Biophysical Journal
Article

Biophysical Society

A multiscale coarse-grained model of the SARS-CoV-2 virion

Alvin Yu,[1] Alexander J. Pak,[1] Peng He,[1] Viviana Monje-Galvan,[1] Lorenzo Casalino,[2] Zied Gaieb,[2] Abigail C. Dommer,[2] Rommie E. Amaro,[2] and Gregory A. Voth[1,*]
[1]Department of Chemistry, Chicago Center for Theoretical Chemistry, Institute for Biophysical Dynamics and James Franck Institute, The University of Chicago, Chicago, Illinois and [2]Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, California

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 are constrained to shorter timescales and require billion-atom simulations for these processes. Here, we report the current status and ongoing development of a largely "bottom-up" coarse-grained (CG) model of the SARS-CoV-2 virion. 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 become 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 This study reports the construction of a molecular model for the SARS-CoV-2 virion and details our multiscale approach toward model refinement. The resulting model and methods can be applied to and enable the simulation of SARS-CoV-2 virions.

INTRODUCTION

The onset of the global coronavirus disease 2019 (COVID-19) pandemic has brought intense investigation into the molecular components of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) encoded by the virus's 30-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 hundreds 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

Submitted October 1, 2020, and accepted for publication October 30, 2020.
*Correspondence: gavoth@uchicago.edu
Editor: Tamar Schlick.
https://doi.org/10.1016/j.bpj.2020.10.048

bilayers of virion membranes and contribute to viral budding (23).

In this work, 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 new multiscale studies of SARS-CoV-2 processes to possibly help find treatment and prevention strategies against COVID-19. Atomistic trajectory and experimental structural data deposited in the National Science Foundation (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.

METHODS

Building models from structural data

We first constructed atomic models of the structural proteins of the SARS-CoV-2 virion (Fig. 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 (Protein Data Bank, PDB: 6VYB, 6VXX) (5), respectively, and atomic models of the N protein were constructed based on the x-ray crystallographic structure of the nucleocapsid N-terminal domain (NTD) (PDB: 6M3M) (27). Glycosylation sites were modeled using Glycan Reader & Modeler in CHARMM-GUI (28) and the site-specific glycoprofile derived from mass spectrometry and cryo-EM analysis (29,30). Homology models for the S-protein stalk, including the HR2 and TM domains, were assembled as α-helical trimeric bundles using MODELER (31) on the basis of secondary structure assignments in JPred4 (32). Homology models for the SARS-CoV-2 N protein C-terminal domain (CTD) were created from the x-ray crystallographic structure of the SARS-CoV N protein CTD (PDB: 2CJR) (33). Missing amino acid backbones in loop regions were built in MODELER, and side chain rotamers were built using SCWRL4 (34). We used atomic models for the M-protein dimer (25) and the pentameric E ion channel (26) that were developed by homology.

AA protein models (see discussion below) were subsequently simulated and coarse grained to generate the CG models (see Fig. 2 and sections below). A previously developed CG model for lipids was used, consisting of three CG beads per lipid and distinct bead types for lipid headgroups and hydrophobic tails (35). A single-component CG lipid bilayer was generated in a spherical configuration and equilibrated using CG MD 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 diameters of the membrane envelope are ~100 and 140 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 × 225 Å lipid bilayer mimicking the composition of the endoplasmic reticulum (ER)-Golgi intermediate compartment (41,42). The lipid patch was built using CHARMM-GUI. The complete protein-membrane system was solvated 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 were performed using the CHARMM36 force field (44,45) and NAMD 2.14 (46). Production runs were performed in the NPT ensemble using a Langevin thermostat at 310 K 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 (3×) and closed (3×) 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 (Fig. 2 *A*) was parameterized from the AA MD simulations described above (47). Reference statistics used conformations sampled equally from both open and closed states,

