Toward an atomistic model of SARS-CoV-2

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.

[*]To whom correspondence should be addressed; E-mail: sai@mail.tsinghua.edu.cn (SL), c.song@pku.edu.cn (CS)

Text

The ongoing Coronavirus disease 2019 (COVID-19) pandemic has infected over 300 million people globally as of January 2022. 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. To understand 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 recombinant proteins by cryo-electron microscopy (cryo-EM) or crystallography, and structures *in situ* for the native proteins by cryo-electron tomography (cryo-ET) [1]. Although recent technical developments of cryo-ET have enabled the reconstruction of intact SARS-CoV-2, the structure has been limited to nanometer resolution [2, 3]. In the meanwhile, computational studies have also provided highly valuable information on the structure of the virus [4–6], but the existing structural models have been limited to a coarse-grained (CG) scale, focusing only on the virus envelope, or constructed without considering the protein localization from the *in situ* cryo-ET data. Therefore, we set out to construct an atomistic model of SARS-CoV-2 by fully employing the latest cryo-ET data [2], in combination with the available experimentally resolved protein structures, structure prediction and modeling methods, as well as molecular dynamics (MD) simulations. The atomistic model will not only provide a 3D structure for scientific demonstration and education, but also offer a framework for future exascale all-atom MD simulations to understand the dynamics of the intact virus, its assembly, and mutations.

To obtain a better equilibrated atomistic model of SARS-CoV-2, we first built a CG model and then equilibrated it with the Martini force field [7], followed by a resolution transformation. To build the CG model, we constructed structural models of each protein component separately, and then we assembled them onto a pre-equilibrated lipid envelope according to the architecture of the intact virus revealed by cryo-ET [2]. As 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 200-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 & S1b), and the final structure, built after a 200-ns CG MD simulation (Fig. S1d).

The viral envelope packages the virus genome and serves as a foundation for structural protein embedding. In this study, the envelope was built as a vesicle consisting of PC:PE:PI:PS:SM:CHOL = 45:20:5:10:5:15, which mimics the natural compositions of a viral envelope (SI 1.1) [8]. The pre-equilibrated vesicle (Fig. S2) served as the main scaffold for the virus construction.

The ~600 kDa, membrane-embedded S-trimer (Fig. 1b and 1c) was shown to be heavily glycosylated with 66 N-linked glycans and 4 O-linked glycans (Table S1) [9], which is responsible for receptor binding and membrane fusion [10]. S protein binds to the cellular surface receptor angiotensin converting enzyme 2 (ACE2) through the receptor-binding domain (RBD) [11], which can adopt two statuses on intact virions: the "RBD down" or the "one RBD up" conformations [2]. The various RBD conformations play vital roles in host recognition and immune evasion, as RBDs can only bind with ACE2 when in an "up" conformation. Statistical analysis of the native S suggested that SARS-CoV-2 harbors Ss with an approximately half-and-half ratio of the 'up/down' conformations [2]. Therefore, both the "RBD down" (Fig. S3a) and "one RBD up" conformations (Fig. S3b) were integrated into our model (SI 1.2). Moreover, the flexible S-trimers are rotatable around their stalks and are mobile on the envelope. These multiple hierarchies of motions are unique among enveloped viral spikes, and they were also enabled in our CG MD simulations during the construction of the atomistic model.

The ~25 kDa, membrane-embedded M (Fig. 1d) is the most abundant structural protein of SARS-CoV-2 and is a key component for membrane curvature formation, intracellular retention of S, and localization of ribonucleoproteins (RNPs)

during viral budding [12]. Owing to the difficulties in discerning M from the SARS-CoV-2 tomograms and the lack of an atomistic structure, we adopted the monomer structure of M predicted by AlphaFold2 [13] (Target C1906x2) as the initial structural model (Fig. S4a), and dimerized it (Fig. S4b) using a protein–protein docking algorithm, as M probably presents in a dimeric form in mature virions [14]. The detailed docking procedure can be found in SI 1.3.

The ~10 kDa, membrane-embedded E (Fig. 1e) is another structural component of mature SARS-CoV-2, although it only presents in a small amount [15]. It participates in viral budding and is likely responsible for scissoring the budded virion [12]. Currently, only the transmembrane domain (TMD) of E is determined. Homologous modeling based on the SARS-CoV E protein generates a structure (Fig. S5a) with its cytoplasmic domain inserted into the viral envelope. The Feig Lab [16] further optimized the modeled structures and obtained a more reasonable inner helix orientation, which is consistent with our membrane contact prediction [17] (SI 1.4). Therefore, Feig Lab's E model (Fig. S5c) was chosen and embedded into our viral envelope.

The ~46 kDa Ns (Fig. 1f to 1h) assemble with viral RNA to form RNPs within the viral envelope. SARS-CoV-2 nucleocapsid protein is composed of multiple domains: the N-terminal domain, an RNA binding domain, a disordered central linker domain, a dimerization domain, and the C-terminal domain (CTD) [18]. It is generally accepted that SARS-CoV-2 N interacts with M through both the dimerization domain and the CTD [12]. The cryo-ET map (EMD-30429) has not only shown the likely N protein dimerization paradigm, but also proposed a plausible RNP constitution, which guided us to build a model of the complex structure of RNPs (Fig. S6). Then the RNP unit models were organized to fit the cryo-ET densities within the viral envelope. In addition, the RNA segments bound to N proteins were included in the RNP model to reproduce the interaction between N proteins and RNA as well (see SI 1.5 for details).

Moreover, the recent cryo-ET map (EMD-30430) described the copy number

and distribution of Ss and RNPs as well as the envelope at nanometer precision. Therefore, the spatial arrangement of these three virus constitutions (S proteins, RNPs, and the envelope) was fitted to EMD-30430. Ms and Es were indiscernible from the map and were allocated uniformly and randomly on the envelope. More details of the assemble protocol can be found in SI 1.6. With the above considerations, we have constructed an authentic SARS-CoV-2 from all four structural proteins and the envelope (Fig. 1a).

With the whole virus constructed (except for the complete RNA), a 200-ns CG simulation was performed in a water environment to relax each component to reach a more equilibrated configuration (Movie S1). Then, both the initial (Fig. S1a) and final (Fig. S1c) CG structures were transformed into atomistic models, as shown in Fig. S1b 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 present the most complete picture of the intact SARS-CoV-2 so far, and they can serve as a framework for future improvements. Also, the atomistic structures of the virus will enable us to investigate and understand the dynamics of the intact virus by using all-atom simulations, which may facilitate the development of new treatments for COVID-19.

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 [19]. Our work has 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. This is one of the first and most complete atomic-level models of SARS-CoV-2 that fully employs the *in situ* cryo-ET data, which 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.

