TRACKING THE MOLECULAR EVOLUTION OF THE SARS-COV-2 VIRUS

By Michael Keith

Supervised by Javier Arsuaga, PhD

Table of Contents

Abstract

SARS-CoV-2 is a highly contagious respiratory virus. Currently, there exist multiple strains of the virus with some variants offering enhanced infectivity. Previously, understanding the structure of the proteins of the virus was crucial in the development of therapeutic solutions – such as vaccines, manufactured antibodies, etc. – to ameliorate the deleterious effects of the virus on humans. With the rise of the new, more infectious variants, the elucidation of the S protein structure is paramount to the continued efficacy of existing therapeutic solutions. For example, the delta variant of the virus exhibits exacerbated infectivity and severity due to genetic alterations, which causes subsequent changes in its S-protein structure relative to the wild type (Wuhan) virus. By and large, differences in infectivity and subsequent severity of the COVID-19 are due to better binding affinities of mutant viral binding domains to the host receptor relative to the wild type. Thus, if we wish to establish a degree of control over the virus, we must continue to understand what conformation changes occur because of mutations (and their combinations) of the viral genome. Ideally, we will be able to synthesize vaccines based on a forecast of the potential SARS-CoV-2 mutants in an analogous fashion to the annual production of vaccines for the influenza virus. Here we generate a high-resolution 3D reconstruction of the S-protein monomer using a cryo-EM analysis software called CryoSPARC. This software applies complex mathematical algorithms to analyze micrograph images taken by an electron microscope to produce 3D models of microscopic particles of interest.

Introduction

The S-protein is a trimeric, metastable glycoprotein [28]. It typically exists as a homotrimer (three identical proteins interacting via a quaternary structure) and has a variable conformation depending on the state of its receptor binding domain (RBD). Each protomer has two main subunits. The S1 subunit of the protein comprises the N-terminal domain, the RBD (C-terminal domain), and the two SD1 and SD2 subdomains [6]. The S2 subunit is the transmembrane portion of the S-protein and incorporates the N-terminal hydrophobic fusion peptide (FP), the heptad repeats (HR1 and HR2), the transmembrane domain (TM), and the cytoplasmic tail (CT) [6, 28]. Overall, the protein undergoes significant changes that result in viral fusion to the host cell. These changes are facilitated by the hinge-like movement of the RBD – the RBD is either up or down, exposing or occluding the binding elements, respectively, which leads to the metastability of the protein. During fusion, the exposed RBD surface binds to cognate receptors in the host organisms [6]. For example, the primary point-of-entry for the SARS-CoV-2 virus in

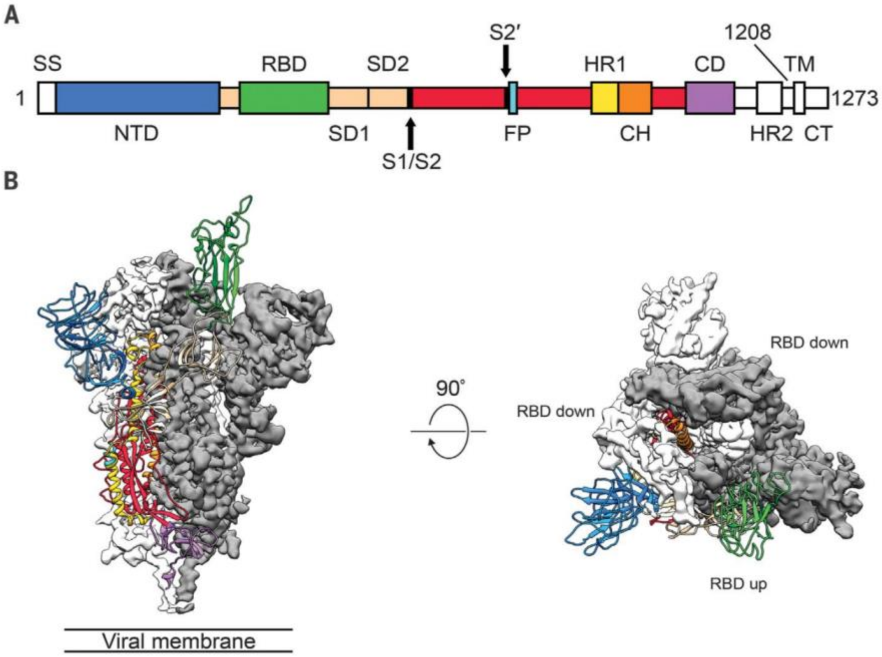
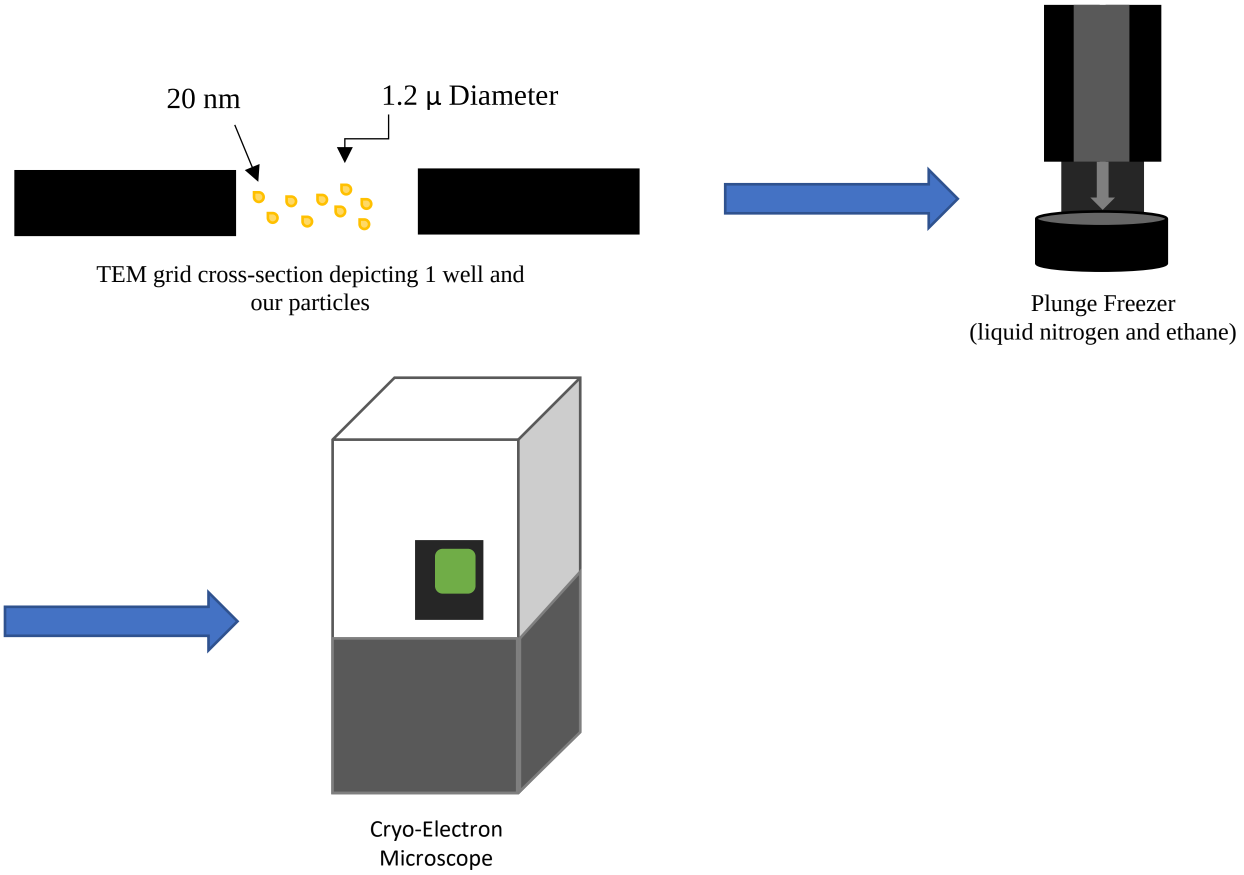


