

# **Phylogenetics for Predicting Virus Evolution**

---

**Anticipating next seasons influenza strains**

**Paul Wiesemeyer**

## Problem



HOUSTON,  
we have a — CHOO!

- Estimated 3 – 11% of the population catch symptomatic influenza each season [Tokars et al., 2018]
- Influenza evolves rapidly, evading immune system recognition
- Epidemics and pandemics caused by RNA recombination events
- Science is slow responding to health issues

# Problem

- $\sim 8 \text{ months}$  from submission to publication of a medical paper [AAMC, 2018]
- $\sim 6 \text{ months}$  from vaccine strain selection to distribution:



Figure 1: from [Bedford, 2015]

- compare:  $\sim 1.6 \text{ days}$  doubling time of A/H1N1 (2009) [Mostaço-Guidolin et al., 2011]

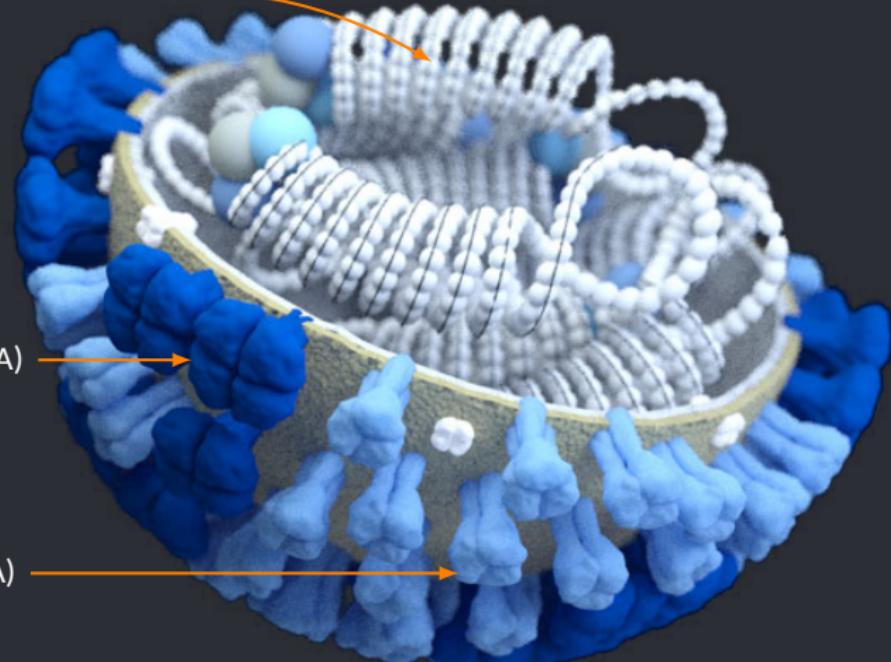
# Outline

1. Influenza
2. Phylogenetics
  - 2.1 The Molecular Clock
3. Predicting the next strain of influenza
  - 3.1 Approaches
  - 3.2 The Hemagglutination Inhibition Assay
  - 3.3 Mapping Antigenicity to the Tree
  - 3.4 Predictions
4. *Nextstrain*
  - 4.1 Hands-On with the Narrative
5. Outlook

# Influenza

*a closer look*

8 single stranded RNA



Neuraminidase (NA)

Hemagglutinin (HA)

Image from [CDC, 2020]

# Influenza jumps far

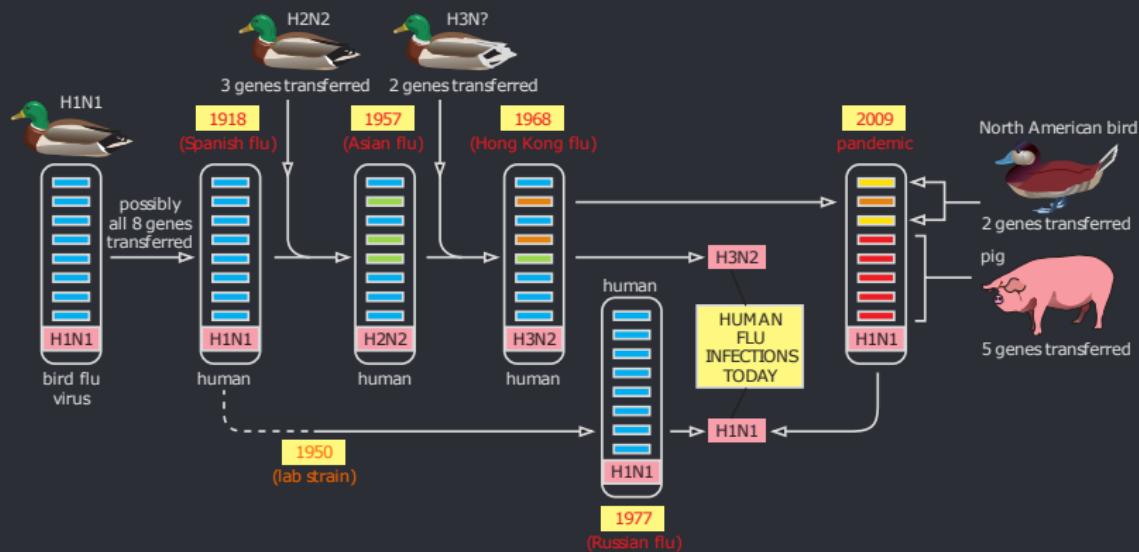


Figure 2: recent influenza A evolutionary shift events [Alberts, 2015]

