scientific reports

Check for updates

OPEN

# Potential neutralizing antibodies discovered for novel corona virus using machine learning

Rishikesh Magar[1], Prakarsh Yadav[2] & Amir Barati Farimani[1,2,3]✉

**The fast and untraceable virus mutations take lives of thousands of people before the immune system can produce the inhibitory antibody. The recent outbreak of COVID-19 infected and killed thousands of people in the world. Rapid methods in finding peptides or antibody sequences that can inhibit the viral epitopes of SARS-CoV-2 will save the life of thousands. To predict neutralizing antibodies for SARS-CoV-2 in a high-throughput manner, in this paper, we use different machine learning (ML) model to predict the possible inhibitory synthetic antibodies for SARS-CoV-2. We collected 1933 virus-antibody sequences and their clinical patient neutralization response and trained an ML model to predict the antibody response. Using graph featurization with variety of ML methods, like XGBoost, Random Forest, Multilayered Perceptron, Support Vector Machine and Logistic Regression, we screened thousands of hypothetical antibody sequences and found nine stable antibodies that potentially inhibit SARS-CoV-2. We combined bioinformatics, structural biology, and Molecular Dynamics (MD) simulations to verify the stability of the candidate antibodies that can inhibit SARS-CoV-2.**

The biomolecular process for recognition and neutralization of viral particles is through the process of viral antigen presentation and recruitment of appropriate B cells to synthesize the neutralizing antibodies[1]. Theoretically, this process allows the immune system to stop any viral invasion, but this response is slow and often requires days, even weeks before adequate immune response can be achieved[2,3]. This poses a challenging question: can the process of antibody discovery be computationally accelerated to counter highly infective viral diseases?

The general paradigm of computational antibody design revolves around doing complex Molecular dynamics (MD) simulations that are computationally expensive. The computational expense of MD simulations makes them inaccessible in scenarios like global pandemic when rapid solutions are needed that can be reliable and accurate. Thus, it is imperative to design and develop techniques that can aid the computational antibody discovery process. With the rapid expansion of available biological data, such as DNA/protein sequences and structures[4], machine learning (ML) approaches have been increasing used in modelling and predicting biological phenomenon[5,6]. Given sufficient training data, ML can be used to learn a mapping between the viral epitope and effectiveness of its complementary antibody. Once such mapping is learnt, it can be used to predict potentially neutralizing antibody for a given viral sequence enabling us to design novel antibodies[7]. Thus, enabling ML models to be used for high throughput screening of antibody sequences which is faster than traditional methods of computational protein design using MD simulations.

ML can learn the complex antigen–antibody interactions much faster than human immune system. This allows rapid generation of a library of synthetic inhibitory antibodies bridge, which can overcome the latency between viral infection and human immune system response. This bridge can potentially save the life of many people during the outbreak of novel viruses for which we lack treatment. One such instance is the spread of coronavirus disease (COVID-19)[8].

With incredibly high infectivity and mortality rate, COVID-19 has become a global scare[9,10]. Although the vaccines against COVID-19 are now available but there are no proven therapeutics, such as antibody serum, to aid the suffering patients[2,9,11–18]. Vaccines are a preventive measure to stop the spread of COVID-19, but do not have a therapeutic effect if a patient has been infected by SARS-CoV-2. Antibody serum based therapies can help patients after they have been infected by the SARS-CoV-2. Only viable treatment at the moment is symptomatic and there is a desperate need for developing therapeutics to counter COVID-19. Recently, the proteomics sequences of 'WH-Human 1' coronavirus became available through Metagenomic RNA sequencing of a patient in Wuhan[4,19]. WH-Human 1 is 89.1% similar to a group of SARS-like coronaviruses[4]. With the availability of

[1]Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [2]Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [3]Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. ✉email: barati@cmu.edu

this sequence, it is possible to find potential inhibitory antibodies by scanning thousands of antibody sequences and discovering the neutralizing ones[20–22]. However, this requires very expensive and time-consuming experimentation to discover the inhibitory responses to SARS-CoV-2 in a timely manner. In addition, computational and physics-based models require the bound crystal structure of antibody-antigen complex, however; only a few of these structures have become available[23–26]. Moreover, in case of SARS-CoV-2, the complex of viral antigen and neutralizing antibody is not available to-date[27,28]. Due to the lack of availability of structural data we aimed to develop an ML model which leverages the information in the antibody-antigen sequences rather than the structures to predict the potential neutralizing antibodies[29].

In this paper, we have collected a dataset comprised of antibody-antigen sequences of variety of viruses including HIV, Influenza, Dengue, SARS, Ebola, Hepatitis, etc. with their patient clinical/biochemical $IC_{50}$ data. Using this dataset (we call it VirusNet), we trained and benchmarked different shallow and deep ML models and selected the best performing model, to predict a set of potentially neutralizing antibodies. Based on SARS 2006 neutralizing antibody scaffold[30], we created thousands of antibody candidates by mutation and screened them with our best performing ML model to find the potentially neutralizing antibodies. Finally, molecular dynamics (MD) simulations were performed on the neutralizing candidates to check their structural stability. We predict nine structures that were stable over the course of simulation and are potential neutralizing antibodies for SARS-CoV-2. In addition, we interpreted the ML method to understand what alterations in the sequence of binding region of the antibody would most effectively counter the viral mutation(s) and restore the ability of the antibody to bind to the virus[31]. This information is critical in terms of antibody design and engineering in order to reduce the dimension of combinatoric mutations needed to find a neutralizing antibody.

This work highlights the merits of leveraging an ML based method for high-throughput discovery of neutralizing antibodies for viruses where only the sequences of viral coat protein-antibody pairs can be obtained. Moreover, this work also proposes a recipe for computational antibody design using ML approaches to work concurrently with the traditional molecular dynamics simulations-based approaches in order to augment each other. Through our computational approach we are able to leverage ML techniques to computationally design antibodies and also take advantage of accepted paradigm of molecular dynamics to validate our ML based approach.

## Methods

**Dataset.** The VirusNet dataset consists of 1933 samples spanning over 15 different types of viruses. Majority of the data in the training set is composed of HIV antibody-antigen complex as it widely studied and readily available. Most of the samples for the HIV training set were obtained from the Compile, Analyze and Tally NAb panels (CATNAP) database from the Los Alamos National Laboratory (LANL)[32,33]. From CATNAP, data was collected for monoclonal antibodies, 2F5, 4E10 and 10E8, which bind with GP41[34–36]. Using CATNAP's functionality for identifying epitope alignment, we selected FASTA sequence of the antigen corresponding to the site of alignment, in the antibody. We generated a dataset with 1831 training examples comprising of antibodies—antigen sequences and their corresponding $IC_{50}$ values. The CATNAP output is comprised of site of antigen sequence alignment for each of the antibodies with respect to 2F5, 4E10 and 10E8. Using the co-crystalized structure of (2F5-ELDKWAS) in (PDB:1TJG)[34] as template, the antibody fragment that comes in contact with the antigen was found by considering amino acids within 7 Å of the antigen in the co-crystallized structure.

To make the dataset more diverse and train a more robust ML model, we included more available antibody-antigen sequences and their neutralization potential. To do this, we compiled the sequences of Influenza, Dengue, Ebola, SARS, Hepatitis, etc.[30,37–90] by searching the keywords of "virus, antibody" on RCSB server[91] and selected the neutralizing complex by reading their corresponding publications. Furthermore, for each neutralizing complex, the contact residues at the interface of antibody and antigen were selected (Fig. S3). To select the antigen contact sequences, all amino acids within 5 Å of corresponding antibody were chosen (Figs. S4, S5). To select the antibody contact sequences, all amino acids within 5 Å of the antigen were chosen. In total, 102 sequences of antibody-antigen complexes were mined, comprising of structures from X-ray diffraction of crystal structure and Cryo-EM experiments, and added to the 1831 samples collected from CATNAP, resulting in total number of 1933 training samples.

