Main text

Statistical inference to optimize anti-SARS-CoV2 antibody design and uncover viral resistance phenotypes

Natalie Dullerud1, Tea Freedman-Susskind 2, Christopher Snow, Anthony P West, Vanessa D. Jonsson3,4

1 Department of Mathematics, University of Southern California, Los Angeles, CA

2 Department of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA

3 Department of Hematology, City of Hope National Medical Center, Duarte, CA

4 Department of Quantitative and Computational Medicine, City of Hope National Medical Center, Duarte, CA

**Abstract**

SARS-CoV2 antibody therapy

**Introduction**

*Background on pandemic. shorten this*

SARS-CoV-2, a novel coronavirus in the SARS-related (sarbecovirus) subgenus of betacoronaviruses, caused the 2020 pandemic outbreak of respiratory disease COVID-19. SARS-CoV-2 infect human hosts by binding to the human angiotensin-converting enzyme 2 (ACE2) via the receptor binding domain (RBD) on the viral spike surface glycoprotein. SARS-CoV-2 has evolved significantly throughout the course of the pandemic, resulting in several recorded mutations across the globe. Such mutations have been tracked through extensive research database systems, such as nextstrain [(1)](https://www.zotero.org/google-docs/?jLMexY). The recorded sequence mutations in the global population have not conferred significant increase in infectivity according to most studies on ACE2 binding and transmissibility [(2)](https://www.zotero.org/google-docs/?bpbAP8). Despite early concerns about mutation D614G, which has been observed to dominate in groups in which the mutation appears, this mutation has been shown to have very little evolutionary advantage over the wild type virus in terms of infectivity [(3)](https://www.zotero.org/google-docs/?COjYUS). Period of immunity in convalescent COVID-19 individuals has been a topic of notable speculation (cite). A recently documented case of re-infection in Hong Kong involved different mutations in the first and second infections (cite). These mutations included E780Q in the first infection period and mutations L18F, A22V, and D614G in the second infection. The case has sparked interest in relative immunity of convalescent individuals against SARS-CoV-2 mutations in potential subsequent infections or exposure.

*Background on antibodies in SARS-COV2.*

* Clinical aspect, antibody titers over course of infection, B cell memory. (Krammer, Crotty, other studies)
* Antibodies as a therapeutic and potential for escape within the host. (cite Bjorkman, Nussenzweig Krammer, Crotty, other studies)

Neutralizing antibodies for SARS-CoV-2 from recovered COVID-19 patients mainly neutralize the virus in human hosts through binding the receptor binding domain on the spike glycoprotein, although some antibodies have been discovered to bind to other parts of the virus [(4)](https://www.zotero.org/google-docs/?ox8nX6). A large number of strong neutralizing antibodies discovered in convalescent COVID-19 patients are in the class of antibodies with heavy V (variable) gene VH3-53 or VH3-66 [(5)](https://www.zotero.org/google-docs/?nR8wJu). In recent studies and neutralization assays, the strength of these neutralizing antibodies has been examined, typically through 50% maximal inhibitory concentrations (IC50 ) for SARS-CoV-2 or 50% maximal response concentrations (EC50 ). Biological and logistical information from studies on SARS-CoV-2 antibodies can be accessed from the aggregated database Coronavirus Antibody Database (CoV-AbDab), run by Oxford University [(17)](https://www.zotero.org/google-docs/?7SaAOD). Several of the studies examining SARS-CoV-2 neutralizing antibodies include structures from cryogenic electron microscopy (Cryo-EM) or have uploaded structures to the Protein DataBank to further analyze these antibodies from a structural perspective. Such structural analyses are crucial in understanding how antibodies will interact with different mutations of the virus through binding energies and structural conformations.

*Structural studies classify antibodies and inform combinatorial strategies.*

Structural studies have recently classified anti SARS cov2 antibodies into three classes (cite) shown that escape is orthogonal in these classes, and have proposed a strategy for combinatorial antibody design. describe these.

*Experimental strategies to understand antibody escape.*

Regeneron study and cite.

