Citizen Scientists Create an Exascale Computer to Combat COVID-19

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], Gregory R. Bowman[1,2,]*

[1]*Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, Missouri 63110, United States*
[2]*Center for Science and Engineering of Living Systems (CSELS), Washington University in St. Louis, St. Louis, Missouri 63130, United States*
[3]*Tri-Institutional PhD Program in Chemical Biology, Memorial Sloan Kettering Cancer Center, New York, New York 10065, United States*
[4]*Computational and Systems Biology Program, Sloan Kettering Institute, New York, New York 10065, United States*
[5]*Department of Chemistry, Temple University, Philadelphia, Pennsylvania 19122, United States*
[6]*Department of Chemistry and Hamilton Institute, Maynooth University, Maynooth, Kildare, Ireland*
[7]*Cauldron Development LLC*

**Corresponding Author: g.bowman@wustl.edu*

Abstract

The SARS-CoV-2/COVID-19 pandemic continues to threaten global health and socioeconomic stability. Experiments have revealed snapshots of many of the viral components but remain blind to moving parts of these molecular machines. To capture these essential processes, over a million citizen scientists have banded together through the Folding@home distributed computing project to create the world's first Exascale computer and simulate protein dynamics. An unprecedented 0.1 seconds of simulation of the viral proteome reveal how the spike complex uses conformational masking to evade an immune response, conformational changes implicated in the function of other viral proteins, and 'cryptic' pockets that are absent in experimental snapshots. These structures and mechanistic insights present new targets for the design of therapeutics.

This living document will be updated as we perform further analysis and make the data publicly accessible.

Introduction

Severe acute respiratory syndrome coronavirus 2 (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 coronavirus disease 2019 (COVID-19) ranging from ~0.1-12% (high temporal variation), SARS-CoV-2/COVID-19 has the potential to spread quickly and endanger the global population.[2-6] As of June 23rd, 2020, there have been over 9.1 million confirmed cases and over 472,000 fatalities, globally. Quarantines and social distancing are effective at slowing the rate of infection; however, they cause significant social and economic disruption. Taken together, it is 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] Towards this end, the structural biology community has made heroic efforts to rapidly build models of SARS-CoV-2 proteins and the complexes they form. 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 proteins in SARS-CoV-2 will provide a clearer picture of how they accomplish their functions, such as infecting cells, evading the 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 significant 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. 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 compute resources on an unprecedented 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. Together, we have run an unprecedented 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/) in accordance with open science principles to accelerate the discovery of new therapeutics.[8,9]

To the Exascale and beyond!

Folding@home (http://foldingathome.org) is a community of citizen scientists, researchers, and tech organizations dedicated to 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. The project was founded in the year 2000 with the intent of understanding how proteins fold.[10] At the time, simulating the folding of even small proteins could easily take thousands of years on a single computer. To overcome this challenge, Folding@home divided this seemingly intractable problems into smaller simulations that could be performed completely independently of one another. They then created the Folding@home project to enable anyone with a computer and an internet connection to volunteer to run these small chunks of simulation, called "work units".

Over the years, the applications of Folding@home have been generalized to address many aspects of protein dynamics, and the algorithms have developed significantly. The Folding@home Consortium now involves eight laboratories around the world studying various aspects of disease from cancer to microbial resistance to membrane protein dysfunction diseases (https://foldingathome.org/about/the-foldinghome-consortium/). The project has provided insight into diverse topics, ranging from signaling mechanisms.[11-13] to the connection between phenotype and genotype.[14-16] Translational applications have included new means to combat antimicrobial resistance, Ebola virus, and SFTS virus.[17-19]