**Graph featurization and machine learning.** For effective representation of molecular structure of amino acids, the individual atoms of amino acids of antibody and antigen were treated as undirected graph, where the atoms are nodes and bonds are edges[92]. In this work, we generate the representations of molecules from their respective molecular graphs. We construct these molecular graphs using RDkit[93]. Embeddings are generated to encode relevant features about the molecular graph[94,95]. These embeddings encode information like the type of atom, valency of an atom, hybridization state, aromaticity etc. (Table S3). First, each antibody and antigen were encoded into separate embeddings and then concatenated into a single embedding for the entire antibody-antigen complex. We then apply mean pooling over the features for this concatenated embedding to ensure dimensional consistency across the training data. The pooled information is then passed to classifier algorithms like XGBoost[96], Random Forest[97], Multilayer perceptron, Support Vector Machine (SVM)[98] and Logistic Regression which then predict whether the antibody is capable of neutralizing the virus. XGBoost is a gradient boosting framework which uses the second order derivative to approximate gradient to learn the features[96]. Random forest is an ensemble machine learning method as it uses multiple decision tress and selects the mode of these decision tress as the output[97]. Multilayer perceptron is a feedforward artificial neural network (ANN) which is composed of fully connected layers of perceptrons with an activation function. SVM is a machine learning algorithm which tries to learn the maximum-margin hyperplane to classify the data[98]. Logistic regression is the estimation of the parameters of a logistic model which is used to model the probability of different classes.

**Hypothetical antibody library generation.** In order to find potential antibody candidates for SARS-CoV-2, 2589 different mutant strains of antibody sequences were generated based on the sequence of SARS neutralizing antibodies. The reason we selected these antibodies as initial scaffolds is that the genome of SARS-CoV-2[4] is 79.8% identical to "Tor2" isolate of SARS (Accession number: AY274119)[99]. Exhaustive search of the RCSB PDB server concluded that 4 structures SARS (PDB: 2GHW, 3BGF, 6NB6, 2DD8) were the only SARS antigen and antibody complexes which have been reported till date. Using 4 different antibody variants of SARS[30,80,85,90], point mutations were applied to all the amino acids in the binding region of antibody. (See "Supporting Information" for SARS-CoV-2 antigen and antibody interactions.) To find out the binding region of these antibodies for sequence generation, all amino acids within 5 Å of their respective antigen were chosen. To assess the biological feasibility of these mutant sequences, we scored each mutation by using the BLOSUM62 matrix[100].

**Molecular dynamics simulations.** To assess the stability of proposed antibody structures, we performed molecular dynamics (MD) simulations of each of antibody structure in a solvated environment[101]. The simulation of solvated antibody was carried out using GROMACS-5.1.4[102–104], and topologies for each antibody were generated according the GROMOS 54a7[105] forcefield. The protein was centered in a box, extending 1 nm from surface of the protein. This box was the solvated by the SPC216 model water atoms, pre-equilibrated at 300 K. The antibody system in general carried a net positive charge and it was neutralized by the counter ions. Energy minimization was carried out using steepest descent algorithm, while restraining the peptide backbone to remove the steric clashes in atoms and subsequently optimize solvent molecule geometry. The cut-off distance criteria for this minimization were forces less than 100.0 kJ/mol/nm or number of steps exceeding 50,000. This minimized structure was the sent to two rounds of equilibration at 300 K. First, an NVT ensemble for 50 picoseconds and a 2-femtosecond time step. Leapfrog dynamics integrator was used with Verlet scheme, neighbor-list was updated every ten steps. All the ensembles were under Periodic Boundary Conditions and harmonic constraints were applied by the LINCS algorithm[106]; under this scheme the long-range electrostatic interactions were computed by Particle Mesh Ewald (PME) algorithm[107]. Berendsen thermostat was used for temperature coupling and pressure coupling was done using the Parrinello-Rahman barostat[108,109]. The last round of NPT simulation ensures that the simulated system is at physiological temperature and pressure. The system volume was free to change in the NPT ensemble but in fact did not change significantly during the course of the simulation. Following the rounds of equilibration, production run for the system was carried out in NPT and no constraints for a total of 150 ns, under identical simulation parameters.

## Results and discussions

The flowchart of SARS-CoV-2 antibody discovery using ML has four major steps (Fig. 1): (1) collecting data and curating the dataset for training set. (2) Featurization, embedding and benchmarking ML models and selecting the best performing one. (3) Hypothetical antibody library generation and ML screening for neutralization and (4) checking the stability of proposed antibodies. This workflow enables the rapid screening of large space of potential antibodies to neutralize COVID-19. In general, this workflow can be used for high throughput screening of antibodies for any type of virus by only knowing the sequences of antigen epitopes.

To better understand the diversity and similarity of the sequences that were used in the training set, we project the graph embeddings encoding the fingerprints of the molecules in the t-Distributed Stochastic Neighbor Embedding (t-SNE) space (Fig. 2a). t-SNE axes shows the directions of the maximum variance in the feature space of the dataset, therefore, the dimensionality of the data can be reduced to lower dimensions (here two). HIV antigen shows the most variations on t-SNE components where viruses such as Influenza, Dengue and H1N1 are very close to each other. The neutralizing antibodies were also projected using t-SNE to show the variations in the available neutralizing sequences (Fig. 2b). Unlike antigen variations, antibody sequences are much closer to the center of t-SNE with a few scattered ones. The comparison of Fig. 2a,b shows that the neutralizing antibodies are not sequence-diverse compared to virus antigens. This difference demonstrates that a large space of potential antibodies can be screened and used for finding novel antibodies. The labels in the dataset are comprised of the neutralization panel data, $IC_{50}$ values for monoclonal antibodies and pseudo-typed viruses (Fig. 2c). The $IC_{50}$ labels were collected from 49 published neutralization studies and were collected from CATNAP Database (for 1831 samples in our training set). For some cases in CATNAP, personal communication with the authors were made to resolve sequence name ambiguities between different laboratories. For 102 samples of various viruses collected from RCSB server, all of them neutralize their antigen based on biochemical assays. These samples were labeled by setting their $IC_{50}$ to zero. Since classification is performed on the training dataset, $IC_{50} \leq 10$ are set to neutralizing and $IC_{50} > 10$ to non-neutralizing (Fig. 2c). To visualize the diversity of the virus types used in the dataset other than HIV, the distribution of 13 more viruses were presented in Fig. 2d. Influenza, Dengue, SARS, Ebola and then Hepatitis have relatively larger samples in the dataset.

To benchmark the performance of different ML models on the VirusNet dataset and select the best performing one, XGBoost, Random Forest (RF), Multilayer perceptron (MLP), Support Vector Machines (SVM), and Logistic Regression (LR) were used (Fig. 3a). The five-fold cross validation on 80–20% split, train, and test was implemented and best accuracy was observed for XG-Boost model. The performance and ranking of models follow the order of XGBoost (90.57%) > RF (89.18%) > LR (81.17%) > MLP (78.23) > SVM (75.49%). Since the featurized training data is sparse in the case of VirusNet (see Fig. 2a,b), XGBoost selects the sparse features input by pruning and learning the underlying sparsity patterns. In order to augment the accuracy as a performance metric we have also added ROC-AUC score as a performance metric in the "Supporting Information" (See Fig. S6). To test the robustness of the XGBoost on completely unseen virus types, for each left-out virus type, the model was trained on all the sequences in the VirusNet except for the left-out. For example, for Influenza, all the sequences