In efforts to find and optimize treatment options for patients with COVID-19, antibody treatment has been examined as a potential avenue in several recent studies [(6,7)](https://www.zotero.org/google-docs/?QVyY8y). For example, Regeneron has completed several studies of antibodies synthesized from convalescent patients. In particular, Regeneron studies have shown that antibody cocktails exceed single antibody treatments in efficacy for treating SARS-CoV-2 in both *in vitro* and *in vivo* studies with rhesus macaques [(6)](https://www.zotero.org/google-docs/?BcST1e).

*Combination of structural and computational based combinatorial antibody design.*  Need to constrain a large design space.

Despite these breakthroughs in identification of neutralizing antibodies and their combinatorial effect, there is an increasing need to constrain the antibody design space as many of these antibodies may have mutations with better binding or neutralization abilities than wild type. In addition, due to the observed advantage of cocktails over single treatments, computational tools need to be used to assess optimal combinations of antibodies and the potential for mutational escape by the virus.

**Main**

Introduce the specific antibodies that we are studying. More specifics

**Satlasso uncovers antibody amino acid residues that confer better binding**

We developed a statistical model to uncover the relationship between anti-SARS-CoV-2 antibody sequence data from convalescent COVID-19 patients and corresponding SARS-CoV-2 neutralization data while mitigating the issue of irregular experimental data corresponding to neutralization assay limits. We obtained our model by extending the well-known least absolute shrinkage and selection operator (Lasso) by incorporating a component to operate on saturated experimental data (data that exceeds the limit of the neutralization assay) (cite). In particular, our model, the saturated Lasso (satlasso), identifies a minimal set of antibody residues that are most significant to virus neutralization and that minimize error between model and experimental data. A mathematical description of satlasso can be found in Materials and Methods. Model selection was performed using 5-fold cross validation.

We applied our model to two experimentally derived datasets of anti-SARS-CoV-2 antibody data. The datasets correspond to: 1) antibody sequence and neutralization data obtained from convalescent COVID-19 patients in a single study 2) *VH3-53* and *VH3-66* antibody sequence and neutralization data obtained from convalescent COVID-19 patients across several studies [(8–16)](https://www.zotero.org/google-docs/?RW0ueE). The single-study dataset was used in order to reduce the potential effect of outliers caused by discrepancies between neutralization assays across different studies. We chose to apply the model to the *VH3-53*/*VH3-66* dataset due to the similarities in SARS-CoV-2 binding within the group and the prevalence of these antibodies in recent literature [(5)](https://www.zotero.org/google-docs/?bnpRT9). We applied the model to these two datasets and obtained sets of antibody residues most relevant to SARS-CoV-2 neutralization.

In order to assess the accuracy of satlasso, we examined the predictions of the model on experimentally derived datasets and the significance of the residues identified by the model. Our satlasso model performed well in prediction of 50% maximal inhibitory concentration (IC50) for the training datasets, with an average R2 score of 0.952 on unsaturated data (figure - predicted, need a table with R2 values, error, and lambda values for these datasets). To test the generalization of our satlasso model, we compare the largest magnitude regressors to binding epitopes for analyzed antibodies that have associated PDB structures. We find that many antibody residues known to be critical to SARS-CoV2 neutralization are chosen (figure - logo plot). In particular, the largest magnitude heavy chain residues unearthed by satlasso on selected antibodies (C105, CV30, CB6, B38, CC12.1) generally correspond to SARS-CoV-2 binding epitopes or are near binding epitopes (figure - logo plot). From these metrics, we deduce that satlasso has strong prediction and residue identification ability.

Satlasso discovered several sites and mutations on anti-SARS-CoV-2 antibodies that may confer better neutralization ability. The sign of the regressors indicate whether or not the residues detract from (negative) or contribute to (positive) 50% maximal inhibitory concentration (IC50). As seen in figure - wild type barplots, on antibody C105, mutations are chosen that detract from IC50, thus increasing neutralization ability, including: YH58F, GH97R, and TH28I. Wild type residues that reduce neutralization ability are also selected by satlasso, such as TH28. Similarly, satlasso identifies beneficial mutation sites on other chosen antibodies that have corresponding PDB structures (CV30, CB6, B38, CC12.1) (figure - wild type barplots).

