Molecular Dynamics Modeling of the SARS-CoV-2 Spike Protein at PH2 Through PH11.5

Ziyuan Niu
Stony Brook University
Peng Zhang
Stony Brook University
Miriam Rafailovich
Stony Brook University
Marcia Simon
Stony Brook University
Meichen Song
Stony Brook University
Yuefan Deng (Yuefan.Deng@StonyBrook.edu)
Stony Brook University https://orcid.org/0000-0002-0068-4874

Research Article

Keywords: SARS-CoV-2, Spike protein, pH values, Conformational state

Posted Date: July 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-665823/v1

Molecular dynamics modeling of the SARS-CoV-2 spike protein at pH2 through pH11.5

Ziyuan Niu[1] · Peng Zhang[1] · Miriam Rafailovich[2] · Marcia Simon[3] · Meichen Song[1] · Yuefan Deng[1,4]

Departments of [1]Applied Mathematics and Statistics, [2]Materials Science and Chemical Engineering, [3]Oral Biology and Pathology, Stony Brook University, Stony Brook, New York 11794, United States. [4]Email: yuefan.deng@stonybrook.edu.

Abstract

The spike glycoprotein (S protein) of the SARS-CoV-2 that has be studied extensively *in vitro* is modeled by all-atom molecular dynamics for its conformational states at six pH values ranging from 2 to 11.5. The MD simulations up to 3.7 μs demonstrate interesting discoveries while confirming known facts. (1). At pH2, the protein's time averaged RMSD is 62.5% higher than that of pH7, as the control group, and the receptor binding domain (RBD) deviates from that of pH7 by 200%. (2). For pH4 through pH10.5, the S protein remains relatively stable evident by the invariance of the side chain H bond counts and RMSD from pH7, suggesting high tolerance of the S protein to a wide range of pH values other than the extreme acidic and basic conditions. (3). For pH2 to pH4, the structure of the S protein alters significantly, suggesting the existence of a critical pH value at which the S protein responds to acid sharply. (4). In the residue-based relative entropy analysis, we identify several RBM and RBD residue clusters with maximum deviations that cause the overall protein structure changes.

Keywords SARS-CoV-2 · Spike protein · pH values · Conformational state

Introduction

With the rapid spread and mutation of SARS-CoV-2, there is an urgent need to develop protective vaccines and targeted cleaners. The coronavirus is postulated to infect cells by attaching its outer membrane S protein [1] to the host cell targets such as ACE2, CD26, Ezrin, and cyclophilins [2]. Structurally altering the S protein may mitigate or even block its infectivity and other physicochemical properties. Surgical changes to the protonation and deprotonation states of each ionizable group to control the pH value [3] is one of such alterations. SARS-CoV-2 is known to spread through aerosol and the pH values of the aerosol droplets' environment are known to change the chemical form, distribution, and reactivity [4-6] of the protein. Performing such study, *e.g.*, immersing the virus, especially the S protein, in a solvent of given pH values, is dangerous and nebulous. Moreover, measuring the pH of a single aerosol droplet is challenging in the traditional or *in vitro* laboratories due to the lack of tools that can detect pH in an individual droplet environment [6, 7]. Our all-atom molecular dynamics *in silico* experiments, with their own unique challenges including excruciatingly long computations, allow us to examine the confirmational states of the S protein, at atomistic resolutions, at the full ranged pH values, for extended time. In strong acids and strong bases, our *in silico* experiments pinpoint individual residues that cause singularly large deviations to the structure of the S protein.

Related Work

S protein, an important determinant of virus virulence, tissue tropism and host range in coronaviruses envelope during infection [8, 9]. S1 and S2 are two sections of the S protein, with S1 being for host cell receptor binding and S2 being for membrane fusion [9, 10]. S protein on the coronavirus envelope binding to the receptors, mediate membrane fusion, virus entry, and syncytia formation, and these help elicit virus-neutralizing antibodies [11]. Coronavirus S proteins will be activated by receptor binding and/or low pH to protease cleavage between the S1 and S2 domains to permit conformational changes in S2, so that the mediate membrane fusion leading to virus entry and syncytia formation [11-14].

Therefore, control the conformational state of the S protein will control the spread of the coronaviruses.

Different pH conditions will affect coronaviruses S protein structure has been widely recognized in recent study. At pH5.5-8.5, the mouse hepatitis virus Type 4 (MHV4) coronavirus causes significant cell-cell fusion. Chloroquine and ammonium chloride, both endosomotropic weak bases, do not prevent the MHV4 infection. However, some selected variants from a neural cell line persistently infected with MHV4 are completely reliant on acid pH to fuse host cells and are substantially inhibited by endosomotropic weak bases [15]. At pH5 buffer, avian coronavirus infectious bronchitis virus (IBV) was still activated, and fusion was unaffected. Virions also have reversible evidence of conformational changes in their surface proteins, indicating the reversibility of the fusion reaction [16]. For human coronavirus 229E, an optimal stability of viral infectivity was observed at pH6 at temperatures of both 4 and 33°C. Indeed, viral infectivity was undetectable after exposure to pH4 or pH9 at 33°C. At 4°C in medium buffered at pH10, infectivity of the virus changes little [17]. The infectivity of the coronavirus (MHV-A59) is extremely sensitive to pH; stable at pH6 and 37°C but inactivated by a short treatment at pH8 at the same temperature. Extreme pH values, such as pH3 and pH9 or pH10, exacerbate virus's inactivation [18].

SARS-CoV-2 attaches to ACE2 as the host cell receptor [2]; this spike's RBD with its receptor binding motif (RBM) capable of attaching to ACE2. SARS-CoV-2 spike RBM can better insert into ACE2 hydrophobic pocket, due to variable ridge loop with a four-residue motif, novel interactions because of Leu472, unique hydrogen bond (H bond) between Lys353 from ACE2 and SARS-CoV-2 RBD main chain [19, 20]. Therefore, compared with other hCoVs, the high-affinity of the SARS-CoV-2 and ACE2 interactions may be an explanation for the greater infectivity of this virus. A more detailed understanding of the H bonds in different parts of the S protein and how they are affected by different pH values is one of the goals of our work.

The pH-induced retraction of RBDs through the spike adopting an all-down conformation can be described as a ''conformational masking'' energy barrier, the SARS-CoV-2 spike evasion from CR3022 neutralization to depend on the reduced affinity of CR3022 [21, 22]. SARS-CoV-2 spikes at serological pH (pH6) bind to ACE2 and CR3022 and at lower pH (pH4.5-5.5) they still bind to ACE2 but not CR3022 [21].

