nature chemistry

ARTICLES

https://doi.org/10.1038/s41557-021-00758-3

Check for updates

A glycan gate controls opening of the SARS-CoV-2 spike protein

Terra Sztain[1,8], Surl-Hee Ahn[1,8], Anthony T. Bogetti[2], Lorenzo Casalino[1], Jory A. Goldsmith[3], Evan Seitz[4], Ryan S. McCool[3], Fiona L. Kearns[1], Francisco Acosta-Reyes[5], Suvrajit Maji[5], Ghoncheh Mashayekhi[6], J. Andrew McCammon[1,7], Abbas Ourmazd[6], Joachim Frank[4,5], Jason S. McLellan[3], Lillian T. Chong[2]✉ and Rommie E. Amaro[1]✉

SARS-CoV-2 infection is controlled by the opening of the spike protein receptor binding domain (RBD), which transitions from a glycan-shielded 'down' to an exposed 'up' state to bind the human angiotensin-converting enzyme 2 receptor and infect cells. While snapshots of the 'up' and 'down' states have been obtained by cryo-electron microscopy and cryo-electron tomagraphy, details of the RBD-opening transition evade experimental characterization. Here over 130 μs of weighted ensemble simulations of the fully glycosylated spike ectodomain allow us to characterize more than 300 continuous, kinetically unbiased RBD-opening pathways. Together with ManifoldEM analysis of cryo-electron microscopy data and biolayer interferometry experiments, we reveal a gating role for the *N*-glycan at position N343, which facilitates RBD opening. Residues D405, R408 and D427 also participate. The atomic-level characterization of the glycosylated spike activation mechanism provided herein represents a landmark study for ensemble pathway simulations and offers a foundation for understanding the fundamental mechanisms of SARS-CoV-2 viral entry and infection.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an enveloped RNA virus and the causative agent of coronavirus disease 2019 (COVID-19), a disease that has caused substantial morbidity and mortality worldwide[1,2]. The main infection machinery of the virus, the spike protein that sits on the outside of the virus, is the first point of contact that the virion makes with the host cell, and is a major viral antigen[3]. A substantial number of cryo-electron microscopy (cryo-EM) structures of the spike protein have been recently reported, collectively informing on structural states of the spike protein. The vast majority of resolved structures fall into either 'down' or 'up' states, as defined by the position of the receptor binding domain (RBD), which modulates interaction with the angiotensin-converting enzyme 2 (ACE2) receptor for cell entry[4–6].

The RBDs must transition from a 'down' to an 'up' state for the receptor binding motif (RBM) to be accessible for ACE2 binding (Fig. 1), and therefore the activation mechanism is essential for cell entry. Lu et al.[7] used single-molecule fluorescence (Förster) resonance energy transfer (smFRET) imaging to characterize spike dynamics in real time. Their work showed that the spike dynamically visits four distinct conformational states, the populations of which are modulated by the presence of the human ACE2 receptor and antibodies. However, smFRET, as well as conventional structural biology techniques, are unable to inform on the atomic-level mechanisms underpinning such dynamical transitions. Recently, all-atom molecular dynamics (MD) simulations of the spike protein, with experimentally accurate glycosylation together with corroborating experiments, indicated the extensive shielding by spike glycans, as well as a mechanical role for glycans at positions N165 and N234 in supporting the RBD in the 'open' conformation[8]. Conventional MD simulations as performed in Casalino et al.[8] also revealed microsecond-timescale dynamics to better characterize the spike dynamics but were limited to sampling configurations that were similar in energy to the cryo-EM structures. Several enhanced sampling MD simulations have been performed to study this pathway; however, these simulations lacked glycosylation for the spike protein[9] or involved the addition of an external force[10] or did not provide mechanistic detail[11].

In this study, we characterized the spike RBD-opening pathway for the fully glycosylated SARS-CoV-2 spike protein to gain a detailed understanding of the activation mechanism. We used the weighted ensemble (WE) path-sampling strategy[12,13] (Supplementary Fig. 1) to enable the simulation of atomistic pathways for the spike-opening process. As a path-sampling strategy, WE focuses computing power on the functional transitions between stable states rather than the stable states themselves[14]. This is achieved by running multiple trajectories in parallel and periodically replicating trajectories that have transitioned from previously visited to newly visited regions of configurational space[15], thus minimizing the time spent waiting in the initial stable state for 'lucky' transitions over the free energy barrier. Given that these transitions are much faster than the waiting times[16,17], the WE strategy can be orders of magnitude more efficient than conventional MD simulations in generating pathways for rare events such as protein folding and protein binding[18,19]. This efficiency is even higher for slower processes, increasing exponentially with the effective free energy barrier[20]. Not only are dynamics carried out without any biasing force or modifications to the free energy landscape, but suitable assignment of statistical weights to

[1]Department of Chemistry and Biochemistry, University of California–San Diego, La Jolla, CA, USA. [2]Department of Chemistry, University of Pittsburgh, Pittsburgh, PA, USA. [3]Department of Molecular Biosciences, The University of Texas at Austin, Austin, TX, USA. [4]Department of Biological Sciences, Columbia University, New York, NY, USA. [5]Department of Biochemistry and Molecular Biophysics, Columbia University Medical Center, New York, NY, USA. [6]Department of Physics, University of Wisconsin–Milwaukee, Milwaukee, WI, USA. [7]Department of Pharmacology, University of California–San Diego, La Jolla, CA, USA. [8]These authors contributed equally: Terra Sztain, Surl-Hee Ahn. ✉e-mail: ltchong@pitt.edu; ramaro@ucsd.edu

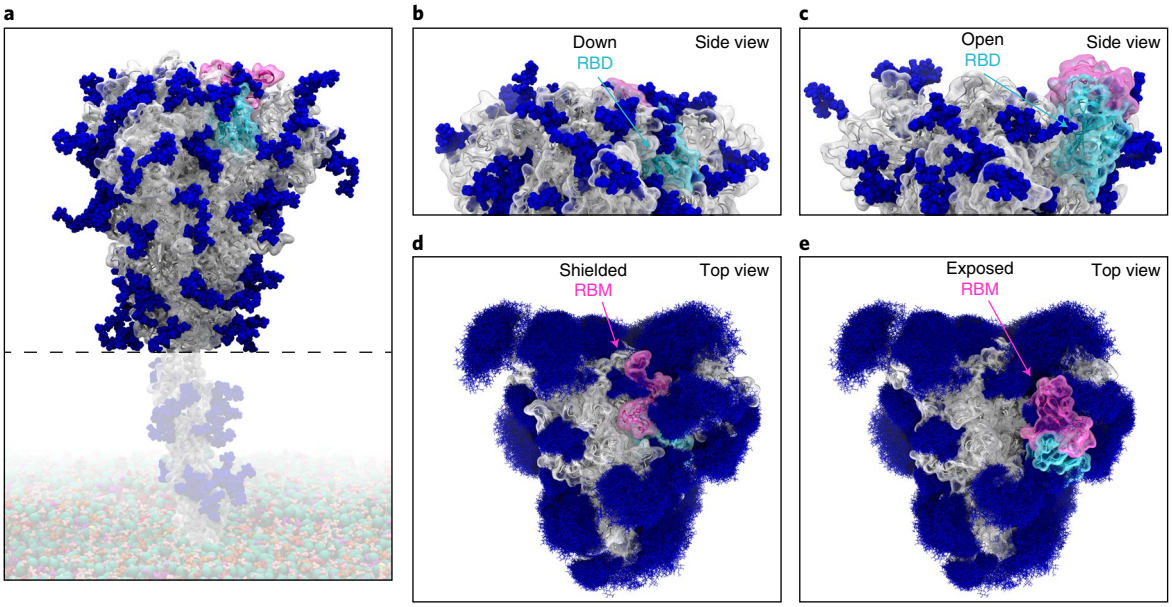


Fig. 1 | Glycosylated spike RBD 'down' and 'open' conformations. a, The SARS-CoV-2 spike head (grey) with glycans (dark blue) as simulated, with the stalk domain and membrane (not simulated here, but shown as transparent for completeness). RBD shown in cyan, RBM in pink. **b**,**c**, Side view of the RBD_{down} (shielded, **b**) and RBD_{open} (exposed, **c**). **d**,**e**, Top view of the RBM_{closed} (shielded, **d**) and RBM_{open} (exposed, **e**). Composite image of glycans (dark blue lines) shows many overlapping snapshots of the glycans over the microsecond simulations.

trajectories provides an unbiased characterization of the system's time-dependent ensemble properties[13]. The WE strategy therefore generates continuous pathways with unbiased dynamics, yielding the most direct, atomistic views for analysing the mechanism of functional transitions, including elucidation of transient states that are too fleeting to be captured by laboratory experiments. Furthermore, while the strategy requires a progress coordinate towards the target state, the definition of this target state need not be fixed in advance when applied under equilibrium conditions[21], enabling us to refine the definition of the target 'open' state of the spike protein on the basis of the probability distribution of protein conformations sampled by the simulation.

Our work characterizes a series of transition pathways of the spike opening, in agreement with conformations detected in the cryo-EM dataset by ManifoldEM[22], and identifies key residues, including a glycan at position N343, that participate in the opening mechanism. Our simulation findings are corroborated by biolayer interferometry (BLI) experiments, which show a reduction in the ability of the spike to interact with ACE2 after mutation of these key residues.

Results and discussion

WE simulations of spike opening. As mentioned above, simulations of the spike-opening process require an enhanced sampling strategy as the process occurs beyond the microsecond timescale (that is, the seconds timescale[7]). We therefore used the WE path-sampling strategy, which enabled the generation of continuous, atomistic pathways for the spike-opening process with unbiased dynamics (Fig. 2a–e and Supplementary Video 1); these pathways were hundreds of nanoseconds long, excluding the waiting times in the initial 'down' state. The protein model was based on the head region (residues 16 to 1,140) of the glycosylated SARS-CoV-2 spike from Casalino et al.[8] (Fig. 1), which in turn was built on the cryo-EM structure of the three-RBD-down spike (Protein Data Bank (PDB) ID, 6VXX (ref. [5])). The entire simulation system, including explicit water and salt ions, reaches almost half a million atoms. We focused sampling along a two-dimensional progress coordinate to track RBD opening: the difference in the centre of mass of the spike core to the RBD and the root-mean-square deviation of the RBD from the RBD_{up} state (Fig. 2f,g). On the San Diego Supercomputer Center (SDSC) Comet and Texas Advanced Computing Center (TACC) Longhorn supercomputers, 100 graphics processing units (GPUs) ran the WE simulations in parallel for over a month, generating over 130 μs of glycosylated spike trajectories and more than 200 TB of trajectory data. We simulated a total of 310 independent pathways, including 204 pathways from the RBD_{down} conformation (PDB ID, 6VXX (ref. [5])) to the RBD_{up} conformation (PDB ID, 6VSB (ref. [4])) and 106 pathways from the RBD_{down} to the RBD_{open} state, in which the RBD twists open beyond the 6VSB (ref. [4]) cryo-EM structure. Remarkably, the RBD_{open} state that we sampled includes conformations that align closely with the ACE2-bound spike cryo-EM structure (PDB ID, 7A95 (ref. [6])) even though this structure was not a target state of our progress coordinate (Fig. 2f,g, Supplementary Video 1 and Supplementary Figs. 2 and 3). This result underscores the value of using (1) equilibrium WE simulations that do not require a fixed definition of the target state and (2) a two-dimensional progress coordinate that allows the simulations to sample unexpected conformational space along multiple degrees of freedom. The ACE2-bound spike conformation has also been sampled by the Folding@home-distributed computing project[11], and RBD rotation has been detected in cryo-EM experiments[6].

Comparison with spike conformations detected by ManifoldEM. To validate our simulated RBD_{down} to RBD_{up} pathway, the ManifoldEM framework[22] was applied using the cryo-EM dataset of PDB 6VSB from McLellan and colleagues[4]. The ManifoldEM method allows characterization of conformational variations as obtained from a single-particle cryo-EM ensemble of a molecule in thermal equilibrium. Two conformational coordinates (that is, collective motion coordinates) CC1 and CC2 were discovered from this dataset, and observed from several exemplary projection directions (PDs) showing a (1) RBD_{down} to RBD_{up} pathway and (2) RBD outward opening pathway (Supplementary Fig. 4 and Supplementary Videos 2 and 3).