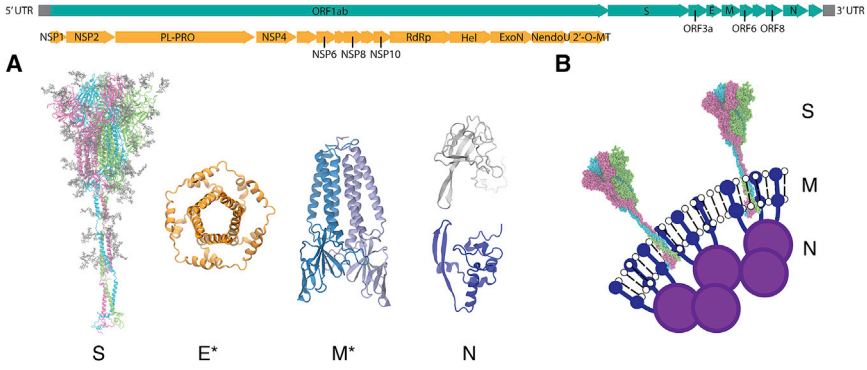

FIGURE 1 Viral proteins of SARS-CoV-2. The genome of SARS-CoV-2 is shown in the top panel. Nonstructural proteins (NSPs) encoded in the open reading frame (ORF) 1ab are colored in orange, and the full genome is in teal. (*A*) AA models of the structural proteins of SARS-CoV-2 consisting of the S, E, M, and N proteins are given. Asterisks indicate homology-modeled protein structures for M and E (25,26). (*B*) A schematic of the virion surface from cryo-EM images of the virion is given, adapted from (19). To see this figure in color, go online.

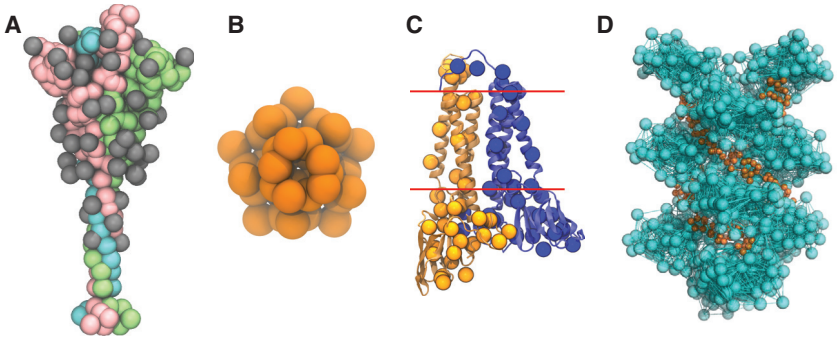

FIGURE 2 CG models of the SARS-CoV-2 structural proteins. (*A*) The CG model of the S-protein trimer in the open state is shown. The protein monomers are depicted as pink, green, and cyan beads, respectively; the monomer in pink has an exposed receptor binding domain. Each of the 22 (× 3) N-linked glycans are depicted as gray 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 is shown. The N protein helix and bonds derived from the hENM are depicted in cyan, and the RNA is depicted as orange beads. To see this figure in color, go online.

with AA trajectories spanning 3.0 and 1.8 μs, 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. Intraprotein interactions were represented as a heteroelastic 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). Interprotein interactions within the S-trimers were composed of excluded volume, attractive, and screened electrostatic terms. For excluded volume interactions, a phenomenological soft cosine potential, $A[1 + \cos\frac{\pi r}{r_c}]$, was used, where $A = 25$ kcal/mol and r_c is the onset for excluded volume. Attractive, nonbonded interactions between interprotein 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, $q_i q_j / (4\pi\varepsilon_r\varepsilon_0 r_{ij})\exp(-\kappa r_{ij})$, where q_i is the aggregate charge of CG site i, $\kappa = 1.274$ 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 ER. The membrane model included PC:PE:PI:PS:Chol lipids (0.45:0.10:0.23:0.10:0.12 mol 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 before a 4-μs production run on Anton 2. All simulations were run in the constant NPT ensemble at 310 K and 1 atm using the CHARMM36m force field. A CG model containing ~5 residues per CG bead was mapped from the reference statistics of the AA MD simulations using the EDCG (Fig. 2 *C*), 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 five amino acids (Fig. 2 *B*).

