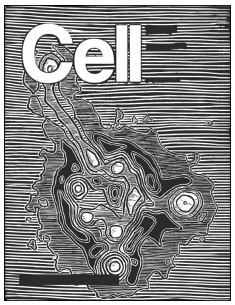
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.



Deep Mutational Learning Predicts ACE2 Binding and Antibody Escape to Combinatorial Mutations in the SARS-CoV-2 Receptor Binding Domain

Joseph M. Taft, Cédric R. Weber, Beichen Gao, Roy A. Ehling, Jiami Han, Lester Frei, Sean W. Metcalfe, Max Overath, Alexander Yermanos, William Kelton, Sai T. Reddy

PII: S0092-8674(22)01119-9

DOI: https://doi.org/10.1016/j.cell.2022.08.024

Reference: CELL 12608

To appear in: *Cell*

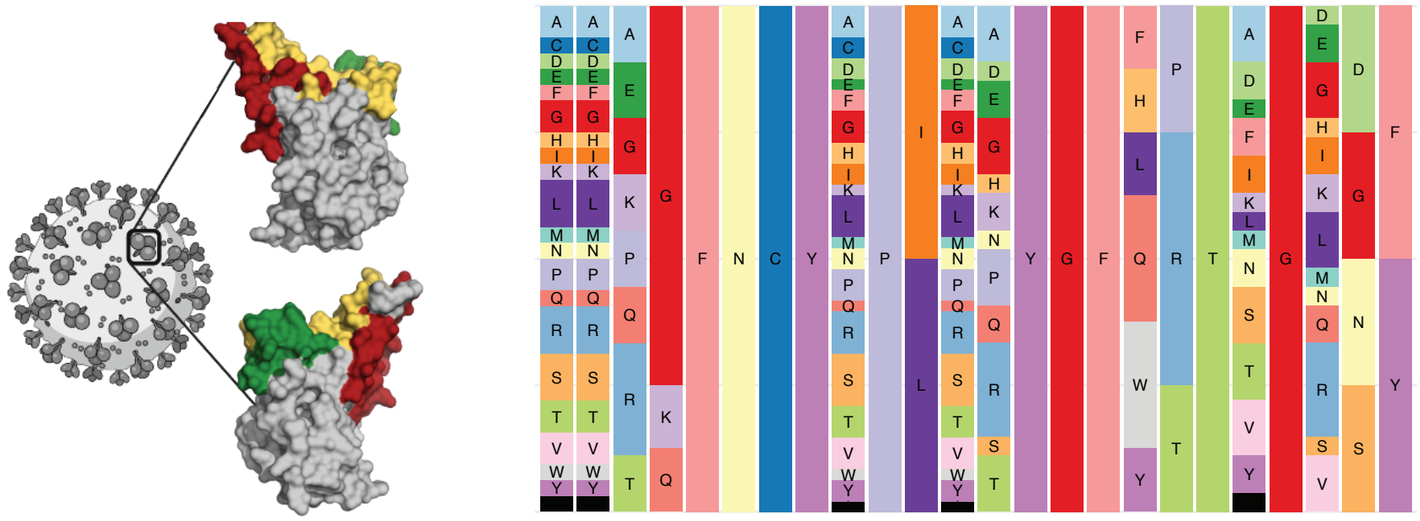
Received Date: 9 December 2021

Revised Date: 22 June 2022

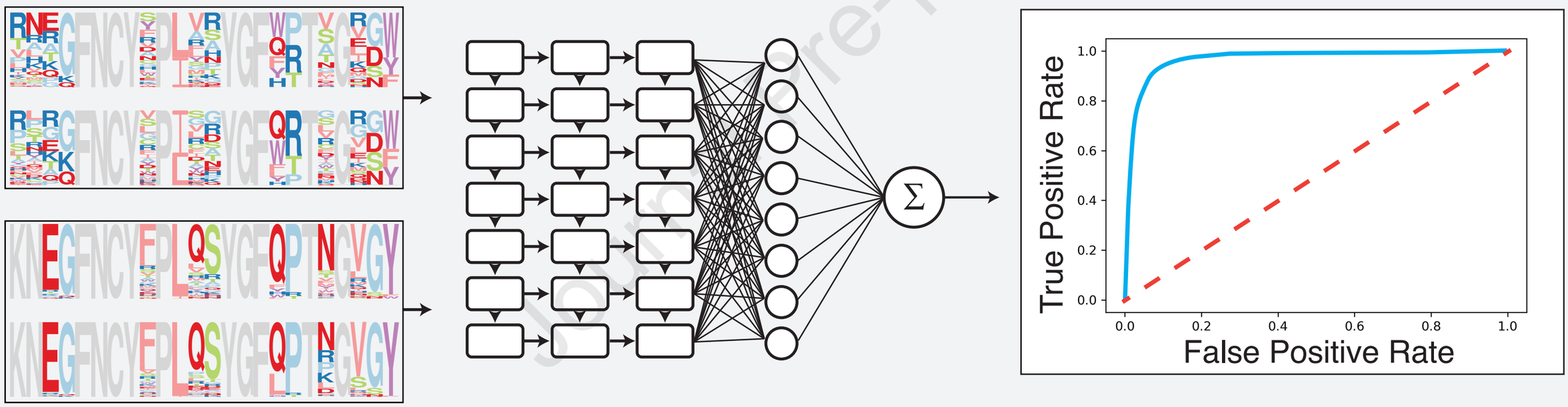
Accepted Date: 25 August 2022

Please cite this article as: Taft, J.M., Weber, C.R., Gao, B., Ehling, R.A., Han, J., Frei, L., Metcalfe, S.W., Overath, M., Yermanos, A., Kelton, W., Reddy, S.T., Deep Mutational Learning Predicts ACE2 Binding and Antibody Escape to Combinatorial Mutations in the SARS-CoV-2 Receptor Binding Domain, *Cell* (2022), doi: https://doi.org/10.1016/j.cell.2022.08.024.

RBD library design and screening

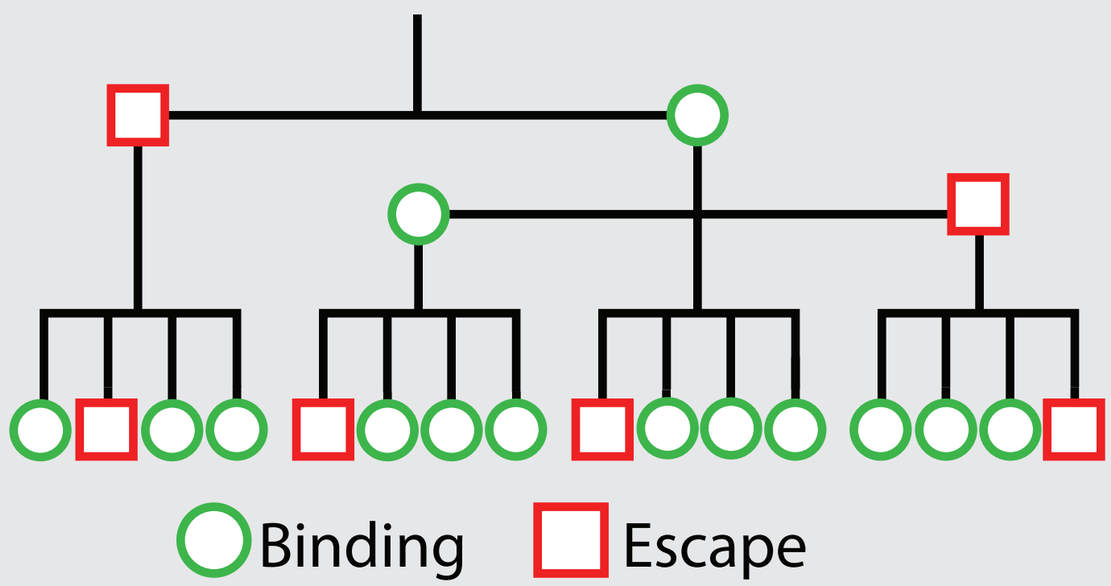


Deep sequencing and machine learning



Prediction of ACE2 binding and antibody escape of RBD variants

Sequence	ACE2 RNN	ACE2 RF	ACE2 Label	mAb RNN	mAb RF	mAb Label
KNEGFNCYYPLQSYGFQ	0.99	0.85	+	0.92	0.72	+
KNKGFNCYIPLPSYGFQ	0.01	0.18	-	0.03	0.44	-
KNQGFNCYFPLPAYGFL	0.02	0.26	-	0.08	0.36	-
KNEGFNCYFPIQSYGFL	0.53	0.27	-	0.67	0.58	+
KNKGFNCYYPLQAYGFH	0.82	0.54	+	0.91	0.82	+
KNKGFNCYFPIKSYGFH	0.15	0.81	-	0.04	0.31	-
KNKGFNCYFPIQTYGFH	0.87	0.79	+	0.56	0.57	+
KNKGFNCYFPLQAYGFH	0.98	0.34	-	0.81	0.80	+
KNQGFNCYLPLEAYGFQ	0.93	0.64	+	0.23	0.35	-
KNQGFNCYLPLHTYGFH	0.68	0.67	+	0.80	0.67	+
KNQGFNCYSPLQAYGFH	0.97	0.76	+	0.95	0.86	+
KNAGFNCYCPIQSYGFH	0.69	0.52	+	0.70	0.69	+
KNEGFNCYFPIQTYGFH	0.76	0.69	+	0.75	0.43	-



Deep Mutational Learning Predicts ACE2 Binding and Antibody Escape to Combinatorial Mutations in the SARS-CoV-2 Receptor Binding Domain

Joseph M. Taft[1, 2]*, Cédric R. Weber[3]*, Beichen Gao[1, 2], Roy A. Ehling[1], Jiami Han[1, 2], Lester Frei[1, 2], Sean W. Metcalfe[1], Max Overath[1], Alexander Yermanos[1, 2, 4, 5], William Kelton[6] and Sai T. Reddy[1,2#].

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel 4058, Switzerland
[2]Botnar Research Centre for Child Health, Basel 4058, Switzerland
[3]Alloy Therapeutics (Switzerland) AG, Basel 4058, Switzerland
[4]Department of Biology, Institute of Microbiology and Immunology, ETH Zurich, Zurich 8093, Switzerland
[5]Department of Pathology and Immunology, University of Geneva, Geneva 1211, Switzerland
[6]Te Huataki Waiora School of Health, University of Waikato, Hamilton 3240, New Zealand

*Equal contribution
#Lead contact: sai.reddy@ethz.ch

Summary

The continual evolution of the SARS-CoV-2 and the emergence of variants that show resistance to vaccines and neutralizing antibodies threaten to prolong the COVID-19 pandemic. Selection and emergence of SARS-CoV-2 variants are driven in part by mutations within the viral spike protein and in particular the ACE2 receptor-binding domain (RBD), a primary target site for neutralizing antibodies. Here, we develop deep mutational learning (DML), a machine learning-guided protein engineering technology, which is used to interrogate a massive sequence space of combinatorial mutations, representing billions of RBD variants, by accurately predicting their impact on ACE2 binding and antibody escape. A highly diverse landscape of possible SARS-CoV-2 variants is identified that could emerge from a multitude of evolutionary trajectories. DML may be used for predictive profiling on current and prospective variants, including highly mutated variants such as Omicron, thus guiding the development of therapeutic antibody treatments and vaccines for COVID-19.

INTRODUCTION

Throughout the course of 2021 and 2022, variants of SARS-CoV-2 associated with higher transmissibility and/or immune evasion (antibody escape) have supplanted the original founder strain (Wu-Hu-1) (Wu et al., 2020a). Such variants often possess at least one mutation in the RBD, which can directly influence binding to ACE2 (Supasa et al., 2021; Yi et al., 2020). For example, Alpha (B.1.1.7), Beta (B.1.351), and Gamma (P.1) variants all possess the N501Y mutation, which is associated with higher affinity binding to ACE2 (Han et al., 2021), suggesting this may represent a possible selective pressure for variant emergence.

Neutralizing antibodies, including monoclonal antibody therapeutics and those induced by vaccination (with the original Wu-Hu-1 spike protein), often display reduced binding and neutralization to variants. Detailed molecular analysis has revealed that many neutralizing antibodies to SARS-CoV-2 share sequence and structural features (Nielsen et al., 2020; Yang et al., 2021), which has led to their categorization into four common classes defined by groups of targeted RBD epitopes (Barnes et al., 2020; Harvey et al., 2021). For example, Class 1 antibodies include the previously clinically used REGN10933 (casirivimab) (Baum et al., 2020; Hansen et al., 2020) and LY-CoV16 (etesevimab) (Shi et al., 2020). Circulating variants with mutations in position K417 [e.g., Beta, Gamma, and Delta plus (B.1.617.2 + K417N)] as well as the mink-selected Y453F mutation (Cluster 5) display decreased neutralization by these class 1 antibodies (Hoffmann et al., 2021; Starr et al., 2021a; Wang et al., 2021b). Class 2 neutralizing antibodies including the clinically used LY-CoV555 (bamlanivimab) also strongly inhibit ACE2 binding, however, variants such as Beta, Gamma, eta (B.1.525), Kappa (B.1617.1), and iota (B1.526) all possess the RBD mutation E484K/Q that can lead to a substantial loss of binding and neutralization. Class 3 antibodies, including the clinically used REGN10987 (imdevimab) and S309 (sotrovimab) (Pinto et al., 2020), bind partially conserved epitopes and are resistant to several variants (e.g., Alpha, Beta, Gamma) (Tzou et al., 2020). Class 4 antibodies such as CR3022 target a highly conserved epitope among sarbecoviruses (ter Meulen et al., 2006; Wu et al., 2020b; Yuan et al., 2020), and have been largely resistant to escape variants, but generally lack neutralizing potency since they do not directly inhibit ACE2 binding.

The emergence of Omicron has revealed variants can evolve with severe immune evasion properties such as escape from several classes of neutralizing antibodies that bind to a diverse set of RBD epitopes (Cao et al., 2022a; Dejnirattisai et al., 2022; Iketani et al., 2022; Liu et al., 2022). Notably, nearly all clinically approved antibody therapies have lost substantial neutralizing activity against Omicron (BA.1 / B.1.1.529), including the multi-class antibody cocktails from Eli Lilly (LY-CoV16+LyCoV555), Regeneron (REGN10933+REGN10987) and AstraZeneca (AZD8895+AZD1061) (Cao et al., 2022a) all of which have subsequently had their clinical authorization revoked (Eli Lilly and Regeneron) or dosage modified (AstraZeneca) by the US FDA (FDA, 2022a, 2022c; NIH, 2022). One exception was S309, originally isolated from B cells of a patient infected with SARS-CoV-1 and possessing cross-reactivity to SARS-CoV-2 (Pinto et al., 2020), S309 has reduced but still potent neutralizing activity against Omicron (BA.1) (Cameroni et al., 2022; Sheward et al., 2022), likely because it binds to a highly conserved epitope found across genetically diverse sarbecoviruses. However, the Omicron sublineage BA.2 shows substantial escape from S309 (Iketani et al., 2022) and became widely circulating in early 2022, thus leading to the loss of clinical authorization of S309 (FDA, 2022b). As of August 2022, LY-CoV1404 (bebtelovimab) is the only clinically approved antibody discovered based on binding to the original Wu-Hu-1 RBD (Westendorf et al., 2022) that has maintained strong neutralizing activity to Omicron BA.1 and BA.2 variants (Iketani et al., 2022), as well as emerging sublineages BA.4 and BA.5 (Cao et al., 2022b).

Bloom and colleagues have performed yeast surface display and deep mutational scanning (DMS) (Fowler and Fields, 2014) on the entire 201 amino acid RBD of SARS-CoV-2 in order to determine the impact of single-position substitutions on binding to ACE2 and escape from monoclonal or serum antibodies (Greaney et al., 2021a, 2021b; Starr et al., 2020, 2021a, 2021b). While DMS has been very effective at single mutation profiling of the RBD, several previously circulating variants (e.g., Beta, Gamma, and Delta) possess multiple RBD mutations, and Omicron and its sublineages possess up to 21 RBD mutations (BA.1.12.1), thus underscoring the urgent need to determine the impact of combinatorial mutations. However, combinatorial sequence space grows exponentially as the number of mutations and amino acid diversity increases, rapidly outpacing the capabilities of experimental screening techniques. For instance, when focusing only on a subset of twenty RBD residues directly involved in ACE2 binding (Lan et al., 2020), theoretical sequence space (20^{20} = 1 x 10^{26}) far exceeds what can be screened by yeast display libraries (~10^9).

Here, we establish deep mutational learning (DML), which integrates experimental yeast display screening of RBD mutagenesis libraries with deep sequencing and machine learning (**Fig. 1**). We perform DML to comprehensively interrogate combinatorial RBD mutations and their impact on ACE2 binding and escape from a panel of neutralizing antibodies, including clinically used therapeutics and other broadly neutralizing and potent antibodies. DML reveals a highly diverse mutational landscape of RBD mutations that can maintain binding to ACE2 while escaping many different classes of neutralizing antibodies. Finally, DML is able to predict antibody robustness to prospective SARS-CoV-2 variants and thus may be a valuable tool in evaluating and selecting the most promising antibody therapeutics for clinical development.

RESULTS

Design and screening of RBD mutagenesis libraries

SARS-CoV-2 RBD mutagenesis libraries were targeted to core regions of the receptor-binding motif (RBM-3: positions 439-452; RBM1: 453-478; RBM-2: 484-505), which are subregions of the RBD that interface with ACE2 and where mutations are commonly observed in viral genome sequencing data [available on GISAID (www.gsaid.org)]. To generate training datasets covering a high mutational sequence space, combinatorial mutagenesis schemes were designed based on DMS data for ACE2 binding, previously published by Starr et al. (Starr et al., 2020). Single mutation fitness values were empirically thresholded and converted to amino acid frequencies, with mutations below the ACE2 binding fitness threshold excluded. For each position, degenerate codons approximating the desired amino acid distribution were selected by minimizing mean-squared error (Mason et al., 2018) (some positions remained fixed due to their inability to tolerate mutations and retain ACE2 binding), resulting in RBM libraries with theoretical amino acid diversities of 3.5 x 10^7 (library 3C), 1.2 x 10^{10} (library 1C) and 1.50 x 10^{10} (library 2C) (**Fig. 2A, Fig. S1A**). An extended version of library 2C was also designed, with fully degenerate codons (NNK) at positions 417 and 439, which are mutated in a number of circulating variants and associated with antibody escape (Thomson et al., 2021; Tsai et al., 2021), resulting in a theoretical amino acid diversity of 5.99 x 10^{12} (library 2CE). To generate training datasets covering a lower mutational sequence space, we constructed tiling mutagenesis libraries, whereby fully degenerate codons (NNK) were tiled across three of the positions in each RBM, resulting in a theoretical amino acid diversities of 3.94 x 10^5 (library 3T), 2.53 x 10^6 (library 1T) and 1.53 x 10^6 (library 2T) (**Fig. 2B**).

Synthetic oligonucleotides encoding the different libraries and spanning the region of interest were amplified by PCR to produce double-stranded DNA with homology to the full RBD sequence. Co-transformation of yeast (*S. cerevisiae* EBY100) with library-encoding DNA and linearized plasmid yielded more than 2 x 10^7 transformants for each library. RBD variants, displayed on the yeast surface as a C-terminal fusion to Aga2 (Boder and Wittrup, 1997), were isolated by fluorescence-activated cell sorting (FACS) based on binding to soluble human ACE2 receptor (Wu-Hu-1 RBD used as a guide for gating) **(Fig. 2C)**. RBD variants which showed a complete loss of binding to ACE2 were also isolated. Importantly, this did not include variants with only partially reduced binding since such an intermediate population could not be assigned as binding or non-binding with sufficient confidence necessary for training supervised machine learning models (**Fig. 2C, Table S1**). Targeted deep sequencing (Illumina) of the RBD gene was performed on all the sorted libraries; protein sequence logos revealed highly similar patterns of amino acid usage between the ACE2-binding and non-binding fractions (**Fig. 2D, Fig. S2 and Table S2,).**

Next, using exclusively the ACE2-binding populations, FACS was performed to isolate variants that maintained binding or showed a complete loss of binding (escape) to a panel of 13 neutralizing antibodies; these include clinically used therapeutic antibodies (REGN10987, REGN10933, LY-CoV16, LY-CoV555, S309 and LY-CoV1404), antibodies demonstrating exceptional sabrecovirus breadth (S2E12, S2H97 and A23-58.1) (Starr et al., 2021c; Wang et al., 2021a) and other potent neutralizers isolated directly from COVID-19 individuals (G32A4, mAb-50, mAb-64 and mAb-82) (Ehling et al., 2022; Tong et al., 2021) (**Table S3, Fig. S2B, C)**. Wu-Hu-1 RBD was again used as a guide for gating antibody binding or escape **(Fig. 2E)**. The proportion of binding and escape (non-binding) for each antibody and library was highly variable, for example with RBM-2 libraries REGN10933 showed a low fraction of escape variants and LY-CoV555 had a very high fraction (**Fig. 2E**, **Table S1**). Deep sequencing was once again performed on antibody binding and escape fractions of all the sorted RBD libraries, and similar to ACE2, protein sequence logos of the two fractions looked highly similar (**Fig. S2**).

Machine learning models accurately predict ACE2 binding and antibody escape

Deep sequencing data from ACE2 selections underwent pre-processing, quality filtering and balancing steps to create the final training sets **(STAR Methods, Table S4)**. Following nucleotide to protein translation, amino acid sequences were converted to an input matrix by one-hot encoding (**Fig. 3A**). Supervised machine learning models were trained for classification of ACE2 binding, which is defined as the probability (*P)* that any given RBD sequence binds to ACE2 (higher *P* correlates with binding). For initial benchmarking, a range of different baseline models (default parameters) were trained using data derived from RBM-2 libraries (2C, 2CE and 2T) and evaluated based on their classification performance across several metrics (accuracy, F1, precision, recall). Machine learning models tested included K-nearest neighbor, logistic regression, naive Bayes, support vector machines and random forests (RF); long-short term memory recurrent neural networks (RNN) were also trained, which are a class of deep learning models that have the ability to learn long-range dependencies in sequential data (Akbar et al., 2021; Hochreiter and Schmidhuber, 1997; Mason et al., 2021; Saka et al., 2021). All baseline models trained on RBM-2 libraries performed effectively (i.e., accuracy scores between 0.87 - 0.94), including some of the more simple models such as logistic regression and naive Bayes, highlighting the potential importance of training data generated with combinatorial mutagenesis libraries to evaluate to which degree mutations are additive. RF and RNN models were selected for further optimization and application since they showed relatively higher performance metrics and could be trained faster (**Fig. S3, Table S5**).

SARS-CoV-2 evolves across a range of mutational trajectories, including variants such as Omicron and its sublineages that have accumulated multiple combinatorial mutations in their RBD. Determining the performance of machine learning models across various mutational edit distances [Edit Distance (ED) from the reference Wu-Hu-1 RBD sequence] is therefore an important criterion. Initially, DMS data consisting of single point mutations (ED_1) were used to train baseline models, with binding/non-binding labels assigned based on minimum affinity required to retain ACE2 binding as previously defined in (Starr et al., 2020). The resulting models exhibited very low performance metrics for ACE2-binding prediction (i.e., accuracy scores of 0.50 and AUC of 0.56 - 0.65) (**Fig. S3)**. This is likely because single point mutations are not additive at higher distances, and therefore are unable to account for the non-linear effects of combinatorial mutations, leading to models that predict nearly all combinatorial escape variants to be ACE2 binders **(Fig. S3)**. Next, we examined model performance on RBM-2 test data that was divided into low mutational distances ($\leq ED_5$), which corresponds to variants such as Beta and Gamma, and high mutational distances ($\geq ED_6$), which corresponds to variants such as Omicron. We found that when models were trained using only the low distance library (2T), their performance on predicting ACE2-binding on high distance data was very poor (accuracy <0.65 and AUC <0.83). **(Fig. 3B)**. However, models trained using high distance only (RBM-2C/CE) or combined low distance and high distance libraries (RBM-2 Full) resulted in vastly improved performance across all distances, with accuracy scores of >0.94 and >0.92 for low and high distances, respectively, and AUC of >0.97 for both models (**Fig. 3B**).

