Supplementary Material
Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies

Jiahui Chen[1,†], Kaifu Gao[1,†], Rui Wang[1,†], and Guo-Wei Wei[1,2,3*]
[1] Department of Mathematics,
Michigan State University, MI 48824, USA.
[2] Department of Electrical and Computer Engineering,
Michigan State University, MI 48824, USA.
[3] Department of Biochemistry and Molecular Biology,
Michigan State University, MI 48824, USA.
[†]First three authors contributed equally.

March 19, 2021

*Corresponding author. E-mail: weig@msu.edu

Contents

S1 Methods

S1.1 Data collection and pre-processing

In this work, we retrieved over 203,246 complete genome sequences with high coverage of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) strains from the infected individuals in the world downloaded from the GISAID database [16] (https://www.gisaid.org/) as of January 20, 2021. The complete genome sequence of SARS-CoV-2 was first released on the GenBank (Access number: NC_045512.2) submitted Zhang's group at Fudan University [24] on January 5, 2020. Since then, there has been a rapid accumulation of SARS-CoV-2 genome sequences. Incomplete records and records without the exact submission date in GISAID were not considered. To rearrange the complete genome sequences according to the reference SARS-CoV-2 genome, multiple sequence alignment (MSA) is carried out by using Clustal Omega [17] with default parameters. The amino acid sequence of NSP2, NSP12, NPS13, spike (S) protein, ORF3a, ORF8, and nucleocapsid were downloaded from the GenBank [1].

S1.2 Sequences, structures and their alignments

All sequences and 3D structures are downloaded from Protein Data Bank (PDB https://www.rcsb.org): sequences are from the FASTA files and 3D structures are from pdb files.

The 3D alignments as well as graphs are created by using PyMOL [5]. The 2D sequence alignments are calculated by clustalw (https://www.genome.jp/tools-bin/clustalw) [20] and 2D alignment graphs are generated by Jalview [23].

S1.3 Secondary structure determination

To guarantee high accuracy, we used a hybrid approach to determine the secondary structure of the S protein. The 3D conformations consisting 1031 of 1273 residues of the S protein are already resolved in PDB structure 7C2L [4], and the secondary structure of these 1031 residues were assigned by PyMOL [5] based on their 3D conformations. The secondary structures of the remaining 242 residues missing in 7C2L were predicted by RaptorX-Property [22].

S1.4 TopNetTree model for protein-protein interaction (PPI) binding free enrrgy changes upon mutation

Mutation-induced protein-protein binding free energy (BFE) changes are an important approach for understanding the impact of mutations on protein-protein interactions (PPIs) and viral infectivity [10]. A variety of advanced methods has been developed [10, 15]. The topology-based network tree (TopNetTree) model [3, 21] is applied to predict mutation-induced BFE changes of PPIs in this work. TopNetTree model was implemented by integrating the topological representation and network tree (NetTree) to predict the BFE changes ($\Delta\Delta G$) of PPIs following mutations [21]. The structural complexity of protein-protein complexes is simplified by algebraic topology [2, 6, 25] and is represented as the vital biological information in terms of topological invariants. NetTree integrates the advantages of convolutional neural networks (CNN) and gradient-boosting trees (GBT), such that CNN is treated as an intermediate model that converts vectorized element- and site-specific persistent homology features into a higher-level abstract feature, and GBT uses the upstream features and other biochemistry features for prediction. The performance test of tenfold cross-validation on the dataset (SKEMPI 2.0 [8]) carried out using gradient boosted regression tree (GBRTs).

The errors with the SKEMPI2.0 dataset are 0.85 in terms of Pearson correlation coefficient (R_p) and 1.11 kcal/mol in terms of the root mean square error (RMSE) [21].

