MSc
2.º CICLO
FCUP
2021

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

U.PORTO

SARS-CoV-2 membrane protein: From genomic data to structural new insights AI-powered

SARS-CoV-2 membrane protein: From genomic data to structural new insights

Manuel N. Pires

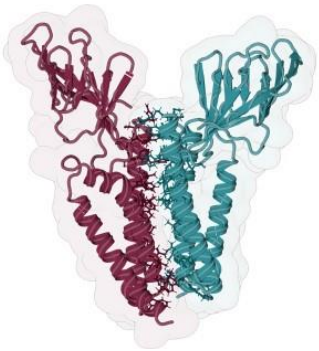
Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Bioinformática e Biologia Computacional

2021

Manuel N. Pires

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc



SARS-CoV-2 membrane protein: From genomic data to structural new insights

Manuel N. Pires
Mestrado em Bioinformática e Biologia Computacional
Departamento de Biologia
2021

Orientador
Vítor Costa, Professor Associado, Faculdade de Ciências da Universidade do Porto

Orientador
Irina Moreira, Professora Auxiliar, Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Acknowledgments

Antes de expor o meu trabalho gostava de agradecer às imensas pessoas que de uma forma ou de outra desempenharam um papel fundamental no desenvolvimento e escrita do mesmo.

Gostaria de começar por agradecer às pessoas que me possibilitaram desenvolver o trabalho aqui apresentado ao aceitarem orientar-me durante este ano, a Professora Doutora Irina Moreira e o Professor Doutor Vitor Costa. Um especial agradecimento à Professora Doutora Irina por me ter possibilitado dar os primeiros passos na bioinformática, bem como por me ter aceite novamente no seu laboratório durante esta etapa. Gostaria também de agradecer às pessoas com quem trabalhei direta ou indiretamente neste grupo, nomeadamente à Nícia por me ter ajudado a guiar o trabalho, assim como à Catarina, à Raquel e à Nádia por terem sido fundamentais neste trabalho que desenvolvemos e por tudo o que aprendi com elas durante este tempo.

De seguida gostaria de agradecer às pessoas que sempre fizeram tudo por mim e me apoiaram em todas as decisões que tomei, ainda que por vezes não compreendessem bem onde iriam dar. Aos meus pais e à minha irmã o meu enorme obrigado por me terem acompanhado nesta desfiante viagem, por festejarem comigo cada um dos meus feitos e por me terem sempre apoiado nos momentos dificeis.

Por fim gostava de agradecer a amigos que de forma direta ou indireta ajudaram a conclusão da minha tese. Agradecer ao Ramalhão por ter estado sempre lá nos momentos de pânico com uma palavra amiga de calma. À Marta e à Raquel, por todas as conversas, passeios e explorações no Porto. À Sara e à Maria que com gosma me aguentarem inteiro até ao fim. À Carolina e à Bárbara por todo o encorajamento que, como em outros momentos, me ofereceram ao longo deste caminho. Ao Piochi, ao Mica, ao Tomé e ao JP pelas várias conversas e partilhas ao longo deste ano, quer virtuais quer com um fino a acompanhar. À Madalena, ao Laulo, ao Tomás, à Cristo, ao Soveral, ao Shmefer, à Rita, ao Pimenta e à Casi por todos os cafés e me fazerem sair de casa para relaxar e me acalmarem. Ao Tozé e à Claúdia por todas as noites de Catan e discussões produtivas que me ajudavam a descontrair nesta altura de stress. Um especial agradecimento ao Tozé por me ter dado a conhecer e ajudado a dar os primeiros passos na bioinformatica e por me ter desafiado a aprender imenso. Ao Chico, à Bea e à Mariana por terem partilhado e ajudado a desmistificar o meu pânico durante este ano.

Abstract

Severe Acute Respiratory Syndrome CoronaVirus-2 (SARS-CoV-2) is responsible for a worldwide pandemic, accounting for 228,000,000 infections and over 4,600,000 deaths. SARS-CoV-2's most abundant structural protein, Membrane (M) protein, has a pivotal role both during viral infection cycle and host interferon antagonism. This is a highly conserved viral protein, thus an interesting and suitable target for drug discovery. As SARS-CoV-2 M protein structure is difficult to experimentally stabilize and crystallize, we developed and applied a detailed and robust *in silico* workflow to predict M protein dimeric structure, membrane orientation, and interface characterization. Mutations in M protein were retrieved from over 1.2 M SARS-CoV-2 genomes and proteins from the Global Initiative on Sharing All Influenza Data (GISAID) database, 91 of which were located at the predicted dimer interface. Among those, we identified mutations in Variants of Concern (VOC) and Variants of Interest (VOI). Binding free energy differences were evaluated for dimer interfacial mutations to infer mutant protein stabilities. A few high-prevalent mutated residues were found to be especially relevant in VOC and VOI. This realization may be a game changer to structure driven formulation of new therapeutics for SARS-CoV-2.

Resumo

O *Severe Acute Respiratory Syndrome CoronaVirus-2* (SARS-CoV-2) é responsável por uma pandemia mundial, sendo responsável por quase 228,000,000 de infeções e mais de 4,600,000 de mortes. A proteína mais abundante do SARS-CoV-2 é a proteína de Membrana (M), que tem um papel fundamental tanto durante o ciclo de infeção viral como enquanto antagonista do interferão do hospedeiro. Esta é uma proteína viral altamente conservada e por isso um alvo interessante e adequado para a descoberta de drogas. Como a estrutura da proteína M do SARS-CoV-2 é difícil de estabilizar e cristalizar experimentalmente desenvolvemos e aplicamos uma detalhada e robusta abordagem *in silico* para prever a estrutura dimérica da proteína M, a sua orientação de membrana e a caracterização da sua interface. Mutações na proteína M foram obtidas de mais de 1.2M de genomas e proteínas da base de dados *Global Initiative on Sharing All Influenza Data* (GISAID), 91 das quais se encontravam localizadas na zona prevista de interface. Entre estas identificamos as mutações em *Variants of Concern* (VOC) e em *Variants of Interest* (VOI). As diferenças nas energias livres de ligação foram avaliadas para as mutações na interface do dímero para inferir a estabilidade das proteínas mutantes. Uns reduzidos números de resíduos mutados muito prevalentes foram descobertos ser especialmente relevantes em VOC e VOI. Esta descoberta pode ser de extrema importância para a formulação de novas terapêuticas baseadas em estrutura para o SARS-CoV-2.

Keywords

SARS-CoV-2, Membrane Protein, Mutations, Dimeric Interface, Protein-Protein Interactions.

Table of Contents

List of Figures and Tables

List of Abbreviations

CoV: CoronaVirus
hCoV: human CoronaVirus
S protein: Spike protein
M protein: Membrane protein
N protein: Nucleocapsid protein
E protein: Envelope protein
HE: Hemagglutinin-Esterase
SARS-CoV: Severe Acute Respiratory Syndrome CoronaVirus
MERS-CoV: Middle East Respiratory Syndrome CoronaVirus
COVID-19: CoronaVirus Disease 19
SARS-CoV-2: Severe Acute Respiratory Syndrome CoronaVirus-2
WHO: World Health Organization
ORF: Open Reading Frame
UTR: Untranslated Region
pp: Polypeptides
nsp: Nonstructural protein
RTC: Replicase Transcriptase Complex
aa: Amino acids
RBD: Receptor Binding Domain
TMH: Transmembrane Helix
ERGIC: Endoplasmic Reticulum Golgi Intermediate Compartment
MHC-I: Major Histocompability Complex Class I
ACE2: Angiotensin-converting enzyme 2 receptor
TMPRSS2: Transmembrane Protease Serine 2
RdRp: RNA-dependent RNA polymerase
VOI: Variants of Interest
VOC: Variants of Concern
GISAID: Global Initiative on Sharing All Influenza Data
NMR: Nuclear Magnetic Resonance
cryo-EM: Cryo-Eletron Microscopy
MD: Molecular Dynamics
$\triangle\triangle G_{binding}$: Free energy variation
WT: Wild Type

OPM: Orientations of Proteins in Membranes

TMHMM: Transmembrane Helix Markov Model

PSIPRED: Prediction of secondary structure

CCTOP: Consensus Constrained TOPology prediction

GPUs: Graphical Processing Units

RMSD: Root Mean Square Deviation

RMSF: Root Mean Square Fluctuation

CHARMM-GUI: Chemistry at HARvard Macromolecular Mechanics Graphical User Interface

GROMACS: GROningen MAchine for Chemical Simulations

CPU: Central Processing Units

HADDOCK: High Ambiguity Driven protein-protein DOCKing

FASTA: FAST-All

vcf-format: Variant Call Format

PISA: Protein Interfaces Surfaces and Assemblies

PRODIGY: PROtein binDIng enerGY

CCA: Cross-Correlation Analysis

Cα: Alpha Carbon

SASA: Solvent-Accessible Surface Area

PPI: Protein-Protein Interaction

Chapter 1 - Introduction

The Coronaviridae family is a group of viruses with spherical shape, club-like spikes and a size of approximately 125 nm. Several CoronaViruses (CoVs) known to infect humans throughout history, were called human CoronaViruses (hCoVs)[1], and contain the common proteins to all CoVs, Spike (S), Membrane (M), Nucleocapsid (N) and Envelope (E) proteins, and some hCoVs also possess hemagglutinin-esterase (HE)[2]. This is a single-stranded positive RNA virus family, organized into four different groups, Alphacoronavirus, Betacoronavirus, Gammacoronavirus and Deltacoronavirus, but only the first two infect humans[2].

1.1 Human Coronavirus History

At the end of 2002, an uncommon pneumonia was detected in Guangdong China. Afterwards, this disease was found to have been caused by a novel CoV from the betacoronavirus group, Severe Acute Respiratory Syndrome CoronaVirus (SARS-CoV). Later, in March 2003 this virus had spread to other countries around the world[3], infecting close to 8,500 humans and killing 916.

In 2012, a man was hospitalized in Saudi Arabia with fever, cough, and shortness of breath. It was found to be due to another betacoronavirus that became known as Middle East Respiratory Syndrome CoronaVirus (MERS-CoV). Since being detected, this virus was found throughout the Arabian Peninsula and in some people that traveled to the region[4]. MERS-CoV was responsible, until now, for 2,468 infection cases and 851 deaths.

In December 2019 another unusual form of pneumonia was detected in Wuhan, China[5]. This disease, later denominated as CoronaVirus Disease 19 (COVID-19), found to be caused by yet another betacoronavirus, Severe Acute Respiratory Syndrome CoronaVirus-2 (SARS-CoV-2)[6], named because of its high identity with SARS-CoV (79%)[7]. As of 21 of September 2021 it was known to have infected nearly 228,000,000 and killed over 4,600,000, around the world, according to the World Health Organization (WHO), distinctively the virus responsible for the most infections and deaths out of the described hCoVs.

1.2 SARS-CoV-2 genome and proteins

The SARS-CoV-2 genome is about 30 kb[8] and typically contains 13 to 15 Open Reading Frames (ORFS)[7] (Figure 1). The genome starts with a 5' Untranslated Region (UTR), followed by ORFs 1a and 1b that encode two long polypeptides (pp), pp1a and pp1b, that are then cleaved into 16 nonstructural proteins (nsps) responsible for multiple viral functions as listed in Table 1.

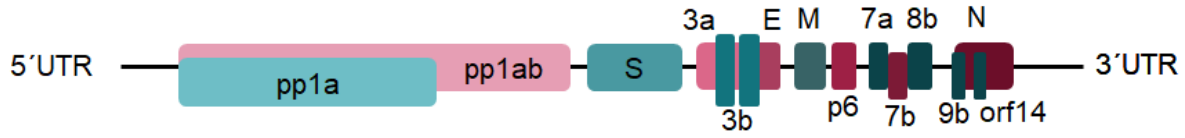


Figure 1 - SARS-CoV-2 genome representation.

Table 1 - Nsp and their functions in Replicase Transcriptase Complex (RTC).

Protein	Function
nsp1	Host mRNA degradation[9]; suppression of Interferon expression[10]; suppression of host antiviral pathways[11].
nsp2	Possible pathogenic activity[12].
nsp3	Immune evasion[13]; Interaction with N vital for viral replication[14]; Viral peptide cleavage[15].
nsp4	Membrane rearrangements vital in viral replication[16].
nsp5	Immune evasion[17].
nsp6	Type I interferon supressor[18].
nsp7	Auxiliary factor nsp12.
nsp8	Auxiliar factor in RNA replication (primase).
nsp9	RNA binding protein[19].
nsp10	Forms a complex with nsp16 essential for immune evasion and RNA methylation capping[20].

nsp11	Unknown function.
nsp12	Functions as RNA dependent RNA polimerase[21].
nsp13	Helicase in SARS CoV and MERS CoV[22].
nsp14	Interacts with RTC, possibly as an error correction mechanism[23].
nsp15	Primary interferon supressor[24].
nsp16	Forms a complex with nsp10 essential for immune evasion and RNA methylation capping[20].

These first two ORFs are followed by the first structural protein, S protein[25]. S protein is essential for viral entry in the host's cell as it contains a class I fusion peptide that requires a cleavage to activate its fusion potential[26]. This protein is 1,273 amino acids (aa) long, starting with a signal peptide in residues 1-13 followed by two subunits, S1 and S2[27]. S1 is located in residues 14-685 and contains the Receptor Binding Domain (RBD) in residues 319-541 and receptor binding motif in residues 437-508. S2 is found in residues 686-1,273 and encompasses the fusion peptide in residues 788-806[28]. This protein shows 22 sites for polysaccharide attachment that may have an effect in protein folding as well as in the capabilities of evasion by the virus to the host's immune system[29].

The next ORF in SARS-CoV-2 genome encodes accessory proteins Orf3a and Orf3b[25]. Orf3a forms a dimeric structure that acts as a cation channel and may trigger cellular apoptosis[30]. This protein may also be responsible for immune evasion, as it is thought to stop the fusion of autophagosomes with lysosomes, incapacitating these from degrading the viral material[31]. Orf3b does not have any information about its structure available either in SARS-CoV-2 or another CoV. However, in SARS-CoV-2 it is probably a more effective Interferon repressor than its homologous in SARS-CoV[32].

The ensuing ORF codes another signature CoV's structural protein, E protein[25]. It's a transmembrane protein containing a small N-terminal domain, followed by a short transmembrane domain with at least one α-helix, and a C-terminal domain that constitutes a major part of the protein. Nevertheless, this protein's topology varies, as it adopts a N-exo C-endo topology in IBV, whereas in SARS-CoV and MHV it adopts a hairpin topology with N-endo and C-endo[33].The topology presented by this protein in SARS-CoV-2 is still unknown. This protein can oligomerize forming a pentamer and creating a pore to conduct ions through the membrane[34]. The pore function of this protein does not affect virus growth and replication, although viruses with ion conduction are more viable. However, this function does affect virus

pathogenicity via inflammatory response, pulmonary damage, and patient outcome[35]. E protein also plays a major role in virion assembly[36].

The next e ORF codes M protein[25]. This structural protein has 223 residues and interacts not only with itself to form a homodimer as well as with all the other structural proteins. Each monomer is composed of a short N-terminal ectodomain, followed by three Transmembrane Helices (TMH 1 to 3) and a C-terminal endodomain[37–39]. M protein plays a key role in viral replication as it interacts with all other structural proteins, and it is the most abundant protein in virions[40]. M protein in SARS-CoV interacts with S protein to keep it in Endoplasmic Reticulum Golgi Intermediate Compartment (ERGIC), incorporating S protein in new virions[41], with N protein to stabilise the N protein-RNA complex and RNA packaging[42] and with E protein to form the viral envelope[36,43]. Finally, this protein is also responsible for immune evasion as it's homologous in SARS-CoV suppresses the activation of nuclear factor kappa B, an important contributor for the immune and inflammatory responses[44].

The subsequent ORF codes Orf6 protein[25]. This protein in SARS-CoV-2 has an amino acid identity of 69% compared to the one from SARS-CoV[24]. Orf6 is an accessory protein, not vital for replication, but suppresses interferon activity, through a different mechanism than nsp1[45].

Following, there are two ORFS that code Orf7a and Orf7b proteins[25]. Orf7a is 121 aa long and it starts with a short N-terminal signal region, followed by an ectodomain like an immunoglobulin, a short transmembrane domain and finally an endoplasmic reticulum retention motif[46]. This transmembrane protein is thought to interact with S protein and with lymphocytes or leukocytes integrin receptors, which modulates cell targeting[47]. Orf7b is a small transmembrane peptide, only 43 residues long. It possesses a single transmembrane domain, a N-terminal domain in the endoplasmic reticulum lumen and a cytoplasmic C-terminal domain[48]. This protein forms multimers that interact with cellular transmembrane leucine zipper proteins, affecting the normal function of the host's cells. This accessory protein is also proposed as a possible cause for some of Covid-19 symptoms, such as heart arrhythmia, loss of smell, odor loss and lung or bowel complications.[48]

The posterior ORF codes Orf8 protein[25], a dimer bonded by a disulfide bond[49]. It can down-regulate Major Histocompatibility Complex Class I (MHC-I), which presents peptides related to different viral proteins in order to lead to the death of the host's infected cells impeding the viral replication through that cell's machinery[50]. Orf8 also activates pro-inflammatory pathways that lead to cytokine storms, a quick production of many cytokines[51] and this can lead to acute respiratory syndrome[52].

The final ORF codes the last structural protein, N protein[25]. It is a homodimer that encompasses two domains, N-terminal, and C-terminal domains. Both domains bind to

Ribonucleic Acids (RNA), however, whereas the N-terminal domain region is thought to move freely, the C-terminal domain forms the dimer[53]. While these domains are very organized, they are flanked by three stretches of unorganized coils: C-terminal domain by the C-tail, N-terminal domain by an N-arm and both by a linkage region. These are rich in positively charged residues and modulate the RNA binding activity of both C-terminal and N-terminal domains[54]. This protein plays a key role in CoVs life cycles, interacting with the RTC, stimulating RNA synthesis[55] It is also known to arrest the host's cellular cycle, by modulating the activity of Cyclin-dependent Kinase 2 and Cyclin-dependent Kinase 4, fundamental kinases for mitotic cylce's S phase progression[56].

In the N gene there is also an alternative ORF that codes accessory protein Orf9b[57]. This protein is 98 aa long forming homodimers through two adjacent β sheets. Within the homodimer exists a hydrophobic pocket capable of binding key lipids[58]. This protein plays a role in immune evasion, through its interaction with mitochondrial anchored receptor TOM70, key to import preproteins to the mitochondria[59] as well as to activate important antiviral pathways, limiting the interferon response by the cell[60].

1.3 SARS-CoV-2 Life Cycle

SARS-CoV-2's life cycle can be split into three main phases - Infection, Replication/Assembly and Release of the virus. The infection process starts with the recognition of the host's Angiotensin-Converting Enzyme 2 (ACE2) by S protein's RBD located in the S1 subunit, promoting viral envelope and cellular membrane fusion[6,61]. The fusion process starts with a S protein cleavage in two sites mediated by a protease, typically Transmembrane Protease Serine 2 (TMPRSS2)[62]. The first cleavage separates RBD and S protein's fusion domains leading to infection of the host's cell[2]. The second cleavage takes place in the S2' subunit and exposes the fusion peptide, leading to its insertion in the membrane. The second cleavage is followed by the formation of an antiparallel six-helix bundle, which results in fusion of the viral and host's cell membranes[28] and subsequent release of the viral genetic information.

After entering the cell, pp1a and pp1b are translated from ORF 1a and ORF 1b, respectively. The cleavage of these two peptides leads to nsps 1 to 16 (Table 1), with a wide range of functions that include the formation of the RTC, responsible for RNA replication[2]. Nsp12 plays a major role in viral replication, since it acts as RNA-dependent RNA polymerase (RdRp), in conjunction with cofactors[63] nsp7 and nsp8, acting as primase. These three proteins form the mini RTC, with the binding of one nsp7 and two nsp8 to one nsp12. During this process two nsp13 bind to the mini RTC complex. One nsp13 functions as an anchor and

allows for a second nsp13 to bind to the mini RTC while the second acts as an helicase during the viral genome replication[64].

After viral RNA production, viral proteins translation takes place by commanding the host's ribosomes. Translation of N protein occurs in the cytosol whereas translation of S, M and E, transmembrane proteins, takes place in the Rough Endoplasmic Reticulum[65]. The assembly and maturation of virions transpire in the Golgi Complex. These virions are then released from the cell, resorting to vesicles[66].

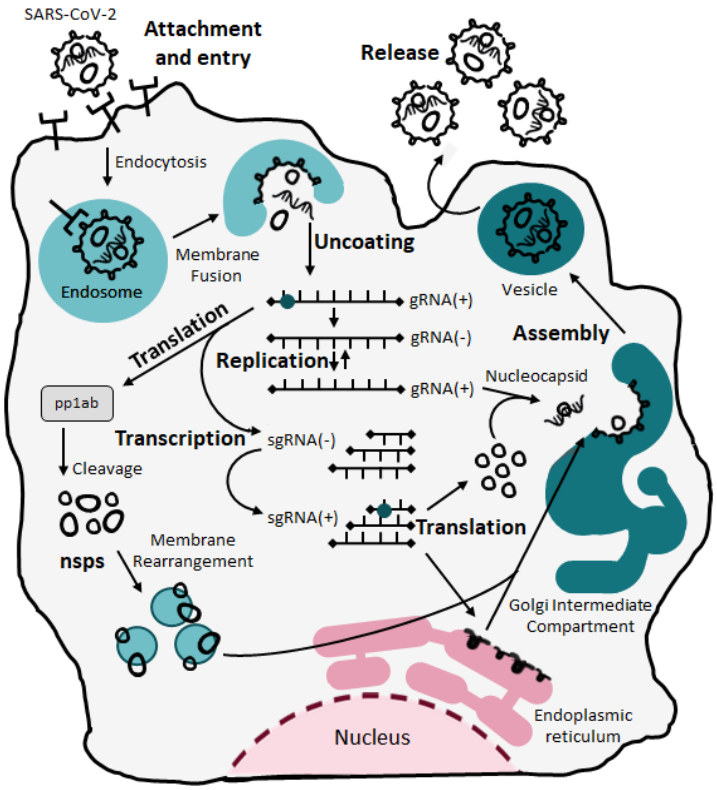


Figure 2 - SARS-CoV-2 life cycle.