Similar to the ACE2 selections, deep sequencing data from antibody selections were pre-processed, quality filtered, balanced and encoded as before. Supervised machine learning models (RF and RNN) were trained to classify antibody escape, which is defined as the probability that a given RBD sequence escapes a defined antibody (lower *P* correlates with escape). As before, we show that using both low and high distance libraries (RBM-2 Full) for training data resulted in models with better performance for predicting escape from a representative antibody (LY-CoV16) as opposed to training models with only low distance (RBM-2T) or only high distance libraries (RBM-2C/CE) (**Fig. 3C**). RBM-2 models for nearly all antibodies showed high performance metrics, with the only exception being LY-CoV555 exhibiting low F1 scores. Initial machine learning training and benchmarking revealed that balanced classification data (similar number of sequence variants in binding vs. non-binding/escape classes) was required for training accurate models (see Methods). Thus, the lower performance of LY-CoV555 models can be explained through its imbalanced classification data (nearly all RBD variants escape LY-CoV555) (**Fig. 3D, Fig. S4B)**. For RBM-1 libraries, most antibodies produced imbalanced classification data at low ED (very few escape variants at $\leq ED_5$) (Fig. S7); therefore RBM-1 models were trained using only high distance data ($\geq ED_6$) and resulted in high performance metrics for most antibodies, with the exception of LY-CoV16, mAb-64 and mAb-82, which had imbalanced classification data (few escape variants) across all distances **(Fig. 3E, Fig. S4A)**. Finally, for RBM-3 libraries, nearly all antibodies produced imbalanced classification data (mostly all binding or all escape variants) (**Fig. S4C and Table S5**); and thus, RBM-3 machine learning models were excluded for future analysis.

Predictive profiling on synthetic lineage variants

Having established that ACE2 binding and antibody escape machine learning models can make highly accurate predictions on test data, we next evaluated their classification performance on defined variants, followed by experimental validation and structural modeling. First, we identified synthetic RBD variants that had single mutations (ED_1 from Wu-Hu-1) predicted to retain binding, whereas the combination of two single mutations (ED_2) were predicted to escape a given antibody. Our machine learning models predict that REGN10933 retains binding to the

single mutation variants E484K or F490I or the single mutation variants G485K and Q493V (**Fig. S5C**), which also correlates with previously published data from DMS (Greaney et al., 2021c). However, the combination of E484K + F490I or G485K + Q493V results in variants predicted to escape REGN10933. We subsequently expressed each variant by yeast surface display and evaluated antibody binding and escape; indeed, we observed that single mutation variants retained binding whereas the combinatorial variants show substantially reduced binding to REGN10933, thus corresponding with the machine learning predictions (**Fig. S5D**).

Next, synthetic lineages were generated *in silico* to simulate plausible evolutionary paths, where RBD variants without predicted ACE2-binding intermediates at each mutational step were excluded (**Fig. 4A**). We focused our analysis on the RBM-2 region and the four neutralizing antibodies (LY-CoV16, LY-CoV555, REGN10933, REGN10987) that were used extensively as clinical therapeutics during the peak of the pandemic in 2021. The lineages were designed to include variants at ED_3, ED_5, and ED_7 from the original Wu-Hu-1 RBD sequence (nucleotide and amino acid). Additionally, the sequences were chosen to form lineages containing mutations observed in circulating variants (e.g., Alpha: N501Y, Beta/Gamma: E484K and N501Y, Kappa: E484Q and N501Y). ACE2 binding was predicted based on a consensus model, whereby a given RBD sequence is predicted to bind ACE2 when both RF and RNN models yield $P > 0.5$, else they were predicted to be non-binders. The 46 synthetic lineage variants were chosen to contain diversity in ACE2 binding prediction (36 predicted binders, 10 predicted non-binders) (**Fig. 5A**). Additionally, predictions for escape from each of the four therapeutic antibodies were made for the synthetic variants using a similar consensus model approach (RBD sequence escapes an antibody when both RF and RNN outputs are $P < 0.5$) (**Fig. S5B**). After having made all machine learning predictions, each synthetic RBD variant was individually expressed on the surface of yeast cells and assessed for ACE2 binding and antibody escape. The consensus model correctly predicted ACE2 binding for 91.67% (33/36) of the synthetic variants, with an accuracy of non-binding prediction of 100% (10/10), resulting in an overall prediction accuracy of 93.48% (43/46) (**Fig. 4B, Fig. S5A**). For the 33 correctly predicted ACE2-binding variants, the combined accuracy of antibody escape predictions across all four therapeutic antibodies was 93.94% (124/132) (LY-CoV16: 31/33, LY-CoV555: 30/33, REGN10933: 31/33, REGN10987: 32/33) (**Fig. 4C, Fig. S5B)**. Additionally, we identified three variants that were only ED_3 (nucleotide and amino acid) from the Wu-Hu-1 RBD and in which consensus models predicted ACE2 binding and escape from all four therapeutic antibodies. One of these variants possessed mutations in positions 493, 498 and 501, which are all mutated in the Omicron variant (Martin et al., 2021). Subsequent yeast display experiments confirmed these machine learning predictions of antibody escape to all four therapeutic antibodies, including escape from the often mutation-resistant REGN10987 (**Fig. S5E, F**). Structural modeling by AlphaFold2 (Jumper et al., 2021) was performed on eight synthetic RBD variants (all variants were accurately classified and experimentally validated for ACE2 binding or non-binding) (**Fig. 4D**). The structural predictions showed that several ACE2 non-binding variants did not differ substantially from the original Wu-Hu-1 RBD. In contrast, the ACE2-binding variants showed a wide diversity of possible structural conformations.

Predicting antibody escape to current and prospective variants

In addition to the selected synthetic lineages, we also performed machine learning to predict ACE2 binding and antibody escape on a panel of 12 naturally-occurring variants of SARS-CoV-2 (**Table S6**). Once again, we focused our analysis on RBM-2 and four extensively used clinical antibodies (LY-CoV16, LY-CoV555, REGN10933, REGN10987). We determined the accuracy of machine learning predictions on antibody escape by using the Stanford SARS-CoV-2 Susceptibility Database as a reference (Tzou et al., 2020). Applying the same RBM-2 consensus model approach (RF and RNN) and thresholds as before, the prediction accuracy for ACE2 binding was 100% (12/12) and

the prediction accuracy for escape across all four therapeutic antibodies was 85.42% (41/48) (antibody escape is defined here as a reported 30-fold reduced neutralization in the Stanford database). Strikingly, when applying a more stringent threshold for antibody escape prediction that requires both RF and RNN models to have high certainty in their prediction (both models $P < 0.25$ for escape and $P > 0.75$ for binding), 100% (30/30) of the machine learning predictions matched the results reported in the Stanford database.

Next, we used machine learning models on RBM-2 to predict antibody escape on prospective ACE2-binding lineages at low mutational distances (ED_1 and ED_2) from the Wu-Hu-1, Alpha, Beta, Kappa, Gamma, and B.1.1523 RBD sequences. (**Fig. 5, Fig. S6, Table S5**). Using a stringent threshold for antibody escape, we identified distinct patterns based on the starting variant. For example, REGN10933 and REGN10987 were largely resilient to escape from ED_1 lineages of Wu-Hu-1, Alpha, Kappa (**Fig. 5A-I and Fig. S6A-I**). While ED_1 lineages of Beta and Gamma almost entirely escape both LY-CoV555 and LY-CoV16. A large fraction of ED_2 lineages from all variants escaped REGN10933, LY-CoV555, and LY-CoV16, revealing an increasing likelihood of escape with an increasing number of mutations. Notably, a small fraction (0.17%) of Beta ED_2 lineages are predicted to escape all four therapeutic antibodies, whereby several of these variants possess mutations in positions 417, 484, 493 and 501, which are all mutated in Omicron variants (**Fig. 5F**). For further visualization, we constructed deep escape networks (**Fig. 5C, F, I, Fig. S6C, F, I**), depicting the vulnerability of the four therapeutic antibodies to low distance mutations (ED_1 and ED_2). Specifically, deep escape networks illustrate the increase in sequence space per mutation while also pointing out the presence of mutations that vastly increase escape from multiple antibodies. For example, there are variants at ED_1 from Wu-Hu-1 that are predicted to not escape any of the four antibodies, however, just one additional mutation (ED_2) can result in variants predicted to escape up to three antibodies. DML enables rapid *in silico* evaluation of new variants that appear on genomic databases (GISAID). For example, we performed a similar analysis on B.1.1.523 variant possessing RBD mutations E484K and S494P (Veer et al., 2021), which revealed complete escape from LY-CoV555 and ED_1 lineages, as well as substantial escape for other antibodies in ED_2 lineages, including three variants that escaped all four of the extensively used therapeutic antibodies (**Fig. S6G, H, I)**.

Determining antibody robustness to SARS-CoV-2 mutational lineages

Determining antibody robustness (maintenance of binding) to potential SARS-CoV-2 variants, including to high distance variants with many combinatorial RBD mutations such as Omicron, may be a critical parameter when selecting candidate antibodies for therapeutic development. To this end, we applied DML to determine if we could prospectively determine the robustness of several neutralizing antibodies (**Table S3**). Initially, we focused on synthetic lineage variants that correspond with Omicron **(Fig. 6A)**. We determined antibody escape against specific single and combinatorial mutations present in Omicron in RBM-2 (K417N, E484A, Q493R, N501Y). Machine learning revealed that some antibodies such as LY-CoV16 and LY-CoV555 are predicted to maintain binding to most single variants but lose binding to nearly all combinatorial variants (**Fig. 6B**), whereas in contrast other antibodies such as REGN10987 and LY-CoV1404 are predicted to bind nearly all single and combinatorial variants. Expanding on this approach, we explored the impact of all mutations in a given position or combination of positions, by calculating the average percentage of escape induced by mutations in that position. This allowed us to construct dynamic antibody escape profiles and identify lineages with mutation orders that may lead to increased escape. For example, lineages derived from an initial mutation in position 493 and even at low distances of ED_2 are predicted to have very high escape fractions for several antibodies (LY-CoV16, LY-CoV555, REGN10933, mAb-50, mAb-82) (**Fig. 6C**). Whereas,

lineages from some positions such as 501 require more mutations such as ED_3 and ED_4 before antibody escape is predicted. Notably, several antibodies such as REGN10987, mAb-50 and LY-CoV1404 showed highly resilient escape profiles across all positional lineages. Determining antibody robustness at high distances is critical given how most antibodies discovered against the original SARS-CoV-2 (Wu-Hu-1) are unable to neutralize the high distance Omicron variant. Therefore, we used machine learning models to predict antibody binding and escape to high distance combinatorial variants (ED_6 - ED_{10}) in RBM-2 (**Fig. 6D, E**), revealing varying levels of robustness for several antibodies. For example, as expected from low distance predictions, LY-CoV555 showed very little robustness, with nearly all high distance variants predicted to escape. In contrast, antibodies such as S2E12 and S2H97 are predicted to be very robust at high distances, which correlates with the fact they show broad neutralization across a diverse clade of sarbecoviruses (Starr et al., 2021c). Finally, when considering strict thresholds for binding and escape (both RF and RNN models $P < 0.25$ for escape and $P > 0.75$ for binding), LY-CoV1404 was determined to be the antibody with the highest robustness, as it is predicted to bind the broadest set of combinatorial RBD variants, thus correlating with the fact that it maintains neutralization to Omicron and all its sublineages (Cao et al., 2022b).

DISCUSSION

Eradication of SARS-CoV-2 appears improbable. Instead, an endemic future likely awaits (Antia and Halloran, 2021; Phillips, 2021). An endemic and continually evolving SARS-CoV-2 poses a perpetual risk for the emergence of new variants that escape from vaccine- or infection-induced antibodies. In this study, we develop DML, a machine learning-guided protein engineering method for determining the impact of combinatorial mutations in the SARS-CoV-2 RBD on ACE2 binding and antibody escape. In DML, machine learning models trained on thousands of labeled RBD variants obtained from library screening make highly accurate predictions across a sequence space of billions of RBD variants, several orders of magnitude larger than what is possible from experimental screening alone.

A combination of future library designs, more elaborate screening strategies based on different binding thresholds and improved machine learning models - perhaps incorporating structural knowledge, could improve predictions across longer lengths of the RBD. For example, a recent study performed a machine learning analysis to predict apparent affinities of ACE2 binding on single and multiple mutation RBD variants (Makowski et al., 2022). Another important consideration is that our DML library was based on the original Wu-Hu-1 RBD sequence, however nearly all of the circulating variants globally (as of August 2022) are Omicron or its sublineages. Bloom and colleagues demonstrated single-amino acid mutations (DMS) result in a shifting of mutational trajectories when on a different background of RBD variants (Starr et al., 2022). Given the extensive mutational changes present in Omicron variants (> 15 RBD mutations), future DML (as well as DMS) studies should use mutagenesis libraries based on an Omicron background RBD sequence, with the caveat that other future high distance variants may still yet emerge, requiring continued updating of library designs.

The evolution and emergence of SARS-CoV-2 variants has created a continuously shifting landscape of clinical authorization for antibody therapeutics: several clinically approved antibodies such as REGN10933, REGN10987, LY-COV555, LY-COV16 and S309 are no longer authorized for clinical use, in most cases due to loss of activity versus Omicron or its sublineages (Cameroni et al., 2022; Cao et al., 2022b). One exception is LY-CoV1404, which to date has maintained effective binding and neutralization to all SARS-CoV-2 variants, including Omicron BA.1 and its sublineages (BA.2, BA.4, BA.5) and is therefore still authorized for clinical use (Cao et al., 2022b). By providing

accurate predictions of antibody escape across a large mutational landscape, DML may enable researchers to select candidate antibody therapeutics and cocktails with the most robustness: broadest efficacy against the spectrum of possible variants, some of which may occur simultaneously and may be highly mutated such as Omicron. Assessing the robustness of candidate antibodies against future variants puts therapeutic development on a proactive rather than reactive footing, potentially avoiding situations where many clinically approved antibodies are only used for short periods of time. Furthermore, such an approach could be used to guide the development of antibodies and cocktails that maximize breadth and potency (Hastie et al., 2021; Starr et al., 2021c) to both current and prospective variants and therefore extending the lifespan of clinical use. For example, of the 13 antibodies that we profiled, all of which were discovered based on binding to the original Wu-Hu-1 RBD and included several clinically approved therapeutics, DML predicted that LY-COV1404 is the most robust to prospective RBD variants, which correlates with its broad neutralization and continued clinical use (Cao et al., 2022b). Therefore going forward, in addition to neutralization potency, antibody robustness to combinatorial mutations in the RBD will be a critical parameter to assess for the therapeutic development of antibodies for COVID-19.

Finally, evidence exists that the receptor-binding domains of other endemic coronaviruses may be undergoing adaptive evolution to escape from human antibody responses (Eguia et al., 2021; Kistler and Bedford, 2021). Consequently, the application of DML to predict SARS-CoV-2 escape from polyclonal antibodies present in serum of vaccinated or convalescent individuals, combined with phylogenetic models of viral evolution (Worobey et al., 2020), may enable the prospective identification of future variants with the highest likelihood of emergence and thus support vaccine development for COVID-19.

Limitations of the Study

To establish DML, we rationally designed our combinatorial mutagenesis libraries using previously published DMS data on the RBD (Starr et al., 2020) in order to improve the probability of isolating variants that maintain binding to ACE2, which is important for generating sufficient training data for machine learning. This led to us leaving some positions fixed since the single mutation DMS data suggested mutations in these positions lead to a complete loss of binding to ACE2 (**Fig. 2A**). While this approach was largely effective in covering the mutational sequence space of most SARS-CoV-2 variants, it did lead to some limitations, as some of the fixed positions in our library design (e.g., 486, 496) are mutated in Omicron or its sublineages, further highlighting the impact of epistasis or combinatorial mutations in SARS-CoV-2 evolution. Most notable is position 486, which is mutated in Omicron BA.4 and BA.5 variants (F486V) and is strongly associated with antibody escape, including to BA.1-specific antibodies (Cao et al., 2022b). Therefore, future mutagenesis library designs for DML will need to consider the impact of epistatic effects and should not only rely on single-mutation DMS data. Additionally, during library construction, we split the RBD into three distinct regions to build RBM-1, -2 and -3 libraries in order to constrain the size of our combinatorial libraries. This leaves us unable to explore epistatic effects of mutations across RBM sites (e.g., S477N in RBM-1 and Q498R in RBM-2 present in Omicron).

ACKNOWLEDGMENTS

We thank the ETH Zurich D-BSSE Single Cell Unit and the Genomics Facility Basel for support. This work was supported by the Botnar Research Centre for Child Health (FTC Covid-19, to STR) and the European Research Council (H2020-SC1-PHE-Coronavirus-2020, CoroNab, to STR).

AUTHOR CONTRIBUTIONS

Conceptualization: JMT, CRW, BG, RAE and STR. Experimental methodology: JMT, RAE, MO and LF. Computational methodology: JMT, CRW, BG, SWM, JH, AY. Supervision: JMT, CRW and STR. Writing - original draft: JMT, CRW, BG, RAE and STR. Writing - review & editing: all authors.

DECLARATION OF INTERESTS

ETH Zurich has filed for patent protection on the technology described herein, and JMT, CRW, BG, RAE and STR. are named as co-inventors. CRW is an employee of Alloy Therapeutics (Switzerland) AG. CRW and STR may hold shares of Alloy Therapeutics. STR is on the scientific advisory board of Alloy Therapeutics.

REFERENCES

Akbar, R., Robert, P.A., Weber, C.R., Widrich, M., Frank, R., Pavlović, M., Scheffer, L., Chernigovskaya, M., Snapkov, I., Slabodkin, A., et al. (2021). In silico proof of principle of machine learning-based antibody design at unconstrained scale. BioRxiv 2021.07.08.451480. https://doi.org/10.1101/2021.07.08.451480.

Antia, R., and Halloran, M.E. (2021). Transition to endemicity: Understanding COVID-19. Immunity 54, 2172–2176. https://doi.org/10.1016/j.immuni.2021.09.019.

Barnes, C.O., West, A.P., Huey-Tubman, K.E., Hoffmann, M.A.G., Sharaf, N.G., Hoffman, P.R., Koranda, N., Gristick, H.B., Gaebler, C., Muecksch, F., et al. (2020). Structures of human antibodies bound to SARS-CoV-2 spike reveal common epitopes and recurrent features of antibodies. Cell 0. https://doi.org/10.1016/j.cell.2020.06.025.

Baum, A., Fulton, B.O., Wloga, E., Copin, R., Pascal, K.E., Russo, V., Giordano, S., Lanza, K., Negron, N., Ni, M., et al. (2020). Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. Science https://doi.org/10.1126/science.abd0831.

Boder, E.T., and Wittrup, K.D. (1997). Yeast surface display for screening combinatorial polypeptide libraries. Nat. Biotechnol. 15, 553–557. https://doi.org/10.1038/nbt0697-553.

Boström, H. (2008). Calibrating Random Forests. In 2008 Seventh International Conference on Machine Learning and Applications, (San Diego, CA, USA: IEEE), pp. 121–126.

Cameroni, E., Bowen, J.E., Rosen, L.E., Saliba, C., Zepeda, S.K., Culap, K., Pinto, D., VanBlargan, L.A., De Marco, A., di Iulio, J., et al. (2022). Broadly neutralizing antibodies overcome SARS-CoV-2 Omicron antigenic shift. Nature 602, 664–670. https://doi.org/10.1038/s41586-021-04386-2.

Cao, Y., Wang, J., Jian, F., Xiao, T., Song, W., Yisimayi, A., Huang, W., Li, Q., Wang, P., An, R., et al. (2022a). Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. Nature 602, 657–663. https://doi.org/10.1038/s41586-021-04385-3.

Cao, Y., Yisimayi, A., Jian, F., Song, W., Xiao, T., Wang, L., Du, S., Wang, J., Li, Q., Chen, X., et al. (2022b). BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. Nature 1–3. https://doi.org/10.1038/s41586-022-04980-y.

Chao, G., Lau, W.L., Hackel, B.J., Sazinsky, S.L., Lippow, S.M., and Wittrup, K.D. (2006). Isolating and engineering human antibodies using yeast surface display. Nat. Protoc. 1, 755–768. https://doi.org/10.1038/nprot.2006.94.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal Complex Systems, 1695.

Dejnirattisai, W., Zhou, D., Ginn, H.M., Duyvesteyn, H.M.E., Supasa, P., Case, J.B., Zhao, Y., Walter, T.S., Mentzer, A.J., Liu, C., et al. (2021). The antigenic anatomy of SARS-CoV-2 receptor binding domain. Cell 184, 2183-2200.e22. https://doi.org/10.1016/j.cell.2021.02.032.

Dejnirattisai, W., Huo, J., Zhou, D., Zahradník, J., Supasa, P., Liu, C., Duyvesteyn, H.M.E., Ginn, H.M., Mentzer, A.J., Tuekprakhon, A., et al. (2022). SARS-CoV-2 Omicron-B.1.1.529 leads to widespread escape from neutralizing antibody responses. Cell 185, 467-484.e15. https://doi.org/10.1016/j.cell.2021.12.046.

Eguia, R.T., Crawford, K.H.D., Stevens-Ayers, T., Kelnhofer-Millevolte, L., Greninger, A.L., Englund, J.A., Boeckh, M.J., and Bloom, J.D. (2021). A human coronavirus evolves antigenically to escape antibody immunity. PLOS Pathog. 17, e1009453. https://doi.org/10.1371/journal.ppat.1009453.

Ehling, R.A., Weber, C.R., Mason, D.M., Friedensohn, S., Wagner, B., Bieberich, F., Kapetanovic, E., Vazquez-Lombardi, R., Di Roberto, R.B., Hong, K.-L., et al. (2022). SARS-CoV-2 reactive and neutralizing antibodies discovered by single-cell sequencing of plasma cells and mammalian display. Cell Rep. 38, 110242. https://doi.org/10.1016/j.celrep.2021.110242.

FDA, C. for D.E. and (2022a). FDA authorizes revisions to Evusheld dosing. FDA.

FDA, C. for D.E. and (2022b). FDA updates Sotrovimab emergency use authorization. FDA.

FDA, O. of the C. (2022c). Emergency Use Authorization. FDA.

