# Testing the Validity of the Use of Genealogical Trees to Predict Virus Evolution

**By Dominique Hughes, Nur Izzati, Ling Yang**

**Mentor: Nicole Fider**

**Math 485**

**Midterm Report**


Figure 1: Image adapted from [1].

## Introduction

The flu vaccine is a prediction of the virus mutation in a given flu season. Because viruses mutate randomly, it is not easy to predict the strains of a virus that may circulate during an outbreak, and researchers can mistakenly predict the flu strains for a given year [2]. In order to predict this, we require knowledge not only about the genetic mutations but also the ways that those mutations affect the "fitness" of the virus. The fitness of a virus is the likelihood that it will continue to survive over time. But what makes a virus fit? And how do we know whether a given mutation will increase or decrease a virus fitness levels?

We could look at the RNA of the flu virus but in order to analyze that we would need a complex understanding of the function of each portion of the RNA. Luckily for us, we do have extensive knowledge about the flu and the ways amino acid sequences affect its phenotype. For most viruses, however, we do not have the genetic information necessary to predict the fitness of various strains [3]. This is why it is critical that we look to other ways to predict future virus strains, so that we can accurately create vaccines and protect the population. The ability to predict the outbreak of viruses is extremely useful in the field of medicine, specifically in the preparation of vaccines.

## Description of the Model

When viruses multiply, they copy their genetic material to make clones. However, the genetic material of clones might be different from the original due to changes during the cloning process. In rare selective sweeps (Figure 2A), these changes dramatically increase the fitness of a strain, leaving one strain far more fit than the others. In continuous adaptation (Figure 2B), small changes to the genetic material create small increases in fitness for various strains as they all

adapt. However, after many generations, these changes would cumulatively lead to large differences in the genetic material, thus it is important to track and identify which strains are most likely to continue and evolve.

