

Article

Delta Variant with P681R Critical Mutation Revealed by Ultra-Large Atomic-Scale Ab Initio Simulation: Implications for the Fundamentals of Biomolecular Interactions

Puja Adhikari ¹ , Bahaa Jawad ^{1,2} , Praveen Rao ³, Rudolf Podgornik ^{4,5,6,7}  and Wai-Yim Ching ^{1,*} 

¹ Department of Physics and Astronomy, University of Missouri-Kansas City, Kansas City, MO 64110, USA; paz67@umkc.edu (P.A.); bajrmd@mail.umkc.edu (B.J.)

² Department of Applied Sciences, University of Technology, Baghdad 10066, Iraq

³ Department of Health Management and Informatics, Department of Electrical Engineering and Computer Science, University of Missouri-Columbia, Columbia, MO 65212, USA; praveen.rao@missouri.edu

⁴ School of Physical Sciences and Kavli Institute of Theoretical Science, University of Chinese Academy of Sciences, Beijing 100049, China; rudipod@gmail.com

⁵ CAS Key Laboratory of Soft Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing 100090, China

⁶ Wenzhou Institute of the University of Chinese Academy of Sciences, Wenzhou 325000, China

⁷ Department of Physics, Faculty of Mathematics and Physics, University of Ljubljana, SI-1000 Ljubljana, Slovenia

* Correspondence: chingw@umkc.edu

Abstract: The SARS-CoV-2 Delta variant is emerging as a globally dominant strain. Its rapid spread and high infection rate are attributed to a mutation in the spike protein of SARS-CoV-2 allowing for the virus to invade human cells much faster and with an increased efficiency. In particular, an especially dangerous mutation P681R close to the furin cleavage site has been identified as responsible for increasing the infection rate. Together with the earlier reported mutation D614G in the same domain, it offers an excellent instance to investigate the nature of mutations and how they affect the interatomic interactions in the spike protein. Here, using ultra large-scale ab initio computational modeling, we study the P681R and D614G mutations in the SD2-FP domain, including the effect of double mutation, and compare the results with the wild type. We have recently developed a method of calculating the amino-acid–amino-acid bond pairs (AABP) to quantitatively characterize the details of the interatomic interactions, enabling us to explain the nature of mutation at the atomic resolution. Our most significant finding is that the mutations reduce the AABP value, implying a reduced bonding cohesion between interacting residues and increasing the flexibility of these amino acids to cause the damage. The possibility of using this unique mutation quantifiers in a machine learning protocol could lead to the prediction of emerging mutations.

Keywords: SARS-CoV-2; spike-protein; delta variant; interatomic interaction; amino-acid–amino-acid bond pair; machine learning



Citation: Adhikari, P.; Jawad, B.; Rao, P.; Podgornik, R.; Ching, W.-Y. Delta Variant with P681R Critical Mutation Revealed by Ultra-Large Atomic-Scale Ab Initio Simulation: Implications for the Fundamentals of Biomolecular Interactions. *Viruses* **2022**, *14*, 465. <https://doi.org/10.3390/v14030465>

Academic Editor: Yoshitaka Sato

Received: 26 January 2022

Accepted: 21 February 2022

Published: 24 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The COVID-19 pandemic started two years ago and continues with unabated intensity, with no clear end in sight, despite many repeated attempts to contain it. The recent emergence of various variants of concern (VOC) in severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) [1], such as Alpha [2], Beta [3], Delta [4], and Gamma [5], and variants of interest (VOI), such as Eta [6], Iota [7], Kappa [8], Lambda [9], and Mu [10], instigates new anxieties. Among the new VOC, the Delta variant causes a more severe infection and spreads faster than previous variants of the SARS-CoV-2 virus, emerging as the dominant strain in the world [11], causing worries among the general population and solidifying the belief that the battle against the pandemic will be a long one. In a broader

context, this historical moment that faces us is a grand one from every conceivable direction. It also introduces a new chapter in the perception and significance of biomedical sciences. However, a successful response to this dire situation crucially implies and promotes not only pandemic-related efforts in biomedical sciences but, even more importantly, deep level collaborations across all scientific fields, guiding a concerted action grounded in different social instruments.

The evolution of viruses in recent decades has been well-documented, including the 1918 flu pandemic [12], the zoonotic HIV [13], and the seasonal flu virus variations, as well as several predecessors of SARS-CoV-2, such as SARS-CoV-1 in 2003 and MERS in 2012 [14]. The emergence of the Delta variant, together with other VOC, are a natural and unavoidable part of the virus evolution, and can be traced to specific mutations of the amino acids in the protein sequence that can result in an enhanced infection rate or can quench the full action of the already developed vaccines and/or other therapeutic drugs [15,16]. Other mutational variants, in addition to the known Alpha, Beta, Gamma, Delta, etc., will continue to emerge as the epidemic rages on, making it imperative to strive for a fundamental understanding of the role of mutations at a deeper molecular and atomic level. This fundamental understanding can enable the design of new strategies and methods to combat the current and emerging variants, such as the AY.4.2 “Delta plus” variant, which seems to be more transmissible than the original Delta variant in the United Kingdom [17].

The spike (S) protein of SARS-CoV-2 has two subunits, S1 and S2, responsible for ACE2 receptor docking and membrane fusion, respectively [18]. In fact, SARS-CoV-2 enters the host cells through its S-protein, which is synthesized as an inactive precursor that must be cleaved to successfully mediate membrane fusion [19]. The cleavage activation mechanism occurs at S1/S2 and S2' cleavage sites [20], the former being located at the boundary between the S1 and S2 subunits, having a unique polybasic insertion furin recognition site ${}_{681}\text{PRRAR}|S_{686}$ ($|$ denotes proteolytic cleavage site) [19]. The S-protein is thus first cleaved at the S1/S2 site, which does not actually result in the complete separation of the S1 and S2 subunits but allows them to remain non-covalently bound [20]. Upon the S1/S2 cleavage and binding of the S-protein to ACE2, a second cleavage site, S2', becomes more exposed to being completely cleaved by host proteases for activating virus–cell membrane fusion [19–25]. Hence, the unique S1/S2 site has been identified as being mainly responsible for its high infectivity and transmissibility [22]. Interestingly, the P681R mutation right at the furin cleavage site of the Delta variant plays an important role in enhancing the S-protein cleavage [26–28] and is hypothesized as the main culprit for the Delta variant infectivity [26–28]. In addition to the P681R mutation, the Delta variant also contains a D614G mutation, which promotes the RBD of the S-protein in an “open” conformation, making its binding with the ACE2 receptor easier [29], as well as enhancing the protease cleavage at the S1/S2 cleavage site [30]. In view of this overarching importance of the P681R and D614G mutations, it is therefore crucial to understand the role that these mutations play in the phenomenology of the Delta variant at the *atomic scale*, which can only be accomplished by unleashing the best that the *ab initio* quantum chemical methodology has to offer. *Ab initio* calculations, combined with a further deep analysis, can offer more in the fundamental understanding of such biomolecules.

The specific aim of this study is to investigate the nature of these two important mutations, P681R and D614G, in the Delta variant using ultra large-scale *ab initio* quantum chemical modeling, combined with an advanced analysis that allows for a quantitative assessment of the impact of mutations on the atomic resolution scale. Specifically, we study the atomically resolved structure and quantify the interatomic impact of P681R and D614G in the SD2 to FP (SD2-FP) domains of the S-protein, together with the effect of the double mutation, and compare the results with the unmutated case or the wild type (WT). We use the recently developed method of calculating the amino-acid–amino-acid bond pairs (AABP) to characterize the quantitative details of the interatomic interactions [31]. Such an unprecedented and detailed analysis of the origin and impact of atomistically

resolved mutations provides many fundamental insights that could lead to a new level of understanding in the development of therapeutic drug design against the SARS-CoV-2 virus and its variants.

2. Model Design and Construction

Large biomolecular systems, such as proteins, have complex structures and contain many amino acids linked together in a specific order. Currently, we are capable of conducting ab initio simulations with up to 5000 atoms for a single calculation by adopting a divide and conquer strategy to investigate the S-protein. We briefly describe the model used in this study.

The SD2 to FP region marked as region 3 in Figure S1 is our SD2-FP model, which is used in the actual atomic-scale calculation in the present work. The other broader structural regions consist of different structural domains in the S-protein of the SARS-CoV-2, which shows the specific mutation sites in the Delta variants. This is fully illustrated in Figure S1 in the Supplementary Materials (SM). The SD2-FP models involved in the present work are: (a) the wild type (WT), (b) the mutated model P681R, (c) the mutated model D614G, and (d) the double mutation with both D614G and P681R.

The initial structure for the regions shown in Figure S1 was obtained from Woo et al. from [PDB ID 6VSB] [32], which originated from the study by Wrapp et al. [18]. Here, it should be mentioned that the 6VSB Cryo-EM structure from Wrapp et al. has many missing amino acids (AAs) due to the limitation in their technique. Different prediction methods to model the missing AAs of 6VSB were used and the details of obtaining the full-length SARS-CoV-2 spike (S) protein structure can be found in Woo et al. [32]. We chose Chain A in its up conformation. Sequence numbers 592–834 in S-protein were used for SD2-FP model (6VSB_1_2_1) [33]. We used our procedure to construct a manageable size model and summarize it as follows. First, we selected all residues of the SD2 and FP region to create the SD2-FP model (residue 592 to 834). Next, we removed the glycans and the associated hydrogen (H) atoms from the SD2-FP model and later added the H atoms using the Leap module with ff14SB force field in the AMBER package [34–36], which then yields the WT model used as a template to generate the mutated models. To prepare the mutated models with a single mutation (P681R or D614G) or a double mutation (P681R and D614G), we used Dunbrack backbone-dependent rotamer library [37], implemented by the UCSF Chimera package [38]. In total, we created four SD2-FP models, including the WT model, mutated with P681R and D614G mutations, and double mutation consisting of both D614G and P681R. These models were minimized with 100 steepest descent steps and 10 conjugate gradient steps using UCSF Chimera to avoid bad clashes. These models (see Figures 1 and S1) were then further optimized using Vienna ab initio simulation package (VASP) [39], as described in Section S1.1 in SM.