Our proposed satlasso model performs variable selection in domains where significant saturation in the training data is present. In the context of understanding the relationship between antibody residues and neutralization ability, satlasso can appropriately overcome irregularity in the experimental data due to limits in neutralization assays. Through use of satlasso on anti-SARS-CoV-2 antibody data, we have validated our model and determined potential mutation sites on antibodies that may enhance neutralization.

**Random forest predicts viral infectivity hotspots**

To reveal infectivity hotspots on the SARS-CoV-2 virus amino acid sequence, we applied a random forest regressor to angiotensin-converting enzyme 2 receptor binding data on SARS-CoV-2 mutations. The random forest regressor uncovered residues on the virus that are important to binding with the ACE2 receptor. We used experimental data from [(2)](https://www.zotero.org/google-docs/?4y4PwU) which provided values for change in ACE2 dissociation constant between wild type and mutant (Δlog(Kd) = log(Kd,wild-type) - log(Kd,mut)) for nearly each possible mutation on the receptor binding domain (RBD) of SARS-CoV-2.

The random forest regressor returns a vector of feature importances, where the features correspond to amino acid residues in the SARS-CoV-2 sequence. In order to test the efficacy of this method in evaluation of residue importance on SARS-CoV-2 amino acid sequence, we studied the accuracy of the prediction of the random forest regressor on our dataset and compared the feature importances selected by the model to known binding epitopes on SARS-CoV-2. We found that the random forest regressor predicted Δlog(Kd) values of SARS-CoV-2 mutants well, with R2 value of 0.941 (figure - predictor plot, also need R2 value for sars-cov-2 listed somewhere). The random forest regressor also discovered feature importances at sites or near sites known to be contacts with the ACE2 receptor, providing support for validation of the model’s efficacy at determining important sites for binding with ACE2 (figure - logo plot). An interesting characteristic of the feature importances returned by the random forest regressor is that the largest feature importances are all wild type SARS-CoV-2 residues, as seen in (figure - wild type barplot, logo plot). In particular, all mutations at these sites are deleterious to binding with the ACE2 receptor (violin plot). These observations replicate and confirm the results of [(2)](https://www.zotero.org/google-docs/?DLBTjk): the wild type virus is largely optimized for binding with the ACE2 receptor and mutations are not likely to enhance binding.

The use of the random forest regressor method to predict viral infectivity hotspots confirmed experimental data indicating that enhancement of ACE2 receptor binding due to single mutations of SARS-CoV-2 is unlikely. In the context of antibody design, the possibility of escape mutations from optimized antibodies may pose less of a problem due to lower infectivity on average of single mutant viruses.

**Using structural data to constrain**

**Computing Gibbs free energy models of antibody and SARS-CoV2 mutations**

*Constraining the antibody design space based on feature selections from estimator.* To do this we develop a biophysical model based on Gibbs free energy of binding derived from energy minimization calculations on structural information to quantify binding energy differences as a surrogate metric for antibody neutralization.

We hypothesize that a virus’ capacity to infect and be neutralized by specific neutralizing antibodies (NAbs) can be approximated by differences in Gibbs free energy of binding associated to de novo point mutations on the spike protein complexed with the ACE2 receptor and antibody structure. To compute fitness landscapes relating to viral replication and antibody neutralization, we apply an empirical force field, Fold X , to evaluate the effect of point mutations on the stability, folding and dynamics on detailed S-protein/ACE2 and S protein/NAb molecular structures.

A computational chemical tool used widely in energy minimization and mutational scanning in recent years, FoldX, can calculate and provide considerable structural information about proteins. FIND EXAMPLES

It has been shown that loss of binding affinity is consistent with RBD mutations that confer escape (cite).

*Coupling mutations on RBD with antibody mutations.* We describe the model, and we constrain this approximation to locations on RBD that are proximal physically to the antibody mutation (methods). we assume that there are no large changes away from the proximity of the antibody mutation /RBD mutation pair.

