Toward Atomistic Models of Intact SARS-CoV-2 via Martini Coarse-Grained Molecular Dynamics Simulations

Dali Wang[1,2,†], Jiaxuan Li[1,†], Lei Wang[1], Yipeng Cao[3,4], Sai Li[2,5,6,*], and Chen Song[1,2,*]

[1]Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

[2]Peking-Tsinghua Center for Life Sciences, Beijing, China

[3]Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin 300060, China

[4]National Supercomputer Center in Tianjin, Tianjin 300457, China

[5]School of Life Sciences, Tsinghua University, Beijing 100084, China

[6]Beijing Advanced Innovation Center for Structural Biology & Frontier Research Center for Biological Structure, Beijing 100084, China

[†]These authors contributed equally to this work.

[*]E-mail: c.song@pku.edu.cn (CS), sai@mail.tsinghua.edu.cn (SL)

Abstract

The causative pathogen of Coronavirus disease 2019 (COVID-19), severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is an enveloped virus assembled by a lipid envelope and multiple structural proteins. In this study, by integrating experimental data, structural modeling, and coarse-grained molecular dynamics simulations, we constructed multiscale models of SARS-CoV-2. Our 500-ns coarse-grained simulation of the intact virion allowed us to investigate the dynamic behavior of the membrane-embedded proteins and the surrounding lipid molecules *in situ*. Our results indicated that the membrane-embedded proteins are highly dynamic, and certain types of lipids exhibit various binding preferences to specific sites of the membrane-embedded proteins. The equilibrated virion model was transformed into atomic resolution, which provided a 3D structure for scientific demonstration and can serve as a framework for future exascale all-atom MD simulations.

1 Introduction

The ongoing Coronavirus disease 2019 (COVID-19) pandemic has infected a massive amount of people globally in the past few years. The causative pathogen, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is an enveloped virus assembled by a lipid envelope, a positive-sense single-stranded RNA, and four structural proteins: the spike (S), membrane (M), envelope (E), and nucleocapsid (N) proteins. For the purpose of understanding the molecular basis of viral functions, assembly, virus-host interactions, and antibody neutralization, extensive studies have been carried out to solve the structures *in vitro* for the SARS-CoV-2 viral proteins by cryo-electron microscopy (cryo-EM) or crystallography, and structures *in situ* for the native proteins by cryo-electron tomography (cryo-ET). Although recent technical developments of cryo-ET have enabled the reconstruction of intact SARS-CoV-2, the structure has been limited to nanometer resolution [1, 2].

In the meanwhile, computational studies have also provided highly valuable information on the structure and dynamics of the virus [3–7], especially the pioneering work by the Voth lab [3] and the ground-breaking AI-enabled multi-scale simulations by a large team led by the Amaro lab

[4, 6]. However, the existing structural models of the virus have been either limited to a coarse-grained (CG) scale, focusing primarily on the virus envelope, or constructed without considering the specific protein localization from the *in situ* cryo-ET data, particularly the N proteins. Therefore, we set out to construct both coarse-grained (CG) and atomistic models of SARS-CoV-2 that are as intact as possible, by fully employing the latest cryo-ET data [1], the available experimentally resolved protein structures, structure prediction and modeling methods, as well as coarse-grained molecular dynamics (MD) simulations. The CG and atomistic models can not only provide 3D structures for scientific demonstration, but also offer a framework for future exascale MD simulations to understand the dynamics of the intact virus, its assembly, and its mutations.

To obtain a better equilibrated model of SARS-CoV-2, we first built a CG model and then equilibrated it with the Martini force field [8], followed by a resolution transformation. To build the CG model, we constructed structural models of each protein component separately. Then we assembled them onto a pre-equilibrated lipid envelope according to the architecture of the intact virus revealed by cryo-ET [1]. Since there is currently no way to solve or predict the full-length RNA structures within the envelope, only the N-bound RNA segments were considered in our model. The CG SARS-CoV-2 model was solved in a water box, equilibrated by a 500-ns CG MD simulation, and then transformed into the atomic resolution. Two atomistic models of the intact virus are provided: the initial structure, built according to the cryo-ET map prior to the CG MD simulation (Fig. 1a & S1c), and the final structure, built after a 500-ns CG MD simulation (Fig. S1d). Although the CG simulation was not able to quantitatively characterize the conformational changes of the viral proteins, we can obtain key information regarding the dynamic properties of the structural proteins on the envelope, such as the interactions between the membrane-embedded proteins, the diffusion coefficients of the membrane-embedded proteins, as well as the lipid clustering around them. Therefore, the CG simulation not only efficiently equilibrated the virion model for the following resolution transformation toward an atomistic model, but also provided valuable insights on the protein-membrane interaction of an intact SARS-CoV-2 virion as well as the overall dynamics of each viral component on the envelope.

2 Results

Based on the cryo-ET map (EMD-30430) [1], we constructed a vesicle with a rough diameter of 85 nm as the viral envelope and assembled a virion model (Fig. 2a). Using the Martini Force Field, we were able to obtain a well-equilibrated system, in which the vesicle reached a converged size within 200 ns in the CG simulation (Fig. 2b). In the previous work, Yao et al. analyzed the SARS-CoV-2 envelope size based on 2,294 virions and demonstrated that the viral envelope shape is ellipsoidal [1]. The virion diameter measured from our CG trajectory differ slightly from the statistical data, but overall matched well with the cryo-ET map (EMD-30430).

The protein-membrane interaction is critical along the whole life cycle of SARS-CoV-2. During the assembly stage of the virus in host cells, the membrane-bound E, M and S proteins on the ERGIC (ER-golgi intermediate compartment) recruit the viral RNPs, together budding into the ERGIC and forming new virions [9]. After the structural proteins insertion, the lipid molecules of the envelope may rearrange to find their favorite positions and cluster around the membrane-embedded proteins. We analyzed the protein-lipid interface on the virion envelope to identify the specific lipid-binding sites. We calculated the radial distribution function (RDF) of lipids around the membrane-embedded proteins. Our analysis showed that each lipid component had a similar RDF profile at the very beginning of the simulation, representing the randomly distributed lipid molecules before equilibration (top panels of Fig. 3b-d). Along with the CG simulation, the RDF profiles of various lipid molecules gradually changed and converged. As shown in Fig. 3, the RDFs calculated from the 300-400 ns trajectory (middle panels in Fig. 3b-d) and 400-500 ns trajectory (bottom panels in Fig. 3b-d) were already indistinguishable, indicating that the lipid distribution had well converged.

Although the transmembrane domains (TMDs) are very different among the M, S, and E proteins, the converged lipid distribution around them showed common features. Relatively, Phosphatidylinositol (PI) and Phosphatidylserine (PS) were more frequently detected around the proteins than phosphatidylcholine (PC) and phosphatidylethanolamine (PE) (Fig. 3), which illustrated that the negatively charged lipids have a stronger binding preference toward the membrane proteins in SARS-CoV-2. It was also noticeable that more PI molecules enrich around the proteins than PS, although they carry the same charge of -1 e. Further analyses showed the PI-enriched region distributes many aromatic residues (Fig. 4), illustrating that the Pi-Pi stack interaction between PI head region and aromatic residue side chain probably plays a critical role in PI-protein binding. In addition, cholesterol (CHOL) showed the second-high probability of surrounding the membrane-embedded proteins (orange lines in Fig. 3b-d). These results are consistent with previous work showing that PI and CHOL prefer concentrating around the membrane proteins [10], and a recent work by Wang et al. also reported that PI and CHOL have tendency to locate near the viral S, M, and E proteins [7]. Therefore, PI, CHOL, and PS molecules constitute the preferred surrounding environment of the TMDs of membrane proteins. The stable protein-lipid interaction interface benefits to the embedding of the structural proteins in the viral envelope and maintains the entire virus architecture during the virus life cycle. Taken together, these results indicate that the massive membrane protein insertion would significantly influence the lipid's distribution on the virion surface, eventually forming a highly heterogeneous distribution on the stable virion envelope.