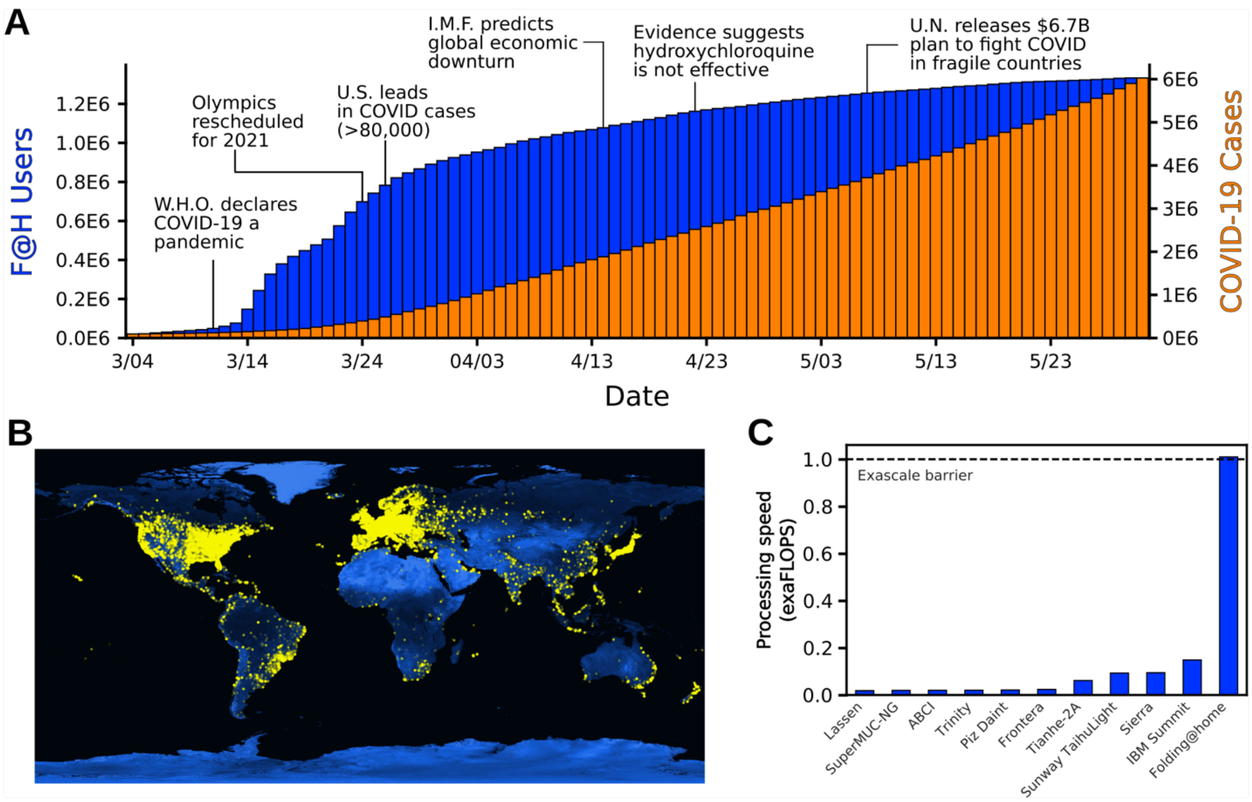


Figure 1: Summary of Folding@home's computational power. A) The growth and usage of Folding@home in response to COVID-19. Users are colored blue and COVID-19 cases are orange. **B)** Location of folding at home users. Each yellow dot represents a unique IP address contributing to Folding@home. **C)** The processing speed of Folding@home and the next 10 fastest supercomputers, in exaFLOPS.

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 at a time when they were otherwise feeling helpless alluring. In less than three months, the project grew from ~30,000 active devices to over a million devices around the globe (Fig. 1A and 1B).

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 to simulate. 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 (see Methods for details). We note that a larger time window is used on our website for historical reasons.

Given the above, we conservatively estimate the peak performance of Folding@home hit 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 the 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 exaFLOPS performance we report for Folding@home is 5-fold greater than the peak performance of the world's fastest traditional supercomputer, called Summit (Fig. 1C). It is also more 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.

Unmasking the spike complex

The spike complex (S) is a prominent vaccine target that is known to undergo substantial conformational changes as part of its function.[20-22] 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.[23,24] The fact that S is exposed on the virion surface makes it an appealing vaccine target. However, it has a number of effective defense strategies. First, S is decorated extensively with glycans that aid in immune evasion by shielding potential antigens.[25,26] S also uses a conformational masking strategy, wherein it predominantly adopts a closed conformation that buries the receptor-binding domains (RBDs) to evade immune surveillance mechanisms. To engage with ACE2, S undergoes rare transitions to an open state that exposes the conserved binding interface of the RBDs. Characterization of the full range of this motion 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 iterates between running a batch of simulations, building a Markov state model (MSM), ranking the MSM states 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.[27,28] The ranking function is designed to balance between favoring structures with a desired geometric feature (in this case opening of S) and broad exploration of conformational space. By balancing exploration-exploitation tradeoffs, 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, totaling over 1 ms of data for SARS-CoV-2 S, enabling us to obtain a statistically sound final model.

Our SARS-CoV-2 S protein simulations capture opening of S and substantial conformational heterogeneity in the open state (Fig. 3). Capturing opening of S is an impressive technical feat given that previous large-scale simulations were unable to observe this essential event for the initiation of infection.[26] Intriguingly, we find that opening occurs only for a single

RBD at a time, akin to that observed in cryoEM structures.[29] Additionally, we find that the scale of this opening can be substantially larger than has been observed in experimental snapshots (Fig. 3). The dramatic opening we observe is consistent with the observation that antibodies can bind to regions of the RBD that are deeply buried and seemingly inaccessible in existing experimental snapshots.[30]

