nature chemistry

ARTICLES

https://doi.org/10.1038/s41557-021-00707-0

Check for updates

SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome

Maxwell I. Zimmerman[1,2], Justin R. Porter[1,2], Michael D. Ward[1,2], Sukrit Singh[1,2], Neha Vithani[1,2], Artur Meller[1,2], Upasana L. Mallimadugula[1,2], Catherine E. Kuhn[1,2], Jonathan H. Borowsky[1,2], Rafal P. Wiewiora[3,4], Matthew F. D. Hurley[5], Aoife M. Harbison[6], Carl A. Fogarty[6], Joseph E. Coffland[7], Elisa Fadda[6], Vincent A. Voelz[5], John D. Chodera[4] and Gregory R. Bowman[1,2] ✉

SARS-CoV-2 has intricate mechanisms for initiating infection, immune evasion/suppression and replication that depend on the structure and dynamics of its constituent proteins. Many protein structures have been solved, but far less is known about their relevant conformational changes. To address this challenge, over a million citizen scientists banded together through the Folding@home distributed computing project to create the first exascale computer and simulate 0.1 seconds of the viral proteome. Our adaptive sampling simulations predict dramatic opening of the apo spike complex, far beyond that seen experimentally, explaining and predicting the existence of 'cryptic' epitopes. Different spike variants modulate the probabilities of open versus closed structures, balancing receptor binding and immune evasion. We also discover dramatic conformational changes across the proteome, which reveal over 50 'cryptic' pockets that expand targeting options for the design of antivirals. All data and models are freely available online, providing a quantitative structural atlas.

SARS-CoV-2 is a novel coronavirus that poses an imminent threat to global human health and socioeconomic stability[1]. With estimates of the basic reproduction number at ~3–4 and a case fatality rate for COVID-19 ranging from ~0.1 to 12% (high temporal variation), SARS-CoV-2/COVID-19 has spread quickly and currently endangers the global population[2–6]. As of 12 September 2020, there have been over 29 million confirmed cases and over 925,000 fatalities globally. Quarantines and social distancing are effective at slowing the rate of transmission; however, they cause substantial social and economic disruption. Taken together, these facts render it crucial that we find immediate therapeutic interventions.

A structural understanding of the SARS-CoV-2 proteins could accelerate the discovery of new therapeutics by enabling the use of rational design[7]. To this end, the structural biology community has made heroic efforts to rapidly build models of SARS-CoV-2 proteins and the complexes they form[8–16]. However, it is well established that a protein's function is dictated by the full range of conformations it can access, many of which remain hidden to experimental methods. Mapping these conformations for SARS-CoV-2 proteins will provide a clearer picture of how they enable the virus to perform diverse functions such as infecting cells, evading a host's immune system and replicating. Such maps may also present new therapeutic opportunities, such as cryptic pockets that are absent in experimental snapshots but provide novel targets for drug discovery.

Molecular dynamics simulations have the ability to capture the full ensemble of structures a protein adopts but require substantial computational resources. Such simulations capture an all-atom representation of the range of motions a protein undergoes. Modern datasets often consist of a few microseconds of simulation for a single protein, with a few noteworthy examples reaching millisecond timescales[17,18]. However, many important processes occur on slower timescales. Moreover, simulating every protein that is relevant to SARS-CoV-2 for biologically relevant timescales would require computational resources on a massive scale.

To overcome this challenge, more than a million citizen scientists from around the world have donated their computer resources to simulate SARS-CoV-2 proteins. This massive collaboration was enabled by the Folding@home distributed computing platform, which has crossed the exascale computing barrier and is now the world's largest supercomputer. Using this resource, we constructed quantitative maps of the structural ensembles of over two dozen proteins and complexes that pertain to SARS-CoV-2 from milliseconds of simulation data generated for each system. Together, we have run 0.1 s of simulation. Our data uncover the mechanisms of conformational changes that are essential for SARS-CoV-2's replication cycle and reveal a multitude of new therapeutic opportunities. The data are supported by a variety of experimental observations and are being made publicly available (https://covid.molssi.org/ and https://osf.io/fs2yv/), in accordance with open science principles, to accelerate the discovery of new therapeutics[19,20].

To the exascale and beyond

Folding@home (http://foldingathome.org) is a community of citizen scientists, researchers and tech organizations dedicated to

[1]Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St Louis, MO, USA. [2]Center for Science and Engineering of Living Systems (CSELS), Washington University in St. Louis, St Louis, MO, USA. [3]Tri-Institutional PhD Program in Chemical Biology, Memorial Sloan Kettering Cancer Center, NY, New York, USA. [4]Computational and Systems Biology Program, Sloan Kettering Institute, NY, New York, USA. [5]Department of Chemistry, Temple University, Philadelphia, PA, USA. [6]Department of Chemistry and Hamilton Institute, Maynooth University, Maynooth, Ireland. [7]Cauldron Development LLC, Petaluma, CA, USA. ✉e-mail: g.bowman@wustl.edu

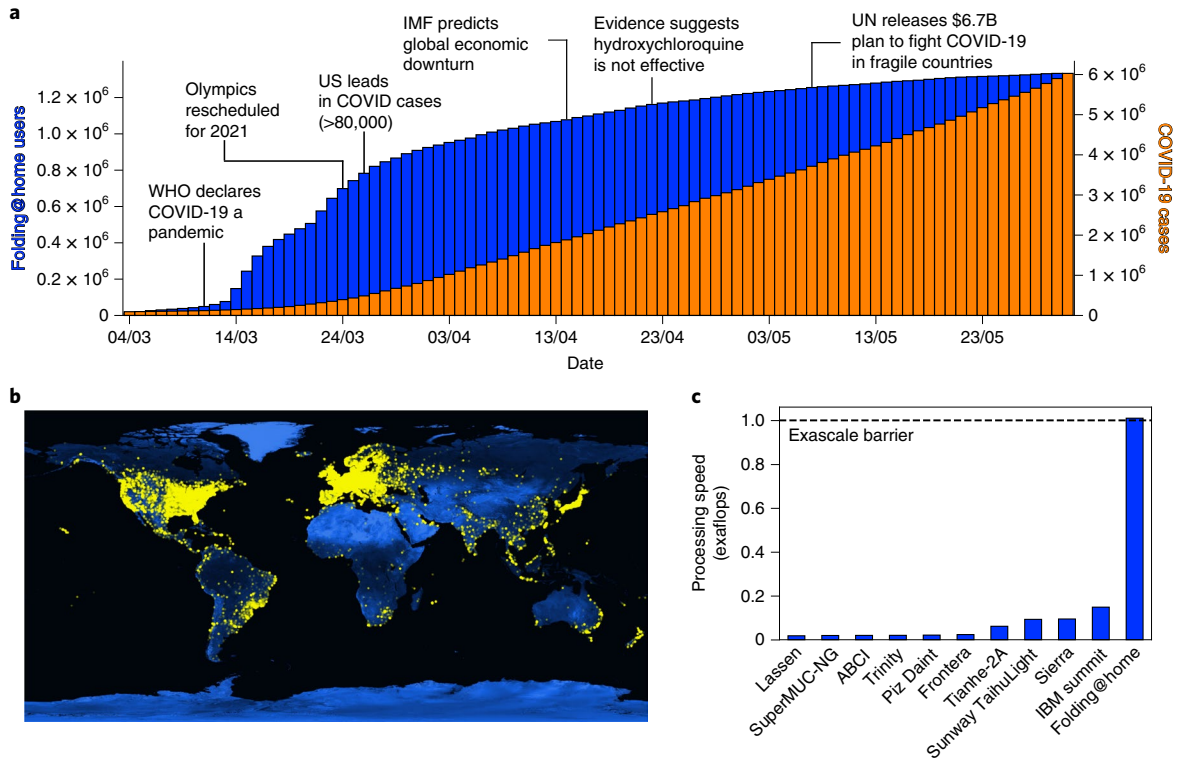

Fig. 1 | Summary of Folding@home's computational power. a, The growth of Folding@home in response to COVID-19. The cumulative number of users is shown in blue and COVID-19 cases are shown in orange. **b**, Global distribution of Folding@home users. Each yellow dot represents a unique IP address contributing to Folding@home. **c**, The processing speed of Folding@home and the next 10 fastest supercomputers in exaflops (one exaflop is 10^{18} floating point operations per second).

applying their collective computational and intellectual resources to understand the role of proteins' dynamics in their function and dysfunction and to aid in the design of new proteins and therapeutics[21]. Folding@home enables anyone with a computer and an Internet connection to contribute to biomedical research by volunteering to run small chunks of simulation called 'work units' that are used to build maps of protein dynamics. The project has provided insight into diverse topics ranging from protein folding to signalling mechanisms[22–24] to the connection between phenotype and genotype[25–27]. Translational applications have included new means to combat antimicrobial resistance, Ebola virus and SFTS virus[28–30].