Figure of landscape of antibody mutations and RBD mutations, with structure

Antibodies that have been characterized, see extended table Barnes: Class I (C102, C105) characterized, class II (C121, C002, C104, C144, binding in 440-444), Class III (C135, S309, N343RBD , N440 binding). Both class I and class II antibodies block ACE2 binding and class III and class IV do not overlap with ACE2.

*Introduce the example*

* We looked at class I and class II antibodies to optimize design of the antibody as these classes of antibodies overlap with ACE2 binding in contrast to class III and class IV antibodies.
* In the combination therapy design, we included the class III antibody REGN10987 as well.

*Describe results specifically for 5 antibodies.*

* Are the results from foldx for the wild type antibody consistent with escape mutants for that class?
* Do the individual designs per class overcome resistance of the individual class?
* Is the combination of optimized class I and class II antibodies sufficient to overcome resistance of all classes or do we have to include the REG10987 antibody as well?

*Conclude (something related to binding domains and overlap of these antibodies)* (*How do RBD feature selection locations by random forest compare in terms of antibody vulnerability hotspots)*

**Combinatorial optimization for antibody selection**

*Motivation*

The potential for intra patient viral evolution is possible due to antibody selective pressure (cite regeneron, cite mouse model B cells).

*Describe the algorithm* (ND)

To estimate how mutations of the virus might affect new antibody design, we calculated differences in Gibbs free energies between mutations of the virus and antibody designs predicted by our algorithm.

*Describe results specifically for 5 antibodies . We ran the algorithm and found that we were able to optimize coverage of x RBD mutations, given …*

*Conclude (something related to binding domains and overlap of these antibodies, talk about viral mutations, rate of mutation, mutations in the pandemic that can overlap with binding domains for different antibodies, justification? nussenzweig paper on evolution experiment of some of the C1.. antibodies )*

**Discussion**

Recent advances in the identification and engineering of anti- HIV-1 antibodies have produced a large set of detailed molecular structures and neutralization data generated against a broad panel of HIV-1 strains. Recent computational analysis of antibody neutralization data has been successful in categorizing antibodies with respect to their neutralization activity [5], extracting the identities of RBD residues that are necessary for neutralization [17] and uncovering antibody epitopes [11]. Here, we report a computational methodology that utilizes antibody neutralization data and structural information to construct SAR fitness landscapes and reason about the dynamics of resistance. It consists of the interpretation of neutralization data using statistical inference, the construction of fitness landscapes using computational chemistry and the development of biophysical and mathematical model to capture the dynamics of replication, mutation and selection.

Our statistical model is able to uncover critical residues involved in antibody neutralization and is consistent with recent studies in antibody resistance. It does not identify the entire structural epitope involved in the protein-protein contact. Rather, satlasso identifies an epitope that is involved in the function of the protein-protein interaction, in our case neutralization.

**Methods**

**Encoding amino acid sequences**

We employ an one-hot encoding for amino acid sequence translation. We consider the list of amino acids *A* of length 20. To encode amino acid *ai*, we construct a vector *V* of the same length as *A*, such that for each element *vj* in vector *V*: *vj* = 1 if i = j and *vj* = 0 otherwise. To encoding a sequence of amino acids of length n, *aj1aj2 … ajn*, we use the described one-hot encoding for each amino acid in the sequence and concatenate these one-hot encodings in the other in which the amino acids appear in the sequence to obtain a vector of length 20n.

In the case that amino acid sequences provided were not equal in length, we used the ClustalOmega multi-sequence alignment tool to align our sequences and then translated the sequences using one-hot encoding.

**Random forest regressor**

We use the random forest regressor method to determine residues in the SARS-CoV-2 virus sequence most relevant to ACE2 receptor binding. We construct our training data by constructing a matrix in which the rows are the one-hot encoded amino acid sequences of the SARS-CoV-2 mutants (method described above). The labels for our training data are corresponding Δlog(Kd) for the encoded SARS-CoV-2 mutants. We used the sklearn RandomForestRegressor python package.