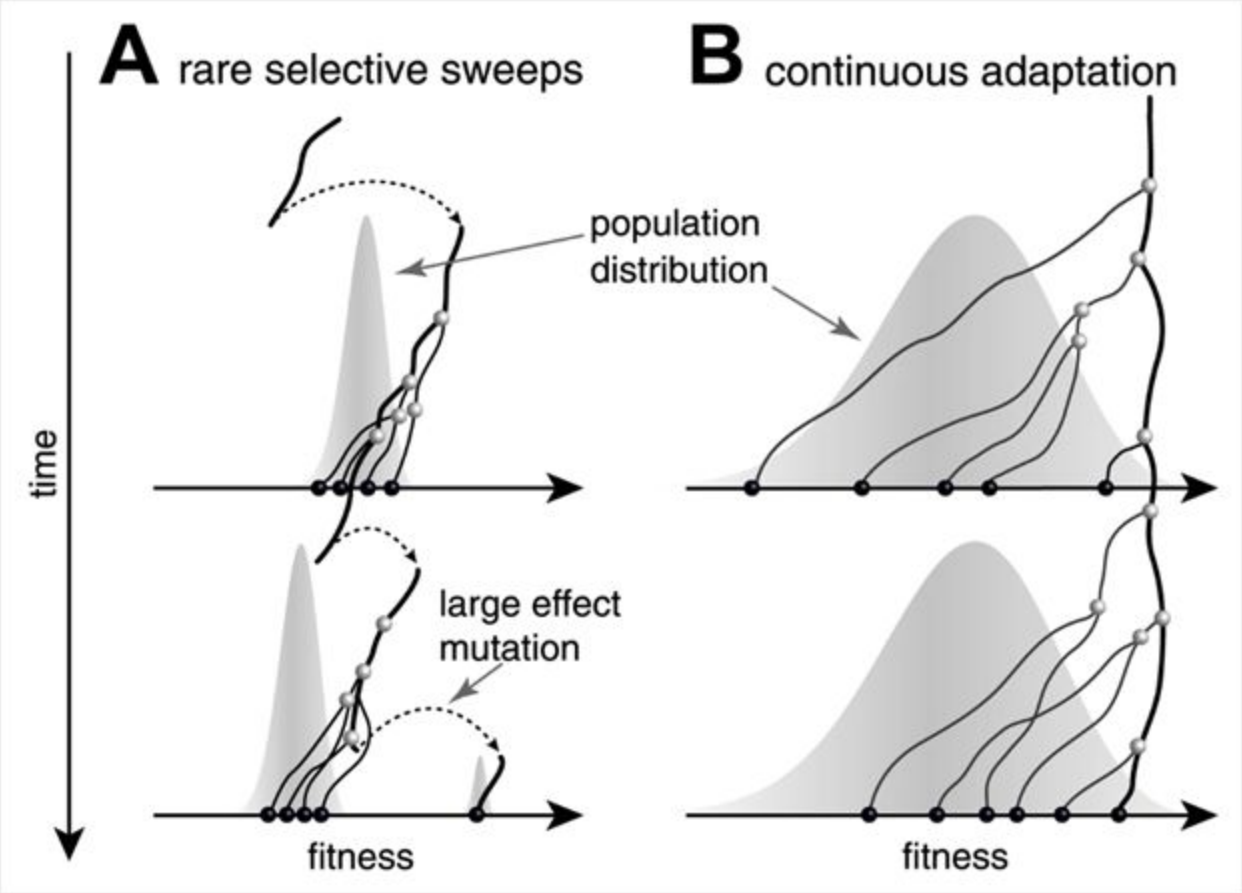


Figure 2: This figure demonstrates the difference between rare selective sweeps and continuous adaptation. Image adapted from [3].

The model created by Neher, Russell, & Shraiman operates under the assumption that most flu viruses occur due to continuous adaptation [3]. This model also operates under the assumption that the virus population is under persistent directional selection (i.e. random mutations that affect the viruses' ability to persist in a given environment, such that viruses whose mutations allow the virus to thrive in said environment are more likely to survive and if the mutation makes them less likely to thrive than that strain will die out).

Unlike other vaccine predictors that use in depth knowledge of the genetic sequence, this model uses a more general method to predict fitness from virus genetic sequences by first creating a "family tree" for a virus population. The genealogical tree of the virus population is constructed by recording the different strains of virus at different times and using those samples to construct a tree which shows likely ancestry patterns connecting the observed strains. This can be an incredibly powerful tool to understand the evolutionary track of a species [4]. Neher et al. has created a mathematical model that uses the shape of the tree and the branching of the tree to rank which virus strain is the most fit and therefore the most likely to be the progenitor of the next year's virus [3]. For example, N2 (Figure 3) would be more fit than N3 based on the extensive branching in comparison to N3. This is called the inferred fitness model (which we abbreviate IFM). Neher et al. also created a second model that is reliant only on branch length called the Local Branching Index (LBI) that they used to corroborate the IFM [3].

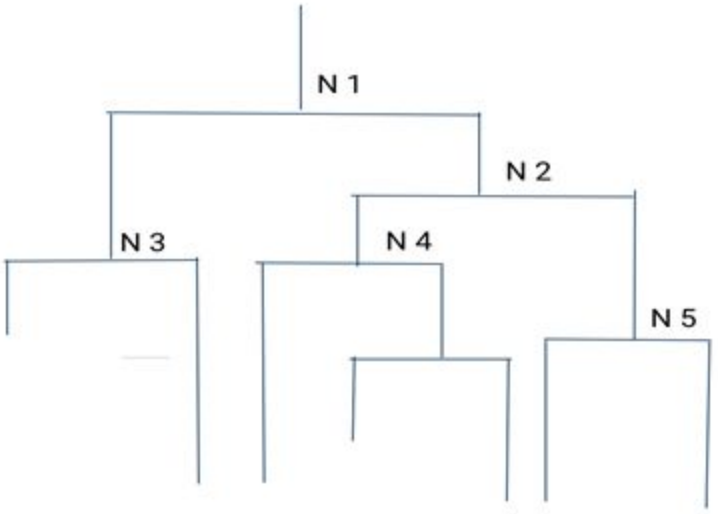


Figure 3: This figure shows an example of a genetic tree with node labelling.

