

# Omicron BA.2 (B.1.1.529.2): high potential to becoming the next dominating variant

Jiahui Chen<sup>1</sup> and Guo-Wei Wei<sup>1,3,4\*</sup>

<sup>1</sup> Department of Mathematics,

Michigan State University, MI 48824, USA.

<sup>2</sup> Department of Electrical and Computer Engineering,

Michigan State University, MI 48824, USA.

<sup>3</sup> Department of Biochemistry and Molecular Biology,

Michigan State University, MI 48824, USA.

February 11, 2022

## Abstract

The Omicron variant of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has rapidly replaced the Delta variant as a dominating SARS-CoV-2 variant because of natural selection, which favors the variant with higher infectivity and stronger vaccine breakthrough ability. Omicron has three lineages or subvariants, BA.1 (B.1.1.529.1), BA.2 (B.1.1.529.2), and BA.3 (B.1.1.529.3). Among them, BA.1 is the currently prevailing subvariant. BA.2 shares 32 mutations with BA.1 but has 28 distinct ones. BA.3 shares most of its mutations with BA.1 and BA.2 except for one. BA.2 is found to be able to alarmingly reinfect patients originally infected by Omicron BA.1. An important question is whether BA.2 or BA.3 will become a new dominating “variant of concern”. Currently, no experimental data has been reported about BA.2 and BA.3. We construct a novel algebraic topology-based deep learning model trained with tens of thousands of mutational and deep mutational data to systematically evaluate BA.2’s and BA.3’s infectivity, vaccine breakthrough capability, and antibody resistance. Our comparative analysis of all main variants namely, Alpha, Beta, Gamma, Delta, Lambda, Mu, BA.1, BA.2, and BA.3, unveils that BA.2 is about 1.5 and 4.2 times as contagious as BA.1 and Delta, respectively. It is also 30% and 17-fold more capable than BA.1 and Delta, respectively, to escape current vaccines. Therefore, we project that Omicron BA.2 is on its path to becoming the next dominating variant. We forecast that like Omicron BA.1, BA.2 will also seriously compromise most existing mAbs, except for sotrovimab developed by GlaxoSmithKline.

Keywords: COVID-19, SARS-CoV-2, Omicron, infectivity, antibody-resistance, vaccine breakthrough,

---

\*Corresponding author. Email: weig@msu.edu

# 1 Introduction

On November 26, 2021, the World Health Organization (WHO) declared the Omicron variant (B.1.1.529) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) initially discovered in South Africa a variant of concern (VOC). Within a few days (i.e., December 1, 2021), an artificial intelligence (AI) model predicted the Omicron variant to be about 2.8 times as infectious as the Delta variant, have a near 90% likelihood to escape current vaccines, and severely compromise the efficacy of monoclonal antibodies (mAbs) developed by Eli Lilly, Regeneron, AstraZeneca, and many others, except for GlaxoSmithKline’s sotrovimab [1]. The subsequent experiments confirm Omicron’s high infectivity [2,3], high vaccine breakthrough rate [4,5], and severe antibody escape rate [6–8]. The U.S. Food and Drug Administration (FDA) halted the use of mAbs from Eli Lilly and Regeneron in January 2022. Due to its combined effects of high infectivity and high vaccine breakthrough rate, the Omicron variant is far more transmissible than the Delta variant and has rapidly become the dominating variant in the world.

Omicron has three lineages, BA.1 (B.1.1.529.1), BA.2 (B.1.1.529.2), and BA.3 (B.1.1.529.3), which were first detected in November 2021 in South Africa [9]. Among them, BA.1 lineage is the preponderance that has ousted Delta. Compared to the reference genome reported in Wuhan, Omicron BA.1 has a total of 60 mutations on non-structure protein (NSP3), NSP4, NSP5, NSP6, NSP12, NSP14, S protein, envelope protein, membrane protein, and nucleocapsid protein. Among them, 32 mutations are on the spike (S) protein, the main antigenic target of antibodies generated by either infection or vaccination. Fifteen of these mutations affect the receptor-binding domain (RBD), whose binding with host angiotensin-converting enzyme 2 (ACE2) facilitates the viral cell entry during the initial infection [10]. BA.2 shares 32 mutations with BA.1 but has 28 distinct ones. On the RBD, BA.2 has four unique mutations and 12 shared with BA.1. In contrast, the Delta variant has only two RBD mutations. BA.3 shares most of its mutations with BA.1 and BA.2, except for one on NSP6 (A88V). It also has 15 RBD mutations, but none is distinct from BA.1 and BA.2. Nationwide Danish data in late December 2021 and early January 2022 indicate that Omicron BA.2 is inherently substantially more transmissible than BA.1 and capable of vaccine breakthrough [11]. Israel reported a handful of cases of patients who were infected with original Omicron BA.1 strain and have reinfectd with BA.2 in a short period [12]. Although BA.2 did not cause worse illness than the original Omicron BA.1 strain, its reinfection is very alarming. It means the antibodies generated from the early Omicron BA.1 were evaded by the BA.2 strain. It is imperative to know whether BA.2 will become the next dominating strain to reinfect the world population.

Currently, there are no experimental results about the infectivity, vaccine breakthrough, and antibody resistance of BA.2 and BA.3 [13]. In this work, we present a comprehensive analysis of Omicron BA.2 and BA.3’s potential of becoming the next prevailing SARS-CoV-2 variant. Our study focuses on the S protein RBD, which is essential for virus cell entry. Studies show that binding free energy (BFE) between the S RBD and the ACE2 is proportional to the viral infectivity [10,14,15]. In July 2020, it was discovered that SARS-CoV-2 evolution is governed by infectivity-based natural selection [10], which was conformed beyond doubt in April 2021 [17]. The RBD is not only crucial for viral infectivity but also essential for vaccines and antibody protections. An antibody that can disrupt the RBD-ACE2 binding would directly neutralize the virus [18–20]. We integrate tens of thousands of mutational and deep mutational data, biophysics, and algebraic topology to construct an AI model. We systematically investigate the binding free energy (BFE) changes of an RBD-ACE2 complex structure and a library of 185 structures of RBD-antibody complexes induced by the RBD mutations of Alpha, Beta, Gamma, Delta, Lambda, Mu, BA.1, BA.2, and BA.3 to reveal their infectivity, vaccine-escape potential, and antibody resistance. Using our comparative analysis, we unveil that the Omicron BA.2 variant is about 1.5 times as infectious as BA.1 and about 4.2 times as contagious as the Delta variant. It also has a 30% higher potential than BA.1 to escape existing vaccines. Therefore, we project the Omicron BA.2 is on its path to becoming the next dominating variant.

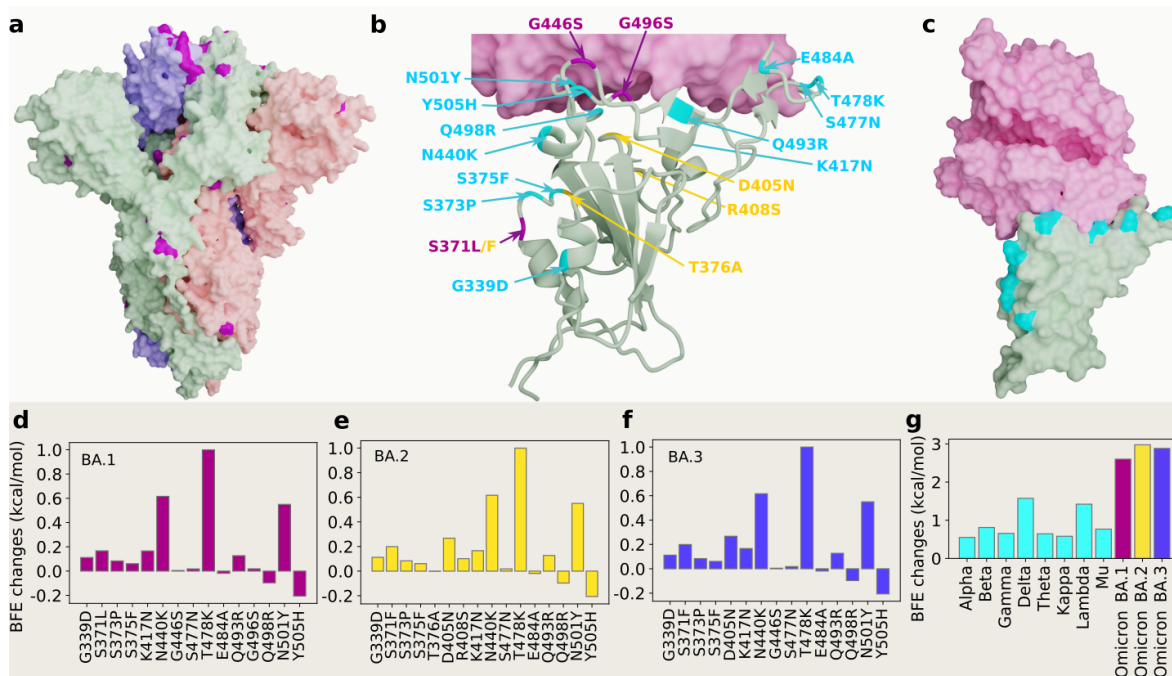


Figure 1: 3D structures of Omicron strains, their ACE2 complexes and their mutation-induced BFE changes. **a** Spike protein (PDB: 7WK2 [3]) with Omicron mutations being marked yellow. **b** BA.1 and BA.2 RBD mutations at the RBD-ACE interface (PDB: 7T9L [21]). The shared 12 mutations are labeled in cyan, BA.1 mutations are marked with magenta, and distinct BA.2 mutations are plotted in yellow. **c** The structure of the RBD-ACE2 complex with mutations on cyan spots. **d**, **e**, **f** and **g** BFE changes induced by mutations of Omicron BA.1, BA.2, BA.3, respectively. **h** a comparison of predicted mutation-induced BFE changes for few SARS-CoV-2 variants.