In response to the COVID-19 pandemic, Folding@home quickly pivoted to focus on SARS-CoV-2 and the host factors it interacts with. Many people found the opportunity to take action alluring at a time when they were otherwise feeling helpless. In less than three months, the project grew from ~30,000 active devices to over a million devices around the globe (Fig. 1a,b).

We conservatively estimate the peak performance of Folding@home reached 1.01 exaflops. This performance was achieved at a point when ~280,000 GPUs and 4.8 million CPU cores were performing simulations. As explained in Methods, to be conservative about our claims, we assume that each GPU/CPU has worse performance than a card released before 2015. For reference, the aggregate 1 exaflop performance we report for Folding@home is fivefold greater than the peak performance of the world's fastest traditional supercomputer at the time, Summit (Fig. 1c). This performance is also greater than the top 100 supercomputers combined. Prior to Folding@home, the first exascale supercomputer was not scheduled to come online until the end of 2021.

Extreme spike opening reveals cryptic epitopes

The spike complex (S) is a prominent vaccine target that is known to undergo substantial conformational changes as part of its function[10,14,31]. Structurally, S is composed of three interlocking proteins, with each chain having a cleavage site separating an S1 and S2 fragment. S resides on the virion surface, where it waits to engage with an angiotensin-converting enzyme 2 (ACE2) receptor on a host cell to trigger infection[32,33]. The fact that S is exposed on the virion surface makes it an appealing vaccine target. However, it has a number of effective defence strategies. First, S is decorated extensively with glycans that aid in immune evasion by shielding potential antigens[34,35]. S also uses a conformational masking strategy, wherein it predominantly adopts a closed conformation (often called the down state) that buries the receptor-binding domains (RBDs) to evade immune surveillance mechanisms. To engage with ACE2, S must somehow expose the conserved binding interface of the RBDs. Characterizing the full range of S opening is important for understanding pathogenesis and could provide insights into novel therapeutic options.

To capture S opening, we employed our goal-oriented adaptive sampling algorithm, FAST, in conjunction with Folding@home. The FAST method[36,37] iterates between running a batch of simulations, building a map of conformational space called a Markov state model (MSM)[38,39] from all the data generated so far, ranking the conformational states of this MSM based on how likely starting a new simulation from that state is to yield useful data and starting a new batch of simulations from the top-ranked states. The ranking function is designed to balance favouring structures with a desired geometric feature (in this case, opening of S) and broad exploration of conformational space. By balancing exploration–exploitation trade-offs, FAST often captures conformational changes with orders of

magnitude less simulation time than alternative methods. Broadly distributed structures from our FAST simulations were then used as starting points for extensive Folding@home simulations, totalling over 1 millisecond of data for SARS-CoV-2 S, enabling us to obtain a statistically sound final model.

Our SARS-CoV-2 S protein simulations predict extreme opening of S and substantial conformational heterogeneity in the open state (Fig. 2). Capturing opening of S is an impressive technical feat. Other large-scale simulations have provided valuable insight into aspects of S but were unable to capture this event, which is essential for the initiation of infection[28,32,33]. For example, Casalino et al.[35] performed ~10 microseconds of simulation to show that one of the glycans helps stabilize a partially open state, and Turoňová et al.[40] performed 2.5 microseconds of simulation that revealed three hinges in the stalk. However, the shorter timescale of these simulations prevented the authors from capturing the opening process at all. With our milliseconds of sampling, we successfully[35,40,41] captured this rare event for both glycosylated and unglycosylated S and found that glycosylation slightly increases the population of the open state, but the difference between glycosylated and unglycosylated S is smaller than that between different spike variants (Supplementary Fig. 1). The closed state is more probable than the open state, explaining the experimental observation that full-length S has a lower affinity for ACE2 than an isolated RBD[42]. Intriguingly, we found that opening occurs only for a single RBD at a time, akin to the up state observed in cryoEM structures[43]. Moreover, we predict that the scale of S opening is often substantially larger than has been observed in experimental snapshots in the absence of binding partners (Supplementary Fig. 2).

The dramatic opening we discovered predicts that antibodies, as well as other therapeutics, can bind to regions of S that are deeply buried and seemingly inaccessible in existing experimental snapshots[9,13,44,45]. Consistent with this prediction, the cryptic epitope for the antibody CR3022 is buried in up and down cryoEM structures but is clearly exposed in our conformational ensemble (Fig. 2c). Indeed, our ensemble captures the exposure of many known epitopes, despite their occlusion in apo experimental snapshots (Fig. 2d). Our models also provide a quantitative estimate of the probability that different epitopes are exposed, are consistent with experimental measures of dynamics and can be used to determine the most suitable regions for the design of neutralizing antibodies.

Our results suggest that S binds ACE2 and many antibodies via a conformational selection mechanism wherein S first opens and then binds to its partners. Previous work based on examining the up and down structures observed by cryoEM also proposed a role for conformational selection, hypothesizing that an S RBD may bind CR3022 by first adopting an up conformation and then twisting to expose the cryptic epitope[8]. To test this hypothesis, we projected the free-energy landscape and the highest-flux pathway for S opening onto two order parameters: the angle of RBD opening and the twist of the RBD (Supplementary Fig. 3). We found that the RBD simultaneously twists and peels off S as it transitions from the closed to open conformation. Furthermore, the motion we observe predicts the exposure of other epitopes that would not be exposed by the mechanism proposed by Yuan et al.[8]. These additional epitopes have now been corroborated by work on the binding sites of other antibodies (Fig. 2d).

To understand the potential role of conformational masking in determining the lethality and infectivity of different coronaviruses, we also simulated the opening of S proteins from two related viruses: SARS-CoV-1 and HCoV-NL63. These viruses were selected because they also bind the ACE2 receptor but are associated with different mortality rates. SARS-CoV-1 caused an outbreak in 2003 with a high case fatality rate but has not become a pandemic[46]. NL63 was discovered the following year and continues to spread around the globe but is substantially less lethal than either SARS virus[47].

We hypothesized that phenotypic differences between coronaviruses may be partially explained by changes to the S conformational ensemble, particularly the probability of spike opening. Specifically, we propose that mutations or other perturbations can increase the S–ACE2 affinity by increasing the probability that S adopts an open conformation or by increasing the affinity between an exposed RBD and ACE2. In contrast, the affinity of S for ACE2 (or antibodies that bind cryptic epitopes) can be reduced by stabilizing the closed state or decreasing the affinity between an exposed RBD and its binding partner(s).

As expected, the propensities of the three S complexes to adopt an open state and bind ACE2 are very different. Structures from each ensemble were classified as competent to bind ACE2 if superimposing an ACE2-RBD structure on S did not result in any steric clashes between ACE2 and the rest of the S complex. We found that SARS-CoV-1 has the highest population of conformations that can bind to ACE2 without steric clashes, followed by SARS-CoV-2, while opening of NL63 is sufficiently rare that we did not observe ACE2-binding-competent conformations in our simulations (Fig. 2b). Interestingly, S proteins that are more likely to adopt structures that are competent to bind ACE2 are also more likely to adopt highly open structures (Fig. 2c).

We also predict a number of interesting correlations between the conformational masking, lethality and infectivity of different coronaviruses. First, more deadly coronaviruses have S proteins with less conformational masking. Second, there is an inverse correlation between S opening and the affinity of an isolated RBD for ACE2 (RBD–ACE2 affinities of ~35 nM, ~44 nM and ~185 nM for HCoV-NL63, SARS-CoV-2 and SARS-CoV-1, respectively)[48,49].

