Configurations of the SARS-CoV-2 nucleocapsid protein

A molecular dynamics approach to understanding its role in genome packaging

Lars van den Biggelaar
4654323

A thesis presented for the degree of
Bachelor of Science

dr. Ireth García Aguilar
dr. Marianne Bauer
dr. Martin Depken
Bionanoscience
Delft University of Technology
November 3rd, 2023

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a respiratory virus capable of infecting humans and causing mild to severe complications during and after infection, which collectively comprise coronavirus disease 2019 (COVID-19). After an initial local epidemic in late 2019, the virus quickly spread on a global scale and was declared a pandemic shortly after. Many research efforts were devoted to SARS-CoV-2 leading to the quick discovery and communication of crucial data, resulting in improved countermeasures and accelerated development and deployment of new vaccines. These achievements notwithstanding, a number of unknowns of the working mechanisms of the virus and its replication inside of a host still remain. The exact function and structure of the nucleocapsid (N) protein have not been clearly established despite its recognized importance throughout the virus' life cycle. N is involved in both the replication and translation of viral RNA, although its most prominent function is the formation of ribonucleoprotein (RNP) for the stable and compact storage of genomic RNA inside virions. Here we present a course-grained model of N for molecular dynamics simulations. Employing simple rules, the dimerisation and RNA-binding capacities of N have been emulated. A range of parameters giving rise to structures resembling those found *in vivo* by balancing attractive and repulsive properties are analysed and described. Lastly, by implementing an RNA substitute, conformations have been found which could give an indication of the structure of the RNP found inside SARS-CoV-2 virions. The model described in this work can be used as a basis for more extensive simulations incorporating further elements to gain a better understanding of the behaviour of this and similar viruses, contributing to the prevention of future large-scale disease outbreaks.

Contents

Chapter 1

Introduction

Viruses are the smallest replicating biological entities, but they make up for their size with sheer numbers. At an estimated 10^{31} viral particles they outnumber the total number of cells on Earth by at least an order of magnitude, making them by far the most abundant biological agent (Hendrix et al. 1999). The debate on whether or not viruses can be considered to be alive is ongoing, and may never be settled entirely (Koonin and Starokadomskyy 2016), as they lack several critical components required for their own replication. Instead, a virus infects a host cell and uses its replication machinery to produce more virions. Almost every (other) known form of life is theorised to have multiple viral counterparts capable of infecting it (Koonin, Dolja, et al. 2020), which has likely played a major role in evolution during every age of life (Forterre and Prangishvili 2009). However, they can also have a significant impact on their hosts on shorter timescales, which becomes especially noticeable if a virus can infect and spread between humans.

Upper respiratory tract viral infections, often simply referred to as "common colds", are the most common diseases affecting humans (Johnston and Holgate 1996). Over 200 different viral types have been identified as the cause of such an infection, though the type of infective agent and its properties seem less important for the appearance and severity of symptoms than aspects of the host such as age and physical health (Kilbourne 1987). Despite the term "flu" often being used interchangeably with "common cold", influenzaviruses only account for 5-15% of infections, with rhinoviruses and coronaviruses occurring more often at 30-50% and 10-15% of infections respectively.

When human coronaviruses were first described and identified as one of the causes of seasonal respiratory infections, the resulting illness usually only saw mild symptoms in healthy individuals (Bradburne, Bynoe, and D. A. Tyrrell 1967; D. A. J. Tyrrell and Bynoe 1965). However, this was not the case when a novel virus caused a limited outbreak in 2002 and 2003. This severe acute respiratory syndrome coronavirus (SARS-CoV) caused more serious illness in the majority of those infected, with 1 in 3 afflicted patients requiring intensive care (Chan 2003; Ksiazek et al. 2003; Peiris et al. 2003). The number of infections was kept limited at 8098, but the disease still spread to 29 different countries and resulted in 774 deaths (Christian et al. 2004; World Health Organization 2015). Although the majority of infections were caused by human-to-human transmission, the first cases in humans are suspected to have originated from interactions with infected raccoon dogs and palm civets at a live-animal market in Guangdong, China (Guan et al. 2003; WHO 2004). The discovery of SARS-like viruses with high sequence similarity in bats indicates a potential source of the disease, as well as a reservoir and source for potential future outbreaks (Ge et al. 2013).

In 2012, the disease vector of an epidemic of severe respiratory illness in Saudi Arabia was identified as a new coronavirus variant, later named Middle East respiratory syndrome coronavirus (MERS-CoV) (Groot et al. 2013; Zaki et al. 2012). Since then there have been 2600 confirmed cases in 27 countries, leading to 935 deaths (WHO 2022a). The significantly increased fatality rate compared to that of SARS was deemed due to the higher prevalence of comorbidities in those suffering from MERS, as symptoms were otherwise highly similar between the two diseases (Assiri, Al-Tawfiq, et al. 2013). The major difference between the two illnesses lay in the mode of transmission: in the case of MERS-CoV, only a limited number of infections were reported to occur among healthcare workers and contacts of patients (Assiri, McGeer, et al. 2013; Drosten et al. 2014), with the majority of afflictions caused by interactions with dromedary camels and their products (Chantal BEM Reusken et al. 2013; WHO 2022a). Related

viruses have been circulating in dromedary populations for at least 30 years before the first case in a human and persist to the current day, meaning there is an ongoing risk for new infections (Corman et al. 2014; Meyer et al. 2014; Müller et al. 2014; Chantal B.E.M. Reusken et al. 2014).

These outbreaks showed how extensive contact between humans and live animals can be hazardous to public health, but without major cultural and societal reforms it seemed to be only a matter of time before a new epidemic would appear. In early December 2019, these fears were confirmed when multiple people contracted pneumonia with a yet-unknown cause in Wuhan, China (C. Huang et al. 2020; Hui et al. 2020; Lu, Stratton, and Tang 2020). Initially referred to as 2019-new coronavirus (2019-nCoV) and 2019-nCoV disease, its genetic and symptomatic resemblance to the earlier strain meant the International Committee on Taxonomy of Viruses (ICTV) would officially name it as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Alexander E. Gorbalenya et al. 2020), with the corresponding disease referred to as coronavirus disease 2019 (COVID-19) by the World Health Organization (WHO) (WHO 2020a). The majority of patients had a fever or cough with a third suffering from shortness of breath (N. Chen et al. 2020). Once more reflecting the earlier related outbreaks, 47 of the first 99 identified patients worked at or lived near the local Huanan seafood market, which also hosted live animal sales, with 2 more patients having recently visited. SARS-CoV-2-related coronaviruses were identified in pangolins and bats (Lam et al. 2020; K. Xiao et al. 2020), though their absence from the market means that other species held there such as civet cats or raccoon dogs could have acted as intermediate hosts for zoonotic transmissions (X. Xiao et al. 2021). Measures taken by local authorities based on recommendations initially appeared to lower the reproduction number, suggesting that the number of infections for this outbreak could be limited as with other strains (Yan et al. 2020). However, millions of inhabitants of Wuhan had left the city before a lockdown was reinforced, including possibly hundreds already infected (Zhong, Guo, and T. Chen 2020). The disease started appearing in other regions of China, and the first cases had been confirmed in Europe and the United States by the end of January 2020 (Arora et al. 2021). The rapid increase in infections despite extensive measures taken by many governments led to the WHO declaring COVID-19 a pandemic on March 11^{th} (WHO 2020b). The number of confirmed infections reached 10 million by June and had further risen to 100 million a year after the first international cases in January 2021 (WHO 2022b). At the time of writing, there have been over 771 million infections reported to the WHO resulting in close to 7 million deaths. However, as the number of COVID-related deaths is likely significantly underreported, the disease has likely actually resulted in between 18.0 and 33.0 million deaths (95% confidence interval) (Economist and Solstad 2021). Outside of the major impact of SARS-CoV-2 on societies worldwide, this would make it the deadliest pandemic caused by a respiratory virus since the Spanish Flu (H1N1 influenza A) pandemic between 1918 and 1920 (Feehan and Apostolopoulos 2021).

Chapter 2

Biological background

It is often said that there are no hard rules in biology, as one can almost always find exceptions when looking closely enough. Still, there seems to be one fundamental exception to this rule of exceptions: the Central Dogma. It states that genetic information may be transferred from nucleic acid to nucleic acid and from nucleic acid to protein, but never in reverse or between proteins (Crick 1958, 1970). In literature, it is more generally applied to describe the transfer of genetic information in all organisms across the tree of life. Information is stored as deoxyribonucleic acid (DNA), which can either be replicated to create more copies of the cell's genome or transcribed into an intermediate or active form as ribonucleic acid (RNA). Specifically, messenger RNA (mRNA) can be translated to create peptides and proteins that give the cell its properties and functions. Although viruses do not violate the rules of the Central Dogma, they also do not fit with these classical descriptions of life and information transfer.

Firstly, viruses have fundamentally different structures than cells. They still carry genetic material, but unlike in cells this can be stored as DNA or RNA depending on the species. The vulnerable nucleic acids are protected by the capsid, a protein "shell" encoded in the viral genome. The capsid consists of one or a few different protein types often resulting in a highly regular structure, often icosahedral in shape. In most viruses, the capsid is surrounded by an envelope, a lipid membrane originating from membranes of the host cell the virus was assembled inside of. The full particle is around 100 nm in diameter for species such as influenza viruses and coronaviruses, making them 100 to 1000 times smaller than their target cells (Louten 2016; Weiss and Navas-Martin 2005).

Secondly, most viruses do not follow the DNA $\rightarrow$ mRNA $\rightarrow$ protein process as seen in cells. There are 7 known approaches employed by different viruses to form mRNA from their genome, corresponding with the 7 groups of the Baltimore classification (Baltimore 1971). A virus species is placed in a group based on whether its genome consists of DNA or RNA, whether its genome is single-stranded (ss) or double-stranded (ds), whether the sequence directly encodes proteins (positive sense, +) or is complementary to that sequence (negative sense, -), and whether the nucleic acid not forming the genome appears as an intermediate during the virus' replication cycle. This means that for certain groups of viruses RNA is replicated instead of DNA, or DNA is reverse transcribed from RNA, or both. As these processes occur rarely or never inside of their target hosts and thus lack the machinery to fulfil them, these viruses need to encode the required enzymes in their genome.