## 2 Results

### 2.1 Infectivity

Figure 1 **a** shows the three-dimensional (3D) structure of Omicron BA.1 [3]. At the RBD, Omicron BA.1, BA.2 and BA.3 share 12 RBD mutations, i.e., G339D, S373P, S375F, K417N, N440K, S477N, T478K, E484A, Q493R, Q498R, N501Y, and Y505H as shown in Figure 1 **b**. However, BA.1 has distinct RBD mutations S371L, G446S, and G496S, BA.2 has S371F, T376A, D405N, and R408S, and BA.3 has S371F, D405N, and G446S. Figures 1 **d**, **e** and **f** present the BFE changes of the RBD-ACE2 complex induced by the RBD mutations of Omicron AB.1, BA.2 and BA.3, respectively. The larger the BFE change is, the higher infectivity will be. Since natural selection favors those mutations that strengthen the viral infectivity [10], the most contagious variant will become dominant in a population under the same competing condition. The accumulated BFE changes are summarized in Figure 1 **g**. A comparison is given to other main SARS-CoV-2 variants Alpha, Beta, Gamma, Delta, Theta, Kappa, Lambda, and Mu. The Delta variant had the highest BFE change among the earlier variants and was the most infectious variant before the occurrence of the Omicron variant, which explains its dominance in 2021. Omicron BA.1, BA.2, and BA.3 have BFE changes of 2.60, 2.98, and 2.88 kcal/mol, respectively, which are much higher than those of other major SRAS-CoV-2 variants. Among them, Omicron BA.2 is the most infectious variant and is about 20 and 4.2 times as infectious as the original SARS-CoV-2 and the Delta variant, respectively. Our model predicts that BA.2 is about 1.5 as contagious BA.2, which is the same as reported in an initial study [12]. Another report confirms that Omicron BA.2 is more contagious than BA.1 [11]. Therefore, Omicron BA.2 may eventually replace the original Omicron strain BA.1 in the world.

## 2.2 Vaccine breakthrough

Omicron BA.1 is well-known for its ability to escape current vaccines [5, 6]. Its 15 mutations at the RBD enable it to not only strengthen its infectivity by a stronger binding to human ACE2 but also create mismatches for most direct neutralization antibodies generated from vaccination or prior infection. Although BA.1, BA.2, and BA.3 share 12 RBD mutations, BA.1 has 3 additional RBD mutations, BA.2 has 4 additional RBD mutations, and BA.3 has one mutation the same as that of BA.1’s additional ones and two mutations the same as those of BA.2’s additional ones. Therefore, it is important to understand their vaccine-escape potentials. Currently, no experimental result has been reported about the vaccine-breakthrough capability of BA.2 and BA.3.

Experimental analysis of the variant vaccine-escape capability over the world’s populations is subject to many uncertainties. Different vaccines may stimulate different immune responses and antibodies for the same person. Different individuals may have different immune responses and antibodies from the same vaccine due to their different races, gender, age, and underlying medical conditions. Uncontrollable experimental conditions and different experimental methods may also contribute to uncertainties. Consequently, it is impossible to accurately characterize a variant’s vaccine-escape capability (or rate) over the world’s populations.

In our work, we take an integrated approach to understanding the intrinsic vaccine-escape capability of SARS-CoV-2 variants. We collect a library of 185 known antibody and S protein complexes and analyze the mutational impact on the binding of these complexes [1, 9]. The results in terms of mutation-induced BFE changes serve as the statistical ensemble analysis of Omicron subvariants’ vaccine-breakthrough potentials. This molecular-level analysis becomes very useful when it is systematically applied to a series of variants.

Figures 2 **a**, **b1**, and **b2** depict the BFE changes of ACE2-RBD and 185 antibody-RBD complexes induced by the RBD mutations from SARS-CoV-2 variants. The first bunch of 7 mutations is associated with Alpha, Beta, Gamma, Delta, Lambda, and Mu. The second bunch of 12 mutations is shared among BA.1, BA.2, and BA.3. The next bunch of 3 mutations is associated with BA.1. The last bunch of 4 mutations belongs to BA.2. Binding-strengthening mutations give rise to positive BFE changes, while binding-weakening mutations lead to negative BFE changes. Obviously, shared Omicron mutations K417N, E484A, and Q493R are very disruptive to many antibodies. BA.1 mutation G496S is also quite disruptive. BA.2 mutations T376A, D405N, and R408S may reduce the efficacy of many antibodies. Apparently, these complexes are significantly impacted by Omicron BA.1, BA.2, and BA.3 RBD mutations. Overall, Figure 2 shows more negative BFE changes than positive ones, suggesting Omicron BA.1, BA.2, and BA.3 mutations enable the breakthrough of current vaccines.

Statistical analysis of the BFE changes of 185 antibody-RBD complexes induced by BA.1, BA.2, BA.3, and Delta RBD mutations is presented in Figure 3 and analysis of Alpha, Beta, Gamma, Lambda, and Mu is presented in Figure S2. Accumulated BFE changes are provided in Figure 3 **a1**, **b1**, and **c1**. Obviously, all Omicron subvariants have more negative accumulated BFE changes than positive ones, showing their antibody resistance. Among them, BA.2’s distribution is extended to a wider negative domain, showing its strongest antibody resistance. In contrast, Delta variant’s statistics is given in Figure 3 **d1**, showing a smaller domain of distribution.

As discussed earlier, it is difficult to obtain a variant’s true vaccine-escape rate over world’s populations. However, a molecular-based comparative analysis can offer desirable information. Figures 3 **a2**, **b2**, **c2**, and **d2** depict the number of antibody-RBD complexes that is regarded as disrupted by BA.1, BA.2, BA.3, and Delta mutations, respectively, under different thresholds ranging from 0 kcal/mol, -0.3 kcal/mol, to <-3 kcal/mol. Previously, threshold -0.3 kcal/mol was used to decide whether a mutation disrupts an antibody-RBD complex [1], which gives rise to 163, 168, and 164 disrupted antibody-RBD complexes, respectively for BA.1, BA.2, and BA.3. The corresponding rates of potential vaccine breakthrough are 0.88, 0.91, and 0.89 for BA.1, BA.2, and BA.3, respectively. Therefore, BA.2 is slightly more antibody resistant than

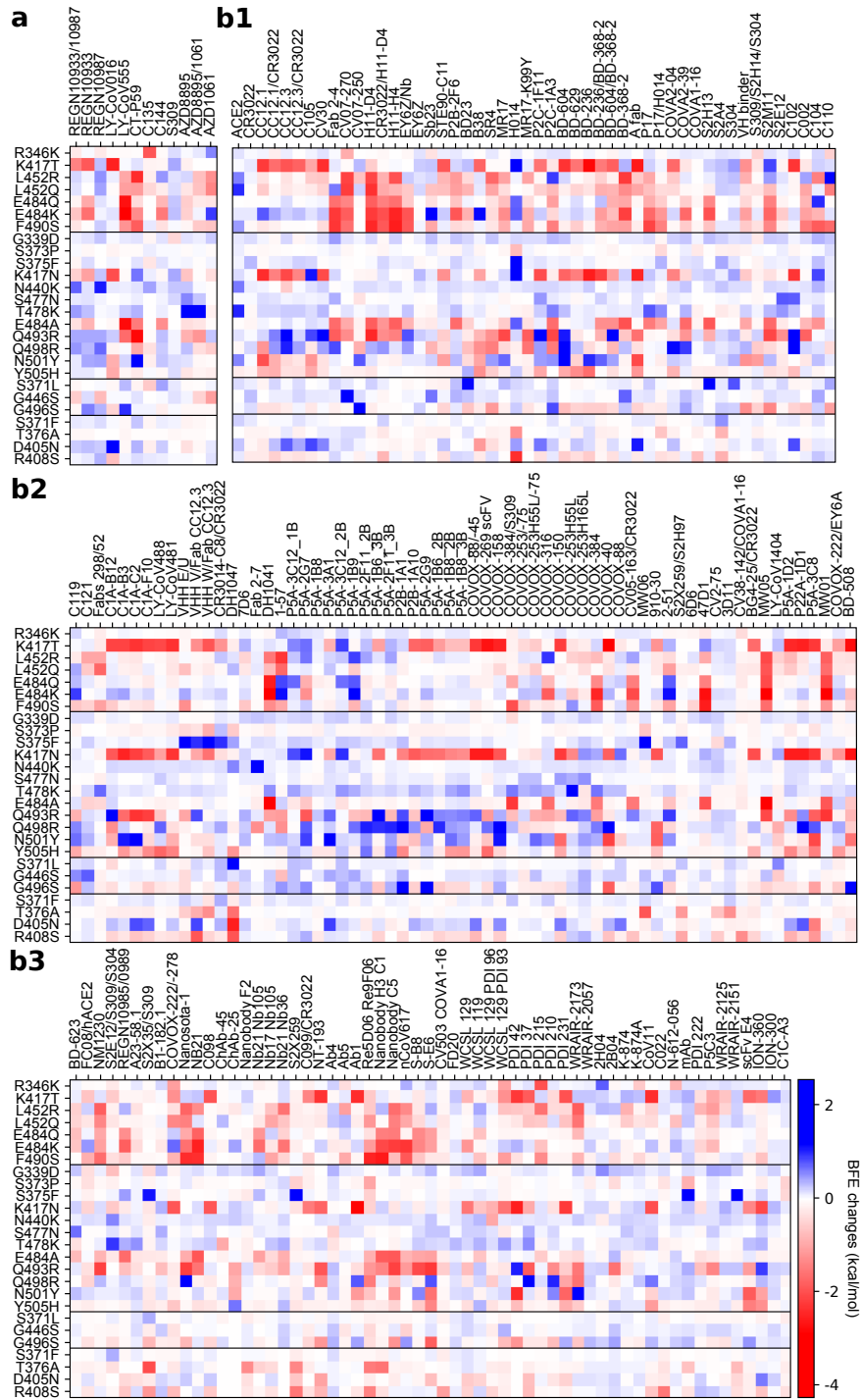


Figure 2: Illustration of mutation-induced BFE changes of 185 antibody-RBD complexes and an ACE2-RBD complex. Positive changes strengthen the binding, while negative changes weaken the binding. **a** Heat map for 12 antibody-RBD complexes in various stages of drug development. Gray color stands for no predictions due to incomplete structures. **b1** Heat map for ACE2-RBD and antibody-RBD complexes. **b2** and **b3** Heat map for antibody-RBD complexes. The first 7 mutations are associated earlier SARS-CoV-2 variants. The next 12 mutations are shared among BA.1, BA.2, and BA.3 strains. The next three mutations are distinct to BA.1, and the final bunch of 4 mutations belong to BA.2.

BA.1. As a reference, the Delta variant may disrupt 70 out of 185 antibody-RBD complexes, suggesting a vaccine-breakthrough rate of 0.37.