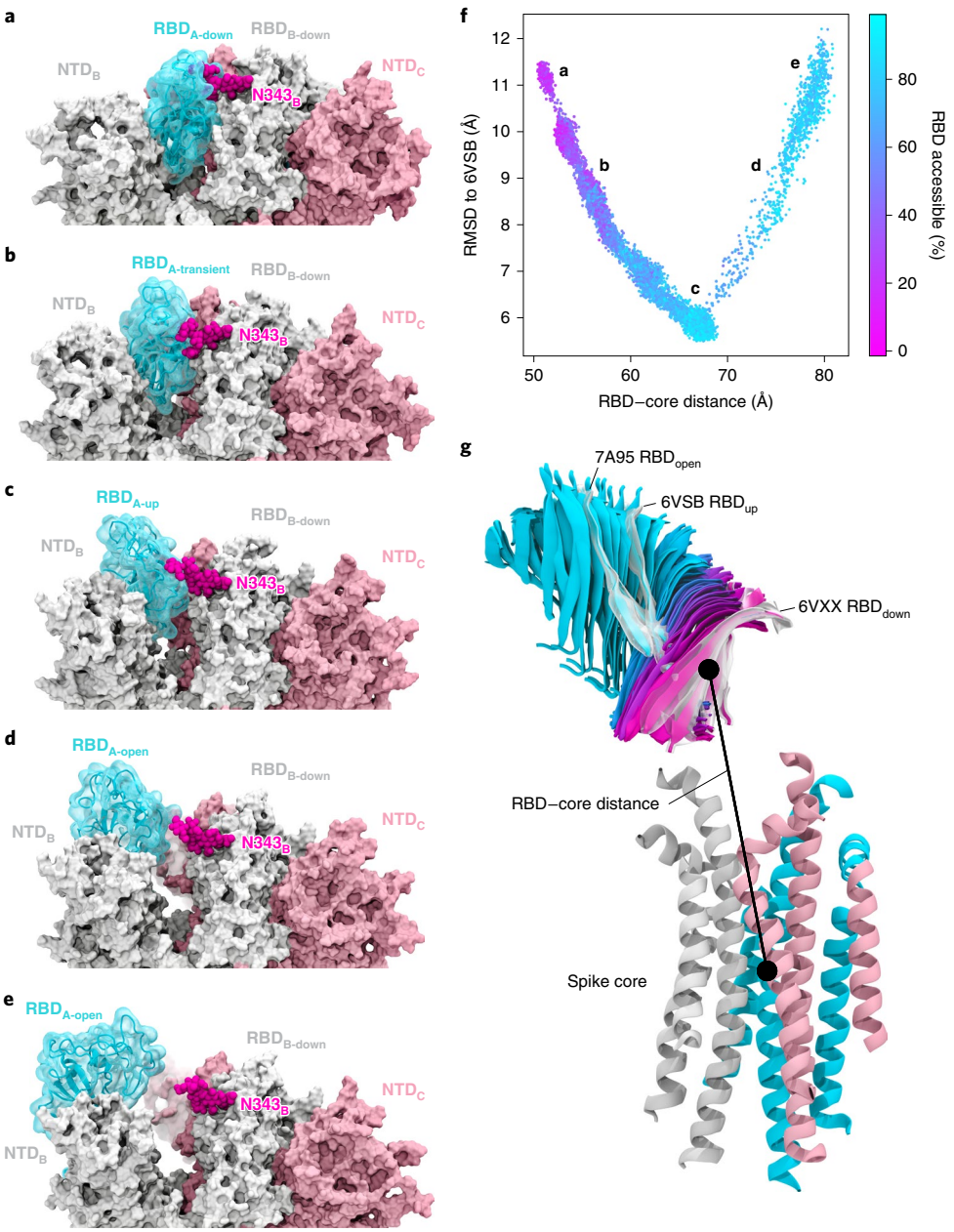


Fig. 2 | Atomically detailed pathways of spike opening. a–e, Snapshot configurations along the opening pathway with chain A shown in cyan, chain B in grey, chain C in pink and the glycan at position N343 in magenta. Each RBD and N-terminal domain (NTD) are subscripted with their chain ID (A, B or C). RBDs are also subscripted with their conformation from initial conformation with all three RBDs in the 'down' state (6VXX) (**a**), RBD_A in a 'transient' state in between the 'down' and 'up' state (6VSB) (**b**) RBD_A in the 'up' state (**c**), RBD_A in the 'open' state (beyond 6VSB) (**d**) and RBD_A in the furthest open state sampled (**e**). **f**, Scatter plot of data from the 310 continuous pathways with the Cα-root-mean-square deviation (RMSD) of the RBD from the RBD_{up} state plotted against the RBD–core distance. Data points are coloured on the basis of the percentage RBD solvent-accessible surface area compared with the RBD_{down} state. The locations of the snapshots shown in **a–e** are labelled. **g**, Primary regions of spike defined for tracking progress of the opening transition. The spike core is composed of three central helices per trimer, coloured according to chains as in **a–e**. The RBD contains a structured pair of antiparallel beta-sheets, and an overlay of snapshots from a continuous WE simulation are shown coloured along a spectrum resembling the palette in **f**. Overlayed cryo-EM structures are highlighted and labelled including the initial RBD_{down} state (6VXX), the target RBD_{up} state and the ACE2-bound RBD_{open} state (7A95).

These projections were next aligned to corresponding two-dimensional projections of coulomb potential maps generated with frames from the WE simulation (Supplementary Fig. 5 and Supplementary Videos 2 and 3). Overall, there was very good agreement between the ManifoldEM conformational coordinates and the WE trajectory, aside from two discrepancies. First, the CC2 observed in the ManifoldEM included concerted opening of all three RBDs, while the WE focused sampling on the opening of a single RBD (Supplementary Video 2). Second, the WE trajectory ultimately opens to an RBD–core distance 11 Å greater than the most open conformation in the ManifoldEM. This is probable because the simulations sample the S1 subunit en route to the post-fusion conformation, whereas the experimental dataset does not.

The N343 glycan gates RBD opening. In the 'down' state, the RBD of the SARS-CoV-2 spike is shielded by glycans at positions

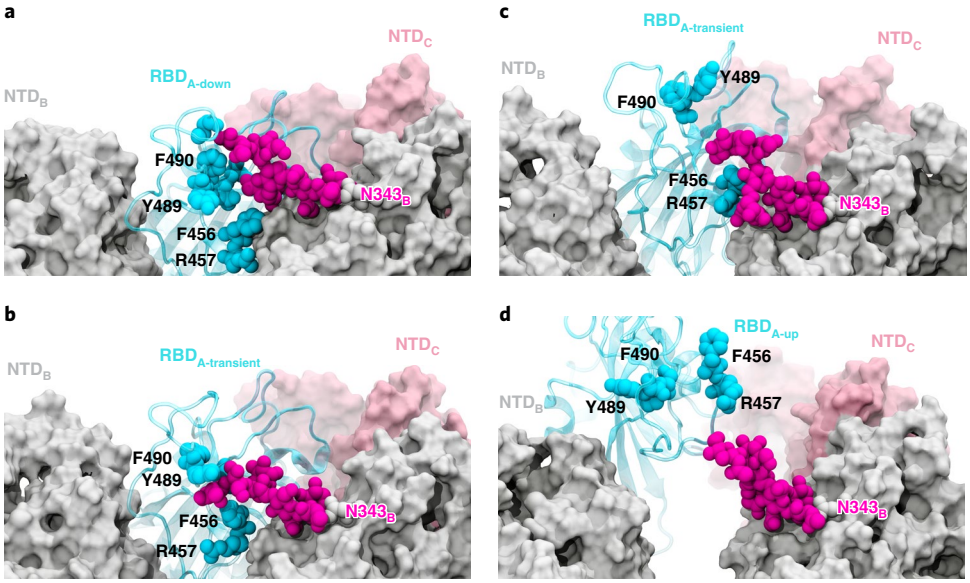


Fig. 3 | Glycan gating by N343. a–d, Snapshot configurations along the opening pathway with chain A shown in cyan, chain B in grey, chain C in pink and the glycan at position N343 in magenta. RBD_A in the 'down' conformation is shielded by the glycan at position N343 of the adjacent RBD_B (**a**). The N343 glycan intercalates between (**b**) and underneath (**c**) residues F490, Y489, F456 and F457 to push the RBD up and open (**d**).

N165, N234 and N343 (ref. [23]). While glycan shielding had been investigated for the RBD_{down} and RBD_{up} states[8], our WE simulations allowed characterization of shielding during the opening process, revealing an abrupt decrease in glycan shielding when the RBD transitions from the 'down' to the 'up' state. The glycans at position N165 and N234 consistently shield the RBM, while shielding by the N343 glycan decreases with RBD opening (Supplementary Fig. 6). Beyond shielding, a structural role for glycans at positions N165 and N234 has been recently reported, stabilizing the RBD in the 'up' conformation through a 'load and lock' mechanism[8].

Our WE simulations reveal an even more specific, critical role of a glycan in the opening mechanism of the spike: the N343 glycan acts as a 'glycan gate' pushing the RBD from the 'down' to the 'up' conformation by intercalating between residues F490, Y489, F456 and R457 of the ACE2 binding motif in a 'hand-jive' motion (Fig. 2a–e, 3 and Supplementary Video 4). Therefore, the N343 glycan plays an active role in initiating the transition, distinct from the stabilizing roles of glycans N165 and N234. This gating mechanism was initially visualized in several successful pathways of spike opening and then confirmed through analysis of all 310 successful pathways in which the N343 glycan was found to form contacts (within 3.5 Å) with each of the aforementioned residues in every successful pathway (Supplementary Fig. 7). The same mechanistic behaviour of the N343 glycan was observed in two fully independent WE simulations, suggesting the result is robust despite potentially incomplete sampling that can challenge WE and other enhanced sampling simulation methods[15].

To test the role of the N343 glycan as a key-gating residue, we performed BLI experiments. BLI experiments assess the binding level of the spike RBM (residues 438 to 508) to ACE2, acting as a proxy for the relative proportion of RBDs in the 'up' position for each spike variant. No residues directly involved in the binding were mutated (that is, at the RBM–ACE2 interface) to ensure controlled detection of the impact of RBD opening in response to mutations. Although previous results have shown reduced binding levels for N165A and N234A variants in the SARS-CoV-2 S-2P protein[8], the N343A variant displayed an even greater decrease in ACE2 binding, reducing the spike binding level by ~56% (Fig. 4 and Supplementary Table 1). As a negative control, the S383C/D985C variant[24], which is expected to be locked by disulfides into the three-RBD-down conformation, showed no association with the ACE2 receptor. These results support the hypothesis that the RBD_{up} conformation is substantially affected by glycosylation at position N343.

Atomic details of the opening mechanism. The RBD_{down} state features a hydrogen bond between T415 of the RBD_A and K986 of chain C, a salt bridge between R457 of RBD_A and D364 of RBD_B, and a salt bridge between K462 of RBD_A and D198 of N-terminal domain C (NTD_C) (Fig. 5a–c,e and Supplementary Fig. 8). The hydrogen bond $T415_A$–$K986_C$ spends an average of 12% of the successful pathways to the 'up' state before $K986_C$ makes a short lived (2% average duration to the 'up' state) salt bridge with RBD_A–D427. (Fig. 5b,e and Supplementary Fig. 8). Next, $K986_C$ forms salt bridges with $E990_C$ and $E748_C$ as the RBD_A continues to open. These contacts are formed in all 310 successful pathways (Supplementary Fig. 8). Mutation of K986 to proline has been used to stabilize the prefusion spike[25,26], including in vaccine development[27], and these simulations provide molecular context to an additional role of this residue in RBD opening.

Subsequently, at an average of 16% of the way through the successful pathways to the 'up' state, the $R457_A$–$D364_B$ salt bridge is broken, prompting the RBD_A to twist upward, away from RBD_B towards RBD_C and forming a salt bridge between R408 of RBD_A and D405 of RBD_C (Fig. 5c,e and Supplementary Fig. 8). This salt bridge persists for 20% of the successful trajectories to the 'up' state and is present in all 310 successful pathways.

A salt bridge between R466 of RBD_A and E132 from NTD_B is present in 189 out of 204 successful pathways to the 'up' state, and all 106 pathways to the 'open' state. This contact is most prevalent during the transition between the 'up' and 'open' state. Finally, the salt bridge between D428 of RBD_A and R454 of RBD_C is present only in all 106 pathways from the 'up' to the 'open' state and is the last salt bridge between the RBD and the spike in the 'open' state before the S1 subunit begins to peel off (Fig. 5d,e and Supplementary Fig. 8), at which point the last remaining contact to the RBD_A is the glycan at position N165 of NTB_B.

Additional BLI experiments of the key identified spike residues R408A, D405A and D427A corroborate the pathways observed in

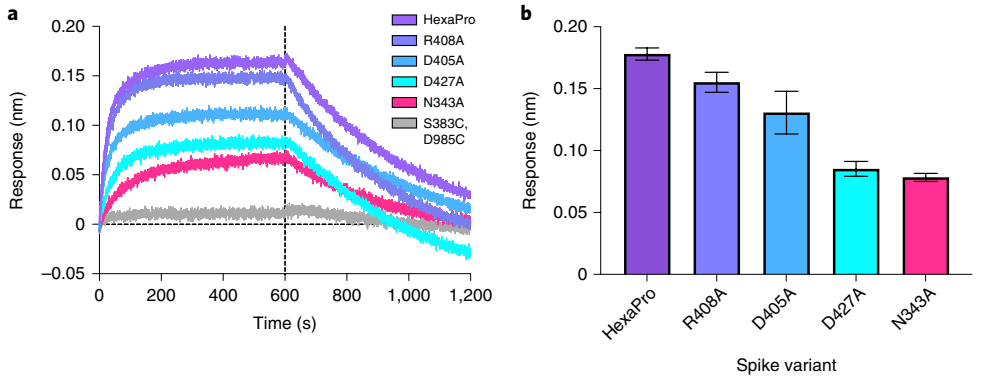


Fig. 4 | ACE2 binding is reduced by mutation of the N343 glycosylation site and key salt bridge residues. a, BLI sensorgrams of HexaPro spike variants binding to ACE2. For clarity, only the traces from the first replicate are shown. **b**, Graph of the binding response for BLI data collected in triplicate with error bars representing the standard deviation from the mean.

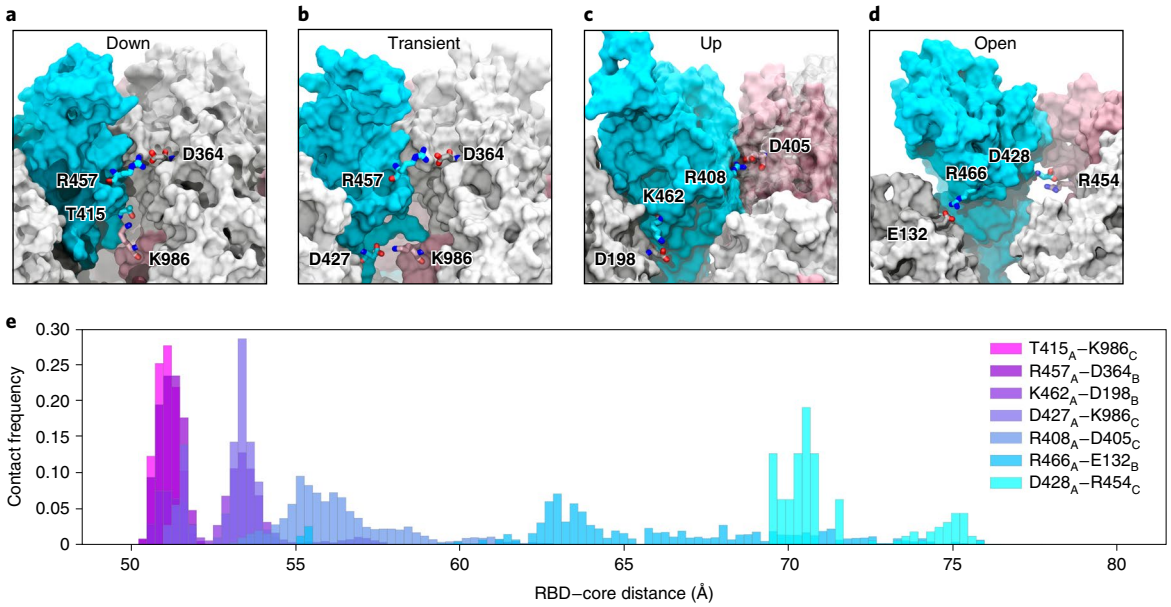


Fig. 5 | Salt bridges and hydrogen bonds along the opening pathway. a–d, Salt-bridge or hydrogen-bond contacts made between RBD_A, shown in blue, and RBD_B, shown in grey, or RBD_C, shown in pink, within the 'down' (**a**), 'transient' (**b**), 'up' (**c**) and 'open' (**d**) conformations. Nitrogen atoms are coloured blue and oxygen atoms are coloured red in the stick representation of amino acids to show hydrogen-bond and salt-bridge contacts. **e**, Histogram showing the frequency at which residues from **a–d** are within 3.5 Å of each other relative to the RBD–core distance. Frequencies are normalized to 1.

our simulations. Each of these reduces the binding interactions of the spike with ACE2 by ~13%, ~27% and ~52%, respectively (Fig. 4 and Supplementary Table 1). We also note that identified residues D198, N343, D364, D405, R408, T415, D427, D428, R454, R457, R466, E748, K986 and E990 are conserved between SARS-CoV and SARS-CoV-2 spikes, supporting their importance in coordinating the primary spike function of RBD opening. The emerging mutant SARS-CoV-2 strains, B.1 (D614G), B.1.1.7 (H69–V70 deletion and Y144–Y145 deletions, N501Y, A570D, D614G, P681H, T716I, S982A and D1118H), B.1.351 (L18F, D80A, D215G, R246I, K417N, E484K, N501Y, D614G and A701V), P.1 (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y and T1027I), and CAL.20 C (L452R and D614G)[28], do not contain mutants in the residues we identified here to facilitate RBD opening. Analysis of neighbouring residues and glycans to those mutated in the emerging strains along the opening pathway is detailed in Supplementary Table 2, and distances between each residue and glycan to RBD_A is summarized in Supplementary Video 5.