These observations suggest a trade-off wherein stabilizing the closed spike enables immune evasion but hampers cell entry, requiring a higher affinity between an exposed RBD and ACE2 to reliably infect a host cell. We propose that the NL63 S complex is probably best at evading immune detection but is not as infectious as the SARS viruses because strong conformational masking reduces the overall affinity for ACE2. In contrast, the SARS-CoV-1 S complex adopts open conformations more readily but is also more readily detected by immune surveillance mechanisms. Finally, SARS-CoV-2 balances conformational masking and the RBD–ACE2 affinity in a manner that allows it to evade an immune response while maintaining its ability to infect a host cell.

Our atomically detailed model of glycosylated S can facilitate structure-based vaccine antigen design through identification of regions minimally protected by conformational masking or the glycan shield[50]. To identify these potential epitopes, we calculated the probability that each residue in S could be exposed to therapeutics (that is, is not shielded by a glycan or buried by conformational masking), as shown in Fig. 3a. Visualizing these values on the protein reveals a few patches of protein surface that the glycan shielding leaves exposed (Fig. 3b). However, another important factor when targeting an antigen is picking a region with a conserved sequence to yield broader and longer-lasting efficacy. Not surprisingly, many of the exposed regions do not have a strongly conserved sequence. Promisingly, though, we do find a conserved area with a larger degree of solvent exposure (Fig. 3c). This region was recently found to be an effective site for neutralizing antibodies[51]. Another possibility for antigen design is to exploit the opening motion. A number of residues surrounding the receptor-binding motif of the RBD show an increase in exposure by ~30% in ACE2-binding-competent structures (Fig. 3c). These regions are hotspots for neutralizing antibody binding[9,52,53], which is consistent with immunoassays and cryoEM structures.

Cryptic pockets and functional dynamics are present throughout the proteome

Every protein in SARS-CoV-2 remains a potential drug target. So, to understand their roles in disease and help further the design of

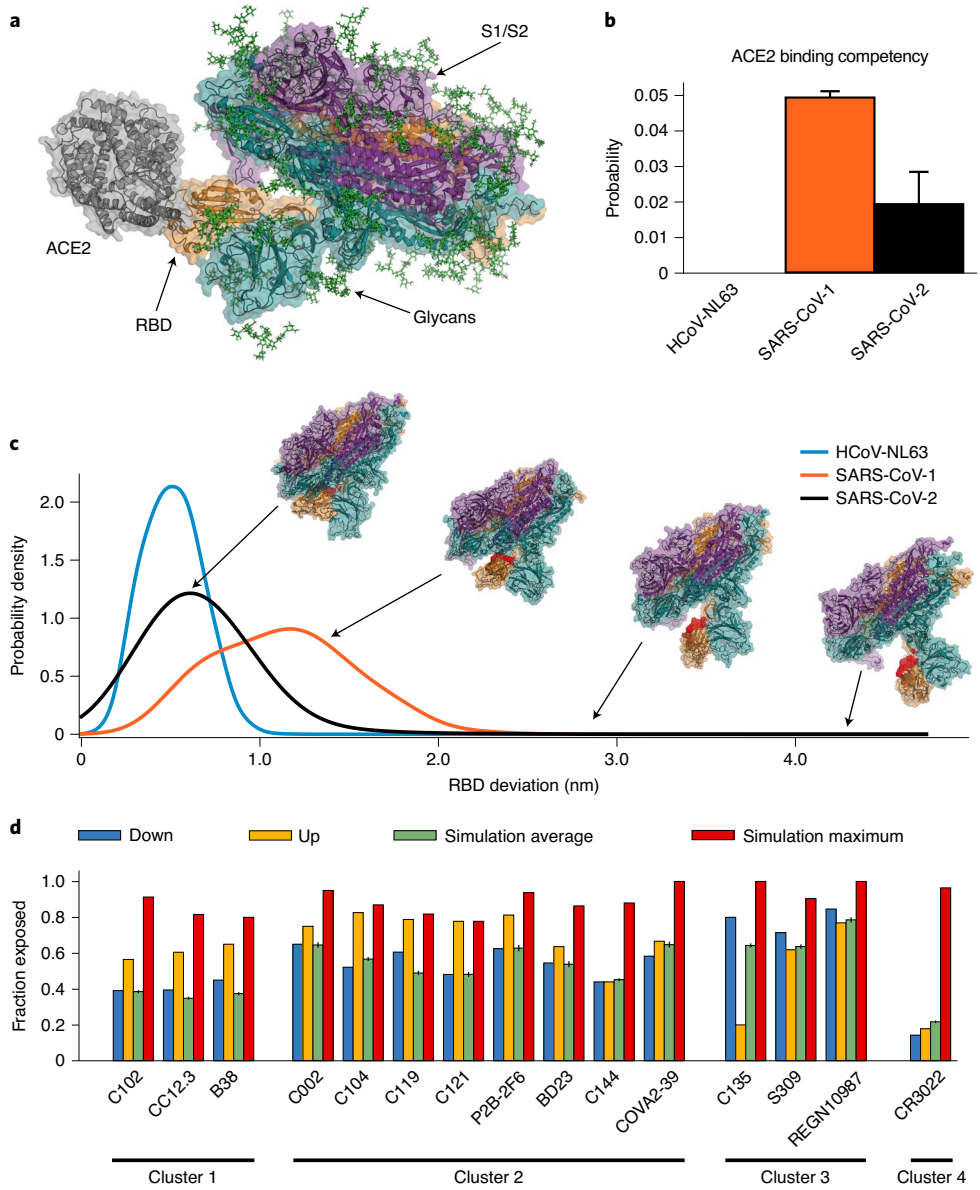

Fig. 2 | Structural characterization of spike opening and conformational masking for three spike homologues. a, An example structure of SARS-CoV-2 spike protein from our simulations that is fully compatible with receptor binding, as shown by superimposing ACE2 (grey). The three chains of the spike protein are illustrated with a cartoon and transparent surface representation (orange, teal and purple), and glycans are shown as sticks (green). **b**, Three spike homologues have very different probabilities of adopting ACE2-binding-competent conformations, likely modulating their affinities for both ACE2 and antibodies that engage the ACE2-binding interface. HCoV-NL63, SARS-CoV-1 and SARS-CoV-2 are shown as light blue, orange and black, respectively. **c**, The probability distribution of the spike opening for each homologue. Opening is quantified in terms of how far the centre of mass of an RBD deviates from its position in the closed (or down) state. The cryptic epitope for the antibody CR3022 (red) is only accessible to antibody binding in extremely open conformations. **d**, Our simulations capture exposure of cryptic epitopes that are buried in the up and down cryoEM structures. The fraction of residues within different epitopes that are exposed to a 0.5 nm radius probe for the down structure (blue), up structure (yellow), the ensemble average from our simulations (green) and the maximum value we observe in our simulations (red). Epitopes are determined as the residues that contact the specified antibody and are clustered by their binding location on the RBD[13].

antivirals, we unleashed the full power of Folding@home to simulate dozens of systems related to pathogenesis. While we are interested in all aspects of a protein's functional dynamics, expanding the number of antiviral targets is of immediate value. To this end, we seeded Folding@home simulations from our FAST-pockets adaptive sampling to aid in the discovery of cryptic pockets. Out of 36 datasets, we briefly discuss 2 illustrative examples.

Non-structural protein number 5 (NSP5, also named the main protease, 3CL^{pro}, or as we will refer to it, M^{pro}) is an essential protein in the life cycle of coronaviruses, cleaving polyprotein 1a into functional proteins, and is a major target for the design of antivirals[11]. It is highly conserved between coronaviruses and shares 96% sequence identity with SARS-CoV-1 M^{pro}. It cleaves polyprotein 1a at no fewer than 11 distinct sites, placing substantial evolutionary constraint on its active site. M^{pro} is only active as a dimer; however, it exists in a monomer–dimer equilibrium with estimates of its dissociation constant falling in the low μM range[54]. Small molecules targeting this protein to inhibit enzymatic activity by either altering its active site or favouring the inactive monomer state would be promising broad-spectrum antiviral candidates[55].