Image by Wrapp, et al [1]

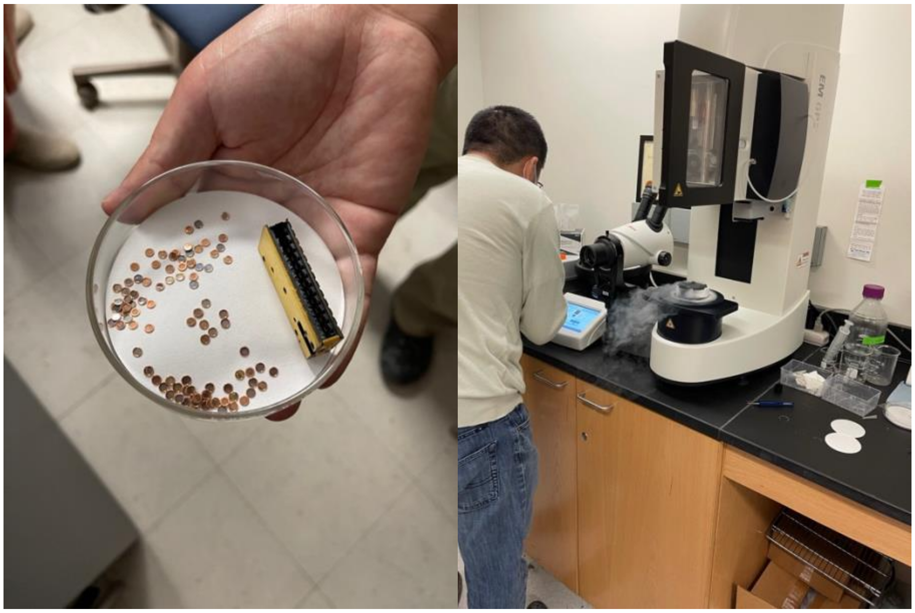
humans is via the angiotensin-converting enzyme 2 (ACE2) cell receptor [10, 15]. The above image depicts the various 2D and 3D motifs of the S protein, and highlights (in green) an RBD in the ‘up’ (binding elements exposed) conformation.

Methods

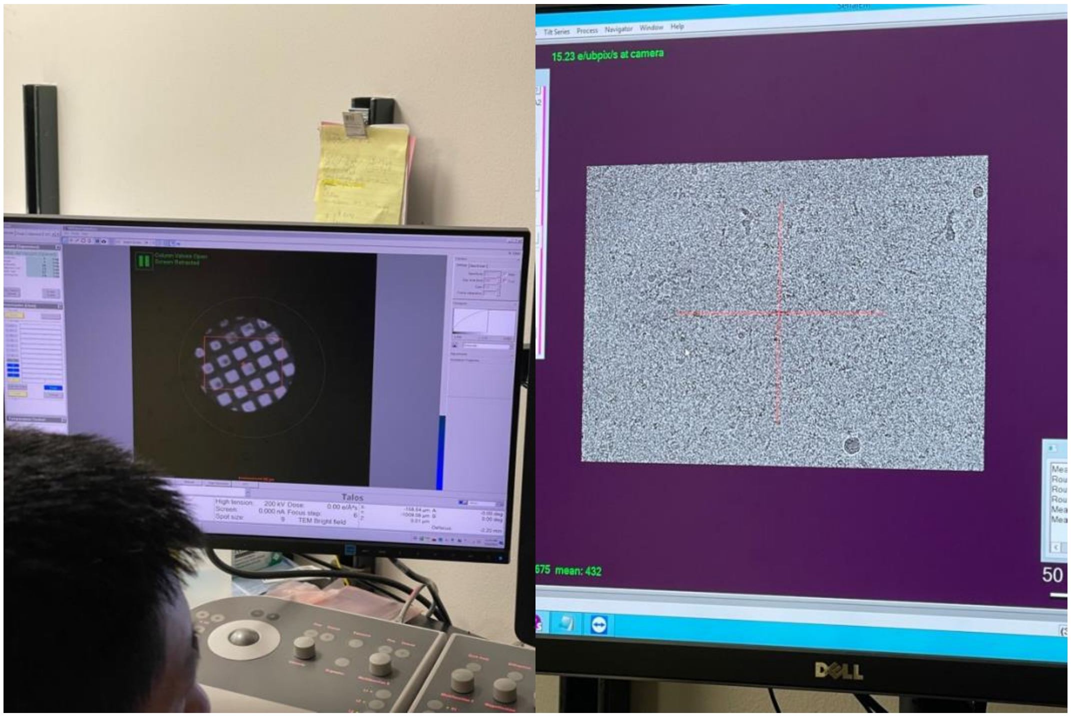
We seek to generate a 3D model of the S-protein monomer, capturing the different possible RBD states. Hence, we performed single-particle cryo-electron microscopy with S-protein monomers using a Thermo Fisher Glacios cryo-transmission electron microscope (cryo-TEM) equipped with a Gatan K3 electron detector followed by analysis in CryoSPARC. A typical workflow for cryo-EM is depicted below, and it summarizes the work we performed.