Conclusions

We report extensive WE MD simulations of the glycosylated SARS-CoV-2 spike head characterizing the transition from the 'down' to 'up' conformation of the RBD. Over 130 μs of simulation provide more than 300 independent RBD-opening transition pathways. The simulated opening pathways align very well to conformations detected from cryo-EM with the ManifoldEM method. Analysis of these pathways from independent WE simulations indicates a clear gating role for the glycan at N343, which lifts and stabilizes the RBD throughout the opening transition. We also characterize an 'open' state of the spike RBD, in which the N165 glycan of chain B is the last remaining contact with the RBD en route to further opening of S1. BLI experiments of residues identified

as key in the opening transitions, including N343, D405, R408 and D427, broadly supported our computational findings. Notably, a 56% decrease in ACE2 binding of the N343A mutant, compared with a 40% decrease in N234A mutant and a 10% decrease in the N165A mutant reported previously[8], evidenced the key role of N343 in gating and assisting the RBD-opening process, highlighting the importance of sampling functional transitions to fully understand mechanistic detail. None of the individual mutations fully abolished ACE2 binding, indicating that the virus has evolved a mechanism involving multiple residues to coordinate spike opening. Our work indicates a critical gating role of the N343 glycan in spike opening and provides new insights into mechanisms of viral infection for this important pathogen.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41557-021-00758-3.

Received: 15 February 2021; Accepted: 21 June 2021;
Published online: 19 August 2021

References

1. Chan, J. F.-W. et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* **395**, 514–523 (2020).
2. Lu, R. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
3. Li, F. Structure, function, and evolution of coronavirus spike proteins. *Annu. Rev. Virol.* **3**, 237–261 (2016).
4. Wrapp, D. et al. Cryo-EM structure of the 2019-NCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).
5. Walls, A. C. et al. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **181**, 281–292.e6 (2020).
6. Benton, D. J. et al. Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature* **588**, 327–330 (2020).
7. Lu, M. et al. Real-time conformational dynamics of SARS-CoV-2 spikes on virus particles. *Cell Host Microbe* **28**, 880–891.e8 (2020).
8. Casalino, L. et al. Beyond shielding: the roles of glycans in the SARS-CoV-2 spike protein. *ACS Cent. Sci.* **6**, 1722–1734 (2020).
9. Gur, M. et al. Conformational transition of SARS-CoV-2 spike glycoprotein between its closed and open states. *J. Chem. Phys.* **153**, 075101 (2020).
10. Fallon, L. et al. Free energy landscapes for RBD opening in SARS-CoV-2 spike glycoprotein simulations suggest key interactions and a potentially druggable allosteric pocket. Preprint at *ChemRxiv* https://doi.org/10.26434/chemrxiv.13502646.v1 (2020).
11. Zimmerman, M. I. et al. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nat. Chem.* **13**, 651–659 (2021).
12. Huber, G. A. & Kim, S. Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophys. J.* **70**, 97–110 (1996).
13. Zhang, B. W., Jasnow, D. & Zuckerman, D. M. The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *J. Chem. Phys.* **132**, 054107 (2010).
14. Chong, L. T., Saglam, A. S. & Zuckerman, D. M. Path-sampling strategies for simulating rare events in biomolecular systems. *Curr. Opin. Struct. Biol.* **43**, 88–94 (2017).
15. Zuckerman, D. M. & Chong, L. T. Weighted ensemble simulation: review of methodology, applications, and software. *Annu. Rev. Biophys.* **46**, 43–57 (2017).
16. Pratt, L. R. A statistical method for identifying transition states in high dimensional problems. *J. Chem. Phys.* **85**, 5045–5048 (1986).
17. Zuckerman, D. M. & Woolf, T. B. Transition events in butane simulations: similarities across models. *J. Chem. Phys.* **116**, 2586–2591 (2002).
18. Adhikari, U. et al. Computational estimation of microsecond to second atomistic folding times. *J. Am. Chem. Soc.* **141**, 6519–6526 (2019).
19. Saglam, A. S. & Chong, L. T. Protein–protein binding pathways and calculations of rate constants using fully-continuous, explicit-solvent simulations. *Chem. Sci.* **10**, 2360–2372 (2019).
20. DeGrave, A. J., Ha, J.-H., Loh, S. N. & Chong, L. T. Large enhancement of response times of a protein conformational switch by computational design. *Nat. Commun.* **9**, 1013 (2018).
21. Suárez, E. et al. Simultaneous computation of dynamical and equilibrium information using a weighted ensemble of trajectories. *J. Chem. Theory Comput.* **10**, 2658–2667 (2014).
22. Dashti, A. et al. Trajectories of the ribosome as a Brownian nanomachine. *Proc. Natl Acad. Sci. USA* **111**, 17492–17497 (2014).
23. Watanabe, Y., Allen, J. D., Wrapp, D., McLellan, J. S. & Crispin, M. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science* **369**, 330–333 (2020).
24. Henderson, R. et al. Controlling the SARS-CoV-2 spike glycoprotein conformation. *Nat. Struct. Mol. Biol.* **27**, 925–933 (2020).
25. Hsieh, C.-L. et al. Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science* **369**, 1501–1505 (2020).
26. Pallesen, J. et al. Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. *Proc. Natl Acad. Sci. USA* **114**, E7348–E7357 (2017).
27. Cross, R. The tiny tweak behind COVID-19 vaccines. *Chem. Eng. News* **98**, 18–20 (2020).
28. Corum, J. & Zimmer, C. Coronavirus variants and mutations. *New York Times* (10 February 2021).
29. Amaro, R. E. & Mulholland, A. J. A community letter regarding sharing biomolecular simulation data for COVID-19. *J. Chem. Inf. Model.* **60**, 2653–2656 (2020).
30. Bogetti, A. T. et al. A suite of tutorials for the WESTPA rare events sampling software. *Living J. Comput. Mol. Sci.* **1**, 10607 (2019).

Data availability

Data supporting the findings of this study are included in the article and its Supplementary Information files. We endorse the community principles around open sharing of COVID-19 simulation data[29]. All simulation input files and data are available at the NSF MolSSI COVID-19 Molecular Structure and Therapeutics Hub at https://covid.molssi.org and the Amaro Lab website http://amarolab.ucsd.edu. Source data are provided with this paper.

Code availability

This study utilized the standard builds of the simulation software WESTPA 2020.02 (https://github.com/westpa/westpa) and AMBER 18 (https://ambermd.org) according to best practices for running WE simulations[30] with no special modifications.

Acknowledgements

We are grateful for the efforts of the Texas Advanced Computing Center (TACC) Longhorn team and for the computing time made available through a Director's Discretionary Allocation (made possible by National Science Foundation (NSF) award OAC-1818253). We thank Z. Gaieb for helpful discussions around system construction. We thank M. Tatineni for help with computing on SDSC Comet, as well as a COVID-19 HPC Consortium Award for computing time. We also thank C. Simmerling and his research group (SUNY Stony Brook), and A. Mulholland and his research group (University of Bristol), for helpful discussions related to the spike protein, as well as D. Zuckerman, J. Copperman, M. Zwier and A. Saglam for helpful methodological discussions. T.S. is funded by an NSF GRFP grant (DGE-1650112). This work was supported by: a National Institutes of Health (NIH) grant (GM132826); an NSF RAPID grant (MCB-2032054); an award from the RCSA Research Corp. and a UC San Diego Moores Cancer Center 2020 SARS-CoV-2 seed grant to R.E.A.; an NIH grant (R01-GM31749) to J.A.M.; an NIH grant (R01-AI127521) to J.S.M.; an NIH grant (R01 GM115805) and an NSF grant (CHE-1807301) to L.T.C.; and NIGMS grants (R01 GM29169 and R35 GM139453) to J.F. A.O. and G.M. acknowledge support by the US Department of Energy, Office of Science, Basic Energy Sciences under award DE-SC0002164 (underlying dynamical algorithms), and by the US National Science Foundation under awards STC 1231306 (underlying data-analytical techniques) and DBI-2029533 (underlying data-analytical models).

Author contributions

T.S. and S.-H.A. contributed equally to this work. R.E.A. and L.T.C. oversaw the project. T.S. and L.C. prepared the simulation model. T.S. and S.-H.A. performed WE simulations and A.T.B. provided WESTPA scripts. S.-H.A., A.T.B., T.S. and L.T.C. carried out WE analysis. T.S. and F.L.K. performed simulation analyses. L.C., T.S. and F.L.K. created figures and videos. J.S.M. designed and oversaw BLI experiments. R.S.M. and J.A.G. performed BLI experiments and wrote the corresponding parts in Results and Methods. J.F. and A.O. directed the ManifoldEM study, and E.S., F.A.-R., S.M. and G.M. performed the ManifoldEM study. E.S. and F.A.-R. analysed the results and compared them with WE simulations. E.S. and J.F. described the ManifoldEM methods and results. T.S., S.H.-A., L.T.C. and R.E.A. wrote the manuscript with contributions from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41557-021-00758-3.

Correspondence and requests for materials should be addressed to L.T.C. or R.E.A.

Peer review information *Nature Chemistry* thanks Syma Khalid and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

nature portfolio

https://doi.org/10.1038/s41557-021-00758-3

Supplementary information

A glycan gate controls opening of the SARS-CoV-2 spike protein

In the format provided by the authors and unedited

Supplementary Information

A glycan gate controls opening of the SARS-CoV-2 spike protein

Terra Sztain[1†], Surl-Hee Ahn[1†], Anthony T. Bogetti[2], Lorenzo Casalino[1], Jory A. Goldsmith[3], Evan Seitz[4], Ryan S. McCool[3], Fiona L. Kearns[1], Francisco Acosta-Reyes[5], Suvrajit Maji[5], Ghoncheh Mashayekhi[6], J. Andrew McCammon[1,7], Abbas Ourmazd[6], Joachim Frank[4,5], Jason S. McLellan[3], Lillian T. Chong[2*], Rommie E. Amaro[1*]

1. Department of Chemistry and Biochemistry, UC San Diego, La Jolla, CA 92093

2. Department of Chemistry, University of Pittsburgh, Pittsburgh, PA 15260

3. Department of Molecular Biosciences, The University of Texas at Austin, Austin, TX 78712

4. Department of Biological Sciences, Columbia University, New York, NY, 10032, USA

5. Department of Biochemistry and Molecular Biophysics, Columbia University Medical Center, New York, NY 10032, USA

6. Department of Physics, University of Wisconsin-Milwaukee, 3135 N. Maryland Ave, Milwaukee, WI 53211, USA

7. Department of Pharmacology, UC San Diego, La Jolla, CA 92093

† These authors contributed equally to this work.

* contact authors: ramaro@ucsd.edu, ltchong@pitt.edu

Table of Contents

1. Supplementary Methods

1.1 Computational Methods

1.1.1 Model preparation of the initial "down" state

A model of the "down" state of the glycosylated spike structure and CHARMM36 force field parameters[31,32] was obtained from Casalino *et al*,[8] modeled using the cryoEM structure (PDB ID: 6VXX);[5] in this model hydrogen atoms were added using ionization states present in solution at pH 7.4. The stalk and membrane were excluded, and only residues 16-1140 of each trimer were used (**Fig. 1A**). The system was solvated in a cubic box of TIP3P[33] explicit water molecules with at least 10 Å between the protein and box edges and 150 mM NaCl using VMD,[34] yielding a system size of 490,621 atoms. The GPU-accelerated Amber18[35,36,37,38] molecular dynamics (MD) engine was used, which gave a 16-fold speedup in dynamics propagation on a GPU vs. CPU. To enable the use of the Amber18 software package, the Chamber program[39] was used to convert the CHARMM36 force field parameters into an Amber readable format.

To relieve unfavorable interactions, the solvated system was subjected to a two-stage energy minimization followed by a two-stage equilibration. To minimize the energy of the system, the solvent was first minimized for 10,000 steps with harmonic position restraints (force constant of 100 kcal/mol /Å^2) applied to the sugars and proteins followed by an unrestrained minimization of the entire system for 100,000 steps. To equilibrate the energy-minimized system, the system was incrementally heated to 300 K over 300 ps in the NVT ensemble followed by a 1-ns equilibration in the NPT ensemble. A production simulation was then carried out in the NPT ensemble for 20 ns on the Triton Shared Computing Cluster at San Diego Supercomputer Center (SDSC). Equilibration and production simulations were carried out with a 2 fs timesteps and SHAKE[40] constraints on bonds to hydrogens. Pressure and temperature were controlled with the Monte Carlo barostat (with 100 fs between attempts to adjust the system volume) and the Langevin thermostat (1 ps^{-1} collision frequency), respectively. Long-range electrostatics were accounted for with the PME method[41] using a 10 Å cutoff for short-range, non-bonded interactions. To provide more extensive sampling of the closed state, we selected a set of 24 equally weighted conformations ("basis states") from the latter 5 ns of the production simulation for a weighted

ensemble (WE) simulation; this portion of the simulation exhibited reasonable convergence of the Cα root-mean-squared deviation (RMSD) from the initial, minimized conformation (**Supplemental Fig. 6**).

1.1.2 Weighted ensemble simulations

The weighted ensemble (WE) path sampling strategy orchestrates an ensemble of parallel trajectories with periodic communication to enhance the sampling of pathways for rare events without biasing the dynamics.[15] In particular, a resampling step is applied at fixed time intervals τ to enrich for promising trajectories that have advanced towards the target state – typically, along a progress coordinate that has been divided into bins. Trajectories are all initially assigned equal statistical weights and rigorously tracked to ensure that all weights sum to one at all times of the simulation, introducing no bias in the dynamics.[12] During the resampling step, trajectories that transition to empty bins are replicated and their corresponding weights split evenly between the resulting child trajectories; trajectories that do not make progress are occasionally terminated with their respective weights merged to other trajectories that will be continued. (**Supplemental Fig. 1**)

WE simulations can be run under non-equilibrium steady state or equilibrium conditions and can therefore provide equilibrium (*e.g.*, state populations) and non-equilibrium observables (e.g., rate constants), respectively. To maintain non-equilibrium steady-state conditions, trajectories that reach the target state are "recycled" by initiating a new trajectory from the initial state with the same trajectory weight; steady-state WE simulations therefore require that the target state be defined in advance of the simulation, but are more efficient in generating successful events than equilibrium WE simulations. On the other hand, equilibrium WE simulations do not require a fixed definition of the target state and therefore enable refinement of the target-state definition at any time during the simulation. Here, we leveraged the advantages of both non-equilibrium steady state and equilibrium WE simulations: steady-state simulations were used to more efficiently generate successful pathways trajectories once the target state could be defined and equilibrium simulations were used to further explore and refine the definition of the target state.

All WE simulations were run using the open-source, highly scalable WESTPA software package[42] (**Supplemental Fig. 7**) with a fixed time interval τ of 100 ps for resampling and a target number of 8 trajectories/bin. Details of the progress coordinate and bin spacing for each WE simulation are provided below.

