# Unified visualization of seasonal influenza evolution in the laboratory and in nature.

**Allison Black**
**Sarah Hilton**
**John Huddleston**
**Khrystyna North**

## 1 INTRODUCTION

Seasonal influenza is an infectious disease that infects between 10 and 20% of the world's population every year [5]. For some diseases, infection confers immunity that protects that person from subsequent infections for many years. In other cases, such as with influenza virus, the surface proteins on the pathogen change quickly, thereby escaping recognition by antibodies elicited by the previous infection. It is for this reason that the influenza vaccine must be updated almost annually, since the immune response elicited by the vaccine needs to match whichever viruses we expect to circulate during the next influenza season. Vaccination remains the primary public health response effort to control the spread of influenza.

While the surface proteins are the main target for immune response, they are also critical for allowing the virus to gain entry into cells and cause infection. This creates a competing dynamic, in which influenza hemagglutinin must change to evade immunity, but also must stay sufficiently stable to function during infection. These competing needs mean that not every site in the influenza surface proteins can change. Rather, there are some sites that can tolerate different amino acids, and there are others that cannot be changed at all. In order to understand possible evolutionary trajectories for influenza virus, molecular biologists test the ability of the virus to tolerate mutations at different sites in the protein. These experiments, called Deep Mutational Scanning experiments (DMS), use PCR mutagenesis to create every possible amino acid mutation at every site in the gene of interest, in our case, in the influenza hemagglutinin gene. These experiments use deep sequencing to look at the sequence diversity at different sites of the gene under different conditions. At its simplest, DMS data will look at which mutations are tolerated in order to have viruses that can successfully replicate in cell culture. In other cases, we can also introduce serum selection and look at which mutations are tolerated for viral function and allow the virus to escape from human immunity.

Despite the value of these data, they are not intuitive to interpret. Sequencing produces linear sequences, yet the protein conformation may have amino acid residues that are close together in 3D space that are quite distant from each other in linear sequence space. Thus it can be hard to interpret how changes visible on a sequence map onto the protein structure. Additionally, as DMS experiments represent controlled laboratory experiments looking at infection in cell culture, there are questions regarding how well DMS results reflect the patterns of influenza evolution that we observe in nature. Here, we introduce an interactive visualization platform that allows unified display of DMS differential selection data, 3-D protein structure, and information about amino acid preferences observed in nature, as unified by their common variable: site in the the sequence.

## 2 RELATED WORK

The Bloom lab at Fred Hutch performs are large amount of DMS work to investigate functional stability and immune escape and selection for important human pathogens such as HIV, influenza, and Zika virus. Originally developed by Thyagarajan and Bloom [6], initial DMS experiments on influenza virus hemagglutinin focused on understanding which sites of the protein could tolerate multiple amino acids and cause a functional infection. These studies have been extended to include not only purifying selective forces, but also immune selection as well, by innate immunity [1] and adaptive immunity involving antibody response [4]. In total, these experiments have helped us to define broad atlases of paths for escaping immunity [3]. Importantly, the development of such atlases, and the visualization of these data, requires an intensive computational workflow that involves iteratively generating static graphs for the relevant streams of data: immune selection as measured by the DMS experiment and the 3D structure of the protein. By default these images also do not link a third relevant data stream, and that is the frequencies of amino acids observed at different sites of the protein in nature. Thus the task of our visualization platform is to provide an easy way to view and interact with these three data streams in a linked fashion that is also easy to use in the absence of coding expertise.

## 3 METHODS

This visualization platform links three data views together by their common shared element, site in the protein. The three data streams are: 1) DMS differential selection value, which is

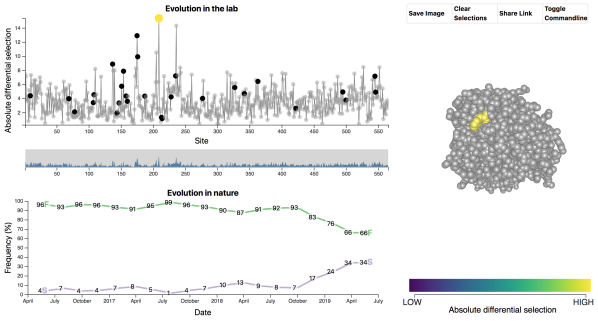


**Figure 1: Full view of visualization platform. The initial view of the visualization shows three panels: the laboratory-based DMS data, influenza variant frequencies in nature, and the 3D protein structure.**