Fontanet, A., Autran, B., Lina, B., Kieny, M.P., Karim, S.S.A., and Sridhar, D. (2021). SARS-CoV-2 variants and ending the COVID-19 pandemic. The Lancet 397, 952–954. https://doi.org/10.1016/S0140-6736(21)00370-6.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat. Methods 11, 801–807. https://doi.org/10.1038/nmeth.3027.
Garcia-Beltran, W.F., Lam, E.C., Denis, K.S., Nitido, A.D., Garcia, Z.H., Hauser, B.M., Feldman, J., Pavlovic, M.N., Gregory, D.J., Poznansky, M.C., et al. (2021). Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. Cell 184, 2372-2383.e9. https://doi.org/10.1016/j.cell.2021.03.013.
Greaney, A.J., Starr, T.N., Gilchuk, P., Zost, S.J., Binshtein, E., Loes, A.N., Hilton, S.K., Huddleston, J., Eguia, R., Crawford, K.H.D., et al. (2021a). Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. Cell Host Microbe 29, 44-57.e9. https://doi.org/10.1016/j.chom.2020.11.007.
Greaney, A.J., Loes, A.N., Crawford, K.H.D., Starr, T.N., Malone, K.D., Chu, H.Y., and Bloom, J.D. (2021b). Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. Cell Host Microbe 29, 463-476.e6. https://doi.org/10.1016/j.chom.2021.02.003.
Greaney, A.J., Starr, T.N., Barnes, C.O., Weisblum, Y., Schmidt, F., Caskey, M., Gaebler, C., Cho, A., Agudelo, M., Finkin, S., et al. (2021c). Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. Nat. Commun. 12, 4196. https://doi.org/10.1038/s41467-021-24435-8.
Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinforma. Oxf. Engl. 32, 2847–2849.
Gustavsen, J.A., Pai, S., Isserlin, R., Demchak, B., and Pico, A.R. (2019). RCy3: Network biology using Cytoscape from within R.
Han, P., Su, C., Zhang, Y., Bai, C., Zheng, A., Qiao, C., Wang, Q., Niu, S., Chen, Q., Zhang, Y., et al. (2021). Molecular insights into receptor binding of recent emerging SARS-CoV-2 variants. Nat. Commun. 12, 6103. https://doi.org/10.1038/s41467-021-26401-w.
Hansen, J., Baum, A., Pascal, K.E., Russo, V., Giordano, S., Wloga, E., Fulton, B.O., Yan, Y., Koon, K., Patel, K., et al. (2020). Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. Science 369, 1010–1014. https://doi.org/10.1126/science.abd0827.
Harvey, W.T., Carabelli, A.M., Jackson, B., Gupta, R.K., Thomson, E.C., Harrison, E.M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S.J., et al. (2021). SARS-CoV-2 variants, spike mutations and immune escape. Nat. Rev. Microbiol. 1–16. https://doi.org/10.1038/s41579-021-00573-0.
Hastie, K.M., Li, H., Bedinger, D., Schendel, S.L., Dennison, S.M., Li, K., Rayaprolu, V., Yu, X., Mann, C., Zandonatti, M., et al. (2021). Defining variant-resistant epitopes targeted by SARS-CoV-2 antibodies: A global consortium study. Science 374, 472–478. https://doi.org/10.1126/science.abh2315.
Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. Neural Comput 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
Hoffmann, M., Arora, P., Groß, R., Seidel, A., Hörnich, B.F., Hahn, A.S., Krüger, N., Graichen, L., Hofmann-Winkler, H., Kempf, A., et al. (2021). SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing antibodies. Cell 184, 2384-2393.e12. https://doi.org/10.1016/j.cell.2021.03.036.
Iketani, S., Liu, L., Guo, Y., Liu, L., Chan, J.F.-W., Huang, Y., Wang, M., Luo, Y., Yu, J., Chu, H., et al. (2022). Antibody evasion properties of SARS-CoV-2 Omicron sublineages. Nature 604, 553–556. https://doi.org/10.1038/s41586-022-04594-4.
Ju, B., Zhang, Q., Ge, J., Wang, R., Sun, J., Ge, X., Yu, J., Shan, S., Zhou, B., Song, S., et al. (2020). Human neutralizing antibodies elicited by SARS-CoV-2 infection. Nature 584, 115–119. https://doi.org/10.1038/s41586-020-2380-z.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2.
Kistler, K.E., and Bedford, T. (2021). Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses OC43 and 229e. ELife 10, e64509. https://doi.org/10.7554/eLife.64509.
Kolde, R. (2019). pheatmap: Pretty Heatmaps.

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., et al. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 581, 215–220. https://doi.org/10.1038/s41586-020-2180-5.

Liu, L., Iketani, S., Guo, Y., Chan, J.F.-W., Wang, M., Liu, L., Luo, Y., Chu, H., Huang, Y., Nair, M.S., et al. (2022). Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. Nature 602, 676–681. https://doi.org/10.1038/s41586-021-04388-0.

Makowski, E.K., Schardt, J.S., Smith, M.D., and Tessier, P.M. (2022). Mutational analysis of SARS-CoV-2 variants of concern reveals key tradeoffs between receptor affinity and antibody escape. PLOS Comput. Biol. 18, e1010160. https://doi.org/10.1371/journal.pcbi.1010160.

Martin, D.P., Lyrtas, S., Lucaci, A.G., Maier, W., Grüning, B., Shank, S.D., Weaver, S., Maclean, O.A., Orton, R.J., Lemey, P., et al. (2021). Selection analysis identifies significant mutational changes in Omicron that are likely to influence both antibody neutralization and Spike function.

Mason, D.M., Weber, C.R., Parola, C., Meng, S.M., Greiff, V., Kelton, W.J., and Reddy, S.T. (2018). High-throughput antibody engineering in mammalian cells by CRISPR/Cas9-mediated homology-directed mutagenesis. Nucleic Acids Res. https://doi.org/10.1093/nar/gky550.

Mason, D.M., Friedensohn, S., Weber, C.R., Jordi, C., Wagner, B., Meng, S.M., Ehling, R.A., Bonati, L., Dahinden, J., Gainza, P., et al. (2021). Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. Nat. Biomed. Eng. 1–13. https://doi.org/10.1038/s41551-021-00699-9.

McCormick, K.D., Jacobs, J.L., and Mellors, J.W. (2021). The emerging plasticity of SARS-CoV-2. Science 371, 1306–1308. https://doi.org/10.1126/science.abg4493.

ter Meulen, J., van den Brink, E.N., Poon, L.L.M., Marissen, W.E., Leung, C.S.W., Cox, F., Cheung, C.Y., Bakker, A.Q., Bogaards, J.A., van Deventer, E., et al. (2006). Human monoclonal antibody combination against SARS coronavirus: synergy and coverage of escape mutants. PLoS Med. 3, e237. https://doi.org/10.1371/journal.pmed.0030237.

Neuwirth, E. (2014). RColorBrewer: ColorBrewer Palettes.

Niculescu-Mizil, A., and Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning - ICML '05, (Bonn, Germany: ACM Press), pp. 625–632.

Nielsen, S.C.A., Yang, F., Jackson, K.J.L., Hoh, R.A., Röltgen, K., Jean, G.H., Stevens, B.A., Lee, J.-Y., Rustagi, A., Rogers, A.J., et al. (2020). Human B Cell Clonal Expansion and Convergent Antibody Responses to SARS-CoV-2. Cell Host Microbe 28, 516-525.e5. https://doi.org/10.1016/j.chom.2020.09.002.

NIH (2022). Anti-SARS-CoV-2 Monoclonal Antibodies.

Phillips, N. (2021). The coronavirus is here to stay — here's what that means. Nature 590, 382–384. https://doi.org/10.1038/d41586-021-00396-2.

Pinto, D., Park, Y.-J., Beltramello, M., Walls, A.C., Tortorici, M.A., Bianchi, S., Jaconi, S., Culap, K., Zatta, F., De Marco, A., et al. (2020). Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. Nature 1–10. https://doi.org/10.1038/s41586-020-2349-y.

Platt, J. (2000). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Adv Large Margin Classif 10.

R Core Team R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

Rossum, G.V., and Drake, F.L.Jr. (2011). The Python Language Reference Manual (Network Theory Ltd).

Saka, K., Kakuzaki, T., Metsugi, S., Kashiwagi, D., Yoshida, K., Wada, M., Tsunoda, H., and Teramoto, R. (2021). Antibody design using LSTM based deep generative model from phage display library for affinity maturation. Sci. Rep. 11, 5852. https://doi.org/10.1038/s41598-021-85274-7.

Schrödinger, L., and DeLano, W. (2020). PyMOL.

Sheward, D.J., Kim, C., Ehling, R.A., Pankow, A., Castro Dopico, X., Dyrdak, R., Martin, D.P., Reddy, S.T., Dillner, J., Karlsson Hedestam, G.B., et al. (2022). Neutralisation sensitivity of the SARS-CoV-2 omicron (B.1.1.529) variant: a cross-sectional study. Lancet Infect. Dis. 22, 813–820. https://doi.org/10.1016/S1473-3099(22)00129-3.

Shi, R., Shan, C., Duan, X., Chen, Z., Liu, P., Song, J., Song, T., Bi, X., Han, C., Wu, L., et al. (2020). A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. Nature 584, 120–124. https://doi.org/10.1038/s41586-020-2381-y.

Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H.D., Dingens, A.S., Navarro, M.J., Bowen, J.E., Tortorici, M.A., Walls, A.C., et al. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell 182, 1295-1310.e20. https://doi.org/10.1016/j.cell.2020.08.012.

Starr, T.N., Greaney, A.J., Dingens, A.S., and Bloom, J.D. (2021a). Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. Cell Rep. Med. 0. https://doi.org/10.1016/j.xcrm.2021.100255.

Starr, T.N., Greaney, A.J., Addetia, A., Hannon, W.W., Choudhary, M.C., Dingens, A.S., Li, J.Z., and Bloom, J.D. (2021b). Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. Science 371, 850–854. https://doi.org/10.1126/science.abf9302.

Starr, T.N., Czudnochowski, N., Liu, Z., Zatta, F., Park, Y.-J., Addetia, A., Pinto, D., Beltramello, M., Hernandez, P., Greaney, A.J., et al. (2021c). SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape. Nature 597, 97–102. https://doi.org/10.1038/s41586-021-03807-6.

Starr, T.N., Greaney, A.J., Hannon, W.W., Loes, A.N., Hauser, K., Dillen, J.R., Ferri, E., Farrell, A.G., Dadonaite, B., McCallum, M., et al. (2022). Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. Science 0, eabo7896. https://doi.org/10.1126/science.abo7896.

Supasa, P., Zhou, D., Dejnirattisai, W., Liu, C., Mentzer, A.J., Ginn, H.M., Zhao, Y., Duyvesteyn, H.M.E., Nutalai, R., Tuekprakhon, A., et al. (2021). Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera. Cell 184, 2201-2211.e7. https://doi.org/10.1016/j.cell.2021.02.033.

Thomson, E.C., Rosen, L.E., Shepherd, J.G., Spreafico, R., da Silva Filipe, A., Wojcechowskyj, J.A., Davis, C., Piccoli, L., Pascall, D.J., Dillen, J., et al. (2021). Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. Cell 184, 1171-1187.e20. https://doi.org/10.1016/j.cell.2021.01.037.

Tong, P., Gautam, A., Windsor, I.W., Travers, M., Chen, Y., Garcia, N., Whiteman, N.B., McKay, L.G.A., Storm, N., Malsick, L.E., et al. (2021). Memory B cell repertoire for recognition of evolving SARS-CoV-2 spike. Cell 184, 4969-4980.e15. https://doi.org/10.1016/j.cell.2021.07.025.

Tsai, K.-C., Lee, Y.-C., and Tseng, T.-S. (2021). Comprehensive Deep Mutational Scanning Reveals the Immune-Escaping Hotspots of SARS-CoV-2 Receptor-Binding Domain Targeting Neutralizing Antibodies. Front. Microbiol. 12, 1812. https://doi.org/10.3389/fmicb.2021.698365.

Tzou, P.L., Tao, K., Nouhin, J., Rhee, S.-Y., Hu, B.D., Pai, S., Parkin, N., and Shafer, R.W. (2020). Coronavirus Antiviral Research Database (CoV-RDB): An Online Database Designed to Facilitate Comparisons between Candidate Anti-Coronavirus Compounds. Viruses 12, 1006. https://doi.org/10.3390/v12091006.

Vazquez-Lombardi, R., Nevoltris, D., Luthra, A., Schofield, P., Zimmermann, C., and Christ, D. (2018). Transient expression of human antibodies in mammalian cells. Nat. Protoc. 13, 99–117.

Veer, B.M.J.W. van der, Dingemans, J., Alphen, L.B. van, Hoebe, C.J.P.A., and Savelkoul, P.H.M. (2021). A novel B.1.1.523 SARS-CoV-2 variant that combines many spike mutations linked to immune evasion with current variants of concern.

Wang, L., Zhou, T., Zhang, Y., Yang, E.S., Schramm, C.A., Shi, W., Pegu, A., Oloniniyi, O.K., Henry, A.R., Darko, S., et al. (2021a). Ultrapotent antibodies against diverse and highly transmissible SARS-CoV-2 variants. Science 373, eabh1766. https://doi.org/10.1126/science.abh1766.

Wang, P., Nair, M.S., Liu, L., Iketani, S., Luo, Y., Guo, Y., Wang, M., Yu, J., Zhang, B., Kwong, P.D., et al. (2021b). Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature 593, 130–135. https://doi.org/10.1038/s41586-021-03398-2.

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., Lu, G., Qiao, C., Hu, Y., Yuen, K.-Y., et al. (2020). Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. Cell 181, 894-904.e9. https://doi.org/10.1016/j.cell.2020.03.045.

Westendorf, K., Žentelis, S., Wang, L., Foster, D., Vaillancourt, P., Wiggin, M., Lovett, E., van der Lee, R., Hendle, J., Pustilnik, A., et al. (2022). LY-CoV1404 (bebtelovimab) potently neutralizes SARS-CoV-2 variants. Cell Rep. 39, 110812. https://doi.org/10.1016/j.celrep.2022.110812.

Wibmer, C.K., Ayres, F., Hermanus, T., Madzivhandila, M., Kgagudi, P., Oosthuysen, B., Lambson, B.E., de Oliveira, T., Vermeulen, M., van der Berg, K., et al. (2021). SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. Nat. Med. 27, 622–625. https://doi.org/10.1038/s41591-021-01285-x.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).

Wickham, H., François, R., Henry, L., and Müller, K. (2021). dplyr: A Grammar of Data Manipulation.

Wickham,H. (2019) stringr: Simple, Consistent Wrappers for Common String Operations.

Worobey, M., Pekar, J., Larsen, B.B., Nelson, M.I., Hill, V., Joy, J.B., Rambaut, A., Suchard, M.A., Wertheim, J.O., and Lemey, P. (2020). The emergence of SARS-CoV-2 in Europe and North America. Science 370, 564–570. https://doi.org/10.1126/science.abc8169.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., et al. (2020a). A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269. https://doi.org/10.1038/s41586-020-2008-3.

Wu, Y., Wang, F., Shen, C., Peng, W., Li, D., Zhao, C., Li, Z., Li, S., Bi, Y., Yang, Y., et al. (2020b). A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. Science https://doi.org/10.1126/science.abc2241.

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., and Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 367, 1444–1448. https://doi.org/10.1126/science.abb2762.

Yang, F., Nielsen, S.C.A., Hoh, R.A., Röltgen, K., Wirz, O.F., Haraguchi, E., Jean, G.H., Lee, J.-Y., Pham, T.D., Jackson, K.J.L., et al. (2021). Shared B cell memory to coronaviruses and other pathogens varies in human age groups and tissues. Science *372*, 738–741. https://doi.org/10.1126/science.abf6648.

Yi, C., Sun, X., Ye, J., Ding, L., Liu, M., Yang, Z., Lu, X., Zhang, Y., Ma, L., Gu, W., et al. (2020). Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ACE2 and neutralizing antibodies. Cell. Mol. Immunol. *17*, 621–630. https://doi.org/10.1038/s41423-020-0458-z.

Yuan, M., Wu, N.C., Zhu, X., Lee, C.-C.D., So, R.T.Y., Lv, H., Mok, C.K.P., and Wilson, I.A. (2020). A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. Science *368*, 630–633. https://doi.org/10.1126/science.abb7269.

Zhou, D., Dejnirattisai, W., Supasa, P., Liu, C., Mentzer, A.J., Ginn, H.M., Zhao, Y., Duyvesteyn, H.M.E., Tuekprakhon, A., Nutalai, R., et al. (2021). Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. Cell *184*, 2348-2361.e6. https://doi.org/10.1016/j.cell.2021.02.037.

Zost, S.J., Gilchuk, P., Chen, R.E., Case, J.B., Reidy, J.X., Trivette, A., Nargi, R.S., Sutton, R.E., Suryadevara, N., Chen, E.C., et al. (2020). Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the SARS-CoV-2 spike protein. Nat. Med. *26*, 1422–1427. https://doi.org/10.1038/s41591-020-0998-x.

MAIN FIGURES

Figure 1. Overview of deep mutational learning of the RBD for prediction of ACE2 binding and antibody escape. The RBD or the SARS-CoV-2 spike protein is expressed on the surface of yeast, mutagenesis libraries are designed on the receptor-binding motif of the RBD (RBM-3, RBM-1, RBM-2), which are the sites of interaction with ACE2 and neutralizing antibodies (e.g., therapeutic antibody drugs). RBD libraries are screened by FACS for binding to ACE2 and neutralizing antibodies, both binding and non-binding (escape) populations are isolated and subjected to deep sequencing. Machine learning models are trained to predict binding status to ACE2 or antibodies based on RBD sequence. Machine learning models are then used to predict ACE2 binding and antibody escape on current and prospective variants and lineages.

Figure 2. Design of RBD mutagenesis libraries and screening by yeast surface display and deep sequencing.
(A) Shown is the amino acid usage in the combinatorial libraries (Library 3C, 1C, 2C). Degenerate codons are derived from DMS data for ACE2 binding (Starr et al., 2020)
(B) Representative examples of degenerate codons tiled across RBM-2, which are pooled together to comprise library 2T.
(C) Flow cytometry dot plots depict yeast display screening of combinatorial (1C, 2C, 2CE, 3C) and tiling (1T, 2T, 3T) RBD libraries and control RBD (Wu-Hu-1); gating schemes correspond to selection of ACE2-binding and non-binding variants.
(D) Amino acid logo plots of the RBD are based on deep sequencing data from ACE2-binding and non-binding selections.
(E) Flow cytometry dot plots depict yeast display screening of pooled RBD libraries (2C and 2CE) after selection for ACE2 binding; gating schemes correspond to selection of variants for binding and escape (non-binding) to monoclonal antibodies (mAbs).
See also Figure S1 and S2 and Table S1, S2 and S3.

Figure 3. Training and testing of machine and deep learning models for prediction of ACE2 binding and antibody escape based on RBD sequence.
(A) Deep sequencing data from ACE2 and monoclonal antibody (mAb) selections is encoded by one-hot encoding and used to train supervised machine learning (e.g., Random Forest, RF) and deep learning models (e.g., recurrent neural network, RNN). Models perform classification by predicting a probability (*P*) of ACE2 binding or non-binding and mAb binding or escape (non-binding) based on the RBD sequence.
(B and C) Performance of RF and RNN models trained on 2T, 2C or Full ACE2 or LY-CoV16 binding data shown by accuracy, F1, and ROC curves. Low and high distance sequences are defined as those $\leq ED_5$ and $\geq ED_6$ from Wu-Hu-1 RBD, respectively.
(D and E) Accuracy, F1, and AUC of all 13 mAb models trained on RBM-2 and RBM-1 data, evaluated on both low and high distance test sequences.
See also Figure S3 and S4 and Table S4, S5 and S6.

Figure 4. Prediction and experimental validation of synthetic lineages of RBD variants.
(A) Workflow to select and test synthetic variants at chosen edit distances (ED_3, ED_5, and ED_7) from Wu-Hu-1 RBD.
(B) Lineage plot of synthetic variants depicts machine learning predictions and experimental validation (Fig. S5) for ACE2 binding and non-binding
(C) Dot plots of synthetic variants correspond to machine learning model (RF and RNN) predictions and experimental validation for antibody binding or escape.
(D) Structural modeling by AlphaFold2 shows predicted structures of RBD variants that are ACE2 binding (green boxes) or non-binding (red boxes); control is Wu-Hu-1 RBD (black box).
See also Figure S5.

Figure 5. Predictive profiling of selected RBD variants for antibody escape across low mutational distances.
(A, D and G) Heatmap depicts monoclonal antibody (mAb) binding as assessed by RF and RNN models of ED_1 and ED_2 variants of Alpha, Beta and Kappa.

(B, E and H) The number of sequences escaping a combination of *n* (number) mAbs for ED_1 and ED_2 (agreement between models, threshold >0.5).
(C, F and I) Deep escape networks display possible evolutionary paths between variants and their escape from mAbs. See also Figure S6.

Figure 6. Determining antibody robustness to synthetic RBD variants and mutational lineages.
(A) Omicron (BA.1) mutations covered by combinatorial library RBM-2.
(B) Binding prediction for single and combinatorial mutations observed in Omicron
(C) Dynamic escape profile along Omicron lineage with percentage escape sequences across all mutations at distance 1–4 from Wu-Hu-1.
(D) Antibody prediction of ACE2 binding RBDs for each antibody at edit distance 6-10 from Wu-Hu-1 (10’000 sequences simulated in triplicate, only confident predictions shown (i.e. *P*(ACE2 binding) > 0.5 and either *P*(antibody binding) > 0.75 or *P*(antibody escape) < 0.25 for both RNN and RF)
(E) Total count of confident predictions across all distances (mean across triplicates).

STAR METHODS

KEY RESOURCES TABLE

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Flow Cytometry Staining Reagents		
PE anti-DYKDDDDK Tag Antibody	Biolegend	637309
anti-human IgG-AlexaFluor647	Jackson Immunoresearch	109-605-098
biotinylated human ACE2	Acro	AC2-H82E6
streptavidin-AlexaFluor 647	Biolegend	405237
Deposited Data		
Raw sequencing data	Starr et al. 2020	NCBI SRA: BioProject PRJNA639956
Raw and processed sequencing data	This study	https://github.com/LSSI-ETH
Oligonucleotides		
Degenerate Ultramers and oPools for RBD library construction	IDT	https://github.com/LSSI-ETH
Recombinant DNA		
pYD1-RBD(wt)	This study	https://github.com/LSSI-ETH
Cell Lines		
EBY100	ATCC	MYA-4941
Software and Algorithms		
bbduk	Joint Genome Institute	https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbduk-guide/
custom scripts for curation, analysis, and visualization	This study	https://github.com/LSSI-ETH

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for reagents and resources should be directed to and will be fulfilled by the Lead Contact, Sai T. Reddy (sai.reddy@ethz.ch)

Materials Availability

SARS-CoV-2 mutagenesis libraries generated in this study will be made available on request to the Lead Contact with a completed Materials Transfer Agreement.

Data and Code Availability

- The main data supporting the results in this study are available within the paper and its Supplementary Information.
- Raw and processed data and code (scripts) used for data curation, analysis and visualization are available at: https://github.com/LSSI-ETH
- Additional data files and code that supports the findings of this study is available from the corresponding authors upon reasonable request.
- Additional Supplemental Items are available from Mendeley Data at **https://data.mendeley.com/datasets/pkg3jk26y6/3**

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Saccharomyces cerevisiae EBY100 harboring the pYD1 plasmid were cultured for 1-2 days at in a 250 rpm shaking incubator at 30C in SD-UT medium (20 g/l glucose, 6.7 g/l yeast nitrogen base without amino acids, 5.4 g/l Na2HPO4, 8.6 g/l NaH2PO4·H2O and 5 g/l casamino acids), or at 23C for 2 days in SG-UT induction medium (SD-UT with 20 g/l galactose instead of glucose).

METHOD DETAILS

Rational design of SARS-CoV-2 RBD mutagenesis libraries

Combinatorial library design.

The design of the combinatorial library 2C consisted of mutating residues within the RBM-2 region (positions 484-505 of the RBD) and was based on previously described results from deep mutational scanning (DMS) experiments (Starr et al., 2020). DMS enrichment ratios described by Starr et al. were thresholded to exclude mutations with decreased ACE2-binding fitness and then converted to amino acid frequencies as described previously (Mason et al., 2018). For each position, degenerate codons approximating the amino acid frequency distribution and diversity were selected, resulting in a library with a theoretical diversity of 1.50×10^{10} amino acid sequences. Library 2CE consisted of the same combinatorial design in positions 484-505 but with additional fully degenerate codons (NNK) in positions 417 and 439, resulting in a theoretical amino acid diversity of 5.95×10^{12}.

Tiling library design.

The tiling library 2T was designed by incorporating three positions with full degenerate codons (NNK) within the RBM-2 (positions 484-505) of the RBD. The degenerate codons are tiled across such that the total sequences of a tiling library, i.e., the number of variants of up to a maximum edit distance (ED) *k* away from the wild-type sequence is determined by the length of sequence (*n*, here = 14 non-fixed positions in RBM-2), the number of NNKs (or max ED) introduced (*k*, here = 3) and the size of the Alphabet (*a*, here = 20):

$$\sum_{i=0}^{k} (a-1)^i \times \binom{n}{i}$$

Similarly, the number of sequences for a given ED *k* is given by:

$$(a-1)^k \times \binom{n}{k}$$

The resulting total diversity of the library 2T is 1,533,035 sequences.

