EVOLUTIONARY BIOLOGY

Phylogenetic identification of influenza virus candidates for seasonal vaccines

Maryam Hayati[1]†, Benjamin Sobkowiak[2]†, Jessica E. Stockdale[2], Caroline Colijn[2]*

The seasonal influenza (flu) vaccine is designed to protect against those influenza viruses predicted to circulate during the upcoming flu season, but identifying which viruses are likely to circulate is challenging. We use features from phylogenetic trees reconstructed from hemagglutinin (HA) and neuraminidase (NA) sequences, together with a support vector machine, to predict future circulation. We obtain accuracies of 0.75 to 0.89 (AUC 0.83 to 0.91) over 2016–2020. We explore ways to select potential candidates for a seasonal vaccine and find that the machine learning model has a moderate ability to select strains that are close to future populations. However, consensus sequences among the most recent 3 years also do well at this task. We identify similar candidate strains to those proposed by the World Health Organization, suggesting that this approach can help inform vaccine strain selection.

INTRODUCTION

Seasonal influenza (flu) epidemics cause substantial serious illness and mortality every year despite preventive efforts. The extent of morbidity and mortality in a given year reflects the degree of genetic drift or shift in the dominant strain of influenza virus and the efficacy and coverage of vaccination. Influenza A virus has two glycoprotein spikes on its virion surface: hemagglutinin (HA) and neuraminidase (NA), which have opposite functions (*1*). HA binds to oligosaccharides containing terminal sialic acid (SA) and initiates the infectious cycle. NA removes terminal SA from oligosaccharides and completes the infectious cycle (*1*). Studies have shown that the interaction between receptor binding and receptor destroying is important in viral transmission (*1*, *2*). HA contains epitopes that are vital to induce B cells to produce neutralizing antibodies. Therefore, the epitope sites on the surface of HA are a determining factor affecting viral mutation and recombination mechanisms (*3*, *4*). Mutations that occur on the surface of HA also allow influenza viruses to evade host population immunity, resulting in seasonal flu epidemics. NA is the second most abundant glycoprotein on the surface of the virus and has a crucial role in viral infection by binding to SA receptors. SA moieties trigger the release of nascent virions and facilitate the spread of influenza viruses (*5*–*7*). Together, these proteins are used to classify influenza into its subtypes, e.g., H3N2.

Vaccination is an important approach in controlling influenza, limiting its potentially serious complications (*8*) and reducing the severity of influenza-associated illness (*9*). For vaccination to be successful, the specific viruses included in seasonal flu vaccines need to be similar to those influenza viruses that will circulate in the upcoming season. Seasonal flu vaccines are not always effective, and this effectiveness varies with several factors, including the patient's medical history and age, the current types of circulating influenza viruses, and the degree of similarity between circulating viruses and those included in the vaccine (*10*, *11*). Recent studies have shown that the effectiveness of flu vaccines in reducing the risk of becoming infected during the influenza season has ranged from 40 to 60% across all ages (*12*–*17*). During the 2018–2019 influenza season, overall adjusted vaccine effectiveness against all influenza virus infection associated with medically attended acute respiratory illness was 47% [95% confidence interval (CI) = 34 to 57%] (*12*). In general, current vaccines are more effective against influenza B and influenza A(H1N1) than influenza A(H3N2) viruses. This is due in part to the fact that genetic changes occur in influenza A(H3N2) viruses more frequently than in other types (*10*), but factors such as differences in glycosylation, immune imprinting, preexisting immunity in a population, and additional strain-specific factors could also moderate effectiveness. Vaccination has been shown to decrease the number of influenza-related illnesses, hospitalizations, and deaths substantially each year; however, there remains scope for improvement in both the immunogenicity and efficacy of flu vaccines.

Phylogenetic trees capture patterns of descent among groups of organisms and should, in principle, capture information about fitness (*18*, *19*). The tree's branch lengths reflect either the time or the genetic distance between branching events, while its shape specifies the patterns of relatedness, ancestry, and descent among the organisms (*18*, *20*, *21*). Phylogenetic trees have become essential tools in phylodynamics and infectious disease: They are used to estimate the basic reproduction number (*22*), parameters of transmission models (*23*), and aspects of underlying contact networks (*24*–*27*) to predict the short-term growth and fitness of influenza virus trees (*18*, *19*, *21*) and, in densely sampled datasets, even to infer person-to-person transmission events and timing (*28*–*31*).

Recent advances in genome sequencing technology mean that it is now feasible to collate large datasets of influenza strains collected over a long time frame. This, together with development of computational resources for reconstructing large phylogenetic trees, enables us to study how a population of influenza viruses changes over time. In this work, we use topological features of influenza trees, together with machine learning tools, to find candidate strains for H3N2 influenza vaccines. We use HA and NA sequences from 1980 to each of February 2016, 2017, 2018, 2019, and 2020 to reconstruct influenza virus phylogenetic trees and propose

[1]School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada. [2]Department of Mathematics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada.
*Corresponding author. Email: ccolijn@sfu.ca
†These authors contributed equally to this work.

candidate strains for the following year's vaccine. For each influenza strain (here, each taxon or tip), we assign a set of features extracted from the shape of the ancestral subtrees of that strain in the reconstructed HA and NA trees. We then use binary classifiers to classify the strains as either successful or unsuccessful according to whether the near-term ancestors of the strain gave rise to a sufficiently large number of taxa in the coming 4 to 5 years. Using this approach, we predict which recently observed taxa are likely to be successful strains. We choose our final candidates for inclusion in the following year's H3N2 vaccine using genetic distances between the taxa's epitope sites and those of past taxa. Focusing on data up to 2020, we compare our proposed vaccine strains for 2020/2021 to those suggested by the World Health Organization (WHO).

MATERIALS AND METHODS

We introduce a method that incorporates trees reconstructed from both HA and NA gene sequences of influenza virus (H3N2) to predict vaccine candidates for inclusion in the following year's seasonal flu vaccine. Our approach is based on the hypothesis that the fitness (the reproductive rate and capacity of a group of organisms) of each sequence (a tip of an influenza tree) can be measured using the topological properties of the ancestral subtrees that it belongs to.

Definitions

Given a phylogenetic tree T, a tip (also called an external node or leaf) of T is a node of degree one. An internal node of T is any non-leaf node of the tree. A rooted tree is a tree in which a particular internal node, called the root, is distinguished from the others; it is usually considered to be a common ancestor of all other nodes in the tree. In a rooted tree T, the parent of a node i is the node preceding it on the unique path from the root r to the node i; all nodes of T except its root have a parent. A child of a node i is a node whose parent is i. A phylogenetic tree is bifurcating if all its internal nodes have two children. In this work, we use rooted bifurcating phylogenetic trees that are reconstructed from either the HA or NA sequences of influenza virus A strains (by maximum likelihood). A subtree of a tree (in general) is any connected subgraph of the tree. In our rooted trees, each subtree will have a node closest to the tree's root. This is the subtree's ancestor. Here, we use subtrees that consist of all tips that (i) descend from a given subtree ancestor and (ii) whose branch length to that ancestor is less than a given time threshold, α, along with the nodes and edges connecting these tips. If this subtree has a sufficient number of tips (above five, in practice) to compute the features that we are interested in (below), then we consider it to be "relevant."

Data and tree reconstruction

We downloaded all HA and NA human H3N2 sequences collected from 1980 to February 2020 from the Global Initiative on Sharing Avian Influenza Data (GISAID) (*32*). From these, we created 10 datasets: containing HA or NA sequences from 1980 to each of February 2016, 2017, 2018, 2019, and 2020 to reconstruct 10 trees (5 HA trees and 5 NA trees, one each per year). We included only HA and NA sequences with lengths of at least 1701 and 1400, respectively. The number of sequences included in each dataset is shown in Fig. 1. We used the standard "augur" pipeline of nextflu (*33*) to align the sequences to the A/Wisconsin/67/2005 reference sequence. We manually removed sequences (<0.1%) with large numbers of insertions/deletions (more than half the sequence length) from the alignment. We used IQ-TREE2 (*34*) to first estimate the best-fitting nucleotide substitution model and then to reconstruct the approximated maximum likelihood tree. We run IQ-TREE with default settings. IQ-TREE2 is appropriate for reconstructing rooted trees. Last, we converted the trees to timed trees using the software "least squares dating" (*35*).

In total, our 2020 dataset included 29,571 tips before March 2017. These tips are used for training and testing. The remaining 3 years of tips (March 2017 to February 2020 and equivalently for the other experiments) are used for the vaccine candidate prediction task: We call these "recent tips." We did not use tips before 2010 for training or testing on the trees reconstructed from sequences up to February 2020 (and we did not use tips before 2009 for the 2019 experiment, and so on) because there were far fewer sequences available in comparison to more recent years. This adds a level of uncertainty in tip response variables ("successful" or not) for tips before 2010. In addition, the pre-2010 tips had fewer relevant subtrees, on average, than the tips from more recent years, so the removal of these from testing and training datasets also helps reduce bias. For the 2020 experiment, after removing the past tips, outliers (25 tips), and duplicated tips (234), we used the remaining 26,418 tips for training and testing our models (and similarly for the other years), as described in the following sections.