## The Mathematical Model

Neher et al. includes in their article the mathematical derivation for both the IFM and the LBI [3]. For brevity, we have only included the final equations for the IFM and the LBI; for a description of the derivation, please look at the Materials and Methods section of their paper, "Predicting evolution from the shape of genealogical trees" [3].

$$P(x|T) = \frac{p_0(x_0)}{Z(T)} \prod_{i=0}^{n_{int}} g(x_{i1}, t_{i1}|x_i, t_i)\, g(x_{i2}, t_{i2}|x_i, t_i) \tag{1}$$

Equation (1) is the IFM equation, where $p_0(x_0)$ is the fitness distribution in the population, and index $i$ runs from 0 through all the internal nodes. The indices $i1$ and $i2$ denote the two children of node $i$, while $Z(T)$ ensures normalization of the distribution. The branch propogatory $g(x_j, t_j|x_i, t_i)$ describes the likelihood of the lineage to connect an ancestor with fitness $x_i$, at time $t_i$ to a child with fitness $x_j$ at a later time $t_i$ (child in a sense of a subclade in the tree, rather than direct offspring, meaning the virus strains connected to a previous virus strain through direct branching). Essentially, this model infers the fitness of nodes and determines how likely they are to be connected. This allows them to determine likely progenitor strains for a given flu year; they look at the ranking of the node fitnesses and then argue that the most fit strain is the likely progenitor for a given year.

$$\lambda_i(\tau) = m_{\downarrow i} + \sum_j m_{\uparrow i_j} \tag{2}$$

Equation (2) is the local branching index (LBI). The LBI was used to corroborate the IFM. The *m* terms in the LBI model are iterative processes that determine the fitness of the nodes by

incorporating data on the progenitor nodes along a branch, or the descendent nodes along the branch. This model essentially uses the length of the branches as a determination of fitness, which makes sense as fit strains are more likely to survive longer.

## Analysis and Investigation of Model

*Previous Work Completed by Neher et al.*

Neher et al. tested their models on influenza virus data from 1995-2013 and claimed that they were able to predict progenitor strains of the influenza virus with high accuracy with the IFM, because the LBI was able to determine the fitness with similar accuracy to the full fitness inference algorithm [3]. This is a huge step forward because these results did not rely on the specific genetic information and how it relates to fitness. This means that this model could be used to attempt to predict other asexual population mutations where we do not understand the specifics of the genetics yet such as other RNA viruses (which, like the flu, only have RNA as their genetic material), or even cancer clusters.

*New Work in Testing the Model*

We attempted to validate the IFM using real influenza data gathered from the Influenza Research Database (IRD) [5]. We used the code for the IFM provided by Neher et al. on the flu virus strain AH3N2, which was narrowed down according to similar parameters as the strains in the Neher et al. study [6,3]. We used the IRD to find AH3N2 influenza strains from the years 1968 to 2014, in humans. We looked specifically at a portion of the genome, the HA portion of the genome, which is what Neher et al. also looked at when comparing virus data. We grouped this data according to states where the strains were recorded. We then used the code they provided to receive a ranking of the virus strains. The strains are ranked in order of fitness based

on the genetic tree generator the program uses (Table 1) [7]. The highest ranked virus strain is meant to be the progenitor for the next year.