Cloning and expression of RBD mutagenesis libraries for yeast surface display

For libraries 2C and 2CE, synthetic single-stranded oligonucleotides (ssODNs) (Integrated DNA Technologies ultramers or oPools) were designed with degenerate codons spanning the region of interest and encoding the desired library diversity, with 30 bp overhangs on each end that were homologous to the yeast display plasmid pYD1. For library 2T, pools of ssODN were designed, where each member of the pool contains one combination of the three 'NNK' codons; in this case, consisting of 120 unique ssODNs. The ssODNs were amplified by PCR to produce double-stranded DNA. The plasmid pYD1 was modified such that the entire C-terminal fusion to Aga2 was replaced with a cassette encoding the RBD (Wu-Hu-1 sequence), expression tags and stop codon (HA Tag-RBD-FLAG-Stop). The RBM-2 residues 484-505 were replaced with an EcoRI recognition site, allowing production of a linearized vector with homology to mutagenesis ssODNs and with no parental background. Insert and EcoRI-linearized plasmids were concentrated and purified by silica spin columns (Zymo D4013) followed by drop dialysis for 1 hour in nuclease-free H_2O (Millipore VSWP02500). The libraries were cloned and expressed in yeast by *in vivo* homologous recombination, as previously described (Boder and Wittrup, 1997; Chao et al., 2006), using 1 µg each of plasmid and insert DNA per 300 µl of electrocompetent EBY100 cells in a 2 mm electroporation cuvette.

Screening RBD libraries for ACE2-binding and non-binding

Surface expression of SARS-Cov2 RBD was induced by growth in SG-UT medium at 23°C for 16-40 hours, as previously described (Boder and Wittrup, 1997). Approximately 10^8 library cells were washed once with 1 mL wash buffer (Dulbecco's PBS+ 0.5% BSA + 0.1% Tween20 + 2 mM EDTA) by centrifugation at 8000 x g for 30 s. Washed cells were stained with 50 nM biotinylated human ACE2 (Acro AC2-H82E6) for 30 minutes at 4 °C, followed by an additional wash. Cells were then stained with 2.5 ng/µl streptavidin-AlexaFluor 647 (Biolegend 405237) and 1 ng/µl PE anti-DYKDDDDK Tag Antibody (Biolegend 637310) for 30 minutes at 4 °C. Cells were subsequently pelleted by centrifugation at 8000 x g for 30s and kept on ice until sorting. Binding (ACE2+/FLAG+) and non-binding (ACE2-/FLAG+) cells were sorted by FACS (BD FACSAria Fusion or Sony MA800 cytometer) (**Fig. 2**). Collected cells were cultured in SD-UT medium for one to two days at 30 °C. Induction and sorting was repeated until the desired populations were pure.

Screening RBD libraries for antibody binding and escape

RBD libraries pre-sorted for ACE2-binding were cultured and induced, as described above. Induced cells were washed once with DPBS wash buffer, followed by incubation with 100 nM monoclonal antibody, or antibody mixtures. In the case of antibody mixtures, 100 nM of each antibody was used. Following an additional wash, cells were resuspended in 5 ng/µl anti-human IgG-AlexaFluor647 (Jackson Immunoresearch 109-605-098) and incubated for 30 minutes at 4°C. Cells were washed once more and resuspended in 1 ng/µl anti-FLAG-PE before 30 minutes of incubation at 4°C. Cells were then pelleted by centrifugation at 8000 x g for 30s and kept on ice until sorting. Cells expressing RBD that maintained antibody-binding (IgG+/FLAG+) or showed a complete loss of antibody binding (escape) (IgG-/FLAG+) were sorted by FACS (BD Aria Fusion or Sony MA800 instrument). Collected cells were cultured in SD-UT medium for 16-40 hours at 30 °C. Induction and sorting was repeated for multiple rounds until the desired populations of RBD variants showed purity for binding and escape (non-binding) to antibodies.

Antibody production and purification

Heavy chain and light chain inserts were cloned into pTwist transient expression vectors by Gibson Assembly. 30 mL cultures of Expi293 cells (Thermo, A14635) were transfected according to the manufacturer's instructions. After 5-7 days, dense Expi293 cultures were centrifuged at 300 x g for 5 minutes to pellet the cells. Supernatant was filtered using Steriflip® 0.22 µm (Merck, SCGP00525) filter units. Using protein G purification, Expi supernatant was directly loaded onto Protein G Agarose (Pierce, Cat# 20399) gravity columns, washed twice with PBS and eluted using Protein G Elution Buffer (Pierce, Cat# 21004). The eluted fractions were immediately neutralized with 1M TRIS-Buffer (pH 8) to physiological pH and quantified by Nanodrop™ 2000c for A280 nm absorption. Protein containing fractions were pooled and buffer exchanged using SnakeSkin™ dialysis tubing (10 MWCO, Pierce Cat#68100) followed by further dialysis and concentration using Amicon Ultra-4 10kDa centrifugal units (Merck, Cat# UFC801096), as described previously (Vazquez-Lombardi et al., 2018).

Deep sequencing of RBD libraries

Plasmid DNA encoding the RBD variants was isolated following the manufacturer's instructions (Zymo D2004). Mutagenized regions of the RBD were amplified using custom oligonucleotides. Illumina Nextera barcode sequences were added in a second PCR amplification step, allowing for multiplexed high-throughput sequencing runs. Populations were pooled at the desired ratios and sequenced using Illumina 2 x 250 PE or 2 x 150 PE protocols (MiSeq or NovaSeq instruments).

Experimental validation of selected RBD variants for ACE2-binding and antibody escape

Individual sequences for RBD variants were ordered as complementary forward and reverse primers (Integrated DNA Technologies) in 96-well plates A single round of annealing and extension was used to produce double-stranded DNA with 14-bp of homology at 5' and 3' ends to the pYD1-RBD entry vector, followed by Gibson Assembly with EcoRI digested vector. Plasmids were transformed into EBY100 prepared with the Frozen-EZ Yeast Transformation Kit II (Zymo) and plated on SD-UT agar. Individual colonies were picked and grown in SD-UT liquid medium overnight at 30°C, then diluted to OD_{600} = 0.5 in SG-UT medium and grown for 40-48 hours at 23°C. Cells were stained with biotinylated ACE2 or purified antibody as described above. Flow cytometry analysis was performed on the BD Fortessa cytometer.

QUANTIFICATION AND STATISTICAL ANALYSIS

Processing of deep sequencing data, statistical analysis and plots

Data preprocessing

Sequencing reads were paired, quality trimmed and assembled using Geneious and BBDuk, with a quality threshold of qphred ≥ 25. Mutagenized regions of interest were then extracted using custom Python scripts, followed by translation to amino acid sequences. The sequences obtained from each of the three libraries (2C, 2CE and 2T) were pre-processed separately before being combined into the final training set used for model training and evaluation. To remove sequencing errors, all libraries were filtered for sequences complying with the initial degenerate codon mutagenesis scheme. Library 2CE was filtered for only those sequences retaining unmutated residues in positions 417/439, to focus on the 484-505 region. Next, library 2T was filtered using a threshold of read counts > 4 and restricted to sequences that were ≤ ED_3 from Wu-Hu-1 RBD sequence.

Duplicate sequences in the full dataset were removed and a balanced dataset was created from the remaining data such that equal numbers of positive (binding) and negative class (non-binding) sequences were present for each ED. We observed significant bias in model performance when predictions are separated by ED from the Wu-Hu-1 RBD sequence, which was likely due to class (binding vs. non-binding/escape) imbalance in the training data**.** Class balancing was thus performed through random subsampling from the majority class at each ED equal to the counts from the minority class. Those that were not sampled from the majority class were then reserved separately as additional "unseen sequences". These were then used for model evaluation to ensure that the models could generalize well even to the sequences removed during balancing.

Statistical analysis and plots

Statistical analysis was performed using R 4.0.1 (R Core Team) and Python 3.8.5 (Rossum and Drake, 2011) . Graphics were generated using the ggplot2 3.3.3 (Wickham, 2009), ComplexHeatmap 2.4.3 (Gu et al., 2016), pheatmap 1.0.12 (Kolde, 2019), igraph 1.2.6 (Csardi and Nepusz, 2006), RCy3 2.8.1 (Gustavsen et al., 2019), stringr 1.4.0 (Wickham, 2019), dplyr 1.0.6 (Wickham et al., 2021), and RColorBrewer 1.1-2 (Neuwirth, 2014) R package.

Escape Networks

Network plots were generated using the igraph package (Csardi and Nepusz, 2006) and Cytoscape software 3.8.2 (Shannon et al., 2003) with edges drawn between every pair of two amino acid sequences from ED 1 and 2, when the pair of sequences share a common mutation on amino acid level. Edges were colored according to the change in number of antibodies that escape. Nodes representing RBD variant sequences were clustered and colored according to the number of antibodies that escape, and the mutational distance from the reference sequence.

Machine learning model training and evaluation

All machine learning (ML) classifier models were built in Python (3.8.5) (Rossum and Drake, 2011). Data was prepared and visualized using numpy (1.19.2), matplotlib (3.3.4), and pandas (1.2.4). Random Forest (RF) and other benchmarking ML models were built using Scikit-Learn (0.24.2), a 80/20 train-test data split (random split) to train baseline models, and a 90/10 train-test data split (random split) for final RF and RNN models. Keras libraries (2.4.3) from Tensorflow (v2.5) were used to build the long-short-term-memory recurrent neural networks (RNN) models.

RBD sequences were one-hot encoded prior to being used as inputs into the models. For the RNN, the 2D one-hot encoded matrix was used as the input, while for all other models, the matrix was flattened into a single dimensional vector beforehand. After selecting the best models, hyperparameter optimization was performed to further improve the performance of the chosen RF and RNN models using 50 rounds of Random Search with 5-fold cross-validation while scoring based on precision **(Table S5)**. All RF models were further calibrated using both the "isotonic" and Platt scaling (Platt, 2000; Boström, 2008; Niculescu-Mizil and Caruana, 2005), and the best model from the three was selected by calculating the overall mean-square error (MSE) from the true labels, with the RF model with the lowest MSE selected as the final model. Models were evaluated on the full test set which is unbalanced and includes the sequences removed during balancing to create the training set. Models were evaluated on the basis of Accuracy, F1, MCC, and AUC-ROC curve using the full test set, over 5 rounds of external cross-validation using different train-test splits. For further detailed evaluation, the test data was separated into two distances: low and high distance sequence sets, which consisted only of sequences $\leq ED_5$ or $\geq ED_6$ from Wu-Hu-1 RBD sequence, respectively. These two sets were then used to evaluate the accuracy, F1, MCC, and AUC-ROC of models to investigate any performance bias at different distances.

In silico sequence generation and evaluation

Synthetic RBD variant sequences were generated *in silico* using custom Python scripts for selected edit distances (ED) from the Wu-Hu-1 RBD sequence. The ED was defined on both the nucleotide and amino acid level, such that each generated nucleotide sequence was categorized by an ED pair (distance_nt, distance_aa). The synthetic variants (*in silico* generated sequences) were evaluated for their probability of ACE2-binding and non-binding using a consensus model (RF and RNN) approach. For a given RBD sequence, ACE2-binding prediction was defined as the case where both models output $P > 0.5$, else the sequence was considered as ACE2 non-binding. Similarly, the sequences were evaluated for binding and escape (non-binding) from monoclonal antibodies. Here the sequences were categorized into one of four categories: escape (both models $P < 0.25$), antibody binding (both models $P > 0.75$), unsure (at least one model gives P between 0.25 and 0.75), and disagree (one model outputs $P < 0.25$ while the other model outputs a $P > 0.75$).

Structural Prediction of RBD variants by AlphaFold2

Structural predictions were generated with the Alphafold v2.1.0 public iPython notebook using residues 331-530 of the spike protein. (https://colab.research.google.com/github/deepmind/Alphafold/blob/main/notebooks/AlphaFold.ipynb) (Jumper et al., 2021). Results were visualized and aligned in PyMol v2.2.3 (Schrödinger and DeLano, 2020).

Deep Mutational Learning Predicts ACE2 Binding and Antibody Escape to Combinatorial Mutations in the SARS-CoV-2 Receptor Binding Domain

Joseph M. Taft[1, 2]*, Cédric R. Weber[3]*, Beichen Gao[1, 2], Roy A. Ehling[1], Jiami Han[1, 2], Lester Frei[1, 2], Sean W. Metcalfe[1], Max Overath[1], Alexander Yermanos[1, 2, 4, 5], William Kelton[6] and Sai T. Reddy[1,2#].

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel 4058, Switzerland
[2]Botnar Research Centre for Child Health, Basel 4058, Switzerland
[3]Alloy Therapeutics (Switzerland) AG, Basel 4058, Switzerland
[4]Department of Biology, Institute of Microbiology and Immunology, ETH Zurich, Zurich 8093, Switzerland
[5]Department of Pathology and Immunology, University of Geneva, Geneva 1211, Switzerland
[6]Te Huataki Waiora School of Health, University of Waikato, Hamilton 3240, New Zealand

*Equal contribution
#Lead Contact: sai.reddy@ethz.ch

SUPPLEMENTARY FIGURES

Figure S1. Design and Screening of RBD Libraries.
(A) Amino acid distribution of combinatorial libraries RBM-1 and RBM-3.
(B) Yeast display of RBD libraries pre-selected for ACE2 binding were sorted by flow cytometry for binding and escape to four therapeutic monoclonal antibodies (mAbs): LY-CoV16, LY-CoV555, REGN10933, and REGN10987.
(C) A further nine monoclonal antibodies were screened for binding and escape. Approximately 10^7 yeast cells were screened for each antibody.
See also Figure 2 and Table S1.

Figure S2. Combinatorial sequence space of RBD libraries following selection.
Sequence logo plots of sorted populations for ACE2 binding and antibody escape. For each population, up to the 10,000 most abundant unique amino acid sequences after read count thresholding are shown.
See also Figure 2 and Table S2.

Figure S3. Performance metrics machine learning models.
(A) K-nearest Neighbours (KNN), Logistic Regression (Log Reg), Naive Bayes (NB), Random Forest (RF), Long-short term memory recurrent neural network (RNN), Support vector machine with linear kernel (SVM Linear), and Support vector machine with radial basis function kernel (SVM RBF) models were trained on the ACE2 deep sequencing data without hyperparameter optimization. Models were then challenged to perform classification by predicting a probability (*P*) of ACE2 binding on test data. Performance of models was evaluated by Accuracy, F1, Precision, and Recall. All models except RNN were trained using Sci-kit Learn, and the RNN was trained using Keras.
(B) K-nearest Neighbours (KNN), Logistic Regression (Log Reg), Naive Bayes (NB), Random Forest (RF), Long-short term memory recurrent neural network (RNN), Support vector machine with linear kernel (SVM Linear), and Support vector machine with radial basis function kernel (SVM RBF) models were trained on the ACE2 deep sequencing data without hyperparameter optimization. Models were then challenged to perform classification by predicting a probability (*P*) of ACE2 binding on test data. Performance of models was evaluated by Accuracy, F1, Precision, and Recall. All models except RNN were trained using Sci-kit Learn, and the RNN was trained using Keras.
(C and D) DMS trained models were evaluated on the larger combinatorial ACE2 binding test data shown by accuracy, F1 graphs, and ROC curves.
See also Figure 3 and Table S4, S5 and S6.

Figure S4. Distribution of binding and non-binding across RBM regions.

Count distributions of unique binding/non-binding sequences from the ACE2 and antibody selection library datasets after pre-processing.
(A) RBM-1,
(B) RBM-2,
(C) RBM-3.
See also Figure 3.

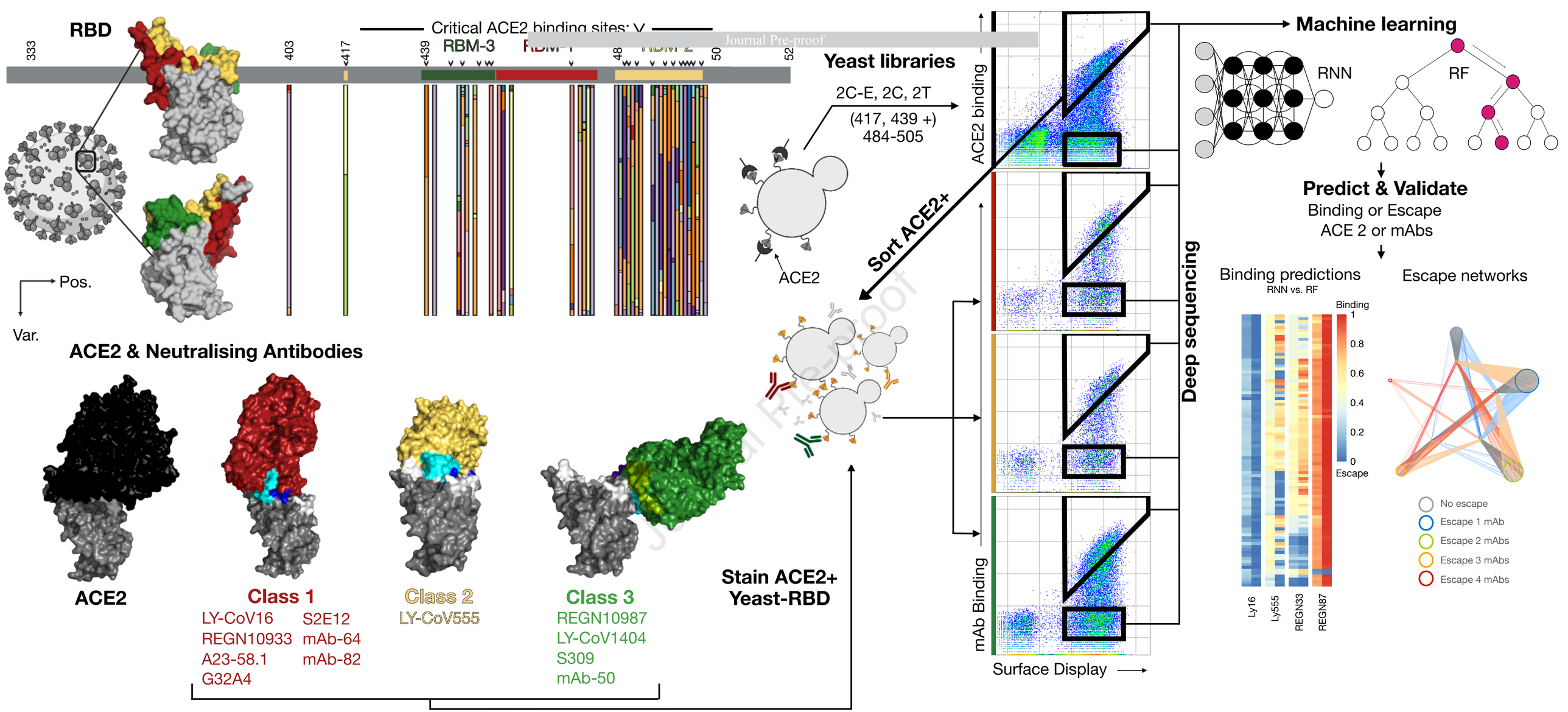
Figure S5. Experimental evaluation of selected RBD variants for antibody escape.
(A) The 46 selected synthetic variants were individually cloned and expressed for yeast display and ACE2 binding by flow cytometry. 43 variants showed ACE2 binding or non-binding that matched machine learning predictions. The ACE2-binding status for two variants (38 and 42) could not be conclusively determined, while one variant (41) showed was incorrectly predicted by machine learning for ACE2 binding.
(B) RBD sequences at chosen EDs (ED_0, ED_3, ED_5, ED_7) from the Wu-Hu-1 RBD were predicted for ACE2 binding and escape from four therapeutic monoclonal antibodies (mAbs). Accuracy for antibody escape predictions are the following: LY-CoV16 = 31/33 (93.94%), LY-CoV555 = 30/33 (90.91%), REGN10933 = 31/33 (93.94%), REGN10987 = 32/33 (96.97%).
(C and D) Two double mutants, and their constituent mutations, which were predicted to display epistasis were assayed individually by yeast surface display
(E) Three synthetic RBD variants of ED_3 from Wu-Hu-1 RBD that were predicted to escape all four therapeutic antibodies by the consensus machine learning model were expressed as individual clones in yeast and evaluated by flow cytometry for binding to antibody or ACE2.
See also Figure 4.

Figure S6. Predictive profiling of additional selected RBD variants for antibody escape across low mutational distances. (A, D and G) Heatmap depicts monoclonal antibody (mAb) binding as assessed by RF and RNN models of ED_1 and ED_2 variants of Wu-Hu-1, Gamma, and B.1.523.
(B, E and H) The number of sequences escaping a combination of *n* (number) mAbs for ED_1 and ED_2 (agreement between models, threshold >0.5).
(C, F and I) Deep escape networks display possible evolutionary paths between variants and their escape from mAbs.
See also Figure 5.

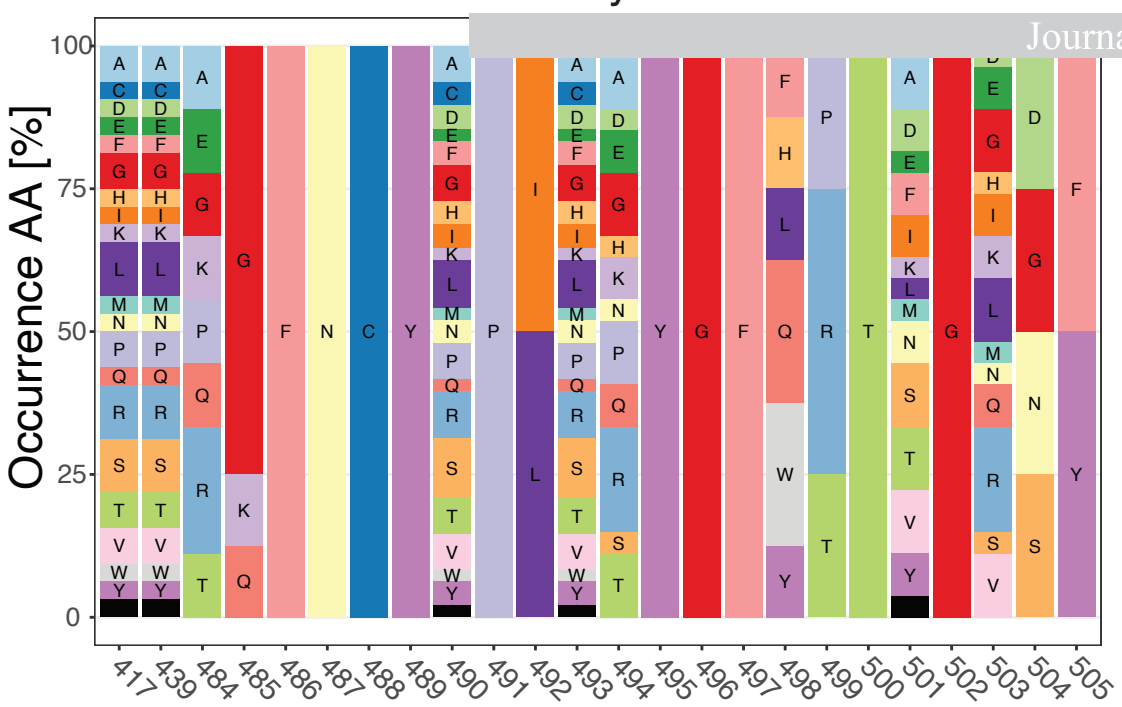
SUPPLEMENTARY TABLES

Table S4 (see Excel File). Detailed sequences used as the training data for individual models. Each dataset combines sequences from all RBM libraries after preprocessing, filtering, and removing duplicates.
See Fig. 3.