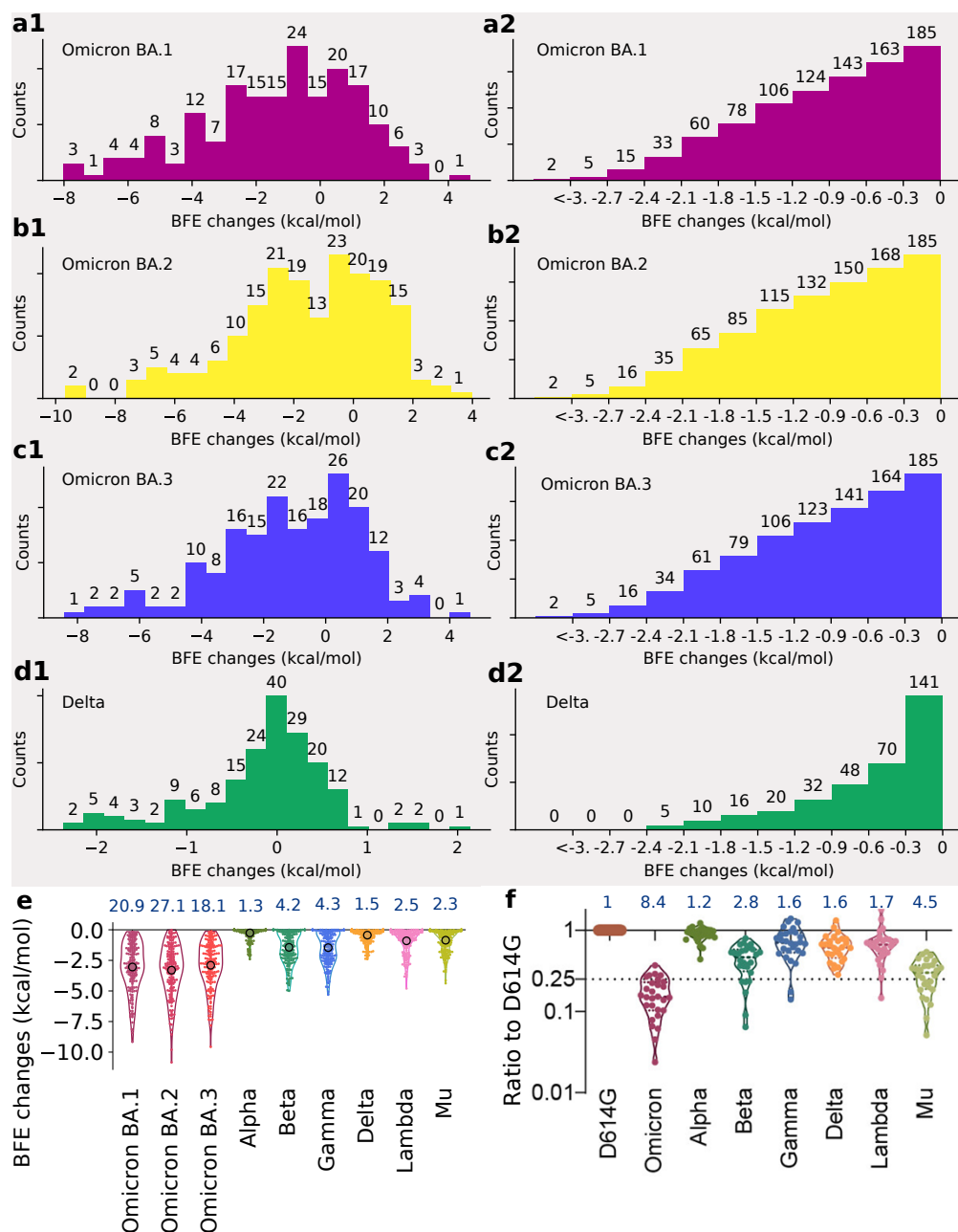


Figure 3: Analysis of variant mutation-induced BFE changes of ACE2-RBD and 185 antibody-RBD complexes. **a1**, **b1**, **c1**, and **d1** The distributions (counts) of accumulated BFE changes induced by Omicron BA.1, BA.2, BA.3, and Delta mutations respectively for 185 antibody-RBD complexes. For each case, there are more mutation-weakened complexes than mutation-strengthened complexes. **a2**, **b2**, **c2**, and **d2** The numbers of antibody-RBD complexes regarded as disrupted by BA.1, BA.2, BA.3, and Delta mutations respectively under different thresholds ranging from 0 kcal/mol, -0.3 kcal/mol, to <-3 kcal/mol. **e** Accumulated negative BFE changes induced by BA.1, BA.2, BA.3, Alpha, Beta, Delta, Gamma, Lambda, and Mu mutations respectively for 185 antibody-RBD complexes. For each variant, the number on the top is the fold of binding affinity reduction computed by  $e^{-\text{BFEchange}_{\text{average}}}$ , where  $\text{BFEchange}_{\text{average}}$ , marked by a circle, is the mean value of negative BFE changes for 185 antibody-RBD complexes. **f** The comparison of neutralization activity against Omicron (BA.1), Alpha, Beta, Delta, Gamma, Lambda, and Mu variants based on 28 convalescence sera [5]. For each variant, the number on the top is the ratio of neutralization  $\text{ED}_{50}$  compared to the reference strain D614G.

It is interesting to compare our analysis with experimental results [5]. In Figure 3 f, the sensitivity of 28 serum samples from COVID-19 convalescent patients infected with an earlier SARS-CoV-2 strain (D614G) was tested against pseudotyped Omicron, Alpha, Beta, Gamma, Delta, Lambda, and Mu [5]. The

results indicate the Omicron (BA.1) and Delta variant have 8.4 and 1.6 fold reductions, respectively, to the mean neutralization ED50 of these sera compared with the D614G reference strain. Figure 3 **e** presents a comparison of accumulated negative BFE changes for variants Omicron BA.1, BA.2, BA.3, Alpha, Beta, Delta, Gamma, Lambda, and Mu. For each antibody-RBD complex, we only consider disruptive effects by setting positive BFE changes to zero and sum over RBD mutations (e.g., 15 mutations for Omicron BA.1 and 2 for Delta) to obtain the accumulated negative BFE change. As such, we have 185 accumulated negative BFE changes for each variant. We use the mean of these 185 values to compute the fold of affinity reduction, which can be compared for different variants against the original virus reported in Wuhan ( $\text{BFE}_{\text{change}_{\text{average}}} = 0$ ). The RBD mutations of the Delta variant cause 1.5 fold reduction in the neutralization capability. In the same setting, Omicron BA.1, BA.2, and BA.3 may lead to about 21, 27, and 18 fold increases in their vaccine-breakthrough capabilities. As such, BA.2 is about 30% more capable to escape existing vaccines than BA.1 and 17 times more than the Delta variant. Our prediction has a correlation coefficient of 0.9 with the experiment. With its highest infectivity and highest vaccine-escape potential, the Omicron BA.2 is set to take over the Omicron BA.1 in infecting the world population.

### 2.3 Antibody resistance

The design and discovery of mAbs are part of an important achievement in combating COVID-19. Unfortunately, like vaccines, mAbs are prone to viral mutations, particularly antibody-resistant ones. Early studies predicted that Omicron BA.1 would compromise the anti-COVID-19 mAbs developed by Eli Lilly, Regeneron, AstraZeneca, Celltrion, and Rockefeller University [1]. However, Omicron BA.1’s impact on GlaxoSmithKline’s mAb, called sotrovimab, was predicted to be mild [1]. These predictions have been confirmed and the FDA has halted the use of Eli Lilly and Regeneron’s COVID-19 mAbs. Currently, GlaxoSmithKline’s sotrovimab is the only antibody-drug authorized in the U.S. for the treatment of COVID-19 patients infected by the Omicron variant. An important question is whether sotrovimab remains effective for the BA.2 subvariant that might drive a new wave of infections in the world population.

In this work, we further analyze the efficacy of these mAbs for BA.2 and BA.3. Our studies focus on Omicron subvariants’ RBD mutations, which appear to be optimized by the virus to evade host antibody protection and infect the host cell. Figure 4 provides a comprehensive analysis of the BFE changes of various antibody-RBD complexes induced by Omicron BA.1, BA.2, and BA.3. Since BA.3 subvariant’s RBD mutations are the subsets of those of BA.1 and BA.2, we only present 19 unique RBD mutations. Impacts of twelve shared RBD mutations are labeled with cyan, those of three additional BA.1 RBD mutations are marked with magenta, and those of four additional BA.2 RBD mutations are plotted in yellow. Figures 4 **a1**, **b1**, **c1**, **d1**, **e1**, **f1** and **g1** depict 3D antibody-RBD complexes for mAbs from Eli Lilly (LY-CoV016 and LY-CoV555), Regeneron (REGN10933, REGN10987, and REGN10933/10987), AstraZeneca (AZD1061 and AZD8895), Celltrion (CT-P59), Rockefeller University (C135, C144), and GlaxoSmithKline (S309), respectively. The ACE2 is included in these plots as a reference.

Figures 4 **a2** and **a3** show that LY-CoV016 is disrupted by shared mutation K417N and LY-CoV555 is weakened by shared mutations E484A and Q493R. Additional mutations from BA.2 may not significantly affect Eli Lilly mAbs. However, if BA.2 become dominant, Eli Lilly mAbs would still be ineffective.

The impacts of BA.1 and BA.2 mutations on Regeneron’s mAbs are illustrated in Figures 4 **b2**, **b3** and **b4**. REGN10933 is undermined by shared mutations N417K and E484A. REGN10987 is disrupted by BA.1 mutation G446S. The antibody cocktail is undermined by shared Omicron mutations as well, which implies Regeneron’s mAbs would still be compromised should Omicron BA.2 become a dominant SRAS-CoV-2 subvariant.