Our simulations not only focused on the trajectory of S protein at different pH values, but also analyzed RBD deviation and RBD-water H bonds separately; therefore, we have a more accurate understanding of RBD and RBM structure variations. We also identify the residues that are responsible for the overall structural deviation of the S protein at a given pH.

Our Main Contributions: We have performed MD simulations of S protein for six pH values in 2, 4, 5, 7, 10.5, 11.5 for up to 3.7 μs. With the current rate of simulations, we reached a record simulated time to obtain the conformational states of the S protein at these six pH values. Methodically controlling the pH values during the *in silico* experiments by protonation and deprotonation state of the amino acid side chain residues, we corroborated the simulated conformational states with the properties obtained in *in vitro* bench experiments. We examined the detailed conformational state changes at different pH values of the entire S protein as well as the individual residues of interest. We separately analyzed the trajectory of RBD under different pH, which gave μs a new understanding of the infectiousness of the virus. Additionally, we identified the residues that are responsible for the S protein structural variations at five of six pH values. Our study may help the development of drugs and vaccines, and even the cleaners, for prevention of the SARS-CoV-2 or other similar viruses, at lower risks.

State Generation and Analysis

The *in silico* Experiment Setup

The open-source Gromacs [23] was adapted to perform the MD simulation on the AiMOS supercomputer configured with IBM POWER9 processors and NVIDIA Volta V100 GPUs [24]. The S protein's structure data is taken from the protein data bank (6VXX.pdb) and the force field for the S protein and SPC/E water molecule is from the CHARMM27. The S protein of size $12\times13\times16$ nm^3 is placed in the explicit solvent, TIP3PBOX water model of size $20\times20\times20$ nm^3. Periodic boundary condition (PBC) is applied to all three Carstensen dimensions of the water box. Every simulation including the reference experiment at the neutral pH7 is controlled at the normal human body temperature of 37°C. The energy is minimized by the steepest descent minimization, and simulations are controlled by the canonical (NVT) and Parrinello-Rahman pressure coupling (NPT) with 2 fs time step. Therefore, the simulations are controlled through the equilibrium of temperature (with NVT) and density (with NPT). On the shared AiMOS supercomputer [24] with 268 nodes and a theoretical peak speed of 11,032 Tflop/s, we perform our experiments on a sub-partition of 4 nodes with which we achieved a running speed of approximately 65 ns/day. A complete each experiment for a given pH takes two months, plus queuing time. We deliberately terminated the runs for pH4 and pH10.5 earlier for their quicker approach to equilibrium and less sensitive for data analysis. More *in silico* experiment details are included in Table 1.

Table 1 Parameter details for our *in silico* experiments of varying pH values

	pH2	pH4	pH5	pH7	pH10.5	pH11.5
Simulation Box (nm^3)	$20 \times 20 \times 20$					
Protein Atoms	45,432	45,300	45,192	45,146	45,090	45,027
Waters Elements	743,397	743,862	744,225	789,480	744,156	743,979
Protein Charges	+261	+129	+21	-15	-81	-144
Aspartate	Protonated	Protonated	Not change	Not change	Not change	Not change
Glutamate	Protonated	Protonated	Not change	Not change	Not change	Not change
Histidine	Protonated	Protonated	Protonated	Not change	Not change	Not change
Lysine	Not change	Not change	Not change	Not change	Deprotonated	Deprotonated
Density (g/cc)	1.016	1.015	1.015	1.015	1.016	1.016
Simulated Time (μs)	3.7	1.7	3.7	3.6	1.9	3.7
Simulating Time (Days)	~62	~29	~62	~60	~32	~62

Protonation State Control

The desired pH values are achieved by protonation and deprotonation states of the amino acid side chain residues in the S protein following the protonation fraction curve and the Henderson–Hasselbalch equation

$$\mathrm{pH} = pKa + \log\frac{[A^-]}{[HA^-]}$$

This method was used to verify another method proposed by Mongan et al. [25] to simulate the constant pH value.

All our simulations start from the 6VXX structure while pH2 and pH4 with aspartates, glutamates, and histidine titrating, pH5 with histidine titrating, while pH10.5 and pH11.5 with lysine titrating.

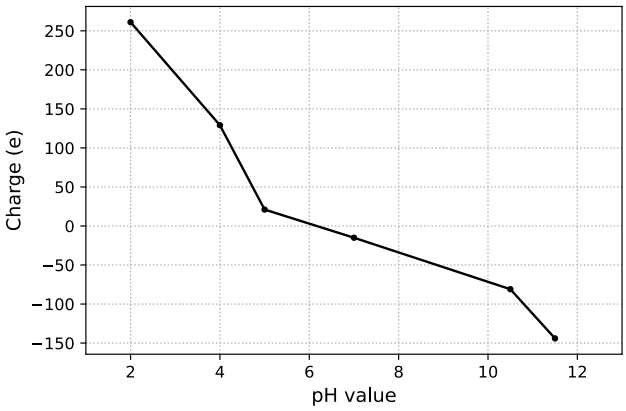


Fig. 1 Total charges of the S protein for pH2 through pH11.5, the range of our *in silico* experiments

Charge neutrality, along with the protonation state accuracy [26], is achieved by addition of the salt ions to the solvent. The total charge of S protein after protonation and deprotonation is shown in Fig.1. For achieving the acidic solvent where the protein accepts the H^+ leading to a net positive charge, Cl^- is added and, conversely, for achieving the basic solvent where the protein donates the H^+ resulting in net negative charge, Na^+ is added.

State Measurements

As the core raw data, $\vec{x}_i(t)$, the time-varying 3D coordinates, or time series, for all particles, are collected during the MD simulations, we analyze them, in space and in time, to understand the structures and dynamics of the S protein. Our analysis involves the following measurements categorized in two groups. For each pH value, the first group is expressed as function of time and second requires statistics in time. In the first group, we measure (1) RMSD of the backbones; (2) the mass deviation for receptor binding domains; (3) the RMSF of individual residues; (4) the numbers of H bonds connecting the protein with water (P-W), protein's main chain with main chain (MC-MC), the side chain with side chain (SC-SC), and RBD to water (RBD-W); and (5) the RBD deviation. In the second group, we perform time statistics to compare the protein's overall structures at various pH values by introducing the relative entropy [27] with which we also identify the outlying residues that dominate changes to the overall protein structure. As a common practice, we represent different pH values in colors (Fig. 2) and these colors will be adopted without further mentioning when pH impact is expressed graphically. Figs. 3 and 4 illustrate the relationships of these measurements and the data analysis workflow, respectively.