2.1 Coronaviruses

Parallel to the Baltimore classification, the ICTV maintains a virus taxonomy based on genetic similarity resembling the systems used for the phylogenetic organisation of organisms (Lefkowitz et al. 2017). The classification of coronaviruses with taxonomic names and defining features is as follows:

- Realm: *Riboviria* - RNA-dependent RNA polymerase (RdRp) or reverse transcriptase
  - Kingdom: *Orthornavirae* - RNA genome, RdRp
    - Phylum: *Pisuviricota* - dsRNA or +ssRNA genome, infect eukaryotes
      - Class: *Pisoniviricetes* - +ssRNA genome
        - Order: *Nidovirales* - Envelope, "nested" mRNAs
          - Family: *Coronaviridae* - Infect amphibians, birds, mammals

As all viruses in the order *Nidovirales*, coronaviruses belong to Baltimore group IV comprising positive-sense single-stranded RNA viruses (+ssRNA), meaning the sequence of their genome directly corresponds with the mRNA encoding the viral proteins. All coronaviruses belong to the subfamily *Orthocoronavirinae* and are further divided into four genera: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus*, and *Deltacoronavirus*. Species capable of infecting humans, other mammals, or both are found in the first two genera, with SARS-CoV, MERS-CoV, and SARS-CoV-2 all classified under *Betacoronavirus*.

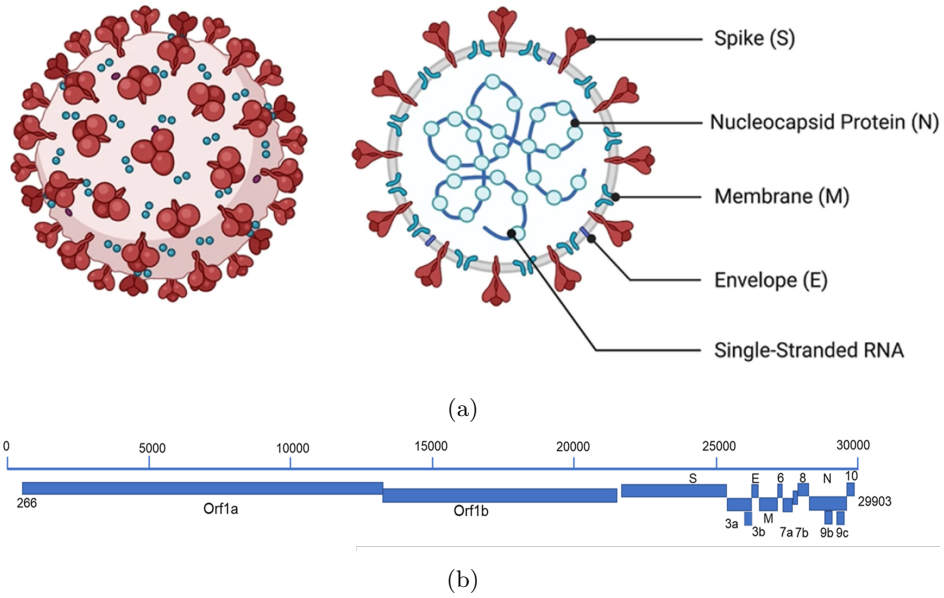
2.2 SARS-CoV-2

SARS-CoV-2 particles are spherical or ellipsoid in shape, generally up to 100 nm in diameter (Yao et al. 2020). Its single-segment genome of 29,903 nucleotides in length is among the largest known for RNA viruses (Weiss and Navas-Martin 2005; F. Wu et al. 2020). The sequence is divided into multiple open reading frames (ORFs), the majority of which encode non-structural or accessory proteins such as those involved in replication (see Figure 2.1b) (Lowery, Sariol, and Perlman 2021; Zhou et al. 2020). Furthermore, four structural proteins are present on and inside SARS-CoV-2 virions (see Figure 2.1a). The spike protein (S) is the largest and binds to the angiotensin-converting enzyme 2 (ACE2) receptor on host cells to facilitate entry of the virus to start an infection cycle (Y. Huang et al. 2020). The envelope protein (E) is an ion channel important for virulence and pathogenesis, and is furthermore involved in the assembly of new viral particles (Nieto-Torres et al. 2014). The membrane protein (M) acts as a scaffold for other structural proteins during the formation of virions, after which it becomes important for avoiding the host's immune system (Fu et al. 2020; Gordon et al. 2020). Finally, the nucleocapsid protein (N) is also involved in assembly, but unlike the other structural proteins is not present on the outer membrane of virions. Instead, it dimerises to bind viral mRNA to aid replication and package genomic RNA inside the viral lumen as ribonucleoprotein (RNP) (Zeng et al. 2020). However, it is unclear how this RNP would be structured and how the protein is able to pack the full 30 kb RNA into the viral lumen. The structure of SARS-CoV-2 RNP seems to be different from that of the highly similar SARS-CoV, meaning models fitting the latter cannot be directly applied to the former (C.-Y. Chen et al. 2007; Yao et al. 2020).

2.3 Nucleocapsid protein

The SARS-CoV-2 N protein is 419 amino acids in length and consists of 5 distinct domains, described here from N- to C-terminus (Bai et al. 2021; Chang et al. 2005; Ye et al. 2020). The N-terminal domain, or N-arm, is highly disordered. Its function is not fully known, but its location suggests it may interact with viral RNA. The ordered and highly conserved RNA-binding domain (RBD) has a positively charged pocket to bind viral RNA either inside of a virion or during replication in a host cell. The linker domain is highly disordered, making it a flexible connection between the two main domains of N which may also be important for RNA-binding and phase separation. It connects to the ordered dimerisation domain (DD), which is able to interact with the same domain on another monomer to create an N dimer. This is achieved by the alignment of a β-hairpin from both proteins to form a single β-sheet, which in addition to hydrophobic interactions with neighbouring α-helices creates a highly stable structure (Peng et al. 2020) Lastly, the C-terminal domain, or C-tail, is similar to the N-arm in its disorder and unclear function. However, some evidence suggests that the C-tail may be important for the formation of RNP itself (Ribeiro-Filho et al. 2022).

Although the RBD and DD have been studied as separate domains, the high degree of disorder in the full protein makes it difficult to study its structure. Apart from the terminal domains, the flexible



(a)

(b)

Figure 2.1: **(a)** The structure of SARS-CoV-2 virions with structural proteins and genomic RNA visible. Adapted from (Jamison et al. 2022). **(b)** Genes in the SARS-CoV-2 genome. Numbers represent the number of nucleotides. Adapted from (C.-r. Wu et al. 2022).

linker allows the RBD and DD to move and rotate freely in relation to each other, making it unclear if there is a single preferred conformation when interacting with RNA and if so what that conformation would be. One possible structure for the RNP would contain 5 N dimers forming a "reverse G-shape" 15 nm wide and 16 nm high (see Figure 2.2) (Yao et al. 2020). A full virion should contain around 30 to 35 of these structures, the majority of which arranged close to the particle's membrane. Although the volume for the shape was experimentally resolved, the exact position, rotation, or even amount of N proteins could not be determined. The researchers did find that a specific arrangement of the ordered domains resulted in a positively charged groove winding around the protein, giving a possible local arrangement for N and RNA.

2.4 Outline

The important role of the SARS-CoV-2 nucleocapsid protein in replication and genome packaging could make it a valuable target for future drug and vaccine development. However, to be able to develop medication capable of effectively targeting and neutralizing the protein, it needs to be better understood first. Although research on SARS-CoV-2 has been massively accelerated over the past few years because of the need to combat this global threat, experimental approaches alone do not seem sufficient to solve the questions that still remain about the N protein. *In silico* methods could prove to be a valuable tool to be used, not instead of, but in conjunction with established techniques to further our collective understanding of SARS-CoV-2. With this in mind, our goal for this project was to develop a computational model for the SARS-CoV-2 nucleocapsid protein to further our understanding of its role in viral genome packaging.

As the structures of full N monomers and dimers are hard to resolve due to their inherent disorder, we first predicted possible 3D structures of N based on currently available partial structures. From there we designed a coarse-grained model of an N monomer to approximate its properties. This model was then used in molecular dynamics simulations, from which we were able to narrow down values for parameters to be applied to the model. Finally, we combined our model of N with an RNA analogue to take the first steps towards a full theoretical system of SARS-CoV-2 genome packaging.

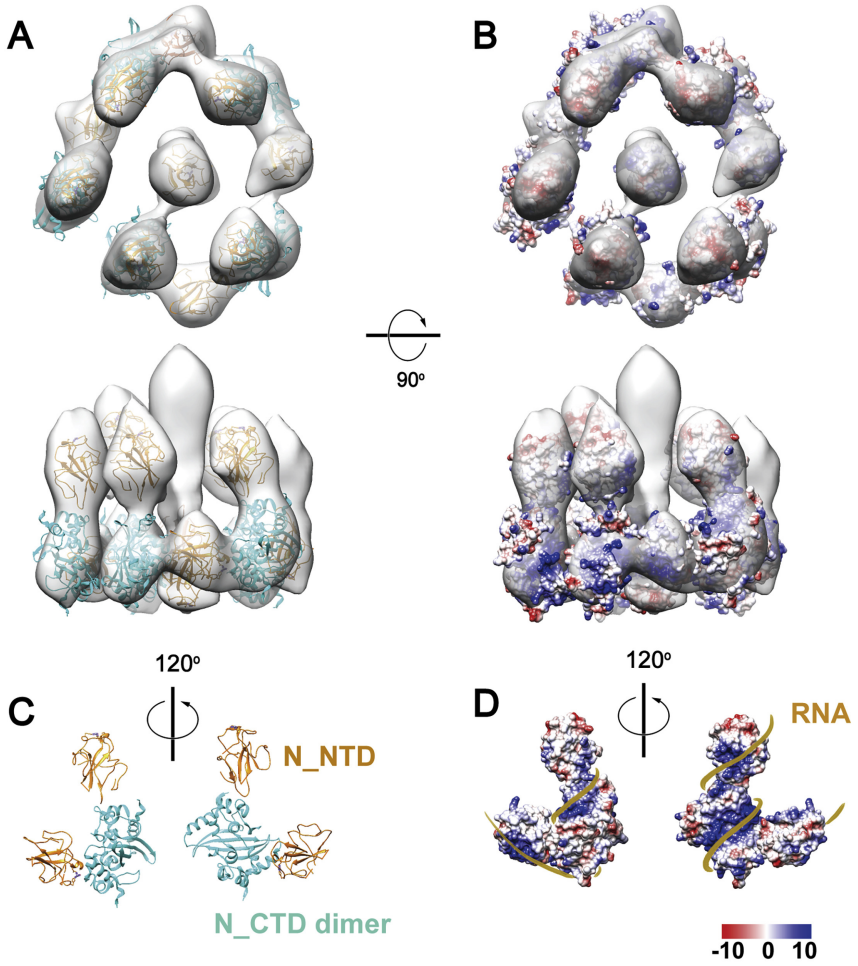


Figure 2.2: The reverse G-shape as seen in SARS-CoV-2 virions. White: experimentally resolved volume. **(a,b)** Possible arrangement of 5 N dimers inside the volume, showing either a cartoon representation or surface charge. **(c)** Relative orientations of the RBDs (NTD, yellow) and DDs (CTD, cyan). **(d)** Surface charges of the dimer configuration, with a possible path of the RNA winding around the protein shown. Image adapted from (Yao et al. 2020).