BA.1 and BA.2’s impacts on AstraZeneca’s AZD1061 and AZD8895 are demonstrated in Figures 4 **c2**, **c3** and **c4**. It is noticed that BA.1 mutation G446S has a disruptive effect on AZD1061. AZD8895 is weakened

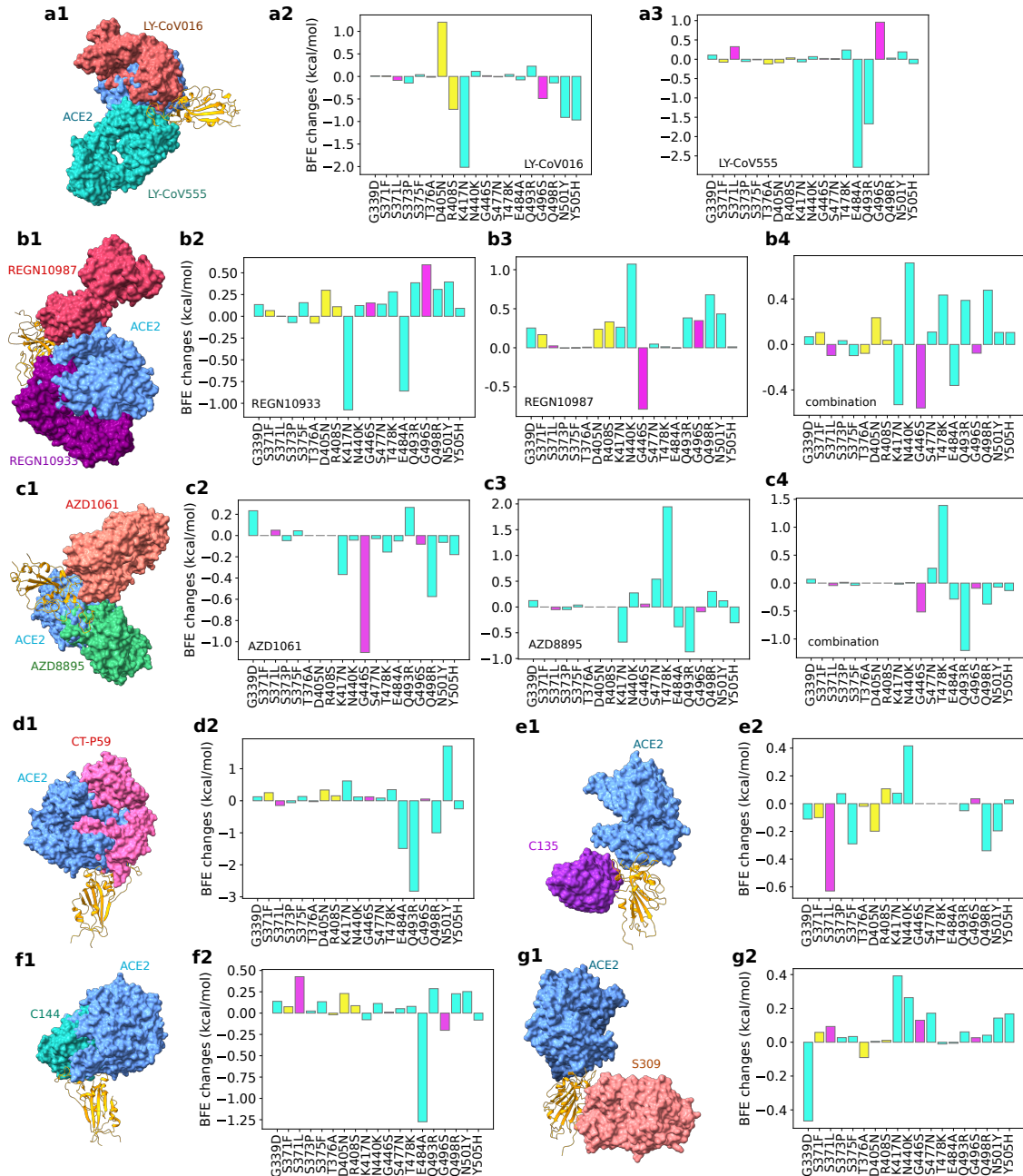


Figure 4: Illustration of Omicron BA.1 and BA.2 RBD mutational impacts on clinical mAbs. **a1**, **b1**, **c1**, **d1**, **e1**, **f1** and **g1** depict the 3D structures of antibody-RBD complexes of Eli Lilly LY-CoV555 (PDB ID: 7KMG [23]) and LY-CoV016 (PDB ID: 7C01 [24]), Regeneron REGN10987 and REGN10933 (PDB ID: 6XDG [25]), AstraZeneca AZD1061 and AZD8895 (PDB ID: 7L7E [26]), Celltrion CT-P59 (aka Regdanvimab, PDB ID: 7CM4), Rockefeller University C135 (PDB ID: 7K8Z) and C144 (PDB ID: 7K90), and GlaxoSmithKline S309 (PDB ID: 6WPS), respectively. In all plots, the ACE2 structure is aligned as a reference. Omicron BA.1 and BA.2 RBD mutation-induced BFE changes (kcal/mol) are given in **a2** and **a3** for Eli Lilly mAbs, **b2**, **b3** and **b4** for Regeneron mAbs, **c2**, **c3**, and **c4** for AstraZeneca mAbs, **d2** for Celltrion CT-P59, **e2** and **f2** for Rockefeller University mAbs, and **g2** for GlaxoSmithKline S309, respectively. Cyan bars label the BFE changes induced by twelve RBD mutations shared by BA.1, BA.2, and BA.3 subvariants. Magenta bars mark the BFE changes induced by three additional BA.1 RBD mutations. Yellow bars denote the BFE changes induced by four additional BA.2 RBD mutations.

by two shared mutations. The AZD1061-AZD8895 combination is also disrupted by shared mutation Q493R. Therefore, the efficacy of AstraZeneca’s mAbs would be reduced should BA.2 prevail in world populations.

As shown in Figure 4 **d2**, Celltrion’s mAb CT-P59 is prone to shared mutations Q493R and E484A. BA.2



mutations may not bring additional destruction. However, the shared mutations pose a threat to Celltrion’s mAb, which implies its efficacy would not restore should BA.2 prevail.

Figures 4 **e2** and **f2** present BA.1 and BA.2’s mutational impacts on Rockefeller University’s mAbs. C135 is mainly disrupted by Omicron BA.1 and its C144 is made ineffective by shared mutation E484A. Therefore, C135 might become effective if BA.2 dominates.

Finally, we plot mutational impacts on antibody S309’s binding with RBD in Figure 4 **g2**. Antibody S309 is the parent antibody for Sotrovimab developed by GlaxoSmithKline and Vir Biotechnology, Inc. It is seen from the figure that there is only one disruptive BFE change of -0.47kcal/mol and the rest of the BFE changes are mostly positive. The BA.2 mutations have little effect on S309. Therefore, we expect a mild effect from Omicron BA.1 and BA.2 on sotrovimab.

It is interesting to understand why S309 is the only antibody that is not significantly affected by Omicron variants. Figure 4 show that all mAbs that compete with the human ACE2 for the receptor-binding motif (RBM) are seriously compromised by Omicron subvariants because most of the RBD mutations locate at the RBM. A possible reason is that Omicron subvariants had optimized RBD mutations at the RBM to strengthen the viral infectivity and evade the direct neutralization antibodies. Consequently, all mAbs that target RBM are seriously compromised by Omicron subvariants. Figures 4 **e1** and **g1** show that antibodies C135 and S309 do not directly compete with ACE2 for the RBM. However, C135 is still very close to the RBM and significantly weakened by some Omicron mutations. In contrast, S309 is further away from the RBM and escapes from Omicron’s RBD mutations.

### 3 Materials and Methods

The deep learning model is designed for predicting mutation-induced BFE changes of the binding between protein-protein interactions. A series of three steps consist of training data preparation, feature generations, and deep neural network training and prediction (see Figure S2). Here, we briefly discuss each step and leave more details in Supporting Information. Readers are also suggested literature [5, 10, 27] for more details.

Firstly, the training data is prepared to comprise experimental BFE changes and next-generation sequencing data. SKEMPI 2.0 [1] is the fundamental BFE change dataset. Additionally, SARS-CoV-2 related datasets are the mutational scanning data of the ACE2-RBD complex [3, 4, 30] and the CTC-445.2-RBD complex [4]. Next is to prepare the features. It is required a variety of biochemical, biophysical, and mathematics features from PPI complex structures, such as surface areas, partial charges, van der Waals interaction, Coulomb interactions, pH values, electrostatics, persistent homology, graph theory, etc. [10, 12] A detailed list and description of these features are provided in Supporting Information. In the following, the key idea of the element-specific and site-specific persistent homology is illustrated briefly. As the persistent homology [34, 35] introduced as a useful tool for data analysis for scientific and engineering applications, it is further applied to molecular studies [27, 36]. For 3D structures, atoms are modeled as vertices in a point cloud. Then edges, faces, etc. can be constructed as simplices  $\sigma$  which form simplicial complexes  $X$ . Groups  $C_k(X)$ ,  $k = 0, 1, 2, 3$  are sets of all chains of  $k$ th dimension, which is defined as a finite sum of simplices as  $\sum_i \alpha_i \sigma_i^k$  with coefficients  $\alpha_i$ . The boundary operator  $\partial_k$  therefore, maps  $C_k(X) \rightarrow C_{k-1}(X)$  as

$$\partial_k \sigma^k = \sum_{i=0}^k (-1)^i [v_0, \dots, \hat{v}_i, \dots, v_k], \tag{1}$$

where  $\sigma^k = \{v_0, \dots, v_k\}$  and  $[v_0, \dots, \hat{v}_i, \dots, v_k]$  is a  $(k-1)$ -simplex excluding  $v_i$  with  $\partial_{k-1} \partial_k = 0$ . The chain complex is given as

$$\dots \xrightarrow{\partial_{k+1}} C_k(X) \xrightarrow{\partial_k} C_{k-1}(X) \xrightarrow{\partial_{k-1}} \dots \xrightarrow{\partial_2} C_1(X) \xrightarrow{\partial_1} C_0(X) \xrightarrow{\partial_0} 0. \tag{2}$$

The  $k$ -th homology group  $H_k$  is defined by  $H_k = Z_k/B_k$  where  $Z_k = \ker \partial_k = \{c \in C_k \mid \partial_k c = 0\}$  and  $B_k = \text{im } \partial_{k+1} = \{\partial_{k+1} c \mid c \in C_{k+1}\}$ . Thus, the Betti numbers can be defined by the ranks of  $k$ -th homology group  $H_k$ . Persistent homology can be devised to track Betti numbers through a filtration where  $\beta_0$  describes the number of connected components,  $\beta_1$  provides the number of loops, and  $\beta_2$  is the number of cavities. Therefore, using persistent homology, the atoms of 3D structures are grouped according to their elements, as well as the atoms from the binding site of antibodies and antibodies. The interactions and their impacts on PPI complex bindings are characterized by the topological invariants, which are further implemented for machine learning training.