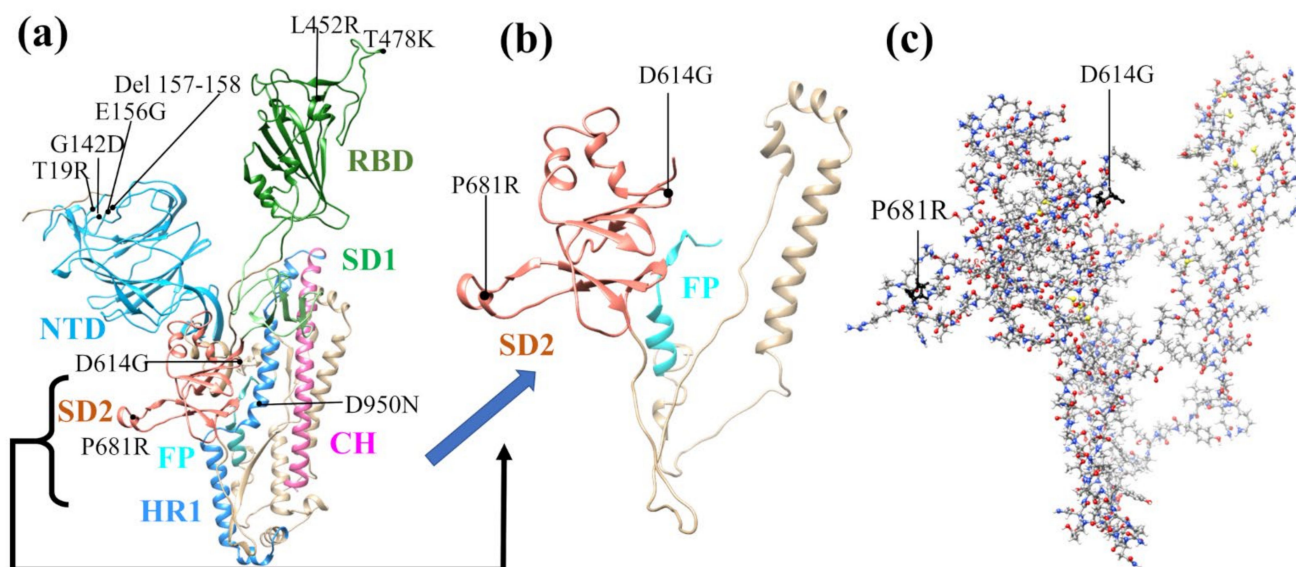


Figure 1. The illustration of SD2-FP model construction. (a) Ribbon of a single protomer in up conformation (chain A) of spike protein SARS-CoV-2 from signal peptide (SP) to central helix (CH) and the associated mutation for Delta variant in different domains, (b) the ribbon structure of SD1-FP model with two marked mutations that is selected for constructing our models, (c) ball and stick of the SD1-FP in (b) and their respective mutations as marked.

3. Amino-Acid–Amino-Acid Bond PAIR (AABP)

The VASP-optimized structures were used as input in the orthogonalized linear combination of atomic orbitals (OLCAO) method [40] to calculate the electronic structure and interatomic interactions in biomolecules. The details of the OLCAO method are described in Section S1.2 of SM. Using OLCAO, we calculated the bond order (BO), which quantifies the strength of the bond between two atoms and usually scales with the bond length (BL). The sum of all BO values within a single structure unit gives the total bond order (TBO). The relatively new concept of BO and TBO in biomolecules quantifies the cohesion of the system. Before we extend our formulation and analysis of BO and TBO to the amino-acid–amino-acid bond pair (AABP), we will first discuss briefly the 20 canonical amino acids with distinct residues listed in Table S1 and illustrated through their functional groups in Figure S2.

Amino acids are the basic structural units of proteins, sharing three common structural elements: an amine group, a carboxyl group, and a side chain residue. Different functional groups comprising the side chain consign to each of the 20 canonical amino acids distinct physical properties that influence the protein structure and function. The peptide bond links two adjacent AAs with a covalent bond between C1 (carbon number one) of one AA and N2 (nitrogen number two) of another, creating a linear chain connecting AAs with chemically distinct side chain residues into different linear sequences that can form long polypeptide chains that are able to fold upon themselves, thereby giving rise to diverse, functionally distinct proteins. Any discussion of the structure, properties, and functionalities of proteins must therefore originate from the unique structures and properties of the 20 canonical AAs [41,42].

Various physical properties characterize and differentiate the canonical AAs in bathing aqueous solutions, such as the number of atoms, size, molecular weight, hydrophathy, polarity, charge, protonation/deprotonation dissociation constants, etc. In view of our aspiration to embark on a detailed *ab initio* investigation, unprecedented in size and scope, of the nature and effects of point mutations on interatomic interactions in the spike protein, most of the listed quantifiers do not seem to be appropriate and some appear to be distinctly missing. Based upon what the ultra-large scale *ab initio* methodology can provide, we

focus on two structural quantifiers: the well-known partial charge (PC) parameter and a modification of the BO parameter referred to as the amino-acid–amino-acid bond pair (AABP). The first one addresses the polarity and charge structure of the molecule and certainly responds to all of the alterations wrought by the mutation in the protein sequence. The second one is a generalization of the BO and TBO and is specifically tailored to proteins by defining the amino-acid–amino-acid bond pair (AABP) as where the summations are over all atoms α in AA u and all atoms β in AA v . AABP embodies all possible bonding between two amino acids, including both covalent and hydrogen bonding (HB). AABP is a single parameter that quantifies the amino-acid–amino-acid interaction, so that the stronger the interaction, the larger the AABP, and vice versa.

$$AABP(u, v) = \sum_{\alpha \in u} \sum_{\beta \in v} \rho_{\alpha i, \beta j} \quad (1)$$

We stress that the use of this novel AABP concept is not the same as using the conventional description of the amino-acid–amino-acid interaction in biomolecules, since the distance of separation and the atomic interactions between two AAs are difficult to quantify accurately. The AABP values are calculated from quantum mechanical wave functions to study different types of interactions in biomolecules, such as nearest neighbor (NN) and non-NN, also designated as off-diagonal or non-local (NL) interactions between AAs. Non-NN AAs are not vicinal in the 1D primary sequence space but are vicinal in the 3D embedding folding space. The AABP concept can help to foment a better understanding of the overall 3D interactions, not only of proteins, but of complex biomolecular systems in general [31].

The AABP defined above is a unique feature in the present study. Simulation methodologies that have been routinely and extensively used by the biomolecular research community [43,44], such as classical molecular dynamics (MD) with its onerous energy or enthalpy calculations, are based on different types of presumably transferable a priori force field specifications. As such, they cannot reveal the atomic details of the real interatomic interactions and mostly rely on the atomic potential parametrizations and assumed geometric structures, which are both inherently limited. On the other hand, the ab initio molecular dynamics methodology [45], intended for a more realistic simulation of complex biomolecular systems and processes from first principles, is, at present, hampered by the excessive computational times and resources, making it inapplicable to the analysis of even modest-sized proteins. The use of the concept of bond order, as implemented in this work, can quantitatively characterize the AA-AA interaction in 3D folding space and can also be applied to larger scale protein–protein interactions, or the interaction of different segments of the same protein, thus providing a promising and valuable alternative (see Section 4.2 for details).

Our approach here will be based on the characterization of the wild type and mutated protein by the PC and AA-AA bond pair parameters. By judiciously labeling each mutation as a data point, with specific details for the different components of the partial charge and AA-AA bond pair parameters, it will furthermore facilitate its application in a machine learning (ML) protocol when many mutation data become available.

4. Results

This section is conveniently divided into four sections, but the key section is Section 4.3 AABP data for mutations in the structural domain of SD2-FP containing the furin cleavage site (Figure S1). The AABP data are based on the results of the model structures using VASP and the OLCAO for the electronic interactions. Table S2 lists the structure information from VASP optimization for the four SD2-FP models in the Delta variant: (a) wild type (WT), (b) mutated P681R (R681), (c) mutated D614G (G614), and (d) double mutation (DM) labeled as G614-R681. In addition, Table S2 also lists two HR1-CH models in the Delta variant: (e) wild type (WT) D950 and (f) mutated D950N (N950). As can be seen in Table S2, the energy is sufficiently converged to the level of 0.03 to 0.04 eV, which is less than 10^{-5} eV

per atom, including H atoms, but the entailed computational resources consumed are humongous. The VASP-optimized structure is used as the input for the DFT calculation using OLCAO.

The results are divided into the following four sections. Sections 4.1 and 4.2 are standard electronic structures routinely present for the analysis in biomolecular systems [31,46–50]. Section 4.2 describes the key data on interatomic interactions between all atoms whose results are used for the main Section 4.3 on the AABP data for the four SD2-FP models (a) to (d) listed in Table S2. Section S2 in SM provides the additional results from the two HR1-CH models (e) and (f), listed in Table S2, that support the observation in Section 4.3.

4.1. Electronic Structure

In an ab initio calculation of any materials, the focus is on the density of states (DOS) or its components, the partial DOS (PDOS). In small molecules, researchers tend to use the list of energy levels separated by HOMO-LUMO gaps. Figure S3 shows the calculated total DOS (TDOS) of the current supercell WT SD2-FP containing P681 and D614 with 3654 atoms. The 0.0 eV energy stands for the HOMO state or the top of the occupied valence band. The LUMO is located at approximately 2.5 eV. There exist some gap states within the HOMO-LUMO gap as expected due to some interacting states within this complex biomolecule. There is virtually no difference in the TDOS between the WT model and those that contain the mutated AA. The only difference is a very minute structure in some peaks, where, presumably, the mutated AA has a slightly different energy level. In principle, the PDOS can be resolved into an individual AA or groups of AAs, which will be very useful if a more detailed analysis is necessary, such as making distinctions between mutated and unmutated AAs. The atomic scale interaction must be revealed by a detailed analysis of the calculated electronic structure on relevant AAs, which will be fully revealed in this Section later.

