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

Natalie Dullerud1, Tea Freedman-Susskind2 , Priyanthi Gnanapragasam, Christopher Snow3, Anthony P West Jr4, Vanessa D Jonsson5,6

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

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](https://www.zotero.org/google-docs/?KEbYqS). To this end, several studies have structurally classified anti-SARS-CoV2 antibodies with respect binding to the virus spike protein receptor binding domain (RBD) (cite), and used directed evolution to quantify antibody and virus fitness and inform antibody combinations [6](https://www.zotero.org/google-docs/?jD4yi4).

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 (LASSO) (cite), to extract features of antibody sequences that contribute or detract from neutralization while accounting for saturated data characteristic of antibody neutralization assays. The application of the saturated LASSO (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.

*Constraining antibody design with machine learning*

We used publicly available sequence and neutralization data from 93 antibodies that were recovered from convalescent COVID-19 patients (cite) 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](https://www.zotero.org/google-docs/?FOcHkv). 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[1](https://www.zotero.org/google-docs/?VHFDTZ). It shares binding characteristics with other investigated antibodies, most notably sharing a binding mode with B38, CB6, CC12.1, and CV30[1](https://www.zotero.org/google-docs/?S9l46J). 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 C105TH28I-YH58F was validated experimentally and found to have significantly increased neutralization compared to non-optimized C105 (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 1fg).

Several studies have shown escape from antibody selective pressure[4,6](https://www.zotero.org/google-docs/?0nlrmq). Structural studies have recently classified anti SARS-COV-2 antibodies into four classes*[2](https://www.zotero.org/google-docs/?V2v9t1)*, shown that escape is orthogonal in these classes, and have proposed a strategy for combinatorial antibody design[*2,7*](https://www.zotero.org/google-docs/?0yJcLE). In efforts to optimize treatment options for patients with COVID-19, antibody polytherapy has been examined as a potential avenue in several recent studies[*4,8*](https://www.zotero.org/google-docs/?bAxX8f) – 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 (cite).

We developed a combinatorial optimization algorithm that utilizes binding energy landscapes generated for mutations of RBD for both ACE2/RBD and RBD/hAb binding reactions 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). 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 provide a comprehensive quantification of both antibody design space and virus mutational space that otherwise would be experimentally intractable.

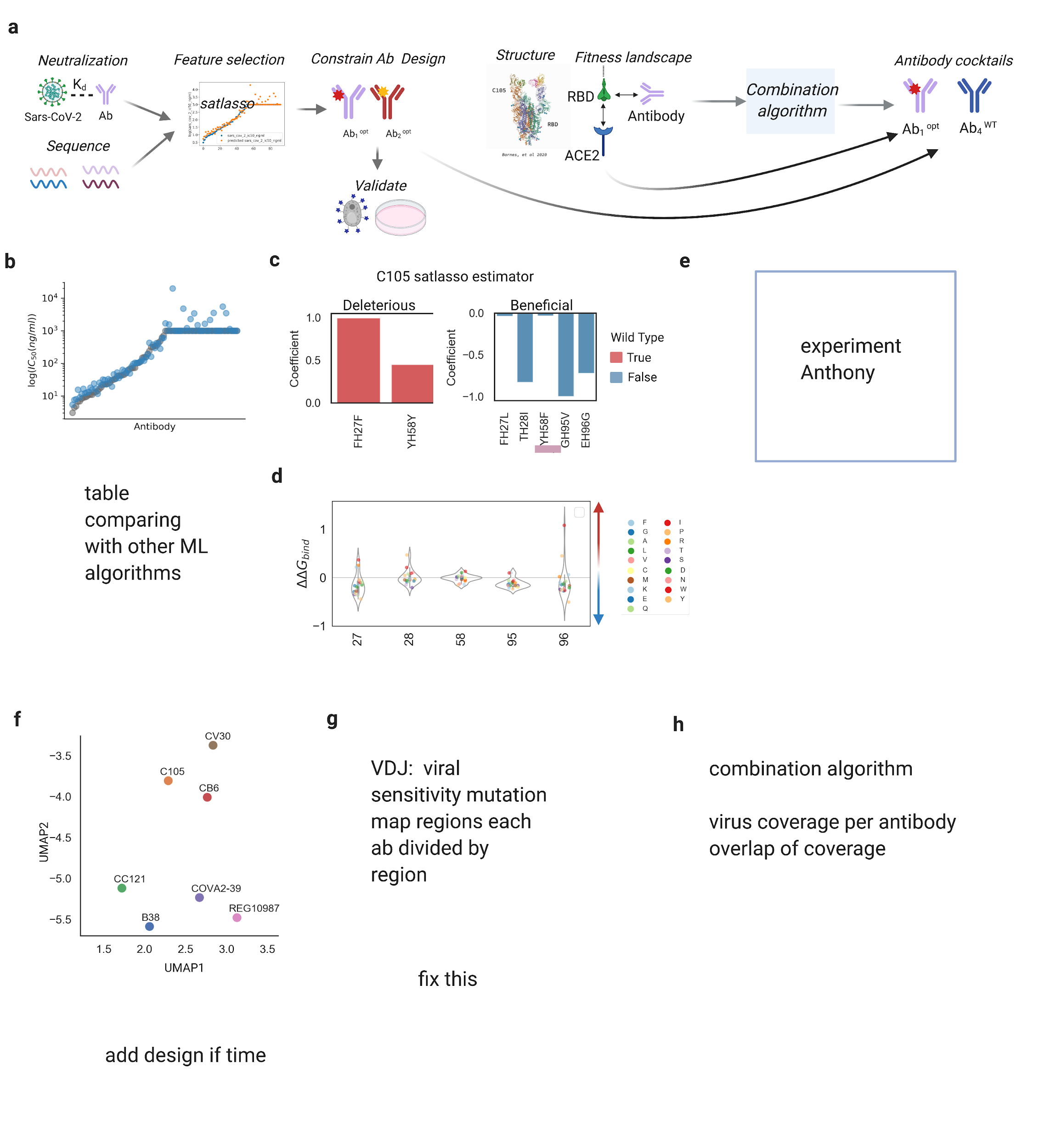
**Data availability**

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. VH3-53/66 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 mutated PDB structures, processed data used in this manuscript is deposited at https://github.com/vdjonsson/gibbsfitness/data/.

**Code availability**

Satlasso code can be found here

Antibody pipeline can be found here



**Figure 1 |** **a.** Illustration of the data and methods used in this study. **b.** Scatter plot of experimentally derived neutralization IC50 (grey) and IC50 prediction using satlasso (blue) for VH3-53-66 antibodies (top) and antibodies found in convalescent COVID-19 patients in (cite) **c.** Top coefficients 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. **d.** 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. **e.** Neutralization curves for C105 WT and C105-TH28I-YH58F. **f.** UMAP representation of RBD/antibody binding (list antibodies in this figure) calculated using mutational scanning and energy minimization for the following antibodies. **g.** Violinplot with binding energies and locations on the RBD genome that are sensitive to antibody binding. **H.** Results of combination therapy algorithm on antibodies described in **f** and optimized antibody C105-TH28I-YH58F: frequency of virus mutational coverage per antibody, neutralization for each antibody.