CG model of the N protein

Several studies suggest that the 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 three copies of the CTD octamer structure (PDB: 2CJR), which is composed of four CTD dimers and homology modeled from the x-ray crystallographic structure of the SARS-CoV N protein CTD (33). 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 (Fig. 2 *D*). 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, $\boldsymbol{M}_{\mathbf{R}}^{N} : \mathbf{r}^n \rightarrow \mathbf{R}^N$, 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:

$$\chi^2 = \frac{1}{3N}\sum_{I=1}^{N}\left\langle \sum_{i,j} \left| \boldsymbol{r}_i - \boldsymbol{r}_j \right|^2 \right\rangle_t : i,j \,\varepsilon\, I,\ j \geq i, \qquad (1)$$

where $I = 1, \ldots, N$ 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, I; and $\boldsymbol{r}_i = \boldsymbol{x}_i - \langle \boldsymbol{x}_i \rangle_t$ is the displacement of atom i from the atom's mean position, $\langle \boldsymbol{x}_i \rangle_t$. Note that the residual is small when the displacements, $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$, 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 particles 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,

$$\boldsymbol{H}\boldsymbol{v}_{\boldsymbol{k}} = \omega_{\boldsymbol{k}}^2 \boldsymbol{M}\boldsymbol{v}_{\boldsymbol{k}}, \tag{2}$$

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

$$\boldsymbol{M}\frac{\mathrm{d}^2\boldsymbol{q}}{\mathrm{d}t^2} + \boldsymbol{H}\boldsymbol{q} = 0, \tag{3}$$

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

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

$V(\boldsymbol{q}_m)$ is a constant, and $\frac{\partial V}{\partial q_i}\Big|_m$ 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.,

$$\frac{1}{k_{ij}^{n+1}} = \frac{1}{k_{ij}^{n}} - \alpha\left(\left\langle r_{ij}^2 \right\rangle_{\mathrm{CG}} - \left\langle r_{ij}^2 \right\rangle_{\mathrm{AA}}\right), \tag{5}$$

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, nonbonded CG interactions are determined either using force matching (a.k.a. multiscale CG) (52,53) or REM approaches (54,55). In multiscale CG, the CG potential is constructed from a linearly independent basis set

$$U\left(\boldsymbol{R}^N\right) = \sum_{D=1}^{N_D} \phi_D U_D\left(\boldsymbol{R}^N\right), \tag{6}$$

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:

$$\chi^2 = \frac{1}{3N}\left\langle \sum_{I=1}^{N} \left| \boldsymbol{f}_{\boldsymbol{I}}(\boldsymbol{r}^n) - \boldsymbol{F}_{\boldsymbol{I}}\left(\boldsymbol{M}_{\mathbf{R}}^N(\boldsymbol{r}^n)\right) \right|^2 \right\rangle_t, \tag{7}$$

where $\boldsymbol{F}_{\boldsymbol{I}}(\boldsymbol{R}^N) = -\nabla U(\boldsymbol{R}^N)$ is the CG force and $\boldsymbol{f}_{\boldsymbol{I}}(\boldsymbol{r}^n)$ 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

$$S_{\mathrm{rel}} = \int \rho_{\mathrm{AA}}(\boldsymbol{r}^n)\log\left(\frac{\rho_{\mathrm{AA}}(\boldsymbol{r}^n)}{\rho_{\mathrm{CG}}(\boldsymbol{R}^N)}\right)d\boldsymbol{r}^n + \left\langle S_{\mathrm{map}}\right\rangle_{\mathrm{AA}}, \tag{8}$$

where $\rho_{\mathrm{AA}}(\boldsymbol{r}^n) = Z_{\mathrm{AA}}^{-1} e^{-\beta U_{\mathrm{AA}}(\boldsymbol{r}^n)}$ and $\rho_{\mathrm{CG}}(\boldsymbol{R}^N) = Z_{\mathrm{CG}}^{-1} e^{-\beta U_{\mathrm{CG}}(\boldsymbol{R}^N)}$ 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:

$$S_{\mathrm{rel}} = \beta\langle U_{\mathrm{CG}} - U_{\mathrm{AA}}\rangle_{\mathrm{AA}} - \beta(A_{\mathrm{CG}} - A_{\mathrm{AA}}) + \left\langle S_{\mathrm{map}}\right\rangle_{\mathrm{AA}}, \tag{9}$$