# Influenza

*runs fast*

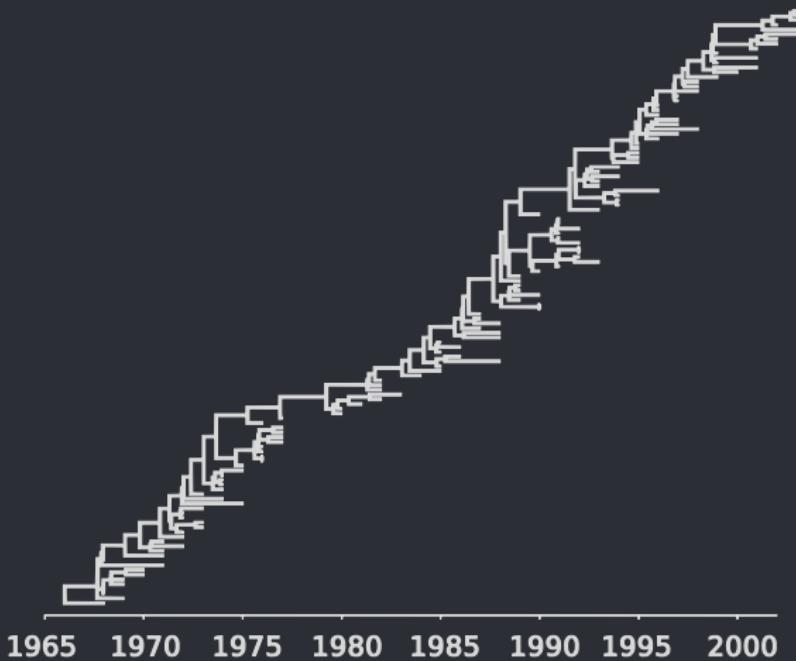


Figure 3: Phylogeny of Influenza A/H3N2 Hemagglutinin [Volz et al., 2013]

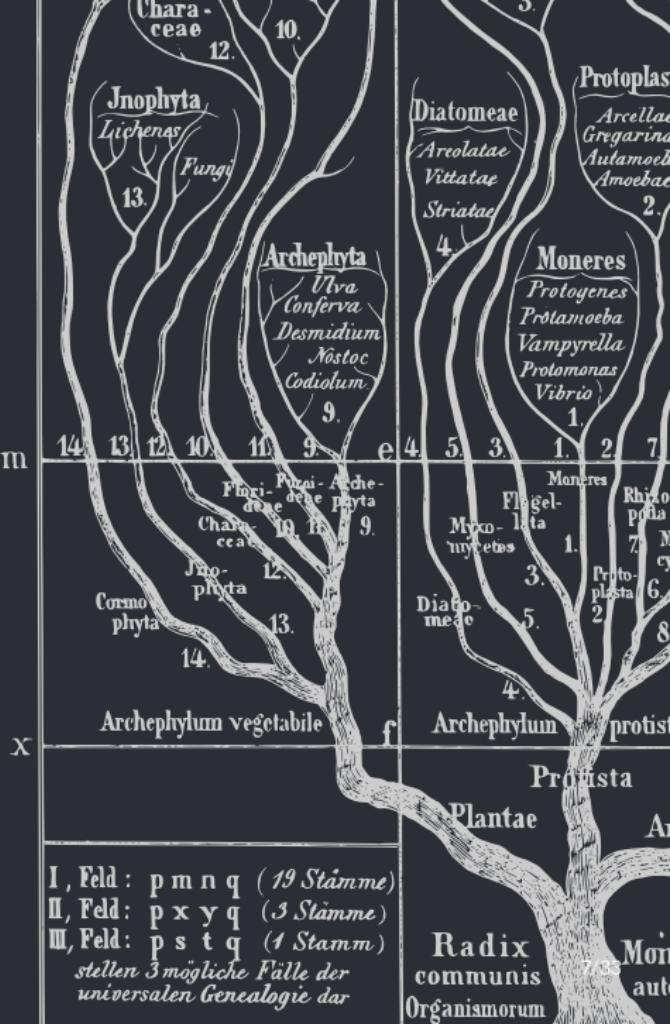
# Phylogenetics

## Terminology

Tree—a simply connected graph

- branches
- nodes (vertices)
- leaves (endpoints)
- root  $\Rightarrow$  parents, children  
 $\Rightarrow$  clades:  
a branch with its children

Image from [Haeckel, 1866]



# Phylogenetics

*Concept: Parsimony*

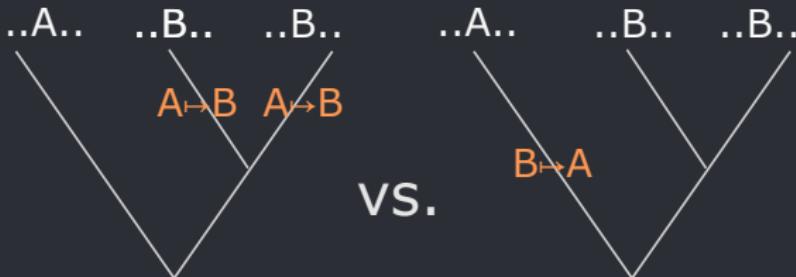


Figure 4: Reconstructing a tree requires likelihood considerations

Minimize the number of mutations to get the most likely tree(s)

Problem: for  $N$  leaves:    # possible trees  $\propto 2^N (N - 1)!$

⇒ We can usually just compute a minute fraction of these.

# Phylogenetics

## *Example algorithm*

	A	B	C	D	E	F
A		9	2	4	9	10
B		9	6	2	10	
C			5	9	10	
D				6	10	
E					10	
F						

$N$  sequences of lengths  $\ell_1 \dots \ell_N$  (not identical due to indels, seq. errors)

- Compare pairwise, assign a distance, make  $N \times N$  table
- Lowest distances group together, form new subunit.
- Iterate. When comparing to sets of sequences, use arithmetic mean.

Result: small number of locally most likely trees,  
not necessarily global maximum.

