A CNN model for predicting binding affinity changes between SARS-CoV-2 spike RBD variants and ACE2 homologues

Chen Chen[a,1], Veda Sheersh Boorla[a,1], Ratul Chowdhury[a], Ruth H. Nissly[b,c], Abhinay Gontu[b,c], Shubhada K. Chothe[b,c], Lindsey LaBella[c], Padmaja Jakka[b,c], Santhamani Ramasamy[b,c], Kurt J. Vandegrift[d,e], Meera Surendran Nair[b,c], Suresh V. Kuchipudi[b,c,e,2], Costas D. Maranas[a,2]

[a] *Department of Chemical Engineering, The Pennsylvania State University, University Park, PA 16802, USA*

[b] *Animal Diagnostic Laboratory, Department of Veterinary and Biomedical Sciences, The Pennsylvania State University, University Park, PA 16802, USA*

[c] *Department of Veterinary and Biomedical Sciences, The Pennsylvania State University, University Park, PA 16802, USA*

[d] *Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA*

[e] *Center for Infectious Disease Dynamics, Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA*

Author contributions: C.C., V.S.B., C.D.M., S.V.K. designed research; C.C., V.S.B. performed research; C.C., V.S.B. analyzed data; C.C., V.S.B., C.D.M., R.C., S.V.K., S.K.C, K.J.V., R.H.N., A.G., M.S.N., L.L., P.J., S.R. wrote the paper

[1] *C.C. and V.S.B. contributed equally to this work.*

[2] *To whom correspondence may be addressed. Email: costas@psu.edu or skuchipudi@psu.edu*

ABSTRACT

The cellular entry of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) involves the association of its receptor binding domain (RBD) with human angiotensin converting enzyme 2 (hACE2) as the first crucial step. Efficient and reliable prediction of RBD-hACE2 binding affinity changes upon amino acid substitutions can be valuable for public health surveillance and monitoring potential spillover and adaptation into non-human species. Here, we introduce a convolutional neural network (CNN) model trained on protein sequence and structural features to predict experimental RBD-hACE2 binding affinities of 8,440 variants upon single and multiple amino acid substitutions in the RBD or ACE2. The model achieves a classification accuracy of 83.28% and a Pearson correlation coefficient of 0.85 between predicted and experimentally calculated binding affinities in five-fold cross-validation tests and predicts improved binding affinity for most circulating variants. We pro-actively used the CNN model to exhaustively screen for novel RBD variants with combinations of up to four single amino acid substitutions and suggested candidates with the highest improvements in RBD-ACE2 binding affinity for human and animal ACE2 receptors. We found that the binding affinity of RBD variants against animal ACE2s follows similar trends as those against human ACE2. White-tailed deer ACE2 binds to RBD almost as tightly as human ACE2 while cattle, pig, and chicken ACE2s bind weakly. The model allows testing whether adaptation of the virus for increased binding with other animals would cause concomitant increases in binding with hACE2 or decreased fitness due to adaptation to other hosts.

Keywords: SARS-CoV-2, Human ACE2, Animal ACE2, Binding Affinity, Convolutional Neural Network, Sequence

1. Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is responsible for the coronavirus disease 2019 (COVID-19) pandemic and has continued to evolve since the initial outbreak.[1] Several variants of the wild-type (WT) virus (Wuhan-Hu-1)[2] have been identified in different countries, including the United Kingdom (Alpha or B.1.1.7, Eta or B.1.525), South Africa (Beta or B.1.351, Omicron or B.1.1.529 or BA.1), Brazil (Gamma or P.1, Zeta or P.2), United States (Epsilon or B.1.429/B.1.427, Iota or B.1.526), India (Kappa or B.1.617.1, Delta or B.1.617.2), Philippines (Theta or P.3), Columbia (Mu or B.1.621), Peru (Lambda or C.37) and France (B.1.640.2).[3–5] The Omicron variant has quickly overtaken Delta as the globally dominant variant thanks in part to 60 mutations which have conferred increased transmissibility and increased reinfection risk as well as significant reductions in vaccine effectiveness.[6,7] The Omicron subvariant BA.2 contains additional amino acid changes, and it has been reported that both naive and immunologically trained individuals exhibit even higher susceptibility to BA.2

than the original Omicron strain.[8] While novel therapeutics and authorized COVID-19 vaccines have been deployed to mitigate the public health impacts of SARS-CoV-2 and limit its spread, the continuous emergence of SARS-CoV-2 variants is complicating the situation and results in periodic surges of new infections.[9]

Many domestic and wild animals have been shown to be susceptible to SARS-CoV-2 either by natural and/or experimental infections,[9,10] revealing an extensive host range of SARS-CoV-2. Cats, dogs, lions, tigers in zoos, minks, and ferrets have been reported to be infected via contact with COVID-19 patients, while snow leopards, pumas, and gorillas have been found to be infected with SARS-CoV-2 by natural means.[10] White-tailed deer have also been confirmed to be susceptible to SARS-CoV-2 through infection studies[11,12] After anti-SARS-CoV-2 antibodies were detected in 33% of the 481 white-tailed deer samples in four different states[13], multiple reports of natural infection of deer in USA[14–17] and Canada[18,19] have been reported. Moreover, recent studies[20] showed that Syrian hamsters in Hong Kong and deer in Canada[18] were able to transmit SARS-CoV-2 to humans, adding to the list of animal to human transmissions after the initial documented case in mink[21] with evidence for animal-to-human spillover. These findings underscore the importance of continual monitoring of potential spillovers of SARS-CoV-2 into non-human species, as the virus could gain a toe-hold and evolve in animal reservoirs, which could constitute a persistent threat of spillover back to humans. This scenario would complicate future mitigation strategies against the SARS-CoV-2 virus.[21]

SARS-CoV-2 is an enveloped single-stranded RNA virus that expresses the spike protein on its surface, which mediates the binding to host cells.[22] The association of the receptor binding domain (RBD) of the spike protein with the human angiotensin converting enzyme-2 (hACE2) receptor represents the first crucial step of viral infection.[2,23] SARS-CoV-2 variants contain single or multiple amino acid substitution as well as indels in the spike protein, and some of these changes have been shown to alter the RBD-hACE2 binding strength.[3,4] Several amino acid changes in the RBD of spike protein have considerably impacted the transmissibility and antigenicity in circulating variants, and the increased frequency of these amino acid changes may indicate a positive association with RBD-hACE2 binding affinity enhancement.[3,24] The amino acid change D614G showed increased prevalence which emerged multiple times in the global SARS-CoV-2 population, and the advantages for infectivity and transmissibility of this mutation have been indicated in several studies.[25–28] The N439K mutation was observed with increased frequency when circulating widely in Europe,[29] and Y453F was detected in Denmark among farmed mink and humans,[21] where both mutations showed enhanced RBD binding affinity to hACE2. Being present in five circulating variants, N501Y can significantly increase the binding affinity by introducing new aromatic stacking interactions and hydrogen bonds.[30,31] E484K is even more widely prevalent and is present in seven circulating variants. It destabilizes the RBD-down state to favor the RBD-up state which is a required conformation for effective binding to hACE2.[32,33] In addition to affinity enhancement, E484K also affords escape from neutralizing

antibodies,[33–36] and the presence of the alternative change E484Q in the Kappa variant further suggests the importance of monitoring amino acid changes at this position. In the Omicron variant, as many as 60 amino acid changes are encoded including 37 in the spike protein (with 15 in the RBD),[6] among which N501Y, T478K and K417N were also identified in earlier variants Alpha, Beta, Gamma, Delta, Theta, and Mu.[3] The immune escape capabilities are possibly conferred by K417N,[37] S477N,[38] E484A[38] and other novel mutations, while the binding affinity between RBD and hACE2 is retained due to strongly binding-improving mutations such as N501Y,[30,31] both contributing to the increased viral transmissibility of Omicron compared to earlier variants.[39–41] Early on during the pandemic, to gain a more complete view of the effect of RBD mutation, Starr et al.[30] exhaustively assessed the impact of single amino acid changes on RBD expression and hACE2 binding. They found that 84.5% of the amino acid changes are detrimental, 7.5% are neutral and the rest 8.0% can lead to enhanced binding of the RBD with hACE2. Chan et al.[42] used deep mutagenesis to systematically evaluate the binding affinity of WT RBD for hACE2 mutants, and the engineered decoy receptor for SARS-CoV-2 showed comparably high affinity to neutralizing antibodies. These findings highlight the importance of monitoring single amino acid changes and their potential for binding affinity improvement and call for prospective methods that can rapidly scan and identify amino acid change combinations that are likely to further boost affinity with ACE2.