We additionally downloaded all 57,339 human H3N2 influenza protein sequences collected between March 2016 and October 2020 from GISAID (*32*). These protein sequences were not used in the tree building or modeling tasks but were used solely for evaluating the success of our vaccine proposal approach (see the "Epitope distance for vaccine proposal" section).

Subtree extraction and features

To compare tips in our influenza trees, we use a set of features defined on subtrees, including both tree shape and patterns in the branch lengths. These topological features are summarized in Table 1. We normalize values of each property to reduce dependence on subtree size. The features associated with each tip are derived from the topological features of all relevant subtrees on the path from that tip to the root of the tree. We call this the "relevant path"; It is uniquely defined for each tip. To find the relevant subtrees in the HA trees, we first extract all subtrees and then trim those tips that occurred more than $\alpha = 3$ years after the ancestral node of each subtree. The relevant subtrees are those trimmed subtrees with at least five tips. We exclude from our datasets any tips that do not have any relevant subtrees on their relevant paths. Among 66,881 HA tips in the dataset ending February 2020, for example, 500 are excluded. The procedure for extracting subtrees and computing features for the NA trees is the same as the HA trees, with the exception of the 3-year cutoff. Because of the different structures of the NA and HA trees, trimming NA subtrees after 3 years results in a much larger proportion of tips with no relevant subtrees on their relevant paths. As a result, we trim the subtrees of the NA trees after 4 years (see the Supplementary Materials for more details). In addition to the topological features, for the H3N2-HA datasets, we also consider a feature derived from the epitope sites of the tips of the subtree. We define the epitope distance between any two sequences as the genetic distance between the epitope sites of the two sequences. Then, for each subtree, we consider the mean epitope distance between the tips of the subtree and

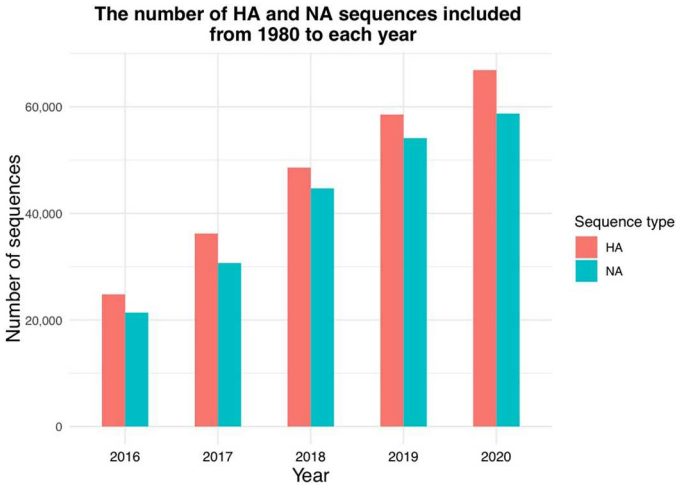

Fig. 1. Volume of data used in each experiment (year). We downloaded all available hemagglutinin (HA) and neuraminidase (NA) human H3N2 sequences collected between 1980 and February 2020 from the GISAID (*32*). These sequences are divided into 10 datasets: Across five experiments ("Year"), we include either HA (red) or NA (blue) sequences with minimum lengths of 1701 and 1400, respectively, collected between 1980 and February of the respective year.

all sequences with dates before the subtree. This is the epitope feature. We use the locations of known antigenic epitopes as mentioned in (*36*), namely, 72 sites in the *HA1* subunit of HA.

Previous studies have found that in fixed-size populations, increasing fitness results in increased asymmetric branching and long terminal branch lengths, and this indicates that fitness leaves traces in genealogical trees (*19*, *21*). Tree structure has been used previously to predict growth using influenza virus trees. Neher and colleagues (*18*) used the local branching index, a measure of the total branch length surrounding a node, in their predictive model. Hayati and colleagues (*19*) used a set of features, including asymmetry, small shape frequencies, measures of local branching, and features derived from network science, that capture global structure of the subtrees in their predictive model. In this work, we expand on this repertoire of tree features by adding weighted network features (*37*), spectral properties of subtrees (*38*), and diversification rate, which is the reciprocal sum of the branches from a tip to the root of a tree (*39*). In total, for each accepted subtree of the HA and NA trees, we compute 39 topological features. We add one epitope feature for subtrees of the HA tree, which results in a total set of 79 features. The feature set therefore incorporates information from both the HA and NA trees, integrating this information at the level of individual strains.

After computing the features of the subtrees of the HA and NA trees, we define the features of each tip based on the features of the subtrees on the tip's relevant path. We calculate the kth feature of a tip t, g_t^k, as a weighted sum of the kth features of t's relevant subtrees

$$g_t^k = \sum_{x \in p} f_x^k \, e^{-d(t,x)/D} \quad (1)$$

where t is a tip, p is the relevant path, x are the ancestors of the relevant subtrees for tip t, f_x is a feature of the subtree descending from x, and $d(t, x)$ is the path length from tip t to node x. D is a scalar parameter with value $D = 5 \times$ (median of tree edge lengths) chosen by tuning. This approach somewhat mirrors the definition of local branching index, which uses a sum of exponentially discounted lengths (*18*). In other words, subtrees that are far from the tip contribute exponentially less to the tip's features than subtrees that are close to the tip. The tip feature computation process is summarized in Fig. 2.

Success and training approach

The success of each tip is defined by the success of its relevant subtrees in both HA and NA trees. We call a subtree of size m successful if its root has a total of more than m tip descendants in the time frame of 4 (HA) or 5 (NA) years from the root of the subtree. To compute the labels of the tips (0 denoting unsuccessful and 1 denoting successful), we first compute the weighted sum of the labels of the relevant subtrees of each tip. We call these values "relevant labels," and the weights for computing the weighted, tip-level labels are the same as those used for computing the features of each tip. We denote the median of the weighted labels among all the tips on the HA and NA trees as μ_{HA} and μ_{NA}, respectively. A tip is defined as successful if its weighted label derived from both the HA and NA trees is greater than both μ_{HA} and μ_{NA}. In other words, for a tip to be labeled as successful, the relevant subtrees for the tip, in both HA and NA, must be such that when appropriately weighted and summed in Eq. 1, it has a success signal from both medians.

Our approach, using subtrees on paths from the tips to the root, induces similar feature sets and labels in sibling and closely related tips. This makes for a challenging training approach, as we cannot

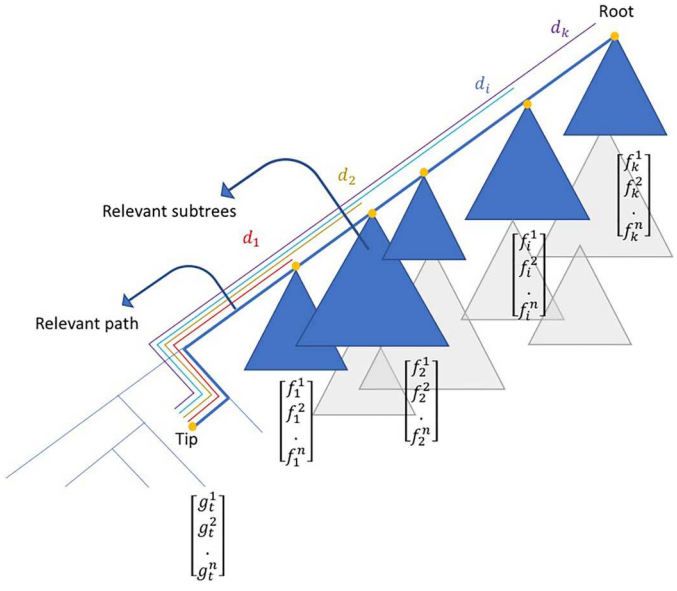

Fig. 2. The procedure for computing the features of each tip. We compute n = 79 features ($f_x^1, f_x^2, ..f_x^n$) for each accepted subtree x. Then, the feature set of a tip t ($g_t^1, g_t^2, ..g_t^n$) is given by the weighted sum of the features of the relevant subtrees. Blue triangles indicate relevant subtrees for the tip marked "tip." Relevant subtrees are located on the path from the tip to the root (this is the relevant path). The weights of the subtrees are reduced according to their distance to the tip, $d(t, x)$. Gray triangles indicate that the full tree continues into the future, and the nodes on the path from a tip t to the root can have descendants that are not in t's relevant subtrees.

