A Multiscale Coarse-grained Model of the SARS-CoV-2 Virion

Authors: Alvin Yu¹, Alexander J. Pak¹, Peng He¹, Viviana Monje-Galvan¹, Lorenzo Casalino², Zied Gaieb², Abigail C. Dommer², Rommie E. Amaro², and Gregory A. Voth^{1,*}

Affiliations:

¹Department of Chemistry, Chicago Center for Theoretical Chemistry, Institute for Biophysical Dynamics, and James Franck Institute, The University of Chicago, Chicago, IL 60637

²Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, CA 92093

Corresponding Author:

*Gregory A. Voth
Department of Chemistry
The University of Chicago
5735 S. Ellis Ave, SCL 123
Chicago, IL 60637

Phone: (773) 702-9092

Fax: (773) 795-9106

E-mail: gavoth@uchicago.edu

Keywords: SARS-CoV-2, virion, coarse-graining, molecular dynamics simulations, multiscale

Abstract

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative agent of the COVID-19 pandemic. Computer simulations of complete viral particles can provide theoretical insights into large-scale viral processes including assembly, budding, egress, entry, and fusion. Detailed atomistic simulations, however, are constrained to shorter timescales and require billion-atom simulations for these processes. Here, we report the current status and on-going development of a largely “bottom-up” coarse-grained (CG) model of the SARS-CoV-2 virion. Structural data from a combination of cryo-electron microscopy (cryo-EM), x-ray crystallography, and computational predictions were used to build molecular models of structural SARS-CoV-2 proteins, which were then assembled into a complete virion model. We describe how CG molecular interactions can be derived from all-atom simulations, how viral behavior difficult to capture in atomistic simulations can be incorporated into the CG models, and how the CG models can be iteratively improved as new data becomes publicly available. Our initial CG model and the detailed methods presented are intended to serve as a resource for researchers working on COVID-19 who are interested in performing multiscale simulations of the SARS-CoV-2 virion.

Significance Statement

This study reports the construction of a molecular model for the SARS-CoV-2 virion and details our multiscale approach towards model refinement. The resulting model and methods can be applied to and enable the simulation of SARS-CoV-2 virions.

Main Text

Introduction

The onset of the global COVID-19 pandemic has brought intense investigation into the molecular components of SARS-CoV-2 encoded by the virus's 30 kilobase (kb) genome. Structural biology efforts using cryo-electron microscopy (cryo-EM) and x-ray crystallographic techniques are currently reporting new structures of viral proteins every week (1–12), and computational structure prediction efforts are targeting unresolved sections of the genome using a variety of protein folding algorithms. Computational and experimental studies are underway to find new molecular therapeutics that can inhibit viral activity or further elucidate the mechanisms of action of SARS-CoV-2 proteins (13–16). The computer simulation of large-scale SARS-CoV-2 processes, such as virion assembly, budding, entry, and fusion, will remain intrinsically challenging to investigate using all-atom (AA) molecular dynamics (MD), owing to the computational cost of meaningfully simulating the hundred of millions to billions of atoms involved.

A holistic model of the SARS-CoV-2 virion can provide insight into the mechanisms of large-scale viral processes and the collective behavior of macromolecules involved in viral replication and infectivity. SARS-CoV-2 virions contain four main structural proteins: the spike (S), membrane (M), nucleocapsid (N), and envelope (E) proteins (17). S proteins are glycosylated trimers that mediate fusion and entry, in part by attaching enclosed fusion peptide sequences into the membranes of host cells (18). M proteins appear as dimeric complexes embedded within the virion envelope, and are believed to anchor ribonucleoprotein complexes to the envelope (19, 20). N proteins associate with and organize RNA into ribonucleoprotein structures found in the interior of virions (21, 22). Lastly, E proteins are thought to form pentameric ion channels that are found at the lipid bilayers of virion membranes and contribute to viral budding (23).

In this paper, we construct a largely “bottom-up” coarse-grained (CG) model of the SARS-CoV-2 virion from the currently available structural and atomistic simulation data on SARS-CoV-2 proteins. In general, this model serves as a resource for researchers working on COVID-19, and as a platform to incorporate computational and experimental data. This model also enables the multiscale study of SARS-CoV-2 processes for the development of treatment and

prevention strategies against COVID-19. Atomistic trajectory and experimental structural data deposited in the NSF Molecular Sciences Software Institute (MolSSI) will be incorporated, as they become publicly available (24). In this work, we detail several of our CG methods used to iteratively develop a CG model for the full SARS-CoV-2 virion, in which molecular interactions between CG particles are derived using a combination of phenomenological, experimental, and atomistic simulation approaches.

Theory and Methodology

Building models from structural data

We first constructed atomic models of the structural proteins of the SARS-CoV-2 virion, (Figure 1). AA models of the open and closed state of the S protein were built based on the cryo-EM structures of the spike ectodomain (PDB ID: 6VYB, 6VXX) (5), respectively, and atomic models of the N protein were constructed based on the x-ray crystallographic structure of the nucleocapsid NTD (PDB ID: 6M3M) (25). Glycosylation sites were modeled using Glycan Reader & Modeler in CHARMM-GUI (26) and the site-specific glycoprofile derived from mass spectrometry and cryo-EM analysis (27, 28). Homology models for the S protein stalk, including the HR2 and TM domains were assembled as α -helical trimeric bundles using MODELLER (29) on the basis of secondary structure assignments in JPred4 (30). Homology models for the SARS-CoV-2 N protein CTD were created from the x-ray crystallographic structure of the SARS-CoV N protein CTD (PDB ID: 2CJR) (31). Missing amino acid backbones in loop regions were built in MODELLER, and side chain rotamers were built using SCWRL4 (32). For the M protein dimer and the pentameric E ion channel, we used atomic models derived in the recent Critical Assessment of Structure Prediction (CASP) Commons competition (33, 34).

