. doi: 10.1186/s13059-019-1845-6.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996. URL `http://dl.acm.org/citation.cfm?id=3001460.3001507`.

Warren J Ewens. *Mathematical population genetics 1: theoretical introduction*. Springer Science & Business Media, 2012.

Walter M Fitch, Robin M Bush, Catherine A Bender, and Nancy J Cox. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl. Acad. Sci. USA*, 94(15): 7712–7718, 1997.

Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8):801–807, August 2014. doi: 10.1038/nmeth.3027.

Sylvain Gandon, Troy Day, C Jessica E Metcalf, and Bryan T Grenfell. Forecasting epidemiological and evolutionary dynamics of infectious diseases. *Trends Ecol. Evol. (Amst.)*, 31 (10):776–788, October 2016.

Hannah Gelman, Jennifer N Dines, Jonathan Berg, Alice H Berger, Sarah Brnich, Fuki M Hisama, Richard G James, Alan F Rubin, Jay Shendure, Brian Shirts, et al. Recommendations for the collection and use of multiplexed functional data for clinical variant interpretation. *Genome Medicine*, 11(1):85, 2019. doi: 10.1186/s13073-019-0698-7.

Lizhi Ian Gong, Marc A Suchard, and Jesse D Bloom. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife*, 2:e00631, 2013.

Sigrid Gouma, Madison Weirick, and Scott E. Hensley. Antigenic assessment of the H3N2 component of the 2019-2020 Northern Hemisphere influenza vaccine. *Nature Communications*, 11(1):2445, May 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-16183-y. URL `https://www.nature.com/articles/s41467-020-16183-y`. Number: 1 Publisher: Nature Publishing Group.

Hugh K Haddox, Adam S Dingens, and Jesse D Bloom. Experimental estimation of the effects of all amino-acid mutations to HIV's envelope protein on viral replication in cell culture. *PLoS Pathogens*, 12(12):e1006114, 2016.

Hugh K Haddox, Adam S Dingens, Sarah K Hilton, Julie Overbaugh, and Jesse D Bloom. Mapping mutational effects along the evolutionary landscape of HIV envelope. *eLife*, 7: e34420, 2018.

J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, page bty407, May 2018. doi: 10.1093/bioinformatics/bty407. URL `http://dx.doi.org/10.1093/bioinformatics/bty407`.

Michael J Harms and Joseph W Thornton. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature*, 512(7513):203–207, 2014.

William T Harvey, Donald J Benton, Victoria Gregory, James PJ Hall, Rodney S Daniels, Trevor Bedford, Daniel T Haydon, Alan J Hay, John W McCauley, and Richard Reeve. Identification of low-and high-impact hemagglutinin amino acid substitutions that drive antigenic drift of influenza A (H1N1) viruses. *PLoS Pathogens*, 12(4):e1005526, 2016.

Sarah K Hilton, Michael B Doud, and Jesse D Bloom. phydms: Software for phylogenetic analyses informed by deep mutational scanning. *PeerJ*, 5:e3657, 2017.

George K Hirst. Studies of antigenic differences among strains of influenza A by means of red cell agglutination. *J Exp Med*, 78(5):407–423, 1943.

John Holland, Katherine Spindler, Frank Horodyski, Elizabeth Grabau, Stuart Nichol, and Scott VandePol. Rapid evolution of RNA genomes. *Science*, 215(4540):1577–1585, 1982.

J. Huddleston, J. R. Barnes, T. Rowe, X. Xu, R. Kondor, D. E. Wentworth, L. Whittaker, B. Ermetal, R. S. Daniels, J. W. McCauley, S. Fujisaki, K. Nakamura, N. Kishida, S. Watanabe, H. Hasegawa, I. Barr, K. Subbarao, P. Barrat-Charlaix, R. A. Neher, and T. Bedford. Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. *eLife*, 9:e60067, Sep 2020. ISSN 2050-084X. doi: 10.7554/eLife.60067. URL `https://doi.org/10.7554/eLife.60067`.

Abbas Jariani, Christopher Warth, Koen Deforche, Pieter Libin, Alexei J Drummond, Andrew Rambaut, Frederick A Matsen IV, and Kristof Theys. SANTA-SIM: simulating viral sequence evolution dynamics under selection and recombination. *Virus Evolution*, 5 (1), March 2019.

Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.

Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 07 2002. ISSN 0305-1048. doi: 10.1093/nar/gkf436. URL https://doi.org/10.1093/nar/gkf436.

Jacqueline M Katz, Kathy Hancock, and Xiyan Xu. Serologic assays for influenza surveillance, diagnosis and vaccine evaluation. *Expert Review of Anti-infective Therapy*, 9(6):669–683, 2011. doi: 10.1586/eri.11.51. URL https://doi.org/10.1586/eri.11.51. PMID: 21692672.

Björn F Koel, David F Burke, Theo M Bestebroer, Stefan van der Vliet, Gerben C M Zondag, Gaby Vervaet, Eugene Skepner, Nicola S Lewis, Monique I J Spronken, Colin A Russell, Mikhail Y Eropkin, Aeron C Hurt, Ian G Barr, Jan C de Jong, Guus F Rimmelzwaan, Albert D M E Osterhaus, Ron A M Fouchier, and Derek J Smith. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science*, 342(6161):976–979, November 2013.

Katia Koelle and David A Rasmussen. The effects of a deleterious mutation load on patterns of influenza A/H3N2's antigenic evolution in humans. *Elife*, 4:e07361, September 2015.

Katia Koelle, Sarah Cobey, Bryan Grenfell, and Mercedes Pascual. Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science*, 314(5807): 1898–1903, December 2006.

Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, August 2012.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 957–966. JMLR.org, 2015. URL http://dl.acm.org/citation.cfm?id=3045118.3045221.

Michael Lässig, Ville Mustonen, and Aleksandra M Walczak. Predicting evolution. *Nat Ecol Evol*, 1(3):77, February 2017.

Adam S Lauring and Raul Andino. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathogens*, 6(7):e1001005, 2010.

Hong Kai Lee, Julian Wei-Tze Tang, Debra Han-Lin Kong, Tze Ping Loh, Donald Kok-Leong Chiang, Tommy Tsan-Yuk Lam, and Evelyn Siew-Chuan Koay. Comparison of mutation patterns in full-genome A/H3N2 influenza sequences obtained directly from clinical samples and the same samples after a single MDCK passage. *PLoS One*, 8(11):e79252, 2013.

Juhye M. Lee, John Huddleston, Michael B. Doud, Kathryn A. Hooper, Nicholas C. Wu, Trevor Bedford, and Jesse D. Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proceedings of the National Academy of Sciences*, 115(35):E8276–E8285, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1806133115. URL `http://www.pnas.org/content/115/35/E8276`.

Juhye M Lee, Rachel Eguia, Seth J Zost, Saket Choudhary, Patrick C Wilson, Trevor Bedford, Terry Stevens-Ayers, Michael Boeckh, Aeron C Hurt, Seema S Lakdawala, Scott E Hensley, and Jesse D Bloom. Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. *Elife*, 8, August 2019.

Chengjun Li, Masato Hatta, David F Burke, Jihui Ping, Ying Zhang, Makoto Ozawa, Andrew S Taft, Subash C Das, Anthony P Hanson, Jiasheng Song, et al. Selection of antigenically advanced variants of seasonal influenza viruses. *Nature Microbiology*, 1:16058, 2016.

Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 09 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr509. URL `https://doi.org/10.1093/bioinformatics/btr509`.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 06 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp352. URL `https://doi.org/10.1093/bioinformatics/btp352`.

Marta Łuksza. Personal Communication, 2020.

Marta Łuksza and Michael Lässig. A predictive fitness model for influenza. *Nature*, 507 (7490):57–61, March 2014.

Jock Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141, April 1986. ISSN 0730-0301. doi: 10.1145/22949.22950. URL `https://doi.org/10.1145/22949.22950`.

Nicolle Marshall, Lalita Priyamvada, Zachary Ende, John Steel, and Anice C Lowen. Influenza virus reassortment occurs with high frequency in the absence of segment mismatch. *PLoS Pathogens*, 9(6):e1003421, 2013.

Mikhail Matrosovich, Tatyana Matrosovich, Jackie Carr, Noel A Roberts, and Hans-Dieter Klenk. Overexpression of the α-2, 6-sialyltransferase in mdck cells increases influenza virus sensitivity to neuraminidase inhibitors. *Journal of virology*, 77(15):8418–8425, 2003.