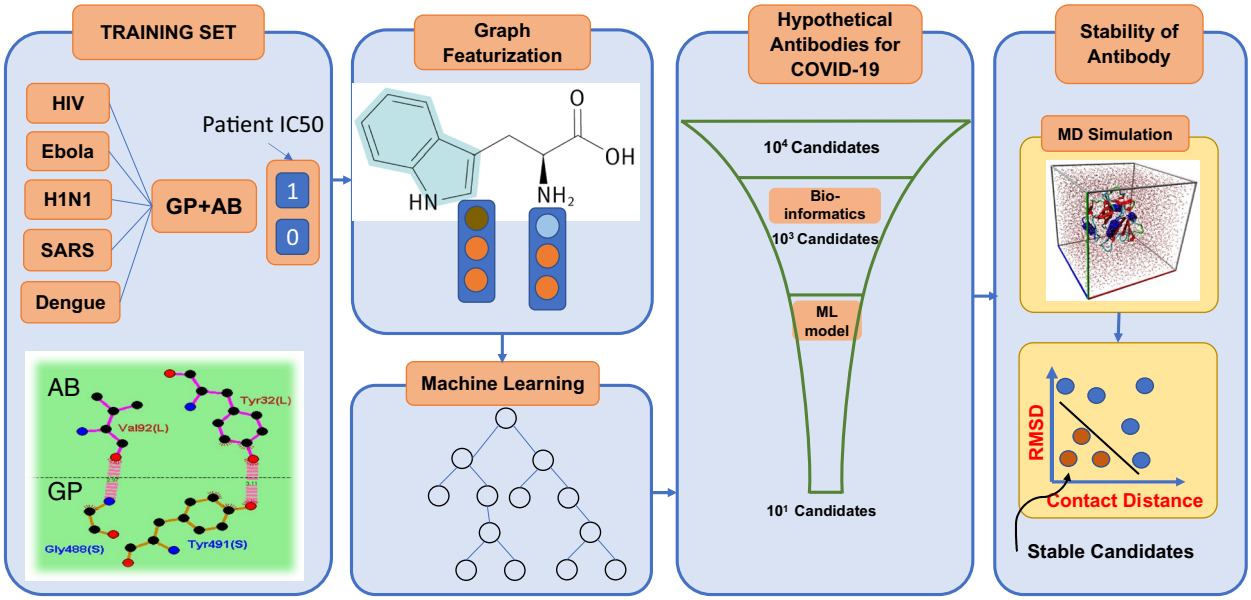


**Figure 1.** Designing antibodies or peptide sequences that can inhibit the SARS-CoV-2 virus requires high throughput experimentation of vastly mutated sequences of potential inhibitors. The screening of thousands of available strains of antibodies are prohibitively expensive, and not feasible due to lack of available structures. However, machine learning models can enable the rapid and inexpensive exploration of vast sequence space on the computer in a fraction of seconds. We collected 1933 virus-antibody sequences with clinical patient $IC_{50}$ data. Graph featurization of antibody-antigen sequences creates a unique molecular representation. Using graph representation, we benchmarked and used a variety of shallow and deep learning models and selected XGBoost because of its superior performance and interpretability. We trained our model using a dataset including 1933 diverse virus epitope and the antibodies. To generate the hypothetical antibody library, we mutated the SARS scaffold antibody of 2006 (PDB:2GHW) and generated thousands of possible candidates. Using the ML model, we classified these sequences and selected the top 18 sequences that will neutralize SARS-CoV-2 with high confidence. We used MD simulations to check the stability of the 18 sequences and rank them based on their stability.

and labels of Influenza were removed from the training set and the trained model on the remaining dataset were tested on all the Influenzas's sequences and consequently the classification accuracy were reported (Fig. 3b). The accuracies for the out of class test is as follows: Influenza (84.61%), Dengue (100%), Ebola (75%), Hepatitis (75%), SARS (100%). From these results, we can conclude that our model performance will be reliable based on the accuracies for out-of-class prediction. The fact that our model prediction is highly accurate for various out of class tests, demonstrate its capability of effectively predicting the antibodies for novel SARS-CoV-2.

Next, using the best performing model (XGBoost), all the hypothetical candidates in the library were evaluated for neutralization. Out of all the candidates, some of them are invalid mutations screened using BLOSUM62 matrix[100] (Fig. 3c). 18 final candidates are both valid mutations and can neutralize SARS-CoV-2 with high confidence probability of 0.9895 as per the ML model are then selected for MD screening (shown with green color in Fig. 3c).

A recent study reports that antibodies which effectively neutralized the previous SARS strains are not able to neutralize WH-Human 1[110]. However, the study also reports that there is "presence of a conserved immunogenic epitope among different Corona viruses". Therefore, we had generated mutant and co-mutant sequences to create a diverse set of antibodies which we could screen through the ML model. The ML model we have developed uses the antigen–antibody interactions and tries to learn the structure-based mapping between the amino acids involved at interaction surface. This was the rationale behind including viruses from various other species as well in the dataset used to train the ML model. The dataset is sufficiently diverse so that the ML model can learn the structurally important features from variety of viruses, Fig. 2A, in addition to the sequence dependent information. This antigen diversity allowed us to overcome the constraint of dissimilarity in WH-Human 1 and SARS ACE2 receptor and yet make accurate predictions.

Interpretability of the ML models is very important in both explaining the underlying biological and chemical understanding of neutralization and providing design guidelines for antibody engineering. One of the significant advantages of ensemble methods such as XGBoost is their interpretability. By taking advantage of this property, the important features that are giving rise to neutralization were ranked and scored (Fig. 3d). The input features to the model contains atomic level attributes such as atom type, valency, hybridization, etc. To collectively translate the important atomic features into important amino acid features, the scores of amino acids with unique atomic features were summed up and ranked (Fig. 3d). Some of the atomic features were common among all the amino acids (e.g. Carbon, implicit valency, Oxygen, etc.) therefore; we ignored them. However,

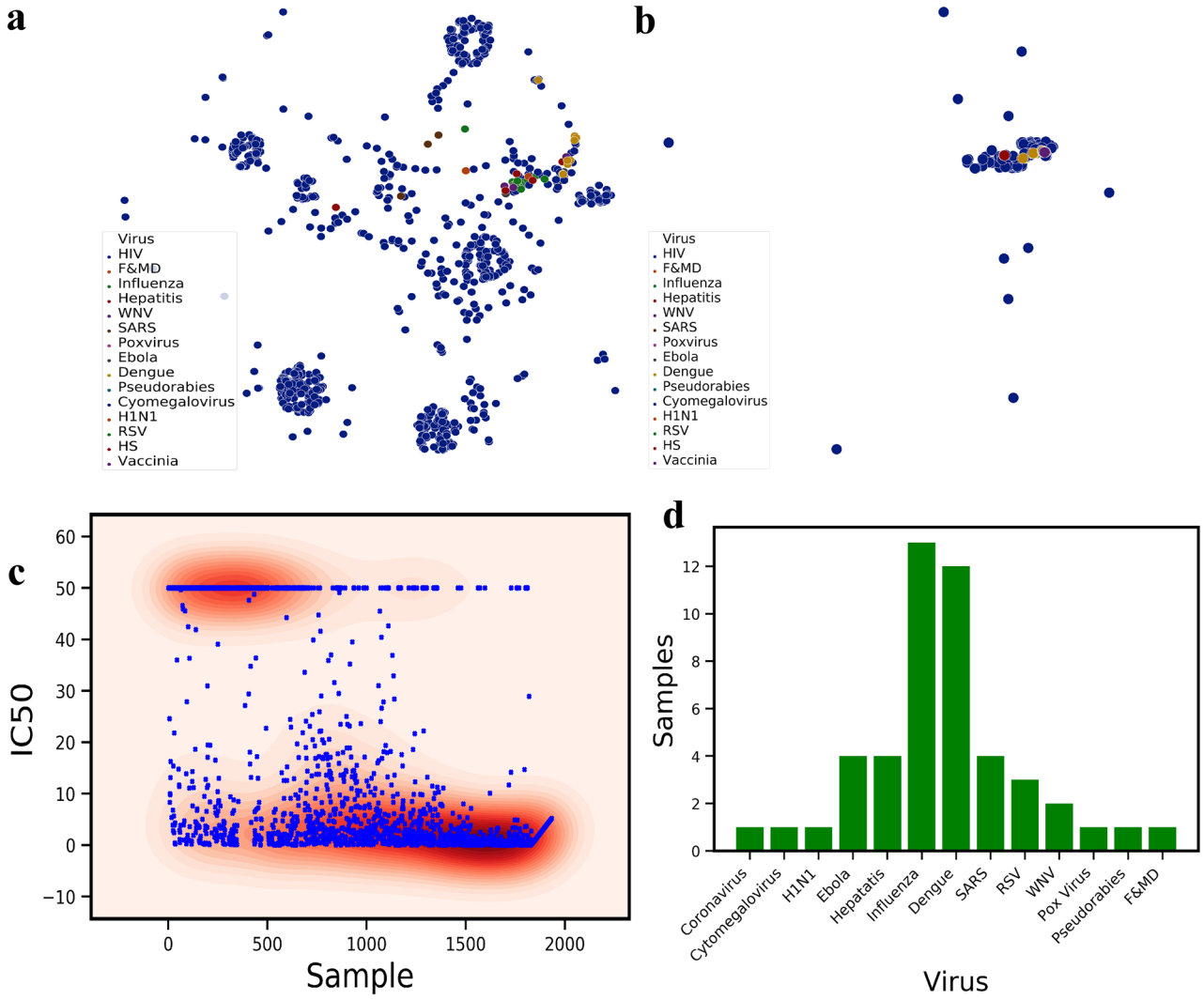


**Figure 2.** (**a**) t-Distributed Stochastic Neighbor Embedding (t-SNE) of all the viruses epitopes used in the training dataset, revealing biological similarity and diversity of the sequences used in the dataset. (**b**) t-SNE of all the therapeutics antibody sequences used in the training set for variety of different virus types. The majority of the broadly neutralizing antibodies such as 2F5 is clustered at the center of this plot. (**c**) Patient clinical IC50 data obtained from various sources and the distribution of the neutralizing ($IC_{50} < 10$) and Non-neutralizing ($IC_{50} > 10$) samples. (**d**) The number of samples for each virus class except HIV. For HIV, we collected 1883 samples. Influenza and Dengue has 10+ samples.