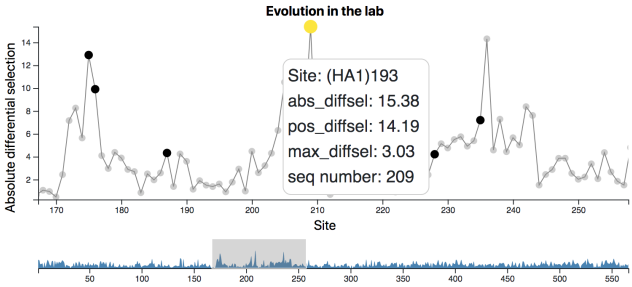


**Figure 2: Visualization of differential selection data generated from DMS experiments. The laboratory data can be explored and zoomed in on by brushing the context panel below the primary panel, giving both detail and overview to the user. Sites that are colored black in this panel represent sites for which there is natural frequency data. For all grey sites, only one variant of influenza at that site is found to circulate naturally. Mousing over a tip brings up details on demand, and clicking on a data point selects that site for visualization in the other panels, and colors the point according to its maximal differential selection value.**

a quantitative measure of the preference for a certain amino acid at a specific site in the protein, 2) the natural amino acid frequencies of a site, a quantitative variable the describes what proportion of all sampled strains of influenza at time T have a specific amino acid at a site, and 3) the 3-D structure of protein itself, which takes linear sequence space and folds that up to reflect where certain sites of the sequence fall on the protein structure. We describe design decisions about how to visualize these data streams below. Our visualization was implemented using D3 [2]

### DMS differential selection preferences

Navigation of our visualization begins with the user selecting a site in the gene that they would like to explore. Because the gene segments can be long, and many measurements are highly similar, it would be challenging to observe all measurements at once and interact with them. To address this challenge, we have two plots within this data stream's panel, one that gives context about the overall positioning in the gene segment, and one that is the zoomed view that can be interacted with. To zoom in, the context panel can be brushed, and this will select the portion of the gene segment that the user would like to zoom in on. Within the zoom panel, the DMS preferences by site, encoded using position, can be interacted with by mousing over the data points. Mousing over brings up a tooltip providing additional details about the DMS data at that site. In particular, the tooltip gives the exact value of the differential selection measurement, since this might be hard to read from looking at the y-axis. The points can also be selected by clicking on them, which allows the user to enter the linked visualization, whereby the natural amino acid frequencies for that site, and that site in the 3D protein structure, will be selected.

### Amino acid frequencies observed in nature

The natural amino acid frequencies are quantitative-ratio data that vary from 0 to 1 over time. They are derived by asking, at different points in time, what proportion of all circulating influenza viruses have a specific amino acid at this site in the gene. Given that these data are proportions, they take on values between 0 and 1, and data at cross-sectional time points sum to 1. In our visualization, we have encoded these frequencies using position. Selection of a site from the DMS preferences panel brings up a plotting space that shows which amino acids were observed at that position over time, and how those data changed through time. We use position to encode the frequencies value, and color to differentiate between different amino acid residues.

### 3D protein structure

Here, a selected site is differentiated from unselected sites by color. All unselected sites are uniformly grey, and the selected site is colored. The highlight color of the selected site is chosen according to the differential selection value of that site observed in the DMS data.

## 4 RESULTS

The intended users of this visualization platform are researchers from a wide-range of disciplines. These include clinicians and statisticians evaluating viral escape and the efficacy of immunotherapies in clinical trials, vaccinologists designing and evaluating immunogens and vaccines, evolutionary biologists studying viral evolution and immune

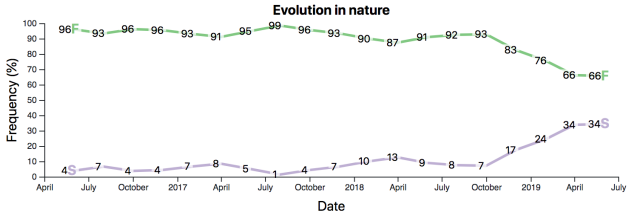


**Figure 3: Frequencies of influenza variants in nature. Once a site has been selected from the laboratory data panel, this panel shows the frequencies of influenza viruses with different amino acids at that site over the last three years. In this example, at site 193 we see that there are two variants circulating, one has a phenylalanine (F) at this site, while the other has a serine (S). We can see that the phenyalanine variant is at higher frequencies, but that the serine variant is rising in frequency.**