**Satlasso**

We define the saturated least absolute shrinkage and selection operator (satlasso) and formulate it as a convex optimization problem. We consider the antibody neutralization data X = Xu+Xs consisting of unsaturated data Xu, and of saturated data Xs, corresponding to sequences with IC50s values equal to the limit of neutralization assays. Observe n predictor response pairs (xi,yi) where xi ∈ Rp and yi ∈ R. Let m be the number of unsaturated predictor response pairs. Forming X ∈ Rn×p, X = Xu + Xs with standardized columns, the saturated lasso, (satlasso) is an estimator defined by the following convex optimization problem:

minimize β∈Rp {(1/m)λ1||yu − Xuβ||22 + (1/m)λ2|β|1 + λ3max(ys − Xsβ, 0) }

Model selection was performed by 5 fold-cross validation. Satlasso was implemented using the CVXPY python package and cross validation was performed using the sklearn KFold python package.

We use the implemented satlasso class to perform variable selection for residues relevant for neutralization. We construct our training data by constructing a matrix in which the rows are the one-hot encoded aligned amino acid sequences of the antibodies (method described above). The labels for our training data are corresponding IC50 (ng/ml) for the encoded antibodies.

**Binding Energy, Mutation and Mutation Coupling Calculations**

Let R, R^i be the reference (wild type) and mutant receptors (RBD), and let L, L^j be the reference or mutant ligand (antibody), with i∈{1,…,m} and j∈{ 1,…,n} and where m,n∈ Z^(>0) represent the number of single point mutations on RBD and the antibody, respectively.

We derive the difference in free energy of binding of the mutant R^i and ligand L, ∆∆G\_bind (R^i L) , given the free energies of folding ∆G\_unfold obtained from FoldX. We assume that the difference between bound and unbound unfolded reference states is approximately zero. The difference between bound and unbounded unfolded states is:

∆∆G\_unfold (R^i L-RL)- ∆∆G\_unfold (R^i-R)= ∆G\_unfold (R^i L)- ∆G\_unfold (RL)-(∆G\_unfold (R^i )- ∆G\_unfold (R))

= ∆∆G\_unfold (R^i L- R^i )-0 = ∆∆G\_bind (R^i L)

The difference in free energy of binding of the reference RBD R and mutant antibody〖 L〗^j, ∆∆G\_bind (RL^j ) is derived similarly. It follows that the effect of a mutation on RBD, R^i on binding to the antibody mutation L^j is

∆∆∆G\_mut (R^i L^j-RL^j )= ∆∆G\_bind (R^i L^j-RL^j )- ∆∆G\_bind (R^i L-RL).

The difference in binding energy due to RBD and antibody mutation coupling R^i and L^j is

∆∆∆∆G\_couple (R^i L^j )= ∆∆∆G\_mut (R^i L^j-RL^j )- ∆∆∆G\_mut (R^i L^j-RL^j ) = ∆G\_unfold (R^i L^j )- ∆G\_unfold (RL^j )-( ∆G\_unfold (R^i L)- ∆G\_unfold (RL))

To ensure that FoldX approximations are driven by changes local to the mutation pair, we consider mutations on ACE2 such that the distance between the antibody mutation and ACE2 is less than 4Å. We assume that coupling terms for mutations R^i and L^j should approximate zero unless R^i and L^j are in physical proximity and that energetically coupled residues are more likely to be compensating evolutionarily.