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 Ministry of Science and Technology of China (2021YFE0108100 to C.S.), the National Natural Science Foundation of China (21873006 and 32071251 to C.S.; 32171195 to S.L.), and Tsinghua University Spring Breeze Fund (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) Hardenbrook, N. J.; Zhang, P. *Current Opinion in Virology* **2021**, *52*, 123–134.

(2) Yao, H. et al. *Cell* **2020**, *183*, 730–738.

(3) Klein, S.; Cortese, M.; Winter, S. L.; Wachsmuth-Melm, M.; Neufeldt, C. J.; Cerikan, B.; Stanifer, M. L.; Boulant, S.; Bartenschlager, R.; Chlanda, P. *Nature Communications* **2020**, *11*, 1–10.

(4) Yu, A.; Pak, A. J.; He, P.; Monje-Galvan, V.; Casalino, L.; Gaieb, Z.; Dommer, A. C.; Amaro, R. E.; Voth, G. A. *Biophysical Journal* **2021**, *120*, 1097–1104.

(5) Pezeshkian, W.; Grünewald, F.; Narykov, O.; Lu, S.; Wassenaar, T. A.; Marrink, S. J.; Korkin, D. *bioRxiv* **2021**, DOI: `10.1101/2021.09.15.459697`.

(6) Dommer, A. et al. *bioRxiv* **2021**, DOI: `10.1101/2021.11.12.468428`.

(7) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; De Vries, A. H. *Journal of Physical Chemistry B* **2007**, *111*, 7812–7824.

(8) Ivanova, P. T.; Myers, D. S.; Milne, S. B.; McClaren, J. L.; Thomas, P. G.; Brown, H. A. *ACS Infectious Diseases* **2015**, *1*, 399–452.

(9) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; Mclellan, J. S.; Fadda, E.; Amaro, R. E. *ACS Central Science* **2020**, *6*, 1722–1734.

(10) Walls, A. C.; Park, Y. J.; Tortorici, M. A.; Wall, A.; McGuire, A. T.; Veesler, D. *Cell* **2020**, *181*, 281–292.

(11) Hoffmann, M.; Kleine-Weber, H.; Schroeder, S.; Krüger, N.; Herrler, T.; Erichsen, S.; Schiergens, T. S.; Herrler, G.; Wu, N. H.; Nitsche, A.; Müller, M. A.; Drosten, C.; Pöhlmann, S. *Cell* **2020**, *181*, 271–280.

(12) Bracquemond, D.; Muriaux, D. *mBio* **2021**, *12*, 1–14.

(13) Jumper, J. et al. *Nature* **2021**, *596*, 583–589.

(14) Neuman, B. W.; Kiss, G.; Kunding, A. H.; Bhella, D.; Baksh, M. F.; Connelly, S.; Droese, B.; Klaus, J. P.; Makino, S.; Sawicki, S. G.; Siddell, S. G.; Stamou, D. G.; Wilson, I. A.; Kuhn, P.; Buchmeier, M. J. *Journal of Structural Biology* **2010**, *174*, 11–22.

(15) Bar-On, Y. M.; Flamholz, A.; Phillips, R.; Milo, R. *eLife* **2020**, *9*, 697–698.

(16) Heo, L.; Feig, M. *bioRxiv* **2020**, DOI: 10.1101/2020.03.25.008904.

(17) Wang, L.; Zhang, J.; Wang, D.; Song, C. *bioRxiv* **2021**, DOI: 10.1101/2021.01.17.426988.

(18) Ye, Q.; West, A. M.; Silletti, S.; Corbett, K. D. *Protein Science* **2020**, *29*, 1890–1901.

(19) Li, S. *Trends in Biochemical Sciences* **2021**, *47*, 173–186.

Figure

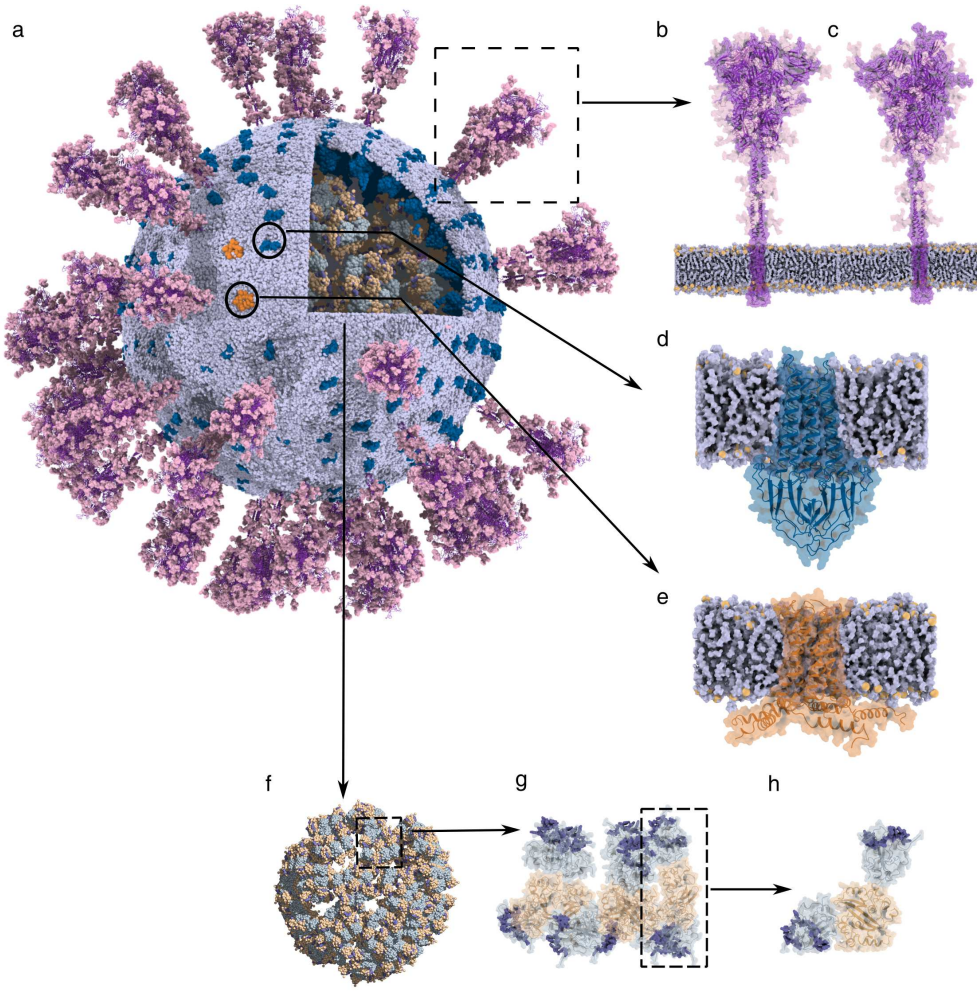


Figure 1: (a) The overview of the virus structure. The viral envelope is colored blue–white. Purple, deep blue, and orange regions indicate the S, M, and E proteins, respectively. RNPs are located within the envelope, and domains near the N-terminal and C-terminal of N proteins are shown in grey–blue 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 of 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.