| Name | Rank | Mean | Standard Deviation |
|---|---|---|---|
| A/Arizona/3165/2013 | 1 | 4.535924712858883 | 1.0507880304684343 |
| A/Arizona/17/2013a | 2 | 4.49878380386077 | 1.078765239875353 |
| A/Arizona/17/2013b | 3 | 4.493411722221582 | 1.0812592375411327 |
| A/Arizona/08/2013 | 4 | 4.481323382503069 | 1.0840298396185797 |
| A/Arizona/11/2012 | 5 | 4.403203016034367 | 1.1212706609593637 |
| A/Arizona/16/2013 | 6 | 4.256500820442292 | 1.1604913892034618 |
| A/Arizona/09/2012a | 7 | 4.172387170159597 | 1.0781140213403744 |
| A/Arizona/09/2012b | 8 | 4.167051937046791 | 1.0806802282469703 |
| A/Arizona/NHRC394047/2013 | 9 | 4.134820211333482 | 1.1219456667155943 |
| A/Arizona/M10/2012 | 10 | 4.020257206987394 | 1.0932912355587239 |
| A/Arizona/M14/2012 | 11 | 4.014900964218635 | 1.0958475862756414 |
| A/Arizona/M18/2012 | 12 | 4.000440472102679 | 1.1008969545951213 |
| A/Arizona/06/2011 | 13 | 3.4264016823924774 | 0.9284119559706921 |
| A/Arizona/03/2011 | 14 | 3.178065907706785 | 1.1950397820214254 |
| A/Arizona/11/2010 | 15 | 3.107614669206091 | 0.9948661048098962 |
| A/Arizona/13/2010 | 16 | 3.0982507010766622 | 1.231901658688871 |
| A/Arizona/M12/2012 | 17 | 3.082002045262339 | 1.2663829907675046 |
| A/Arizona/07/2012 | 18 | 3.0093585554071134 | 1.3008072834235416 |
| A/Arizona/M11/2012 | 19 | 2.9425922366056723 | 1.0869868056117409 |
| A/Arizona/09/2007 | 20 | 2.8806338512790126 | 1.0721440114751448 |
| A/Arizona/10/2007 | 21 | 2.7651530773528337 | 1.176980495189009 |
| A/Arizona/12/2007 | 22 | 2.7651530773528337 | 1.176980495189009 |
| A/Arizona/WRAIR1562P/2009 | 23 | 2.7262024737449426 | 0.7723499246571361 |

| | | | |
|---|---|---|---|
| A/Arizona/11/2009 | 24 | 2.4815285459518135 | 1.1350842785926718 |
| A/Arizona/12/2010 | 25 | 2.3023097269923403 | 1.2397696865434429 |

Table 1: This figure shows the listed ranking of the strains of the AH3N2 influenza virus in humans in Arizona, during the years 1968 to 2014.

In order to validate the ranking, we used the IRD to retrieve the same type of influenza data as before, except now we extended the data collection to the year 2015. We used the IRD to create a phylogenetic tree of this data, so we could see a likely lineage for these strains [5]. This would allow us to determine if the top ranked virus strain from the IFM was indeed a progenitor for future strains. This was done by determining what we called the "genealogical distance" between the top ranked strains and the 2015 strains; this was measured by the number of generations away from a progenitor strain for the 2015 strains. Two example phylogenetic trees generated from the IRD for the data from New Hampshire and Arizona are used to demonstrate this system (Figure 4, Figure 5). The New Hampshire phylogenetic tree demonstrates a high accuracy prediction (Figure 4). The highlighted strain is a direct and immediate progenitor of the 2015 strains for New Hampshire (so this would be considered zero generations away from connection) (Figure 4). The Arizona phylogenetic tree shows a poor prediction, as the top ranked strains are multiple generations away from connection with the 2015 strains for Arizona (as the top ranked strain is three generations away from connection with a progenitor for the 2015 strains) (Figure 5).