**Optimization algorithm**

Let the fitness function fv: R ⇒ R define the fitness of a virus, and fa: R ⇒ R define the fitness of an antibody. Then, we define the combined fitness of antibody *a* with respect to virus *v* as f(a,v) = fv(v) - fa(a), where f(a,v) > 0 indicates good fitness of the antibody with respect to the virus and f(a,v) < 0 indicates poor fitness of the antibody with respect to the virus. Given antibodies *a1*, ..., *an* and viruses *v1, …, vm*, define fitness matrix *F* = ( fij ), where fij = f(*ai* ,*vj* ) 1in, 1jm. Define Sk : Rp ⇒ R be a function that sums the smallest k elements of a vector in Rp, where k p. Let *I* be the incidence matrix of positive elements in *F*, i.e. *I* = ( Iij ) where Iij = 1 if fij > 0 and Iij = 0 otherwise. Then, we formulate the following optimization problem:

minimize c∈{0,1}^n { λ|c|1- ci fij } subject to ci 1 and Sk(Ic) 1,

where k is the maximum number of viruses permitted to be not covered by the chosen antibodies, and λ is a regularizing parameter. The solution to the optimization problem returns c ∈ {0,1}n where ci = 1 denotes that antibody ai is included for the optimal antibody cocktail, and ci = 0 indicates antibody ai is not included in the optimal cocktail.

**Data availability**

Molecular structures 6xcm (C105), 7bz5 (B38), 7c01 (CB6), 6xc3 (CC12.1), 6xe1 (CV30) and 6m0j (RBD) were downloaded from the PDB. Antibody neutralization and sequence data including C105 was downloaded from. VH3-55/66 antibody neutralization and sequence data was downloaded from . RBD mutational scanning and binding data was downloaded from . Sequence data for each antibody was downloaded from . All processed data used in this manuscript is deposited at [https://github.com/vdjonsson/gibbs\_fitness](https://github.com/vdjonsson/gibbs_fitnesssc)/data/ .

**Code availability**

satlasso code is available at https://github.com/vdjonsson/satlasso. Gibbs energy analysis code is available at [https://github.com/vdjonsson/gibbs\_fitness](https://github.com/vdjonsson/gibbs_fitnesssc). Code to generate figures for the paper can be found at https://github/com/vdjonsson/paper\_covab.

**References**

[1. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018 Dec 1;34(23):4121–3.](https://www.zotero.org/google-docs/?4PVi8I)

[2. Starr TN, Greaney AJ, Hilton SK, Crawford KHD, Navarro MJ, Bowen JE, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. bioRxiv [Internet]. 2020 Jun 17 [cited 2020 Aug 28]; Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7310626/](https://www.zotero.org/google-docs/?4PVi8I)

[3. Isabel S, Graña-Miraglia L, Gutierrez JM, Bundalovic-Torma C, Groves HE, Isabel MR, et al. Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. Sci Rep. 2020 Aug 20;10(1):14031.](https://www.zotero.org/google-docs/?4PVi8I)

[4. Liu L, Wang P, Nair MS, Yu J, Rapp M, Wang Q, et al. Potent neutralizing antibodies directed to multiple epitopes on SARS-CoV-2 spike. Nature [Internet]. 2020 Jul 22 [cited 2020 Jul 28]; Available from: http://www.nature.com/articles/s41586-020-2571-7](https://www.zotero.org/google-docs/?4PVi8I)

[5. Yuan M, Liu H, Wu NC, Lee C-CD, Zhu X, Zhao F, et al. Structural basis of a public antibody response to SARS-CoV-2 [Internet]. Microbiology; 2020 Jun [cited 2020 Aug 28]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.06.08.141267](https://www.zotero.org/google-docs/?4PVi8I)

[6. Baum A, Copin R, Ajithdoss D, Zhou A, Lanza K, Negron N, et al. REGN-COV2 antibody cocktail prevents and treats SARS-CoV-2 infection in rhesus macaques and hamsters [Internet]. Microbiology; 2020 Aug [cited 2020 Aug 28]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.08.02.233320](https://www.zotero.org/google-docs/?4PVi8I)

[7. Baum A, Fulton BO, Wloga E, Copin R, Pascal KE, Russo V, et al. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. Science. 2020 Aug 21;369(6506):1014–8.](https://www.zotero.org/google-docs/?4PVi8I)

[8. Robbiani DF, Gaebler C, Muecksch F, Lorenzi JCC, Wang Z, Cho A, et al. Convergent antibody responses to SARS-CoV-2 in convalescent individuals. Nature. 2020 Aug;584(7821):437–42.](https://www.zotero.org/google-docs/?4PVi8I)

