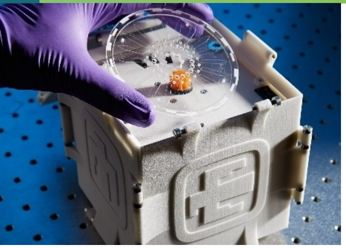
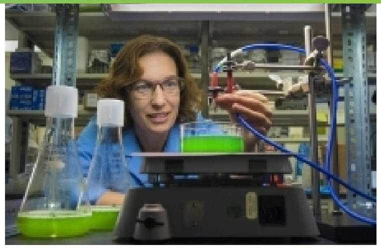
This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2022-14961C

Sandia National Laboratories

Experimental and Modeling Informed Data Analytics Platform to Identify Viral Features Indicative of Pandemic Potential

PRESENTED BY

Thomas Sheffield

CBD S&T

December 8, 2022

U.S. DEPARTMENT OF ENERGY NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Emerging infectious disease pose an imminent threat to human health, economic and national security

"Pandemics are for the most part disease outbreaks that become widespread as a result of the spread of human-to-human infection. Beyond the debilitating, sometimes fatal, consequences for those directly affected, pandemics have a range of negative social, economic and political consequences. These tend to be greater where the pandemic is a novel pathogen, has a high mortality and/or hospitalization rate and is easily spread. According to Lee Jong-wook, former Director-General of the World Health Organization (WHO), pandemics do not respect international borders. **Therefore, they have the potential to weaken many societies, political systems and economies simultaneously**."

United Nations Chronical, 2008 (https://www.un.org/en/chronicle/article/national-security-and-pandemics)

1918: Influenza

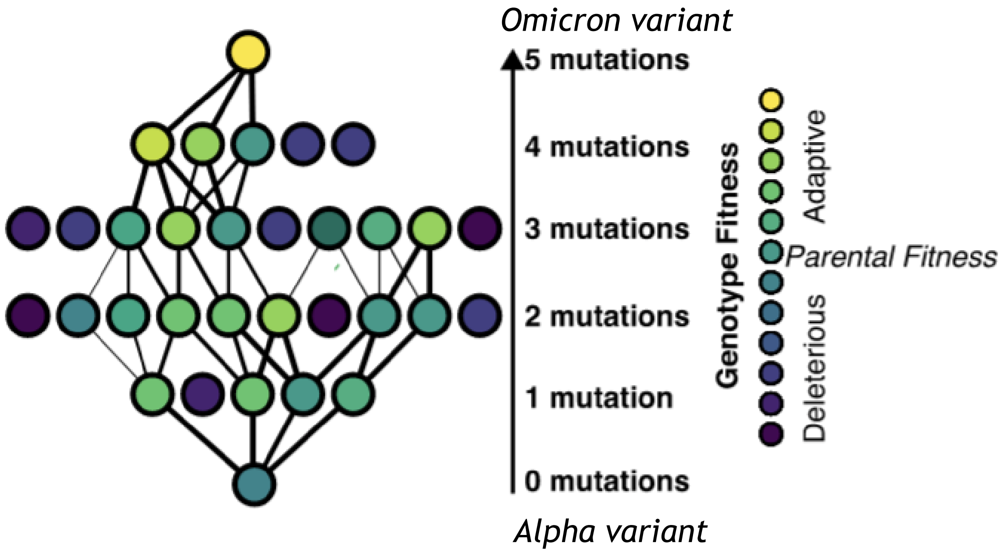
2002: West Nile Virus
2003: SARS
2005: Bird flu
2009: Swine flu
2014: Ebola
2016: Zika virus
2019 - : Covid 19

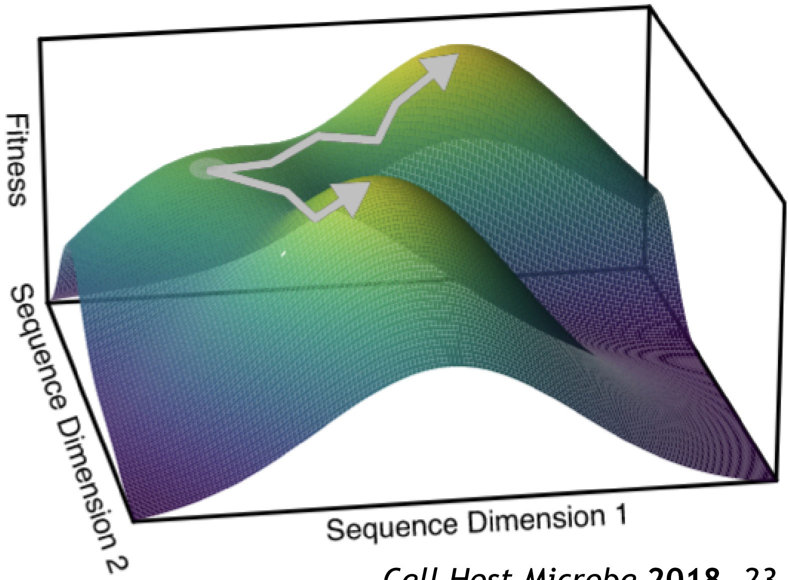
The current Covid-19 pandemic highlights the devastating potential of new and emerging infectious diseases.

And the need to develop methods to predict the pandemic potential of emerging pathogens.

We are trying to understand and predict viral evolution

Evolution occurs along a genotype/phenotype – fitness landscape





Cell Host Microbe **2018**, *23* (4), 435-446.

Simple 5 mutation site, 2 possible mutations network

- 2^5 (32) genotype combinations
- Connected by single mutations
- Prob{mutation} ~ line width
- Network complexity ~ Interactions among mutations (epistasis)

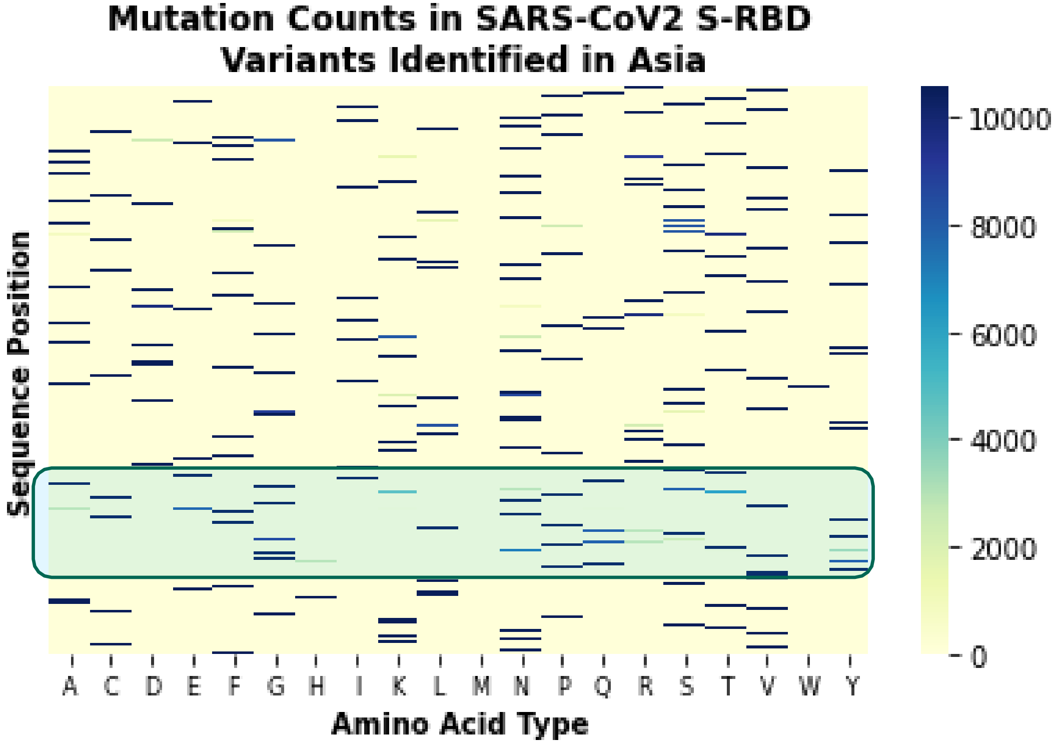
Populations explore the topography of the fitness landscape