AA protein models (see discussion below) were subsequently simulated and coarse-grained to generate the CG models (Figure 2 and see sections below). A previously developed CG model for lipids was used, consisting of three CG beads per lipid and distinct bead types for lipid head groups and hydrophobic tails (35). A single component CG lipid bilayer was generated in a spherical configuration and equilibrated using CG molecular dynamics simulations under constant NVT conditions in LAMMPS (36). We note that in the future more complex CG lipid

models (37) can be added. Transmembrane segments of component membrane proteins were visually identified and assigned based on secondary structure motifs. Individual lipids on the outer leaflet of the spherical bilayer were randomly selected and used as initial positions for embedding spike, membrane, and envelope proteins. For each initial position, the center-of-mass of the transmembrane domain was aligned with the center of the lipid bilayer, and the principal axis of the protein was aligned with the vector normal to the lipid bilayer. Transmembrane regions were then substituted for the overlapping CG lipids to embed the proteins. The procedure was iterated until a spike, membrane, and envelope protein density on the virion surface was achieved that was approximately consistent with current available experimental estimates of ~25, 1000, and 20 per virion, respectively, from cryo-EM and biochemical data (38–40). The diameter of the membrane envelope is approximately 100 nm and 120 nm including the S proteins on the virion surface. As higher-resolution experimental data are released, the overall structure of this model can be refined.

AA MD simulations of the S protein

Two glycosylated models of the open and closed spike were inserted into a symmetric 225 Å x 225 Å lipid bilayer mimicking the composition of the endoplasmic reticulum-Golgi intermediate compartment (41, 42). The lipid patch was built using CHARMM-GUI. The complete protein-membrane system was solvated with using the TIP3P water model (43), and neutralized with chloride and sodium ions to maintain a 150 mM concentration. Each system contained ~1.7 million atoms. Minimization and equilibration was performed using the CHARMM36 forcefield (44, 45) and NAMD 2.14 (46). Production runs were performed in the NPT ensemble using a Langevin thermostat at 310K and Nosé-Hoover Langevin barostat at 1 atm. All production runs used a 2-fs timestep and the SHAKE algorithm. Multiple replicas of AA MD simulations of the open (3x) and closed (3x) systems were performed on NSF Frontera at the Texas Advanced Computing Center (TACC), achieving an aggregate sampling of 3.0 and 1.8 μs, respectively.

CG model of the S protein

The CG model of the glycosylated S protein (Figure 2 A) was parameterized from the AA MD simulations described below (47). Reference statistics used conformations sampled equally from both open and closed states with AA trajectories spanning the 3.0 and 1.8 μ s, respectively. First, the protein was mapped to CG beads using essential dynamics coarse-graining (EDCG) (48). We used 60 and 50 CG beads for the S_1 and S_2 domains, respectively, and the 22 N-linked glycans were each mapped to a single bead. Intra-protein interactions were represented as a hetero-elastic network model (hENM) with bond energies $k (r - r_0)^2$ where k is the spring constant of a particular CG bond and r_0 is the equilibrium bond length. These parameters were optimized using the hENM method (49). Inter-protein interactions within the S-trimers were composed of excluded volume, attractive, and screened electrostatic terms. For excluded volume interactions, a phenomenological soft cosine potential, $A \left[1 + \cos \frac{\pi r}{r_c} \right]$, was used, where $A = 25$ kcal/mol and r_c is the onset for excluded volume. Attractive, non-bonded interactions between inter-protein contacts were modeled as the sum of two Gaussian potentials, $A_1 \exp \left[-\frac{(r_{ij}-r_1)^2}{2\sigma_1^2} \right] + A_2 \exp \left[-\frac{(r_{ij}-r_2)^2}{2\sigma_2^2} \right]$, where r_1 and σ_1 are the mean and standard deviation determined by a fit to the pair correlation between CG sites i and j through least-squares regression. The constants, A_1 and A_2 , were optimized using relative-entropy minimization (REM). Screened electrostatics were modeled using Yukawa potentials, $\frac{q_i q_j}{4\pi\epsilon_r\epsilon_0 r_{ij}} \exp(-\kappa r_{ij})$, where q_i is the aggregate charge of CG site i , $\kappa = 1.274 \text{ nm}^{-1}$ is the inverse Debye length for 0.15 M NaCl, and ϵ_r is the effective dielectric constant of the protein environment, approximated as 17.5 (50).

CG models of the M and E proteins

AA simulations of the M protein dimer were performed using homology models and a membrane model based on the endoplasmic reticulum (ER). The membrane model included PC:PE:PI:PS:Chol lipids (0.45:0.10:0.23:0.10:0.12 mole fraction) as an initial approximation to the ER-Golgi intermediate compartment (ERGIC). The protein-membrane systems were solvated and neutralized in a similar fashion as described previously. The simulations were equilibrated for 400 ns prior to a 4 μ s production run on Anton2. All simulations were run in the constant NPT ensemble at 310K and 1 atm using the CHARMM36m forcefield. A CG model containing

approximately 5 residues per CG bead was mapped from the reference statistics of the AA MD simulations using the EDCG (Figure 2C), and hENM approaches. A CG model for the E protein was developed by linearly mapping the amino acid sequence to particles at a resolution of 1 CG bead per 5 amino acids (Figure 2B).

CG model of the N protein

Several studies suggest that the C-terminal domain (CTD) of the N protein assembles into a helix that contains two RNA binding grooves (21, 51). Based on these studies, we constructed atomic and CG models of the viral ribonucleoprotein complex (vRNP) by iterating between CG and AA simulations. We first constructed an atomic model of the N protein CTD helix with two RNA binding grooves by stacking 3 copies of the CTD octamer structure (PDB: 2CJR), which is composed of 4 CTD dimers and homology modeled from the X-ray crystallographic structure of the SARS-CoV N protein CTD (31). The CTD helix was simulated in the CHARMM36m force field for 400 ns. We then constructed the CTD helix model using EDCG combined with hENM followed by manually placing CG RNA beads into the groove of the helix (Figure 2D). The positions of the CG beads were used as restraints to build an atomic model of the vRNP complex. Finally, the vRNP model was relaxed and simulated in the CHARMM36m force field for 400 ns. It is important to note that recent cryo-EM studies have found granule-like densities within the virion for the vRNP complex (22). Structural detail into how CTD oligomers (including the previously proposed helical model) and RNA fit into these densities will likely require higher-resolution images.

