bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Predicting the short-term success of human influenza A variants with machine learning

Maryam Hayati[1] Priscila Biller[2] Caroline Colijn[2,3]

[1]Simon Fraser University, Department of Computing Science
[2]Simon Fraser University, Department of Mathematics
[3]Imperial College London, Department of Mathematics

April 14, 2019

Abstract

Seasonal influenza viruses are constantly changing, and produce a different set of circulating strains each season. These small genetic changes can accumulate over time and result in antigenically different viruses. Accordingly this may prevent the body's immune system to recognize those viruses. Due to the rapid mutations in the hemagglutinin gene, vaccines against seasonal influenza have to be updated frequently. This requires choosing strains to include in the updates to maximize the vaccines' benefits, according to estimates of which strains will be circulating in upcoming seasons. This is a challenging prediction task. In this paper we use longitudinally sampled phylogenetic trees based on hemagglutinin sequences, together with counts of epitope site polymorphisms in hemagglutinin, to predict which influenza strains are likely to be successful. We extract small groups of taxa (subtrees) and use a suite of features of these subtrees as key inputs to the machine learning tools. Using a range of training and testing strategies, including training on H3N2 and testing on H1N1, we find that successful prediction of future expansion of small subtrees is possible from these data, with accuracies of 0.71-0.85 and AUC 0.75-0.9.

1 Introduction

Human influenza A virus remains a substantial global public health challenge, causing considerable illness and mortality despite the availability of effective vaccines. Influenza viruses are categorized according to features of two surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA), with types such as H3N2 and H1N1 indicating the variant of HA and NA characterizing the strain. Influenza A is prone to variability, both in the form

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

of so-called antigenic drift, and in the form of reassortment. Reassortments can give rise to new variants with distinct antigenic properties compared to previous strains; resulting pandemic influenza strains may be highly pathogenic. In contrast to pandemic strains arising from reassortment, seasonal influenza primarily arises through genetic drift, as influenza has a high propensity for generating antigenic variation. This allows influenza to evade the previous immunity built up in host populations. As a consequence, seasonal influenza vaccines need to be regularly updated.

Influenza vaccines typically focus on preventing influenza A infection by raising antibodies specific to the hemagglutinin (HA) protein. In order to update a seasonal influenza vaccine, currently-circulating strains must be selected for inclusion. This relies on surveillance and sequencing of circulating influenza genotypes, and on measured antigenic properties of circulating strains. These data do not, in and of themselves, describe future circulating strains, and sometimes the strains selection process does not reflect the near-future composition of the influenza viruses well enough to achieve the desired reductions in illness and mortality. Predictive models are now being used in conjunction with sequencing and immunological surveillance in order to improve the strain selection process.

Phylogenetic trees encode patterns of ancestry and descent in a group of organisms, and so they necessarily include information about differences in these patterns between different subsets of taxa [35, 12]. Trees contain both branch lengths and information in the form of the tree topology or shape. Phylogenetic trees have been used in infectious disease to estimate the basic reproduction number [46], parameters of transmission models [51], aspects of underlying contact networks [39, 23, 40, 29] and in densely sampled datasets even person-to-person transmission events and timing [13, 19, 2, 52]. It is therefore natural to hypothesize that phylogenetic tree structures and branching patterns contain information about short-term growth and fitness. Tree information is central in some predictive models for short-term influenza evolution and models of fitness [35, 12]. However, the mapping between the phylogenetic tree structure and interpretable biological information can be subtle, [40, 29, 9, 27] and trees do not directly reveal the short-term evolutionary trajectories of groups of taxa.

Improvements in influenza surveillance, sequencing, data sharing and visualization [33, 14] mean that sequence data over considerable time frames is now available to the community alongside intuitive and interactive displays showing how the population of influenza viruses has changed over time. Computational systems to reconstruct large-scale phylogenetic trees from sequence data have also been developed. Machine learning models are well-suited to systematically explore subtle relationships between a suite of features and an outcome. These together present the opportunity to integrate information from different sources to improve short-term influenza predictions, using phylogenetic trees as a framework. Here, we use a convolution-like approach to identify small subtrees within a large global H3N2 phylogeny derived from HA sequences sampled between 1980 and July 2018. We fit classification models to detect early signs of growth and hence to predict the

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

short-term success or failure of these subtrees. We validate the predictions on a portion of the data not included in the fit procedure. We relate our predictions to the WHO defined clades [3, 4] for sequences sampled from 2015-2018. Our approach could be performed in real time, is computationally efficient and can be continually refined to improve the quality of predictions as more data are gathered. We suggest that small groups of closely-related influenza sequences and the phylogenetic trees that capture their recent shared ancestry patterns can complement other approaches to better predict short-term seasonal influenza evolution.

2 Results

Briefly, we extract subtrees from the H3N2 phylogeny. Each subtree corresponds to an internal node of the tree and the tip descendants that have occurred within a fixed time frame (1.4 years). The remaining tips occur after the fixed time frame following the relevant internal node, and help to define whether the subtree has successfully grown into the future.

The approach results in a total of 391 subtrees, overlapping to some extent, containing 7615 of the 12785 tips in the full phylogeny. We use a wide range of features of the subtrees, focusing largely on tree structure but also including some branch length features, and the number of changes in the epitope sites of HA compared to previous sequences – see Material and Methods. We train supervised machine learning models to use this information to predict whether subtrees will succeed. Figure1 shows the H3N2 hemagluttinin phylogeny and highlights in yellow the tips that belong to at least one subtree.

The trained models successfully predict which subtrees will grow sufficiently, as measured within 3.4 years of a subtree's originating node (2 years after the last possible tip in a subtree). Using support vector machine classification with a linear kernel, our overall 10-fold cross validation accuracy in H3N2 (using the HA sequences) was 74% with an average AUC of 0.82 (range 0.73-0.9); see figure S3. We found an accuracy (portion correctly classified) of 79% and AUC of 0.89 when training on 75% of the subtrees chosen uniformly at random, and testing on the rest (Figure 2).

This train/test split gives rise to the model used for the predictions in Table 1. Figure 2 shows receiver-operator characteristic curves illustrating the trade-off between sensitivity and specificity. AUC ranges were obtained by training 10 models each on 90% of the subtrees; see Figure S3. We obtained a 79% accuracy and 0.86 AUC when we trained a linear kernel SVM model on a training portion of the subtrees (75% of the subtrees chosen uniformly at random) obtained from the H1N1 phylogeny, reconstructed using sequences from 2009 to 2018-05 (see Methods; we did not use epitope features in any H1N1 analyses as the HA protein differs in H1N1). We performed 10-fold cross-validation on H1N1 subtrees, which resulted in 0.76 average AUC (range 0.52-0.95). We also pooled the subtrees of H1N1 and H3N2 and divided the pooled subtrees into training and test

