Feature selection and combinatorial optimization on fitness landscapes to constrain anti-SARS-CoV2 antibody design and address viral escape

Natalie Dullerud[1], Tea Freedman-Susskind[2], Priyanthi Gnanapragasam[3], Christopher Snow[4], Anthony P West Jr[3], Vanessa D Jonsson[5,6]

1: Department of Mathematics, University of Southern California, Los Angeles, CA, USA.
2: Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA.
3: Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA.
4: Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO, USA.
5: Department of Hematology, City of Hope, Duarte, CA, USA.
6: Department of Computational and Quantitative Medicine, City of Hope, Duarte, CA, USA.

Extended Abstract

Effective therapeutic strategies to mitigate the extent of the COVID-19 pandemic — with more than 1 million fatalities as of October 23, 2020 — are crucial prior to the implementation of a readily available vaccine. Passive transfer of anti-SARS-CoV-2 antibodies for prophylaxis or therapy has been proposed as an interim solution: the isolation, design and prototyping of antibodies with enhanced neutralization against a broad range of viral mutations is of critical importance as is the rational design of antibody cocktails to target viral escape[1–5]. To this end, several studies have structurally classified anti-SARS-CoV2 antibody binding to the receptor-binding-domain (RBD) of the virus spike protein[4], and used directed evolution to quantify antibody and virus fitness and inform antibody combinations[6].

Antibodies that most effectively neutralize viruses are developed by searching for antibody mutations that can simultaneously maximize binding while being robust to viral escape. This amounts to jointly optimizing virus and antibody fitness landscapes, and link the concept of antibody amino acid sequence to antibody fitness subject to mutations in virus sequence. We propose a combined machine learning and combinatorial optimization method to jointly constrain antibody design space and predict antibody cocktails to address viral escape.

We describe a new statistical model based on the least absolute shrinkage and selection operator[7] (LASSO) to extract features of antibody sequences that contribute or detract from neutralization while accounting for saturated data characteristic of antibody neutralization assays. Given neutralization information partitioned into two data sets — measurements within the limits of the assay (unsaturated) and at the limit of the assay (saturated) — the saturated LASSO (satlasso) extends the LASSO objective function by penalizing predictions for unsaturated data using squared error and L_1-regularization and, for saturated data, if the prediction underestimates the true value. The application of satlasso to antibody sequence and neutralization data combined with mutational scanning on antibody structures identifies sequence positions and mutations that have the potential to enhance antibody neutralization. Saturated LASSO significantly outperforms standard regression algorithms used for feature selection, such as least-angles regression, LASSO, and random forest regression, when applied to several neutralization data sets (Table 1).

Constraining antibody design space with feature selection and mutational scanning

We used publicly available sequence and neutralization data from 93 antibodies that were recovered from convalescent COVID-19 patients[8] and performed satlasso selection to uncover antibody sequence features amenable to optimization (Fig 1e). We focused on the C105 antibody, whose antigen-complex structure has been solved[1]. C105 can bind to the S trimer in two states: with all three RBDs in the “up” state or with two RBDs “up” and one “down” by resting its three heavy chain regions and two of its light chain regions against the receptor-binding ridge of the RBD. It shares binding characteristics with other investigated antibodies, most notably sharing binding mode with B38, CB6, CC12.1, and CV30[1]. The relatively wide range of binding as well as shared characteristics makes C105 a good antibody for case study. Predictions for the C105 antibody yielded five dominant sequence locations, two of which were found to be common in other VH3-53 anti-SARS-CoV2 antibodies (Fig. 1c). In particular, heavy chain mutations TH28I and YH58F were predicted to significantly contribute to SARS-CoV2 neutralization; this optimized antibody $C105^{TH28I-}$

^{YH58F} was validated experimentally and found to have a 15-fold increase in neutralization potential compared to non-optimized C105 (0.0055 ug/mL vs 0.0354 ug/mL IC_{50}) (Fig. 1e).

Combinatorial optimization on fitness landscapes to design antibody cocktails

To explore antibody sensitivity in the context of viral escape, we developed a biophysical antibody/RBD binding model that utilizes data derived from energy minimization calculations on structural information – and assessed binding energy differences due to mutations on RBD. Clustering and dimensionality reduction on resulting fitness landscapes clustered antibodies with similar viral sensitivities and highlighted viral sensitivities unique to certain antibodies (Fig. 2ab).