By pre-screening mutations in search of potentially contagious variants, computational methods can contribute to understanding alterations in the characteristics of the circulating variants and identify particularly problematic ones. The binding free energy can be reasonably approximated using molecular-mechanics-based empirical force fields such as Rosetta.[43–45] It can also be calculated by performing molecular mechanics-generalized Born surface area (MM-GBSA)[46–49] analysis on configurations generated by molecular dynamics (MD) simulations.[50] The hybrid quantum mechanics/molecular mechanics (QM/MM) approaches[51] can also be used when critical chemical reactions are involved in binding activities, but the significant computational cost restricts QM treatment for large protein-protein complexes.[52] Reweighting of energy terms can be used to reach better prediction accuracy, where the weights are trained to accurately reproduce experimentally determined binding free energies ($\Delta\Delta G_{\mathrm{bind}}$).[53,54] As a powerful complement, machine learning algorithms can effectively "learn" highly complex relationships among energy components and the target binding affinity from training samples.[55,56] Using a data-driven approach, higher-level correlations can be captured and later be used to improve the accuracy of predictions.[57]

Recently, we developed a two-step framework to quantitatively predict binding affinity change upon amino acid substitutions.[58] The first step consists of 48 parallel 4-ns MD simulations of each RBD variant complexed with hACE2, followed by MM-GBSA analysis to extract decomposed binding energy terms. The second step involves implementing a neural network (NN) to predict the apparent dissociation constant ($K_{\mathrm{D,app}}$) ratios between the variants and the

WT using the obtained decomposed energy terms as descriptors. Agreement between experimental values and the NN_MM-GBSA model predictions was significantly better than predictions made directly using raw energies, reaching a correlation coefficient of 0.73 and an accuracy of 82.8% for the classification of improving or worsening the binding affinity.[58] Albeit encouraging, the computational demand in dataset preparation involving expensive MD simulations prevents the NN_MM-GBSA model from being expanded to a much larger training set and/or from being extensively used for large-scale genomic screening. Therefore, developing a tool with comparable accuracy yet better computational efficiency remains relevant. Readily accessible descriptors are crucial for machine learning processes, and the protein sequence is a promising candidate. Recent breakthroughs by AlphaFold2[59] and RoseTTAFold[60] demonstrated that the amino acid sequence contains a wealth of information for protein structure prediction, which can serve as the basis for protein-protein binding affinity prediction.[61–65]

Herein, we introduce a convolutional neural network (CNN) regression model (CNN_seq) based on protein sequence and WT complex structural features. Various features for (i) individual residue identities including hydropathy index,[66] volume,[67] zScales,[57] and VHSE[68] (principal components score Vectors of Hydrophobic, Steric, and Electronic properties), (ii) residue-pair interactions in AAIndex[69], and (iii) residue properties such as normalized vdW volume, polarity, charge, polarizability, secondary structure, and solvent accessibility[69] were used in the feature encoding procedure. The training was performed against all 8,440 variants with single and multiple amino acid changes in the RBD and hACE2. The predictive capability of this CNN_seq model was assessed by comparing it with the experimental $K_{D,app}$ ratios, achieving a classification accuracy of 83.28% and a correlation coefficient r of 0.85 in five-fold cross-validation tests. The model was further tested against many circulating variants that were blind to the model during training, and most were predicted to show improved binding affinity, as demonstrated by other experimental studies.[23–25,29,30,32,37,70–91] Furthermore, we randomly chose 1,667 of the 8,440 variants to serve as the blind test set and used the rest to perform a five-fold cross-validation test. Similar performance (%VC = 83.47%, r = 0.84) was observed for the blind test set, reconfirming the robustness of the CNN model predictions.

Given the accuracy and efficiency of this CNN_seq model, we were able to exhaustively screen over 220,000 RBD variants for each host with combinations of up to four amino acid substitutions and suggest candidates with the highest improvement in RBD-ACE2 binding affinity for monitoring. The predicted binding affinity of RBD variants for deer ACE2 was found to be similar to humans, which was lower for cattle and pigs and lowest for chickens. The computational model can be accessed from the GitHub repository (https://github.com/maranasgroup/CNN_seq_CoV2).

2. Results

Dataset preparation. We curated the training dataset by combining experimental results of deep mutational scanning of RBD (Starr et al.[30]) and random mutational scanning of hACE2 (Chan et al.[42]). The curated dataset consisted of 8,440 variants (see Methods for details of variants selection and distribution). To quantify the change in RBD binding affinity for hACE2, Starr et al.[30] reported the apparent dissociation constant $K_{D,app}$ ratio which was defined as $K_{D,app,variant}/K_{D,app,WT}$. For each variant, as compared to the WT RBD, a $K_{D,app}$ ratio greater than 1 indicates stronger (or improving) binding to hACE2 whereas a value less than 1 implies weaker (or worsening) binding. Note that, the $K_{D,app,WT}$ in present work always refers to the $K_{D,app}$ for the complex formed by WT hACE2 and WT SARS-CoV-2 RBD proteins, and such designation allows for the direct comparison across human and animal hosts.

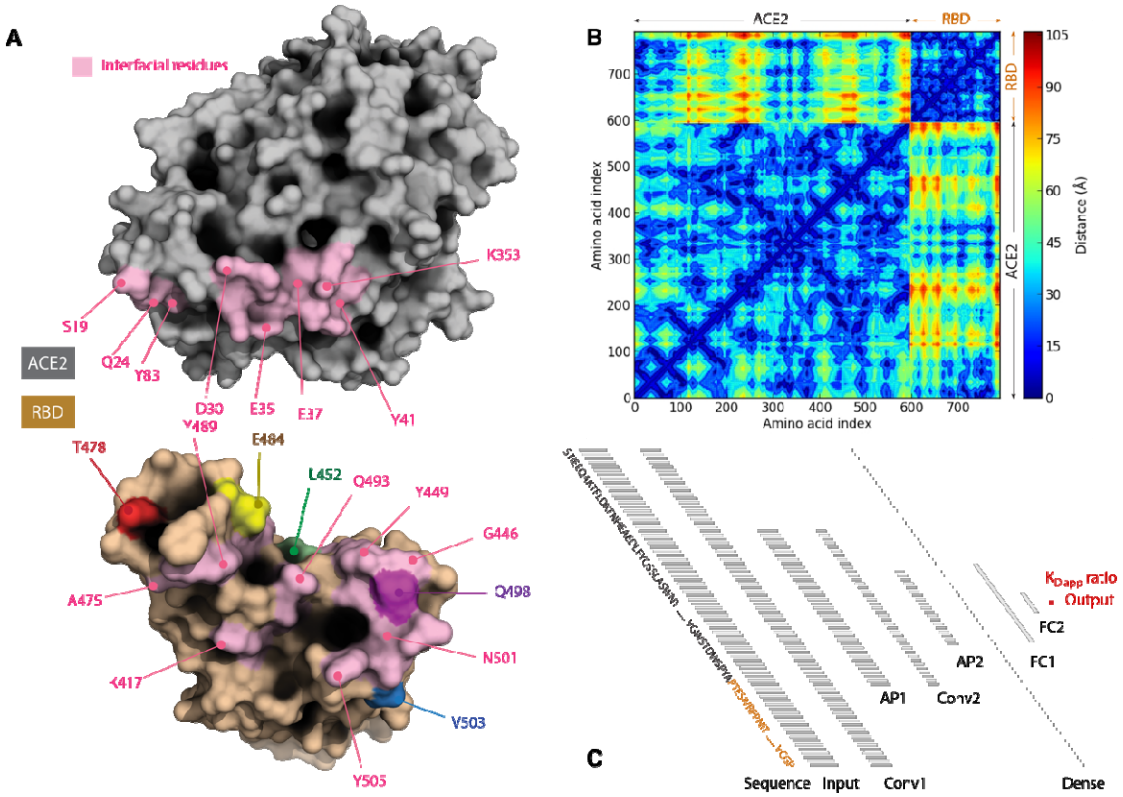

Figure 1. Schematic representation of the CNN_seq workflow. (A) Identification of interfacial residues, hydrogen bonds, salt bridges and disulfide bonds from the structure of receptor binding domain (RBD) complexed with angiotensin converting enzyme-2 (ACE2). Interfacial residues are colored pink on protein surfaces, and some key non-interfacial residues are highlighted. (B) Generation of the distance map recording the distances between all residue pairs in the complex structure. (C) The structure of CNN_seq model which takes RBD-ACE2 complex sequence as the input and apparent dissociation constant ($K_{D,app}$) ratio as the output.

Feature encoding. The CNN_seq model uses both sequence-based and structure-based features in model construction to maximize the utilization of available data. For the human case, the sequences and three-dimensional (3D) structure of SARS-CoV-2 spike RBD in complex with hACE2 were obtained from the Protein Data Bank[92] (PDB ID: 6LZG), for the animal cases, the sequences of ACE2 proteins were collected from UniProt[93] for four species: deer (*Odocoileus virginianus*, ID: A0A6J0Z472), cattle (*Bos indicus* × *Bos taurus*, ID: A0A4W2H6E0), pig (*Sus scrofa*, ID: A0A220QT48), and chicken (*Gallus gallus*, ID: F1NHR4). Due to the lack of experimentally determined structures for these animals, we used SWISS-Model[94] to perform homology modeling and generated 3D structures for the complexes. Subsequent structural refinement for each complex was performed through an MD simulation with explicit water, and the interface equilibration was demonstrated by reaching a standard deviation (SD) in the root mean square deviation (RMSD) of less than 1 Å in the final 100-ns (see Methods for details).