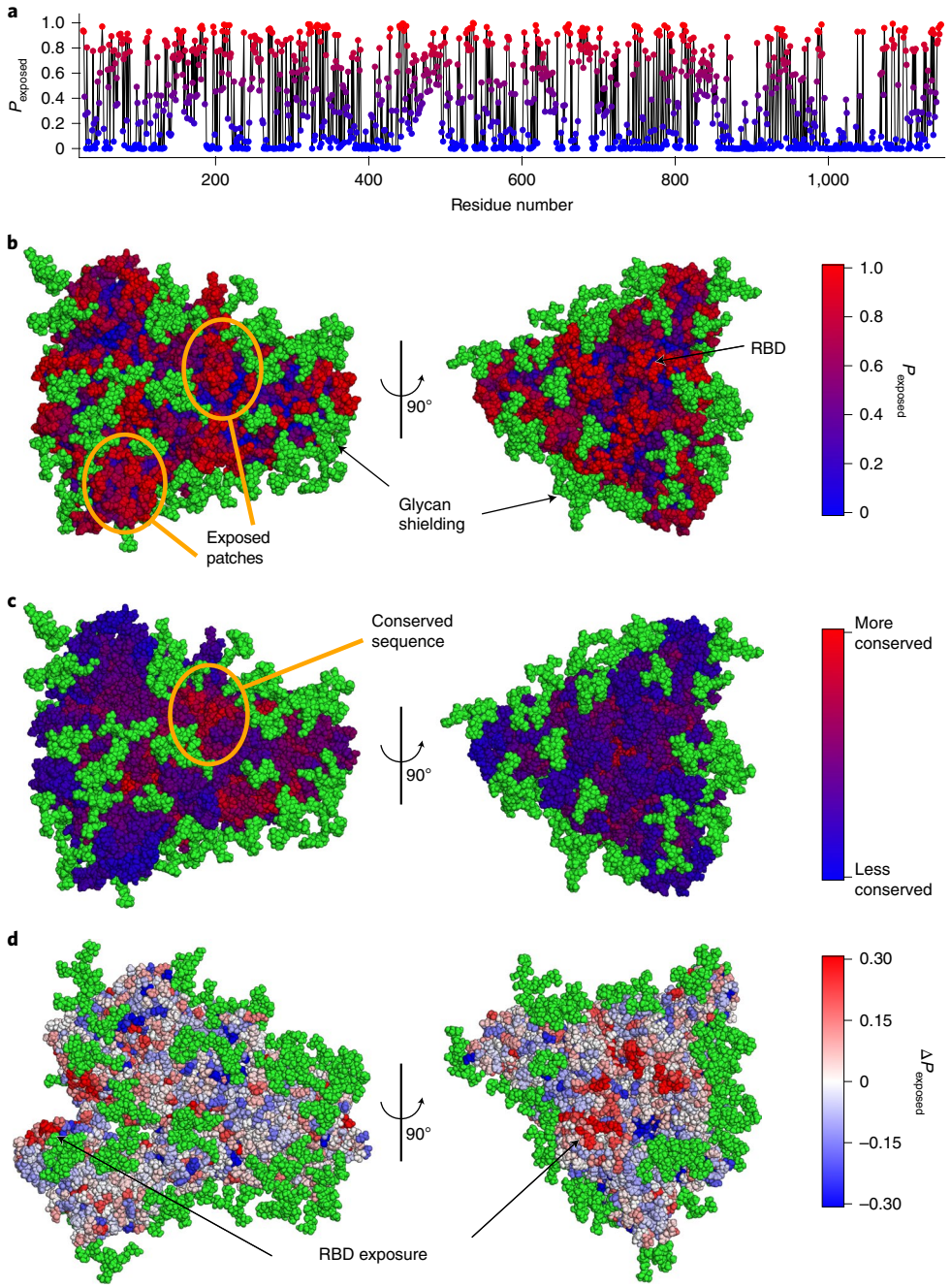

Fig. 3 | Effects of glycan shielding and conformational masking on the accessibility of different parts of the spike to potential therapeutics. a, The probability that a residue is exposed to potential therapeutics ($P_{exposed}$), as determined from our structural ensemble. Red indicates a high probability of being exposed, and blue indicates a low probability of being exposed. **b**, Surface of the spike protein coloured by exposure probability. Exposed patches are circled in orange. Red residues have a higher probability of being exposed, whereas blue residues have a lower probability of being exposed. Atoms belonging to glycans are shown in green. **c**, Spike protein coloured by sequence conservation score. A conserved patch on the protein is circled in orange. Red residues have higher conservation, whereas blue residues have lower conservation. **d**, The difference in the probability that each residue is exposed between the ACE2-binding-competent conformations and the entire ensemble ($\Delta P_{exposed}$). Red residues have a higher probability of being exposed upon opening, whereas blue residues have a lower probability of being exposed. Exposure data can be found online at https://osf.io/fs2yv/ under the SARS-CoV-2 spike project in the analysis folder.

Our simulations predict two novel cryptic pockets on M^{pro} that expand our current therapeutic options. These are shown in Fig. 4a, which projects states from our MSM onto the solvent exposure of residues that make up the pockets. The first cryptic pocket is an expansion of NSP5's catalytic site. We predict that the loop bridging domains II and III is highly dynamic and can fully undock from the rest of the protein. This motion may impact catalysis—for example, by sterically regulating substrate binding—and is similar to motions we have observed previously for the enzyme β-lactamase[56]. Owing to the pocket's location, a small molecule bound in this pocket is likely to prevent catalysis by obstructing polypeptide association with catalytic residues. The second pocket is a large opening between domains I/II and domain III. Located at the dimerization interface, this pocket offers the possibility of finding small molecules or