Several studies have shown escape from antibody selective pressure[2,3] : structural studies have recently classified anti SARS-COV-2 antibodies into four classes[4], shown that escape is orthogonal in these classes, and have proposed a strategy for combinatorial antibody design[4,5]. In efforts to optimize treatment options for patients with COVID-19, antibody polytherapy has been examined as a potential avenue — notably, antibody cocktails exceed single antibody treatments in efficacy for neutralizing SARS-CoV-2 and targeting RBD mutations both *in vitro* and in a rhesus macaque model [2,6].

We developed a combinatorial optimization algorithm that utilizes binding energy landscapes generated for mutations of RBD with respect to infection (angiotensin converting enzyme 2 (ACE2) and RBD) and neutralization (RBD and antibody) to determine antibody combinations that most effectively target mutations on RBD. This algorithm inputs a desired virus mutation coverage and outputs an antibody cocktail that jointly maximizes combined antibody neutralization and virus mutation coverage (Fig. h). The algorithm takes as parameters a fitness matrix that represents antibodies against virus mutations, and a maximum bound on the proportion of viruses allowed to be unmanaged by the chosen combination of antibodies. The algorithm then jointly maximizes the virus mutations covered and minimizes the number of antibodies in the cocktail, such that the proportion of unmanaged virus mutations is below the given bound. This combinatorial optimization algorithm yields combinations of antibodies to simultaneously constrain the antibody design space and address viral escape mutations. Our method can be used in combination with high throughput experimental platforms like directed evolution or yeast display, and provides a method to jointly constrain antibody design using available neutralization and sequence data, and propose antibody combinations to address virus escape.

Data availability

Molecular structures 6xcm (C105), 7bz5 (B38), 7c01 (CB6), 6xc3 (CC12.1), 6xe1 (CV30) and 6m0j (RBD) were downloaded from the PDB (https://www.rcsb.org). The following data was publicly availble an obtained from referenced studies; antibody neutralization and sequence data including C105[9]; VH3-53/66 antibody neutralization and sequence data[9–16]; RBD mutational scanning and binding data[3]; sequence data for each antibody[9–16] . All mutated PDB structures, processed data used in this study is deposited at https://github.com/vdjonsson/sars-cov2-ab/data/.

Code availability

satlasso code and tutorial: https://github.com/vdjonsson/satlasso . code to reproduce results in this study is found at https://github.com/vdjonsson/sars-cov2-ab/data/.

Tables

		MSE	$MSE_{unsaturated}$	Saturated Loss	R^2	$R^2_{unsaturated}$
Neutralization data for C002-C215	Lars	$4.4x10^{49}$	$7.9x10^{49}$	$7.9x10^{49}$	$-2.1x10^{44}$	$-3.1x10^{45}$
	Lasso	$2.4x10^{5}$	$2.1x10^{4}$	$2.2x10^{4}$	-0.19	0.19
	RF	$7.1x10^{4}$	$4.3x10^{3}$	$5.1x10^{3}$	0.66	0.83
	satlasso	$4.4x10^{6}$	$2.5x10^{3}$	$2.5x10^{3}$	0.99	0.91
Neutralization data for VH3-53/66	Lars	$1.1x10^{24}$	$1.2x10^{24}$	$1.2x10^{24}$	$-4.0x10^{18}$	$-4.8x10^{18}$
	Lasso	$2.8x10^{5}$	$2.5x10^{5}$	$2.6x10^{5}$	0.0	0.0
	RF	$4.6x10^{4}$	$4.1x10^{4}$	$4.2x10^{4}$	0.83	0.83
	satlasso	$3.4x10^{2}$	$3.7x10^{2}$	$3.7x10^{2}$	0.99	0.99

Table 1 | Metrics for regression models trained on the two datasets considered in the paper. Regression models considered are least-angle regression (Lars), least absolute shrinkage and selection operator (Lasso), random forest regressor (RF) and saturated LASSO (satlasso). The metrics calculated include: mean squared error on full dataset; mean squared error on unsaturated partition of dataset; saturated loss defined by sum of mean squared error on unsaturated partition of dataset and maximum over difference of predicted values from true values and zero on saturated partition of dataset; R^2 score on full dataset; and R^2 score on unsaturated partition of dataset. Neutralization data for antibodies C002-215 from [9]. Neutralization data for antibodies with heavy V gene VH3-53 or VH3-66 from [9–15].