Claire D McWhite, Austin G Meyer, and Claus O Wilke. Sequence amplification via cell passaging creates spurious signals of positive adaptation in influenza virus H3N2 hemagglutinin. *Virus Evol*, 2(2), July 2016.

Matthew J Memoli, Brett W Jagger, Vivien G Dugan, Li Qi, Jadon P Jackson, and Jeffery K Taubenberger. Recent human influenza A/H3N2 virus evolution driven by novel selection factors in addition to antigenic drift. *Journal of Infectious Diseases*, 200(8):1232–1241, 2009.

Dylan H Morris, Katelyn M Gostic, Simone Pompei, Trevor Bedford, Marta Łuksza, Richard A Neher, Bryan T Grenfell, Michael Lässig, and John W McCauley. Predictive modeling of influenza shows the promise of applied evolutionary biology. *Trends Microbiol.*, October 2017.

Chandrasekhar Natarajan, Noriko Inoguchi, Roy E Weber, Angela Fago, Hideaki Moriyama, and Jay F Storz. Epistasis among adaptive mutations in deer mouse hemoglobin. *Science*, 340(6138):1324–1327, 2013.

Richard A. Neher. Genetic draft, selective interference, and population genetics of rapid adaptation. *Annual Review of Ecology, Evolution, and Systematics*, 44(1):195–215, 2013. doi: 10.1146/annurev-ecolsys-110512-135920.

Richard A Neher and Trevor Bedford. nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*, 31(21):3546–3548, November 2015.

Richard A Neher, Colin A Russell, and Boris I Shraiman. Predicting evolution from the shape of genealogical trees. *Elife*, 3:e03568, November 2014.

Richard A Neher, Trevor Bedford, Rodney S Daniels, Colin A Russell, and Boris I Shraiman. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proc. Natl. Acad. Sci. U.S.A.*, 113(12):E1701–9, March 2016.

Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, 11 2014. ISSN 0737-4038. doi: 10.1093/molbev/msu300. URL `https://doi.org/10.1093/molbev/msu300`.

Y Okuno, K Tanaka, K Baba, A Maeda, N Kunita, and S Ueda. Rapid focus reduction neutralization test of influenza A and B viruses in microtiter system. *J. Clin. Microbiol.*, 28(6):1308–1313, June 1990.

Velislava N. Petrova and Colin A. Russell. The evolution of seasonal influenza viruses. *Nature Reviews Microbiology*, 16(1):47–60, Jan 2018. ISSN 1740-1534. doi: 10.1038/nrmicro.2017.118. URL https://doi.org/10.1038/nrmicro.2017.118.

David D Pollock, Grant Thiltgen, and Richard A Goldstein. Amino acid coevolution induces an evolutionary stokes shift. *Proc. Natl. Acad. Sci. USA*, 109(21):E1352–E1359, 2012.

David Posada and Thomas R Buckley. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793–808, 2004.

Barney I Potter, Rebecca Kondor, James Hadfield, John Huddleston, John Barnes, Thomas Rowe, Lizheng Guo, Xiyan Xu, Richard A Neher, Trevor Bedford, and David E Wentworth. Evolution and rapid spread of a reassortant A(H3N2) virus that predominated the 2017–2018 influenza season. *Virus Evolution*, 5(2), 12 2019. ISSN 2057-1577. doi: 10.1093/ve/vez046. URL https://doi.org/10.1093/ve/vez046. vez046.

Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. Fasttree 2 – approximately maximum-likelihood trees for large alignments. *PLOS ONE*, 5(3):1–10, 03 2010. doi: 10.1371/journal.pone.0009490. URL https://doi.org/10.1371/journal.pone.0009490.

Oliver G Pybus, Andrew Rambaut, Robert Belshaw, Robert P Freckleton, Alexei J Drummond, and Edward C Holmes. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. *Molecular Biology and Evolution*, 24(3):845–852, 2007.

Hangfei Qi, Nicholas C Wu, Yushen Du, Ting-Ting Wu, and Ren Sun. High-resolution genetic profile of viral genomes: why it matters. *Current Opinion in Virology*, 14:62–70, 2015.

Jayna Raghwani, Robin N Thompson, and Katia Koelle. Selection on non-antigenic gene segments of seasonal influenza A virus and its impact on adaptive evolution. *Virus Evolution*, 3(2), 2017.

Andrew Rambaut, Oliver G Pybus, Martha I Nelson, Cecile Viboud, Jeffery K Taubenberger, and Edward C Holmes. The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453(7195):615–619, April 2008.