S1.4.1 Training set for TopNetTree model

The TopNetTree model is trained by several important training sets. The most important dataset which provides the information for binding free energy changes upon mutations is the SKEMPI 2.0 dataset [8]. The SKEMPI 2.0 is an updated version of the SKEMPI database, which contains new mutations and data from other three databases: AB-Bind [18], PROXiMATE [9], and dbMPIKT [12]. There are 7,085 elements including single- and multi-point mutations in SKEMPI 2.0. 4,169 variants in 319 different protein complexes are filtered as single-pint mutations are used for TopNetTree model training. Moreover, SARS-CoV-2 related datasets are also included to improve the prediction accuracy after a label transformation. They are all deep mutation enrichment ratio data, mutational scanning data of ACE2 binding to the receptor binding domain (RBD) of the S protein [14], mutational scanning data of RBD binding to ACE2 [11, 19], and mutational scanning data of RBD binding to CTC-445.2 and of CTC-445.2 binding to the RBD [11]. Note the training dataset used in the validation in main text does not include the test dataset, which the mutational data scanning data of RBD binding to CTC-445.2.

S1.4.2 Topology-based feature generation of PPIs

To construct the algebraic topological analysis on protein-protein interactions, we first preset the constructions for a PPI complex into various subsets.

1. $\mathcal{A}_{\mathrm{m}}$: atoms of the mutation sites.
2. $\mathcal{A}_{\mathrm{mn}}(r)$: atoms in the neighbourhood of the mutation site within a cut-off distance r.
3. $\mathcal{A}_{\mathrm{Ab}}(r)$: antibody atoms within r of the binding site.
4. $\mathcal{A}_{\mathrm{Ag}}(r)$: antigen atoms within r of the binding site.
5. $\mathcal{A}_{\mathrm{ele}}(\mathrm{E})$: atoms in the system that has atoms of element type E. The distance matrix is specially designed such that it excludes the interactions between the atoms form the same set. For interactions between atoms a_i and a_j in set $\mathcal{A}$ and/or set $\mathcal{B}$, the modified distance is defined as

$$D_{\mathrm{mod}}(a_i, a_j) = \begin{cases} \infty, \text{ if } a_i, a_j \in \mathcal{A}, \text{ or } a_i, a_j \in \mathcal{B}, \\ D_e(a_i, a_j), \text{ if } a_i \in \mathcal{A} \text{ and } a_j \in \mathcal{B}, \end{cases} \tag{1}$$

   where $D_e(a_i, a_j)$ is the Euclidian distance between a_i and a_j.

In algebraic topology, molecular atoms of different can be constructed as points presented by v_0, v_1, v_2, ..., v_k as k+1 affinely independent points in simplicial complex. Simplicial complex is a finite collection of sets of points $K = \{\sigma_i\}$, and σ_i are called linear combinations of these points in $\mathbb{R}^n$ ($n \geq k$). To construct simplicial complex, two that are widely used for point clouds are the Vietoris-Rips (VR) complex and alpha complex which are applied in this model [6]. The boundary operator for a k-simplex would transfer a k-simplex to a $k-1$-simplex. Consequently, the algebraic construction to connect a sequence of complexes by boundary maps is called a chain complex

$$\cdots \xrightarrow{\partial_{i+1}} C_i(X) \xrightarrow{\partial_i} C_{i-1}(X) \xrightarrow{\partial_{i-1}} \cdots \xrightarrow{\partial_2} C_1(X) \xrightarrow{\partial_1} C_0(X) \xrightarrow{\partial_0} 0$$

and the kth homology group is the quotient group defined by

$$H_k = Z_k / B_k. \tag{2}$$

Then the Betti numbers are defined by the ranks of kth homology group H_k which counts k-dimensional invariants, especially, $\beta_0 = \text{rank}(H_0)$ reflects the number of connected components, $\beta_1 = \text{rank}(H_1)$ reflects the number of loops, and $\beta_2 = \text{rank}(H_2)$ reveals the number of voids or cavities. Together, the set of Betti numbers $\{\beta_0, \beta_1, \beta_2, \cdots\}$ indicates the intrinsic topological property of a system.

Persistent homology is devised track the multiscale topological information over different scales along a filtration [6] and is significant important for constructing feature vectors for the machine learning method. Features generated by binned barcode vectorization can reflect the strength of atom bonds, van der Waals interactions, and can be easily incorporated into a CNN, which captures and discriminates local patterns. Another method of vectorization is to get the statistics of bar lengths, birth values, and death values, such as sum, maximum, minimum, mean, and standard derivation. This method is applied to vectorize Betti-1 (H_1) and Betti-2 (H_2) barcodes obtained from alpha complex filtration based on the facts that higher-dimensional barcodes are sparser than H_0 barcodes.