- By acquiring mutations
- Natural selection drives populations toward local maxima
- Swarms of variants simultaneously exploring the fitness landscape

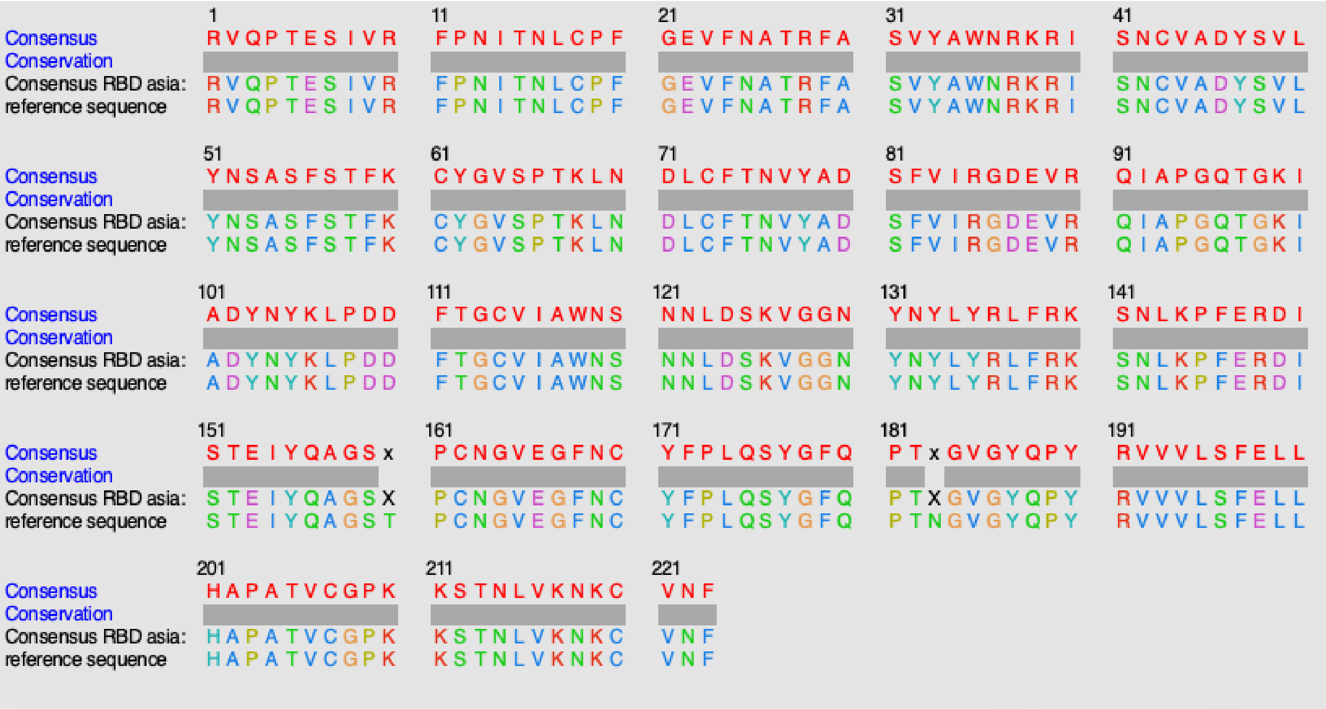
We are essentially trying to identify (predict) the local maxima on the fitness landscape

Why is it so challenging?

Sequence alignment of 10,570 SARS-CoV2-S-RBD variants identified in Asia



Most sequence positions are relatively invariant



```
                       1           11          21          31          41
Consensus              RVQPTESIVR  FPNITNLCPF  GEVFNATRFA  SVYAWNRKRI  SNCVADYSVL
Conservation
Consensus RBD asia:    RVQPTESIVR  FPNITNLCPF  GEVFNATRFA  SVYAWNRKRI  SNCVADYSVL
reference sequence     RVQPTESIVR  FPNITNLCPF  GEVFNATRFA  SVYAWNRKRI  SNCVADYSVL

                       51          61          71          81          91
Consensus              YNSASFSTFK  CYGVSPTKLN  DLCFTNVYAD  SFVIRGDEVR  QIAPGQTGKI
Conservation
Consensus RBD asia:    YNSASFSTFK  CYGVSPTKLN  DLCFTNVYAD  SFVIRGDEVR  QIAPGQTGKI
reference sequence     YNSASFSTFK  CYGVSPTKLN  DLCFTNVYAD  SFVIRGDEVR  QIAPGQTGKI

                       101         111         121         131         141
Consensus              ADYNYKLPDD  FTGCVIAWNS  NNLDSKVGGN  YNYLYRLFRK  SNLKPFERDI
Conservation
Consensus RBD asia:    ADYNYKLPDD  FTGCVIAWNS  NNLDSKVGGN  YNYLYRLFRK  SNLKPFERDI
reference sequence     ADYNYKLPDD  FTGCVIAWNS  NNLDSKVGGN  YNYLYRLFRK  SNLKPFERDI

                       151         161         171         181         191
Consensus              STEIYQAGSx  PCNGVEGFNC  YFPLQSYGFQ  PTxGVGYQPY  RVVVLSFELL
Conservation
Consensus RBD asia:    STEIYQAGSX  PCNGVEGFNC  YFPLQSYGFQ  PTXGVGYQPY  RVVVLSFELL
reference sequence     STEIYQAGST  PCNGVEGFNC  YFPLQSYGFQ  PTNGVGYQPY  RVVVLSFELL

                       201         211         221
Consensus              HAPATVCGPK  KSTNLVKNKC  VNF
Conservation
Consensus RBD asia:    HAPATVCGPK  KSTNLVKNKC  VNF
reference sequence     HAPATVCGPK  KSTNLVKNKC  VNF
```

Consensus sequence generated from alignment of the 10,570 variants identified in Asia

- Conservation does reduce the search space or space of possible variants.
- Space of variants is still huge ~ 20^n, where n is the number of possible mutation sites.
- Highly fit variants are very rare and predicting the pathway along the fitness landscape difficult.
- Fitness landscape is very high dimension with multiple objective functions.