With the CG MD trajectory, we further analyzed how the lipids distribute around each residue of the membrane proteins to identify the specific binding sites. We calculated the lipid contact probability of each residue, and the results are shown in Fig. 4. For the S proteins (Fig. 4a), the residues in the inner and outer leaflets prefer different lipid neighbors. The CHOLs tend to locate near the inner side (M1233, L1234, C1235, C1236), while the PI lipids distribute more densely on the outer side, around W1212, P1213, W1214, and Y1215. Interestingly, we also detected that a few PC molecules were gathering near the W1212 and W1214, indicating that PC and PI may share these common binding sites. The same phenomena were observed around the E proteins (Fig. 4b), for which PI and PC lipid concentrate in the outer leaflet around S4, F6, E7, E8, while the inner-leaflet residues A36, C40, and A41 attract CHOLs. As for the lipid distribution around the M proteins (Fig. 4c), we did not observe the same asymmetric distribution of PI and CHOL as for the S and E proteins above. The PI binding sites (G6, L206, N207, T208, D209) and CHOL binding sites (L17, E18, Q19, N21, L22, S94, I97) distribute on both sides of the M proteins. However, as our structural model of the M protein dimer was mispredicted, this analysis should be interpreted with caution.

The diffusion of the membrane-embedded proteins on the virus envelope is also of high interest, which reflects how dynamic each protein is, in addition to their internal flexibility. To obtain the diffusion coefficients of M and S proteins, we analyzed the motion of each protein by calculating the mean squared deviation of all the M and S proteins in the CG trajectory (Fig. 5). The spherical coordinates θ and φ were used to describe the position of viral proteins, as shown in Fig. 5a. The motion of each protein's center of geometry is shown in Fig. 5b & c, based on which we calculated the diffusion coefficient of S and M proteins to characterize their diffusion abilities (Fig. 5d & e). Our analyses showed that the M and S proteins share similar diffusion coefficients: $7.1 \pm 0.2\ \mu m^2/s$ for M proteins, and $8.2 \pm 1.1\ \mu m^2/s$ for S proteins, respectively. These values are close to previous CG MD simulation results, which demonstrates that the membrane protein's diffusion coefficient in the MARTINI force field ranges from 3.3 to 12.0 $\mu m^2/s$ [11–13].

3 Discussion

With the whole virion model constructed (except for the complete RNA), a 500-ns CG simulation was performed in a water environment to relax each component to reach a more equilibrated configuration (Movie S1). Our simulation and analyses showed that the Martini CG simulations can be used to efficiently equilibrate such a complex system. It took at least 200-300 ns to equilibrate the virion system to reach a stable size and converged lipid distribution around membrane-embedded proteins. With such a CG equilibration, the transformed atomistic model would be more relaxed and require less computation time for further equilibration.

According to our RDF analysis, the PI lipids and CHOL were found to be more concentrated around the membrane-embedded proteins, which is consistent with another recent simulation study [7] and an earlier systematic analysis based on extensive simulations of membrane proteins [10]. The PS lipids also showed a moderate binding affinity to the S and M proteins, while the PC and PE lipids exhibited the least binding preference. The sphingomyelin (SM), DPSM, did not show binding preference to any membrane-embedded proteins either. Overall, the lipid distribution in the envelope is in line with the previous work by Corradi et al. [10]. In addition, our results showed that the PI lipids tend to concentrate on the outer leaflet, while CHOLs prefer to bind with proteins in the inner leaflet. The residues M1233, L1234, C1235, C1236 in the S protein, and A36, C40, and A41 in the E protein, located on the inner leaflet of the envelope, can recruit CHOLs. On the outer leaflet, the aromatic residues W1212, W1214, and Y1215 in the S protein, and F6 in the E protein, are PI attractive sites, indicating that the aromatic interactions may be one of the reasons for the enrichment of PIs around proteins.

Unlike a planar bilayer, the curvature of a spherical envelope may influence the diffusion of embedded proteins. Our analyses showed that the diffusion coefficients of M and S protein are $7.1 \pm 0.2\ \mu m^2/s$ and $8.2 \pm 1.1\ \mu m^2/s$, respectively. These diffusion coefficients are close to the values calculated from a planar bilayer system [11–13], indicating that membrane proteins in a spherical membrane may have similar diffusion ability with that in a planar bilayer. Due to the smoother energy landscape in MARTINI force field, the protein in CG force field diffuses faster than in AA force field. Previous studies compared the diffusion coefficient of proteins and lipids in the CG and atomistic models [11, 14], which showed that proteins and lipids in CG may diffuse four to ten times faster than in AA models. Based on this estimation, the diffusion coefficient of M and S proteins in AA models are estimated to be around $1.8 \pm 0.1\ \mu m^2/s$ and $2.1 \pm 0.3\ \mu m^2/s$, which are close to previously measured diffusion coefficients of membrane proteins ($4 - 10\ \mu m^2/s$) [15, 16]. It appears that the S proteins are rather dynamic, which can diffuse to form clusters (Fig. S2). These S-protein clusters may provide a more infectious condition for multiple spikes binding to one host cell receptor, which has been reported in previous studies [1, 17].

After the CG simulation completed, both the initial (Fig. S1a) and final (Fig. S1b) CG structures were transformed into atomistic models, as shown in Fig. S1c and Fig. S1d. Therefore, we are able to provide both the coarse-grained (CG) and atomistic model structures of SARS-CoV-2 here. Although there are some unavoidable uncertainties introduced by the prediction and modeling procedure, these models represent one of the most complete architectures of the intact SARS-CoV-2 so far, and they can serve as a framework for future improvements. For example, when more accurate protein structures are obtained, the structural models used here can be updated into the more reliable ones.

Unlike non-enveloped viruses, enveloped viruses are assembled by multiple structural proteins together with the lipid envelope. The presence of lipid bilayers in their assembly imposes significant challenges in the determination and simulation of intact enveloped viral structures [18]. This computational work has tried to efficiently tackled these challenges in heterogeneity through the development of an atomistic model of an authentic SARS-CoV-2 virion based on its low-resolution cryo-ET map and multi-scale modeling and simulations. Hopefully, the models will not only provide a foundation for future all-atom simulations of the intact virus, but also provide essential and intuitive information for the structural studies of enveloped viruses.

4 Materials and Methods

Our structural models of SARS-CoV-2 were based on recent structural biology studies, particularly the Cryo-ET density map of the virus [1], as well as protein structure prediction methods and molecular dynamics (MD) simulations. Constructing an atomistic model of such a large and complex system directly from scratch may produce massive bad contacts between atoms, which will cost a long time to relax and equilibrate. Therefore, we first built a coarse-grained (CG) model of the virion and equilibrated the system with the Martini force field [8, 19]. Then the CG system was transformed into an atomistic model. The details of the system construction are as follows:

4.1 Construction of the membrane envelope

We set up the initial CG vesicle with the CHARMM-GUI Vesicle Maker [20]. Since the vesicle would shrink after equilibration [20], we extended the initial vesicle diameter ($D_{init} = 109\,\mathrm{nm}$) to ensure it will reach the target diameter ($D_{target} \approx 85\,\mathrm{nm}$) after equilibration (Fig. S3) to match that observed in the Cryo-ET density map [1].

The detailed composition of the membrane envelope remains elusive. Previous MD simulation studies adopted various membranes with distinct lipid ratios to investigate the dynamics of SARS-CoV-2 spike protein embedded in a lipid bilayer. Hyeonuk et al. used a lipid bilayer composed of PC:PE:PS:SM:CHOL = 10:30:10:20:30, of which PE and CHOL are the majority [21]. Whereas Mateusz et al. (PC:PE:PI:PS:SM:CHOL = 50:20:15:5:5:5) and Casalino et al. (PC:PE:PI:PS:CHOL = 47:20:11:7:15) chose the membrane composition mimicking the lipid ratio of the ERGIC (ER-Golgi intermediate compartment) membrane, where PC and PE are predominant [22, 23]. In this work, we followed the latter strategy to construct a complex vesicle with the composition PC:PE:PI:PS:SM:CHOL = 45:20:5:10:5:15 [24].

The CG vesicle system was pre-equilibrated in a water box of $130 \times 130 \times 130\ \mathrm{nm}^3$ with the Martini2.2 force field [8, 19]. After a 10000-step energy minimization, the system was equilibrated in the NPT (isothermal-isobaric) ensemble for 200 ns. The long-range electrostatics was calculated by the reaction-field method. The van der Waals interaction and Coulomb interaction were considered within 1.1 nm. The v-rescale method and Berendsen method were used to maintain the system temperature at 310 K and pressure at 1.0 bar, respectively [25, 26]. The pressure coupling was isotropic. The coupling time constants for both the pressure and temperature were set to 1.0 ps.