The structure of the RBD-ACE2 complex encodes information as to the local environment that is accessible by each residue, and the interactions between residues in ACE2 and RBD underpin the overall binding strength of the complex. We used the PISA (Proteins, Interfaces, Structures and Assemblies) tool[95] provided by Protein Data Bank in Europe (PDBe)[96] to help identify the interfacial residues as well as key interactions such as hydrogen bonds, salt bridges, and disulfide bonds formed by compatible residue pairs (Figure 1A). In addition, the Protein Contact Maps tool developed by Benjamin et al.[97] was employed to generate the contact map (Figure 1B), which is a two-dimensional (2D) matrix recording the distances between all possible amino acid residue pairs in the complex. For each complex, the structure, contact map and the interfacial residue information are not changed significantly when mutations are subsequently introduced.

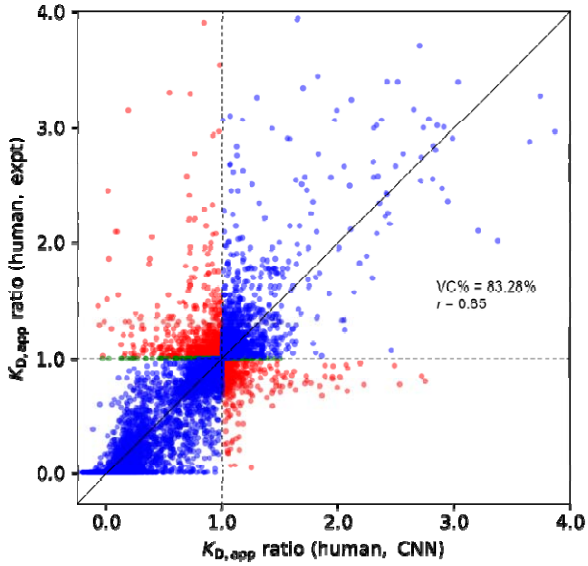

Figure 2. Comparison of $K_{D,app}$ ratio between experiments and CNN_seq models predictions from five-fold cross-validation tests on 8,440 variants. Correctly classified variants are colored in blue, incorrectly classified variants are colored in red, and variants with unchanged binding affinities are colored in green. Horizontal and vertical dashed lines are drawn to indicate the dividing line where $K_{D,app}$ ratio equals to 1, and a diagonal solid line is drawn to indicate perfect correlation.

CNN model trained on experimental $K_{D,app}$ ratio. A CNN regression model was constructed by taking the RBD-ACE2 sequence for variants as the input. As illustrated in Figure 1C, the CNN_seq model contains one input layer, two 1D convolutional (Conv) layers, two 1D average pooling (AP) layers, one dense layer, two fully-connected (FC) layers, and an output layer. There is a rectified linear unit (ReLU) activation function following each Conv and FC layer, and dropout regularization method[98] is applied to each FC layer to reduce overfitting (see Methods for details of the CNN structure). The matrix of input features pass through a series of layers and functions continuously, and the final output is the apparent dissociation constant $K_{D,app}$ ratio. Such architecture takes into account the interactions between individual residues and their local environment, and allows higher level correlations between domains to be captured.

For model evaluation, we followed the five-fold cross-validation procedure, where the entire dataset was split into five subsets, with each subset considered as the test set once, and the rest of the four subsets are used together to form the training set. A complete cycle of five-fold cross-validation produced five individual models, and each model made predictions on the subset of variants that the model itself did not see. Combining predictions of the five subsets, all the variants in the original dataset were predicted once, and the performance of the model can be fairly evaluated. Two metrics were used to evaluate the performance of the CNN_seq model, the percent recovery of correct variant classification (%VC) and the Pearson correlation coefficient (r). The %VC calculates in percentage the accuracy of classifying the direction of change in the binding affinity compared to WT, while r measures the strength of the linear correlation between the predicted and experimental $K_{D,app}$ ratio values (see Methods for details).

The averaged results from the five-fold cross-validation tests of the CNN_seq model achieved a %VC of 83.28% and a correlation coefficient r of 0.85 for a complete dataset of 8,440 variants. Compared to the performance of the NN_MM-GBSA model (%VC = 82.8%, r = 0.73),[58] the %VC of the CNN_seq model remained similar whereas there was a slight improvement in the correlation coefficient r. It appears that on balance the information content embedded in the 8,440 variants dataset approximately matches the information contained in the energy terms of the GBSA molecular dynamic simulation of the 108 variants used in model NN_MM-GBSA.[58] Figure 2 compares the $K_{D,app}$ ratio values between experiments and CNN_seq predictions, where 54.94% of the incorrectly classified variants appear to be for variants with experimental $K_{D,app}$ ratio values within [0.9-1.1]. The performance of the CNN_seq model was also evaluated separately on ACE2 and RBD variants (Figure S1 in *SI Appendix*). We found that CNN performed better for the 6,105 RBD variants (i.e., %VC = 85.2% and r = 0.87) compared to the 2,335 hACE2 variants (%VC = 72.19% and r = 0.63). The apparent difference in the performance of CNN in the two datasets alludes to systematic differences in the way affinities were calculated or different error margins. However, since there are nearly three times more RBD variants (72.10% of the dataset) than ACE2 variants (27.90% of the dataset), it is not surprising to see that the performance bias leans towards RBD variants. Although a model built on solely RBD variants could achieve better overall performance, we chose to include both sets of data in the training set because this model was ultimately designed to predict the binding affinity changes for animal hosts, whose ACE2s differ slightly from the human version.

Table 1. Comparison of $K_{D,app}$ ratio from experiments (Expt), CNN_seq (CNN), and NN_MM-GBSA (NN) model predictions on circulating RBD variants complexed with human and animal ACE2 proteins.

WHO label[99]	Pango lineage[100]	Amino acid change(s)	$K_{D,app}$ ratio						
			Human (Expt)	Human (NN[58])	Human (CNN)	Deer (CNN)	Cattle (CNN)	Pig (CNN)	Chicken (CNN)
		WT	1.00	1.00	1.00	1.01	0.91	0.82	0.60
Alpha	B.1.1.7	N501Y	4.36	1.22	1.66	1.30	1.21	1.11	1.09
		E484K*+N501Y	4.38	1.22	2.06	1.57	1.42	1.34	1.47
		S494P*+N501Y	–	1.25	1.64	1.25	1.23	1.07	1.04
		E484K*+S494P*+N501Y	–	1.22	2.00	1.49	1.42	1.30	1.40
Beta	B.1.351	K417N+E484K+N501Y	2.40	1.22	1.81	1.34	1.20	1.06	1.17
Gamma	B.1.1.28/P.1	K417T+E484K+N501Y	5.51	1.22	1.74	1.40	1.25	1.07	1.25
Delta	B.1.617.2	L452R+T478K	1.32	1.12	1.19	1.16	1.02	1.10	0.77
		K417N*+L452R+T478K	0.27	1.24	0.96	1.01	0.84	0.80	0.55
Epsilon	B.1.429/B.1.427	L452R	1.05	1.09	1.12	1.16	1.01	0.97	0.81
Zeta	P.2	E484K	1.15	1.21	1.32	1.22	1.04	0.95	0.87
Eta	B.1.525	E484K	1.15	1.21	1.32	1.22	1.04	0.95	0.87
Theta	P.3	E484K+N501Y	4.38	1.25	2.06	1.57	1.42	1.34	1.47
Iota	B.1.526	E484K	1.15	1.21	1.32	1.22	1.04	0.95	0.87
		L452R*+E484K	2.28	1.22	1.38	1.36	1.11	1.10	1.02
		S477N*+E484K	–	1.10	1.53	1.35	1.12	1.03	0.99
		L452R*+S477N*+E484K	–	1.20	1.59	1.49	1.20	1.18	1.12
Kappa	B.1.617.1	L452R+E484Q	1.67	1.21	1.28	1.26	1.06	1.04	0.92
Lambda	C.37	L452Q+F490S	–	1.11	1.18	1.15	1.04	0.93	0.72
Mu	B.1.621	R346K+E484K+N501Y	1.69	1.25	2.14	1.55	1.46	1.40	1.54
Omicron	B.1.1.529/BA.1	G339D+S371L+S373P+ S375F+K417N+N440K+ G446S+S477N+T478K+ E484A+Q493R+G496S+ Q498R+N501Y+Y505H	1.40	1.25	1.23	0.65	0.96	0.73	0.66
	BA.2	G339D+S371F+S373P+ S375F+T376A+D405N+ R408S+K417N+N440K+ S477N+T478K+E484A+ Q493R+G496S+Q498R+ N501Y+Y505H	–	1.25	1.18	0.87	0.97	0.91	0.77
	B.1.1.298	Y453F	1.78	1.21	1.41	1.23	1.05	0.98	0.73
	B.1.1.519	T478K	1.05	1.11	1.05	1.03	0.93	0.84	0.59
	B.1.36	N440K	1.17	1.11	1.13	1.11	0.97	0.86	0.65
	B.1.1.317	S477N+A522S	–	1.11	1.15	1.10	0.96	0.84	0.67
	B.1.1.141	T385I	0.91	1.12	1.02	1.05	0.93	0.76	0.54
	B.1.640.2	R346S+N394S+Y449N+ E484K+F490S+N501Y	–	1.25	1.73	1.31	1.33	1.06	1.15
	C.1.2	449H+484K+501Y	–	1.21	1.88	1.40	1.37	1.09	1.32

* Amino acid change detected in some sequences but not all.