1.4 Variants of Interest and Variants of Concern

According to WHO, some variants of SARS-CoV-2 represent a higher risk for public health, and these were classified in two tiers, according to its threat: Variants of Interest (VOI) and Variants of Concern (VOC). VOIs are variants that have acquired mutations predicted or known to enhance transmissibility and pathogenicity. VOCs add to the description of VOI a decrease in effectivity of treatment and diagnostic methods employed for SARS-CoV-2. To identify variants, a few solutions were proposed like Pango Lineage and Global Initiative on Sharing All Influenza Data (GISAID) Clades. Pango Lineage is a dynamic nomenclature proposition in which descent lineages show phylogenetic relations with the ancestor lineage

and emerge in a different geographic population. This nomenclature not only takes into account the possibility of rapidly increasing number of variants but also distinguishes between active variants, seen in the last month, unobserved variants, seen less than 3 months ago but not in the last month, and inactive variants, not seen in the last three months[67]. GISAID clades were formed by performing statistical clustering based on phylogenetic distance of sequences and smaller lineages were merged with major ones based on shared markers. The letters chosen to represent clusters were not generic, but instead chosen for this virus and according to markers considered for different clades[68].

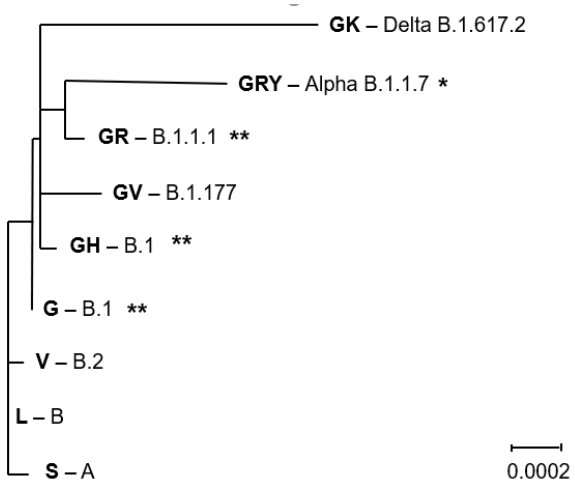


Figure 3 - Phylogenetic tree with GISAID clades S, L, V, G, G, GV, GR, GRY and K with the respective Pango Lineages. *- Clades with VOC, **- Clades with VOC and VOI. Retrieved from GISAID.

1.5 Membrane protein interactions

M protein is the most expressed protein during viral replication and assembly and its interactions with CoVs structural proteins play a key role in viral replication cycle of CoVs[69], acting as a scaffolding protein, interacting with the other structural proteins. The interaction with S protein in SARS-CoV-2 is essential to keep these viral proteins in the ERGIC during virion assembly, playing a pivotal role in the incorporation and glycosylation of spikes in the virions to be released[70]. The interaction between M and E proteins in SARS-CoV is sufficient for the formation of smooth virions, hence this interaction is absolutely necessary for viral replication[36], with M protein as the major constituent of the viral envelope[38]. M and N proteins are also known to interact in other CoVs not only to stabilize the N-RNA complex facilitating RNA packaging[42] but also to help stabilize the viral envelope[71]. M protein also interacts with itself, originating dimers that are thought to interact through all transmembrane helices as well as through the endodomain, forming most of the viral envelope[69,72]. However, these

interactions are not sufficient for the formation of stable virions and E protein is also required, as stated above.

A study in SARS-CoV's M protein, that shares 90.5% sequence identity and 90% homology with SARS-CoV-2 M protein[73], has found that substitutions in residues W19, W57, P58, W91, Y94 and F95 all resulted in a diminished viral replication. Authors postulated it to be caused by deficient M proteins, hypothesizing that these residues may play a pivotal role in M-M interactions[38]. M dimers may also interact, forming a matrix-like structure, probably through endodomain interactions[69,72,74].

CoVs' M protein has two different conformations, Mcompact and Mlong, named due to the elongation or compression of the endodomain, which has an impact on membrane curvature. Both conformations can interact with S protein and present viral spikes, but these have been observed clustering in regions rich in Mlong conformation. The Mlong conformation is hypothesized to be related with the presence of N protein and its interaction with M protein. This led to the hypothesis that Mlong conformation may be stabilized by the other three structural proteins[72].

However, due to being a transmembrane protein, there is no experimentally resolved structure available by either Nuclear Magnetic Resonance (NMR), X-ray crystallography or cryo-Electron Microscopy (cryo-EM)[75], as it is hard stabilize these kind of structures[76]. While there are computationally predicted structures for the monomer[73,77,78], neither membrane orientation or dimer structure has been previously predicted or experimentally resolved, to the extent of our knowledge.

1.6 Objectives

The main objective of this work was to study the impact of mutations in SARS-CoV-2's M protein homodimer. Due to the lack of information and homologous structures for this protein, it was necessary to obtain a prediction of the monomeric structure and predict the dimeric structure before the analysis of the impact of mutations. The overall pipeline can be observed in Figure 4.

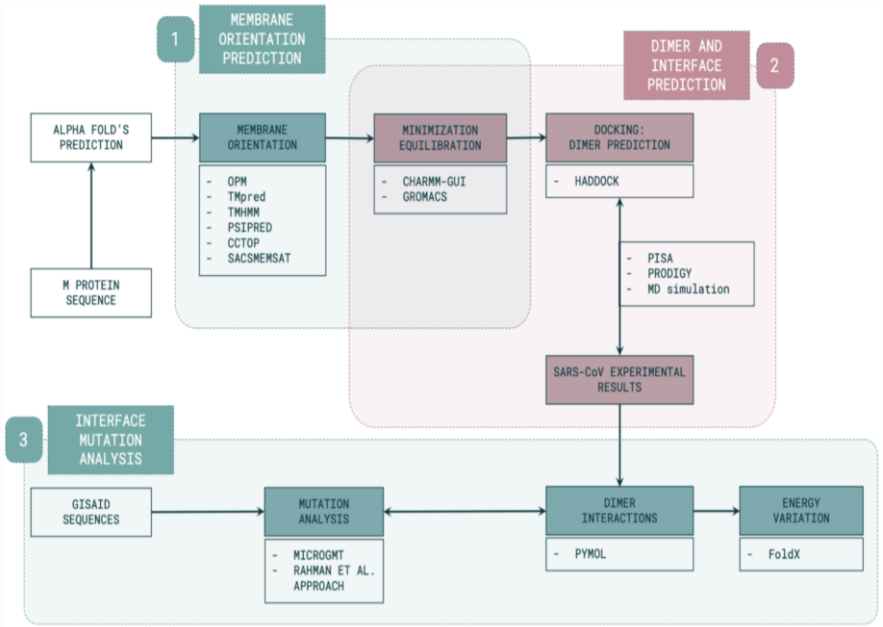


Figure 4 - Overview of the pipeline developed and employed. M protein structure was predicted by AlphaFOLD[24]. Membrane orientation was predicted with OPM[79], TMpred[80], TMHMM[81,82], PSIPRED[83,84], CCTOP[85,86] and SACSMEMSAT[87]. Protein-membrane systems were constructed with CHARMM-GUI[88] and minimization and equilibration were conducted using GROMACS[89,90]. M protein dimer was predicted with HADDOCK[91] and results were compared to SARS-CoV experimental data. Gene and protein mutations were analyzed with MicroGMT[92] and Rahman et al.[93] programs and energy variation of mutations in dimer interaction residues were calculated with FoldX[94].

1.6.1 Prediction of membrane orientation for the monomer

As stated before, M protein is a transmembrane protein capable of forming a homodimer. The AlphaFold[95] team has proposed a structure for the monomer of this protein. Hence, the first step of this work was to predict how this structure would insert and orient itself on the membrane, resorting to six different membrane orientation predictors.

1.6.2 Prediction of dimer structure and membrane orientation

Having predicted how the monomer would orient within the membrane, the next step was to obtain a prediction for M protein dimeric structure, resorting to docking, and deciding on which would be the membrane orientation used for further study as well as understanding what the interacting interface for the dimer would be. Finally, it was necessary to understand how the dimer would orient itself on the membrane to obtain the proposed structure, studying its behavior in Molecular Dynamics (MD) simulations.

1.6.3 Mutation detection for the M gene and protein

Having obtained the dimeric structure for M protein the conditions to start studying how the mutations would impact the stability were fulfilled. The first step in this study was to detect the mutations in M gene and protein to then further study them, resorting to two different software's, one developed for genomes and another for proteins.

1.6.4 Mutation impact analysis on protein stability

With the different identified mutations and the predicted structure of M protein dimer, we were in conditions to go through with the main objective of the study, studying how the different mutations would impact the structure of the protein dimer and how that could have no impact on stability, stabilize or destabilize M protein homodimer. For this an empirical force field was deployed to study the binding free energy difference ($\triangle\triangle G_{binding}$) between Wild Type (WT) and the mutated protein.

Chapter 2 - Methodology - the theoretical foundations

2.1 Membrane orientation and protein structure in Molecular Dynamics

Several components constitute the MD system and different cautions should be taken when building different systems, such as the existence of the tridimensional structure of the protein to study, the composition of the membrane bi-layer and the insertion of the protein in the membrane. If a protein possesses a solved tridimensional crystallographic structure there are a few possible artifacts to consider, usually caused by crystal lattice packing. When using experimental techniques to obtain tridimensional structures it can lead to artifacts as the membrane is not resolved. However, MD simulations can be performed if this is considered of this possibility[96]. In case there is not an experimental structure available three approaches are usually employed to predict the protein structure to study. Homology modeling considers the tridimensional structures of proteins that share at least 35-40% similarity and uses these as a template to predict the folded structure of a new protein. *Ab initio* methods only consider the characteristics of the amino acids present in the protein primary sequence and predict how they would interact after protein folding. Finally, fold recognition considers similar regions of a wide set of proteins and uses these regions as template for the folding of the similar region in the chosen protein[97].

When studying membrane proteins, it is of the utmost importance to determine their orientation within a lipidic membrane. Several predictors with different approaches can be employed for this task. Here are described six different web-based tools that were used: Orientations of Proteins in Membranes (OPM)[79], TMpred[80], Transmembrane Helix Markov Model (TMHMM)[81,82], Prediction of secondary structure (PSIPRED)[83,84], Consensus Constrained TOPology prediction (CCTOP)[85,86] and SACSMEMSAT[87]. OPM optimizes protein membrane position resorting to the interactions between protein and membrane and predicts how the protein structure is inserted in the membrane[79]. TMpred is a tool that predicts the regions of proteins inserted in membranes and its orientation[80]. TMHMM can predict the position of α-helices in both soluble and membrane proteins, also being able to distinguish between them[81,82]. PSIPRED uses scoring matrices that are position specific to predict the membrane secondary structures orientation in the membrane[83,84]. CCTOP takes advantage of both experimental and computational membrane topologies to predict membrane

orientation[85,86]. SACSMEMSAT resorts to transmembrane protein data to predict how the secondary structures from the protein are inserted in the membrane[87].

2.2 Molecular Dynamics

MDs are not a recent method, as the first report of a study using this technique dates back to 1957[98], and the first simulation of a protein system happened almost 20 years later[99]. MD has been a field with big developments, as software's are becoming more accessible and user friendly and the computational cost of simulations is decreasing due to advances in hardware and the possibility to use Graphical Processing Units (GPUs) to perform these tasks[100].

MD simulations calculate the positions of atoms in a biomolecular system throughout time, considering the force of interactions between atoms within the system under study. This information is then used to update each atom's position for each time step[96]. Simulations can have different levels of detail, more detailed simulations, such as atomic systems are more precise in the calculations made and, in the obtained trajectories. However, they are not suitable for simulating very large systems or very long simulations as they require immense computational power. On the other hand, coarse grain systems are a simpler representation of the system and are better suited for simulation of large systems and/or long run times. On the downside, MD calculations in these systems are more inaccurate, when compared with atomic systems[101]. MD simulations are used to study a wide range of structural characteristics that span from understanding a system reaction to different perturbations to understanding protein interactions with other proteins or different ligands. MD techniques are also employed to complement cryo-EM and X-ray crystallography to obtain the tridimensional structure of proteins. For example, as the experimental methods average the position of the atoms to determine the structure it is not possible to detect flexible regions in structure. MD simulations counter these shortcomings, further refining the obtained experimental structures.

The generation of the simulation system is essential to perform MD simulations. This is usually the biomolecule or biomolecules to study and other interacting molecules. In the case of transmembrane proteins it is also necessary to build and incorporate a membrane into the system that should have a similar constitution to the biological membrane involved *in vivo*, as different membrane compositions will have different interactions with the same protein[96]. After these components are built in initial conformation ions are added, and the system is surrounded by a solvent. The solvent can be represented explicitly or implicitly. Usually, the

chosen representation is explicit in the form of, for example, water molecules, even though this type of representation increases the system size even further[101].

The energy of a system is dependent on the position for the atoms that make up the constituents of the system, hence, this is a necessary calculation to perform MD simulations. Force fields are the mathematical expressions used to perform these calculations[102]. These aim to account for the different interactions that may happen between atoms. In classical MD no covalent bonds form or break, as such, covalent bonds are usually represented as springs. There are other more precise representations, however they present a steep computational cost. Two types of force fields exist nowadays. Classic force fields, also called addictive or non-polarizable, do not account for the movement of the electron cloud of the molecules and cannot interact with the environment as they are averaged out. New non-addictive or polarizable force fields are being developed to surpass this shortcoming, but they are still costly and difficult to develop[103]. Many addictive force fields (Equation 1) are available to use, and they take different parameterizations to better describe certain systems[104]. For example, some force fields are not well parameterized for lipids and are not advisable for systems with a membrane. In this case only force fields with adequate parameterization for lipids should be used, such as CHARMM36[105,106]. According to Equation 1, several intermolecular and intramolecular interactions are calculated with different parameterizations for different force fields. System potential energy (V) is calculated taking into account covalent bond's length ($\sum_{bonds}\frac{1}{2}\,k_b(r-r_0)^2$), covalent bond's angles ($\sum_{angles}\frac{1}{2}\,k_\theta(\theta-\theta_0)^2$), dihedral and improper dihedral angles ($\sum_{dihedrals}K_\phi[1+cos(n\phi-\delta)]$ and $\sum_{improper\ dihedrals}\frac{1}{2}\,k_\xi(\xi-\xi_0)^2$, respectively) and interactions between non bonded atoms ($\sum_{atom\ pairs\ i.j}\left(\frac{1}{4\pi\varepsilon_0}q_iq_jr_{ij}^{-1}+A_{ij}r_{ij}^{-12}+B_{ij}r_{ij}^{-6}\right)$)[104]. Choosing the right force field for the system submitted to MD simulation is paramount and may have a big impact on accuracy, as it may be better parameterized for certain types of molecules[97].

$$V=\sum_{bonds}\frac{1}{2}\,k_b(r-r_0)^2+\sum_{angles}\frac{1}{2}\,k_\theta(\theta-\theta_0)^2+\sum_{improper\ dihedrals}\frac{1}{2}\,k_\xi(\xi-\xi_0)^2$$

$$+\sum_{dihedrals}K_\phi[1+cos(n\phi-\delta)]+\sum_{atom\ pairs\ i.j}\left(\frac{1}{4\pi\varepsilon_0}q_iq_jr_{ij}^{-1}+A_{ij}r_{ij}^{-12}+B_{ij}r_{ij}^{-6}\right)$$

Equation 1 - General mathematical description of classical force fields.

When studying membrane proteins, building an adequate membrane for the system is essential. As such, the membrane composition will play a pivotal role in the MD simulation. When working with bi-layer membranes it is also important to confirm that the inner and outer leaflet have the adequate composition, which sometimes might be different between leaflets in some types of membranes[107], as this will also affect the simulation. Membrane orientation of the protein is also an important characteristic for the system. Some membrane builders can orient the protein in the membrane, although the most used MD simulation software's have auxiliary tools to perform this task[97].

As the systems used in MD simulations are small and do not represent entire biological systems, it is necessary to counter surface effects that may have a negative impact on the accuracy of the study. The most common way to surpass this problem is to use periodic boundary conditions, which surround the system's cube with replicas of the system that are also subjected to the simulation. Some atoms from other periodic images can be used for calculations if an atom from a certain periodic image leaves the simulation system, leading to an atom from a different periodic image entering the system being studied[108].

After having a system ready for the MD simulation, it is important to perform minimization and equilibration steps. The aim of the minimization step is to find the conformation for the protein and its side chains that maximizes the net attractive forces between all atoms and is important to avoid structural clashes during the simulation[104]. After minimization the system is static, however, the equilibration step will simulate experimental conditions defined for MD simulation of the system. The system heating prevents the atoms from being static and follows a velocity, according to Newton's second law of movement (Equation 2). This velocity is updated in certain time intervals until the desired temperature for the system is attained[104]. This heating step is done according to an ensemble, for example the canonical ensemble which keeps temperature, volume, and number of molecules for the system constant. This ensemble is kept by using a thermostat to control the temperature. It is also necessary to get pressure in experimental conditions, accomplished resorting to a barostat and according to a constant pressure ensemble, for example the isothermal-isobaric ensemble which keeps the number of molecules, temperature, and pressure for the system constant.

$$F_i = m_i a_i$$

Equation 2 - Newton's second law of movement. F_i is the force acting on the atom, m_i the mass of the atom and a_i the acceleration of the atom.

Finally, in the production phase, the atoms will interact according to the chosen force field for the MD simulation. During this phase the temperature and pressure are manipulated according to the objective of the study, resorting again to thermostats and barostats and

according to the adequate ensemble. Positions for the atoms are updated in a time step defined by the user, until the stipulated total simulation time is achieved.

The analysis of a MD simulation can provide insight into a wide range of system characteristics. Some information, such as the average structure throughout the simulation, the Root Mean Square Distance (RMSD) of residues in comparison to the initial positions, the Root Mean Square Fluctuation (RMSF) of residues in comparison to the initial structure, or the energy in the system are easily calculated with the information gathered about the system after MD simulation[104]. The occlusion and accessibility of certain protein residues can also be calculated and may be of interest when studying protein interactions. These are only some examples of what information can be retrieved from studies using MD simulation techniques.

2.2.1 CHARMM-GUI

Chemistry at HARvard Macromolecular Mechanics Graphical User Interface (CHARMM-GUI)[88,109–114] was developed with the aim of providing researchers with a graphical user interface tool capable of generating systems to be used in several contexts of several simulation techniques. CHARMM-GUI possesses several tools, such as a solvator, an implicit solvent modeler, a PEBQ solver and a membrane builder, as well as a PDB reader that converts PDB files into CHARMM readable files, the start point for the other tools mentioned[88]. Within these tools, in this work, only CHARMM-GUI´s membrane builder was used to create the systems necessary for the MD simulations. This tool generates all components necessary for MD simulations of a system as complex as one with a lipid bi-layer and a transmembrane protein[88,112].

To create this kind of system, the membrane builder needs a membrane orientation for the protein. In CHARMM-GUI this can be done automatically resorting to the OPM[79] database if the protein membrane orientation is available in the database, resorting to CHARMM-GUI membrane insertion tools or it can be provided by the user. Several lipid types are available, which makes this tool versatile as different types of membranes can be constructed for the simulation system[112]. If the system is being built for long MD simulations membrane orientation is not a major concern, as the protein will potentially equilibrate in the membrane throughout the simulation.

The membrane can be built through two different methods, insertion, or replacement. The insertion method creates a hole in the membrane to insert the protein, through weak repulsive radial forces[112]. This method is much faster at generating the membrane but is limited in protein shapes and system size. The replacement method surrounds the protein by

lipid-like pseudo atoms and randomly replaces them by the actual lipids, according to the chosen membrane composition[112]. This method is slower than the insertion, but it is more versatile in terms of system size and protein shape[112]. The membrane builder can generate files to use in several simulation softwares, such as GROningen MAchine for Chemical Simulations (GROMACS)[89,115].

2.2.2 GROMACS

GROMACS is a widely used free open-source software designed to perform MD simulations of biochemical systems, although it is also used for non-biological systems. It allows simulations with several force fields and computation to be run in Central Processing Units (CPUs) with GPU acceleration, making the calculations for the MD simulation of the system faster[116].

2.3 HADDOCK

The High Ambiguity Driven protein-protein DOCKing (HADDOCK) web server[91] is a user-friendly implementation of HADDOCK[117], in order to make its use simpler while providing computational resources for everyone to use within an adequate time frame[117]. HADDOCK's docking protocol can be divided into three main steps: randomization of orientations and rigid body energy minimization, semirigid simulated annealing in torsion angle space and final refinement in cartesian space with explicit solvent. Each step has a different scoring function, described in Equations 3 - 5. E_{vdw} represents the energy for van der Waals interactions, E_{elec} the energy for electrostatic interactions, E_{desol} the energy related with desolvation, E_{air} the energy related with distance restraints and BSA the buried surface área.