Computing scales with  $N^2$

Better: Markov Chain Monte Carlo

# Phylogenetics

Concept: *The Molecular Clock*

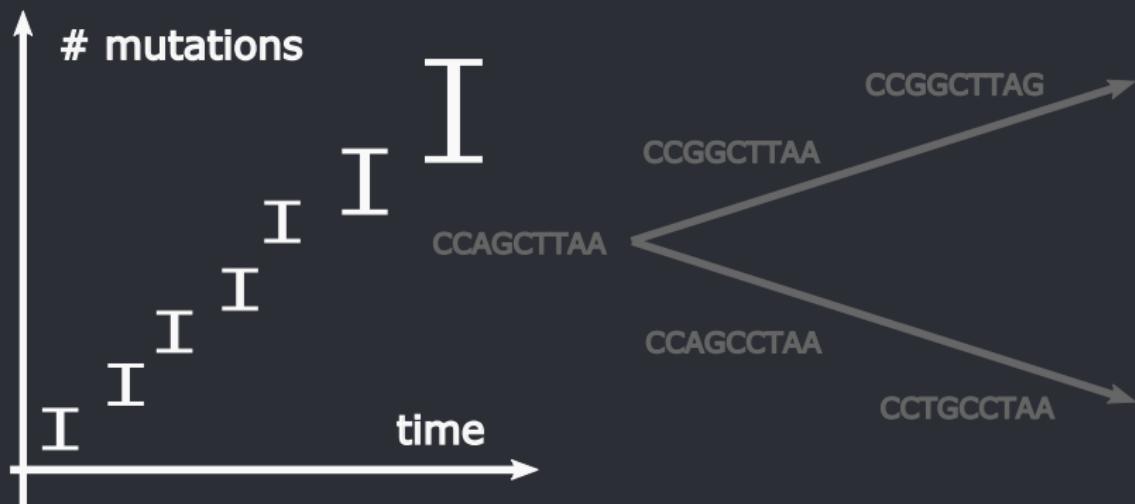


Figure 5: Linear time-mutation relationship

# Phylogenetics

Concept: The Molecular Clock

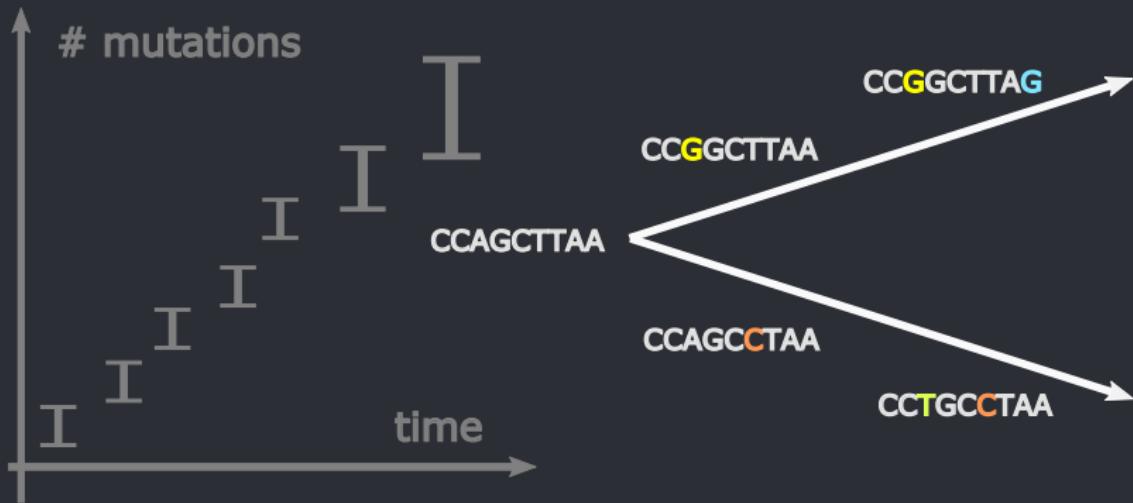
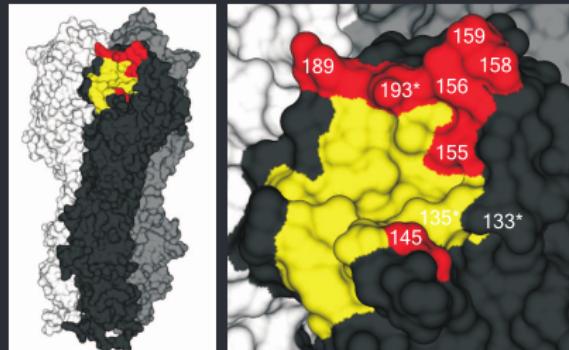
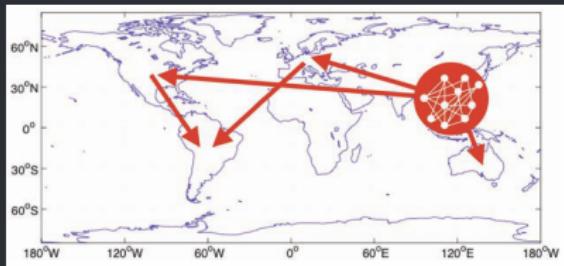


Figure 5: Linear time-mutation relationship

Distinguish between *silent* and *non-silent*—amino acid changing—mutations, the latter are under evolutionary pressure.

# Predicting the next strain of influenza

## Approaches



Epidemiology: Geographical predictor?

Molecular Biology: Is there a telltale mutation site?