4.2 Structural model of the spike (S) protein

The initial atomistic structures of the 'one RBD up' (PDB ID: 6xm3) [27] and 'RBD down' (PDB ID: 6xr8) [28] conformations were downloaded from the Protein Data Bank (PDB). These two high-resolution structures contained most of the S protein architecture, yet there are still some residues missing. These missing residues can be categorized into two types, and we adopted distinct protocols to fill these residues with MODELLER [29, 30]:

1. The residues located in the edge of the S protein ectodomain are too flexible to determine their exact positions, which leads to unresolved gaps in the cryo-EM structures. Therefore, we modeled these disordered regions with loops to maintain the integrity of the S protein ectodomain.

2. The other unresolved structures are around the membrane envelope (residue number 1148–

1273 of 'one RBD up' and 1163–1273 of 'RBD down'), including the Heptad Repeat-2 region, the transmembrane domain (TMD), and the endodomain. These structures were modeled with MODELLER [29, 30], based on the secondary structure predictions by the web server SPIDER3 [31].

As the S protein is a homo-trimer, the C_3 symmetry constraint was applied in the above modeling procedure. Then the atomistic structure was converted to the CG model for CG simulations.

Glycosylation of the specific sites on the S protein promotes the interaction between the virus and the host cell receptors, facilitating the fusion of the viral envelope, and the host cell membrane [32, 33]. Therefore, determining the specific glycosylation sites is important for atomistic modeling. Although glycosylation was not considered in the CG model or the CG simulation, we took it into account when constructing the atomistic models.

According to the previous experimental data [34, 35], numerous glycan types can be detected in each glycosylation site with different possibilities. Apart from the main N-glycosylation sites, few O-glycans are located on the three chains [23, 35]. All the glycosylation sites taken into account are listed in Table S1. Here, we built the glycosylated residue sites with two criteria:

1. If one glycan type shows dominant probability, then this particular type was used to set up the corresponding glycosylated residue.

2. If multiple glycan types show similar possibilities at one site, we picked the top two probable glycan types in the glycosylated residue to represent the complex glycosylation state.

The topology file of the full-length S protein with all glycosylated sites was generated by the CHARMM-GUI GLYCAN MODELER [36] with the CHARMM36m force field [37]. The structure of the glycosylated full-length S proteins is shown in Fig. S4. The details of the glycan types on each glycosylated site are shown in Table S1.

4.3 Structural model of the membrane (M) protein

Previous studies showed that M proteins may form dimers on the virus envelope [38], so we built a dimeric structure of the M protein based on previous studies [38, 39] with the docking software ZDOCK [40].

The structural topology of the M protein of SARS-CoV-2 and SARS-CoV (UniProtKB-P59596) should be identical or similar since they share high sequence similarity (about 96%). Previous studies on both proteins showed that the M protein of SARS-CoV and SARS-CoV-2 can be divided into two domains — the transmembrane (TM) domain and C-terminal domain (CTD) [41]. But the full-length structure of the SARS-CoV/SARS-CoV-2 M protein was not resolved. Therefore, we had to use structure prediction tools to build the M protein model. Multiple protein structure prediction methods/groups (trRosetta, FEIG-lab, AlphaFold2) give consistent two-domain architectures, but the specific predicted models vary.

As AlphaFold2 [42] was best scored in CASP14, we picked the monomeric structure predicted by AlphaFold2 to construct the M protein dimer (Fig. S5a). To obtain a rational dimer structure, we need to determine the dimer interface between the M protein monomers. A previous study illustrated that the TM domain of the M protein, which is comprised of three alpha-helices (residue 1–100), might be responsible for dimerization as well as for interacting with S proteins [38]. The CTD (residue 101-222) locates at the intracellular domain and may interact with other structural proteins such as N proteins, and is therefore excluded from the dimer interface. We limited the M-M binding area when using ZDOCK 3.0.2 and blocked the CTD atoms by changing their ACE type to 19 in the PDB file. Then we followed the common procedure of ZDOCK and predicted 2000 possible complexes for evaluation and selection. The most probable and reasonable dimeric model for the construction of the virus structure (Fig. S5b) was chosen under these criteria: the TM domain and CTD in the dimer maintain the same 'up' and 'down' orientations; the two monomers keep certain symmetry, especially the TM domains and their parallel helices; the CTD should not intrude to the membrane region nor crash with intracellular proteins such as the RNPs in the cryo-ET density map.

After our modeling and simulations completed, the dimeric structure of M protein was resolved [43]. There are differences between the predicted structure and the resolved structure, such as the relative positions between the three transmembrane helices and the tilt angle between the transmembrane domain and the CTD (Fig. S5c & d). However, the overall scaffold is similar. Not only the secondary structure of the transmembrane domain but also the CTD are coincided between the predicted and the resolved structures. In addition, the size of the transmembrane domain in the predicted structure is similar to the resolved structure, so we think the predicted M structure can be used for rough modeling and CG simulations of M proteins, which can then be replaced with the experimental structure for further simulations.

4.4 The envelope (E) protein

The E protein structure was published (PDB ID: 7k3g) [44], in which the transmembrane domain was resolved, whereas the N-terminal loop and endodomain structure remain uncertain. The secondary structure prediction by the RaptorX and SPIDER3 web servers [31, 45] indicates that the endodomain of E protein may form an alpha-helix, but the orientation of this inner helix cannot be determined. The homology modeling structure (Fig. S6a) based on the SARS-CoV E protein looks strange as the endodomain helices roll up toward the TMD helices, meaning that the endodomain helices are inserted into the viral envelope. However, our recently developed membrane contact probability (MCP) predictor [46] showed that while the residues 8–34 (forming the transmembrane helices in resolved E structure) entirely interact with membrane with high probability (Fig. S6b, green curve), the inner helices (residue 38–60) (Fig. S6b, blue curve) show discrete MCP signal, reflecting that the inner helix may be adsorbed onto the membrane surface rather than being embedded into the membrane. Interestingly, the E protein structure predicted by Feig's Lab [47] showed that the endo helix is optimized to touch the viral envelope, which is consistent with our MCP prediction (Fig. S6c). Also, the E protein structure predicted by the Feig Lab has been proven to be stable in microsecond MD simulations [48]. As for the oligomerization state, previous studies showed that the E protein of coronaviruses (like MHV and SARS-CoV) is able to self-oligomerize to create a pentameric ion channel, making this protein a viroporin [9, 49]. Therefore, we picked the predicted structural model by the Feig Lab as the initial E protein structure for our virus construction.

4.5 The nucleocapsid (N) protein

The N protein monomer contains five domains: the N-terminal domain (NTD), RNA binding domain (RBD), central Ser/Arg (SR)-rich linker, a dimerization domain and C-terminal domain (CTD) [9]. Previous studies have reported that the critical residues responsible for RNA binding are located in the N terminal region of N proteins (NTD and RBD) in multiple coronaviruses [50–53]. The dimerization domain is thought to mediate the formation of the N protein dimer. The RBD and dimerization domain are separated by the SR-rich linker, which is an intrinsically disordered region (IDR). In addition to the SR-rich linker, the N-terminal loop and C-terminal loop of N protein are both IDRs as well [54].

The N-terminal RBD (PDB ID: 7act) [55] and C-terminal dimerization domain (PDB ID: 6yun) [56] structures of the SARS-CoV-2 N protein have been determined separately. While these two isolated structures cannot tell the interface between them. The viral ribonucleoprotein (RNP) cryo-ET density map (EMD-30429) [1] provides a paradigm about how these two domains bind, which guided us to perform protein-protein docking with ZDOCK [40] to construct an N protein dimer (Fig. S7a).

We did not fill the IDRs of the N protein in our model. Instead, we utilized distance restraint to maintain the N protein structure (details in the next section). The full-length RNA was not included in our model because there is no way to determine the whole 3D RNA structure at the moment. However, the RNA fragment (10 bps) with a definite structure resolved together with the N protein (PDB ID: 7act) was included in our model. As the recent cryo-ET density map (EMD-30429) showed, the viral RNP unit was composed of five N protein dimers. Thus, the RNP unit structure was obtained by aligning five N protein dimers into the density map (EMD-30429) (Fig. S7b).

4.6 Assemble of the SARS-CoV-2 virus

After the envelope and structural proteins were set up as described above, we assembled all the components into one piece according to the Cryo-ET density map (EMD-30430) that clearly identified the architecture of the entire virus [1].