Nicholas G Reich, Logan C Brooks, Spencer J Fox, Sasikiran Kandula, Craig J McGowan, Evan Moore, Dave Osthus, Evan L Ray, Abhinav Tushar, Teresa K Yamana, Matthew Biggerstaff, Michael A Johansson, Roni Rosenfeld, and Jeffrey Shaman. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc. Natl. Acad. Sci. U.S.A.*, 116(8):3146–3154, February 2019.

Nicholas G Reich, Matthew Cornell, Evan L Ray, Katie House, and Khoa Le. The Zoltar forecast archive: a tool to facilitate standardization and storage of interdisciplinary prediction research, 2020.

Alan F Rubin, Hannah Gelman, Nathan Lucas, Sandra M Bajjalieh, Anthony T Papenfuss, Terence P Speed, and Douglas M Fowler. A statistical framework for analyzing deep mutational scanning data. *Genome Biology*, 18(1):150, 2017. doi: 10.1186/s13059-017-1272-5.

Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66, Jan 1998. doi: 10.1109/ICCV.1998.710701.

C A Russell, T C Jones, I G Barr, N J Cox, R J Garten, V Gregory, I D Gust, A W Hampson, A J Hay, A C Hurt, J C de Jong, A Kelso, A I Klimov, T Kageyama, N Komadina, A S Lapedes, Y P Lin, A Mosterin, M Obuchi, T Odagiri, A D M E Osterhaus, G F Rimmelzwaan, M W Shaw, E Skepner, K Stohr, M Tashiro, R A M Fouchier, and D J Smith. The global circulation of seasonal influenza A (H3N2) viruses. *Science*, 320(5874): 340–346, April 2008.

Pavel Sagulenko, Vadim Puller, and Richard A Neher. TreeTime: Maximum-likelihood

phylodynamic analysis. *Virus Evolution*, 4(1), 01 2018. ISSN 2057-1577. doi: 10.1093/ve/vex042. URL https://doi.org/10.1093/ve/vex042.

Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

Premal Shah, David M McCandlish, and Joshua B Plotkin. Contingency and entrenchment in protein evolution under purifying selection. *Proceedings of the National Academy of Sciences*, 112(25):E3226–E3235, 2015.

Arthur Chun-Chieh Shih, Tzu-Chang Hsiao, Mei-Shang Ho, and Wen-Hsiung Li. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proceedings of the National Academy of Sciences*, 104(15):6283–6288, April 2007.

Yuelong Shu and John McCauley. Gisaid: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, 22(13), 2017.

Derek J Smith, Alan S Lapedes, Jan C de Jong, Theo M Bestebroer, Guus F Rimmelzwaan, Albert D M E Osterhaus, and Ron A M Fouchier. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376, July 2004.

Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 01 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu033. URL https://doi.org/10.1093/bioinformatics/btu033.

Lea M Starita, Nadav Ahituv, Maitreya J Dunham, Jacob O Kitzman, Frederick P Roth, Georg Seelig, Jay Shendure, and Douglas M Fowler. Variant interpretation: functional assays to the rescue. *The American Journal of Human Genetics*, 101(3):315–325, 2017. doi: 10.1016/j.ajhg.2017.07.014.

Tyler N Starr and Joseph W Thornton. Epistasis in protein evolution. *Protein Science*, 25 (7):1204–1218, 2016.

Tyler N Starr, Lora K Picton, and Joseph W Thornton. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature*, 549(7672):409–413, 2017.

L Steinbrück, T R Klingen, and A C McHardy. Computational prediction of vaccine strains for human influenza A (H3N2) viruses. *J. Virol.*, 88(20):12123–12132, October 2014.

DA Steinhauer and JJ Holland. Rapid evolution of RNA viruses. *Annual Reviews in Microbiology*, 41(1):409–431, 1987.

Natalja Strelkowa and Michael Lässig. Clonal interference in the evolution of influenza. *Genetics*, 192(2):671–682, 2012.

Hailiang Sun, Jialiang Yang, Tong Zhang, Li-Ping Long, Kun Jia, Guohua Yang, Richard J Webby, and Xiu-Feng Wan. Using sequence data to infer the antigenicity of influenza virus. *MBio*, 4(4):e00230–13, 2013.