Immunology: Hemagglutinin inhibition cartography

Vision: Bring these levels together

# The Hemagglutination Inhibition Assay

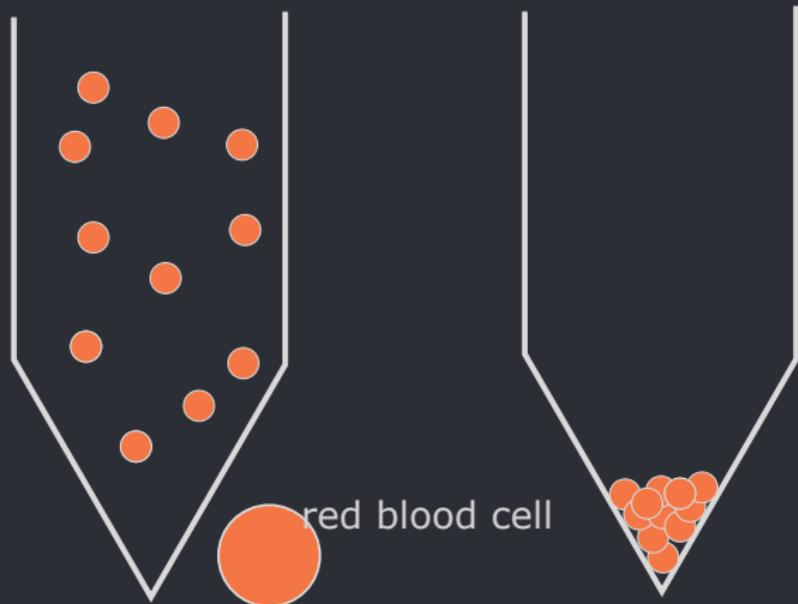


Figure 6: Red blood cells (RBC) precipitate.

## The Hemagglutination Inhibition Assay

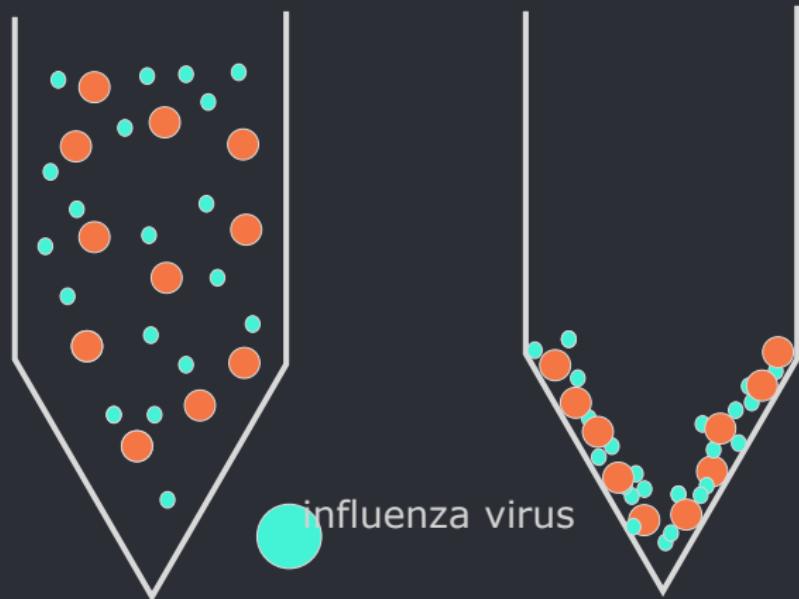


Figure 7: Influenza Hemagglutinin (HA) coagulates the RBC, forming a mat.

## The Hemagglutination Inhibition Assay

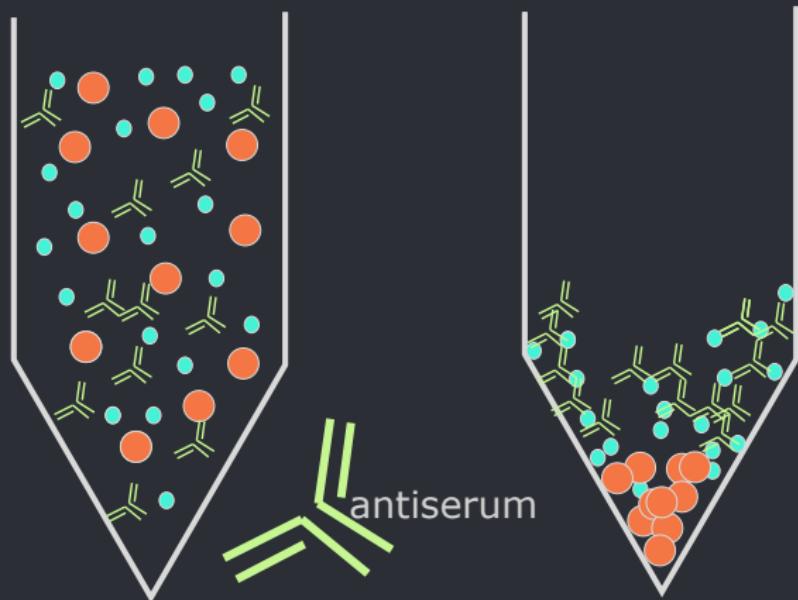


Figure 8: Antisera that fit the HA's epitope site bind to it, letting the RBC sink to the bottom. Effect works up to a certain antigenic distance and antiserum concentration.