Deriving CG molecular interactions from AA simulations

Several computational approaches have been developed to build or refine CG models using data from AA or fine-grained simulations. Our approach to coarse-graining the SARS-CoV-2 virion is to couple several CG methods in a hierarchical fashion. CG sites or “beads” are mapped from atomic structures using EDCG, a method designed to preserve the principal modes of motion sampled during atomic-level simulations (48). In EDCG, a given CG mapping operator, $\mathbf{M}_{\mathbf{R}}^{\mathbf{N}}: \mathbf{r}^n \rightarrow \mathbf{R}^N$, that relates the configurations of the atomistic trajectory ($\mathbf{r}^n$) to that of the CG model ($\mathbf{R}^N$), is variationally optimized using simulated annealing. Typically, the mapping is constructed so that

contiguous segments of a protein's primary amino acid sequence are mapped to distinct CG sites. For a fixed number of CG beads, N , the sets of atoms that are mapped to CG sites are adjusted to minimize the target residual:

$$\chi^2 = \frac{1}{3N} \sum_{l=1}^N \left\langle \sum_{i,j} |\mathbf{r}_i - \mathbf{r}_j|^2 \right\rangle_t \quad : i, j \in l, j \geq i \quad (1)$$

where, $l = 1, \dots, N$ is the CG site index, the brackets, $\langle \cdot \rangle_t$, denote a time-averaged quantity, the sum over i, j is a sum over all unique pairs in the set of atoms belonging to the CG site, l , and $\mathbf{r}_i = \mathbf{x}_i - \langle \mathbf{x}_i \rangle_t$ is the displacement of atom i from the atom's mean position, $\langle \mathbf{x}_i \rangle_t$. Note that the residual is small when the displacements, $\mathbf{r}_i$ and $\mathbf{r}_j$, are similar, i.e. the motions of atoms in the same CG site are correlated. A new map is constructed and either accepted or rejected according to a Metropolis-Hastings criterion (i.e., accepted if $\chi_n^2 < \chi_{n+1}^2$; otherwise accepted or rejected such that the new map has probability $\rho = \exp [-(\chi_{n+1}^2 - \chi_n^2)/T]$, where n is the number of iterations for simulated annealing, and T is the coupling to a fictitious temperature, that is gradually lowered during optimization).

After defining the AA $\leftrightarrow$ CG map, intramolecular interactions within a single polypeptide chain are treated using elastic network models (ENMs) to capture protein flexibility. In the hENM method (49), effective harmonic bonds are assigned to all pairs of particles in the CG model within a tunable distance cutoff that all initially have the same force constant, k_{ij} , between particle i and j to construct the bonded topology of the CG model. The harmonic force constants are optimized by first computing the normal modes of this model. In other words, solving the eigenvalue problem,

$$\mathbf{H} \mathbf{v}_k = \omega_k^2 \mathbf{M} \mathbf{v}_k \quad (2)$$

where, $\mathbf{H}$ is the Hessian: $H_{i,j} = \frac{\partial^2 V}{\partial q_i \partial q_j} \Big|_m$, $\mathbf{M}$ is the diagonal matrix for the masses of the particles, and ω_k frequency for the mode of motion. Note that this is the solution to the equation of motion

$$\mathbf{M} \frac{d^2 \mathbf{q}}{dt^2} + \mathbf{H} \mathbf{q} = 0 \quad (3)$$

where $\mathbf{q}$ is the generalized coordinate, and that for N classically, interacting particles near the potential energy minimum, $\mathbf{q}_m$:

$$V(\mathbf{q}) = V(\mathbf{q}_m) + \sum_i \left. \frac{\partial V}{\partial q_i} \right|_m (q_i - q_{i,m}) + \frac{1}{2} \sum_{i,j} \left. \frac{\partial^2 V}{\partial q_i \partial q_j} \right|_m (q_i - q_{i,m})(q_j - q_{j,m}) + O(\mathbf{q} - \mathbf{q}_m)^3 \quad (4)$$

$V(\mathbf{q}_m)$ is a constant, and $\left. \frac{\partial V}{\partial q_i} \right|_m$ is zero. Using the normal modes, mean-squared fluctuations $\langle r_{ij}^2 \rangle = \langle (x_{ij} - \langle x_{ij} \rangle)^2 \rangle$ for each i, j pair can be computed by rescaling the amplitudes according to an equipartition of energy that reflects the temperature of the atomistic data. Harmonic force constants for each bond in the CG ENM are then iteratively adjusted so that fluctuations in the CG model, match that of the atomistic data, i.e.,

$$\frac{1}{k_{ij}^{n+1}} = \frac{1}{k_{ij}^n} - \alpha (\langle r_{ij}^2 \rangle_{\text{CG}} - \langle r_{ij}^2 \rangle_{\text{AA}}) \quad (5)$$

where n is the number of iterations and α is a parameter that controls the magnitude of the adjustment for each iteration.

For the intermolecular interactions between proteins, non-bonded CG interactions are determined either using force matching (aka multiscale coarse-graining, or MS-CG) (52, 53) or REM approaches (54, 55). In MS-CG, the CG potential is constructed from a linearly independent basis set

$$U(\mathbf{R}^N) = \sum_{D=1}^{N_D} \phi_D U_D(\mathbf{R}^N) \quad (6)$$

where the functional forms for the basis potentials, U_D (e.g., B-splines, Lennard-Jones, etc), and the number of them, N_D , are determined by the user. The coefficients, $\{\phi_D\}$, are variationally optimized such that the following residual is minimized:

$$\chi^2 = \frac{1}{3N} \left\langle \sum_{l=1}^N |f_l(\mathbf{r}^n) - \mathbf{F}_l(\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}^n))|^2 \right\rangle_t \quad (7)$$

where $F_I(\mathbf{R}^n) = -\nabla U(\mathbf{R}^n)$ is the CG force and $f_I(\mathbf{r}^n)$ is the atomistic force on the CG site I . Similarly, in the REM approach, the objective function for minimization is the Kullback-Liebler divergence, which provides a metric for the differences between the atomistic and CG probability distributions