Using Published Data

Starr, Tyler N., et al. "Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding." Cell 182.5 (2020): 1295-1310.

- Receptor binding domain (RBD) expression on cell surface of yeast
  - RBD consists of spike amino acids 331-531 (201 total)
- PCR-based mutagenesis introduces mutations

Binding

- Titration curves for 16 ACE2 concentrations
- Binding Endpoint: Change in log10(Ka) from wildtype (Ka is inverse dissociation constant)

Expression

- Fluorescence-activated cell sorting
- Expression Endpoint: Change in mean fluorescence intensity from wildtype

Global epistasis models predict effects on expression/binding for single mutations

Using Published Data Cont'd

Greaney, Allison J., et al. "Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies." Cell host & microbe 29.3 (2021): 463-476.

- Similar experiment to before
- 10 neutralizing antibodies
  - 9 from SARS-CoV-2 patients, 1 from SARS-CoV-1
- Also used to build global epistasis models

Endpoint: log10(Escape Fraction)

- Escape Fraction $E_v = \mathrm{F} * \frac{n_v^{post}/n_v^{pre}}{N^{post}/N^{pre}}$
  - F is total fraction of library that escapes antibody binding
  - n_v^{pre}, n_v^{post} are the counts for variant v before and after enriching for antibody escape plus a pseudo-count .5
  - $N^{pre} = \sum n_v^{pre}, N^{post} = \sum n_v^{post}$

Variants in Data

Binding

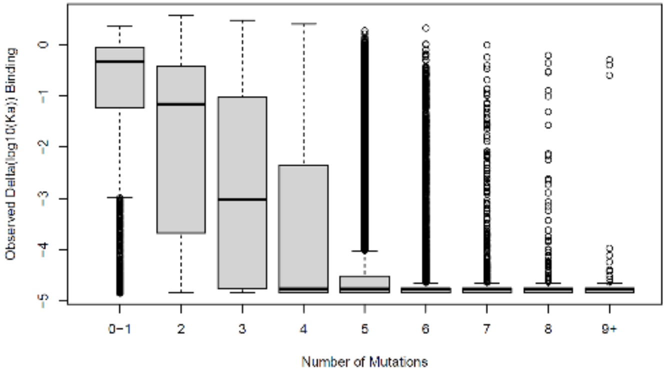
- 146,437 observations of 105,526 unique variants
- 0-10 mutations per variant with median 3
- 3,802 unique mutations represented (out of 19*201 = 3,819 possible)
  - No deletions/insertions

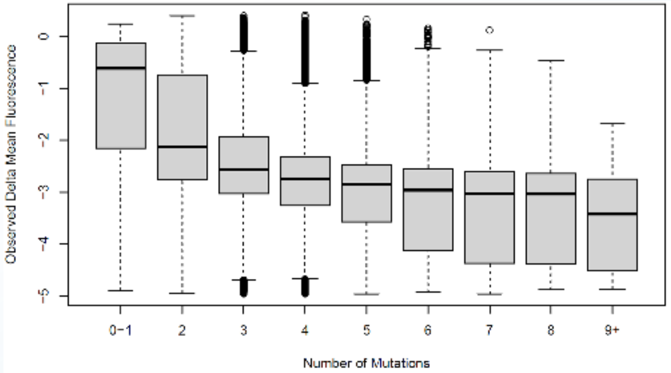
Expression

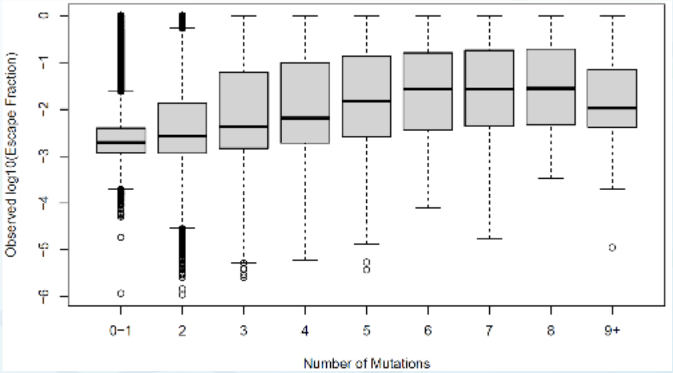
- 177,759 observations of 135,386 unique variants
- 0-12 mutations per variant with median 3
- 4,002 unique mutations represented (of 4,020 possible)
  - Deletions included, but no insertions

Antibody Escape

- 714,797 observations of 50,795 unique variants
- 10 antibodies with 66,403 - 79,126 observations each
- 0-10 mutations per variant with median 2
- 3,954 unique mutations represented (of 4,020 possible)
  - Deletions included, but no insertions







Data Engineering

One Variant per row

Features

- Individual mutations: e.g. N1A, N1C, N1D, …
- Antibody type
- Currently no feature selection
- Use sparse matrices
- Also tried:
  - Location and types of mutations: e.g. 1,2,3,…,201, AC, AD, …
  - Derived features: e.g. Moreau-Broto autocorrelation, conjoint triad descriptors, etc.

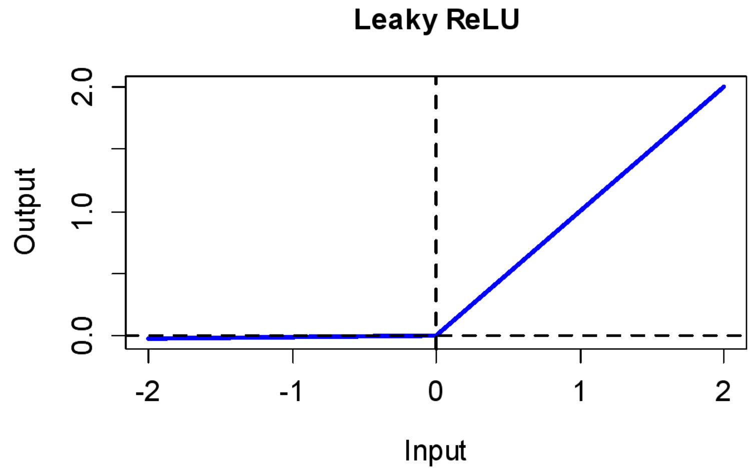
Variant	N1*	N1A	N1C	…	T201Y	COV2-2082_400	…	CR3022_400	Endpoint
1	0	0	0		0	0		0	-2.93
2	0	0	0		0	1		0	-2.57
3	0	0	0		0	0		0	-2.81
4	0	0	0		0	0		0	-3.19