Firstly, we used the ‘fitmap’ tool within Chimera [57] to fit the equilibrated vesicle and S proteins into the cryo-ET density map (EMD-30430). Since most of the S proteins ectodomain show significant tilt with respect to the normal axis of the envelope in the cryo-ET density map, rigidly aligning our S protein model structure into the corresponding density often results in inappropriate orientations of S proteins, where their transmembrane domains were not embedded in the lipid bilayer (Fig. S8a). Therefore, after the rigid alignment, we optimized the orientation of each S protein to make its first principal axis parallel to the normal axis of the envelope and moved each S protein along the membrane normal direction to embed the transmembrane domain into the viral envelope properly. After the optimization, S proteins were located at the viral surface with an initial orientation perpendicular to the membrane surface (Fig. S8b). Then the optimized S protein structures were transformed to CG model in the Martini force field [8, 19]. Usually, the elastic network (ELN) algorithm is used to maintain the global protein conformation during the CG MD simulations. A longer ELN cutoff will enlarge the ELN intensity and make the protein more rigid. From the cryo-ET data [1], it was observed that the S proteins tend to tilt 40° relative to the normal axis of the viral envelope. To reproduce this flexibility of S proteins, we performed a series of simulations with an S protein embedded into a lipid bilayer with different ELN cutoffs. From the tilt angle analysis (Fig. S9), an ELN cutoff of 0.8 nm showed the largest flexibility and reasonable orientation angles of the S protein on the lipid bilayer surface within the simulation time, which was therefore adopted in our CG MD simulations for the S proteins. Please note that the utilization of an ELN may introduce some artifacts to the dynamics of S proteins, which is an intrinsic limitation of the CG MD simulations, but this would not be an issue for the model constructing purpose at this stage.

Next, 32 RNP units (Fig. S7b) were fitted into the density map (EMD-30430), where all the RNPs are nestled up to the inner surface of the viral envelope (Fig. S10a). Then we transformed all the RNP structures into a CG model. Without full-length RNA binding, the assembled RNPs may be unstable, so we applied distance restraints (force constant was set to 1000 kJ mol^{-1} nm^{-2}) between each pair of N protein dimer to maintain the relative positions of the RNP units during the following simulations (Fig. S10b). To maintain the entire RNPs architecture we also applied distance restraints (force constant 1000 kJ mol^{-1} nm^{-2}) between the center of mass (COM) of each RNP unit and the COM of all the RNPs. In addition, as all the IDRs of N protein are absent in the dimer structure, which may cause the N protein structure dissociation during the CG MD simulations, we utilized ELN (cutoff = 2.0 nm) to maintain the overall stability of the N protein dimers as well (Fig. S10c).

M proteins are located in an intricate lipid environment and are hard to be distinguished from the density map (EMD-30430). Therefore, it is difficult to directly fit the M proteins into the Cryo-ET density as had been done for S proteins and RNPs. Previous studies showed that the ratio of M:N proteins ranges from 1:1 to 3:1 [38, 58]. In the Cryo-ET density map (EMD-30430), there are 32 RNPs (32×5 N protein dimers) per virus. Therefore, 320 M protein dimers (M:N = 2:1) were initially inserted into the viral envelope uniformly with random orientations, and then the M protein dimers orientations are adjusted to ensure that the transmembrane domains are fully inserted into the envelope, and the first principal axis of M dimer is parallel to the normal direction of the envelope. After optimizing the orientations of the M proteins, there were 66 M protein dimers showing bad contacts with other structural proteins. As this will lead to infinite energy in the following energy minimization and equilibration procedure, we removed these 66 M protein dimers with bad contacts. As a consequence, there were 254 M protein dimers left in the system, still resulting in a reasonable ratio of M:N $\approx$ 3:2.

Like M proteins, E proteins are also embedded into the viral envelope, whereas far fewer E proteins are detected in a mature virus, as previous studies showed that the ratio of M:E $\approx$ 100:1 [59]. Therefore, we replaced two M protein dimers with E protein pentamers with proper orientation.

Following the above procedure, we assembled all the structural proteins (50 Spikes, 160 N dimers, 252 M dimers, and 2 E pentamers) into the viral envelope to form a SARS-CoV-2 virus in the absence of the complete RNA. We removed the lipid molecules within 0.1 nm of the proteins and solvated the protein–vesicle system into a cubic box of water molecules. In the end, the $155 \times 155 \times 155\ \mathrm{nm}^3$ sized simulation box contained 31,226,794 CG beads in total.

4.7 Coarse-grained molecular dynamics simulations

All the MD simulations were performed with the software GROMACS 2018.4. The CG simulation system was first performed an energy minimization using the steepest descent algorithm for 30,000 steps, followed by equilibration in the NPT ensemble (constant pressure and constant temperature) for 25 ns with the time step gradually enlarged from 1 fs to 5 fs. The Berendsen algorithm was utilized to maintain the system temperature at 310 K and pressure at 1.0 bar [26]. The coupling constant Tau-T and Tau-P were set to 1.0 ps. The pressure coupling type was set to isotropic, and the compressibility was $4.5 \times 10^{-5}\ \mathrm{bar}^{-1}$. The electrostatic interactions were calculated with the reaction-field method. The van der Waals interaction was cut off at 1.1 nm. After the equilibration, we performed a 500-ns CG MD production simulation with the time step of 5 fs.

4.8 Trajectory analysis

4.8.1 Vesicle size measurement

To evaluate the transformation of the viral envelope shape in the CG trajectory, we analyzed the vesicle diameter along the X, Y, and Z axes during the simulation. Through the center of geometry of the vesicle (COG_vesicle) and along the X axis direction, we delimited a cylinder whose radius was set to 1 nm. Then we extracted the lipid PO4 bead within the cylinder and classified these beads into two categories: one category contains the beads whose X coordinates are less than the x coordinate of the COG_vesicle, while the other category contains the rest. The distance between the COG of each category was used to character the vesicle size in the X direction. The same procedure was adopted to assess the vesicle size along the Y and Z direction.

4.8.2 Radial distribution function (RDF) analysis

The RDF profiles were generated with gmx_rdf, a built-in analysis tool of GROMACS. The lipids within 5 nm of the proteins were considered into the RDF calculation. We performed RDF analysis towards every S, M, E proteins and averaged the corresponding results to represent the lipid distribution around these embedded structural proteins. In addition, to examine whether the RDF has converged we picked three trajectory segments 0-5 ns, 300-400 ns, 400-500 ns for RDF calculation.

4.8.3 Protein-lipid interaction

The protein-lipid interaction was considered when the distance between protein residue and lipid head group is less than 6 Å. We counted the frames (frames_interact) that generic lipid L can interact with protein residue R in the last 100-ns trajectory. Then the ratio between the frames_interact and total frames of the trajectory (frames_total) was used to reflect the probability that residue R contacts with the lipid L. Fig. 4 showed the average results among 50 Ss, 252 Ms and 2 Es.

4.8.4 Diffusion coefficient

The protein motion on the vesicle surface can be viewed as a 2-dimensional diffusion. The position of each protein can be described by two coordinates, the latitude (θ) and longitude (φ) with respect to the COG of the vesicle (Fig. 5a). In Fig. 5b & 5c, we plotted these coordinates with Mollweide projection and colored these data points by their time stamps. The displacement of a protein, r, can be characterized by the arc length on the sphere surface, which was utilized to further calculate the mean squared deviation ($< r^2 >$) of all the proteins. Given the positions of a protein before and after a short interval, (θ_1, φ_1) and (θ_0, φ_0), we can calculate the displacement of the protein by:

$$r = R \times arccos[cos(\theta_1)cos(\theta_0)cos(\varphi_1 - \varphi_0) + sin(\theta_1)sin(\theta_0)] \quad (1)$$

where R is the radius of the viral envelope.

Then the mean squared deviation ($< r^2 >$) can be calculated, and so is the diffusion coefficient according to:

$$< r^2 >= 4Dt \tag{2}$$

where t is the simulation time.

As our analysis showed that the system needed 200 ns to reach equilibrium in size, we only used the trajectories from 200 to 500 ns for the diffusion coefficient calculation. The data from 210 to 250 ns (delimited by dashed lines in Fig. 5d & 5e) was extracted to perform a linear fit to calculate the diffusion coefficient of proteins.

4.9 Conversion from the CG system to the atomistic system

Here we present two all-atom virus models transformed from the first and last frames of CG simulation by the CG2AT2 tool [60]: (1) The all-atom model of SARS-CoV-2 virion converted from the initial CG structure contains 15,526,323 atoms, in which all the S proteins are full-glycosylated. Totally, there are 278,131,974 atoms in the entire atomistic system of the simulation box. (2) Because the glycosylation did not be considered in the previous CG simulation, directly transformed the final CG virus structure to atomistic resolution results in the loss of glycosylated residues in this atomistic virus model, whose atom number is 14,873,073. Correspondingly, the simulation system involves 266,063,412 atoms after solved the virus structure into a water box. Although challenging at the moment, these atomistic systems can serve as the initial structure for future all-atom simulations.