Chapter 3

Results

3.1 Nucleocapsid structure

Due to the inherent disorder of the N-terminal, C-terminal, and central linker domain of the nucleocapsid protein, it is difficult to experimentally resolve its full tertiary structure. On the online Protein Data Bank (PDB) (Berman et al. 2000) a single model is available of a full N monomer (PDB ID 8FD5 (Casasanta et al. 2022), which has been resolved through a combination of electron microscopy and molecular modelling. Other available structures are the dimerisation domain (DD) in dimer configuration (PDB ID 6YUN (Zinzula et al. 2021), and the RNA-binding domain (RBD) without (PDB ID 6YI3) and with RNA (PDB ID 7ACT) (Dinesh et al. 2020) As no resolved structure is available of full-length N in dimer configuration, a prediction of this structure was instead made using ColabFold (Mirdita et al. 2022) (for settings see Section 3.7). ColabFold is a cloud-based alternative for the structure prediction model AlphaFold (Jumper et al. 2021) which can also resolve multimers (Evans et al. 2021) Using NCBI Reference Sequence: YP_009724397.2 as the sequence for N, multiple possible structures for an N-dimer were predicted containing two monomers referred to as chains A and B. Each dimer was ranked by a template modelling score reflecting their similarity to structures in the PDB (Zhang 2005) The highest-ranking structure, seen in Figure 3.1, was then compared to the experimentally resolved structures described above by aligning the corresponding domains in PyMOL. Table 3.1 shows the resulting root-mean-square deviations after outlier rejection.

Table 3.1: RMSD between the predicted structure of N and multiple experimentally resolved structures. Comparison of a domain to one of either monomer chain is denoted by the corresponding letter after the RMSD value.

PDB ID (domain)	**RMSD [Å] (chain)**
6YUN (DD)	0.575
6YI3 (RBD)	0.833 (A)
	0.829 (B)
7ACT (RBD + RNA)	1.057 (A)
	1.073 (B)
8FD5 (full N)	3.068 (A)
	3.049 (B)

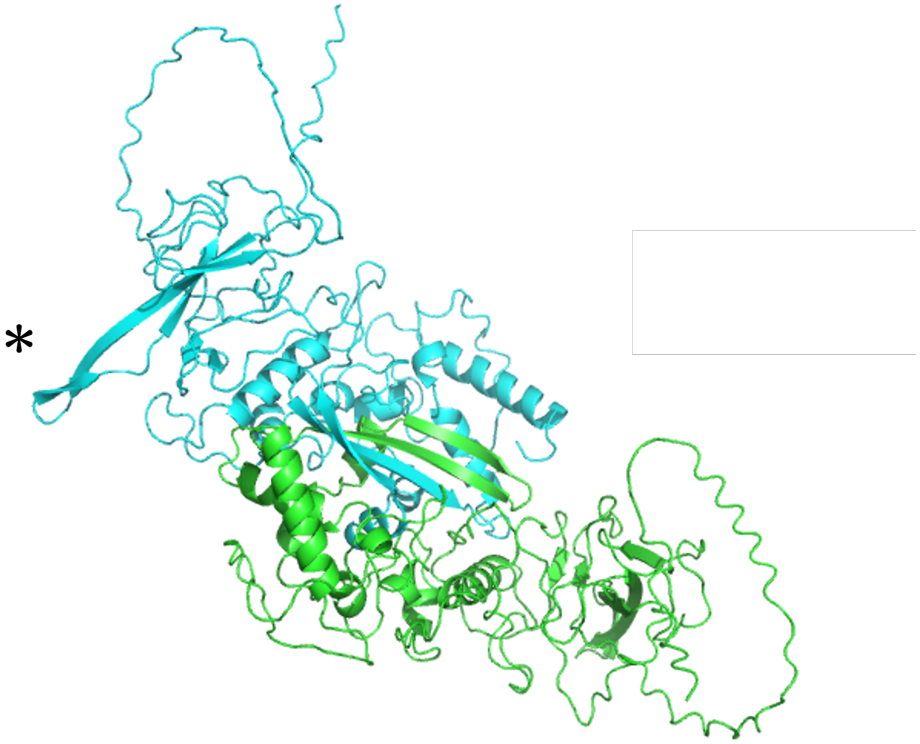


Figure 3.1: Structure of the SARS-CoV-2 nucleocapsid dimer as predicted using ColabFold. Secondary structures are depicted with cartoon helices and ribbons. The RNA-binding arm of the RBD (*) and combined β-sheet of the two dimerisation domains (centre) are visible.

The predicted and resolved structures are the most similar for the dimerisation domains, likely because of the highly stable nature of the conjugated domains minimizing strand movement. The difference between models 6YI3 and 7ACT comes from the conformational changes of the RBD due to the introduction of RNA in the latter model; as the predicted structure reflects the relaxed conformation, it is more similar to 6YI3. Finally, the increased dissimilarity when compared to 8FD5 is mostly due to the flexible linker domain, which means that even if the ordered domains are highly similar the RMSD will still increase due to their relative position having been shifted by the flexible linker.

3.2 Coarse-grained model

As described earlier in Section 2.3, experiments have been unable to determine the exact structure and formation of the SARS-CoV-2 ribonucleoprotein. We here describe an alternative approach based on molecular dynamics simulations on a simplified model of the N protein that could aid in shedding light on this conformation and the underlying processes.

The simulations are run using HOOMD-blue, a package for the Python programming language for molecular dynamics (MD) simulations (Anderson, Glaser, and Glotzer 2020) Instead of SI units, HOOMD-blue uses an internal system where all units are derived from the 'base units' [energy], [length], and [mass]. As such, a physical property may be considered to be expressed in arbitrary units (a.u.) in simulations if not otherwise defined. By employing a simplified model of N and a set of physical forces and constraints, the behaviour of the protein as observed in *in vitro* experiments can be emulated and further analysed. This requires a so-called coarse-grained model of the protein, in which spheres of corresponding size and charge represent the domains.

The coarse-grained model of N used throughout this project is partly based on the one described in (Li and Zandi 2022) This model assumes that the protein consists of three major domains: RBD (RNA-binding domain, representing residues 1 through 174 comprising the N-terminal intrinsically disordered

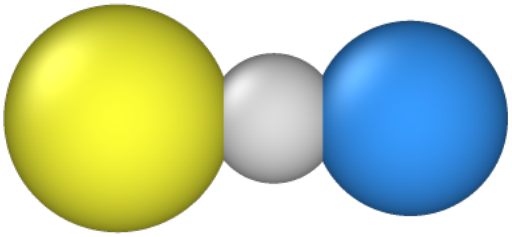


Figure 3.2: Graphical view of the nucleocapsid protein model. Yellow: RNA-binding domain (RBD), 35 Å diameter, +10 charge. White: linker domain (LINK), 20 Å diameter, +16 charge. Blue: dimerisation domain (DD), 30 Å diameter, ±0 charge. Image created in OVITO.

region and subsequent ordered region), LINK (linker domain, representing residues 175 through 246 comprising the disordered linker domain), and DD (dimerisation domain, representing residues 247 through 364 comprising the C-terminal ordered region). The C-terminal disordered tail is considered to be possibly involved in higher-order RNP formation (Ribeiro-Filho et al. 2022), but as the aim of this work lies mainly in N dimer formation and low-order interactions with RNA this region is not explicitly defined in our model.

Using the Measurement tool in PyMOL, we estimated the diameter for a spherical approximation of each of the mentioned domains; RBD being 35 Å, LINK being 20 Å, and DD being 30 Å. All three domains are initially arranged in a line with a distance of 45 Å between RBD and DD, with LINK placed roughly halfway between the other domains at 24 Å from RBD. Note that due to its highly disordered nature the linker domain could stretch to over 200 Å in length, but all models available in the Protein Data Bank show the domain in a more relaxed state with the centres of RBD and DD at roughly 40 to 50 Å from each other. All domains have the same mass of 1 [mass]. Neighbouring domains are connected by a harmonic spring with a rest length equal to the initial distance between domains ($r_0 = 24$ Å between RBD and LINK, $r_0 = 21$ Å between LINK and DD). Both bonds have a spring constant $k = 100$, allowing for some extension and contraction to represent the flexibility of the linker domain. Domains are not restricted by bond angle and may thus move freely in relation to each other. Note that bonds between particles are not considered physical entities in simulations and may thus freely cross and overlap each other and particles if this does not violate other constraints of the system.

The other pre-defined property for each particle is their electrical charge. To determine the charges to be used in the model, the sequence of each domain was analysed for basic and acidic amino acids, respectively providing a positive and negative charge to the domain. With that, the charges were found to be +4 for N-arm, +6 for NTD, +7 for LINK, and +9 for CTD. However, as also noted in (Li and Zandi 2022), most positively-charged residues of CTD are located at its N-terminal side, so in our model they will be grouped with the linker domain. The charges used in the simulations are therefore +10 for RBD, +16 for LINK, and 0 for DD.

3.3 Simulation interactions

Particles in the system move and behave in accordance with the Langevin equations of motion, which give an expression for the total force on a given particle as:

$$m\frac{d\vec{v}}{dt} = \vec{F}_{\mathrm{P}} - \gamma \cdot \vec{v} + \vec{F}_{\mathrm{R}} \tag{3.1}$$

where m is the particle's mass, $\vec{v}$ is the particle's velocity, $\vec{F}_{\mathrm{P}}$ is the force on the particle from all potentials, $\gamma = 50$ is the drag coefficient, and $\vec{F}_{\mathrm{R}}$ is a uniform random force. The latter represents the effect of Brownian motion on all particles and can be further defined according to:

$$\langle \vec{F}_{\mathrm{R}} \rangle = 0 \tag{3.2}$$

$$\langle |\vec{F}_{\mathrm{R}}| \rangle = \frac{6kT\gamma}{\delta t} \tag{3.3}$$

where $kT = 1.5$ [energy] is the temperature of the system and δt denotes that the force at any given time step is uncorrelated from that at every other time.

$\vec{F}_{\mathrm{P}}$ is the sum of forces on a particle from all potentials acting on it. For any given force between particles, a negative value implies that the force acts attractively whereas positive values result in repulsion. Particles experience a Hooke force from the springs connecting them to adjacent domains based on their relative distance. Additionally, the Lennard-Jones (LJ) and Coulomb (C) pair potentials are defined for every pair of particles in the system, giving rise to a number of attractive and repulsive interactions in and between N monomers. The Lennard-Jones pair potential and corresponding force are defined as:

$$U_{LJ}(r) = 4\varepsilon \left(\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right) \tag{3.4}$$

$$\vec{F}_{LJ} = \nabla U_{LJ} = 24\varepsilon \left(\frac{\sigma^6}{r^7} - 2\frac{\sigma^{12}}{r^{13}} \right) \tag{3.5}$$

The Lennard-Jones potential is negative and near-zero at long distances and decreases further as r decreases. The potential has a minimum, or well, of $-\varepsilon$ at $r = 2^{\frac{1}{6}}\sigma$, is zero at $r = \sigma$, and increases to ∞ at lower values for r (see Figure 3.3). In the simulations, we implement a cut-off radius where $U_{LJ} = 0$ if $r > 250$ Å as the influence of the potential is considered negligible at longer ranges. σ roughly represents the size of both particles in the pair potential, and as such differs depending on the particle types involved (see Table 3.2). All values of σ have been multiplied with a correction factor $2^{-\frac{1}{6}}$ to effectively shift the potential minimum from $r = 2^{\frac{1}{6}}\sigma$ to $r = \sigma$, meaning two particles now experience no force from this potential when their volumes touch instead of at slightly larger distances. This was deemed more representative of situations where domains of two monomers would have physical contact, such as in interactions between dimerisation domains.