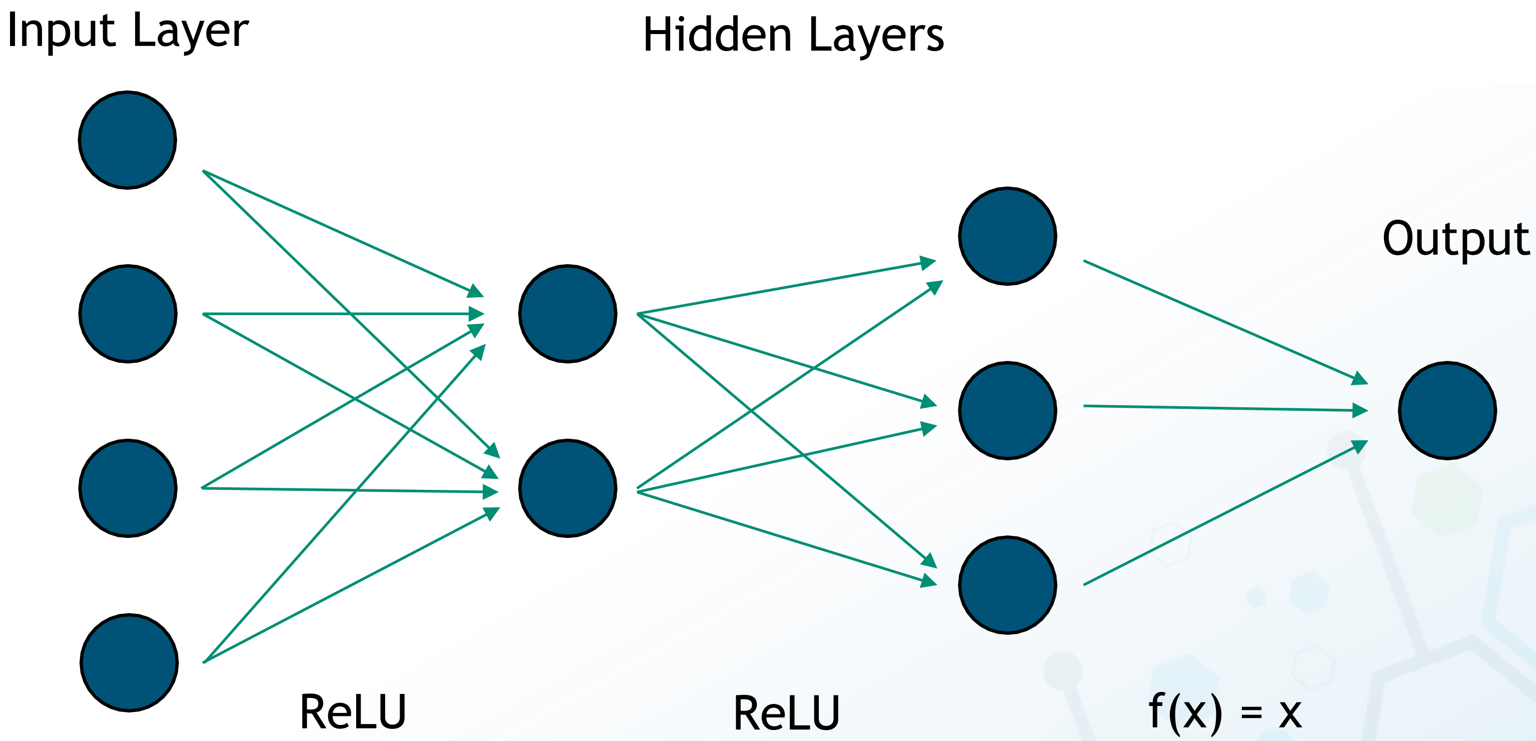
Modeling

Use Keras neural net machine learning model

- Tensorflow backend
- Thousands of parameters
- Also tried:

  Xgboost, random forest, support vector machines (too much data)





Tuning

Five-fold cross-validation

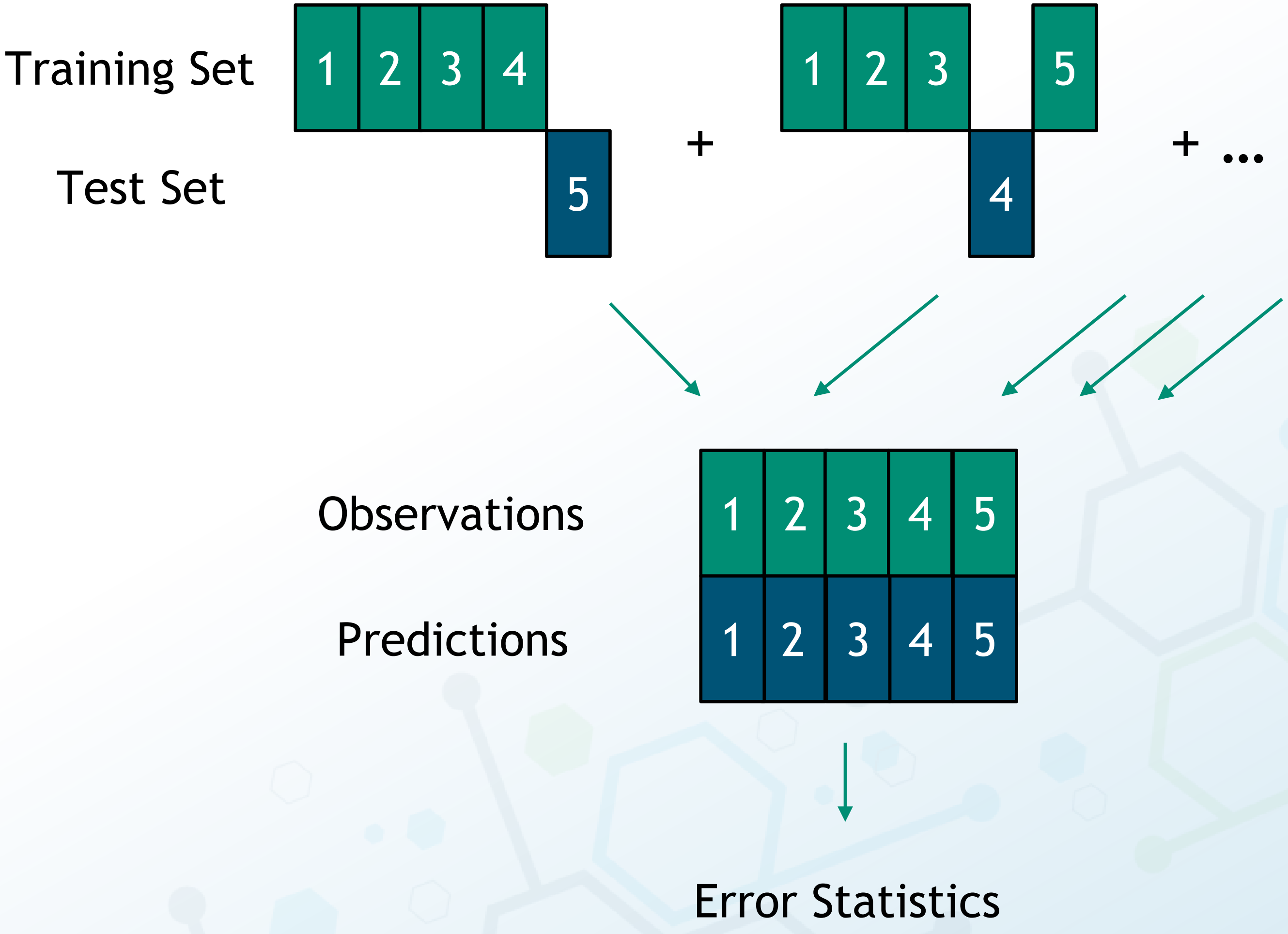
- Estimates predictive ability on data outside of training set
- Hidden assumption: future data is similar to data you have