Table 1. Brief definition of the tree shape statistics. Here, r_i and s_i are the number of tips of the left and right subtrees of an internal node i. n is the number of tips of a subtree. n_i is the number of nodes at depth i, M_i represents the height of the subtree rooted at an internal node i, and N_i is equal to the depth of node i. A ladder in a tree is a set of consecutive nodes with one tip child. We represent the set of all internal nodes of a tree as $\mathscr{I}$, the set of all tips (or external nodes) as $\mathscr{L}$, and the root of a subtree as r. In "generalized branching next," we chose $m = 2$. Skewness and kurtosis are two measures to describe the degree of asymmetry of a distribution (*63*). The tree shape statistics induced by betweenness centrality, closeness centrality, and eigenvector centrality are defined as the maximum values of each centrality over all the nodes of a tree, and distances are in units of number of edges (without branch lengths). The network science properties were computed in R using the treeCentrality package (*64*), and the tree-wide summaries were primarily obtained using the phyloTop package (*65*).

Name	Description	Reference
Properties from network science		
Betweenness centrality	Maximum number of shortest paths through nodes	(*66*)
Weighted betweenness	As above, but with weighted edges	(*37*)
Closeness centrality	Maximum total distance to all other nodes	(*66*)
Weighted closeness	As above, but with weighted edges	(*37*)
Eigenvector centrality	Maximum value in Perron-Frobenius vector	(*66*)
Weighted eigenvector	As above, but with weighted edges	(*37*)
Diameter	Largest distance between two nodes	(*67*)
WienerIndex	Sum of all distances between two nodes	(*68*)
Mean tips pairwise distance	Average distance between 2 tips	(*19*)
Maximum tips pairwise distance	Maximum distance between 2 tips (with branch lengths)	(*19*)
Spectral properties		
Minimum adjacency	Minimum adjacency matrix eigenvalue >0	(*38*)
Maximum adjacency	Maximum adjacency matrix eigenvalue	(*38*)
Minimum Laplacian	Minimum Laplacian matrix eigenvalue >0	(*38*)
Maximum Laplacian	Maximum Laplacian matrix eigenvalue	(*38*)
Numbers of small configurations		
Cherry number	Number of nodes with two tip children	(*69*)
Normalized Pitchforks	3 × (number of nodes with 3 tip descendants)/n	(*70*)
Tree-wide summaries		
Normalized Colless imbalance	$\frac{1}{n^{3/2}}\sum_{i\in\mathscr{I}\cup\{r\}}\lvert r_i - s_i\rvert$	(*71*)
Normalized Sackin imbalance	$\frac{1}{n^{3/2}}\sum_{i\in\mathscr{L}}N_i$	(*72*)
Normalized maximum height	The maximum height of tips in the tree/(n − 1)	(*65*)
Maximum width	Maximum number of nodes at the same depth	(*73*)
Stairs1	The portion of imbalanced subtrees	(*74*)
Stairs2	The average of $\frac{\min(s_i,r_i)}{\max(s_i,r_i)}$ over all internal nodes	(*74*)
Maximum difference in widths	$\max_i(n_{i+1} - n_i)$	(*73*)
Variance	The variance of internal node depth	(*75*)
I2	$\sum_{\substack{i\in\mathscr{I}\cup\{r\}\\ r_i+s_i>2}}\frac{\lvert r_i-s_i\rvert}{\lvert r_i+s_i-2\rvert}$	(*75*)
B1	$\sum_{i\in\mathscr{I}}M_i^{-1}$	(*75*)
B2	$\sum_{i\in\mathscr{L}}\frac{N_i}{2^{N_i}}$	(*75*)
Normalized average ladder	The mean size of ladders in the tree/(n − 2)	(*65*)
Normalized ILnumber	Number of internal nodes with a single tip child/(n − 2)	(*65*)
Branching speed	The ratio of the number of tips to the height of the tree	(*19*)
Measures from edge length		
Branching next index	Mean of indicator: Does the next branching event descend from this node	(*19*)
Generalized branching next	Number of next two branching events descending from this node	(*19*)
Skewness	The skewness of the internal branch lengths	(*19*)
Kurtosis	The kurtosis of the internal branch lengths	(*19*)
Diversification rate	The reciprocal sum of the branches from a tip to the root of the tree	(*39*)

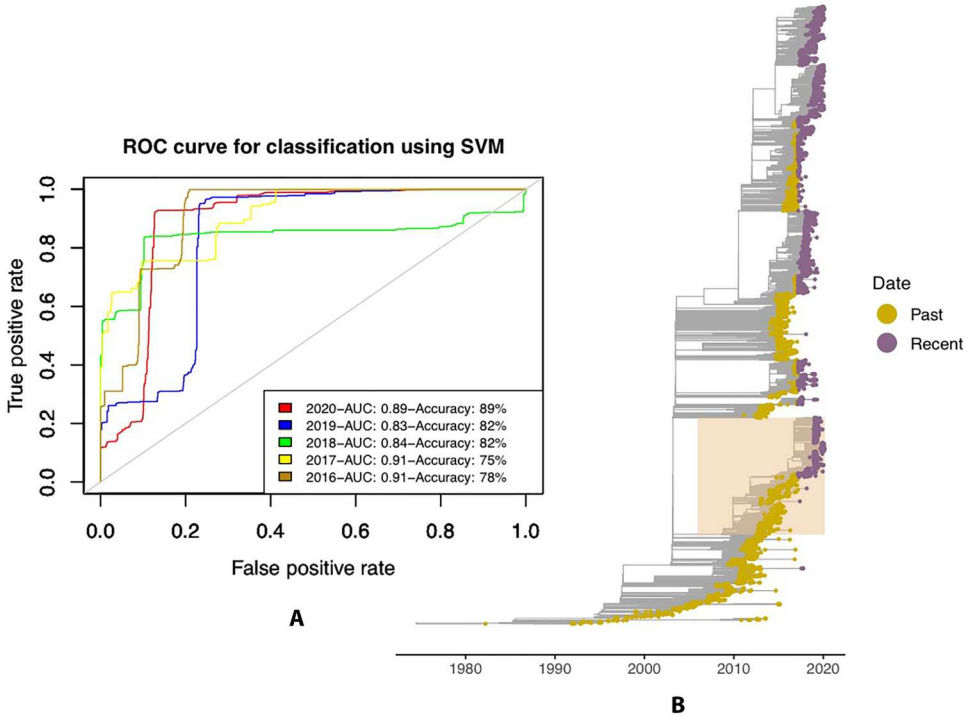

Fig. 3. Input and performance of the support vector machine. (**A**) ROC curve showing the performance of support vector machine (SVM) with a polynomial kernel on trees reconstructed from sequences up to February 2016, 2017, 2018, 2019, and 2020. For each tree, we extract one subtree for testing the model and use the remainder of the tree for training our model. See the Supplementary Materials (table S3) for more details on accuracy calculation. (**B**) Tree reconstructed from influenza sequences from 1980 up to February 2020. Orange highlight shows the test subtree. Tips are colored by date: yellow for "past" tips sampled before March 2017 and purple for tips sampled after March 2017, labeled "recent" and used to predict the successful tips that are more probable to circulate in the next (2020/2021) flu season.

uniformly randomly divide each dataset into training and testing sets. To ensure that the training and testing sets are not dependent because of relatedness, we take two approaches. In the first (results in the main text), we choose one large clade of the tree to test the model and train on the remainder of the tree. For each experiment culminating in years 2016, 2017, 2018, 2019, and 2020, we choose a test clade such that there is an approximate 10 to 20% ratio between training and test datasets (see Fig. 3B). In the second approach (results in the Supplementary Materials), we train our model on the "past tips" (before August 2016 in the tree reconstructed from sequences up to 2020; see the Supplementary Materials for full details) and test on those tips that remain, i.e., those sampled between the past tips and the recent tips. We include a gap of 4 months between the training and testing set in this approach to avoid dependency between tips in different sets.

Classification

We compare several binary classification tools: support vector machines (SVMs) with a range of kernel choices (*40*), random forests (*41*), and gradient boosting (*42*). Among the different binary classification tools used, an SVM with a polynomial kernel had the best performance (see the Supplementary Materials). Methods were implemented in R, with the packages e1071 (*43*), randomForest (*41*), and caret (*44*). Outliers can affect the training process; we use the local outlier factor algorithm (*45*) implemented in the DMwR package (*46*) in R to identify and remove outliers.

To carry out a sensitivity analysis on the performance of the SVM classifier, we repeat the experiment on the dataset of sequences up to February 2020, including the same test clade, and geographically downsample the dataset by two methods. In the first, we randomly select 10, 20, and 50% of all sequences collected from each continent. In the second, we downsample sequences proportional to the population per continent in 2019 (*47*). Our dataset contains the fewest sequences proportionally from Africa (N = 1047), which contains 17.2% of the world population. Therefore, we set the total size of the downsampled dataset to (1047/0.172) 6087 sequences and randomly select sequences from each continent proportional to their share of world population. Each analysis is repeated 100 times per sampling percentage, randomly selecting different sequences using the given method and rerunning the prediction model to produce receiver operating characteristic (ROC) curves. We also calculate the proportion of successful sequences in each downsampled analysis that are predicted to be successful using the full dataset.