CNN_seq model prediction of $K_{D,app}$ of circulating SARS-CoV-2 variants. The circulating variants include amino acid substitutions and/or indels in the SARS-CoV-2 spike protein such as the amino-terminal domain, the RBD, and the furin cleavage sequence.[24] In total, spike variants

for 21 circulating variants were chosen to be examined by the CNN_seq model, including 13 that are assigned WHO labels[99] and eight that only have Pango[100] classifications. We also assessed different instances of variants containing amino acid changes only present in some sequences. The experimental values were collected from available experimental results obtained through surface plasmon resonance,[29,32,70,71,89–91] bio-layer interferometry,[23,25,29,30,37,72–85,90] enzyme-linked immunosorbent assay,[74,82,86,87,90] or microscale thermophoresis.[88] For all variants with single amino acid changes, in the spike protein, we defaulted to the $K_{D,app,variant}/K_{D,app,WT}$ value from the work of Starr et al.[30] For experimentally tested variants with multiple amino acid changes we used the reported $K_{D,variant}/K_{D,WT}$ to make comparisons. Given the often-large differences in the reported values, the median calculated from the available experiment results was used.

As shown in Table 1, both the CNN_seq and NN_MM-GBSA models predicted improved binding affinity for most circulating variants relative to the WT, consistent with experimental observations. To further assess the performance of CNN_seq and NN_MM-GBSA models on blinded datasets, we also prepared a list of 15 variants containing multiple amino acid changes (Table S1 in *SI Appendix*). For this small blinded test set, the CNN_seq model achieved a %VC of 92.9% and r of 0.60 whereas NN_MM-GBSA model performance was somewhat worse with %VC of 75.7% and r of 0.28. This performance gap could be because NN_MM-GBSA model was trained using only 108 single amino acid change variants, whereas the blinded dataset involved single amino acid changes with 66.67% not present in the dataset used for training NN_MM-GBSA. Roughly, all circulating variants can be separated into two families, ones with $K_{D,app}$ ratios above 1.5 including Alpha, Beta, Gamma, Theta, Mu, B.1.640.2, C.1.2 variants, and the others with $K_{D,app}$ ratios between 1 and 1.5 including Delta, Epsilon, Zeta, Eta, Iota, Kappa, Lambda, Omicron, B.1.1.298, B.1.1.519, B.1.1.36, B.1.1.317, B.1.1.141, BA.2. This classification of the variants agrees well with the antigenic map,[1] alluding to the partial correlation between RBD-ACE2 binding affinity and evolution trajectories.

Specifically for the Omicron variant, both CNN_seq and NN_MM-GBSA models predict improved binding affinity with the $K_{D,app}$ ratio values of 1.23 ± 0.58 and 1.25 ± 0.09, respectively. Notably, only 13 out of a total of 25 amino acid change predictions from individual CNN_seq models were improving while the rest were worsening. This alludes to the hypothesis that highly improving towards binding amino acid changes enables neighboring positions to assume binding worsening amino acid changes to evade dominant immune recognition sites.[39,71,87,101–104] We also calculated the binding free energy ΔG of the RBD-hACE2 complex using the Rosetta force-field (see Methods for details). The resulting averaged ΔG of the Omicron- and WT-RBD-hACE2 complexes were -36.2 ± 2.4 and -44.3 ± 1.0 kcal/mol, respectively. Experimentally, there does not seem to be a consensus on the relative binding strength between the Omicron variants.[71,87,101] Surface plasmon resonance analysis by Cameroni et al.[71] showed that Omicron and Delta bind 2.4- and 1.2-fold stronger than WT, whereas Mannar et al.[101] suggest 1.5- and

1.4-fold improvement over WT for Omicron and Delta, respectively. Moreover, a combined experimental and simulation study by Wu et al.[87] concluded that Omicron has a lower binding affinity than Delta and WT. Nevertheless, with as many as 15 amino acid changes in the RBD of the Omicron variant, there could be formation/interruption of multiple residue-residue pairs and substantial structural rearrangement that might be responsible for the inconsistent predictions from different methods.[71,87,101]

As the subvariant of Omicron, BA.2 variant carries one less (G446S), one altered (S371L vs. S371F), and three additional amino acid changes (T376A, D405N, R408S). Notably, these five different mutations were all known to decrease the binding affinity for ACE2.[30] The CNN_seq model predicts just slightly lower $K_{D,app}$ ratio of 1.18 for BA.2 than 1.23 for the original Omicron. In contrast, the B.1.640.2 variant adds six amino acid changes comprised of four binding-improving (R346S, N394S, E484K, N501Y), one neutral (F490S), and one binding-decreasing (Y449N) mutation,[30] consistent with the predicted high $K_{D,app}$ ratio of 1.73 by the model.

Using the CNN_seq model built on human data to predict binding with animal ACE2. The inclusion of ACE2 mutations in the training set allows the CNN_seq model to learn how amino acid changes on ACE2 in addition to the RBD could affect the binding affinity change. The model predicted $K_{D,app,WT,animal}/K_{D,app,WT}$ values are summarized in Table 1, where $K_{D,app}$ ratio values for cattle, pig, and chicken were smaller than one, indicating weaker binding affinity of SARS-CoV-2 RBD for these animals than human, consistent with the experimental observations.[9,10] For deer, the WT RBD showed slightly higher binding affinity than human, which seems to agree with the high susceptibility as recently reported.[11–15] As shown in Table 1, circulating variants generally show enhanced binding affinity for animal ACE2 in all cases in line with corresponding improvements for human. However, the quantitative improvement generally lags behind the one achieved for human ACE2 alluding that adaptation to human ACE2 remains the main driving force of diversity generation. Delta and Kappa variants exhibit slightly higher binding affinity for deer ACE2 than human. Theta, Mu and Alpha+E484K variants show consistently high binding affinity ($K_{D,app}$ ratio > 1.3) whereas Gamma, Alpha+E484K+S494P, Iota+L452R, and Iota+L452R+S477N show slightly elevated binding affinity ($K_{D,app}$ ratio > 1.0). These CNN_seq model predictions suggest that just as variants achieve tighter binding with hACE2 which may explain gains in infectivity, affinity gains are also predicted against animal ACE2s. The gains in affinity for animal ACE2s appear to track the infection susceptibility of the four assessed animals with deer being very near to human, cattle and horse being significantly less, and chicken only moderately increased from the originally low affinity values.

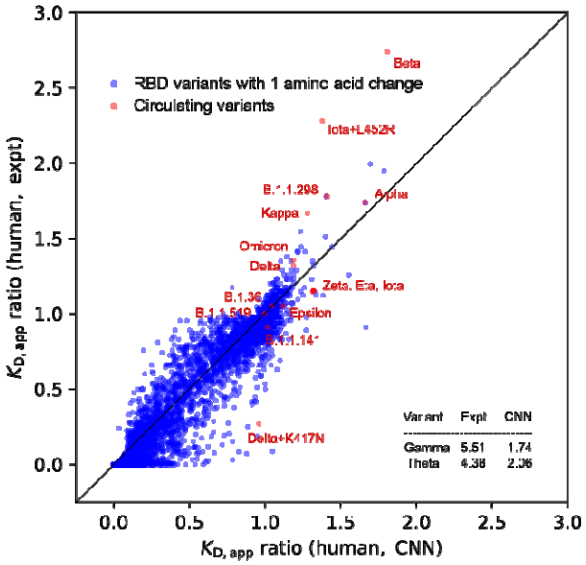

Figure 3. Comparison of $K_{D,app}$ ratio between experiments and CNN_seq models predictions from scanning on 3900 variants with one amino acid change. Scanning results are colored in blue, and circulating variants are labeled in red or indicated in the inset table. A diagonal dashed line is drawn to indicate the perfect positive correlation.

Computational scanning for novel variants with up to four amino acid changes using CNN_seq model for human and animals ACE2. On average, it takes approximately 250 ms to predict the $K_{D,app}$ ratio for a single variant on an NVIDIA Tesla P100 GPU card. This permits the exhaustive exploration of multiple amino acid changes for possibly further increases in binding affinity with ACE2. Based on the sequence range of the experimentally determined structure of RBD-hACE2 complex,[105] we focused on RBD variants in the range of residues from 333 to 527. We adopted a hierarchical approach by first exhaustively assessing all single amino acid changes and then selecting the top 20 variants with the highest $K_{D,app}$ ratio to exhaustively assess the addition of a second amino acid change. This procedure is repeated until variants with up to four amino acid changes are assessed. This procedure relies on the observation that affinity gains seen so far have been largely, though not exclusively, additive in the contribution of individual amino acid changes in the spike.[106]

Figure 3 compares the CNN_seq predicted and experimental $K_{D,app}$ ratio values for variants formed by hACE2 and RBD with single amino acid changes. In total, 195 × 20 = 3,900 RBD variants were scanned, where 3,883 of them have experimental referenced data. Among these variants, 1,344 (34.6%) are in the training set, and the rest 2,539 (65.4%) are unknown to the model. These 2,539 variants are all worsening examples and forms a blind test set, for which the prediction shows a VC% of 94.62% and r of 0.90, even higher than the results from training/validation process, further implying the robustness of the model. The binding affinity of circulating variants are reasonably well predicted by the model, showing general agreement with experiments.