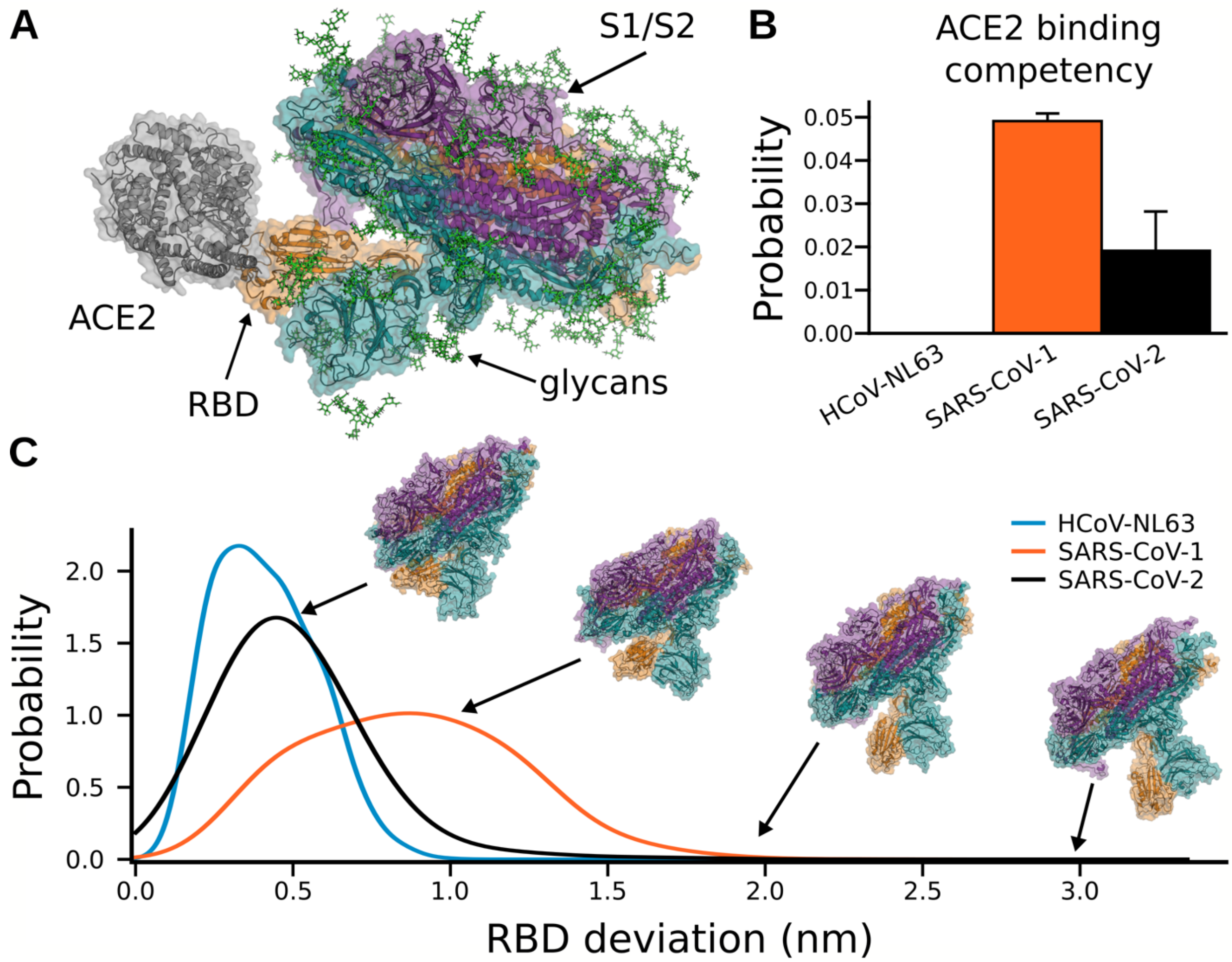


Figure 2: Structural characterization of conformational masking in different spike complexes. A) A representative structure of SARS-CoV-2 spike protein in an open conformation, as pulled from our molecular dynamic simulations. ACE2 (gray) is superimposed onto the structure to highlight binding compatibility. The three chains of Spike are illustrated with a cartoon and transparent surface representation (orange, teal, and purple), and glycans are shown as sticks (green). **B)** The probability that each sequence adopts an ACE2 binding competent pose. HCoV-NL63, SARS-CoV-1, and SARS-CoV-2 are shown as light-blue, orange, and black, respectively. **C)** The probability that the center of mass of an RBD deviates from its position in the closed state for HCoV-NL63, SARS-CoV-1, and SARS-CoV-2. To highlight the protein determinants of opening, comparison of spike openings is performed without glycosylation.