Epitope distance for vaccine proposal

To select candidate vaccine strains from among the sequences identified as successful by the SVM with polynomial kernel, we compare three methods using epitope distance, as well as the consensus amino acid sequence among successful strains. The consensus sequences are found by assigning the amino acid code at each site that is found at the highest frequency within the successful strains. For the methods using epitope distance, we first calculate the epitope distances between the successful (i.e., our model classifies them as successful) sequences and all other sequences in the dataset that were sampled in the 5 years before the sampling of

the successful sequences. We then use these distances to determine the vaccine candidates.

We define the set of successful sequences as $\mathscr{S}$. Then, the epitope distance $H(s, \mathscr{R})$ between a successful tip $s \in \mathscr{S}$ and the set of sequences from the prior 5 years $\mathscr{R}$ is defined as the mean of the distances $h(s, r)$ from s to all sequences r in $\mathscr{R}$

$$H(s, \mathscr{R}) = \frac{1}{|\mathscr{R}|} \sum_{r \in \mathscr{R}} h(s, r)$$

Here, h is restricted to the epitope sites in the HA sequence according to (*36*).

We consider three scenarios for using these distances $H(s, \mathscr{R})$, $s \in \mathscr{S}$ to select final vaccine candidates. In the first scenario ("epitope-furthest-from-previous"), we choose the successful sequences that are most distant to the set of tips from the prior 5 years as the final vaccine candidates [i.e., we find the set of s that maximizes $H(s, \mathscr{R})$]. The motivation here is that the most distinct sequences may be the best candidates, as humans' immune systems will not have developed antibodies against them. In the second scenario ("epitope-nearest-to-previous"), we choose the set of successful sequences that are the least distant to the set of tips from the prior 5 years [we find the set of s that minimizes $H(s, \mathscr{R})$]. This can be motivated by the fact that it is more probable that sequences from fast growing clades will persist into the future. In the third scenario ("epitope-central-to-previous"), we select the set of successful sequences that are closest to the median distance to the set of tips from the prior 5 years [we find the set of s that are closest to the median of $H(s, \mathscr{R})$]. We might expect for this scenario to result in a set of sequences that are different enough from past sequences to avoid preexisting immunity, while still being likely to be situated in growing clades.

To assess the performance of our candidate selection approach for each season (2016–2020), we compute both the epitope and full amino acid hamming distances between the vaccine candidate sequences and the following season's circulating sequences. To evaluate these approaches further, we compute these distances between the following season's circulating sequences and the WHO vaccine candidates for each year and a consensus sequence derived from all circulating strains in the previous 3 years ("Consensus-from-all").

For this task, we used 57,339 protein sequences collected from March 2016 to October 2020, divided by flu season March to February. For the final year in the study period, i.e., proposal of the 2021 influenza vaccine, we only used sequences from March through October 2020. However, given the low number of flu cases detected in the Northern Hemisphere 2020/2021 flu season during the COVID-19 pandemic (*48*), this should not come at great loss of diversity.

Genome-wide association study

We use treeWAS (*49*) to conduct genome-wide association studies (GWAS) to identify single-nucleotide polymorphisms (SNPs) that were associated with successful sequences in the 2020/2021 dataset. We conduct this analysis separately for the NA and HA sequences. We use the "terminal" option in treeWAS to identify correlation between the successful strain phenotype and the presence of SNPs, with significance thresholds corrected for multiple testing.

RESULTS

Classifier performance and features

We use our chosen SVM with a polynomial kernel to predict successful sequences among the recent tips (those sampled after March 2017) of the trees reconstructed from sequences up to February 2020. This provides a set of sequences that we would expect to have been more likely to circulate in the following 2020/2021 flu season. We also perform this prediction for the other datasets (2016, 2017, 2018, and 2019).

The performance of the model is found to be good. Figure 3A shows ROC curves for each of the experiments 2016 to 2020, with "area under the curve" (AUC) between 0.83 and 0.91. The accuracy of the classifier is between 75% (2017) and 89% (2020). The influenza tree and its subtree used for testing in the 2020 experiment are shown alongside in Fig. 3B. The model performance is similar for the second approach to dividing into train/test datasets; results are included in the Supplementary Material.

The performance of the SVM model is found to be robust to downsampling the data in two ways: first, to explore sensitivity to the overall data volume and second, to explore sensitivity to geographic sampling differences. Reducing the dataset of sequences up to February 2020 by 50% results in very good model performance overall (fig. S5B), with median AUC = 0.88 (SD ± 0.01) and median accuracy = 0.88 (SD ± 0.01). Even when including only 10% of the dataset in the SVM classification, we find that the model performance remains high (fig. S5D), with median AUC = 0.9 (SD ± 0.03) and median accuracy = 0.82 (SD ± 0.03). Model performance is also good when downsampling by population rather than the number of sequences collected, which accounts for potential sampling bias [median AUC = 0.88 (SD ± 0.02), median accuracy = 0.79 (SD ± 0.01); fig. S5A]. We find a high concordance overall between the successful strains identified in the full dataset and the strains predicted to be successful with each downsampling approach (fig. S6).

We used feature selection, repeating on the downsampled cases to explore robustness, and found no consistent signal, indicating that some features contribute more heavily to the predictions than others (see the Supplementary Materials). This can happen when features contribute overlapping information, and this is the case here: Figure 4 shows the pairwise correlations among the tip-level features in our data. There is a high level of duplicated information, resulting in feature selection methods not being able to consistently rank features. For example, the maximum width feature and the divergence rate feature are positively correlated and naturally show very similar patterns of correlation with the other features in the data. Some of the network science features are correlated, e.g., the "maxAdj" and "eigenvector" features. The epitope features were not correlated with the others, suggesting that they are an independent line of evidence. In a related study in 2020, we found that including epitope and tree features together resulted in the strongest performance (*19*). However, recently, Barrat-Charlaix *et al.* (*50*) found that mutations at the epitope sites in (*36*) fixed in the population less often than would be expected under neutral drift.

Vaccine selection

Among the set of sequences identified by the SVM as likely to be successful, we choose several to propose as candidates for inclusion in a potential 2020/2021 flu vaccine. We first compare the three