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

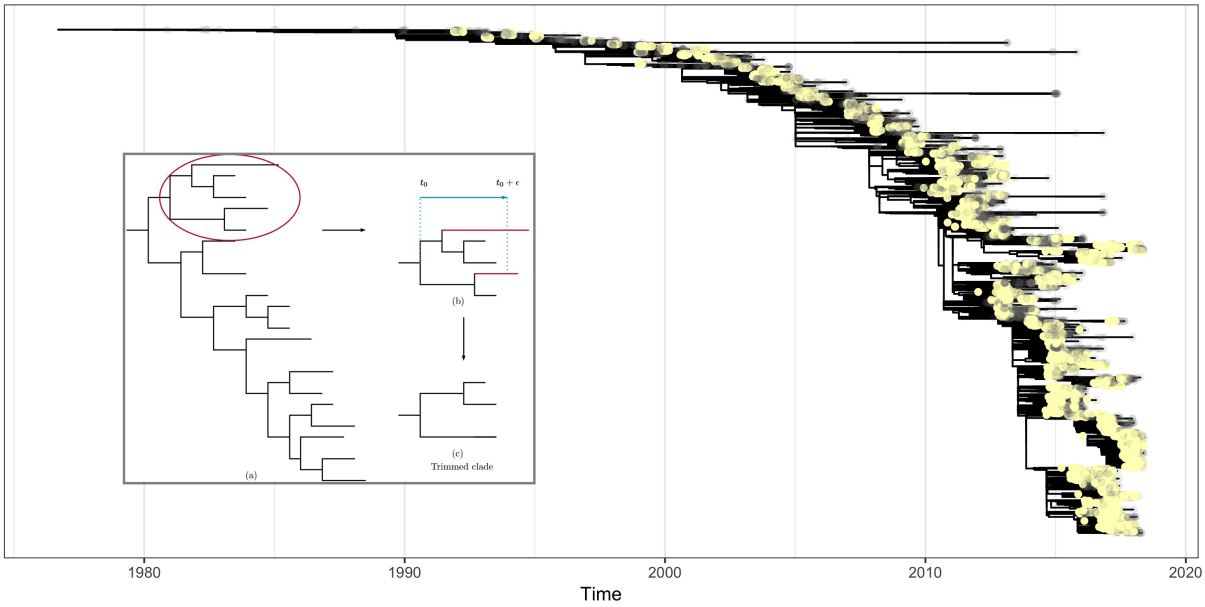
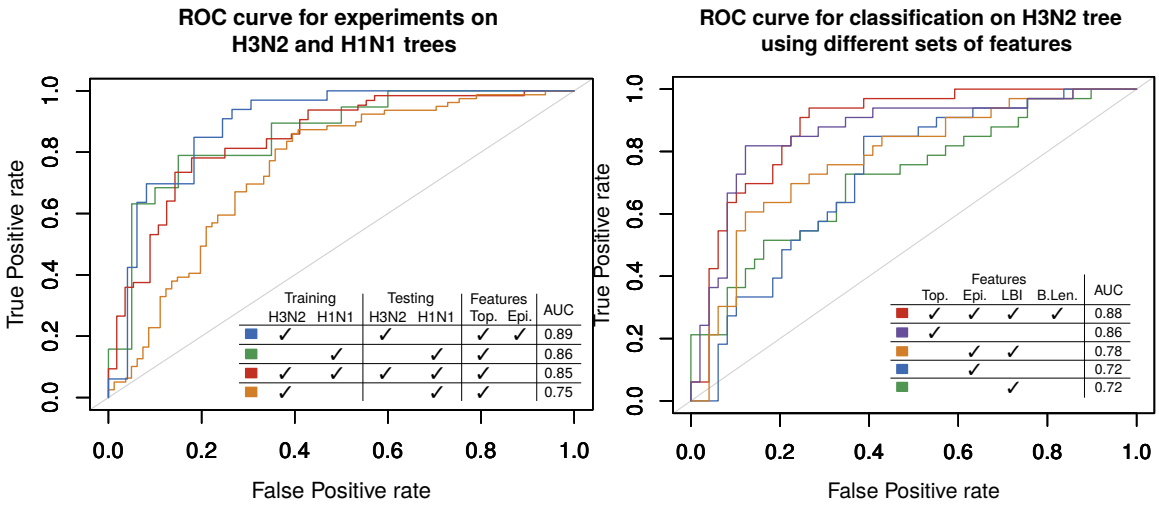


Figure 1: Influenza tree reconstructed from H3N2 subtype sequences using RAxML, with tips highlighted. Each yellow tip is in a trimmed subtree (7615 out of 12785 tips); grey tips are not. The sequences are downloaded from GenBank from 1980 to 2018-5. Long branches in this timed tree did not appear as long branches in the RAxML tree and were left in, though their tips are not in any trimmed subtree. Inset: illustration of formation of trimmed subtrees: (a) the circled clade contains a subtree; (b) red branches reach tips that occur after the trimming time period and so are pruned out; (c) the resulting trimmed subtree.

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



(a) (b)

Figure 2: (a) ROC showing performance of linear kernel *SVM* trained and tested on H3N2, trained and tested on H1N1, trained and tested on the merged subtrees of H3N2 and H1N1 and trained on H3N2 and tested on H1N1. Figure S3 shows variation in these curves and their AUCs over 10 models trained on 90% of the test data for each case. (b) ROC showing classification performance for restricted sets of features. Top: tree shape features (no branch lengths). Epi: epitope features. LBI: local branching index. B.Len: some features include tree branch lengths.

data sets; this resulted in an accuracy of 75% and an AUC of 0.85 (10-fold cross-validation AUC range 0.75-0.88). We also compared classifier performance using only portions of the data, and find that combining the tree shape features with epitope and local branching index (LBI) [35] gives the highest quality, with AUC of 0.88 compared to 0.72 for either epitope features [24] or LBI alone, and 0.78 for these combined.

Our subtrees are based on internal nodes in long-time phylogenies, and these nodes are present as a consequence of the relatedness patterns in all the data that are passed in to the tree reconstruction algorithm (in particular, including sequences from the entire time range). In consequence, a node's existence and local structure may be conditional on sequences occurring chronologically long after the node. We took several approaches to ensure that our models were not influenced by some such subtle knowledge of the future. We trained models on H3N2 HA phylogeny but tested on an H1N1 HA phylogeny. We obtained an accuracy of 72% and an AUC of 0.75 (range 0.72-0.75 when training 10 models each on 90% of H3N2 subtrees and testing on H1N1; Figures 2 and S3). The reduced accuracy is natural given that the HA proteins differ between the two types. We also created an H3N2 tree using only tips occurring prior to May 2017, extracted subtrees, and tested their success using tips from our entire dataset including those after May 2017 (see Materials and Methods and Supplementary Information). This mimics a 'real-time' analysis in which subtrees cannot depend on sequences arising after the fixed time. This

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

approach performs comparably to our other tests, with an accuracy of 76% and an AUC of 0.82.

Subtrees originating recently have had less opportunity to grow into the future than older subtrees, and consequently, we may not know whether they are successful. To accommodate for this, throughout our analysis, we only trained and tested models on subtrees that either originated prior to January 2015, or had sufficient growth to determine their known outcome (success). The cross-validation models can then be used to make predictions for recent trimmed subtrees whose outcomes are not known.

Our recent subtrees originated between 2015-03 and 2017-02, and in this period, contained tips as listed in Table 1, with the majority of tips in clade A1b, A1, A1a, A2 and A2/re. This reflects the sequences in GenBank, and is likely not globally representative [14]. There is a substantial portion of the early A2/re clade that is predicted to be successful in our model, and indeed the A2/re clade did become more successful than was been predicted at the time, probably due to a re-assortment event [4]. Besides, our model also predicted that other areas of A2 and A3 will continue to rise; there is a no clear signal for A1b and descending clades A1b/135K and N. In the time frame we had, there were relatively few sequences in our GenBank data that were mapped into the A1b/135K and A1b/135N clades by the 'augur' pipeline [14], and in our predictions these are not consistently predicted to be growing. Of the known successful subtrees in this latter part of the data, 10 of 10 A2/re subtrees were predicted (correctly) to be successful by our model, 9 of 10 in 3c3.A and 6 of 6 in clade A1. Table 1 includes these as well as subtrees for which we do not yet know the outcome.

	A1b	A1b/135K	A1b/135N	A1	3c2.A	A1a	A2	A2/re	A3	A4	3c3.A
Frac tips: succeed	0.89	0.93	0.56	0.45	0.57	0.22	0.982	0.89	0.942	0	0.80
Frac tips: fail	0.25	0.80	0.60	0.81	0.86	0.80	0.055	0.16	0.058	1	0.27
Frac subtrees: succeed	0.43	0.50	0.50	0.38	0.33	0.27	0.778	0.74	0.875	0	0.69
Frac subtrees: fail	0.57	0.50	0.50	0.62	0.67	0.73	0.222	0.26	0.125	1	0.31
Total # subtrees	14	2	4	26	6	11	9 0	19	8 0	1	16
Total # tips	296	15	48	394	88	323	326 0	806	139 0	20	196
# known outcomes	7	1	1	12	4	4	2	10	5	1	10

Table 1: Predictions for subtrees arising after 2016. These trees' success is only known if the subtree has already had sufficient growth. 'Frac tips: succeed' refers to the fraction of tips in subtrees predicted to succeed, and so on. The prediction is correct in 40/57 Of these known successful subtrees. All successful subtrees in this period had at least one prediction of success from the 10 cross-validation models, compared to only 66% of the subtrees whose ultimate success is not yet known (and which did not, in the brief time available, have sufficient descendants to be termed successful). Since tips can occur in more than one subtree due to overlap, the fraction of tips in successful and unsuccessful subtrees does not add to 1.