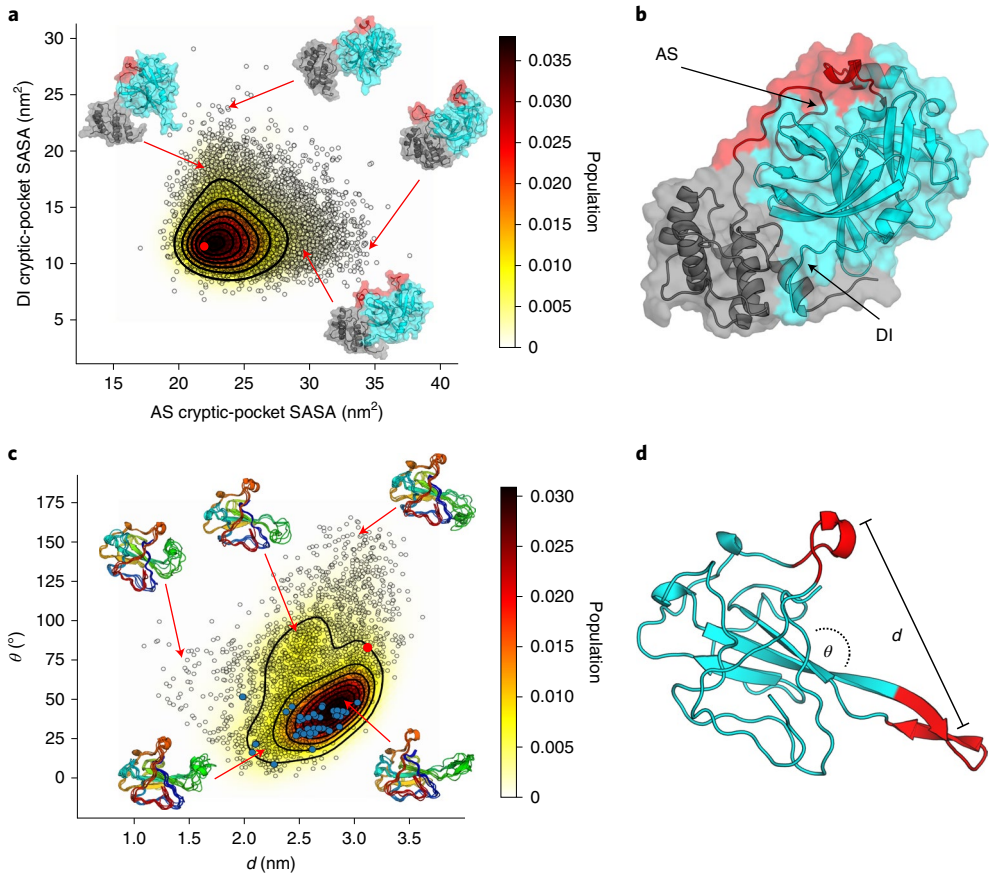

Fig. 4 | Examples of cryptic pockets and functionally relevant dynamics. a,**b**, Conformational ensemble of M^{pro} (monomeric) predicts cryptic pockets near the active site (AS) and dimerization interface (DI). Conformational states (black circles) are projected onto the solvent-accessible surface areas (SASAs) of residues surrounding either the active site or the dimerization interface. The starting structure for simulations (6Y2E) is shown as a red dot. Representative structures are depicted by cartoons and transparent surfaces. Domains I and II are coloured cyan and domain III is coloured grey. The loop of domain III, which covers the active-site residues and is seen to be highly dynamic, is coloured red. **c**,**d**, The conformational ensemble from our simulations of nucleoproteins is similar to the distribution of structures seen experimentally. Conformational states are projected onto the distance and angle between the positive finger and a nearby loop. Angles θ were calculated between vectors that point along each red segment in **d**, and distances d were calculated between their centres of mass. Cluster centres are represented as black circles, the starting structure for simulations (6VYO) is shown as a red dot and NMR structures are shown as solid blue dots. Representative structures are shown as cartoons.

peptides that favour the inactive monomer state. We repeated these calculations and found that the discovery of cryptic pockets is robust to the choice of force field (Supplementary Fig. 4).

In addition to cryptic pockets, our data capture many potentially functionally relevant motions within the SARS-CoV-2 proteome. We illustrate this with the SARS-CoV-2 nucleoprotein. The nucleoprotein is a multifunctional protein responsible for major life-cycle events such as viral packaging, transcription and physically linking RNA to the envelope[57,58]. As such, we expect the protein to accomplish these goals through a highly dynamic and rich conformational ensemble, akin to context-dependent regulatory modules observed in Ebola virus nucleoprotein[59,60]. Investigating the RNA-binding domain, we predict both cryptic pockets and an incredibly dynamic beta-hairpin, referred to as a 'positive finger', that hosts the RNA-binding site (Fig. 4c,d). The conformational heterogeneity of the positive finger we observe is consistent with a structural ensemble determined using solution-state NMR spectroscopy[61]. Our simulations also capture numerous states of the putative RNA-binding pose, where the positive finger curls up to form a cradle for RNA. These states can provide a structural basis for the design of small molecules that would compete with RNA binding, preventing viral assembly.

The data we present in this paper represent the single largest collection of all-atom simulations. Table 1 is a comprehensive list of the systems we have simulated. Systems span various oligomerization states, include important complexes and include representation from multiple coronaviruses. We also include human proteins that are targets for supportive therapies and preventative treatments. To accelerate the discovery of new therapeutics and promote open science, our MSMs and structures of cryptic pockets are available online (https://covid.molssi.org/ and https://osf.io/fs2yv/). For each system analysed, we provide a detailed Markov model and relevant analysis. For cryptic pockets, we provide two directories, 'model' and 'cryptic_pockets', as well as a README.dat that details all hyperparameters used for model construction. The model directory contains the following files: full_centers.xtc (GROMACS binary of cluster centres), populations.npy (numpy binary file of equilibrium populations), prot_masses.pdb (PDB topology file), tcounts.npy (numpy binary of the transition count matrix) and tprobs.npy (numpy binary of the transition probability matrix). For each cryptic pocket 'X' that we characterize, there exist cryptic_pockets/pocketX_resis.dat and cryptic_pockets/pocketX_rankings.dat files, which detail the residues that are present in the cryptic pocket and present a list of states with

Table 1 | Summary of protein systems we have simulated on Folding@home, organized by viral strain

System name	Oligomerization	Initial structure	Residues	Atoms in system	Aggregate simulation time (μs)	Cryptic pockets discovered
SARS-CoV-2						
NSP3 (macrodomain 'X')	Monomer	6W02	167	23,907	10,906	—
NSP3 (papain-like protease 2, $PL2^{pro}$)	Monomer	3E9S[a]	306	97,285	731	2
NSP5 (main protease, M^{pro}, $3CL^{pro}$)	Monomer	6Y2E	306	64,791	6,405	2
NSP5 (main protease, M^{pro}, $3CL^{pro}$)	Dimer	6Y2E	612	77,331	2,902	2
NSP7	Monomer	5F22[a]	79	20,094	3,722	3
NSP8	Monomer	2AHM[a]	191	156,282	1,776	3
NSP9	Dimer	6W4B[b]	226	49,885	8,939	2
NSP10	Monomer	6W4H[b]	131	29,560	6,141	2
NSP12 (polymerase)	Monomer	6NUR[a]	891	186,622	3,330	3
NSP13 (helicase)	Monomer	6JYT[a]	596	129,368	3,407	3
NSP14	Monomer	5C8S[a]	527	216,380	2,384	2
NSP15	Monomer	6VWW	347	67,345	3,674	4
NSP15	Hexamer	6VWW	2,082	230,339	4,270	—
NSP16	Monomer	6W4H[b]	298	45,672	2,382	5
Nucleoprotein (RBD)	Monomer	6VYO	173	29,125	9,493	3
Nucleoprotein dimerization domain	Monomer	6YUN[b]	118	34,905	6,782	—
Nucleoprotein dimerization domain	Dimer	6YUN[b]	236	72,733	1,458	2
Spike	Trimer	6VXX[c]	3,363	442,881	1,109	—
NSP7/NSP8/NSP12	Trimer complex	6NUR[a]	1,184	215,694	1,686	—
NSP10/NSP14	Dimer complex	5C8S[a]	688	226,672	689	3
NSP10/NSP16	Dimer complex	6W4H[b]	429	63,752	3,463	2
SARS-CoV-1						
NSP3 (macrodomain 'X')	Monomer	2FAV	172	33,117	659	—
NSP9	Dimer	1QZ8[b]	226	49,599	7,763	—
NSP15	Monomer	2H85	345	67,345	4,734	—
NSP15	Hexamer	2H85	2,070	230,339	1,130	—
Nucleoprotein RBD	Monomer	2OFZ	174	29,125	4,088	—
Nucleoprotein dimerization domain	Monomer	2GIB	370	34,905	1,626	—
Nucleoprotein dimerization domain	Dimer	2GIB	740	72,733	4,221	—
Spike	Trimer	5X58[c]	3,261	375,851	741	—
NSP10/NSP16	Dimer complex	6W4H[a]	425	69,589	518	—
Human						
IL6	Monomer	1ALU	166	26,855	1,593	2
IL6-R	Monomer	1N26	299	149,764	196	5
ACE2	Monomer	6LZG	596	75,787	664	2
MERS						
NSP13	Monomer	5WWP	596	121,134	719	—
NSP10/NSP16	Dimer complex	6W4H[a]	424	69,127	518	—
HCoV-NL63						
Spike	Trimer	5SZS[c]	3,606	453,348	651	—

[a]Structural model was generated from a homologous sequence using Swiss model[65]. [b]Missing residues were modelled using Swiss model[64]. [c]Missing residues were modelled using CHARMM-GUI[66,67].

cryptic pockets ranked from most open to most closed. Other contemporary works are already building on these data, providing new insight into multiple systems (for example, NSP16, spike protein and nucleoprotein) and making new connections with experiments[59,62,63].