20 nm
1.2 μ Diameter
TEM grid cross-section depicting 1 well and our particles
Plunge Freezer
(liquid nitrogen and ethane)
Cryo-Electron Microscope

We employed transmission electron microscope (TEM) grids with wells of 1.2 microns in diameter. Next, we loaded our S-protein monomers on the grids via micropipettes – our S protein monomers are approximately 20 nm in length. Next, we inserted our prepared grids into a Leica EM GP2 plunge freezer, which plunged them into a Dewar of liquid nitrogen cooled by liquid ethane. The following are images of unprepared grids and the plunge freezer prior to loading the samples:



Next, we took images of the grids using the electron microscope. Afterward, we processed the micrograph images in CryoSPARC. Below are pictures of the cryo-EM imaging steps:



The left-most picture depicts the focusing of the electron beam near a well in one of the grids. The right picture demonstrates the micrograph captured by the microscope within the well. We introduced defocus by positioning the crosshair in the left image away from the center of the well. We corrected for this defocus later when processing the micrographs in CryoSPARC. Additionally, there are several large blobs in the micrograph on the right. These are 'junk' particles – dust, pieces of ice, or other foreign materials that are not our S-protein monomer. Hence, we excluded these micrographs and particles whenever possible in our later analyses.

CryoSPARC Jobs

The CryoSPARC software operates via workflow consisting of numerous jobs. Each job represents a particular pre-processing or refinement method that aids in the selection of useable micrographs and particles. By 'useable,' we mean particles that are not cut-off in the micrograph, exhibit high resolution, and take on different 3D orientations; and micrographs that do not contain numerous junk particles, contain large ice crystals or holes from the electron beam, and demonstrate the particles we wish to analyze. The following is a summary of the workflow we utilized:

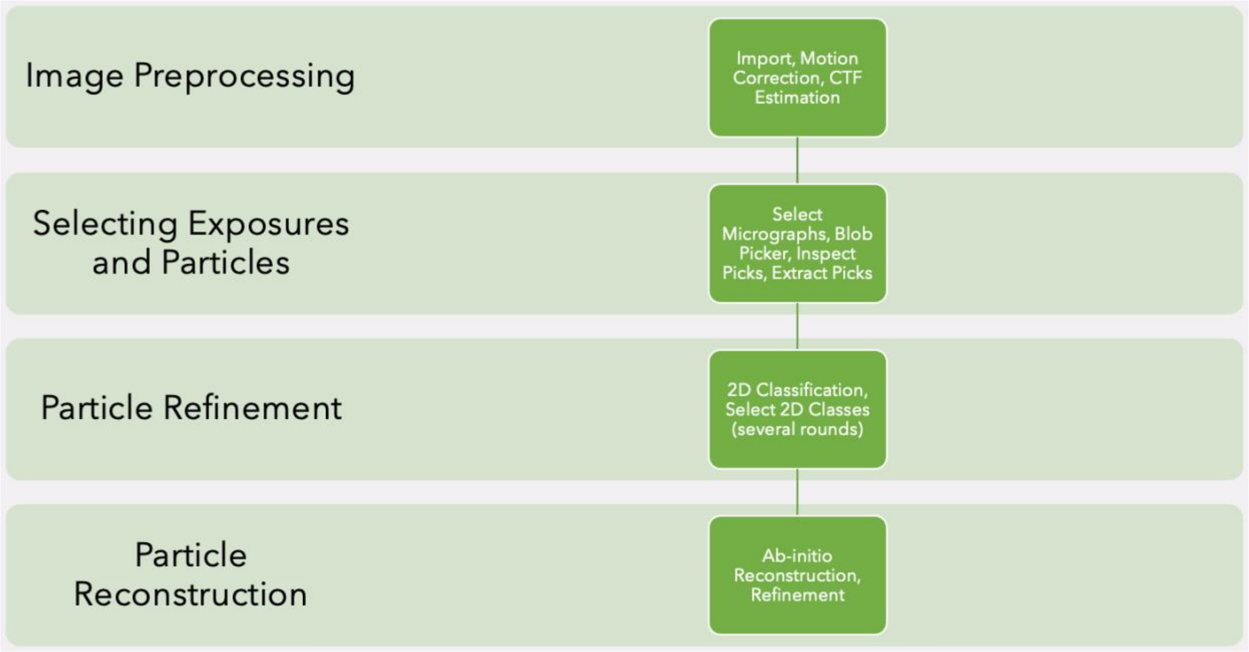


Image Preprocessing
Import, Motion Correction, CTF Estimation
Selecting Exposures and Particles
Select Micrographs, Blob Picker, Inspect Picks, Extract Picks
Particle Refinement
2D Classification, Select 2D Classes (several rounds)
Particle Reconstruction
Ab-initio Reconstruction, Refinement

Each of the jobs on the right are categorized on the left by the general type of processing performed.

(1) Image Pre-processing:

We first imported the micrographs taken during cryo-EM into CryoSPARC. We had a raw pixel size of 0.44 Å with spherical aberration of 2.7 mm. Additionally, we used an accelerating voltage of 200 kV and a total exposure dose of 60 e/Å^2. Next, we ran a motion correction job to detect and correct for the motion of the particles during the imaging process, binning to half the resolution. We then ran a Patch CTF Estimation job to correct for the interference inherent to the electron beam.