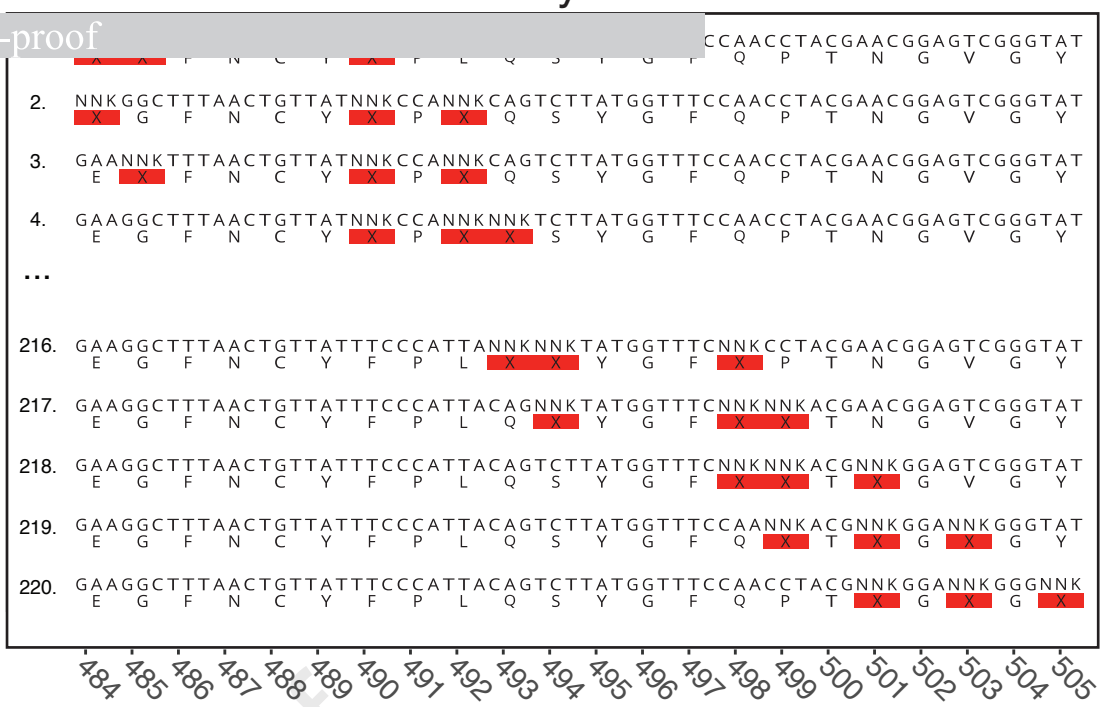
Table S6 (see Excel File). Machine and deep learning model predictions compared to susceptibility data from the Stanford Database (https://covdb.stanford.edu/page/susceptibility-data/, 2021-10-19). RF and RNN model predictions are compared to previously published susceptibility data. The sequences had previously been excluded from the training datasets.
See Fig. 3.



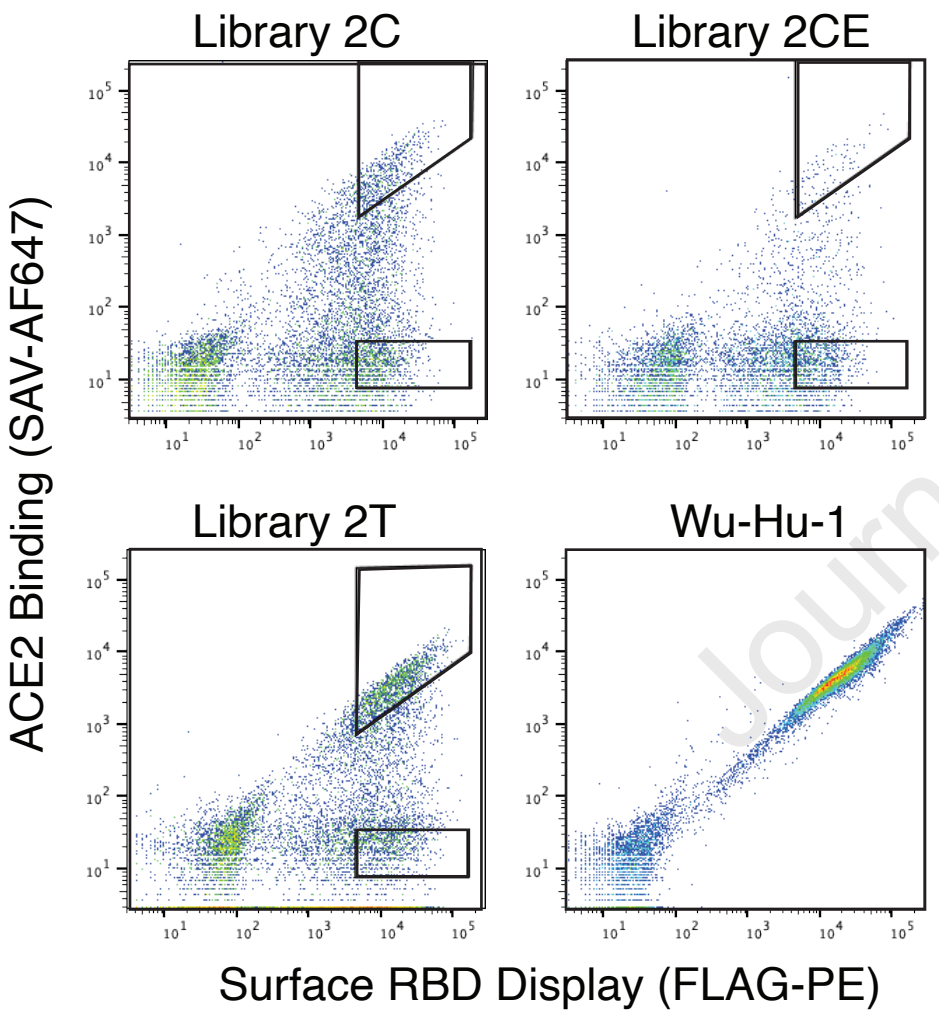
A



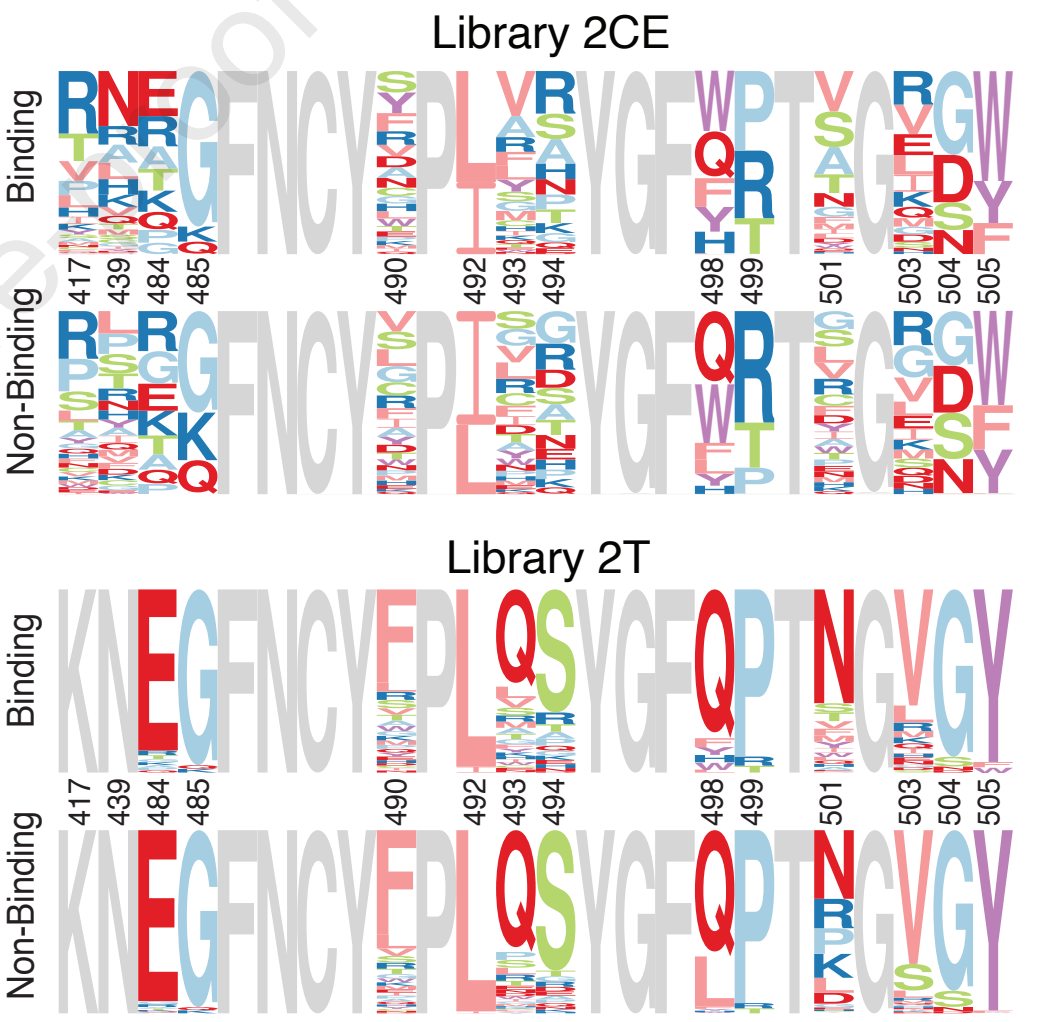
B



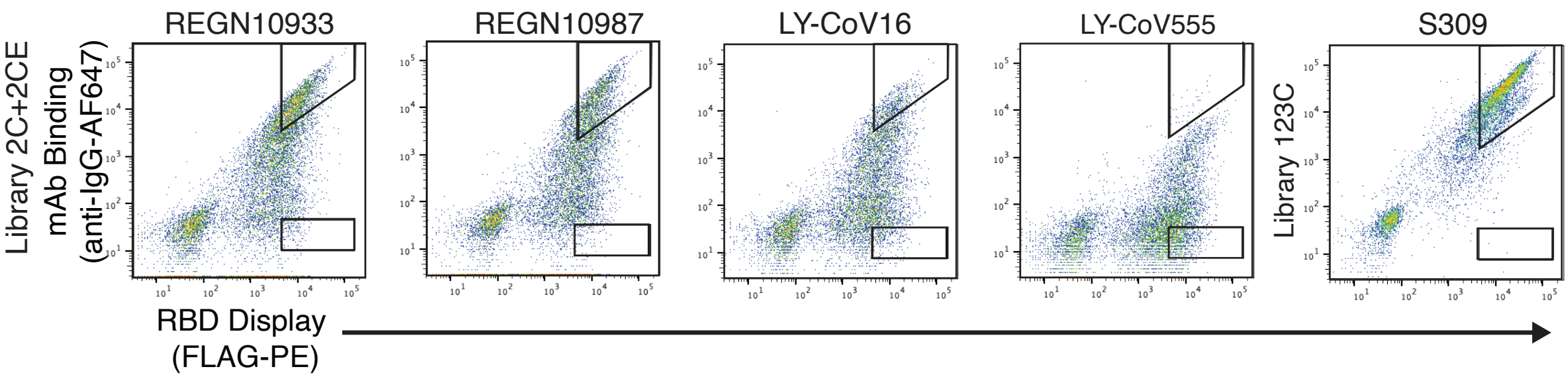
C

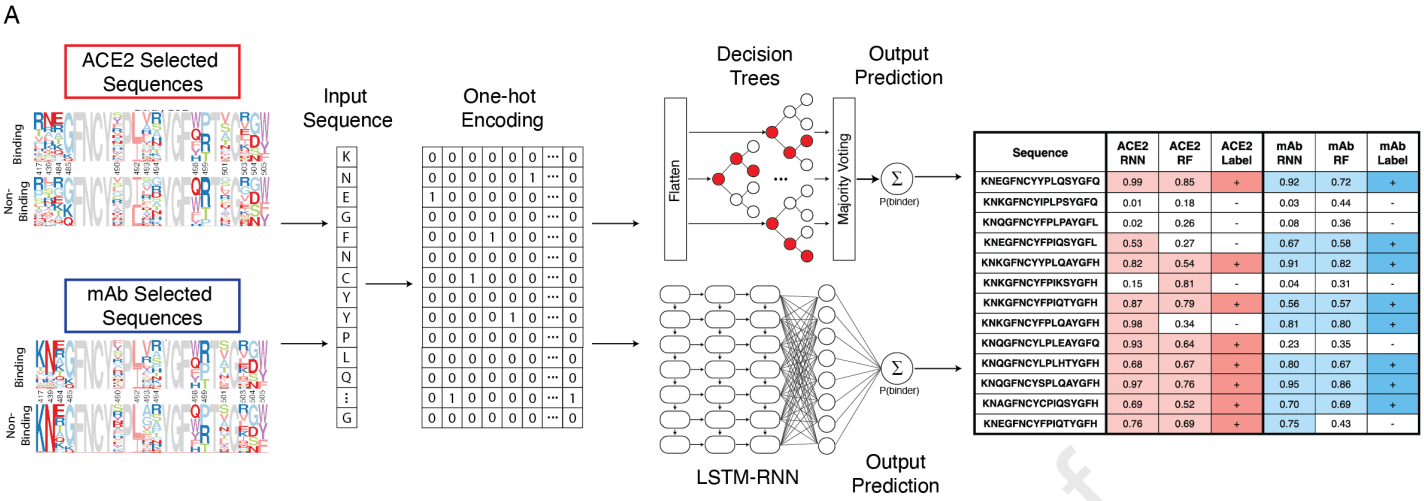


D

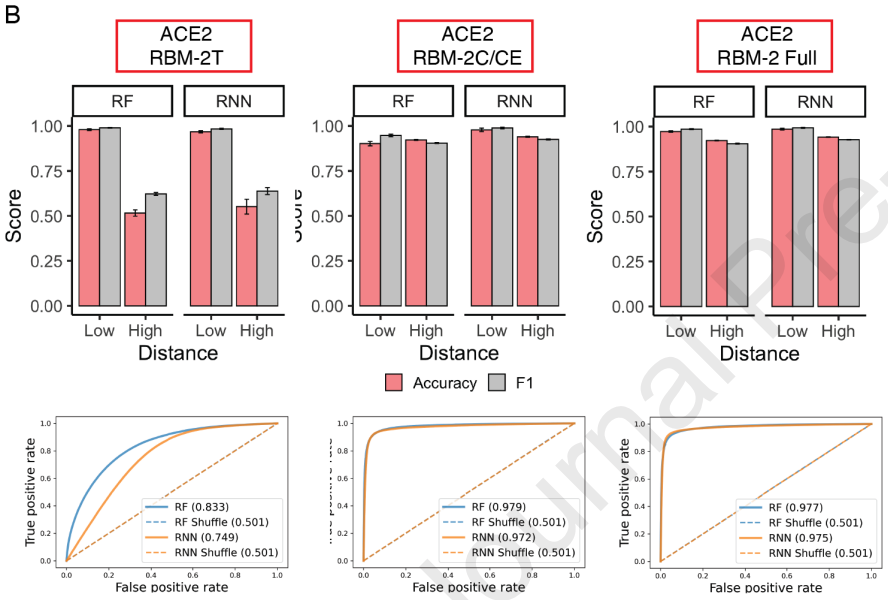


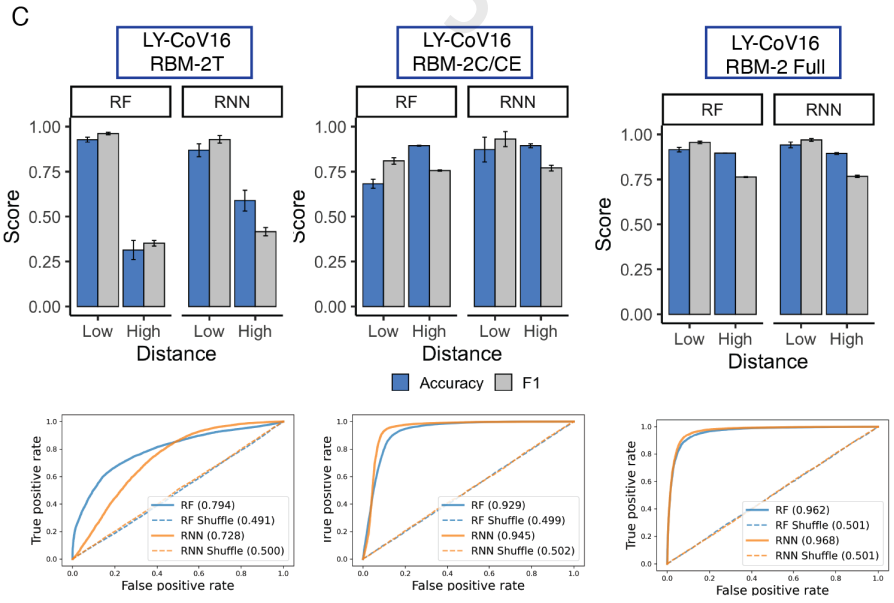
E

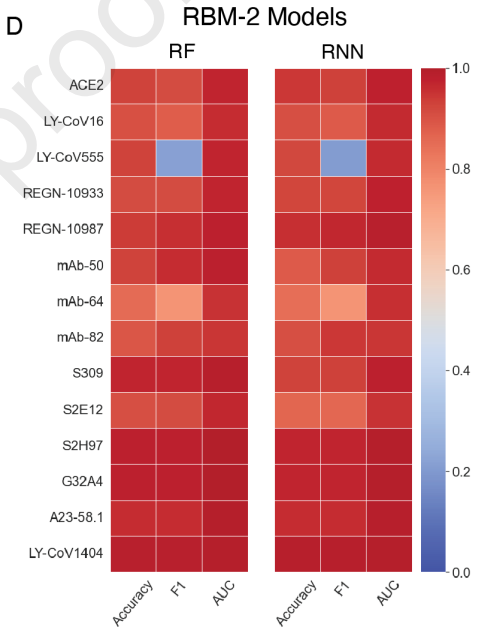


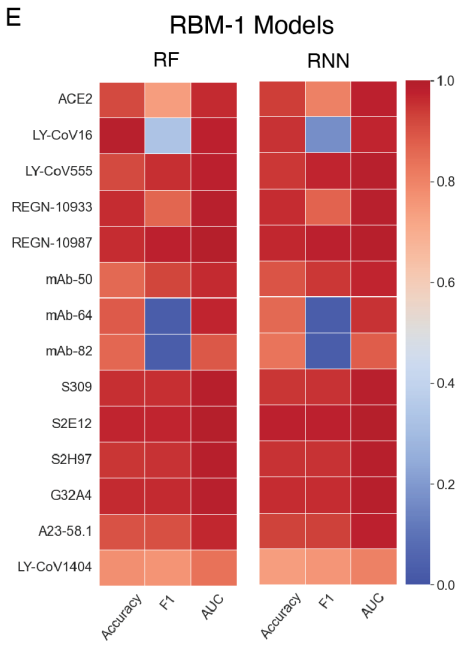


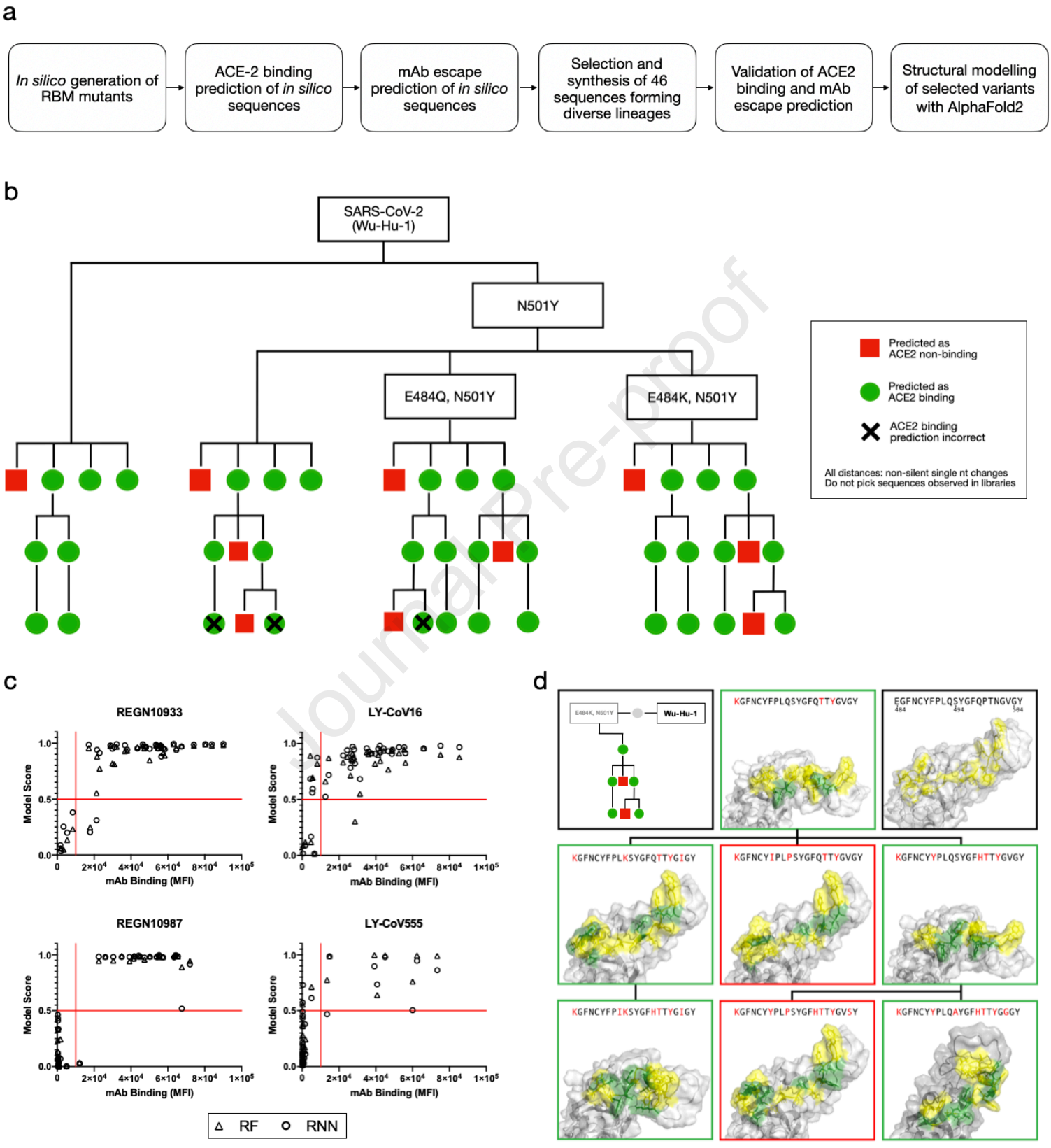
Sequence	ACE2 RNN	ACE2 RF	ACE2 Label	mAb RNN	mAb RF	mAb Label
KNEGFNCYYPLQSYGFQ	0.99	0.85	+	0.92	0.72	+
KNKGFNCYIPLPSYGFQ	0.01	0.18	-	0.03	0.44	-
KNQGFNCYFPLPAYGFL	0.02	0.26	-	0.08	0.36	-
KNEGFNCYFPIQSYGFL	0.53	0.27	-	0.67	0.58	+
KNKGFNCYYPLQAYGFH	0.82	0.54	+	0.91	0.82	+
KNKGFNCYFPIKSYGFH	0.15	0.81	-	0.04	0.31	-
KNKGFNCYFPIQTYGFH	0.87	0.79	+	0.56	0.57	+
KNKGFNCYFPLQAYGFH	0.98	0.34	-	0.81	0.80	+
KNQGFNCYLPLEAYGFQ	0.93	0.64	+	0.23	0.35	-
KNQGFNCYLPLHTYGFH	0.68	0.67	+	0.80	0.67	+
KNQGFNCYSPLQAYGFH	0.97	0.76	+	0.95	0.86	+
KNAGFNCYCPIQSYGFH	0.69	0.52	+	0.70	0.69	+
KNEGFNCYFPIQTYGFH	0.76	0.69	+	0.75	0.43	-