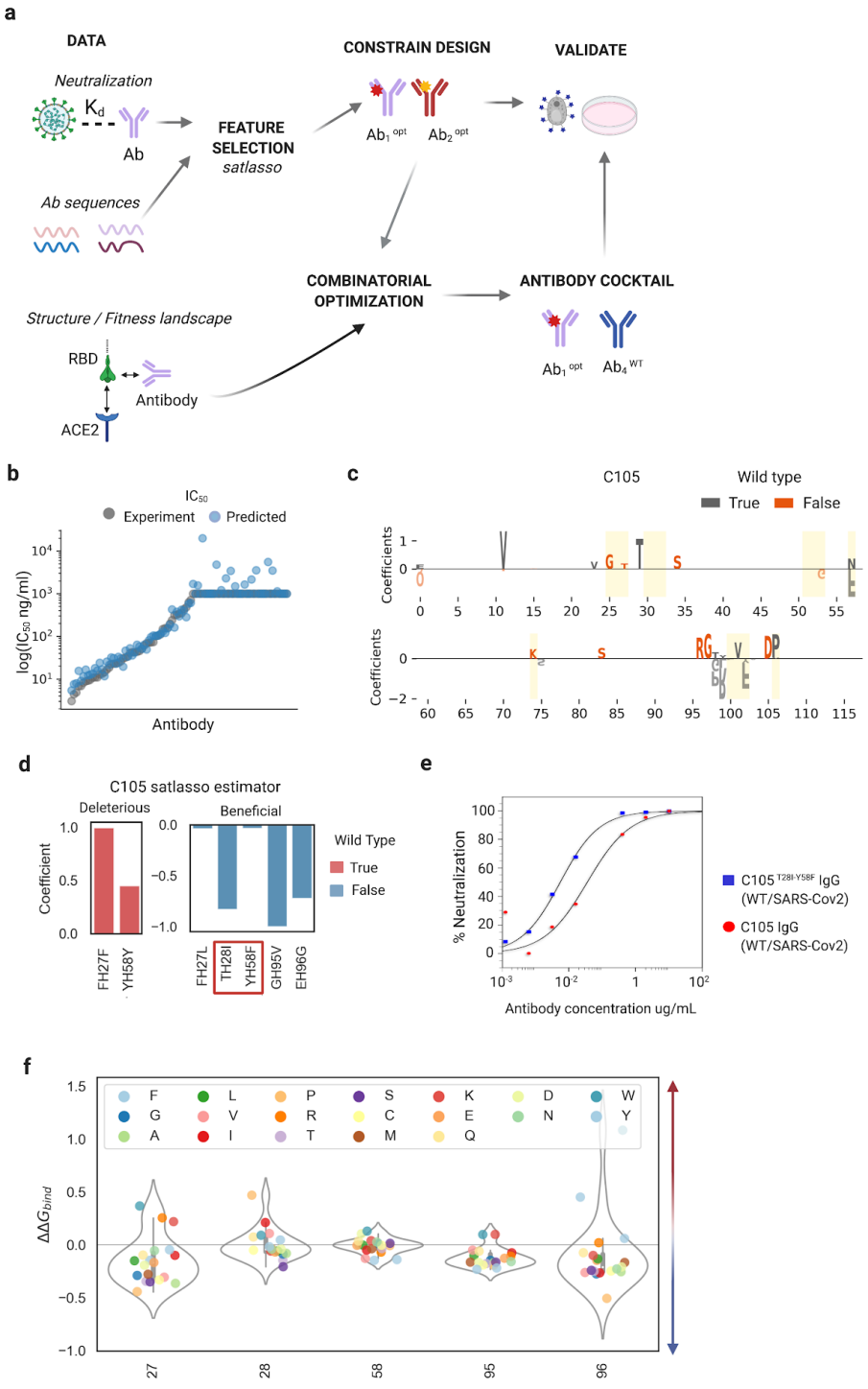

Figure 1 | a. Illustration of the data and methods used in this study. **b.** Scatter plot of experimentally derived neutralization IC_{50} (grey) from 93 antibodies derived from convalescent COVID-19 patients[9] and IC_{50} prediction using the satlasso model (blue). **c.** Logo plot of antibody sequence features selected by satlasso for C105 antibody. Negative coefficients indicate sequence features that enhance neutralization and positive coefficients detract from neutralization.