In the randomization of orientations and rigid body energy minimization step the two molecules chosen to dock are positioned far from each other and each molecule is randomly rotated around its center of mass. After this initial step, the system is submitted to four cycles in which each molecule is allowed rotational movements to minimize the system energy function. Following this, each protein is allowed rotational and translational movements to minimize the system energy (Equation 3).

$$\mathrm{HARDDOCK}_{score} - \mathrm{it}_0 = 0.01\,E_{vdw} + 1.0\,E_{elec} + 1.0\,E_{desol} + 0.01\,E_{air} - 0.01\,BSA$$

Equation 3 – Score function used in it0 step in HADDOCK.

$$\mathrm{HARDDOCK}_{score} - \mathrm{it}_1 = 1.0\,E_{vdw} + 1.0\,E_{elec} + 1.0\,E_{desol} + 0.1\,E_{air} - 0.01\,BSA$$

Equation 4 – Score function used in it1 step in HADDOCK.

$$\text{HARDDOCK}_{score} - \text{water} = 1.0\ E_{vdw} + 0.2\ E_{elec} + 1.0\ E_{desol} + 0.1\ E_{air}$$

Equation 5 - Score function used in water step in HADDOCK.

The structures with both molecules that best minimize the energy function are then submitted to semirigid simulated annealing in torsion angle space step. In this step three annealing refinement simulations are performed. In the first both proteins are still considered rigid bodies and its objective is to further optimize the orientation of both molecules. In the second annealing refinement, only the side chains of the interface residues are allowed to move. In the third and final annealing refinement both side chains and backbone residues are allowed to move, and this potentiates possible conformational rearrangements that minimize the energy function. Finally, the structures with both molecules obtained from the three annealing refinement simulations are subjected to a steep descent energy minimization.

The third and final step in this docking pipeline is final refinement in cartesian space with explicit solvent. In this step, water molecules are added to the system and three different steps of MD are performed. In the first step, the system is heated and only the side chains on the interface are allowed to move. In the second step temperature is constant and heavy atoms are restrained from moving. Finally, in the third step, the system is cooled, and all atoms are allowed movement except molecules backbone residues and atoms outside of the interface.

2.4 Mutation analysis

2.4.1 Genome based mutation detection

For mutation detection in genomes, we used MicroGMT, a mutation detection tool developed and optimized for tracking the mutations that are happening in the SARS-CoV-2 genome, although it can be used for other microbiomes. MicroGMT requires three inputs: a FAST-all (FASTA) reference genome sequence to analyze and compare, an annotation file for the reference genome and the sequences the user intends to submit to the pipeline in an assembled genome FASTA or raw FASTA format, both sequences containing formats. The reference genome and its annotations are not necessary for SARS-CoV-2 as they are made available with the software.

This tool works through a pairwise alignment between the reference sequence and each of the genomes presented by the user. However, to reduce the computational cost, there are a set of sequences already precomputed and only new sequences are submitted to this process. MicroGMT starts by creating a variant call format (vcf-format) variant file for each input that is enriched with annotations and generates a new annotated vcf-format file for each

entry. Following this, it presents summary tables for the mutations identified and these can be enriched with regional information of the input genomes.

2.4.2 Protein based mutation detection

For mutation identification in protein sequences *Rahman et al* proposed a tool in a previous study on SARS-CoV-2 S protein mutations[93]. This tool accepts the reference sequence within a multiple sequence alignment with the protein sequence the user intends to study. Then, through pairwise alignment the mutations are identified and saved in a text file. This method was also used in a different study to analyze SARS-CoV-2 N protein mutations[118].

2.4.3 FoldX

FoldX is an empirical force field developed to allow in depth studies of mutations effects in proteins from a PDB structure of the WT protein. FoldX applies a force field based on Equation 6. The letters (a to l) correspond to weights applied to different components necessary for the energy calculations. The rest of the terms present in Equation 3 account for several types of interactions observed in protein systems. ΔG_{vdw} represents the effect of van der Waals interactions and is calculated as a desolvation, scaling with the burial of atoms, and taking into account energy transferred from water to vapor. ΔG_{solvH} and ΔG_{solvP} are the contributors of the interactions with the solvent, ΔG_{solvH} represents the effect of the hydrophobic groups and ΔG_{solvP} the contribution of the polar groups. These are calculated like the van der Waals interactions, but without considering the energy transferred from water to vapor. There are also some water molecules that formed more than one lasting hydrogen bond with the protein, and these are represented by term ΔG_{wb}. The impact of the remaining hydrogen bonds is calculated through geometric parameters only and is represented by term ΔG_{hbond}. ΔG_{el} accounts for the impact of electrostatic interactions and is calculated by a simple Coulomb law. The burial of the bond in consideration leads to a scale of the dielectric constant in these calculations. The force field also considers the electrostatic interactions between different peptides in the applicable cases, represented by ΔG_{kon}. This term is calculated resorting to an empirical formula that describes the complex formation association rate. FoldX accounts for the entropic penalties of fixing the backbone of the protein in a certain conformation, ΔS_{mc}, as well as fixing the entropic penalties associated with fixing a side chain in a certain conformation, ΔS_{sc}. ΔS_{mc} is a result of the observation of a range of high resolution structures. ΔS_{sc} is calculated by adapting a set of predetermined entropy parameters. The final

term, ΔG_{clash}, accounts for the repulsive energy that is generated by the steric overlap of atoms.

$$\Delta G = a \cdot \Delta G_{vdw} + b \cdot \Delta G_{solvH} + c \cdot \Delta G_{solvP} + d \cdot \Delta G_{wb} + e \cdot \Delta G_{hbond} + f \cdot \Delta G_{el} + g \cdot \Delta G_{kon} + h \cdot T\Delta S_{mc} + k \cdot T\Delta S_{sc} + l \cdot \Delta G_{clash}$$

Equation 6 - Empirical force field applied by FoldX.

FoldX allows the study of different protein characteristics that may be impacted by mutations, among these, protein stability. The impact of mutations in protein stability can be studied resorting to Gibbs free energy (Equation 7) of folding calculations.

$$\Delta\Delta G_{binding} = \Delta G_{mutant} - \Delta G_{WildType}$$

Equation 7 - Free energy variation.

Chapter 3 - Methods

The work described in this dissertation encompasses three main components to study SARS-CoV-2 M protein dimeric structure. First, to predict membrane orientation for the monomer, followed by the prediction of the dimeric structure and its orientation in the membrane. Lastly, we analyzed mutations in possible interacting residues that may lead to the formation of the dimer and study its effect in the stability of the dimeric structure.

3.1 M protein monomer structure and membrane orientation

As stated previously, there are no experimentally resolved structures for SARS-CoV-2 M monomer. There are also no homologous proteins with available 3D structures, hence, for this work, the monomeric structure used was the one predicted and made available by the AlphaFold[95] team, resorting to their own tool[78]. However, the predicted structure only accounts for residues 11 to 203, from the 223 aa that are present in SARS-CoV-2 M protein. We decided to restrict our structure to these crucial regions as the missing residues are inherently disordered residues, more difficult to study and not essential for this study[78]. This monomeric structure was then used for prediction of how M protein monomer would insert itself in the membrane, resorting six different membrane orientation predictors: OPM[79], TMpred[80], TMHMM[81,82], PSIPRED[83,84], CCTOP[85,86] and SACSMEMSAT[87].

To perform initial minimization for each obtained membrane orientation we resorted to MD simulations. Before performing the MD, it was necessary to build the systems with membrane inserted protein to use in the simulations. For this we used CHARMM-GUI membrane builder[88]. Since the M protein is translated in the ER and viral assembly also takes place in this organelle, the membrane built was a replica of the lipid composition present in ER bilayer membrane made by POPC:POPE:PI:POPS:PSM:Cholesterol, with constitution, according to Table 2, and complemented with TIP3 waters, 0.9 M Na^+ and Cl^- ions.

Table 2 - Equilibration systems details for: i) M protein monomer; systems with M protein and membrane were equilibrated with OPM[79], TMpred[80], TMHMM[81,82], PSIPRED[83,84], CCTOP[85,86] and SACSMEMSAT[87] membrane orientation prediction; ii) M protein dimer, two M monomers with TMHMM[81,82] membrane orientation and membrane were equilibrated. Systems sizes, solvents (water and ions- Na^+ and Cl^-) and membrane lipids (POPC- phosphatidylcholine, Cholesterol, SAPI24 and SAPI25- phosphatidylinositol, POPE- phosphatidylethanolamine, POPS- phosphatidylserine and PSM- sphingolipid) constitution are listed herein.

		Monomer						**Dimer**
		OPM	**TMPRED**	**TMHMM**	**PSIPRED**	**CCTOP**	**SACSMEMSAT**	
System size	x	14.21	14.21	14.21	14.21	14.21	14.21	11.94
	y	14.21	14.21	14.21	14.21	14.21	14.21	11.94
	z	12.21	12.21	12.21	12.21	12.21	12.21	11.60
Solvent	H2O	48,949	48,918	48,954	48,935	49,005	48,935	31,319
	Na+	1,045	1,047	1,045	1,046	1,044	1,043	235
	Cl-	797	799	798	798	796	795	83
Membrane Lipids	POPC	370	370	370	370	370	370	245
	POPE	132	132	132	132	132	132	88
	SAPI24	30	30	30	30	30	30	26
	SAPI25	30	30	30	30	30	30	26
	POPS	18	18	18	18	18	18	12
	PSM	18	18	18	18	18	18	12
	Cholesterol	2	2	2	2	2	2	2

MD simulations for minimization of the systems for each membrane orientation were performed using GROMACS[89,115] and the CHARMM36 force field[105]. Minimization was performed with the steepest descent algorithm. Equilibration was performed by heating the systems with a Berendsen-thermostat at 310 K in the canonical ensemble (NVT) over 7 ns, keeping the pressure constant at one bar with isothermal-isobaric ensemble (NPT) over 20 ns with semi-isotropic pressure coupling algorithm[119]. The fast smooth Particle-Mesh Ewald

method was used to process long-range electrostatic interactions. No positional restraints were used during the minimization and equilibration process. To perform RMSD analysis between the positions of Cα residues in the AlphaFold structure and after minimization and equilibration, regarding the whole protein and only considering the transmembrane region, we resorted to Pymol, version 1.2r3pre[120].

3.2 M protein dimer and interface prediction

Not all orientations were selected for the steps that follow monomer membrane prediction, due to the disparity in results. Henceforth only OPM, TMpred and TMHMM membrane orientation predictions were used. To predict M protein dimeric structure, we resorted to the web-based docking tool HADDOCK[91] version 2.4. This tool uses experimental data as a basis to predict quaternary structures of proteins. To perform the docking step, we had to determine the monomer's active residues. To do this we used CPORT[121], an atomic level predictor for interacting residues between proteins. For the docking step only, transmembrane residues predicted by this tool were used as hot-spots.

Docking was then performed and non-crystallographic symmetry restrictions were imposed for TMH2 and TMH3, due to the high sequence identity and homology of M protein in SARS-CoV and SARS-CoV-2[73], and the knowledge that most homodimeric proteins are symmetric[122]. In the first phase of docking, rigid body docking, 5,000 structures were generated and in the following phase, semi-flexible refinement, 1,000 structures were obtained. Since this is a transmembrane protein the last phase, final refinement, was not executed.

Docking results were analyzed based on how well the bilayer membrane was aligned for both monomers. This led us to select 20 structures from the original 3,000. These 20 structures were then submitted to a web-based tool that takes analyses and models macromolecular interactions, Protein Interfaces Surfaces and Assemblies (PISA)[123]. They were further selected based on the interface interacting residues that were known from SARS-CoV M protein dimer. For that protein homologous residues W20, W58, P59, W92, Y95, F96 and C159 in SARS-CoV-2 play a pivotal role in the interactions that form the dimeric structure[38]. Two dimers were selected from this process and in order to select only one structure both were submitted to PROtein binDIng enerGY (PRODIGY)[124,125], a tool that predicts if the structures being studied are biologically relevant, using a random forest approach with structural features as input[126].

This structure was then used in MD simulations with systems built in a similar fashion to the described above for the minimization of the monomers, in CHARMM-GUI and according

to Table 2. Minimization and equilibration for the systems containing the dimers was also performed as described above for the monomer simulations using GROMACS. Following this initial step GROMACS was used to perform three independent MD simulations over 0.5 μs for the final structure. Simulations were produced with constant pressure and temperature (isothermal-isobaric ensemble). A Nose-Hoover thermostat with a time constant of 1 ps was used for temperature coupling and a semi-isotropic Parrinello-Rahman barostat with time constant 5 ps and compressibility of 4.5 × 10-5 bar-1 was used to ensure the pressure was constant. Linear constraint solver was used to limit hydrogen bonds and electrostatic interactions were again processed with fast smooth Particle-Mesh Ewald but with a cutoff of 1.2 nm.

Alpha Carbon (Cα) atoms were used for the calculation of Cross-Correlation Analysis (CCA), that tracks the movement of two or more time series data regarding one another, carried out with Bio3D R package[127]. GROMACS packages were then used to calculate both RMSD and RMSF for Cα atoms. GROMACS packages were also used for Solvent-Accessible Surface Area (SASA) analysis of the dimeric protein ($SASA_{complex}$), for each separate monomer ($SASA_{monomerA}$ and $SASA_{monomerB}$) and finally ΔSASA (Equation 8). These values were assessed on a residue ($_i$) based. Even though this last metric provides a good quantitative measure to show conformational changes after protein coupling relSASA (Equation 9) was also calculated to further explore how each residue behaves after the dimeric structure is formed. Possible interacting residues were detected by an in-house script that detected residues with a side chain in a 5 Å neighborhood of another residue's side chain for at least 90% of the simulation. For this a structure was retrieved every 2 ns of simulation from 300 ns until 500 ns in each replica.

$$\Delta SASA_i = SASA_{complex_i} - \left(SASA_{monomerA_i} + SASA_{monomerB_i}\right)$$

Equation 8 - ΔSASA calculation.

$$relSASA_i = \Delta SASA_i \,/\, SASA_{monomer_i}$$

Equation 9 - relSASA calculation.

3.3 M protein mutation analysis

The first step in mutation analysis was genome and protein retrieval, from the GISAID[128] database. Genomes were then analyzed using MicroGMT[92]. In this work, from genome mutations, only non-synonymous substitution mutations in the M gene region were considered. For the retrieved protein mutation detection, we resorted to an approach developed by Rahman *et al*.[93]. The reference sequence used for both tools was the first SARS-

CoV-2 sequenced genome (NC_045512.2) and its M protein (YP_009724393) and both tools were used with default parameterization. Only single residue substitution mutations were considered, and the further use of the mutation term will be to refer to this type of mutation. The mutations were detected in both protein sequences and M gene region of the genome and all detected mutations were used.

Finally, FoldX[94], an empiric force field was used to calculate Gibbs energy difference and understand the impact of the different mutations in the stability of M protein dimer, based on $\triangle\triangle G_{binding}$ (Equation 7), according to the differences in hydrophobic, polar, Van der Waals, hydrogen bonds and electrostatic interactions in between the mutated protein and the reference protein. Superior and inferior cutoff values of 0.5 kcal/mol and -0.5 kcal/mol, respectively, were defined to avoid considering $\triangle\triangle G_{binding}$ values that were close to 0 kcal/mol as an impact in protein stability.

Throughout this work residues considered as polar were R, N, D, C, E, N, H, K, S, T, Q and Y, residues considered as non-polar were A, G, I, L, M, F, P, W and V, residues considered as aromatic were F, W and Y and residues considered as non-aromatic were R, N, D, C, E, N, H, K, S, T, Q, A, G, I, L, M, P and V. Furthermore the original images presented throughout this dissertation were made resorting to Protein Imager[129], ggplot2 R package[130] and Bio3D R package[127].

Chapter 4 - Results

4.1 M protein monomer membrane orientation

M protein monomer membrane orientation is key to understanding the structure and interactions of this protein as well as its biological function. As stated in the previous chapter, six different predictions for M protein orientation were obtained resorting to six different tools: OPM[79], TMpred[80], TMHMM[81,82], PSIPRED[83,84], CCTOP[85,86] and SACSMEMSAT[87].

RMSD values were calculated to compare the different orientation predictions for the whole monomer and for only the TMH domains. RMSD values regarding the whole protein were 1.74 Å for OPM, 1.47 Å for TMpred, 1.42 Å for TMHMM, 2.50 Å for PSIPRED, 1.43 Å for CCTOP and 1.59 Å for SACSMEMSAT and the values regarding only the TMH sections of the protein were 0.40 Å for TMpred, 0.44 Å for SACSMEMSAT, 0.69 Å for OPM, 0.74 Å for TMHMM, 0.81 Å for CCTOP and 0.98 Å for PSIPRED (Figure 3).

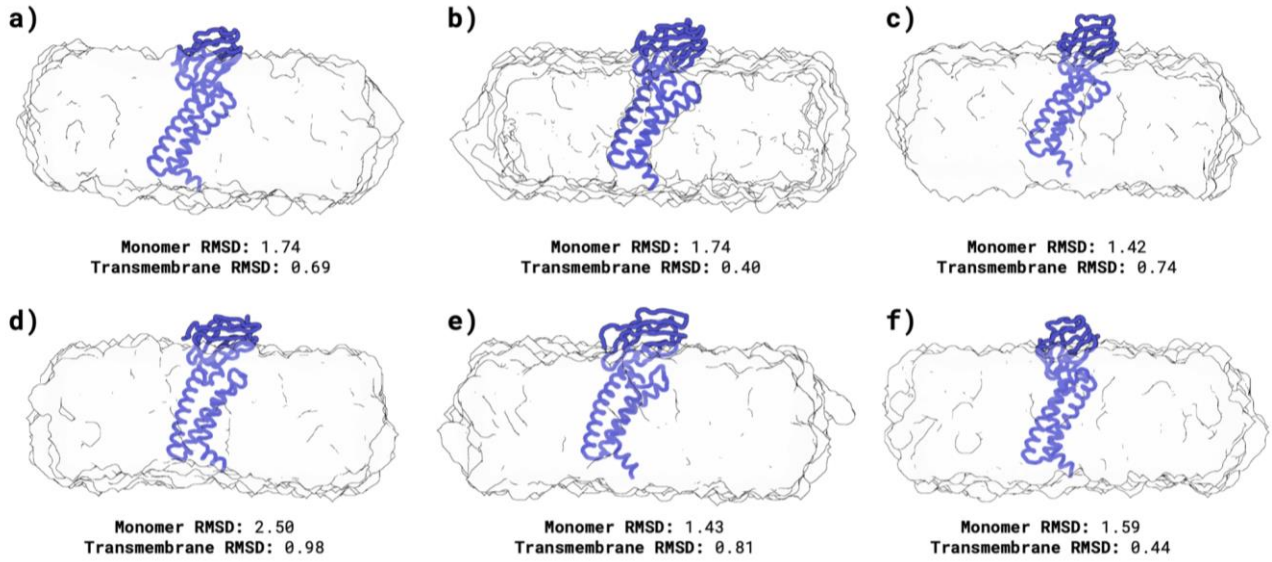


Figure 3 - SARS-CoV-2 M protein monomer transmembrane prediction: a) OPM[79], b) TMpred[80], c) TMHMM[81,82], d) PSIPRED[83,84], e) CCTOP[85,86] and f) SACSMEMSAT[87]. The RMSD was calculated between each structure after minimization and equilibration and the initial AlphaFold structure for the monomer.

PSIPRED was not used for further studies due to the significantly higher values of RMSD that its prediction presented when considering both TMH domains, as well as for the whole protein. SACSMEMSAT and CCTOP presented an arched shape in TMH1 after equilibration and minimization through MD simulations, that could have an impact on dimer stability (Figure 3). Hence, these two predictors were also not used anymore, although the

RMSD results were like the other predictors. From this step forward only the remaining three membrane orientation predictions - OPM, TMHMM and TMpred - were considered.

4.2 M protein dimer and interface prediction

Previously selected M protein monomer membrane orientation predictions, OPM, TMpred and TMHMM, were used to predict dimer tridimensional structure through docking using HADDOCK[91]. 20 structures that respected membrane orientation for both monomers in the dimeric structure were selected from 3,000 total structures proposed after docking.

Out of the selected structures 11 were from OPM, 4 from TMpred and 5 from TMHMM. Only two structures were selected through a preliminary interaction analysis, both with TMHMM membrane orientation predictions. The selection of these structures was made considering experimentally known M-M interactions that take place in SARS-CoV's TMH2, in residue P59, and TMH3, in residues W92, L93 and F96[38]. To select between the final two dimeric structures, we resorted to PRODIGY for its biological probability and predicted binding affinity metrics. The final selected dimeric structure chosen to use in the following work, presented 85.6% biological probability and -6.3 kcal/mol of predicted binding affinity and the discarded structure presented 74.8% of biological probability and -5.9 kcal/mol of binding affinity. According to the selected monomeric structure and its membrane orientation prediction obtained by TMHMM residues 11 to 19 belong to the ecto N-terminal domain, residues 100 to 203 to the endo C-terminal domain, residues 20 to 38 to TMH1, residues 46-70 to TMH2 and residues 76 to 100 to TMH3 (Figure 4).