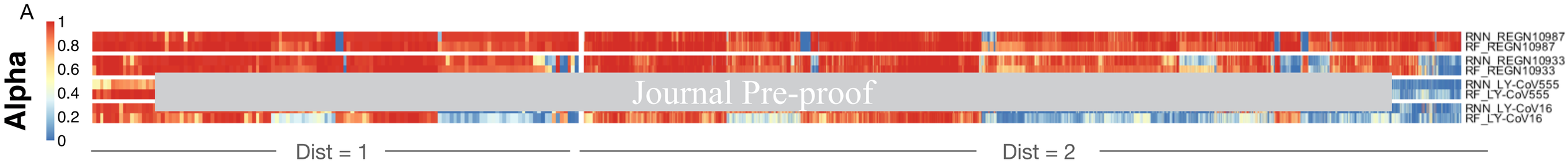




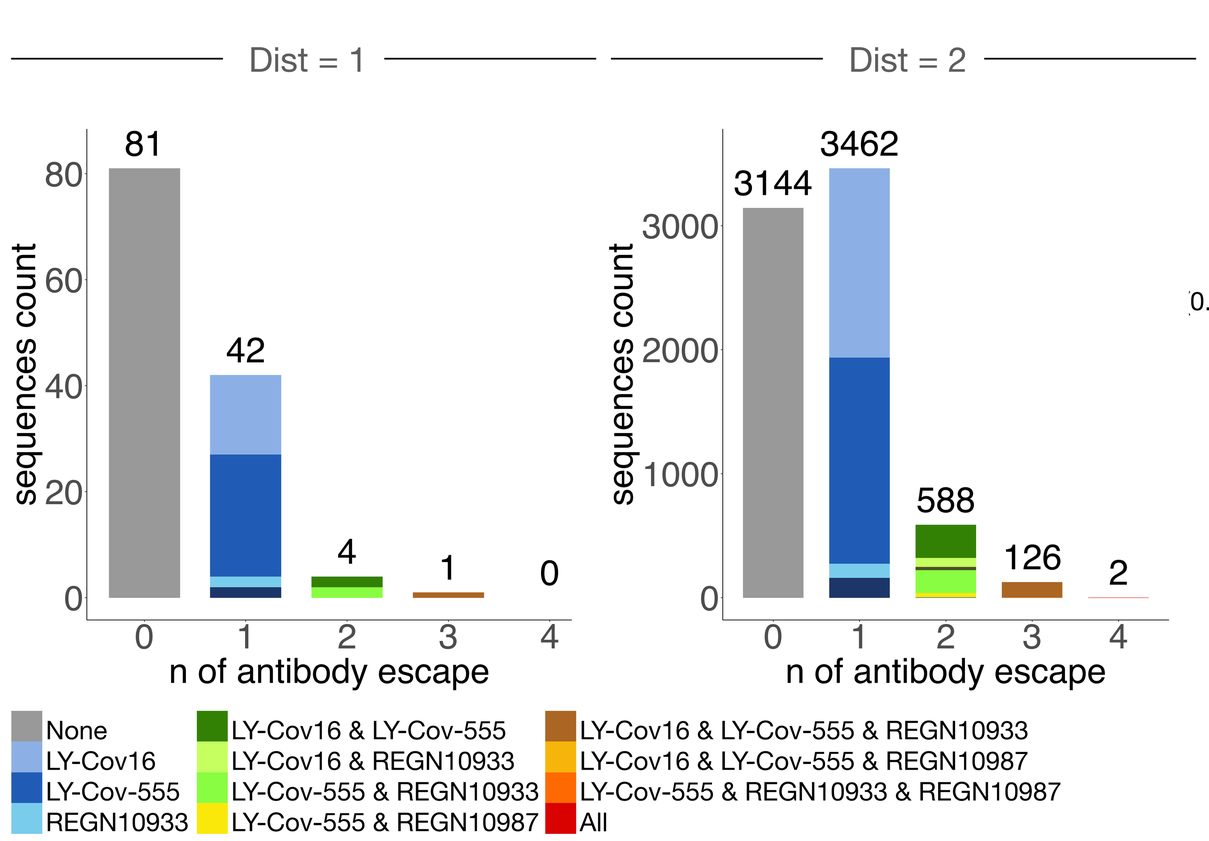




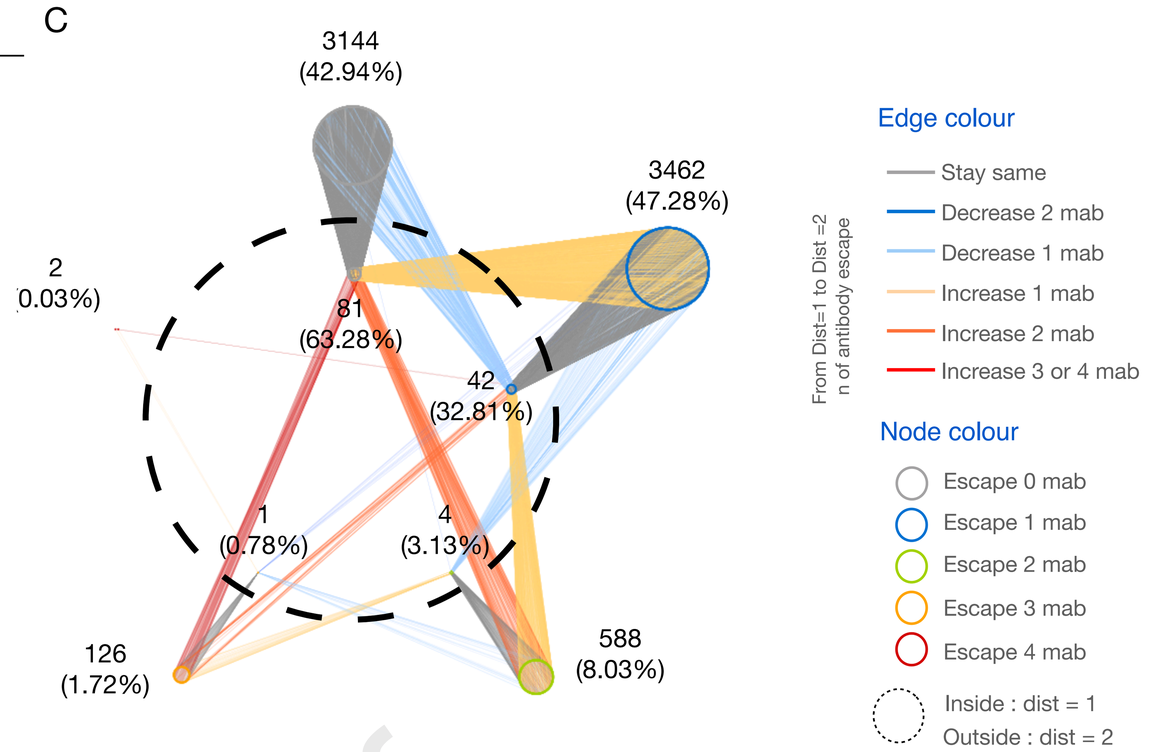
A



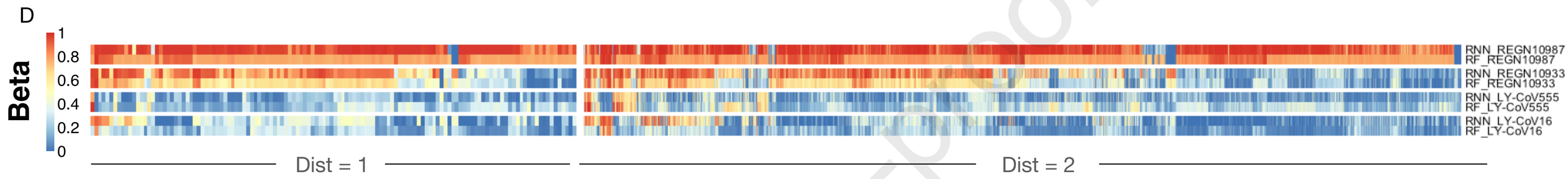
B



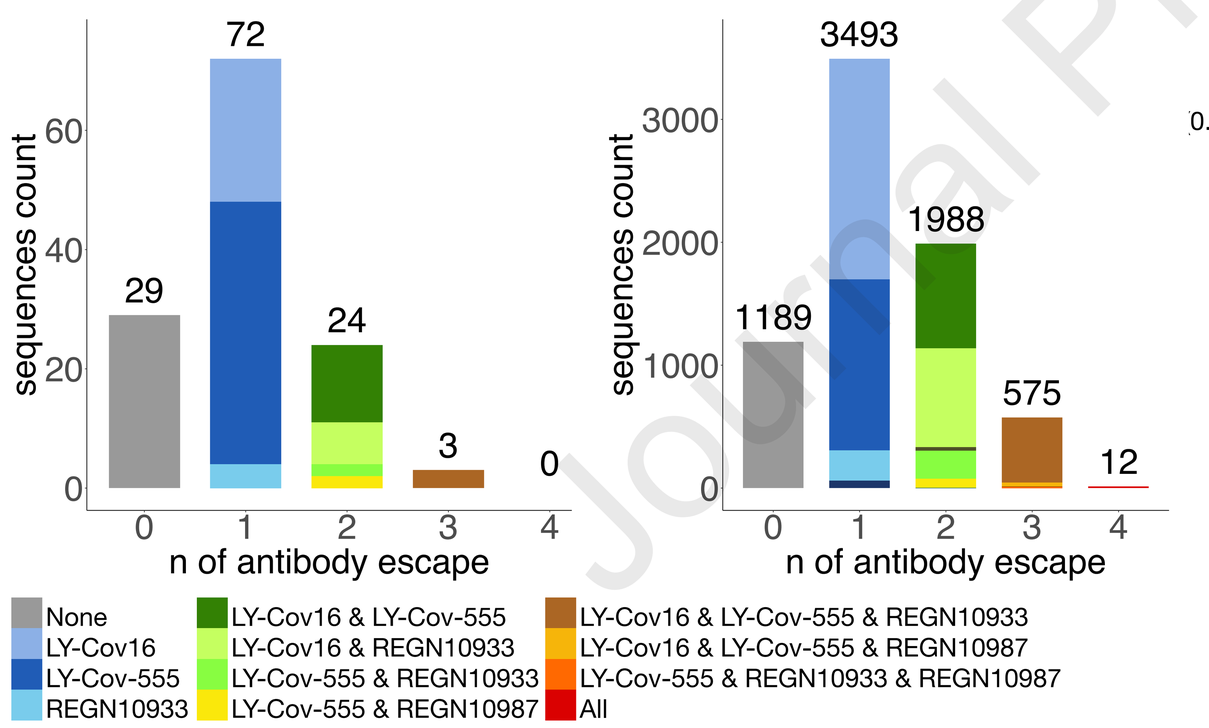
C



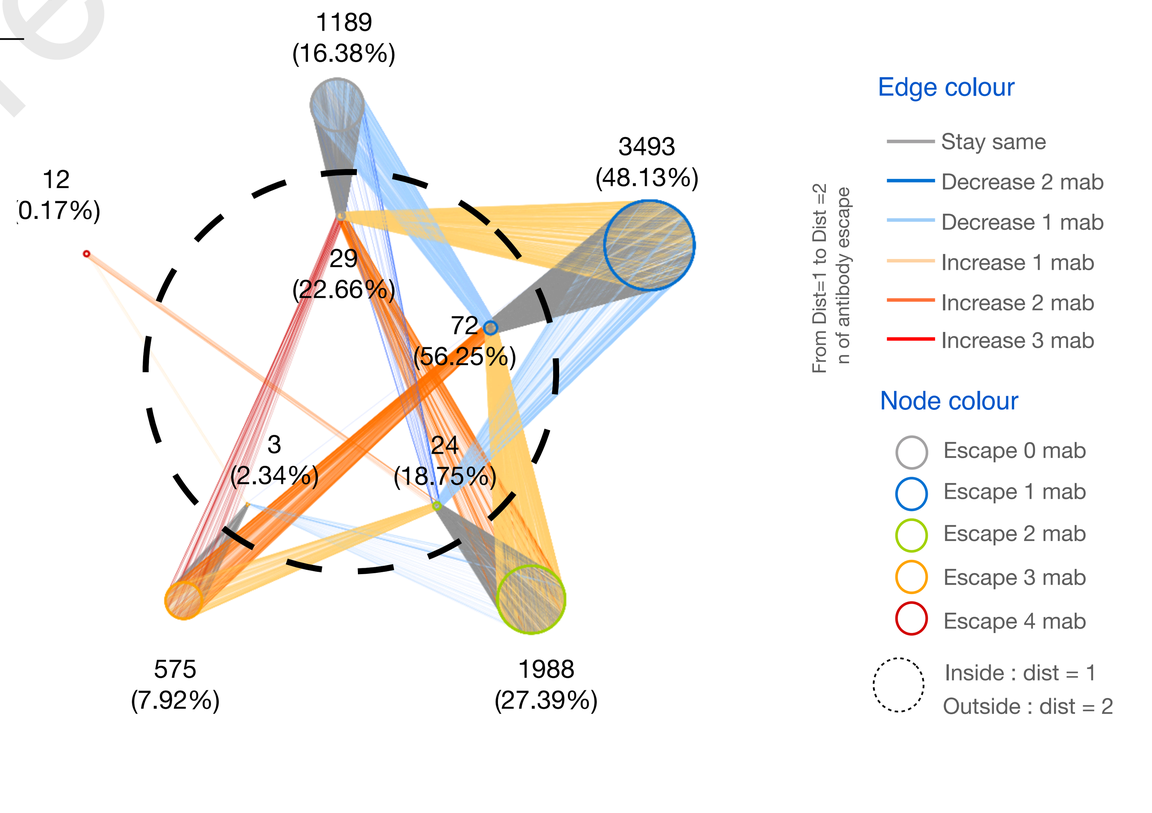
D



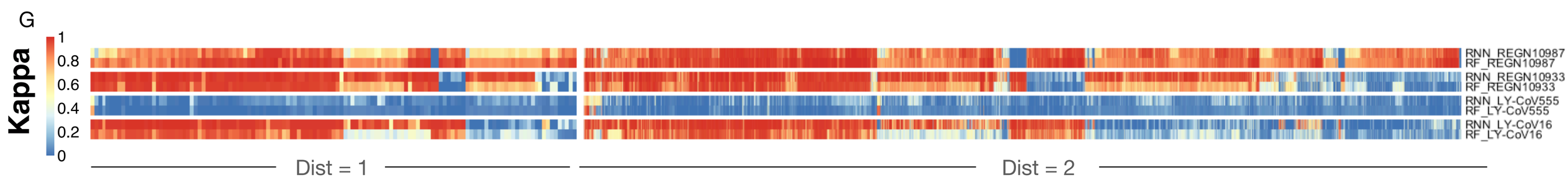
E



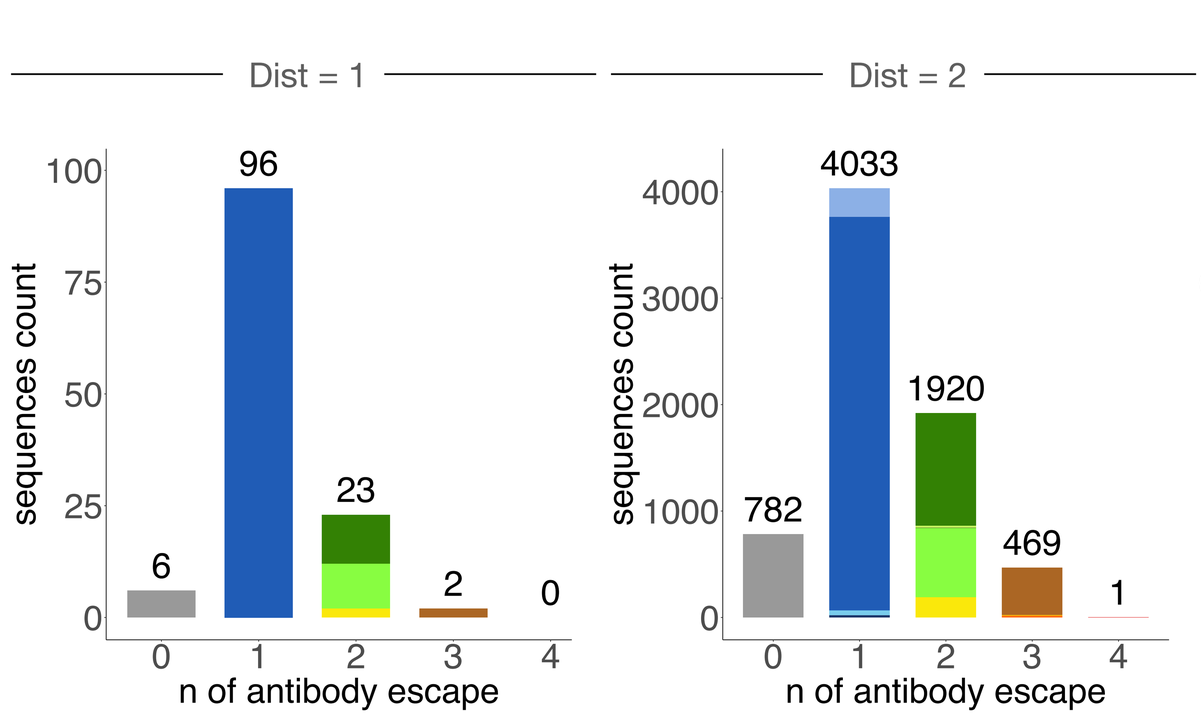
F



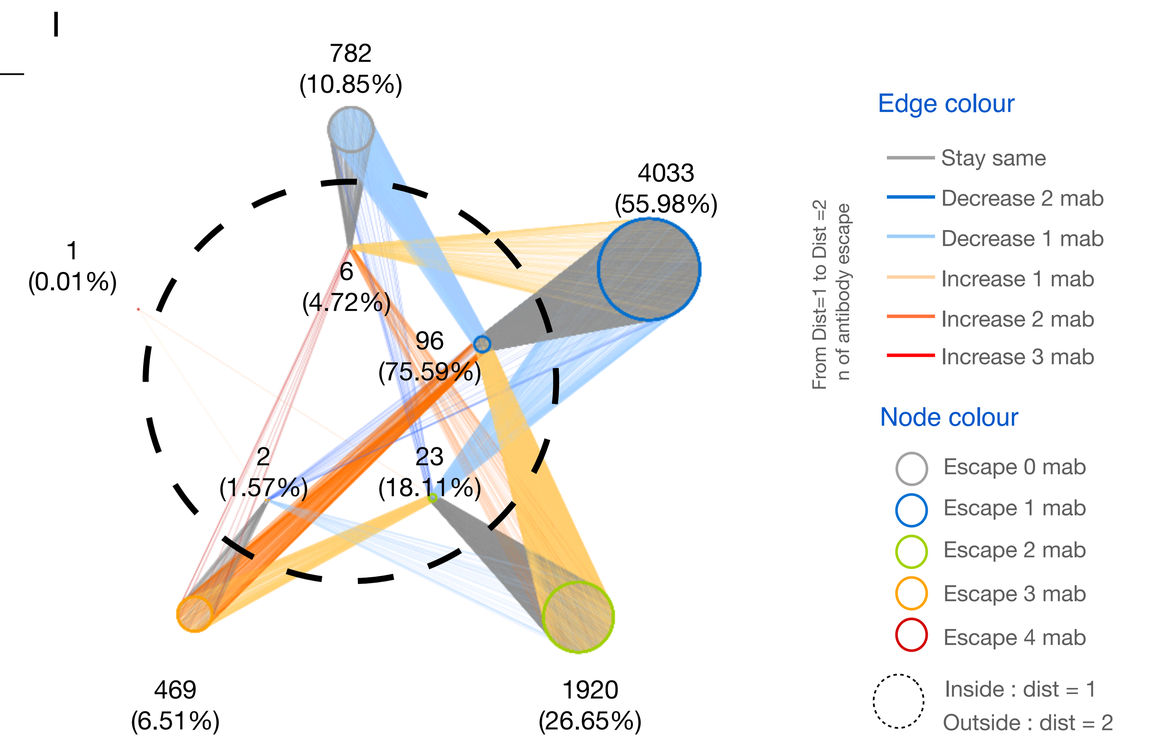
G

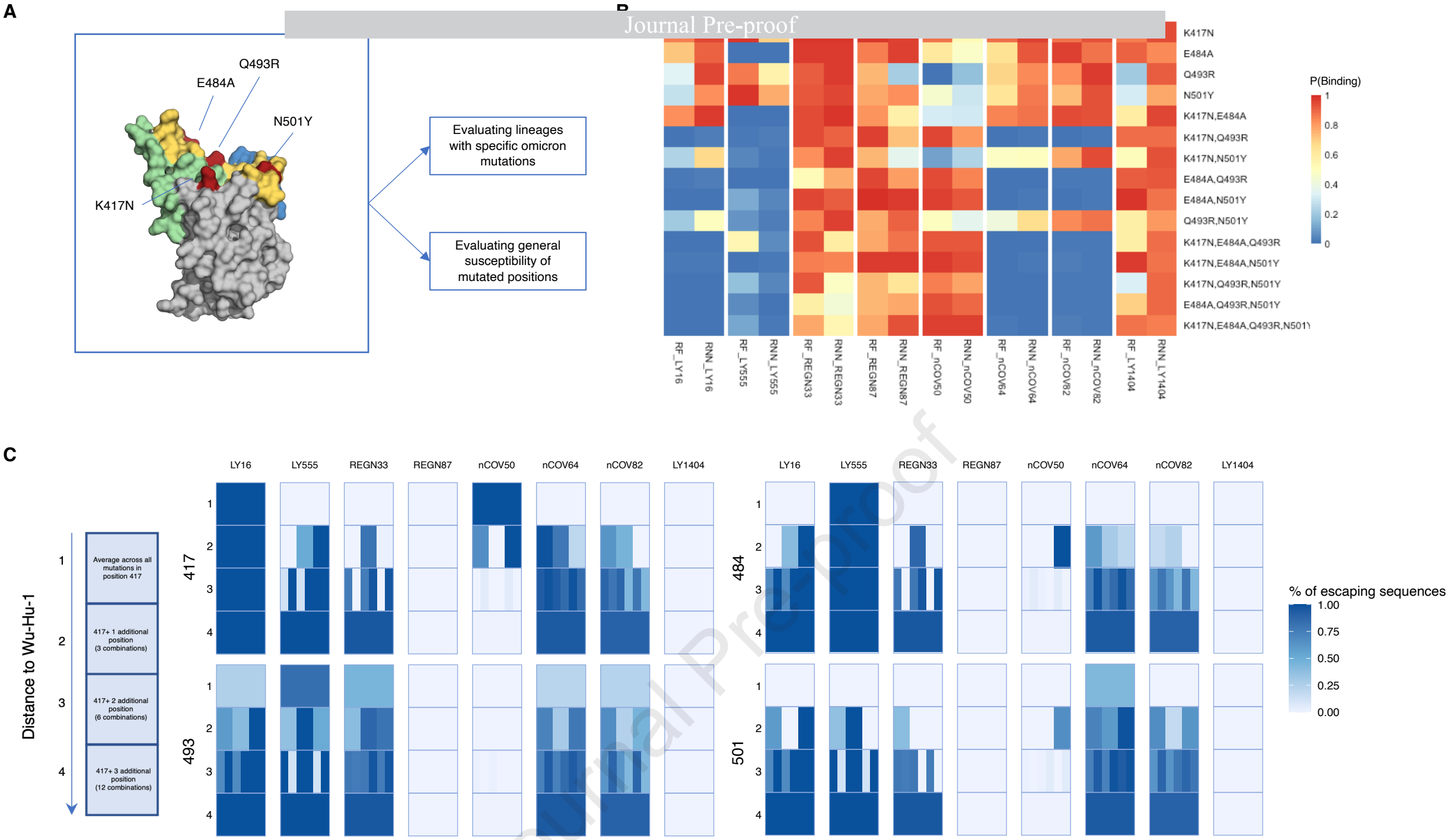


H



I





Deep Mutational Learning Predicts ACE2 Binding and Antibody Escape to Combinatorial Mutations in the SARS-CoV-2 Receptor Binding Domain

Joseph M. Taft[1, 2]*, Cédric R. Weber[3]*, Beichen Gao[1, 2], Roy A. Ehling[1], Jiami Han[1, 2], Lester Frei[1, 2], Sean W. Metcalfe[1], Max Overath[1], Alexander Yermanos[1, 2, 4, 5], William Kelton[6] and Sai T. Reddy[1,2#].