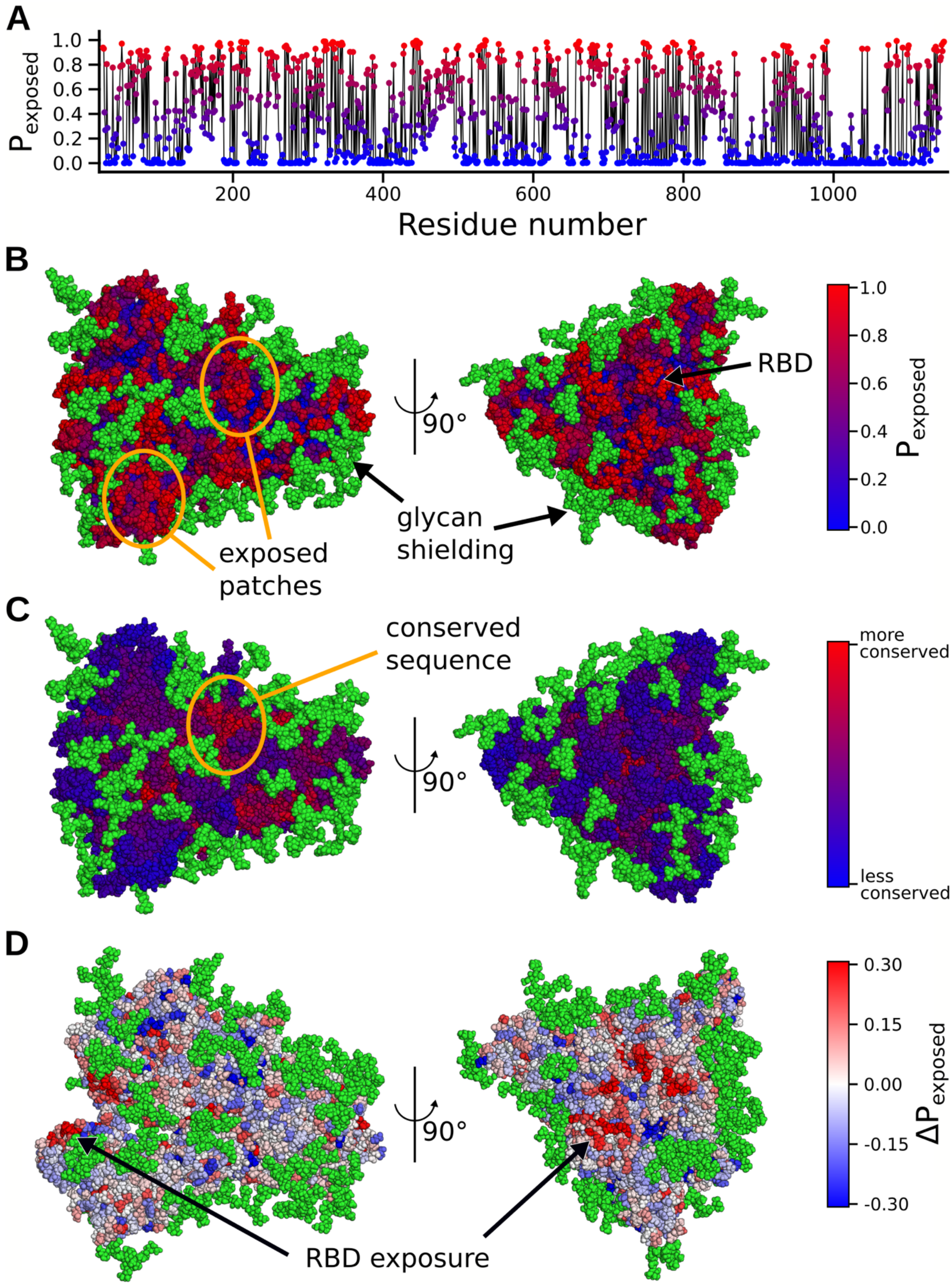
To understand the potential role of conformational masking in determining the lethality and infectivity of different coronaviruses, we also simulated the opening of S proteins from two related viruses: SARS-CoV-1 and HCoV-NL63. These viruses were selected because they also bind the ACE2 receptor but are associated with varying mortality rates. SARS-CoV-1 caused an outbreak in 2003 with a high case fatality rate but has not become a pandemic.[31] NL63 was discovered the following year and continues to spread around the globe, although it is significantly less lethal than either SARS virus.[32] We hypothesized that these phenotypic differences may be explained by changes to the S conformational ensemble. Specifically, we propose 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.

As expected, the three S complexes have very different propensities to adopt an open state and bind ACE2. 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 find 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 observe a number of interesting correlations between 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).[33,34]

These observations suggest a tradeoff wherein greater conformational masking enables immune evasion but requires a higher affinity between an exposed RBD and ACE2 to successfully 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. Based on this model, we predict that mutations that increase the probability that the SARS-CoV-2 S complex adopts open conformations may be more lethal but spread less readily.

Our atomically detailed model of S can enable rapid structure-based vaccine antigen design through identification of regions minimally protected by conformational masking or the glycan shield.[35] To identify these potential epitopes, we calculated the probability that each residue in S could be exposed to therapeutics (e.g. 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 are exposed through the glycan shielding (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). Another possibility for antigen design is to exploit the opening motion. A number of residues surrounding the receptor binding motif (RBM) of the RBD show an increase in exposure by ~30% in ACE2 binding competent structures (Fig. 3C). Consistent with immunoassays, this region was recently found to be a cluster for neutralizing antibody binding.[36,37]



A
1.0
0.8
0.6
0.4
0.2
0.0
P_exposed
200
400
600
800
1000
Residue number
B
exposed patches
90°
glycan shielding
RBD
1.0
0.8
0.6
0.4
0.2
0.0
P_exposed
C
conserved sequence
90°
more conserved
less conserved
D
90°
RBD exposure
0.30
0.15
0.00
-0.15
-0.30
ΔP_exposed

Figure 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, as determined from our structural ensemble. **B)** Exposure probabilities colored on the surface of the spike protein. 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. Green atoms denote glycans. **C)** Sequence conservation score colored onto the Spike protein. 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. Red residues have a higher probability of being exposed upon opening, whereas blue residues have a lower probability of being exposed.