Table 3.2: Particle size σ for the Lennard-Jones pair potential per particle type pair.

Particle types in pair	$\boldsymbol{\sigma}$ **(Å)**
RBD, RBD	$3.5 \cdot 2^{-\frac{1}{6}}$
RBD, LINK	$2.4 \cdot 2^{-\frac{1}{6}}$
RBD, DD	$3.25 \cdot 2^{-\frac{1}{6}}$
LINK, LINK	$2.0 \cdot 2^{-\frac{1}{6}}$
LINK, DD	$2.1 \cdot 2^{-\frac{1}{6}}$
DD, DD	$3.0 \cdot 2^{-\frac{1}{6}}$

In vivo, two monomers dimerise by a β-hairpin from either dimerisation domain aligning to form a single β-sheet (see Figure 3.1), which in addition to hydrophobic interactions with neighbouring α-helices creates a highly stable structure (Peng et al. 2020Zinzula et al. 2021) In our simulations, we set ε between DD domains such that the Lennard-Jones potential allows the domains to be attracted when in close vicinity, and remain stably close together once the domains have reached each other's potential well. All other type pairs have a fixed $\varepsilon = 0.1 < kT$ to characterise the van der Waals force. This means particles experience a weak attraction to other particles at short distances, which is of lesser magnitude than $\vec{F}_{\mathrm{R}}$, and strong repulsion at shorter distances to prevent overlap.

In addition to the Lennard-Jones potential, a Coulomb pair potential is also calculated for all charged particles in the system, which here are the LINK and RBD domains. The Coulomb potential and corresponding force between particles a and b are given by the following formulas:

$$U_C(r) = \alpha \cdot \frac{q_a q_b}{r} \tag{3.6}$$

$$\vec{F}_C = \nabla U_C = -\alpha \cdot \frac{q_a q_b}{r^2} \tag{3.7}$$

where α is a scaling factor representing the Coulomb constant, q_a and q_b are the charges of particles a and b respectively, and r is the distance between said particles. The magnitude of the potential is near-zero at long ranges and increases exponentially as r decreases. Similar to with LJ, we set a cut-off radius such that $U_C = 0$ if $r > 150$ Å. As a negative potential implies an attractive interaction, the Coulomb potential will cause particles to repel if their charges have the same sign and attract if the signs of their charges differ. Depending on the magnitude of the scaling factor, the Coulomb pair potential is usually stronger at medium to long distances than LJ. When the distance between particles closes, the potential well and subsequent strong repulsion of LJ become more significant.

If only the Lennard-Jones potential is applied to particles in the system, monomers will randomly clump in large groups around their dimerisation domains. This behaviour does not reflect how N proteins form dimers *in vivo*, and as such the attractive interactions need to be counteracted by repulsion. By adding the Coulomb potential, which is solely repulsive given that all LINK and RBD have positive charges and DD are neutral, monomers will tend to form smaller clusters or completely repel each other depending on the ratio between attractive and repulsive forces.

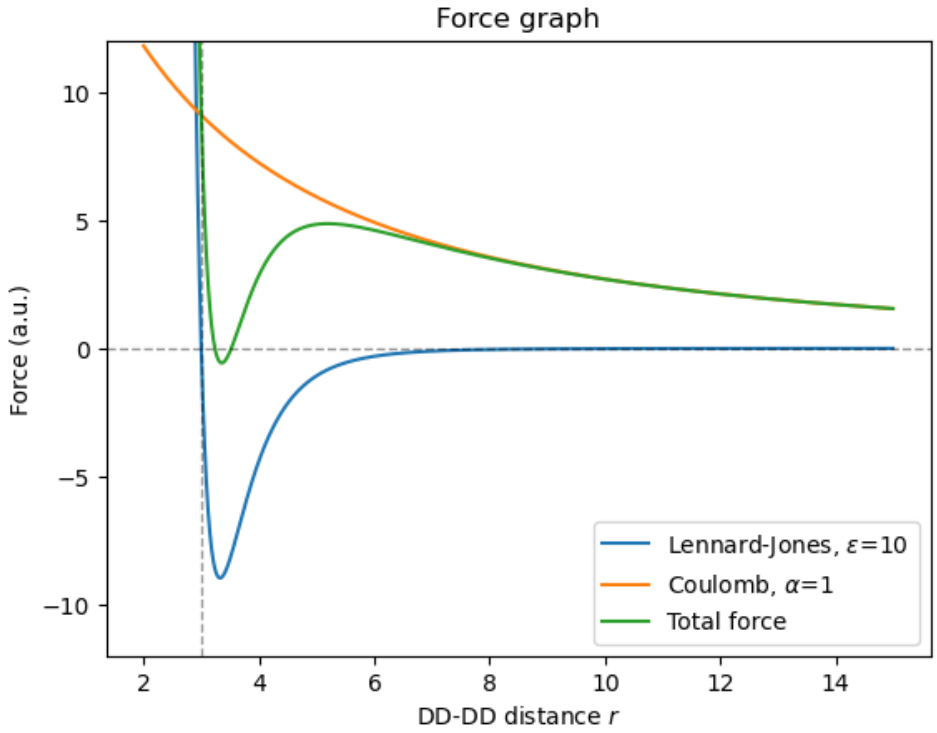


Figure 3.3: Graph showing the magnitude of the Lennard-Jones (Equation 3.5), Coulomb (Equation 3.7), and summed forces for a given distance between the dimerisation domains of two monomers. Calculations for the forces assume that the monomers are antiparallel and that the domains of both monomers are positioned on a single axis at the relative rest lengths as defined in Section 3.2.

3.4 Preliminary simulations of N

The aim of the simulations containing varying numbers and arrangements of N monomers was to find combinations of ε and α that would allow for the formation of dimers, while minimizing the amount of larger oligomers appearing throughout runs. To find an initial set of these combinations, several preliminary tests were run on a limited number of N proteins with different starting arrangements and parameter values. From these first simulations, it became clear that a ratio of roughly 10:1 between ε

for the dimerisation domains and α for all interactions respectively gave the best results. Therefore, subsequent simulations with fixed parameters were run with $\varepsilon = 40$ and $\alpha = 4$.

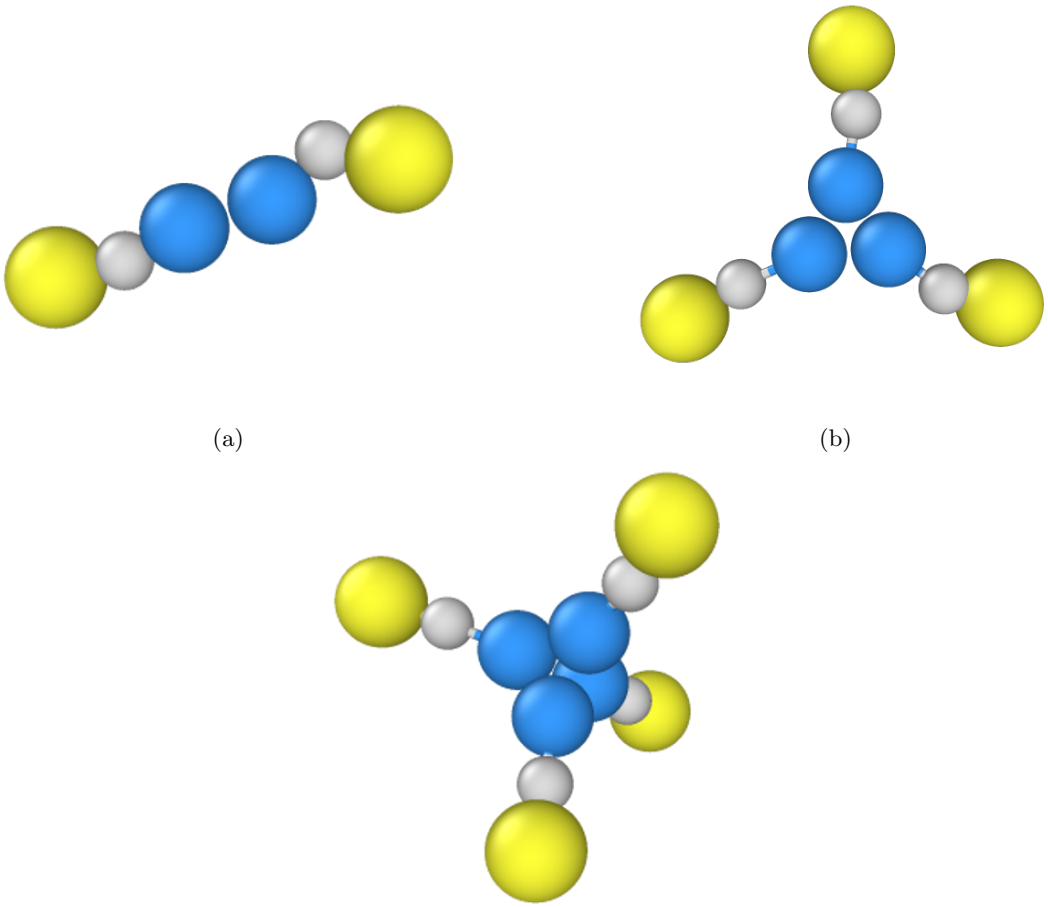
Although simulations run with the values for the parameters described above did create dimers in some cases, for example when placing monomers in pairs with their dimerisation domains close together, other simulations showed that the final outcome could be almost entirely dependent on the initial arrangement of proteins. When monomers were placed too far apart, Brownian motion combined with the repulsive Coulomb interactions meant monomers would spread out through the system space without pairing up. Theoretically, dimers could form given a sufficient amount of proteins and time, but this was not seen as plausible on a reasonable timescale given the limited number of monomers used in simulations here. Alternatively, placing many monomers close together and oriented parallelly did show a significant fraction of them forming dimers, but also allowed for the formation of trimers and tetramers. As described in section 3.1, nucleocapsid proteins dimerise *in vivo* by combining their dimerisation domains into a single stable structure. To our knowledge, this should not allow for the formation of larger oligomers, hence why those groupings are deemed undesirable for our purposes.