Figure S4 displays the PC on each of the 243 AAs from F592 to I834 in the WT SD2-FP model. The PC values of the mutated AAs G614, R681, and the double mutation G614-R681 are also highlighted and marked. The distribution of PC can be divided into three groups: (1) 20 AAs that are largely positively charged with PC values above 0.2 e^- ; (2) 21 AAs that are largely negatively charged with PC values lower than -0.2 e^- ; and (3) a large group of 202 AAs with small PC between 0.2 e^- and -0.2 e^- . The most positively charged AAs are W633 (2.01 e^-) and Y764 (1.90 e^-), and the majority of the highly negatively charged 13 AAs have a nearly equal PC of around -0.80 e^- to -0.99 e^- . The data in Figure S4 clearly show that the PC of D614 and P681 have dramatic changes in PC values under mutation and double mutation. D614 changes from the -0.96 e^- in WT to -0.02 e^- and -0.01 e^- for mutated G614 and double mutation G614-R681, respectively. Similarly, P681 changes from 0.15 e^- for WT to 0.93 e^- and 1.03 e^- for mutated R681 and double mutation G614-R681. Besides the famous 614th and 681st residues, there exists another site that shows a high variation in PC: Y756. Y756 changes from a highly positive charge of 1.90 e^- in WT to -0.10 e^- in all mutated models.

Figure S5 displays the PC of AAs on the solvent-excluded surface of the SD1-FP model (Figure 1c), with the location of key AAs that undergo mutation marked D614 and P681. Interestingly, D614 is highly negatively charged and P681 is near neutral (0.15 e^-). Other relatively positively charged AAs (colored blue) shown in the Figure S5 are F592, W633, R634, R646, R682, R683, R685, Y756, R765, K776, K786, K790, and K814, whereas the negatively charged AAs (colored red) consist of E619, D627, E654, D663, E702, E725, D745, E748, E773, E780, D830, and I834. These results indicate that the atomic-scale calculations can provide a charge distribution of protein subunits impacting the long-range electrostatic interaction between different structural units of the protein. The actual PC of selected AAs are listed in Table S3. There are 141 AAs with a PC range of -0.182 e^- to 0.019 e^- (grey color) and 63 AAs with a PC range in between 0.020 e^- and 0.222 e^- (green color). This type of charge distribution is common in such biomolecules. The PC distribution of the

HR1-CH WT model is displayed in Figure S6. The WT D950 has a largely negative PC, which changes to a positive PC when it mutates to N950. The PC for each AA for the HR1-CH model is listed in Table S4.

4.2. Interatomic Bonding

In contrast to the TDOS discussed above, the actual interatomic interaction in the form of bond order (BO) values (see Computational Methods Section S1 in SM) are fully available for all of the atoms in the supercell used in the DFT calculation with the OLCAO method. These BO values are the basic ingredients for calculating the AABP values, central to this paper.

Figure 2 shows the BO vs. BL distribution for every pair in the WT model for BL less than 3.5 Å, including all covalently bonded pairs, as well as the HB. The inset shows the distribution for BL from 2.0 Å to 4.5 Å. This is a very busy figure containing many interesting facts. We succinctly summarize them below:

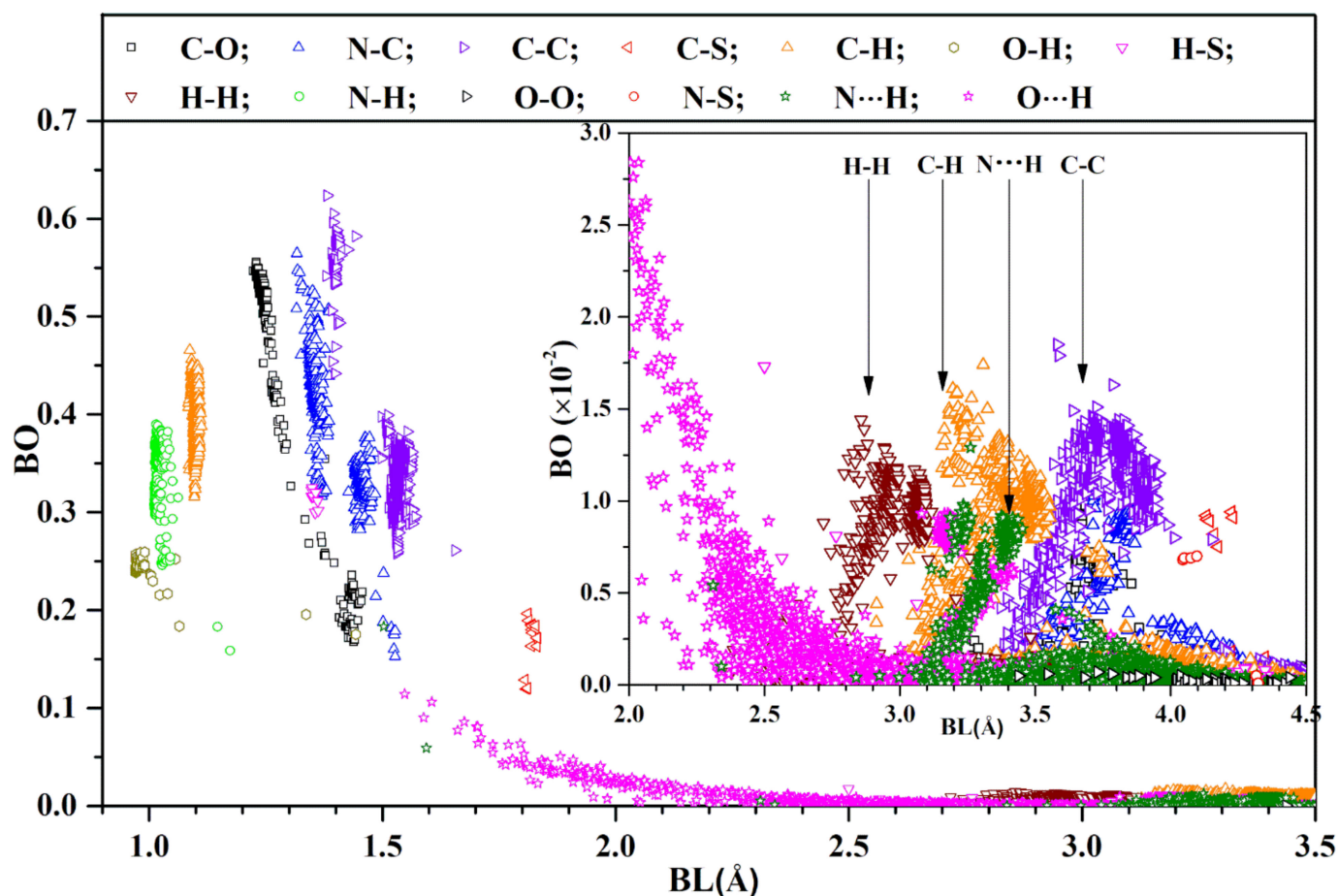


Figure 2. Distribution of BO vs. BL for the WT SD2-FP model (1–3.5 Å). Inset: detailed distribution in range (2.0–4.5 Å) with reduced BO scale. Most of them are in 4 types of atomic pairs marked with vertical arrows.

1. The first group of covalent bonds are O-H, N-H, and C-H, with BL ranging from less than 1 Å to less than 1.2 Å, and with BO ranging from 0.15 e^- to 0.48 e^- , depending on the actual structure of the AAs listed in Figure S2; it may even be between certain AAs. There are two O-H bonds at 1.33 Å and 1.44 Å. The former bond occurs between two AAs (D808 and K811) and the latter one is from the same AA F592;
2. The second group of covalent bonds is the usual covalent bond between C, O, and N. Their BO values can be very large, ranging from 0.16 e^- for N-C up to 0.62 e^-

for C=C, which are strong double bonds. The relatively weaker C-C bonds have a slightly larger BL. A similar observation can be seen for bonds between (C, O) and (N, C) pairs;

3. The next important bond is the HBs: mostly O···H and a few N···H. HBs are much weaker than covalent bonds, but are ubiquitous, ranging all the way to a ‘BL’ of close to 4 Å (see inset). According to a detailed analysis by Lei et al. on a super-cold network of water [51], the maximum BO for O···H is around 0.1 e[−];
4. The next interesting bonds are the covalent H-S and C-S bonds from the only S-containing AAs Cys and Met. Some H-S bonds have a short BL (1.4 Å) with a strong BO (0.3 e[−]), whereas the others have a large BL with a weak BO (more than 2.5 Å and less than 0.05 e[−]). The C-S bonds are located at a BL of around 1.8 Å and with a relatively strong BO value between 0.1 e[−] and 0.2 e[−] at this longer BL. Our results show that there are no disulfide bonds in the SD2-FP model;
5. We now focus our observations on the inset of Figure 2 for the BL ranging from 2.0 Å to 4.5 Å. It reveals many weaker HBs, with a BO less than 0.03 e[−]. Even more surprising is the presence of many atomic pairs (H-H, C-H, C-C, N···H) that contribute to BO values with weak but nontrivial values of less than 0.02 e[−]. These bonds are obviously formed between the non-local AAs, which play a critical role in the total AABP values, to be discussed in the next section;
6. One point we must emphasize is that the use of BO is a relatively new concept advocated by us. The BLs must be interpreted as the distance of separations between a pair of atoms, with the proviso that their interatomic interaction can go beyond the actual atomic pairs labeled as ‘BL’ due to the quantum effects arising from overlapping orbitals of their nearby atoms. Such subtle issues are usually ignored in biomolecular systems, since they are seldom discussed in the context of quantum mechanical wave functions, but rely on the distances between two atoms quantified by ‘BL’. Similar issues have been raised recently in the literature regarding the nature of C-H and C-C bonds [52].

4.3. AABP Data for Mutations in Delta Variant

The mutations on the Delta variant have been a hot topic that has attracted a lot of attention [26,28,53–57]. Most of these studies focus on the clinical or experimental observations to demonstrate the danger of mutations, especially the P681R near the furin cleavage site in the SD2-FP domain of the S-protein, but, to the best of our knowledge, no theoretical explanation or computational studies have been reported so far. Based on the detailed atomic-scale electronic structure calculation described in the above two sections, we extend the calculation of interactions between AAs involved in the mutations in the form of AABP described in the methods section. The calculated AABP values of mutations are summarized in Table 1. Each calculation is considered as a data point labeled by the specifically designed notation that will be instrumental in data mining and machine learning (ML) applications (see Section 5.2). The main observations of Table 1 are as follows.