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel 4058, Switzerland
[2]Botnar Research Centre for Child Health, Basel 4058, Switzerland
[3]Alloy Therapeutics (Switzerland) AG, Basel 4058, Switzerland
[4]Department of Biology, Institute of Microbiology and Immunology, ETH Zurich, Zurich 8093, Switzerland
[5]Department of Pathology and Immunology, University of Geneva, Geneva 1211, Switzerland
[6]Te Huataki Waiora School of Health, University of Waikato, Hamilton 3240, New Zealand

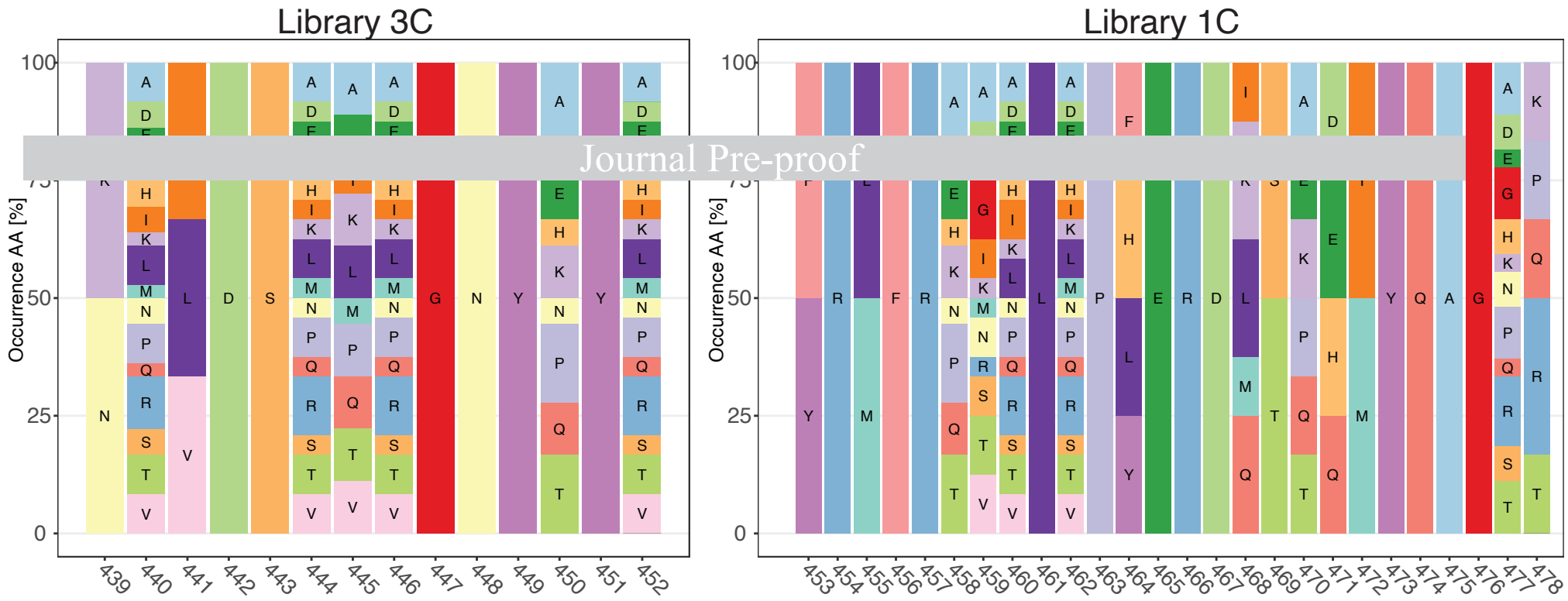
*Equal contribution
#Lead contact: sai.reddy@ethz.ch

Highlights

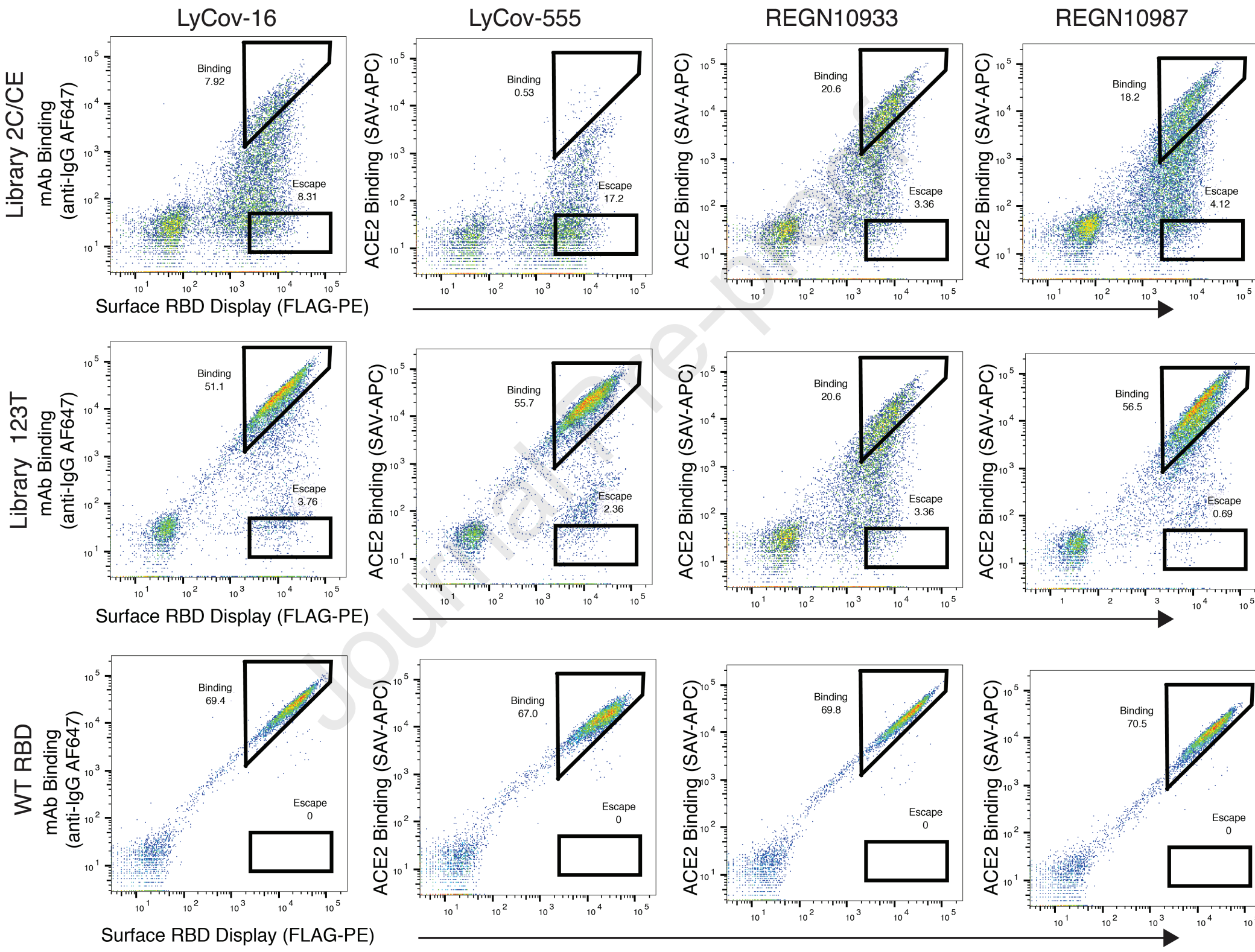
- Millions of combinatorial SARS-Cov-2-RBD variants screened by yeast surface display
- Machine learning models accurately predict ACE2 binding and antibody escape
- Identification of combinatorial mutations that drive escape to multiple antibodies
- Assessment of antibody robustness to billions of prospective RBD variants

In Brief:
A machine learning-guided, protein engineering method enables the prediction of how SARS-CoV-2 RBD combinatorial mutations will impact therapeutic antibody escape and ACE2 affinity. This method facilitates the identification of multisite mutations that are major drivers of antibody escape and the evaluation of neutralizing antibody efficacy on heavily mutated viral variants.

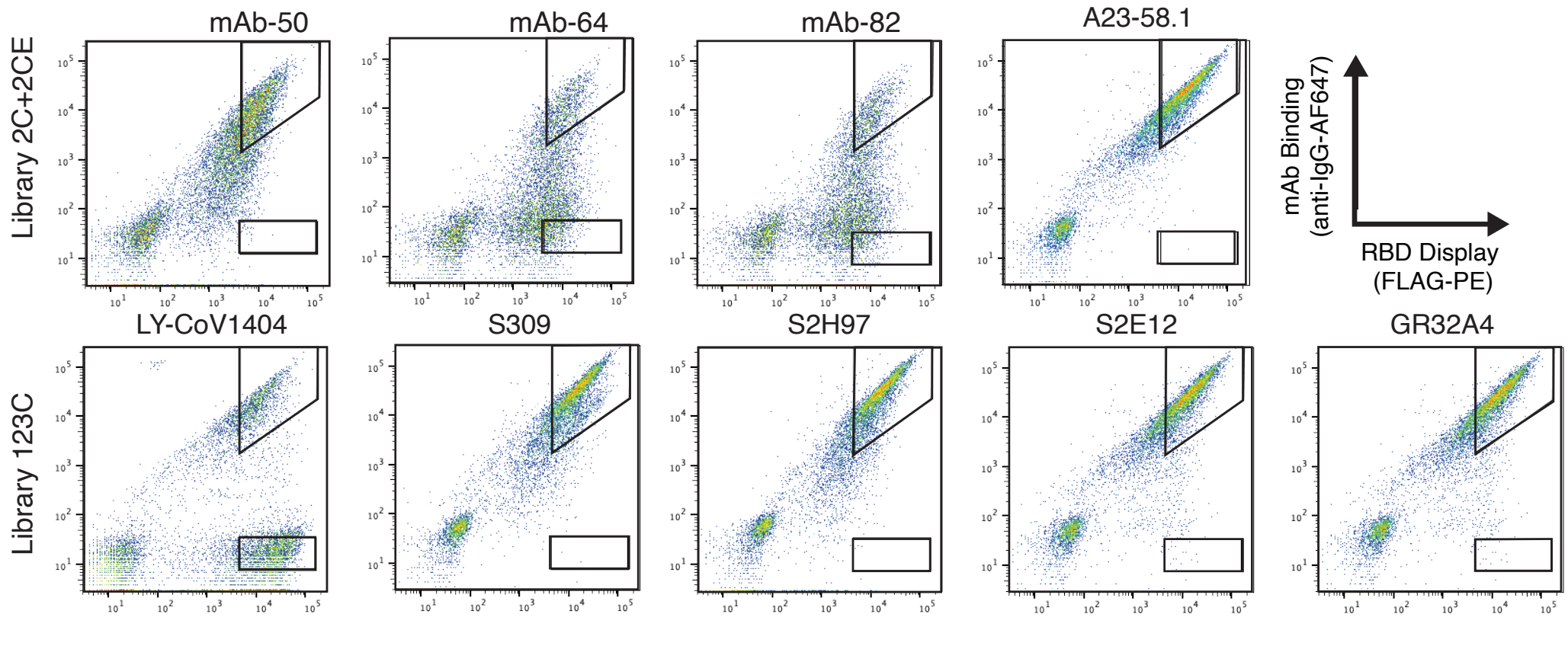
A



B

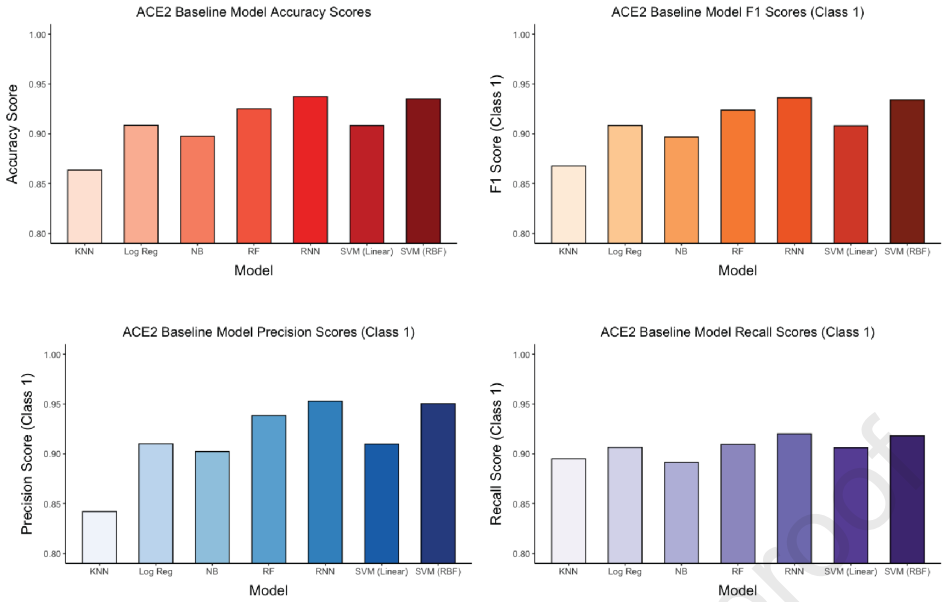


C

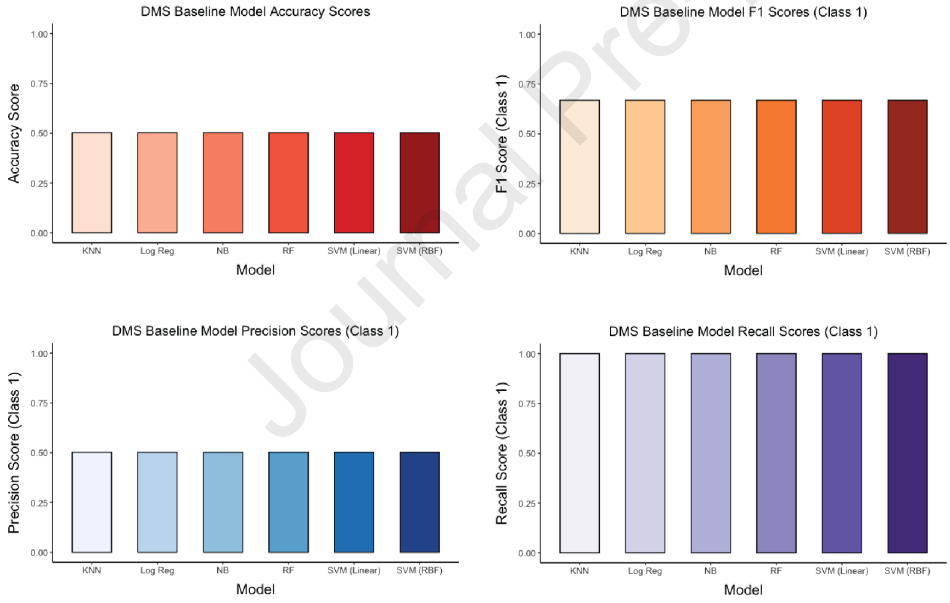



ACE2
Library 3C
Library 1C
Library 2CE
Library 2C
Library 3T
Library 1T
Library 2T
Non-Binding
Binding
REGN10933
mAb-50
REGN10987
mAb-64
LY-CoV16
mAb-62
LY-CoV555
LY-CoV1404
S309
S2E12
S2H97
GR32A4
A23-58.1
Escape