Fig. 2 The coloring scheme of the pH values to be used throughout the manuscript

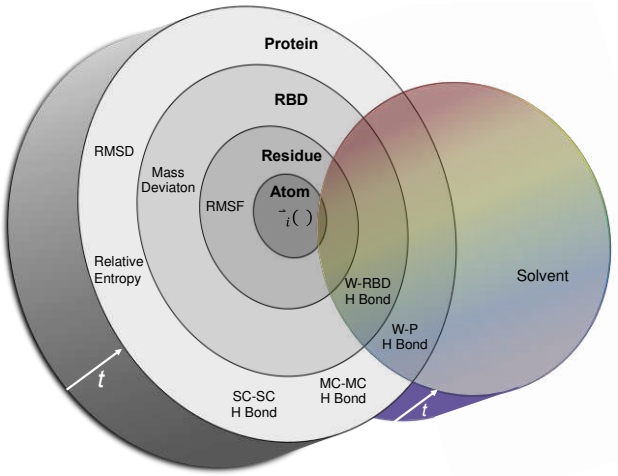


Fig. 3 A pictorial summary of the measurements and a graphical representation of these measurements in the relevant domains: atoms, residues, RBD, and the whole protein, as well as the solvent, etc.

The RMSDs in Various pH Solvents

Our RMSD is calculated by averaging over the backbone atoms in reference to their starting structure. In our simulations, the atoms' coordinates were recorded every 0.1 ns and a moving average of window size 2 ns were used. For all six pH values, we present the time series of the RMSD in Fig. 4 (R.1) and its time average, during the last 1.5 μs while the protein attained equilibrium, in Fig. 4 (R.2). In both figures, we note the following: (1) For pH2 and pH11.5, the RMSD is significantly higher than that of pH7 by approximately 60% after 2 μs time mark, signaling dramatic protein structure variation for these two pH values, as highly expected; (2) For pH4, pH5 and pH10.5, we observe negligible differences from pH7, a result corroborating *in vitro* results of other coronaviruses at different pH we mentioned above and will compare later [15-18, 28, 29].

H bonds Counts

We also calculated the numbers of H bonds of various setting and their time averages, shown in two sets of figures. Fig. 4 (H.PW.1) and (H.PW.2) shows the H bonds for the S protein to water while Fig. 4 (H.MS.1) and (H.MS.2) for MC-MC and SC-SC. From these figures, we note the following: (1) At pH2, approximately 800, or 14%, protein-water H bonds broke, as expected for strong acid; (2) At pH4, approximately 400, or 7%, protein-water H bonds broke, as expected for weaker acid; (3) At pH5 and higher, we see no visible change of protein-water H bonds; (4) For any pH values, we see no visible change of MC-MC H bonds while, for pH2 and pH11.5 we see major and mild change of SC-SC H bonds, respectively.

The Measurements in RBDs

We performed detailed statistical analysis of RBD deviations and RBD-W H bonds provide a comparison of the movement of the residues in RBD (ARG319-PHE541), and even RBM (SER438-GLN506) [30, 31]. These analyses intend to address the following widely known observations. S protein contains a variety of defensive mechanisms. S protein is decorated extensively with glycans that aid in immune evasion by shielding potential antigens [32, 33]. S protein uses a conformational masking strategy, wherein it predominantly adopts a closed conformation that retract the RBDs to escape

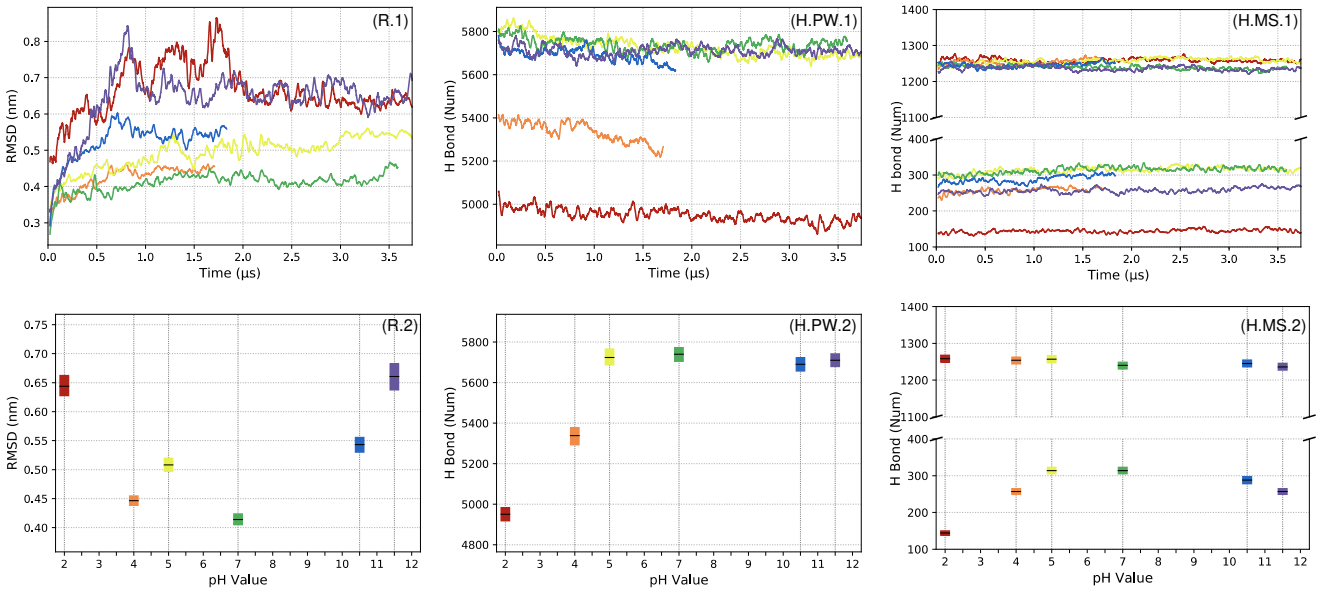


Figure 4. The RMSDs (R.1) evolving as a function of time and time-averaged (R.2) values; the number of P-W H bonds evolving with time (H.PW.1) and time-averaged values (H.PW.2), and the numbers of MC-MC (upper portion) and SC-SC (lower potion) H bonds evolving with time (H.MS.1) and time-averaged values (H.MS.2), all at pH2-11.5

immune surveillance mechanisms, and then the RBD needs to open to intersect with ACE2 [31].

To assess the ACE2 binding, we measure the RBD mass deviation by the distance between the center of mass of the RBD and its position in the initial frame closed (or down) state

$$D_t = \sqrt{(x_t - x_0)^2 + (y_t - y_0)^2 + (z_t - z_0)^2}$$

where D_t measures the Euclidian distance of RBD's center of mass (x_t, y_t, z_t) at frame t from the initial frame (x_0, y_0, z_0).