Dimers are usually positioned on a single axis, with the attracting dimerisation domains close together and touching, and the other domains facing away due to their repulsive charges (see Figure 3.4a). Although some bending and turning of the monomers is still expected, we observed that in the majority of cases these dimer structures remained stable without falling apart over the used time spans.

Trimers formed stable structures with three monomers approximately 120° apart on a single plane (see Figure 3.4b). Although the addition of another dimerisation domain seemed to create an overall more stable structure than a dimer, certain cases with relatively strong repulsive interactions combined with random Brownian motion could cause one of the three to be pushed away from the others, creating a dimer and separate monomer instead.

Though more rare than smaller groupings, some tetramers were observed in simulations as well. Four monomers would arrange into a 'saddle shape', with two monomers curving in one direction, and two more curving in the other and rotated 90° (see Figure 3.4c). Theoretically, four monomers could also be arranged in a plane similar to what was seen for trimers, but any random motion out of plane would likely result in either the saddle configuration as described here. Tetramers were never observed to divide into smaller groupings, likely owing to the strong attractive interactions at their cores.

Larger oligomers (pentamers, hexamers, etc.) were never observed under any condition tested in this series of simulations. Unless the initial arrangement would be specifically designed to create them, it seems unlikely that they would form spontaneously under the conditions used here. For example, the repulsive 'shell' of a tetramer from the charged domains pointing outwards would make it difficult for an additional monomer to approach close enough to be attracted into the core.



(a) (b)

(c)

Figure 3.4: Graphical view of possible oligomers. **(a)** Dimer in extended configuration. **(b)** Trimer with all monomers on a single plane and at approximately 120° from each other. **(c)** Tetramer with monomers in a 'saddle' configuration.

3.5 Potential parameter sweep

To find which combinations for parameters ε and α allow for dimerisation of N without creating larger clusters, a parameter sweep was performed. 10 initialisation files were created, each containing 125 N monomers in a 5x5x5 grid (see Figure 3.5). The RBD domains were spaced 100 Å apart from each other, with the LINK and DD of each monomer at a random orientation to the corresponding RBD. Simulations were run for all combinations of $\varepsilon = \{10, 20, 30, ..., 100\}$ and $\alpha = \{1, 2, 3, ..., 10\}$ repeating over all initialisation files, giving a total of 1,000 simulations. A fixed simulation duration was chosen based on a number of preliminary runs, in which the groupings of N monomers were observed to remain unchanged after some time. Based on this, the number of timesteps was chosen to ensure that all clustering interactions would have sufficient time to resolve while limiting the amount of computational resources required to complete the simulations.

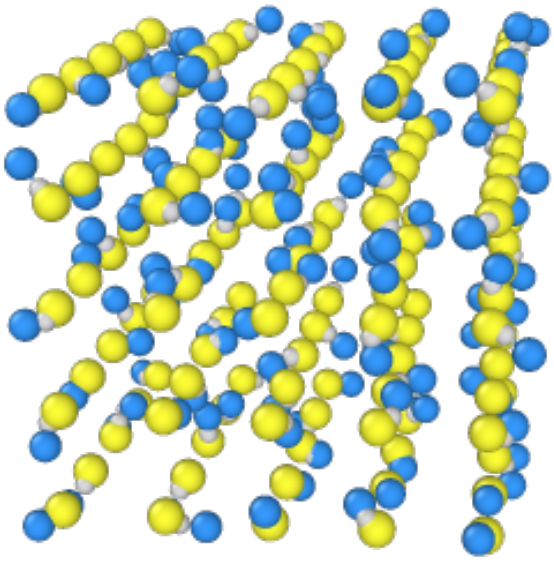
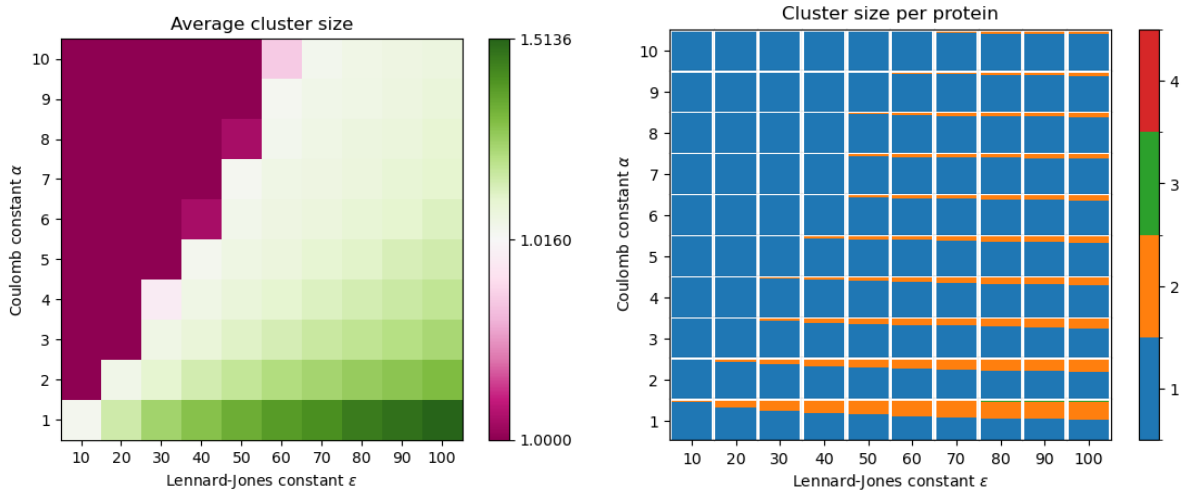


Figure 3.5: Graphical overview of one of the 10 possible initialisation states for the parameter sweep simulations. Yellow RBDs are positioned in a 5x5x5 grid pattern at 100 Å distance, with the LINK and DD at a random orientation for each monomer. Image created in OVITO.

To count the size and amount of clusters in each simulation, the number of dimerisation domains within 5 Å was determined for each DD on the final frame; free monomers would be assigned a cluster size of 1, both N in a dimer would be assigned a cluster size of 2, etc. With 125 monomers per simulation, all 10 simulations for a given parameter combination contain at least 1 dimer on average for a mean cluster size of ($\frac{123 \cdot [monomer] + 2 \cdot [dimer]}{125} = \frac{123 \cdot 1 + 2 \cdot 2}{125} =$) 1.016 (see Figure 3.6a). Combinations of strong Coulomb forces with weak Lennard-Jones forces caused little to no oligomer formation; for certain combinations, e.g. $\varepsilon = 40$ and $\alpha = 6$, oligomers were found in some but not all simulations. In Figure 3.6b a trend becomes noticeable where dimers start to form more consistently in repeat simulations with identical variables as the ratio between them moves towards $\varepsilon > 10\alpha$, or where Lennard-Jones attraction becomes stronger relative to Coulomb repulsion.

Simulations run with 5 of the 10 initialisation files only contained monomers and dimers in all runs. Trimers were observed in some, but not all, simulations where $\alpha = 2$ and $\varepsilon = 100$, and $\alpha = 1$ and $\varepsilon = \{50, 70, 80, 90, 100\}$. Tetramers were seen in 1 out of 10 simulations where $\alpha = 1$ and $\varepsilon = \{60, 70, 80, 90, 100\}$ (see Table 3.3). Trimers and tetramers appearing in subsequent simulations from the same initialisation file consisted of the same particles and appeared at roughly the same position. Exceptions to this are file 3 where a trimer is formed for $\varepsilon = 80$ and $\varepsilon = 100$ but not $\varepsilon = 90$, and file 6 where the sole trimer from $\varepsilon = 50$ became the trimer observed for $\varepsilon = 60$ onwards.



(a) (b)

Figure 3.6: **(a)** Heatmap showing the average clustering size per monomer per combination of parameters ε and α. **(b)** Stacked area plot showing the actual distribution of cluster size per monomer per combination of ε and α.

Table 3.3: Amount of larger oligomers (trimers;tetramers) observed during the parameter sweep for a given initialisation file per given combination of ε and α.

File number \ ε	50	60	70	80	90	100	$\boldsymbol{\alpha}$
0	0;0	0;0	0;0	1;0	1;0	1;0	1
1	0;0	1;0	1;0	2;0	2;0	2;0	1
2	0;0	0;0	1;0	1;0	3;0	4;0	1
3	0;0	0;0	0;0	2;0	1;0	3;0	1
	0;0	0;0	0;0	0;0	0;0	1;0	2
6	1;0	0;1	0;1	1;1	1;1	3;1	1

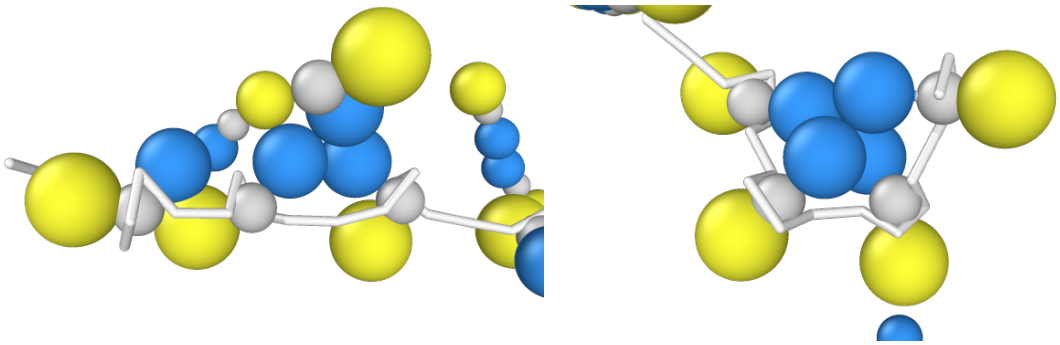
3.6 N-RNA simulations

As a first step towards a full model to study SARS-CoV-2 genome packaging, we performed exploratory simulations on a system containing both N and a longer bead-spring model representing RNA. The chain consists of beads 4 Å in diameter connected by springs 20 Å long, using that the persistence length of RNA is $l_p \approx 10$ Å and a Kuhn segment length is twice the persistence length. The springs have a spring constant $k = 5000$ to limit the stretching and contracting of segments. As a single base of RNA is 3.4 Å in height and has a charge of -1, each segment represents six nucleotides and thus has a charge of -6.

To enable N and RNA to interact with each other, the RBD and LINK domains of N now serve an additional function. By implementing Coulomb potentials between the positively charged domains and negatively charged RNA, N monomers will be attracted to the chain while still able to interact with other N as before. In these simulations, the potential parameters were set as $\varepsilon = 40$ between dimerisation domains, $\alpha = 4$ for interactions between monomers, and $\alpha = 3$ for interactions between monomers and RNA, and among RNA particles. σ between domains and RNA was set at the size of the domain and $\sigma = 1.0 \cdot 2^{-\frac{1}{6}}$ for RNA-RNA interactions; particle size for other type combinations remain as described in Table 3.2.

For the initial setup, the RNA was positioned in a straight line surrounded by N monomers parallel to the strand at varying distances. Monomers were positioned in pairs with dimerisation domains facing each other, meaning the majority of particles would form dimers from the very start of the simulation.