# The Hemagglutination Inhibition Assay

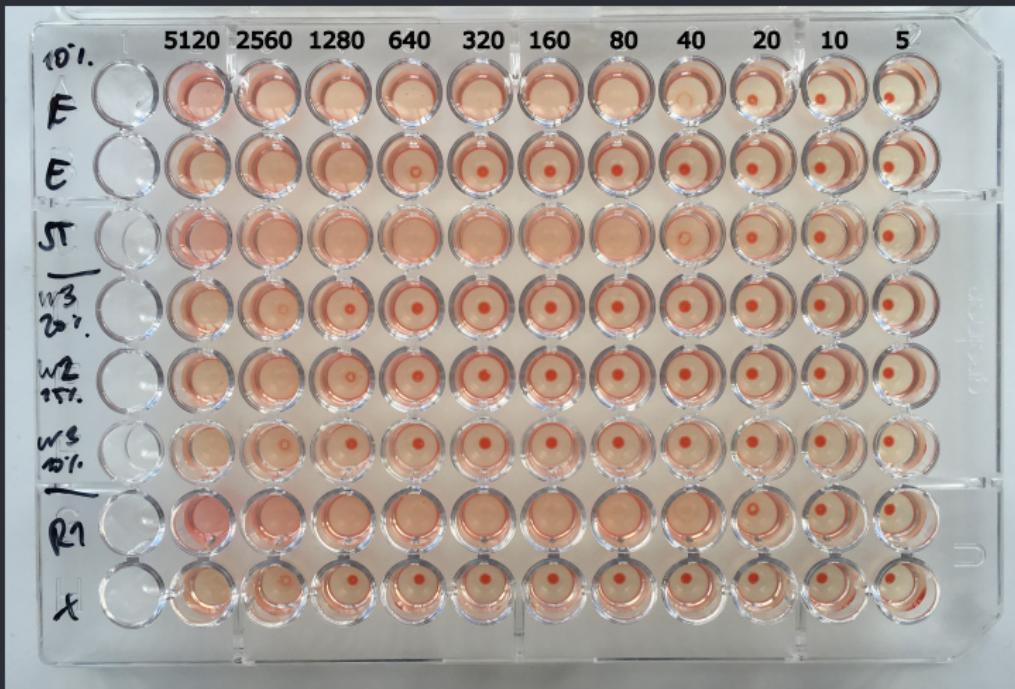


Figure 9: Test 12 different dilutions of one antiserum against 8 different virus strains. The highest dilution that prevents agglutination is called the titer.

# The Hemagglutination Inhibition Assay

*How this used to be looked at*

Post infection ferret sera				
A/Bris	A/Uru	A/HK	A/Perth	
10/07	716/07	1985/09	16/09	
F29/08	F26/08	F21/09	F25/09	

**REFERENCE VIRUSES**

A/Brasilia/10/2007	2560	2560	80	<
A/Uruguay/716/2007	1280	2560	<	<
A/Hong Kong/1985/2009	80	160	1280	640
A/Perth/16/2009	<	40	640	640

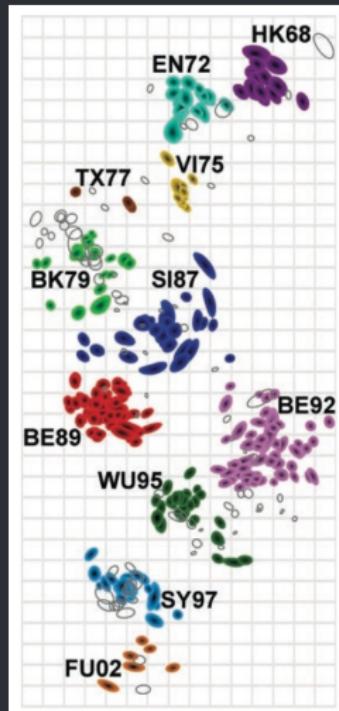
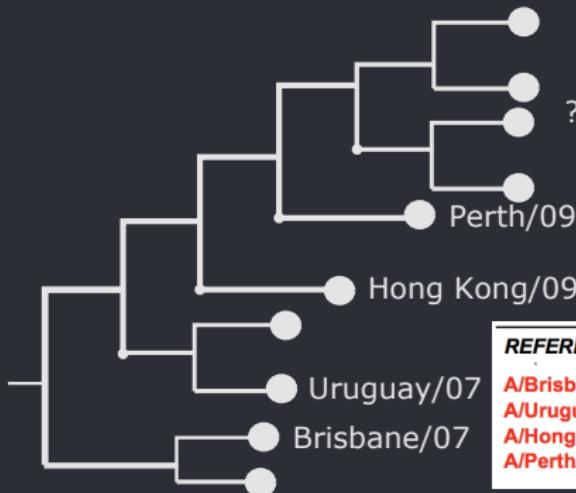


Table from [Bedford, 2015],

cartography from [Smith et al., 2004]

(one gridline is one  $\log_2$  titer i.e. two-fold dilution)

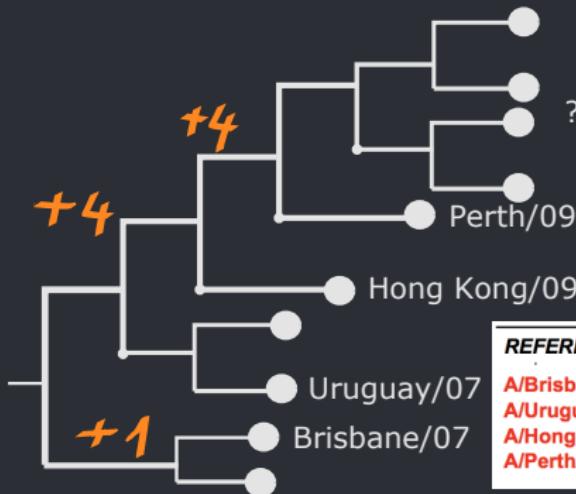
# Mapping Antigenicity to the Tree



REFERENCE VIRUSES	A/Bris	A/Uru	A/HK	A/Perth
A/Brisbane/10/2007	2560	2560	80	<
A/Uruguay/716/2007	1280	2560	<	<
A/Hong Kong/1985/2009	80	160	1280	640
A/Perth/16/2009	<	40	640	640

Figure 10: Mapping the chart to the tree inferred from sequences. [Bedford, 2015]

# Mapping Antigenicity to the Tree



REFERENCE VIRUSES	A/Bris	A/Uru	A/HK	A/Perth
A/Brisbane/10/2007	2560	2560	80	<
A/Uruguay/716/2007	1280	2560	<	<
A/Hong Kong/1985/2009	80	160	1280	640
A/Perth/16/2009	<	40	640	640