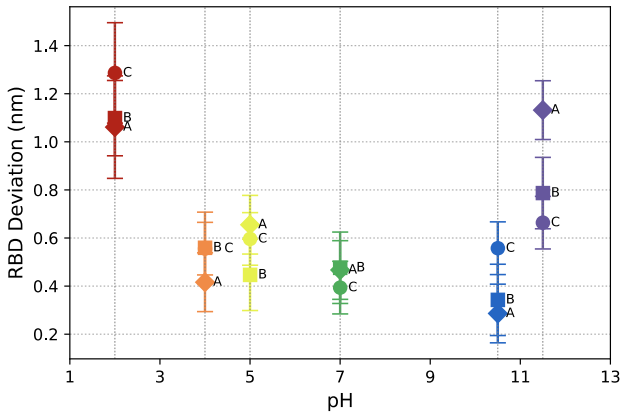


Fig. 5 RBD mass deviations for three chains at pH2-11.5

Mass deviations: Fig. 5 shows RBD mass deviations for all three chains at our tested pH values. At all three chains of pH2 and chain A of pH11.5, the RBD mass deviations nearly doubles those of the middle range of pH values and the mass deviations significantly deviate among the three chains, even at the same pH values.

The numbers of **H bonds in RBD**: Here again we examine the numbers of H bonds. This time, we analyze the RBD-W H bonds. As shown in Fig. 6, we note significant differences in H bond counts for pH2 (~360) with the rest that are similar (all ~410), leading µs to suggest an extreme acidic solvent can break meaningful number of the RBD-W H bonds. Also, for chain C, the numbers of H bonds vary rather significantly across the pH values, we counted 20-40 H bonds broken relative to the cases of pH5 and 7.

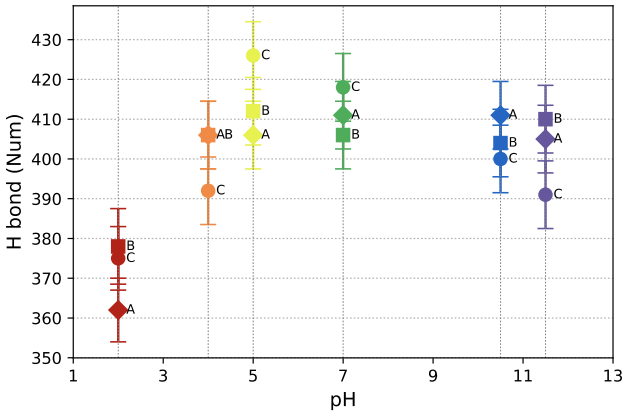
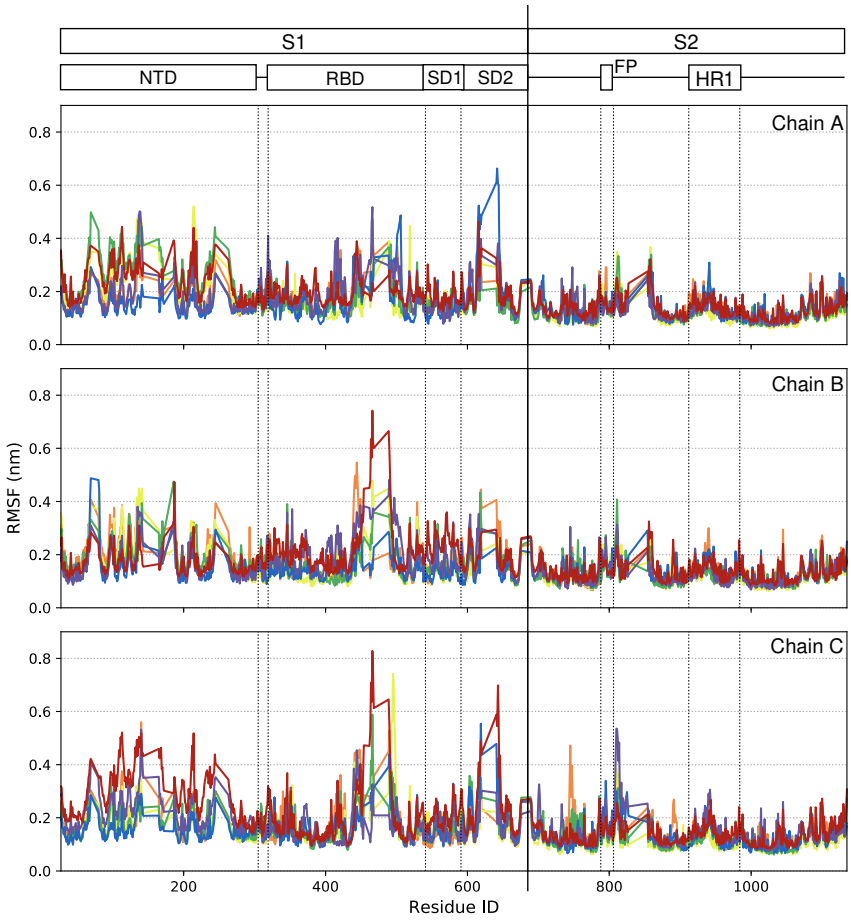


Fig. 6 The numbers of RBD-W H bonds at pH2-11.5

Fig. 7 RMSF at pH2-11.5 in three chains



Relative Entropy for Structure Divergence

For individual residues, we use root mean square fluctuation (RMSF) to measure the time-averaged fluctuations from their initial states. Obviously, while most of the residues stay intact, a few outlier members deviate and cause significant structural changes of the entire protein, in each pH solvent. Fig. 7 shows the RMSF for every residue of every chain for our tested pH values marked by the colors of the curves, as usual. The curves in this figure are central to understanding the impacts of the pH solvents on the proteins. We observe higher fluctuations of clusters of residues for each chain, for example, at pH2, the RMSF's of PHE464-PHE490 in chain B and LYS463-PHE490 in chain C, located in the RBM of the S1 domain, and ASN641-GLN644 in chain C, and, at pH11.5, of ASP138, ARG466 chain A and PHE140, SER810 LYS811 of chain C, are higher than 0.5 nm. For the mild pH5, except the RMSF's of GLN493-GLN498 in RBM of chain C higher than 0.5nm, no residues fluctuate.

For a better understanding of the structural changes of the S protein at different pH values, we introduced the idea of relative entropy [27]. We consider the entropy are calculated by the Kullback–Leibler divergence

$$D_{KL}(x||7) = \sum_i r_i(x) \log\left(\frac{r_i(x)}{r_i(7)}\right)$$

where $r_i(x)$ is the RMSF for residue i in pH values x. Therefore, to compare the RMSF with that of pH7, we use the relative entropy, a measure of overall conformational state of the entire protein at a given pH values relative to that of the protein at the pH7.