While dimers moving away from the RNA would remain stable for the full length of the simulation, those approaching the strand rarely remained as dimers. On occasion only one of the two monomers would attach to the RNA with the other becoming a free monomer. When a full dimer did bind to the RNA, it would frequently interact with other nearby N to form trimers and tetramers, despite these arrangements not appearing during the parameter sweep (see Figure 3.6a). Zooming in on the interaction surface between N and RNA, the strand displayed winding behaviour around the LINK domains by forming half- to full turns around a monomer (see Figure 3.7).



(a) (b)

Figure 3.7: Graphical views of N monomers interacting with RNA (white strand). **(a)** The RNA winds around the LINK domains of three monomers, including two from a trimer. **(b)** Four monomers bound to the RNA are able to form a tetramer. Images created in OVITO.

3.7 Appendix

The code used for the simulations in Sections 3.5 and 3.6 is available at `https://github.com/LarsvdBiggelaar/sarscov2-bep/`.

ColabFold using MMseqs2 is available at `https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb`.

- Version 1.5.2
- num_relax = 0
- template_mode = pdb100
- msa_mode = mmseqs2_uniref_env
- pair_mode = unpaired_paired
- model_type = auto
- num_recycles = auto
- recycle_early_stop_tolerance = auto
- pairing_strategy = greedy
- max_msa = auto
- num_seeds = 1

Chapter 4

Discussion

4.1 On the appearance of trimers and tetramers

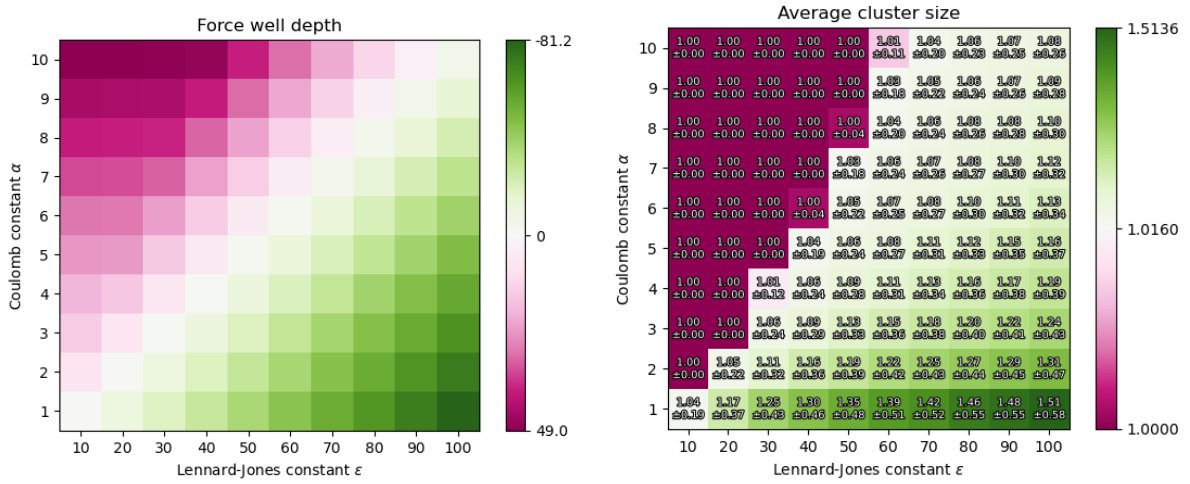
In section 3.5 we showed that as ε increases and α decreases, and thus the force balance shifts from repulsive to attractive, the number of dimers rose in line with expectations. This change unfortunately also caused an increase in the formation of trimers and tetramers. Further analysis shows that these trimers and tetramers are not evenly found throughout simulations, however. Similar to what was noted in 3.4, the starting positions of particles have a considerable influence on the outcome of the simulation regardless of other parameters and simulation time. To be able to draw general conclusions based on other parts of a system independent of its starting state, significant care should be taken to minimize this effect. Although the approach described here with multiple initialisation files could be taken as a starting point, this could be further expanded upon by for example increasing the number of repeats with different random seeds, or by randomizing the system more before starting full simulations.

We hypothesised that the sum of the Lennard-Jones force (Equation 3.5) and the Coulomb force (Equation 3.7) acting on a pair of monomers could act as an indication for whether or not dimers could stably form similar to what was shown in the graph of Figure 3.3. More specifically, a negative total force in the 'well' indicates that the monomers experience a net attractive force. The well depth was calculated over the same ranges for ϵ and α as in the parameter sweep, with the resulting heatmap shown in Figure 4.1a. Two regions are distinctly visible where the minimal force is either positive or negative, split by the diagonal combinations. Despite showing a similar general trend, there are clear differences when comparing this result to that of the actual cluster sizes of the simulations (Figure 4.1b).

Although the most striking difference is that for $\epsilon \leq 50$ and $\epsilon < 10\alpha$ little to no dimers are observed, this result was mostly expected. If the net force is always repulsive regardless of the distance between monomers, the exact magnitude of the repulsion should be arbitrary. Yet for a number of parameter combinations where the minimal force is still repulsive, especially with higher values of ϵ, clustering was still observed. This could be due to a number of factors that have not been taken into account for the well depth calculation. Firstly, the calculation assumes that the two monomers in question are isolated, whereas in the simulations they each have a number of neighbours with which they also interact. If any two monomers are pushed towards each other by their surroundings, they will end up closer together than if they had no surrounding monomers at all. Secondly, the calculation assumes that the distances between the domains of a single monomer remain constant. This does not correspond fully with the simulations, where Coulomb forces between charged domains combined with the relatively low spring constants allows said domains to move away from the attracting dimerisation domains. The increased distances cause lower repulsion, which could mean the minimal net force becomes negative in some cases for which the calculation predicts a positive minimum. Lastly, the drag experienced by all particles in the system further counteracts the repulsive forces experienced by a dimer, aiding in keeping the two monomers together.

The second region of interest is the domain of high ϵ and low α. Unlike in the repulsive domain where the weight of repulsion is less important, the magnitude of the attractive force can be quite significant in determining the amount and size of clusters observed. For $\alpha = 1$ the calculations and simulations seem to correlate fairly well, with an increased net attractive force appearing in an increase

in cluster count and size. However, for higher values of α there is a sudden change in the final outcome of simulations, whereas the calculations indicate a more gradual change. These differences can be at least partially explained by two other aspects of the system not taken into account here. As described in Section 3.3 and Equation 3.1, particles in the system always experience Brownian motion in the form of a random force $\vec{F}_{\mathrm{R}}$. While in most cases this random force is of lower strength than the parameters we are interested in, it becomes more important in cases where the net force is only slightly attractive and thus of comparable magnitude as the random force. Even if an attractive force is possible and dimers are created, random movement could shift monomers to a distance where the net force drives them apart and their dimer is split. Furthermore, regardless of the net force minimum if dimerisation domains are close enough together, sufficient Coulomb forces could create a 'repulsive barrier' at slightly longer distances. If this barrier is large enough, separate monomers simply cannot approach close enough together to interact, meaning they are not able to dimerise even if such a dimer would remain stable once formed.



(a) (b)

Figure 4.1: **(a)** Heatmap showing the 'well depth' after summation of the Lennard-Jones and Coulomb forces for given combinations of ε and α. Negative numbers indicate a net attractive force, positive numbers indicate a net repulsive force. **(b)** Heatmap showing the average clustering size per monomer per combination of parameters ε and α. Numbers indicate the average $\pm$ standard devation cluster size. At size 1.0160, at least one dimer has formed for every repeat for that parameter combination.

4.2 On the implementation of RNA

The simulations described in Section 3.6 already showed some interesting and promising results. For example, the winding of RNA seen in Figure 3.7a bears resemblance to what has previously been described in Section 2.3. The results are a first step towards a more complete system of SARS-CoV-2 genome packaging, but should be seen as a preliminary test rather than an accurate representation of actual behaviour. While interactions between a single monomer and the RNA strand show potential winding and packaging around the protein, the system still seems to lack structures of a higher order. As the RNA attracts proteins regardless of dimerisation state, groups of different sizes seem to be randomly distributed along the strand. Although the strand can act as a scaffold to allow dimerisation of monomers, some monomers get stuck between two clusters unable to interact with either, yet trimers appear at an increased rate along the RNA when compared to free-floating proteins. Monomers initially placed closer to the RNA quickly attach to it, but after this first stage of interactions remaining free monomers are either pushed away by those already connected or are already beyond the cut-off distance. The RNA is compacted significantly, reducing the end-to-end distance from 160 nm to roughly 90 nm, but this co-occurs rapidly with the initial binding of protein and does not change significantly after this time. Lastly, apart from local bends due to winding the strand remains mostly straight, whereas the final RNP structure is expected to consist of curled or spherical substructures each formed by a small

number of dimers.

4.3 Future prospect

The model described throughout this work was developed as a novel theoretical approach for the increased understanding of the SARS-CoV-2 nucleocapsid protein. While the pair potentials between monomers have been analysed with a parameter sweep and interactions between N and RNA have been preliminarily explored, there is room to investigate these aspects further in-depth and yet more aspects not taken into consideration here can be inspected as well.

The coarse-grained model of N is partially based on the assumption that the N- and C-terminal disordered regions have little effect on the protein's function. Although this assumption is recuperated by most literature at least for the formation of N-dimers, results from (Ribeiro-Filho et al. 2022) suggest that the C-terminal tail could be important for the development of larger RNP-structures by interacting with tails from other monomers. As such, the addition of this tail in the model, potentially as simple as a new bead attached to the dimerisation domain capable of interacting with other identical beads, could already give further insight into the formation of higher order structures in the genome packaging process.

A feature of HOOMD-blue not employed in this project is the class of updater functions, which can alter the properties of particles during a simulation. In the context of expanding on this work, an interesting application would be to change the behaviour of dimerisation domains after the formation of a dimer. For example, a permanent bond could be created between two neighbouring domains similar to those already present within a monomer if they move close enough together, while simultaneously reducing the ϵ of the Lennard-Jones potential of both domains to 0.1 such that it only prevents particle overlap as with other particle types. Alternatively, associative bond swaps (Ciarella and Ellenbroek 2022) could be implemented between DDs. This would mean that if two domains are in close proximity, other particles only experience a repulsive force until the domains move away again. While the two approaches offer flexibility in terms of creating a permanent bond between dimerisation domains or keeping it nonpermanent, they could both simplify future parameter sweeps as the formation of trimers and tetramers would become more difficult or impossible. In turn, this would allow finetuning for other goals, such as interactions between N and RNA, and the formation of larger RNP structures. Furthermore, as N seems to only interact with RNA when in dimer form, simulations could be set up such that electrostatic interactions between protein and RNA are only enabled once dimerisation has taken place.