Supplementary Information for:

Toward an atomistic model of SARS-CoV-2

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.

[*]To whom correspondence should be addressed; E-mail: sai@mail.tsinghua.edu.cn (SL), c.song@pku.edu.cn (CS)

1 Materials and Methods

Our atomistic model of SARS-CoV-2 was 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 [2, 3]. Then the CG system was transformed into an atomistic model. The details of the system construction are as follows:

1.1 Construction of the membrane envelope

We set up the initial CG vesicle with the CHARMM-GUI Vesicle Maker [4]. Since the vesicle would shrink after equilibration [4], we extended the initial vesicle diameter (D_{init} = 109 nm) to ensure it will reach the target diameter ($D_{target} \approx$ 85 nm) after equilibration (Fig. S2) 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 [5]. 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 mimicing the lipid ratio of the ERGIC (ER-Golgi intermediate compartment) membrane, where PC and PE are predominant [6, 7]. 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 [8].

The CG vesicle system was pre-equilibrated in a water box of $130 \times 130 \times 130$

nm^3 with the Martini2.2 force field [2, 3]. After a 10000-step energy minimization, the system was equilibrated in the NPT (isothermal-isobaric) ensemble for 300 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 [9, 10]. The pressure coupling was isotropic. The coupling time constants for both the pressure and temperature were set to 1.0 ps.

1.2 Structural model of the spike (S) protein

The initial atomistic structures of the 'one RBD up' (PDB ID: 6xm3) [11] and 'RBD down' (PDB ID: 6xr8) [12] 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 [13, 14]:

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 [13, 14], based on the secondary structure predictions by the web server SPIDER3 [15].

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 [16, 17]. 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 [18, 19], 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 [7, 19]. All of 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 [20] with the CHARMM36m force field [21]. The structure of the glycosylated full-length S proteins is shown in Fig. S3. The details of the glycan types on each glycosylated site are shown in Table S1.

1.3 Structural model of the membrane (M) protein

Previous studies showed that M proteins may form dimers on the virus envelope [22], so we built a dimeric structure of the M protein based on previous studies [22, 23] with the docking software ZDOCK [24].

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) [25]. 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 [26] was best scored in CASP14, we picked the monomeric structure predicted by AlphaFold2 to construct the M protein dimer (Fig. S4a). 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 [22]. The CTD (residue 101-222) locates at the intracellular domain and may interact with other structural proteins such as N proteins, therefore excluded for the M protein dimer construction. We limited the 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 complexes for evaluation and selection. The most probable and reasonable dimeric model for the construction of the virus structure (Fig. S4b) 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.

1.4 The envelope (E) protein

The E protein structure was published (PDB ID: 7k3g) [27], 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 [15, 28] 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. S5a) 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 [29] showed that while the residues 8–34 (forming the transmembrane helices in resolved E structure) entirely interact with membrane with high probability (Fig. S5b, green curve), the inner helices (residue 38–60) (Fig. S5b, 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 [30] showed that the endo helix is optimized to touch the viral envelope, which is consistent with our MCP prediction (Fig. S5c). Also, the E protein structure predicted by the Feig Lab has been proven to be stable in microsecond MD simulations [31]. 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 [32, 33]. Therefore, we picked the predicted structural model by the Feig Lab as the initial E protein structure for our virus construction.

1.5 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) [33]. 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 [34–37]. 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 [38].

The N-terminal RBD (PDB ID: 7act) [39] and C-terminal dimerization domain (PDB ID: 6yun) [40] 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 [24] to construct an N protein dimer (Fig. S6a).

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. S6b).

1.6 Assemble of the SARS-CoV-2 virus

After the envelope and structural proteins were set up as described above, we assembled all of 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 [41] 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. S7a). 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. S7b). Then the optimized S protein structures were transformed to CG model in the Martini force field [2, 3]. 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. S8), 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. S6b) were fitted into the density map (EMD-30430), where all the RNPs are nestled up to the inner surface of the viral envelope (Fig. S9a). 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. S9b). 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 pro-

tein 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. S9c).

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 protein:N protein ranges from 1:1 to 3:1 [22, 42]. 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 [43]. 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 $155 \times 155 \times 155$ nm^3 sized simulation box contained 31,226,794 CG beads in total.

1.7 Coarse-grained molecular dynamics simulations

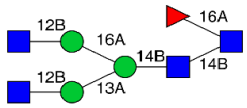
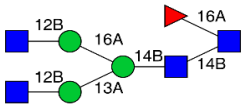
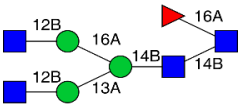
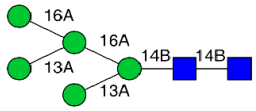
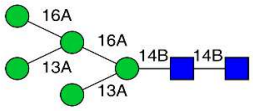
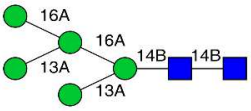
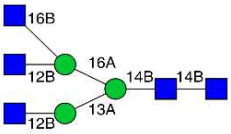
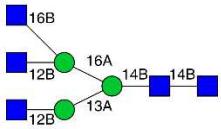
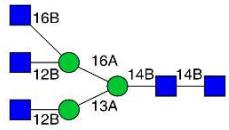
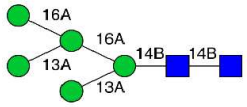
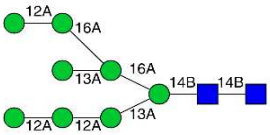
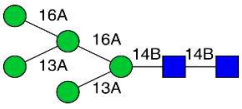
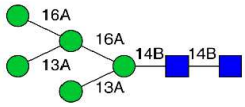
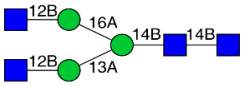
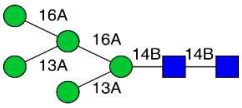
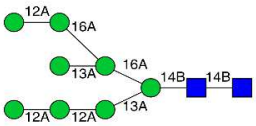
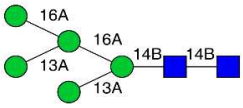
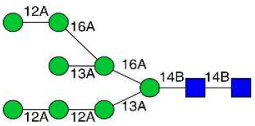
All of the MD simulations were performed with the software GROMACS 2018.4. The CG simulation system was first performed an energy minimization using the steepest descent algorithm for 30,000 steps, followed by equilibration in the NPT ensemble (constant pressure and constant temperature) for 25 ns with 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 [10]. 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 4.5×10^{-5} 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 200-ns CG MD production simulation with the time step of 5 fs.