Competing Interests

The authors declare no competing interests.

Author Contributions

C.S. and S.L. conceived the idea and supervised the project. D.W. and J.L. built the models and performed molecular dynamics simulations. L.W. and Y.C. participated in protein modeling. D.W., J.L., S.L., and C.S. wrote the original manuscript.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (21873006 and 32071251 to C.S.; 32171195 to S.L.), the Ministry of Science and Technology of China (2021YFE0108100 to C.S.), and Tsinghua University Spring Breeze Fund (#2021Z99CFZ004 to S.L.). D.W. was supported in part by the Postdoctoral Fellowship of Peking-Tsinghua Center for Life Sciences. The MD simulations were performed on the Computing Platform of the Center for Life Sciences at Peking University and the National Supercomputer Center in Tianjin.

References

[1] Hangping Yao, Yutong Song, Yong Chen, Nanping Wu, Jialu Xu, Chujie Sun, Jiaxing Zhang, Tianhao Weng, Zheyuan Zhang, Zhigang Wu, Linfang Cheng, Danrong Shi, Xiangyun Lu, Jianlin Lei, Max Crispin, Yigong Shi, Lanjuan Li, and Sai Li. Molecular Architecture of the SARS-CoV-2 Virus. *Cell*, 183(3):730–738, 2020.

[2] Steffen Klein, Mirko Cortese, Sophie L. Winter, Moritz Wachsmuth-Melm, Christopher J. Neufeldt, Berati Cerikan, Megan L. Stanifer, Steeve Boulant, Ralf Bartenschlager, and Petr Chlanda. SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. *Nature Communications*, 11(1):1–10, 2020.

[3] Alvin Yu, Alexander J. Pak, Peng He, Viviana Monje-Galvan, Lorenzo Casalino, Zied Gaieb, Abigail C. Dommer, Rommie E. Amaro, and Gregory A. Voth. A multiscale coarse-grained model of the SARS-CoV-2 virion. *Biophysical Journal*, 120(6):1097–1104, 2021.

[4] Lorenzo Casalino, Abigail C Dommer, Zied Gaieb, Emilia P Barros, Terra Sztain, Surl-Hee Ahn, Anda Trifan, Alexander Brace, Anthony T Bogetti, Austin Clyde, Heng Ma, Hyungro Lee, Matteo Turilli, Syma Khalid, Lillian T Chong, Carlos Simmerling, David J Hardy, Julio DC Maia, James C Phillips, Thorsten Kurth, Abraham C Stern, Lei Huang, John D McCalpin, Mahidhar Tatineni, Tom Gibbs, John E Stone, Shantenu Jha, Arvind Ramanathan, and Rommie E Amaro. AI-driven multiscale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics. *The International Journal of High Performance Computing Applications*, 35(5):432–451, September 2021.

[5] Weria Pezeshkian, Fabian Grünewald, Oleksandr Narykov, Senbao Lu, Tsjerk A Wassenaar, Siewert J. Marrink, and Dmitry Korkin. Molecular architecture of sars-cov-2 envelope by integrative modeling. *bioRxiv*, 2021.

[6] Abigail Dommer, Lorenzo Casalino, Fiona Kearns, Mia Rosenfeld, Nicholas Wauer, Surl-Hee Ahn, John Russo, Sofia Oliveira, Clare Morris, Anthony Bogetti, Anda Trifan, Alexander Brace, Terra Sztain, Austin Clyde, Heng Ma, Chakra Chennubhotla, Hyungro Lee, Mat-

teo Turilli, Syma Khalid, Teresa Tamayo-Mendoza, Matthew Welborn, Anders Christensen, Daniel G. A. Smith, Zhuoran Qiao, Sai Krishna Sirumalla, Michael O'Connor, Frederick Manby, Anima Anandkumar, David Hardy, James Phillips, Abraham Stern, Josh Romero, David Clark, Mitchell Dorrell, Tom Maiden, Lei Huang, John McCalpin, Christopher Woods, Alan Gray, Matt Williams, Bryan Barker, Harinda Rajapaksha, Richard Pitts, Tom Gibbs, John Stone, Daniel Zuckerman, Adrian Mulholland, Thomas Miller, Shantenu Jha, Arvind Ramanathan, Lillian Chong, and Rommie Amaro. #covidisairborne: Ai-enabled multiscale computational microscopy of delta sars-cov-2 in a respiratory aerosol. *bioRxiv*, 2021.

[7] Beibei Wang, Changqing Zhong, and D. Peter Tieleman. Supramolecular organization of sars-cov and sars-cov-2 virions revealed by coarse-grained models of intact virus envelopes. *Journal of Chemical Information and Modeling*, 62(1):176–186, 2022.

[8] Siewert J. Marrink, H. Jelger Risselada, Serge Yefimov, D. Peter Tieleman, and Alex H. De Vries. The MARTINI force field: Coarse grained model for biomolecular simulations. *Journal of Physical Chemistry B*, 111(27):7812–7824, 2007.

[9] David Bracquemond and Delphine Muriaux. Betacoronavirus Assembly: Clues and Perspectives for Elucidating SARS-CoV-2 Particle Formation and Egress. *mBio*, 12(5):1–14, 2021.

[10] Valentina Corradi, Eduardo Mendez-Villuendas, Helgi I. Ingólfsson, Ruo Xu Gu, Iwona Siuda, Manuel N. Melo, Anastassiia Moussatova, Lucien J. Degagné, Besian I. Sejdiu, Gurpreet Singh, Tsjerk A. Wassenaar, Karelia Delgado Magnero, Siewert J. Marrink, and D. Peter Tieleman. Lipid-Protein Interactions Are Unique Fingerprints for Membrane Proteins. *ACS Central Science*, 4(6):709–717, 2018.

[11] Chenyi Liao, Xiaochuan Zhao, Jiyuan Liu, Severin T. Schneebeli, John C. Shelley, and Jianing Li. Capturing the multiscale dynamics of membrane protein complexes with all-atom, mixed-resolution, and coarse-grained models. *Physical Chemistry Chemical Physics*, 19(13):9181–9188, 2017.

[12] Sivaramakrishnan Ramadurai, Andrea Holt, Lars V. Schäfer, Victor V. Krasnikov, Dirk T.S. Rijkers, Siewert J. Marrink, J. Antoinette Killian, and Bert Poolman. Influence of hydrophobic

mismatch and amino acid composition on the lateral diffusion of transmembrane peptides. *Biophysical Journal*, 99(5):1447–1454, 2010.

[13] Xavier Periole, Thomas Huber, Siewert Jan Marrink, and Thomas P. Sakmar. G protein-coupled receptors self-assemble in dynamics simulations of model bilayers. *Journal of the American Chemical Society*, 129(33):10126–10132, 2007.

[14] Elena Ermakova and Yuriy Zuev. Effect of ergosterol on the fungal membrane properties. All-atom and coarse-grained molecular dynamics study. *Chemistry and Physics of Lipids*, 209(November):45–53, 2017.

[15] Sivaramakrishnan Ramadurai, Andrea Holt, Victor Krasnikov, Geert Van Den Bogaart, J. Antoinette Killian, and Bert Poolman. Lateral diffusion of membrane proteins. *Journal of the American Chemical Society*, 131(35):12650–12656, 2009.

[16] Sivaramakrishnan Ramadurai, Ria Duurkens, Victor V. Krasnikov, and Bert Poolman. Lateral diffusion of membrane proteins: Consequences of hydrophobic mismatch and lipid composition. *Biophysical Journal*, 99(5):1482–1489, 2010.

[17] Yeol Kyo Choi, Yiwei Cao, Martin Frank, Hyeonuk Woo, Sang Jun Park, Min Sun Yeom, Tristan I. Croll, Chaok Seok, and Wonpil Im. Structure, Dynamics, Receptor Binding, and Antibody Binding of the Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein in a Viral Membrane. *Journal of Chemical Theory and Computation*, 17(4):2479–2487, 2021.

[18] Sai Li. Cryo-electron tomography of enveloped viruses. *Trends in Biochemical Sciences*, 47(2):173–186, 2021.