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

3 Discussion

We efficiently predict the success of individual influenza subtrees using machine learning tools applied to phylogenetic trees. Our method allows binary classifiers to be trained to predict which currently circulating subtrees will persist into the future based primarily on a suite of phylogenetic features.

Our approach is complementary to previous approaches, including the fitness based model proposed by Łuksza and Lässig [24] and the tree-centred work of Neher and colleagues [35, 34]. Our approach requires a reconstructed timed phylogenetic tree, and can accommodate additional data (e.g. we have used epitope mutations) easily. Other approaches often require additional data such as HI titers and estimates of the ancestral sequences, introducing experimental and computational costs and uncertainty. Our approach makes use of the reconstructed phylogeny in two distinct ways, first in obtaining the groups of taxa (subtrees) considered together for analysis, and second in that the tree shape and length features are derived from the phylogeny, and capture features of the complex branching patterns within subtrees, as opposed to their overall rates.

Our approach is rooted in the hypothesis that fitness and early success leave signatures in the branching time and structure of phylogenetic trees, which can be complemented with additional relevant information such as epitope diversification. With only slight modifications, our model could be applied to other organisms. We could also extend the approach and train regression models to predict the number of tips arising from subtrees. However, a natural limitation of this (and other tree-based approaches) is that it detects signs of early growth - if an adaptive new mutation arose in the population and was sampled before that early growth could occur and be sampled, then we could not detect early signs of growth and would not see the new adaptive mutation. In contrast, a principled modelling approach based on an understanding of both what makes an influenza virus fit and on the current composition of population immunity would likely be able to detect fit novel mutations without relying on such viruses already having begun to spread.

Influenza A can be categorized based on the presence of different proteins on the surface of the viruses: hemagglutinin (HA) and neuraminidase (NA). We use trees reconstructed from HA sequences; relatedness in the HA tree corresponds to similar HA sequences and hence to similar immunity profiles, as antibodies are induced by HA. Indeed, path lengths in the HA phylogenetic tree provide a good model for antigenic differences modeled by serological assays [35]. Trees describe the relative number of recent descendants of a lineage compared to closely-related lineages, the timing and asymmetry in the descent patterns and the short- and long-term future populations that are related to the lineages. Our approach allows this information to be included in predictive efforts.

We did not include information about proximity of strains to current or recent vaccines, which might have led to false positives in our results, if a subtree showed early signs of success but was later suppressed by vaccination. We also did not explicitly include immunological assay data, as these are not generally available. We do not have good

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

estimates of the current frequencies of subtrees or strains - indeed, if up-to-date global frequencies were available at high resolution it would greatly facilitate short-term prediction. We used epitope sites following the approach of [24]; a model reflecting the impact of polymorphisms across more locations in HA and in other genes, if this were available, would also potentially improve predictions.

We used RAxML to infer the trees; it uses a maximum likelihood approach and is considered a state-of-the art reconstruction algorithm [53, 21]. Due to the large numbers of isolates we did not perform Bayesian tree reconstruction to accommodate tree uncertainty; in addition to each required MCMC this would also have required exploration of different priors and assumptions, and it is computationally unfeasible for thousands of tips. In order to check the robustness of our approach, we used different training and testing trees including training on H3N2 but testing on H1N1, pooling H3N2 and H1N1 and using distinct time slices and consistently obtained successful predictions.

Tree shape statistics are dependent on the size of a subtree, so in order to have a more robust comparison between the subtrees, it would be best to select subtrees with approximately the same size. In our data the size of a trimmed subtree was a poor predictor of the fractional growth. Furthermore, constraining the sizes of subtrees reduces both the number of subtrees and the number of tips that can be included in the analysis, and size may in fact be a valuable predictor. We chose the approach here to balance these contrasting issues; the ongoing sampling and sequencing and the natural passing of time will ultimately provide more data – more years, and more samples per year – such that subtrees of more consistent size can be used. We anticipate that this will increase the quality of predictions. Another next logical step will be to model competition or other interactions between major lineages. We have not explicitly modelled ancestral states, key individual mutations, serological data or estimates of the fitness of sequences, but our approach could easily integrate results from models that include these features. Ideally, all relevant sources of information would be integrated and updated in real time [34, 24]. However, while short-term forecasting based on various data sources is feasible and is required to update seasonal vaccines, perfect short-term prediction and accurate long-term prediction are likely not possible because evolutionary events are fundamentally stochastic.

Recent studies on the viral isolates from vaccinated individuals indicate that they are significantly distinct from the vaccine strain and are broadly distributed on the tree, resulting in accelerated antigenic evolution [48]. Researchers have been working to develop a universal vaccine that would provide broad protection against both seasonal and pandemic influenza. Recent studies have also indicate that universal vaccines could decelerate the speed of evolution [32, 48]. If successful, such universal vaccines would eliminate the need for continual updates to seasonal influenza vaccines, but we would suggest that even so, current efforts to make short-term predictions based on surveillance and sequence data gathered over time can yield both practical results and broader insights into short-term patterns of evolution. Our approach indicates that fitness can affect the phylogenetic properties of a tree reconstructed from influenza viruses and that properties of small subtrees

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

can be used as a set of predictors to estimate which groups of sequences are showing signs of success.

4 Materials and Methods

Our approach is rooted in the hypothesis that fitness – the reproductive rate and capacity of a group of organisms – affects tree structure and branching patterns (including timing) and that this information can be extracted using machine learning tools.

4.1 Definitions

Given a phylogenetic tree T, a tip (also called an external node) 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 the 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 node to the root r of T; 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. We use a rooted bifurcating phylogenetic tree that is reconstructed from the hemagglutinin sequences of influenza A strains (by maximum likelihood).

4.2 Reconstructing Influenza trees

We collected full-length HA sequences from human H3N2 and H1N1 strains on GenBank [5, 44]. We used unique sequences of H3N2-HA from human cases for years between 1980 and 2018-5, excluding laboratory strains. This results in approximately 12919 sequences. For influenza H1N1 tree we collected pandemic sequences from 2009 until 2018-5, giving 10652 unique sequences. We aligned each set of sequences using MAFFT [15, 16] and then we used RAxML [47] to reconstruct maximum-likelihood phylogenetic trees. The reconstructed trees using RAxML are neither rooted nor dated; in some cases they include very long branches, from which descending tips were removed from our analysis. We rooted trees with the *rtt* function in the *ape* package [38] in *R* and then converted them to timed trees using the Least Squares Dating (*LSD*) software [49].

4.3 Subtree extraction

We use a convolution-style approach to identify subtrees of the global timed phylogeny that serve as units of analysis. For each internal node i in the tree, we find the tips that occur within a fixed time window (1.4 years by default) chronologically following i; this is node i's "trimmed subtree". We cannot train machine learning models on the subtree descending

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

from every internal node in the tree, because these subtrees will overlap substantially. We use the notion of a node's "relevant ancestor" (described below) to control the overlap, and select subtrees in a convolution-like way.

We first initialize each node's relevant ancestor to be its parent. We traverse the nodes of the tree in a depth-first search order. If a node's complete subtree is too small (fewer than 8 nodes by default), we reject the node and all its descendants, as none of the descending nodes can have a larger subtree than the node itself. If the node's trimmed subtree is too small but its complete subtree is large enough, we reject the node but not its descendants, since they may have subtrees that are large enough. If the node's trimmed subtree is large enough, we check the overlap between the node's subtree and its relevant ancestor's subtree. The overlap is the portion of node i's trimmed subtree that is contained in the relevant ancestor's trimmed subtree. If this overlap is not too large (under 80% of the subtree size by default), the subtree is included in our analysis. If the overlap is too large, we reject the node, and we set the relevant ancestor of the node's children to be the node's relevant ancestor. In that way, when we decide whether to accept the subtree of the node's child, we will control the correct overlap.