Discussion

To tackle a global threat, the Folding@home community has created one of the largest computational resources in the world. Over a million citizen scientists have pooled their computer resources to help understand and combat COVID-19, generating more than

0.1 seconds of simulation data. The colossal scale of these simulations has helped to characterize crucial stages of infection. We predict that spike proteins have a strong trade-off between making ACE2-binding interfaces accessible to infiltrate cells and conformationally masking epitopes to subvert immune responses. SARS-CoV-2 represents a more optimal trade-off than related coronaviruses, which may explain its success in spreading globally. Our simulations also provide an atomically detailed roadmap for designing vaccines and antivirals. For example, we have made a comprehensive atlas and repository of cryptic pockets hosted online to accelerate the development of novel therapeutics. Many groups are already using our data, including the COVID Moonshot[64], an international collaboration between multiple computational and experimental groups working to develop a patent-free inhibitor of the main protease.

Beyond SARS-CoV-2, we expect this work to aid in a better understanding of the roles of proteins in the Coronaviridae family. Coronaviruses are not new—indeed, they have been around for millennia—yet many of their proteins are still poorly understood. Because climate change has made zoonotic transmission events more commonplace, it is imperative that we continue to perform basic research on these viruses to better protect us from future pandemics. For each protein system in Table 1, an extraordinary amount of sampling has led to the generation of a quantitative map of its conformational landscape. There is still much to learn about coronavirus function, and these conformational ensembles contain a wealth of information to pull from.

While we have aggressively targeted research on SARS-CoV-2, Folding@home is a general platform for running molecular dynamics simulations at scale. Before the COVID-19 pandemic, Folding@home was already generating datasets that were orders of magnitude greater than some of those generated by conventional means. With our explosive growth, our compute power has increased by around 100-fold. Our work here highlights the incredible utility this compute power has to enable rapid understanding of health and disease, providing a rich source of structural data for accelerating the design of therapeutics. With the continued support of the citizen scientists that have made this work possible, we have the opportunity to make a profound impact on other global health crises such as cancer, neurodegenerative diseases and antibiotic resistance.

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/s41557-021-00707-0.

Received: 6 October 2020; Accepted: 14 April 2021;
Published online: 24 May 2021

References

1. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
2. Liu, Y., Gayle, A. A., Wilder-Smith, A. & Rocklöv, J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J. Travel Med.* **27**, taaa021 (2020).
3. Sorci, G., Faivre, B. & Morand, S. Why does COVID-19 case fatality rate vary among countries? *SSRN Electron. J.* https://doi.org/10.2139/ssrn.3576892 (2020).
4. Khafaie, F. R. M. A. Cross-country comparison of case fatality rates of COVID-19/SARS-CoV-2. *Osong Public Health Res. Perspect.* **11**, 74–80 (2020).
5. Mahase, E. Coronavirus: COVID-19 has killed more people than SARS and MERS combined, despite lower case fatality rate. *Br. Med. J.* **368**, m641 (2020).
6. Onder, G., Rezza, G. & Brusaferro, S. Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. *JAMA* **323**, 1775–1776 (2020).
7. Ferreira, L. G., Santos, R. N. D., Oliva, G. & Andricopulo, A. D. Molecular docking and structure-based drug design strategies. *Molecules* **20**, 13384–13421 (2015).
8. Yuan, M. et al. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science* **368**, 630–633 (2020).
9. Zhou, T. et al. A pH-dependent switch mediates conformational masking of SARS-CoV-2 spike. Preprint at *bioRxiv* https://doi.org/10.1101/2020.07.04.187989 (2020).
10. Wrapp, D. et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).
11. Zhang, L. et al. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* **368**, 409–412 (2020).
12. Lu, M. et al. Real-time conformational dynamics of SARS-CoV-2 spikes on virus particles. *Cell Host Microbe* **28**, 880–891 (2020).
13. Barnes, C. O. et al. Structural classification of neutralizing antibodies against the SARS-CoV-2 spike receptor-binding domain suggests vaccine and therapeutic strategies. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.30.273920 (2020).
14. Walls, A. C. et al. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **181**, 281–292.e6 (2020).
15. 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).
16. Cai, Y. et al. Distinct conformational states of SARS-CoV-2 spike protein. *Science* **369**, 1586–1592 (2020).
17. Voelz, V. A., Bowman, G. R., Beauchamp, K. & Pande, V. S. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1−39). *J. Am. Chem. Soc.* **132**, 1526–1528 (2010).
18. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
19. Stodden, V. Enabling reproducible research: open licensing for scientific innovation. *Int. J. Commun. Law Policy* **13** (2009).
20. 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).
21. Shirts, M. & Pande, V. S. COMPUTING: screen savers of the world unite! *Science* **290**, 1903–1904 (2000).
22. Kohlhoff, K. J. et al. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat. Chem.* **6**, 15–21 (2014).
23. Shukla, D., Meng, Y., Roux, B. & Pande, V. S. Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nat. Commun.* **5**, 3397 (2014).
24. Sun, X., Singh, S., Blumer, K. J. & Bowman, G. R. Simulation of spontaneous G protein activation reveals a new intermediate driving GDP unbinding. *eLife* **7**, e38465 (2018).
25. Hart, K. M., Ho, C. M. W., Dutta, S., Gross, M. L. & Bowman, G. R. Modelling proteins' hidden conformations to predict antibiotic resistance. *Nat. Commun.* **7**, 12965 (2016).
26. Chen, S. et al. The dynamic conformational landscape of the protein methyltransferase SETD8. *eLife* **8**, e45403 (2019).
27. Porter, J. R., Meller, A., Zimmerman, M. I., Greenberg, M. J. & Bowman, G. R. Conformational distributions of isolated myosin motor domains encode their mechanochemical properties. *eLife* **9**, e55132 (2020).
28. Hart, K. M. et al. Designing small molecules to target cryptic pockets yields both positive and negative allosteric modulators. *PLoS ONE* **12**, e0178678 (2017).
29. Cruz, M. A. et al. Discovery of a cryptic allosteric site in Ebola's 'undruggable' VP35 protein using simulations and experiments. Preprint at *bioRxiv* https://doi.org/10.1101/2020.02.09.940510 (2020).
30. Wang, W. et al. The cap-snatching SFTSV endonuclease domain is an antiviral target. *Cell Rep.* **30**, 153–163.e5 (2020).
31. Kirchdoerfer, R. N. et al. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Sci. Rep.* **8**, 15701 (2018).
32. Zhang, H., Penninger, J. M., Li, Y., Zhong, N. & Slutsky, A. S. Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Med.* **46**, 586–590 (2020).
33. Hoffmann, M. et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**, 271–280.e8 (2020).
34. Watanabe, Y. et al. Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nat. Commun.* **11**, 2688 (2020).
35. Casalino, L. et al. Beyond shielding: the roles of glycans in the SARS-CoV-2 spike protein. *ACS Cent. Sci.* **6**, 1722–1734 (2020).
36. Zimmerman, M. I. & Bowman, G. R. FAST conformational searches by balancing exploration/exploitation trade-offs. *J. Chem. Theory Comput.* **11**, 5747–5757 (2015).