Lastly, a deep learning algorithm, artificial/deep neural networks (ANNs or DNNs), is used to tackle the features with datasets for training and predictions [5]. A trained model is available at [TopNetmAb](#), a SARS-CoV-2-specific model, whose early model was integrating convolutional neural networks (CNNs) with gradient boosting trees (GBTs) and was trained only on the SKEMPI 2.0 dataset with a high accuracy [12].

Recent work with predictions from TopNetmAb [5, 9, 37] is highly consistent with experimental results. One should notice it is important with the help of the aforementioned deep mutational datasets related to SARS-CoV-2. The Pearson correlation of our predictions for the binding of CTC-445.2 and RBD with experimental data is 0.7 [4, 5]. Meanwhile, a Pearson correlation of 0.8 is observed of the predictions of clinical trial antibodies against SARS-CoV-2 induced by emerging mutations in the same work [5] compared to the natural log of experimental escape fractions [38]. Moreover, the prediction of single mutations L452R and N501Y for the ACE2-RBD complex have a perfect consistency with experimental luciferase data [5, 39]. More detailed validations are in Supporting Information.

## 4 Conclusion

The Omicron variant has three subvariants BA.1, BA.2, and BA.3. The Omicron BA.1 has surprised the scientific community by its large number of mutations, particularly those on the spike (S) protein receptor-binding domain (RBD), which enable its unusual infectivity and high ability to evade antibody protections induced by viral infection and vaccination. Viral RBD interacts with host angiotensin-converting enzyme 2 (ACE2) to initiate cell entry and infection and is a major target for vaccines and monoclonal antibodies (mAbs). Omicron BA.1 exploits its 15 RBD mutations to strengthen its infectivity and disrupt mAbs generated by prior viral infection or vaccination. Omicron BA.2 and BA.3 share 12 RBD mutations with BA.1 but differ by 4 and 3 RBD mutations, respectively, suggesting potentially serious threats to human health. However, no experimental result has been reported for Omicron BA.2 and BA.3, although BA.2 is found to be able to alarmingly reinfect patients originally infected by Omicron BA.1 [12]. In this work, we present deep learning predictions of BA.2’s and BA.3’s potential to become another dominating variant. Based on an intensively tested deep learning model trained with tens of thousands of experimental data, we investigate Omicron BA.2’s and BA.3’s RBD mutational impacts on the RBD-ACE2 binding complex to understand their infectivity and a library of 185 antibodies to shed light on their threats to vaccines and existing mAbs. We unveil that BA.2 is about 1.5 and 4.2 times as contagious as BA.1 and Delta, respectively. It is also 30% and 17-fold more capable than BA.1 and Delta, respectively, to escape current vaccines. It is predicted to undermine most existing mAbs, except for sotrovimab developed by GlaxoSmithKline. We forecast Omicron BA.2 will become another prevailing variant by infecting populations with or without antibody protection.

## Data and model availability

The structural information of 185 antibody-RBD complexes with their corresponding PDB IDs and the results of BFE changes of PPI complexes induced by mutations can be found in Section S2 of the Supporting

Information. The TopNetTree model is available at [TopNetmAb](#). The detailed methods can be found in the Supporting Information S3 and S4. The validation of our predictions with experimental data can be located in Supporting Information S5.

## Supporting information

The supporting information is available for

- S1 Supplementary figures: analysis of variant mutation-induced BFE changes for Alpha, Beta, Gamma, Lambda, and Mu variants (the extension of Figure 3).
- S2 Supplementary data: The Supplementary\_Data.zip contains two files: the BFE changes of antibodies disrupted by Omicron subvariant mutations and the list of antibodies with corresponding PDB IDs
- S3 Supplementary feature generation methods
- S4 Supplementary machine learning methods
- S5 Supplementary validation: validations of our machine learning predictions with experimental data

## Acknowledgment

This work was supported in part by NIH grant GM126189, NSF grants DMS-2052983, DMS-1761320, and IIS-1900473, NASA grant 80NSSC21M0023, Michigan Economic Development Corporation, MSU Foundation, Bristol-Myers Squibb 65109, and Pfizer.

## References

- [1] Jiahui Chen, Rui Wang, Nancy Benovich Gilby, and Guo-Wei Wei. Omicron variant (b. 1.1. 529): Infectivity, vaccine breakthrough, and antibody resistance. *Journal of chemical information and modeling*, 2022.
- [2] Huiping Shuai, Jasper Fuk-Woo Chan, Bingjie Hu, Yue Chai, Terrence Tsz-Tai Yuen, Feifei Yin, Xiner Huang, Chaemin Yoon, Jing-Chu Hu, Huan Liu, et al. Attenuated replication and pathogenicity of sars-cov-2 b. 1.1. 529 omicron. *Nature*, pages 1–1, 2022.
- [3] Qin Hong, Wenyu Han, Jiawei Li, Shiqi Xu, Yifan Wang, Zuyang Li, Yanxing Wang, Chao Zhang, Zhong Huang, and Yao Cong. Molecular basis of sars-cov-2 omicron variant receptor engagement and antibody evasion and neutralization. *bioRxiv*, 2022.
- [4] Sandile Cele, Laurelle Jackson, David S Khoury, Khadija Khan, Thandeka Moyo-Gwete, Houriiyah Tegally, James Emmanuel San, Deborah Cromer, Cathrine Scheepers, Daniel G Amoako, et al. Omicron extensively but incompletely escapes pfizer bnt162b2 neutralization. *Nature*, pages 1–5, 2021.
- [5] Li Zhang, Qianqian Li, Ziteng Liang, Tao Li, Shuo Liu, Qianqian Cui, Jianhui Nie, Qian Wu, Xiaowang Qu, Weijin Huang, et al. The significant immune escape of pseudotyped sars-cov-2 variant omicron. *Emerging microbes & infections*, 11(1):1–5, 2022.
- [6] Lihong Liu, Sho Iketani, Yicheng Guo, Jasper FW Chan, Maple Wang, Liyuan Liu, Yang Luo, Hin Chu, Yiming Huang, Manoj S Nair, et al. Striking antibody evasion manifested by the omicron variant of sars-cov-2. *Nature*, pages 1–8, 2021.
- [7] Lu Lu, Bobo Wing-Yee Mok, Linlei Chen, Jacky Man-Chun Chan, Owen Tak-Yin Tsang, Bosco Hoi-Shiu Lam, Vivien Wai-Man Chuang, Allen Wing-Ho Chu, Wan-Mui Chan, Jonathan Daniel Ip, et al.

- Neutralization of sars-cov-2 omicron variant by sera from bnt162b2 or coronavac vaccine recipients. *Clin Infect Dis*, doi:10.1093/cid/ciab1041, 2021.
- [8] Markus Hoffmann, Nadine Krüger, Sebastian Schulz, Anne Cossmann, Cheila Rocha, Amy Kempf, Inga Nehlmeier, Luise Graichen, Anna-Sophie Moldenhauer, Martin S Winkler, et al. The omicron variant is highly resistant against antibody-mediated neutralization—implications for control of the covid-19 pandemic. *Cell*, 2021.
- [9] Perumal Arumugam Desingu, K Nagarajan, and Kuldeep Dhama. Emergence of omicron third lineage ba. 3 and its importance. *Journal of Medical Virology*, 2022.
- [10] Alexandra C Walls, Young-Jun Park, M Alejandra Tortorici, Abigail Wall, Andrew T McGuire, and David Veesler. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, 2020.
- [11] Frederik Plesner Lyngse, Carsten Thure Kirkeby, Matthew Denwood, Lasse Engbo Christiansen, Kåre Mølbak, Camilla Holten Møller, Robert Leo Skov, Tyra Grove Krause, Morten Rasmussen, Raphael Niklaus Sieber, et al. Transmission of sars-cov-2 omicron voc subvariants ba. 1 and ba. 2: Evidence from danish households. *medRxiv*, 2022.
- [12] BA2reinfection. <https://www.timesofisrael.com/several-cases-of-omicron-reinfection-said-detected-in-israel-with-new-ba2-strain/>.
- [13] World Health Organization et al. Enhancing readiness for omicron (b. 1.1. 529): technical brief and priority actions for member states, 2021.
- [14] Wendong Li, Zhengli Shi, Meng Yu, Wuze Ren, Craig Smith, Jonathan H Epstein, Hanzhong Wang, Gary Cramer, Zhihong Hu, Huajun Zhang, et al. Bats are natural reservoirs of SARS-like coronaviruses. *Science*, 310(5748):676–679, 2005.
- [15] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*, 181(2):271–280, 2020.
- [16] Jiahui Chen, Rui Wang, Mengjun Wang, and Guo-Wei Wei. Mutations strengthened SARS-CoV-2 infectivity. *Journal of molecular biology*, 432(19):5212–5226, 2020.
- [17] Rui Wang, Jiahui Chen, Kaifu Gao, and Guo-Wei Wei. Vaccine-escape and fast-growing mutations in the United Kingdom, the United States, Singapore, Spain, India, and other COVID-19-devastated countries. *Genomics*, 113(4):2158–2170, 2021.
- [18] Chunyan Wang, Wentao Li, Dubravka Drabek, Nisreen MA Okba, Rien van Haperen, Albert DME Osterhaus, Frank JM van Kuppeveld, Bart L Haagmans, Frank Grosveld, and Berend-Jan Bosch. A human monoclonal antibody blocking SARS-CoV-2 infection. *Nature communications*, 11(1):1–6, 2020.
- [19] Fei Yu, Rong Xiang, Xiaoqian Deng, Lili Wang, Zhengsen Yu, Shijun Tian, Ruiying Liang, Yanbai Li, Tianlei Ying, and Shibo Jiang. Receptor-binding domain-specific human neutralizing monoclonal antibodies against SARS-CoV and SARS-CoV-2. *Signal Transduction and Targeted Therapy*, 5(1):1–12, 2020.
- [20] Cheng Li, Xiaolong Tian, Xiaodong Jia, Jinkai Wan, Lu Lu, Shibo Jiang, Fei Lan, Yinying Lu, Yanling Wu, and Tianlei Ying. The impact of receptor-binding domain natural mutations on antibody recognition of SARS-CoV-2. *Signal Transduction and Targeted Therapy*, 6(1):1–3, 2021.
- [21] X Zhu, D Mannar, JW Saville, SS Srivastava, AM Berezuk, KS Tuttle, and S Subramaniam. Cryo-em structure of sars-cov-2 omicron spike protein in complex with human ace2 (focused refinement of rbd and ace2), 2021.