Figure 10: Mapping the chart to the tree inferred from sequences. [Bedford, 2015]

# Mapping Antigenicity to the Tree

*A literal toy model:*



Figure 11: [rosipaw, 2010]

## The Model — [Neher et al., 2016] — I

$T_{a\beta}$  ..... HI titer of virus  $a$  against antiserum  $\beta$  (virus  $b$ )

$H_{a\beta}$  .....  $\log_2$  relative titer (we'll use this one)

$$H_{a\beta} = \log_2(T_{b\beta}) - \log_2(T_{a\beta}) \quad (1)$$

$\hat{H}_{a\beta}$  ..... predicted  $\log_2$  relative titer

$v_a$  ..... avidity of virus  $a$  (=greediness)

$p_\beta$  ..... potency of antiserum  $\beta$  (=effectiveness)

$D_{ab}$  ..... genetic component of titer drop

$$\hat{H}_{a\beta} = v_a + p_\beta + D_{ab} \quad (2)$$

## The Model — [Neher et al., 2016] — II

What remains is split up into a sum over individual branch contributions  $d_i \geq 0$ :

$$D_{ab} = \sum_{i \in (a \dots b)} d_i \quad (3)$$

Where the sum is over the path connecting virus  $a$  and virus  $b$  corresponding to antiserum  $\beta$ . Now we want

$$\hat{H}_{a\beta} \stackrel{!}{=} H_{a\beta} \quad (4)$$

to that end we minimize a cost function  $C$  of the whole tree:

$$C := \sum_{a,\beta} (\hat{H}_{a\beta} - H_{a\beta})^2 + \lambda \sum_i d_i + \gamma \sum_a v_a^2 + \delta \sum_\alpha p_\alpha^2 \quad (5)$$

# Predictions of antigenicity

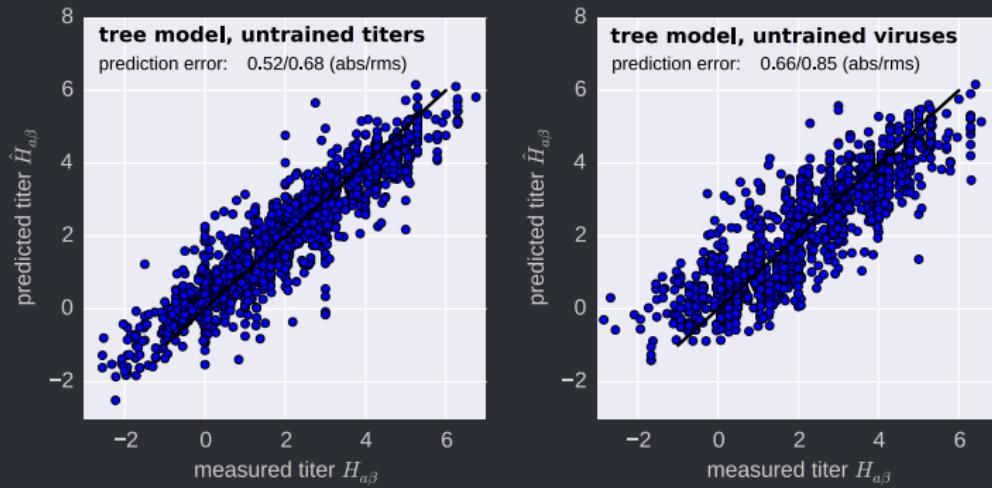


Figure 12: On the left, 10% randomly picked measurements were omitted, on the right, 10 % of entire titer columns were held back, as if it was a new clade. Dataset: Influenza A/H3N2 (12y) [Neher et al., 2016]

# Predictions

*corroborating "tree-likeness"*

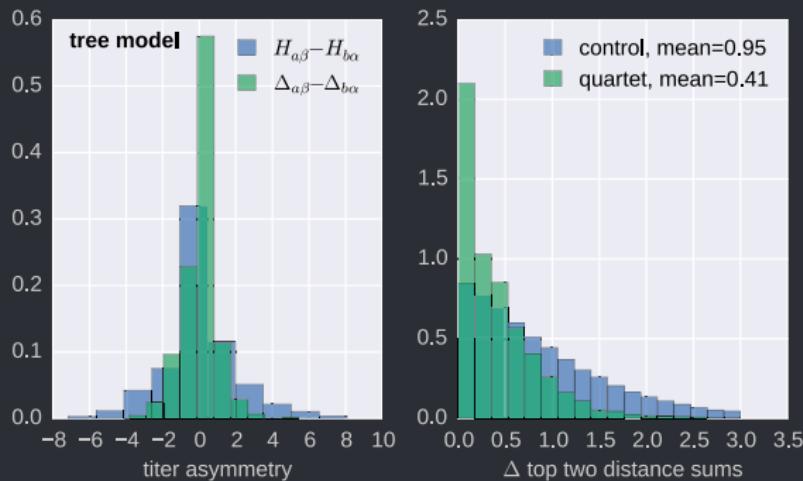


Figure 13: Left: Does subtracting  $v_a$  and  $p_\beta$  enforce symmetry? ( $\Delta_{a\beta} = D_{ab}$ )  
Right: showing tree-likeness, employing the quartet rule. [Neher et al., 2016]