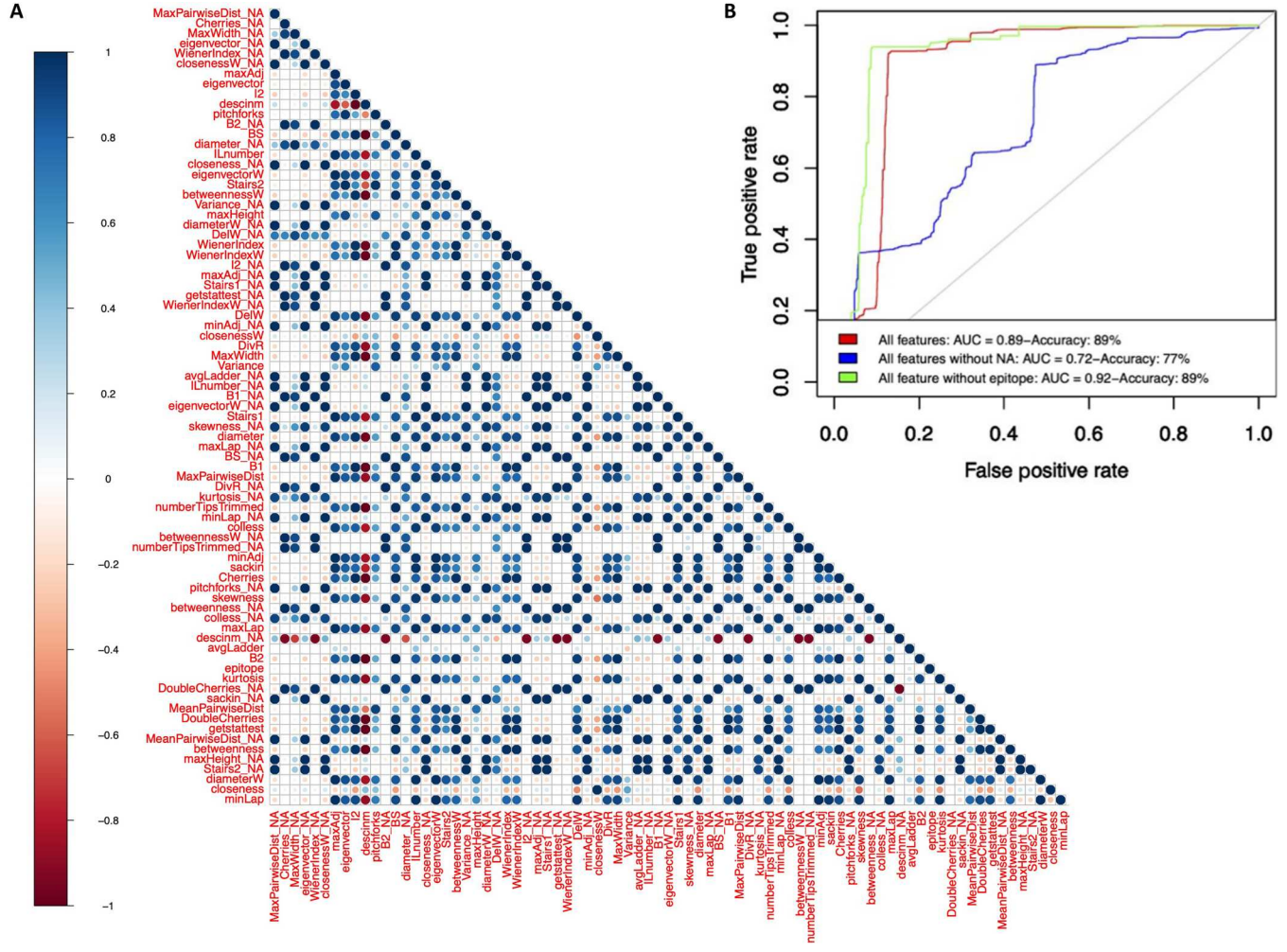

Fig. 4. Correlations between features, and comparative model performance when only subsets of the features are used. (**A**) Feature correlations, in which color illustrates the Pearson correlation between the tip-based features (computed as in Eq. 1). Blue indicates positive correlation (darkest blue: correlation of 1), and red indicates negative (darkest red: −1). There is considerable correlation in the data. (**B**) Comparative performance shown by receiver operating characteristic (ROC) curves with the full feature set (red), without the epitope features (green), and without the neuraminidase (NA)-derived features (blue). Note, however, that NA success is used to define the "success" label for a strain [as is the hemagglutinin (HA) tree; see Materials and Methods].

epitope distance approaches for candidate proposal described in the "Epitope distance for vaccine proposal" section, across all experiments 2016 to 2020. While the epitope features did not come up as consistently important in the classification, for vaccination, they may be relevant because of their antigenic signal, and the epitope distances provide one rationale for filtering the large number (3962) of sequences to a smaller set of candidates. We then ask whether our predicted successful strains are antigenically close to the sequences circulating in the following season. We compare four approaches to choosing sequences informed by our machine learning results to two other approaches: the consensus sequence ("consensus from all") and the WHO-selected strains from the given year. Barrat-Charlaix *et al.* (*50*) recently found that consensus strains are as close or closer to future populations than strains with a high local branching index, motivating the comparison to consensus strains. Figure 5 shows the epitope and amino acid distances (epitope site–only amino acid Hamming distance and whole-sequence amino acid Hamming distance, respectively). One clear signal is that strains that had the furthest epitope distance from cocirculating strains are further from the future population than the other choices (purple bars). Consensus strains do reasonably well, but our model does provide some advantage: In 2020, the consensus among our "successful strains" was closer in both the epitope and whole-sequence amino acid distances to the future strains than the WHO choice and the consensus sequence, and in the most recent 3 years, it consistently did well. The WHO vaccine choice was never the closest to the future population in the whole-sequence amino acid distance. In 2019, the epitope-central strains were closer to the future than the consensus strain and the WHO strain in both distance measures. In 2018, all but the epitope-furthest-from-previous choices were strong in terms of epitope distance, and the consensus strain was furthest in amino acid distance. In 2016 and 2017 (where there is also less data), the consensus sequence did better than the consensus-from-successful strains, but there, our epitope-nearest-to-previous choice did as well as the consensus sequence overall. However, this choice

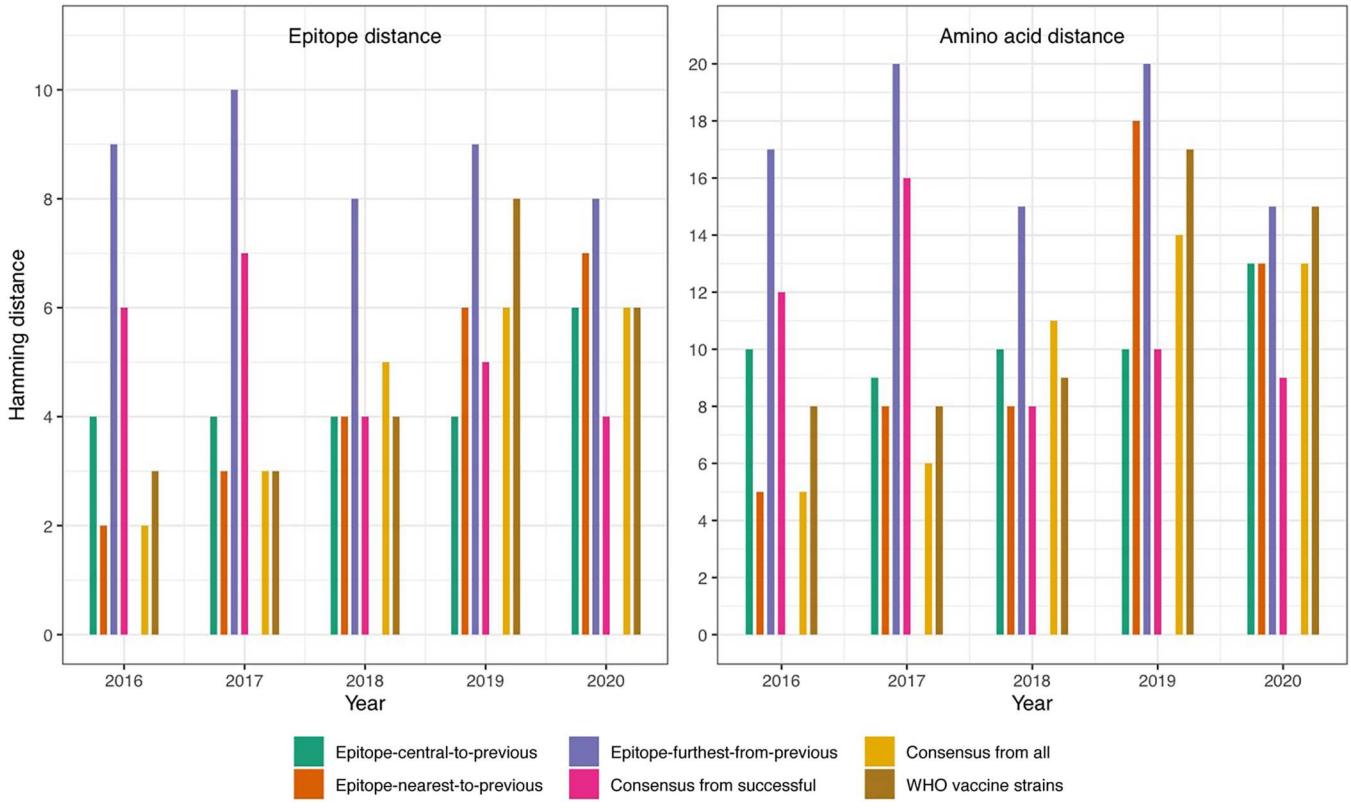

Fig. 5. Genetic distances between our vaccine candidate sequences and the sequences that circulated during the flu season in which the vaccine would have been implemented. We use the Hamming distance between amino acid sequences for our genetic distance, calculating the median distance between both epitope sites and the full amino acid sequence. Results are shown for each of the four scenarios relevant to candidate selection. We also include distances for the WHO vaccine candidate sequences from each year and a consensus sequence derived from all strains circulating in the 3 years before the end of the experiment data. "Year" in the axis refers to the final year of experiment data; vaccine candidate sequences are compared to observed protein sequences from the following flu season (e.g., "2016" uses data through February 2016 and is compared to sequences from the 2016/2017 flu season).

performed less well for amino acid distances in 2019. In the Supplementary Materials (fig. S7), we show the variability in the distances.

We select 17, 21, 37, 13, and 17 vaccine candidates for the years 2020, 2019, 2018, 2017, and 2016, respectively, from among the successful sequences using the epitope-central-to-previous criterion, which had moderate performance overall in Fig. 5. The strains suggested by the WHO for vaccine inclusion in 2020–2021 are among the set of successful sequences as predicted by our model, and all final vaccine candidates suggested by our model are very close to those suggested by the WHO; they are located in the same subtree (see Fig. 6).

Genome-wide association study

We conduct a GWAS to identify SNPs in the HA and NA sequences that are associated with successful sequences in the 2020/2021 dataset. In the NA sequences, we find three significant SNPs that are correlated with increased presence in the successful sequences predicted from our analysis in the 2020/2021 dataset. These SNPs encode a nonsynonymous mutation at position 1015 that results in the amino acid change N339D and synonymous SNPs at position 237 (codon 79) and position 870 (codon 290). Mutations at codon 356 in NA have been previously reported to be permissive mutations linked to a highly deleterious mutation at codon 336, which affects viral fitness, although we find no evidence of this association and our predicted success of sequences (*51*).

We find no SNPs in the HA sequences that are significantly associated with presence in successful sequences (fig. S4). Previous studies have identified seven key amino acid changes in HA that are involved in major antigenic change in H3N2 (*52*). While we do not find evidence of SNPs in these codons that are significantly associated with the predicted successful sequences, we manually searched for differences in these amino acid changes between the predicted successful and not successful sequences. There is no observable difference between sequences at most sites, although we do find some heterogeneity in the amino acid sequence at codon 158, with a higher proportion of successful strain carrying a glycine or arginine at this position and a lower proportion carrying lysine.