$$S_{\text{rel}} = \int \rho_{\text{AA}}(\mathbf{r}^n) \log \left(\frac{\rho_{\text{AA}}(\mathbf{r}^n)}{\rho_{\text{CG}}(\mathbf{R}^n)} \right) d\mathbf{r}^n + \langle S_{\text{map}} \rangle_{\text{AA}} \quad (8)$$

where $\rho_{\text{AA}}(\mathbf{r}^n) = Z_{\text{AA}}^{-1} e^{-\beta U_{\text{AA}}(\mathbf{r}^n)}$ and $\rho_{\text{CG}}(\mathbf{R}^n) = Z_{\text{CG}}^{-1} e^{-\beta U_{\text{CG}}(\mathbf{R}^n)}$ in the canonical ensemble, and Z is the configurational partition function. Furthermore, the relative entropy can be expressed as a difference between the potential energy and free energy of the atomistic and CG ensembles:

$$S_{\text{rel}} = \beta \langle U_{\text{CG}} - U_{\text{AA}} \rangle_{\text{AA}} - \beta (A_{\text{CG}} - A_{\text{AA}}) + \langle S_{\text{map}} \rangle_{\text{AA}} \quad (9)$$

where $A = -k_{\text{B}} T \log Z$. Minimization of the relative entropy is performed using iterative Newton-Raphson techniques. It is important to note, however, that the quality and fidelity of such CG models is determined by the molecular behavior sampled in the underlying AA simulations.

Incorporating new behavior in CG simulations

Macromolecular complexes, such as virions, undergo a wide range of behavior including physical and chemical transitions that will be difficult to capture through AA simulations alone, as well as experimental techniques. This is especially true for processes that involve large conformational changes that are not sampled effectively in AA simulations, whether due to the long timescales required, free energy barriers, or inherent limitations of the simulation force field. For instance, the S protein of SARS-CoV-2 exhibits two proteolytic cleavage sites (at the S1/S2 and S2' locations) as well as binding to the host cell receptor, angiotensin-converting enzyme 2 (ACE-2). Cleavage and binding events trigger dramatic conformational changes in the spike that result in the insertion of the fusion peptide into the host cell membrane. High-resolution structural studies of the S-ACE-2 complex have made protein binding simulations amenable to enhanced sampling techniques at the AA level (4, 56). The proteolytic cleavage of the spike and large-scale conformational shift towards fusion peptide insertion, however, are more difficult to sample in

atomistic simulations. To address these issues, one can use CG molecular simulation techniques that allow CG particles to adaptively switch discrete “states” and interactions, such as “Ultra-Coarse-Graining” (UCG) (57–59). In the limit of infrequent internal state switches, UCG implements microscopically reversible state changes are coupled to a Metropolis Hastings like criterion:

$$K_{i \rightarrow j} = k_{i \rightarrow j} \min \left[\frac{k_{j \rightarrow i}}{k_{i \rightarrow j}} e^{-\beta(U_j - U_i)}, 1 \right] \quad (1)$$

where $K_{i \rightarrow j}$ is the instantaneous switching rate from state i to j , $U_j - U_i$ is the CG effective potential energy difference between state j and i , and the rates, $k_{i \rightarrow j}$ and $k_{j \rightarrow i}$ are model parameters either treated as input or calculated from atomistic simulations. This approach is similar to hybrid kinetic Monte-Carlo and MD methods, but with a spatial kinematic component, and it can be used to examine the transitions of the spike (its “states”) that lead to the fusion of SARS-CoV-2 with host cells.

Experiments can probe longer timescales than are available from AA MD simulations. In recent cryo-EM images of SARS-CoV-2 particles, the S1 domain of the S protein was found to transiently open and close in order to bind the ACE-2 receptor (3, 5), which are subtle conformational changes that are difficult to sample in atomistic simulations. For these conformational changes – in the case that they cannot be treated as discrete state switches – plastic network models (60) or multi-configuration coarse graining (MCCG) methods (61) can be used to construct a CG model that continuously transitions from one state to the next. For plastic network models, two known experimental configurations of the protein are used to build a multi-basin ENM that represents deviations away from each of the individual conformational minima. A phenomenological interaction Hamiltonian is constructed that couples and mixes the ENMs between the two structural endpoints. In MCCG, the primary difference is that the coupling terms in the Hamiltonian are constructed from a two-state mixing approach, derived on the basis of a mapped potential of mean force which is explicitly computed from AA simulation data along collective variables that distinguish between the two (or more) conformational states at a CG level.

Phenomenological CG models

An alternative (and sometimes necessary “top-down”) route to deriving CG models is to construct a model Hamiltonian and then analyze the model’s resulting behavior in the context of the assumed interactions. Typically, parameterization of such models is designed to fit or reproduce particular observables measured in experimental data and perhaps particular sets of AA simulation data. These can be performed based on variational optimization of some system-specific functional that depends on the experimental observable. Model Hamiltonian approaches have the advantage that physical intuition is clearer, but are not systematic, since each new problem requires a different treatment for the set of interactions involved. Furthermore, these approaches often require orthogonal experiments to validate the underlying model. Such coarse-graining methods are, however, especially useful in cases where atomistic simulation and/or experimental data is difficult or infeasible to obtain on the system or if the bottom-up methods described above are not expected to yield converged results for the CG effective potential, owing to limited atomistic sampling.

Early Results

Here we present results from the first CG simulations of the SARS-CoV-2 virion (Figure 3). It should be noted that these are early results and we can thus expect additional simulations to become available from this model as more experimental data and AA simulations become available for the various virion components. In addition, the overall CG methodology and modeling of the virion will continue to evolve, and are works in progress.

A CGMD simulation was performed on the complete CG virion model using LAMMPS for 10×10^6 CG timesteps (see Supplementary Movie 1). The system was energy minimized using conjugate gradient descent. A temperature of 300K was maintained with a Langevin thermostat, with a damping constant ($t_{damp} = 10$ ps) and 100 fs timestep. Statistics were collected every 100 CG timesteps. Several radial distributions functions (RDFs) or pair correlations between CG particles were computed for the MD trajectories of individual S proteins and compared to the mapped AA reference statistics from which the models were derived (Figure 4 A). In general, the

CG model captured the positions and peaks in the pair correlation functions; however, error in the fine structure of the peaks was also present, indicating that refinement involving the addition of more expressive basis CG potentials (e.g., splines) may be necessary.