# Predictions

## *fixation implications*

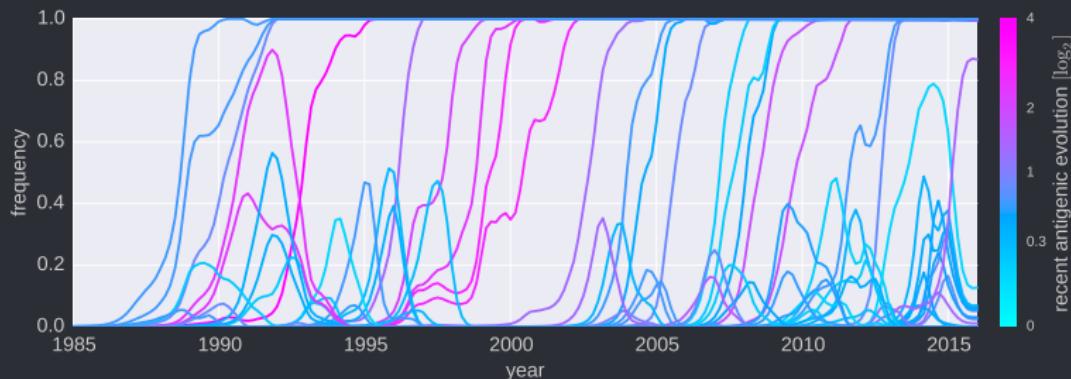


Figure 14: Fraction of samples having a certain mutation plotted over time.  
Strains with frequencies smaller than 0.01 were omitted. [Neher et al., 2016]

# Predictions

*recall:*

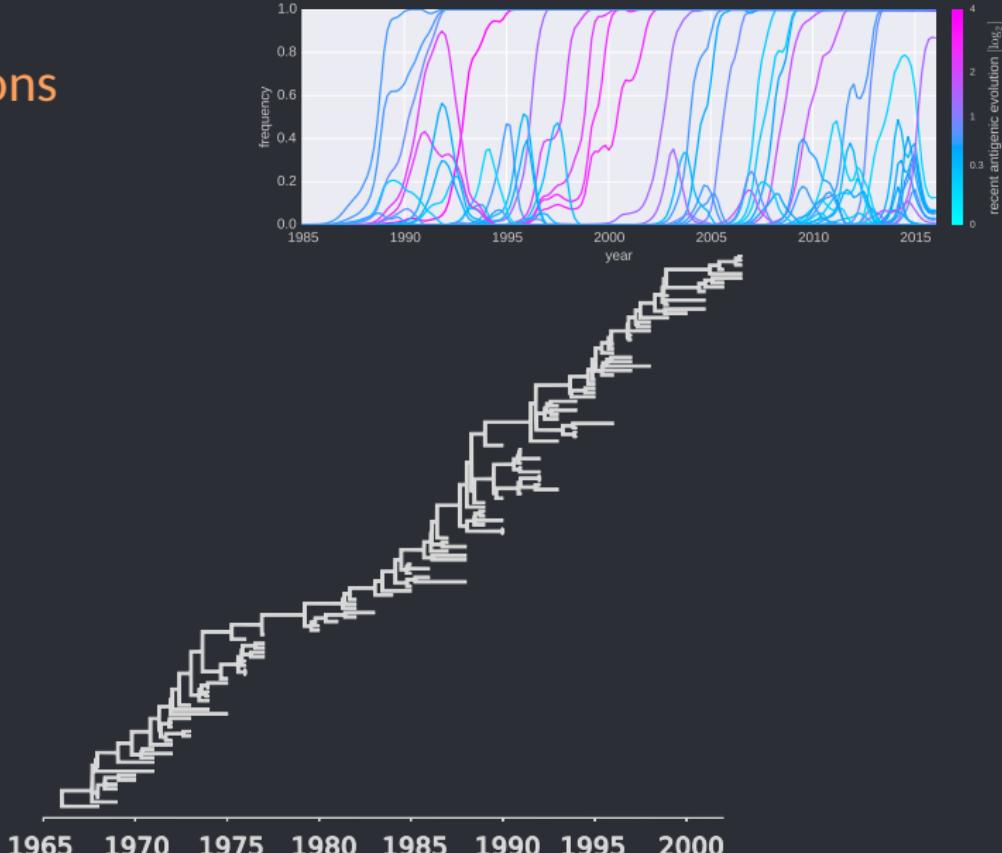


Figure 15: A/H3N2 phylogeny [Volz et al., 2013]

# Predictions fixation jumps

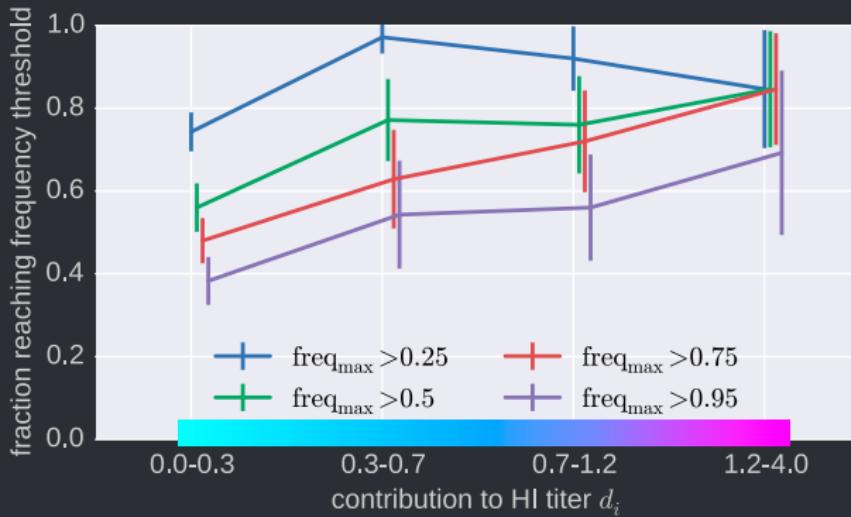
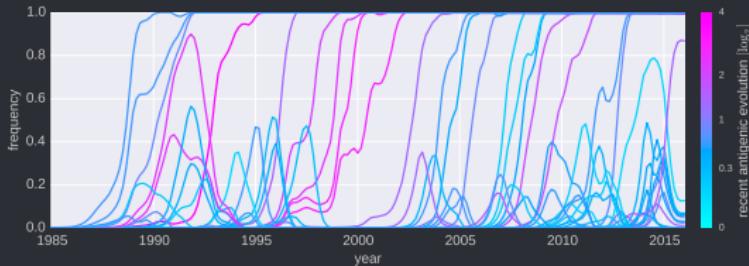
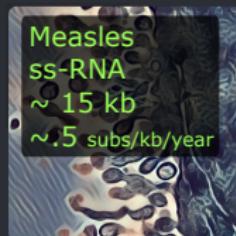
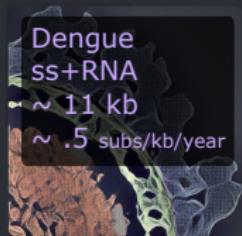
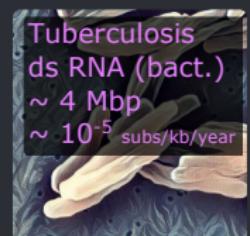
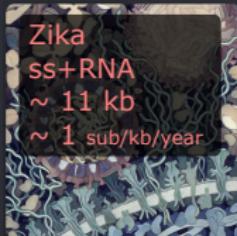
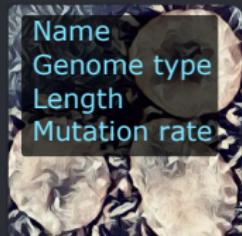