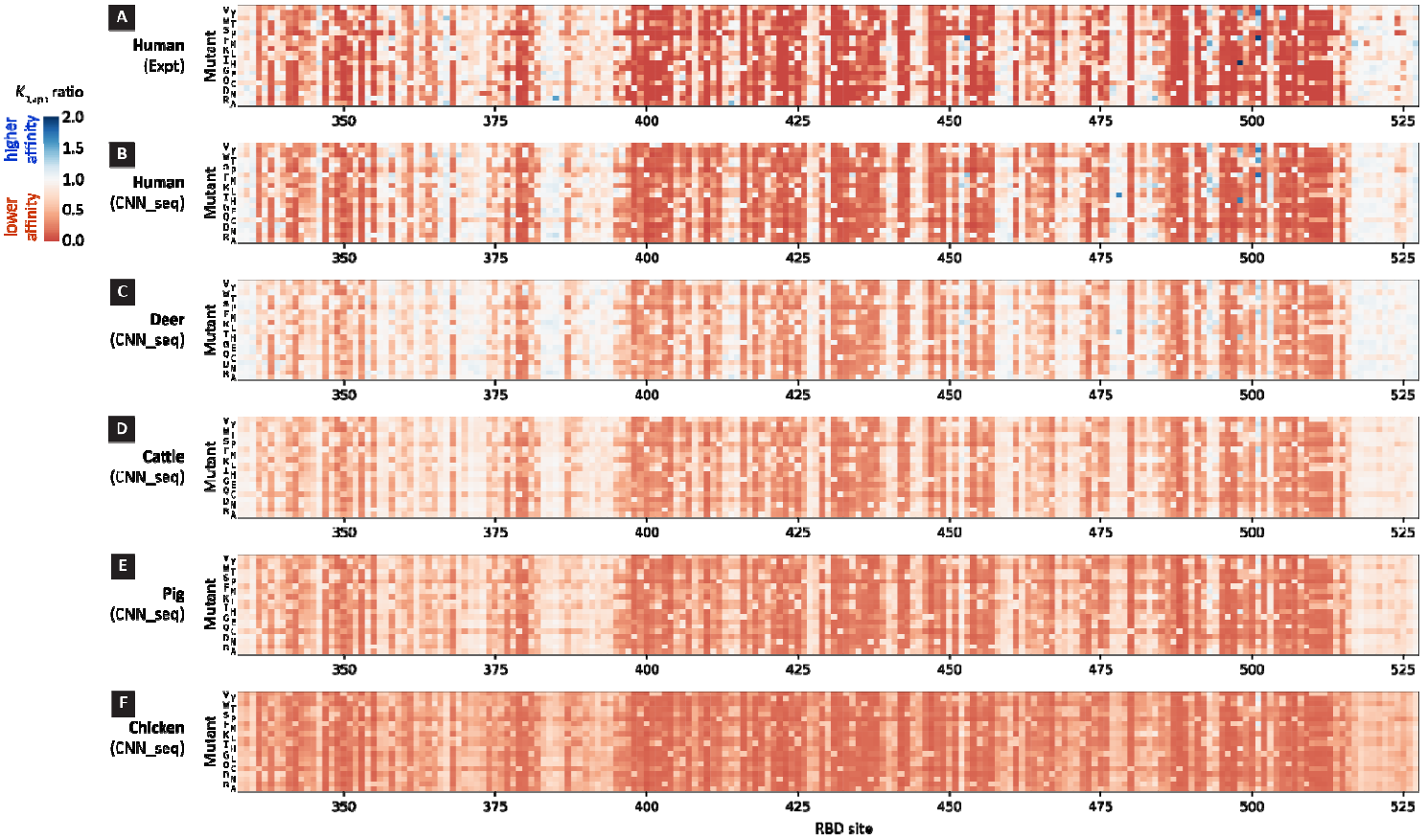

Figure 4. Binding affinity change heatmaps of RBD variants for ACE2 proteins from human and animal hosts. Experimental values and CNN_seq predictions for human are shown in (A) and (B), and CNN_seq predictions for animals are depicted in (C) deer, (D) cattle, (E) pig, (F) chicken. Squares are colored according to the $K_{D,app}$ ratio values.

To illustrate the variant distribution across the RBD sites under investigation, in Figure 4 we constructed the binding affinity change heatmaps for all 3,900 variants with one amino acid changes followed the pictorial style introduced in Starr et al.[30] Each one of the stripes represents the complete scanning results for one of the species, where the horizontal and vertical axes indicate the RBD sites and amino acid, respectively. The small squares are colored according to the CNN_seq predicted $K_{D,app}$ ratio values, with red for lower affinity, blue for higher affinity, and white for neutral or similar affinity. The general patterns of the heatmap for the experimental data (Figure 4A) are well captured in the CNN_seq predictions (Figure 4B) for the human case, in line with the VC% and *r* values obtained from data shown in Figure 3A.

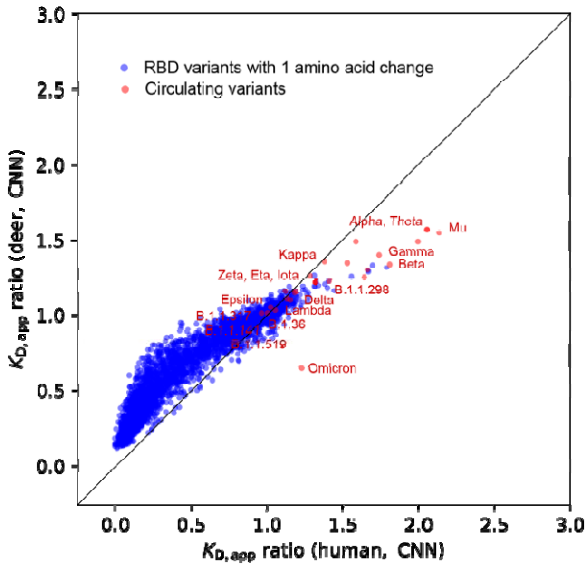

Figure 5. Comparison of binding affinities of RBD variants for ACE2 proteins of human and deer. All the RBD variants with single amino acid changes are colored in blue, circulating variants are labeled in red. A diagonal line is drawn to indicate the perfect positive correlation.

To examine the effects of RBD mutations on the change of binding affinity for animal ACE2 proteins, we further scanned the 3,900 RBD variants with one amino acid changes for all four animal hosts. As illustrated in Figure 4C to Figure 4F, the heat maps largely resemble the ones for human in proportion to the sequence similarity between their ACE2 proteins. In Figure 5 we compare the deer vs. human $K_{D,app}$ ratio values for all scanned 3,900 data and circulating variants. Results for other animals are shown in Figure S2 in *SI Appendix*. Correlations are observed between human and all animals RBD variants (see Figure 4).

With the scanning results, we could examine whether adaptation of the virus for increased binding with animals would cause concomitant increases to binding with human ACE2 (animal/human converging amino acid changes) or decreased fitness. The top deer/human converging and diverging amino acid changes are tabulated in Table S2 in *SI Appendix*. Among the deer/human converging amino acid changes, there are many candidates that show significant binding affinity enhancement for both deer and human including N501Y, E484K, Y453F, L452R that are contained in multiple circulating variants. For deer/human diverging amino acid changes, most candidates show moderated binding affinity increase for deer and mild decrease for human. Notably, change A522S was identified in B.1.1.317 variant, which circulated in Russia, UK, and Thailand in late 2020.[107]

Table 2. Scanning results for human with up to 4 amino acid changes on SARS-CoV-2 RBD.