- [22] Jiahui Chen, Kaifu Gao, Rui Wang, and Guo-Wei Wei. Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies. *Chemical Science*, 12(20):6929–6948, 2021.
- [23] Bryan E Jones, Patricia L Brown-Augsburger, Kizzmekia S Corbett, Kathryn Westendorf, Julian Davies, Thomas P Cujec, Christopher M Wiethoff, Jamie L Blackbourne, Beverly A Heinz, Denisa Foster, et al. The neutralizing antibody, ly-cov555, protects against sars-cov-2 infection in nonhuman primates. *Science translational medicine*, 13(593), 2021.
- [24] Rui Shi, Chao Shan, Xiaomin Duan, Zhihai Chen, Peipei Liu, Jinwen Song, Tao Song, Xiaoshan Bi, Chao Han, Lianao Wu, et al. A human neutralizing antibody targets the receptor binding site of SARS-CoV-2. *Nature*, pages 1–8, 2020.
- [25] Johanna Hansen, Alina Baum, Kristen E Pascal, Vincenzo Russo, Stephanie Giordano, Elzbieta Wloga, Benjamin O Fulton, Ying Yan, Katrina Koon, Krunal Patel, et al. Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. *Science*, 369(6506):1010–1014, 2020.
- [26] Jinhui Dong, Seth J Zost, Allison J Greaney, Tyler N Starr, Adam S Dingens, Elaine C Chen, Rita E Chen, James Brett Case, Rachel E Sutton, Pavlo Gilchuk, et al. Genetic and structural basis for sars-cov-2 variant neutralization by a two-antibody cocktail. *Nature Microbiology*, 6(10):1233–1244, 2021.
- [27] Zixuan Cang and Guo-Wei Wei. Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics*, 33(22):3549–3557, 2017.
- [28] Jiahui Chen, Kaifu Gao, Rui Wang, and Guo-Wei Wei. Revealing the threat of emerging SARS-CoV-2 mutations to antibody therapies. *Journal of Molecular Biology*, 433(7744), 2021.
- [29] Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019.
- [30] Kui K Chan, Danielle Dorosky, Preeti Sharma, Shawn A Abbasi, John M Dye, David M Kranz, Andrew S Herbert, and Erik Procko. Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science*, 369(6508):1261–1265, 2020.
- [31] Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine HD Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, 182(5):1295–1310, 2020.
- [32] Thomas W Linsky, Renan Vergara, Nuria Codina, Jorgen W Nelson, Matthew J Walker, Wen Su, Christopher O Barnes, Tien-Ying Hsiang, Katharina Esser-Nobis, Kevin Yu, et al. De novo design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2. *Science*, 370(6521):1208–1214, 2020.
- [33] Menglun Wang, Zixuan Cang, and Guo-Wei Wei. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence*, 2(2):116–123, 2020.
- [34] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.
- [35] Herbert Edelsbrunner, John Harer, et al. Persistent homology—a survey. *Contemporary mathematics*, 453:257–282, 2008.
- [36] Zixuan Cang, Lin Mu, and Guo-Wei Wei. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS computational biology*, 14(1):e1005929, 2018.

- [37] Rui Wang, Jiahui Chen, Yuta Hozumi, Changchuan Yin, and Guo-Wei Wei. Emerging vaccine-breakthrough SARS-CoV-2 variants. *ACS Infect. Dis.*, 2021.
- [38] Tyler N Starr, Allison J Greaney, Amin Addetia, William W Hannon, Manish C Choudhary, Adam S Diggins, Jonathan Z Li, and Jesse D Bloom. Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science*, 371(6531):850–854, 2021.
- [39] Xianding Deng, Miguel A Garcia-Knight, Mir M Khalid, Venice Servellita, Candace Wang, Mary Kate Morris, Alicia Sotomayor-González, Dustin R Glasner, Kevin R Reyes, Amelia S Gliwa, et al. Transmission, infectivity, and antibody neutralization of an emerging SARS-CoV-2 variant in California carrying a L452R spike protein mutation. *MedRxiv*, 2021.

# Supporting information for Omicron BA.2 (B.1.1.529.2): high potential to becoming the next dominating variant

Jiahui Chen<sup>1</sup> and Guo-Wei Wei<sup>1,3,4\*</sup>

<sup>1</sup> Department of Mathematics,  
Michigan State University, MI 48824, USA.

<sup>2</sup> Department of Electrical and Computer Engineering,  
Michigan State University, MI 48824, USA.

<sup>3</sup> Department of Biochemistry and Molecular Biology,  
Michigan State University, MI 48824, USA.

February 11, 2022

## **S1 Supplementary figures**

Figure S1 provides the statistic analysis of BFE changes of RBD-ACE2 induced by mutations from Alpha, Beta, Gamma, Lambda, and Mu.

## **S2 Supplementary data**

The Supplementary\_Data.zip contains four files as listed in the following subsection.

### **S2.1 Disrupted antibodies**

File antibodies.BFEs.csv shows the BFE changes of antibodies disrupted by Omicron mutations.

### **S2.2 List of antibodies**

File antibodies.csv lists the Protein Data Bank (PDB) IDs for all of the 185 SARS-CoV-2 antibodies.

## **S3 Supplementary feature generation methods**

In this section, the workflow of the deep learning-based BFE change predictions of protein-protein interactions induced by mutations for the present SARS-CoV-2 variant analysis and prediction will be firstly introduced, which includes four steps as shown in Figure S2: (1) training data preparation; (2) feature generations of protein-protein interaction complexes; and (3) prediction of protein-protein interactions by deep neural networks. Next, the validation of our machine learning-based model will be demonstrated, suggesting consistent and reliable results compared to the experimental deep mutations data.

---

\*Corresponding author. Email: weig@msu.edu

\*Corresponding author. Email: weig@msu.edu

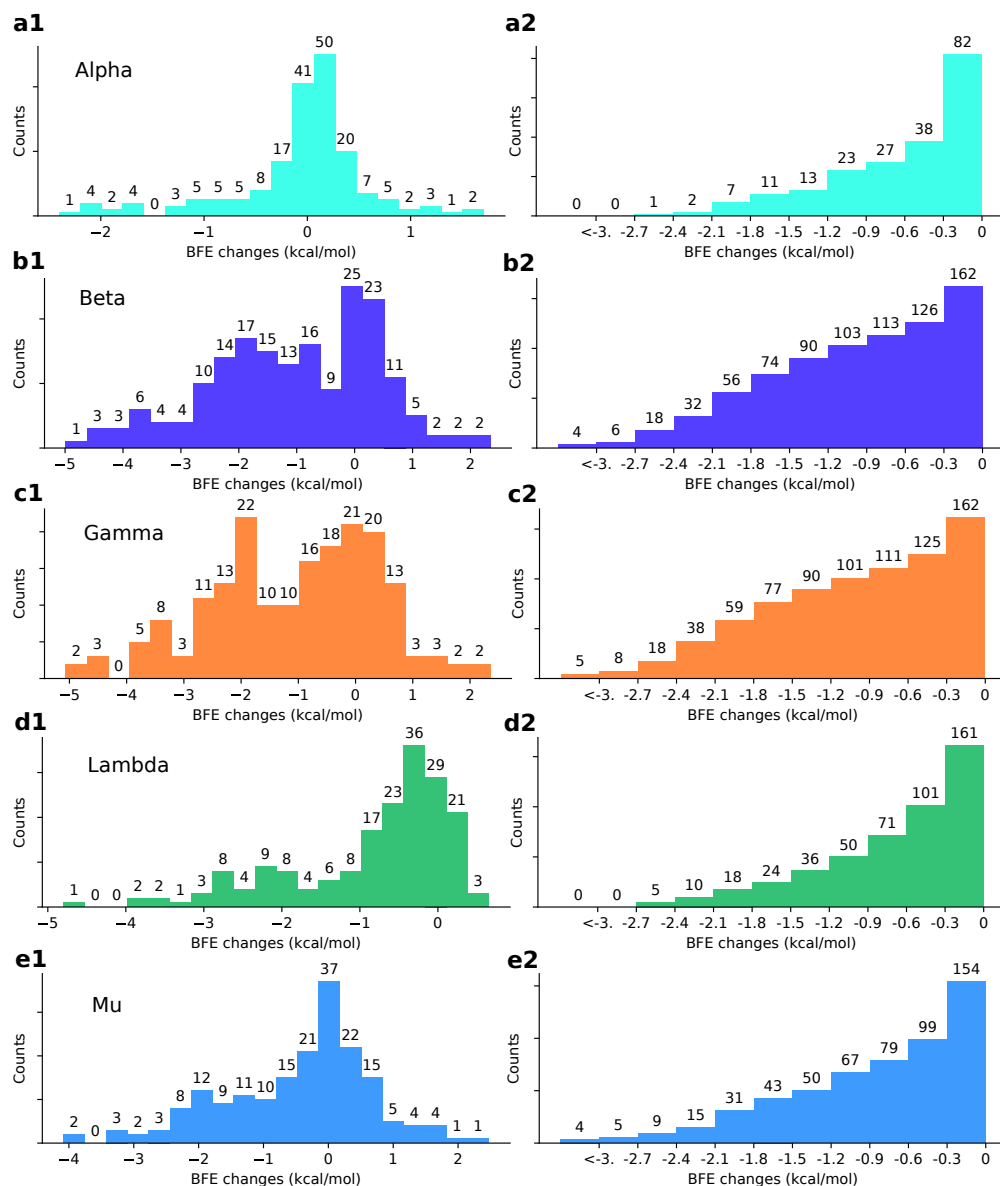


Figure S1: The extension of Figure 3. Analysis of variant mutation-induced BFE changes of 185 antibody-RBD complexes. **a1**, **b1**, **c1**, **d1** and **e1** The distributions (counts) of accumulated BFE changes induced by Alpha, Beta, Gamma, Lambda, and Mu mutations respectively for 185 antibody-RBD complexes. For each case, there are more mutation-weakened complexes than mutation-strengthened complexes. **a2**, **b2**, **c2**, **d2** and **e2** The numbers of antibody-RBD complexes regarded as disrupted by Alpha, Beta, Gamma, Lambda, and Mu mutations respectively under different thresholds ranging from 0 kcal/mol, -0.3 kcal/mol, to <math>-3</math> kcal/mol.