S1.4.3 Machine learning models

It is very challenging to predict binding affinity changes following mutation for PPIs due to the complex dataset and 3D structures. A hybrid machine learning algorithm that integrates a CNN and GBT is designed to overcome difficulties, such that partial topologically simplified descriptions are converted into concise features by the CNN module and a GBT module is trained on the whole feature set for a robust predictor with effective control of overfitting [21]. The gradient boosting tree (GBT) method produces a prediction model as an ensemble method which is a class of machine learning algorithms. It builds a popular module for regression and classification problems from weak learners. By the assumption that the individual learners are likely to make different mistakes, the method using a summation of the weak learners to eliminate the overall error. Furthermore, a decision tree is added to the ensemble depending on the current prediction error on the training dataset. Therefore, this method (a topology-based GBT or TopGBT) is relatively robust against hyperparameter tuning and overfitting, especially for a moderate number of features. The GBT is shown for its robustness against overfitting, good performance for moderately small data sizes, and model interpretability. The current work uses the package provided by scikit-learn (v 0.23.0) [13]. A supervised CNN model with the PPI $\Delta\Delta G$ as labels is trained for extracting high-level features from H_0 barcodes. Once the model is set up, the flatten layer neural outputs of CNN are feed into a GBT model to rank their importance. Based on the importance, and ordered subset of CNN-trained features is combined with features constructed from high-dimensional topological barcodes, H_1 and H_2 into the final GBT model.

S2 Multiple sequence alignments of antibodies and pairwise identity scores

Through the sequence clustering algorithm in CD-HIT suite [7], the 46 antibodies were classified into 28 clusters. Among them, the first five clusters contain more than one antibody. Figures S1-S5 are the multiple sequence alignments of these five clusters. The pairwise identity scores inside each of these five clusters are over 0.9, especially clusters 2 and 4 have such scores over 0.95. Their pairwise identity scores are deposited in the file "antibody-2d-score-matrix.csv".