In this way, we obtain subtrees containing tips that are within a specified time window after their originating node, have at least a minimum number of tips, and have a limited overlap with other subtrees. We varied the minimum size, time interval and permitted overlap (See table S1 and the Supplement). We obtain a total of 392 subtrees in H3N2, and 198 subtrees in H1N1. After removing subtrees with big size in relation to the average size, and recent subtrees with insufficient growth to determine their outcome, we obtained 329 subtrees for H3N2 and 160 subtrees for H1N1.

4.4 Features

We use a set of features defined on subtrees, including both tree shape and patterns in the branch lengths. The topological features are summarized in Table 2. For the H3N2-HA dataset, we also consider some features derived from the epitope sites of the tips of the subtree. For each subtree, we consider the mean, median and the maximum genetic distances between the epitope sites of the tips of a subtree and the epitope sites of the sequences with dates prior to the subtree. We used the locations of known antigenic epitopes as mentioned in [45], namely 72 sites in the *HA1* subunit of HA.

Our features cover a wide range of global and local structures in trees, expanding considerably over previous approaches which largely focus on tree asymmetry and a few properties of branch lengths [35, 12]. Previous authors have noted that fitness leaves traces in genealogical trees [12] by observing in fixed-size populations that increased fitness resulted in increased asymmetric branching and long terminal branch lengths; Neher and colleagues used the local branching index (LBI), a measure of the total branch length surrounding a node, in their predictive model [35]. We significantly expand on the repertoire of tree features, including asymmetry and measures of local branching but also including

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

features derived from network science [8] that capture global structure of the subtrees, small shape frequencies and others – see Table 2.

In Table 2, r_i and s_i are the number of tips of the left and right subtrees of an internal node respectively. 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 by $\mathcal{I}$, the set of all tips (or external nodes) by $\mathcal{L}$, and the root of a subtree by r. In "generalized branching next" we chose $m = 2$. Skewness and Kurtosis are two measures to describe the degree of asymmetry of a distribution [25]. 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). Features called "natural" may not have been used as tree features previously but are natural extensions of simple features (eg skewness is a natural quantity to compute). The network science properties are described in [8] and were computed in R using the treeCentrality package [22]. The tree-wide summaries were primarily obtained using the phyloTop package [1].

For comparison purposes we implemented Neher et al.'s local branching index (LBI) [35], which is a measure of rapid branching near a node in the tree. In doing this we noted that there are strong parallels between the LBI and the weighted version of the Katz centrality, a classic measure from network science [17]. Figure S6 shows the correspondence. We performed the main classification task (H3N2) using all our features, only the topological tree features, only the epitope and LBI, only the epitopes and only the LBI (Figure 2(b)). We found that the combined features gave the best performance, followed by the tree features.

4.5 Success and training approach

We call a subtree of size n "successful" if its root has a total of more than αn tip descendants in the time frame of 3.4 years from the root of the subtree. The threshold of $\alpha = 1.1$ results in a good balance of successful and unsuccessful subtrees, which facilitates training the machine learning models (see Supplement). We chose to use fractional growth as our outcome rather than proximity to tips of the following season, because proximity to the following season fluctuates depending on when in the season the subtree originates, the definition of the season (i.e. the cutoff dates) and the subtree's location (tropical vs temperate).

The potential overlap between subtrees could induce dependence in the outcome variable (success), i.e., if nodes n_i and n_j have overlap, and n_i is successful, then n_j may be more likely to be successful. Notice that having some tips in common does not mean that overall subtree features are similar, but we hypothesize that the chance increases as the proportion of shared tips grows. If n_i was in the training data and n_j in the test data,

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Name	Description	Reference
Properties from network science		
Betweenness centrality	Max number of shortest paths through nodes	[36]
Closeness centrality	Max total distance to all other nodes	[36]
Eigenvector centrality	Max value in Perron-Frobenius vector	[36]
Diameter	Largest distance between 2 nodes	[6]
WienerIndex	Sum of all distances between 2 nodes	[31]
Mean tips pairwise distance	Average distance between 2 tips	natural
Max tips pairwise distance	Max distance between 2 tips (with branch lengths)	weighted diameter
Numbers of small configurations		
Cherry number	Number of nodes with 2 tip children	[28]
Normalized Pitchforks	3*(Number of nodes with 3 tip descendants) / n	[41]
Tree-wide summaries		
Normalized Colless imbalance	$\frac{1}{n^{3/2}} \sum_{i \in \mathcal{I} \cup \{r\}} \lvert r_i - s_i \rvert$	[10]
Normalized Sackin imbalance	$\frac{1}{n^{3/2}} \sum_{i \in \mathcal{L}} N_i$	[42]
Normalized Maximum height	The maximum height of tips in the tree. / $(n-1)$	[18]
Maximum width	Max number of nodes at the same depth	[9]
Stairs1	The portion of imbalanced subtrees	[37]
Stairs2	The average of $\frac{min(s_i,r_i)}{max(s_i,r_i)}$ over all internal nodes	[37]
Max difference in widths	$\max_i(n_{i+1} - n_i)$	[9]
Variance	The variance of internal node depth	[26]
I2	$\sum_{\substack{i \in \mathcal{I} \cup \{r\} \\ r_i + s_i > 2}} \frac{\lvert r_i - s_i \rvert}{\lvert r_i + s_i - 2 \rvert}$	[26]
B1	$\sum_{i \in \mathcal{I}} M_i^{-1}$	[26]
B2	$\sum_{i \in \mathcal{L}} \frac{N_i}{2^{N_i}}$	[26]
Normalized Average ladder	The mean size of ladders in the tree / $(n-2)$	[18]
Normalized ILnumber	Number of internal nodes with a single tip child / $(n-2)$	[18]
Branching speed	The ratio of the number of tips to the height of the tree	new
Measures from edge length		
Branching next index	Mean of indicator: does the next branching event descend from this node	new
Generalized branching next	Number of next two branching events descending from this node	new
Skewness	The skewness of the internal branch lengths	natural
Kurtosis	The kurtosis of the internal branch lengths	natural

Table 2: Brief definitions for tree shape statistics.

and if the two subtrees had similar features, then correlations in the success of these two subtrees could result in overfitting the data. Controlling this potential effect is one reason to use a cutoff of 3.4 years for success (meaning that n_j could be unsuccessful with n_i successful). In one of our experiments, we trained on H3N2 and tested on H1N1 (see Figure 2(a)) to ensure that test and training data are completely distinct. We also explored the performance of our approach on a set of pool subtrees from both H1N1 and H3N2.

Our data are censored, because we cannot observe the future of subtrees beginning in 2018; furthermore, we have limited knowledge of the true success of subtrees beginning in the most recent 3.4 years of our data (since it takes approximately 2 years following the

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

end of the trimmed subtree before its success is known). We only know whether a subtree has been successful if it has already had a sufficient number of tips; other subtrees may yet do so. Accordingly, we could not train our models using the last few years of data. We used 10-fold cross validation on the set of subtrees whose ancestral nodes arose before 2015-1, together with those whose ancestral nodes arose after 2015-1 but had sufficient growth that we already know their outcome. We refer to this as the set of "known subtrees". This results in one "out of fold" prediction for each such subtree. We trained an additional "general" model on a training set consisting of 75% of the known subtrees, leaving the remaining 25% for testing. We then had 10 cross-validation models and one additional model that we could use to make predictions on the subtrees arising after 2015 (for which the true success is only partially known).

The structure (and hence internal nodes) of a large phylogeny depends on all tips, not only the tips prior to the node chronologically. To avoid having the "future" tips affect the nodes on which we based our analysis, we also used a "time slicing" approach that is amenable to use in real-time, season by season. Here, we extracted subtrees not from the full phylogeny but from a tree containing only tips prior to a fixed time. We then assessed the success of these subtrees with reference to later tips (see Supplement).