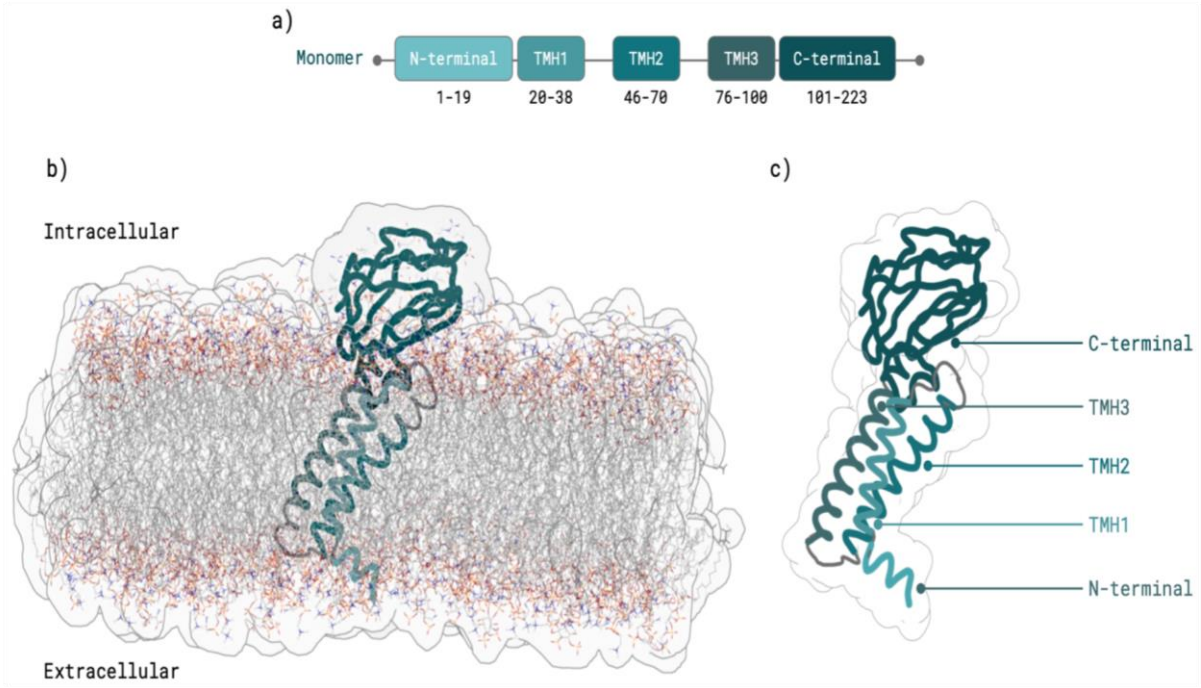


Figure 6 - SARS-CoV-2 M protein monomer. a) M protein domains predicted by TMHMM[81,82] membrane predictor. b) TMHMM[81,82] M protein monomer structure prediction after equilibration in membrane with ER membrane composition. c) M protein structure with domains highlighted.

The selected dimeric structure (Figure 5) revealed polar contacts between M protein and membrane lipids in residues K14, Y39, R42, N43, R44, F45, Y71, R72, W75, S94, R101, R107, W110, S173 and R174, which supports the prediction made for TMH domains (Figure5).

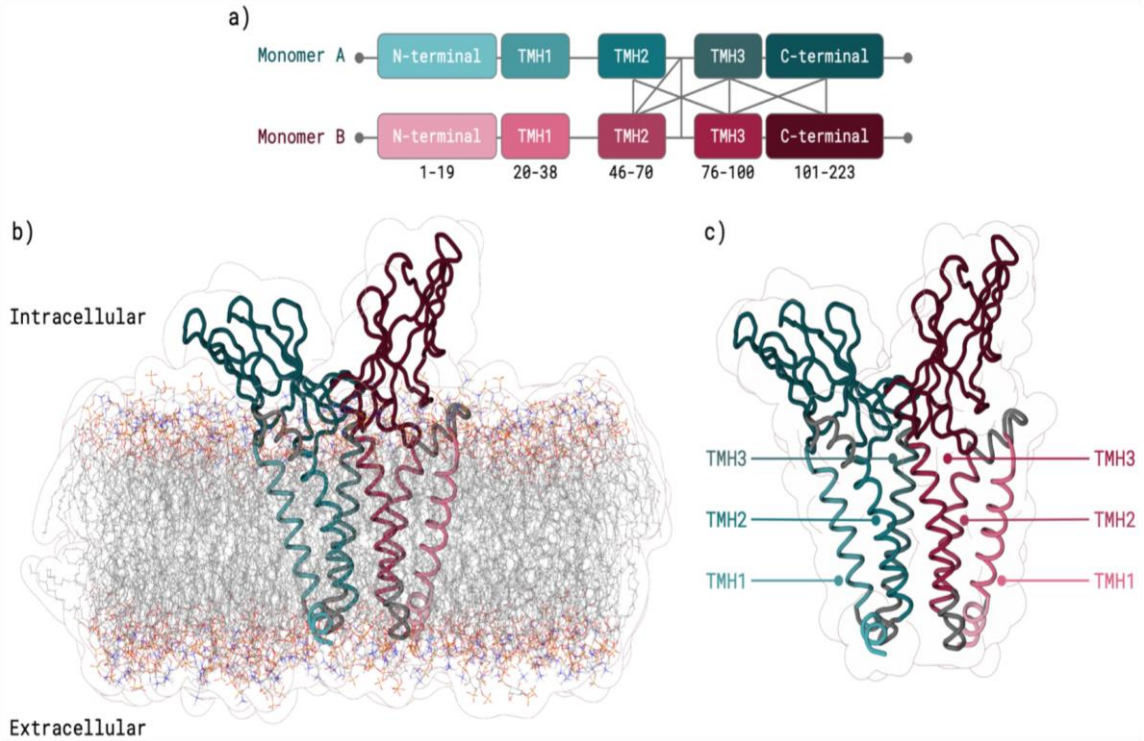


Figure 7 - SARS-CoV-2 M protein dimer HADDOCK[91] prediction using TMHMM[81,82] based monomers. a) Interaction representation between Monomer A (teal) and Monomer B (garnet) domains. b) M protein dimer within the membrane: Monomer A (teal), Monomer B (garnet). c) M protein dimer with TMH domains highlighted: Monomer A (teal), Monomer B (garnet).

After minimization and equilibration, the dimeric structure was submitted to three independent 0.5 μs MD simulation replicas. Both monomers behaved slightly differently during the MD simulation, according to the RMSD results (Figure 6). For monomer A TMH2 was rather less stable than TMH1 and TMH3 was the most stable domain (Figure 6A). TMH2 from monomer A interacted with monomer B. Domains with TMH were generally stable in monomer B, with the most unstable regions being the N-terminal and C-terminal domains (Figure 6B).

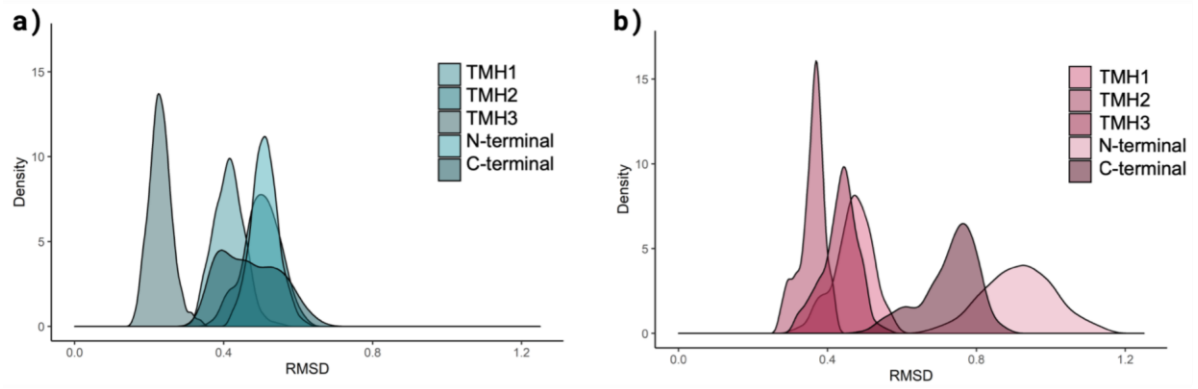


Figure 8 - RMSD results for MD simulations split by TMH1, TMH2, TMH3, N-terminal and C-terminal: a) Monomer A, and b) Monomer B.

Both monomers presented mostly similar RMSF results (Figure 7) in TMH domains that are largely composed of α-helices, showing a lower fluctuation when compared with other regions. This is particularly true for random coils, like the C-terminal domain that showed a higher fluctuation. Additionally, RMSF values in the N-terminal region is higher for B monomer and lower in the TMH2 region, corroborating the RMSD analysis. In CCA results it's possible to observe that TMH1 and TMH2 of opposite monomers showed a negative correlation, meaning they moved in opposite directions (Figure 8). TMH2 from monomer A showed a very weak, almost neglectable, positive correlation with TMH2 from monomer B. TMH2 from monomer A also showed a weak negative correlation with TMH3 from monomer B (Figure 8). TMH3 from each monomer showed a positive correlation between them. This leads us to hypothesize that TMH3 may play an important role in dimer formation.

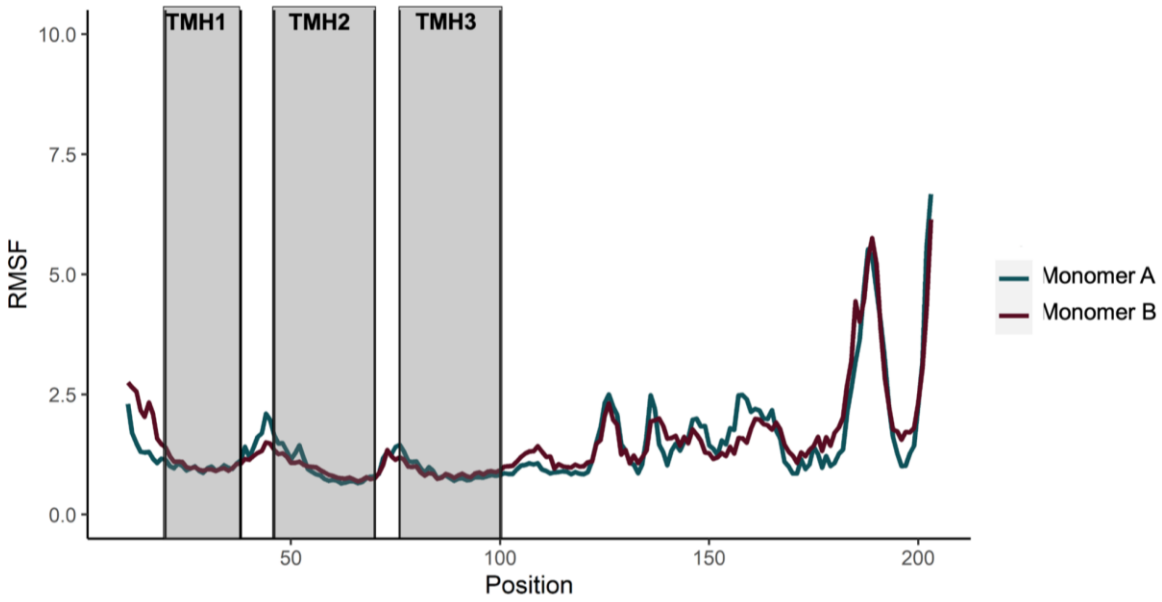


Figure 9- RMSF results for MD simulations split by monomers A and B. TMH1, TMH2 and TMH3 correspondent residues are highlighted as grey areas.

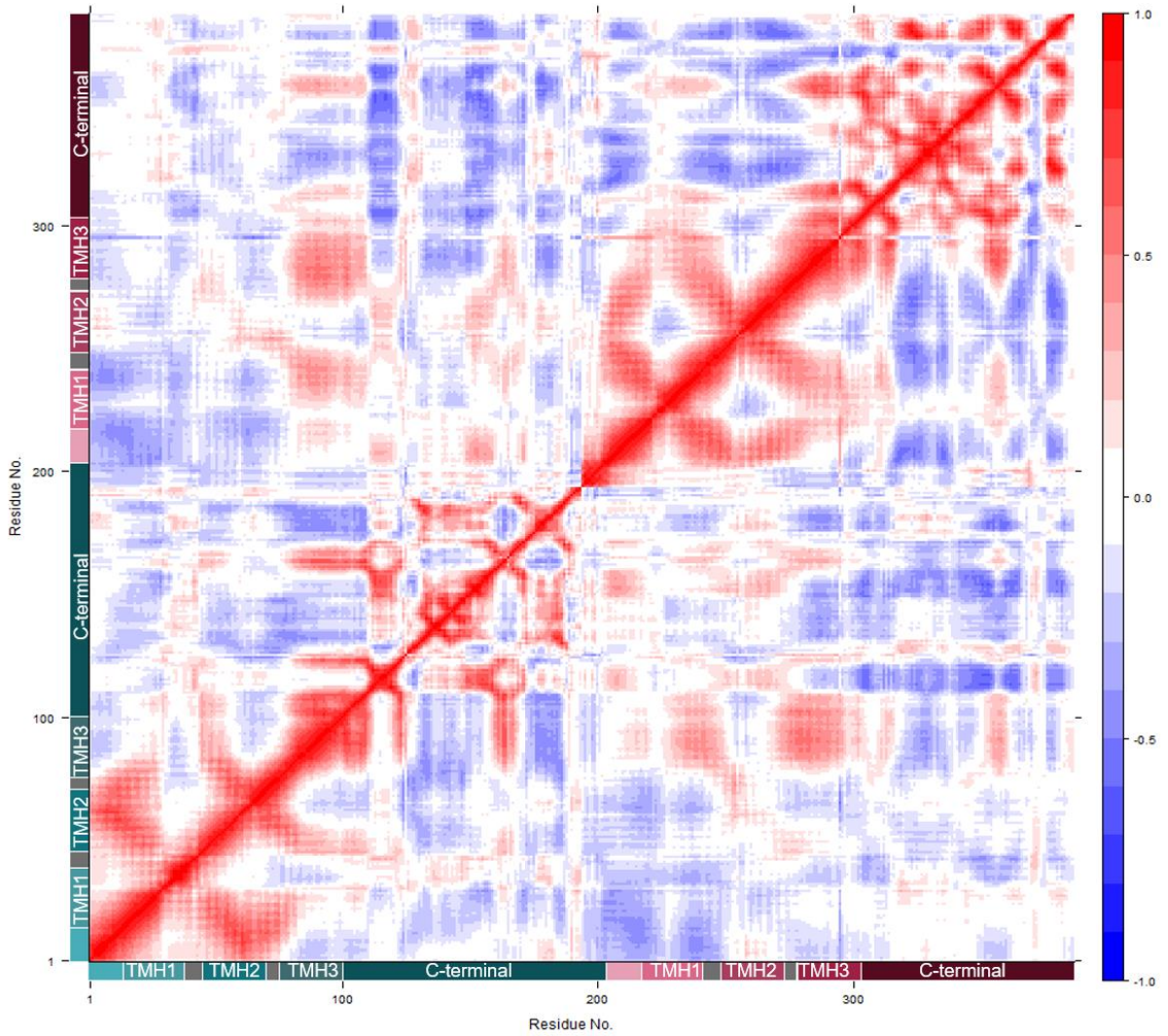


Figure 10 - Residue CCA between M protein dimer Monomer A and Monomer B.

Dimer interface in the selected structure was composed of 17 residues from monomer A - W55, P59, L62, V66, A69, V70, W75, I82, A85, W92, L93, F96, F100, F103, R107, M109 and F112 - and 21 residues from monomer B - P59, L62, V66, A69, V70, Y71, I82, A85, W92, L93, F96, I97, F100, F103, A104, R107, S108, M109, S111 and F112. These were considered as interface residues as they were within 5Å during at least 90% of the simulation time, with a total of 34 interactions identified (Table 3). The mean distance between Cα was 9.57 ± 0.60 Å, with the farthest being in the interaction between W92 from both monomers with a distance of 12.58 Å and the closest interaction was between V70 from both monomers with a distance of 5.25 Å. Out of the 34 total interactions, 12 were comprised by the same residue from each monomer - P59, V66, A69, V70, I82, L93, F96, F100, F103, R107, M109 and F112. These interactions were composed by 23 unique residues, 8 aromatic - Y71, W55, W75, W92, F96, F100, F103 and F112 -, 20 non polar - W55, P59, L62, V66, L67, A69, V70, Y71, W75, I82, A85, W92, L93, F96, I97, A104, F100, F103, M109 and F112 -, 3 polar - S108, S111 and R107

- and only one positively charged residue - R107. These interacting residues are in several regions of the protein, seven on the TMH2 domain - W55, P59, L62, L67, V66, A69 and V70 - , seven on the TMH3 domain - I82, W92, L93, I97, A85, F96, F100 -, seven on the C-terminal region - F103, A104, R107, S108, M109, S111, F112 -, and two on TMH2-TMH3 ectodomain loop region - Y71, W75.

Table 3 - SARS-CoV-2 M protein dimer interacting residues, using a prevalence time cut-off of 90% (all results were listed as mean values ± standard deviation).

Monomer A	ΔSASA A (Å²)	relSASA	Monomer B	ΔSASA B (Å²)	relSASA B	Percentage (%)	Cα distance (Å)
W55	69.00 ± 23.91	0.62 ± 0.16	L93	70.51 ± 16.96	0.67 ± 0.14	100.00	10.93 ± 0.65
V66	57.26 ± 11.49	0.80 ± 0.13	V66	58.18 ± 11.81	0.80 ± 0.10	100.00	7.11 ± 0.32
A69	15.96 ± 6.92	0.93 ± 0.15	V70	87.47 ± 11.74	0.88 ± 0.08	100.00	6.31 ± 0.41
V70	83.79 ± 16.05	0.81 ± 0.16	A69	14.78 ± 9.16	0.78 ± 0.43	100.00	6.70 ± 0.50
V70	83.79 ± 16.05	0.81 ± 0.16	V70	87.47 ± 11.74	0.88 ± 0.08	100.00	5.25 ± 0.51
W75	66.06 ± 38.00	0.33 ± 0.18	Y71	8.53 ± 40.92	0.09 ± 0.57	100.00	11.42 ± 0.74
I82	63.02 ± 18.62	0.65 ± 0.14	V70	87.47 ± 11.74	0.88 ± 0.08	100.00	8.62 ± 0.65
W92	64.08 ± 12.99	0.87 ± 0.10	W92	48.02 ± 16.01	0.76 ± 0.17	100.00	12.58 ± 0.49
L93	67.87 ± 23.73	0.62 ± 0.20	P59	11.83 ± 23.49	0.20 ± 0.53	100.00	8.62 ± 0.61
F96	67.33 ± 16.53	0.90 ± 0.09	F96	52.22 ± 15.81	0.89 ± 0.12	100.00	9.67 ± 0.65
F103	66.38 ± 15.37	0.88 ± 0.10	F103	78.66 ± 15.91	0.95 ± 0.07	100.00	10.79 ± 0.58
M109	89.25 ± 27.84	0.54 ± 0.14	F103	78.66 ± 15.91	0.95 ± 0.07	100.00	8.31 ± 0.44
P59	32.47 ± 25.51	0.50 ± 0.27	L93	70.51 ± 16.96	0.67 ± 0.14	99.67	09.01 ± 0.62
F112	76.09 ± 25.84	0.84 ± 0.08	F100	64.39 ± 26.02	0.50 ± 0.19	99.67	9.13 ± 0.49
V70	83.79 ± 16.05	0.81 ± 0.16	I82	45.82 ± 20.01	0.50 ± 0.20	99.34	9.08 ± 0.66
F100	83.51 ± 28.35	0.62 ± 0.14	F112	38.18 ± 31.03	0.52 ± 0.41	99.34	9.16 ± 0.55
W55	69.00 ± 23.91	0.62 ± 0.16	I97	22.90 ± 22.99	0.23 ± 0.24	99.01	11.45 ± 0.68
W92	64.08 ± 12.99	0.87 ± 0.10	L93	70.51 ± 16.96	0.67 ± 0.14	99.01	11.78 ± 0.60

R107	71.92 ± 29.68	0.36 ± 0.13	M109	86.63 ± 28.09	0.49 ± 0.14	99.01	7.72 ± 0.77
L62	24.35 ± 18.78	0.44 ± 0.30	L62	17.37 ± 14.70	0.34 ± 0.30	98.35	11.78 ± 0.44
M109	89.25 ± 27.84	0.54 ± 0.14	F100	64.39 ± 26.02	0.50 ± 0.19	97.36	8.9 ± 0.57
M109	89.25 ± 27.84	0.54 ± 0.14	A104	17.84 ± 14.34	0.26 ± 0.21	97.36	7.78 ± 0.50
I82	63.02 ± 18.62	0.65 ± 0.14	L67	14.01 ± 19.51	0.17 ± 0.24	96.37	8.69 ± 0.55
F103	66.38 ± 15.37	0.88 ± 0.10	S108	9.79 ± 12.38	0.28 ± 0.76	95.05	10.8 ± 0.67
F112	76.09 ± 25.84	0.84 ± 0.08	F103	78.66 ± 15.91	0.95 ± 0.07	94.72	11.13 ± 0.55
W75	66.06 ± 38.00	0.33 ± 0.18	V70	87.47 ± 11.74	0.88 ± 0.08	94.39	10.58 ± 0.65
F103	66.38 ± 15.37	0.88 ± 0.10	F112	38.18 ± 31.03	0.52 ± 0.41	94.39	10.28 ± 0.69
I82	63.02 ± 18.62	0.65 ± 0.14	V66	58.18 ± 11.81	0.80 ± 0.10	93.07	9.02 ± 0.49
W55	69.00 ± 23.91	0.62 ± 0.16	F96	52.22 ± 15.81	0.89 ± 0.12	93.07	11.66 ± 0.63
V66	57.26 ± 11.49	0.80 ± 0.13	A85	0.92 ± 6.38	0.00 ± 0.00	92.08	9.78 ± 0.44
F103	66.38 ± 15.37	0.88 ± 0.10	S111	-3.07 ± 3.85	0.00 ± 0.00	92.08	11.02 ± 0.76
A85	1.49 ± 6.45	0.00 ± 0.00	V66	58.18 ± 11.81	0.80 ± 0.10	91.42	9.62 ± 0.45
F100	83.51 ± 28.35	0.62 ± 0.14	F96	52.22 ± 15.81	0.89 ± 0.12	91.42	11.66 ± 0.72
M109	89.25 ± 27.84	0.54 ± 0.14	R107	53.28 ± 38.80	0.26 ± 0.18	91.42	8.96 ± 0.68

Out of the 34 total interactions considered, 9 were predicted to be between C-terminal domain residues of each monomer - F103-F103, M109-F103, R107-M109, M109-A104, F103-S108, F112-F103, F103-F112, F103-S111 and M109-R107 -, 6 between TMH2 domain of monomer A and TMH3 domain of B monomer - W55-L93, P59-L93, V70-I82, W55-I97, V66-A85 and W55-F96 -, 5 between the TMH2 domain of each monomer - V66-V66, A69-V70, V70-A69, V70-V70 and L62-L62 -, 5 between TMH3 domain of monomer A and TMH2 domain of monomer B - I82-V70, L93-P59, I82-L67, I82-V66 and A85-V66 -, 4 between the TMH3 domain of each monomer - W92-W92, F96-F96, W92-L93, and F100-F96 -, 2 between C-terminal domain of A monomer and TMH3 domain of monomer B -F112-F100 and M109-F100 -, 2 interactions between W75, located in the TMH2-TMH3 extracellular loop region, from monomer A and Y71, located in TMH2-TMH3 extracellular loop residue, and V70, located in

the TMH2 region, from monomer B and 1 interaction between F100, located at the TMH3 domain, from monomer A and F112, located in the C-terminal domain, from monomer B. From all the interactions only 12 - W59-L93, V66-V66, A69-V70, V70-A69, V70-V70, W75-Y71, I82-V70, W92-W92, L93-P59, F96-F96, F103-F103 and M109-F103 - were prevalent throughout 100% of the simulation (Table 3) and the regions which these residues occupy also showed low RMSF values.

