Determining the contribution of glycosylation to SARS-CoV-2 S-protein conformational dynamics

~~PI: James C. Gumbart, Georgia Institute of Technology~~

Abstract: We propose to resolve the role of covalently attached oligosaccharides in the conformational transition of the SARS-CoV-2 S protein from a closed, non-infectious state to an open one that can bind to its receptor ACE2 on human cells. The oligosaccharides form a so-called "glycan shield", which comprises 20% of the mass of the system and contributes to immune system evasion. We will use highly scalable replica-exchange umbrella sampling to map the free-energy landscape of the conformational change in non- and fully-glycosylated states. Intermediate states along the pathway for the native, glycosylated S protein will be used in an ongoing high-throughput virtual screening effort at ORNL.

Scientific Goals

The S protein, which forms spikes on the surface of the viral envelope, must first go through a conformational change in which it "opens up", exposing the binding motif for its human target, the protein ACE2, on the receptor binding domain (RBD). High-resolution structures of the S-protein trimer in both its closed (down) and open (one monomer up) states have been determined [1], but the transition pathway connecting them is completely unknown. Nonetheless, **it is precisely the intermediate states on the transition pathway that present attractive targets for small-molecule inhibition**, as something that could stabilize the interface between the RBD and the rest of the S-protein trimer would render the virus innocuous.

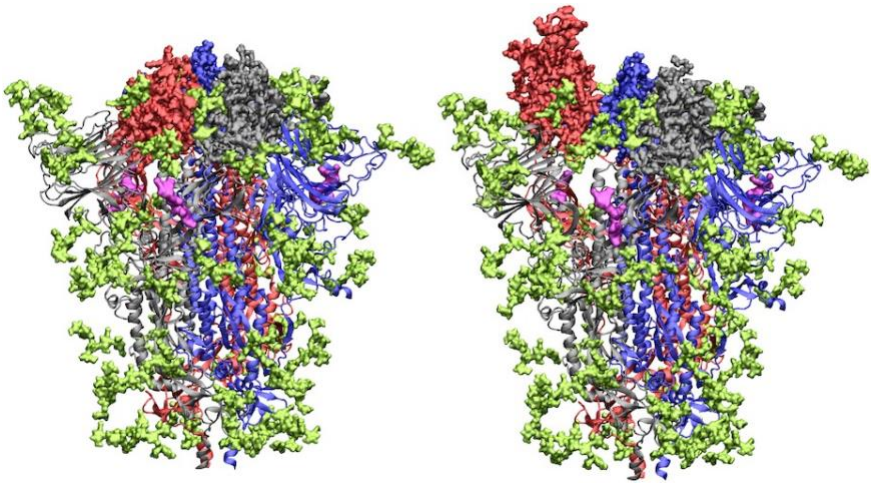

Figure 1: **Initial model of the S-protein trimer in its closed (left) and open (right) states.** The three monomers are shown in blue, red, and grey, respectively. N- and O-linked glycans surrounding the protein are shown in green and magenta, respectively.

One of the key features of the S protein is its high degree of glycosylation. Specifically, the S-protein trimer possesses numerous oligosaccharides covalently attached to 18 N- and 1 O-linked glycosylation sites per monomer, comprising ~20% of the system's total mass. The specific distribution of glycans at each site has been determined by site-specific mass spectrometry [2] (see Fig. 1). The importance of this so-called "glycan shield" cannot be overstated as it is one of the primary mechanisms of immune evasion by viruses [3, 4]. **Thus, it is of critical importance to understand how the presence of the glycan shield affects the S-protein's conformational dynamics.** Indeed, there are seven oligosaccharides attached to, or in the vicinity of, the RBD, which will alter the opening pathway and energetics.

In order to resolve the contribution of bound oligosaccharides to S-protein opening, we will map the free-energy landscape of the conformational change between down and up states of the S protein with and without glycosylation. To this end, we have already created one of the most detailed models of the S-protein trimer, having modeled the numerous missing protein segments, mutated residues back to wild-type where necessary, and **added oligosaccharides to 57 N- and O-linked glycosylation sites**, including those missing from the original structures [1, 2, 5]. To provide sufficient space for the desired conformational changes, the fully solvated system is 0.7 million atoms. **We emphasize that our proposed calculations are likely to be distinct from existing ones due to the inclusion of the glycan shield.**