To better explain the information present in Table 1 regarding the nature of total AABP values and its components of nearest neighbor (NN) AABP and non-local AABP (NL) values, we show in Figure 3 the sequence of AAs from E592 to S689. Figure 3a in the ribbon form and Figure 3b in the sequential form both show the location of the main mutation sites P681 and D614 (pink circle), the NN AAs to these two mutations are in yellow circles, and the non-local interacting AAs (green circle) connected by lines indicate that they are interacting.

Table 1. AABP of four SD2-FP models and other information. DM denotes double mutation.

Models	Total AABP	NN AABP	Non-Local AABP	AABP from HB	No. of NL AAs	Data Notation
WT P681	1.117	1.064	0.054	0.064	5	P681-1.117-1.064-0.054-0.064-0
WT D614	0.917	0.912	0.005	0.040	5	D614-0.917-0.912-0.005-0.040-0
Mutated R681	1.082	0.971	0.111	0.122	6	R681-1.082-0.971-0.111-0.122-1
Mutated G614	0.904	0.904	0.001	0.040	2	G614-0.904-0.904-0.001-0.040-1
DM G614-R681						
R681	1.023	0.975	0.047	0.066	6	R681-1.023-0.975-0.047-0.066-1
G614	0.901	0.901	0.001	0.041	2	G614-0.901-0.901-0.001-0.041-1

The main observations of the data in Table 1 are as follows: 1. AABP values provide the quantitative information on each AA position in the protein as the baseline comparisons to assess the mutation effect. 2. Significant differences in AABP values between site 681 and site 614 are noted. P681 has a much larger AABP than D614 due to their locations and interactions with other AAs. 3. Mutated R681 decreases the AABP by $1.082\text{ e}^- - 1.117\text{ e}^- = -0.035\text{ e}^-$. 4. Mutated G614 decreases the AABP by $0.904\text{ e}^- - 0.917\text{ e}^- = -0.013\text{ e}^-$. 5. The double mutation affects the changes in AABP for both sites: R681: $1.023\text{ e}^- - 1.117\text{ e}^- = -0.095\text{ e}^-$. G614: $0.901\text{ e}^- - 0.917\text{ e}^- = -0.015\text{ e}^-$. When single and double mutations are compared, the non-local AABP of R681 decreases by $0.047\text{ e}^- - 0.111\text{ e}^- = -0.064\text{ e}^-$ in case of double mutation. 6. Please note that the contribution from the NL and HB part is a substantial portion of the total AABP.

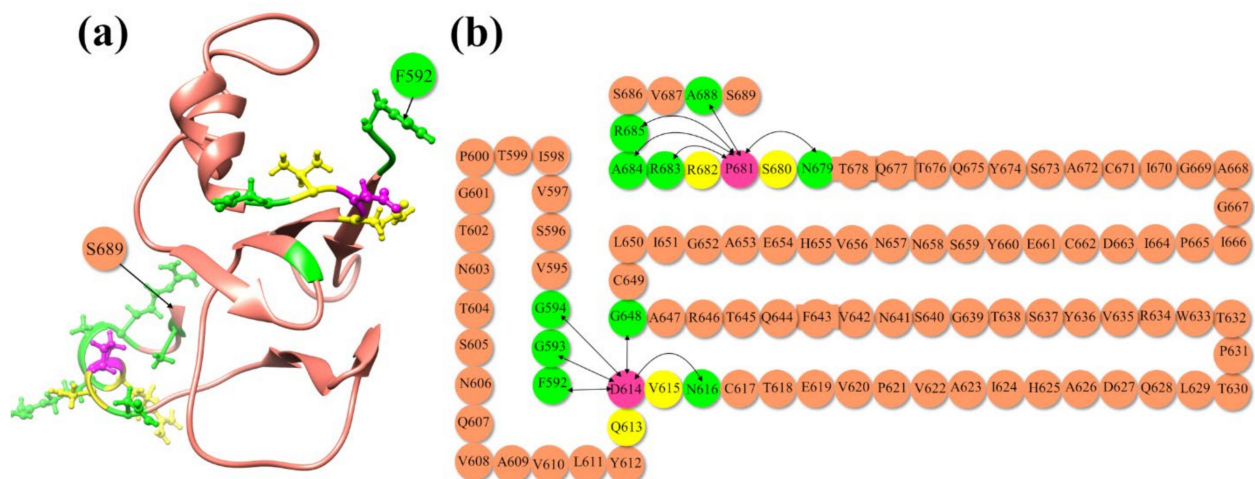


Figure 3. Interaction of D614 and P681 to their NN and NL AAs in WT SD2-FP model. (a) Ribbon structure from residue F592 to S689. (b) Sketch of AA sequence from F592 to S689 showing AABP interaction for D614 and P681 with joining lines. Both D614 and P681 are shown in pink color with their NN in yellow and NL interaction in green.

Figure 4 shows more vividly the non-local AA-AA interactions of mutations in the Delta variant. They are divided into six panels in two columns: (a) WT D614 and (b) WT P681; (c) mutated G614 and (d) mutated R681; and (e) double mutation G614-R681 G614 and (f) R681. In each case, the ball and stick sketch of all participating AAs is shown (red, O, grey, C, blue N, white, H). The focused AA is marked light pink. Its two NNs are marked light yellow, and its interacting NL AAs are marked light green. All interactions are marked by solid lines and the dashed lines show HBs. All of these NL interactions with the bonds formed are listed in Tables S5–S7. At the lower part of each figure, the three smaller figures show the same figure rotated for 90° , 180° , 270° from left to right. These figures show some of the most detailed information on the AA-AA interaction at the atomic-scale summarized as follows.

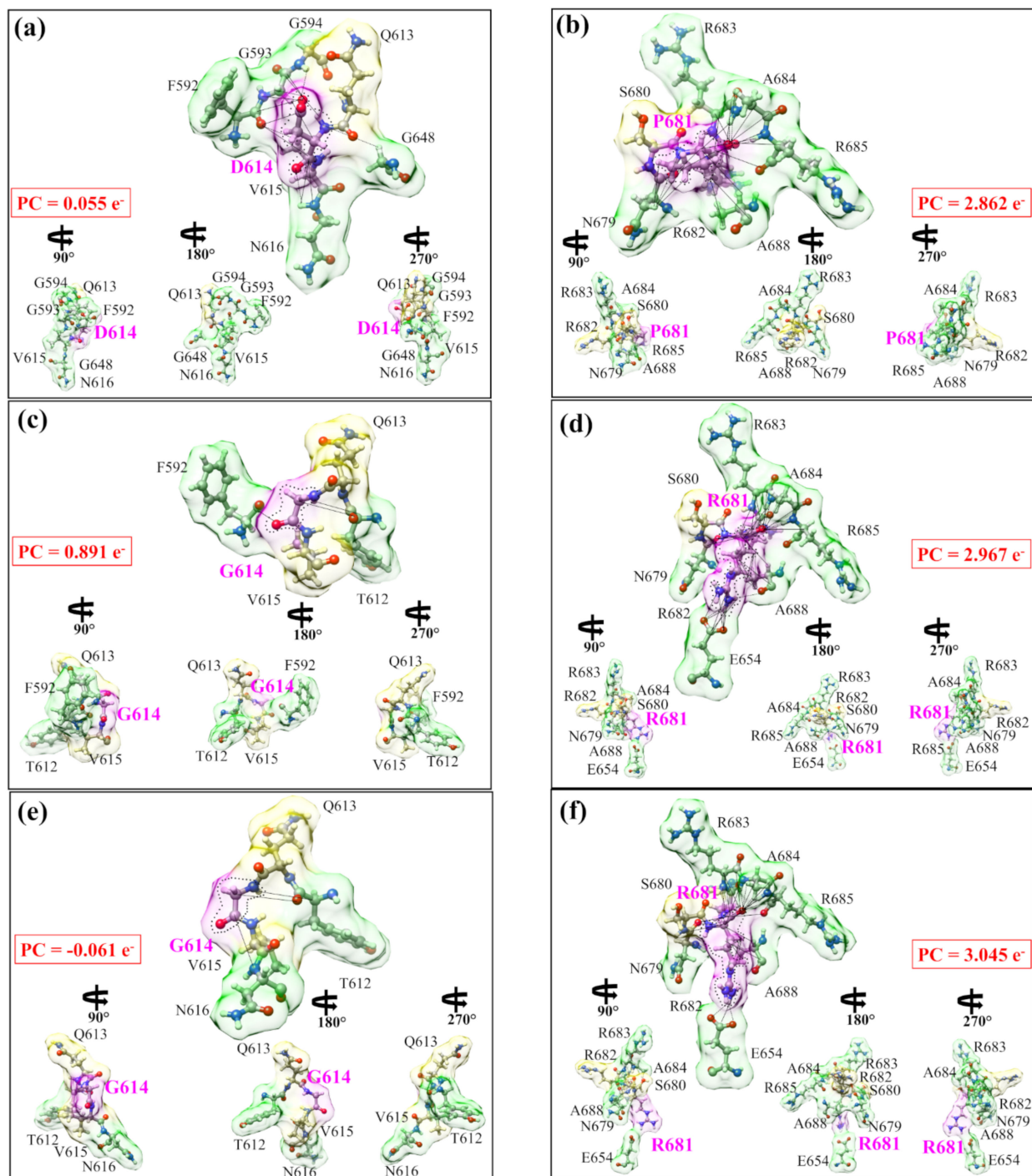


Figure 4. Details of the actual distribution of the three cases of interactions in the AABP calculation: (a,b) for the WT, (c,d) for the single mutation, and (e,f) for the double mutation. The left panel in each case is centered on 614th site and the right panel is centered on 681st site, and are both colored pink. The NN AAs are colored light yellow, and the NL AAs are colored light green. All the AAs involved in the interaction are marked. The lines show the NL bonded pairs. For each case, we show the NN AAs and the NL AAs that contribute to the AABP values. It also shows drastic difference between the mutation D614G and P681R in the shape, size, and orientation, and the total number of AAs involved in each case. The PCs for each groups of AAs in the AABP calculation are listed in the red box, showing large difference both in the sign and the magnitude.