Cryptic pockets and functional dynamics

Every protein in SARS-CoV-2 remains a potential drug target. So, to understand their role in disease and help progress the design of 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 proteins' functional dynamics, expanding on the number of antiviral targets is of immediate value. Towards this end, we seeded Folding@home simulations from our FAST-pockets adaptive sampling to aid in the discovery of cryptic pockets. We briefly discuss two illustrative examples, out of 36 datasets.

Nonstructural 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 lifecycle of coronaviruses, cleaving polyprotein 1a into functional proteins, and is a major target for the design of antivirals.[38] 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 significant 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 in the low µM range.[39] Small molecules targeting this protein to inhibit enzymatic activity, either by altering its active site or favoring the inactive monomer state, would be promising broad-spectrum antiviral candidates.[40]

Our simulations reveal two novel cryptic pockets on M^{pro} that expand our current therapeutic options. These are shown in figure 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 find 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—i.e. by sterically regulating substrate binding—and is similar to motions we have observed previously for the enzyme β-lactamase.[41] Owing to its 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 to find small molecule or peptide stabilizers that favor the inactive monomer state.

In addition to cryptic pockets, our data captures 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 lifecycle events such as viral packaging, transcription, and physically linking RNA to the envelope.[42,43] 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.[44,45] Investigating the RNA-binding domain, we observe both cryptic pockets and an incredibly dynamic beta-hairpin, which hosts the RNA binding site, referred to as a "positive finger" (Fig. 4C-D). Our observed conformational heterogeneity of the positive finger is consistent with a structural ensemble determined using solution-state nuclear magnetic resonance spectroscopy.[46] 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. Additionally, knowledge of these probabilities can provide further insight into the mechanisms and regulation of genome compaction/release.