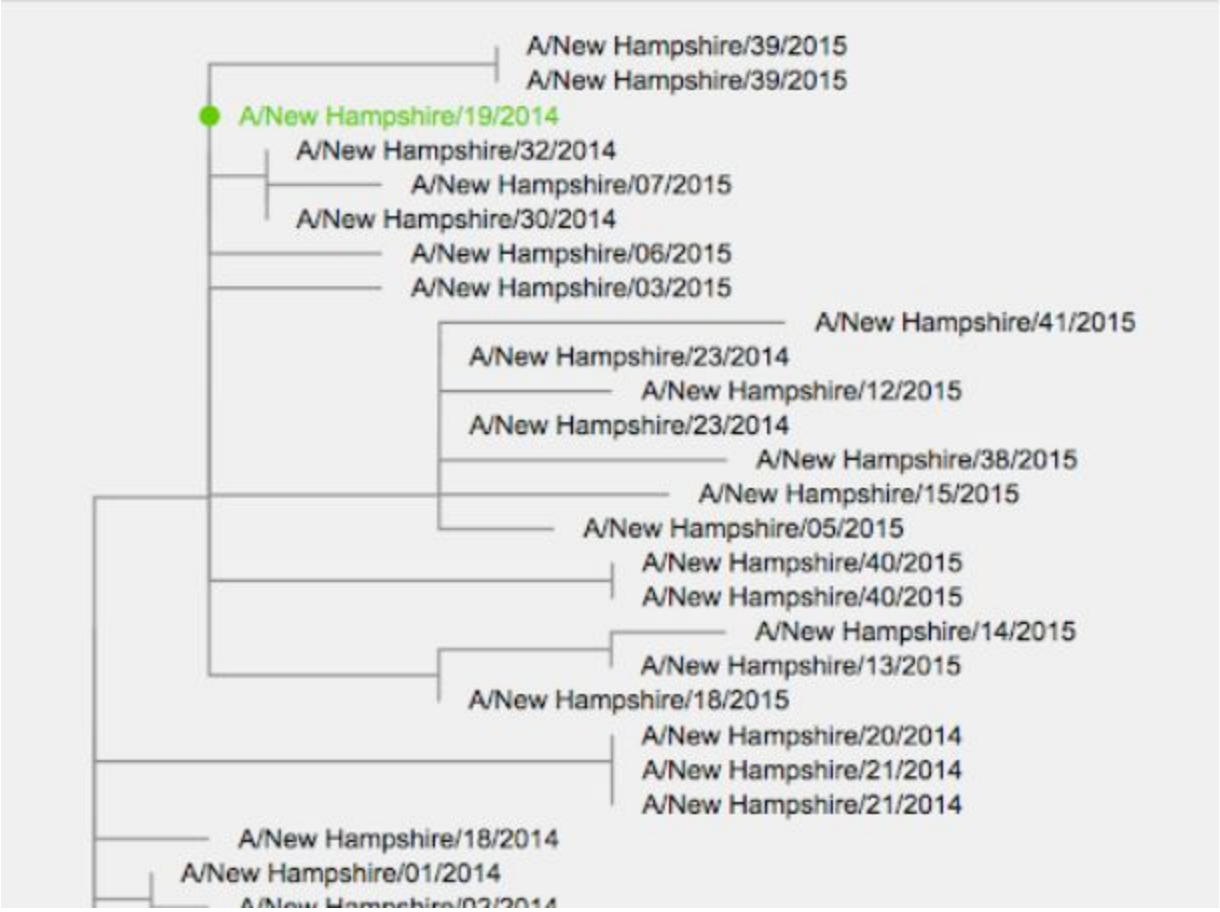

Figure 4: This shows a portion of the phylogenetic tree created for the state of New Hampshire. The top ranked virus strain for New Hampshire is highlighted in green. This strain is zero generations from the 2015 mutations, giving this prediction a high accuracy.