(1) The WT D614 and WT P681 in (a) and (b) at the two different locations in the S-protein have very different features in size, shape, and volume controlled by their inter-atomic interactions with the two NN and five NL AAs. (2) The mutated AAs G614 and R681 in (c) and (d) are drastically different from the WT case in (a) and (b). (3) The double mutation in (e) and (f) also show significant differences with WT and interacts with the same number of NL AAs as in a single mutation.

These graphical illustrations demonstrate the complex structural features of mutations and interaction details never revealed before. In particular, the presence or the lack of HBs within the same AA or with other AAs have not been sufficiently elaborated in the current literature in biochemical interactions, except in a few isolated cases. In the same figure, the partial charge on each group is shown in the red boxes. These PC values are obtained by summing PC values of individual AAs in each interacting group. In the WT, both groups involving D614 and P681 have a positive PC. After mutation, G614 and its interacting group have their PC significantly increase, whereas R681 and its interacting group have their positive PC only slightly increase. More surprisingly, in the case of double mutation, G614 with its group have their PC slightly negative and P681 with its group have their positive PC continue to increase. This is another solid instance of evidence for the strong effect of mutation on different AAs that has never been discussed or revealed before. Similarly, the detailed interaction of the 950th site in two HR1-CH models is shown in Figure S7, with their AABP listed in Table S8, their NL bonding listed in Table S9, and their results discussed in Section S2 in SM.

5. Discussions

5.1. The Origins of Mutation

Viruses can undergo frequent genetic mutations, including point mutations (the source of genetic variation) and recombination [58]. Mutations are nucleotide changes that result in AA sequence changes implying new phenotype variants, whereas recombination allows for these variants to move across genomes to produce new haplotypes. Recombination occurs when viruses containing variants with different mutations infect the same host cell and exchange the genetic segments [58]. The fate of these genetic changes will be ultimately determined by natural selection and genetic drift, so it is very difficult to forecast when a viral mutation will become globally dominant. Although coronaviruses have an exonuclease enzyme that reduces their replication error rates, they accumulate mutations and generate more diversity via recombination [57,59,60]. SARS-CoV-2 itself is most likely the result of a recombination between different SARS-related coronaviruses [61], with different subsequent mutations affecting their many biological and biomedical properties, such as pathogenicity, infectivity, transmissibility, and/or antigenicity, even though they tend to be either deleterious and quickly purged, or relatively neutral [62].

One of the most urgent tasks in the virus research is the origin of virus mutations and how to predict new variants even before they occur. We assert that the first task must be related to the interaction between AAs at the atomic level that results in the structural modification in the S-protein due to mutated AAs. It must involve the interaction with non-local AAs in addition to the NN AAs. The contribution to AABP from hydrogen bonding is critical, and the role of HB has been recognized by all researchers but seldom explored in detail. In a much broader sense, the origin of mutation is not limited just to SARS-CoV-2 research per se, but is related to broader themes of evolutionary biology, such as the origin of species [63]. This accentuates the importance of a fundamental understanding of biomolecular interactions.

The data in Table 1 reveal that D614G and P681R mutations have lower AABP values than the unmutated WT case. Our results make sense for the following reasons. First, the substitution of D614 with G results in losing the sidechain, leading to the elimination of many intramolecular interactions in the same protomer, as shown in Figure 4. More specifically, our result elucidates that this mutation disrupts the non-local network interactions (Table 1). This could have large structure consequences in other domains of the S-protein,

such as in promoting the up conformation of the S-protein or enhancing the cleavage site, as reported before [29,30]. This enhancement in the cleavage site could be due to an increase in the flexibility of the mutated 614 and 681 sites. In addition to these intra-protomer interactions, it has been structurally demonstrated that the D614G mutation destroys an inter-protomer hydrogen bond between D614 (chain A) and T859 (chain B) [64]. However, the SD2-FP model alone is insufficient in assessing these significant conformational changes. It is necessary to include all atoms of the S protein trimer, which is currently impossible to perform in a single ab initio calculation. Second, our decomposition of the total AABP into NN and NL AABPs reveals that the R681 increases the NL AABP by forming new HBs and decreases the NN AABP as compared to P681 (Table 1 and Figure 4). Importantly, the P681R mutation reduces the local rigidity, as evidenced by the NN AABP values (Table 1). Additionally, the proline is well-known as the most rigid AA, and when it is mutated to arginine at position 681, it loses its rigidity. Furthermore, the positive charge associated with R681 appears to alter virus tropism via enhancing S1/S2 cleavage, as previously demonstrated in human airway epithelial cells [28]. This coincides with our conclusion that mutation decreases AABP but also increases the flexibility of AAs.

Such understanding of the molecular and atomic origins of the individual mutations or their combinations in SARS-CoV-2 can provide deep information to prepare and prevent future outbreaks, such as those reported for AY.4.2. or the just-emerging “Omicron” VOC. It can also play an important role in guiding the development of new drugs. It would also be desirable to have a quantitative scale from 1 (insignificant) to 10 (most dangerous) to quantify the nature of mutations by linking the mutation to specific clinical data or research using other methods, such as experimental or computational.

5.2. Extension to Machine Learning (ML)

Over the last decade or so, machine learning (ML) has become a very powerful tool that is being applied to many different areas, including image, speech, text and facial recognition, autonomous vehicles, medical image classification, instruction detection, finances, drones, national defense, etc. [65,66]. In the present work, the word ML is strictly used only for the calculated data of AABP between AAs in the S-protein to predict potential unknown mutations.

One of the major challenges we face in applying ML techniques to our problem is the size of the dataset. This is because each data instance of a Delta variant model is computationally expensive to generate. When dealing with small datasets, deep learning techniques (based on artificial neural networks) will tend to underperform. Hence, conventional ML techniques should be preferred. Our first step is to prepare the data by constructing feature vectors for different Delta variant models. We can represent each model as a five-dimensional feature vector: TAABP (total AABP), NN (NN AABP), NL (non-local AABP), HB (AABP from HB), and>NNL (no. of NN AAs). Each feature is a continuous variable. The target vector is denoted by a binary variable MT to represent whether a Delta variant model is unmutated (0) or mutated (1). Table 2 shows how the data can be represented for the Delta variant models, shown in Table 1 for SD1-FP and Table S8 for HR1-CH. Several extensions can be made to the target vector depending on the prediction task. Suppose we wish to predict the site of the Delta variant model (e.g., 614, 681). A categorical variable can be introduced as the target vector. To predict different kinds of mutations (e.g., single, double), another categorical variable can be introduced in the target vector.

Suppose we wish to predict whether a new Delta variant model is mutated or not. Using the feature and target vectors in Table 2, a binary classifier can be learned on the data. Classifiers can be built using different techniques, such as (a) a logistic regression classifier [67], (b) Gaussian naïve Bayes classifier [67], (c) support vector machine (SVM) classifier [68], (d) decision tree classifier [69], (e) random forest classifier [70], and (f) the extreme gradient boosting (XGBoost) classifier [71]. Hyperparameter tuning is necessary for the different classifiers to achieve the best accuracy. Feature importance is another

task that can be valuable to researchers. For instance, the XGBoost classifier trained on data in Table 2 showed that TAABP and HB were the most important features to predict a mutation. XGBoost uses the notion of gain, which is the relative contribution of a feature to the model, to compute feature importance. The above ML techniques assume that the data instances are independently and identically distributed (i.i.d). However, if this assumption is not appropriate, then ML techniques, such as Markov logic networks [72], should be considered.

Table 2. Data representation for ML.

TAABP	NN	Feature Vector			Target MT
		NL	HB	NNL	
1.117	1.064	0.054	0.064	5	0
0.917	0.912	0.005	0.040	5	0
1.082	0.971	0.111	0.122	6	1
0.904	0.904	0.001	0.040	2	1
1.023	0.975	0.047	0.066	6	1
0.901	0.901	0.001	0.041	2	1
1.163	1.021	0.143	0.154	6	0
1.093	1.009	0.084	0.106	6	1

We can also learn causal relationships among the features using a probabilistic graphical model, such as a Bayesian network [73]. By better understanding the cause–effect relationships among the features, better prediction models can be developed. A Bayesian network can compactly encode the joint distribution of a set of random variables as the product of the conditional probability of the nodes given its parents in the network. Using a Bayesian network, we can also simulate new data that follow the distribution learned by the network. Table 3 shows a set of four samples obtained by simulating random samples from a Bayesian network. (This network was learned using the Max-Min Hill Climbing algorithm [74] over the data shown in Table 2.) This algorithm first learns the skeleton of a Bayesian network. It then uses a greedy hill-climbing search to orient the edges using a Bayesian scoring function [74]).

Table 3. Data obtained by simulating random samples from a Bayesian network.

TAABP	NN	Feature Vector			Target MT
		NL	HB	NNL	
1.064	1.024	0.041	0.145	7	0
1.181	1.098	0.084	0.089	4	1
1.068	0.993	0.075	0.055	4	0
1.113	1.057	0.056	0.119	5	1

Figure 5 shows the general steps involved in applying supervised ML to our Delta variant data. In summary, ML can offer unprecedented opportunities to predict mutations in the Delta variant.

5.3. Looking Forward

Looking forward, we would like to speculate where our methodology could lead to in the future. Obviously, it can be extended to other mutations in the S-protein, such as RBD and NTD and their interfaces with ACE2, or maybe even those newly emerged mutations, such as AY.4.2 [17] and B.1.1.529 [75]. The obvious goal is to increase the mutation data points so that a reasonable size of the AABP database can be used in the ML, as discussed in the previous section. It is totally possible to calculate other potential mutations for more data points in different domains based on the insights obtained with the calculations on

known mutations. The list is endless, restricted by the large computational resources it demands. Fortunately, the emergence of the next generation of the exa-scale supercomputer is already available [76] to meet these challenges.