1-AA change	$K_{D,app}$ ratio	2-AA change	$K_{D,app}$ ratio	3-AA change	$K_{D,app}$ ratio	4-AA change	$K_{D,app}$ ratio
N501F	1.79	478I+386S	2.45	501F+386S +493M	2.79	365W+390R+369W+498H	3.59
Q498H	1.70	365W+390R	2.26	365W+390R +369W	2.78	365W+390R+493L+498H	3.51
T478I	1.67	501F+386S	2.25	365W+390R +453F	2.77	365W+390R+453F+498H	3.49
N501Y	1.66	501F+358W	2.22	501F+386S+493V	2.77	365W+390R+484K+498H	3.47
N501T	1.55	498H+386S	2.16	478I+386S+453F	2.76	365W+390R+493M+498H	3.46
Q498Y	1.44	501Y+358W	2.13	365W+390R+493L	2.73	501F+386S+493M+498H	3.42
Y453F	1.41	501Y+386S	2.12	501F+386S+493A	2.73	478I+386S+453F+498H	3.41
Q493M	1.40	501F+358F	2.11	501F+386S+493L	2.72	365W+390R+453K+498H	3.41
Q493L	1.38	501F+391S	2.11	478I+386S+494K	2.72	501F+386S+493L+498H	3.40
Q493V	1.34	478I+358F	2.10	365W+390R+493M	2.71	501F+386S+493V+498H	3.40
E484K	1.32	501F+392W	2.10	501F+358W+493M	2.71	501F+358W+493M+498H	3.38
Q493Y	1.32	478I+378Q	2.09	365W+390R+453K	2.71	365W+390R+453K+384R	3.36
Q493A	1.31	498H+358W	2.09	365W+390R+484K	2.71	501F+358W+493V+498H	3.36
L452K	1.31	501F+518T	2.08	501Y+386S+493M	2.70	501Y+386S+493M+498H	3.36
Y365W	1.27	478I+386A	2.07	501F+358W+493V	2.69	501Y+386S+493L+498H	3.34
Y369W	1.27	501F+473F	2.06	501Y+386S+493V	2.69	501Y+358W+493M+498H	3.34
N501V	1.27	478I+527M	2.06	501Y+386S+493L	2.68	501Y+386S+493V+498H	3.33
Q498W	1.26	478I+358W	2.06	478I+386S+414A	2.68	365W+390R+369W+498Y	3.31
S494K	1.25	478I+386T	2.06	501F+358W+493A	2.66	478I+386S+494K+498H	3.31
Q414A	1.24	478I+339D	2.06	501Y+358W+493M	2.66	365W+390R+453K+384K	3.30

The effect of multiple amino acid changes (see Table 2) is difficult to interpret as they may cause significant re-organization of the RBD. It is unclear whether continually improving binding affinity would translate to increased infectivity as other biological processes underpinning productive cell infection and proliferation may become limiting (e.g., furin cleavage, RBD presentation, or internalization efficiency). However, the computationally predicted potential to reach even more increased binding affinity may suggest a still untapped potential of SARS-CoV-2 to undergo additional immune evasion changes whose potentially detrimental effect on binding could be ameliorated by a set of improving amino acid changes. As seen in Table 2, the effect of adding amino acids changes is not always additive. There are many instances where the best double change variant combines an amino acid change with very high affinity and another with mild or even lower affinity. Tables S3, S4, S5 and S6 list corresponding results for different animals in SI Appendix.

3. Discussion

The introduced CNN_seq model overcomes the computational barriers associated with our previous NN_MD-MMGBSA procedure[58] and at the same time unlocks the opportunity to consider amino acid changes throughout the entire RBD as well as ACE2 receptor during model training. The enlarged (by about 80-fold) training set affords the CNN_seq model to learn more effectively and potentially capture distal correlations. Meanwhile, the improved efficiency and accessibility of feature encoding expand the range of possible variants that can be assessed for binding affinity changes with human and animal ACE2s.

Focusing on the Omicron variant as an exemplar application, the CNN_seq model delineated the role of mutations as binding improving and/or immune escaping[108,109]. In addition, given that it is intractable to exhaustively assess the binding affinity changes in response to two or more amino acid changes experimentally, the CNN_seq model could serve as a computational alternative for the *a priori* assessment of possible epistasis and synergistic effects.[104] In addition, pre-calculation of very high binding affinity variants involving multiple amino acid changes using CNN and continuous surveillance of circulating variant databases could offer an alert system whenever an emerging variant is becoming sequence-adjacent to a computationally predicted problematic variant. This could become part of a surveillance system to flag variant amino acid changes seen in human or animal hosts that upon one or more additional changes could become a particularly problematic variant.

4. Methods

Homology modeling using Swiss-Model. Homology modelling was performed using SWISS-Model. For each animal, the sequence of specific ACE2 was concatenated with that of SARS-CoV-2 spike RBD, and loaded into the server interface. A representative subset of the template structures can be automatically extracted from the database through multiway alignment.

Refinement of homology models using molecular dynamics simulations. The structures of homology models were refined through MD simulations. Each RBD-ACE2 complex structure was first prepared using protein preparation wizard of Maestro in Schrödinger suite (v2019.4), and subsequently solvated in an orthorhombic box with 10 Å buffer in all three dimensions. The TIP3P model[110] was chosen for water molecules and the whole system was neutralized by adding Na^+ and Cl^- ions to reach a salt concentration of 0.15 M. The Amber99SB-ILDN force field[111] was used to account for the interactions of proteins, while the interactions between protein and water molecules were automatically generated by the Desmond force field building tool Viparr. The solvated system was minimized and equilibrated following the default relaxation protocol of Desmond[112], and a 100-ns production run was subsequently conducted to fully relax and optimize the structure of the RBD-ACE2 complex. The production run was performed in isothermal-isobaric (NPT) ensemble with periodic boundary conditions applied in all three dimensions, and a temperature of 300 K and a pressure of 1.0 atm was maintained. A time step of 2.0 fs was set to integrate the equations of motion, particle mesh Ewald method was used to describe the long-range interactions, and a cutoff distance of 9.0 Å was applied in the calculation of non-bonded short-range interactions. The convergence of the system was verified by monitoring the RMSD of the simulation run. For all RBD-ACE2 complexes, the RMSD values kept below 1 Å in the final 100-ns of the production run, suggesting convergence of the structure optimization.

Generate distance map based on structure using Protein Contact Maps. Since the base structure of each RBD-ACE2 complex requires only one optimization, we prepared the protein

contact map and used it as the constant input during CNN model construction. The protein contact map can be generated by utilizing the Protein Contact Maps tool developed by Benjamin et al. Each optimized RBD-ACE2 complex structure was loaded into the online server and the outputs were the contact map and the distance map associated with the individual input structure. The obtained binary two-dimensional matrix distance map is indexed following the provided protein sequence and will be accessed in CNN feature encoding process.

Identification of protein-protein interfacial contacts using PDBePISA. Protein Data Bank in Europe (PDBe) offers a web tool PISA (i.e., Proteins, Interfaces, Structures and Assemblies) to help access the interfacial interactions between proteins. By loading the optimized RBD-ACE2 complex structure into the PISA server, the interfacial contact information can be promptly generated. Among the rich output content, we take the interfacial residues, the contact pairs and the distances for hydrogen bonds or salt bridges. This information is used in CNN feature encoding process to label whether a specific residue is at interface or involved in hydrogen bonds and salt bridges, such that the weight of the feature could be adjusted independently.

Dataset generation. The total number of all available ACE2 and RBD variants sum up to 93,669. There are predominantly more worsening (97.02%) than improving (2.38%) or neutral (0.60%) variants. Absorbing all of the variants will create imbalanced dataset that may result in models with poor performance. To tackle this problem, we chose to include all the improving variants with up to six amino acid changes from the RBD and ACE2 variants, and randomly picking three times as many worsening or neutral variants accordingly. This process creates a dataset of 8,440 variants where the number (percentage) of improving, worsening, neutral variants are 2,212 (26.21%), 5,667 (67.14%), 561 (6.65%), respectively. This dataset contains 6,105 (72.10%) variants with RBD mutations, 2335 (27.90%) variants with ACE2 mutations, 1341 (15.89%) variants with single amino acid changes, and 7099 (84.11%) variants with multiple amino acid changes.

Hyperparameter optimization. The Bayesian hyperparameter optimization was performed by utilizing the Hyperopt package (http://hyperopt.github.io/hyperopt) to achieve optimal performance of the CNN_seq model. The loss function for minimization was defined as:

$$loss = (\%\mathrm{VC}_{\mathrm{training}} - \%\mathrm{VC}_{\mathrm{validation}})/10 + (\mathrm{MSE}_{\mathrm{validation}} - \mathrm{MSE}_{\mathrm{training}})/0.05 \quad (1)$$

The loss function numerically evaluates the gap in performance of CNN_seq model on training and validation set, and the constants 10 and 0.05 were chosen to match the difference in %VC and MSE with a roughly equal contribution. A total of 50 iterations of optimization were performed to achieve the final set of hyperparameters that are summarized in Table S7 in *SI Appendix*.

Model training, evaluation, and prediction. The model was trained with the objective function as the MSE between predicted and target $K_{\mathrm{D,app}}$ ratio values. The Adam optimizer[113] was used to perform the backpropagation and the training was performed for 5,000 epochs including the

entire training data in each batch. To measure the strength of correlation between experimental and predicted $K_{D,app}$ ratios, we calculate the Pearson correlation coefficient *r*, which is defined as:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (2)$$

where x_i and y_i are the target and predicted value for the i^{th} sample, and $\bar{x}$ and $\bar{y}$ are the mean value for all x_i targets and y_i predictions. Note that, due to classification ambiguity, variants with either target value or predicted value of $K_{D,app}$ ratio equals to 1 were excluded during the calculation of %VC, but all variants were included during the calculation of *r*.

The final predictions of CNN_seq model in Table 1 are calculated using a single model trained on 100% of the variants selected from the full database. The variants in Table 1 were excluded from the training data so that a fair evaluation can be made.