[19] Luca Monticelli, Senthil K. Kandasamy, Xavier Periole, Ronald G. Larson, D. Peter Tieleman, and Siewert Jan Marrink. The MARTINI coarse-grained force field: Extension to proteins. *Journal of Chemical Theory and Computation*, 4(5):819–834, 2008.

[20] Yifei Qi, Helgi I. Ingólfsson, Xi Cheng, Jumin Lee, Siewert J. Marrink, and Wonpil Im. CHARMM-GUI Martini Maker for Coarse-Grained Simulations with the Martini Force Field. *Journal of Chemical Theory and Computation*, 11(9):4486–4494, 2015.

[21] Hyeonuk Woo, Sang Jun Park, Yeol Kyo Choi, Taeyong Park, Maham Tanveer, Yiwei Cao, Nathan R. Kern, Jumin Lee, Min Sun Yeom, Tristan I. Croll, Chaok Seok, and Wonpil Im. Developing a fully glycosylated full-length SARS-COV-2 spike protein model in a viral membrane. *Journal of Physical Chemistry B*, 124(33):7128–7137, 2020.

[22] Beata Turoňová, Mateusz Sikora, Christoph Schürmann, Wim J.H. Hagen, Sonja Welsch, Florian E.C. Blanc, Sören von Bülow, Michael Gecht, Katrin Bagola, Cindy Hörner, Ger van Zandbergen, Jonathan Landry, Nayara Trevisan Doimo de Azevedo, Shyamal Mosalaganti, Andre Schwarz, Roberto Covino, Michael D. Mühlebach, Gerhard Hummer, Jacomine Krijnse Locker, and Martin Beck. In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges. *Science*, 370(6513):203–208, 2020.

[23] Lorenzo Casalino, Zied Gaieb, Jory A. Goldsmith, Christy K. Hjorth, Abigail C. Dommer, Aoife M. Harbison, Carl A. Fogarty, Emilia P. Barros, Bryn C. Taylor, Jason S. Mclellan, Elisa Fadda, and Rommie E. Amaro. Beyond shielding: The roles of glycans in the SARS-CoV-2 spike protein. *ACS Central Science*, 6(10):1722–1734, 2020.

[24] Doralicia Casares, Pablo V Escrib, and Catalina Ana Rosselló. Membrane Lipid Composition Effect on Membrane and Organelle Structure, Function and Compartmentalization and Therapeutic Avenues. *International Journal of Molecular Science*, 20(2167):2167, 2019.

[25] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *Journal of Chemical Physics*, 126:1–8, 2007.

[26] H. J.C. Berendsen, J. P.M. Postma, W. F. Van Gunsteren, A. Dinola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81:3684–3690, 1984.

[27] Tongqing Zhou, Yaroslav Tsybovsky, Jason Gorman, Micah Rapp, Gabriele Cerutti, Gwo-Yu Chuang, Phinikoula S. Katsamba, Jared M. Sampson, Arne Schön, Jude Bimela, Jeffrey C. Boyington, Alexandra Nazzari, Adam S. Olia, Wei Shi, Mallika Sastry, Tyler Stephens, Jonathan Stuckey, I-Ting Teng, Pengfei Wang, Shuishu Wang, Baoshan Zhang, Richard A.

Friesner, David D. Ho, John Mascola, Lawrence Shapiro, and Peter D. Kwong. Cryo-EM Structures Delineate a pH-Dependent Switch that Mediates Endosomal Positioning of SARS-CoV-2 Spike Receptor-Binding Domains. *SSRN Electronic Journal*, 28:867–880, 2020.

[28] Yongfei Cai, Jun Zhang, Tianshu Xiao, Hanqin Peng, Sarah M. Sterling, Richard M. Walsh, Shaun Rawson, Sophia Rits-Volloch, and Bing Chen. Distinct conformational states of SARS-CoV-2 spike protein. *Science*, 369(6511):1586–1592, 2020.

[29] A Sali and T L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, 1993.

[30] Andras Fiser, Richard Kinh Gian Do, and A Sali. Modeling Loops in Protein Structures. *Protein Science*, 9:1753–1773, 2000.

[31] Rhys Heffernan, Yuedong Yang, Kuldip Paliwal, and Yaoqi Zhou. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*, 33(18):2842–2849, September 2017.

[32] Tzu Jing Yang, Yen Chen Chang, Tzu Ping Ko, Piotr Draczkowski, Yu Chun Chien, Yuan Chih Chang, Kuen Phon Wu, Kay Hooi Khoo, Hui Wen Chang, and Shang Te Danny Hsu. Cryo-EM analysis of a feline coronavirus spike protein reveals a unique structure and camouflaging glycans. *Proceedings of the National Academy of Sciences of the United States of America*, 117(3):1438–1446, 2020.

[33] Yasunori Watanabe, Zachary T. Berndsen, Jayna Raghwani, Gemma E. Seabright, Joel D. Allen, Oliver G. Pybus, Jason S. McLellan, Ian A. Wilson, Thomas A. Bowden, Andrew B. Ward, and Max Crispin. Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nature Communications*, 11(1):1–10, 2020.

[34] Yasunori Watanabe, Joel D. Allen, Daniel Wrapp, Jason S. McLellan, and Max Crispin. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science*, 369(6501):330–333, 2020.

[35] Asif Shajahan, Nitin T. Supekar, Anne S. Gleinich, and Parastoo Azadi. Deducing the N- And O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. *Glycobiology*, 30(12):981–988, 2020.

[36] Sang Jun Park, Jumin Lee, Yifei Qi, Nathan R. Kern, Hui Sun Lee, Sunhwan Jo, Insuk Joung, Keehyung Joo, Jooyoung Lee, and Wonpil Im. CHARMM-GUI Glycan Modeler for modeling and simulation of carbohydrates and glycoconjugates. *Glycobiology*, 29(4):320–331, 2019.

[37] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L. De Groot, Helmut Grubmüller, and Alexander D. MacKerell. CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14:71–73, 2016.

[38] Benjamin W. Neuman, Gabriella Kiss, Andreas H. Kunding, David Bhella, M. Fazil Baksh, Stephen Connelly, Ben Droese, Joseph P. Klaus, Shinji Makino, Stanley G. Sawicki, Stuart G. Siddell, Dimitrios G. Stamou, Ian A. Wilson, Peter Kuhn, and Michael J. Buchmeier. A structural analysis of M protein in coronavirus assembly and morphology. *Journal of Structural Biology*, 174(1):11–22, 2010.

[39] Dewald Schoeman and Burtram C. Fielding. Coronavirus envelope protein: current knowledge. *Virology Journal*, 16(1):69–90, December 2019.

[40] Brian G. Pierce, Yuichiro Hourai, and Zhiping Weng. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS ONE*, 6(9):1–5, 2011.

[41] Cornelis A. M. de Haan, Harry Vennema, and Peter J. M. Rottier. Assembly of the Coronavirus Envelope: Homotypic Interactions between the M Proteins. *Journal of Virology*, 74(11):4967–4978, 2000.

[42] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray

Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[43] Zhikuan Zhang, Norimichi Nomura, Yukiko Muramoto, Toru Ekimoto, Tomoko Uemura, Kehong Liu, Moeko Yui, Nozomu Kono, Junken Aoki, Mitsunori Ikeguchi, Takeshi Noda, So Iwata, Umeharu Ohto, and Toshiyuki Shimizu. Structure of SARS-CoV-2 membrane protein essential for virus assembly. *Nature Communications*, 13(1):1–7, 2022.

[44] Venkata S. Mandala, Matthew J. McKay, Alexander A. Shcherbakov, Aurelio J. Dregni, Antonios Kolocouris, and Mei Hong. Structure and drug binding of the SARS-CoV-2 envelope protein transmembrane domain in lipid bilayers. *Nature Structural and Molecular Biology*, 27(12):1202–1208, 2020.

[45] Sheng Wang, Wei Li, Shiwang Liu, and Jinbo Xu. RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Research*, 44(W1):430–435, 2016.

[46] Lei Wang, Jiangguo Zhang, Dali Wang, and Chen Song. Lipid contact probability: an essential and predictive character for the structural and functional studies of membrane proteins. *bioRxiv*, January 2021.

[47] Lim Heo and Michael Feig. Modeling of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) proteins by machine learning and physics-based refinement. *bioRxiv*, 2020.

[48] Viviana Monje-Galvan and Gregory A. Voth. Molecular interactions of the M and E integral membrane proteins of SARS-CoV-2. *Faraday Discussions*, 232:49–67, 2021.