DISCUSSION

In this work, we have trained machine learning models to predict which sequences are most likely to grow during the upcoming flu season on the basis of features of phylogenetic trees. We explore choices for which sequences should be considered for inclusion in the following year's flu vaccine. The practical availability and feasibility of prospective vaccine viruses is also a key consideration (*53*, *54*). For example, candidate viruses must be identified in time to be

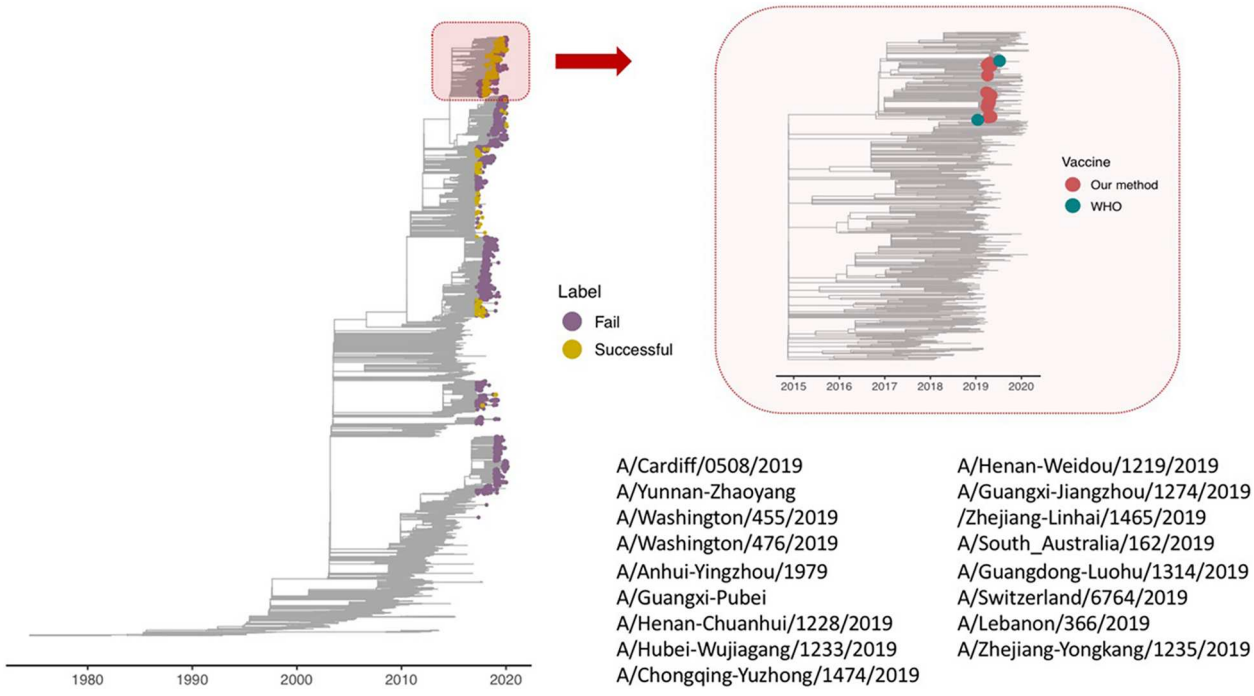

Fig. 6. Influenza tree reconstructed from sequences up to February 2020, with the subtree including our candidates for vaccine inclusion in 2020–2021 highlighted in red. Recent tips are colored yellow/purple on the basis of the prediction of our support vector machine (SVM) model. The highlighted subtree also includes the World Health Organization (WHO) candidates, as colored in the inset alongside the candidates from our SVM. The suggested vaccines by WHO for 2020–2021 are A/Hong Kong/45/2019 and A/Hong Kong/2671/2019, and our suggested candidates are listed in this figure.

approved for seasonal vaccine development, and some influenza viruses grow poorly under the conditions that regulators allow for vaccine production, limiting the available choice. The approach we have introduced in this work does not incorporate these factors explicitly, although it motivates our selection of several vaccine candidates. We have focused on predictions for the Northern Hemisphere H3N2 season. There are fewer sequences available for the Southern Hemisphere, limiting the development of methods such as this that are specifically tailored to Southern Hemisphere influenza. Our downsampling analyses indicate, however, that the accuracy is likely to be robust to uniformly lower sequencing and to differences in sampling in different regions, so our method may be applicable to Southern Hemisphere dynamics.

In this work, we used sequences (tips, in the trees) as the unit of prediction, compiling features derived from both HA and NA trees, despite reassortment. This approach is amenable to including other data at the sequence level. Influenza virus vaccines typically develop immunity by causing the host to produce antibodies specific to the HA protein, although NA also has a crucial role in viral infection in binding to SA receptors and accordingly facilitating the spread of influenza viruses (*5*). Our approach could be expanded to incorporate strain-specific data for nonphylogenetic sources and/or information from trees derived from other segments.

The findings in this paper are subject to several limitations. We did not explicitly include immunological assay data, as these are not generally available. We used specific epitope sites from (*36*) following the approach of (*55*), although a model reflecting the impact of polymorphisms across more locations in HA and other genes, if available, might further improve predictions. We also do not have good estimates of the current global circulating frequencies of strains. All influenza sequences were downloaded from GISAID; the likelihood of an infection resulting in a sequence in the data reflects geographical differences in testing, sequencing, and deposition in GISAID. Furthermore, the seasonal flu vaccine is designed to protect against common influenza viruses (H3N2, H1N1, and B) that are highly likely to circulate during the upcoming flu season. Here, we consider the H3N2 subtype and therefore predict only H3N2 sequences that are likely to spread during the upcoming flu season.

Recent work by Barrat-Charlaix *et al.* (*50*) suggests that despite signals of strong selection in surface proteins in influenza viruses (*56*) and ladder-like phylogenies usually understood to be shaped by immune selection (*57*), there is limited evidence that mutation frequency trajectories are predictable. These trajectories are more consistent with neutral evolution, in which the fixation probability is simply the mutation's frequency. These authors also explore proximity of strains to a future population, and whereas previous work had indicated that the local branching index could identify fit strains (*18*), consensus sequences are just as close (or closer) to future populations as strains that are fit according to the local branching index. This is an important result, calling into question whether neutral evolution is just as good, or better, a model for H3N2 as a pattern of selective sweeps. In a neutral model, we should not expect much predictability, and consensus sequences will be the best predictors of the future population (*50*). Here, we trained a machine model to do a task: predict subtree growth (and moderate growth at that, doubling in size). We then asked whether the results from that model are informative for the

broader question of proximity to the future population. These questions are conceptually related but quite different: A strain could grow into the future, with subtrees doubling from size 50 to 100 (for example) but remain distant from the majority of (the thousands of) strains in that future. Hence, we would not expect that a machine learning model tuned to solve the first problem would necessarily perform well at the second, although if it was highly specific to strains that grew markedly, those strains would be close to the future population. We found that the machine learning results were slightly helpful for the broader problem but that consensus sequences also do well at being near the broader future population. This is consistent with the overall evolutionary pattern of H3N2 being well modeled by neutral drift.

The COVID-19 pandemic has caused a notable change in influenza transmission dynamics. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus that causes COVID-19, and flu are immensely different pathogens, but they do have similar behaviors in some important areas (*58*, *59*). Because both viruses are transmitted by the respiratory route, the adoption of nonpharmaceutical interventions, such as mandated face masks in public, school closures, restrictions on movement, enhanced personal hygiene, and reduced travel, has had a huge impact on influenza infections and transmission (*59*). It has been argued that these interventions likely resulted in a more substantial interruption of influenza transmission than SARS-CoV-2 transmission, in part a consequence of the lower transmissibility of seasonal influenza virus [R_0 = 1.28, interquartile range: 1.19 to 1.37 (*60*)] compared with that of SARS-CoV-2 (R_0 = 2 to 3.5) (*61*). After recognition of widespread transmission of SARS-CoV-2 by mid-February 2020, influenza activity declined sharply in the Northern Hemisphere. Studies on data from clinical laboratories in the United States during March to May 2020 showed a 61% decrease in the number of submitted specimens and a 98% decrease in the percentage of positive patient tests for influenza (from a median 19.34 to 0.33% positive tests) (*61*). Influenza data reported to the WHO from Australia, South America, and Southern Africa also indicated very low influenza activity during the 2020 Southern Hemisphere flu season (June to August 2020) (*61*). In the 2022–2023 season, influenza levels recovered to levels comparable to previous years (*62*). The long-term impact of the pandemic-related disruption to influenza virus transmission during the pandemic is as yet uncertain. Although this may prove problematic for our methodology and existing approaches for vaccine candidate selection, having a wider range of tools available, including the tool we have introduced here that integrates tree-based information from different genes (and potentially additional information as well), could be increasingly useful.