[9. Barnes CO, West AP, Huey-Tubman KE, Hoffmann MAG, Sharaf NG, Hoffman PR, et al. Structures of Human Antibodies Bound to SARS-CoV-2 Spike Reveal Common Epitopes and Recurrent Features of Antibodies. Cell. 2020 Aug;182(4):828-842.e16.](https://www.zotero.org/google-docs/?4PVi8I)

[10. Seydoux E, Homad LJ, MacCamy AJ, Parks KR, Hurlburt NK, Jennewein MF, et al. Characterization of neutralizing antibodies from a SARS-CoV-2 infected individual [Internet]. Immunology; 2020 May [cited 2020 Aug 28]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.05.12.091298](https://www.zotero.org/google-docs/?4PVi8I)

[11. Hansen J, Baum A, Pascal KE, Russo V, Giordano S, Wloga E, et al. Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. Science. 2020 Aug 21;369(6506):1010–4.](https://www.zotero.org/google-docs/?4PVi8I)

[12. Brouwer PJM, Caniels TG, Straten K van der, Snitselaar JL, Aldon Y, Bangaru S, et al. Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. Science. 2020 Aug 7;369(6504):643–50.](https://www.zotero.org/google-docs/?4PVi8I)

[13. Shi R, Shan C, Duan X, Chen Z, Liu P, Song J, et al. A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. Nature. 2020 Aug;584(7819):120–4.](https://www.zotero.org/google-docs/?4PVi8I)

[14. Zost SJ, Gilchuk P, Chen RE, Case JB, Reidy JX, Trivette A, et al. Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the SARS-CoV-2 spike protein. Nat Med. 2020 Jul 10;1–6.](https://www.zotero.org/google-docs/?4PVi8I)

[15. Rogers TF, Zhao F, Huang D, Beutler N, Burns A, He W, et al. Isolation of potent SARS-CoV-2 neutralizing antibodies and protection from disease in a small animal model. Science. 2020 Aug 21;369(6506):956–63.](https://www.zotero.org/google-docs/?4PVi8I)

[16. Wu Y, Wang F, Shen C, Peng W, Li D, Zhao C, et al. A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. Science. 2020 Jun 12;368(6496):1274–8.](https://www.zotero.org/google-docs/?4PVi8I)

[17. Raybould MIJ, Kovaltsuk A, Marks C, Deane CM. CoV-AbDab: the Coronavirus Antibody Database. Bioinformatics [Internet]. [cited 2020 Aug 28]; Available from: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa739/5893556](https://www.zotero.org/google-docs/?4PVi8I)

**Acknowledgements**

We thank Anthony West and Pamela Bjorkman. This work was supported by an award from the National Cancer Institute of the National Institutes of Health (K12CA001727). N.D. and T.F.S. were supported by the Norton Anonymous donor fund. T.F.S. was also supported by the Donald Voet and Jerome Vinograd SURF Fellowship. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**Author Information**

**Affiliations**

**Contributions**

V.D.J. conceived the project. V.D.J., T.F.S. and N.D. analyzed data. V.D.J., T.F.S and N.D. implemented analytic and computational pipelines. First manuscript draft: V.D.J. and N.D. Final manuscript: V.D.J. and N.D. with contributions for all authors.

**Corresponding authors**

Correspondence to Vanessa D. Jonsson.

**Ethics declarations**

**Competing interests**

**Extended Data**

**Supplementary Information**

Tables:

(Note: had to clip the saturated values to the saturated point in order to make R^2 make sense for all data)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | Method used | Lambda1 | Lambda2 | Lambda3 | R^2 unsaturated | R^2 all data |
| bjorkman | satlasso | 5.00 | 10.00 | 6.25 | 0.9051526686895268 | 0.9933961240684538 |
| vh3-53/66 | satlasso | 1.00 | 10.00 | 10.00 | 0.9985044559888621 | 0.9987680200117586 |
| bloom | random forest regressor | - | - | - | - | 0.9412980841486871 |