Extensive sampling of the initial "down" state

To extensively sample the initial "down" state, we ran an equilibrium WE simulation starting from randomly selected conformations from the basis states discussed above. A two-dimensional progress coordinate was used. One dimension consisted of the distance between the centers of mass (COM) of (i) Cα atoms of the entire system **and** all atoms in the four main beta strands of the RBD (residues 375-380, 394-404, 431-438, 508-517; refers to RBD from chain A unless otherwise specified), and (ii) Cα atoms of the entire system **and** all atoms in the structured region of the helical core domain (residues 747-784, 946-967, 986-1034 from each of the three chains). The second dimension consisted of the Cα RMSD of the entire system and all atoms in the four main beta strands of the RBD from the initial model of the "down"-state structure after 1 ns equilibration. Progress coordinates were calculated using CPPTRAJ.[43] This initial WE simulation was run for 8.77 days on 80 P100 GPUs on Comet at the San Diego Supercomputer Center (SDSC) collecting a comprehensive sampling of ~7.5 μs aggregate simulation time. Bin spacing was periodically monitored and adjusted to maximize efficient sampling.

Due to a typo in the CPPTRAJ atom selection *(i.e.,* "and" instead of "of"), the progress coordinate above was not the one we originally intended. Our intention was to use 1) the COM distance between the Cα atoms **of** the four main beta sheets of the RBD and the Cα atoms of the structured region of the helical core domain and 2) the Cα RMSD **of** the four main beta sheets of the RBD from the initial model of the "down"-state structure. As shown in **Figs. 2F and S2**, our WE simulations with this progress coordinate nonetheless capture the large-scale protein transitions that are evident with the intended progress coordinate, but on a more compressed scale.

Simulations of spike opening

After extensive sampling of the "down" state, exploratory WE simulations were run to determine effective progress coordinates and binning to capture the opening of the spike protein. Based on these simulations, we found that taking the RMSD from the target "up" state was much more effective than taking the RMSD from the initial "down" state. The target state, with one RBD in the "up" conformation, modeled by Casalino *et al.*[8] using the cryoEM structure (PDB ID: 6VSB),[4] was subject to 1 ns of equilibration using identical methods as described above for the closed structure. The RMSD of the initial state from the target state was calculated as 11.5 Å.

Next, an independent, equilibrium WE simulation was conducted using the two-dimensional progress coordinate described above for sampling the "down" state, but taking the RMSD from the target "up" state instead of the initial "down" state and using the bin spacing determined by the exploratory simulations. The WE simulation was stopped for analysis after 1729 iterations, 19.64 days on 100 NVIDIA V100 GPUs on Longhorn at TACC, collecting an aggregate of ~51.5 μs of sampling and 106 pathways from the "down" to the "open" state. Finally, another WE simulation that was under non-equilibrium steady-state conditions was conducted to maximize sampling of transitions from the "down" to the "up" states. This WE simulation started from iteration 1576 of the previous WE simulation, which was the last iteration before the RBD-COM distance was 9.0 Å or greater, was stopped for analysis after 3000 iterations, 25.03 days later, on 100 NVIDIA V100 GPUs on Longhorn at TACC, collecting an additional ~69.2 μs of sampling and 204 pathways from the "down" to the "open" state. The WESTPA software was shown to scale almost linearly on these 100 NVIDIA V100 GPUs on Longhorn (**Supplemental Fig. 7**), which enabled fast and efficient simulation of the spike.

1.1.3 Analysis of weighted ensemble simulations

Number of successful pathways

The successful pathways that reached the "up" state (8.9 Å ≤ RBD-COM distance) or the "open" state (9.9 Å ≤ RBD-COM distance) were obtained by counting all arrivals to that particular state at every WE iteration, which yielded 204 and 106 pathways, respectively. We consider these pathways to be statistically independent pathways. The splitting trees for the 204 and 106 pathways, respectively, can be seen in **Supplemental Figs. 8 and 9**, respectively, which shows trajectory segments shared by the pathways and points of splitting the pathways. The number of

pathways is similar to that obtained from calculating the autocorrelation function of arrivals to the "up" and "open" states at a particular WE iteration. For instance, at the end of the WE simulation that sampled the "open" state, there were 1824 trajectories in total and 1193 trajectories that were part of the "open"-state ensemble (defined in later sections as 9.0 Å ≤ RBD-COM distance). Out of the 1193 trajectories that reached the "open"-state ensemble, 133 trajectories were calculated to be statistically independent from calculating the autocorrelation function of the number of arrivals to the "open"-state ensemble[19] (**Supplemental Fig. 10**). The correlation time was calculated to be 16 WE iterations or 1.6 ns so the trajectories that did not share a common segment for 16 iterations from the last point in the trajectory were considered to be statistically independent. By checking these multiple independent pathways that reached the "up" or "open" states, we were able to confirm reproducibility of the identified glycan and residue interactions involved in the particular transition. For calculating the shortest and longest transition times, all successful pathways were taken into account. The first 25% of all successful pathways were disregarded to obtain the most probable transition times, however, since the initial transitions can skew the transition time to be shorter than it is normally (**Supplemental Figs. 11 and 12**).

State definitions

Based on our WE simulations, key states were defined as follows. The "down"-state ensemble consisted of structures with RMSD ≥ 11.0 Å and RBD-COM distance ≤ 7.5 Å, ~13 μs aggregate simulation time. Note that the entire progress coordinate array had to satisfy the criteria to be counted as part of the ensemble. The "up"-state ensemble was defined as 8.5 Å ≤ RBD-COM distance < 9.0 Å, ~ 6.5 μs aggregate simulation time. The "open"-state ensemble was defined as having an RBD-COM distance ≥ 9.0 Å, ~ 4.9 μs aggregate simulation time.

Trajectory analysis

Trajectories were visualized using VMD.[34] Glycans, salt bridge, and hydrogen bonding interactions involved in the "down" to "up" and "open" transition were first visually identified. Next, distances between the identified residues were calculated using cpptraj[43] for all 310 successful pathways, and plotted with matplotlib.[44] To obtain the percentage of the most probable transition time that had a certain salt bridge, the distance between the atoms/residues of

the salt bridge was measured, and the total time in which the distance was less than 3.5 Å was calculated. The total time for each pathway was calculated and averaged to obtain the final percentage. To obtain the number of successful pathways that had a certain quantity, *e.g.,* salt bridge, glycan-residue contact, the pathway was counted if the distance was less than 3.5 Å in at least one of the conformations, sampling conformations every 100 ps. Contact maps calculating the distance between the RBD (from chain A) and all other residues and glycans were generated using MDAnalysis[45,46] (**Supplemental Video 5**). Structures for figures and movies were generated using VMD, including NanoShaper[47] surface representation.

Solvent accessible surface area (SASA) was calculated using a protocol presented in Casalino *et al.*[8] involving the *measure sasa* command within VMD and a solvent probe radius of 1.4 Å. The surface area of the Receptor Binding Motif (RBM, residues 438-508 in chain A) that was shielded by glycans was calculated by taking the difference between the SASA of the "naked" spike (without glycans) and the SASA of the glycosylated spike (with glycans). Individual contributions to shielding of the RBM by glycans at positions N165-B, N234-B, N343-B were also calculated by considering only the respective glycans in the SASA calculation of the glycosylated spike.

Analysis of residues mutated in emerging SARS-CoV-2 strains

To date, the following SARS-CoV-2 variants have been identified (with mutations to spike noted in parentheticals): B.1 (D614G), B.1.1.7 (H69-V70 deletion, Y144-Y145 deletions, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H), B.1.351: (L18F, D80A, D215G, R246I, K417N, E484K, N501Y, D614G, A701V), P1 (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y ,D614G, H655Y, T1027I) and CAL.20C (L452R, D614G).[28] To examine potential implications of these mutations on Spike opening mechanics, we have monitored the neighboring residues of key WT residues as a function of the opening mechanism. MDAnalysis[45,46] was used to identify residues whose center of mass was within 10 Å of the center of mass of the key residue of interest. For each contact, the fraction of conformations in the "down", "up", and "open" ensembles containing the contact is provided. Contacts were only considered if they exist within > 5% of all conformations and if the contacting pairs were separated by more than three peptide bonds in one-dimensional sequence.

1.2 ManifoldEM method

1.2.1 Background

The set of algorithms now under the name ManifoldEM[48] employ a three-step procedure[22] to characterize conformational variations in a dataset from single-particle cryo-EM of a molecule in thermal equilibrium. In the first step, which can be performed on any of the existing cryo-EM platforms, data are classified by orientation, and prepared as aligned image stacks. In the second step, for each projection direction (PD) data falling into the angular aperture are analyzed as a manifold and represented in a low-dimensional space spanned by what is now termed "conformational coordinates,"[48] equivalent to collective motion coordinates. In the third step, the manifold representations resulting from the second step, one for each projection direction, are reconciled and combined across the angular sphere to obtain a consolidated representation. From this an energy landscape can be obtained, enabling a functional analysis of the molecule,[48] and 3D volumes can be captured along inferred trajectories.

1.2.2 Preprocessing

The initial image-stack we received from McLellan and colleagues corresponding to PDB ID: 6VSB[4] contained 631,920 snapshots. This initial image stack was pruned by approximately 10% (from 631,920 to 578,588 particles) to remove artifacts. Additional 3D Auto-Refinement via RELION[49] was performed to realign all images. Next RELION 2D Classification was used to remove an additional 1% of particles, leaving the final count of 574,324. The consensus refinement in RELION displayed a Fourier Shell Correlation ($FSC_{0.143}$) of 4.3 Å. In parallel, this stack was separately refined using CryoSPARC[50] non-uniform refinement with a GSFSC resolution of 3.5 Å.

These two refinements were next compared within the preliminary steps of ManifoldEM. Although both reconstructions appeared fine, we found upon closer examination that the RELION refinement encountered a problem of preferred orientations, where thousands of particles had been clumped within nearly the same local area (*i.e.*, nearly identical Euler coordinates) of the 2-sphere. In contrast, the CryoSPARC non-uniform refinement produced

much more uniformly-distributed angular assignments, albeit with a lower average occupancy per PD. 2D conformational coordinate movies obtained in ManifoldEM from the CryoSPARC alignment proved superior to those using RELION. While the CryoSPARC alignment was chosen for all subsequent steps in ManifoldEM, the RELION protocol was not altogether without its own merit. We additionally ran RELION focused 3D Classification using the angular alignment from CryoSPARC with a mask around the RBDs. We obtained classes with different configurations of the RBD, including one class in the RBD-"down" conformation (**Supplemental Fig. 4)**. The original study,[4] in contrast, found no such particles - nor did other labs to which the data were sent for further analysis. Importantly, the discovery of these missing particles explains the presence of RBD-"down" volumes constructed along the 3DVA[51] "reaction coordinate" discovered in that study.[4]

1.2.3 Manifold embedding

We next set up a more thorough ManifoldEM analysis using the cryoSPARC alignment. First, a number of initial inputs are required for the ManifoldEM pipeline to tessellate the orientational 2-sphere into a finite number of PDs. These are (1) Pixel size: 1.047 Å; (2) Resolution: 3.5 Å; (3) Object diameter: 335 Å (taken as the maximum width of the average volume); and (4) Aperture index: {1-5}. The aperture index is a flexible parameter that controls the angular width of each PD, such that a larger aperture index corresponds to more images assigned to each PD from a larger region of angular space. After experimenting with several aperture indices and evaluating the corresponding PD statistics and 2D movie qualities, we chose aperture index 5 for all future computations. This measure provided us with 1678 PDs thoroughly spread out in angular space, with a handful of regions with heightened PD-occupancy. When displayed as a histogram, the occupancy of PDs exhibited a chi-squared distribution, with the majority of PDs housing around 230 images and a rightward tail reaching approximately 800 images in the most highly-occupied PD.

Following the ManifoldEM framework, 1678 manifolds were constructed from the images in each corresponding PD via the Diffusion Maps[52] framework. Following Dashti et al.,[22] Nonlinear Laplacian Spectral Analysis (NLSA)[53] was then performed on the eigenvectors of these high-dimensional manifolds to extract a set of possible reaction coordinates from each. In

sum, these steps were programmed to produce eight 2D movies per PD, with each 2D movie corresponding to one of the PD-manifold's eigenvectors.

Conformational analysis

Upon completion, our task was to next classify the type of motions seen in each 2D movie per PD, noting that not all 2D movies extracted must correspond to valid conformational information; this is especially true of those obtained with smaller singular values. Our approach was to initiate a search to detect all PDs housing 2D movies with above-average visual appearance. In this search, many PD-manifolds were found to have extremely noisy or otherwise insensible information. This was a predictable scenario given the known deficiencies in the dataset[4] (*i.e.*, orientational bias leading to low occupancies in many PDs), and beyond remediation by ManifoldEM. As a result, only a subset of PDs where the images therein met the prerequisites for the manifold embedding approach could be analyzed. Of these above-threshold PDs, we found 216 PDs of the 1678 PDs (13%) with above average quality and 73 high-quality PDs (4%), as judged by visual inspection relative to the whole. Thus, overall, a relatively small percentage of the data as partitioned into these PDs met the prerequisite conditions for displaying the highest-quality conformational variation signals.

We next organized all above-average PDs into 22 well-spaced groups on the 2-sphere, and selected several of the best PDs from each angular region. Detailed analysis was performed on the 64 PDs chosen, including classification of conformational motion type in each of the eight 2D NLSA movies per PD. As shown in **Supplementary Videos 2 and 3**, we predominantly observed two conformational motions: (1) RBD-"down" to RBD-"up"; and (2) trimer-claw close to open, which we call conformational coordinate 1 (CC1) and conformational coordinate 2 (CC2), respectively. However, PDs where a clear distinction existed between CC1 and CC2 were rare. Specifically, CC1 alone could only be clearly established in 31 of 64 PDs (48%); while both CC1 and CC2 were found occupying separate 2D movies in only 6 of 64 PDs (9%). In the remaining PDs, these conformational motions were not cleanly separated but were present in hybrid form.

This discrepancy arises from the nature of our analysis, where we define Euclidean distances between images that are 2D projections of the molecule. As a result, from a given viewing direction, a 3D motion projected onto 2D will appear more or less pronounced than it does in some other, depending on the type of motion and PD. For example, we found that the CC2 trimer claw motion was most pronounced only when observed from the "top-down view", the PD aligned with the axis of the protein's central alpha helices (PD 112).

Transformation of structures along WE trajectory

We next aimed to compare the conformational coordinates discovered by ManifoldEM from experimental cryo-EM ensembles with the WE motions observed in the spike-opening trajectory detailed in the main text. To this end, we converted the PDB files from the WE into a collection of 2D projections. We first selected 20 frames from the WE trajectory spanning conformations from the RBD-"down" to the RBD-"up" state. We next imported these files into Chimera[54] along with a coarse 3D map obtained from ManifoldEM to be used for alignment reference. In order to place both frameworks in the same coordinate system for subsequent analysis, we translated and rotated the PDB files to coincide with the ManifoldEM map, using the Chimera fitmap command. Each PDB was then saved in Chimera. Next, these fitted PDBs were re-centered using Phenix[55] pdbtools and converted into MRC-formatted Coulomb potential maps via EMAN2[56] e2pdb2mrc. For this last step, a resolution of 5 Å was chosen based on visual assessment of the EMAN2 outputs relative to those from ManifoldEM. Projections of these 20 MRCs were then taken using the standard projection operator in e2project3d with C1 symmetry in EMAN2. Importantly, the Euler coordinates for these projections were supplied by those representing the 64 ManifoldEM anchors (after correcting for a coordinate transformation from ManifoldEM to ZXZ' convention). Finally, these projections were combined into sequences for each PD to form 64 20-frame 2D movies of the WE trajectory.