some of the other features like aromaticity or having Sulfur are unique and we considered those in amino acid features. Through this criteria some of the important mutations that we obtain include Cysteine, Methionine, Tyrosine, Phenylalanine and Tryptophan. Mutations to cysteine we concluded to not be viable as introduction of additional cysteine to the antibody structure, which heavily relies on disulfide bridges, would be detrimental as it can cause misfolding of the antibody structure. Further validation of this was done by the BLOSUM62 matrix, which put very heavy penalties on mutations which convert amino acids to cysteine. These observations cumulatively led us to the conclusion that Methionine is an important amino acid for antibody interface whereas cysteine is not. Methionine is known to be playing a crucial role in antigen recognition by antibody and further protein–protein interaction[111,112]. In addition, oxidative damage to Methionine is reported to have negatively impact the pharmacokinetic properties of antibodies[113]. This information validated the features learnt by the ML model, allowing us to definitively conclude that Methionine is indeed one of the important amino acids in antigen recognition by antibodies (Fig. 3d).

To validate the biological feasibility of the ML model predicted antibodies, we assessed the stability of the predicted antibody by Molecular Dynamics (MD) simulations. We assessed the antibodies based on two criteria's Root Mean Square Deviation (RMSD) and Mean Contact Distance[26]. RMSD is a measure of the deviation in the structure of the protein over the course of the simulation, a higher RMSD indicates that the structure is changing with respect to the initial structure for simulation. Contact Distance is the distance between the interacting amino acids of the protein, a higher Mean Contact Distance is indicative of an unstable protein as the amino acids are moving further apart. The combination of these features from simulation data of potentially neutralizing

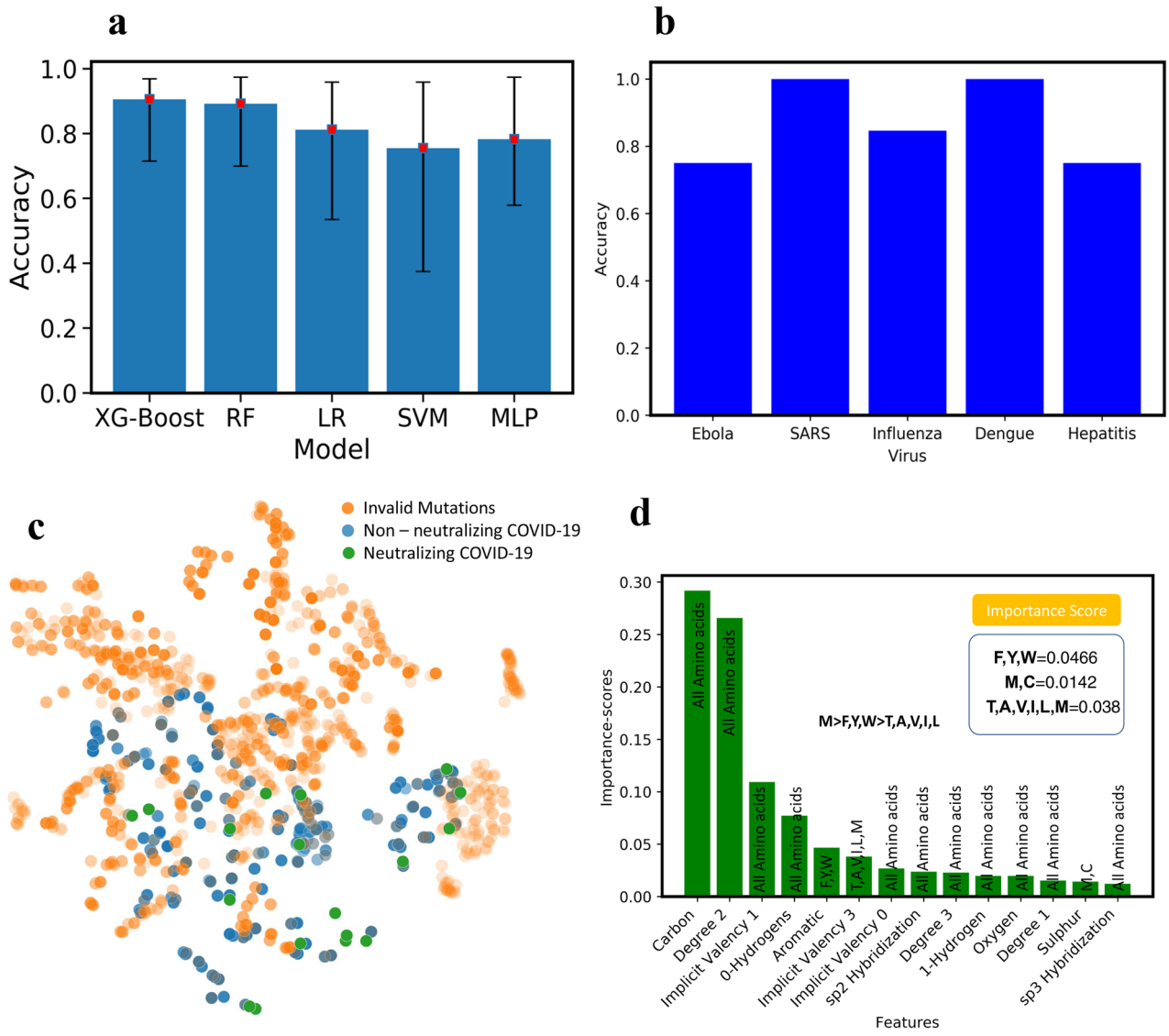


**Figure 3.** (**a**) The test accuracy with five-fold cross validation for XG-Boost, Random Forrest (RF), Logistic Regression (LR), Support Vector Machine (SVM) and Deep Learning (Multilayer Perceptron. XGBoost has the highest performance with (90.75%). (**b**) Out of training class test accuracy for influenza, Dengue, Ebola, Hepatitis, and SARS. To perform this test, for example for influenza, all the influenza virus-antibody sequences were removed from the training set and the obtained model were tested on all samples of Influenza and the accuracy is reported here. (**c**) Blosum validated mutations, non-neutralizing and neutralizing antibody sequences. To achieve more confidence, we set the threshold of prediction probability to 0.9895 in XGBoost and found 18 neutralizing antibody sequences (the green points). (**d**) Interpretability of ML model: to understand what mutations are playing the key roles in neutralization, XGBoost feature importance used with ranked atomic level features. Through connecting the atomic features with each of 20 amino acids, M was found to be the most important amino acids in neutralization followed by F, Y, W. The ML model predicted the presence of hydrophobicity and Sulfur as an important feature in antibody-antigen interaction. We concluded that M was the most important amino acid as it has both the characteristics of hydrophobicity and the presence of Sulfur.

antibodies allowed us to validate their stability and select most stable candidates. The predicted sequences from the ML model were then used to model the novel structure of potentially neutralizing antibodies. The predicted sequences were projected onto their progenitor antibody and the changes in amino acid sequence were modelled as follows: simple point mutations were introduced by modifying the target amino acid using Coot[114,115] (Crystallographic Object-Oriented Toolkit). Coot environment allowed us to predict the stereochemical effect of each point mutation and appropriately compensate for it. Using such an approach, we were able to accurately model the putative structures of the antibodies. The modelled structures were then passed to MD simulations for stability check.