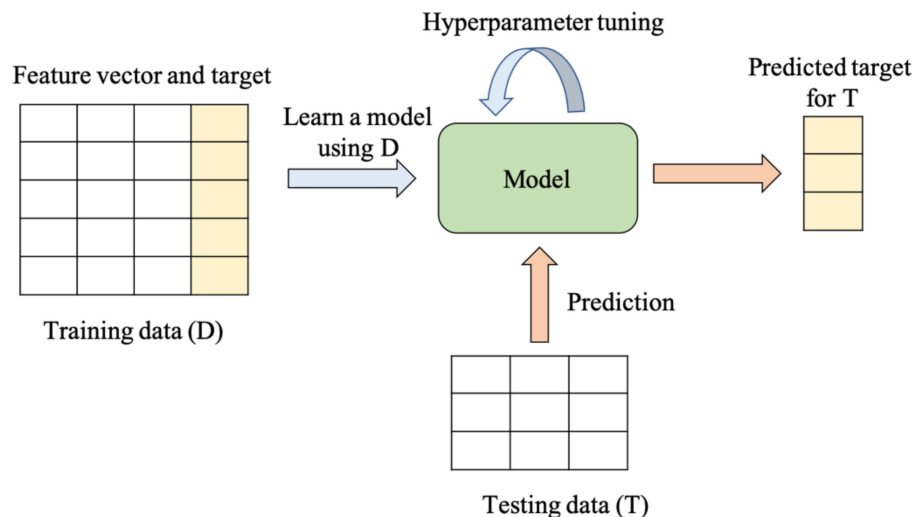


Figure 5. The general workflow of how supervised ML works is shown. Here, D and T represent training data and testing data, respectively.

However, even more importantly, we posit that our methodology could be extremely valuable for the VOC that appear to be characterized by an unprecedented number of concurrent mutations—as appears to be the case in the latest Omicron VOC [75]—with more than 30 mutations per S-protein. Clearly, in such cases, there must be collective effects tying together several mutations, possibly also enhancing the susceptibility of different AAs to mutations. The mutation quantifier based on AABP, as introduced in this paper, which, by its very nature, already embodies the collective effects of mutations, is clearly a good candidate for such an analysis. In view of this striking development of the SARS-CoV-2 mutational capacity, our research seems to represent a well-placed origin for further elaboration, not only of the effects of single mutations, but, maybe even more importantly, their interaction and synergy.

One of the fundamental questions in biology is to ask if mutations are random or if there are specific reasons for each mutation. We may be able to also shed some light on this issue by accumulating a large database for known mutations and applying the ML protocol to see if any successful prediction can be verified with a global database or clinical data. Thus far, we have six data points, plus another four from RBD, or ten data points altogether.

The current calculation and analysis are restricted to chain A of the S-protein of SARS-CoV-2. An extension to cases with chains A, B, and C in up and down conformations would be highly desirable and important. It is also opportune to start looking at mutations in the RBD of the S-protein, such as those involved in binding to the ACE2 receptor, to provide a deeper understanding of how the virus can mutate and overcome the human defense mechanisms of immune response, intimately related to vaccination. Most neutralizing monoclonal antibodies (mAbs) target either the RBD or the NTD of the S-protein, and mutations in these domains have the greatest impact on neutralization.

The broader implications of the present work would also be in protein–protein interactions, where one could define similar *ab initio* interaction quantifiers based on single-point calculations as applied to a protein–protein network, protein–protein mapping, application to cancer research, etc. This broader phenomenology also revolves around mutations (replacing AA at a specific site, some AAs are more important than others), but on a much larger scale and with much more detailed interactions. While, so far, such studies were

essentially based on experimental or clinical observations, the lack of a firm fundamental theoretical basis is noticeable.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/v14030465/s1>, Figure S1: Detailed schematic of S-protein in SARS-CoV-2 colored by domains: SP, NTD, RBD, SD2, furin cleavage site (S1/S2), FP, HR1, CH, CD, HR2, TM, and CT. The delta variant mutation sites are marked by gray solid line in the bottom. In the present work, our calculations have been mainly carried out on region 3 of SD2 and FP domains; Figure S2: The 20 amino acids in Table S1 are divided into 6 functional groups based on the nature of their sidechains: (a,b) for hydrophobic AAs, (c) for neutral polar AAs, (d,e) for charged polar AAs, and (f) for unique AAs; Figure S3: Calculated total density of states (TDOS) for WT SD2-FP model; Figure S4: Bar graph with PC distribution for WT SD2-FP. PC values are marked for the two mutation sites 614 and 681 showing values in different color for different mutation cases; Figure S5: (a) Comparison of PC on the solvent excluded surface between P681 and D614. (b) 180° orientation of (a); Figure S6: Bar graph with PC distribution for WT HR1-CH. PC values are marked for the two mutation sites 950 showing values in different color for D950N mutation; Figure S7: Interactions of 950th site in two HR1-CH models in Delta variant: (a) WT D950 and (b) mutated N950. In each case, the ball and stick sketch of all participating AAs are shown (red, O, grey, C, blue N, white, H). The focused 950th AA is marked light pink. Its two NNs are marked light yellow, and its interacting NL AAs are marked light green. All NL interactions are marked by solid lines and dashed lines show HBs. All these NL interactions with the bonds formed are shown in Table S9. At the lower part of each of the figure, three smaller figures show the same figure rotated for 90°, 180°, 270° from left to right. These figures show some of the most detailed information on the AA-AA interaction at atomic scale. In the same figure, the partial charge on each group is shown in the red boxes. These PC values are obtained summing PC values of individual AAs in each interacting group. The mutated case has distinctly higher positive PC; Table S1: 20 canonical amino acids in alphabetical order. The last column shows their functional group; Table S2: Four SD2-FP models (a–d) and two HR1-CH models (e,f) with the number of atoms, total energy and time used in Cori; Table S3: List of PC value for each amino acids with their sequence number for WT SD2-FP; Table S4: List of PC value for each amino acids with their sequence number for WT HR1-CH; Table S5: NL bonding information for WT SD2-FP shown in Figure 4a,b; Table S6: NL bonding information for mutated R681 shown in Figure 4c,d; Table S7: NL bonding information for double mutation R681 & G614 shown in Figure 4e,f; Table S8: AABP of two HR1-CH models; Table S9: Bonding information for WT and mutated HR1-CH shown in Figure S7. References [77–89] are cited in the Supplementary Materials.

Author Contributions: W.-Y.C. and P.A. conceived the project. W.-Y.C. and P.A. performed the calculations. P.A. and B.J. made most of the figures. W.-Y.C., P.A. and B.J. drafted the paper with inputs from R.P. and P.R. All authors participated in the discussion and interpretation of the results. All authors edited and proofread the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This project is funded partly by the National Science Foundation of USA: RAPID DMR/CMMT-2028803. P.R. was partly funded by the National Science Foundation of USA: RAPID CISE/CNS-2034247.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are listed in tables or presented in the figures in main text or Supplementary Materials.