We performed principal component analysis (PCA) on a subset of the CG particles to examine collective modes of motion of the virion (Figure 4B). The Cartesian coordinates of 1 particle for every 15 CG lipids, 1 for every M and E protein, and 1 for every 3 S particles were extracted from the trajectory data and used to compute the covariance matrix, $c_{i,j} = \frac{1}{N-1} \sum_{t=1}^N r_i(t) r_j(t)$, where $r_i(t)$ is the mean-free position vector, $r_i(t) = x_i - \langle x_i \rangle_t$ of particle i . The highest variance eigenmode, PC1, (see Supplementary Movie 2) corresponds to splaying motions in the S1/S2 domain of the S protein, and accounts for 51% of the total variance seen during the simulations. Similarly, PC2 (see Supplementary Movie 3) accounts for 12.5% of the variance and corresponds to rocking motions of the S1/S2 domain, while PC3 (see Supplementary Movie 4) accounts for 7.0% of the total variance and corresponds to twisting of the S1/S2 during CGMD. In general, there was a high degree of variance in the S protein, and these correlated modes of motion are likely informative of how the virion collectively utilizes spike proteins to explore and detect receptors. Longer CG simulations with more expressive CG models will likely be required to uncover additional modes of motion in the virion, including modes that involve the structural M, N, and E proteins.

Conclusions and Outlook

This work provides an initial CG molecular model of the SARS-CoV-2 virion and details a bottom-up CG approach capable of further refining the model as new atomistic and experimental data becomes available. Currently, the lipid envelope is described using a particle-based phenomenological model with a soft tunable bending modulus well suited for large-scale membrane deformations, whereas the M and E proteins are modeled as rigid bodies. Intra-spike interactions were developed using relative entropy minimization approaches on the basis of extensive, microsecond AA simulations of the spike protein. The N protein is modeled on the basis of AA simulations of helical oligomers in complex with RNA. Cross-interactions between the lipids and structural proteins are modeled using attractive Gaussian potentials between the

hydrophobic lipid tails and the transmembrane domains of membrane proteins. This virion model will be iteratively refined and improved as structural, biochemical, and AA trajectory data are publicly released. The construction of an integrated CG model from individual atomistic simulations will also be benefited by new developments in systematic methods for ensuring consistency between CG models developed from the reference statistics of those simulations. In particular, methods that variationally optimize in a “divide and conquer” fashion on the basis of joint statistics will likely improve model fidelity. Nonetheless, despite these noted challenges, we find that the behavior of SARS-CoV-2 structural proteins is coupled in the virion.

Coarse-grained simulations of viral processes have helped elucidate a wide range of mechanisms in viruses. For example in HIV, CG simulations contributed to the understanding of the self-assembly of the capsid (62) and innate immune sensor recognition and block of viral activity (63), as well as its inhibition by drug molecules (64). Atomistic simulations of ligand binding have also revealed a variety of unexpected, drug-targetable protein-ligand interaction sites (65–71). It is likely that molecular probes into the processes the involving a holistic model of the SARS-CoV-2 virion can reveal new therapeutic strategies that exploit viral mechanisms involving large-scale behavior.

References

1. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.
2. Berman, H., K. Henrick, and H. Nakamura. 2003. Announcing the worldwide Protein Data Bank. *Nature Structural & Molecular Biology.* 10:980–980.
3. Wrapp, D., N. Wang, K.S. Corbett, J.A. Goldsmith, C.-L. Hsieh, O. Abiona, B.S. Graham, and J.S. McLellan. 2020. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science.* 367:1260–1263.
4. Yan, R., Y. Zhang, Y. Li, L. Xia, Y. Guo, and Q. Zhou. 2020. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science.* 367:1444–1448.
5. Walls, A.C., Y.-J. Park, M.A. Tortorici, A. Wall, A.T. McGuire, and D. Veasler. 2020. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell.* 181:281-292.e6.
6. Shang, J., G. Ye, K. Shi, Y. Wan, C. Luo, H. Aihara, Q. Geng, A. Auerbach, and F. Li. 2020. Structural basis of receptor recognition by SARS-CoV-2. *Nature.* 581:221–224.
7. Hillen, H.S., G. Kokic, L. Farnung, C. Dienemann, D. Tegunov, and P. Cramer. 2020. Structure of replicating SARS-CoV-2 polymerase. *Nature.* 584:154–156.
8. Kern, D.M., B. Sorum, C.M. Hoel, S. Sridharan, J.P. Remis, D.B. Toso, and S.G. Brohawn. 2020. Cryo-EM structure of the SARS-CoV-2 3a ion channel in lipid nanodiscs. *bioRxiv.* 2020.06.17.156554.
9. Surya, W., Y. Li, and J. Torres. 2018. Structural model of the SARS coronavirus E channel in LMPG micelles. *Biochimica et Biophysica Acta (BBA) - Biomembranes.* 1860:1309–1317.
10. Frick, D.N., R.S. Viridi, N. Vuksanovic, N. Dahal, and N.R. Silvaggi. 2020. Molecular Basis for ADP-Ribose Binding to the Mac1 Domain of SARS-CoV-2 nsp3. *Biochemistry.* 59:2608–2615.
11. Rut, W., Z. Lv, M. Zmudzinski, S. Patchett, D. Nayak, S.J. Snipas, F.E. Oualid, T.T. Huang, M. Bekes, M. Drag, and S.K. Olsen. 2020. Activity profiling and structures of inhibitor-bound SARS-CoV-2-PLpro protease provides a framework for anti-COVID-19 drug design. *bioRxiv.* 2020.04.29.068890.
12. Zhang, L., D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox, and R. Hilgenfeld. 2020. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α -ketoamide inhibitors. *Science.* 368:409–412.
13. Sztain, T., R. Amaro, and J.A. McCammon. 2020. Elucidation of cryptic and allosteric pockets within the SARS-CoV-2 protease. *bioRxiv.* 2020.07.23.218784.
14. Babuji, Y., B. Blaiszik, T. Brettin, K. Chard, R. Chard, A. Clyde, I. Foster, Z. Hong, S. Jha, Z. Li, X. Liu, A. Ramanathan, Y. Ren, N. Saint, M. Schwarting, R. Stevens, H. van Dam, and R. Wagner. 2020. Targeting SARS-CoV-2 with AI- and HPC-enabled Lead Generation: A First Data Release. *arXiv:2006.02431 [cs, q-bio, stat].*
15. Zimmerman, M.I., J.R. Porter, M.D. Ward, S. Singh, N. Vithani, A. Meller, U.L. Mallimadugula, C.E. Kuhn, J.H. Borowsky, R.P. Wiewiora, M.F.D. Hurley, A.M. Harbison, C.A. Fogarty, J.E. Coffland, E. Fadda, V.A. Voelz, J.D. Chodera, and G.R. Bowman. 2020. Citizen Scientists Create an Exascale Computer to Combat COVID-19. *bioRxiv.* 2020.06.27.175430.