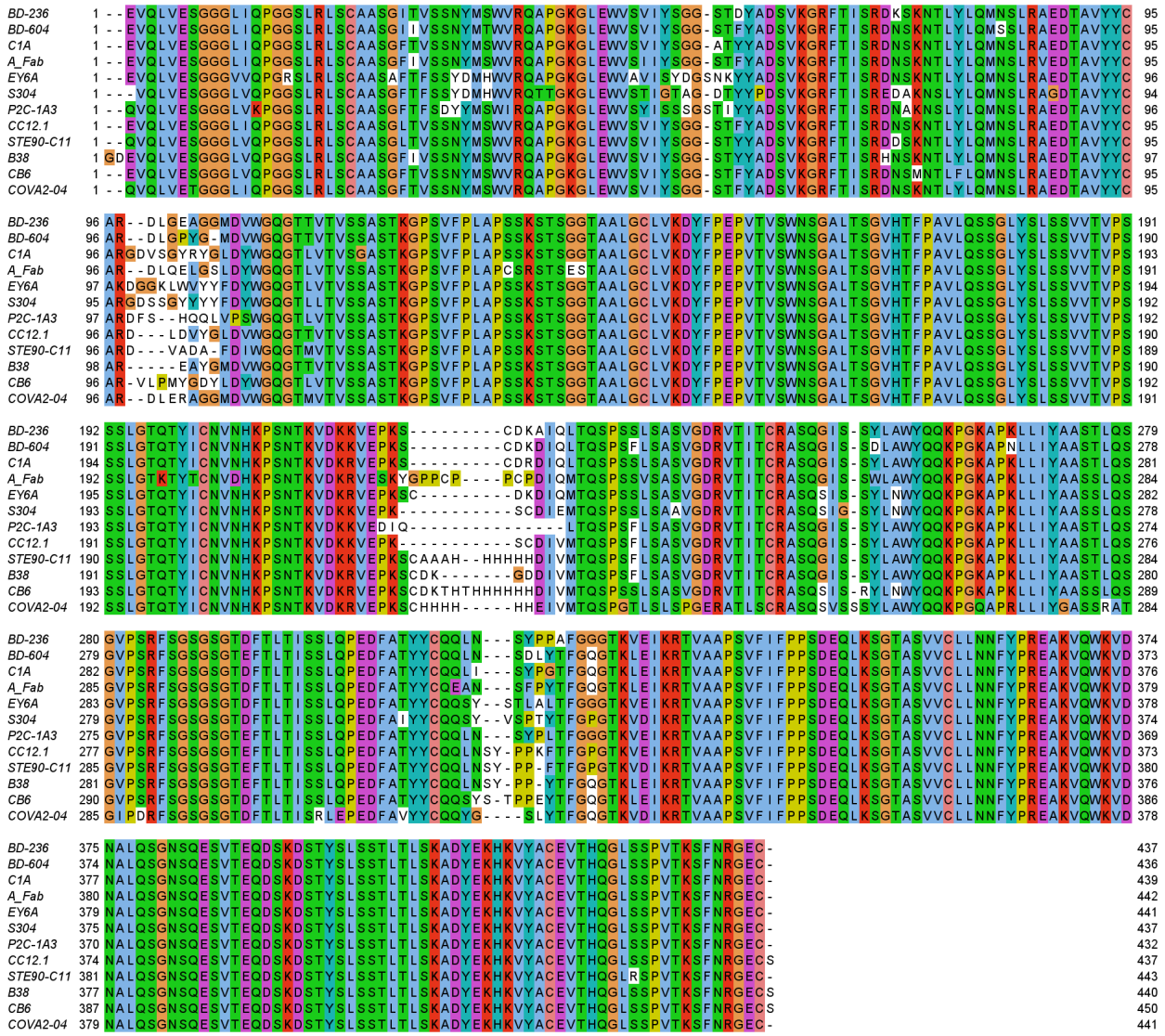

Figure S1: The 2D sequence alignment of the antibodies in cluster 1: BD-236, BD-604, C1A, a fab, EY6A, S304, P2C-1A3, CC12.1, STE90-C11, B38, CB6, COVA2-04.

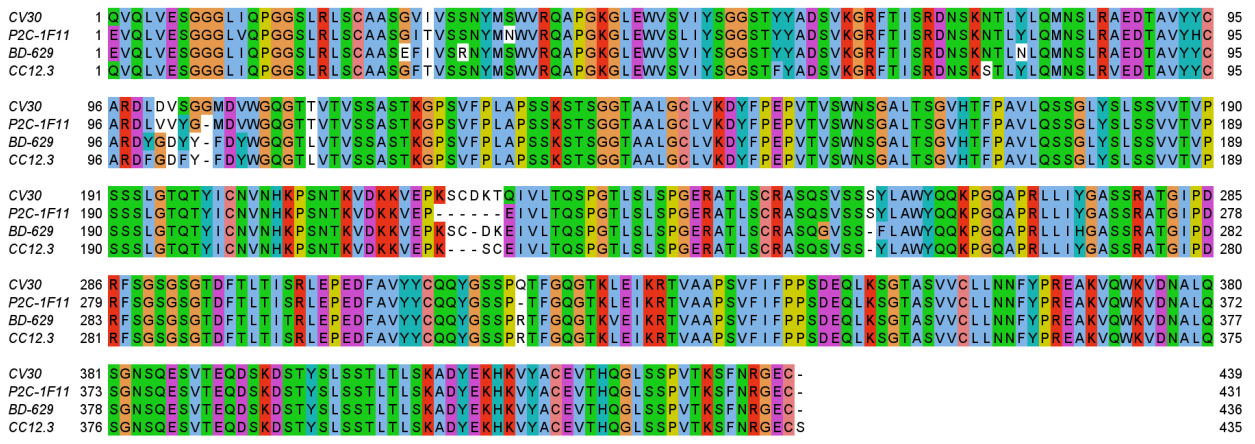

Figure S2: The 2D sequence alignment of the antibodies in cluster 2: CV30, P2C-1F11, BD-629, CC12.3.

Figure S3: The 2D sequence alignment of the antibodies in cluster 3: COVA2-39, CV07-270.