where $A = -k_{\mathrm{B}}T\log Z$. 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 are 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 or even with experimental techniques. This is especially true for processes that involve large conformational changes that are not sampled effectively in AA simulations, whether because of the long timescales required, free energy barriers, or inherent limitations of the simulation force field. For instance, the S protein of SARS-CoV-2 has two proteolytic cleavage sites (at the S1-S2 and S2′ locations), and binds 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 toward 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 that are coupled to a Metropolis-Hastings-like criterion:

$$K_{i\to j} = k_{i\to j}\min\left[\frac{k_{j\to i}}{k_{i\to j}}e^{-\beta\left(U_j - U_i\right)}, 1\right] \tag{10}$$

where $K_{i\to j}$ is the instantaneous switching rate from state i to j, $U_j - U_i$ is the CG effective potential energy difference between states j and i, and the rates $k_{i\to j}$ and $k_{j\to i}$ 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 (i.e. "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 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 multiconfiguration 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 multibasin 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 that 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 because 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 for which atomistic simulation 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 effective CG potential, owing to limited atomistic sampling.

RESULTS AND DISCUSSION

Here, we present results from the first CG simulations of the SARS-CoV-2 virion (Fig. 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 CG MD simulation was performed on the complete CG virion model using LAMMPS for 10×10^6 CG time steps (see Video S1). The system was energy minimized using conjugate gradient descent. A temperature of 300 K was maintained with a Langevin thermostat, with a damping constant ($t_{damp} = 10$ ps) and 100-fs timestep. Statistics were collected every 100 CG time steps. Several radial distribution 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 (Fig. 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 CG basis 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 (Fig. 4 *B*). The Cartesian coordinates of one particle for every 15 CG lipids, one for every M and E protein, and one 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 Video S2), 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 Video S3) accounts for 12.5% of the variance and corresponds to rocking motions of the S1-S2 domain, whereas PC3 (see Video S4) accounts for 7.0% of the total variance and corresponds to twisting of the S1-S2 during CG MD. 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.

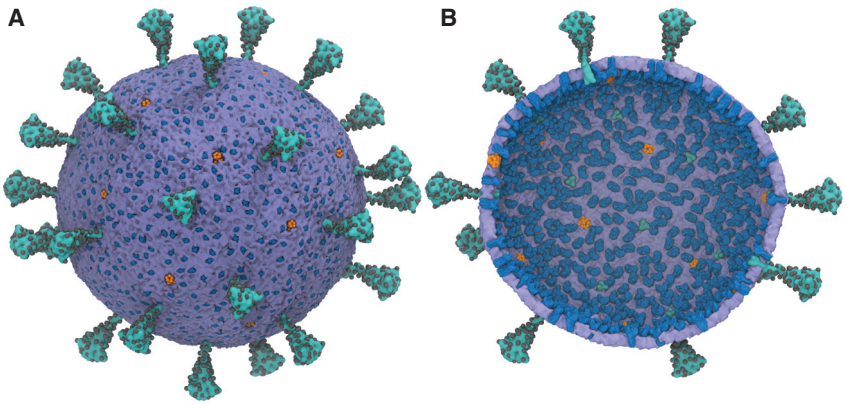

FIGURE 3 A multiscale model of the SARS-CoV-2 virion. (*A*) Exterior view of the SARS-CoV-2 virion is given. (*B*) Interior view of the SARS-CoV-2 virion is given. S-protein trimers are depicted in teal, with the glycosylation sites represented as black spheres. M-protein dimers are in blue, with pentameric E ion channels in orange. The density of S, M, and E proteins was chosen to be consistent with experiments (38–40). The diameters of the membrane envelope are ~100 and 140 nm including the S proteins on the virion surface. To see this figure in color, go online.