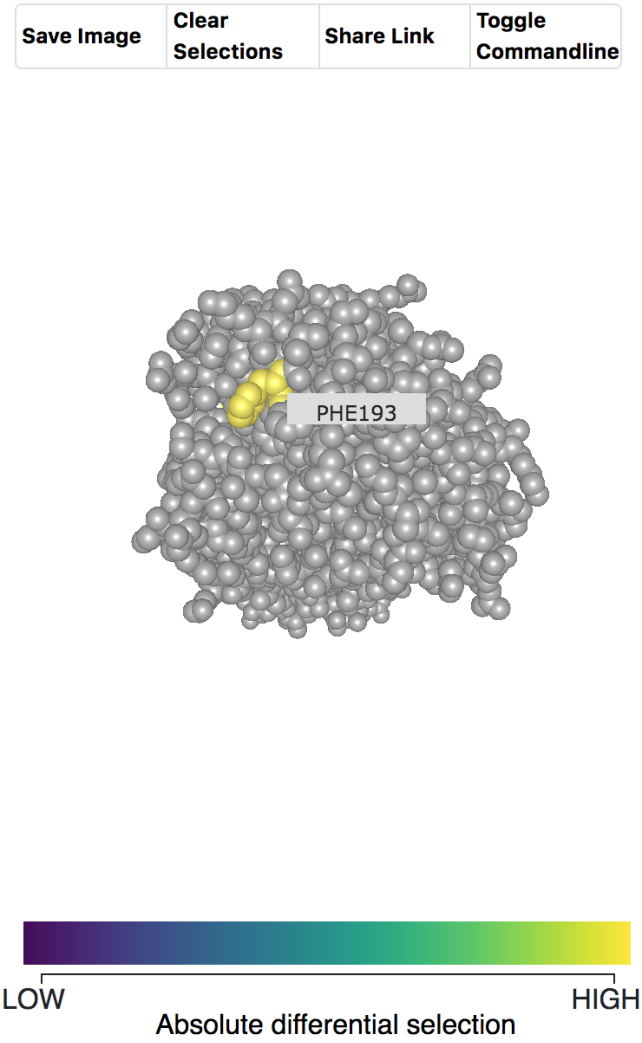


**Figure 4: 3D protein structure with selected site highlighted. 3D protein structure panel, with the selected site colored according to its differential selection value. The tooltip also provides information about which amino acid is at highest frequency at the site. In this image, site 193 has a PHE, or phenylalanine.**

evasion, and structural biologists interpreting the biological significance protein structures. By building a platform that unifies relevant pieces of data, such as DMS preference data, naturally-occurring amino acid preferences, and protein structure information, the platform enables researchers to explore the interplay between viral evolution and antibody immunity, contextualize experimental data, and quickly address preliminary hypothesis with the data. Most importantly, this type of data unification and presentation usually poses significant data curation and computational hurdles that non-computational biologists may struggle to perform. The platform developed here lowers the bar necessary to explore the data, making access to these data, and the inferences that can be made from them, more broad.

## 5 DISCUSSION

We created a browser-based visualization tool that allows the user to explore data generated from DMS experiments on influenza HA gene within the context of the 3D protein structure and of patterns of amino acid preferences observed in nature. This platform, to our knowledge, provides the first such unified and interactive visualization of these data. This unification helps to provides new insights to DMS preference data. Firstly, hotspots of mutational tolerance or immune selection may not be easily intuited from the linear sequence data, since sites that are close to each other in the 3D protein may be quite distal in the sequence space. Thus, tying these two views together improves the ease with which researchers can see whether changes at specific locations on the protein are responsible for immune escape and protein function. Additionally, given the controlled environment of the lab, some researchers wonder to what extent DMS data recapitulates the evolutionary pathways observed in nature. Our platform allows users to ask that exact question, by looking at both the frequencies of multiple amino acids observed at a site (that sites mutational tolerance in the lab and in nature), and also which amino acids specifically are observed (is the preferred amino acid in nature the same as the preferred amino acid in the lab). Taken together, the ability to make new inferences across these data streams improves our ability to assess, analyze, and contextualize DMS studies.

## 6 FUTURE WORK

DMS experiments are conducted for various different genes, different pathogens, and outside of the field of infectious disease research as well. As such, in the future we would like to expand this visualization platform to accept data sets other than from influenza. To facilitate wider use, we would like to build functionality into the visualization platform such that it can take in drag-and-drop files for other protein structures and preferences. Additionally, we would like to expand the number of panels that would allow joint visualization, for instance, by including phylogenetic trees that would allow querying of specific viruses with linking to relevant information about the virus and its proteins to relevant DMS data.

## REFERENCES

[1] Orr Ashenberg, Jai Padmakumar, Michael B Doud, and Jesse D Bloom. 2017. Deep mutational scanning identifies sites in influenza nucleoprotein that affect viral inhibition by MxA. *PLoS pathogens* 13, 3 (2017), e1006288.

[2] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. $D^3$ data-driven documents. *IEEE transactions on visualization and computer graphics* 17, 12 (2011), 2301–2309.

[3] Adam S Dingens, Dana Arenz, Haidyn Weight, Julie Overbaugh, and Jesse D Bloom. 2019. An Antigenic Atlas of HIV-1 Escape from Broadly Neutralizing Antibodies Distinguishes Functional and Structural Epitopes. *Immunity* 50, 2 (2019), 520–532.

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

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

[6] Bargavi Thyagarajan and Jesse D Bloom. 2014. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *Elife* 3 (2014), e03300.