16. Woo, H., S.-J. Park, Y.K. Choi, T. Park, M. Tanveer, Y. Cao, N.R. Kern, J. Lee, M.S. Yeom, T.I. Croll, C. Seok, and W. Im. 2020. Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein Model in a Viral Membrane. *J. Phys. Chem. B.* 124:7128–7137.
17. Masters, P.S. 2006. The Molecular Biology of Coronaviruses. In: *Advances in Virus Research*. Academic Press. pp. 193–292.
18. Cai, Y., J. Zhang, T. Xiao, H. Peng, S.M. Sterling, R.M. Walsh, S. Rawson, S. Rits-Volloch, and B. Chen. 2020. Distinct conformational states of SARS-CoV-2 spike protein. *Science*. 369:1586–1592.
19. Neuman, B.W., B.D. Adair, C. Yoshioka, J.D. Quispe, G. Orca, P. Kuhn, R.A. Milligan, M. Yeager, and M.J. Buchmeier. 2006. Supramolecular Architecture of Severe Acute Respiratory Syndrome Coronavirus Revealed by Electron Cryomicroscopy. *Journal of Virology*. 80:7918–7928.
20. Siu, Y.L., K.T. Teoh, J. Lo, C.M. Chan, F. Kien, N. Erciou, S.W. Tsao, J.M. Nicholls, R. Altmeyer, J.S.M. Peiris, R. Bruzzone, and B. Nal. 2008. The M, E, and N structural proteins of the severe acute respiratory syndrome coronavirus are required for efficient assembly, trafficking, and release of virus-like particles. *J. Virol.* 82:11318–11330.
21. Chang, C., M.-H. Hou, C.-F. Chang, C.-D. Hsiao, and T. Huang. 2014. The SARS coronavirus nucleocapsid protein--forms and functions. *Antiviral Res.* 103:39–50.
22. Yao, H., Y. Song, Y. Chen, N. Wu, J. Xu, C. Sun, J. Zhang, T. Weng, Z. Zhang, Z. Wu, L. Cheng, D. Shi, X. Lu, J. Lei, M. Crispin, Y. Shi, L. Li, and S. Li. 2020. Molecular architecture of the SARS-CoV-2 virus. *Cell*. S0092867420311594.
23. Schoeman, D., and B.C. Fielding. 2019. Coronavirus envelope protein: current knowledge. *Virology Journal*. 16:69.
24. Amaro, R.E., and A.J. Mulholland. 2020. A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19. *J. Chem. Inf. Model.* 60:2653–2656.
25. Kang, S., M. Yang, Z. Hong, L. Zhang, Z. Huang, X. Chen, S. He, Z. Zhou, Z. Zhou, Q. Chen, Y. Yan, C. Zhang, H. Shan, and S. Chen. 2020. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharmaceutica Sinica B*. 10:1228–1238.
26. Jo, S., T. Kim, V.G. Iyer, and W. Im. 2008. CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*. 29:1859–1865.
27. Watanabe, Y., J.D. Allen, D. Wrapp, J.S. McLellan, and M. Crispin. 2020. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science*. 369:330–333.
28. Shajahan, A., N.T. Supekar, A.S. Gleinich, and P. Azadi. Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. *Glycobiology*.
29. Webb, B., and A. Sali. 2016. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics*. 54:5.6.1-5.6.37.
30. Drozdetskiy, A., C. Cole, J. Procter, and G.J. Barton. 2015. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* 43:W389–W394.
31. Chen, C.-Y., C. Chang, Y.-W. Chang, S.-C. Sue, H.-I. Bai, L. Riang, C.-D. Hsiao, and T. Huang. 2007. Structure of the SARS Coronavirus Nucleocapsid Protein RNA-binding