Supplementary Materials

This PDF file includes:

Figs. S1 to S7
Tables S1 to S5

REFERENCES AND NOTES

1. D. Kobasa, S. Kodihalli, M. Luo, M. R. Castrucci, I. Donatelli, Y. Suzuki, T. Suzuki, Y. Kawaoka, Amino acid residues contributing to the substrate specificity of the influenza A virus neuraminidase. *J. Virol.* **73**, 6743–6751 (1999).
2. I. Kosik, J. W. Yewdell, Influenza hemagglutinin and neuraminidase: Yin–Yang proteins coevolving to thwart immunity. *Viruses* **11**, 346 (2019).
3. X. Chen, S. Liu, M. U. Goraya, M. Maarouf, S. Huang, J.-L. Chen, Host immune response to influenza A virus infection. *Front. Immunol.* **9**, 320 (2018).
4. X. Wang, Q. Sun, Z. Ye, Y. Hua, N. Shao, Y. Du, Q. Zhang, C. Wan, Computational approach for predicting the conserved B-cell epitopes of hemagglutinin H7 subtype influenza virus. *Exp. Ther. Med.* **12**, 2439–2446 (2016).
5. X. Zhu, R. M. Bride, C. M. Nycholat, W. Yu, J. C. Paulson, I. A. Wilson, Influenza virus neuraminidases with reduced enzymatic activity that avidly bind sialic acid receptors. *J. Virol.* **86**, 13371–13383 (2012).
6. Y. P. Lin, V. Gregory, P. Collins, J. Kloess, S. Wharton, N. Cattle, A. Lackenby, R. Daniels, A. Hay, Neuraminidase receptor binding variants of human influenza A (H3N2) viruses resulting from substitution of aspartic acid 151 in the catalytic site: A role in virus attachment? *J. Virol.* **84**, 6769–6781 (2010).
7. S. Wang, H. Li, Y. Chen, H. Wei, G. F. Gao, H. Liu, S. Huang, J.-L. Chen, Transport of influenza virus neuraminidase (NA) to host cell surface is regulated by ARHGAP21 and Cdc42 proteins. *J. Biol. Chem.* **287**, 9804–9816 (2012).
8. Centers for Disease Control and Prevention, Prevention and control of seasonal influenza with vaccines: Recommendations of the Advisory Committee on Immunization Practices—United States, 2013-2014. *MMWR Recomm. Rep* **62**, 1–43 (2013).
9. C. Arriola, S. Garg, E. J. Anderson, P. A. Ryan, A. George, S. M. Zansky, N. Bennett, A. Reingold, M. Bargsten, L. Miller, K. Yousey-Hindes, L. Tatham, S. R. Bohm, R. Lynfield, A. Thomas, M. L. Lindegren, W. Schaffner, A. M. Fry, S. S. Chaves, Influenza vaccination modifies disease severity among community-dwelling adults hospitalized with influenza. *Clin. Infect. Dis.* **65**, 1289–1297 (2017).
10. Centers for Disease Control and Prevention, How flu vaccine effectiveness and efficacy is measured: Questions and answers. Atlanta, GA: US Department of Health and Human Services, CDC, www.cdc.gov/flu/vaccines-work/effectivenessqa.htm. [accessed 15 March 2021].
11. L. A. Grohskopf, E. Alyanak, K. R. Broder, L. H. Blanton, A. M. Fry, D. B. Jernigan, R. L. Atmar, Prevention and control of seasonal influenza with vaccines: Recommendations of the advisory committee on immunization practice—United States, 2019–20 influenza season. *MMWR Recomm. Rep.* **68**, 1–24 (2019).
12. L. Blanton, V. G. Dugan, A. I. A. Elal, N. Alabi, J. Barnes, L. Brammer, A. P. Budd, E. Burns, C. N. Cummings, S. Garg, R. Garten, L. Gubareva, K. Kniss, N. Kramer, A. O'Halloran, C. Reed, M. Rolfes, W. Sessions, C. Taylor, X. Xu, A. M. Fry, D. E. Wentworth, J. Katz, D. Jernigan, Update: Influenza activity—United States, September 30, 2018–February 2, 2019. *MMWR Morb. Mortal Wkly. Rep.* **68**, 125–134 (2019).
13. M. G. Thompson, J. C. Kwong, A. K. Regan, M. A. Katz, S. J. Drews, E. Azziz-Baumgartner, N. P. Klein, H. Chung, P. V. Effler, B. S. Feldman, K. Simmonds, B. E. Wyant, F. S. Dawood, M. L. Jackson, D. B. Fell, A. Levy, N. Barda, L. W. Svenson, R. V. Fink, S. W. Ball, A. Naleway; PREVENT Workgroup, Influenza vaccine effectiveness in preventing influenza-associated hospitalizations during pregnancy: A multi-country retrospective test negative design study, 2010–2016. *Clin. Infect. Dis.* **68**, 1444–1453 (2019).
14. B. Flannery, S. B. Reynolds, L. Blanton, T. A. Santibanez, A. O'Halloran, P.-J. Lu, J. Chen, I. M. Foppa, P. Gargiullo, J. Bresee, J. A. Singleton, A. M. Fry, Influenza vaccine effectiveness against pediatric deaths: 2010–2014. *Pediatrics* **139**, e20164244 (2017).
15. D. M. Skowronski, S. Leir, S. Sabaiduc, M. Murti, J. A. Dickinson, R. Olsha, J. B. Gubbay, M. A. Croxen, H. Charest, T. Chan, N. Bastien, Y. Li, M. Krajden, G. De Serres, Interim estimates of 2018/19 vaccine effectiveness against influenza A(H1N1)pdm09, Canada, January 2019. *Euro Surveill.* **24**, 1900055 (2019).
16. M. L. Jackson, J. R. Chung, L. A. Jackson, C. H. Phillips, J. Benoit, A. S. Monto, E. T. Martin, E. A. Belongia, H. Q. McLean, M. Gaglani, K. Murthy, R. Zimmerman, M. P. Nowalk, A. M. Fry, B. Flannery, Influenza vaccine effectiveness in the United States during the 2015–2016 season. *N. Engl. J. Med.* **377**, 534–543 (2017).
17. E. A. Belongia, M. D. Simpson, J. P. King, M. E. Sundaram, N. S. Kelley, M. T. Osterholm, H. Q. McLean, Variable influenza vaccine effectiveness by subtype: A systematic review and meta-analysis of test-negative design studies. *Lancet Infect. Dis.* **16**, 942–951 (2016).
18. R. A. Neher, C. A. Russell, B. I. Shraiman, Predicting evolution from the shape of genealogical trees. *eLife* **3**, e03568 (2014).
19. M. Hayati, P. Biller, C. Colijn, Predicting the short-term success of human influenza virus variants with machine learning. *Proc. R. Soc. B* **287**, 20200319 (2020).
20. L. Chindelevitch, M. Hayati, A. F. Y. Poon, C. Colijn, Network science inspires novel tree shape statistics. *PLOS ONE* **16**, e0259877 (2019).
21. A. Dayarian, B. I. Shraiman, How to infer relative fitness from a sample of genomic sequences. *Genetics* **197**, 913–923 (2014).
22. T. Stadler, D. Kühnert, S. Bonhoeffer, A. J. Drummond, Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad. Sci. U.S.A.* **110**, 228–233 (2013).