(2) Selecting Exposure and Picking Particles:

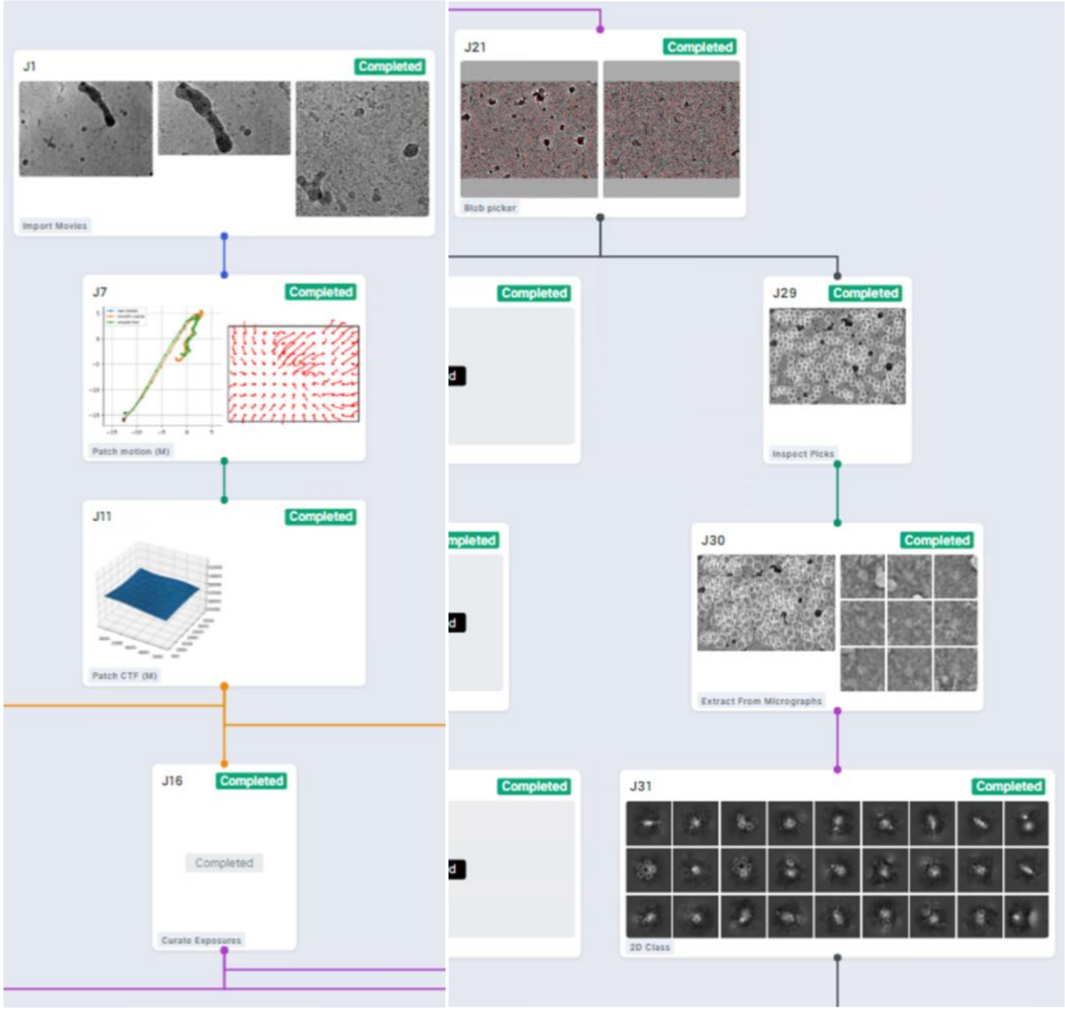
The set of jobs in this category relied on user-input to choose the best micrographs and particles. We chose micrographs that have a CTF resolution cut-off of 5 Å and whose full-frame motion distance was within 22-32 pixels. For Blob Picking (choosing particles), we let the software automatically choose particles given an elliptical blob with major axis length 150 Å and minor axis length 100 Å. Next, we ran an Inspect Picks job that allowed us to eliminate picks the software made that had low-resolution data using a power threshold of 552-800 and a normalized cross correlation (NCC) score of 0.18. Finally, we ran an Extract Picks job, which recorded the coordinates of the selected particles on the micrograph.

(3) Particle Refinement:

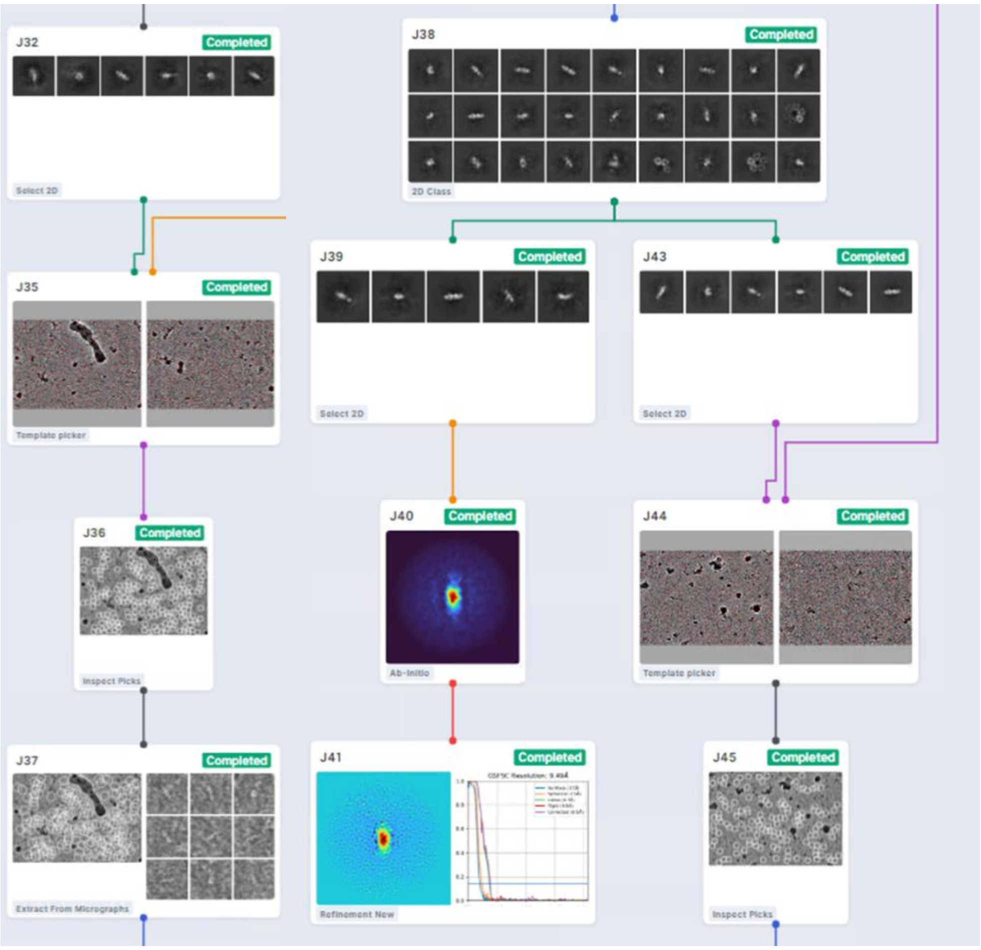
This set of jobs allowed the chosen particles from the previous batch of jobs to be classified into different classes based on their different symmetries and 3D orientations. A class average resolution was assigned by the software to aid in the selection of the best (highest resolution) classes. Initially, we permitted CryoSPARC generate 40 different classes of particles, of which we chose the best 6. Following this, we re-ran the jobs from (2) and (3) using this subset of micrographs. We then selected the 5 best 2D classes.

(4) Particle Reconstruction:

Taking the information from the previous sets of jobs, we performed an *ab-initio* 3D reconstruction of the particle of interest. Further non-uniform refinement jobs were ran using the output of the *ab-initio* job, which resulted in ~9.5 Å and ~8 Å resolution 3D models. Following the images below demonstrating the Cryo-SPARC workflow and the first set of 2D classes, we outline the general mathematical principles behind the *ab-initio* reconstruction job.



Figures 1 and 2 First set of CryoSPARC jobs depicting pre-processing and selection of particles steps as well as the first 2D classification



Figures 3 and 4 CryoSPARC jobs representing the particle refinement steps and an initial check of the ab-initio reconstruction

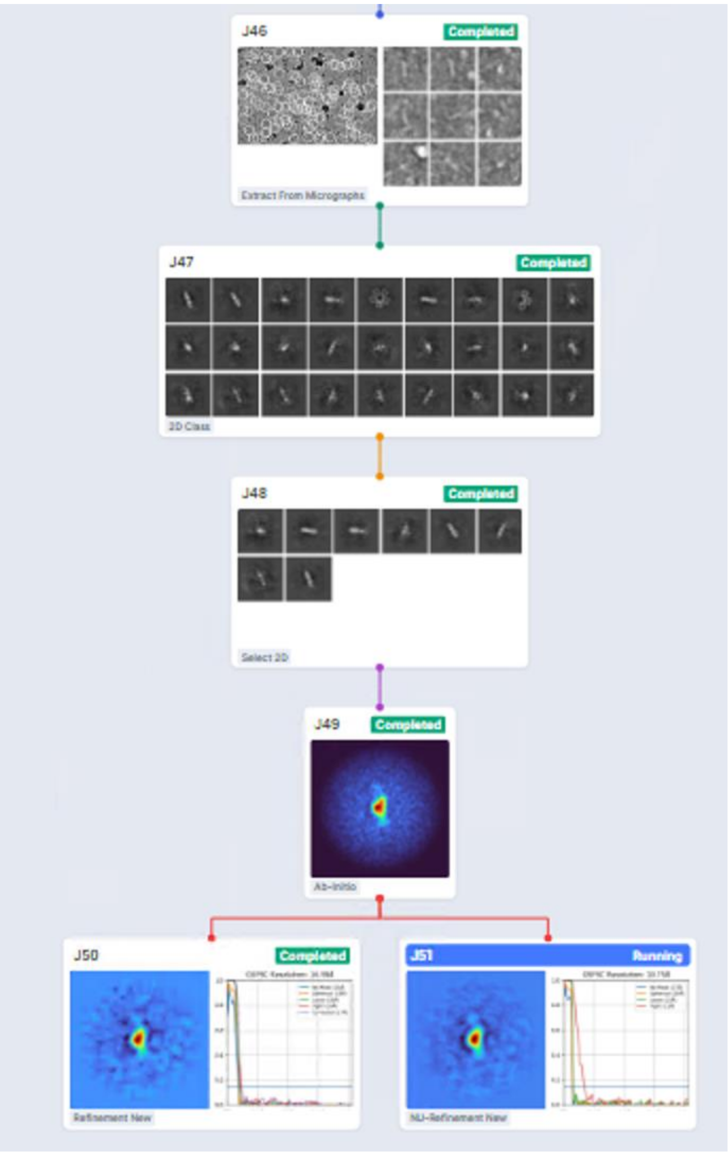


Figure 5 The final ab-initio reconstruction and subsequent non-uniform refinement step