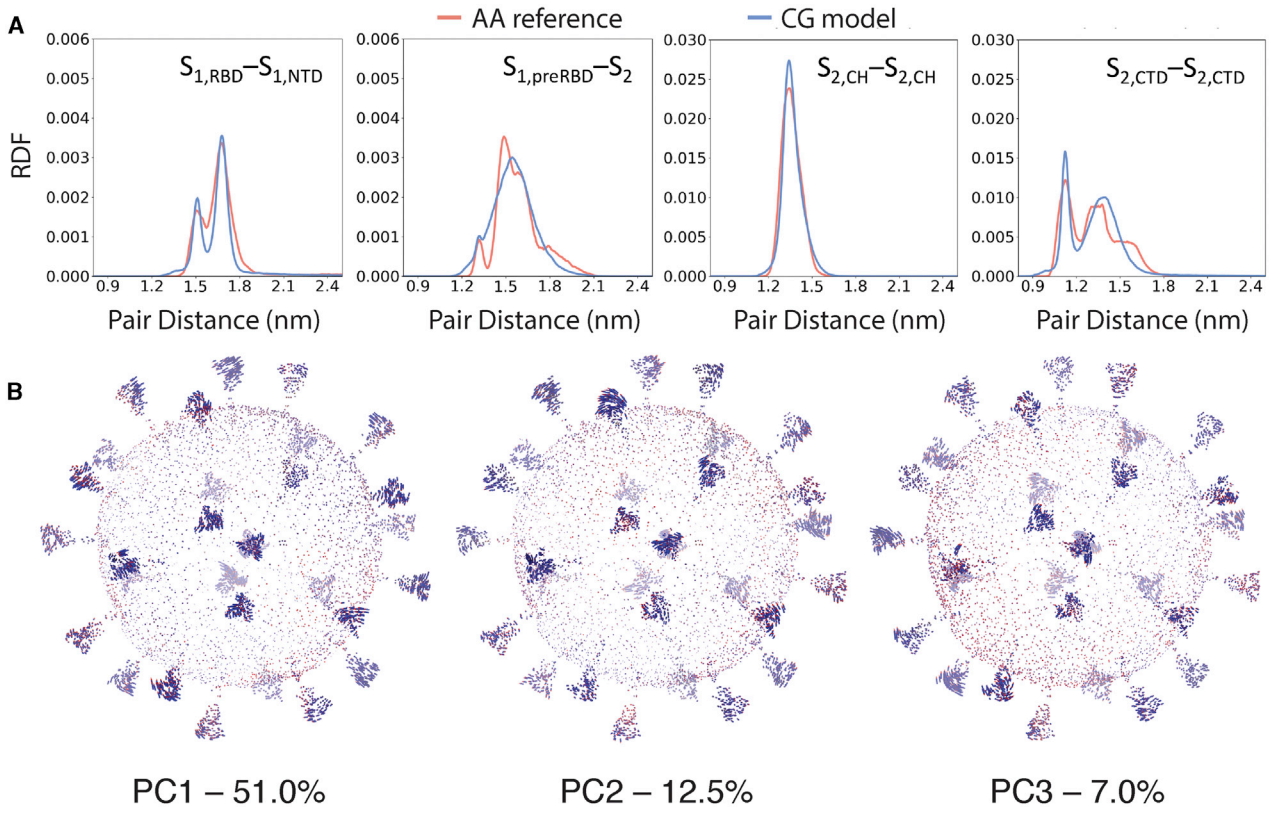

FIGURE 4 Analysis of the CG MD simulations of the SARS-CoV-2 virion. (*A*) RDFs showing the comparison between mapped AA reference statistics and the CG spike model during the MD simulations are given. The measured RDFs are for CG particles that were mapped from the following AA residues of 1) $S_{1,RBD}$ [S459-D467] and $S_{1,NTD}$ [W104-L118], 2) $S_{1,preRBD}$ [E309-R319] and S_2 [A852-L861], 3) $S_{2,CH}$ [A1015-K1028], and 4) $S_{2,CTD}$ [Y1215-V1228]. (*B*) Principal modes of motion of the SARS-CoV-2 virion computed from the CG MD simulation are shown. Arrows are colored from blue to red, indicating the direction of movement (see Videos S2, S3, and S4 for PC1–3, respectively). 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. To see this figure in color, go online.