4.6 Classification

We use several different binary classification tools, including support vector machine (SVM) with a range of kernel choices [11]. We use *R* implementations in the packages *e1071* [30]. We randomly choose 75% of our subtrees for training the model and leave the rest for testing. We selected a train and test datasets (75% for training and the rest for testing) then perform 10 fold cross validation on the training dataset alone; this is to find the best gamma and cost parameters without using all the data to do so. Among different binary classification tools that we used, SVM with a linear kernel had the best performance on this 75% training set, so we proceeded with this option for the remaining results. In order to find the best hyper-parameters of the model such as the training method and kernel, we performed 10-fold cross validation on the training data (not the test data). Datasets can have outliers that affect the training process. We use the local outlier factor (LOF) algorithm [7] implemented in the *DMwR* package [50] in *R*. For classification on the H3N2 tree, we removed 5 outliers which mostly were large subtrees, most of whose descending tips were contained in other subtrees. For the H1N1 tree, we had 198 subtrees; we removed the largest 11 as outliers, for the classification. For the experiment on the merged H1N1 and H3N2 subtrees we removed both big subtrees of H3N2 and H1N1 to have a set of subtrees of approximately the same size.

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

4.7 Clade assignment

We implemented a Python script using the same approach as the one found in *augur* (`https://bedford.io/projects/nextflu/augur/`) to assign a clade to a sequence. A clade is defined by a set of amino-acid or nucleotide mutations specific to that sublineage. All mutations used as markers occur in the HA1 and HA2 proteins of H3N2-HA sequences, and the detailed list of mutations can be found at `https://github.com/nextstrain/seasonal-flu/blob/master/config/clades_h3n2_ha.tsv`. A sequence is assigned to a clade only if it satisfies all criteria of the clade, i.e., only if the sequence contains all the specific mutations.

It is possible for a sequence to satisfy the criteria of more than one clade, since some clades are derived from previous ones. For instance, the sequence KU289613_A/Corsica/33-02/2015_2015/02/20 satisfies requirements for clades 3c, 3c3 and 3c3.B. In this scenario, the most specific clade is assigned (3c3.B). Among the 12919 sequences analyzed, 1106 sequences were assigned to exactly one clade, 7487 sequences were assigned to more than one clade, and 4326 sequences had no clade assigned. Most of the unassigned cases correspond to older sequences, as presented in Figure S5 in the Supplementary Material. Among the sequences assigned to more than one clade, 4225 could be solved automatically by analyzing the clade names, as in the previous example. In the remaining 3262 cases we used the tree at `https://nextstrain.org/flu/seasonal/h3n2/ha/3y` to manually determine the most specific clade (e.g. A1a is contained within 3c2.A, though this information is not encoded in the clade names).

Acknowledgments: CC and PB would like to acknowledge the Canada 150 Research Chair program. CC was also funded in part by the Engineering and Physical Sciences Research Council of the United Kingdom (EP/K026003/1). MH would like to thank Dr. Leonid Chindelevitch for providing this visiting scholar opportunity through funding from an NSERC Discovery Grant as well as the CANSSI Collaborative Research Team. This research was undertaken, in part, thanks to funding from the Canada 150 Research Chairs Program.

References

[1] Michelle Kendall [aut, cre], Michael Boyd [aut], and Caroline Colijn [aut]. phyloTop: Calculating Topological Properties of Phylogenies version 2.1.1 from CRAN. `https://rdrr.io/cran/phyloTop/`, February 2018. Accessed: 2019-3-28.

[2] Joëlle Barido-Sottani, Veronika Bošková, Louis Du Plessis, Denise Kühnert, Carsten Magnus, Venelin Mitov, Nicola F Müller, Jūlija PečErska, David A Rasmussen, Chi Zhang, et al. Taming the beast—a community teaching material resource for beast 2. Systematic biology, 67(1):170–174, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[3] Ian G Barr, Colin Russell, Terry G Besselaar, Nancy J Cox, Rod S Daniels, Ruben Donis, Othmar G Engelhardt, Gary Grohmann, Shigeyuki Itamura, Anne Kelso, et al. Who recommendations for the viruses used in the 2013–2014 northern hemisphere influenza vaccine: Epidemiology, antigenic and genetic characteristics of influenza a (h1n1) pdm09, a (h3n2) and b influenza viruses collected from october 2012 to january 2013. Vaccine, 32(37):4713–4725, 2014.

[4] Trevor Bedford and Richard Neher. Seasonal influenza circulation patterns and projections for feb 2018 to feb 2019. bioRxiv, 2018.

[5] Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. Nucleic acids research, 41(D1):D36–D42, 2012.

[6] Béla Bollobás. Modern graph theory, volume 184. Springer Science & Business Media, 2013.

[7] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In ACM sigmod record, volume 29, pages 93–104. ACM, 2000.

[8] Leonid Chindelevitch, Maryam Hayati, Art Poon, and Caroline Colijn. Network science inspires novel tree shape statistics. April 2019.

[9] Caroline Colijn and Jennifer Gardy. Phylogenetic tree shapes resolve disease transmission patterns. Evolution, medicine, and public health, 2014(1):96–108, 2014.

[10] Donald H Colless. Relative symmetry of cladograms and phenograms: an experimental study. Systematic Biology, 44(1):102–108, 1995.

[11] Nello Cristianini and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.

[12] Adel Dayarian and Boris I Shraiman. How to infer relative fitness from a sample of genomic sequences. Genetics, pages genetics–113, 2014.

[13] Xavier Didelot, Christophe Fraser, Jennifer Gardy, and Caroline Colijn. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Molecular biology and evolution, 34(4):997–1007, 2017.

[14] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics, 2018.

[15] Kazutaka Katoh, George Asimenos, and Hiroyuki Toh. Multiple alignment of dna sequences with mafft. Bioinformatics for DNA sequence analysis, pages 39–64, 2009.

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[16] 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, 2002.

[17] Leo Katz. A new status index derived from sociometric analysis. Psychometrika, 18(1):39–43, 1953.

[18] ML Kendall, M Boyd, and C Colijn. phylotop, 2016.

[19] Don Klinkenberg, Jantien A. Backer, Xavier Didelot, Caroline Colijn, and Jacco Wallinga. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLOS Computational Biology, 13(5), 2017.

[20] Teuvo Kohonen. Learning vector quantization. In Self-organizing maps, pages 175–189. Springer, 1995.

[21] John A Lees, Michelle Kendall, Julian Parkhill, Caroline Colijn, Stephen D Bentley, and Simon R Harris. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. Wellcome Open Research, 3, 2018.

[22] Leonardini. treeCentrality: 'treeCentrality': A package for computing tree shape... in Leonardini/treeCentrality: Computation of network science statistics on trees in linear time. `https://rdrr.io/github/Leonardini/treeCentrality/man/treeCentrality.html`, July 2018. Accessed: 2019-3-28.

[23] Gabriel E Leventhal, Alison L Hill, Martin A Nowak, and Sebastian Bonhoeffer. Evolution and emergence of infectious diseases in theoretical and real-world networks. Nature communications, 6:6101, 2015.

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

[25] Kanti V Mardia. Measures of multivariate skewness and kurtosis with applications. Biometrika, 57(3):519–530, 1970.

[26] Frederick A Matsen. A geometric approach to tree shape statistics. Systematic biology, 55(4):652–661, 2006.

[27] Rosemary M McCloskey, Richard H Liang, and Art FY Poon. Reconstructing contact network parameters from viral phylogenies. Virus evolution, 2(2), 2016.

[28] Andy McKenzie and Mike Steel. Distributions of cherries for two models of trees. Mathematical biosciences, 164(1):81–92, 2000.

[29] Cornelia Metzig and Caroline Colijn. Preferential attachment in systems and networks of constant size. arXiv preprint arXiv:1811.04972, 2018.

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[30] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, 2017. R package version 1.6-8.

[31] Bojan Mohar and Tomaž Pisanski. How to compute the wiener index of a graph. Journal of Mathematical Chemistry, 2(3):267–277, Jul 1988.

[32] 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 in microbiology, 26(2):102–118, 2018.

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

[34] 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. Proceedings of the National Academy of Sciences, 113(12):E1701–E1709, 2016.

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

[36] M. Newman. Networks: An Introduction. OUP Oxford, 2010.