Xiangjie Sun, V Tse Longping, A Damon Ferguson, and Gary R Whittaker. Modifications to the hemagglutinin cleavage site control the virulence of a neurotropic H1N1 influenza virus. *Journal of virology*, 84(17):8683–8690, 2010.

Bargavi Thyagarajan and Jesse D Bloom. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *Elife*, 3:e03300, July 2014.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL `http://www.jmlr.org/papers/v9/vandermaaten08a.html`.

Jacob VanderPlas, Brian E. Granger, Jeffrey Heer, Dominik Moritz, Kanit Wongsuphasawat, Arvind Satyanarayan, Eitan Lees, Ilia Timofeev, Ben Welsh, and Scott Sievert. Altair:

Interactive statistical visualizations for python. *Journal of Open Source Software*, 3(32): 1057, 2018. doi: 10.21105/joss.01057. URL https://doi.org/10.21105/joss.01057.

Mara Villa and Michael Lässig. Fitness cost of reassortment in human influenza. *PLoS Pathog.*, 13(11):e1006685, November 2017.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17 (3):261–272, March 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0686-2. URL https://www.nature.com/articles/s41592-019-0686-2. Number: 3 Publisher: Nature Publishing Group.

WHO. Recommended viruses for influenza vaccines for use in the 2010-2011 northern hemisphere influenza season. *http://www.who.int/influenza/vaccines/virus/recommendations/201002_Recommendation.pdf?ua=1*, 2010.

WHO. Recommended composition of influenza virus vaccines for use in the 2011-2012 northern hemisphere influenza season. *http://www.who.int/influenza/vaccines/2011_02_recommendation.pdf?ua=1*, 2011.

D C Wiley, I A Wilson, and J J Skehel. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature*, 289(5796):373–378, January 1981.

Yuri I Wolf, Cecile Viboud, Edward C Holmes, Eugene V Koonin, and David J Lipman. Long

intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biol. Direct*, 1(1):34, October 2006.

John M Wood, Diane Major, Alan Heath, Robert W Newman, Katja Höschler, Iain Stephenson, Tristan Clark, Jacqueline M Katz, and Maria C Zambon. Reproducibility of serology assays for pandemic influenza H1N1: Collaborative study to evaluate a candidate WHO International Standard. *Vaccine*, 30(2):210–217, January 2012.

World Health Organization. Influenza fact sheet. `http://www.who.int/mediacentre/factsheets/fs211/en/`, 2009. Accessed: 2017-05-02.

World Health Organization. *Seasonal influenza fact sheet*. Available at `http://www.who.int/mediacentre/factsheets/fs211/en/`, 2014.

Emily E Wrenbeck, Matthew S Faber, and Timothy A Whitehead. Deep sequencing methods for protein engineering and design. *Current Opinion in Structural Biology*, 45:36–44, 2017. doi: 10.1016/j.sbi.2016.11.001.

NC Wu, SJ Zost, AJ Thompson, D Oyen, CM Nycholat, R McBride, JC Paulson, SE Hensley, and IA Wilson. A structural explanation for the low effectiveness of the seasonal influenza H3N2 vaccine. *PLoS Pathogens*, 13(10):e1006682, 2017.

Nicholas C Wu, Arthur P Young, Laith Q Al-Mawsawi, C Anders Olson, Jun Feng, Hangfei Qi, Shu-Hwa Chen, I-Hsuan Lu, Chung-Yen Lin, Robert G Chin, Harding H Luan, Nguyen Nguyen, Stanley F Nelson, Xinmin Li, Ting-Ting Wu, and Ren Sun. High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. *Sci Rep*, 4:4942, May 2014.

Katherine S. Xue, Louise H. Moncla, Trevor Bedford, and Jesse D. Bloom. Within-host evolution of human influenza virus. *Trends in Microbiology*, 26(9):781 – 793, 2018. ISSN 0966-842X. doi: https://doi.org/10.1016/j.tim.2018.02.007. URL `http://www.sciencedirect.com/science/article/pii/S0966842X1830043X`.

Ziheng Yang, Rasmus Nielsen, Nick Goldman, and Anne-Mette Krabbe Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155 (1):431–449, 2000.

Fabio Zanini and Richard A. Neher. FFPopSim: an efficient forward simulation package for the evolution of large populations. *Bioinformatics*, 28(24):3332–3333, 10 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts633. URL `https://doi.org/10.1093/bioinformatics/bts633`.