### S3.1 Methods for BFE change predictions

In this section, the process of the machine learning-based BFE change predictions is introduced. Once the data pre-processing and SNP genotyping are carried out, we will firstly proceed with the training data preparation process, which plays a key role in reliability and accuracy. A library of 185 antibodies and RBD complexes, as well as an ACE2-RBD complex, are obtained from Protein Data Bank (PDB). RBD mutation-induced BFE changes of these complexes are evaluated by the following machine learning model. According to the emergency and the rapid change of RNA virus, it is rare to have massive experimental BFE change data of SARS-CoV-2, while, on the other hand, next-generation sequencing data is relatively easy to collect. In the training process, the dataset of BFE changes induced by mutations of the SKEMPI 2.0



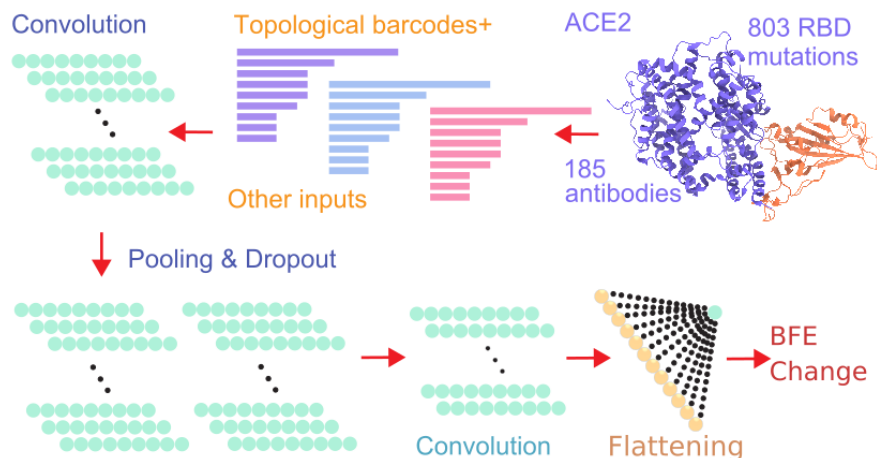


Figure S2: Illustration of genome sequence data pre-processing and BFE change predictions.

dataset [1] is used as the basic training set, while next-generation sequencing datasets are added as assistant training sets. The SKEMPI 2.0 contains 7,085 single- and multi-point mutations and 4,169 elements of that in 319 different protein complexes used for the machine learning model training. The mutational scanning data consists of experimental data of the binding of ACE2 and RBD induced mutations on ACE2 [2] and RBD [3, 4], and the binding of CTC-445.2 and RBD with mutations on both protein [4].

Next, the feature generations of protein-protein interaction complexes are performed. The element-specific algebraic topological analysis on complex structures is implemented to generate topological bar codes [5–8]. In addition, biochemistry and biophysics features such as Coulomb interactions, surface areas, electrostatics, et al., are combined with topological features [9]. The detailed information about the topology-based models will be demonstrated in subsection S3.2. Lastly, deep neural networks for SARS-CoV-2 are constructed for the BFE change prediction of protein-protein interactions [5]. The detailed descriptions of dataset and machine learning model are found in the literature [5, 10, 11] and are available at [TopNetmAb](#).

## S3.2 Feature generation for machine learning model

### S3.2.1 Topology features

Among all features generated for machine learning prediction, the application of topology theory makes the model to a whole new level. Those summarized as other inputs are called as auxiliary features and are described in Section S3.3.2 and S3.3.3. In this section, a brief introduction about the theory of topology will be discussed. Algebraic topology [6, 7] has achieved tremendous success in many fields including biochemical and biophysical properties [8]. Special treatment should be implemented for biology applications to describe element types and amino acids in poly-peptide mathematically, which element-specific and site-specific persistent homology [10, 12]. To construct the algebraic topological features on protein-protein interaction model, a series of element subsets for complex structures should be defined, which considers atoms from the mutation sites, atoms in the neighborhood of the mutation site within a certain distance, atoms from antibody binding site, atoms from antigen binding site, and atoms in the system that belong to type of  $\{C, N, O\}$ ,  $\mathcal{A}_{\text{ele}}(\mathbf{E})$ . Under the element/site-specific construction, simplicial complexes is constructed on point clouds formed by atoms. For example, a set of independent  $k+1$  points is from one element/site-specific set  $U = \{u_0, u_1, \dots, u_k\}$ . The  $k$ -simplex  $\sigma$  is a convex hull of  $k+1$  independent points  $U$ , which is a convex combination of independent points. For example, a 0-simplex is a point and a 1-simplex is an edge. Thus, a  $m$ -face of the  $k$ -simplex with  $m+1$  vertices forms a convex hull in a lower dimension  $m < k$  and is a subset

of the  $k+1$  vertices of a  $k$ -simplex, so that a sum of all its  $(k-1)$ -faces is the boundary of a  $k$ -simplex  $\sigma$  as

$$\partial_k \sigma = \sum_{i=1}^k (-1)^i \langle u_0, \dots, \hat{u}_i, \dots, u_k \rangle, \quad (3)$$

where  $\langle u_0, \dots, \hat{u}_i, \dots, u_k \rangle$  consists of all vertices of  $\sigma$  excluding  $u_i$ . The collection of finitely many simplices is a simplicial complex. In the model, the Vietoris-Rips (VR) complex (if and only if  $\mathbb{B}(u_{i_j}, r) \cap \mathbb{B}(u_{i_{j'}}, r) \neq \emptyset$  for  $j, j' \in [0, k]$ ) is for dimension 0 topology, and alpha complex (if and only if  $\cap_{u_{i_j} \in \sigma} \mathbb{B}(u_{i_j}, r) \neq \emptyset$ ) is for point cloud of dimensions 1 and 2 topology [8].

The  $k$ -chain  $c_k$  of a simplicial complex  $K$  is a formal sum of the  $k$ -simplices in  $K$ , which is  $c_k = \sum \alpha_i \sigma_i$ , where  $\alpha_i$  is coefficients and is chosen to be  $\mathbb{Z}_2$ . Thus, the boundary operator on a  $k$ -chain  $c_k$  is

$$\partial_k c_k = \sum \alpha_i \partial_k \sigma_i, \quad (4)$$

such that  $\partial_k : C_k \rightarrow C_{k-1}$  and follows from that boundaries are boundaryless  $\partial_{k-1} \partial_k = \emptyset$ . A chain complex is

$$\dots \xrightarrow{\partial_{i+1}} C_i(K) \xrightarrow{\partial_i} C_{i-1}(K) \xrightarrow{\partial_{i-1}} \dots \xrightarrow{\partial_2} C_1(K) \xrightarrow{\partial_1} C_0(K) \xrightarrow{\partial_0} 0, \quad (5)$$

as a sequence of complexes by boundary maps. Therefore, the Betti numbers are given as the ranks of  $k$ th homology group  $H_k$  as  $\beta_k = \text{rank}(H_k)$ , where  $H_k = Z_k/B_k$ ,  $k$ -cycle group  $Z_k$  and the  $k$ -boundary group  $B_k$ . The Betti numbers are the key for topological features, where  $\beta_0$  gives the number of connected components, such as number of atoms,  $\beta_1$  is the number of cycles in the complex structure, and  $\beta_2$  illustrates the number of cavities. This presents abstract properties of the 3D structure.

Finally, only one simplicial complex couldn't give the whole picture of the protein-protein interaction structure. A filtration of a topology space is needed to extract more properties. A filtration is a nested sequence such that

$$\emptyset = K_0 \subseteq K_1 \subseteq \dots \subseteq K_m = K. \quad (6)$$

Each element of the sequence could generate the Betti numbers  $\{\beta_0, \beta_1, \beta_2\}$  and consequentially, a series of Betti numbers in three dimensions is constructed and applied to be the topological fingerprints in Figure S2.

### S3.2.2 Residue-level features

**Mutation site neighborhood amino acid composition** Neighbor residues are the residues within 10 Å of the mutation site. Distances between residues are calculated based on residue  $C_\alpha$  atoms. Six categories of amino acid residues are counted, which are hydrophobic, polar, positively charged, negatively charged, special cases, and pharmacophore changes. The count and percentage of the 6 amino acid groups in the neighbor site are regrading as the environment composition features of the mutation site. The sum, average, and variance of residue volumes, surface areas, weights, and hydropathy scores are used but only the sum of charges is included.

**pKa shifts** The pKa values are calculated by the PROPKA software [13], namely the values of 7 ionizable amino acids, namely, ASP, GLU, ARG, LYS, HIS, CYS, and TYR. The maximum, minimum, sum, the sum of absolute values, and the minimum of the absolute value of total pKa shifts are calculated. We also consider the difference of pKa values between a wild type and its mutant. Additionally, the sum and the sum of the absolute value of pKa shifts based on ionizable amino acid groups are included.

**Position-specific scoring matrix (PSSM)** Features are computed from the conservation scores in the position-specific scoring matrix of the mutation site for the wild type and the mutant as well as their difference. The conservation scores are generated by PSI-BLAST [14].

**Secondary structure** The SPIDER2 software is used to compute the probability scores for residue torsion angle and residues being in a coil, alpha helix, and beta strand based on the sequences for the wild type and the mutant [15].

### S3.2.3 Atom-level features

Seven groups of atom types, including C, N, O, S, H, all heavy atoms, and all atoms, are considered when generating the element-type features. Meanwhile, other three atom types, i.e., mutation site atoms, all heavy atoms, and all atoms, are used when generating the general atom-level features.

**Surface areas** Atom-level solvent excluded surface areas are computed by ESES [16].

**Partial charges** Partial change of each atom is generated by pdb2pqr software [17] using the Amber force field [18] for wild type and CHARMM force field [19] for mutant. The sum of the partial charges and the sum of absolute values of partial charges for each atomic group are collected.