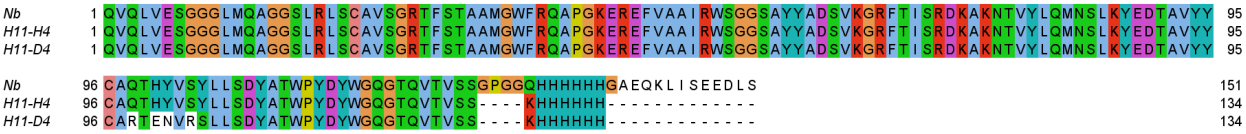

Figure S4: The 2D sequence alignment of the antibodies in cluster 4: Nb, H11-H4, H11-D4.

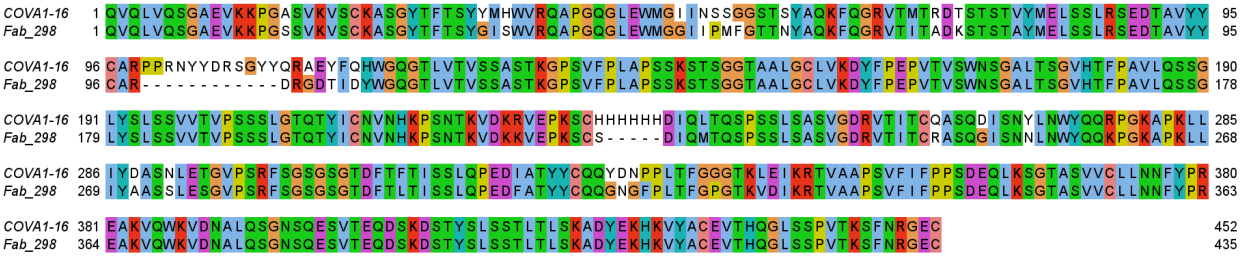

Figure S5: The 2D sequence alignment of the antibodies in cluster 5: COVA1-16, Fab 52.

S3 Random coil percentages of antibody paratopes

Table S1 depicts the random coil percentages of antibody paratopes on S protein, which indicates antibodies predominantly contact residues in random coils of the S protein.

S4 Additional analysis of antibody-S protein complexes

Three antibodies, i.e., 4A8, FC05, and 2G12, do not bind to the RBD. Among them, 4A8 has been analyzed in the main text of the paper and 2G12 involves small molecules at its binding site with the S2 domain of the S protein, which cannot be handled by the present model. Antibody FC05 has two complexes with the S protein (i.e., 7CWU and 7CDJ). Both of them share the same antibody at the N-terminal domain (NTD).

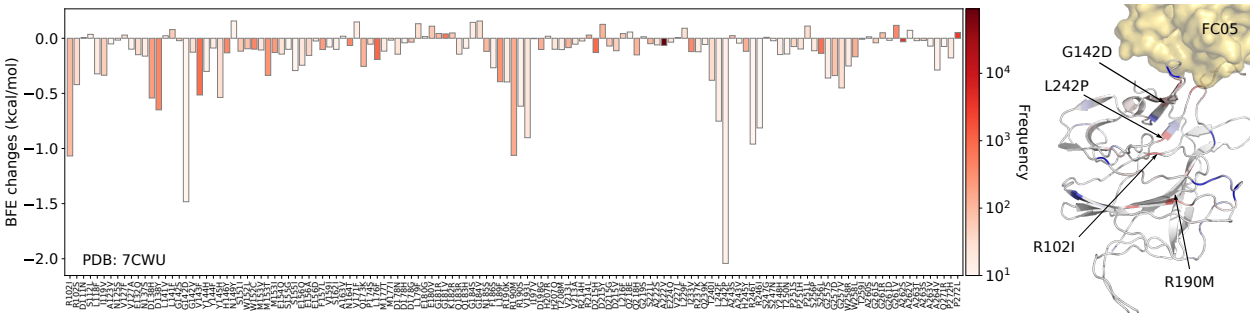

Figure S6: Illustration of SARS-CoV-2 mutation-induced binding free energy changes for the complexes of S protein and FC05 (PDB: 7CWU). Blue color in the structure plot indicates a positive BFE change while red color indicates a negative BFE change, and toning indicates the strength. Here, mutations R102I, G142D, R190M, and L242P could potentially disrupt the binding of antibody FC05 and the S protein.