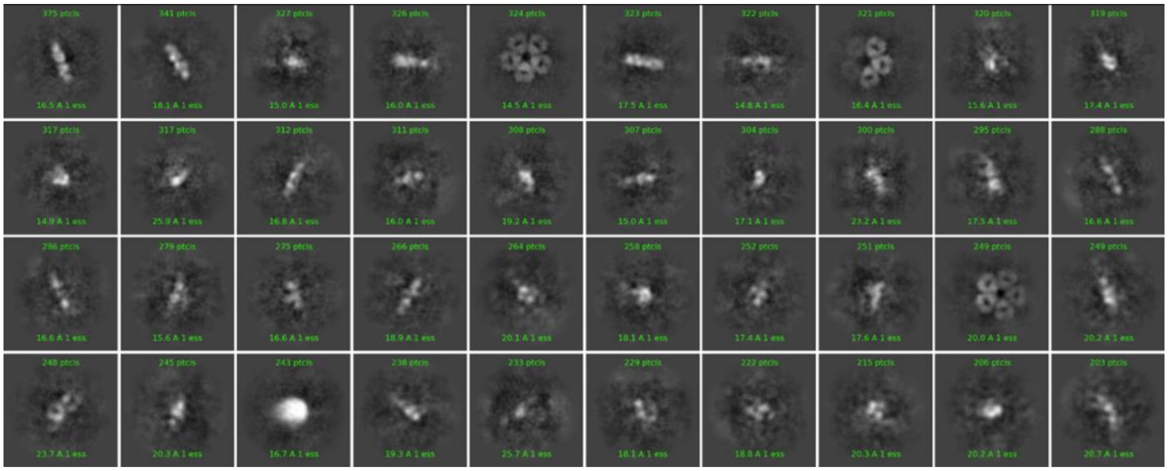


Figure 6 Initial 2D classes

Above is an image of the first 40 2D classes. Notice that some classes contain junk information (5-fold symmetry, white blobs, low resolution particles), and many classes have low-resolution information on average. We did not select these junk classes, and we re-ran the 2D classification several times using only the good classes to get cleaner data to work with.

Mathematical Methods

The *ab-initio* reconstruction relies on a Bayesian framework and makes use of the following function

$$\mathrm{argmax}_{\sigma_{1\ldots K}}\log(p(\sigma_{1\ldots K}|X_1,\ X_2,\ X_3,\ \ldots,\ X_N))$$

$$= \mathrm{argmax}_{\sigma_{1\ldots K}}\sum_{i=1}^{N}\log(\sum_{j=1}^{K}\frac{1}{K}\int p(X_i,\ \phi_i|\sigma_i)d\phi_i + \log(p(\sigma_{1\ldots K}))),$$

where p is the probability of generating some 3D structure, σ, given the images X_i taken; argmax is the set of points/coordinates where p is maximized; and ϕ is the pose or the 3D rotation and 2D translation of the particle [20]. A stochastic gradient descent algorithm is utilized to find the most probable 3D structure given the data [20]. Other methods may result in the same conclusion (3D structure), but they often ‘get stuck’ at local optima, which do not represent the typical 3D structure of the particle [20]. Below is a basic graph that represents this in more detail.

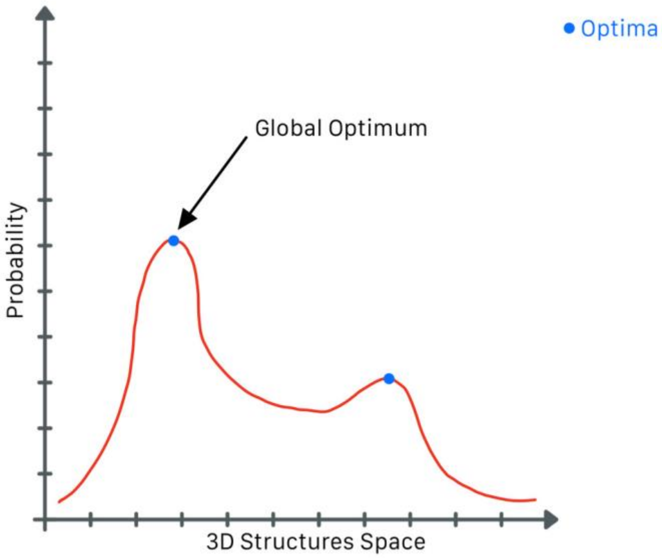


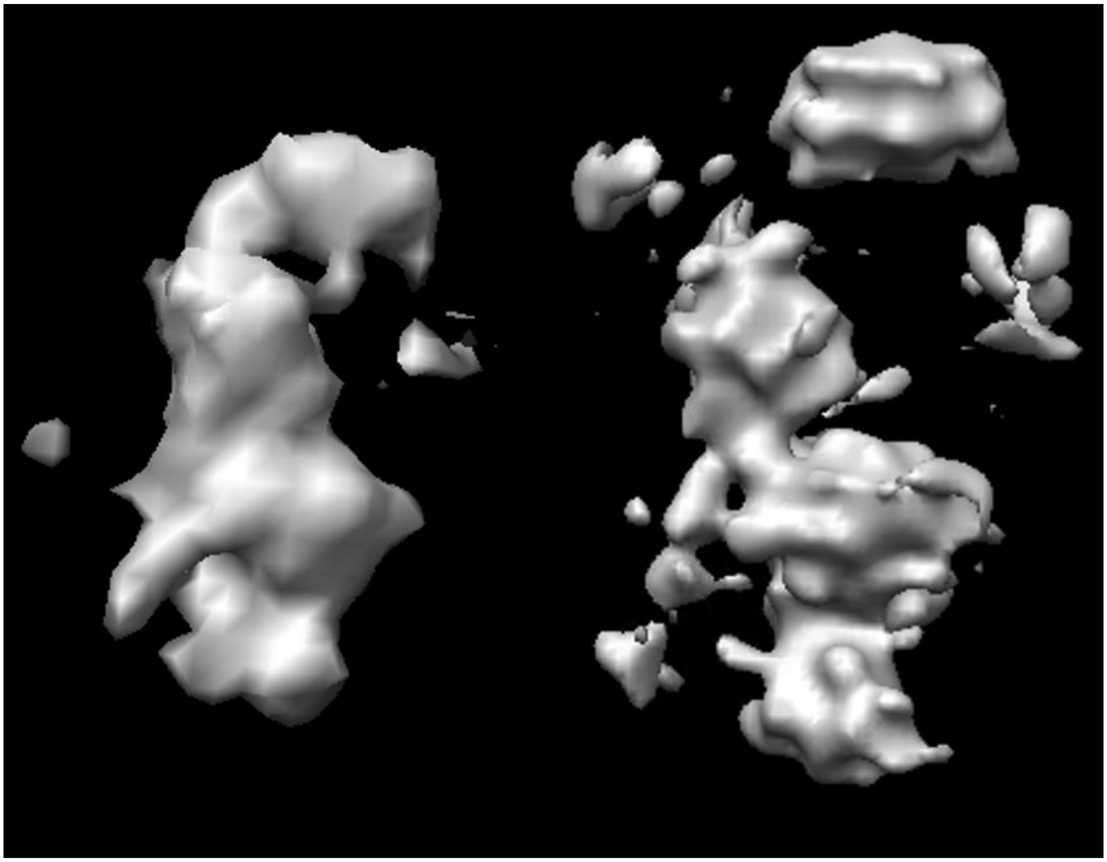
Figure 7 General representation of possible particle confirmations vs their probabilities