[37] Melissa M Norström, Mattia CF Prosperi, Rebecca R Gray, Annika C Karlsson, and Marco Salemi. Phylotempo: a set of r scripts for assessing and visualizing temporal clustering in genealogies inferred from serially sampled viral sequences. Evolutionary bioinformatics online, 8:261, 2012.

[38] E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. Bioinformatics, 20:289–290, 2004.

[39] Art FY Poon. Phylodynamic inference with kernel abc and its application to hiv epidemiology. Molecular biology and evolution, 32(9):2483–2495, 2015.

[40] Katy Robinson, Nick Fyson, Ted Cohen, Christophe Fraser, and Caroline Colijn. How the dynamics and structure of sexual contact networks shape pathogen phylogenies. PLoS computational biology, 9(6):e1003105, 2013.

[41] Noah A Rosenberg. The mean and variance of the numbers of r-pronged nodes and r-caterpillars in yule-generated genealogical trees. Annals of Combinatorics, 10(1):129–146, 2006.

[42] M. J. Sackin. “good” and “bad” phenograms. Systematic Biology, 21(2):225–226, 1972.

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[43] Yvan Saeys, Thomas Abeel, and Yves Van de Peer. Robust feature selection using ensemble feature selection techniques. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 313–325. Springer, 2008.

[44] Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Scott Federhen, et al. Database resources of the national center for biotechnology information. Nucleic acids research, 40(D1):D13–D25, 2012.

[45] 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, 2007.

[46] Tanja Stadler, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv). Proceedings of the National Academy of Sciences, 110(1):228–233, 2013.

[47] Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9):1312–1313, 2014.

[48] Rahul Subramanian, Andrea L Graham, Bryan T Grenfell, and Nimalan Arinaminpathy. Universal or specific? a modeling-based comparison of broad-spectrum influenza vaccines against conventional, strain-matched vaccines. PLoS computational biology, 12(12):e1005204, 2016.

[49] Thu-Hien To, Matthieu Jung, Samantha Lycett, and Olivier Gascuel. Fast dating using least-squares criteria and algorithms. Systematic biology, 65(1):82–97, 2015.

[50] L. Torgo. Data Mining with R, learning with case studies. Chapman and Hall/CRC, 2010.

[51] Erik M Volz. Complex population dynamics and the coalescent under neutrality. Genetics, 190(1):187–201, 2012.

[52] Chris Wymant, Matthew Hall, Oliver Ratmann, David Bonsall, Tanya Golubchik, Mariateresa de Cesare, Astrid Gall, Marion Cornelissen, Christophe Fraser, The Maela Pneumococcal Collaboration STOP-HCV Consortium, and The BEEHIVE Collaboration. Phyloscanner: inferring transmission from within-and between-host pathogen genetic diversity. Molecular biology and evolution, 35(3):719–733, 2017.

[53] Xiaofan Zhou, Xing-Xing Shen, Chris Todd Hittinger, and Antonis Rokas. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. Molecular biology and evolution, 35(2):486–503, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

5 Supplement

5.1 Model parameters and results

Because we have used a fixed time frame to select subtrees, the subtrees vary considerably in size (since we do not control size). Successful subtrees are larger (median 18) than unsuccessful ones (median 12). Figure S1 shows the sizes of the trimmed subtrees from H3N2 that were used in the main analysis, with bars shaded to the outcome. We observed that size alone does not successfully classify the success of subtrees, though it contains some of the information (AUC 0.63). We chose to control time frame rather than subtree size, as time frame has a clear biological meaning and the size of subtrees may in fact be useful information for the classification; trees with rapid branching can achieve more tips in the fixed time frame. However, size (and apparent rapid branching) may also reflect sampling differences.

We used three sets of parameters to extract the subtrees of influenza H3N2 tree. These three models are summarized in Table S1, and the results of prediction using each of these models are shown in Figure S2.

We also explored changing the success threshold, defining a subtree of size n as successful if its ancestor eventually has at least one more tip, more than $1.1n$ tips and more than $1.2n$ tips (ie if size $> \alpha n$ for $\alpha = 1, 1.1, 2$). Figure S2 shows the performance under these variations.

Model	MinTotalSize	MinTrimSize	OverlapCutoff	TimeFrame
ALT0	8	8	0.8	1.4
ALT1	12	12	0.95	2
ALT2	7	7	0.7	1

Table S1: Our subtree selection algorithm uses three parameters: a minimum subtree size, a maximum allowable overlap and the length of the time window (Figure 1). We denote our default as "ALT0" and our alternatives as "ALT1" and "ALT2". There are natural trade-offs: a larger minimum size, lower overlap and longer time frame all result in fewer accepted subtrees. We found good performance for each of these three alternate setups.

5.2 Cross-validation on different experiments

To determine how variable the AUC and accuracy figures are for the analyses reported in the main text, we performed 10-fold cross validation for the main prediction on H3N2, the analysis on H1N1 only (trained on a portion of the H1N1 subtrees), the pooled analysis of H3N2 and H1N1 and the predictions on H1N1 using a model trained on H3N2. The resulting ROC curves are shown in Figure S3, (a)-(d) respectively. AUC values are consistently above the random classifier's expected value of 0.5, with ranges given in the caption of Figure S3. It is particularly encouraging that the test on H1N1 (Figure S3(d)) has a highly

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

consistent set of AUC and accuracy values (range of AUC 0.75-0.88) because this test is in some sense the most challenging; no subtree in the test data (H1N1) ever shares a tip with a subtree in the training data (all H3N2) and the risk of overfitting is low (also note that we did not select a model or method based on this test). The most variable set of AUCs arises from H1N1 alone, which is likely due to the lower volume of data.

5.3 Time slices

To ensure that the method does not rely on internal nodes whose existence depends on the full dataset, we reconstructed the influenza H3N2 tree using only the sequences observed prior to 2017.5; call this tree $T_{2017.5}$. We extracted the subtrees of this tree using the ALT0 parameters. Naturally, the 2017.5 tree does not contain the information as to whether its later-occuring subtrees grow into the following season. To find the remaining success information, we used the tree reconstructed from the sequences any time up to 2018.5 ($T_{2018.5}$. Consider a subtree c in the $T_{2017.5}$, with tip set S_c and size n. First we find the most recent common ancestor (*MRCA*) of S_c in the $T_{2018.5}$. Then, we compare the size of the subtree c (n) with the size of the subtree rooted at the MRCA of S_c in $T_{2018.5}$ (m). If $m > \alpha n$ ($\alpha \in \{1.3, 1.5, 2\}$) then we say that subtree c is successful (see Figure S4). Again we tried different α cutoffs to obtain a balanced dataset (see Figure S4). We randomly choose 75% of the subtrees for training our model and leave the rest to test the model. In order to find the hyper-parameters of the model we performed 10 fold cross-validation on the training set. We tried *linear*, *radial* and *polynomial* kernel of SVM and, among these, the linear kernel resulted in the best performance (data not shown).

5.4 Best Features

For robust feature selection we followed the ensemble technique introduced in [43]. They show that combining multiple (unstable) feature selectors yields more robust feature selection than using a single selection method. We use 4 models including logistic regression, random forests, SVM with linear kernel and learning vector quantization (LVQ)[20] to rank the features based on their contribution in the classification task. In the general classification of the H3N2 subtrees, the epitope features are among the most important. However, classification based only on epitope features reduces the AUC from 0.89 to 0.72, and our classifiers perform well on H1N1 (0.86 AUC, and 0.76 average AUC in the 10-fold cross validation) despite not having the epitope features. No single feature or small group of features that we have identified can perform as well as the combined phylogenetic and epitope features. We did not attempt to reduce the feature set to obtain a minimal set of features with the optimal performance.