[49] Christian Arias-Reyes, Natalia Zubieta-DeUrioste, Liliana Poma-Machicao, Fernanda Aliaga-Raduan, Favio Carvajal-Rodriguez, Mathias Dutschmann, Edith M. Schneider-Gasser, Gustavo Zubieta-Calleja, and Jorge Soliz. Does the pathogenesis of SARS-CoV-2 virus decrease at high-altitude? *Respiratory Physiology and Neurobiology*, 277:1–22, 2020.

[50] Sarah C. Keane, Pinghua Lius, Julian L. Leibowitzs, and David P. Giedroc. Functional Transcriptional Regulatory Sequence (TRS) RNA binding and helix destabilizing determinants of Murine Hepatitis Virus (MHV) Nucleocapsid (N) protein. *Journal of Biological Chemistry*, 287(10):7063–7073, 2012.

[51] Yong Wah Tan, Shouguo Fang, Hui Fan, Julien Lescar, and D. X. Liu. Amino acid residues critical for RNA-binding in the N-terminal domain of the nucleocapsid protein are essential determinants for the infectivity of coronavirus in cultured cells. *Nucleic Acids Research*, 34(17):4816–4825, 2006.

[52] Nicholas E. Grossoehme, Lichun Li, Sarah C. Keane, Pinghua Liu, Charles E. Dann, Julian L. Leibowitz, and David P. Giedroc. Coronavirus N Protein N-Terminal Domain (NTD) Specifically Binds the Transcriptional Regulatory Sequence (TRS) and Melts TRS-cTRS RNA Duplexes. *Journal of Molecular Biology*, 394(3):544–557, 2009.

[53] Kumar Singh Saikatendu, Jeremiah S. Joseph, Vanitha Subramanian, Benjamin W. Neuman, Michael J. Buchmeier, Raymond C. Stevens, and Peter Kuhn. Ribonucleocapsid Formation of Severe Acute Respiratory Syndrome Coronavirus through Molecular Action of the N-Terminal Domain of N Protein. *Journal of Virology*, 81(8):3913–3921, 2007.

[54] Shan Lu, Qiaozhen Ye, Digvijay Singh, Yong Cao, Jolene K. Diedrich, John R. Yates, Elizabeth Villa, Don W. Cleveland, and Kevin D. Corbett. The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nature Communications*, 12(1):502, 2021.

[55] Dhurvas Chandrasekaran Dinesh, Dominika Chalupska, Jan Silhan, Eliska Koutna, Radim Nencka, Vaclav Veverka, and Evzen Boura. Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein. *PLoS Pathogens*, 16(12):1–16, 2020.

[56] Luca Zinzula, Jerome Basquin, Stefan Bohn, Florian Beck, Sven Klumpe, Andreas Bracher, F Ulrich Hartl, and Wolfgang Baumeister. High-resolution structure and biophysical characterization of the nucleocapsid phosphoprotein dimerization domain from the Covid-19 severe acute respiratory syndrome coronavirus 2. *Biochemical and Biophysical Research Communications*, 538(January):54–62, 2020.

[57] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Elaine C. Meng, Gregory S. Couch, Tristan I. Croll, John H. Morris, and Thomas E. Ferrin. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science*, 30(1):70–82, 2021.

[58] Benjamin W. Neuman, Brian D. Adair, Craig Yoshioka, Joel D. Quispe, Gretchen Orca, Peter Kuhn, Ronald A. Milligan, Mark Yeager, and Michael J. Buchmeier. Supramolecular Architecture of Severe Acute Respiratory Syndrome Coronavirus Revealed by Electron Cryomicroscopy. *Journal of Virology*, 80(16):7918–7928, 2006.

[59] Yinon M Bar-On, Avi Flamholz, Rob Phillips, and Ron Milo. SARS-CoV-2 (COVID-19) by the numbers. *eLife*, 9(10):697–698, April 2020.

[60] Owen N. Vickery and Phillip J. Stansfeld. CG2AT2: an Enhanced Fragment-Based Approach for Serial Multi-scale Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation*, 17(10):6472–6482, 2021.

Figures

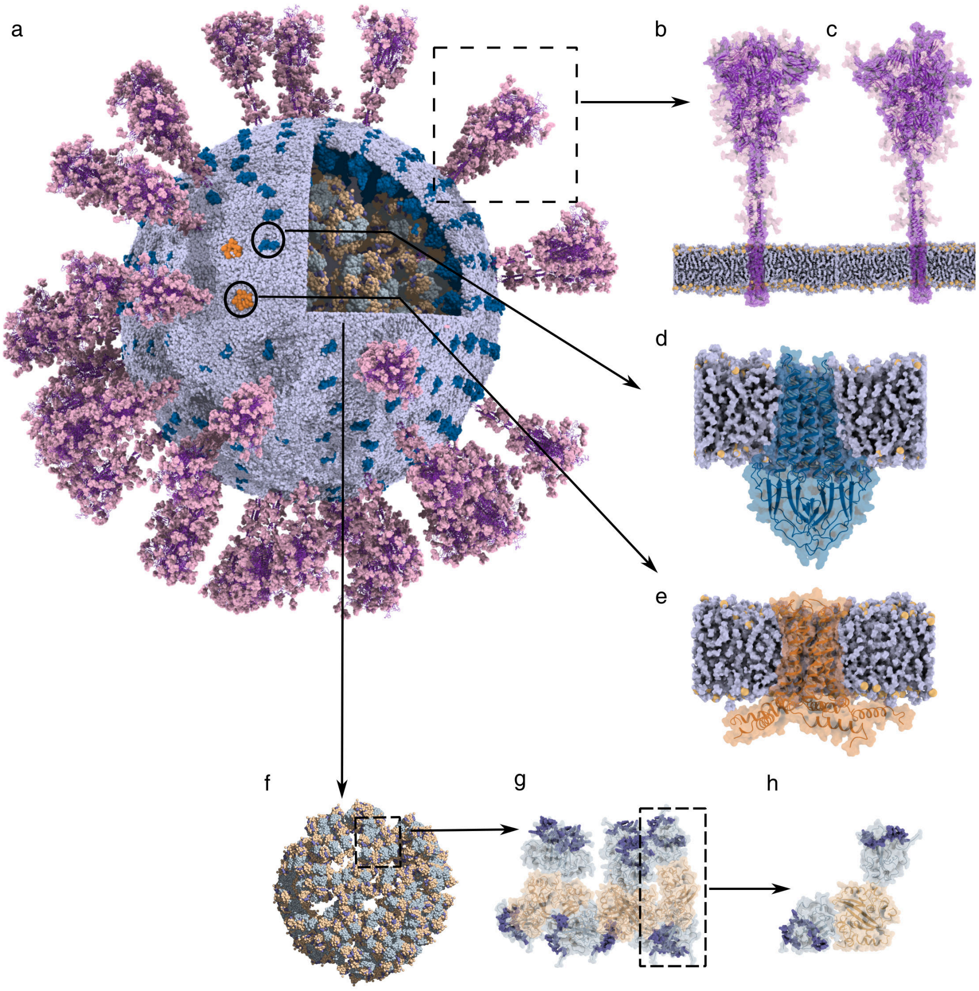

Figure 1: (a) The overview of the virus structure. The viral envelope is colored blue–white. Purple, deep blue, and orange regions indicate the S, M, and E proteins, respectively. RNPs are located within the envelope, and domains near the N-terminal and C-terminal of N proteins are shown in grey–blue and wheat, respectively. (b)–(c) The ‘RBD down’ and ‘one RBD up’ conformations of the S protein. The S proteins are purple, and the light-pink surface shows the glycans. The blue–white surface represents the viral envelope where S proteins are embedded, and the orange spheres indicate the lipid head groups. (d)–(e) The zoom-in view of the M protein (d) and E protein (e). (f)–(h) The architecture of RNPs: arrangement of all the RNPs within the virus (f), a single RNP unit (g), and an N protein dimer (h). The RNA segments bound to the N protein are blue–purple.