1.2.4 Comparison of WE simulations to ManifoldEM outputs

As shown in **Supplementary Videos 2 and 3**, and described in detail within our main text, a striking visual resemblance emerged between conformational motions obtained by WE simulation and experiment. For heightened visual aid, 2D movies from the WE simulation and

the ManifoldEM corresponding to the same PD (and RC therein) were next overlaid to directly highlight similarities and differences. For this procedure, we first layered the ManifoldEM movie over a homogenous red backdrop and applied a Linear Dodge blend mode, with a similar effect applied on the WE movie over a blue backdrop (see **Supplemental Fig. 4** for the results of these operations). We next multiplied the ManifoldEM composite image and the WE composite image together. As an outcome of this multiplication, pixels that are white (signal) in both movies retain their whiteness in the composite. In this way, whiteness in the composite movie becomes a qualitative measure of similarity between conforming domains, while non-white regions emphasize differences.

Finally, this overlaying approach was used to estimate the total extent of the RBD motion as expressed in the ManifoldEM and WE frameworks. For this comparison, CC1 from a side view (PD 1386) was chosen based on its highly prominent view of RBD-"up" to RBD-"down" motion. Next, the ManifoldEM movie was time-remapped to align it optimally in time with the motions observed in the corresponding WE movie (**Supplementary Video 2**). Using the multiplication-composite as a guide, it was determined that the ManifoldEM RBD domain reaches its full extent in the "up" position at the 14th frame out of the 20 frames from the WE trajectory, before the WE trajectory moves onward to a more fully open state. With this knowledge, the total difference in conformational extents was estimated at 11 Å as calculated via RBD — core distance.

1.3 Experimental Methods

Protein Expression and Purification

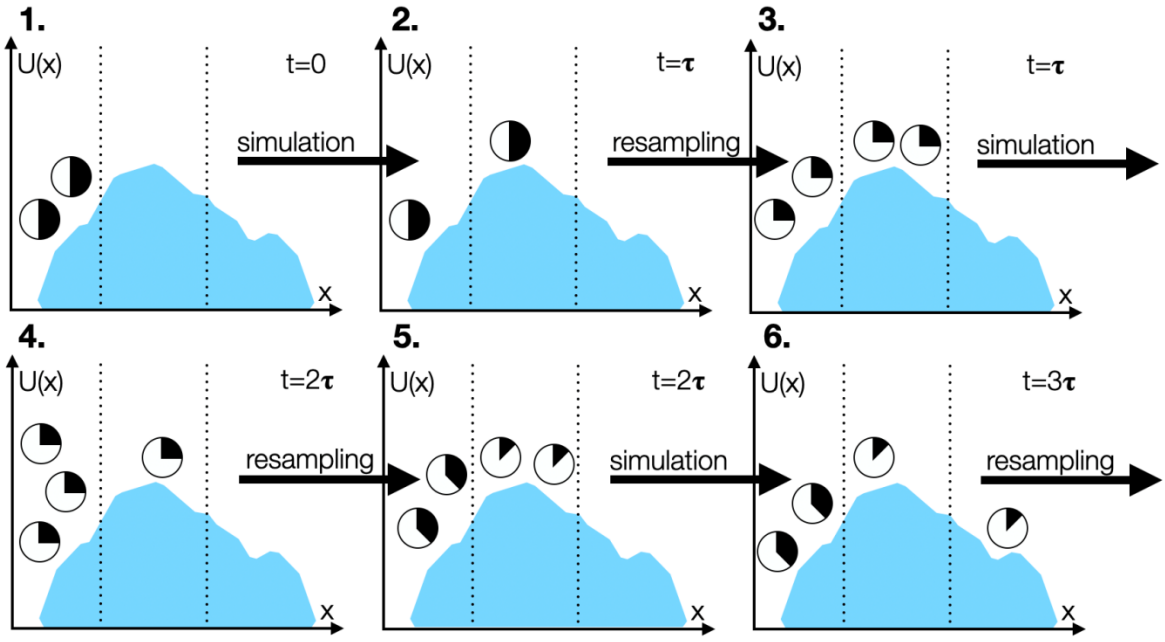
Substitutions N343A, D405A, R408A, and D427A were cloned into the HexaPro SARS-CoV-2 spike background.[23] A spike variant with all RBDs locked in the "down" position through the introduction of a disulfide bond was similarly produced through cysteine substitutions at residues S383C and D985C in the HexaPro protein.[25] All variants were expressed through polyethyleneimine-induced transient transfection of FreeStyle 293-F cells (Thermo Fisher). After 4 days, cell supernatant was clarified by centrifugation, passed through a 0.22 μm filter, and

purified over StrepTactin resin (IBA). Variants were further purified by size-exclusion chromatography on a Superose 6 10/300 column (GE Healthcare) in a buffer consisting of 2 mM Tris pH 8.0, 200 mM NaCl and 0.02% NaN_3. Soluble ACE2 was produced and purified as previously described.[8]

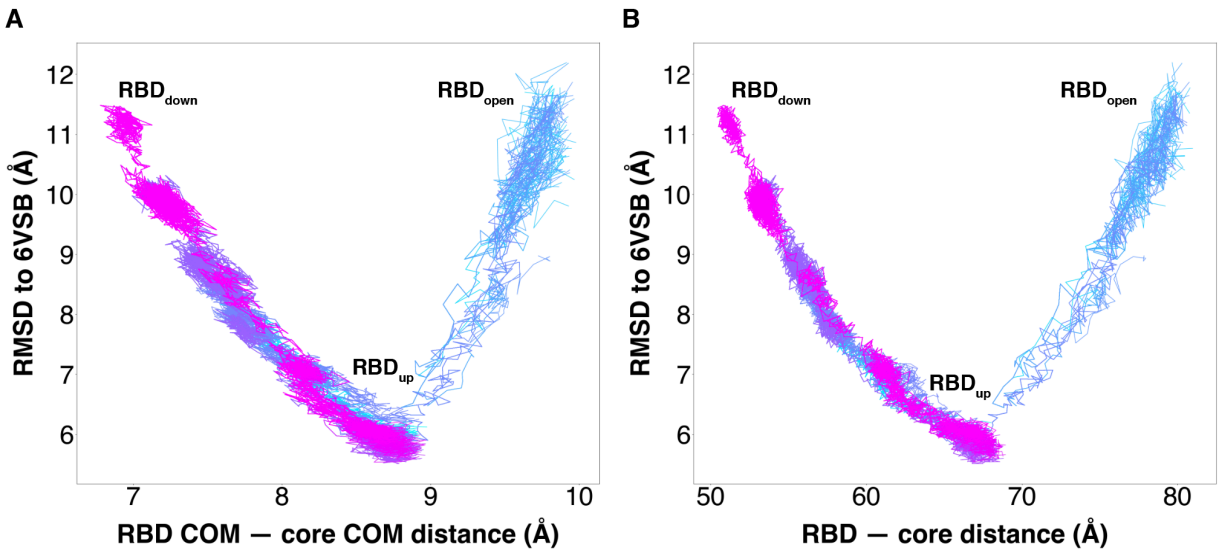
Biolayer Interferometry

Anti-foldon IgG was immobilized to an anti-human Fc (AHC) Octet biosensor (FortéBio). Tips were then submerged into the specified HexaPro variants before being subsequently dipped into 200 nM ACE2 to observe variant association, followed by dissociation in buffer consisting of 20 mM Tris pH 7.5, 150 mM NaCl, 1 mg/mL bovine serum albumin, and 0.01% Tween-20. The relative proportion of RBD in an accessible state was quantified based on the binding level as previously described.[8] The S383C, D985C variant was used as a negative control. Data were collected in triplicate and replicate sensorgrams are shown in **Supplemental Fig. 16.**

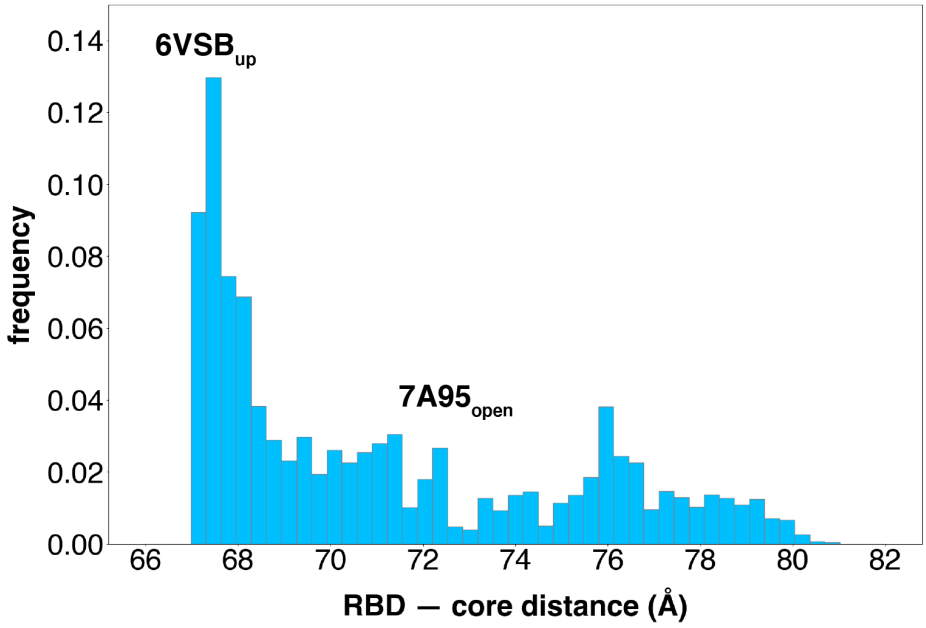
2. Supplementary Figures



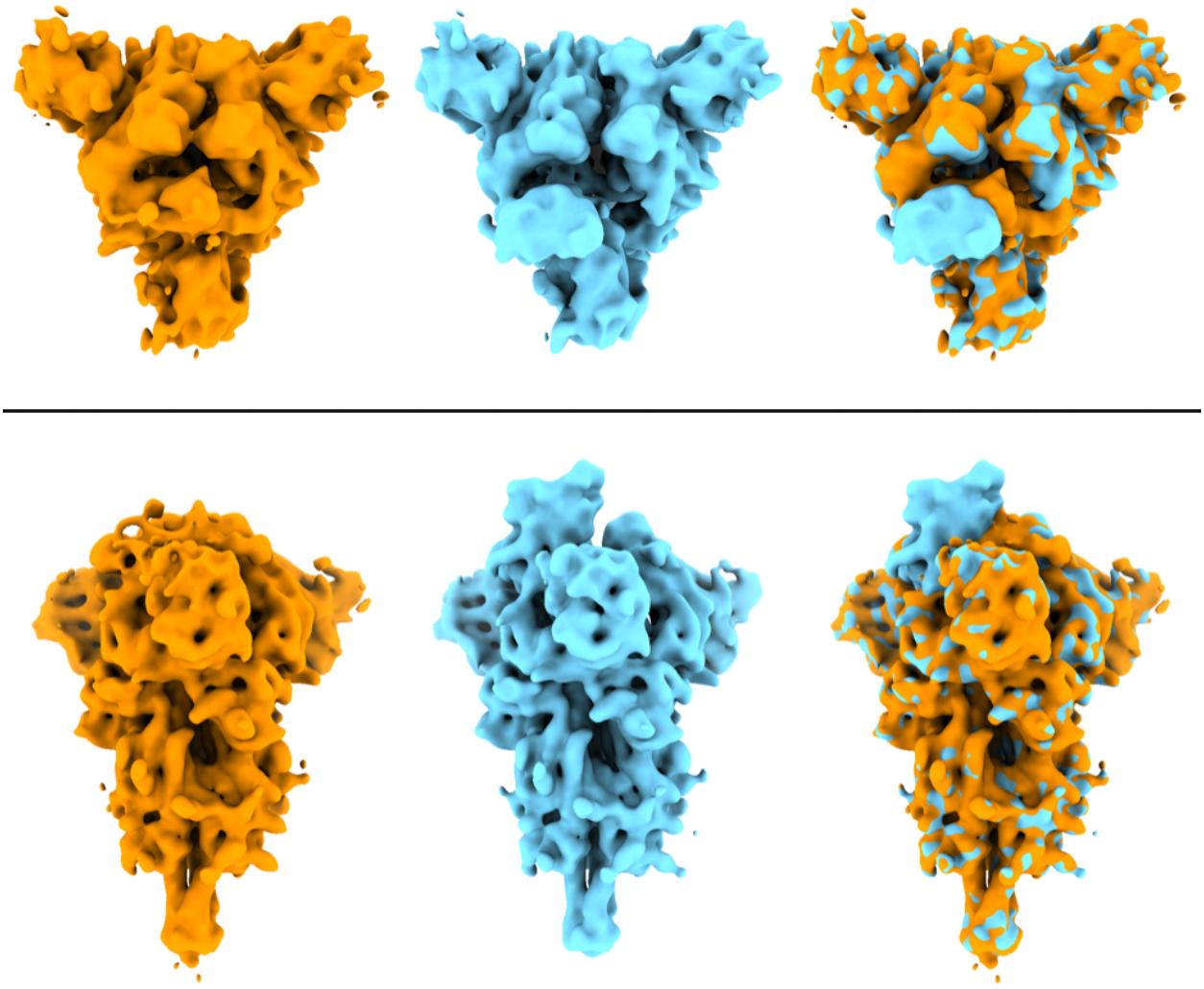
Supplemental Fig. 1 Schematic of the weighted ensemble (WE) strategy. The WE strategy is illustrated for a three-state system with a one-dimensional progress coordinate x that is divided into bins. U(x) represents the potential of the system dependent on x, which can be seen from the curve of the shaded region. 1. WE initiates two equally weighted trajectories (represented as circles) from the first bin, each with a statistical weight of 0.5 (represented as filled parts of the circles), for a fixed time interval **τ**. 2. Resampling is then performed, replicating or terminating trajectories to maintain a target number of two trajectories in each bin (*e.g.,* in the first and second bins, splitting the weight among the two child trajectories with a weight of 0.25 for each trajectory). 3. Trajectories are run for another fixed time interval **τ**. 4. After running, resampling is performed (*e.g.,* in the first bin, terminating two of the three trajectories and in the second bin, replicating the one trajectory to yield two trajectories). 5. The system ends up with two trajectories in each of the visited bins. 6. One of the trajectories ends up in the third bin. Rounds of simulation and resampling are performed until a desired number of continuous pathways into the target state are generated.