To check the stability of predicted structures energetically, 20 MD simulations (18 point mutations + 2 wildtype (WT)) in total were performed (Fig. 4a). Structures with low Root Mean Square Deviation (RMSD) and low contact distance are in a stable conformation, whereas structures with high RMSD and high contact

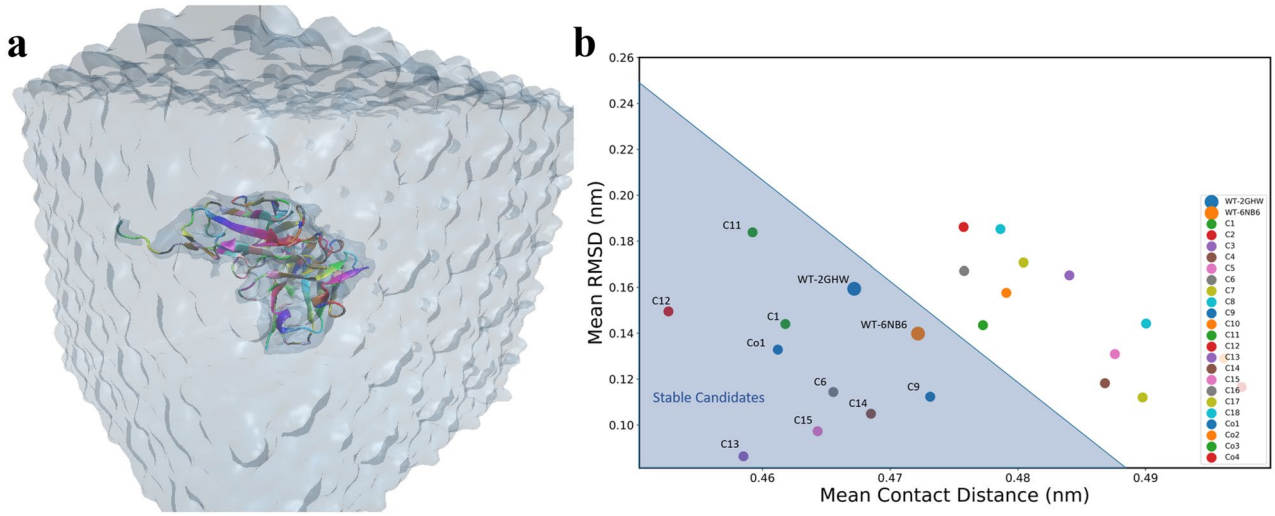


**Figure 4.** (**a**) The snapshot of MD simulation of mutated proteins. Each protein is solvated in a box of water and simulated to collect the statistical data on the stability of mutants and co-mutants. (**b**) Mean Root Mean Square Deviation (RMSD) versus Mean contact distances for each candidate averaged over the whole trajectory.

| Structure | Mutation |
|---|---|
| C1 | 2GHW-A33C |
| C6 | 2GHW-R100H |
| C9 | 2GHW-R162H |
| C11 | 2GHW-T285N |
| C12 | 2GHW-R286H |
| C13 | 6NB6-F203M |
| C14 | 2GHW-T204N |
| C15 | 2GHW-T206N |
| Co1 | 6NB6-I51M, R150H, T204N |

**Table 1.** The final neutralizing candidates obtained through screening with ML model, MD simulation for stability and bioinformatics. The detailed list of sequences is available in the "Supporting Information".

distances are in an unstable conformation. RMSD (Fig. S2) and contact distance (Fig. S3) for WT structures have lower values, demonstrating stability, therefore; the contact distances versus RMSD is a good indicator of the stability of a protein over the course of a simulation (Fig. 4b).

Once mutation introduced in the crystallographic structure, it will cause it to deviate from WT structure's RMSD and contact distance. We performed simulations for all the 18 point-mutant structures (Table S1) and their mean contact distance versus RMSD[116,117] were computed for their respective trajectories (Fig. 4b) (see "Supporting Information"). Based on the two WT structures mean RMSD and contact distances, we selected the mutations which have mean contact distance and RMSD values less than 0.488 nm and 0.25 nm, respectively (the shaded triangle region in Fig. 4b). Candidates with higher values of mean RMSD and contact distances are unstable and will potentially fail to neutralize the SARS-CoV-2.

In order to be more comprehensive and take into account the effect of co-mutation, we created 5 co-mutations that are listed in Table S2 (Co1, Co2, Co3, Co4, Co5). These five co-mutations were screened using XGBoost for neutralization. Among all five co-mutations, Co5 did not neutralize. To check the stability of these four neutralizing co-mutations, MD simulations were performed and Co1 was found to be stable (Fig. 4b). The list of the final nine stable mutations and co-mutations are tabulated in Table 1 and the PDB structures are available in PDB format as "Supporting Information".

## Conclusion

We have developed a machine learning model for high throughput screening of synthetic antibodies to discover antibodies that potentially inhibit SARS-CoV-2. Our approach can be widely applied to other viruses where only the sequences of viral coat protein-antibody pairs can be obtained. The ML models were trained on 14 different virus types and achieved over 90% fivefold test accuracy. The out of class prediction is 100% for SARS and 84.61% for Influenza, demonstrating the power of our model for neutralization prediction of antibodies for novel viruses like COVID-19. Using this model, the neutralization of thousands of hypothetical antibodies was predicted, and 18 antibodies were found to be highly efficient in neutralizing SARS-CoV-2. Using MD simulations, the stability of predicted antibodies were checked and nine stable antibodies were found that can

neutralize SARS-CoV-2. In addition, the interpretation of ML model revealed that mutating to Methionine and Tyrosine is highly efficient in enhancing the affinity of antibodies to SARS-CoV-2. Further validation of the predicted antibodies can be carried out by future work involving in vitro experiments to assess the efficacy of the predicted antibodies at neutralizing the SARS-CoV-2 virus. In our work we assume only point mutations in the antibody sequence of SARS-CoV-1 when generating the potential antibody candidates for SARS-CoV-2, it is possible that the sequences have multiple point mutations and many different combinatorics. We would like to investigate such excluded combinations in the future and create a comprehensive dataset and a more robust protocol for discovering neutralizing antibodies.

Received: 20 May 2020; Accepted: 17 February 2021
Published online: 04 March 2021