CONCLUSIONS

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 become 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. Intraspike interactions were developed using REM 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 used 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 benefit from 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.

CG 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 involving a holistic model of the SARS-CoV-2 virion can help reveal new routes to combat the virus by exploiting viral mechanisms involving large-scale behavior.

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

SUPPORTING MATERIAL

Supporting Material can be found online at https://doi.org/10.1016/j.bpj.2020.10.048.

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 AA 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 manuscript.

ACKNOWLEDGMENTS

This work was supported in part by the NSF 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 National Institutes of Health GM132826, NSF RAPID MCB-2032054, an award from RCSA Research, 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.). A.Y. acknowledges support from the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under grant F32 AI150208. A.J.P. 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 NSF 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.

REFERENCES

1. Berman, H. M., J. Westbrook, …, 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. *Nat. Struct. Biol.* 10:980.
3. Wrapp, D., N. Wang, …, 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, …, 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, …, D. Veesler. 2020. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell.* 181:281–292.e6.
6. Shang, J., G. Ye, …, F. Li. 2020. Structural basis of receptor recognition by SARS-CoV-2. *Nature.* 581:221–224.
7. Hillen, H. S., G. Kokic, …, P. Cramer. 2020. Structure of replicating SARS-CoV-2 polymerase. *Nature.* 584:154–156.
8. Kern, D. M., B. Sorum, …, S. G. Brohawn. 2020. Cryo-EM structure of the SARS-CoV-2 3a ion channel in lipid nanodiscs. *bioRxiv* https://doi.org/10.1101/2020.06.17.156554.
9. Surya, W., Y. Li, and J. Torres. 2018. Structural model of the SARS coronavirus E channel in LMPG micelles. *Biochim. Biophys. Acta Biomembr.* 1860:1309–1317.
10. Frick, D. N., R. S. Virdi, …, 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, …, 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* https://doi.org/10.1101/2020.04.29.068890.
12. Zhang, L., D. Lin, …, 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* https://doi.org/10.1101/2020.07.23.218784.
14. Babuji, Y., B. Blaiszik, …, R. Wagner. 2020. Targeting SARS-CoV-2 with AI- and HPC-enabled lead generation: a first data release. *arXiv*, arXiv:2006.02431 https://arxiv.org/abs/2006.02431.
15. Zimmerman, M. I., J. R. Porter, …, G. R. Bowman. 2020. Citizen scientists create an exascale computer to combat COVID-19. *bioRxiv* https://doi.org/10.1101/2020.06.27.175430.
16. Woo, H., S.-J. Park, …, 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. K. Maramorosch and A. J. Shatkin, eds. Academic Press, pp. 193–292.
18. Cai, Y., J. Zhang, …, B. Chen. 2020. Distinct conformational states of SARS-CoV-2 spike protein. *Science.* 369:1586–1592.
19. Neuman, B. W., B. D. Adair, …, M. J. Buchmeier. 2006. Supramolecular architecture of severe acute respiratory syndrome coronavirus revealed by electron cryomicroscopy. *J. Virol.* 80:7918–7928.
20. Siu, Y. L., K. T. Teoh, …, 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. K., M.-H. Hou, …, T. H. Huang. 2014. The SARS coronavirus nucleocapsid protein–forms and functions. *Antiviral Res.* 103:39–50.
22. Yao, H., Y. Song, …, S. Li. 2020. Molecular architecture of the SARS-CoV-2 virus. *Cell.* 183:730–738.e13.
23. Schoeman, D., and B. C. Fielding. 2019. Coronavirus envelope protein: current knowledge. *Virol. J.* 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. 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* https://doi.org/10.1101/2020.03.25.008904.
26. Srinivasan, S., H. Cui, …, D. Korkin. 2020. Structural genomics of SARS-CoV-2 indicates evolutionary conserved functional regions of viral proteins. *Viruses.* 12:360.
27. Kang, S., M. Yang, …, S. Chen. 2020. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharm. Sin. B.* 10:1228–1238.
28. Jo, S., T. Kim, …, W. Im. 2008. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* 29:1859–1865.
29. Watanabe, Y., J. D. Allen, …, M. Crispin. 2020. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science.* 369:330–333.
30. Shajahan, A., N. T. Supekar, …, P. Azadi. 2020. Deducing the N- and O- glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. *Glycobiology*, cwaa042. Published online May 4, 2020.
31. Webb, B., and A. Sali. 2016. Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinformatics.* 54:5.6.1–5.6.37.
32. Drozdetskiy, A., C. Cole, …, G. J. Barton. 2015. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* 43:W389–W394.