- Dimerization Domain Suggests a Mechanism for Helical Packaging of Viral RNA. *Journal of Molecular Biology*. 368:1075–1086.
32. Krivov, G.G., M.V. Shapovalov, and R.L. Dunbrack. 2009. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics*. 77:778–795.
 33. Moulton, J., J.T. Pedersen, R. Judson, and K. Fidelis. 1995. A large-scale experiment to assess protein structure prediction methods. *Proteins*. 23:ii–v.
 34. Heo, L., and M. Feig. 2020. Modeling of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Proteins by Machine Learning and Physics-Based Refinement. *bioRxiv*.
 35. Grime, J.M.A., and J.J. Madsen. 2019. Efficient Simulation of Tunable Lipid Assemblies Across Scales and Resolutions. *arXiv:1910.05362 [cond-mat, physics:physics]*.
 36. Plimpton, S. 1995. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics*. 117:1–19.
 37. Pak, A.J., T. Dannenhoffer-Lafage, J.J. Madsen, and G.A. Voth. 2019. Systematic Coarse-Grained Lipid Force Fields with Semiexplicit Solvation via Virtual Sites. *J Chem Theory Comput*. 15:2087–2100.
 38. Ke, Z., J. Otonari, K. Qu, M. Cortese, V. Zila, L. McKeane, T. Nakane, J. Zivanov, C.J. Neufeldt, B. Cerikan, J.M. Lu, J. Peukes, X. Xiong, H.-G. Kräusslich, S.H.W. Scheres, R. Bartenschlager, and J.A.G. Briggs. 2020. Structures and distributions of SARS-CoV-2 spike proteins on intact virions. *Nature*. 1–7.
 39. Turoňová, B., M. Sikora, C. Schürmann, W.J.H. Hagen, S. Welsch, F.E.C. Blanc, S. von Bülow, M. Gecht, K. Bagola, C. Hörner, G. van Zandbergen, J. Landry, N.T.D. de Azevedo, S. Mosalaganti, A. Schwarz, R. Covino, M.D. Mühlebach, G. Hummer, J.K. Locker, and M. Beck. 2020. In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges. *Science*.
 40. Bar-On, Y.M., A. Flamholz, R. Phillips, and R. Milo. 2020. SARS-CoV-2 (COVID-19) by the numbers. *eLife*. 9:e57309.
 41. van Meer, G., D.R. Voelker, and G.W. Feigenson. 2008. Membrane lipids: where they are and how they behave. *Nature Reviews Molecular Cell Biology*. 9:112–124.
 42. Casares, D., P.V. Escribá, and C.A. Rosselló. 2019. Membrane Lipid Composition: Effect on Membrane and Organelle Structure, Function and Compartmentalization and Therapeutic Avenues. *International Journal of Molecular Sciences*. 20:2167.
 43. Jorgensen, W.L., J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.
 44. Best, R.B., X. Zhu, J. Shim, P.E.M. Lopes, J. Mittal, M. Feig, and A.D. MacKerell. 2012. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ , ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *J. Chem. Theory Comput*. 8:3257–3273.
 45. Huang, J., S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B.L. de Groot, H. Grubmüller, and A.D. MacKerell. 2017. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*. 14:71–73.

46. Phillips, J.C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé, and K. Schulten. 2005. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*. 26:1781–1802.
47. Casalino, L., Z. Gaieb, J.A. Goldsmith, C.K. Hjorth, A.C. Dommer, A.M. Harbison, C.A. Fogarty, E.P. Barros, B.C. Taylor, J.S. McLellan, E. Fadda, and R.E. Amaro. 2020. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent. Sci.*
48. Zhang, Z., L. Lu, W.G. Noid, V. Krishna, J. Pfandtnr, and G.A. Voth. 2008. A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules. *Biophys J.* 95:5073–5083.
49. Lyman, E., J. Pfandtnr, and G.A. Voth. 2008. Systematic Multiscale Parameterization of Heterogeneous Elastic Network Models of Proteins. *Biophys J.* 95:4183–4192.
50. Li, L., C. Li, Z. Zhang, and E. Alexov. 2013. On the Dielectric “Constant” of Proteins: Smooth Dielectric Function for Macromolecular Modeling and Its Implementation in DelPhi. *J. Chem. Theory Comput.* 9:2126–2136.
51. Klein, S., M. Cortese, S.L. Winter, M. Wachsmuth-Melm, C.J. Neufeldt, B. Cerikan, M.L. Stanifer, S. Boulant, R. Bartenschlager, and P. Chlanda. 2020. SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. *bioRxiv*. 2020.06.23.167064.
52. Izvekov, S., and G.A. Voth. 2005. A Multiscale Coarse-Graining Method for Biomolecular Systems. *J. Phys. Chem. B.* 109:2469–2473.
53. Noid, W.G., J.-W. Chu, G.S. Ayton, V. Krishna, S. Izvekov, G.A. Voth, A. Das, and H.C. Andersen. 2008. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* 128:244114.
54. Shell, M.S. 2008. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* 129:144108.
55. Chaimovich, A., and M.S. Shell. 2011. Coarse-graining errors and numerical optimization using a relative entropy framework. *J. Chem. Phys.* 134:094112.
56. Lan, J., J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, and X. Wang. 2020. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*. 581:215–220.
57. Dama, J.F., A.V. Sinitskiy, M. McCullagh, J. Weare, B. Roux, A.R. Dinner, and G.A. Voth. 2013. The Theory of Ultra-Coarse-Graining. 1. General Principles. *J. Chem. Theory Comput.* 9:2466–2480.
58. Davtyan, A., J.F. Dama, A.V. Sinitskiy, and G.A. Voth. 2014. The Theory of Ultra-Coarse-Graining. 2. Numerical Implementation. *J. Chem. Theory Comput.* 10:5265–5275.
59. Katkar, H.H., A. Davtyan, A.E.P. Durumeric, G.M. Hocky, A.C. Schramm, E.M. De La Cruz, and G.A. Voth. 2018. Insights into the Cooperative Nature of ATP Hydrolysis in Actin Filaments. *Biophysical Journal*. 115:1589–1602.
60. Maragakis, P., and M. Karplus. 2005. Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *J. Mol. Biol.* 352:807–822.
61. Sharp, M.E., F.X. Vázquez, J.W. Wagner, T. Dannenhoffer-Lafage, and G.A. Voth. 2019. Multiconfigurational Coarse-Grained Molecular Dynamics. *J. Chem. Theory Comput.* 15:3306–3315.