Yellow lines indicate RBD binding locations. **d.** Coefficients larger than 10^{-3} for satlasso prediction for C105 antibody optimization, locations on the C105 antibody where the wild type amino acid is deleterious (left) and antibody features that are predicted to be beneficial for antibody neutralization of SARS-CoV2. Red box indicates mutations selected for experimental validation. **e.** Neutralization assay performed for C105 and predicted optimized antibody $C105^{TH28I-YH58F}$. **f.** Difference in binding energy between C105 mutated antibody and C105 WT antibody calculated by mutational scanning and energy minimization on the C105 molecular structure (PDB:6xdg) using FoldX. Red arrow and blue arrow indicate decreased or increased antibody binding with respect RBD mutation.

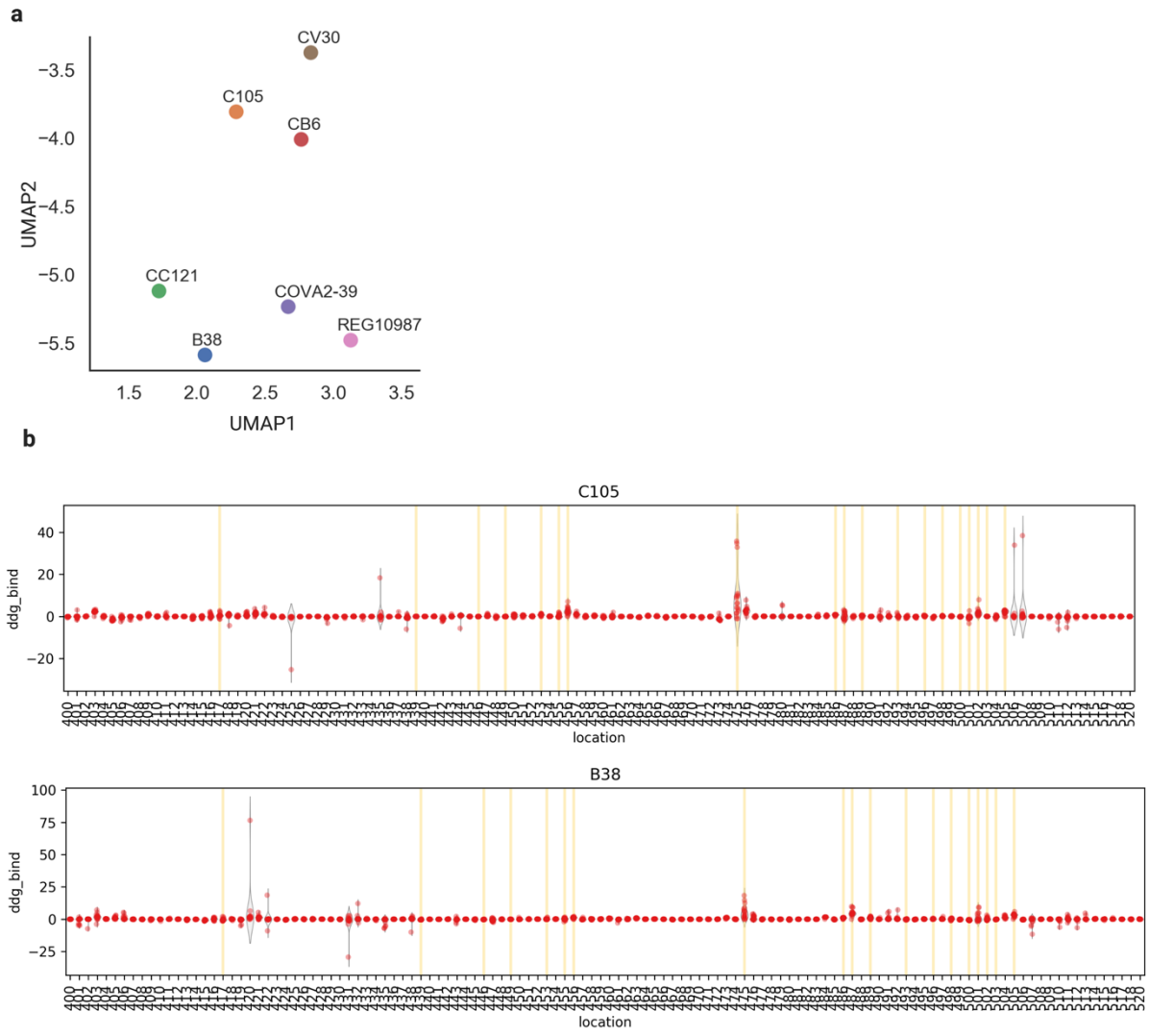

Figure 2 |. **a.** Uniform manifold approximation (UMAP) embedding of RBD/antibody binding fitness landscape calculated using mutational scanning and energy minimization on antibody structures for seven different antibodies. **b.** Violinplot with binding energies and locations on the RBD genome that are sensitive to antibody binding. Yellow lines indicate binding with RBD.

References

1. Barnes, C. O. *et al.* Structures of Human Antibodies Bound to SARS-CoV-2 Spike Reveal Common Epitopes and Recurrent Features of Antibodies. *Cell* **182**, 828-842.e16 (2020).
2. Baum, A. *et al. REGN-COV2 antibody cocktail prevents and treats SARS-CoV-2 infection in rhesus macaques and hamsters*. http://biorxiv.org/lookup/doi/10.1101/2020.08.02.233320 (2020) doi:10.1101/2020.08.02.233320.
3. Starr, T. N. *et al.* Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *bioRxiv* (2020) doi:10.1101/2020.06.17.157982.
4. Barnes, C. O. *et al. Structural classification of neutralizing antibodies against the SARS-CoV-2 spike receptor-binding domain suggests vaccine and therapeutic strategies*. http://biorxiv.org/lookup/doi/10.1101/2020.08.30.273920 (2020) doi:10.1101/2020.08.30.273920.
