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

**Catarina Marques-Pereira**
Center for Neuroscience and Cell Biology, University of Coimbra https://orcid.org/0000-0001-6840-8991

**Manuel Pires**
Center for Neuroscience and Cell Biology, University of Coimbra https://orcid.org/0000-0002-0416-0787

**Raquel Gouveia**
Center for Neuroscience and Cell Biology, University of Coimbra https://orcid.org/0000-0001-5092-4373

**Nadia Pereira**
Center for Neuroscience and Cell Biology, University of Coimbra

**Ana Caniceiro**
Center for Neuroscience and Cell Biology, University of Coimbra https://orcid.org/0000-0002-4074-9142

**Nicia Rosário-Ferreira**
Center for Neuroscience and Cell Biology, University of Coimbra https://orcid.org/0000-0002-7225-9287

**Irina Moreira** ( irina.moreira@cnc.uc.pt )
University of Coimbra, Center for Neuroscience and Cell Biology https://orcid.org/0000-0003-2970-5250

**Research Article**

**Keywords:** SARS-CoV-2, Membrane Protein, Mutations, Dimeric Interface, Protein-Protein Interactions

**Posted Date:** January 25th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-702792/v2

# Abstract

*Severe Acute Respiratory Syndrome CoronaVirus-2* (SARS-CoV-2) is composed by four structural proteins and several accessory non-structural proteins. 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.

In this paper, we explain the structural and dynamic nature of M protein homodimer. To do so, we developed and applied a detailed and robust *in silico* workflow to predict M protein dimeric structure, membrane orientation, and interface characterization. *Single Nucleotide Polymorphisms* (SNPs) 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 SNPs in *Variants of Concern* (VOC) and *Variants of Interest* (VOI). Binding free energy differences were evaluated for dimer interfacial SNPs 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.


# Introduction

COronaVIrus Disease 2019 (COVID-19) is currently a worldwide pandemic that was first reported in December 2019 in Wuhan, China and, since then, led to more than 187 M infected people and over 4.0 M deaths[1] (as of July 11$^{th}$, 2021). COVID-19 is caused by Severe Acute Respiratory Syndrome CoronaVirus-2 (SARS-CoV-2), which is a Coronaviridae family, positive single-stranded RiboNucleic Acid (ssRNA) virus[2,3]. Since the beginning of this pandemic, SARS-CoV-2 has mutated overtime leading to the identification of several variants that, based on phylogeny[4], have been organized into clades named L, S, V, G, GH, GR, GV, GRY and O (clade based on exclusion encompassing sequences that do not fit into other clades)[5,6].

According to the World Health Organization (WHO), there are Variants Of Interest (VOI), variants that have been recognized as being able to acquire community transmission causing clusters and being further identified in several countries, or assessed as a VOI by WHO's SARS-CoV-2 Virus Evolution Group. On the other hand, Variants Of Concern (VOC) are variants that, adding to the characterization as VOI, are linked to increased transmissibility or virulence, and/or a decrease in the effectiveness of treatment, prevention and diagnosis approaches currently used. VOI are distributed among clades G (lineages B.1.525 and B.1.617.1), GH (lineages B.1.427/B.1.429 and B.1.526), and GR (lineages C.37, P.2 and P.3). Moreover, VOC are distributed among clades G (lineage B.1.617.2), GH (lineage B.1.351), GR (lineage P.1), and GRY (lineage B.1.1.7).

SARS-CoV-2 genes encode four major structural proteins: Spike (S) protein, Membrane (M) protein, Nucleocapsid (N) protein, and Envelope (E) protein. Along with these structural proteins, SARS-CoV-2 genes also encode sixteen non-structural proteins (nsp) and accessory proteins[7]. One of the most

conserved structural proteins in SARS-CoV-2 is the M protein, as it has a smaller mutation rate sharing structural and functional similarities with M proteins from another coronavirus[8].