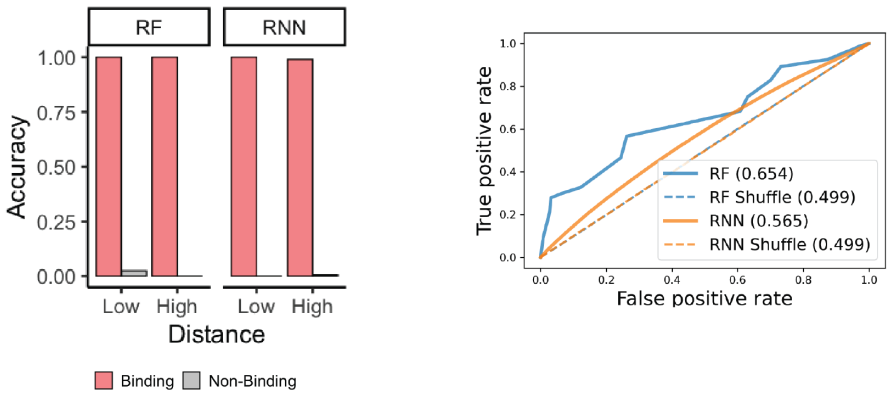
A

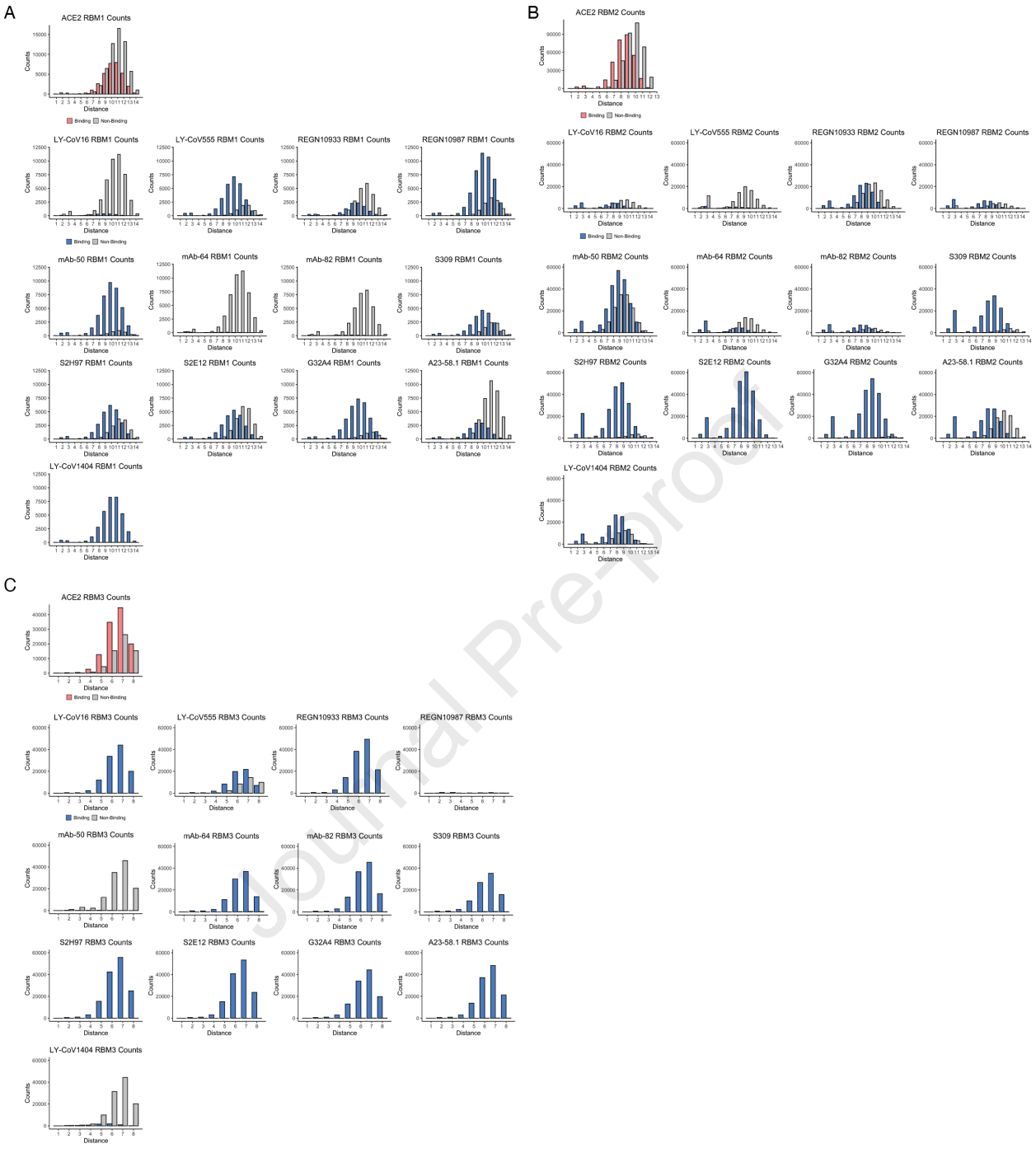


B

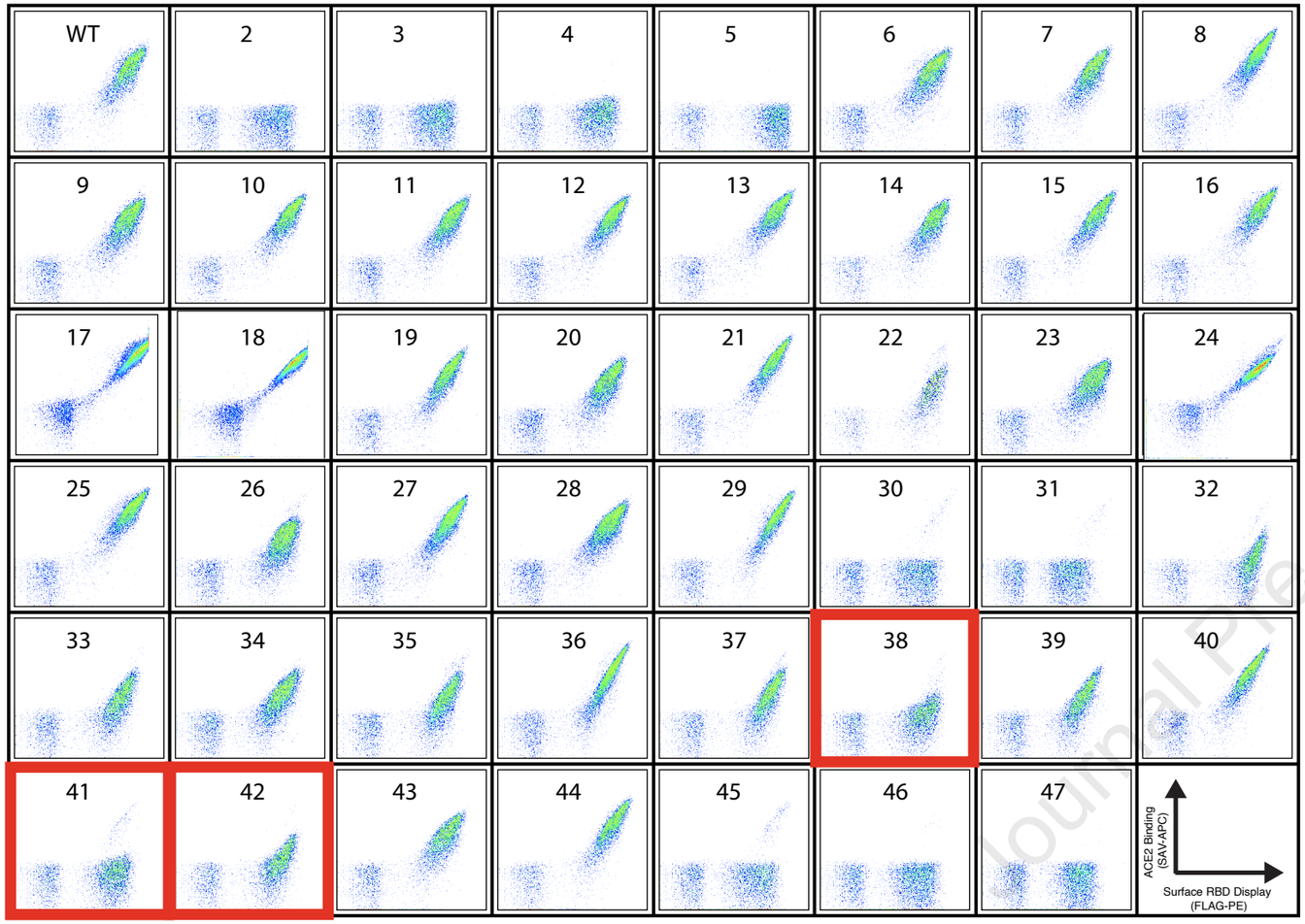


C

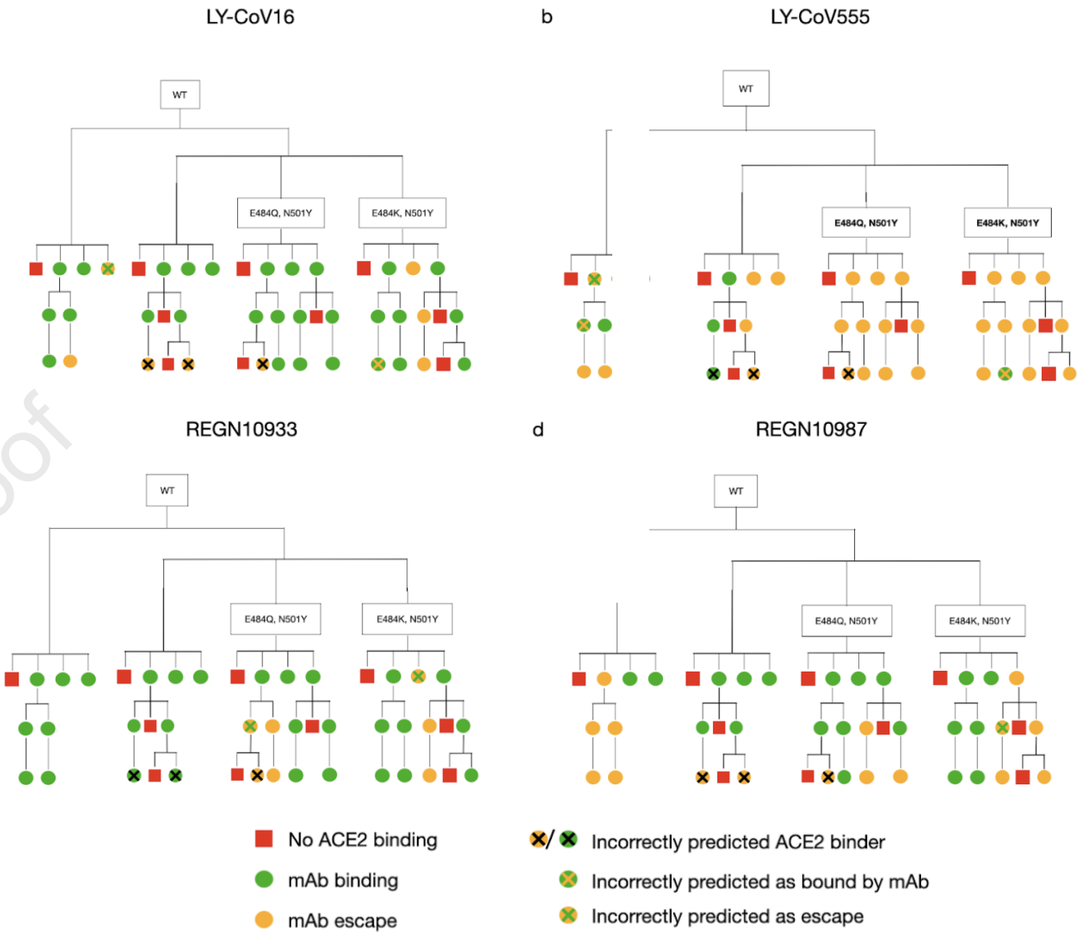




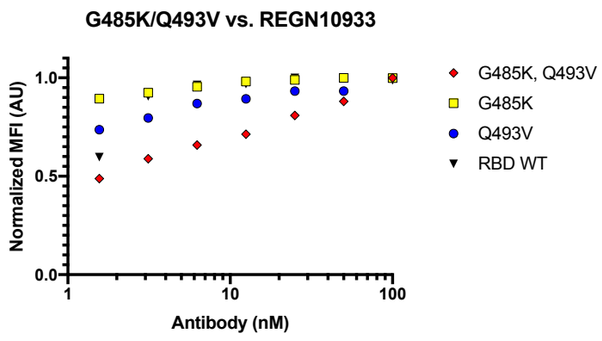
A



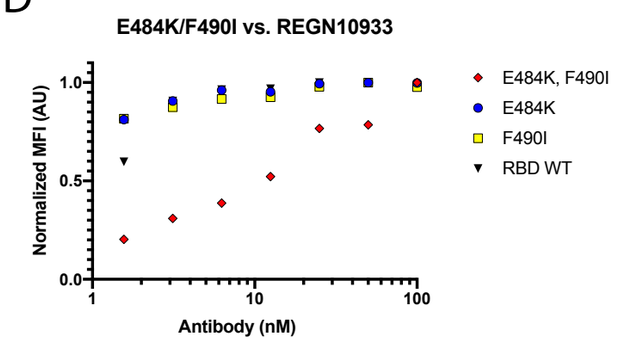
B



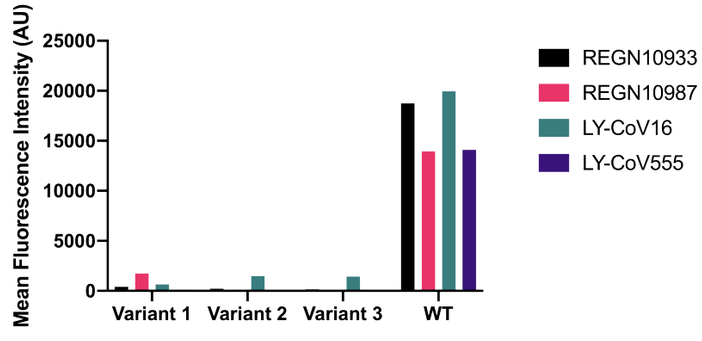
C



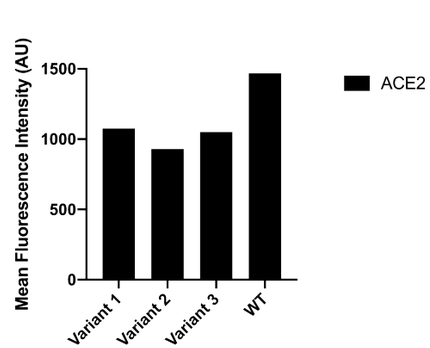
D

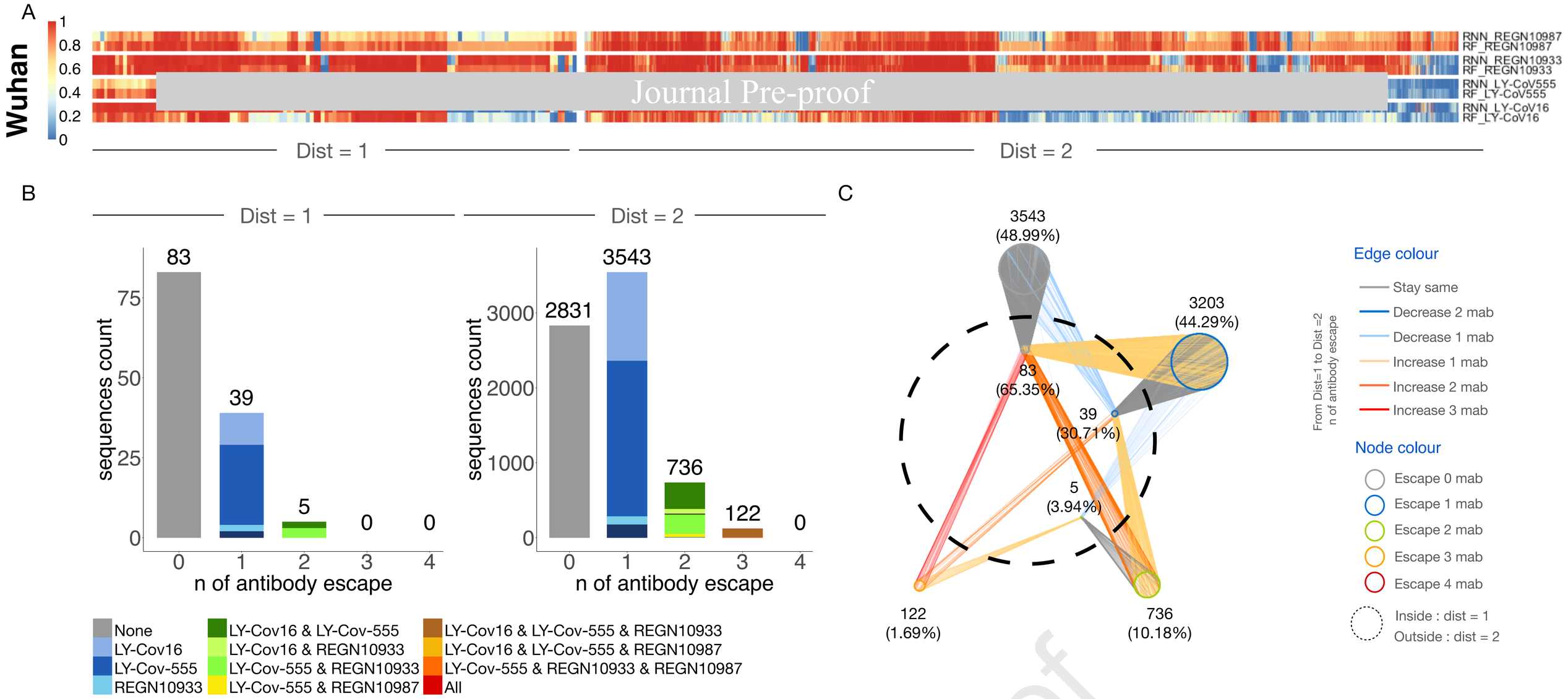


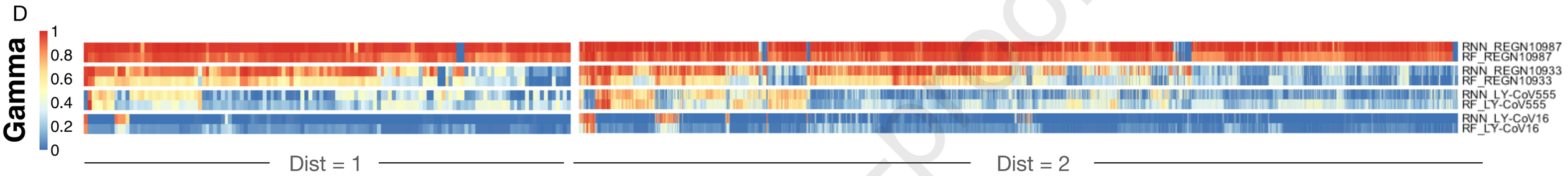
E

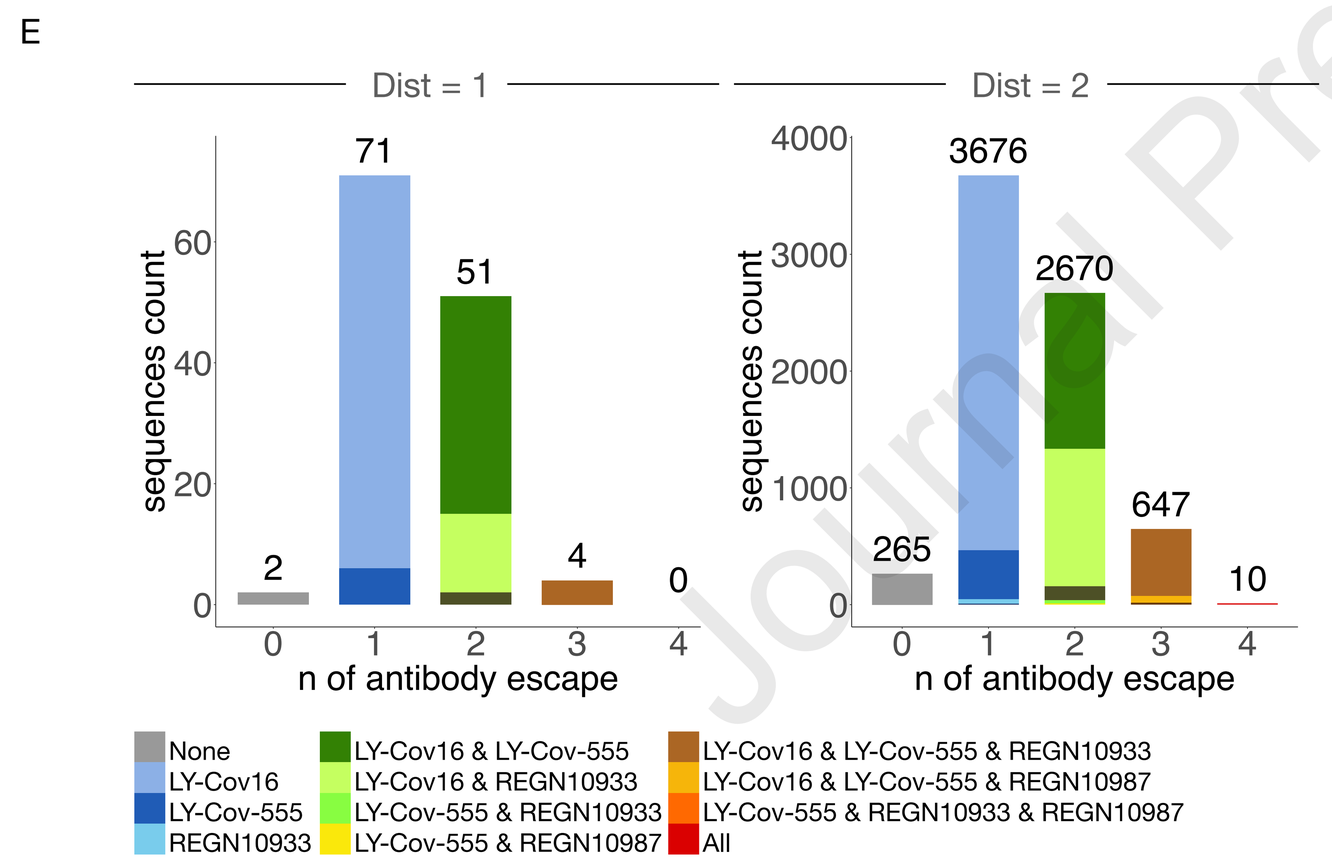


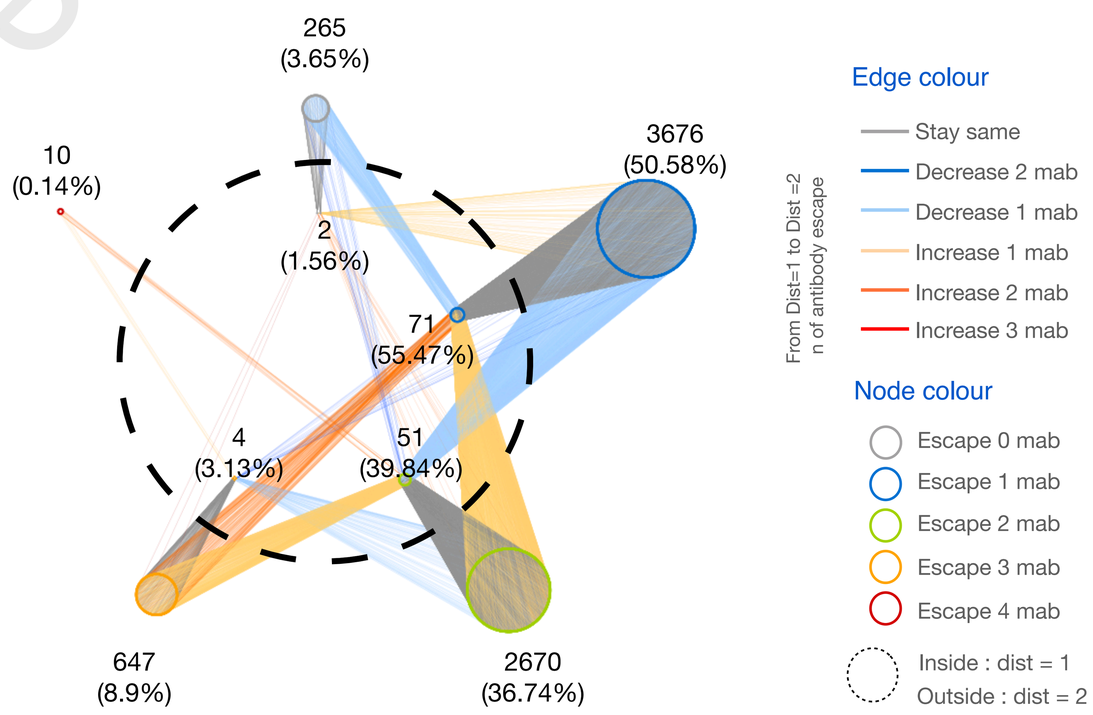
F

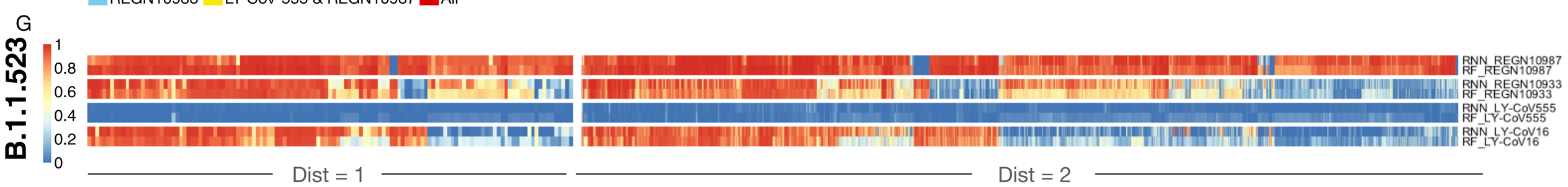


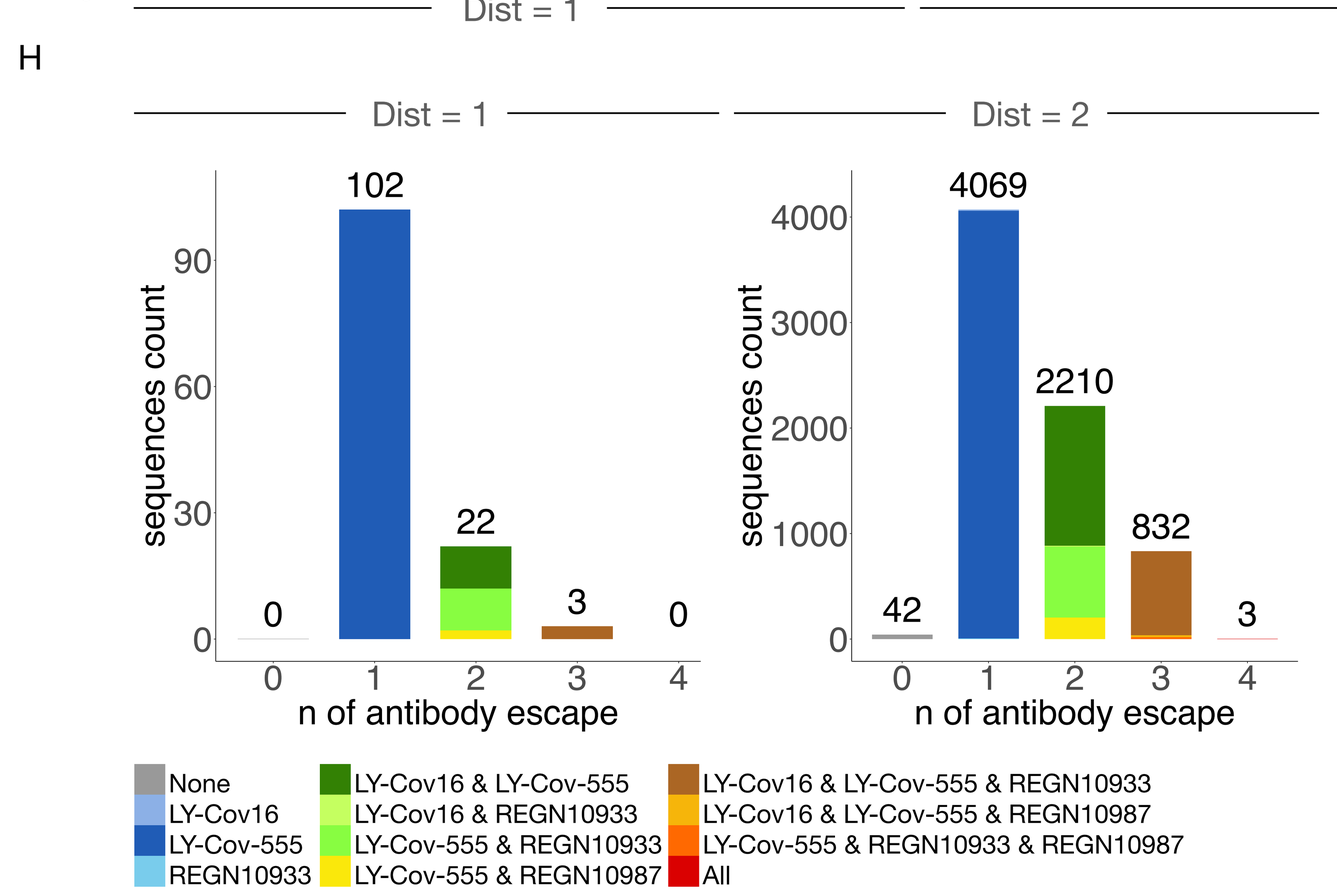


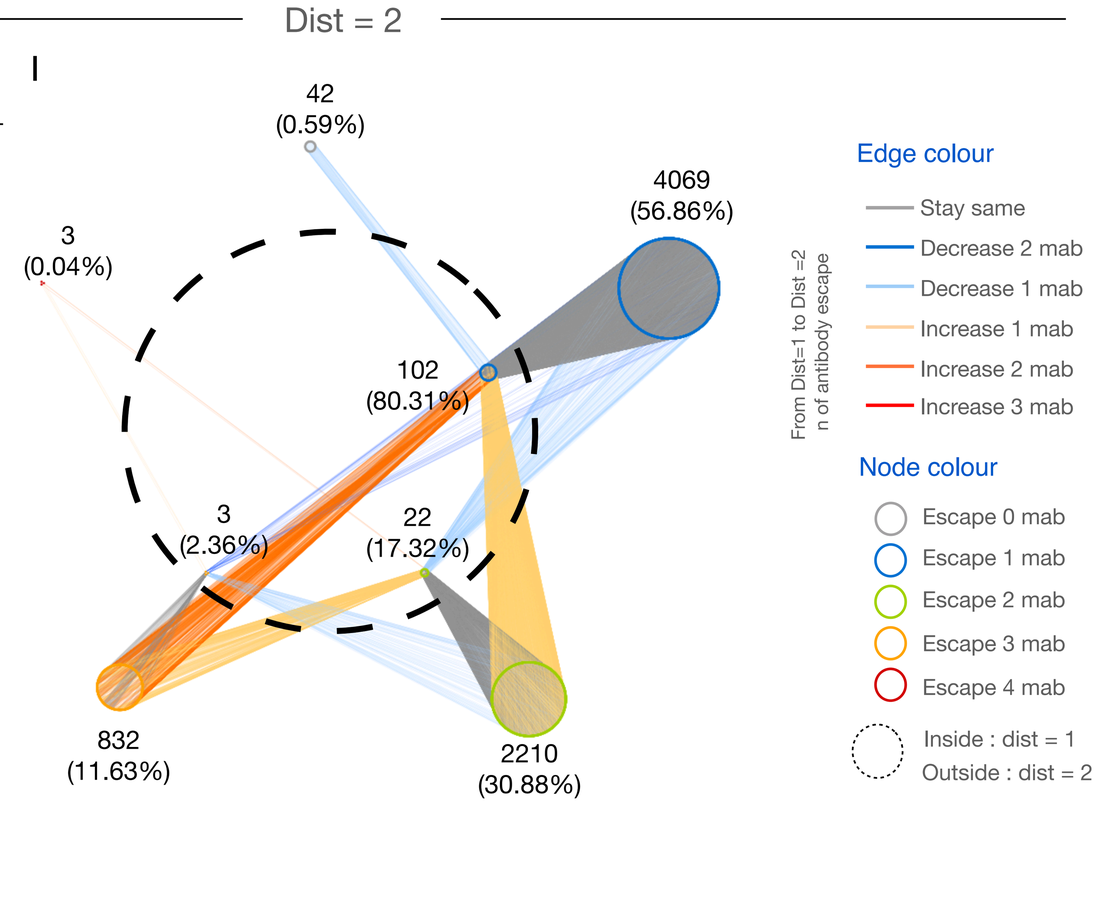












REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Flow Cytometry Staining Reagents		
PE anti-DYKDDDDK Tag Antibody	Biolegend	637309
anti-human IgG-AlexaFluor647	Jackson Immunoresearch	109-605-098
biotinylated human ACE2	Acro	AC2-H82E6
streptavidin-AlexaFluor 647	Biolegend	405237
Deposited Data		
Raw sequencing data	Starr et al. 2020	NCBI SRA: BioProject PRJNA639956
Raw and processed sequencing data	This study	https://github.com/LSSI-ETH
Oligonucleotides		
Degenerate Ultramers and oPools for RBD library construction	IDT	https://github.com/LSSI-ETH
Recombinant DNA		
pYD1-RBD(wt)	This study	https://github.com/LSSI-ETH
Cell Lines		
EBY100	ATCC	MYA-4941
Software and Algorithms		
bbduk	Joint Genome Institute	https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbduk-guide/
custom scripts for curation, analysis, and visualization	This study	https://github.com/LSSI-ETH