We will use atomistic molecular dynamics (MD) simulations combined with Replica Exchange Umbrella

Sampling (REUS) in NAMD to determine the potential of mean force (PMF) for opening along two reaction coordinates. These coordinates have been identified from our analysis of the available high resolution structures, augmented with the early-release 10-μs MD simulations from D.E. Shaw Research [6]. The two coordinates are an angle between the RBD and a neighboring one (which will remain in the down position), as well as a dihedral angle capturing the twist of the RBD. We will first carry out two 100-ns metadynamics simulations with aggressive settings, initiated once from the closed state and once from the open state, to efficiently explore the space of these two coordinates as we have done previously [7]. We have found this approach to be less susceptible to force-induced artifacts than the more commonly used targeted or steered MD. We will use conformations identified to seed the initial windows for REUS. The two reaction coordinates θ and ϕ cover a range of 50° and 42°, respectively. Based on our previous experience sampling the conformational dynamics of the protein AcrA on OLCF's Titan [8], to match the spacing along the arc length, we estimate a necessary window spacing of 2°. This means we will need as many as 25 × 21 = 525 windows for the full 2D PMF. However, as we will not sample very high-energy regions not explored by metadynamics, we estimate this will be reduced by at least 5% [7, 8], giving **500 windows**.

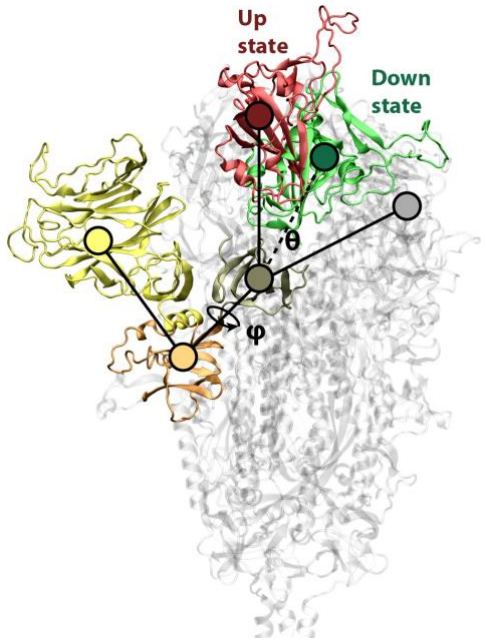

Figure 2: **Reaction coordinates.** An angle θ and a dihedral ϕ describe the conformational transition of the RBD between down and up states.

We will calculate PMFs for opening of (1) the S protein with no glycosylation and (2) the fully glycosylated S protein. For the non-glycosylated state, we will run REUS for 30 ns/window, which is sufficient to reach convergence based on our previous work [7–10]. For the fully glycosylated state, we will run REUS for 50 ns/window, in order to allow adequate time to sample multiple conformations of the attached glycan chains.

From our free-energy calculations, we can map the most probable transition pathway between closed and open states with and without glycosylation to determine the role of oligosaccharides in opening. In particular, from this pathway, we can identify key interactions that either assist with or oppose opening of the RBD. These interactions will serve as focal points for investigations into targeted inhibition of the closed→open S-protein transition. Representative structures along this pathway will be taken as input for subsequent small-molecule docking, similar to what we recently did for Hepatitis B Virus capsid protein [11]. Small-molecule docking will be carried out as part of a separate, on-going project by our collaborators, Dr. Micholas Smith and Dr. Jeremy Smith at ORNL [12]. In vitro testing of their first hits for a different system is already underway by Dr. Colleen Jonsson at the UT Health Science Center.

Estimate of Compute, Storage, and Other Resources

REUS takes advantage of the multiple-copy algorithms within NAMD [13], which permits all windows to be run simultaneously and exchange biases according to a Metropolis criterion, thereby accelerating convergence of the PMF [14]. **Due to the large scale of the proposed calculations (500 windows and 0.7 million atoms), we believe they can only be carried out on Summit.** Using Summit would also provide synergy with the efforts of our collaborators at ORNL for subsequent docking to intermediate structures identified along the minimum free-energy pathway between closed and open states.