33. Chen, C.-Y., C. K. Chang, …, T. H. Huang. 2007. Structure of the SARS coronavirus nucleocapsid protein RNA-binding dimerization domain suggests a mechanism for helical packaging of viral RNA. *J. Mol. Biol.* 368:1075–1086.

34. Krivov, G. G., M. V. Shapovalov, and R. L. Dunbrack, Jr. 2009. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins.* 77:778–795.

35. Grime, J. M. A., and J. J. Madsen. 2019. Efficient simulation of tunable lipid assemblies across scales and resolutions. *arXiv*, arXiv:1910.05362 https://arxiv.org/abs/1910.05362.

36. Plimpton, S. 1995. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* 117:1–19.

37. Pak, A. J., T. Dannenhoffer-Lafage, …, 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. Oton, …, J. A. G. Briggs. 2020. Structures and distributions of SARS-CoV-2 spike proteins on intact virions. *Nature*, Published online August 17, 2020.

39. Turoňová, B., M. Sikora, …, M. Beck. 2020. In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges. *Science.* 370:203–208.

40. Bar-On, Y. M., A. Flamholz, …, 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. *Nat. Rev. Mol. Cell Biol.* 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. *Int. J. Mol. Sci.* 20:2167.

43. Jorgensen, W. L., J. Chandrasekhar, …, 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, …, A. D. Mackerell, Jr. 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, …, A. D. MacKerell, Jr. 2017. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods.* 14:71–73.

46. Phillips, J. C., R. Braun, …, K. Schulten. 2005. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26:1781–1802.

47. Casalino, L., Z. Gaieb, …, R. E. Amaro. 2020. Beyond shielding: the roles of glycans in the SARS-CoV-2 spike protein. *ACS Cent. Sci.* 6:1722–1734.

48. Zhang, Z., L. Lu, …, G. A. Voth. 2008. A systematic methodology for defining coarse-grained sites in large biomolecules. *Biophys. J.* 95:5073–5083.

49. Lyman, E., J. Pfaendtner, 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, …, 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, …, P. Chlanda. 2020. SARS-CoV-2 structure and replication characterized by *in situ* cryo-electron tomography. *bioRxiv* https://doi.org/10.1101/2020.06.23.167064.

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, …, 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, …, 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, …, 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, …, 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, …, G. A. Voth. 2018. Insights into the cooperative nature of ATP hydrolysis in actin filaments. *Biophys. J.* 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, …, G. A. Voth. 2019. Multiconfigurational coarse-grained molecular dynamics. *J. Chem. Theory Comput.* 15:3306–3315.

62. Grime, J. M. A., J. F. Dama, …, G. A. Voth. 2016. Coarse-grained simulation reveals key features of HIV-1 capsid self-assembly. *Nat. Commun.* 7:11568.

63. Yu, A., K. A. Skorupka, …, G. A. Voth. 2020. TRIM5α self-assembly and compartmentalization of the HIV-1 viral capsid. *Nat. Commun.* 11:1307.

64. Pak, A. J., J. M. A. Grime, …, 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. 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, …, 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. Y. Lau. 2018. Neurotransmitter funneling optimizes glutamate receptor kinetics. *Neuron.* 97:139–149.e4.

70. Yu, A., R. Alberstein, …, A. Y. Lau. 2016. Molecular lock regulates binding of glycine to a primitive NMDA receptor. *Proc. Natl. Acad. Sci. USA.* 113:E6786–E6795.

71. Yu, A., E. M. Y. Lee, …, G. A. Voth. 2020. Atomic-scale characterization of mature HIV-1 capsid stabilization by inositol hexakisphosphate (IP_6). *Sci. Adv.* 6:eabc6465.