In examining individual residue's contribution to the entropy, we found reasonable threshold of 0.07 to identify the outliers and, except for chain C at pH2, all chains have around 20 outliers that dominate the conformational changes for each pH solvent. We identify the outlier residues, whose IDs are tabulating in Tables 2-4, that cause the overall protein structure changes. We also pinpoint them structurally in the protein (Fig. 8) and zoom in to view the outliers (Figs. 9 and 10).

Fig. 11 summarizes the central discovery of this study that the relative entropy, or structural, variations as an impact the pH solvent. One may infer from here functions of the protein induced by the pH values. Here is a list of observations. (1) The protein varies the most in strong acid (relative entropy is 27 for pH2) and base (relative entropy is 12 for pH11.5) solvents and stays rather stable in solvent of pH4 through pH10 with relative entropy around 0. This result is consistent with *in vitro* experiments [29]. Of the extreme solvents (pH2 and pH11.5), chain C of the protein varies the most, signaling loss of the infectivity as chain C is in the S1 with RBD and NTD that direct bind to ACE2 [9, 34]. (2) Comparing the strong acidic solvent of pH2 and strong base solvent of pH11.5, we note that chains A and B relative entropies of 6.2 and 10.3 for pH2 and 0.3 and 5.1 for pH11.5. This suggest the S protein is more affected by a strong acid solvent than strong base solvent. (3) When examining the relative entropy outliers with RBD deviations, we note expected consistency. In pH2, the RBD deviation of chain C is larger than that of chain A and B. At the same time, chain C has the largest number of entropy outliers concentrated in RBD and NTD. At the same time, our RMSF data is measured between 2.2-2.7μs (pH2, 5, 7, 11.5) and 1.2-1.7 μs (pH4 and 10), but chain C's RBD deviation after 2.7μs still maintains a larger transformation compared to the other two chains. This demonstrates the power of the entropy outliers in spotting changes in protein structure.

Table 2 The residue IDs of the entropy outliers in chain A

chain A	S1 Domain Residue ID | S2 Domain Residue ID
pH2	PRO384-LYS386 ASP442-LYS444 HIS519 ASN544 GLY545 VAL615 ASN616 VAL620 | GLN784
pH4	TYR421 SER443 GLY447 PHE464 SRG466 ASP467 PRO499 ASN501 VAL642-THR645 ALA647 |ASP796 HIS1083
pH5	VAL213-PRO217 ARG466 HIS519
pH10.5	ASN439 LEU441-SER443 PHE497 ASN501 VAL503-PRO507 ASP614-VAL620 ASN641-TR645 ALA647 | TYR707
pH11.5	GLY413 GLN414 PRO463-ARG466 HIS519 THR604 VAL620 VAL642-GLN644 | LYS733 TYR741 LEU753

Table 3 The residue IDs of the entropy outliers in chain B

chain B	S1 Domain Residue ID | S2 Domain Residue ID
pH2	LYS187 ARG328 VAL367 ASP427 ASP442 TYR451 ARG454 LYS462-ILE468 TYR489-PRO491 TYR495 ASN544 GLY545 GLU554 SER555 LYS557 PHE559 ALA570-VAL576 PRO579 GLN580 | GLN787 ASN856 VAL860 ALA890 ARG983
pH4	ARG102 SER112-GLN115 GLN239 ASN440-LYS444 GLY447-TYR449 SER530 THY531 | SER939 SER940
pH5	ASN81 SER112-GLN115 GLN134-ASN137 GRO139 LYS417-ILE418 ARG466 SER530
pH10.5	HIS 69 ASP80 ASN81 ARG408 | LYS854 ARG1014
pH11.5	PHE135 ASN137 TPR353 ILE402-GLU406 ARG408 GLN409 THR415-ILE418 ASP420 ASP442 SER443 ASN448 TYR449 TYR451 LYS462 SER494-GLN498 GLY504 TYR505 PRO507 GLN580 THR604 | ASP745

Table 4 The residue IDs of the entropy outliers in chain C

chain C	S1 Domain Residue ID | S2 Domain Residue ID
pH2	PHE43 ALA67-HIS69 ASN81-ASN87 GLY89 PHE92 ILE101 GLY103-VAL143 ASN165-172 ASN188 ARG190 PHE192 PHE194-ASN196 GLR199-ILE203 SER205 HIS207 ASN211-ASP215 LEU226 VAL227 LEU229-ILE233 ILE235-LEU244 ALA263-TYR266 ARG319 ASN343 ARG408 TYR449 TYR453 ARG454 LYS462-ILE468 TYR489-PRO491 GLN580 VAL615-CYS617 VAL620 ASN641-ARG646 | ARG983 ARG1091 ASN1135
pH4	PRO139-GLY142 GLY413 THR415 GLY416 ASN422 ASP427 LEU441-SER443 GLY447-TYR449 TYR451 TYR489-PRO491 GLN644-ARG646 | GLY744 ASP745 THR747 PHE855 ALA890 GLY891
pH5	GRO139 TRP353 ASN439-LEU441 LYS444 PHE490 PRO491 GLN493-PRO499 HIS519
pH10.5	ARG328 TYR489 PHE490 GLN580 ASN616 THE618-VAL620 ASN641 PHE643 THR645-ALA647 | ASN824-VAL826 PHE855
pH11.5	ASN122 ALA123 GRO139-GLY142 LEU216 TYR351 ASN439-LEU441 LYS444 TYR449 | GLY700 ALA701 LYS733 PRO809-ASP820 LEU822-ASN824 VAL826 THR827 ALA942 ARG1014 ARG1019 GLN1036 LYS1045

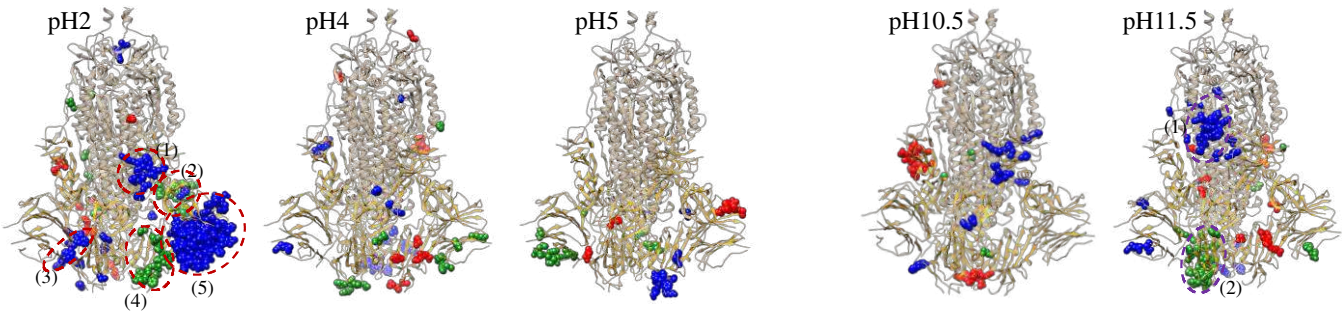


Fig. 8 Identifications of residue outliers causing protein structure divergences and these outliers for experiments with pH2 and pH11.5 to be zoomed further. (In this and the zoomed pictures, the colors are not related to the pH color scheme.)