Hyperparameters

- Layers: 2 or 3
- Sizes: $2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8$
- ReLU: Leaky or Regular
- 784 possible combinations

Tuning

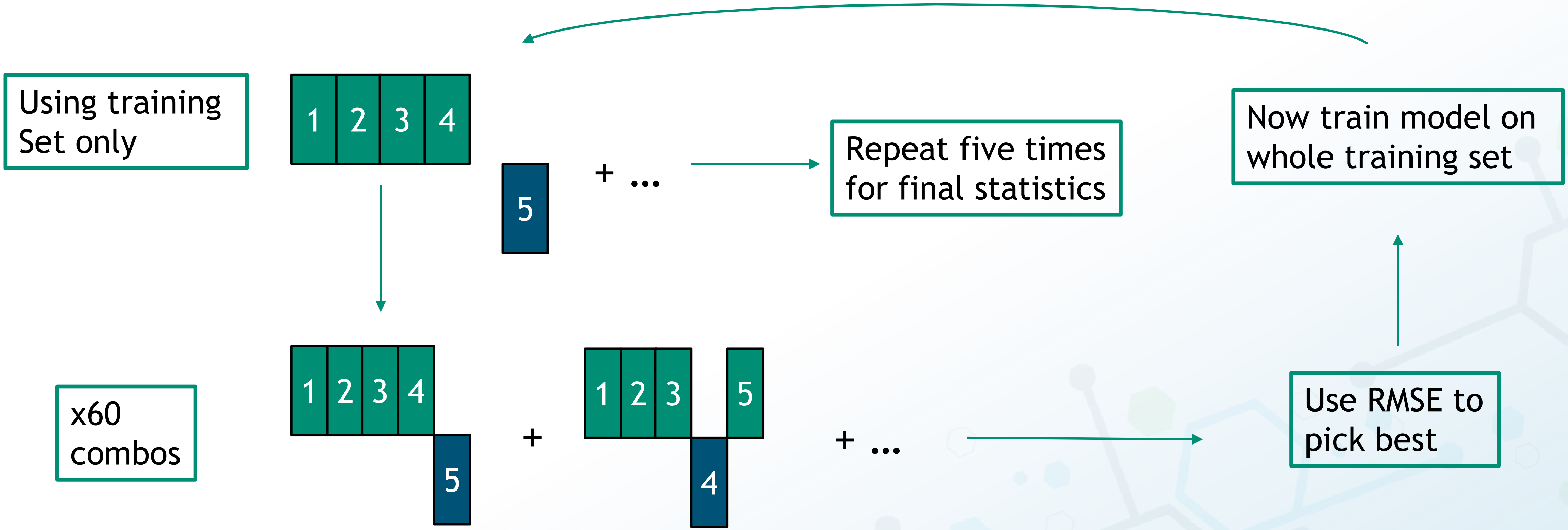
- 60 randomly selected combinations
- Choose parameters with best root-mean-square-error (RMSE)



Tuning in Loop

Tuning in Loop

- Tune independently within each training set
- Avoids overfitting in results



Using training Set only

Repeat five times for final statistics

Now train model on whole training set

x60 combos

Use RMSE to pick best

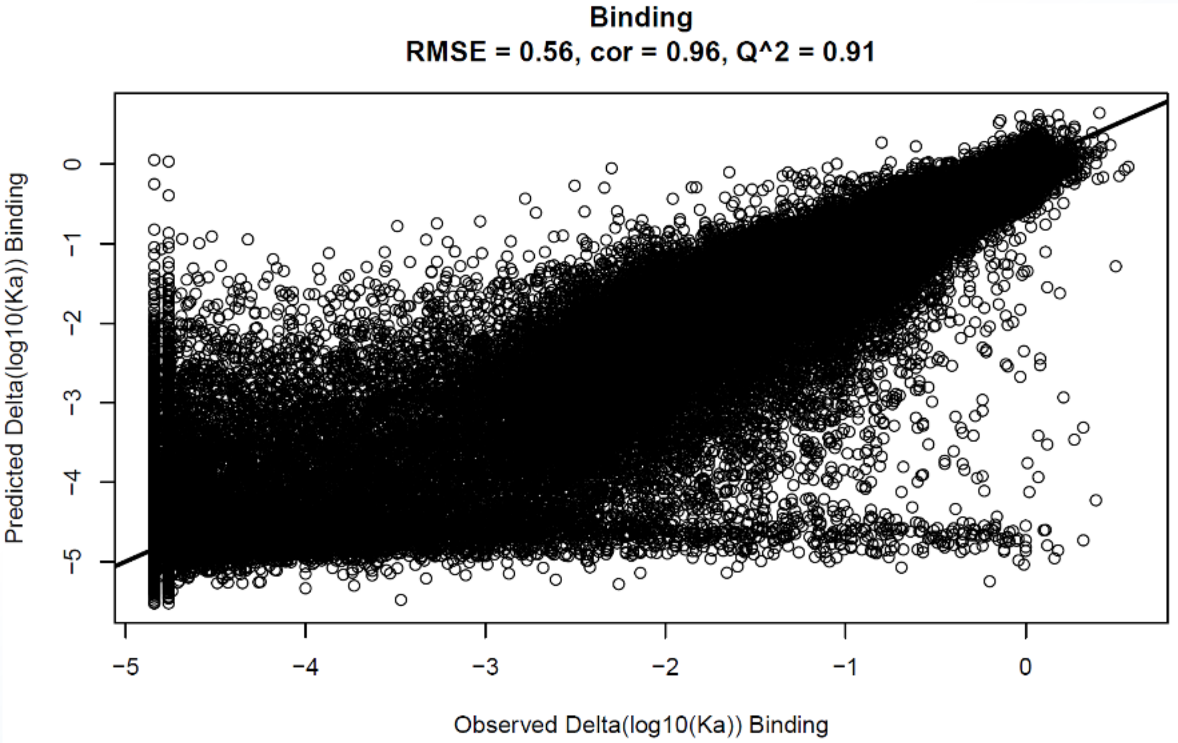
Results

Statistics used

- Root-mean-square error (RMSE)
  - $\sqrt{\Sigma_i (y_i - f_i)^2 / n}$
  - Where y_i are observed endpoints and f_i are predictions based on other data points
- Pearson Correlation
- Cross-validated coefficient of determination (R^2)
  - $Q^2 = 1 - \Sigma_i (y_i - f_i)^2 / \Sigma_i (y_i - \bar{y})^2$

Binding results:

- RMSE = 0.56 $\Delta\log_{10}(K_a)$
- Pearson correlation of 0.96
- $Q^2 = 0.91$