Interactions within the same monomer were also detected, for monomer A hydrophilic interactions took place between residues L62-V66, V66-V69, W92-F96, F96-F100 and F103-R107 and π-π stack interactions occur between W92-F96 and F100-F112. Monomer B only presented hydrophilic interactions within itself, in residues L62-V66, V66-V69, L92-I97, F100-A104, A104-R107, S106-M107 and M107-F112, but π-π stack interactions were observed between residues W55-F100, W92-W92, F100-F112 and F103-F103, in monomer A and monomer B, respectively.

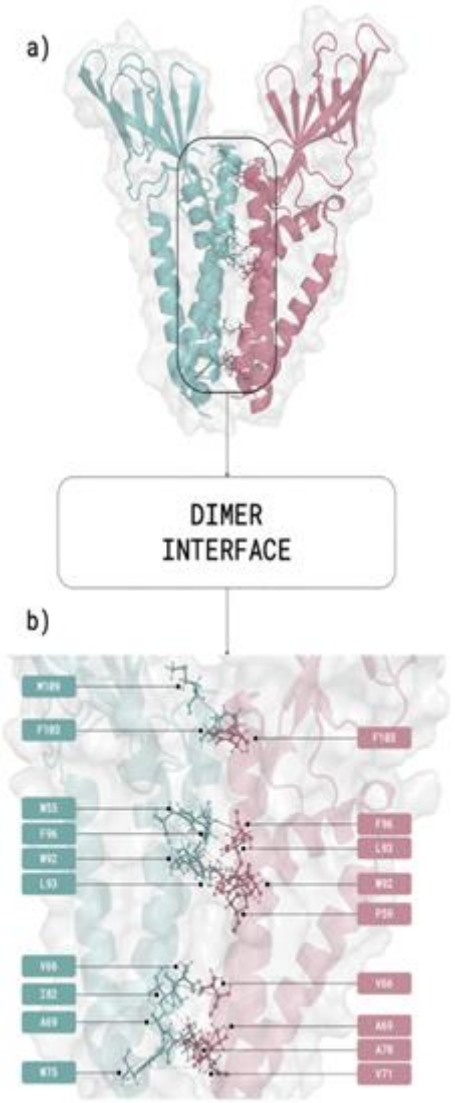


Figure 11 - **a)** SARS-CoV-2 M protein dimer via HADDOCK[30] prediction using TMHMM[20,21] based monomers with interfacial residues represented as sticks, and **b)** interface zoom-in featuring interfacial residues identified with the color code of teal for Monomer A and garnet for Monomer B.

4.3 M protein mutation analysis

4.3.1 Sequence analysis and exploration and mutation detection

Using the GISAID database 1,271,550 M protein sequences were retrieved, submitted from 10/01/2020 to 03/05/2021, collected in 180 different countries. 1,338,747 genomes with more than 29,000 bases and less than 5% missing values were also retrieved from the same database. From the retrieved genomes Clades S, G, GH and GR were composed of sequences with higher prevalence in North America, with GR also having a high prevalence

in Oceania (Figure 9). GV and GRY clades were predominantly in Europe and Clades O and L were distributed around the world without a region of significantly higher prevalence (Figure 9).

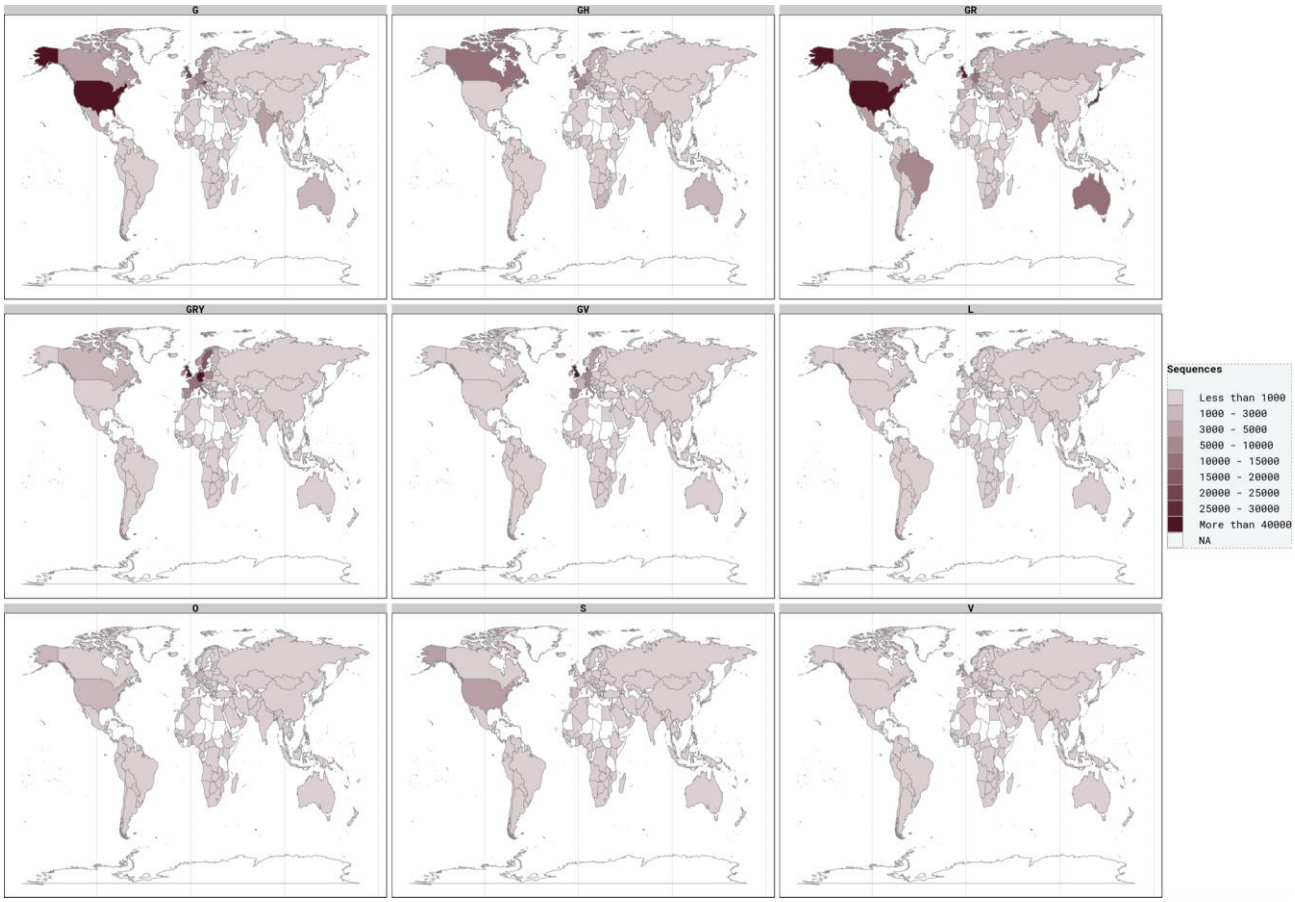


Figure 12 - GISAID data analysis by clades. Clade S includes variants A, clade V variants B.2, clade L variants B, clade G variants B.1, clade GH variants B.1.*, clade GV variants B.1.177, clade GR variants B.1.1.1 and clade GRY variants B.1.1.7.

91 different mutations were detected throughout 21,868 sequences in residues predicted as dimer interface in the previously described work. To assess the possible impact of each mutation in the dimeric protein we resorted to FoldX to predict binding free energy differences between mutated and WT proteins and studied them by analyzing the physio-chemical properties in the mutation (Figure 10).

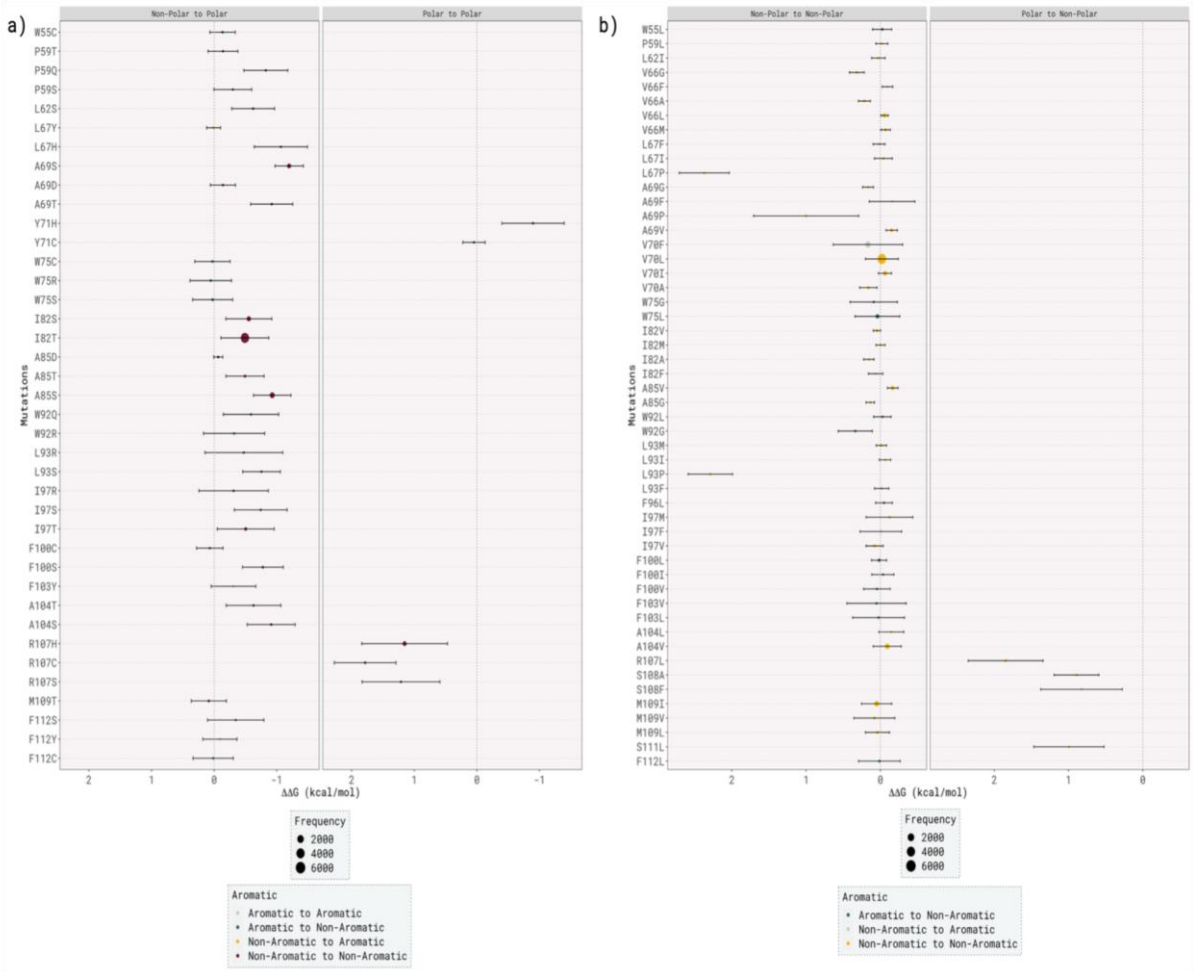


Figure 13 - $\triangle\triangle G_{binding}$ values of predicted interfacial residues split into a) Non-polar to Non-polar residues and Polar to Non-polar residues, and b) Non-Polar to Polar residues and Polar to Polar residues. Size corresponds to the number of times a SNP occurred in all 1,271,550 analysed sequences. Color represents the alteration from Aromatic to Aromatic (sage), Aromatic to Non-aromatic (teal), Non-aromatic to Aromatic (yellow) and Non-aromatic to Non-aromatic (garnet) (all the presented results are mean values ± standard deviation).

Regarding mutations in the considered interface residues 606 (2.77%) had $\langle\triangle\triangle G_{binding}\rangle$ values superior to 0.50 kcal/mol, 2,683 (12.27%) had $\langle\triangle\triangle G_{binding}\rangle$ values inferior to -0.50 kcal/mol and 18,579 (84.96%) had $\langle\triangle\triangle G_{binding}\rangle$ values between -0.50 and 0.50 kcal/mol. The $\langle\triangle\triangle G_{binding}\rangle$ was -0.01 ± 0.62 kcal/mol. Out of all considered mutations 55.53% were mutations from a non-polar residue to another non-polar residue, with a value for $\langle\triangle\triangle G_{binding}\rangle$ of 0.14 ± 0.49 kcal/mol in this mutation type; 41.68% were from a non-polar to a polar residue, with a value for $\langle\triangle\triangle G_{binding}\rangle$ of -0.42 ± 0.36 kcal/mol in this mutation type; 2.68% were from a polar residue to another polar residue, with a value for $\langle\triangle\triangle G_{binding}\rangle$ of 0.65 ± 1.07 kcal/mol in this mutation type; and 0.11% were from a polar residue to a non-polar residue, with a value for $\langle\triangle\triangle G_{binding}\rangle$ of 1.14 ± 0.48 kcal/mol in this mutation type (Figure 10).

The majority of the 91 studied mutations were from a non-aromatic residue to another

aromatic residue (90.01%), with a value for $\langle \triangle\triangle G_{binding} \rangle$ of 0.04 ± 0.77 kcal/mol in this mutation type; 7.27% were from a non-aromatic to an aromatic residue, with a value for $\langle \triangle\triangle G_{binding} \rangle$ of 0.09 ± 0.29 kcal/mol in this mutation type; 2.69% were from an aromatic residue to a non-aromatic residue, with a value for $\langle \triangle\triangle G_{binding} \rangle$ of -0.11 ± 0.30 kcal/mol in this mutation type; and 0.03% were from an aromatic residue to another aromatic residue, with a mean for $\langle \triangle\triangle G_{binding} \rangle$ of -0.20 ± 0.15 kcal/mol in this mutation type (Figure 10).

The most detected mutation was I82T, from a non-polar residue into a polar residue in the TMH3 domain. This mutation was detected in 6,316 (28.88%) of the retrieved sequences. The following most common mutation was V70L, from a non-polar residue into another non-polar residue in the TMH2 domain. This mutation was detected in 6,303 (28.82%) of the retrieved sequences. The third most occurring mutation was only detected in 1,455 of the retrieved sequences, so the two mutations described above were by far the most common ones (Annexed Table 1).

4.3.2 Single and co-occurring mutation analysis in different clades

The same mutation type analysis was also made for each clade (Figure 11) for both single mutated sequences, proteins only mutated in one residue, (Annexed Table 1) and sequences with multiple mutations, proteins with mutations in multiple residues (Annexed Table 2). GRY, a clade that contains VOC, was the most mutated in M protein, accounting for 36.69% of the mutations detected. For sequences of this clade the most frequent mutation was V70L, detected in 73.30% of clade GRY sequences with a $\langle \triangle\triangle G_{binding} \rangle$ value of -0.02 ± 0.22 kcal/mol. Within the GRY clade this mutation was detected co-occurring with M109L in 8 instances, with A104V in 2 instances and with A69F and A85V in 1 instance each, without any major identifiable energetic advantage, as $\langle \triangle\triangle G_{binding} \rangle$ was around 0 kcal/mol.

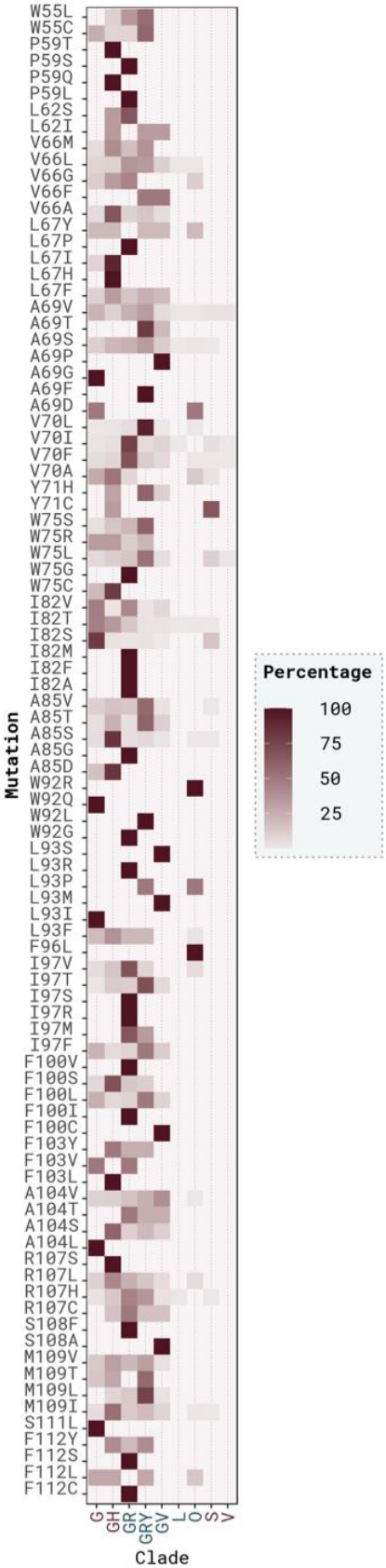


Figure 14 - Distribution across Clades of SARS-CoV-2 M protein sequences. Clade color is related to it encompassing VOC (garnet) and VOI (teal).

The following clade with the most mutations was GH, which also contains VOC and encompasses 21.25% of all sequences retrieved. In this clade the most common mutation was I82T and it accounted for 47.23% of clade GH mutated sequences with a $\langle \triangle\triangle G_{binding} \rangle$ value of -0.49 ± 0.38 kcal/mol. This mutation also had some co-occurring mutations but with a much lower frequency, as for example A85S that induced a more stable dimer interface with $\langle \triangle\triangle G_{binding} \rangle$ value of -1.47 ± 0.47 kcal/mol.

G clade, which contains both VOC and VOI, was the third with the most mutated sequences encompassing 19.05% of all mutated sequences. The most detected mutation in this clade was I82T, in 71.26% of the sequences contained in G clade with a $\langle \triangle\triangle G_{binding} \rangle$ value of -0.49 ± 0.38 kcal/mol. This mutation co-occurred with some other mutations in interface residues, such as with R107L in 4 instances, with V70F in 2 instances, with M109I in 2 instances, with V66M in 2 instances, with A85S in 2 instances and with R107H in 2 instances. None of those co-occurring mutations led to an increase in stability of the dimeric structure, as there were no significant changes in binding free energy.

GR clade, which contains both VOC and VOI, encompasses 17.27% of the retrieved sequences that contain mutations. In this clade the most detected mutation was V70F, present in 26.32% of sequences within the clade with a $\langle \triangle\triangle G_{binding} \rangle$ value of 0.17 ± 0.47 kcal/mol. A85S was detected co-occurring with V70F in 3 instances, with $\langle \triangle\triangle G_{binding} \rangle$ value of -0.72 ± 0.64 kcal/mol and A104V in 1 instance, with $\langle \triangle\triangle G_{binding} \rangle$ value of 0.10 ± 0.54 kcal/mol. Clades GV, S, O, L and V, that contain neither VOC nor VOI, are not as well represented, encompassing only 4.36% in GV, 0.90% in S, 0.38% in O and 0.05% for both L and V clades.

4.3.3 Single mutation analysis in VOC and VOI pango lineages

Analysis of single mutations was also performed for pango lineages considered VOC or VOI. Out of all the mutated sequences retrieved 8,951 (40.93%) were sequences within pango lineages that are considered as VOC and 2,757 (12.61%) were sequences within pango lineages considered as VOI. Most VOC sequences were encompassed in pango lineage B.1.1.7 (8,474 sequences, corresponding to 94.67% of VOC sequences). The most common mutation for this variant was V70L, detected in 6,136 sequences (72.41%) of B.1.1.7 lineage mutated sequences (Figure 12). In VOI the most common variant was pango lineage B1.525 (2,142 sequences, corresponding to 72.59% of VOI sequences), with the most frequent mutation, I82T, detected in 2,139 sequences (72.48%) (Figure 12).