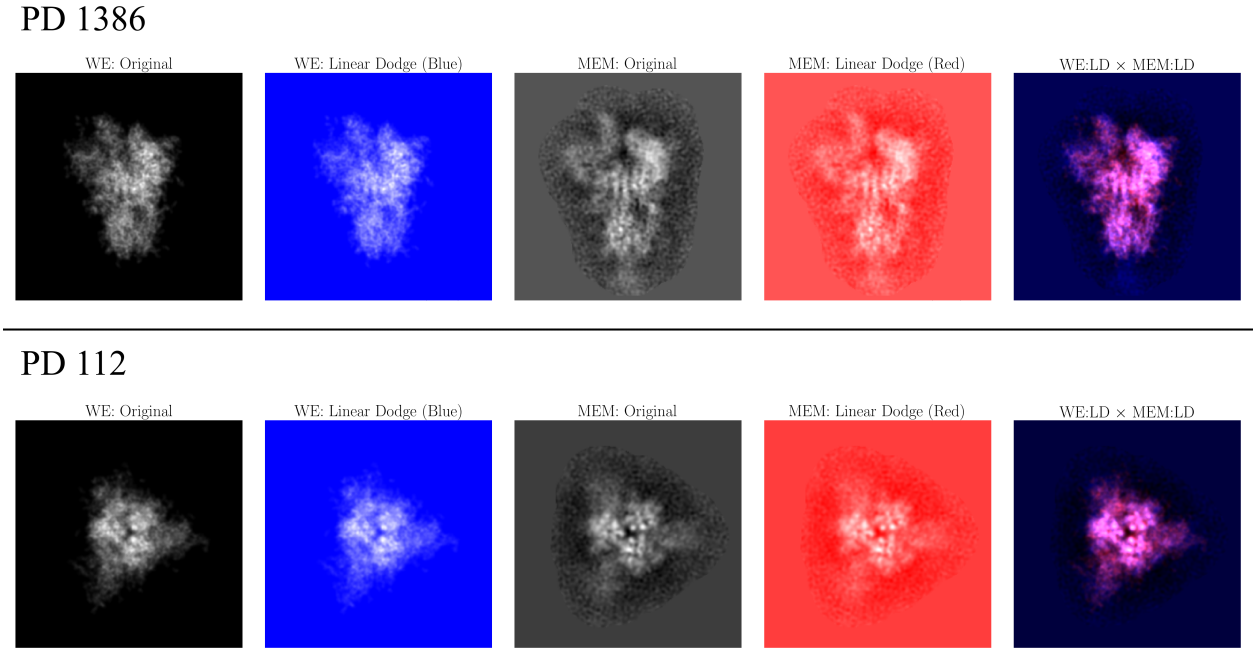
Supplemental Fig. 2 Successful pathways of spike opening for the (A) actual and (B) intended progress coordinate. Overlay of 310 successful pathways including 204 pathways of the RBD transitioning from the "down" state to the "up" state (magenta-purple) and 106 pathways from the "down" to the "open" states (purple to cyan). Continuous trajectories plotted with the Cα RMSD of the RBD to the 6VSB "up" state versus the RBD — core distance.



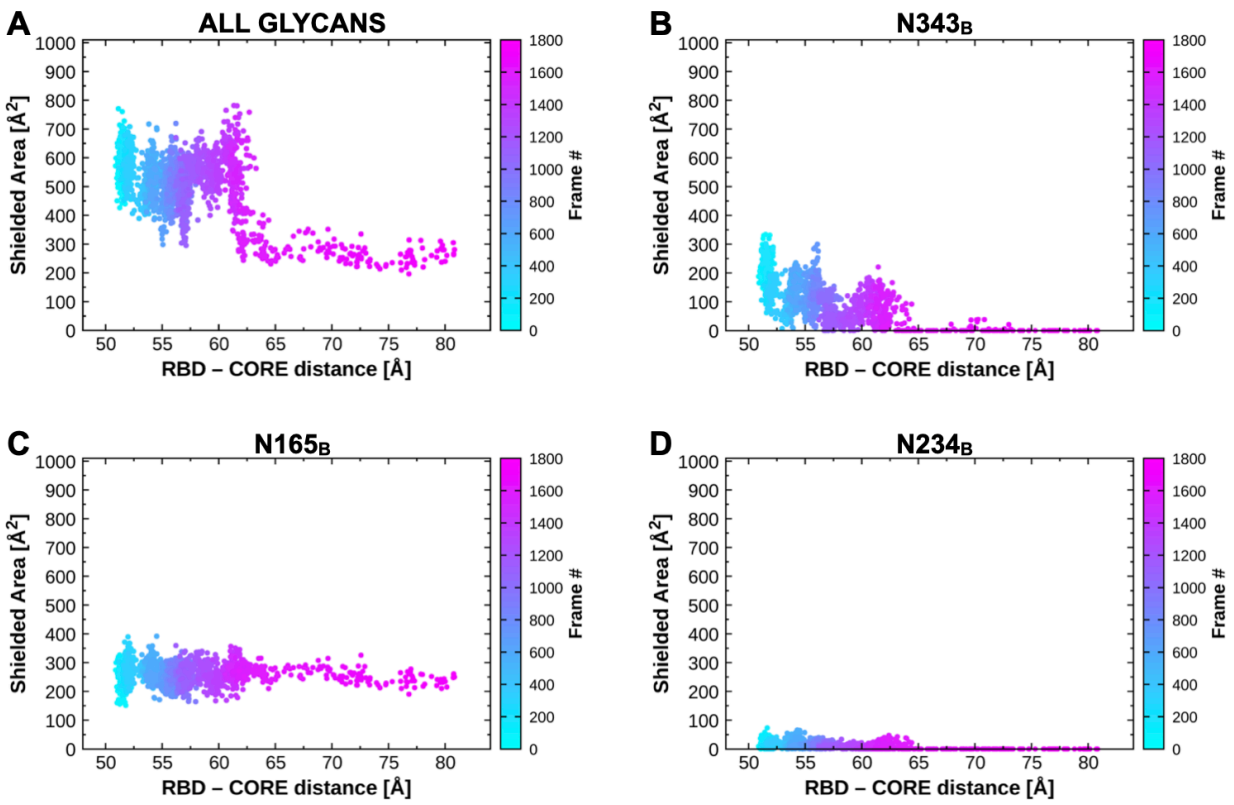
Supplemental Fig. 3 Diversity of the simulated RBD "open" state ensemble. Probability distribution of RBD — core distances greater than the RBD "up" conformation defined by PDB 6VSB (67.2 Å). The ACE2-bound structure from PDB 7A95 distance is 72.1 Å.



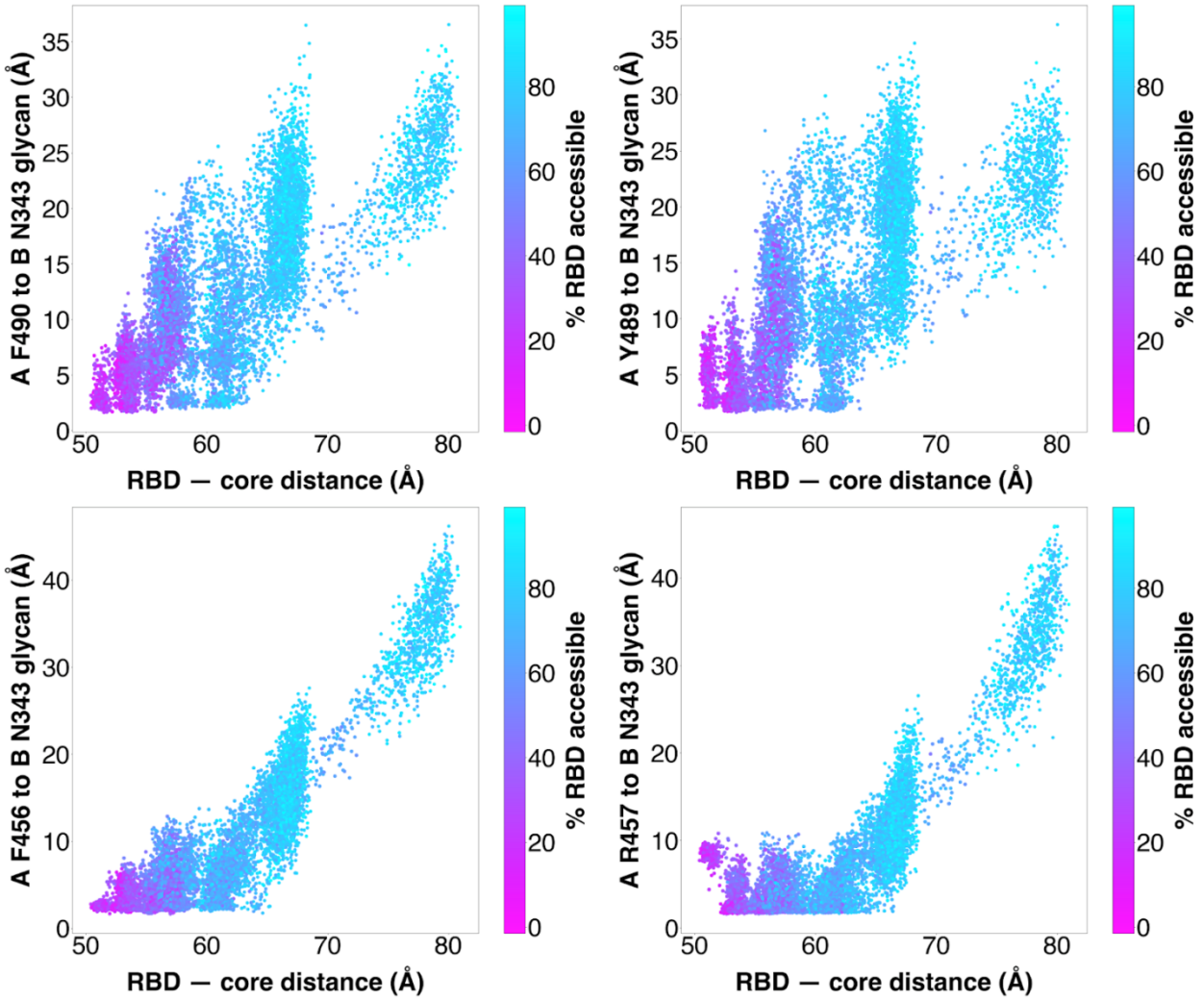
Supplemental Fig. 4 Comparison of two classes from the focused 3D classification in RELION with top and side views of the reconstructed classes. EM density maps are low pass filtered to 8 Å for display purposes. The class with the RBD "down" conformation is displayed with orange on the left, the class with the RBD "up" is displayed with cyan in the center, and the superposition of both maps is shown on the right side to highlight their differences.



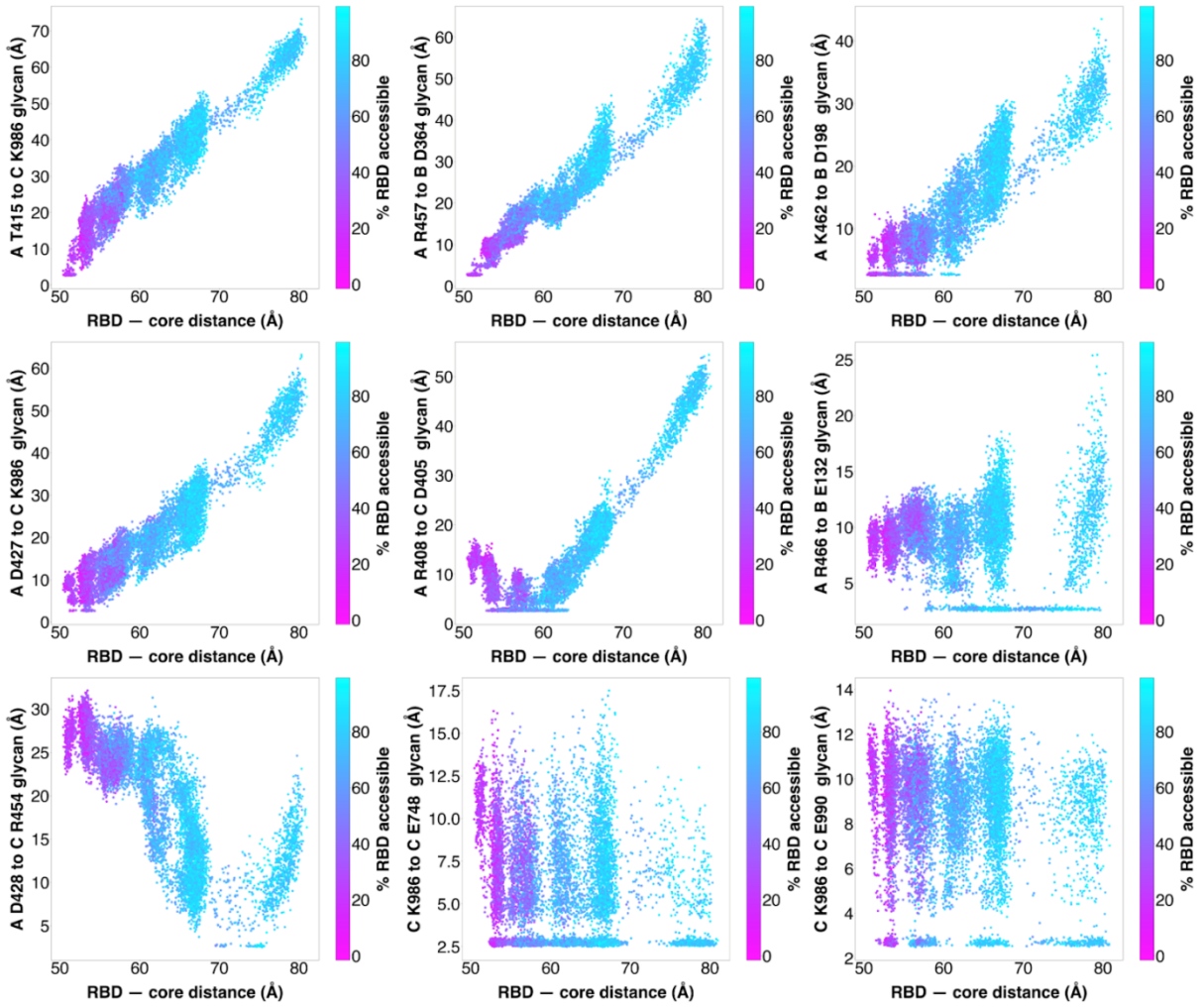
Supplemental Fig. 5 Comparison of a frame from the WE and ManifoldEM (MEM) trajectory as seen from a side view (PD 1386) and top-down view (PD 112). For this comparison, image compositing techniques are applied on the outputs of each method as shown in the columns, including Linear Dodge and Multiply. As an example of its utility, after performing this operation on RC2 from a top-down view (PD 112), it can be seen that a collection of white pixels emerged in the composite movie (bottom-right entry), which strongly emphasize the similarities in positions of RBD and spike core helices between frameworks.