1.8 Conversion from the CG system to the atomistic system

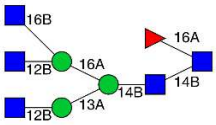
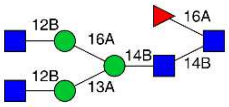
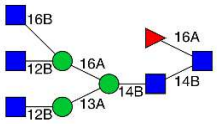
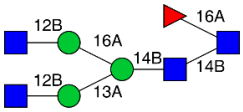
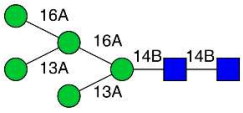
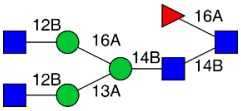
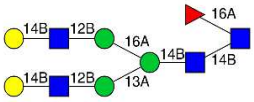
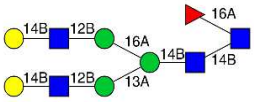
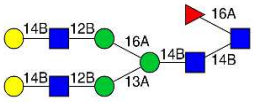
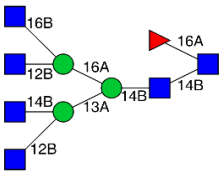
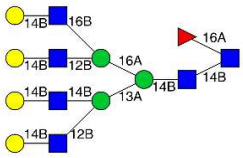
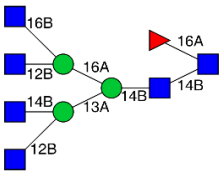
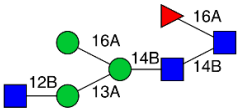
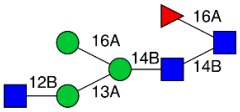
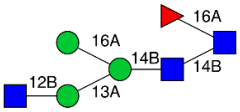
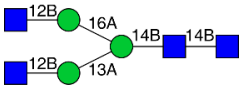
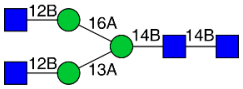
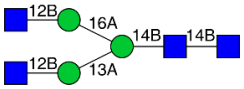
Lastly, the structures of the first and last frames of CG MD simulations were transformed to atomistic systems by the CG2AT2 tool [44]. There are 372,617,567 atoms in the entire atomistic system of the simulation box, in which the SARS-CoV-2 contains 14,182,463 atoms. Although challenging at the moment, the atomistic system can serve as the initial structure for future all-atom simulations.

2 Supplementary Tables

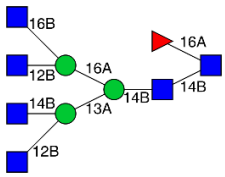
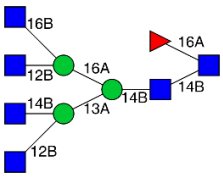
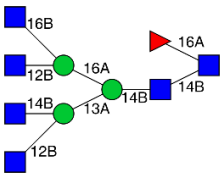
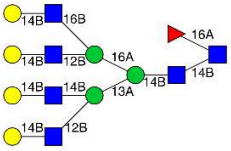
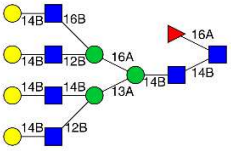
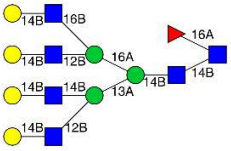
Table S1. The glycan types of the glycosylated sites of the S protein

	Chain A	Chain B	Chain C
N17, N149 N1098			
N61, N603 N709, N717			
N74			
N122, N801			
N165			
N234			

GlcNAc Mannose Fucose

	Chain A	Chain B	Chain C
N282			
N331, N343 N616			
N657			
N1074			
N1134			
N1158			