Figure S6 illustrates the common binding complex of FC05 with the S protein NTD. A total 131 out of 501 mutations on residues ID from 14 to 226 have their frequencies larger than 10. Only 13 of these 131 mutations have their magnitudes of BFE changes large than 0.5 kcal/mol. In particular, the largest magnitude of binding-strengthening mutation has a BFE change of 0.16 kcal/mol. Moreover, 99 out of the 131 mutations have negative BFE changes, including R102I with the frequency of 89.

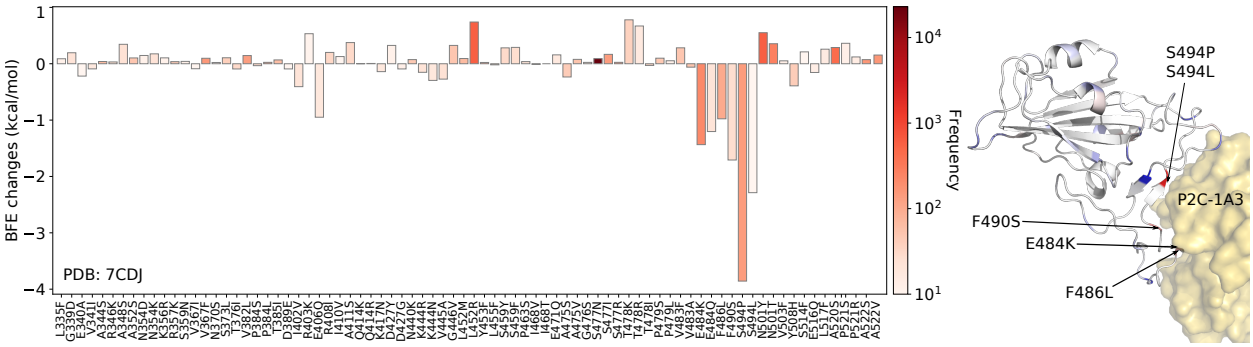

Figure S7: Illustration of SARS-CoV-2 mutation-induced binding free energy changes for the complexes of S protein and P2C-1A3 (PDB: 7CDJ). Blue color in the structure plot indicates a positive BFE change while red color indicates a negative BFE change, and toning indicates the strength. Here, mutations E383K, F486L, F490S, S494P, and S494L could potentially disrupt the binding of antibody P2C-1A3 and the S protein.

We also present in a detailed study of antibody P2C-1A3 because it can be disrupted by a relatively high frequency mutation S494P with a large negative BFE change. Figure S7 illustrates the mutation-induced BFE changes for antibody P2C-1A3 (PDB: 7CDJ), which also shares the binding domain with ACE2. Note that mutation S494P has a BFE change of -3.9 kcal/mol with a frequency of 123. This antibody has mild BFE changes outside the binding motif but dramatic negative changes at mutations on the binding motif.

Antibody	The number of residues inside the paratope	The number of random-coil residues inside the paratope	Percentage
BD-629	27	25	92.6 %
CB6	34	31	91.8 %
COVA2-04	32	31	96.9 %
CV30	29	28	96.6 %
CC12.1	38	36	94.7 %
CC12.3	26	25	96.2 %
BD-236	33	31	93.9 %
BD-368-2	18	17	94.4 %
BD-604	33	31	93.9 %
H11-H4	18	17	94.4 %
H11-D4	18	18	100.0 %
COVA2-39	17	17	100.0 %
H014	26	20	76.9 %
P2B-2F6	19	18	94.7 %
SR4	21	21	100.0 %
BD23	19	18	94.7 %
S309	21	17	81.0 %
CR3022	28	23	82.1 %
B38	34	32	94.1 %
Fab2-4	17	17	100.0 %
MR17	20	20	100.0 %
EY6A	27	23	85.2 %
Nb	17	16	94.1 %
S2H13	13	13	100.0 %
S2A4	19	18	94.7 %
S304	12	12	100.0 %
VH binder	26	25	96.2 %
S2H14	22	22	100.0 %
S2M11	18	18	100.0 %
CV07-250	22	22	100.0 %
CV07-270	22	21	95.5 %
SB23	12	12	100.0 %
P2C-1F11	24	23	95.8 %
P2C-1A3	17	17	100.0 %
A fab	33	32	97.0 %
COVA1-16	24	22	91.7 %
S2E12	16	15	93.8 %
Fab 52	19	16	84.2 %
Fab 298	13	13	100.0 %
C1A	34	34	100.0 %
STE90-C11	35	32	91.4 %
P17	14	14	100.0 %
4A8	16	14	87.5 %
FC05	15	15	100.0 %
2G12	24	21	87.5 %

Table S1: The random coil percentages of antibody paratopes on the S protein.