Figure 16: For high recent antigenic evolution traits, 25% prevalence directly entails 75% [Neher et al., 2016]

# Nextstrain

# Nextstrain— A Growing Platform



## *Nextstrain*— Hands-On with the Narrative

<https://nextstrain.org/community/narratives/gitchhiker/virophyle>

# Outlook

*and closing remarks*

- Difficult to disentangle levels, therefore integrate visualization
- Include more meta data (symptoms, severity, & c.)
- Interesting: larger evolutionary timescale
- Include other pathogens, like bacteria, but careful: HGT
- Nice to see instance of Red Queen Hypothesis
- The nextstrain workings were very transparent and user friendly.

# References I

- [AAMC, 2018] AAMC (2018).  
Journal of the Association of American Medical Colleges Frequently Asked Questions.
- [Alberts, 2015] Alberts, B. (2015).  
*Molecular Biology of the Cell*.
- [Bedford, 2015] Bedford, T. (2015).  
Real-time evolutionary forecasting for influenza vaccine strain selection.
- [CDC, 2020] CDC (2020).  
Images of Influenza Viruses | CDC.
- [Haeckel, 1866] Haeckel, E. (1866).  
*Generelle Morphologie Der Organismen. Allgemeine Grundzüge Der Organischen Formen-Wissenschaft, Mechanisch Begründet Durch Die von C. Darwin Reformirte Descendenz-Theorie, Etc*, volume 2.
- [Koel et al., 2013] Koel, B. F., Burke, D. F., Bestebroer, T. M., van der Vliet, S., Zondag, G. C. M., Vervaet, G., Skepner, E., Lewis, N. S., Spronken, M. I. J., Russell, C. A., Eropkin, M. Y., Hurt, A. C., Barr, I. G., de Jong, J. C., Rimmelzwaan, G. F., Osterhaus, A. D. M. E., Fouchier, R. A. M., and Smith, D. J. (2013).  
Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution.  
*Science*, 342(6161):976–979.
- [Lee et al., 2020] Lee, J., Moncla, L., Anderson, B., Black, A., Sagulenko, P., Huddleston, J., Neher, R. A., Bedford, T., Hadfield, J., Megill, C., Callender, C., Fay, K., Potter, B., Bell, S., Hodcroft, E., Sibley, T., Ilcisin, M., and Grubaugh, N. (2020).  
Nextstrain.org.  
<https://nextstrain.org/>.
- [Mostaço-Guidolin et al., 2011] Mostaço-Guidolin, L. C., Greer, A., Sander, B., Wu, J., and Moghadas, S. M. (2011).  
Variability in transmissibility of the 2009 H1N1 pandemic in Canadian communities.  
*BMC Research Notes*, 4(1):537.

## References II

- [Neher et al., 2016] Neher, R. A., Bedford, T., Daniels, R. S., Russell, C. A., and Shraiman, B. I. (2016).  
Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses.  
*Proceedings of the National Academy of Sciences*, 113(12):E1701–E1709.
- [rosipaw, 2010] rosipaw (2010).  
146/365 square peg into a round hole.
- [Russell et al., 2008] Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., de Jong, J. C., Kelso, A., Klimov, A. I., Kageyama, T., Komadina, N., Lapedes, A. S., Lin, Y. P., Mosterin, A., Obuchi, M., Odagiri, T., Osterhaus, A. D. M. E., Rimmelzwaan, G. F., Shaw, M. W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R. A. M., and Smith, D. J. (2008).  
The Global Circulation of Seasonal Influenza A (H3N2) Viruses.  
*Science*, 320(5874):340–346.
- [Smith et al., 2004] Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D. M. E., and Fouchier, R. A. M. (2004).  
Mapping the Antigenic and Genetic Evolution of Influenza Virus.  
*Science*, 305(5682):371–376.
- [Tokars et al., 2018] Tokars, J. I., Olsen, S. J., and Reed, C. (2018).  
Seasonal Incidence of Symptomatic Influenza in the United States.  
*Clinical Infectious Diseases*, 66(10):1511–1518.
- [Volz et al., 2013] Volz, E. M., Koelle, K., and Bedford, T. (2013).  
Viral Phylodynamics.  
*PLoS Computational Biology*, 9(3).