Electrostatic interactions have so far been simulated through the use of Coulomb potentials. While this approach was deemed appropriate for the aims and extent of this work, the relatively simple Coulomb equation does not necessarily accurately reflect electrostatic interactions of real N proteins. Depending on the value of α the equations are solved as if the particles exist in a vacuum, with the number of interactions of a single particle essentially only limited by a given cutoff distance r_{cut}. However, a nucleocapsid protein inside a host cell or *in vitro* experiment would exist in an aqueous medium containing ions. This medium would cause electrostatic screening between charged monomer domains, reflected as a steeper drop-off in interaction strength over distance and a weaker force overall. The stronger Coulomb potential was chosen here to create effective repulsion between monomers to counteract the formation of trimers and tetramers. However, if the suggestion regarding interactions between dimerisation domains as described above would be incorporated, this would allow the role of the other domains to be focused more on interactions with RNA as with the real protein, in which case the screening effect could be important to achieve realistic results. An effect comparable to screening could be achieved with the Coulomb potential by lowering α and r_{cut}, or alternatively it could be replaced by a potential which does take screening into account, such as the Ewald pair potential or DLVO pair potential.

As described in the previous Section, our implementation of RNA in the simulations showed interesting results which are unfortunately not at the level of detail to be directly compared to experimental observations. Yet, as we have seen time and time again throughout this project, even small changes to parameters or interactions can have a significant impact on the final state of a simulation. With the suggestions described here, we anticipate a far more extensive model of N and RNA to be developed during projects in the (not too distant) future, potentially already mirroring some of the behaviour

seen in actual experiments and answering questions we currently hold about the process. Hopefully, by increasing our understanding of SARS-CoV-2 and similar viruses, the impact of future disease outbreaks can be increasingly reduced for a healthier future for everyone.

Chapter 5

Conclusion

In this report, we have created a coarse-grained model for the SARS-CoV-2 nucleocapsid for use in molecular dynamics simulations to gain a better understanding of the protein's role in viral genome packaging. Given a lack of existing 3D protein structures due to its inherent disorder, we predicted possible structures for the N dimer. From these structures, we were able to design a coarse-grained model to get a simplified representation of the actual protein. To emulate the dimerisation behaviour of the *in vivo* protein, we implemented a Lennard-Jones potential between dimerisation domains to get an attractive interaction. Coulomb potentials were integrated to create repulsion between monomers' charged domains to prevent unwanted larger groupings of N from appearing. A number of small-scale simulations provided us with initial insight into the relative strengths of both potentials. From there, we simulated a range of parameter combinations to find those that enabled dimerisation without creating high numbers of undesired trimers and tetramers. However, analysing the final states of these simulations also showed us that the outcome was dependent on the initial setup. This means our observations can still hold when applied to comparable starting situations, but we cannot yet determine the most optimal combination of parameters for any general situation. Nonetheless, we were able still to see some general trends in the amount of N clusters based on the chosen parameter values alone. By adding a model for RNA to the system, we could observe interesting interactions resembling a previously proposed structure for RNP. The novel theoretical approach we have described in this report will hopefully be further expanded upon in the near future as we move towards a full *in silico* system of SARS-CoV-2 genome packaging.

Bibliography

Alexander E. Gorbalenya, and et al. (Mar. 2020). "The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2". In: *Nature Microbiology* 5.4, pp. 536–544. URL: https://doi.org/10.1038/s41564-020-0695-z.

Anderson, Joshua A., Jens Glaser, and Sharon C. Glotzer (Feb. 2020). "HOOMD-blue: A Python package for high-performance molecular dynamics and hard particle Monte Carlo simulations". In: *Computational Materials Science* 173, p. 109363. URL: https://doi.org/10.1016/j.commatsci.2019.109363.

Arora, Pooja et al. (July 2021). "New Coronavirus (SARS-CoV-2) Crossing Borders Beyond Cities, Nations, and Continents: Impact of International Travel". In: *Balkan Medical Journal* 38.4, pp. 205–211. URL: https://doi.org/10.5152/balkanmedj.2021.21074.

Assiri, Abdullah, Allison McGeer, et al. (Aug. 2013). "Hospital Outbreak of Middle East Respiratory Syndrome Coronavirus". In: *New England Journal of Medicine* 369.5, pp. 407–416. URL: https://doi.org/10.1056/nejmoa1306742.

Assiri, Abdullah, Jaffar A Al-Tawfiq, et al. (Sept. 2013). "Epidemiological, demographic, and clinical characteristics of 47 cases of Middle East respiratory syndrome coronavirus disease from Saudi Arabia: a descriptive study". In: *The Lancet Infectious Diseases* 13.9, pp. 752–761. URL: https://doi.org/10.1016/s1473-3099(13)70204-4.

Bai, Zhihua et al. (June 2021). "The SARS-CoV-2 Nucleocapsid Protein and Its Role in Viral Structure, Biological Functions, and a Potential Target for Drug or Vaccine Mitigation". In: *Viruses* 13.6, p. 1115. URL: https://doi.org/10.3390/v13061115.

Baltimore, David (Sept. 1971). "Expression of animal virus genomes". In: *Bacteriological Reviews* 35.3, pp. 235–241. URL: https://doi.org/10.1128/br.35.3.235-241.1971.

Berman, Helen M. et al. (Jan. 2000). "The Protein Data Bank". In: *Nucleic Acids Research* 28.1, pp. 235–242. ISSN: 0305-1048. eprint: https://academic.oup.com/nar/article-pdf/28/1/235/9895144/280235.pdf. URL: https://doi.org/10.1093/nar/28.1.235.

Bradburne, A. F., M. L. Bynoe, and D. A. Tyrrell (Sept. 1967). "Effects of a "new" human respiratory virus in volunteers." In: *BMJ* 3.5568, pp. 767–769. URL: https://doi.org/10.1136/bmj.3.5568.767.

Casasanta, Michael A et al. (Dec. 2022). "Structural Insights of the SARS-CoV-2 Nucleocapsid Protein: Implications for the Inner-workings of Rapid Antigen Tests". In: *Microscopy and Microanalysis* 29.2, pp. 649–657. ISSN: 1431-9276. eprint: https://academic.oup.com/mam/article-pdf/29/2/649/49764799/ozac036.pdf. URL: https://doi.org/10.1093/micmic/ozac036.

Chan, J W M (Aug. 2003). "Short term outcome and risk factors for adverse clinical outcomes in adults with severe acute respiratory syndrome (SARS)". In: *Thorax* 58.8, pp. 686–689. URL: https://doi.org/10.1136/thorax.58.8.686.

Chang, Chung-ke et al. (Oct. 2005). "Modular organization of SARS coronavirus nucleocapsid protein". In: *Journal of Biomedical Science* 13.1, pp. 59–72. URL: https://doi.org/10.1007/s11373-005-9035-9.

Chen, Chun-Yuan et al. (May 2007). "Structure of the SARS Coronavirus Nucleocapsid Protein RNA-binding Dimerization Domain Suggests a Mechanism for Helical Packaging of Viral RNA". In: *Journal of Molecular Biology* 368.4, pp. 1075–1086. URL: https://doi.org/10.1016/j.jmb.2007.02.069.

Chen, Nanshan et al. (Feb. 2020). "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study". In: *The Lancet* 395.10223, pp. 507–513. URL: https://doi.org/10.1016/s0140-6736(20)30211-7.

Christian, M. D. et al. (May 2004). "Severe Acute Respiratory Syndrome". In: *Clinical Infectious Diseases* 38.10, pp. 1420–1427. URL: https://doi.org/10.1086/420743.

Ciarella, Simone and Wouter Ellenbroek (Apr. 2022). "Associative bond swaps in molecular dynamics". In: *SciPost Physics* 12.4. URL: https://doi.org/10.21468/scipostphys.12.4.128.

Corman, Victor M. et al. (Aug. 2014). "Antibodies against MERS Coronavirus in Dromedary Camels, Kenya, 1992–2013". In: *Emerging Infectious Diseases* 20.8. URL: https://doi.org/10.3201/eid2008.140596.

Crick, Francis H. C. (1958). "On protein synthesis". en. In: *Symposia of the Society for Experimental Biology* 12, pp. 138–163.

— (Aug. 1970). "Central Dogma of Molecular Biology". In: *Nature* 227.5258, pp. 561–563. URL: https://doi.org/10.1038/227561a0.

Dinesh, Dhurvas Chandrasekaran et al. (Dec. 2020). "Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein". In: *PLOS Pathogens* 16.12. Ed. by Michael S. Diamond, e1009100. URL: https://doi.org/10.1371/journal.ppat.1009100.

Drosten, Christian et al. (Aug. 2014). "Transmission of MERS-Coronavirus in Household Contacts". In: *New England Journal of Medicine* 371.9, pp. 828–835. URL: https://doi.org/10.1056/nejmoa1405858.

Economist, The and Sondre Ulvund Solstad (May 2021). "The pandemic's true death toll". In: *The Economist*.

Evans, Richard et al. (Oct. 2021). "Protein complex prediction with AlphaFold-Multimer". In: URL: https://doi.org/10.1101/2021.10.04.463034.

Feehan, Jack and Vasso Apostolopoulos (July 2021). "Is COVID-19 the worst pandemic?" In: *Maturitas* 149, pp. 56–58. URL: https://doi.org/10.1016/j.maturitas.2021.02.001.

Forterre, Patrick and David Prangishvili (Oct. 2009). "The Great Billion-year War between Ribosome- and Capsid-encoding Organisms (Cells and Viruses) as the Major Source of Evolutionary Novelties". In: *Annals of the New York Academy of Sciences* 1178.1, pp. 65–77. URL: https://doi.org/10.1111/j.1749-6632.2009.04993.x.

Fu, Yu-Zhi et al. (Oct. 2020). "SARS-CoV-2 membrane glycoprotein M antagonizes the MAVS-mediated innate antiviral response". In: *Cellular & Molecular Immunology* 18.3, pp. 613–620. URL: https://doi.org/10.1038/s41423-020-00571-x.

Ge, Xing-Yi et al. (Oct. 2013). "Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor". In: *Nature* 503.7477, pp. 535–538. URL: https://doi.org/10.1038/nature12711.

Gordon, David E. et al. (Apr. 2020). "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing". In: *Nature* 583.7816, pp. 459–468. URL: https://doi.org/10.1038/s41586-020-2286-9.

Groot, Raoul J. de et al. (July 2013). "Commentary: Middle East Respiratory Syndrome Coronavirus (MERS-CoV): Announcement of the Coronavirus Study Group". In: *Journal of Virology* 87.14, pp. 7790–7792. URL: https://doi.org/10.1128/jvi.01244-13.

Guan, Y. et al. (Oct. 2003). "Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Southern China". In: *Science* 302.5643, pp. 276–278. URL: https://doi.org/10.1126/science.1087139.

Hendrix, Roger W. et al. (Mar. 1999). "Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage". In: *Proceedings of the National Academy of Sciences* 96.5, pp. 2192–2197. URL: https://doi.org/10.1073/pnas.96.5.2192.