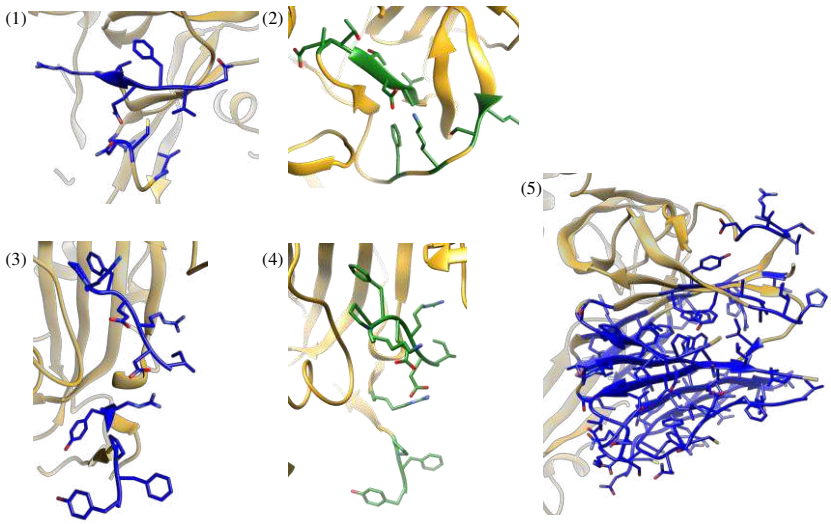


Fig. 9 The residue outlier clusters zoomed structure for pH2, (1) ASN641-ARG646 outlier cluster in chain C, (2) SD2 outlier cluster in chain B, (3) RBD outliers cluster in chain C, (4) RBD outliers cluster in chain B, (5) NTD outliers cluster in chain C

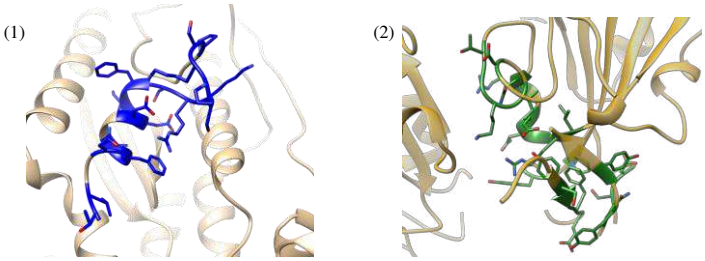


Fig. 10 The residue outlier clusters zoomed structure for pH11.5. (1) S2 domain cluster in chain C. (2) RBD outliers cluster in chain B

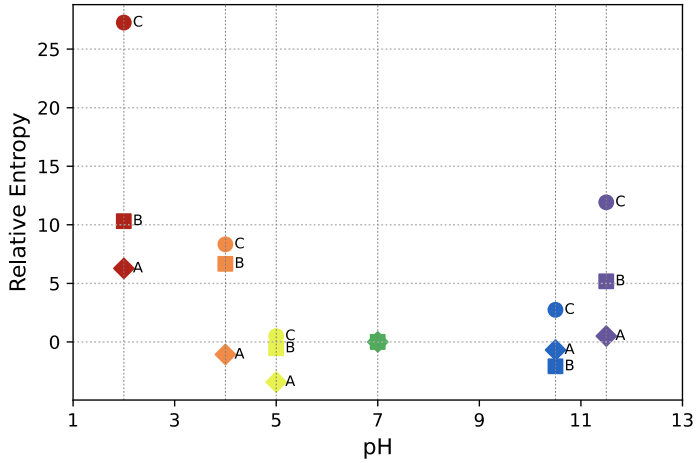


Fig. 11 Relative entropies for pH2-11.5 and A, B, C in the figure label the corresponding chains

The Correlations of the Measurements

For correlating the measurements performed on different pH values, we made scatter plots (Fig. 12) for RMSD, P-W H bonds, MC-MC H bonds, and SC-SC H bonds. In the figure, each dot represents a pair of correlating measures for a given pH value marked in a color scheme in Fig. 2. Moreover, the population correlation coefficient $\mathrm{corr}(X,Y)$ between two sets X and Y is defined as

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

where μ_X and μ_Y are expected values, σ_X and σ_Y are standard deviations. The E is the expected value operator, and the cov is covariance. Since the correlation coefficient is symmetric,

$$\mathrm{corr}(X,Y) = \mathrm{corr}(Y,X)$$

we use the upper left blocks to represent the scatter plot between the two analysis methods at different pH, and the symmetrical position in the lower right blocks to represent the correlation coefficients. Apparently, diagonal blocks, showing the self-correlations of RMSD, P-W H bonds, MC-MC H bonds, and SC-SC H bonds, show four colorful straight lines indicating the dependence of these measures on the pH values.