GlcNAc Mannose Fucose Galactose

	Chain A	Chain B	Chain C
N1173			
N1194			
T323			
S325		—	—

GlcNAc Mannose Fucose Galactose GalNAc

3 Supplementary Figures

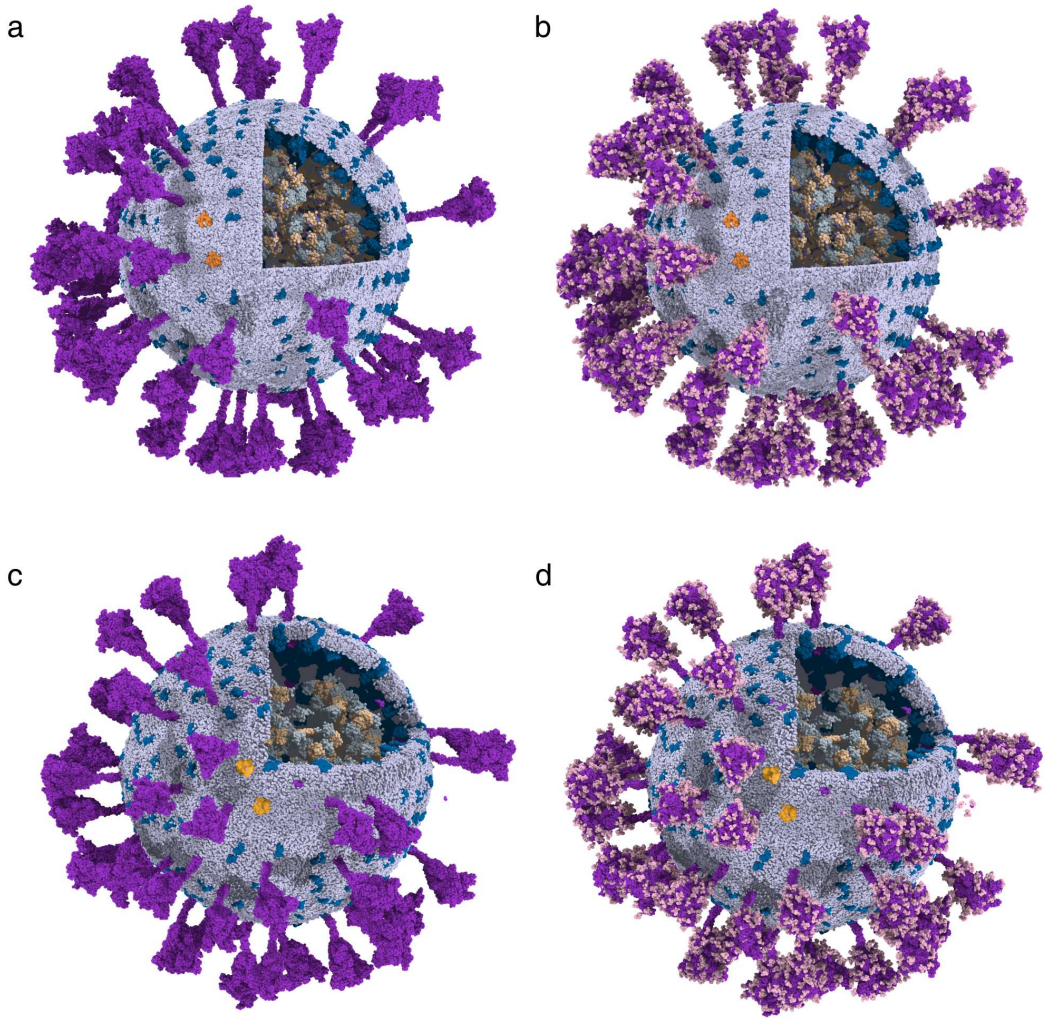


Figure S1: Overview of the coarse-grained and atomistic SARS-CoV-2. The viral structural proteins were colored in purple (S), deep-blue (M), orange (E), grey-blue (NTD and RBD of N), and wheat (CTD and dimerization domain of N). The viral envelope was rendered in blue-white. RNA segments were in blue-purple. In the atomistic system, the glycans were colored in light pink. (a) The initial coarse-grained structure of the virus; (b) The atomistic virus structure transformed from (a); (c) The final coarse-grained virus structure after the 200-ns CG MD simulation; (d) The atomistic virus structure transformed from (c).

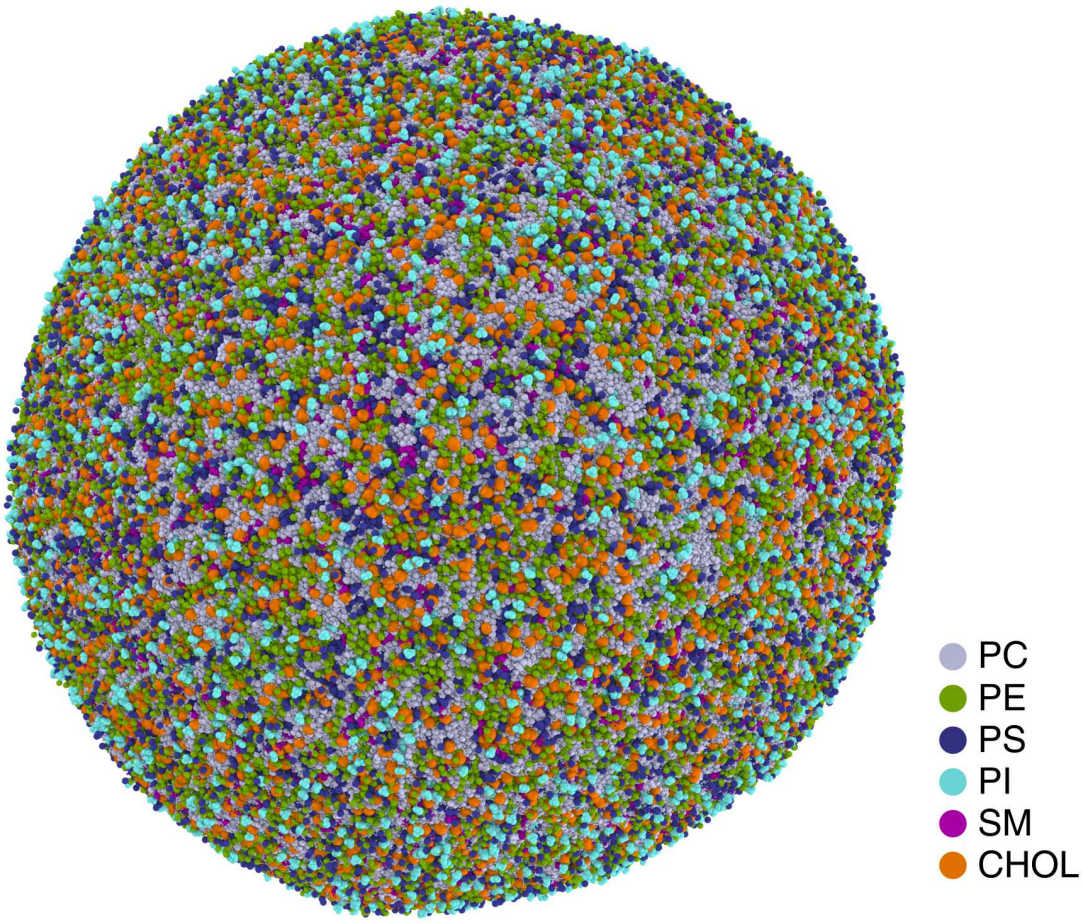


Figure S2: The overview of the equilibrated vesicle.

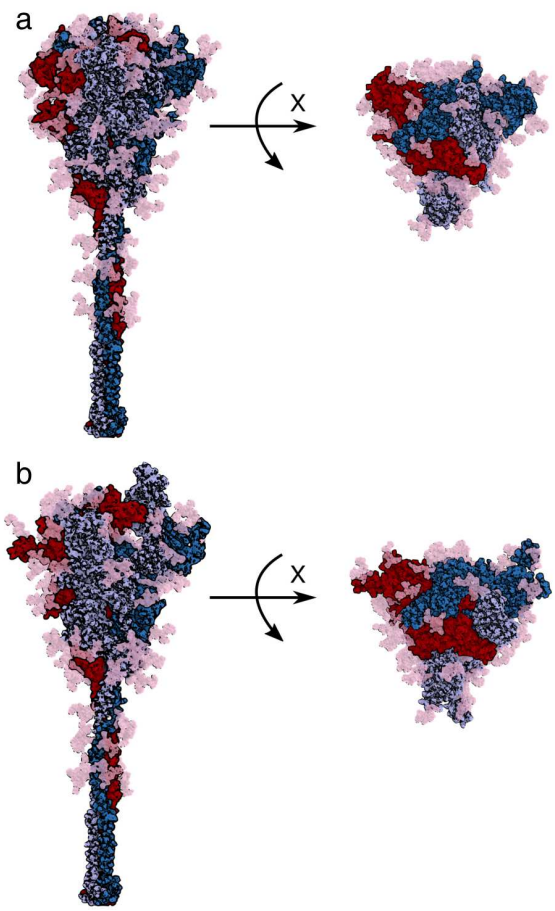


Figure S3: Structures of the spike protein. (a) Full-length 'RBD down' state. The three chains were colored blue-white, sky blue, and red, respectively. The glycans were colored in light pink. (b) same as (a) but for the 'one RBD up' state.