Rosetta calculations for $\Delta\Delta G_{bind}$ prediction. To computationally assess the binding affinity of the Omicron-RBD with hACE2, we used rigorous molecular-mechanics based calculations using the Rosetta force-field. Assuming that the Omicron-RBD still binds to hACE2 at the same binding site as the WT-RBD, we first made the initial complex of Omicron-RBD-hACE2 by making all the 15 amino-acid changes.[6] To account for structural re-arrangements of the mutated RBD in the complex, the initial complex was subject to 100 independent *Relax* trajectories. Harmonic constraints were used to prevent the structure from deviating significantly from the crystal structure. At the end of *Relax*, a gradient minimization is performed using *lbfgs_armijo* algorithm for 2000 steps after which the relevant metrics of binding were calculated using *InterfaceAnalyzer*. The binding energy, ΔG of the Omicron-RBD-hACE2 complex was then calculated as the average of *dG_separated* scores obtained from the 100 *Relax* simulations.

Data Availability. Computational codes were developed in Python using the PyTorch library. All data pertaining to the results discussed in the paper are available either in the main text and *SI Appendix*. Relevant simulation codes for generating the models are deposited in the GitHub repository (https://github.com/maranasgroup/CNN_seq_CoV2).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This activity was primarily supported by the United States Department of Agriculture (USDA) NIFA Award 2020-67015-32175 and also partially enabled by funding provided by The Center for Bioenergy Innovation a U.S. Department of Energy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science (DE-AC05-000R22725).

Computations for this research were performed on the Pennsylvania State University's Institute for Computational and Data Sciences' Roar supercomputer. We also acknowledge Seed grant funding from the Penn State Huck Institutes of life sciences (to SVK).

Kurt Vandegrift was partially supported by a National Science Foundation Ecology and Evolution of Infectious Diseases program grant (# 1619072) as well as National Institute of Allergy and Infectious Diseases, National Institutes of Health grants (#1R21AI156406-01 & #1R01AI134911).

Author Approvals

All the authors approve of the work.

Competing financial interests

The authors declare no competing financial interests.

Appendix A. Supplementary data

Supplementary data to this article are given in "Supplementary Information.pdf"

References

1. Kupferschmidt, K. Evolving threat. *Science* **373**, 844–849 (2021).
2. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
3. Tao, K. *et al.* The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet.* **22**, 757–773 (2021).
4. Winger, A. & Caspari, T. The Spike of Concern—The Novel Variants of SARS-CoV-2. *Viruses* **13**, 1002 (2021).
5. Colson, P. *et al.* Emergence in Southern France of a new SARS-CoV-2 variant of probably Cameroonian origin harbouring both substitutions N501Y and E484K in the spike protein. *medRxiv* 2021.12.24.21268174 (2021). doi:10.1101/2021.12.24.21268174
6. Centre, E. Implications of the emergence and spread of the SARS-CoV-2 B . 1 . 1 . 529 variant of concern (Omicron) for the EU / EEA Event background. 1–7 (2021).
7. Pulliam, J. R. C. *et al.* Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa. *Science* (2022). doi:10.1126/science.abn4947
8. Lyngse, F. P. *et al.* Transmission of SARS-CoV-2 Omicron VOC subvariants BA.1 and BA.2: Evidence from Danish Households. *medRxiv* 2022.01.28.22270044 (2022). doi:10.1101/2022.01.28.22270044
9. Meekins, D. A., Gaudreault, N. N. & Richt, J. A. Natural and Experimental SARS-CoV-2 Infection in Domestic and Wild Animals. *Viruses* **13**, 1993 (2021).
10. World Organisation for Animal Health (OIE). Technical Factsheet Infection With Sars-Cov-2 in Animals. *World Organ. Anim. Heal.* **2**, 1–4 (2020).
11. Cool, K. *et al.* Infection and transmission of ancestral SARS-CoV-2 and its alpha variant in pregnant white-tailed deer. *Emerg. Microbes Infect.* **11**, 95–112 (2022).
12. Palmer, M. V. *et al.* Susceptibility of White-Tailed Deer (Odocoileus virginianus) to SARS-CoV-2. *J. Virol.* **95**, (2021).
13. Chandler, J. C. *et al.* SARS-CoV-2 exposure in wild white-tailed deer (Odocoileus virginianus). *Proc. Natl. Acad. Sci.* **118**, e2114828118 (2021).
14. Kuchipudi, S. V. *et al.* Multiple spillovers from humans and onward transmission of SARS-CoV-2 in white-tailed deer. *Proc. Natl. Acad. Sci.* **119**, e2121644119 (2022).
15. Hale, V. L. *et al.* SARS-CoV-2 infection in free-ranging white-tailed deer. *Nature* **602**, 481–486 (2022).
16. Marques, A. D. *et al.* Evolutionary Trajectories of SARS-CoV-2 Alpha and Delta Variants in White-Tailed Deer in Pennsylvania. *medRxiv* 2022.02.17.22270679 (2022). doi:10.1101/2022.02.17.22270679
17. USDA. Confirmed cases of SARS-CoV-2 in animals in the United States. *United States Department of Agriculture* (2022). Available at: https://www.aphis.usda.gov/aphis/dashboards/tableau/sars-dashboard.

18. Pickering, B. *et al.* Highly divergent white-tailed deer SARS-CoV-2 with potential deer-to-human transmission. *bioRxiv* 2022.02.22.481551 (2022). doi:10.1101/2022.02.22.481551

19. Kotwa, J. D. *et al.* First detection of SARS-CoV-2 infection in Canadian wildlife identified in free-ranging white-tailed deer (Odocoileus virginianus) from southern Québec, Canada. *bioRxiv* 2022.01.20.476458 (2022). doi:10.1101/2022.01.20.476458

20. Yen, H. *et al.* Transmission of SARS-CoV-2 (Variant Delta) from Pet Hamsters to Humans and Onward Human Propagation of the Adapted Strain: A Case Study. *SSRN Electron. J.* **2**, (2022).

21. Munnink, B. B. O. *et al.* Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* **371**, 172–177 (2021).

22. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914-921.e10 (2020).

23. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).

24. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **19**, 409–424 (2021).

25. Daniloski, Z. *et al.* The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. *Elife* **10**, 1–16 (2021).

26. Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **592**, 116–121 (2021).

27. Zhang, L. *et al.* SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat. Commun.* **11**, 6013 (2020).

28. Ozono, S. *et al.* SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. *Nat. Commun.* **12**, 1–9 (2021).

29. Thomson, E. C. *et al.* Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell* **184**, 1171-1187.e20 (2021).

30. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310.e20 (2020).

31. Zhu, X. *et al.* Cryo-electron microscopy structures of the N501Y SARS-CoV-2 spike protein in complex with ACE2 and 2 potent neutralizing antibodies. *PLOS Biol.* **19**, e3001237 (2021).

32. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).

33. Gobeil, S. M. C. *et al.* Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity. *Science* **373**, (2021).

34. Ku, Z. *et al.* Molecular determinants and mechanism for antibody cocktail preventing SARS-CoV-2 escape. *Nat. Commun.* **12**, 1–13 (2021).

35. Baum, A. *et al.* Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational

escape seen with individual antibodies. *Science* **369**, 1014–1018 (2020).

36. Weisblum, Y. *et al.* Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *Elife* **9**, (2020).
37. Zhou, D. *et al.* Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell* **184**, 2348-2361.e6 (2021).
38. Liu, Z. *et al.* Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell Host Microbe* **29**, 477-488.e4 (2021).
39. McCallum, M. *et al.* Structural basis of SARS-CoV-2 Omicron immune evasion and receptor engagement. *Science* **375**, 864–868 (2022).
40. Jalali, N., Brustad, H. K., Frigessi, A., Macdonald, E. & Meijerink, H. Increased household transmission and immune escape of the SARS-CoV-2 Omicron variant compared to the Delta variant : 1–9 (2022).
41. Hu, J. *et al.* Increased immune escape of the new SARS-CoV-2 variant of concern Omicron. *Cell. Mol. Immunol.* **19**, 293–295 (2022).
42. Chan, K. K. *et al.* Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science* **369**, 1261–1265 (2020).
43. Chowdhury, R., Boorla, V. S. & Maranas, C. D. Computational biophysical characterization of the SARS-CoV-2 spike protein binding with the ACE2 receptor and implications for infectivity. *Comput. Struct. Biotechnol. J.* **18**, 2573–2582 (2020).
44. Leaver-Fay, A. *et al.* Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. in *Methods in Enzymology* **487**, 545–574 (Academic Press Inc., 2011).
45. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
46. Jacobson, M. P. *et al.* A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins Struct. Funct. Genet.* **55**, 351–367 (2004).
47. Jacobson, M. P., Friesner, R. A., Xiang, Z. & Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **320**, 597–608 (2002).
48. Li, J. *et al.* The VSGB 2.0 model: A next generation energy model for high resolution protein structure modeling. *Proteins Struct. Funct. Bioinforma.* **79**, 2794–2812 (2011).
49. Wang, E. *et al.* End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chem. Rev.* **119**, 9478–9508 (2019).
50. Beard, H., Cholleti, A., Pearlman, D., Sherman, W. & Loving, K. A. Applying physics-based scoring to calculate free energies of binding for single amino acid mutations in protein-protein complexes. *PLoS One* **8**, e82849 (2013).
51. Warshel, A. & Levitt, M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **103**, 227–249 (1976).