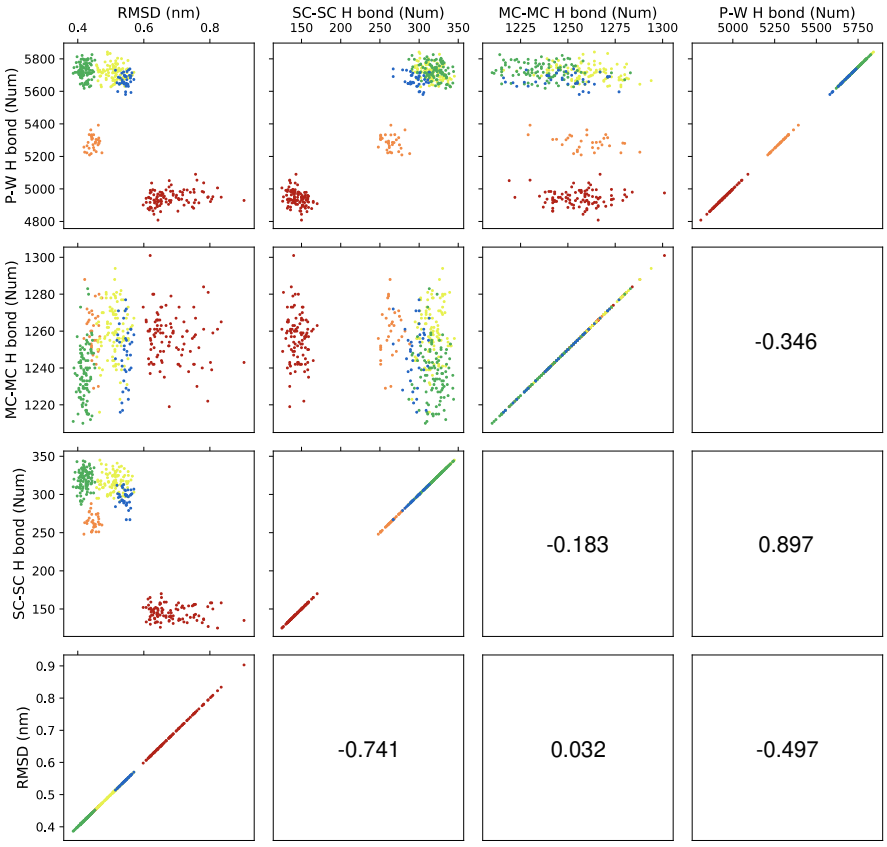


Fig. 12 The scatter plots, and the correlation coefficients, between pairs of measurement of pH2-11.5

The patterns, or the clustering, and the color positioning are quite informative for the correlations of the measurements. For example, the SC-SC H bonds positively correlate with the P-W H bonds (correlation coefficient 0.897) with three distinctive clusters according to the pH values. Another example is the negative correlation of the RMSD with the SC-SC H bonds (correlation coefficient -0.741), as expected. Other pairs appear to be highly uncorrelated.

Corroborations with *in vitro* Experiments

The SARS-CoV titer was unaffected by moderate pH changes ranging from 5 to 9. However, the virus fully inactivated after being exposed to strongly acidic (pH1-3) and basic conditions (pH12-14), These indicate that SARS-CoV infectivity is sensitive to pH extremes [28]. For the high homology between SARS-CoV and SARS-CoV-2, at room temperature, the SARS-CoV-2 virus is highly stable over a wide pH range (pH 3–10) [29]. At pH4-10.5, the S protein is in a relatively stable state; however, for extreme pH (pH2 and 11.5) in our *in silico* experiments, the overall protein structures alter substantially, corroborating the *in vitro* reality. It is not hard to project that under longer exposure, S protein will denature at these extreme conditions.

The density of all RBD domains in the SARS-CoV-2 S protein was highly resolved, revealing multiple RBD orientations in the spike at pH5.5. The S protein conformational heterogeneity to be reduced between pH5.5 and pH4.5, and then to remain unchanged as pH was reduced further (from pH4.5 to pH4) [21]. On the other hand, weak acid pH (pH6-6.5) increases SARS-CoV-2 viral load and infection with increased expression of ACE2 [35]. These data imply that, in weak acid, the S protein is more likely to bind with ACE2, which is consistent with our results (the RBD mass deviation of pH5 is slightly higher than that of pH7).

Conclusions and Discussions

We demonstrated the first microsecond MD studies of the S protein's conformational state changes induced by immersing the protein to pH solvents with six different pH values ranging in pH2 through pH11.5. We discovered that the S-protein of the SARS-CoV-2 is structurally stable at pH4-10.5, consistent with the observations that the SARS-CoV-2 is highly stable across a wide range of pH values (pH3-10) at room temperature [29], and it is extreme unstable, denature quickly, at pH2 and pH11.5, evident by the atypically high RMSDs and the relative entropies.

All-atom simulations are computationally expensive. Each μs-scale *in silico* experiment for a set of parameters at a given pH value costs two months of computing time on a supercomputer. The development of a coarse-grained model can reduce the computation loads while preserving sufficient

simulation accuracy. Moreover, an efficient and intelligent simulation algorithm potentiated by the latest machine learning efforts can further accelerate the simulations [36, 37]. Together, these efforts will help enable the more desirable and more realistic μs -scale *in silico* experiments and more discoveries of the structures of new viruses, more quickly and more accurately.

Acknowledgements The project is supported by the SUNY-IBM Consortium Award, PI: Y. Deng. All simulations were conducted on the AiMOS supercomputer at Rensselaer Polytechnic Institute and the WSC Cluster at the IBM T. J. Watson Research Center through an IBM Faculty Award FP0002468 (PIs: Y. Deng and P. Zhang). Stimulating and useful discussions with the following individuals are appreciated: K. Hasegawa, J. Myers, and K. Swayze.

Funding The project is sponsored by Stony Brook University's OVPR & IEDM COVID-19 Seed Grant, PI: P. Zhang. Co-PIs: Y. Deng, M. Rafailovich, and M. Simon.

Availability of data and material All original data are available upon request.

Code availability The data generation is done through the open-source Gromacs while data analysis is done by our own code that is available upon request.

Authors' Contributions Ziyuan Niu: Data collection and analysis, Conceptualization, Writing- Original draft. Peng Zhang: Resources, Supervision. Marcia Simon: Resources. Miriam Rafailovich: Resources. Meichen Song: Data analysis. Yuefan Deng: Data analysis, Resources, Supervision, Writing- Reviewing and Editing.

Declarations

Ethics approval The manuscript is prepared in compliance with the Ethics in Publishing Policy as described in the Guide for Authors.

Consent to participate The manuscript is approved by all authors for publication.