37. Zimmerman, M. I. & Bowman, G. R. How to run FAST Simulations. *Methods Enzymol.* **578**, 213–225 (2016).
38. Pande, V. S., Beauchamp, K. & Bowman, G. R. Everything you wanted to know about Markov state models but were afraid to ask. *Methods* **52**, 99–105 (2010).
39. Wang, X., Unarta, I. C., Cheung, P. P.-H. & Huang, X. Elucidating molecular mechanisms of functional conformational changes of proteins via Markov state models. *Curr. Opin. Struct. Biol.* **67**, 69–77 (2021).
40. Turoňová, B. et al. In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges. *Science* **370**, 203–208 (2020).
41. Sikora, M. et al. Computational epitope map of SARS-CoV-2 spike protein. *PLoS Comput. Biol.* **17**, e1008790 (2021).
42. Shang, J. et al. Cell entry mechanisms of SARS-CoV-2. *Proc. Natl Acad. Sci. USA* **117**, 11727–11734 (2020).
43. Yuan, Y. et al. Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nat. Commun.* **8**, 15092 (2017).
44. Guo, L. et al. Engineered trimeric ACE2 binds viral spike protein and locks it in "three-up" conformation to potently inhibit SARS-CoV-2 infection. *Cell Res.* **31**, 98–100 (2021).
45. Huo, J. et al. Neutralization of SARS-CoV-2 by destruction of the prefusion spike. *SSRN Electron. J.* https://doi.org/10.2139/ssrn.3613273 (2020).
46. Zhong, N. S. et al. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *Lancet* **362**, 1353–1358 (2003).
47. van der Hoek, L. et al. Identification of a new human coronavirus. *Nat. Med.* **10**, 368–373 (2004).
48. Wu, K., Li, W., Peng, G. & Li, F. Crystal structure of NL63 respiratory coronavirus receptor-binding domain complexed with its human receptor. *Proc. Natl Acad. Sci. USA* **106**, 19970–19974 (2009).
49. Shang, J. et al. Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).
50. Graham, B. S., Gilman, S. A. & McLellan, J. S. Structure-based vaccine antigen design. *Annu. Rev. Med.* **70**, 91–104 (2019).
51. Li, Y. et al. Linear epitopes of SARS-CoV-2 spike protein elicit neutralizing antibodies in COVID-19 patients. *Cell Mol. Immunol.* **17**, 1095–1097 (2020).
52. Brouwer, P. J. M. et al. Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. *Science* **38**, eabc5902 (2020).
53. Hansen, J. et al. Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. *Science* **369**, 1010–1014 (2020).
54. Graziano, V., McGrath, W. J., Yang, L. & Mangel, W. F. SARS CoV main proteinase: the monomer–dimer equilibrium dissociation constant. *Biochemistry* **45**, 14632–14641 (2006).
55. Goyal, B. & Goyal, D. Targeting the dimerization of the main protease of coronaviruses: a potential broad-spectrum therapeutic strategy. *ACS Comb. Sci.* **22**, 297–305 (2020).
56. Porter, J. R. et al. Cooperative changes in solvent exposure identify cryptic pockets, switches, and allosteric coupling. *Biophys. J.* **116**, 818–830 (2019).
57. McBride, R., Zyl, M. V. & Fielding, B. C. The coronavirus nucleocapsid is a multifunctional protein. *Viruses* **6**, 2991–3018 (2014).
58. Masters, P. S. Coronavirus genomic RNA packaging. *Virology* **537**, 198–207 (2019).
59. Cubuk, J. et al. The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *Nat. Commun.* **12**, 1936 (2021).
60. Su, Z. et al. Electron cryo-microscopy structure of ebola virus nucleoprotein reveals a mechanism for nucleocapsid-like assembly. *Cell* **172**, 966–978.e12 (2018).
61. Dinesh, D. C., Chalupska, D., Silhan, J., Veverka, V. & Boura, E. Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein. *PLoS Pathog.* **16**, e1009100 (2020).
62. Kucherova, A., Strango, S., Sukenik, S. & Theillard, M. Modeling the opening SARS-CoV-2 spike: an investigation of its dynamic electro-geometric properties. Preprint at *bioRxiv* https://doi.org/10.1101/2020.10.29.361261 (2020).
63. Vithani, N. et al. SARS-CoV-2 Nsp16 activation mechanism and a cryptic pocket with pan-coronavirus antiviral potential. *Biophys. J.* https://doi.org/10.1016/j.bpj.2021.03.024 (2021).
64. Chodera, J., Lee, A. A., London, N. & von Delft, F. Crowdsourcing drug discovery for pandemics. *Nat. Chem.* **12**, 581–581 (2020).
65. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
66. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).
67. Lee, J. et al. CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *J. Chem. Theory Comput.* **12**, 405–413 (2016).

Methods

System preparation. All simulations were prepared using Gromacs 2020 (ref. [68]). Initial structures were placed in a dodecahedral box that extends 1.0 nm beyond the protein in any dimension. Systems were then solvated and energy was minimized with a steepest descents algorithm until the maximum force fell below 100 kJ mol^{-1} nm^{-1}, using a step size of 0.01 nm and a cut-off distance of 1.2 nm for the neighbour list, Coulomb interactions and van der Waals interactions. The AMBER03 force field was used for all systems except spike protein with glycans, which used CHARMM36 (refs. [69,70]). We chose to use the AMBER03 force field for the discovery of cryptic pockets since we have had extensive success experimentally confirming predictions based on simulations using this force field on other systems[71]. We have also found that AMBER03 gives comparable results to other force fields, given sufficient sampling[72]. Furthermore, we found that discovery of cryptic pockets on NSP5 is robust to the choice of force field (Supplementary Fig. 4). All simulations were performed with explicit TIP3P solvent[73].

Systems were then equilibrated for 1.0 ns, where all bonds were constrained with the LINCS algorithm and virtual sites were used to allow a 4 fs time step[74]. Cut-offs of 1.1 nm were used for the neighbour list with 0.9 for Coulomb and van der Waals interactions. The particle mesh Ewald method was employed for treatment of long-range interactions with a Fourier spacing of 0.12 nm. The Verlet cut-off scheme was used for the neighbour list. The stochastic velocity rescaling (v-rescale) thermostat was used to hold the temperature at 300 K (ref. [75]).

Adaptive sampling simulations. The FAST algorithm was employed for each protein in Table 1 to enhance conformational sampling and quickly explore dominant motions. The procedure for FAST simulations is as follows: (1) run initial simulations, (2) build MSM, (3) rank states based on FAST ranking, (4) restart simulations from the top ranked states and (5) repeat steps 2–4 until ranking is optimized. For each system, MSMs were generated after each round of sampling using a *k*-centres clustering algorithm based on the root-mean-square deviation between select atoms. Clustering continued until the maximum distance of a frame to a cluster centre fell within a predefined cut-off. In addition to the FAST ranking, a similarity penalty was added to promote conformational diversity in starting structures, as has been described previously[76]. The code used to run FAST simulations can be found online (https://github.com/bowman-lab/fast).

FAST-distance simulations of all spike proteins were run at 310 K on the Microsoft Azure cloud computing platform. The FAST-distance ranking favoured states with greater RBD openings using a set of distances between atoms. Each round of sampling was performed with 22 independent simulations that were 40 ns in length (0.88 µs aggregate sampling per round), where the number of rounds totalled 13 (11.44 µs), 22 (19.36 µs) and 17 (14.96 µs) for SARS-CoV-1, SARS-CoV-2 and HCoV-NL63, respectively.

For all other proteins, FAST-pocket simulations were run at 300 K for six rounds, with 10 simulations per round, where each simulation was 40 ns in length (2.4 µs aggregate simulation). The FAST-pocket ranking function favoured restarting simulations from states with large pocket openings. Pocket volumes were calculated using the LIGSITE algorithm[77].

Folding@home simulations. For each adaptive sampling run, a conformationally diverse set of structures was selected to be run on Folding@home. Structures came from the final *k*-centres clustering of adaptive sampling, as described above. Simulations were deployed using a simulation core based on either GROMACS 5.0.4 or OpenMM 7.4.1 (refs. [68,78]).

Estimating the aggregate compute power of Folding@home is non-trivial due to factors like hardware heterogeneity, measures to maintain volunteers' anonymity and the fact that volunteers can turn their machines on and off at will. Furthermore, volunteers' machines only communicate with the Folding@home servers at the beginning and end of a work unit, with the intervening time taking anywhere from tens of minutes to a few days depending on the volunteer's hardware and the protein being simulated. Therefore, we chose to estimate the performance by counting the number of GPUs and CPUs that participated in Folding@home during a three-day window and making a conservative assumption about the computational performance of each device. We note that a larger time window has been used on our website for historical reasons. We make the conservative assumption that each CPU core performs at 0.0127 TFLOPS and each GPU at 1.672 native TFLOPS (or 3.53 x86-equivalent TFLOPS), as explained in our long-standing performance estimate (https://stats.foldingathome.org/os). For reference, a GTX 980 (which was released in 2014) can achieve 5 native TFLOPS (or 10.56 x86-equivalent TFLOPS). An Intel Core i7 4770K (released in 2013) can achieve 0.046 TFLOPS per core. We report x86-equivalent FLOPS.