52. Senn, H. M. & Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chemie Int. Ed.* **48**, 1198–1229 (2009).

53. Barlow, K. A. *et al.* Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. *J. Phys. Chem. B* **122**, 5389–5399 (2018).

54. Wang, M., Cang, Z. & Wei, G.-W. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nat. Mach. Intell.* **2**, 116–123 (2020).

55. Geng, C., Xue, L. C., Roel-Touris, J. & Bonvin, A. M. J. J. Finding the ΔΔG spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **9**, 1–14 (2019).

56. Gao, W., Mahajan, S. P., Sulam, J. & Gray, J. J. Deep Learning in Protein Structural Modeling and Design. *Patterns* **1**, 100142 (2020).

57. Xu, Y. *et al.* Deep Dive into Machine Learning Models for Protein Engineering. *J. Chem. Inf. Model.* **60**, 2773–2790 (2020).

58. Chen, C. *et al.* Computational prediction of the effect of amino acid changes on the binding affinity between SARS-CoV-2 spike RBD and human ACE2. *Proc. Natl. Acad. Sci.* **118**, e2106480118 (2021).

59. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

60. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

61. Blanco, J. D., Hernandez-Alias, X., Cianferoni, D. & Serrano, L. In silico mutagenesis of human ACE2 with S protein and translational efficiency explain SARS-CoV-2 infectivity in different species. *PLoS Comput. Biol.* **16**, e1008450 (2020).

62. Zahradník, J. *et al.* SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. *Nat. Microbiol.* **6**, 1188–1198 (2021).

63. Piplani, S., Singh, P. K., Winkler, D. A. & Petrovsky, N. In silico comparison of SARS-CoV-2 spike protein-ACE2 binding affinities across species and implications for virus origin. *Sci. Rep.* **11**, 13063 (2021).

64. Laurini, E., Marson, D., Aulic, S., Fermeglia, A. & Pricl, S. Computational Mutagenesis at the SARS-CoV-2 Spike Protein/Angiotensin-Converting Enzyme 2 Binding Interface: Comparison with Experimental Evidence. *ACS Nano* **15**, 6929–6948 (2021).

65. Pavlova, A. *et al.* Machine Learning Reveals the Critical Interactions for SARS-CoV-2 Spike Protein Binding to ACE2. *J. Phys. Chem. Lett.* **12**, 5494–5502 (2021).

66. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).

67. Zamyatnin, A. A. Amino Acid, Peptide, and Protein Volume in Solution. *Annu. Rev. Biophys. Bioeng.* **13**, 145–165 (1984).

68. Mei, H., Liao, Z. H., Zhou, Y. & Li, S. Z. A new set of amino acid descriptors and its application in peptide QSARs. *Biopolym. - Pept. Sci. Sect.* **80**, 775–786 (2005).

69. Kawashima, S. *et al.* AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* **36**, 202–205 (2008).

70. Barton, M. I. *et al.* Effects of common mutations in the SARS-CoV-2 Spike RBD and its ligand, the human ACE2 receptor on binding affinity and kinetics. *Elife* **10**, 1–19 (2021).

71. Cameroni, E. *et al.* Broadly neutralizing antibodies overcome SARS-CoV-2 Omicron antigenic shift. *Nature* **602**, 664–670 (2021).

72. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).

73. Tian, X. *et al.* Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. *Emerg. Microbes Infect.* **9**, 382–385 (2020).

74. Tanaka, S. *et al.* An ACE2 Triple Decoy that neutralizes SARS-CoV-2 shows enhanced affinity for virus variants. *Sci. Rep.* **11**, 12740 (2021).

75. Chan, K. K., Tan, T. J. C., Narayanan, K. K. & Procko, E. An engineered decoy receptor for SARS-CoV-2 broadly binds protein S sequence variants. *Sci. Adv.* **7**, 1–9 (2021).

76. Supasa, P. *et al.* Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera. *Cell* **184**, 2201-2211.e7 (2021).

77. Yuan, M. *et al.* Structural and functional ramifications of antigenic drift in recent SARS-CoV-2 variants. *Science* **373**, 818–823 (2021).

78. Liu, C. *et al.* Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. *Cell* **184**, 4220-4236.e13 (2021).

79. Dejnirattisai, W. *et al.* Antibody evasion by the P.1 strain of SARS-CoV-2. *Cell* **184**, 2939-2954.e9 (2021).

80. Amanat, F. *et al.* SARS-CoV-2 mRNA vaccination induces functionally diverse antibodies to NTD, RBD, and S2. *Cell* **184**, 3936-3948.e10 (2021).

81. Augusto, G. *et al.* In vitro data suggest that Indian delta variant B.1.617 of SARS□CoV□2 escapes neutralization by both receptor affinity and immune evasion. *Allergy* **77**, 111–117 (2022).

82. Yi, C. *et al.* 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 (2020).

83. Bayarri-Olmos, R. *et al.* The SARS-CoV-2 Y453F mink variant displays a pronounced increase in ACE-2 affinity but does not challenge antibody neutralization. *J. Biol. Chem.* **296**, 100536 (2021).

84. Lopez, E. *et al.* Simultaneous evaluation of antibodies that inhibit SARS-CoV-2 variants via multiplex assay. *JCI Insight* **6**, 1–13 (2021).

85. Chatterjee, D. *et al.* Antigenicity of the Mu (B.1.621) and A.2.5 SARS-CoV-2 Spikes. *Viruses* **14**, 144 (2022).

86. Zhang, Z. *et al.* Potent prophylactic and therapeutic efficacy of recombinant human

ACE2-Fc against SARS-CoV-2 infection in vivo. *Cell Discov*. **7**, 65 (2021).

87. Wu, L. *et al.* SARS-CoV-2 Omicron RBD shows weaker binding affinity than the currently dominant Delta variant to human ACE2. *Signal Transduct. Target. Ther.* **7**, 2021–2023 (2022).

88. Ramanathan, M., Ferguson, I. D., Miao, W. & Khavari, P. A. SARS-CoV-2 B.1.1.7 and B.1.351 spike variants bind human ACE2 with increased affinity. *Lancet Infect. Dis.* **21**, 1070 (2021).

89. Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).

90. McCallum, M. *et al.* Molecular basis of immune evasion by the Delta and Kappa SARS-CoV-2 variants. *Science* **374**, 1621–1626 (2021).

91. Ren, W. *et al.* Characterization of SARS-CoV-2 variants B.1.617.1 (Kappa), B.1.617.2 (Delta) and B.1.618 on cell entry, host range, and sensitivity to convalescent plasma and ACE2 decoy receptor. *bioRxiv* 2021.09.03.458829 (2021). doi:10.1101/2021.09.03.458829

92. Burley, S. K. *et al.* RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **49**, D437–D451 (2021).

93. Bateman, A. *et al.* UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

94. Waterhouse, A. *et al.* SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).

95. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372**, 774–797 (2007).

96. Velankar, S. *et al.* PDBe: Improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res.* **44**, D385–D395 (2016).

97. Rafferty, B., Flohr, Z. C., & Martini, A. Protein Contact Maps. (2010). doi:10.21981/02DQ-MT84

98. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. 1–18 (2012).

99. WHO. SARS-CoV-2 variants of concern and variants of interest. *World Health Organization* (2021). Available at: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/.

100. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).

101. Mannar, D. *et al.* SARS-CoV-2 Omicron variant: Antibody evasion and cryo-EM structure of spike protein–ACE2 complex. *Science* **375**, 760–764 (2022).

102. VanBlargan, L. A. *et al.* An infectious SARS-CoV-2 B.1.1.529 Omicron virus escapes neutralization by therapeutic monoclonal antibodies. *Nat. Med.* (2022).

doi:10.1038/s41591-021-01678-y

103. Ford, C. T., Jacob Machado, D. & Janies, D. A. Predictions of the SARS-CoV-2 Omicron Variant (B.1.1.529) Spike Protein Receptor-Binding Domain Structure and Neutralizing Antibody Interactions. *Front. Virol.* **2**, 1–11 (2022).

104. Starr, T. N. *et al.* Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. 1–27 (2022).

105. Wang, Q. *et al.* Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell* **181**, 894-904.e9 (2020).

106. Rochman, N. D. *et al.* Epistasis at the SARS-CoV-2 RBD Interface and the Propitiously Boring Implications for Vaccine Escape. *bioRxiv* (2021). doi:10.1101/2021.08.30.458225

107. Li, Q. *et al.* The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* **182**, 1284-1294.e9 (2020).

108. Dejnirattisai, W. *et al.* SARS-CoV-2 Omicron-B.1.1.529 leads to widespread escape from neutralizing antibody responses. *Cell* **185**, 467-484.e15 (2022).

109. Cao, Y. *et al.* Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, (2021).

110. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

111. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinforma.* **78**, 1950–1958 (2010).

112. Bowers, K. J. *et al.* Scalable algorithms for molecular dynamics simulations on commodity clusters. *Proc. 2006 ACM/IEEE Conf. Supercomput. SC'06* (2006). doi:10.1145/1188455.1188544

113. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization. in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (International Conference on Learning Representations, ICLR, 2015).