## References

1. Dörner, T. & Radbruch, A. Antibodies and B cell memory in viral immunity. *Immunity* **27**(3), 384–392. https://doi.org/10.1016/j.immuni.2007.09.002 (2007).
2. Li, Z. *et al.* Development and clinical application of a rapid IgM–IgG combined antibody test for SARS-CoV-2 infection diagnosis. *J. Med. Virol.* https://doi.org/10.1002/jmv.25727 (2020).
3. Hewitt, E. W. The MHC class I antigen presentation pathway: Strategies for viral immune evasion. *Immunology* **110**(2), 163–169. https://doi.org/10.1046/j.1365-2567.2003.01738.x (2003).
4. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* https://doi.org/10.1038/s41586-020-2008-3 (2020).
5. Ardabili, S. F. *et al.* COVID-19 outbreak prediction with machine learning. *medRxiv.* https://doi.org/10.1101/2020.04.17.20070094 (2020).
6. Pinter, G., Felde, I., Mosavi, A., Ghamisi, P. & Gloaguen, R. COVID-19 pandemic prediction for hungary; a hybrid machine learning approach. *Mathematics* **8**(6), 890. https://doi.org/10.3390/math8060890 (2020).
7. Continuous cultures of fused cells secreting antibody of predefined specificity|Nature. https://www.nature.com/articles/256495a0 (Accessed 9 Mar 2020).
8. The species severe acute respiratory syndrome-related coronavirus: Classifying 2019-NCoV and naming it SARS-CoV-2. *Nat. Microbiol.* (2020). https://doi.org/https://doi.org/10.1038/s41564-020-0695-z.
9. Fu, Y., Cheng, Y. & Wu, Y. Understanding SARS-CoV-2-mediated inflammatory responses: from mechanisms to potential therapeutic tools. *Virol. Sin.* https://doi.org/10.1007/s12250-020-00207-4 (2020).
10. Secondary attack rate and superspreading events for SARS-CoV-2—The Lancet. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30462-1/fulltext. (Accessed 9 Mar 2020).
11. Cao, Y. *et al.* Comparative genetic analysis of the novel coronavirus (2019-NCoV/SARS-CoV-2) receptor ACE2 in different populations. *Cell Discov.* **6**(1), 1–4. https://doi.org/10.1038/s41421-020-0147-1 (2020).
12. Chang, Y.-C., Tung, Y.-A., Lee, K.-H., Chen, T.-F., Hsiao, Y.-C., Chang, H.-C., Hsieh, T.-T., Su, C.-H., Wang, S.-S., Yu, J.-Y., Shih, S., Lin, Y.-H., Lin, Y.-H., Tu, Y.-C.E., Tung, C.-W., Chen, C.-Y. Potential therapeutic agents for COVID-19 based on the analysis of protease and RNA polymerase docking. (2020). https://doi.org/10.20944/preprints202002.0242.v1.
13. Hoffmann, M. *et al.* SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* https://doi.org/10.1016/j.cell.2020.02.052 (2020).
14. Nguyen, T. M., Zhang, Y. & Pandolfi, P. P. Virus against virus: A potential treatment for 2019-NCov (SARS-CoV-2) and other RNA viruses. *Cell Res.* **30**(3), 189–190. https://doi.org/10.1038/s41422-020-0290-0 (2020).
15. Tian, X. *et al.* Potent binding of 2019 Novel Coronavirus Spike Protein by a SARS coronavirus-specific human monoclonal antibody. *Emerg. Microbes Infect.* **9**(1), 382–385. https://doi.org/10.1080/22221751.2020.1729069 (2020).
16. Jackson, L. A. *et al.* An MRNA Vaccine against SARS-CoV-2 — Preliminary Report. *N. Engl. J. Med.* **383**(20), 1920–1931. https://doi.org/10.1056/NEJMoa2022483 (2020).
17. Folegatti, P. M. *et al.* Safety and Immunogenicity of the ChAdOx1 NCoV-19 Vaccine against SARS-CoV-2: A Preliminary Report of a Phase 1/2, Single-Blind. *Randomised Controlled Trial. Lancet* **396**(10249), 467–478. https://doi.org/10.1016/S0140-6736(20)31604-4 (2020).
18. Polack, F. P. *et al.* Safety and efficacy of the BNT162b2 MRNA COVID-19 vaccine. *N. Engl. J. Med.* **383**(27), 2603–2615. https://doi.org/10.1056/NEJMoa2034577 (2020).
19. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet* **395**(10224), 565–574. https://doi.org/10.1016/S0140-6736(20)30251-8 (2020).
20. Karthick, V. *et al.* Virtual screening of the inhibitors targeting at the viral protein 40 of ebola virus. *Infect. Dis. Poverty* https://doi.org/10.1186/s40249-016-0105-1 (2016).
21. Miller, C. R. *et al.* Initiating a watch list for ebola virus antibody escape mutations. *PeerJ* **4**, e1674. https://doi.org/10.7717/peerj.1674 (2016).
22. Murin, C. D. *et al.* Structures of protective antibodies reveal sites of vulnerability on ebola virus. *Proc. Natl. Acad. Sci.* **111**(48), 17182–17187. https://doi.org/10.1073/pnas.1414164111 (2014).
23. Tiller, K. E. & Tessier, P. M. Advances in antibody design. *Annu. Rev. Biomed. Eng.* **17**(1), 191–216. https://doi.org/10.1146/annurev-bioeng-071114-040733 (2015).
24. Computational predictions of protein structures associated with COVID-19/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19. (Accessed 9 Mar 2020).
25. Jakobsson, E., Farimani, A. B., Tajkhorshid, E. & Aluru, N. Combining physics-based and evolution-based methods to design antibodies against an evolving virus. *Biophys. J.* **118**(3), 482a. https://doi.org/10.1016/j.bpj.2019.11.2669 (2020).
26. Barati Farimani, A., Aluru, N. R., Tajkhorshid, E. & Jakobsson, E. Computational approach to designing antibody for ebola virus. *Biophys. J.* **110**(3), 537a. https://doi.org/10.1016/j.bpj.2015.11.2877 (2016).
27. Yan, R. *et al.* Structural basis for the recognition of the SARS-CoV-2 by full-length human ACE2. *Science* https://doi.org/10.1126/science.abb2762 (2020).
28. Novel antibody epitopes dominate the antigenicity of spike glycoprotein in SARS-CoV-2 compared to SARS-CoV|Cellular & Molecular Immunology. https://www.nature.com/articles/s41423-020-0385-z. (Accessed 9 Mar 2020).
29. JCI Insight—Predicting the broadly neutralizing antibody susceptibility of the HIV reservoir. https://insight.jci.org/articles/view/130153. (Accessed 13 Mar 2020).
30. Hwang, W. C. *et al.* Structural basis of neutralization by a human anti-severe acute respiratory syndrome spike protein antibody, 80R. *J. Biol. Chem.* **281**(45), 34610–34616. https://doi.org/10.1074/jbc.M603275200 (2006).