To generate the initial windows using metadynamics, we will run 100 ns initiated from each of the open and closed conformations, for both the glycosylated and non-glycosylated forms of the S-protein trimer. While this is not a petascale-level simulation, it is a necessary precursor for the REUS simulations. If

deemed inappropriate for Summit, it can be run using local resources.

$$100\ \text{ns} \times \frac{1.14\ \text{hour}}{\text{ns}} \times \frac{2\ \text{nodes}}{\text{run}} \times 4\ \text{runs} = 912\ \text{node hours} \quad (1)$$

For the REUS simulations, we propose to run for 30 ns/window for the non-glycosylated state and 50 ns/window for the fully glycosylated state. The latter is longer to allow for sampling of multiple conformations of the glycan chains.

$$(30+50)\ \text{ns} \times \frac{1.14\ \text{hour}}{\text{ns}} \times \frac{2\ \text{nodes}}{\text{replica}} \times 500\ \text{replicas} = 91200\ \text{node hours} \quad (2)$$

Trajectory data will be output every 100 ps, which limits I/O to once every seven minutes. All systems generate a certain fixed amount of restart data along with a growing amount of trajectory data. The 0.7-million-atom system requires 32 MB of restart data per replica and 7.9 MB of trajectory data per window per frame. Based on the runs proposed, the data usage can be calculated as follows:

For the metadynamics,

$$\frac{32\ \text{MB}}{\text{run}} \times 4\ \text{runs} + \frac{7.9\ \text{MB}}{\text{run}-\text{frame}} \times 4\ \text{runs} \times \frac{10\ \text{frames}}{\text{ns}} \times 100\ \text{ns} = 0.32\ \text{TB} \quad (3)$$

For the REUS,

$$\frac{32\ \text{MB}}{\text{replica}} \times 500\ \text{replicas} + \frac{7.9\ \text{MB}}{\text{replica}-\text{frame}} \times 500\ \text{replicas} \times \frac{10\ \text{frames}}{\text{ns}} \times (50+30)\ \text{ns} = 3.18\ \text{TB} \quad (4)$$

Therefore, accounting for test runs, **we request 95,000 node-hours and 3.5 TB on Summit for the proposed calculations.**

Support Needs

Our support needs are minimal. We are in collaboration with current Summit users (Dr. Micholas Smith and Dr. Jeremy Smith, ORNL) who can assist with general job submission issues. We are also in close contact with Abhishek Singharoy (Arizona St.), a current INCITE user on Summit who can assist with any NAMD-related issues. Finally, we are also in contact with Dr. Josh Vermaas, an OLCF computational scientist, who ran the benchmarks provided on the right.

All of the data used and to be generated are already or will be made publicly available and are not restricted in any way.

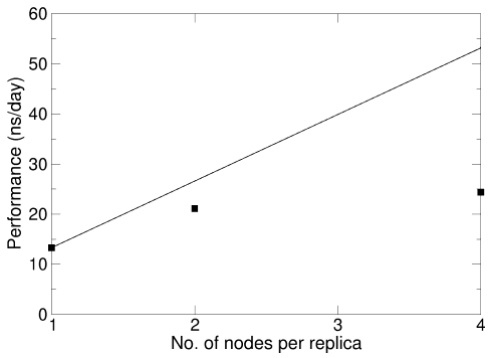

Figure 3: **Benchmarks on Summit.** Runs are for a single replica of the system using NAMD and all GPUs. While scaling for one replica is modest, Hamiltonian replica exchange, as proposed here, scales linearly with the number of replicas [13]. The individual points at 1, 2, and 4 nodes are from benchmarks run by Dr. Josh Vermaas, OLCF, while the line represents linear scaling based on one-node performance. The single-replica run is 80% efficient on two nodes, **suggesting the full REUS run will be equally efficient on 1000 nodes.**

Team and Team Preparedness

Our local team is composed of Dr. James C. Gumbart (PI); Dr. Anna Pavlova (research scientist, GT) and Andrew Pang (Physics graduate student), who are focused on free-energy calculations; along with Dr. Atanu Acharya (postdoc, GT) and Dr. Diane Lynch (research scientist, GT), who are focused on structural modeling and analysis. We are also collaborating with the drug docking team of Dr. Jeremy Smith (Director of the UT/ORNL Center for Molecular Biophysics) and Dr. Colleen Jonsson (Professor, University of Tennessee Health Science Center and Director of the Regional Biocontainment Laboratory).

Our team is prepared to begin immediately. We are currently equilibrating the system using local resources. We have experience running large-scale replica-exchange simulations, including at the petascale, e.g., REUS on a 193k-atom system with 1573 windows on Titan as part of an INCITE award [8].