5. Weisblum, Y. *et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants*. http://biorxiv.org/lookup/doi/10.1101/2020.07.21.214759 (2020) doi:10.1101/2020.07.21.214759.
6. Baum, A. *et al.* Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science* **369**, 1014–1018 (2020).
7. Regression shrinkage and selection via the lasso: a retrospective - Tibshirani - 2017 - Journal of the Royal Statistical Society: Series B (Statistical Methodology) - Wiley Online Library. https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9868.2011.00771.x%4010.1111/%28ISSN%291467-9868.TOP_SERIES_B_RESEARCH?casa_token=tm8UadS4OZ8AAAAA%3AzOqx14a2pY3kt2pEnRaAy-E42LgCdnf4dw29OFnBrO6PT2ko8_S9Syihgn1DcN5V_VbaO3yr55I6T64.
8. Raybould, M. I. J., Kovaltsuk, A., Marks, C. & Deane, C. M. CoV-AbDab: the Coronavirus Antibody Database. *Bioinformatics* doi:10.1093/bioinformatics/btaa739.
9. Robbiani, D. F. *et al.* Convergent antibody responses to SARS-CoV-2 in convalescent individuals. *Nature* **584**, 437–442 (2020).
10. Seydoux, E. *et al. Characterization of neutralizing antibodies from a SARS-CoV-2 infected individual*. http://biorxiv.org/lookup/doi/10.1101/2020.05.12.091298 (2020) doi:10.1101/2020.05.12.091298.

11. Hansen, J. *et al.* Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. *Science* **369**, 1010–1014 (2020).

12. Brouwer, P. J. M. *et al.* Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. *Science* **369**, 643–650 (2020).

13. Shi, R. *et al.* A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature* **584**, 120–124 (2020).

14. Zost, S. J. *et al.* Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the SARS-CoV-2 spike protein. *Nature Medicine* 1–6 (2020) doi:10.1038/s41591-020-0998-x.

15. Rogers, T. F. *et al.* Isolation of potent SARS-CoV-2 neutralizing antibodies and protection from disease in a small animal model. *Science* **369**, 956–963 (2020).

16. Wu, Y. *et al.* A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. *Science* **368**, 1274–1278 (2020).