23. E. M. Volz, Complex population dynamics and the coalescent under neutrality. *Genetics* **190**, 187–201 (2012).
24. A. F. Y. Poon, Phylodynamic inference with kernel ABC and its application to HIV epidemiology. *Mol. Biol. Evol.* **32**, 2483–2495 (2015).
25. G. E. Leventhal, A. L. Hill, M. A. Nowak, S. Bonhoeffer, Evolution and emergence of infectious diseases in theoretical and real-world networks. *Nat. Commun.* **6**, 6101 (2015).
26. K. Robinson, N. Fyson, T. Cohen, C. Fraser, C. Colijn, How the dynamics and structure of sexual contact networks shape pathogen phylogenies. *PLOS Comput. Biol.* **9**, e1003105 (2013).
27. C. Metzig, C. Colijn, A maximum entropy method for the prediction of size distributions. *Entropy* **22**, 312 (2020).
28. X. Didelot, C. Fraser, J. Gardy, C. Colijn, Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* **34**, 997–1007 (2017).
29. D. Klinkenberg, J. A. Backer, X. Didelot, C. Colijn, J. Wallinga, Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLOS Comput. Biol.* **13**, e1005495 (2017).
30. J. Barido-Sottani, V. Bošková, L. D. Plessis, D. Kühnert, C. Magnus, V. Mitov, N. F. Müller, J. P. Erska, D. A. Rasmussen, C. Zhang, A. J. Drummond, T. A. Heath, O. G. Pybus, T. G. Vaughan, T. Stadler, Taming the BEAST—A community teaching material resource for beast 2. *Syst. Biol.* **67**, 170–174 (2018).
31. C. Wymant, M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, M. Cornelissen, C. Fraser; STOP-HCV Consortium, The Maela Pneumococcal Collaboration, and The BEEHIVE Collaboration, PHYLOSCANNER: Inferring transmission from within-and between-host pathogen genetic diversity. *Mol. Biol. Evol.* **35**, 719–733 (2017).
32. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data–from vision to reality. *Euro Surveill.* **22**, 30494 (2017).
33. J. Huddleston, J. Hadfield, T. R. Sibley, J. Lee, K. Fay, M. Ilcisin, E. Harkins, T. Bedford, R. A. Neher, E. B. Hodcroft, Augur: A bioinformatics toolkit for phylogenetic analyses of human pathogens. *J. Open Source Softw.* **6**, 2906 (2021).
34. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
35. T. H. To, M. Jung, S. Lycett, O. Gascuel, Fast dating using least-squares criteria and algorithms. *Syst. Biol.* **65**, 82–97 (2016).
36. A. C. Shih, T.-C. Hsiao, M.-S. Ho, W.-H. Li, Simultaneous amino acid substitutions at antigenic sites drive influenza a hemagglutinin evolution. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 6283–6288 (2007).
37. M. E. J. Newman, Analysis of weighted networks. *Phys. Rev. E* **70**, 056131 (2004).
38. C. Godsil, G. Royle, *Algebraic Graph Theory* (Springer, 2001).
39. A. M. Laughlin, P. Sereda, N. Oliveira, R. Barrios, C. J. Brumme, Z. L. Brumme, J. S. G. Montaner, J. B. Joy, Detection of HIV transmission hotspots in British Columbia, Canada: A novel framework for the prioritization and allocation of treatment and prevention resources. *EBioMedicine* **48**, 405–413 (2019).
40. N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* (Cambridge Univ. Press, 2013).
41. A. Liaw, M. Wiener, Classification and regression by randomforest. *R. News* **2**, 18–22 (2002).
42. J. H. Friedman, Stochastic gradient boosting. *Comput. Stat. Data. Anal.* **38**, 367–378 (2002).
43. D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8 (2017); https://CRAN.R-project.org/package=e1071.
44. M. Kuhn, J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, T. Hunt, caret: Classification and regression training. R package version 6.0-84 (2019); https://CRAN.R-project.org/package=caret.
45. M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, LOF: Identifying density-based local outliers, in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (ACM, 2000), pp. 93–104.
46. L. Torgo, Data Mining with R, learning with case studies (Chapman and Hall/CRC, 2010).
47. United Nations, Department of Economic and Social Affairs, Population Division, World population prospects 2019, https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf [accessed 14 May 2023].
48. World Health Organisation, Influenza update–390, www.who.int/influenza/surveillance_monitoring/updates/2021_03_29_surveillance_update_390.pdf [accessed 1 April 2021].
49. C. Collins, X. Didelot, A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLOS Comput. Biol.* **14**, e1005958 (2018).
50. P. Barrat-Charlaix, J. Huddleston, T. Bedford, R. A. Neher, Limited predictability of amino acid substitutions in seasonal influenza viruses. *Mol. Biol. Evol.* **38**, 2767–2777 (2021).
51. R. Lei, T. J. C. Tan, A. H. Garcia, Y. Wang, M. Diefenbacher, C. Teo, G. Gopan, Z. T. Dargani, Q. W. Teo, C. S. Graham, C. B. Brooke, S. K. Nair, N. C. Wu, Prevalence and mechanisms of evolutionary contingency in human influenza h3n2 neuraminidase. *Nat. Commun.* **13**, 6443 (2022).
52. B. F. Koel, D. F. Burke, T. M. Bestebroer, S. van der Vliet, G. C. M. Zondag, G. Vervaet, E. Skepner, N. S. Lewis, M. I. J. Spronken, C. A. Russell, M. Y. Eropkin, A. C. Hurt, I. G. Barr, J. C. de Jong, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, R. A. M. Fouchier, D. J. Smith, Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science* **342**, 976–979 (2013).
53. Centers for Disease Control and Prevention, Vaccine effectiveness: How well do the flu vaccines work? www.cdc.gov/flu/vaccines-work/vaccineeffect.htm [accessed 29 June 2020].
54. 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, D. J. Smith, Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine* **26**, D31–D34 (2008).
55. M. Łuksza, M. Lässig, A predictive fitness model for influenza. *Nature* **507**, 57–61 (2014).
56. S. Bhatt, E. C. Holmes, O. G. Pybus, The genomic rate of molecular adaptation of the human influenza a virus. *Mol. Biol. Evol.* **28**, 2443–2451 (2011).
57. B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, J. A. Mumford, E. C. Holmes, Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004).
58. R. Ozaras, R. Cirpin, A. Duran, H. Duman, O. Arslan, Y. Bakcan, M. Kaya, H. Mutlu, L. Isayeva, F. Kebanlı, B. A. Deger, E. Bekeshev, F. Kaya, S. Bilir, Influenza and COVID-19 coinfection: Report of six cases and review of the literature. *J. Med. Virol.* **92**, 2657–2665 (2020).
59. D. A. Solomon, A. C. Sherman, S. Kanjilal, Influenza in the COVID-19 era. *JAMA* **324**, 1342–1343 (2020).
60. M. Biggerstaff, S. Cauchemez, C. Reed, M. Gambhir, L. Finelli, Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: A systematic review of the literature. *BMC Infect. Dis.* **14**, 480 (2014).
61. S. J. Olsen, E. Azziz-Baumgartner, A. P. Budd, L. Brammer, S. Sullivan, R. F. Pineda, C. Cohen, A. M. Fry, Decreased influenza activity during the COVID-19 pandemic—United States, Australia, Chile, and South Africa, 2020. *Am. J. Transplant.* **20**, 3681–3685 (2020).
62. Flunet, www.who.int/tools/flunet [accessed 28 May 2023].
63. K. V. Mardia, Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519–530 (1970).
64. L. Chindelevitch, 'treeCentrality': Computation of Network Science Statistics on Trees in Linear Time. 2018. R package version 0.1.0, https://rdrr.io/github/Leonardini/treeCentrality [accessed 28 March 2019].
65. M. Kendall, M. Boyd, C. Colijn. phyloTop: Calculating Topological Properties of Phylogenies. 2018. R package version 2.1.1, https://CRAN.R-project.org/package=phyloTop. [accessed 28 March 2019].
66. M. Newman, *Networks: An Introduction* (Oxford Univ. Press, 2010).
67. B. Bollobás, *Modern Graph Theory* (Springer Science & Business Media, 2013), vol. 184.
68. B. Mohar, T. Pisanski, How to compute the Wiener index of a graph. *J. Math. Chem.* **2**, 267–277 (1988).
69. A. McKenzie, M. Steel, Distributions of cherries for two models of trees. *Math. Biosci.* **164**, 81–92 (2000).
70. N. A. Rosenberg, The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. *Ann. Comb.* **10**, 129–146 (2006).
71. D. H. Colless, Relative symmetry of cladograms and phenograms: An experimental study. *Syst. Biol.* **44**, 102–108 (1995).
72. M. J. Sackin, "Good" and "bad" phenograms. *Syst. Biol.* **21**, 225–226 (1972).
73. C. Colijn, J. Gardy, Phylogenetic tree shapes resolve disease transmission patterns. *Evol. Med. Public Health* **2014**, 96–108 (2014).
74. M. M. Norström, M. C. F. Prosperi, R. R. Gray, A. C. Karlsson, M. Salemi, PhyloTempo: A set of R scripts for assessing and visualizing temporal clustering in genealogies inferred from serially sampled viral sequences. *Evol. Bioinform. Online* **8**, 261–269 (2012).
75. F. A. Matsen, A geometric approach to tree shape statistics. *Syst. Biol.* **55**, 652–661 (2006).

Acknowledgments: We gratefully acknowledge the originating and submitting laboratories of sequences deposited in GISAID that were used in this study. We thank J. W. McCauley for helpful comments on this manuscript. **Funding:** This work was supported by the Federal Government of Canada's Canada 150 Research Chair Programme (C.C.). **Author contributions:** Conceptualization: M.H., B.S., and C.C. Methodology: M.H., B.S., and C.C. Investigation: M.H., B.S., and C.C. Visualization: M.H., B.S., and J.E.S. Supervision: C.C. Writing (original draft): M.H., J.E.S., and C.C. Writing (review and editing): M.H., B.S., J.E.S., and C.C. **Competing interests:** The

authors declare that they have no competing interests. **Data and materials availability:** Code and derived data are available at https://github.com/HayatiMar/flu_vaccine and https://datadryad.org/stash/landing/show?id=doi%3A10.5061%2Fdryad.x95x69pqh. Accession numbers and references to the GISAID submitting laboratories for the sequences used in this study are available in this repository. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 6 March 2022
Accepted 5 October 2023
Published 3 November 2023
10.1126/sciadv.abp9185

Science Advances

Phylogenetic identification of influenza virus candidates for seasonal vaccines

Maryam Hayati, Benjamin Sobkowiak, Jessica E. Stockdale, and Caroline Colijn

Sci. Adv. **9** (44), eabp9185. DOI: 10.1126/sciadv.abp9185

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.