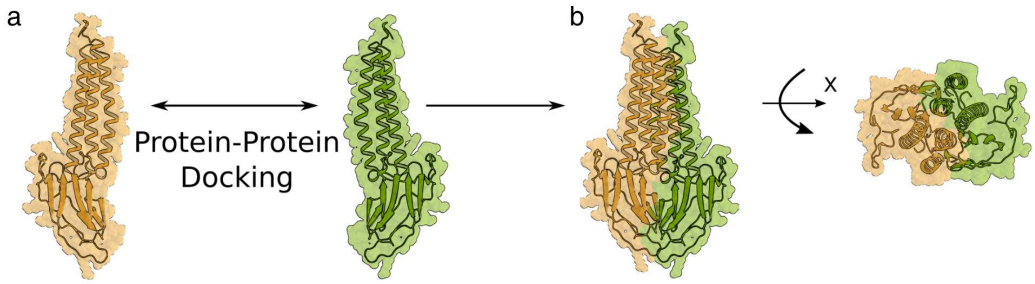


Figure S4: (a) The M protein monomer structure (green and orange) predicted by AlphaFold2; (b) The side and top views of the M dimer structure generated by Protein–Protein docking (ZDOCK).

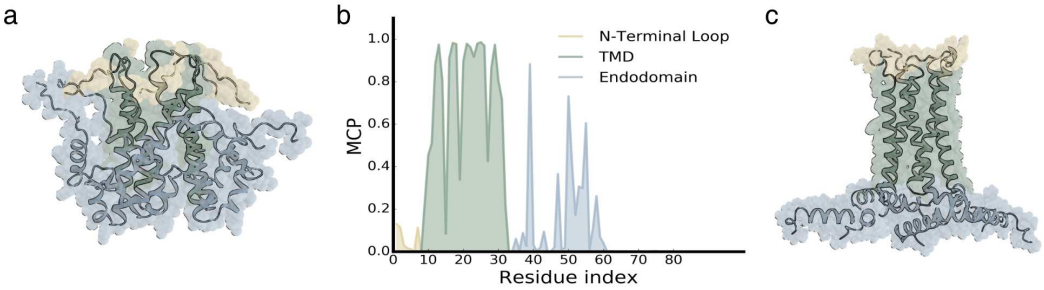


Figure S5: The structures of the E protein pentamer. (a) The homology modeling structure based on the SARS-CoV E protein; (b) The predicted membrane contact probability (MCP) of the E protein. (c) The E protein pentameric structure optimized by Feig's Lab. The brown, green, and blue regions represent the N-terminal loop, the transmembrane domain, and endodomain of the E protein, respectively.

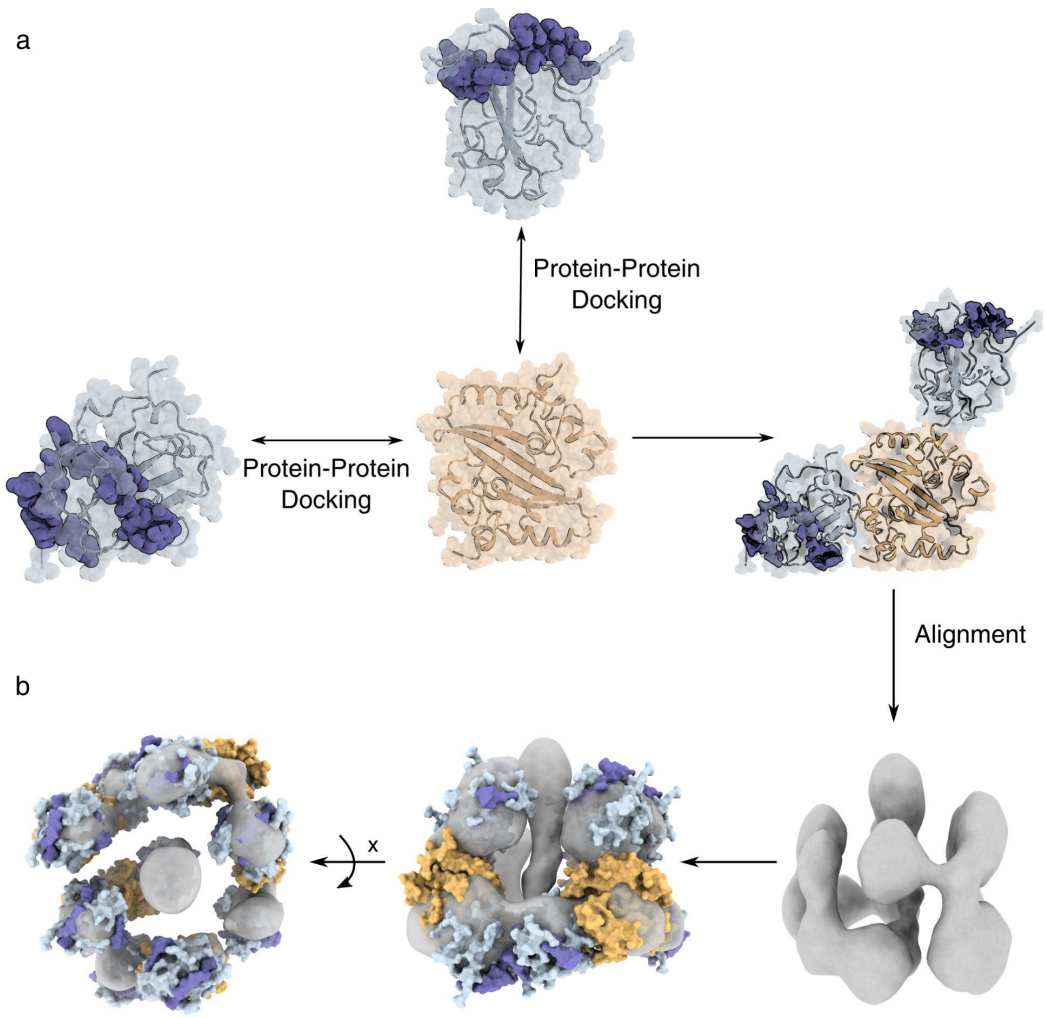


Figure S6: Construction of the RNP unit structure. (a) The N protein dimer (right panel) was generated by docking two NTD structures (light blue) to the dimerized CTD structure (light orange). The RNA segments bound to N proteins were colored deep and opaque blue; (b) Five N protein dimers are fitted into the Cryo-ET density (grey surface) to form a single RNP unit (left panel).