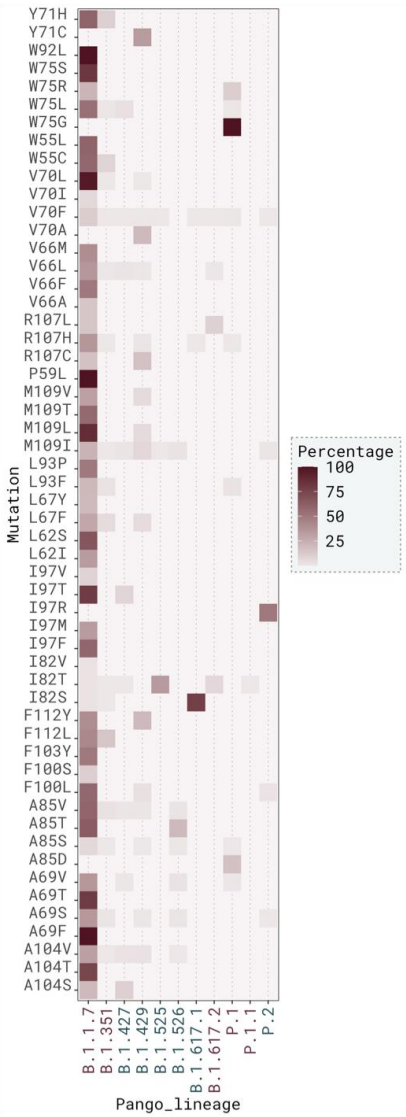


Figure 15 - Distribution across VOC (garnet) and VOI (teal) of SARS-CoV-2 M protein sequences.

As for impact in protein stability, mutations A69P, R107C, R107L, R107S and R107H have a negative impact in protein stability as they have $\langle \triangle\triangle G_{binding} \rangle$ values higher than 0.50 kcal/mol, with the R107H appearing in several VOC, such as lineages B1.1.7, B.1.351, P.1, and VOI lineage B.1.617.1, although R107H had a relatively low frequency compared to some other mutations. Contrarily to these mutations, other have a positive impact on protein stability, possessing $\langle \triangle\triangle G_{binding} \rangle$ values below -0.50 kcal/mol, such as A69S, A69T, A104S, A104T, I82S and I82T, with I82T commonly detected in VOC lineages B.1.617.2 and B.1.1.7 frequently but also in lineage P.1.1 and B.1.351 and in VOI lineage B.1.525. I82S was detected in VOCs lineages B1.1.7 and B.1.351 somewhat infrequently and in VOI lineage B.1.617.1. A69S was detected in VOC lineage B.1.1.7 commonly, in VOC lineage B.1.351, in VOI lineage B.1.526 rarely and in VOI lineage P.2 one only once. A69T was detected only in VOC lineage B.1.1.7 and finally A104S and A104T were only detected in VOC lineages B.1.1.7 two and

three times, respectively.

4.3.4 Interface residue analysis

As SASA values for interface residues are considerably lower when the complex is formed compared to other interface residues, solvent occlusion is known to be an important feature of Protein Protein Interactions (PPIs)[131–135]. Residues that were more commonly mutated showed higher values for ⟨△SASA⟩ and ⟨relSASA⟩. This is an indicator that these residues are being occluded after the complex formation, as shown by $SASA_{complex}$ values tending to zero, ⟨relSASA⟩ values closer to 1 and higher values for ⟨△SASA⟩. Furthermore, residue I82 had a ⟨△SASA⟩ value of 54.42 ± 13.27 Å2 and a ⟨relSASA⟩ value of 0.58 ± 0.12; residue V70 showed a ⟨△SASA⟩ value of 85.63 ± 11.01 Å2 and a ⟨relSASA⟩ value of 0.84 ± 0.10; and A69 displayed a ⟨△SASA⟩ value of 15.37 ± 4.26 Å2 and a ⟨relSASA⟩ value of 0.90 ± 0.12. These were the three most common mutated residues in the sequences retrieved.

Some other residues, with a relatively high mutation frequency, lost accessibility to the solvent, although they remained attainable after complex formation, such as mutation M109, with a ⟨△SASA⟩ value of 87.94 ± 14.46 Å2 and a ⟨relSASA⟩ value of 0.52 ± 0.07; A104, with a ⟨△SASA⟩ value of 13.69 ± 8.89 Å2 and a ⟨relSASA⟩ value of 0.21 ± 0.13; R107, with a ⟨△SASA⟩ value of 62.60 ± 21.03 Å2 and a ⟨relSASA⟩ value of 0.32 ± 0.10; and W75, with a ⟨△SASA⟩ value of 49.28 ± 20.33 Å2 and a ⟨relSASA⟩ value of 0.27 ± 0.11. While some residues only take place in one interaction, such as A104, and other in two, such as A69, R107 and W75, others may play a role in more, such as I82, that takes place in 4 interactions, M109, that takes place in 5 interactions and V70, that takes place in 6 interactions (Table 4).

Chapter 5 - Discussion and future work

M protein plays a pivotal role in SARS-CoV-2 viral assembly, in its viral envelope and plays a role in host immunity evasion. However, the interactions necessary for the formation of its homodimer are still not thoroughly studied. The work described in this dissertation started with AlphaFold's M protein monomer prediction, for which we predicted its membrane orientation through six different web-based tools. We chose the membrane orientation predicted by TMHMM, following the minimization and equilibration steps, due to its low RMSD values compared with the initial AlphaFold predicted structure, with no major changes in key transmembrane regions.

Although little information is available about SARS-CoV-2 M protein, the dimeric structure and it's interactions were previously experimentally studied for SARS-CoV's M protein, showing that the TMH1 region would be constituted by residues 15 to 37, TMH2 region by residues 50 to 72 and TMH3 by residues 77 to 99[38]. In this work, the first reliable study, to the extent of our knowledge, about SARS-CoV-2's M protein, we propose a membrane orientation in which TMH1 is composed of residues 20 to 38, TMH2 by residues 46 to 70 and TMH3 by residues 76 to 100, that align with the available results of SARS-CoV M protein.

SARS-CoV's M protein experimental study also proceeded to show that residues W19, W57, P58, W91, Y94, F95 and C158 were important for the formation of M protein dimer[38], leading us to believe that SARS-CoV-2 homologous residues W20, in the TMH1 domain, W58 and P59, in the TMH2 domain, W92, L93, Y95 and F96, in the TMH3 domain, and C159 in the endo-domain may also play a role in the interactions that result in the dimeric structure. This SARS-CoV study still lead to the hypothesis that residues C63 and C85 do not play a role in the interactions necessary for dimer formation[38], which suggests that the homologous residues C64, C86 in SARS-CoV-2 M protein may not have an impact in dimer formation.

All this information about the dimeric structure in SARS-CoV was considered for the docking steps, as detailed in sections 2.2 and 3.2. As also mentioned in section 3.2 a dimeric structure obtained from the monomer with membrane orientation predicted by TMHMM was chosen due to its complementary membrane orientation, as the structure allowed for both monomers to be correctly inserted in the same membrane while interacting to form a dimer. This structure was then inserted in a membrane with similar composition as ER membrane, as previously described, and this system was subjected to MD simulations. As the simulations showed different RMSD values for monomer A and B, it led us to believe that the conformational stability was slightly different. CCA values corroborate this hypothesis as TMH1 and TMH2 in monomer A are positively correlated and in monomer B the same domains

are negatively correlated. They also showed similar RMSF values, particularly in the TMH regions that allowed for the formation of stable interactions between monomers. We identified 38 interacting residues: 17 in monomer A and 21 in monomer B, responsible for the 34 interactions detected in M protein dimeric structure. As expected in transmembrane proteins such as M protein, most of these interactions - 73.53% - are between transmembrane residues. Out of the 34 interactions, 12 were present in 100% of the simulation time, including the interactions between homologous residues to the residues known to play a role in SARS-CoV M protein interactions - W92-W92, L93-P59 and F96-F96[38] - leading to the suggestion that these interactions might play key roles in the stabilization of the dimeric structure. However, these are not the only residues involved in interactions present throughout the whole simulation and experimental studies are necessary to further investigate the role of residues W55, V66, A69, V70, Y71, W75, I82, F103 and M109.

As for mutation analysis, out of 1,271,550 retrieved sequences, 21,868 (1.7%) sequences had mutations in M protein residues that were predicted as interacting residues in this work, leading us to believe that these are extremely conserved regions, and may be important for viral fitness[136]. We detected 91 different single mutations from the predicted interface residues, 12.27% of which had a positive impact in M protein dimer protein stability, such as: I82T, I97T, I82S, W92Q, L62S, A104T, I97S, L93S, F100S, P59Q, Y71H, A104S, A69T, A85S, L67H and A69S ($\triangle\triangle G_{binding}$ values of -0.49, -0.50, -0.55, -0.59, -0.62, -0.63, -0.74, -0.76, -0.78, -0.83, -0.90, -0.91, -0.92, -0.93, -1.07, -1.20 kcal/mol, respectively). All $\triangle\triangle G_{binding}$ values considered were lower than -0.50 kcal/mol. As I82T had a $\triangle\triangle G_{binding}$ value very close to the cut-off value, we thought it was pertinent to include it. A negative impact in dimer stability was noted in 2.77% mutations as they had a $\triangle\triangle G_{binding}$ higher than 0.50 kcal/mol. About 85% of the detected mutations in the considered region had no impact on protein stability.

As for the changes in the physio-chemical properties of the mutated residues, in more than half of mutations (55.53%) the residues remained non-polar, in 41.68% the residues change from non-polar to polar and most residues were non-aromatic and remained as such. As a transmembrane protein, M protein was expected to be rich in non-polar residues, especially in the regions that are inserted in the membrane, and as expected most of the residues predicted to interact were non-polar residues. This prevalence of non-polar residues in membrane proteins is consistent with previous studies[137].

On the other hand, 99.36% of the detected non-polar to polar mutations had negative $\triangle\triangle G_{binding}$ values, which may occur due to the increase in conformation stability as they can establish hydrogen bonds. From the homologous residues to the experimentally studied in SARS-CoV three mutations had a major impact in protein stability. Mutations L93S (-0.76 ±

0.30 kcal/mol) and W92Q (-0.59 ± 0.49 kcal/mol) had a positive impact, with $\triangle\triangle G_{binding}$ values lower than -0.5 kcal/mol and L93P, which had the second highest $\triangle\triangle G_{binding}$ value out of all mutations (+2.29 ± 2.29 ± 0.30 kcal/mol), had a negative impact, probably due to the destabilization typically caused by Prolines in α-helices. The rest of the mutations detected in residues P59, W92, L93 and F96 were neutral in dimer stability as their $\triangle\triangle G_{binding}$ values close to zero.

Mutations I82T and V70L were by far the most frequent mutations detected in this work, accounting for 28.88% and 28.82% of the detected mutations, respectively. I82T had a positive impact on protein stability as it showed a $\triangle\triangle G_{binding}$ of -0.49 ± 0.38 kcal/mol. V70L didn't impact structure stability as it displayed a $\triangle\triangle G_{binding}$ value of -0.02 ± 0.22 kcal/mol. These are both key residues in the formation of the dimer as both are responsible for interactions that were detected throughout the whole simulations and even interact between them with a mean distance of 8.62 ± 0.65 Å for IA82-VB70 and 9.08 ± 0.66 Å for VA70-IB82. The pair of residues also had ΔSASA values between 45 and 88 $Å^2$, which lead us to believe the interaction is protected, as the residues are occluded from the solvent when the complex is formed. Both residues also showed little fluctuation, according to their RMSF values.

The most represented clades in the mutated sequences retrieved were GRY (36.69%), which contains VOC, as well as GH (21.25%), G (19.06%) and GR (17.27%), that contain both VOC and VOI. The most frequent mutations, I82T and V70L, are present in sequences that are in clades that contain VOC and/or VOI with occurrences of 99.5% and 97.64%, respectively. This leads us to believe that M protein functions in the SARS-CoV-2 life cycle may be impacted by the mutations in the interface region of M protein dimer. V70L was the most detected mutation in VOC, identified in 6137 sequences, and most of the times (97.35%) this mutation was detected in sequences that belong to pango lineage B.1.1.7, a VOC encompassed by GRY clade.

Out of all sequences only 25 had more than one mutation and, of these, only 12 were in two residues that were predicted to play a role in interactions present throughout the whole simulations. The most frequent clades for co-occurring mutations were G in 27.45% of occurrences, GRY in 23.53% of occurrences, GH in 23.53% of occurrences and GR in 19.61% of occurrences. Co-occurrences were also detected, although in much lower frequency, in clades GV (3.92%) and S (1.96%). Most of the co-occurring mutations were present in VOC and VOI containing clades.

V70L was an interesting study subject, as although it co-occurred with other mutations in 9 instances, all of those belonged to GRY clade, that contains VOC. Although by itself the V70L mutation does not seem to have a major impact in dimer formation it may have an impact when co-occurring with other interfacial mutations as it is often found in several VOC.

M protein dimer in SARS-CoV-2 had little knowledge available other than its key role in the viral life cycle, with a critical role for viral assembly and envelope formation. In this work we studied how the protein would interact with the membrane and with itself, proposing not only a membrane orientation but also a dimeric structure and an analysis of its interface. This structure can be used for further studies of the protein and its interactions with other viral proteins as well as host proteins. The study of the extremely conserved interface also provided some insights on the key residues for dimer formation. This may be a steppingstone for further interface study through experimental methods and a starting point for drug development for this essential and well conserved protein.

Furthermore, an *in silico* pipeline was employed to study how the different mutations occurring in SARS-CoV-2 virus would impact the interface and stability of the dimeric structure. This may be important to analyze new variants that cause an increase in viral fitness and may provide insights in dangerous future mutations. To the best of our knowledge this is the first computational or experimental in-depth study on the dimeric structure and mutations of this well conserved protein. Experimental studies to validate proposed interface residues presented in this work are still necessary. The study of other M protein PPIs, namely with other structural proteins, the interface between proteins as well as the mutations that may occur in those regions are an interesting follow up to our work. All this information of protein interactions necessary for viral replication may be useful for drug development targeting the SARS-CoV-2 M protein.

Resulting publications from the dissertation

From the work developed and described in this dissertation a scientific article was submitted. The preprint is available at https://www.researchsquare.com/article/rs-702792/v1.

Besides this main publication, I was also involved in the following publications: https://doi.org/10.3390/biochem1020007 and https://doi.org/10.1016/B978-0-12-820472-6.00048-7.

References

1. Fung, T. S. & Liu, D. X. Human Coronavirus: Host-Pathogen Interaction. *Annu. Rev. Microbiol.* **73**, 529–557 (2019).
2. Malik, Y. A. Properties of Coronavirus and SARS-CoV-2. *Malays. J. Pathol.* **42**, 3–11 (2020).
3. Peiris, J. S. M., Yuen, K. Y., Osterhaus, A. D. M. E. & Stöhr, K. The severe acute respiratory syndrome. *N. Engl. J. Med.* **349**, 2431–2441 (2003).
4. Kraaij – Dirkzwager, M. *et al.* Middle East respiratory syndrome coronavirus (MERS-CoV) infections in two returning travellers in the Netherlands, May 2014. *Eurosurveillance* **19**, 20817 (2014).
5. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
6. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
7. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
8. Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* **27**, 325–328 (2020).
9. Kamitani, W. *et al.* Severe acute respiratory syndrome coronavirus nsp1 protein suppresses host gene expression by promoting host mRNA degradation. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 12885–12890 (2006).
10. Narayanan, K. *et al.* Severe acute respiratory syndrome coronavirus nsp1 suppresses host gene expression, including that of type I interferon, in infected cells. *J. Virol.* **82**, 4471–4479 (2008).
11. Wathelet, M. G., Orr, M., Frieman, M. B. & Baric, R. S. Severe acute respiratory syndrome coronavirus evades antiviral signaling: role of nsp1 and rational design of an

attenuated strain. *J. Virol.* **81**, 11620–11633 (2007).

12. Angeletti, S. *et al.* COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis. *J. Med. Virol.* **92**, 584–588 (2020).
13. Hussain, I. *et al.* Evolutionary and structural analysis of SARS-CoV-2 specific evasion of host immunity. *Genes Immun.* **21**, 409–419 (2020).
14. Khan, M. T. *et al.* SARS-CoV-2 nucleocapsid and Nsp3 binding: an in silico study. *Arch. Microbiol.* **203**, 59–66 (2021).
15. Qiu, Y. & Xu, K. Functional studies of the coronavirus nonstructural proteins. *STEMedicine* **1**, e39–e39 (2020).
16. Sakai, Y. *et al.* Two-amino acids change in the nsp4 of SARS coronavirus abolishes viral replication. *Virology* **510**, 165–174 (2017).
17. Moustaqil, M. *et al.* SARS-CoV-2 proteases PLpro and 3CLpro cleave IRF3 and critical modulators of inflammatory pathways (NLRP12 and TAB1): implications for disease presentation across species. *Emerging Microbes & Infections* vol. 10 178–195 (2021).
18. Xia, H. *et al.* Evasion of Type I Interferon by SARS-CoV-2. *Cell Rep.* **33**, 108234 (2020).
19. Sutton, G. *et al.* The nsp9 replicase protein of SARS-coronavirus, structure and functional insights. *Structure* **12**, 341–353 (2004).
20. Rosas-Lemus, M. *et al.* The crystal structure of nsp10-nsp16 heterodimer from SARS-CoV-2 in complex with S-adenosylmethionine. *bioRxiv* (2020) doi:10.1101/2020.04.17.047498.
21. Ziebuhr, J. The Coronavirus Replicase. in *Coronavirus Replication and Reverse Genetics* (ed. Enjuanes, L.) 57–94 (Springer Berlin Heidelberg, 2005).
22. Jia, Z. *et al.* Delicate structural coordination of the Severe Acute Respiratory Syndrome coronavirus Nsp13 upon ATP hydrolysis. *Nucleic Acids Res.* **47**, 6538–6550 (2019).
23. Ogando, N. S. *et al.* The Enzymatic Activity of the nsp14 Exoribonuclease Is Critical for Replication of MERS-CoV and SARS-CoV-2. *J. Virol.* **94**, (2020).
24. Yuen, C.-K. *et al.* SARS-CoV-2 nsp13, nsp14, nsp15 and orf6 function as potent

interferon antagonists. *Emerg. Microbes Infect.* **9**, 1418–1428 (2020).

25. Chan, J. F.-W. *et al.* Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg. Microbes Infect.* **9**, 221–236 (2020).
26. Bosch, B. J., van der Zee, R., de Haan, C. A. M. & Rottier, P. J. M. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. *J. Virol.* **77**, 8801–8811 (2003).
27. Huang, Y., Yang, C., Xu, X.-F., Xu, W. & Liu, S.-W. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol. Sin.* **41**, 1141–1149 (2020).
28. Xia, S. *et al.* Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. *Cell. Mol. Immunol.* **17**, 765–767 (2020).
29. Watanabe, Y., Allen, J. D., Wrapp, D., McLellan, J. S. & Crispin, M. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science* **369**, 330–333 (2020).
30. Kern, D. M. *et al.* Cryo-EM structure of SARS-CoV-2 ORF3a in lipid nanodiscs. *Nat. Struct. Mol. Biol.* **28**, 573–582 (2021).
31. Zhang, Y. *et al.* The SARS-CoV-2 protein ORF3a inhibits fusion of autophagosomes with lysosomes. *Cell Discov* **7**, 31 (2021).
32. Konno, Y. *et al.* SARS-CoV-2 ORF3b Is a Potent Interferon Antagonist Whose Activity Is Increased by a Naturally Occurring Elongation Variant. *Cell Rep.* **32**, 108185 (2020).
33. Yuan, Q., Liao, Y., Torres, J., Tam, J. P. & Liu, D. X. Biochemical evidence for the presence of mixed membrane topologies of the severe acute respiratory syndrome coronavirus envelope protein expressed in mammalian cells. *FEBS Lett.* **580**, 3192–3200 (2006).
34. Pervushin, K. *et al.* Structure and inhibition of the SARS coronavirus envelope protein ion channel. *PLoS Pathog.* **5**, e1000511 (2009).
35. Nieto-Torres, J. L. *et al.* Severe acute respiratory syndrome coronavirus envelope

protein ion channel activity promotes virus fitness and pathogenesis. *PLoS Pathog.* **10**, e1004077 (2014).

36. Mortola, E. & Roy, P. Efficient assembly and release of SARS coronavirus-like particles by a heterologous expression system. *FEBS Lett.* **576**, 174–178 (2004).
37. Arndt, A. L., Larson, B. J. & Hogue, B. G. A conserved domain in the coronavirus membrane protein tail is important for virus assembly. *J. Virol.* **84**, 11418–11428 (2010).
38. Tseng, Y.-T., Chang, C.-H., Wang, S.-M., Huang, K.-J. & Wang, C.-T. Identifying SARS-CoV membrane protein amino acid residues linked to virus-like particle assembly. *PLoS One* **8**, e64013 (2013).
39. Satarker, S. & Nampoothiri, M. Structural Proteins in Severe Acute Respiratory Syndrome Coronavirus-2. *Arch. Med. Res.* **51**, 482–491 (2020).
40. Mousavizadeh, L. & Ghasemi, S. Genotype and phenotype of COVID-19: Their roles in pathogenesis. *J. Microbiol. Immunol. Infect.* **54**, 159–163 (2021).
41. Fehr, A. R. & Perlman, S. Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol. Biol.* **1282**, 1–23 (2015).
42. Narayanan, K., Chen, C.-J., Maeda, J. & Makino, S. Nucleocapsid-independent specific viral RNA packaging via viral envelope protein and viral RNA signal. *J. Virol.* **77**, 2922–2927 (2003).
43. Regla-Nava, J. A. *et al.* Severe acute respiratory syndrome coronaviruses with mutations in the E protein are attenuated and promising vaccine candidates. *J. Virol.* **89**, 3870–3887 (2015).
44. Fang, X. *et al.* The membrane protein of SARS-CoV suppresses NF-kappaB activation. *J. Med. Virol.* **79**, 1431–1439 (2007).
45. Miorin, L. *et al.* SARS-CoV-2 Orf6 hijacks Nup98 to block STAT nuclear import and antagonize interferon signaling. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 28344–28354 (2020).
46. Zhou, Z. *et al.* Structural insight reveals SARS-CoV-2 ORF7a as an immunomodulating

factor for human CD14+ monocytes. *iScience* **24**, 102187 (2021).