Seth J. Zost, Kaela Parkhouse, Megan E. Gumina, Kangchon Kim, Sebastian Diaz Perez, Patrick C. Wilson, John J. Treanor, Andrea J. Sant, Sarah Cobey, and Scott E. Hensley. Contemporary H3N2 influenza viruses have a glycosylation site that alters binding of antibodies elicited by egg-adapted vaccine strains. *Proceedings of the National Academy of Sciences*, 114(47):12578–12583, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1712377114. URL `https://www.pnas.org/content/114/47/12578`.

Appendix A

AUGUR: A BIOINFORMATICS TOOLKIT FOR PHYLOGENETIC ANALYSES OF HUMAN PATHOGENS

A.1 Summary and statement of need

The analysis of human pathogens requires a diverse collection of bioinformatics tools. These tools include standard genomic and phylogenetic software and custom software developed to handle the relatively numerous and short genomes of viruses and bacteria. Researchers increasingly depend on the outputs of these tools to infer transmission dynamics of human diseases and make actionable recommendations to public health officials [Black et al., 2020, Bedford et al., 2020]. Under these circumstances, bioinformatics tools must scale rapidly with the number of disease samples to enable real-time analyses of pathogen evolution. To meet these needs, we developed Augur, a bioinformatics toolkit designed for phylogenetic analyses of human pathogens.

Augur originally existed as an internal component of the nextflu [Neher and Bedford, 2015] and Nextstrain [Hadfield et al., 2018] applications. In its original form, Augur consisted of two monolithic Python scripts, "prepare" and "process", that performed most operations in memory. These scripts prepared a subset of pathogen sequences and metadata and then processed those data to produce an annotated phylogeny that could be viewed at `nextstrain.org`. The original nextflu scripts only supported seasonal influenza viruses. When nextflu was replaced with Nextstrain and expanded to support multiple viral and bacterial pathogens, each pathogen received its own copy of the original scripts. The resulting redundancy of these large scripts complicated efforts to debug analyses, add new features for all pathogens, and add support for new pathogens. Critically, this software architecture led

to long-lived, divergent branches of untested code in version control that Nextstrain team members could not confidently merge without potentially breaking existing analyses.

A.2 Implementation

To address these issues, we refactored the original Augur scripts into a toolkit of individual subcommands wrapped by a single command line executable, `augur`. With this approach, we followed the pattern established by samtools [Li et al., 2009] and bcftools [Li, 2011] where subcommands perform single, tightly-scoped tasks (e.g., "view", "sort", "merge", etc.) that can be chained together in bioinformatics pipelines. We migrated or rewrote the existing functionality of the original Augur scripts into appropriate corresponding Augur subcommands. To enable interoperability with existing bioinformatics tools, we designed subcommands to accept inputs and produce outputs in standard bioinformatics file formats wherever possible. For example, we represented all raw sequence data in FASTA format, alignments in either FASTA or VCF format, and phylogenies in Newick format. To handle the common case where a standard file format could not represent some or all of the outputs produced by an Augur command, we implemented a lightweight JSON schema to store the remaining data. The "node data" JSON format represents one such Augur-specific file format that supports arbitrary annotations of phylogenies indexed by the name assigned to internal nodes or tips. To provide a standard interface for our own analyses, we also designed several Augur subcommands to wrap existing bioinformatics tools including `augur align` (mafft [Katoh et al., 2002]), `augur tree` (FastTree [Price et al., 2010], RAxML [Stamatakis, 2014], and IQ-TREE [Nguyen et al., 2014]), and `augur refine` (TreeTime [Sagulenko et al., 2018]).

By implementing the core components of Augur as a command line tool, we were able to rewrite our existing pathogen analyses as straightforward bioinformatics workflows using existing workflow management software like Snakemake [Köster and Rahmann, 2012]. Most pathogen workflows begin with user-curated sequences in a FASTA file (e.g., `sequences.fasta`) and metadata describing each sequence in a tab-delimited text file (e.g., `metadata.tsv`). Users

can apply a series of Augur commands and other standard bioinformatics tools to these files to create annotated phylogenies that can be viewed in Auspice, the web application that serves Nextstrain (Figure A.1). This approach allows users to leverage the distributed computing abilities of workflow managers to run multiple steps of the workflow in parallel and also run individual commands that support multiprocessing in parallel.