Consent for publication The consent for publication was obtained from all participants.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Xiong X, Tortorici MA, Snijder J, Yoshioka C, et al. (2018). Glycan shield and fusion activation of a deltacoronavirus spike glycoprotein fine-tuned for enteric infections. J Virol 92(4): e01628-01617. 10.1128/JVI.01628-17
2. Song W, Gui M, Wang X, Xiang Y (2018). Cryo-em structure of the sars coronavirus spike glycoprotein in complex with its host cell receptor ace2. PLoS Pathog 14(8): e1007236. 10.1371/journal.ppat.1007236
3. Donnini S, Tegeler F, Groenhof G, Grubmuller H (2011). Constant ph molecular dynamics in explicit solvent with lambda-dynamics. J Chem Theory Comput 7(6): 1962-1978. 10.1021/ct200061r
4. Schwarzenbach RP, Gschwend PM, Imboden DM (2016). Environmental organic chemistry. John Wiley & Sons.
5. Weber RJ, Guo HY, Russell AG, Nenes A (2016). High aerosol acidity despite declining atmospheric sulfate concentrations over the past 15 years. Nature Geoscience 9(4): 282-+. 10.1038/Ngeo2665
6. Wei H, Vejerano EP, Leng W, Huang Q, et al. (2018). Aerosol microdroplets exhibit a stable ph gradient. Proc Natl Acad Sci U S A 115(28): 7272-7277. 10.1073/pnas.1720488115
7. Ault AP, Axson JL (2017). Atmospheric aerosol chemistry: Spectroscopic and microscopic advances. Anal Chem 89(1): 430-452. 10.1021/acs.analchem.6b04670
8. Li F (2016). Structure, function, and evolution of coronavirus spike proteins. Annu Rev Virol 3(1): 237-261. 10.1146/annurev-virology-110615-042301
9. He J, Tao H, Yan Y, Huang SY, et al. (2020). Molecular mechanism of evolution and human infection with sars-cov-2. Viruses 12(4): 428. 10.3390/v12040428
10. Gui M, Song W, Zhou H, Xu J, et al. (2017). Cryo-electron microscopy structures of the sars-cov spike glycoprotein reveal a prerequisite conformational state for receptor binding. Cell Res 27(1): 119-129. 10.1038/cr.2016.152
11. Qian Z, Dominguez SR, Holmes KV (2013). Role of the spike glycoprotein of human middle east respiratory syndrome coronavirus (mers-cov) in virus entry and syncytia formation. PLoS One 8(10): e76469. 10.1371/journal.pone.0076469
12. Graham RL, Baric RS (2010). Recombination, reservoirs, and the modular spike: Mechanisms of coronavirus cross-species transmission. J Virol 84(7): 3134-3146. 10.1128/JVI.01394-09
13. White JM, Delos SE, Brecher M, Schornberg K (2008). Structures and mechanisms of viral membrane fusion proteins: Multiple variations on a common theme. Crit Rev Biochem Mol Biol 43(3): 189-219. 10.1080/10409230802058320
14. Frana MF, Behnke JN, Sturman LS, Holmes KV (1985). Proteolytic cleavage of the e2 glycoprotein of murine coronavirus: Host-dependent differences in proteolytic cleavage and cell fusion. J Virol 56(3): 912-920. 10.1128/JVI.56.3.912-920.1985
15. Gallagher TM, Escarmis C, Buchmeier MJ (1991). Alteration of the ph dependence of coronavirus-induced cell fusion: Effect of mutations in the spike glycoprotein. J Virol 65(4): 1916-1928. 10.1128/JVI.65.4.1916-1928.1991
16. Chu VC, McElroy LJ, Chu V, Bauman BE, et al. (2006). The avian coronavirus infectious bronchitis virus undergoes direct low-ph-dependent fusion activation during entry into host cells. J Virol 80(7): 3180-3188. 10.1128/JVI.80.7.3180-3188.2006
17. Lamarre A, Talbot PJ (1989). Effect of ph and temperature on the infectivity of human coronavirus 229e. Can J Microbiol 35(10): 972-974. 10.1139/m89-160
18. Sturman LS, Ricard CS, Holmes KV (1990). Conformational change of the coronavirus peplomer glycoprotein at ph 8.0 and 37 degrees c correlates with virus aggregation and virus-induced cell fusion. J Virol 64(6): 3042-3050. 10.1128/JVI.64.6.3042-3050.1990
19. Shang J, Ye G, Shi K, Wan Y, et al. (2020). Structural basis of receptor recognition by sars-cov-2. Nature 581(7807): 221-224. 10.1038/s41586-020-2179-y
20. Costa LB, Perez LG, Palmeira VA, Macedo ECT, et al. (2020). Insights on sars-cov-2 molecular interactions with the renin-angiotensin system. Front Cell Dev Biol 8: 559841. 10.3389/fcell.2020.559841
21. Zhou T, Tsybovsky Y, Gorman J, Rapp M, et al. (2020). Cryo-em structures of sars-cov-2 spike without and with ace2 reveal a ph-dependent switch to mediate endosomal positioning of receptor-binding domains. Cell Host Microbe 28(6): 867-879 e865. 10.1016/j.chom.2020.11.004
22. Wu NC, Yuan M, Bangaru S, Huang D, et al. (2020). A natural mutation between sars-cov-2 and sars-cov determines neutralization by a cross-reactive antibody. PLoS Pathog 16(12): e1009089. 10.1371/journal.ppat.1009089
23. Abraham MJ, Murtola T, Schulz R, Páll S, et al. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1: 19-25.
24. Hanson WA (2019). The coral supercomputer systems. IBM Journal of Research and Development 64(3/4): 1: 1-1: 10.
25. Mongan J, Case DA, McCammon JA (2004). Constant ph molecular dynamics in generalized born implicit solvent. J Comput Chem 25(16): 2038-2048. 10.1002/jcc.20139
26. Chen W, Shen JK (2014). Effects of system net charge and electrostatic truncation on all-atom constant ph molecular dynamics. J Comput Chem 35(27): 1986-1996. 10.1002/jcc.23713
27. Kullback S, Leibler R (1951). On information and sufficiency, annals maths. Statist 22: 79-86.
28. Darnell ME, Subbarao K, Feinstone SM, Taylor DR (2004). Inactivation of the coronavirus that induces severe acute respiratory syndrome, sars-cov. J Virol Methods 121(1): 85-91. 10.1016/j.jviromet.2004.06.006
29. Chin A, Chu J, Perera M, Hui K, et al. (2020). Stability of sars-cov-2 in different environmental conditions. MedRxiv.
30. Lan J, Ge J, Yu J, Shan S, et al. (2020). Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. Nature 581(7807): 215-220. 10.1038/s41586-020-2180-5

31. Zimmerman MI, Porter JR, Ward MD, Singh S, et al. (2021). Sars-cov-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. Nat Chem: 1-9. 10.1038/s41557-021-00707-0
32. Watanabe Y, Berndsen ZT, Raghwani J, Seabright GE, et al. (2020). Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. Nat Commun 11(1): 2688. 10.1038/s41467-020-16567-0
33. Casalino L, Gaieb Z, Goldsmith JA, Hjorth CK, et al. (2020). Beyond shielding: The roles of glycans in the sars-cov-2 spike protein. ACS Cent Sci 6(10): 1722-1734. 10.1021/acscentsci.0c01056
34. Zimmerman M, Porter J, Ward M, Singh S, et al. (2020). Sars-cov-2 simulations go exascale to capture spike opening and reveal cryptic pockets across the proteome. Biorxiv..[google scholar].
35. Jimenez L, Codo AC, Sampaio VS, Oliveira AE, et al. (2020). The influence of ph on sars-cov-2 infection and covid-19 severity. MedRxiv.
36. Han C, Zhang P, Bluestein D, Cong G, et al. (2021). Artificial intelligence for accelerating time integrations in multiscale modeling. Journal of Computational Physics 427: 110053.
37. Zhu Y, Zhang P, Han C, Cong G, et al. (2021). Enabling ai-accelerated multiscale modeling of thrombogenesis at millisecond and molecular resolutions on supercomputers. International Conference on High Performance Computing, Springer.