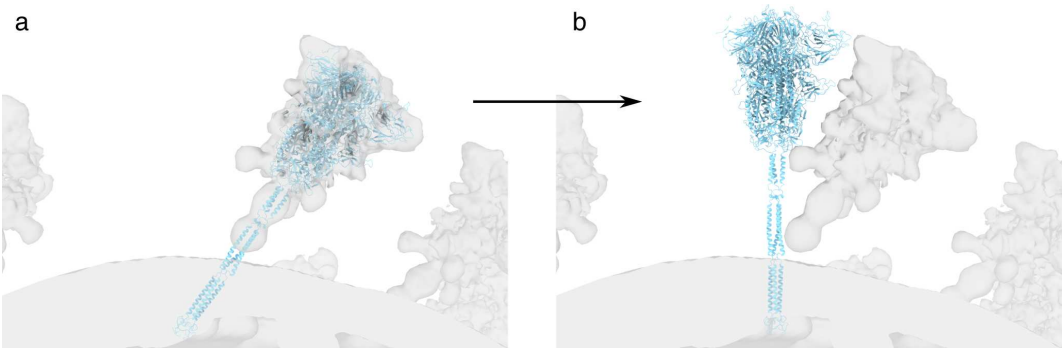


Figure S7: (a) Aligning the S protein into the Cryo-ET density (Chimera –> Fit into density) may lead to inappropriate positioning of the transmembrane domain of the S protein. (b) The adjusted S protein orientation with the transmembrane domain properly embedded into the lipid bilayer. The light blue cartoon indicates the structure of the S protein. The grey surface shows the Cryo-ET density.

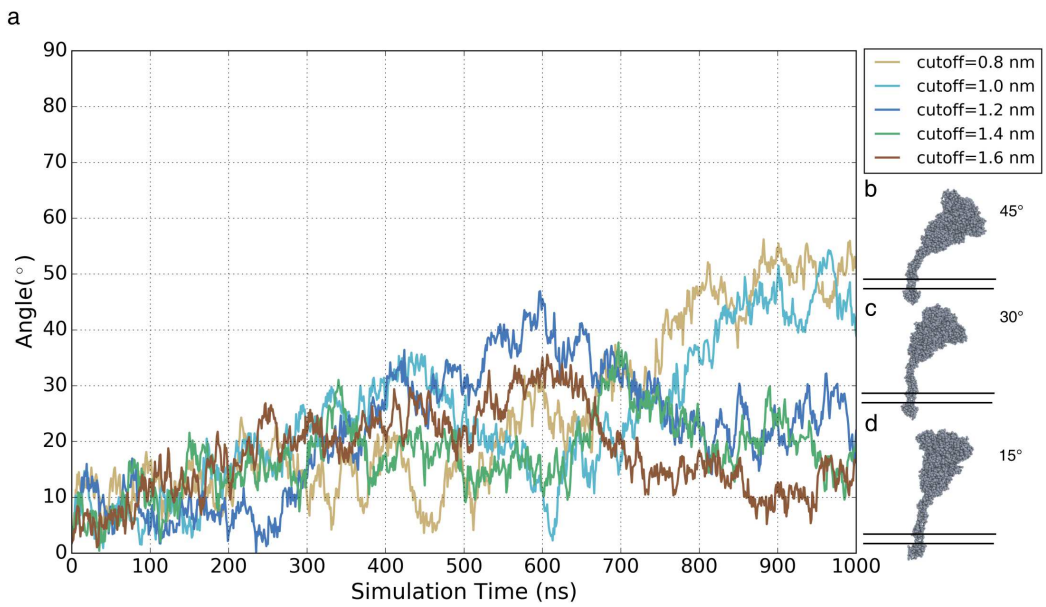


Figure S8: The orientation of the spike protein on the lipid bilayer depends on the elastic network cutoff. (a) Each curve elucidates the spike protein orientation distribution in CG MD simulations with different elastic network cutoff values. The angle was measured between the axis of the spike protein and the normal axis of the membrane surface. (b)–(d) The typical S protein orientations during the simulations.

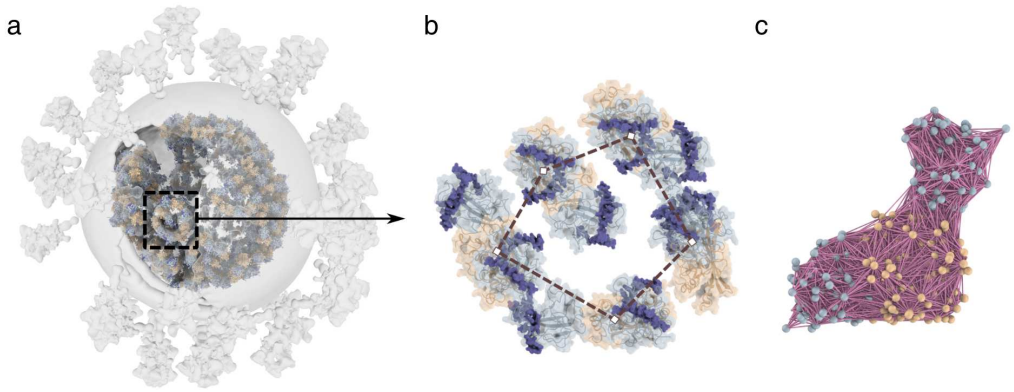


Figure S9: The RNPs and the distance restraints within them. (a) Alignment of the RNP units into the Cryo-ET density map. (b) The zoom-in view of the black frame in (a). The dashed lines represent the distance restraints applied to the N protein dimers in each RNP unit. (c) Magenta lines indicate the pairwise elastic network to maintain the N protein dimer conformation. Light blue and light orange represent domains near the NTD and CTD of the N proteins, respectively. Deep blue indicates the RNA segments bound to N proteins.

References

(1) Yao, H. et al. *Cell* **2020**, *183*, 730–738.

(2) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; De Vries, A. H. *Journal of Physical Chemistry B* **2007**, *111*, 7812–7824.

(3) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. *Journal of Chemical Theory and Computation* **2008**, *4*, 819–834.

(4) Qi, Y.; Ingólfsson, H. I.; Cheng, X.; Lee, J.; Marrink, S. J.; Im, W. *Journal of Chemical Theory and Computation* **2015**, *11*, 4486–4494.

(5) Woo, H.; Park, S. J.; Choi, Y. K.; Park, T.; Tanveer, M.; Cao, Y.; Kern, N. R.; Lee, J.; Yeom, M. S.; Croll, T. I.; Seok, C.; Im, W. *Journal of Physical Chemistry B* **2020**, *124*, 7128–7137.

(6) Turoňová, B. et al. *Science* **2020**, *370*, 203–208.

(7) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; Mclellan, J. S.; Fadda, E.; Amaro, R. E. *ACS Central Science* **2020**, *6*, 1722–1734.

(8) Casares, D.; Escrib, P. V.; Rosselló, C. A. *International Journal of Molecular Science* **2019**, *20*, 2167.

(9) Bussi, G.; Donadio, D.; Parrinello, M. *Journal of Chemical Physics* **2007**, *126*, 1–8.