The modular Augur interface has enabled a proliferation of phylogenetic and genomic epidemiological analyses by academic researchers, public health laboratories, and private companies. Most recently, these tools have supported the real-time tracking of SARS-CoV-2 evolution at global and local scales. This success has attracted contributions from the open source community that have allowed us to improve Augur's functionality, documentation, and test coverage. Augur can be installed from PyPI (nextstrain-augur) and Bioconda (augur). See the full documentation for more details about how to use or contribute to development of Augur.

A.3 Figures

A.4 Acknowledgments

Thank you to all of the open source community members who have contributed to Augur. Thank you to Dan Fornika from BCCDC Public Health Laboratory for creating the first conda recipe for Augur in Bioconda.

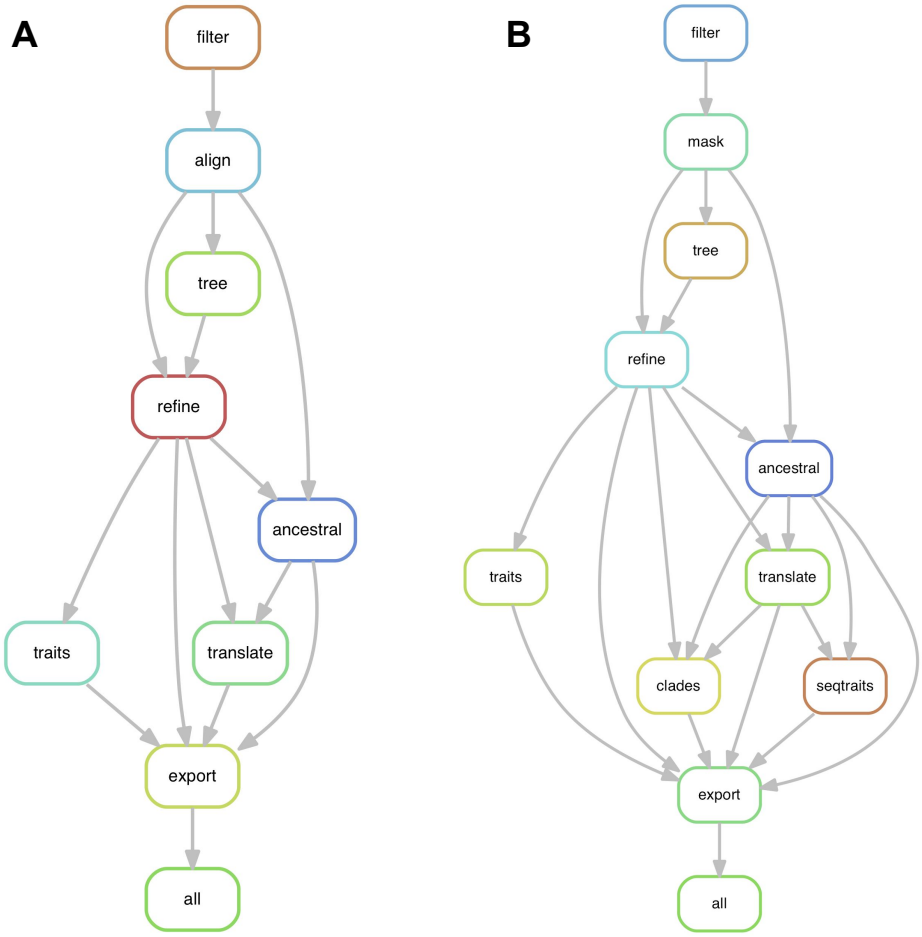

Figure A.1: **Example workflows composed with Snakemake from Augur commands for A) Zika virus and B) tuberculosis. Each node in the workflow graph represents an Augur command than performs a specific part of the analysis (e.g., aligning sequences, building a tree, etc.). A typical workflow starts by filtering sequences and metadata to a desired subset for analysis followed by inference of a phylogeny, annotation of that phylogeny, and export of the annotated phylogeny to a JSON that can be viewed on Nextstrain. Workflows for viral (A) and bacterial (B) pathogens follow a similar structure but also support custom pathogen-specific steps. Multiple outgoing edges from a single node represent opportunities to run the workflow in parallel. See the full workflows at** `https://github.com/nextstrain/zika-tutorial` **and** `https://github.com/nextstrain/tb`.