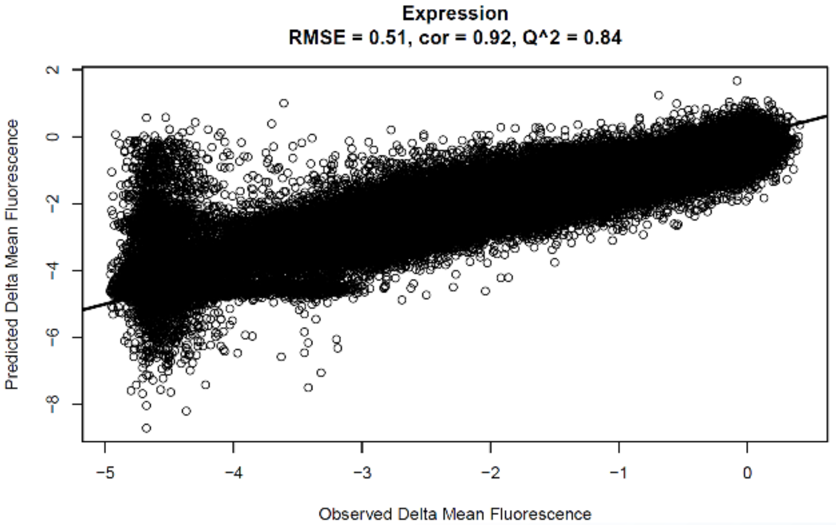
Results Cont'd

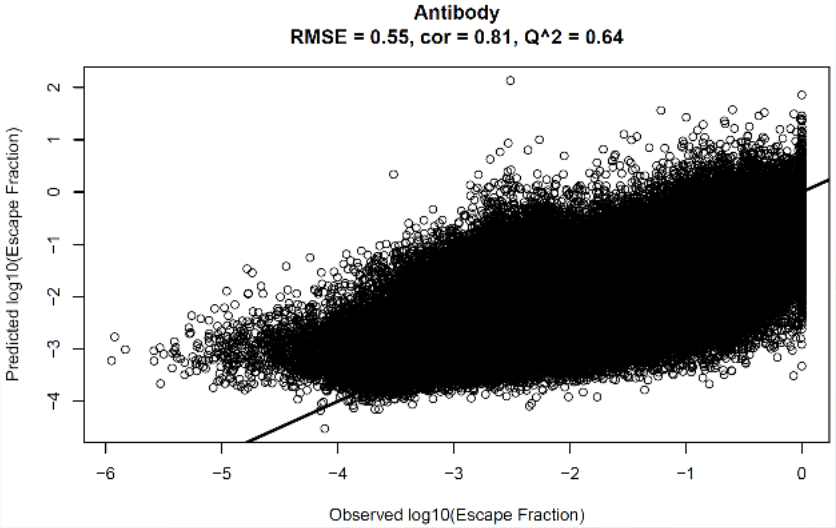
Expression results:

- RMSE = 0.51 ΔMean Fluorescence
- Pearson correlation = 0.92
- Q^2 = 0.84

Antibody results:

- Untuned, using 128 x 32 hidden layers and regular ReLU
- RMSE = 0.55 $\Delta\log_{10}$(Escape Fraction)
- Pearson correlation = 0.81
- Q^2 = 0.64





Antibody Model Issues

Originally made one model for each antibody

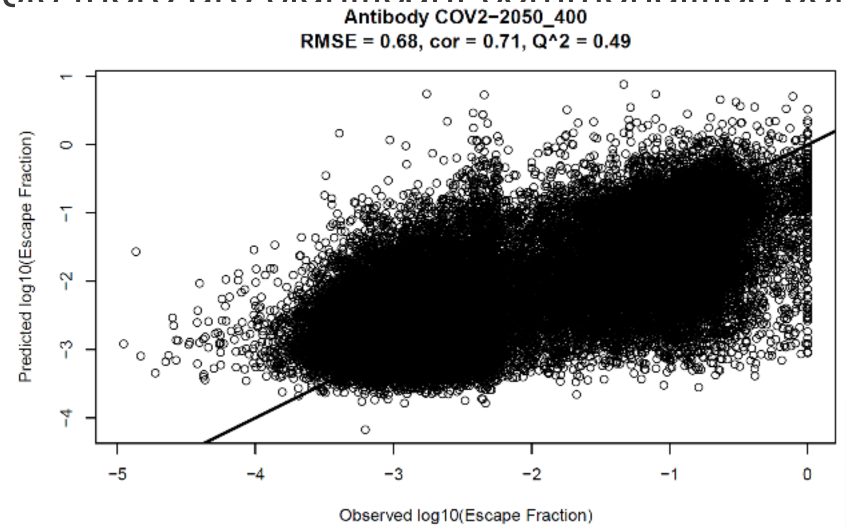
- But half of models were mediocre and other half were poor

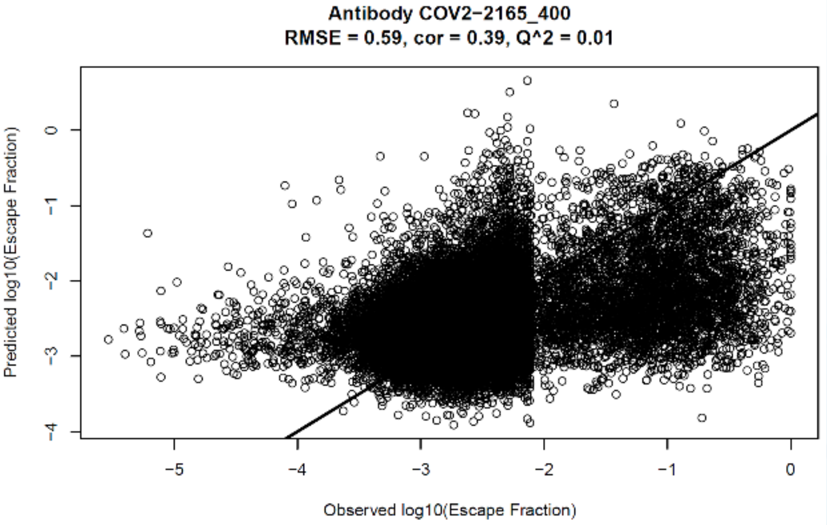
Why?

- Antibody escape is based on counts before and after an antibody is applied
- Lowest count observations are discarded, but the bulk of observations are low-count and high uncertainty
- Removing them degrades model quality even further
- Weighting observations did not help
- The five poor antibody models are dominated by these observations