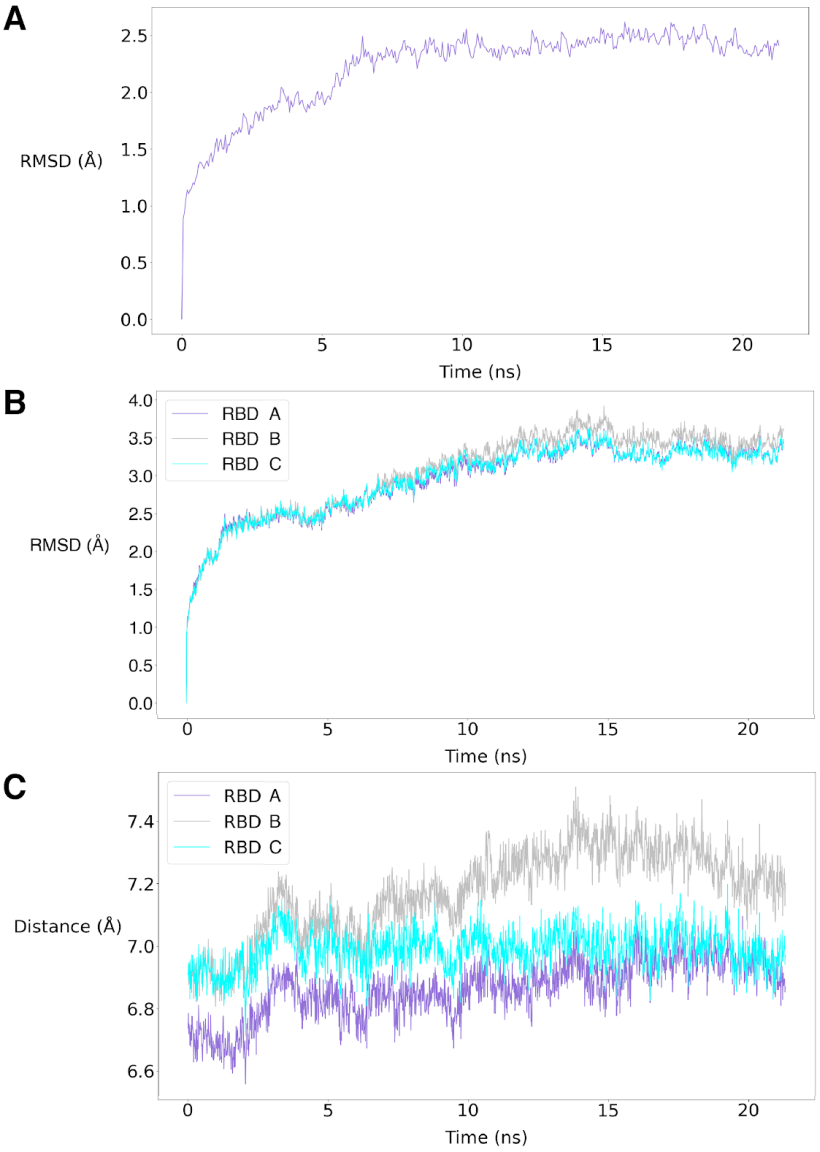
Supplemental Fig. 6 Contribution of glycans shielding receptor binding motif along RBD opening pathway. Shielded area represents the difference between the solvent accessible surface area of the receptor binding motif in the presence and absence of (A) all three glycans, (B) N343, (C) N165, or (D) N234.



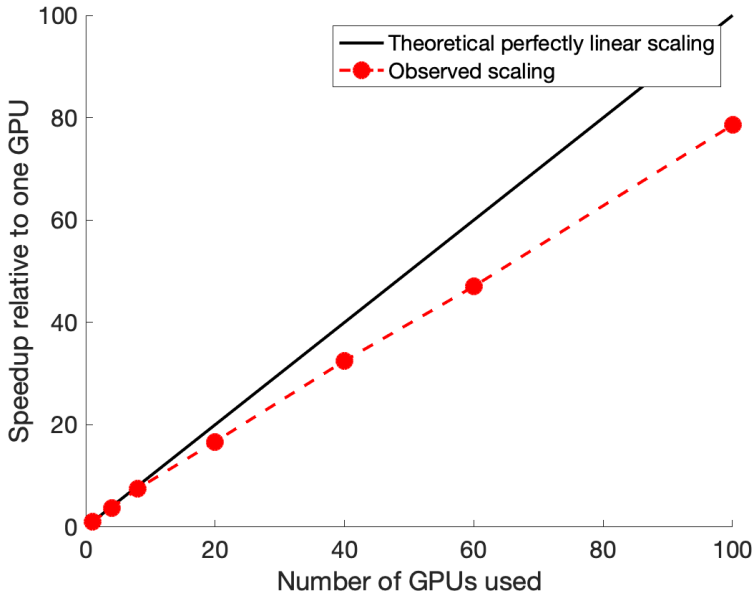
Supplemental Fig. 7 Distance between N343 glycan and RBD residues. Scatter plot of data from the 310 continuous pathways with the minimum distance between the N343 glycan and RBD A residues F490, Y489, F456, or R457 plotted against RBD — core distance. Data points are colored based on % RBD solvent accessible surface area compared to the RBD "down" state 6VXX.



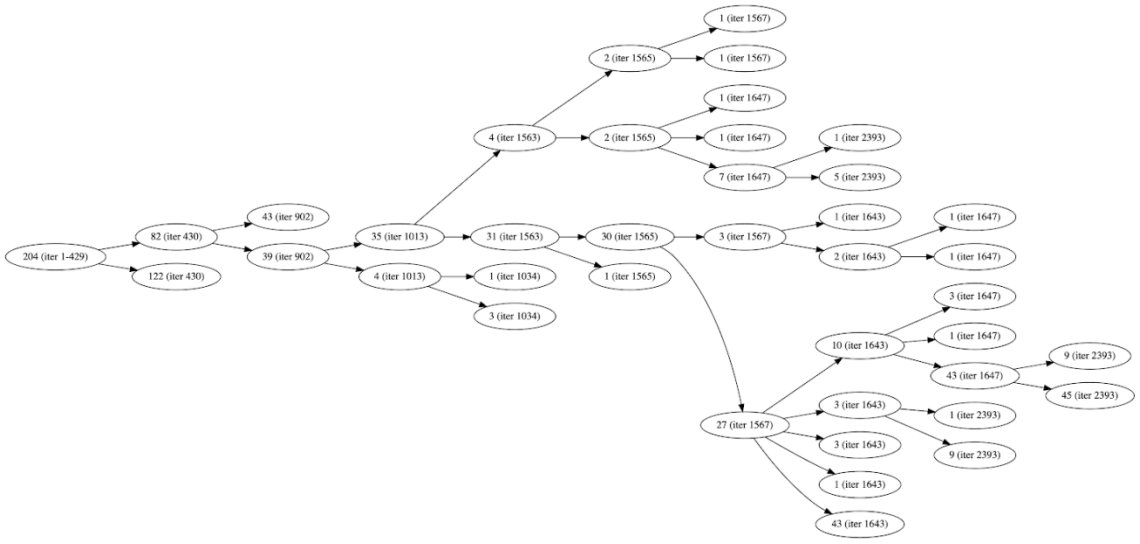
Supplemental Fig. 8 Distance between salt-bridge and hydrogen bonding residues along the spike opening pathway. Scatter plot of data from the 310 continuous pathways with the minimum distance between the residues shown in **Figure 4** plotted against RBD-core distance. Data points are colored based on % RBD solvent accessible surface area compared to the RBD "down" state 6VXX.



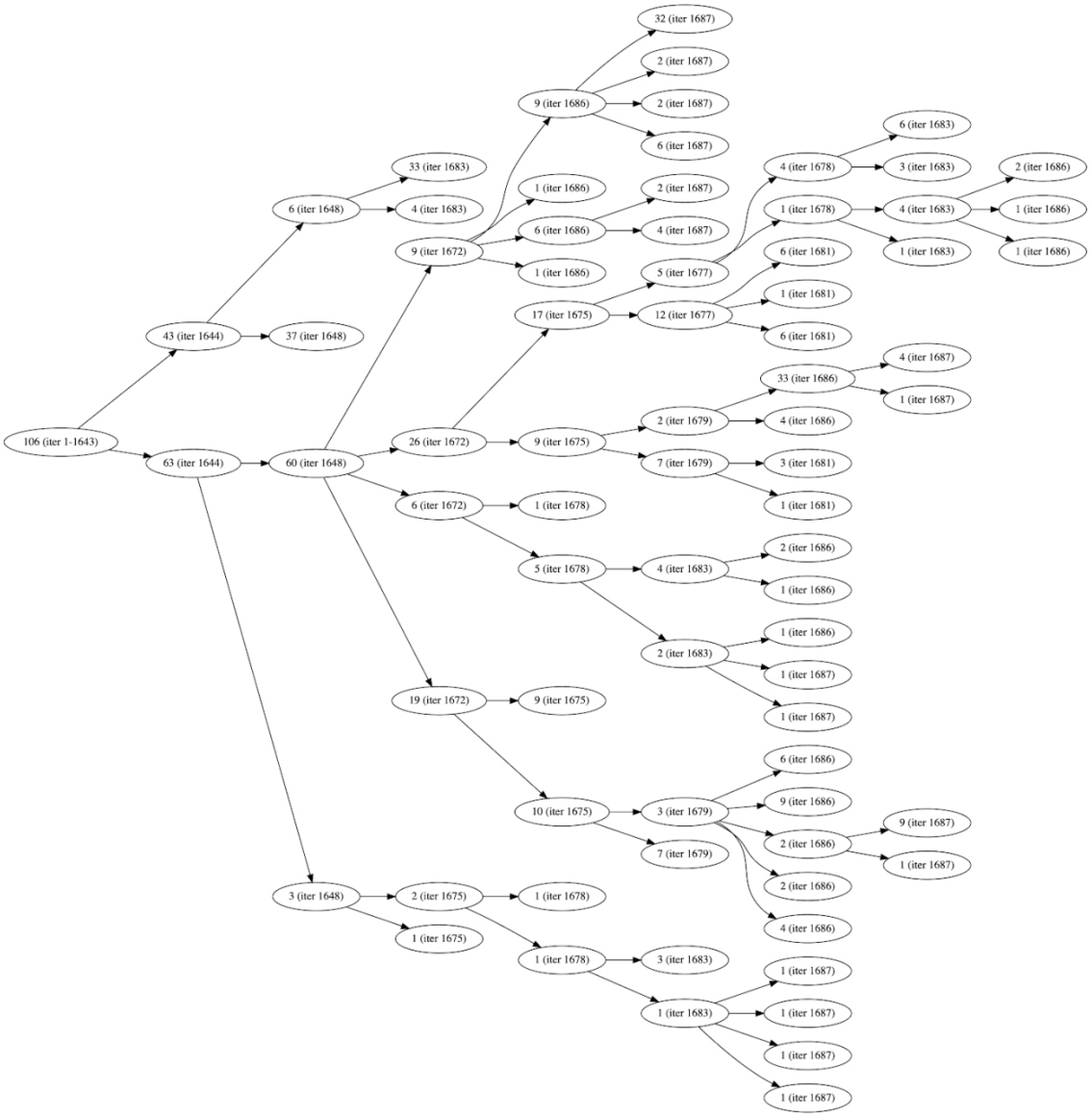
Supplemental Fig. 9 Initial equilibration of a "down"-state structure using a standard MD simulation. Time evolution of (A) Cα RMSD of protein residues , (B) Cα RMSD of structured region of RBD after alignment of core domain to the initial structure and (C) Distance between centers of mass of the RBD and core domain.



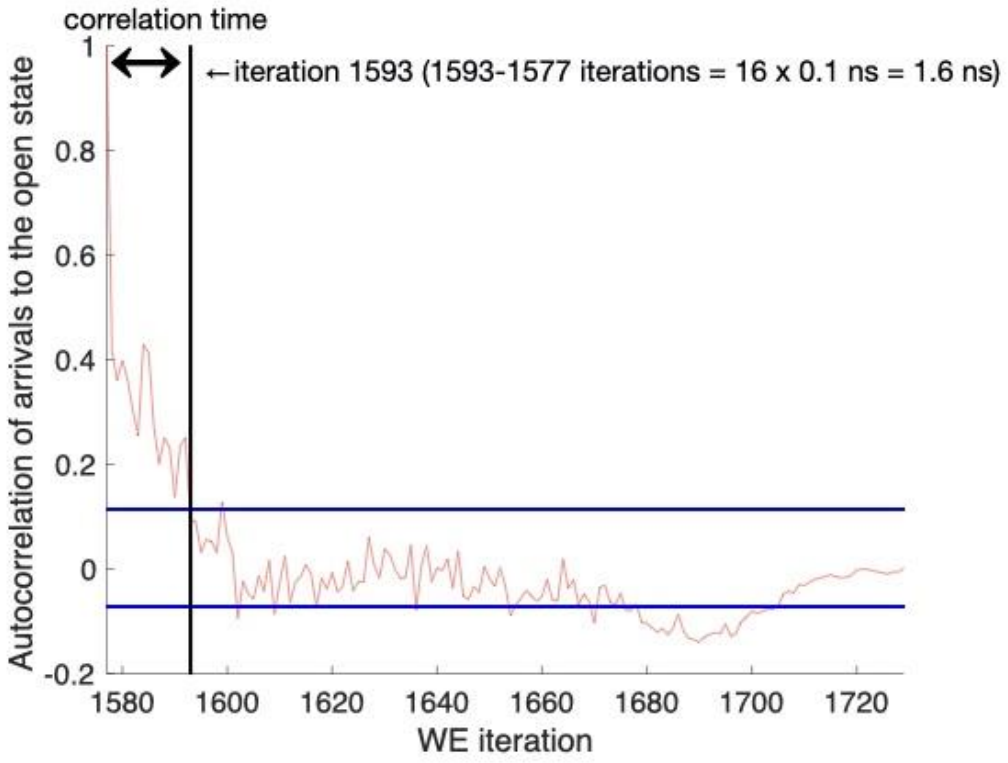
Supplemental Fig. 10 Scaling of the WESTPA software using NVIDIA V100 GPUs on the TACC Longhorn supercomputer vs. theoretical perfectly linear scaling.



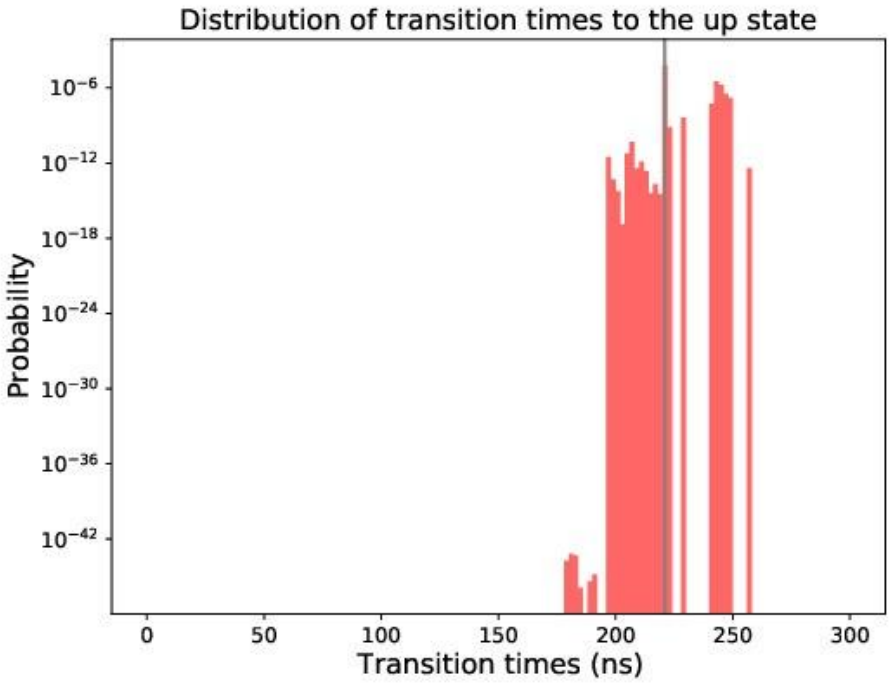
Supplemental Fig. 11 Trajectory splitting tree of the 204 independent pathways that reached the "up" state. The number of each node indicates the number of pathways at the given WE iteration in parentheses. All trajectories shared the same parents until iteration 429, with the first splitting of trajectories occurring at iteration 430. Subsequent splitting occurred at later iterations. Note that the sum of the child pathways does not necessarily match up with the parent's number of pathways due to splitting and merging with other trajectories (not shown).