Huang, Chaolin et al. (Feb. 2020). "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China". In: *The Lancet* 395.10223, pp. 497–506. URL: https://doi.org/10.1016/s0140-6736(20)30183-5.

Huang, Yuan et al. (Aug. 2020). "Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19". In: *Acta Pharmacologica Sinica* 41.9, pp. 1141–1149. URL: https://doi.org/10.1038/s41401-020-0485-4.

Hui, David S. et al. (Feb. 2020). "The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health — The latest 2019 novel coronavirus outbreak in Wuhan, China". In: *International Journal of Infectious Diseases* 91, pp. 264–266. URL: https://doi.org/10.1016/j.ijid.2020.01.009.

Jamison, David A. et al. (May 2022). "A comprehensive SARS-CoV-2 and COVID-19 review, Part 1: Intracellular overdrive for SARS-CoV-2 infection". In: *European Journal of Human Genetics* 30.8, pp. 889–898. URL: https://doi.org/10.1038/s41431-022-01108-8.

Johnston, Sebastian and Stephen Holgate (1996). "Epidemiology of Viral Respiratory Tract Infections". In: *Viral and Other Infections of the Human Respiratory Tract*. Springer Netherlands, pp. 1–38. URL: https://doi.org/10.1007/978-94-011-7930-0_1.

Jumper, John et al. (July 2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873, pp. 583–589. URL: https://doi.org/10.1038/s41586-021-03819-2.

Kilbourne, Edwin D. (1987). "Influenza in Man". In: *Influenza*. Springer US, pp. 157–218. URL: https://doi.org/10.1007/978-1-4684-5239-6_7.

Koonin, Eugene V., Valerian V. Dolja, et al. (May 2020). "Global Organization and Proposed Megataxonomy of the Virus World". In: *Microbiology and Molecular Biology Reviews* 84.2. URL: https://doi.org/10.1128/mmbr.00061-19.

Koonin, Eugene V. and Petro Starokadomskyy (Oct. 2016). "Are viruses alive? The replicator paradigm sheds decisive light on an old but misguided question". In: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 59, pp. 125–134. URL: https://doi.org/10.1016/j.shpsc.2016.02.016.

Ksiazek, Thomas G. et al. (May 2003). "A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome". In: *New England Journal of Medicine* 348.20, pp. 1953–1966. URL: https://doi.org/10.1056/nejmoa030781.

Lam, Tommy Tsan-Yuk et al. (Mar. 2020). "Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins". In: *Nature* 583.7815, pp. 282–285. URL: https://doi.org/10.1038/s41586-020-2169-0.

Lefkowitz, Elliot J et al. (Oct. 2017). "Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV)". In: *Nucleic Acids Research* 46.D1, pp. D708–D717. URL: https://doi.org/10.1093/nar/gkx932.

Li, Siyu and Roya Zandi (2022). "Biophysical Modeling of SARS-CoV-2 Assembly: Genome Condensation and Budding". In: *Viruses* 14.10. ISSN: 1999-4915. URL: https://www.mdpi.com/1999-4915/14/10/2089.

Louten, Jennifer (2016). "Virus Structure and Classification". In: *Essential Human Virology*. Elsevier, pp. 19–29. URL: https://doi.org/10.1016/b978-0-12-800947-5.00002-8.

Lowery, Shea A., Alan Sariol, and Stanley Perlman (July 2021). "Innate immune and inflammatory responses to SARS-CoV-2: Implications for COVID-19". In: *Cell Host & Microbe* 29.7, pp. 1052–1062. URL: https://doi.org/10.1016/j.chom.2021.05.004.

Lu, Hongzhou, Charles W. Stratton, and Yi-Wei Tang (Feb. 2020). "Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle". In: *Journal of Medical Virology* 92.4, pp. 401–402. URL: https://doi.org/10.1002/jmv.25678.

Meyer, Benjamin et al. (Apr. 2014). "Antibodies against MERS Coronavirus in Dromedary Camels, United Arab Emirates, 2003 and 2013". In: *Emerging Infectious Diseases* 20.4, pp. 552–559. URL: https://doi.org/10.3201/eid2004.131746.

Mirdita, Milot et al. (May 2022). "ColabFold: making protein folding accessible to all". In: *Nature Methods* 19.6, pp. 679–682. URL: https://doi.org/10.1038/s41592-022-01488-1.

Müller, Marcel A. et al. (Dec. 2014). "MERS Coronavirus Neutralizing Antibodies in Camels, Eastern Africa, 1983–1997". In: *Emerging Infectious Diseases* 20.12. URL: https://doi.org/10.3201/eid2012.141026.

Nieto-Torres, Jose L. et al. (May 2014). "Severe Acute Respiratory Syndrome Coronavirus Envelope Protein Ion Channel Activity Promotes Virus Fitness and Pathogenesis". In: *PLoS Pathogens* 10.5. Ed. by Mark R. Denison, e1004077. URL: https://doi.org/10.1371/journal.ppat.1004077.

Peiris, JSM et al. (Apr. 2003). "Coronavirus as a possible cause of severe acute respiratory syndrome". In: *The Lancet* 361.9366, pp. 1319–1325. URL: https://doi.org/10.1016/s0140-6736(03)13077-2.

Peng, Ya et al. (Sept. 2020). "Structures of the SARS-CoV-2 nucleocapsid and their perspectives for drug design". In: *The EMBO Journal* 39.20. URL: https://doi.org/10.15252/embj.2020105938.

Reusken, Chantal B.E.M. et al. (Aug. 2014). "Geographic Distribution of MERS Coronavirus among Dromedary Camels, Africa". In: *Emerging Infectious Diseases* 20.8, pp. 1370–1374. URL: https://doi.org/10.3201/eid2008.140590.

Reusken, Chantal BEM et al. (Oct. 2013). "Middle East respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: a comparative serological study". In: *The Lancet Infectious Diseases* 13.10, pp. 859–866. URL: https://doi.org/10.1016/s1473-3099(13)70164-6.

Ribeiro-Filho, Helder Veras et al. (May 2022). "Structural dynamics of SARS-CoV-2 nucleocapsid protein induced by RNA binding". In: *PLOS Computational Biology* 18.5. Ed. by Alexey Onufriev, e1010121. URL: https://doi.org/10.1371/journal.pcbi.1010121.

Tyrrell, D. A. J. and M. L. Bynoe (June 1965). "Cultivation of a Novel Type of Common-cold Virus in Organ Cultures". In: *BMJ* 1.5448, pp. 1467–1470. URL: https://doi.org/10.1136/bmj.1.5448.1467.

Weiss, Susan R. and Sonia Navas-Martin (Dec. 2005). "Coronavirus Pathogenesis and the Emerging Pathogen Severe Acute Respiratory Syndrome Coronavirus". In: *Microbiology and Molecular Biology Reviews* 69.4, pp. 635–664. URL: https://doi.org/10.1128/mmbr.69.4.635-664.2005.

WHO (Jan. 2004). *Severe acute respiratory syndrome (SARS) Report by the Secretariat*. Tech. rep.

— (Feb. 2020a). *Novel Coronavirus (2019-nCoV) Situation Report – 22*. Tech. rep.

— (Mar. 2020b). *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*. Tech. rep.

— (Nov. 2022a). *Middle East respiratory syndrome: global summary and assessment of risk*. Tech. rep.

— (2022b). URL: https://covid19.who.int/ (visited on 10/26/2022).

World Health Organization (2015). *Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003 — who.int*. https://www.who.int/publications/m/item/summary-of-probable-sars-cases-with-onset-of-illness-from-1-november-2002-to-31-july-2003. [Accessed 25-10-2023].

Wu, Can-rong et al. (Jan. 2022). "Structure genomics of SARS-CoV-2 and its Omicron variant: drug design templates for COVID-19". In: *Acta Pharmacologica Sinica* 43.12, pp. 3021–3033. URL: https://doi.org/10.1038/s41401-021-00851-w.

Wu, Fan et al. (Feb. 2020). "A new coronavirus associated with human respiratory disease in China". In: *Nature* 579.7798, pp. 265–269. URL: https://doi.org/10.1038/s41586-020-2008-3.

Xiao, Kangpeng et al. (May 2020). "Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins". In: *Nature* 583.7815, pp. 286–289. URL: https://doi.org/10.1038/s41586-020-2313-x.

Xiao, Xiao et al. (June 2021). "Animal sales from Wuhan wet markets immediately prior to the COVID-19 pandemic". In: *Scientific Reports* 11.1. URL: https://doi.org/10.1038/s41598-021-91470-2.

Yan, Yuxin et al. (Mar. 2020). "The First 75 Days of Novel Coronavirus (SARS-CoV-2) Outbreak: Recent Advances, Prevention, and Treatment". In: *International Journal of Environmental Research and Public Health* 17.7, p. 2323. URL: https://doi.org/10.3390/ijerph17072323.

Yao, Hangping et al. (Oct. 2020). "Molecular Architecture of the SARS-CoV-2 Virus". In: *Cell* 183.3, 730–738.e13. URL: https://doi.org/10.1016/j.cell.2020.09.018.

Ye, Qiaozhen et al. (Aug. 2020). "Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein". In: *Protein Science* 29.9, pp. 1890–1901. URL: https://doi.org/10.1002/pro.3909.

Zaki, Ali M. et al. (Nov. 2012). "Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia". In: *New England Journal of Medicine* 367.19, pp. 1814–1820. URL: https://doi.org/10.1056/nejmoa1211721.

Zeng, Weihong et al. (June 2020). "Biochemical characterization of SARS-CoV-2 nucleocapsid protein". In: *Biochemical and Biophysical Research Communications* 527.3, pp. 618–623. URL: https://doi.org/10.1016/j.bbrc.2020.04.136.

Zhang, Y. (Apr. 2005). "TM-align: a protein structure alignment algorithm based on the TM-score". In: *Nucleic Acids Research* 33.7, pp. 2302–2309. URL: https://doi.org/10.1093/nar/gki524.

Zhong, Ping, Songxue Guo, and Ting Chen (Mar. 2020). "Correlation between travellers departing from Wuhan before the Spring Festival and subsequent spread of COVID-19 to all provinces in China". In: *Journal of Travel Medicine* 27.3. URL: https://doi.org/10.1093/jtm/taaa036.

Zhou, Peng et al. (Feb. 2020). "A pneumonia outbreak associated with a new coronavirus of probable bat origin". In: *Nature* 579.7798, pp. 270–273. URL: https://doi.org/10.1038/s41586-020-2012-7.

Zinzula, Luca et al. (Jan. 2021). "High-resolution structure and biophysical characterization of the nucleocapsid phosphoprotein dimerization domain from the Covid-19 severe acute respiratory syndrome coronavirus 2". In: *Biochemical and Biophysical Research Communications* 538, pp. 54–62. URL: https://doi.org/10.1016/j.bbrc.2020.09.131.