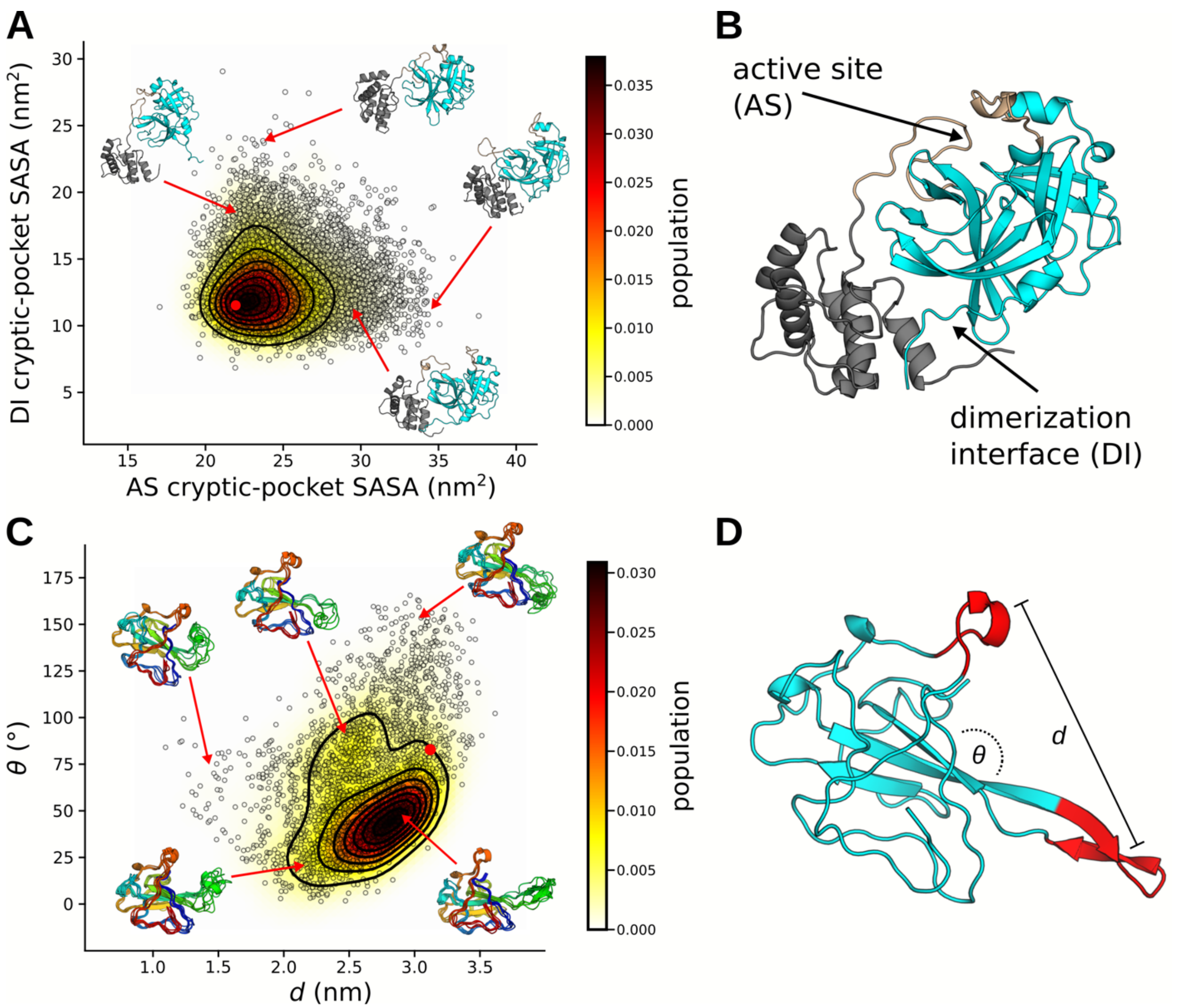


Figure 4: Examples of cryptic pockets and functionally-relevant dynamics. A-B) Conformational ensemble of M^{pro} (monomeric) projected onto the solvent accessible surface areas (SASAs) of residues surrounding either the active-site or the interface of dimerization. Cluster centers are represented as black circles with representative structures depicted with cartoon. The starting structure for simulations (6Y2E) is shown as a red dot. Domains I and II are colored cyan and domain III is colored gray. The loop of domain III, which covers the active-site residues and is seen to be highly dynamic, is colored tan. **C-D)** Conformational ensemble of Nucleoprotein projected onto the distance and angle between the positive finger and a nearby loop. Angles were calculated between vectors that point along each red segment in panel D and distances were calculated between their centers of mass. Cluster centers are represented as black circles and representative structures are depicted with cartoon. The starting structure for simulations (6VYO) is shown as a red dot.

The data we present in this paper represents 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, we are posting all of our data online (https://covid.molssi.org/).

Table 1: **A list of protein systems we have simulated on Folding@home.** Systems are organized by viral strain and include name, oligomerization state, starting structure, number of residues, number of atoms in the system, aggregate simulation time, and the number of cryptic pockets we have identified (TBD: simulations are still being analyzed and pockets are "to be determined").
*Missing residues were modeled using Swiss model.[47]
**Structural model was generated from a homologous sequence using Swiss model.[47]
***Missing residues were modeled using CHARMM-GUI.[48,49]

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	23907	11,266	TBD
NSP3 (Papain-like protease 2, PL2^{pro})	Monomer	3E9S**	306	97285	663	TBD
NSP5 (main protease, M^{pro}, 3CL^{pro})	Monomer	6Y2E	306	64791	6,449	2
NSP5 (main protease, M^{pro}, 3CL^{pro})	Dimer	6Y2E	612	77331	2,466	TBD
NSP7	Monomer	5F22**	79	20094	8,192	TBD
NSP8	Monomer	2AHM**	191	156282	1,404	TBD
NSP9	Dimer	6W4B*	226	49885	7,570	TBD
NSP10	Monomer	6W4H*	131	29560	3,313	2
NSP12 (polymerase)	Monomer	6NUR**	891	186622	1,229	TBD
NSP13 (helicase)	Monomer	6JYT**	596	129368	1,102	3
NSP14	Monomer	5C8S**	527	216380	439	TBD
NSP15	Monomer	6VWW	347	67345	4,032	4
NSP15	Hexamer	6VWW	2082	230339	1,723	TBD
NSP16	Monomer	6W4H*	298	45672	2,812	5
Nucleoprotein (RBD)	Monomer	6VYO	173	29125	9,160	3
Nucleoprotein Dimerization Domain	Monomer	6YUN*	118	34905	2,742	TBD
Nucleoprotein Dimerization Domain	Dimer	6YUN*	236	72733	1,433	TBD
Spike	Trimer	6VXX***	3363	442881	1,073	TBD
NSP7 / NSP8 / NSP12	Trimer complex	6NUR**	1184	215694	347	TBD
NSP10 / NSP14	Dimer complex	5C8S**	688	226672	527	TBD
NSP10 / NSP16	Dimer complex	6W4H*	429	63752	2,585	TBD
SARS-CoV-1						
NSP3 (Macrodomain "X")	Monomer	2FAV	172	33117	507	TBD
NSP9	Dimer	1QZ8*	226	49599	6,736	TBD
NSP15	Monomer	2H85	345	67345	3,037	TBD
NSP15	Hexamer	2H85	2070	230339	941	TBD
Nucleoprotein RBD	Monomer	2OFZ	174	29125	4,286	TBD
Nucleoprotein Dimerization Domain	Monomer	2GIB	370	34905	1,029	TBD
Nucleoprotein	Dimer	2GIB	740	72733	501	TBD

Dimerization Domain						
Spike	Trimer	5X58***	3261	375851	859	TBD
NSP10 / NSP16	Dimer complex	6W4H**	425	69589	205	TBD
Human						
IL6	Monomer	1ALU	166	26855	1,337	TBD
IL6-R	Monomer	1N26	299	149764	147	TBD
ACE2	Monomer	6LZG	596	75787	408	TBD
MERS						
NSP13	Monomer	5WWP	596	121134	335	TBD
NSP10 / NSP16	Dimer Complex	6W4H**	424	69127	204	TBD
HCoV-NL63						
Spike	Trimer	5SZS***	3606	453348	651	TBD

Discussion

In this work, we have utilized the largest computational resource in the world to tackle a global threat. Over a million citizen scientists have pooled their computer resources to combat COVID-19, generating more than 0.1 seconds of simulation data. The unprecedented scale of these simulations has helped to characterize crucial stages of infection. We find 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 tradeoff than related coronaviruses, which may explain its success in spreading globally. Our simulations also provide an atomically detailed roadmap for targeting proteins for vaccines and antivirals. Furthermore, we are working on making a comprehensive atlas and repository of cryptic pockets hosted online to accelerate the development of novel therapeutics.