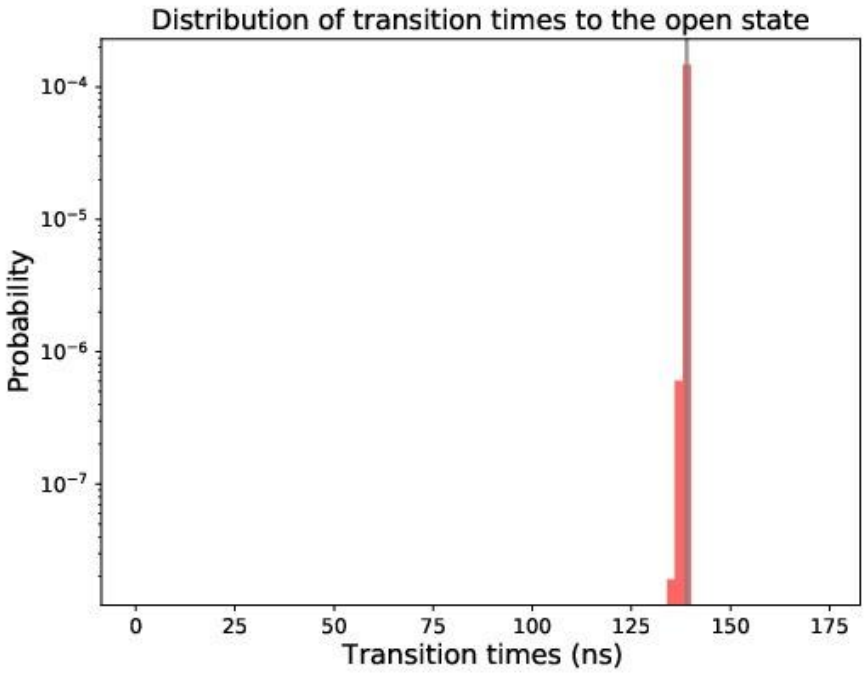
Supplemental Fig. 12 Trajectory splitting tree of the 106 pathways that reached the "open" state. The number of each node indicates the number of pathways at the given WE iteration in parentheses. All trajectories shared the same parents until iteration 1643, the first splitting of trajectories occurring at iteration 1644. Note that the sum of the child pathways does not necessarily match up with the parent's number of pathways at subsequent iterations due to splitting and merging with other trajectories (not shown).



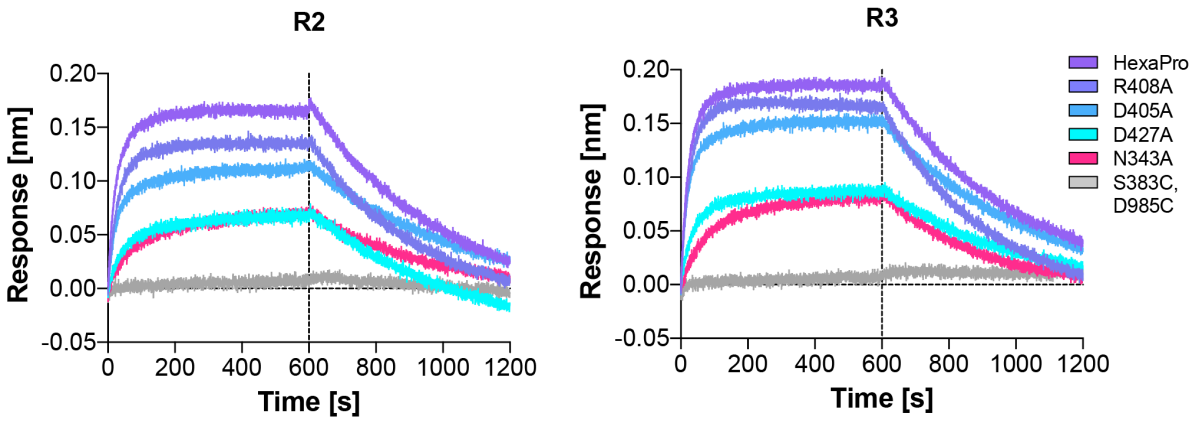
Supplemental Fig. 13 Autocorrelation of arrivals from the "down" state to the "open" state (red) with a 95% confidence interval (blue). The confidence interval was generated using a Monte Carlo bootstrapping strategy where a bootstrap consisted of 1000 randomly drawn datasets (with replacement) from all "down"-to-"open" flux values. The vertical line marks the first point at which the autocorrelation falls within the confidence interval and is used to calculate the correlation time.



Supplemental Fig. 14 Probability distribution of transition times from the "down" state to the "up" state. The most probable transition time is marked in grey. Note that the first 25% of the "fast" transitions are discarded here to calculate the most probable transition time.



Supplemental Fig. 15 Probability distribution of transition times from the "down" state to the "open" state. The most probable transition time is marked in grey. Note that the first 25% of the "fast" transitions are discarded here to calculate the most probable transition time.



Supplemental Fig. 16 BLI sensorgrams of spike variants binding to ACE2 from duplicate (R2) and triplicate (R3) experiments.

3. Supplementary Tables

Supplemental Table 1 Biolayer interferometry data of spike variants binding to ACE2.

VARIANT	HEXAPRO	R408A	D405A	D427A	N343A
R1 - Binding level (nm)	0.1733	0.1560	0.1206	0.0913	0.0783
R2 - Binding level (nm)	0.1776	0.1467	0.1208	0.0793	0.0751
R3 - Binding level (nm)	0.1831	0.1629	0.1506	0.0849	0.0816
Minimum (nm)	0.1733	0.1467	0.1206	0.07932	0.07512
Maximum (nm)	0.1831	0.1629	0.1506	0.09131	0.0816
Range (nm)	0.0098	0.0162	0.03	0.01199	0.00648
Mean (nm)	0.1780	0.1552	0.1307	0.0852	0.0783
Std. Deviation (± nm)	0.0049	0.0081	0.0173	0.0060	0.0032
Response (% to HexaPro)	**100.00**	**87.19**	**73.43**	**47.85**	**44.01**
Response decrease (%)	**0.00**	**12.81**	**26.57**	**52.15**	**55.99**

4. Supplementary Videos

Supplemental Video 1 Continuous pathway of RBD opening. This movie shows one of the continuous, unbiased pathways obtained from the WE simulations. All glycans are shown in blue except the N343 glycan which is colored magenta. Starting from all three RBDs in the "down" conformation, the chain A RBD lifts and twists counterclockwise into the "up" conformation, facilitated through interactions with the two adjacent RBDs, especially the N343 glycan gate on the chain B RBD. Upon reaching the "up" conformation, the RBD continues to twist into an "open" conformation en route to S1 dissociation.

Supplemental Video 2 A comparison of the WE trajectory and ManifoldEM (MEM) CC1 and CC2 for a side view (PD 1386). It can be seen that there is strong agreement between the full WE trajectory and the sequential, piecewise combination of both CCs. Red arrows indicate direction of motion.

Supplemental Video 3 A comparison of the WE trajectory and ManifoldEM (MEM) CC2 for a top-down view (PD 112). A strong agreement can be seen between the outputs of these two frameworks. To note, CC1 was not readily achievable from this view via manifold embedding, since the RBD-"down" to RBD-"up" trajectory from this view is orthogonal to the plane of the projection. Red arrows indicate direction of motion.

Supplemental Video 4 Glycan gate at position N343 intercalates with residues to facilitate RBD opening. This movie zooms in closer to the glycan at position N343 to show how RBD opening is facilitated through intercalation between and underneath the residues F490, Y489, F456, F457 of RBD A. The glycan also transiently interacts with other residues of the RBD which are shown when they are within Å from the glycan.

Supplemental Video 5 Mapping of residue contacts to RBD throughout opening pathway. Distances between residues throughout a continuous opening pathway calculated for the trajectory shown in **Supplemental Videos 1 and 2**. Distances to each residue from RBD_A are shown for each chain in panels A-C and each of the glycans in panel D. Select regions are labeled, and N165, N234, and N343 are labeled with +, ++, +++, respectively.

4. Supplementary References

(31) Huang, J.; MacKerell, A. D. CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. *J Comput Chem* **2013**, *34* (25), 2135–2145. https://doi.org/10.1002/jcc.23354.

(32) Guvench, O.; Hatcher, E. R.; Venable, R. M.; Pastor, R. W.; Mackerell, A. D. CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses. *J Chem Theory Comput* **2009**, *5* (9), 2353–2370. https://doi.org/10.1021/ct900242e.

(33) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(34) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J Mol Graph* **1996**, *14* (1), 33–38, 27–28.

(35) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *Journal of Chemical Theory and Computation* **2012**, *8* (5), 1542–1555. https://doi.org/10.1021/ct200909j.

(36) Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation* **2013**, *9* (9), 3878–3888. https://doi.org/10.1021/ct400314y.

(37) D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden,; R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, Y. Huang,; S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J.; Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R.; Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. SalomonFerrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman. AMBER 2018. *University of California, San Francisco.* **2018**.

(38) Lee, T.-S.; Cerutti, D. S.; Mermelstein, D.; Lin, C.; LeGrand, S.; Giese, T. J.; Roitberg, A.; Case, D. A.; Walker, R. C.; York, D. M. GPU-Accelerated Molecular Dynamics and Free Energy Methods in Amber18: Performance Enhancements and New Features. *J Chem Inf Model* **2018**, *58* (10), 2043–2050. https://doi.org/10.1021/acs.jcim.8b00462.

(39) Crowley, M. F.; Williamson, M. J.; Walker, R. C. CHAMBER: Comprehensive Support for CHARMM Force Fields within the AMBER Software. *International Journal of Quantum Chemistry* **2009**, *109* (15), 3767–3772. https://doi.org/10.1002/qua.22372.

(40) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *Journal of Computational Physics* **1977**, *23* (3), 327–341. https://doi.org/10.1016/0021-9991(77)90098-5.

(41) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092. https://doi.org/10.1063/1.464397.

(42) Zwier, M. C.; Adelman, J. L.; Kaus, J. W.; Pratt, A. J.; Wong, K. F.; Rego, N. B.; Suárez, E.; Lettieri, S.; Wang, D. W.; Grabe, M.; Zuckerman, D. M.; Chong, L. T. WESTPA: An Interoperable, Highly Scalable Software Package for Weighted Ensemble Simulation and Analysis. *J. Chem. Theory Comput.* **2015**, *11* (2), 800–809. https://doi.org/10.1021/ct5010615.

(43) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095. https://doi.org/10.1021/ct400341p.

(44) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **2007**, *9* (3), 90–95. https://doi.org/10.1109/MCSE.2007.55.

(45) R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, D. L. Dotson, J. Domanski, S. Buchoux, I. M. Kenney, and O. Beckstein. MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations. In S. Benthall and S. Rostrup, editors, *Proceedings of the 15th Python in Science Conference*. **2016**, 98-105. Austin, TX, SciPy, doi:10.25080/majora-629e541a-00e.

(46) N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* **2011**, 32 (10), 2319-2327. doi:10.1002/jcc.21787.

(47) Decherchi, S.; Spitaleri, A.; Stone, J.; Rocchia, W. NanoShaper-VMD Interface: Computing and Visualizing Surfaces, Pockets and Channels in Molecular Systems. *Bioinformatics* **2019**, *35* (7), 1241–1243. https://doi.org/10.1093/bioinformatics/bty761.

(48) Dashti, A.; Mashayekhi, G.; Shekhar, M.; Ben Hail, D.; Salah, S.; Schwander, P.; des Georges, A.; Singharoy, A.; Frank, J.; Ourmazd, A. Retrieving Functional Pathways of Biomolecules from Single-Particle Snapshots. *Nat. Commun.* **2020**, *11* (1), 1–14. https://doi.org/10.1038/s41467-020-18403-x.

(49) Scheres, S. H. W. RELION: Implementation of a Bayesian Approach to Cryo-EM Structure Determination. *J. Struct. Biol.* **2012**, *180* (3), 519–530. https://doi.org/10.1016/j.jsb.2012.09.006.

(50) Punjani, A.; Rubinstein, J. L.; Fleet, D. J.; Brubaker, M. A. CryoSPARC: Algorithms for Rapid Unsupervised Cryo-EM Structure Determination. *Nat. Methods* **2017**, *14* (3), 290–296. https://doi.org/10.1038/nmeth.4169.

(51) Punjani, A.; Fleet, D. J. 3D Variability Analysis: Resolving Continuous Flexibility and Discrete Heterogeneity from Single Particle Cryo-EM. *J. Struct. Biol.* **2021**, *213* (2), 107702. https://doi.org/10.1016/j.jsb.2021.107702.

(52) Coifman, R. R.; Lafon, S. Diffusion Maps. *Appl. Comput. Harmon. Anal.* **2006**, *21* (1), 5–30. https://doi.org/10.1016/j.acha.2006.04.006.

(53) Giannakis, D.; Majda, A. J. Nonlinear Laplacian Spectral Analysis for Time Series with Intermittency and Low-Frequency Variability. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (7), 2222–2227. https://doi.org/10.1073/pnas.1118984109.

(54) Huang, C.C., Couch, G.S., Pettersen, E.F., and Ferrin, T.E. "Chimera: An Extensible Molecular Modeling Application Constructed Using Standard Components." *Pacific Symposium on Biocomputing* **1**:724 (1996). http://www.cgl.ucsf.edu/chimera

(55) Liebschner, D.; Afonine, P. V.; Baker, M. L.; Bunkoczi, G.; Chen, V. B.; Croll, T. I.; Hintze, B.; Hung, L. W.; Jain, S.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R. D.; Poon, B. K.; Prisant, M. G.; Read, R. J.; Richardson, J. S.; Richardson, D. C.; Sammito, M. D.; Sobolev, O. V.; Stockwell, D. H.; Terwilliger, T. C.; Urzhumtsev, A. G.; Videau, L. L.;

Williams, C. J.; Adams, P. D. Macromolecular Structure Determination Using X-Rays, Neutrons and Electrons: Recent Developments in Phenix. *Acta Crystallogr. Sect. D Struct. Biol.* **2019**, *75* (Pt 10), 861–877. https://doi.org/10.1107/S2059798319011471.

(56) Tang, G.; Peng, L.; Baldwin, P. R.; Mann, D. S.; Jiang, W.; Rees, I.; Ludtke, S. J. EMAN2: An Extensible Image Processing Suite for Electron Microscopy. *J. Struct. Biol.* **2007**, *157* (1), 38–46. https://doi.org/10.1016/j.jsb.2006.05.009.