Acknowledgments: This research used the resources of the National Energy Research Scientific Computing Center supported by DOE under Contract No. DE-AC03-76SF00098 and the Research Computing Support Services (RCSS) of the University of Missouri System. We thank Richard Gerber, Senior Science Advisor and HPC Department Head for special allocations. This project is funded partly by the National Science Foundation of USA: RAPID DMR/CMMT-2028803. P.R. was partly funded by the National Science Foundation of USA: RAPID CISE/CNS-2034247. RP acknowledges funding from the Key project #12034019 of the National Natural Science Foundation of China.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhong, N.; Zheng, B.; Li, Y.; Poon, L.; Xie, Z.; Chan, K.; Li, P.; Tan, S.; Chang, Q.; Xie, J. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *Lancet* **2003**, *362*, 1353–1358. [CrossRef]
2. Rambaut, A.; Loman, N.; Pybus, O.; Barclay, W.; Barrett, J.; Carabelli, A.; Connor, T.; Peacock, T.; Robertson, D.L.; Volz, E.; et al. Preliminary Genomic Haracterization of an Emergent SARS-CoV-2 Lineage in the UK Defined by a Novel Set of Spike Mutations. SARS-CoV-2 Coronavirus nCoV-2019 Genomic Epidemiology. Virological 2020. Available online: <https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563> (accessed on 22 September 2021).
3. Tegally, H.; Wilkinson, E.; Giovanetti, M.; Iranzadeh, A.; Fonseca, V.; Giandhari, J.; Doolabh, D.; Pillay, S.; San, E.J.; Msomi, N.; et al. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *MedRxiv* **2020**. [CrossRef]
4. Singh, J.; Rahman, S.A.; Ehtesham, N.Z.; Hira, S.; Hasnain, S.E. SARS-CoV-2 variants of concern are emerging in India. *Nat. Med.* **2021**, *27*, 1131–1133. [CrossRef] [PubMed]
5. Faria, N.R.; Claro, I.M.; Candido, D.; Franco, L.M.; Andrade, P.S.; Coletti, T.M.; Silva, C.A.; Sales, F.C.; Manuli, E.R.; Aguiar, R.S.; et al. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: Preliminary findings. *Virological* **2021**, *372*, 815–821.
6. Ozer, E.A.; Simons, L.M.; Adewumi, O.M.; Fowotade, A.A.; Omoruyi, E.C.; Adeniji, J.A.; Dean, T.J.; Taiwo, B.O.; Hultquist, J.F.; Lorenzo-Redondo, R. High prevalence of SARS-CoV-2 B. 1.1. 7 (UK variant) and the novel B. 1.5. 2.5 lineage in Oyo State, Nigeria. *MedRxiv* **2021**. [CrossRef]
7. Annavajhala, M.K.; Mohri, H.; Zucker, J.E.; Sheng, Z.; Wang, P.; Gomez-Simmonds, A.; Ho, D.D.; Uhlemann, A.-C. A novel SARS-CoV-2 variant of concern, B. 1.526, identified in New York. *MedRxiv* **2021**. [CrossRef]
8. Liu, C.; Ginn, H.M.; Dejnirattisai, W.; Supasa, P.; Wang, B.; Tuekprakhon, A.; Nutalai, R.; Zhou, D.; Mentzer, A.J.; Zhao, Y. Reduced neutralization of SARS-CoV-2 B. 1.617 by vaccine and convalescent serum. *Cell* **2021**, *184*, 4220.e13–4236.e13. [CrossRef]
9. Kimura, I.; Kosugi, Y.; Wu, J.; Yamasoba, D.; Butlertanaka, E.P.; Tanaka, Y.L.; Liu, Y.; Shirakawa, K.; Kazuma, Y.; Nomura, R.; et al. SARS-CoV-2 Lambda variant exhibits higher infectivity and immune resistance. *BioRxiv* **2021**. [CrossRef]
10. Laiton-Donato, K.; Franco-Munoz, C.; Alvarez-Diaz, D.A.; Ruiz-Moreno, H.; Usme-Ciro, J.; Prada, D.; Reales, J.; Corchuelo, S.; Herrera-sepulveda, M.; Naizaque, J.; et al. Characterization of the emerging B. 1.621 variant of interest of SARS-CoV-2. *MedRxiv* **2021**. [CrossRef]
11. Reardon, S. How the Delta variant achieves its ultrafast spread. *Nature* **2021**, *21*. [CrossRef]
12. Krishnan, L.; Ogunwole, S.M.; Cooper, L.A. Historical Insights on Coronavirus Disease 2019 (COVID-19), the 1918 Influenza Pandemic, and Racial Disparities: Illuminating a Path Forward. *Ann. Intern. Med.* **2020**, *173*, 474–481. [CrossRef] [PubMed]
13. A Timeline of HIV and AIDS. Available online: <https://www.hiv.gov/hiv-basics/overview/history/hiv-and-aids-timeline> (accessed on 22 September 2021).
14. Hemida, M.; Perera, R.; Wang, P.; Alhammedi, M.; Siu, L.; Li, M.; Poon, L.; Saif, L.; Alnaeem, A.; Peiris, M. Middle East Respiratory Syndrome (MERS) coronavirus seroprevalence in domestic livestock in Saudi Arabia, 2010 to 2013. *Eurosurveillance* **2013**, *18*, 20659. [CrossRef]
15. Preventing the Spread of the Coronavirus. Available online: <https://www.health.harvard.edu/diseases-and-conditions/preventing-the-spread-of-the-coronavirus> (accessed on 9 December 2020).
16. Understanding How COVID-19 Vaccines Work. Available online: <https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines/how-they-work.html> (accessed on 2 November 2020).
17. COVID-19 Genomic Surveillance. Available online: <https://covid19.sanger.ac.uk/lineages/raw> (accessed on 1 November 2021).
18. Wrapp, D.; Wang, N.; Corbett, K.S.; Goldsmith, J.A.; Hsieh, C.-L.; Abiona, O.; Graham, B.S.; McLellan, J.S. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **2020**, *367*, 1260–1263. [CrossRef] [PubMed]
19. Peacock, T.P.; Goldhill, D.H.; Zhou, J.; Baillon, L.; Frise, R.; Swann, O.C.; Kugathasan, R.; Penn, R.; Brown, J.C.; Sanchez-David, R.Y.; et al. The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. *Nat. Microbiol.* **2021**, *6*, 899–909. [CrossRef] [PubMed]
20. Walls, A.C.; Park, Y.-J.; Tortorici, M.A.; Wall, A.; McGuire, A.T.; Veisler, D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **2020**, *181*, 281–292. [CrossRef]
21. Hoffmann, M.; Kleine-Weber, H.; Schroeder, S.; Krüger, N.; Herrler, T.; Erichsen, S.; Schiergens, T.S.; Herrler, G.; Wu, N.-H.; Nitsche, A.; et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **2020**, *181*, 271.e8–280.e8. [CrossRef]
22. Coutard, B.; Valle, C.; de Lamballerie, X.; Canard, B.; Seidah, N.G.; Decroly, E. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antivir. Res.* **2020**, *176*, 104742. [CrossRef]
23. Xia, S.; Lan, Q.; Su, S.; Wang, X.; Xu, W.; Liu, Z.; Zhu, Y.; Wang, Q.; Lu, L.; Jiang, S. The role of furin cleavage site in SARS-CoV-2 spike protein-mediated membrane fusion in the presence or absence of trypsin. *Signal. Transduct. Target. Ther.* **2020**, *5*, 1–3. [CrossRef]
24. Jaimes, J.A.; Millet, J.K.; Whittaker, G.R. Proteolytic cleavage of the SARS-CoV-2 spike protein and the role of the novel S1/S2 site. *iScience* **2020**, *23*, 101212. [CrossRef]

25. Papa, G.; Mallery, D.L.; Albecka, A.; Welch, L.G.; Cattin-Ortolá, J.; Luptak, J.; Paul, D.; McMahon, H.T.; Goodfellow, I.G.; Carter, A. Furin cleavage of SARS-CoV-2 Spike promotes but is not essential for infection and cell-cell fusion. *PLoS Pathog.* **2021**, *17*, e1009246. [CrossRef]
26. Liu, Y.; Liu, J.; Johnson, B.A.; Xia, H.; Ku, Z.; Schindewolf, C.; Widen, S.G.; An, Z.; Weaver, S.C.; Menachery, V.D. Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant. *BioRxiv* **2021**. [CrossRef]
27. Peacock, T.P.; Sheppard, C.M.; Brown, J.C.; Goonawardane, N.; Zhou, J.; Whiteley, M.; de Silva, T.I.; Barclay, W.S.; Consortium, P.V. The SARS-CoV-2 variants associated with infections in India, B. 1.617, show enhanced spike cleavage by furin. *BioRxiv* **2021**. [CrossRef]
28. Saito, A.; Nasser, H.; Uriu, K.; Kosugi, Y.; Irie, T.; Shirakawa, K. SARS-CoV-2 spike P681R mutation enhances and accelerates viral fusion. *BioRxiv* **2021**, *10*, 17.448820.
29. Zhang, L.; Jackson, C.B.; Mou, H.; Ojha, A.; Peng, H.; Quinlan, B.D.; Rangarajan, E.S.; Pan, A.; Vanderheiden, A.; Suthar, M.S. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat. Commun.* **2020**, *11*, 6013. [CrossRef]
30. Gobeil, S.M.-C.; Janowska, K.; McDowell, S.; Mansouri, K.; Parks, R.; Manne, K.; Stalls, V.; Kopp, M.F.; Henderson, R.; Edwards, R.J.; et al. D614G mutation alters SARS-CoV-2 spike conformation and enhances protease cleavage at the S1/S2 junction. *Cell Rep.* **2021**, *34*, 108630. [CrossRef]
31. Adhikari, P.; Ching, W.-Y. Amino acid interacting network in the receptor-binding domain of SARS-CoV-2 spike protein. *RSC Adv.* **2020**, *10*, 39831–39841. [CrossRef]
32. Woo, H.; Park, S.-J.; Choi, Y.K.; Park, T.; Tanveer, M.; Cao, Y.; Kern, N.R.; Lee, J.; Yeom, M.S.; Croll, T.I. Developing a fully glycosylated full-length SARS-CoV-2 spike protein model in a viral membrane. *J. Phys. Chem. B* **2020**, *124*, 7128–7137. [CrossRef]
33. CHARMM-GUI Archive—COVID-19 Proteins Library. Available online: <https://charmm-gui.org/?doc=archive&lib=covid19> (accessed on 1 November 2021).
34. Case, D.A.; Betz, R.; Cerutti, D.; Cheatham, T.; Darden, T.; Duke, R.; Giese, T.; Gohlke, H.; Goetz, A.; Homeyer, N. AMBER 2020 Reference Manual. University of California, San Francisco, 2020. Available online: <https://ambermd.org/Manuals.php> (accessed on 1 November 2021).
35. Case, D.A.; Cheatham, T.E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K.M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R.J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688. [CrossRef] [PubMed]
36. Pearlman, D.A.; Case, D.A.; Caldwell, J.W.; Ross, W.S.; Cheatham, T.E., III; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **1995**, *91*, 1–41. [CrossRef]
37. Dunbrack, R.L., Jr. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **2002**, *12*, 431–440. [CrossRef]
38. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612. [CrossRef] [PubMed]
39. VASP—Vienna Ab Initio Simulation Package. Available online: <https://www.vasp.at/> (accessed on 1 November 2021).
40. Ching, W.-Y.; Rulis, P. *Electronic Structure Methods for Complex Materials: The Orthogonalized Linear Combination of Atomic Orbitals*; Oxford University Press: Oxford, UK, 2012.
41. Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A. PubChem substance and compound databases. *Nucleic Acids Res.* **2016**, *44*, D1202–D1213. [CrossRef] [PubMed]
42. Ryadnov, M.; Hudecz, F. *Amino Acids, Peptides and Proteins*; Royal Society of Chemistry: Cambridge, UK, 2017; Volume 42.
43. Khan, R.J.; Jha, R.K.; Amera, G.M.; Jain, M.; Singh, E.; Pathak, A.; Singh, R.P.; Muthukumaran, J.; Singh, A.K. Targeting SARS-CoV-2: A systematic drug repurposing approach to identify promising inhibitors against 3C-like proteinase and 2'-O-ribose methyltransferase. *J. Biomol. Struct. Dyn.* **2021**, *39*, 2679–2692. [CrossRef] [PubMed]
44. Wang, Y.; Liu, M.; Gao, J. Enhanced receptor binding of SARS-CoV-2 through networks of hydrogen-bonding and hydrophobic interactions. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 13967–13974. [CrossRef] [PubMed]
45. Marx, D.; Hutter, J. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*; Cambridge University Press: New York, NY, USA, 2009.
46. Adhikari, P.; Li, N.; Shin, M.; Steinmetz, N.F.; Twarock, R.; Podgornik, R.; Ching, W.-Y. Intra- and intermolecular atomic-scale interactions in the receptor binding domain of SARS-CoV-2 spike protein: Implication for ACE2 receptor binding. *Phys. Chem. Chem. Phys.* **2020**, *22*, 18272–18283. [CrossRef]
47. Ching, W.-Y.; Adhikari, P.; Jawad, B.; Podgornik, R. Ultra-Large-Scale Ab Initio Quantum Chemical Computation of Bio-Molecular Systems: The Case of Spike Protein of SARS-CoV-2 Virus. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1288–1301. [CrossRef]
48. Jawad, B.; Adhikari, P.; Podgornik, R.; Ching, W.-Y. Key interacting residues between RBD of SARS-CoV-2 and ACE2 receptor: Combination of molecular dynamic simulation and density functional calculation. *J. Chem. Inf. Model.* **2021**, *61*, 4425–4441. [CrossRef]
49. Adhikari, P.; Podgornik, R.; Jawad, B.; Ching, W.-Y. First-Principles Simulation of Dielectric Function in Biomolecules. *Materials* **2021**, *14*, 5774. [CrossRef]
50. Baral, K.; Adhikari, P.; Jawad, B.; Podgornik, R.; Ching, W.-Y. Solvent Effect on the Structure and Properties of RGD Peptide (1FUV) at Body Temperature (310 K) Using Ab Initio Molecular Dynamics. *Polymers* **2021**, *13*, 3434. [CrossRef]