**Atomic pairwise interaction interactions** Coulomb energy of the  $i$ th single atom is calculated as the sum of pairwise coulomb energy with every other atom as

$$C_i = \sum_{j, j \neq i} k_e \frac{q_i q_j}{r_{ij}}, \quad (7)$$

where  $k_e$  is the Coulomb’s constant,  $r_{ij}$  is the distance of  $i$ th atom to  $j$ th atom, and  $q_i$  is the charge of  $i$ th atom. The van der Waals energy of the  $i$ th atom is modeled as the sum of pairwise Lennard-Jones potentials with other atoms as

$$V_i = \sum_{j, j \neq i} \epsilon \left[ \left( \frac{r_i + r_j}{r_{ij}} \right)^{12} - 2 \left( \frac{r_i + r_j}{r_{ij}} \right)^6 \right], \quad (8)$$

where  $\epsilon$  is the depth of the potential well, and  $r_i$  is van der Waals radii.

In atomic pairwise interaction, 5 groups (C, N, O, S, and all heavy atoms) are counted both for Coulomb interaction energy and van der Waals interaction energy.

**Electrostatic solvation free energy** Electrostatic solvation free energy of each atom is calculated using the Poisson-Boltzmann equation via MIBPB [20] and are summed up by atom groups.

## S4 Supplementary machine learning methods

The topology-based network model for BFE change predictions induced mutations on SARS-CoV-2 studying applies a deep neural network structure. Similar approaches have been widely implemented in the energy prediction of protein-ligand binding [21] and protein-protein interactions [12]. The neural network method maps an input feature layer to output layer and mimics biological brains for solving problems where numerous neuron units are involved and weights of neurons are updated by backpropagation methods. To make more complicated structure in order to extract abstract properties, more layers and more neurons in each layer can be constructed. In the training process, optimization methods are applied to avoid overfitting issue, such as dropout methods [22] that a partial of computed neurons of each layer is dropped. For the model cross validations, the Pearson correlation of 10-fold cross-validation is 0.864, and the root mean square error is 1.019 kcal/mol.

### S4.1 Deep learning algorithms

A deep neural network is a neural network method with multi-layers (hidden layers) of neurons between the input and output layers. In each layer, the single neuron gets fully connected with the neurons in the next layer. It should preserve the consistency of all labels when applying the model for mutation-induced BFE change predictions. The loss function is constructed as follows:

$$\operatorname{argmin}_{W, b} L(W, b) = \operatorname{argmin}_{W, b} \frac{1}{2} \sum_{i=1}^N (y_i - f(x_i; \{W, b\}))^2 + \lambda \|W\|^2 \quad (9)$$

where  $N$  is the number of samples,  $f$  is a function of the feature vector  $x_i$  parameterized by a weight vector  $W$  and bias term  $b$ , and  $\lambda$  represents a penalty constant.

## S4.2 Optimization

The backpropagation is applied to evaluate the loss function starting from the output layer and propagates backward through the network structure to update the weight vector  $W$  and bias term  $b$ . Since gradient calculation is required, therefore, we apply the stochastic gradient descent method with momentum, which only evaluates a small part of training data and can be considered as calculating exponentially weighted averages, which is given as

$$\begin{aligned} V_i &= \beta V_{i-1} + \eta \nabla_{W_i} L(W_i, b_i) \\ W_{i+1} &= W_i - V_i, \end{aligned} \tag{10}$$

where  $W_i$  is the parameters in the network,  $L(W_i, b_i)$  is the objective function,  $\eta$  is the learning rate,  $X$  and  $y$  are the input and target of the training set, and  $\beta \in [0, 1]$  is a scalar coefficient for the momentum term. The momentum term involved accelerates the converging speed.

## S5 Supplementary validation

In the main content, we briefly summarized validations of our machine learning predictions and experimental data. For large quantitative validations, we compared the BFE change prediction for mutations on S protein RBD to the experimental deep mutational enrichment data on RBD binding to human ACE2 and CTC-445.2 induced by RBD mutations [4, 5, 9]. To make these validations, we eliminated the experimental deep mutational enrichment data of RBD binding to human ACE2 and CTC-445.2 from the training sets and set them as testing sets, which have 1539 and 1500 samples, respectively. In the validation of RBD and CTC-445.2 complex, there is a very high correlation between the enrichment data and machine learning predictions, as well as the validation of RBD binding to ACE2, with Pearson correlations are 0.69 and 0.70, respectively. The deep mutational enrichment data can give a proportional descriptor of the affinity strength of protein-protein interactions induced by mutations. The machine learning methods, however, give stable and equalized evaluations, while experimental data might be different dramatically due to conditions and environments.

In addition, we compared our machine learning results with other experimental data, which are escape fraction, pseudovirus infection changes, and IC<sub>50</sub> fold changes [5]. In the comparison of 35 cases to experimental escape fractions on RBD binding to clinical trial antibodies induced by emerging mutations, our machine learning predictions have a Pearson correlation of 0.80. Especially, those high escaping mutations E484K and E484Q on LY-CoV555, and mutations K417T and K417N on LY-CoV016, are indicated by both our predictions and the experimental data [5]. We also use the pattern comparisons of our prediction to experimental data. Lastly, we collected experimental data from different literature [23–26]. According to variations from different research groups, they were summarized in increasing/decreasing patterns of emerging variant (including co-mutations) impacts on antibody therapies in clinical trials. In total, there are 20 pattern comparisons with an excellent agreement between various experimental data and our predictions, except for a minor discrepancy [5].

## References

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

- [2] Erik Procko. The sequence of human ace2 is suboptimal for binding the s spike protein of sars coronavirus 2. *BioRxiv*, 2020.
- [3] Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine HD Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, 182(5):1295–1310, 2020.
- [4] Thomas W Linsky, Renan Vergara, Nuria Codina, Jorgen W Nelson, Matthew J Walker, Wen Su, Christopher O Barnes, Tien-Ying Hsiang, Katharina Esser-Nobis, Kevin Yu, et al. De novo design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2. *Science*, 370(6521):1208–1214, 2020.
- [5] Jiahui Chen, Kaifu Gao, Rui Wang, and Guo-Wei Wei. Revealing the threat of emerging SARS-CoV-2 mutations to antibody therapies. *Journal of Molecular Biology*, 433(7744), 2021.
- [6] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.
- [7] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *Proceedings 41st annual symposium on foundations of computer science*, pages 454–463. IEEE, 2000.
- [8] Kelin Xia and Guo-Wei Wei. Persistent homology analysis of protein structure, flexibility, and folding. *International journal for numerical methods in biomedical engineering*, 30(8):814–844, 2014.
- [9] Jiahui Chen, Kaifu Gao, Rui Wang, and Guo-Wei Wei. Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies. *Chemical Science*, 12(20):6929–6948, 2021.
- [10] Jiahui Chen, Rui Wang, Menglun Wang, and Guo-Wei Wei. Mutations strengthened SARS-CoV-2 infectivity. *Journal of molecular biology*, 432(19):5212–5226, 2020.
- [11] Rui Wang, Yuta Hozumi, Changchuan Yin, and Guo-Wei Wei. Mutations on COVID-19 diagnostic targets. *Genomics*, 112(6):5204–5213, 2020.
- [12] Menglun Wang, Zixuan Cang, and Guo-Wei Wei. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence*, 2(2):116–123, 2020.
- [13] Delphine C Bas, David M Rogers, and Jan H Jensen. Very fast prediction and rationalization of pka values for protein–ligand complexes. *Proteins: Structure, Function, and Bioinformatics*, 73(3):765–783, 2008.
- [14] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.
- [15] Yuedong Yang, Rhys Heffernan, Kuldip Paliwal, James Lyons, Abdollah Dehzangi, Alok Sharma, Jihua Wang, Abdul Sattar, and Yaoqi Zhou. Spider2: A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. In *Prediction of protein secondary structure*, pages 55–63. Springer, 2017.
- [16] Beibei Liu, Bao Wang, Rundong Zhao, Yiying Tong, and Guo-Wei Wei. Eses: software for eulerian solvent excluded surface, 2017.
- [17] Todd J Dolinsky, Jens E Nielsen, J Andrew McCammon, and Nathan A Baker. Pdb2pqr: an automated pipeline for the setup of poisson–boltzmann electrostatics calculations. *Nucleic acids research*, 32(suppl\_2):W665–W667, 2004.

- [18] David A Case, Tom A Darden, Thomas E Cheatham, Carlos L Simmerling, Junmei Wang, Robert E Duke, Ray Luo, MRCW Crowley, Ross C Walker, Wei Zhang, et al. Amber 10. Technical report, University of California, 2008.
- [19] Bernard R Brooks, Charles L Brooks III, Alexander D Mackerell Jr, Lennart Nilsson, Robert J Petrella, Benoît Roux, Youngdo Won, Georgios Archontis, Christian Bartels, Stefan Boresch, et al. Charmm: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–1614, 2009.
- [20] Duan Chen, Zhan Chen, Changjun Chen, Weihua Geng, and Guo-Wei Wei. Mibpb: a software package for electrostatic analysis. *Journal of computational chemistry*, 32(4):756–770, 2011.
- [21] Duc Duy Nguyen and Guo-Wei Wei. AGL-Score: Algebraic graph learning score for protein–ligand binding scoring, ranking, docking, and screening. *Journal of chemical information and modeling*, 59(7):3291–3304, 2019.
- [22] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- [23] FACT SHEET FOR HEALTH CARE PROVIDERS EMERGENCY USE AUTHORIZATION (EUA) OF REGEN-COV (fda.gov).
- [24] Yiska Weisblum, Fabian Schmidt, Fengwen Zhang, Justin DaSilva, Daniel Poston, Julio CC Lorenzi, Frauke Muecksch, Magdalena Rutkowska, Hans-Heinrich Hoffmann, Eleftherios Michailidis, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *Elife*, 9:e61312, 2020.
- [25] Pengfei Wang, Manoj S Nair, Lihong Liu, Sho Iketani, Yang Luo, Yicheng Guo, Maple Wang, Jian Yu, Baoshan Zhang, Peter D Kwong, et al. Antibody resistance of SARS-CoV-2 variants B. 1.351 and B. 1.1. 7. *Nature*, 10, 2021.
- [26] Delphine Planas, David Veyer, Artem Baidaliuk, Isabelle Staropoli, Florence Guivel-Benhassine, Maaran Michael Rajah, Cyril Planchais, Françoise Porrot, Nicolas Robillard, Julien Puech, et al. Reduced sensitivity of SARS-CoV-2 variant delta to antibody neutralization. *Nature*, pages 1–7, 2021.