M protein is constituted by 223 amino acids and has three major domains: a short-glycosylated N-terminal ecto-domain, three TransMembrane Helices (labelled as TMH1, TMH2, and TMH3) and a long C-terminal endo-domain[9–11]. SARS-CoV M protein is known to acquire two different conformations: one elongated conformation associated with rigidity, S clustering and a narrow range of membrane curvature, and a more compact conformation associated with greater flexibility, a lower S density and M-S Protein-Protein Interactions (PPIs)[12]. In addition to these heterotypic interactions, M protein can acquire a homodimeric form. Since M protein is essential in the SARS-CoV-2 viral life cycle, a complete understanding of the structure-function relationship will help the development of more efficient therapeutics[12]. However, this task has been affected by the difficulty to stabilize and crystallize the M protein[13,14], and there are no available experimentally acquired structures. Moreover, mutations can impact M protein's structure and, consequently, affect its homotypic interactions. Bioinformatic tools are well established methodologies that allow to attain a structural and functional characterization of relevant biomedical targets[15,16]. In this work, through a in house developed *in silico* approach (Figure 1), we elucidated the M protein monomer and dimer three-dimensional (3D)-structures along with predictions for their membrane orientation and homodimeric interface. We also determined the impact of mutations in the homodimeric interface, paving the way to structure-driven formulation of new drugs.

## Results

# M protein monomer structure and membrane orientation

M protein is a membrane protein and the determination of its correct orientation in the lipid bilayer membrane is needed to understand its main interactions, and therefore its biological function. To this end, six different web-based resources for membrane orientation prediction were used: OPM[18], TMpred[19], TMHMM[20,21], PSIPRED[22,23], CCTOP[24,25] and SACSMEMSAT[26].

M protein Root-Mean-Square-Deviation (RMSD) results were obtained considering residues from the whole protein (monomer RMSD) and only transmembrane residues (transmembrane RMSD). Monomer RMSD values were 1.42 Å for TMHMM, 1.43 Å for CCTOP, 1.47 Å for TMpred, 1.59 Å for SACSMEMSAT, 1.74 Å for OPM, and 2.50 Å for PSIPRED predictions (Supplementary Figure 1). Transmembrane RMSD values 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 predictions (Supplementary Figure 1). M protein monomer predicted residue domains, after system equilibration, were very similar for all membrane orientation predictions. For the following dimer prediction study, PSIPRED results were not used as RMSD values were higher for both monomer and transmembrane RMSD. Despite SACSMEMSAT and CCTOP having comparable values to the other predictors, they showed an arched TMH1 after an equilibration Molecular Dynamic (MD) simulation that could influence dimer stability (Supplementary Figure 1). Hence, out of the six

membrane predictors used initially, OPM, TMHMM and TMpred M protein monomers were chosen for further analysis.

# M protein dimer and interface prediction

OPM, TMpred and TMHMM M monomers from the previous step were used to model dimer 3D structures using a well-established protein-protein docking software: HADDOCK[30]. From 3000 proposed docking decoys, 1000 for each membrane orientation, 20 dimer structures that respected the membrane orientation prediction were selected: 11 from OPM, 4 from TMpred and 5 from TMHMM. From these 20 dimers, two structures from the TMHMM membrane predictor were chosen based on their similarity with SARS-CoV experimental detected interactions, namely in TMH2 (P59) and TMH3 (W92, L93, F96) regions[10]. From these two TMHMM M protein dimers, the final choice was based on PROtein binDIng enerGY (PRODIGY)'s metrics of biological probability and predicted binding affinity. Hence, the M protein dimer structure chosen for the proceeding studies showed 85.6% biological probability and a predicted binding affinity of -6.3 kcal/mol in comparison to 74.8% biological probability and -5.9 kcal/mol binding affinity results from the other available structure. Regarding the TMHMM monomer membrane prediction that served as template for the final chosen dimer, M protein monomer residues 11-19 were shown to stably belong to N-terminal domain, residues 100-203 to C-terminal domain, residues 20-38 to TMH1, residues 46-70 to TMH2 and residues 76-100 to TMH3 (Figure 2).

The final dimer 3D structure (Figure 3) was subjected to three independent dimer system MD replicas of 0.5 μs. After equilibration, polar contacts between M protein monomer and membrane lipids occurred in M monomer residues K14, Y39, R42, N43, R44, F45, Y71, R72, W75, S94, R101, R107, W110, S173, R174. Transmembrane regions were within membrane lipids throughout the entire equilibration and several M protein residues were able to establish polar contacts with membrane lipids, supporting our transmembrane prediction (Figure 3).

RMSD results (Supplementary Figure 2) showed that monomer A and monomer B behaved differently throughout the MD simulation. In monomer A, TMH3 domain was the most stable region. Monomer A TMH2 domain interacted with monomer B and was a bit more unstable when compared with TMH1 domain (Supplementary Figure 2A). In monomer B, TMH domains were also very stable, and the major difference observed was a much higher deviation and lower stability of the N- and C-terminus compared with other domains (Supplementary Figure 2B). Root-Mean-Square-Fluctuation (RMSF) results (Supplementary Figure 3) for monomer A and monomer B were very similar. As expected, TMH residues, in large majority α-helices, showed low fluctuation, whilst C-terminus residues, present in a random coil, presented higher fluctuation. Cross-Correlation Analysis (CCA) results (Supplementary Figure 4) showed that within both monomers, TMH2 is highly positively correlated (moves in the same direction) with

TMH1 and TMH3 within the same protein. On the contrary, between monomers, TMH1 and TMH2 showed a negative correlation (moving in opposite directions) with remaining helices of the opposite monomer.