Beyond SARS-CoV-2, we expect this work to aid in a better understanding of the roles of proteins in the *coronaviridae* family. Coronaviruses 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 from conventional means. With our explosive growth, our compute power has increased around 100-fold. Our work here highlights the incredible utility this compute power has to rapidly understand health, disease, and aid in drug design. Both in terms of scale and approach, we are in a new frontier of using molecular dynamics simulations to understand biophysics; the complex task of simulating an organism's entire proteome could become commonplace. 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.

Methods

System preparation

All simulations were prepared using Gromacs 2020.[50] 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 minimized with a steepest descents algorithm until the maximum force fell below 100 kJ/mol/nm using a step size of 0.01 nm and a cutoff distance of 1.2 nm for the neighbor 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.[51,52] All simulations were simulated with explicit TIP3P solvent.[53]

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.[54] Cutoffs of 1.1 nm were used for the neighbor 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 cutoff scheme was used for the neighbor list. The stochastic velocity rescaling (*v*-rescale) thermostat was used to hold the temperature at 300 K.[55]

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, 5) repeat steps 2-4 until ranking is optimized. For each system, MSMs were generated after each round of sampling using a *k*-centers clustering algorithm based on the RMSD between select atoms. Clustering continued until the maximum distance of a frame to a cluster center fell within a predefined cutoff. In addition to the FAST ranking, a similarity penalty was added to promote conformational diversity in starting structures, as has been described previously.[56]

FAST-distance simulations of all Spike proteins were run at 310 K on the Microsoft Azure cloud computing platform. The FAST-distance ranking favored 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 totaled 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 6 rounds, with 10 simulations per round, where each simulation was 40 ns in length (2.4 μs aggregate simulation). The FAST-pocket ranking function favored restarting simulations from states with large pocket openings. Pocket volumes were calculated using the LIGSITE algorithm.[57]

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*-centers clustering of adaptive sampling, as is described above. Simulations were deployed using a simulation core based on either GROMACS 5.0.4 or OpenMM 7.4.1.[50,58]

To estimate the performance of Folding@home, 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/core. We report x86-equivalent FLOPS.

Markov state models
A Markov state model is a network representation of a free energy landscape and is a key tool for making sense of molecular dynamics simulations.[59] All MSMs were built using our python package, enspara.[60] 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 RMSD between backbone C_α coordinates. States were generated as the top 3000 centers from a k-centers clustering algorithm. All other proteins were clustered based on the euclidean distance between the solvent accessible surface area of residues, as is described previously.[41] Systems generated either 2500, 5000, 7500, or 10000 cluster centers from a k-centers 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 $1/n_{states}$, and row-normalizing, as is described previously.[61] 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: 3D0G, 6M0J, and 3KBH, for SARS-CoV-1, SARS-CoV-2, and HCoV-NL63, respectively. The RBD of the bound complex was superimposed onto each RBD for 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.

Cryptic pockets and solvent accessible surface area
For ease of detecting cryptic pockets and other functional motions, we employed our exposon analysis method.[41] 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.[62] 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.[41]

Spike protein solvent accessible surface areas for SARS-CoV-2 were computed with glycan chains modeled onto each cluster center. 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.[63] Sequences between 30% - 90% were pulled and aligned with the Muscle algorithm.[64] The entropy at each position was calculated to quantify variability of amino acids. Conservation was defined as one minus the entropy.

Acknowledgements
We are extremely grateful to all the citizen scientists who contributed their compute power to make this work possible, and 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 compute 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 and 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. GRB 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, a Burroughs Wellcome Fund Career Award at the Scientific Interface, and a Packard Fellowship for Science and Engineering. JDC acknowledges support from NIH grant P30 CA008748 and NIH grant R01 GM121505. VAV and MFDH acknowledge support from NIH grant R01 GM123296, NIH grant S10-OD020095, and NSF MRI grant CNS-1625061.

Disclosures

JDC 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.