MSMs. An MSM is a network representation of a free-energy landscape and is a key tool for making sense of molecular dynamics simulations[39,79]. All MSMs were built using our python package, enspara (https://github.com/bowman-lab/enspara)[80]. Each system was clustered with the combined FAST and Folding@home datasets. In the case of spike proteins, states were defined geometrically based on the root-mean-square deviation between backbone C-α coordinates. States were generated as the top 3,000 centres from a *k*-centres clustering algorithm. All other proteins were clustered based on the Euclidean distance between the solvent-accessible surface area of residues, as has been described previously[56]. Systems generated either 2,500, 5,000, 7,500 or 10,000 cluster centres from a *k*-centres clustering algorithm. Select systems were refined with 1–10 *k*-medoid sweeps. Transition probability matrices were produced by counting transitions between states, adding a prior count of $1/n$ (where n is the number of states) and row-normalizing, as has been described previously[81]. Equilibrium populations were calculated as the eigenvector of the transition probability matrix with an eigenvalue of one.

Spike–ACE2 binding competency. To determine spike protein binding competency to ACE2, the following structures of the RBD bound to ACE2 were used for SARS-CoV-1, SARS-CoV-2 and HCoV-NL63, respectively: 3D0G, 6M0J and 3KBH. The RBD of the bound complex was superimposed onto each RBD of the structures in our MSM. Steric clashes were then determined between backbone atoms on the ACE2 molecule and the rest of the spike protein. If any of the structures had a superposition that resulted in no clashes, it was deemed binding competent. The final population of binding-competent states was determined as the sum of state populations that were deemed binding competent. Error bars were obtained from bootstrapping the MSM equilibrium populations, as implemented in enspara.

Cryptic pockets and solvent-accessible surface area. For ease of detecting cryptic pockets and other functional motions, we employed our exposon analysis method[56]. This method correlates the solvent exposure between residues to find concerted motions that tend to represent cryptic pocket openings. Solvent-accessible surface area calculations were computed using the Shrake–Rupley algorithm as implemented in the python package MDTraj (ref. [82]). For all proteins and complexes, a solvent probe radius of 0.28 nm was used, which has been shown to produce a reasonable clustering and exposon map[56].

Spike protein solvent-accessible surface areas for SARS-CoV-2 were computed with glycan chains modelled onto each cluster centre. Multiple glycan rotamers were sampled for each state, and accessible surface areas for each residue were weighted based on MSM equilibrium populations.

Sequence conservation. Sequence conservation of spike proteins was calculated using the Uniprot database[83]. Sequences between 30 and 90% were pulled and aligned with the Muscle algorithm[84]. The entropy at each position was calculated to quantify variability of amino acids. Conservation was defined as one minus the entropy.

Data availability

Data supporting the findings of this study are available within the article and its Supplementary Information. The datasets generated and/or analysed during the current study are available at https://covid.molssi.org/ and https://osf.io/fs2yv/.

Code availability

GROMACS (https://github.com/gromacs/gromacs), OpenMM (https://github.com/openmm/openmm), our FAST adaptive sampling method (https://github.com/bowman-lab/fast), mdtraj (https://github.com/mdtraj/mdtraj) and our enspara code (https://github.com/bowman-lab/enspara) are all open source.

References

68. Abraham, M. J. et al. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
69. Duan, Y. et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **24**, 1999–2012 (2003).
70. Huang, J. & MacKerell, A. D. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J. Comput. Chem.* **34**, 2135–2145 (2013).
71. Knoverek, C. R., Amarasinghe, G. K. & Bowman, G. R. Advanced methods for accessing protein shape-shifting present new therapeutic opportunities. *Trends Biochem. Sci.* **44**, 351–364 (2018).
72. Bowman, G. R. Accurately modeling nanosecond protein dynamics requires at least microseconds of simulation. *J. Comput. Chem.* **37**, 558–566 (2016).
73. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
74. Hess, B. P-LINCS: a parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.* **4**, 116–122 (2008).
75. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
76. Zimmerman, M. I. et al. Prediction of new stabilizing mutations based on mechanistic insights from Markov state models. *ACS Cent. Sci.* **3**, 1311–1321 (2017).
77. Hendlich, M., Rippmann, F. & Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* **15**, 359–363, 389 (1997).

78. Eastman, P. et al. OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659 (2017).
79. Husic, B. E. & Pande, V. S. Markov state models: from an art to a science. *J. Am. Chem. Soc.* **140**, 2386–2396 (2018).
80. Porter, J. R., Zimmerman, M. I. & Bowman, G. R. Enspara: modeling molecular ensembles with scalable data structures and parallel computing. *J. Chem. Phys.* **150**, 044108 (2019).
81. Zimmerman, M. I., Porter, J. R., Sun, X., Silva, R. R. & Bowman, G. R. Choice of adaptive sampling strategy impacts state discovery, transition probabilities, and the apparent mechanism of conformational changes. *J. Chem. Theory Comput.* **14**, 5459–5475 (2018).
82. McGibbon, R. T. et al. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
83. Consortium, U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
84. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

Acknowledgements

We are extremely grateful to all the citizen scientists who contributed their compute power to make this work possible and to members of the Folding@home community who volunteered to help with everything from technical support to translating content into multiple languages. Thanks to Microsoft AI for Health for helping us use Azure to run adaptive sampling simulations and to UKRI for providing computational resources to parallelize data analysis. Thanks to Pure Storage for providing a FlashBlade system to store our large datasets, to Seagate and Micron for additional storage and to MolSSI for helping organize public datasets. Thanks to Avast, AWS, Cisco, Linus Tech Tips, Microsoft Azure, Oracle and VMware for helping us to scale up Folding@home's server-side infrastructure to keep up with the tremendous growth we experienced in such a short time. Thanks to AMD, ARM, Neocortix and Intel for helping to improve the performance of Folding@home on their hardware. Thanks to all of these companies for helping to spread the word about Folding@home and also to A16Z, Best Buy, CCP, CoreWeave, Daimler Truck AG, Dell, GitHub, HP, La Liga, Media Monks, Microcenter, NVIDIA and Telefonica. Thanks to CERN and the particle physics community for helping with data management and to DataDog for server monitoring services. Thanks to C. O. Barnes for providing the epitope contacts used for Fig. 2d. J.R.P. acknowledges support from F30HL146052. G.R.B. and his lab were supported by funding from Avast, the Center for the Science and Engineering of Living Systems (CSELS), an NSF RAPID award, NSF CAREER Award MCB-1552471, NIH R01 GM124007, NIH RF1 AG067194, a Burroughs Wellcome Fund Career Award at the Scientific Interface and a Packard Fellowship for Science and Engineering. J.D.C. acknowledges support from NIH grants P30 CA008748 and R01 GM121505. V.A.V. and M.F.D.H. acknowledge support from NIH grant R01 GM123296, NIH grant S10-OD020095 and NSF MRI grant CNS-1625061.

Author contributions

M.I.Z. and G.R.B. contributed to the conception and design of the work. M.I.Z., J.R.P., M.D.W., S.S., N.V., A.M., U.L.M., C.E.K., J.H.B., R.P.W., M.F.D.H., A.M.H., C.A.F., J.E.C., E.F., V.A.V., J.D.C. and G.R.B. aided in the acquisition of data. M.I.Z., J.R.P., M.D.W., S.S., N.V., A.M., U.L.M., C.E.K. and J.H.B. analysed and interpreted data. M.I.Z. and G.R.B. drafted the manuscript.

Competing interests

J.D.C. is a current member of the scientific advisory board of OpenEye Scientific Software and a consultant to Foresite Laboratories. The Chodera laboratory receives or has received funding from multiple sources, including the National Institutes of Health, the National Science Foundation, the Parker Institute for Cancer Immunotherapy, Relay Therapeutics, Entasis Therapeutics, Silicon Therapeutics, EMD Serono (Merck KGaA), AstraZeneca, Vir Biotechnology, Bayer, XtalPi, the Molecular Sciences Software Institute, the Starr Cancer Consortium, the Open Force Field Consortium, Cycle for Survival, a Louis V. Gerstner Young Investigator Award and the Sloan Kettering Institute. A complete funding history for the Chodera lab can be found at http://choderalab.org/funding.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41557-021-00707-0.

Correspondence and requests for materials should be addressed to G.R.B.

Peer review information *Nature Chemistry* thanks the anonymous reviewers for their contribution to the peer review of this work.