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

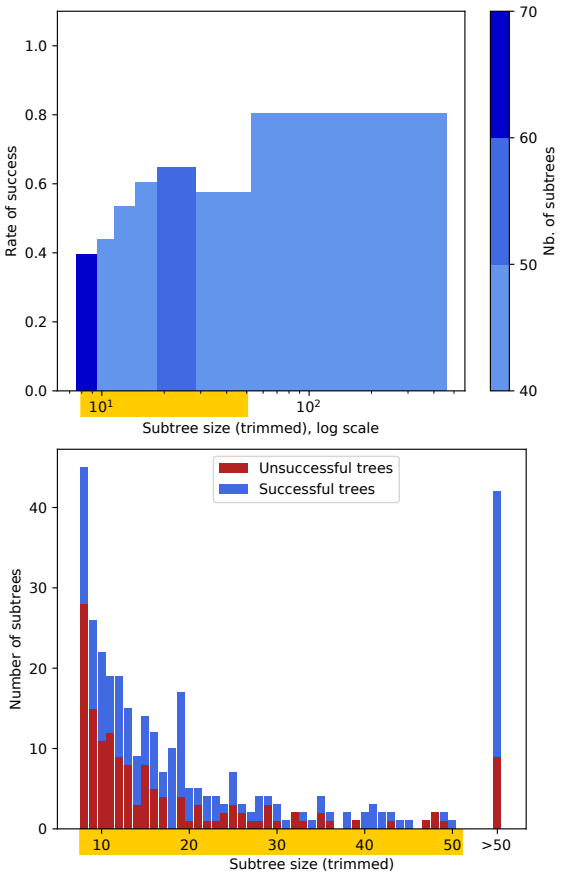


Figure S1: (Top) Correlation between the size of trimmed subtrees (x-axis) and the rate of success (y-axis) for the 391 subtrees from the H3N2-HA dataset. The rate of success is defined as the number of succesful trees divided by the total number of trees for a certain range of sizes. The ranges were computed in order to encompass approximately the same number of subtrees, and the color of the bars represent how many subtrees were taken into account for the computation of the success rate. The subtrees vary from size 8 up to 460, but as most of the dataset is composed by small subtrees, we used a log scale to better visualize the information. The subtree sizes highlighted in yellow are detailed in the bottom panel. (Bottom) Subtree size distribution for trees up to size 50, which corresponds to 87.4% of the dataset. For each size, the graph shows how many subtrees were succesful (blue blar) and unsuccesful (red bar).

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

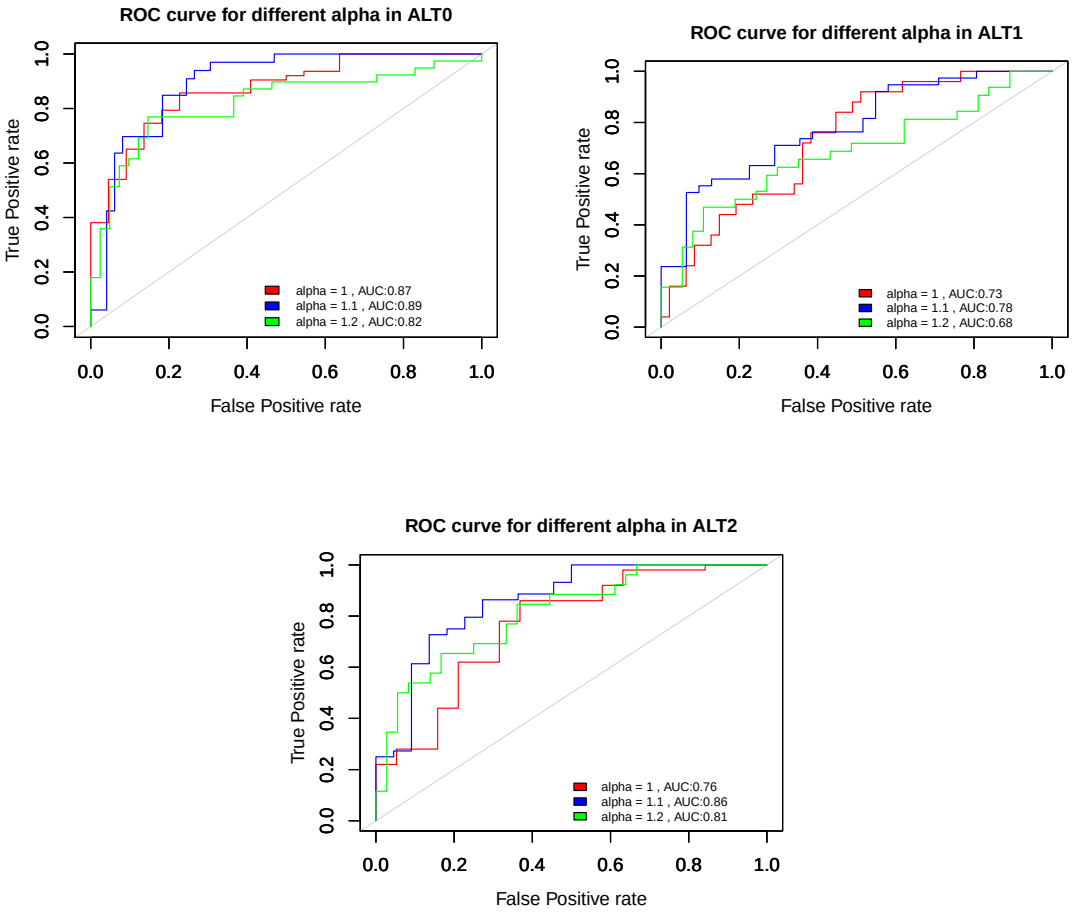


Figure S2: The ROC curves from prediction for the H3N2-HA dataset using different models and different definitions of a successful subtree. In all models, setting the threshold to an appropriate value allows a balance between successful and unsuccessful outcomes, resulting in better performance. Among different sets of parameters for extracting the subtrees, ALT0 is the most powerful model resulting in 0.89 AUC. In all of these experiments we used the topological and epitope features.