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,** (2020).
3. Sorci, G., Faivre, B. & Morand, S. Why Does COVID-19 Case Fatality Rate Vary Among Countries? *SSRN Journal* (2020). doi:10.2139/ssrn.3576892
4. Morteza Abdullatif Khafaie, F. R. Cross-Country Comparison of Case Fatality Rates of COVID-19/SARS-COV-2. *Osong Public Health and Research Perspectives* **11,** 74–80 (2020).
5. Mahase, E. Coronavirus: covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate. *BMJ* **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, Dos, R. N., Oliva, G. & Andricopulo, A. D. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules 2015, Vol. 20, Pages 13384-13421* **20,** 13384–13421 (2015).
8. Stodden, V. Enabling Reproducible Research: Open Licensing for Scientific Innovation. (2009).
9. Amaro, R. E. & Mulholland, A. J. A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19. *Journal of Chemical Information and Modeling* **60,** 2653–2656 (2020).
10. Shirts, M. & Pande, V. S. COMPUTING: Screen Savers of the World Unite! *Science* **290,** 1903–1904 (2000).
11. Kohlhoff, K. J. *et al.* Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nature Chem* **6,** 15–21 (2014).
12. 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,** 1–11 (2014).
13. 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,** 19 (2018).
14. 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,** 1–10 (2016).
15. Chen, S. *et al.* The dynamic conformational landscape of the protein methyltransferase SETD8. *eLife* **8,** 213 (2019).
16. 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,** 19 (2020).
17. Hart, K. M. *et al.* Designing small molecules to target cryptic pockets yields both positive and negative allosteric modulators. *PLOS ONE* **12,** e0178678 (2017).
18. Cruz, M. A. *et al.* Discovery of a cryptic allosteric site in Ebola's 'undruggable' VP35 protein using simulations and experiments. *bioRxiv* **17,** 2020.02.09.940510 (2020).
19. Wang, W. *et al.* The Cap-Snatching SFTSV Endonuclease Domain Is an Antiviral Target. *Cell Reports* **30,** 153–163.e5 (2020).
20. Kirchdoerfer, R. N. *et al.* Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Sci Rep* **8,** 1–11 (2018).
21. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181,** 281–292.e6 (2020).
22. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367,** 1260–1263 (2020).
23. 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).
24. 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).
25. Watanabe, Y. *et al.* Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nat Commun* **11,** 1–10 (2020).
26. Casalino, L. *et al.* Shielding and Beyond: The Roles of Glycans in SARS-CoV-2 Spike Protein. **9,** 221–27 (2020).
27. Zimmerman, M. I. & Bowman, G. R. FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *J. Chem. Theory Comput.* **11,** 5747–5757 (2015).

28. Zimmerman, M. I. & Bowman, G. R. How to Run FAST Simulations. *Methods in Enzymology* **578,** 213–225 (2016).
29. Yuan, Y. *et al.* Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nat Commun* **8,** 1–9 (2017).
30. Huo, J. *et al.* Neutralization of SARS-CoV-2 by Destruction of the Prefusion Spike. *SSRN Journal* (2020). doi:10.2139/ssrn.3613273
31. Zhong, N. S. *et al.* Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *The Lancet* **362,** 1353–1358 (2003).
32. van der Hoek, L. *et al.* Identification of a new human coronavirus. *Nat Med* **10,** 368–373 (2004).
33. Wu, K., Li, W., Peng, G. & Li, F. Crystal structure of NL63 respiratory coronavirus receptor-binding domain complexed with its human receptor. *PNAS* **106,** 19970–19974 (2009).
34. Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581,** 221–224 (2020).
35. Graham, B. S., Gilman, M. S. A. & McLellan, J. S. Structure-Based Vaccine Antigen Design. *Annu. Rev. Med.* **70,** 91–104 (2019).
36. Brouwer, P. J. M. *et al.* Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. *Science* **38,** eabc5902 (2020).
37. Hansen, J. *et al.* Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. *Science* eabd0827 (2020). doi:10.1126/science.abd0827
38. 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).
39. Vito Graziano, William J McGrath, Lin Yang, A.Walter F Mangel. *SARS CoV Main Proteinase: The Monomer−Dimer Equilibrium Dissociation Constant*. *Biochemistry* **45,** 14632–14641 (American Chemical Society, 2006).
40. Goyal, B. & Goyal, D. Targeting the Dimerization of the Main Protease of Coronaviruses: A Potential Broad-Spectrum Therapeutic Strategy. *ACS Combinatorial Science* **22,** 297–305 (2020).
41. Porter, J. R. *et al.* Cooperative Changes in Solvent Exposure Identify Cryptic Pockets, Switches, and Allosteric Coupling. *Biophysical Journal* **116,** 818–830 (2019).
42. McBride, R., Van Zyl, M. & Fielding, B. C. The Coronavirus Nucleocapsid Is a Multifunctional Protein. *Viruses 2014, Vol. 6, Pages 2991-3018* **6,** 2991–3018 (2014).
43. Masters, P. S. Coronavirus genomic RNA packaging. *Virology* **537,** 198–207 (2019).
44. Cubuk, J. *et al.* The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *bioRxiv* **53,** 171–39 (2020).
45. 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).
46. Dinesh, D. C., Chalupska, D., Silhan, J., Veverka, V. & Boura, E. Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein. **73,** 213–13 (2020).
47. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46,** W296–W303 (2018).
48. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of computational chemistry* **29,** 1859–1865 (2008).

49. 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).
50. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2,** 19–25 (2015).
51. 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).
52. 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).
53. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79,** 926–935 (1983).
54. Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **4,** 116–122 (2008).
55. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126,** 014101 (2007).
56. 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).
57. Hendlich, M., Rippmann, F. & Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* **15,** 359–63–389 (1997).
58. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **13,** e1005659 (2017).
59. Husic, B. E. & Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **140,** 2386–2396 (2018).
60. Porter, J. R., Zimmerman, M. I. & Bowman, G. R. Enspara: Modeling molecular ensembles with scalable data structures and parallel computing. *The Journal of Chemical Physics* **150,** 044108 (2019).
61. 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).
62. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **109,** 1528–1532 (2015).
63. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47,** D506–D515 (2019).
64. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).