47. Nizamudeen, Z. A. *et al.* Structural assessment of SARS-CoV2 accessory protein ORF7a predicts LFA-1 and Mac-1 binding potential. *Biosci. Rep.* **41**, (2021).
48. Fogeron, M.-L. *et al.* SARS-CoV-2 ORF7b: is a bat virus protein homologue a major cause of COVID-19 symptoms? doi:10.1101/2021.02.05.428650.
49. Flower, T. G. *et al.* Structure of SARS-CoV-2 ORF8, a rapidly evolving immune evasion protein. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
50. Harty, J. T., Tvinnereim, A. R. & White, D. W. CD8+ T cell effector mechanisms in resistance to infection. *Annu. Rev. Immunol.* **18**, 275–308 (2000).
51. Lin, X. *et al.* ORF8 contributes to cytokine storm during SARS-CoV-2 infection by activating IL-17 pathway. *iScience* **24**, 102293 (2021).
52. Wu, C. *et al.* Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. *JAMA Intern. Med.* **180**, 934–943 (2020).
53. Zeng, W. *et al.* Biochemical characterization of SARS-CoV-2 nucleocapsid protein. *Biochem. Biophys. Res. Commun.* **527**, 618–623 (2020).
54. Chang, C.-K., Hou, M.-H., Chang, C.-F., Hsiao, C.-D. & Huang, T.-H. The SARS coronavirus nucleocapsid protein--forms and functions. *Antiviral Res.* **103**, 39–50 (2014).
55. Cong, Y. *et al.* Nucleocapsid Protein Recruitment to Replication-Transcription Complexes Plays a Crucial Role in Coronaviral Life Cycle. *J. Virol.* **94**, (2020).
56. Surjit, M., Liu, B., Chow, V. T. K. & Lal, S. K. The nucleocapsid protein of severe acute respiratory syndrome-coronavirus inhibits the activity of cyclin-cyclin-dependent kinase complex and blocks S phase progression in mammalian cells. *J. Biol. Chem.* **281**, 10669–10681 (2006).
57. Xu, K. *et al.* Severe acute respiratory syndrome coronavirus accessory protein 9b is a virion-associated protein. *Virology* **388**, 279–285 (2009).

58. Meier, C. *et al.* The crystal structure of ORF-9b, a lipid binding protein from the SARS coronavirus. *Structure* **14**, 1157–1165 (2006).
59. Neupert, W. & Herrmann, J. M. Translocation of proteins into mitochondria. *Annu. Rev. Biochem.* **76**, 723–749 (2007).
60. Han, L. *et al.* SARS-CoV-2 ORF9b antagonizes type I and III interferons by targeting multiple components of the RIG-I/MDA-5-MAVS, TLR3-TRIF, and cGAS-STING signaling pathways. *J. Med. Virol.* **93**, 5376–5389 (2021).
61. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271–280.e8 (2020).
62. Bestle, D. *et al.* TMPRSS2 and furin are both essential for proteolytic activation of SARS-CoV-2 in human airway cells. *Life Sci Alliance* **3**, (2020).
63. Wang, Q. *et al.* Structural Basis for RNA Replication by the SARS-CoV-2 Polymerase. *Cell* **182**, 417–428.e13 (2020).
64. Yan, L. *et al.* Architecture of a SARS-CoV-2 mini replication and transcription complex. *Nat. Commun.* **11**, 5874 (2020).
65. Khade, S. M., Yabaji, S. M. & Srivastava, J. An update on COVID-19: SARS-CoV-2 life cycle, immunopathology, and BCG vaccination. *Prep. Biochem. Biotechnol.* **51**, 650–658 (2021).
66. Shereen, M. A., Khan, S., Kazmi, A., Bashir, N. & Siddique, R. COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. *J. Advert. Res.* **24**, 91–98 (2020).
67. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403–1407 (2020).
68. GISAID - Clade and lineage nomenclature aids in genomic epidemiology of active hCoV-19 viruses. https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/.
69. de Haan, C. A., Vennema, H. & Rottier, P. J. Assembly of the coronavirus envelope:

homotypic interactions between the M proteins. *J. Virol.* **74**, 4967–4978 (2000).

70. Boson, B. *et al.* The SARS-CoV-2 envelope and membrane proteins modulate maturation and retention of the spike protein, allowing assembly of virus-like particles. *J. Biol. Chem.* **296**, 100111 (2021).
71. He, R. *et al.* Characterization of protein-protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. *Virus Res.* **105**, 121–125 (2004).
72. Neuman, B. W. *et al.* A structural analysis of M protein in coronavirus assembly and morphology. *J. Struct. Biol.* **174**, 11–22 (2011).
73. Thomas, S. The Structure of the Membrane Protein of SARS-CoV-2 Resembles the Sugar Transporter SemiSWEET. *Pathog Immun* **5**, 342–363 (2020).
74. Neuman, B. W. *et al.* Supramolecular architecture of severe acute respiratory syndrome coronavirus revealed by electron cryomicroscopy. *J. Virol.* **80**, 7918–7928 (2006).
75. Almeida, J. G., Preto, A. J., Koukos, P. I., Bonvin, A. M. J. J. & Moreira, I. S. Membrane proteins structures: A review on computational modeling tools. *Biochim. Biophys. Acta Biomembr.* **1859**, 2021–2039 (2017).
76. Carpenter, E. P., Beis, K., Cameron, A. D. & Iwata, S. Overcoming the challenges of membrane protein crystallography. *Curr. Opin. Struct. Biol.* **18**, 581–586 (2008).
77. Mahtarin, R. *et al.* Structure and dynamics of membrane protein in SARS-CoV-2. *J. Biomol. Struct. Dyn.* 1–14 (2020).
78. Computational predictions of protein structures associated with COVID-19. https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19.
79. Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I. & Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.* **40**, D370–6 (2012).
80. HOFMANN & K. TMbase-a database of membrane spanning proteins segments. *Biol.*

Chem. Hoppe Seyler **374**, 166 (1993).

81. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182 (1998).
82. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
83. Buchan, D. W. A. & Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407 (2019).
84. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
85. Dobson, L., Reményi, I. & Tusnády, G. E. The human transmembrane proteome. *Biol. Direct* **10**, 31 (2015).
86. Dobson, L., Reményi, I. & Tusnády, G. E. CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res.* **43**, W408–12 (2015).
87. Jones, D. T., Taylor, W. R. & Thornton, J. M. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33**, 3038–3049 (1994).
88. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).
89. Bekker, H. *et al.* Gromacs-a parallel computer for molecular-dynamics simulations. in *4th International Conference on Computational Physics (PC 92)* 252–256 (World Scientific Publishing, 1993).
90. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43–56 (1995).
91. van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative

Modeling of Biomolecular Complexes. *J. Mol. Biol.* **428**, 720–725 (2016).

92. Xing, Y., Li, X., Gao, X. & Dong, Q. MicroGMT: A Mutation Tracker for SARS-CoV-2 and Other Microbial Genome Sequences. *Front. Microbiol.* **11**, 1502 (2020).
93. Rahman, M. S. *et al.* Comprehensive annotations of the mutational spectra of SARS-CoV-2 spike protein: a fast and accurate pipeline. *Transbound. Emerg. Dis.* (2020) doi:10.1111/tbed.13834.
94. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382–8 (2005).
95. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
96. Hollingsworth, S. A. & Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **99**, 1129–1143 (2018).
97. Shiref, H., Bergman, S., Clivio, S. & Sahai, M. A. The fine art of preparing membrane transport proteins for biomolecular simulations: Concepts and practical considerations. *Methods* **185**, 3–14 (2021).
98. Alder, B. J. & Wainwright, T. E. Phase Transition for a Hard Sphere System. *J. Chem. Phys.* **27**, 1208–1209 (1957).
99. McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of folded proteins. *Nature* **267**, 585–590 (1977).
100. Stone, J. E. *et al.* Evaluation of Emerging Energy-Efficient Heterogeneous Computing Platforms for Biomolecular and Cellular Simulation Workloads. *IEEE Int Symp Parallel Distrib Process Workshops Phd Forum* **2016**, 89–100 (2016).
101. Hospital, A., Goñi, J. R., Orozco, M. & Gelpí, J. L. Molecular dynamics simulations: advances and applications. *Adv. Appl. Bioinform. Chem.* **8**, 37–47 (2015).
102. González, M. A. Force fields and molecular dynamics simulations. *Éc. thémat. Soc. Fr. Neutron.* **12**, 169–200 (2011).
103. Marrink, S. J. *et al.* Computational Modeling of Realistic Cell Membranes. *Chem. Rev.*

119, 6184–6226 (2019).

104. Patodia, S. Molecular dynamics simulation of proteins: A brief overview. *J. Phys. Chem. Biophys.* **4**, (2014).

105. Huang, J. & MacKerell, A. D., Jr. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J. Comput. Chem.* **34**, 2135–2145 (2013).

106. Klauda, J. B. *et al.* Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J. Phys. Chem. B* **114**, 7830–7843 (2010).

107. Marquardt, D., Geier, B. & Pabst, G. Asymmetric lipid membranes: towards more realistic model systems. *Membranes* **5**, 180–196 (2015).

108. Allen, M. P. & Others. Introduction to molecular dynamics simulation. *Computational soft matter: from synthetic polymers to proteins* **23**, 1–28 (2004).

109. Brooks, B. R. *et al.* CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614 (2009).

110. Wu, E. L. *et al.* CHARMM-GUI Membrane Builder toward realistic biological membrane simulations. *J. Comput. Chem.* **35**, 1997–2004 (2014).

111. Jo, S., Lim, J. B., Klauda, J. B. & Im, W. CHARMM-GUI Membrane Builder for mixed bilayers and its application to yeast membranes. *Biophys. J.* **97**, 50–58 (2009).

112. Jo, S., Kim, T. & Im, W. Automated builder and database of protein/membrane complexes for molecular dynamics simulations. *PLoS One* **2**, e880 (2007).

113. Lee, J. *et al.* CHARMM-GUI Membrane Builder for Complex Biological Membrane Simulations with Glycolipids and Lipoglycans. *J. Chem. Theory Comput.* **15**, 775–786 (2019).

114. Lee, J. *et al.* CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.* **12**, 405–413 (2016).

115. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43–

56 (1995).

116. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25 (2015).

117. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).

118. Rahman, M. S. *et al.* Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein and its consequences. *J. Med. Virol.* **93**, 2177–2195 (2021).

119. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).

120. Schrödinger, L. L. C. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC (2017). *Google Scholar There is no corresponding record for this reference*.

121. de Vries, S. J. & Bonvin, A. M. J. J. CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PLoS One* **6**, e17695 (2011).

122. Blundell, T. L. & Srinivasan, N. Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 14243–14248 (1996).

123. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).

124. Vangone, A. & Bonvin, A. M. Contacts-based prediction of binding affinity in protein-protein complexes. *Elife* **4**, e07454 (2015).

125. Xue, L. C., Rodrigues, J. P., Kastritis, P. L., Bonvin, A. M. & Vangone, A. PRODIGY: a web server for predicting the binding affinity of protein-protein complexes. *Bioinformatics* **32**, 3676–3678 (2016).

126. Prodigy - prodigy_bp.Method. https://wenmr.science.uu.nl/prodigy/method.

127. Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A. & Caves, L. S. D.

Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **22**, 2695–2696 (2006).

128. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, (2017).
129. Tomasello, G., Armenia, I. & Molla, G. The Protein Imager: a full-featured online molecular viewer interface with server-side HQ-rendering capabilities. *Bioinformatics* **36**, 2909–2911 (2020).
130. Villanueva, R. A. M. & Chen, Z. J. ggplot2: Elegant Graphics for Data Analysis (2nd ed.). *Measurement* **17**, 160–167 (2019).
131. Preto, A. J. & Moreira, I. S. SPOTONE: Hot Spots on Protein Complexes with Extremely Randomized Trees via Sequence-Only Features. *Int. J. Mol. Sci.* **21**, (2020).
132. Moreira, I. S. The Role of Water Occlusion for the Definition of a Protein Binding Hot-Spot. *Curr. Top. Med. Chem.* **15**, 2068–2079 (2015).
133. Munteanu, C. R. *et al.* Solvent accessible surface area-based hot-spot detection methods for protein-protein and protein-nucleic acid interfaces. *J. Chem. Inf. Model.* **55**, 1077–1086 (2015).
134. Martins, J. M., Ramos, R. M., Pimenta, A. C. & Moreira, I. S. Solvent-accessible surface area: How well can be applied to hot-spot detection? *Proteins* **82**, 479–490 (2014).
135. Moreira, I. S., Ramos, R. M., Martins, J. M., Fernandes, P. A. & Ramos, M. J. Are hot-spots occluded from water? *J. Biomol. Struct. Dyn.* **32**, 186–197 (2014).
136. Majumdar, P. & Niyogi, S. SARS-CoV-2 mutations: the biological trackway towards viral fitness. *Epidemiol. Infect.* **149**, e110 (2021).
137. Matos-Filipe, P. *et al.* MENSAdb: a thorough structural analysis of membrane protein dimers. *Database* **2021**, (2021).

Annexes

Annex Table 1 **-** Analysis for each detected mutation in the predicted interface residues for SARS-CoV-2 M protein. This table describes the frequency of each mutation (Frequency), their mean $\triangle\triangle G_{binding}$ values ($\triangle\triangle G$), RMSF for each original residue (RMSF), solvent-accessible surface area (SASA) for each original residue in the complex ($SASA_{cpx}$), SASA for each original residue in the monomer ($SASA_{mon}$), ΔSASA for each original residue (ΔSASA), relative SASA for each original residue ($_{rel}$SASA), the number of interactions each original residues establish (Interactions) and the distribution of the mutation across the GISAID clades (Clade) (all the presented results are mean values ± standard deviation).

		$\Delta\Delta G$	RMSF	$SASA_{cpx}$	$SASA_{mon}$	$\Delta SASA$	$relSASA$		Clade (%)									
Mutation	**Frequency**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Interactions**	L	S	V	G	GH	GR	GV	GRY	O
I82T	6316	-0.49 ± 0.38	91.22 ± 19.94	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.06	0.24	0.00	47.02	34.74	14.06	1.71	1.82	0.35
V70L	6303	-0.021 ± 0.22	86.27 ± 18.53	15.82 ± 13.93	101.45 ± 11.56	85.63 ± 11.01	0.84 ± 0.10	6	0.00	0.00	0.00	0.68	1.08	4.43	0.21	93.30	0.30
V70F	1455	0.17 ± 0.47	86.27 ± 18.53	15.82 ± 13.93	101.45 ± 11.56	85.63 ± 11.01	0.84 ± 0.10	6	0.00	0.27	0.20	4.95	7.90	68.32	6.25	11.62	0.49
A85S	1215	-0.93 ± 0.30	79.81 ± 13.46	6.65 ± 4.25	7.85 ± 4.73	1.21 ± 1.56	0.14 ± 0.17	2	0.00	0.25	0.00	4.44	85.03	3.62	1.15	6.26	0.24
M109I	1005	0.05 ± 0.20	118.10 ± 20.29	80.28 ± 22.41	168.22 ± 20.18	87.94 ± 14.46	0.52 ± 0.07	5	0.00	0.10	0.00	4.38	54.03	13.03	8.56	19.50	0.40
A104V	959	-0.09 ± 0.19	105.19 ± 27.34	48.72 ± 14.21	62.41 ± 13.29	13.69 ± 8.89	0.21 ± 0.13	1	0.00	0.00	0.00	9.28	10.22	16.16	39.31	24.82	0.21
I82S	712	-0.55 ± 0.37	91.22 ± 19.94	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.00	14.75	0.00	80.20	2.25	0.98	0.14	1.68	0.00

A69S	574	-1.12 ± 0.23	83.55 ± 16.47	1.59 ± 3.05	16.96 ± 6.77	15.37 ± 4.26	0.90 ± 0.12	2	0.17	0.17	0.00	9.93	20.38	23.34	11.85	33.28	0.88
R107H	566	1.15 ± 0.68	122.51 ± 37.92	134.74 ± 27.60	197.33 ± 18.99	62.60 ± 21.03	0.32 ± 0.10	2	0.35	0.18	0.00	1.77	13.78	45.23	2.65	36.04	0.00
V70I	526	-0.06 ± 0.09	86.27 ± 18.53	15.82 ± 13.93	101.45 ± 11.56	85.63 ± 11.01	0.84 ± 0.10	6	0.19	3.80	0.19	1.33	3.42	76.81	8.75	5.51	0.00
V66L	430	-0.06 ± 0.05	72.96 ± 15.00	14.52 ± 8.56	72.23 ± 9.42	57.72 ± 5.87	0.80 ± 0.07	4	0.47	0.00	0.00	8.37	9.77	36.74	10.00	33.95	0.70
W75L	382	0.04 ± 0.30	134.37 ± 26.17	136.35 ± 44.73	185.63 ± 24.99	49.28 ± 20.33	0.27 ± 0.11	2	0.00	10.21	1.05	6.28	10.99	14.40	3.93	53.14	0.00
A85V	352	-0.17 ± 0.07	79.81 ± 13.46	6.65 ± 4.25	7.85 ± 4.73	1.21 ± 1.56	0.14 ± 0.17	2	0.00	0.28	0.00	10.23	16.48	14.20	2.27	56.53	0.00
A69V	158	-0.15 ± 0.08	83.55 ± 16.47	1.59 ± 3.05	16.96 ± 6.77	15.37 ± 4.26	0.90 ± 0.12	2	0.63	1.90	1.90	22.78	9.49	24.68	5.06	32.91	0.63
I97T	116	-0.50 ± 0.46	89.75 ± 19.33	77.20 ± 19.03	96.61 ± 17.39	19.41 ± 6.87	0.20 ± 0.06	1	0.00	0.00	0.00	0.86	12.93	10.34	6.03	69.83	0.00
F100L	64	0.02 ± 0.10	92.63 ± 24.89	55.87 ± 22.80	129.82 ± 17.14	73.95 ± 16.79	0.57 ± 0.10	4	0.00	0.00	0.00	25.00	6.25	9.38	9.38	50.00	0.00
I82V	61	0.05 ± 0.05	91.22 ± 19.94	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.00	0.00	0.00	47.54	1.64	42.62	6.56	1.64	0.00
I97F	56	-0.01 ± 0.28	89.75 ± 19.33	77.20 ± 19.03	96.61 ± 17.39	19.41 ± 6.87	0.20 ± 0.06	1	0.00	0.00	0.00	23.21	5.36	8.93	12.50	50.00	0.00
L93F	48	-0.02 ± 0.10	80.02 ± 12.90	38.00 ± 19.35	107.19 ± 12.26	69.19 ± 9.59	0.65 ± 0.08	4	0.00	0.00	0.00	18.75	37.50	20.83	0.00	20.83	2.08

I97V	47	0.08 ± 0.11	89.75 ± 19.33	77.20 ± 19.03	96.61 ± 17.39	19.41 ± 6.87	0.20 ± 0.06	1	0.00	0.00	0.00	4.26	14.89	68.09	0.00	8.51	4.26
V70A	45	0.16 ± 0.12	86.27 ± 18.53	15.82 ± 13.93	101.45 ± 11.56	85.63 ± 11.01	0.84 ± 0.10	6	0.00	2.22	0.00	26.67	51.11	6.67	0.00	0.00	13.33
V66M	40	-0.07 ± 0.06	72.96 ± 15.00	14.52 ± 8.56	72.23 ± 9.42	57.72 ± 5.87	0.80 ± 0.07	4	0.00	0.00	0.00	5.00	40.00	17.50	0.00	37.50	0.00
M109L	38	0.04 ± 0.16	118.10 ± 20.29	80.28 ± 22.41	168.22 ± 20.18	87.94 ± 14.46	0.52 ± 0.07	5	0.00	0.00	0.00	0.00	7.89	13.16	2.63	76.32	0.00
M109V	38	0.08 ± 0.28	118.10 ± 20.29	80.28 ± 22.41	168.22 ± 20.18	87.94 ± 14.46	0.52 ± 0.07	5	0.00	0.00	0.00	13.16	31.58	21.05	2.63	31.58	0.00
A85T	30	-0.49 ± 0.30	79.81 ± 13.46	6.65 ± 4.25	7.85 ± 4.73	1.21 ± 1.56	0.14 ± 0.17	2	0.00	0.00	0.00	3.33	23.33	3.33	10.00	60.00	0.00
F100S	29	-0.78 ± 0.33	92.63 ± 24.89	55.87 ± 22.80	129.82 ± 17.14	73.95 ± 16.79	0.57 ± 0.10	4	0.00	0.00	0.00	6.90	68.97	13.79	0.00	10.34	0.00
W75S	28	0.02 ± 0.32	134.37 ± 26.17	136.35 ± 44.73	185.63 ± 24.99	49.28 ± 20.33	0.27 ± 0.11	2	0.00	0.00	0.00	3.57	14.29	21.43	0.00	60.71	0.00
R107L	21	1.85 ± 0.51	122.51 ± 37.92	134.74 ± 27.60	197.33 ± 18.99	62.60 ± 21.03	0.32 ± 0.10	2	0.00	0.00	0.00	9.52	42.86	23.81	4.76	14.29	4.76
L67F	21	0.02 ± 0.08	76.67 ± 17.80	65.01 ± 14.59	80.66 ± 13.06	15.65 ± 5.30	0.19 ± 0.06	1	0.00	0.00	0.00	9.52	33.33	14.29	19.05	23.81	0.00
V66A	21	0.21 ± 0.08	72.96 ± 15.00	14.52 ± 8.56	72.23 ± 9.42	57.72 ± 5.87	0.80 ± 0.07	4	0.00	0.00	0.00	4.76	66.67	9.52	4.76	14.29	0.00
W55L	20	-0.03 ± 0.13	90.56 ± 16.30	45.24 ± 23.00	112.80 ± 19.40	67.56 ± 16.61	0.60 ± 0.15	3	0.00	0.00	0.00	0.00	10.00	35.00	0.00	55.00	0.00