Combining model had better statistics than any single antibody model

- Suggests there are significant commonalities between antibodies

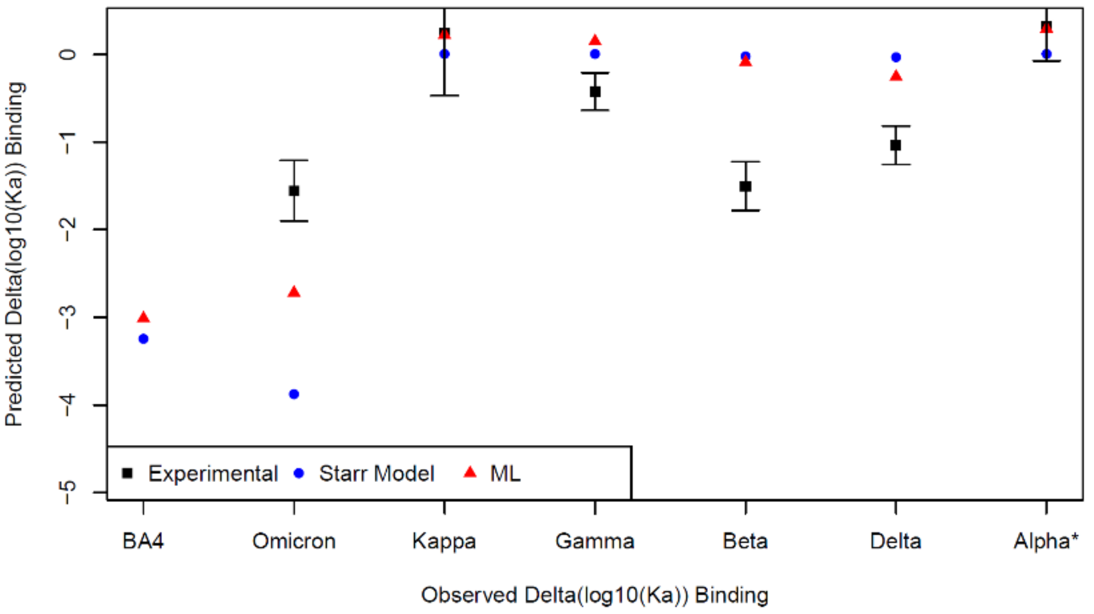




Real Variant Prediction

Our group performed binding assays on wild variants

- Only six can be compared so far
  - Can't predict insertions
  - Alpha is only variant present in training data
- Omicron BA.1.1 has 16 RBD mutations
- BA4/5 has 17, and only 11 in common with BA.1.1



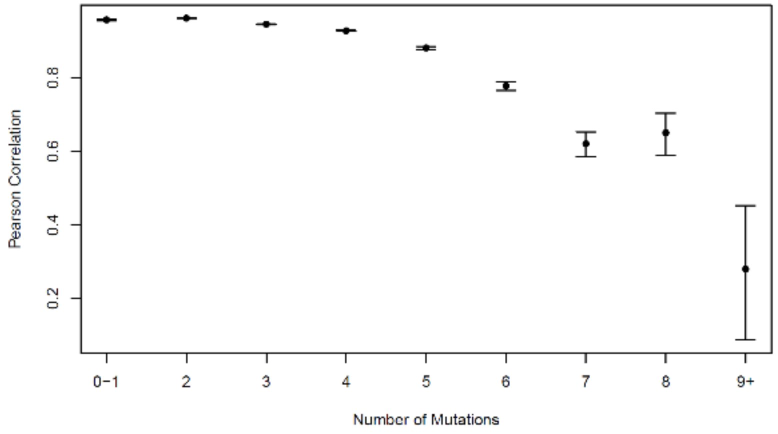
Results By Mutation

Model quality tends to drop with increasing numbers of mutations for all models

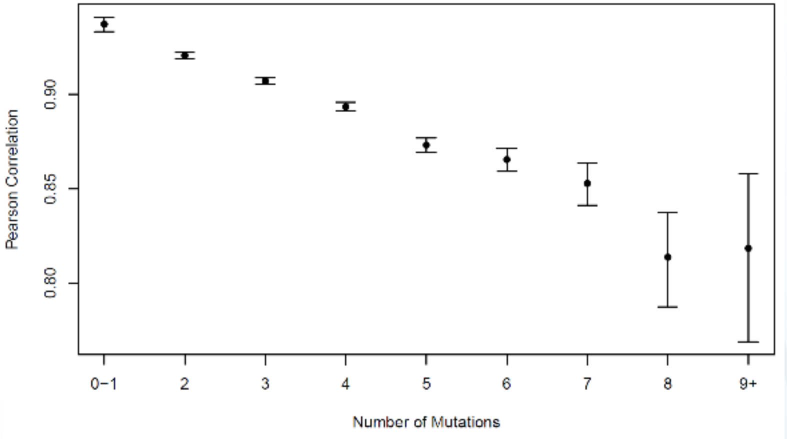
Single mutations and small combinations are well-represented in the data

More complicated mutation combinations are not present, and hard to predict the effect of

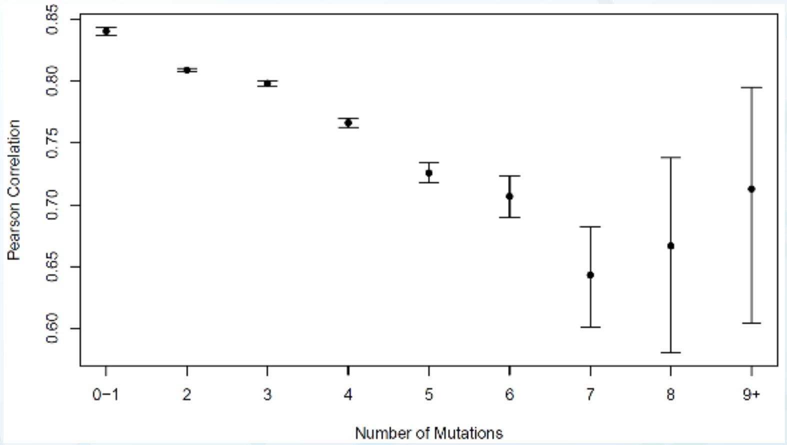
Binding



Expression



Antibody Escape

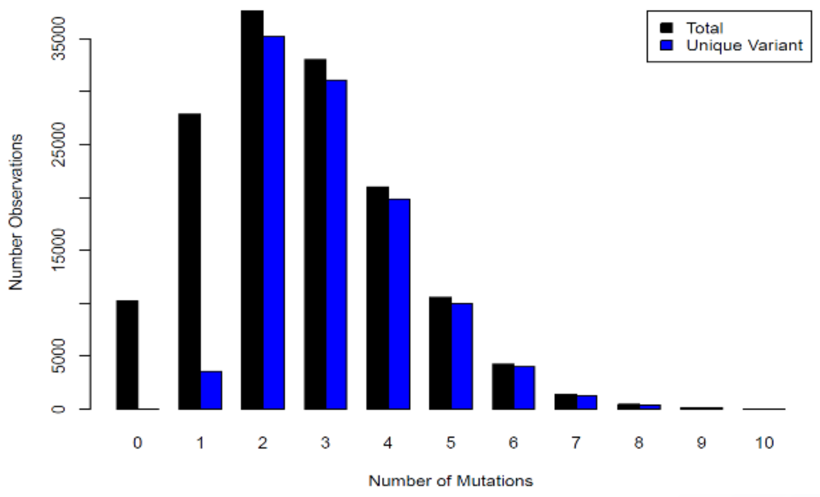


Mutation Coverage

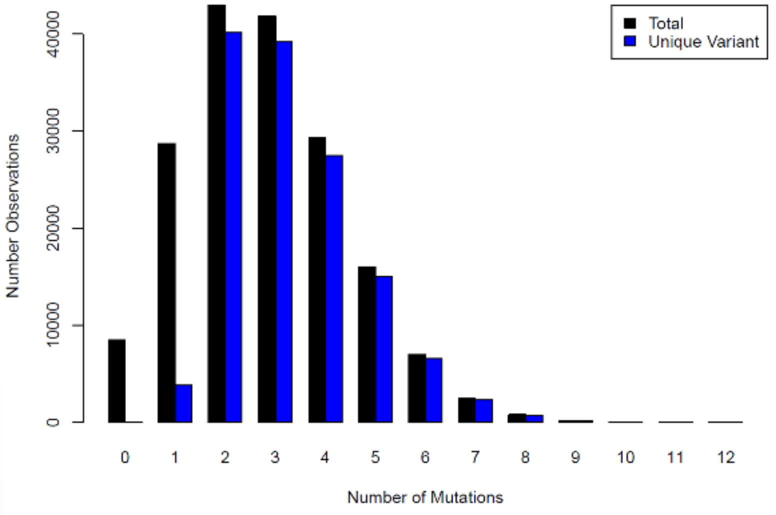
Original experiment was designed to find single mutation effects