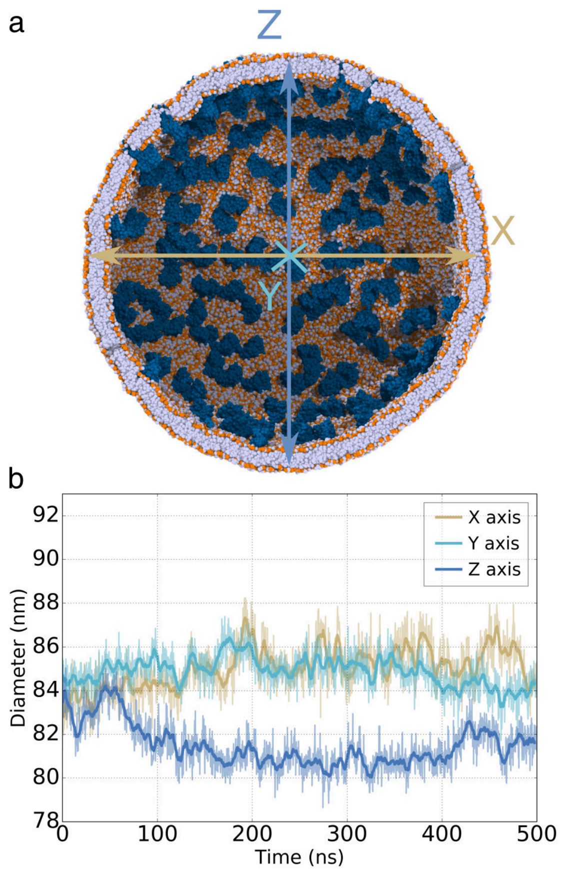

Figure 2: (a) The sectional view of the viral vesicle. The vesicle boundary was marked by the lipid head groups, which were colored in orange. The deep blue spheres represent the M proteins. The RNP and S proteins were hidden for clarity view. (b) Vesicle size evolution along the X (brown), Y (palegreen), and Z (lightblue) axis during the simulation. The colors of the three curves are corresponding to (a).

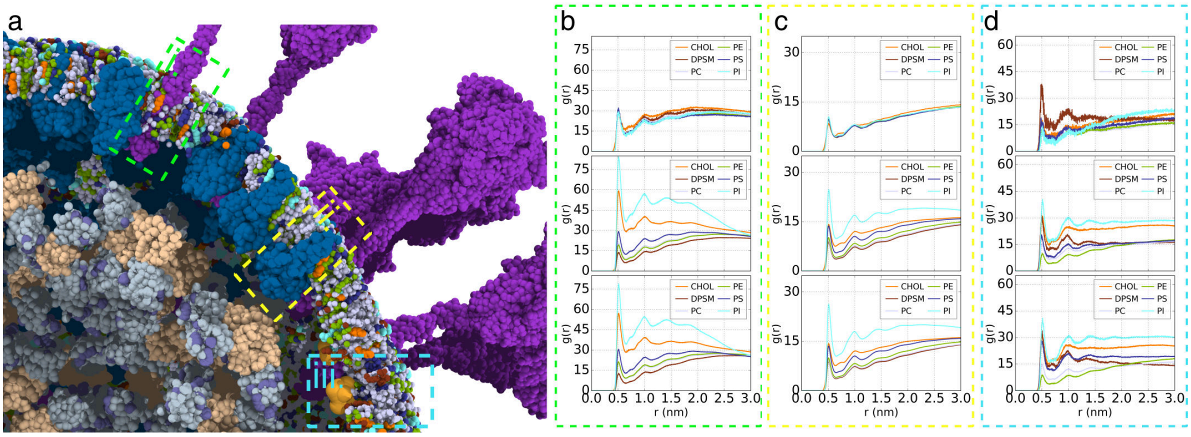

Figure 3: (a) A zoom view reflects the relative position between the S, M, and E trans-membrane domains and the vesicle. The colors of proteins are matched with Fig. 1. The three boxes point out the S (i), M (ii), and E (iii) trans-membrane domain, respectively. (b) The lipid radial distribution function (RDF) refers to the S trans-membrane domain. The panels from top to bottom show the RDF results generated from the 0-5 ns, 300-400 ns, and 400-500 ns trajectories. The different vesicle components were colored in orange (CHOL), chocolate (DPSM), bluewhite (PC), splitpea (PE), deepblue (PS), cyan (PI). (c)-(d) same as (b) but for M and E.

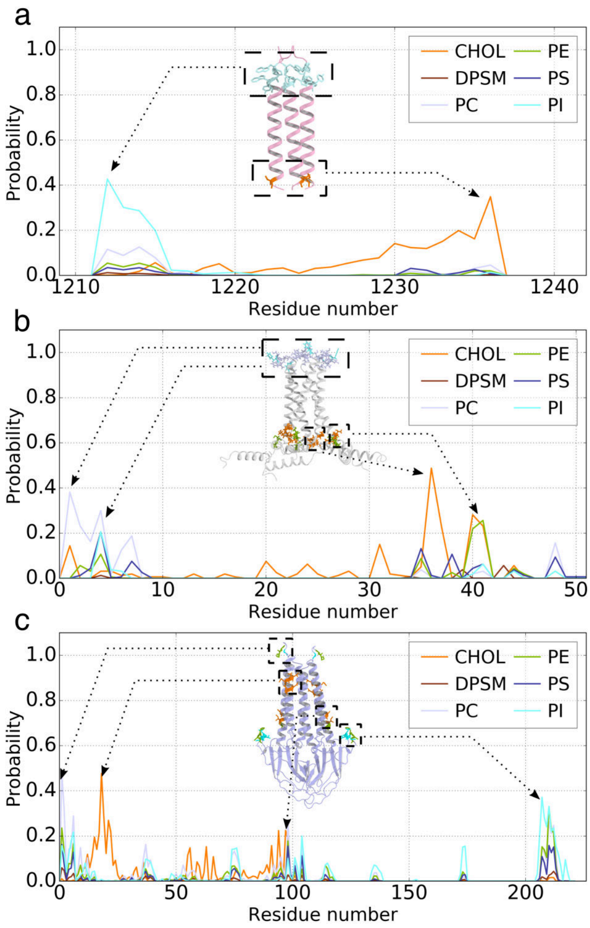

Figure 4: (a) Each profile represents the contact probability between S protein TMD and a kind of lipid. Various colors were applied to distinguish these profiles: orange (CHOL), chocolate (DPSM), bluewhite (PC), splitpea (PE), deepblue (PS), cyan (PI). The ribbon cartoon shows the S protein TMD structure, which is colored in lightpink. The stick representation highlights the residues with a high probability of binding to the lipid. The residues are colored corresponding to the binding lipid. (b)–(c) same as (a) but for E and M.

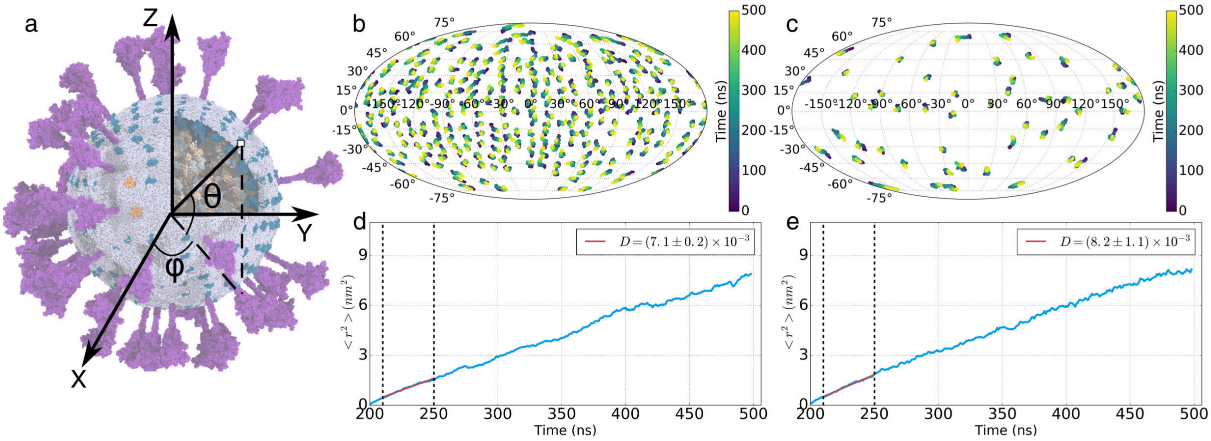

Figure 5: (a) Illustration of θ and φ angle. (b)–(c) The coordinate variation of the M (b), S (c) proteins transmembrane domain during the simulation trajectory. (d)–(e) The correlation between the mean squared position deviation of M (d), S (e) proteins transmembrane domain and simulation time. The dashed lines delimit the curves to do linear fitting. The diffusion coefficients of M, S proteins are $7.1 \pm 0.2\ \mu m^2/s$, $8.2 \pm 1.1\ \mu m^2/s$, respectively.