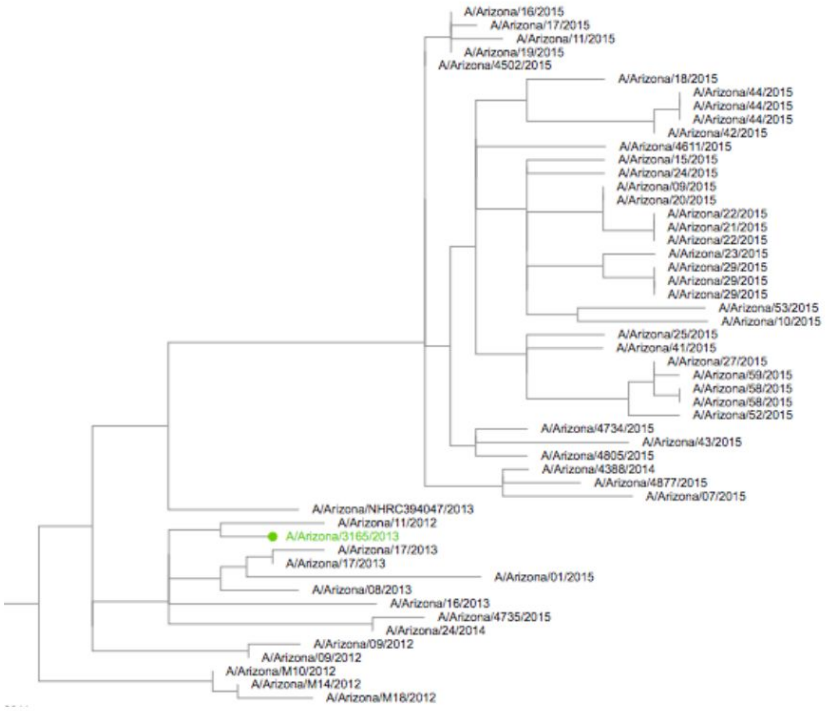


Figure 5: This is a portion of the phylogenetic tree created for the state of Arizona.The top ranked virus strain for Arizona is highlighted in green. This strain is three generations from a connection to the 2015 mutations, giving this prediction a low accuracy.

We used this method of collecting data, finding a potential progenitor, and determining the progenitor's accuracy for 40 states. Some states had to be left out because there was too little data for use, or because there was too much data for the systems to run on our home computers. High accuracy predictions (zero or one generations away) were generated for 65% of the states (Figure 6). Medium accuracy predictions (two generations away) were generated for 10% of the states (Figure 6). Low accuracy predictions (three or four generations away) were generated for 25% of the states (Figure 6).