Given a space of possible 3D structures for a particle and the probability of generating each structure, the above algorithm stochastically samples the 3D space to arrive at different optima.

Some of these are local optima, and while they may be valid 3D structures for the particle given particular states (different pH values, temperatures, etc.…), they are not representative of the general 3D structure (likely the lowest energy state) of the particle we wish to understand. Multiple iterations are performed using random subsets of the images X_i [20]. This allows the global optimum to be discovered, providing a highly accurate 3D model.

Results

Following the work in CryoSPARC, we imported the produced volumes into UCSF Chimera. The images below are the S-protein models viewable in UCSF Chimera. On the left is the ~9.5 Å resolution model of the S-protein monomer based on the above work. While on the right is the ~8 Å resolution model.



We note that the detail on the right was much more apparent – there was significantly more smoothness in the model on the left. There also appeared to be odd blobs on the top and to the right of the protein in both models.

Discussion

Although we expected a higher-resolution model, we still achieved results depicting clusters of several atoms constituting the S-protein monomer. Indeed, further analysis and data cleaning are needed to obtain the near-atomic resolutions others have published, e.g., Wrapp et al. Nevertheless, these models still illustrate the generalized structure of the S-protein monomer and allow for the visualization the variability of the RBD. Moreover, we hypothesize that the models depict an average of the RBD-up and RBD-down configurations possible in the S-protein, and any other abnormalities in the models likely arise from noise in the data. Again, further work is needed to separate these possible states into different models and to improve the amount of noise. However, our work is still applicable to the various SARS-CoV-2 mutants and other single-particle analyses. Furthermore, models of the variants based on our work can be compared to confirm predicted atomic structures, illuminating the moieties that confer enhanced infectivity to the virus. The resulting differences in phenotype may then be corroborated with the genetic sequences of the mutants.

Conclusions and Further Work

Overall, cryo-EM is a powerful tool for imaging miniscule particles, and it has the power to offer high-resolution (near-atomic distance) images. Additionally, CryoSPARC is a unique software reliant on advanced mathematical methods/algorithms to solve complex optimization problems. Given more time, we will further refine the data we have generated to achieve a higher-resolution reconstruction. Moreover, we will differentiate the 2D classes more to reveal different models for the RBD-up and RBD-down configurations. Following this, we will create several ab-initio classes that permit the elucidation of the two RBD conformations. Also, we would like to compare various mutants to see if there exists a computationally modellable conformational difference. Nevertheless, this research highlights the complexity of cryo-EM imaging and analysis, and it presents a great application of mathematics intersecting with the biological, physical, and biochemical fields.

Acknowledgements

I would like to extend my deepest gratitude to Dr Javier Arsuaga for allowing me to have the opportunity to work with him during Summer 2021. I have enjoyed a thoroughly fascinating experience at the intersection of mathematics, biochemistry, and physics. Dr Arsuaga was an amazing mentor, and this project has helped cement my interest in mathematical modelling and mathematical biology. I would also like to thank Dr Greg Kuperberg and Jennifer Brown for their organizing and careful logistical planning of the REU program. Additionally, I would like to give special thanks to Dr Mariel Vázquez and Tamara Christiani of the Arsuaga-Vázquez Lab. Dr Vázquez has been excellent in leading the research group along with Dr Arsuaga, and Tamara guided me extensively in my first approaches with the CryoSPARC software. Finally, I would like to express gratitude toward Georgina González and Sofia Jakovcevic for their assistance with biostatistics and profile Hidden Markov Modelling, respectively. This work was made possible by funding from the NSF.

References

[1] Barton, M. I., MacGowan, S. A., Kutuzov, M. A., Dushek, O., Barton, G. J., & van der Merwe, P. A. (2021). Effects of common mutations in the SARS-COV-2 spike RBD and its ligand, the human ACE2 receptor on binding affinity and kinetics. *ELife*, *10*. https://doi.org/10.7554/elife.70658

[2] Benton D.J., Wrobel A.G., Xu P., et al. (2020). Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature*, *588*(7837), 327-330. doi:10.1038/s41586-020-2772-0

[3] Cai Y., Zhang J., Xiao T., Peng H., Sterling S.M., Walsh R.M. Jr., et al. (2020). Distinct conformational states of SARS-CoV-2 spike protein. *Science*, *369*(6511), 1586–1592

[4] Cheng Y., Grigorieff N., Penczek P.A., Walz T., (2015). A Primer to Single-Particle Cryo-Electron Microscopy, *Cell*, *161*(3), 438-449. https://doi.org/10.1016/j.cell.2015.03.050

[5] Deng, X., Garcia-Knight, M. A., Khalid, M. M., Servellita, V., et al. (2021). Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. *Cell*, *184*(13), 3426–3437.e8. https://doi.org/10.1016/j.cell.2021.04.025

[6] Duan L, Zheng Q, Zhang H, Niu Y, Lou Y and Wang H (2020). The SARS-CoV-2 Spike Glycoprotein Biosynthesis, Structure, Function, and Antigenicity: Implications for the Design of Spike-Based Vaccine Immunogens. *Front. Immunol. 11*, 576622. doi: 10.3389/fimmu.2020.576622

[7] Goddard T.D., Huang C.C., Meng E.C., Pettersen E.F., Couch G.S., Morris J.H., et al. (2018). UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci*, *27*(1), 14–25