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

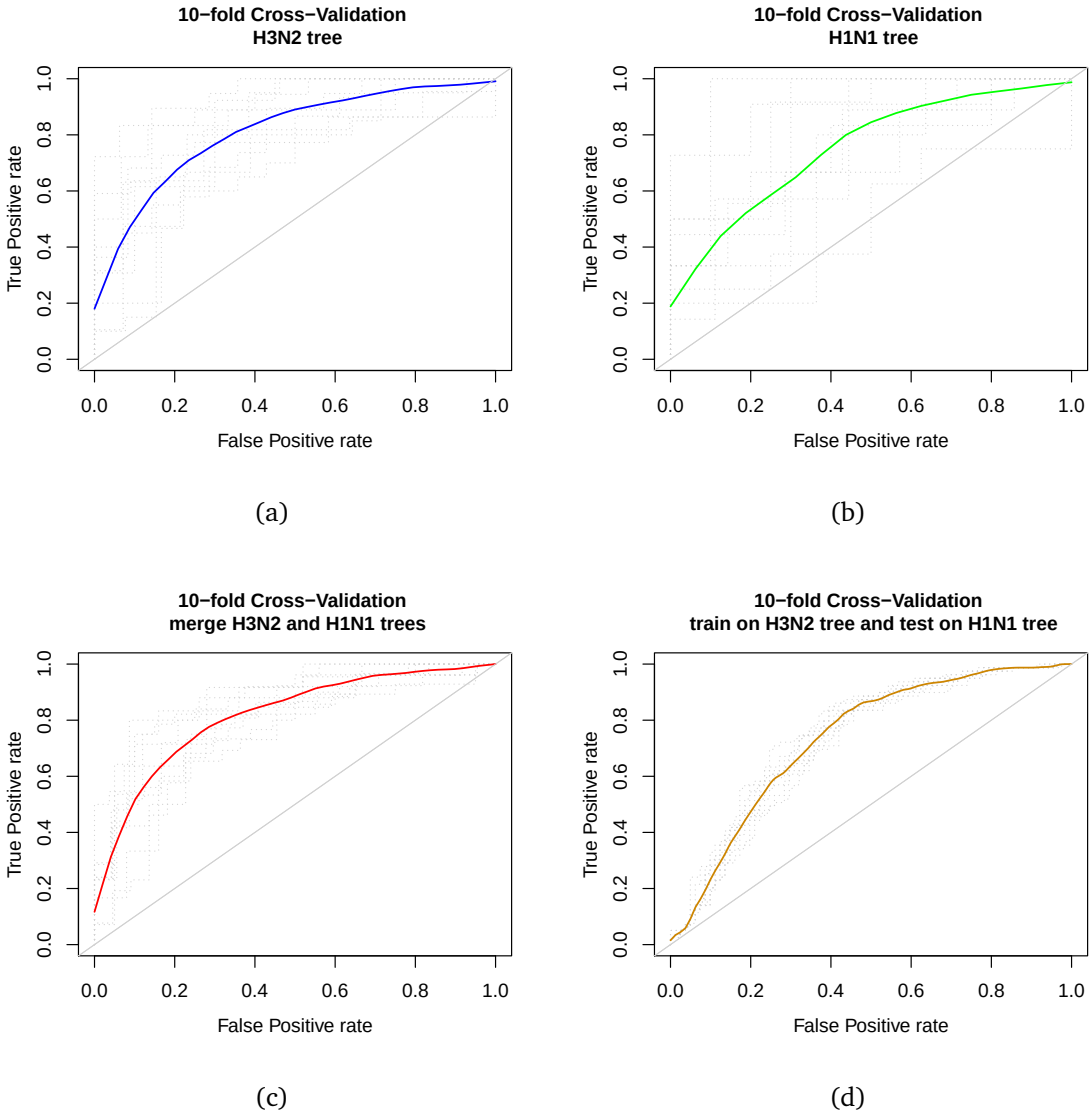
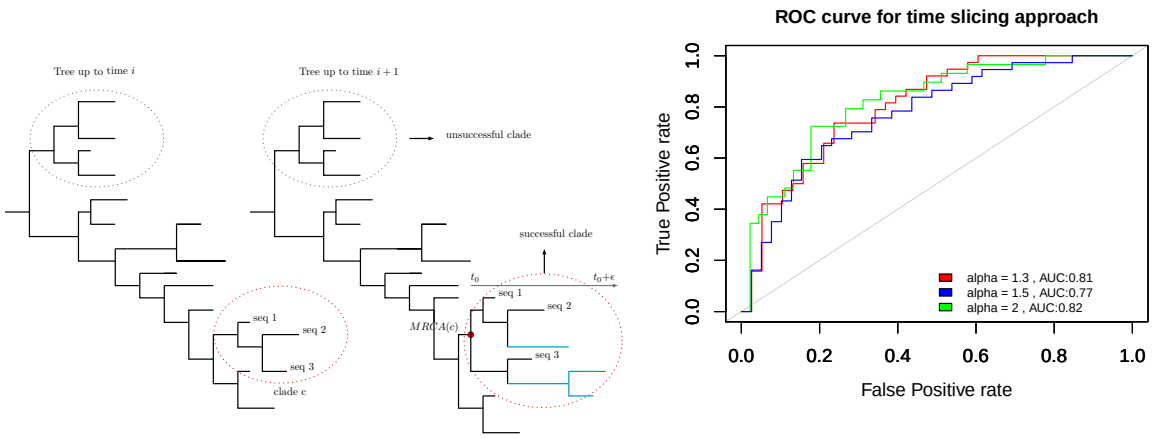


Figure S3: (a) The result of 10-fold cross validation for *SVM* models trained on the subtrees whose ancestral nodes occurred pre-2015-1 or which began later but whose outcomes are known. AUC values range from 0.73 to 0.90 (average 0.82) with 80% of the folds resulting in more than 0.75 AUC. (b) 10-fold cross-validation on the H1N1 tree using *SVM* with linear kernel and a set of topological properties of the clades. AUC values range from 0.52 to 0.95 (average 0.76) with AUC more than 0.75 in 70% of the folds. (c) The result of 10-fold cross-validation on the merged dataset of H3N2 and H1N1 trees. The minimum and maximum values of AUCs among the folds are 0.75 and 0.88 respectively (average 0.82). (d) 10-fold cross-validation for training on the H3N2 tree and testing on the H1N1 tree; we removed 10% of the training data at each fold and evaluated the model on the test data. AUC values range from 0.72 to 0.75 (average 0.74).

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



(a) (b)

Figure S4: (a) This figure depicts the definition of a successful subtree in the time slicing approach. The tree on the left, T_i, represents a tree reconstructed from a set of sequences up to time i and the tree on the right shows the same tree after one year (T_{i+1}). We compare these two trees to predict the successful subtrees in the time slicing approach. For each subtree c we compare the size of the subtree in T_i (X) with the size of the subtree rooted at the most recent common ancestor of the tips of subtree c in tree T_{i+1} (Y), in an interval of 3.4 years following the root of the subtree. If $Y > \alpha X$ ($\alpha \in \{1.3, 1.5, 2\}$) then we say subtree c is successful. This overcomes the challenge that T_{i+1} does not contain the root of c; it does contain a node that is the MRCA of the tips in subtree c. (b) We classified the subtrees of the influenza H3N2 tree reconstructed from the sequences up to 2017-5 by comparing with the tree reconstructed from sequences up to 2018-5. In this figure the result of classification using different ratio between the size of a subtree in H3N2-2017 and the size of the corresponding subtree in H3N2-2018 is shown.

bioRxiv preprint doi: https://doi.org/10.1101/609248; this version posted April 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

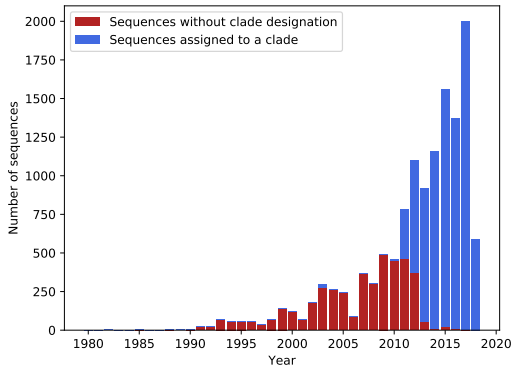
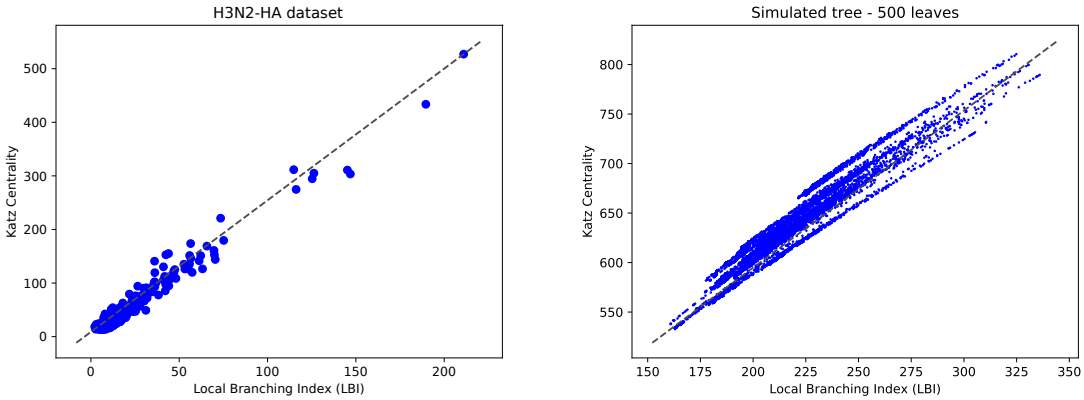


Figure S5: Distribution of H3N2-HA sequences between 1980 and 2018. The blue bar corresponds to sequences assigned to one or more clades, whereas the red bar corresponds to sequences without clade designation. The rate of unassigned cases drops considerably for recent sequences (from 2011 on), which encompass most of the analyzed dataset (73.45%).



(a) (b)

Figure S6: Correlation between Local Branching Index (LBI) and Katz Centrality measures (weighted version, where the distance between two nodes is defined as the sum of the branch lengths in the path separating them). (a) Each point corresponds to the LBI and Katz Centrality measures computed for the root of the 391 subtrees from the H3N2-HA dataset. Parameters used: $\tau = 50$ (LBI), $\alpha = 0.95$ (Katz Centrality). (b) Correlation for 10 trees simulated with a pure birth process. All trees have 500 leaves, and each point corresponds to the LBI and Katz Centrality measures computed for a node of those trees. Parameters used: $\tau = 10$ (LBI), $\alpha = 0.95$ (Katz Centrality).