62. Grime, J.M.A., J.F. Dama, B.K. Ganer-Pornillos, C.L. Woodward, G.J. Jensen, M. Yeager, and G.A. Voth. 2016. Coarse-grained simulation reveals key features of HIV-1 capsid self-assembly. *Nature Communications*. 7:11568.
63. Yu, A., K.A. Skorupka, A.J. Pak, B.K. Ganer-Pornillos, O. Pornillos, and G.A. Voth. 2020. TRIM5 α self-assembly and compartmentalization of the HIV-1 viral capsid. *Nature Communications*. 11:1307.
64. Pak, A.J., J.M.A. Grime, A. Yu, and G.A. Voth. 2019. Off-Pathway Assembly: A Broad-Spectrum Mechanism of Action for Drugs That Undermine Controlled HIV-1 Viral Capsid Formation. *J. Am. Chem. Soc.* 141:10214–10224.
65. Schames, J.R., R.H. Henchman, J.S. Siegel, C.A. Sotriffer, H. Ni, and J.A. McCammon. 2004. Discovery of a Novel Binding Trench in HIV Integrase. *J. Med. Chem.* 47:1879–1881.
66. Dror, R.O., H.F. Green, C. Valant, D.W. Borhani, J.R. Valcourt, A.C. Pan, D.H. Arlow, M. Canals, J.R. Lane, R. Rahmani, J.B. Baell, P.M. Sexton, A. Christopoulos, and D.E. Shaw. 2013. Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs. *Nature*. 503:295–299.
67. Yu, A., and A.Y. Lau. 2018. Glutamate and Glycine Binding to the NMDA Receptor. *Structure*. 26:1035-1043.e2.
68. Yu, A., and A.Y. Lau. 2017. Energetics of Glutamate Binding to an Ionotropic Glutamate Receptor. *J. Phys. Chem. B*. 121:10436–10442.
69. Yu, A., H. Salazar, A.J.R. Plested, and A.Y. Lau. 2018. Neurotransmitter Funneling Optimizes Glutamate Receptor Kinetics. *Neuron*. 97:139-149.e4.
70. Yu, A., R. Alberstein, A. Thomas, A. Zimet, R. Grey, M.L. Mayer, and A.Y. Lau. 2016. Molecular lock regulates binding of glycine to a primitive NMDA receptor. *PNAS*. 113:E6786–E6795.
71. Yu, A., E.M.Y. Lee, J. Jin, and G.A. Voth. 2020. Atomic-scale characterization of mature HIV-1 capsid stabilization by inositol hexakisphosphate (IP6). *Science Advances*. 6:eabc6465.

Author Contributions

A.Y., A.J.P., P.H., V.M.G. and G.A.V. designed research. A.Y. performed modeling and analysis on the virion. A.Y. and A.J.P. performed modeling and analysis on the S protein. A.Y. and P.H. performed modeling and analysis on the N protein. A.Y. and V.M.G. performed modeling and analysis on the M protein. L.C., Z. G., A.C.D, and R.E.A. contributed all-atom simulation data on the S protein. A.Y., A.J.P., P.H., V.M.G., L.C., Z.G., A.C.D., R.E.A., and G.A.V. wrote the paper.

Acknowledgements

This work was supported in part by the National Science Foundation through NSF RAPID grant CHE-2029092 (A.J.P, P.H., and G.A.V), in part by the National Institute of General Medical Sciences of the National Institutes of Health through grant R01 GM063796 (V.M.G. and G.A.V.), and in part by NIH GM132826, NSF RAPID MCB-2032054, an award from the RCSA Research Corp., and a UC San Diego Moore's Cancer Center 2020 SARS-COV-2 seed grant (L.C., Z. G., A.C.D, and R.E.A.). Alvin Yu gratefully acknowledges support from the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under grant F32 AI150208. Alexander J. Pak gratefully acknowledges support from the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under grant F32 AI150477. Computational resources were provided by the Research Computing Center at the University of Chicago, Frontera at the Texas Advanced Computer Center funded by the National Science Foundation grant (OAC-1818253), and the Pittsburgh Super Computing Center through the Anton 2 machine under Grant R01GM116961 from the National Institutes of Health, and the specific allocation PSCA17046P. The Anton 2 machine at PSC was generously made available by D.E. Shaw Research.

Data Availability

The CG virion model is available at: <https://doi.datacite.org/doi/10.34974%2Fq8ya-wh69> and <https://github.com/alvinyu33/sars-cov-2-public>. The model will be periodically updated with new versions as data is added and the model refined,

Figure 1. Viral proteins of SARS-CoV-2. The genome of SARS-CoV-2 is shown in the top panel. Nonstructural proteins (NSPs) encoded in the open reading frame (ORF) 1ab are colored in orange, and the full genome is in teal. **(A)** All-atom models of the structural proteins of SARS-CoV-2 consisting of the S, E, M and N proteins. Asterisks indicate homology modeled protein structures for E and M (34). **(B)** Schematic of the virion surface from cryo-EM images of the virion, adapted from Ref. (19).

Figure 2. CG models of the SARS-CoV-2 structural proteins. (A) The CG model of the S protein trimer in the open state. The protein monomers are depicted as pink, green, and cyan beads, respectively; the monomer in pink has an exposed receptor binding domain. Each of the 22(x3) N-linked glycans are depicted as grey beads. (B) The CG model of the pentameric E protein is depicted as orange beads. (C) The CG M dimer model is depicted as yellow and blue spheres, overlaid on top of the AA model of the M dimer. Each monomer has 36 CG sites, and the red lines indicate the approximate positions of the transmembrane region. (D) The CG model of the N protein CTD helix in complex with viral RNA. The N protein helix and bonds derived from the hENM are depicted in cyan, while the RNA is depicted as orange beads.

Figure 3. A multiscale model of the SARS-CoV-2 virion. (A) Exterior view of the SARS-CoV-2 virion. **(B)** Interior view of the SARS-CoV-2 virion. Spike (S) protein trimers are depicted in teal with the glycosylation sites represented as black spheres. Membrane (M) protein dimers are in blue, with pentameric envelope (E) ion channels in orange. The density of S,M, and E proteins was chosen to be consistent experiments (38–40). N proteins are not shown. The diameter of the membrane envelope is approximately 100 nm and 120 nm including the S proteins on the virion surface.

Figure 4: Analysis of the CGMD simulations of the SARS-CoV-2 virion. **(A)** Radial distribution functions (RDFs) showing the comparison between mapped all-atom reference statistics and the CG spike model during the MD simulations. The measured RDFs are for CG particles that were mapped from the following all-atom residues of (i) $S_{1,RBD}$ [S459–D467] and $S_{1,NTD}$ [W104–L118], (ii) $S_{1,preRBD}$ [E309–R319] and S_2 [A852–L861], (iii) $S_{2,CH}$ [A1015–K1028], and (iv) $S_{2,CTD}$ [Y1215–V1228]. **(B)** Principal modes of motion of the SARS-CoV-2 virion computed from the CGMD simulation (see also Supplementary Movie 2–4). The first principal component (PC1) accounts for 51% of the total variation observed during simulation, whereas the second (PC2) and third (PC3) account for 12.5% and 7%, respectively.