L67I	12	-0.04 ± 0.12	76.67 ± 171.80	65.01 ± 14.59	80.66 ± 13.06	15.65 ± 5.30	0.19 ± 0.06	1	0.00	0.00	0.00	8.33	91.67	0.00	0.00	0.00	0.00
W55C	12	-0.13 ± 0.20	90.56 ± 16.30	45.24 ± 23.00	112.80 ± 19.40	67.56 ± 16.61	0.60 ± 0.15	3	0.00	0.00	0.00	25.00	8.33	8.33	0.00	58.33	0.00
A104S	10	-0.91 ± 0.38	105.19 ± 27.34	48.72 ± 14.21	62.41 ± 13.29	13.69 ± 8.89	0.21 ± 0.13	1	0.00	0.00	0.00	0.00	60.00	10.00	10.00	20.00	0.00
Y71H	10	-0.90 ± 0.50	92.35 ± 19.94	56.84 ± 28.93	84.97 ± 34.74	28.12 ± 10.60	0.34 ± 0.13	1	0.00	0.00	0.00	0.00	30.00	0.00	10.00	60.00	0.00
W75R	9	0.05 ± 0.33	134.37 ± 26.17	136.35 ± 44.73	185.63 ± 24.99	49.28 ± 20.33	0.27 ± 0.11	2	0.00	0.00	0.00	33.33	33.33	11.11	0.00	22.22	0.00
V66G	9	0.32 ± 0.10	72.96 ± 15.00	14.52 ± 8.56	72.23 ± 9.42	57.72 ± 5.87	0.80 ± 0.07	4	0.00	0.00	0.00	11.11	33.33	44.44	0.00	0.00	11.11
L93M	8	-0.01 ± 0.07	80.02 ± 12.90	38.00 ± 19.35	107.19 ± 12.26	69.19 ± 9.59	0.65 ± 0.08	4	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
F112L	7	0.01 ± 0.28	103.70 ± 34.74	23.38 ± 21.40	80.51 ± 27.54	57.14 ± 10.26	0.73 ± 0.15	4	0.00	0.00	0.00	28.57	28.57	0.00	0.00	28.57	14.29
M109T	7	0.08 ± 0.28	118.10 ± 20.29	80.28 ± 22.41	168.22 ± 20.18	87.94 ± 14.46	0.52 ± 0.07	5	0.00	0.00	0.00	14.29	28.57	0.00	0.00	57.14	0.00
R107C	6	1.78 ± 0.49	122.51 ± 37.92	134.74 ± 27.60	197.33 ± 18.99	62.60 ± 21.03	0.32 ± 0.10	2	0.00	0.00	0.00	0.00	16.67	50.00	16.67	16.67	0.00
I97M	6	-0.12 ± 0.31	89.75 ± 19.33	77.20 ± 19.03	96.61 ± 17.39	19.41 ± 6.87	0.20 ± 0.06	1	0.00	0.00	0.00	0.00	0.00	66.67	0.00	33.33	0.00
A85D	6	-0.06 ± 0.07	79.81 ± 13.46	6.65 ± 4.25	7.85 ± 4.73	1.21 ± 1.56	0.14 ± 0.17	2	0.00	0.00	0.00	16.67	83.33	0.00	0.00	0.00	0.00

F112Y	5	-0.09 ± 0.27	103.70 ± 34.74	23.38 ± 21.40	80.51 ± 27.54	57.14 ± 10.26	0.73 ± 0.15	4	0.00	0.00	0.00	0.00	40.00	20.00	0.00	40.00	0.00
W75C	5	0.03 ± 0.28	134.37 ± 26.17	136.35 ± 44.73	185.63 ± 24.99	49.28 ± 20.33	0.27 ± 0.11	2	0.00	0.00	0.00	20.00	80.00	0.00	0.00	0.00	0.00
A69T	5	-0.92 ± 0.34	83.55 ± 16.47	1.59 ± 3.05	16.96 ± 6.77	15.37 ± 4.26	0.90 ± 0.12	2	0.00	0.00	0.00	0.00	0.00	0.00	20.00	80.00	0.00
A69P	5	1.00 ± 0.71	83.55 ± 16.47	1.59 ± 3.05	16.96 ± 6.77	15.37 ± 4.26	0.90 ± 0.12	2	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
L67Y	5	0.01 ± 0.11	76.67 ± 17.80	65.01 ± 14.59	80.66 ± 13.06	15.65 ± 5.30	0.19 ± 0.06	1	0.00	0.00	0.00	20.00	20.00	0.00	20.00	20.00	20.00
A104T	4	-0.63 ± 0.44	105.19 ± 27.34	48.72 ± 14.21	62.41 ± 13.29	13.69 ± 8.89	0.21 ± 0.13	1	0.00	0.00	0.00	0.00	0.00	50.00	25.00	25.00	0.00
F103Y	4	-0.31 ± 0.36	95.69 ± 21.66	6.41 ± 6.87	78.93 ± 14.74	72.52 ± 8.95	0.92 ± 0.04	6	0.00	0.00	0.00	0.00	50.00	25.00	0.00	25.00	0.00
F100C	3	0.07 ± 0.21	92.63 ± 24.89	55.87 ± 22.80	129.82 ± 17.14	73.95 ± 16.79	0.57 ± 0.10	4	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
I82M	3	0.00 ± 0.06	91.22 ± 19.94	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
Y71C	3	0.05 ± 0.18	92.35 ± 19.94	56.84 ± 28.93	84.97 ± 34.74	28.12 ± 10.60	0.34 ± 0.13	1	0.00	66.67	0.00	0.00	33.33	0.00	0.00	0.00	0.00
L62S	3	-0.62 ± 0.34	71.67 ± 9.75	25.27 ± 7.76	46.13 ± 12.86	20.86 ± 7.36	0.45 ± 0.12	1	0.00	0.00	0.00	0.00	33.33	66.67	0.00	0.00	0.00
L62I	3	0.03 ± 0.09	71.67 ± 9.75	25.27 ± 7.76	46.13 ± 12.86	20.86 ± 7.36	0.45 ± 0.12	1	0.00	0.00	0.00	0.00	33.33	0.00	33.33	33.33	0.00

R107S	2	1.21 ± 0.62	122.51 ± 37.92	134.74 ± 27.60	197.33 ± 18.99	62.60 ± 21.03	0.32 ± 0.10	2	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
F103V	2	0.05 ± 0.40	95.69 ± 21.66	6.41 ± 6.87	78.93 ± 14.74	72.52 ± 8.95	0.92 ± 0.04	6	0.00	0.00	0.00	50.00	0.00	50.00	0.00	0.00	0.00
I97R	2	-0.31 ± 0.55	89.75 ± 19.33	77.20 ± 19.03	96.61 ± 17.39	19.41 ± 6.87	0.20 ± 0.06	1	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
L93P	2	2.29 ± 0.30	80.02 ± 12.90	38.00 ± 19.35	107.19 ± 12.26	69.19 ± 9.59	0.65 ± 0.08	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	50.00	50.00
A69D	2	-0.14 ± 0.20	83.55 ± 16.47	1.59 ± 3.05	16.96 ± 6.77	15.37 ± 4.26	0.90 ± 0.12	2	0.00	0.00	0.00	50.00	0.00	0.00	0.00	0.00	50.00
V66F	2	-0.10 ± 0.07	72.96 ± 15.00	14.52 ± 8.56	72.23 ± 9.42	57.72 ± 5.87	0.80 ± 0.07	4	0.00	0.00	0.00	0.00	0.00	0.00	50.00	50.00	0.00
F112C	1	0.02 ± 0.32	103.70 ± 34.74	23.38 ± 21.40	80.51 ± 27.54	57.14 ± 10.26	0.73 ± 0.15	4	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
F112S	1	-0.34 ± 0.45	103.70 ± 34.74	23.38 ± 21.40	80.51 ± 27.54	57.14 ± 10.26	0.73 ± 0.15	4	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
S111L	1	0.99 ± 0.47	102.15 ± 24.07	1.94 ± 3.03	3.21 ± 4.80	1.27 ± 1.80	0.34 ± 0.36	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
S108F	1	0.82 ± 0.55	120.22 ± 27.48	13.54 ± 7.55	18.53 ± 8.82	4.98 ± 4.53	0.27 ± 0.23	1	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
S108A	1	0.89 ± 0.30	120.22 ± 27.48	13.54 ± 7.55	18.53 ± 8.82	4.98 ± 4.53	0.27 ± 0.23	1	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
A104L	1	-0.15 ± 0.17	105.19 ± 27.34	48.72 ± 14.21	62.41 ± 13.29	13.69 ± 8.89	0.21 ± 0.13	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00

F103L	1	0.02 ± 0.35	95.69 ± 21.66	6.41 ± 6.87	78.93 ± 14.74	72.52 ± 8.95	0.92 ± 0.04	6	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
F100V	1	0.05 ± 0.18	92.63 ± 24.89	55.87 ± 22.80	129.82 ± 17.14	73.95 ± 16.79	0.57 ± 0.10	4	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
F100I	1	-0.03 ± 0.15	92.63 ± 24.89	55.87 ± 22.80	129.82 ± 17.14	73.95 ± 16.79	0.57 ± 0.10	4	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
I97S	1	-0.74 ± 0.42	89.75 ± 19.33	77.20 ± 19.03	96.61 ± 17.39	19.41 ± 6.87	0.20 ± 0.06	1	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
F96L	1	-0.05 ± 0.11	85.95 ± 21.44	13.54 ± 7.55	18.53 ± 8.82	4.98 ± 4.53	0.27 ± 0.23	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
L93S	1	-0.76 ± 0.30	80.02 ± 12.90	38.00 ± 19.35	107.19 ± 12.26	69.19 ± 9.59	0.65 ± 0.08	4	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
L93R	1	-0.47 ± 0.62	80.02 ± 12.90	38.00 ± 19.35	107.19 ± 12.26	69.19 ± 9.59	0.65 ± 0.08	4	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
L93I	1	-0.06 ± 0.08	80.02 ± 12.90	38.00 ± 19.35	107.19 ± 12.26	69.19 ± 9.59	0.65 ± 0.08	4	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
W92G	1	0.34 ± 0.23	77.61 ± 11.10	11.91 ± 8.36	67.96 ± 16.30	56.05 ± 9.67	0.83 ± 0.08	2	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
W92L	1	-0.03 ± 0.12	77.61 ± 11.10	11.91 ± 8.36	67.96 ± 16.30	56.05 ± 9.67	0.83 ± 0.08	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
W92R	1	-0.32 ± 0.49	77.61 ± 11.10	11.91 ± 8.36	67.96 ± 16.30	56.05 ± 9.67	0.83 ± 0.08	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
W92Q	1	-0.59 ± 0.49	77.61 ± 11.10	11.91 ± 8.36	67.96 ± 16.30	56.05 ± 9.67	0.83 ± 0.08	2	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00

A85G	1	0.13 ± 0.06	79.81 ± 13.46	6.65 ± 4.25	7.85 ± 4.73	1.21 ± 1.56	0.14 ± 0.17	2	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
I82F	1	0.06 ± 0.10	91.22 ± 19.94	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
I82A	1	0.15 ± 0.07	91.22 ± 19.94	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
W75G	1	0.09 ± 0.32	134.37 ± 26.17	11.91 ± 8.36	67.96 ± 16.30	56.05 ± 9.67	0.83 ± 0.08	2	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
A69F	1	-0.16 ± 0.31	83.55 ± 16.47	1.59 ± 3.05	16.96 ± 6.77	15.37 ± 4.26	0.90 ± 0.12	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
A69G	1	0.16 ± 0.07	83.55 ± 16.47	1.59 ± 3.05	16.96 ± 6.77	15.37 ± 4.26	0.90 ± 0.12	2	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
L67H	1	-1.07 ± 0.43	76.67 ± 17.80	65.01 ± 14.59	80.66 ± 13.06	15.65 ± 5.30	0.19 ± 0.06	1	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
L67P	1	2.37 ± 0.34	76.67 ± 17.80	65.01 ± 14.59	80.66 ± 13.06	15.65 ± 5.30	0.19 ± 0.06	1	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
P59S	1	-0.30 ± 0.30	80.69 ± 16.51	27.90 ± 14.82	50.05 ± 18.94	22.15 ± 6.64	0.45 ± 0.13	2	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
P59L	1	-0.02 ± 0.08	80.69 ± 16.51	27.90 ± 14.82	50.05 ± 18.94	22.15 ± 6.64	0.45 ± 0.13	2	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
P59Q	1	-0.83 ± 0.35	80.69 ± 16.51	27.90 ± 14.82	50.05 ± 18.94	22.15 ± 6.64	0.45 ± 0.13	2	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
P59T	1	-0.14 ± 0.24	80.69 ± 16.51	27.90 ± 14.82	50.05 ± 18.94	22.15 ± 6.64	0.45 ± 0.13	2	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00

Annex Table 2 - Analysis for detected co-occurring mutations in the predicted interface residues for SARS-CoV-2 M protein. This table describes the frequency of co-occurring mutations (Frequency), their $\triangle\triangle G_{binding}$ values ($\triangle\triangle G$), RMSF for each original residue (RMSF), solvent-accessible surface area (SASA) for each original residue in the complex (SASAcpx), SASA for each original residue in the monomer (SASAmon), ΔSASA for each original residue (ΔSASA), relative SASA for each original residue (relSASA), the number of interactions each original residues establish (Interactions) and the distribution of co-occurring mutations across the GISAID clades (Clade) (all the presented results are mean values ± standard deviation).

			ΔΔG	**RMSF**	**$SASA_{cpx}$**	**$SASA_{mon}$**	**ΔSASA**	**relSASA**		**Clade (%)**								
Co-occurrence	**Frequency**	**Mutation**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Interactions**	L	S	V	G	GH	GR	GV	GRY	O
M109L,V70L	8	M109L	0.03 ± 0.35	118 ± 20	80.28 ± 22.41	168.22 ± 20.18	87.94 ± 14.46	0.52 ± 0.07	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
		V70L		86 ± 19	15.82 ± 13.93	101.45 ± 11.56	85.63 ± 11.01	0.84 ± 0.10	6									
I82T,L93F	7	I82T	-0.5 ± 0.41	91 ± 20	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
		L93F		80 ± 13	38.00 ± 19.35	107.19 ± 12.26	69.19 ± 9.59	0.65 ± 0.08	4									
I82T,R107L	4	I82T	1.35 ± 0.64	91 ± 20	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
		R107L		123 ± 38	134.74 ± 27.60	197.33 ± 18.99	62.60 ± 21.03	0.32 ± 0.10	2									
A85S,V70F	3	A85S	-0.72 ± 0.64	80 ± 13	6.65 ± 4.25	7.85 ± 4.73	1.21 ± 1.56	0.14 ± 0.17	2	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
		V70F		86 ± 19	15.82 ± 13.93	101.45 ± 11.56	85.63 ± 11.01	0.84 ± 0.10	6									
A104V,V70L	3	A104V	-0.10 ± 0.32	105 ± 27	48.72 ± 14.21	62.41 ± 13.29	13.69 ± 8.89	0.21 ± 0.13	1	0.00	0.00	0.00	0.00	33.33	0.00	0.00	66.67	0.00

		V70L		86 ± 19	15.82 ± 13.93	101.45 ± 11.56	85.63 ± 11.01	0.84 ± 0.10	6										
I82T,M109V	2	I82T	-0.42 ± 0.50	91 ± 20	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
		M109V		118 ± 20	80.28 ± 22.41	168.22 ± 20.18	87.94 ± 14.46	0.52 ± 0.07	5										
I82T,V70F	2	I82T	-0.22 ± 0.63	91 ± 20	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
		V70F		86 ± 19	15.82 ± 13.93	101.45 ± 11.56	85.63 ± 11.01	0.84 ± 0.10	6										
I82T,M109I	2	I82T	-0.43 ± 0.44	91 ± 20	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
		M109I		118 ± 20	80.28 ± 22.41	168.22 ± 20.18	87.94 ± 14.46	0.52 ± 0.07	5										
I82T,V66M	2	I82T	-0.5 ± 0.37	91 ± 20	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.00	0.00	0.00	50.00	0.00	50.00	0.00	0.00	0.00
		V66M		73 ± 15	14.52 ± 8.56	72.23 ± 9.42	57.72 ± 5.87	0.80 ± 0.07	4										
A69V,A85S	2	A69V	-1.32 ± 0.27	84 ± 16	1.59 ± 3.05	16.96 ± 6.77	15.37 ± 4.26	0.90 ± 0.12	2	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
		A85S		80 ± 13	6.65 ± 4.25	7.85 ± 4.73	1.21 ± 1.56	0.14 ± 0.17	2										
I82S,R107H	2	I82S	0.57 ± 0.74	91 ± 20	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
		R107H		123 ± 38	134.74 ± 27.60	197.33 ± 18.99	62.60 ± 21.03	0.32 ± 0.10	2										
I97V,R107C	2	I97V	1.91 ± 0.55	90 ± 19	77.20 ± 19.03	96.61 ± 17.39	19.41 ± 6.87	0.20 ± 0.06	1	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00

		R107C		123 ± 38	134.74 ± 27.60	197.33 ± 18.99	62.60 ± 21.03	0.32 ± 0.10	2										
A104V,I82T	1	A104V	-0.59 ± 0.40	105 ± 27	48.72 ± 14.21	62.41 ± 13.29	13.69 ± 8.89	0.21 ± 0.13	1	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
		I82T		91 ± 20	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4										
I82T,V66L	1	I82T	-0.57 ± 0.39	91 ± 20	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
		V66L		73 ± 15	14.52 ± 8.56	72.23 ± 9.42	57.72 ± 5.87	0.80 ± 0.07	4										
A85S,I82T	1	A85S	-1.47 ± 0.47	80 ± 13	6.65 ± 4.25	7.85 ± 4.73	1.21 ± 1.56	0.14 ± 0.17	2	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
		I82T		91 ± 20	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4										
I82S,V70F	1	I82S	-0.35 ± 0.64	91 ± 20	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
		V70F		86 ± 19	15.82 ± 13.93	101.45 ± 11.56	85.63 ± 11.01	0.84 ± 0.10	6										
A69V,I82S	1	A69V	-0.65 ± 0.39	84 ± 16	1.59 ± 3.05	16.96 ± 6.77	15.37 ± 4.26	0.90 ± 0.12	2	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		I82S		91 ± 20	38.67 ± 15.10	93.09 ± 13.33	54.42 ± 13.27	0.58 ± 0.12	4										
A104V,V70F	1	A104V	0.09 ± 0.54	105 ± 27	48.72 ± 14.21	62.41 ± 13.29	13.69 ± 8.89	0.21 ± 0.13	1	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
		V70F		86 ± 19	15.82 ± 13.93	101.45 ± 11.56	85.63 ± 11.01	0.84 ± 0.10	6										
A85V,V70F	1	A85V	0.02 ± 0.47	80 ± 13	6.65 ± 4.25	7.85 ± 4.73	1.21 ± 1.56	0.14 ± 0.17	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00

		V70F		86 ± 19	15.82 ± 13.93	101.45 ± 11.56	85.63 ± 11.01	0.84 ± 0.10	6										
A85S,M109V	1	A85S	-0.86 ± 0.40	80 ± 13	6.65 ± 4.25	7.85 ± 4.73	1.21 ± 1.56	0.14 ± 0.17	2	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
		M109V		118 ± 20	80.28 ± 22.41	168.22 ± 20.18	87.94 ± 14.46	0.52 ± 0.07	5										
A69S,M109L	1	A69S	-1.16 ± 0.30	84 ± 16	1.59 ± 3.05	16.96 ± 6.77	15.37 ± 4.26	0.90 ± 0.12	2	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
		M109L		118 ± 20	80.28 ± 22.41	168.22 ± 20.18	87.94 ± 14.46	0.52 ± 0.07	5										
A104V,A69S	1	A104V	-1.32 ± 0.28	105 ± 27	48.72 ± 14.21	62.41 ± 13.29	13.69 ± 8.89	0.21 ± 0.13	1	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
		A69S		84 ± 16	1.59 ± 3.05	16.96 ± 6.77	15.37 ± 4.26	0.90 ± 0.12	2										
A69F,V70L	1	A69F	-0.14 ± 0.36	84 ± 16	1.59 ± 3.05	16.96 ± 6.77	15.37 ± 4.26	0.90 ± 0.12	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
		V70L		86 ± 19	15.82 ± 13.93	101.45 ± 11.56	85.63 ± 11.01	0.84 ± 0.10	6										
A69S,L93M	1	A69S	-1.22 ± 0.23	84 ± 16	1.59 ± 3.05	16.96 ± 6.77	15.37 ± 4.26	0.90 ± 0.12	2	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
		L93M		80 ± 13	38.00 ± 19.35	107.19 ± 12.26	69.19 ± 9.59	0.65 ± 0.08	4										