- Space of variants near Wuhan is well-covered
- Coverage drops quickly with more mutations
  - Especially relative to all possible combinations
- Space near omicron variants is unexplored

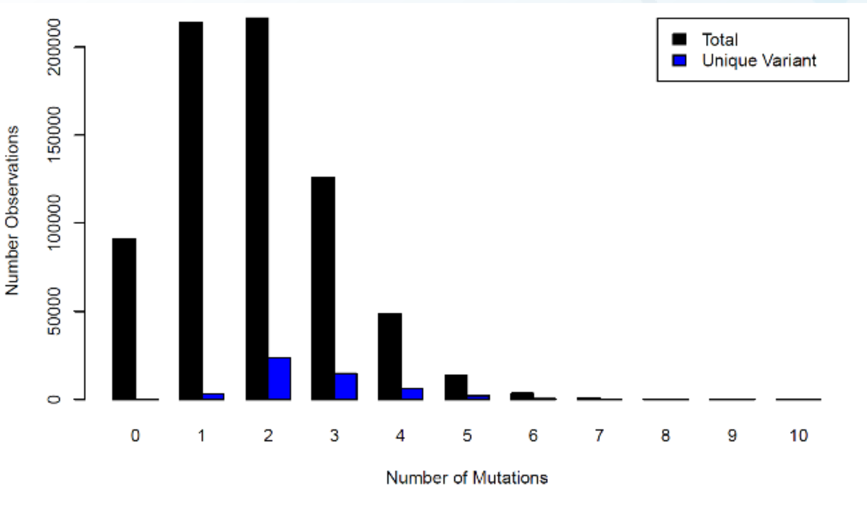
Binding



Expression



Antibody



Site Region Selection For Future Experiments

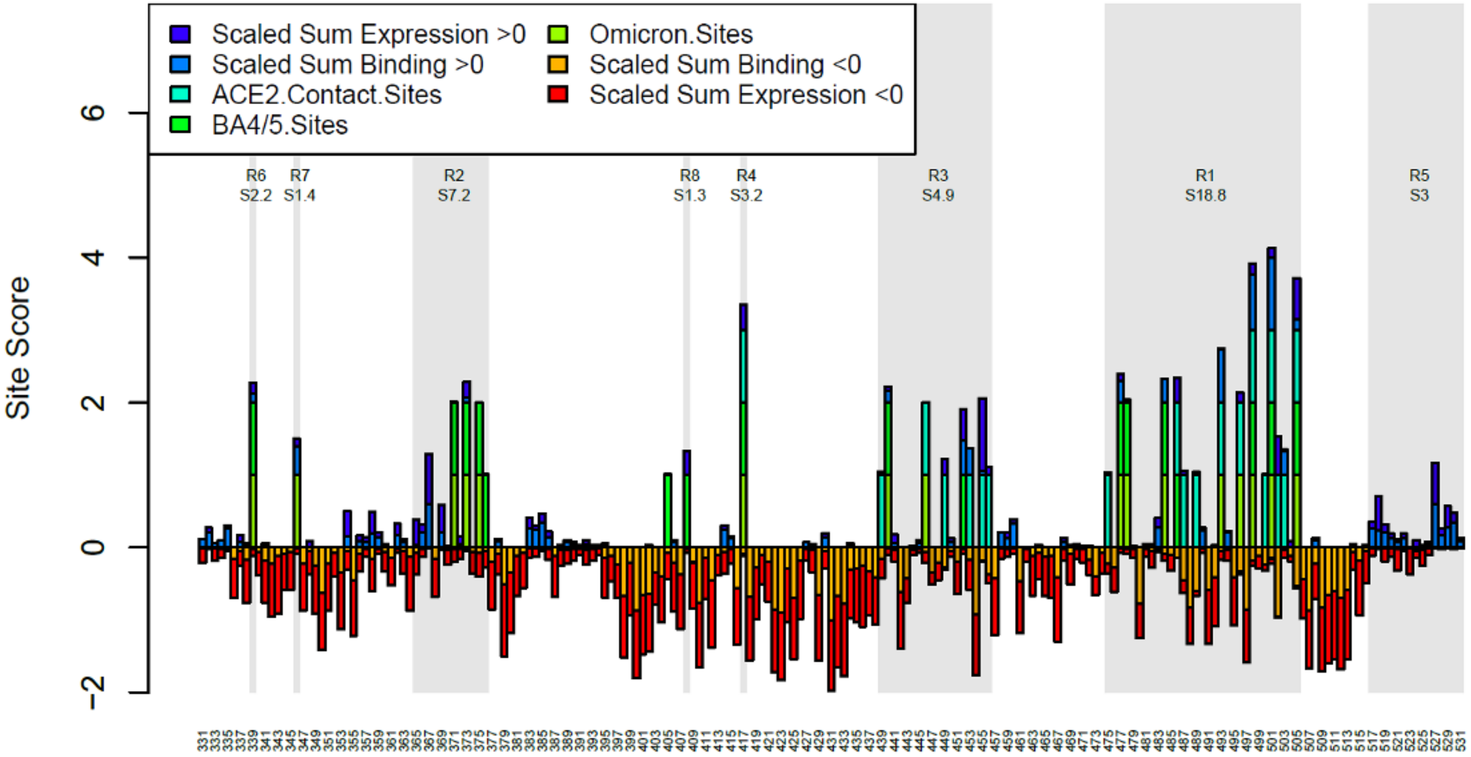
Use Starr model single mutation binding/expression effects

Sum positive and negative binding/expression effects for each site and scale to a maximum of 1 over all sites

Increase score by one for site presence in Omicron BA1.1, BA4/5, or ACE2 contact site

Find regions with maximum total scores iteratively

- Top four regions: 475-505, 365-376, 439-356, 417
- Top four scores: 18.8, 7.2, 4.9, 3.2



Future Work

Further binding experiments centered on BA4/5

- Train wider mutation set
- Closer to current state of virus
- Further model validation

Antibody tuning

Antibody binding based on antibody sequence/characteristics

Acknowledgements

Ken Sale

Mai Pham

Brooke Harmon