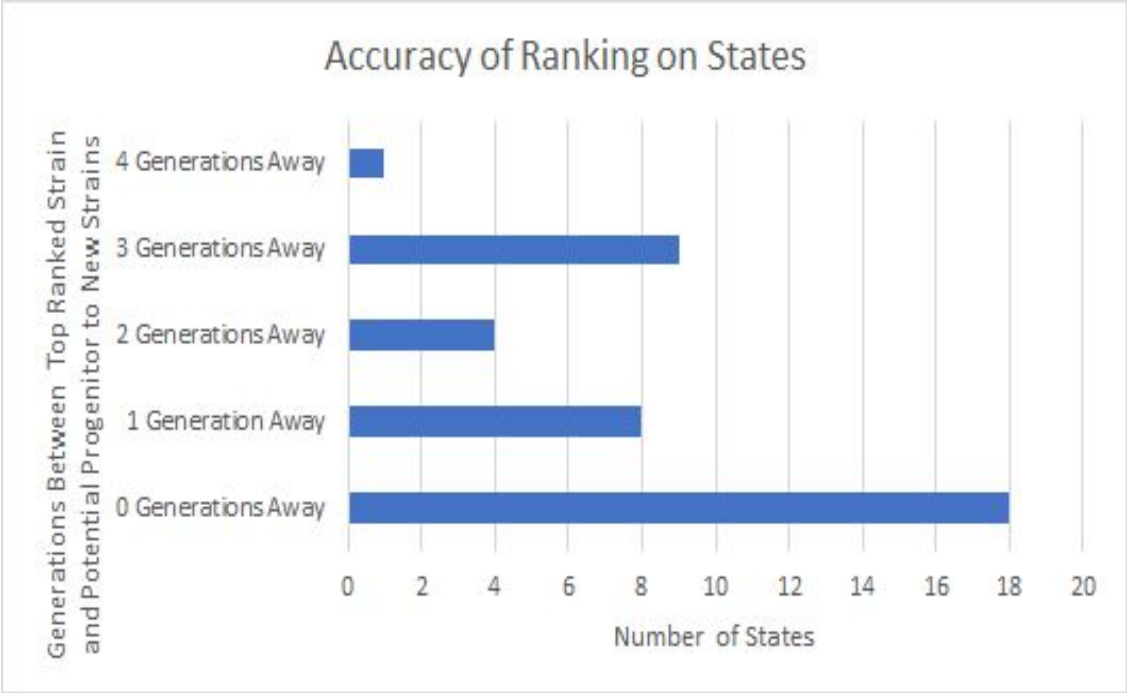


Figure 6: Results of the comparison between top ranked virus strain and it's place in the phylogenetic tree. Strains that are zero generations indicate a high accuracy prediction, whereas strains that four generations away indicate a low accuracy prediction.

## Conclusion

The conclusion of our experiment is that the inferred fitness model is fairly accurate (about 65%). We could not completely reproduce the original experimental settings, so the model may indeed be more accurate if it could be edited as the authors did in specific settings. The program might have had improved accuracy if we had been able to edit it to work more functionally on smaller data sets. Neher et al. state in their paper that they edited their code for use with smaller data sets but did not mention specifically how they had edited it so we were unable to recreate this aspect of their experiment. There was a lot of variability in the amount of data available to us for each state, which can affect how accurate the model and the tree can be (for both, more data generally leads to more accuracy because there is more data to use as parameters). Additionally, the phylogenetic tree is a theoretical method of determining virus

mutations. While it is a likely account of how the virus strains are related to each other, it is not without it's own possibility for error. All of these factors may account for some of the low accuracy predictions.

The Neher et al. method of determining virus evolution is fairly effective and it could have wonderful applications in vaccine prediction and creation. We would require less knowledge about a virus in order to create vaccines which could speed up the vaccine creation process. The vaccines that we are currently creating could potentially even be more accurate as they could be validated with this method as well. We only tested one of the models provided; a possible future direction of this project would be to test the accuracy of the secondary model, so that there is more than one method available to confirm virus prediction methods. We think everyone can see how applicable this could be, especially given the current pandemic.

## References


1. Flu views: Should you get a flu shot? (n.d.). Retrieved from https://www.aoa.org/news/clinical-eye-care/flu-views-should-you-get-a-flu-shot
2. Branswell, H. (2019). Flu Vaccine Selections Suggest This Year's Shot May Be Off the Mark. Retrieved from https://www.scientificamerican.com/article/flu-vaccine-selections-suggest-this-years-shot-may-be-off-the-mark/
3. Neher, R. A., Russell, C. A., & Shraiman, B. I. (2014). Predicting evolution from the shape of genealogical trees. *ELife*, *3*. doi: 10.7554/elife.03568
4. Gregory, T. R. (2008). Understanding Evolutionary Trees. *Evolution: Education and Outreach*, *1*(2), 121–137. doi: 10.1007/s12052-008-0035-x
5. Influenza Research Database - Influenza genome database with visualization and analysis tools. (2020, March 22). Retrieved from https://www.fludb.org/brc/home.spg?decorator=influenza
6. Rneher. (2015, May 7). rneher/FitnessInference. Retrieved from https://github.com/rneher/FitnessInference
7. Price, M. N. (n.d.). FastTree. Retrieved from http://www.microbesonline.org/fasttree/