31. Zhang, C.; Zheng, W.; Huang, X.; Bell, E. W.; Zhou, X.; Zhang, Y. Protein structure and sequence re-analysis of 2019-NCoV genome does not indicate snakes as its intermediate host or the unique similarity between its spike protein insertions and HIV–. 11.
32. Yoon, H. *et al.* CATNAP: a tool to compile, analyze and tally neutralizing antibody panels. *Nucleic Acids Res.* **43**(W1), W213–W219. https://doi.org/10.1093/nar/gkv404 (2015).
33. CATNAP Tools. https://www.hiv.lanl.gov/components/sequence/HIV/neutralization/. (Accessed 10 Mar 2020).
34. Ofek, G. *et al.* Structure and mechanistic analysis of the anti-human immunodeficiency virus type 1 antibody 2F5 in complex with its Gp41 epitope. *J. Virol.* **78**(19), 10724–10737. https://doi.org/10.1128/JVI.78.19.10724-10737.2004 (2004).
35. Kwon, Y. D. *et al.* Optimization of the solubility of HIV-1-neutralizing antibody 10E8 through somatic variation and structure-based design. *J. Virol.* **90**(13), 5899–5914. https://doi.org/10.1128/JVI.03246-15 (2016).
36. Irimia, A., Sarkar, A., Stanfield, R. L. & Wilson, I. A. Crystallographic identification of lipid as an integral component of the epitope of HIV broadly neutralizing antibody 4E10. *Immunity* **44**(1), 21–31. https://doi.org/10.1016/j.immuni.2015.12.001 (2016).
37. Fleury, D. *et al.* A complex of influenza hemagglutinin with a neutralizing antibody that binds outside the virus receptor binding site. *Nat. Struct. Biol.* **6**(6), 530–534. https://doi.org/10.1038/9299 (1999).
38. Pejchal, R. *et al.* A conformational switch in human immunodeficiency virus Gp41 revealed by the structures of overlapping epitopes recognized by neutralizing antibodies. *J. Virol.* **83**(17), 8451–8462. https://doi.org/10.1128/JVI.00685-09 (2009).
39. Ekiert, D. C. *et al.* A highly conserved neutralizing epitope on group 2 influenza a viruses. *Science* **333**(6044), 843–850. https://doi.org/10.1126/science.1204839 (2011).
40. Ochoa, W. F. *et al.* A multiply substituted G-H loop from foot-and-mouth disease virus in complex with a neutralizing antibody: A role for water molecules. *J. Gen. Virol.* **81**(6), 1495–1505. https://doi.org/10.1099/0022-1317-81-6-1495 (2000).
41. Corti, D. *et al.* A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza a hemagglutinins. *Science* **333**(6044), 850–856. https://doi.org/10.1126/science.1205669 (2011).
42. Pejchal, R. *et al.* A potent and broad neutralizing antibody recognizes and penetrates the HIV glycan shield. *Science* **334**(6059), 1097–1103. https://doi.org/10.1126/science.1213256 (2011).
43. Dias, J. M. *et al.* A shared structural solution for neutralizing ebolaviruses. *Nat. Struct. Mol. Biol.* **18**(12), 1424–1427. https://doi.org/10.1038/nsmb.2150 (2011).
44. Barbey-Martin, C. *et al.* An antibody that prevents the hemagglutinin low PH fusogenic transition. *Virology* **294**(1), 70–74. https://doi.org/10.1006/viro.2001.1320 (2002).
45. Venkatramani, L. *et al.* An epidemiologically significant epitope of a 1998 human influenza virus neuraminidase forms a highly hydrated interface in the NA–antibody complex. *J. Mol. Biol.* **356**(3), 651–663. https://doi.org/10.1016/j.jmb.2005.11.061 (2006).
46. Ekiert, D. C. *et al.* Antibody recognition of a highly conserved influenza virus epitope. *Science* **324**(5924), 246–251. https://doi.org/10.1126/science.1171491 (2009).
47. Fleury, D., Wharton, S. A., Skehel, J. J., Knossow, M. & Bizebard, T. Antigen distortion allows influenza virus to escape neutralization. *Nat. Struct. Biol.* **5**(2), 119–123. https://doi.org/10.1038/nsb0298-119 (1998).
48. Lok, S.-M. *et al.* Binding of a neutralizing antibody to dengue virus alters the arrangement of surface glycoproteins. *Nat. Struct. Mol. Biol.* **15**(3), 312–317. https://doi.org/10.1038/nsmb.1382 (2008).
49. Chi, S.-W. *et al.* Broadly neutralizing anti-hepatitis B virus antibody reveals a complementarity determining region H3 lid-opening mechanism. *Proc. Natl. Acad. Sci.* **104**(22), 9230–9235. https://doi.org/10.1073/pnas.0701279104 (2007).
50. Lee, J. E. *et al.* Complex of a protective antibody with its ebola virus GP peptide epitope: Unusual features of a Vλx light chain. *J. Mol. Biol.* **375**(1), 202–216. https://doi.org/10.1016/j.jmb.2007.10.017 (2008).
51. Azoitei, M. L. *et al.* Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold. *Science* **334**(6054), 373–376. https://doi.org/10.1126/science.1209368 (2011).
52. Ekiert, D. C. *et al.* Cross-neutralization of influenza a viruses mediated by a single antibody loop. *Nature* **489**(7417), 526–532. https://doi.org/10.1038/nature11414 (2012).
53. Lyumkis, D. *et al.* Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer. *Science* **342**(6165), 1484–1490. https://doi.org/10.1126/science.1245627 (2013).
54. Ménez, R. *et al.* Crystal structure of a hydrophobic immunodominant antigenic site on hepatitis C virus core protein complexed to monoclonal antibody 19D9D6. *J. Immunol.* **170**(4), 1917–1924. https://doi.org/10.4049/jimmunol.170.4.1917 (2003).
55. Rini, J. M. *et al.* Crystal structure of a human immunodeficiency virus type 1 neutralizing antibody, 50.1, in complex with its V3 loop peptide antigen. *Proc. Natl. Acad. Sci.* **90**(13), 6325–6329. https://doi.org/10.1073/pnas.90.13.6325 (1993).
56. Momany, C. *et al.* Crystal structure of dimeric HIV-1 capsid protein. *Nat. Struct. Biol.* **3**(9), 763–770. https://doi.org/10.1038/nsb0996-763 (1996).
57. Stanfield, R. L., Gorny, M. K., Zolla-Pazner, S. & Wilson, I. A. Crystal structures of human immunodeficiency virus type 1 (HIV-1) neutralizing antibody 2219 in complex with three different V3 peptides reveal a new binding mode for HIV-1 cross-reactivity. *J. Virol.* **80**(12), 6093–6105. https://doi.org/10.1128/JVI.00205-06 (2006).
58. Bryson, S., Julien, J.-P., Hynes, R. C. & Pai, E. F. Crystallographic definition of the epitope promiscuity of the broadly neutralizing anti-human immunodeficiency virus type 1 antibody 2F5: Vaccine design implications. *J. Virol.* **83**(22), 11862–11875. https://doi.org/10.1128/JVI.01604-09 (2009).
59. McLellan, J. S. *et al.* Design and characterization of epitope-scaffold immunogens that present the motavizumab epitope from respiratory syncytial virus. *J. Mol. Biol.* **409**(5), 853–866. https://doi.org/10.1016/j.jmb.2011.04.044 (2011).
60. Frey, G. *et al.* Distinct conformational states of HIV-1 Gp41 are recognized by neutralizing and non-neutralizing antibodies. *Nat. Struct. Mol. Biol.* **17**(12), 1486–1491. https://doi.org/10.1038/nsmb.1950 (2010).
61. Wu, X. *et al.* Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* **333**(6049), 1593–1602. https://doi.org/10.1126/science.1207532 (2011).
62. Thomson, C. A. *et al.* Germline V-genes sculpt the binding site of a family of antibodies neutralizing human cytomegalovirus. *EMBO J.* **27**(19), 2592–2602. https://doi.org/10.1038/emboj.2008.179 (2008).
63. Diskin, R. *et al.* Increasing the potency and breadth of an HIV antibody by using structure-based rational design. *Science* **334**(6060), 1289–1293. https://doi.org/10.1126/science.1213782 (2011).
64. Cockburn, J. J. B. *et al.* Mechanism of dengue virus broad cross-neutralization by a monoclonal antibody. *Structure* **20**(2), 303–314. https://doi.org/10.1016/j.str.2012.01.001 (2012).
65. Tugarinov, V. *et al.* NMR structure of an anti-Gp120 antibody complex with a V3 peptide reveals a surface important for co-receptor binding. *Struct. Lond. Engl.* **8**(4), 385–395. https://doi.org/10.1016/s0969-2126(00)00119-2 (2000).
66. Midgley, C. M. *et al.* Structural analysis of a dengue cross-reactive antibody complexed with envelope domain III reveals the molecular basis of cross-reactivity. *J. Immunol.* **188**(10), 4971–4979. https://doi.org/10.4049/jimmunol.1200227 (2012).
67. Reguera, J. *et al.* structural bases of coronavirus attachment to host aminopeptidase N and its inhibition by neutralizing antibodies. *PLOS Pathog.* **8**(8), e1002859. https://doi.org/10.1371/journal.ppat.1002859 (2012).
68. Zhou, T. *et al.* Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science* **329**(5993), 811–817. https://doi.org/10.1126/science.1192819 (2010).
69. Pancera, M. *et al.* Structural basis for diverse N-glycan recognition by HIV-1-neutralizing V1-V2-directed antibody PG16. *Nat. Struct. Mol. Biol.* **20**(7), 804–813. https://doi.org/10.1038/nsmb.2600 (2013).