[8] Gui M, et al. (2017). Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell Res*, *27*, 119–129

[9] Hatmal, M. M., Alshaer, W., Al-Hatamleh, M. A., Hatmal, M., Smadi, O., Taha, M. O., Oweida, A. J., Boer, J. C., Mohamud, R., & Plebanski, M. (2020). Comprehensive structural and molecular comparison of spike proteins of SARS-COV-2, SARS-COV and MERS-COV, and their interactions with ACE2. *Cells*, *9*(12), 2638. https://doi.org/10.3390/cells9122638

[10] 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. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*, *181*(2), 271–280.e8. https://doi.org/10.1016/j.cell.2020.02.052

[11] Korber B., et al. (2020). Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*, *182*, 812–827

[12] Laha, S., Chakraborty, J., Das, S., Manna, S. K., Biswas, S., & Chatterjee, R. (2020). Characterizations of SARS-COV-2 mutational profile, Spike protein stability and viral transmission. *Infection, Genetics and Evolution*, *85*, 104445. https://doi.org/10.1016/j.meegid.2020.104445

[13] Lan J, et al. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, *581*, 215–220.

[14] Li, M. Y., Li, L., Zhang, Y., & Wang, X. S. (2020). Expression of the SARS-CoV-2 cell receptor gene ACE2 in a wide variety of human tissues. Infectious diseases of poverty, *9*(1), 45. https://doi.org/10.1186/s40249-020-00662-x

[15] Masre, S. F., Jufri, N. F., Ibrahim, F. W., & Abdul Raub, S. H. (2021). Classical and alternative receptors for SARS-CoV-2 therapeutic strategy. Reviews in medical virology, *31*(5), 1–9. https://doi.org/10.1002/rmv.2207

[16] Morais, I. J., Polveiro, R. C., Souza, G. M., Bortolin, D. I., Sassaki, F. T., & Lima, A. T. (2020). The global population of SARS-COV-2 is composed of six major subtypes. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-74050-8

[17] Penczek, P.A., Image restoration in cryo-electron microscopy. *Methods in enzymology*, *482*, 35-72.

[18] Pettersen E.F., et al. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem*, *25*, 1605–1612.

[19] Punjani, A., & Fleet, D. J. (2021). 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *Journal of Structural Biology*, *213*(2), 107702.

[20] Punjani, A., Rubinstein, J., Fleet, D.J., *et al.* (2017). cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods*, *14*, 290–296. https://doi.org/10.1038/nmeth.4169

[21] Punjani, A., Zhang, H. & Fleet, D.J. (2020). Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nat Methods 17*, 1214–1221. https://doi.org/10.1038/s41592-020-00990-8

[22] Rosenthal, P. B., & Henderson, R. (2003). Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of Molecular Biology*, *333*(4), 721–745. https://doi.org/10.1016/j.jmb.2003.07.013

[23] Scheres, S. H., & Chen, S. (2012). Prevention of overfitting in cryo-EM structure determination. *Nature methods*, *9*(9), 853–854. https://doi.org/10.1038/nmeth.2115

[24] Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., & Li, F. (2020). Cell entry mechanisms of SARS-COV-2. *Proceedings of the National Academy of Sciences*, *117*(21), 11727–11734. https://doi.org/10.1073/pnas.2003138117

[25] Song, W., Gui, M., Wang, X., & Xiang, Y. (2018). Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. *PLOS Pathogens*, *14*(8). https://doi.org/10.1371/journal.ppat.1007236

[26] Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veesler, D., & Bloom, J. D. (2020). Deep mutational scanning of SARS-COV-2 receptor binding domain reveals constraints on folding and Ace2 binding. *Cell*, *182*(5). https://doi.org/10.1016/j.cell.2020.08.012

[27] Walls, A. C., Park, Y. J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*, *181*(2), 281–292.e6. https://doi.org/10.1016/j.cell.2020.02.058

[28] Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C.-L., Abiona, O., Graham, B. S., & McLellan, J. S. (2020). Cryo-EM structure of the 2019-ncov spike in the prefusion conformation. *Science*, *367*(6483), 1260–1263. https://doi.org/10.1101/2020.02.11.944462

[29] Xing Zhu et al. (2021). Cryo-electron microscopy structures of the N501Y SARS-CoV-2 spike protein in complex with ACE2 and 2 potent neutralizing antibodies, *PLOS Biology*. DOI: 10.1371/journal.pbio.3001237

[30] Yang, T. J., Yu, P. Y., Chang, Y. C., Liang, K. H., Tso, H. C., Ho, M. R., Chen, W. Y., Lin, H. T., Wu, H. C., & Hsu, S. D. (2021). Effect of SARS-CoV-2 B.1.1.7 mutations on spike protein structure and function. *Nature structural & molecular biology*, *28*(9), 731–739. https://doi.org/10.1038/s41594-021-00652-z

[31] Yip, K.M., Fischer, N., Paknia, E. et al. (2020). Atomic-resolution protein structure determination by cryo-EM. *Nature 587*, 157–161. https://doi.org/10.1038/s41586-020-2833-4

[32] Yuan, Y., Cao, D., Zhang, Y., Ma, J., Qi, J., Wang, Q., Lu, G., Wu, Y., Yan, J., Shi, Y., Zhang, X., & Gao, G. F. (2017). Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nature communications*, *8*, 15092. https://doi.org/10.1038/ncomms15092

[33] Zhang, K., Pintilie, G. D., Li, S., Schmid, M. F., & Chiu, W. (2020). Resolving individual atoms of protein complex by cryo-electron microscopy. *Cell research*, *30*(12), 1136–1139. https://doi.org/10.1038/s41422-020-00432-2

[34] Zhang, L., Jackson, C. B., Mou, H., Ojha, A., Peng, H., Quinlan, B. D., Rangarajan, E. S., Pan, A., Vanderheiden, A., Suthar, M. S., Li, W., Izard, T., Rader, C., Farzan, M., & Choe, H. (2020). SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nature communications*, *11*(1), 6013. https://doi.org/10.1038/s41467-020-19808-4