(10) Berendsen, H. J.; Postma, J. P.; Van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *The Journal of Chemical Physics* **1984**, *81*, 3684–3690.

(11) Zhou, T. et al. *SSRN Electronic Journal* **2020**, *28*, 867–880.

(12) Cai, Y.; Zhang, J.; Xiao, T.; Peng, H.; Sterling, S. M.; Walsh, R. M.; Rawson, S.; Rits-Volloch, S.; Chen, B. *Science* **2020**, *369*, 1586–1592.

(13) Sali, A.; Blundell, T. L. *Journal of Molecular Biology* **1993**, *234*, 779–815.

(14) Fiser, A.; Kinh Gian Do, R.; Sali, A. *Protein Science* **2000**, *9*, 1753–1773.

(15) Heffernan, R.; Yang, Y.; Paliwal, K.; Zhou, Y. *Bioinformatics* **2017**, *33*, 2842–2849.

(16) Yang, T. J.; Chang, Y. C.; Ko, T. P.; Draczkowski, P.; Chien, Y. C.; Chang, Y. C.; Wu, K. P.; Khoo, K. H.; Chang, H. W.; Danny Hsu, S. T. *Proceedings of the National Academy of Sciences of the United States of America* **2020**, *117*, 1438–1446.

(17) Watanabe, Y.; Berndsen, Z. T.; Raghwani, J.; Seabright, G. E.; Allen, J. D.; Pybus, O. G.; McLellan, J. S.; Wilson, I. A.; Bowden, T. A.; Ward, A. B.; Crispin, M. *Nature Communications* **2020**, *11*, 1–10.

(18) Watanabe, Y.; Allen, J. D.; Wrapp, D.; McLellan, J. S.; Crispin, M. *Science* **2020**, *369*, 330–333.

(19) Shajahan, A.; Supekar, N. T.; Gleinich, A. S.; Azadi, P. *Glycobiology* **2020**, *30*, 981–988.

(20) Park, S. J.; Lee, J.; Qi, Y.; Kern, N. R.; Lee, H. S.; Jo, S.; Joung, I.; Joo, K.; Lee, J.; Im, W. *Glycobiology* **2019**, *29*, 320–331.

(21) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. *Nature Methods* **2016**, *14*, 71–73.

(22) Neuman, B. W.; Kiss, G.; Kunding, A. H.; Bhella, D.; Baksh, M. F.; Connelly, S.; Droese, B.; Klaus, J. P.; Makino, S.; Sawicki, S. G.; Siddell, S. G.; Stamou, D. G.; Wilson, I. A.; Kuhn, P.; Buchmeier, M. J. *Journal of Structural Biology* **2010**, *174*, 11–22.

(23) Schoeman, D.; Fielding, B. C. *Virology Journal* **2019**, *16*, 69–90.

(24) Pierce, B. G.; Hourai, Y.; Weng, Z. *PLoS ONE* **2011**, *6*, 1–5.

(25) De Haan, C. A. M.; Vennema, H.; Rottier, P. J. M. *Journal of Virology* **2000**, *74*, 4967–4978.

(26) Jumper, J. et al. *Nature* **2021**, *596*, 583–589.

(27) Mandala, V. S.; McKay, M. J.; Shcherbakov, A. A.; Dregni, A. J.; Kolocouris, A.; Hong, M. *Nature Structural and Molecular Biology* **2020**, *27*, 1202–1208.

(28) Wang, S.; Li, W.; Liu, S.; Xu, J. *Nucleic Acids Research* **2016**, *44*, 430–435.

(29) Wang, L.; Zhang, J.; Wang, D.; Song, C. *bioRxiv* **2021**, DOI: 10.1101/2021.01.17.426988.

(30) Heo, L.; Feig, M. *bioRxiv* **2020**, DOI: 10.1101/2020.03.25.008904.

(31) Monje-Galvan, V.; Voth, G. A. *Faraday Discussions* **2021**, *232*, 49–67.

(32) Arias-Reyes, C.; Zubieta-DeUrioste, N.; Poma-Machicao, L.; Aliaga-Raduan, F.; Carvajal-Rodriguez, F.; Dutschmann, M.; Schneider-Gasser, E. M.; Zubieta-Calleja, G.; Soliz, J. *Respiratory Physiology and Neurobiology* **2020**, *277*, 1–22.

(33) Bracquemond, D.; Muriaux, D. *mBio* **2021**, *12*, 1–14.

(34) Keane, S. C.; Lius, P.; Leibowitzs, J. L.; Giedroc, D. P. *Journal of Biological Chemistry* **2012**, *287*, 7063–7073.

(35) Tan, Y. W.; Fang, S.; Fan, H.; Lescar, J.; Liu, D. X. *Nucleic Acids Research* **2006**, *34*, 4816–4825.

(36) Grossoehme, N. E.; Li, L.; Keane, S. C.; Liu, P.; Dann, C. E.; Leibowitz, J. L.; Giedroc, D. P. *Journal of Molecular Biology* **2009**, *394*, 544–557.

(37) Saikatendu, K. S.; Joseph, J. S.; Subramanian, V.; Neuman, B. W.; Buchmeier, M. J.; Stevens, R. C.; Kuhn, P. *Journal of Virology* **2007**, *81*, 3913–3921.

(38) Lu, S.; Ye, Q.; Singh, D.; Cao, Y.; Diedrich, J. K.; Yates, J. R.; Villa, E.; Cleveland, D. W.; Corbett, K. D. *Nature Communications* **2021**, *12*, 502.

(39) Dinesh, D. C.; Chalupska, D.; Silhan, J.; Koutna, E.; Nencka, R.; Veverka, V.; Boura, E. *PLoS Pathogens* **2020**, *16*, 1–16.

(40) Zinzula, L.; Basquin, J.; Bohn, S.; Beck, F.; Klumpe, S.; Bracher, A.; Hartl, F. U.; Baumeister, W. *Biochemical and Biophysical Research Communications* **2020**, *538*, 54–62.

(41) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Meng, E. C.; Couch, G. S.; Croll, T. I.; Morris, J. H.; Ferrin, T. E. *Protein Science* **2021**, *30*, 70–82.

(42) Neuman, B. W.; Adair, B. D.; Yoshioka, C.; Quispe, J. D.; Orca, G.; Kuhn, P.; Milligan, R. A.; Yeager, M.; Buchmeier, M. J. *Journal of Virology* **2006**, *80*, 7918–7928.

(43) Bar-On, Y. M.; Flamholz, A.; Phillips, R.; Milo, R. *eLife* **2020**, *9*, 697–698.

(44) Vickery, O. N.; Stansfeld, P. J. *Journal of Chemical Theory and Computation* **2021**, *17*, 6472–6482.