70. Luftig, M. A. *et al.* Structural basis for HIV-1 neutralization by a Gp41 fusion intermediate-directed antibody. *Nat. Struct. Mol. Biol.* **13**(8), 740–747. https://doi.org/10.1038/nsmb1127 (2006).
71. Lee, C.-C. *et al.* Structural basis for the antibody neutralization of herpes simplex virus. *Acta Crystallogr. D Biol. Crystallogr.* **69**(10), 1935–1945. https://doi.org/10.1107/S0907444913016776 (2013).
72. Su, H.-P., Golden, J. W., Gittis, A. G., Hooper, J. W. & Garboczi, D. N. Structural basis for the binding of the neutralizing antibody, 7D11, to the poxvirus L1 protein. *Virology* **368**(2), 331–341. https://doi.org/10.1016/j.virol.2007.06.042 (2007).
73. Tang, X. *et al.* Structural basis for the neutralization and genotype specificity of hepatitis E virus. *Proc. Natl. Acad. Sci.* **108**(25), 10266–10271. https://doi.org/10.1073/pnas.1101309108 (2011).
74. Cherrier, M. V. *et al.* Structural basis for the preferential recognition of immature flaviviruses by a fusion-loop antibody. *EMBO J.* **28**(20), 3269–3276. https://doi.org/10.1038/emboj.2009.245 (2009).
75. Austin, S. K. *et al.* Structural basis of differential neutralization of DENV-1 genotypes by an antibody that recognizes a cryptic epitope. *PLOS Pathog.* **8**(10), e1002930. https://doi.org/10.1371/journal.ppat.1002930 (2012).
76. Kong, L. *et al.* Structural basis of hepatitis C virus neutralization by broadly neutralizing antibody HCV1. *Proc. Natl. Acad. Sci.* **109**(24), 9499–9504. https://doi.org/10.1073/pnas.1202924109 (2012).
77. Nybakken, G. E. *et al.* Structural basis of west nile virus neutralization by a therapeutic antibody. *Nature* **437**(7059), 764–769. https://doi.org/10.1038/nature03956 (2005).
78. Zhou, T. *et al.* Structural definition of a conserved neutralization epitope on HIV-1 Gp120. *Nature* **445**(7129), 732–737. https://doi.org/10.1038/nature05580 (2007).
79. Fleury, D., Daniels, R. S., Skehel, J. J., Knossow, M. & Bizebard, T. Structural evidence for recognition of a single epitope by two distinct antibodies. *Proteins* **40**(4), 572–578 (2000).
80. Pak, J. E. *et al.* Structural insights into immune recognition of the severe acute respiratory syndrome coronavirus S protein receptor binding domain. *J. Mol. Biol.* **388**(4), 815–823. https://doi.org/10.1016/j.jmb.2009.03.042 (2009).
81. Wu, Y. *et al.* Structural insight into distinct mechanisms of protease inhibition by antibodies. *Proc. Natl. Acad. Sci.* **104**(50), 19784–19789. https://doi.org/10.1073/pnas.0708251104 (2007).
82. Cockburn, J. J. *et al.* Structural insights into the neutralization mechanism of a higher primate antibody against dengue virus. *EMBO J.* **31**(3), 767–779. https://doi.org/10.1038/emboj.2011.439 (2012).
83. Backovic, M. *et al.* Structure of a core fragment of glycoprotein H from pseudorabies virus in complex with antibody. *Proc. Natl. Acad. Sci.* **107**(52), 22635–22640. https://doi.org/10.1073/pnas.1011507107 (2010).
84. McLellan, J. S. *et al.* Structure of a major antigenic site on the respiratory syncytial virus fusion glycoprotein in complex with neutralizing antibody 101F. *J. Virol.* **84**(23), 12236–12244. https://doi.org/10.1128/JVI.01579-10 (2010).
85. Prabakaran, P. *et al.* Structure of severe acute respiratory syndrome coronavirus receptor-binding domain complexed with neutralizing antibody. *J. Biol. Chem.* **281**(23), 15829–15836. https://doi.org/10.1074/jbc.M600697200 (2006).
86. Lee, J. E. *et al.* Structure of the ebola virus glycoprotein bound to an antibody from a human survivor. *Nature* **454**(7201), 177–182. https://doi.org/10.1038/nature07082 (2008).
87. Hu, G., Liu, J., Roux, K. H. & Taylor, K. A. Structure of simian immunodeficiency virus envelope spikes bound with CD4 and monoclonal antibody 36D5. *J. Virol.* https://doi.org/10.1128/JVI.00134-17 (2017).
88. Huang, C. *et al.* Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 Gp120 and CD4. *Science* **317**(5846), 1930–1934. https://doi.org/10.1126/science.1145373 (2007).
89. Malby, R. L. *et al.* The structure of a complex between the NC10 antibody and influenza virus neuraminidase and comparison with the overlapping binding site of the NC41 antibody. *Structure* **2**(8), 733–746. https://doi.org/10.1016/S0969-2126(00)00074-5 (1994).
90. Walls, A. C. *et al.* Unexpected receptor functional mimicry elucidates activation of coronavirus fusion. *Cell* **176**(5), 1026-1039.e15. https://doi.org/10.1016/j.cell.2018.12.028 (2019).
91. Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Res.* **28**(1), 235–242. https://doi.org/10.1093/nar/28.1.235 (2000).
92. Duvenaud, D. K. *et al.* Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems* Vol. 28 (eds Cortes, C. *et al.*) 2224–2232 (Curran Associates Inc., 2015).
93. RDKit. https://www.rdkit.org/. (Accessed 14 Mar 2020).
94. Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., & Pande, V. MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* **9**(2), 513–530 (2018).
95. Ramsundar, B. *et al. Deep Learning for the Life Sciences* (O'Reilly Media, 2019).
96. Chen, T., Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; KDD'16. (Association for Computing Machinery, San Francisco, California, USA, 2016, pp 785–794). https://doi.org/10.1145/2939672.2939785.
97. Breiman, L. Random forests. *Mach. Lang.* **45**(1), 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
98. Neffati, S., Abdellafou, K. B., Taouali, O. & Bouzrara, K. Enhanced SVM–KPCA method for brain MR image classification. *Comput. J.* https://doi.org/10.1093/comjnl/bxz035 (2020).
99. He, R. *et al.* Analysis of multimerization of the SARS coronavirus nucleocapsid protein. *Biochem. Biophys. Res. Commun.* **316**(2), 476–483. https://doi.org/10.1016/j.bbrc.2004.02.074 (2004).
100. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA.* **89**(22), 10915–10919 (1992).
101. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**(9), 646–652. https://doi.org/10.1038/nsb0902-646 (2002).
102. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**(1), 43–56. https://doi.org/10.1016/0010-4655(95)00042-E (1995).
103. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25. https://doi.org/10.1016/j.softx.2015.06.001 (2015).
104. Lindahl; Abraham; Hess; van der Spoel. GROMACS 2020.1 Source Code; Zenodo (2020). https://doi.org/10.5281/zenodo.3685919.
105. Schmid, N. *et al.* Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J.* **40**(7), 843. https://doi.org/10.1007/s00249-011-0700-9 (2011).
106. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**(12), 1463–1472. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12%3c1463::AID-JCC4%3e3.0.CO;2-H (1997).
107. Darden, T., York, D. M. & Pedersen, L. G. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* https://doi.org/10.1063/1.464397 (1993).
108. Lingenheil, M., Denschlag, R., Reichold, R. & Tavan, P. The, "Hot-Solvent/Cold-Solute" problem revisited. *J. Chem. Theory Comput.* **4**(8), 1293–1306. https://doi.org/10.1021/ct8000365 (2008).
109. Simulations of proteins with inhomogeneous degrees of freedom: The effect of thermostats. *J. Comput. Chem. Wiley Online Library*. (2008). https://onlinelibrary.wiley.com/doi/full/10.1002/jcc.20951. (Accessed 13 Mar 2020).
110. Ou, X. *et al.* Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat. Commun.* **11**(1), 1–12. https://doi.org/10.1038/s41467-020-15562-9 (2020).

111. Morgan, R. S. & McADON, J. M. Predictor for sulfur-aromatic interactions in globular proteins. *Int. J. Pept. Protein Res.* **15**(2), 177–180. https://doi.org/10.1111/j.1399-3011.1980.tb02566.x (1980).
112. Morgan, R. S., Tatsch, C. E., Gushard, R. H., Mcadon, J. M. & Warme, P. K. Chains of alternating sulfur and π-bonded atoms in eight small proteins. *Int. J. Pept. Protein Res.* **11**(3), 209–217. https://doi.org/10.1111/j.1399-3011.1978.tb02841.x (1978).
113. Stracke, J. *et al.* A novel approach to investigate the effect of methionine oxidation on pharmacokinetic properties of therapeutic antibodies. *mAbs* **6**(5), 1229–1242. https://doi.org/10.4161/mabs.29601 (2014).
114. (IUCr) Coot: model-building tools for molecular graphics. https://onlinelibrary.wiley.com/iucr/doi/10.1107/S0907444904019158. (Accessed 9 Mar 2020).
115. (IUCr) Features and development of Coot. https://onlinelibrary.wiley.com/iucr/doi/10.1107/S0907444910007493. (Accessed 13 Mar 2020).
116. McGibbon, R. T. *et al.* MDTraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109**(8), 1528–1532. https://doi.org/10.1016/j.bpj.2015.08.015 (2015).
117. Native contacts determine protein folding mechanisms in atomistic simulations|PNAS. https://www.pnas.org/content/110/44/17874. (Accessed 13 Mar 2020).


## Acknowledgements

The authors gratefully acknowledge the use of the supercomputing resource Arjuna provided by the Pittsburgh Supercomputing Center (PSC). This work is supported by Center for Machine Learning in Health (CMLH) (47247.1.5007162) at Carnegie Mellon University (https://www.cs.cmu.edu/cmlh-cfp) and start-up fund from Mechanical Engineering Department at CMU. The authors would like to thank Prof. Reeja Jayan for her support and Junhan Li for his help in collecting the data.

## Author contributions

R.M. collected the dataset and performed Machine Learning models. P.Y. ran the simulations and performed the interpretability. A.B.F. conceived the research and supervised the research. R.M., P.Y and A.B.F. wrote the paper.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-84637-4.

**Correspondence** and requests for materials should be addressed to A.B.F.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.