51. Liang, L.; Rulis, P.; Ouyang, L.; Ching, W. Ab initio investigation of hydrogen bonding and network structure in a supercooled model of water. *Phys. Rev. B* **2011**, *83*, 024201. [\[CrossRef\]](#)
52. Vermeeren, P.; van Zeist, W.-J.; Hamlin, T.A.; Guerra, C.F.; Bickelhaupt, F.M.; Bickelhaupt, F.; Guerra, C.F. Not Carbon s-p Hybridization, but Coordination Number Determines C–H and C–C Bond Length. *Chem. A Eur. J.* **2021**, *27*, 7074–7079. [\[CrossRef\]](#) [\[PubMed\]](#)
53. Mlcochova, P.; Kemp, S.; Dhar, M.S.; Papa, G.; Meng, B.; Ferreira, I.A.; Datir, R.; Collier, D.A.; Albecka, A.; Singh, S. SARS-CoV-2 B. 1.617. 2 Delta variant replication and immune evasion. *Nature* **2021**, *599*, 114–119. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Lopez Bernal, J.; Andrews, N.; Gower, C.; Gallagher, E.; Simmons, R.; Thelwall, S.; Stowe, J.; Tessier, E.; Groves, N.; Dabrera, G. Effectiveness of COVID-19 vaccines against the B. 1.617. 2 (Delta) variant. *N. Engl. J. Med.* **2021**, *385*, 585–594. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Kannan, S.R.; Spratt, A.N.; Cohen, A.R.; Naqvi, S.H.; Chand, H.S.; Quinn, T.P.; Lorson, C.L.; Byrareddy, S.N.; Singh, K. Evolutionary analysis of the Delta and Delta Plus variants of the SARS-CoV-2 viruses. *J. Autoimmun.* **2021**, *124*, 102715. [\[CrossRef\]](#)
56. Rajah, M.M.; Hubert, M.; Bishop, E.; Saunders, N.; Robinot, R.; Grzelak, L.; Planas, D.; Dufloo, J.; Gellenoncourt, S.; Bongers, A. SARS-CoV-2 Alpha, Beta, and Delta variants display enhanced Spike-mediated syncytia formation. *EMBO J.* **2021**, *40*, e108944. [\[CrossRef\]](#)
57. Tao, K.; Tzou, P.L.; Nouhin, J.; Gupta, R.K.; de Oliveira, T.; Kosakovsky Pond, S.L.; Fera, D.; Shafer, R.W. The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet.* **2021**, *22*, 757–773. [\[CrossRef\]](#)
58. Pérez-Losada, M.; Arenas, M.; Galán, J.C.; Palero, F.; González-Candelas, F. Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences. *Infect. Genet. Evol.* **2015**, *30*, 296–307. [\[CrossRef\]](#)
59. Duffy, S.; Shackelton, L.A.; Holmes, E.C. Rates of evolutionary change in viruses: Patterns and determinants. *Nat. Rev. Genet.* **2008**, *9*, 267–276. [\[CrossRef\]](#)
60. Graham, R.L.; Baric, R.S. Recombination, reservoirs, and the modular spike: Mechanisms of coronavirus cross-species transmission. *J. Virol.* **2010**, *84*, 3134–3146. [\[CrossRef\]](#)
61. Boni, M.F.; Lemey, P.; Jiang, X.; Lam, T.T.-Y.; Perry, B.W.; Castoe, T.A.; Rambaut, A.; Robertson, D.L. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* **2020**, *5*, 1408–1417. [\[CrossRef\]](#)
62. Harvey, W.T.; Carabelli, A.M.; Jackson, B.; Gupta, R.K.; Thomson, E.C.; Harrison, E.M.; Ludden, C.; Reeve, R.; Rambaut, A.; Peacock, S.J. SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **2021**, *19*, 409–424. [\[CrossRef\]](#) [\[PubMed\]](#)
63. On the origin of Species. *The Economist*, 25 August 2021.
64. Yurkovetskiy, L.; Wang, X.; Pascal, K.E.; Tomkins-Tinch, C.; Nyalile, T.P.; Wang, Y.; Baum, A.; Diehl, W.E.; Dauphin, A.; Carbone, C.; et al. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* **2020**, *183*, 739–751.e8. [\[CrossRef\]](#) [\[PubMed\]](#)
65. Eshete, B. Making machine learning trustworthy. *Science* **2021**, *373*, 743–744. [\[CrossRef\]](#) [\[PubMed\]](#)
66. Goodfellow, I.; McDaniel, P.; Papernot, N. Making machine learning robust against adversarial inputs. *Commun. ACM* **2018**, *61*, 56–66. [\[CrossRef\]](#)
67. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*; Springer: New York, NY, USA, 2013; Volume 112.
68. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
69. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [\[CrossRef\]](#)
70. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
71. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 13 August 2016.
72. Richardson, M.; Domingos, P. Markov logic networks. *Mach. Learn.* **2006**, *62*, 107–136. [\[CrossRef\]](#)
73. Pearl, J. *Causality: Models, Reasoning and Inference*; Cambridge University Press: Cambridge, UK, 2000; p. 19.
74. Tsamardinos, I.; Brown, L.E.; Aliferis, C.F. The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* **2006**, *65*, 31–78. [\[CrossRef\]](#)
75. Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern. 2021. Available online: [https://www.who.int/news/item/26-11-2021-classification-of-omicron-\(b.1.1.529\)-sars-cov-2-variant-of-concern](https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern) (accessed on 28 November 2021).
76. NERSC Perlmutter. 2021. Available online: <https://www.nersc.gov/systems/perlmutter/> (accessed on 28 November 2021).
77. Perdew, J.P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865. [\[CrossRef\]](#)
78. Poudel, L.; Steinmetz, N.F.; French, R.H.; Parsegian, V.A.; Podgornik, R.; Ching, W.-Y. Implication of the solvent effect, metal ions and topology in the electronic structure and hydrogen bonding of human telomeric G-quadruplex DNA. *Phys. Chem. Chem. Phys.* **2016**, *18*, 21573–21585. [\[CrossRef\]](#) [\[PubMed\]](#)
79. Poudel, L.; Twarock, R.; Steinmetz, N.F.; Podgornik, R.; Ching, W.-Y. Impact of hydrogen bonding in the binding site between capsid protein and MS2 bacteriophage ssRNA. *J. Phys. Chem. B* **2017**, *121*, 6321–6330. [\[CrossRef\]](#)
80. Eifler, J.; Podgornik, R.; Steinmetz, N.F.; French, R.H.; Parsegian, V.A.; Ching, W.Y. Charge distribution and hydrogen bonding of a collagen α 2-chain in vacuum, hydrated, neutral, and charged structural models. *Int. J. Quantum Chem.* **2016**, *116*, 681–691. [\[CrossRef\]](#)

81. Poudel, L.; Wen, A.M.; French, R.H.; Parsegian, V.A.; Podgornik, R.; Steinmetz, N.F.; Ching, W.Y. Electronic structure and partial charge distribution of doxorubicin in different molecular environments. *ChemPhysChem* **2015**, *16*, 1451–1460. [[CrossRef](#)] [[PubMed](#)]
82. Poudel, L.; Rulis, P.; Liang, L.; Ching, W.-Y. Electronic structure, stacking energy, partial charge, and hydrogen bonding in four periodic B-DNA models. *Phys. Rev. E* **2014**, *90*, 022705. [[CrossRef](#)] [[PubMed](#)]
83. Adhikari, P.; Xiong, M.; Li, N.; Zhao, X.; Rulis, P.; Ching, W.-Y. Structure and electronic properties of a continuous random network model of an amorphous zeolitic imidazolate framework (a-ZIF). *J. Phys. Chem. C* **2016**, *120*, 15362–15368. [[CrossRef](#)]
84. Ching, W.Y.; Yoshiya, M.; Adhikari, P.; Rulis, P.; Ikuhara, Y.; Tanaka, I. First-principles study in an inter-granular glassy film model of silicon nitride. *J. Am. Ceram. Soc.* **2018**, *101*, 2673–2688. [[CrossRef](#)]
85. Ching, W.-Y.; San, S.; Brechtel, J.; Sakidja, R.; Zhang, M.; Liaw, P.K. Fundamental electronic structure and multiatomic bonding in 13 biocompatible high-entropy alloys. *npj Comput. Mater.* **2020**, *6*, 45. [[CrossRef](#)]
86. Jawad, B.; Poudel, L.; Podgornik, R.; Ching, W.-Y. Thermodynamic Dissection of the Intercalation Binding Process of Doxorubicin to dsDNA with Implications of Ionic and Solvent Effects. *J. Phys. Chem. B* **2020**, *124*, 7803–7818. [[CrossRef](#)]
87. Baral, K.; Li, A.; Ching, W.-Y. Ab Initio Study of Hydrolysis Effects in Single and Ion-Exchanged Alkali Aluminosilicate Glasses. *J. Phys. Chem. B* **2020**, *124*, 8418–8433. [[CrossRef](#)]
88. Mulliken, R.S. Electronic population analysis on LCAO–MO molecular wave functions. I. *J. Chem. Phys.* **1955**, *23*, 1833–1840. [[CrossRef](#)]
89. Mulliken, R. Electronic population analysis on LCAO–MO molecular wave functions. II. Overlap populations, bond orders, and covalent bond energies. *J. Chem. Phys.* **1955**, *23*, 1841–1846. [[CrossRef](#)]