The South Africa variant E484K and mutations F486L, F490S, S494P, and S494L will reduce P2C-1A3's competitiveness with ACE2.

References

[1] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. GenBank. *Nucleic acids research*, 37(suppl_1):D26–D31, 2009.

[2] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[3] J. Chen, R. Wang, M. Wang, and G.-W. Wei. Mutations strengthened SARS-CoV-2 infectivity. *Journal of Molecular Biology*, 2020.

[4] X. Chi, R. Yan, J. Zhang, G. Zhang, Y. Zhang, M. Hao, Z. Zhang, P. Fan, Y. Dong, Y. Yang, et al. A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science*, 369(6504):650–655, 2020.

[5] W. L. DeLano et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography*, 40(1):82–92, 2002.

[6] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. In *Proceedings 41st annual symposium on foundations of computer science*, pages 454–463. IEEE, 2000.

[7] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, 2010.

[8] J. Jankauskaitė, B. Jiménez-García, J. Dapkūnas, J. Fernández-Recio, and I. H. Moal. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019.

[9] S. Jemimah, K. Yugandhar, and M. Michael Gromiha. PROXiMATE: a database of mutant protein–protein complex thermodynamics and kinetics. *Bioinformatics*, 33(17):2787–2788, 2017.

[10] G. Li, S. Pahari, A. Krishna Murthy, S. Liang, R. Fragoza, H. Yu, and E. Alexov. SAAMBE-SEQ: A Sequence-based Method for Predicting Mutation Effect on Protein-protein Binding Affinity. *Bioinformatics*, 2020.

[11] T. W. Linsky, R. Vergara, N. Codina, J. W. Nelson, M. J. Walker, W. Su, C. O. Barnes, T.-Y. Hsiang, K. Esser-Nobis, K. Yu, et al. De novo design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2. *Science*, 370(6521):1208–1214, 2020.

[12] Q. Liu, P. Chen, B. Wang, J. Zhang, and J. Li. dbMPIKT: a database of kinetic and thermodynamic mutant protein interactions. *Bmc Bioinformatics*, 19(1):1–7, 2018.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, 12:2825–2830, 2011.

[14] E. Procko. The sequence of human ACE2 is suboptimal for binding the S spike protein of SARS coronavirus 2. *BioRxiv*, 2020.

[15] C. H. Rodrigues, Y. Myung, D. E. Pires, and D. B. Ascher. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic acids research*, 47(W1):W338–W344, 2019.

[16] Y. Shu and J. McCauley. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13), 2017.

[17] F. Sievers and D. G. Higgins. Clustal omega, accurate alignment of very large numbers of sequences. In *Multiple sequence alignment methods*, pages 105–116. Springer, 2014.

[18] S. Sirin, J. R. Apgar, E. M. Bennett, and A. E. Keating. AB-Bind: antibody binding mutational database for computational affinity predictions. *Protein Science*, 25(2):393–409, 2016.

[19] T. N. Starr, A. J. Greaney, S. K. Hilton, D. Ellis, K. H. Crawford, A. S. Dingens, M. J. Navarro, J. E. Bowen, M. A. Tortorici, A. C. Walls, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, 182(5):1295–1310, 2020.

[20] J. D. Thompson, T. J. Gibson, and D. G. Higgins. Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics*, (1):2–3, 2003.

[21] M. Wang, Z. Cang, and G.-W. Wei. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence*, 2(2):116–123, 2020.

[22] S. Wang, W. Li, S. Liu, and J. Xu. RaptorX-Property: a web server for protein structure property prediction. *Nucleic acids research*, 44(W1):W430–W435, 2016.

[23] A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton. Jalview Version 2a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 2009.

[24] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, et al. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–269, 2020.

[25] K. Xia and G.-W. Wei. Persistent homology analysis of protein structure, flexibility, and folding. *International journal for numerical methods in biomedical engineering*, 30(8):814–844, 2014.