After dimer equilibration in an ER membrane to mimic the expected biological environment, we showed that the dimer interface was composed of 38 residues, 17 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 residues established 34 pairwise interactions, showing high proximity and high prevalence time (90% cut-off) (Table 1). Carbon Alpha (Cα) distances of interacting residues varied between 5.25 Å (V70-V70 residues interaction) and 12.58 Å (W92-W92 residues interaction), with a mean Cα distance of 9.57 ± 0.60 Å. From these residues, 12 (P59, V66, A69, V70, I82, L93, F96, F100, F103, R107, M109 and F112) interacted in both monomers. From these 38 residues, 23 were unique residues, seven from TMH2 (W55, P59, L62, L67, V66, A69 and V70), two from TMH2-TMH3 extracellular loop (Y71, W75), seven from TMH3 (I82, W92, L93, I97, A85, F96, F100) and seven from C-terminal (F103, A104, R107, S108, M109, S111, F112) (Table 1). From these, 8 were 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, R107) and 1 was a positively charged residue (R107).

Interactions between monomer A and monomer B residues 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 interactions throughout 100% of MDs simulation time, with side chain distances lower than 5 Å (Table 1, Figure 4). These regions also showed a low fluctuation (e.g., low RMSF values). Hydrophilic interactions occurred between monomer A residues L62-V66, V66-V69, W92-F96, F96-F100 and F103-R107 and between monomer B residues L62-V66, V66-V69, L92-I97, F100-A104, A104-R107, S106-M107 and M107-F112. π-π stack interactions occurred between monomer A residues W92-F96 and F100-F112 and between monomer A and monomer B residues W55-F100, W92-W92, F100-F112 and F103-F103, respectively. Within these 34 interactions: 9 were established between monomer A and monomer B C-terminal residues (F103-F103, M109-F103, R107-M109, M109-A104, F103-S108, F112-F103, F103-F112, F103-S111 and M109-R107), 6 between monomer A TMH2 and monomer B TMH3 residues (W55-L93, P59-L93, V70-I82, W55-I97, V66-A85 and W55-F96), 5 between monomer A and monomer B TMH2 residues (V66-V66, A69-V70, V70-A69, V70-V70 and L62-L62), 5 between monomer A TMH3 and monomer B TMH2 residues (I82-V70, L93-P59, I82-L67, I82-V66 and A85-V66), 4 between monomer A and monomer B TMH3 residues (W92-W92, F96-F96, W92-L93, and F100-F96), 2 between monomer A C-terminal and monomer B TMH3 residues (F112-F100 and M109-F100), 2 between monomer A residue W75 from TMH2-TMH3 extracellular loop and monomer B TMH2-TMH3 extracellular loop residue Y71 and with monomer B TMH2 residue V70, respectively, and 1 between monomer A TMH3 domain and monomer B C-terminal domain (F100-F112) (Table 1).

# M protein mutation analysis

We retrieved 1271550 M protein sequences, submitted between 10/01/2020 and 03/05/2021 from 180 countries, from the Global Initiative on Sharing All Influenza Data (GISAID)[34,35] database. Genomic sequences were obtained from human hosts, with more than 29,000 bases per sequence, and less than 5% missing values. The sequence distribution retrieved across GISAID clades and across the world can be observed in Figure 5.

Clades S, G, GH and GR encompass sequences that are most prevalent in North America. The latter clade is also well represented in the Oceania region. Clades GV and GRY are most prevalent in Europe and clades O and L are sparse across the world. Within the M protein interfacial residues from analyzed sequences, 91 Single Nucleotide Polymorphisms (SNPs) were retrieved from 21868 sequences. FoldX was used to assess the binding free energy differences between mutated and Wild-Type (WT) proteins ($\triangle\triangle G_{binding}$) and the respective values by physio-chemical character of the analyzed mutation are illustrated in Figure 6 and with higher detail in Supplementary Figure 5.

In these considered regions, the overall $\triangle\triangle G_{binding}$ was -0.01 ± 0.62 kcal/mol, in which 606 (2.77%) of the mutated sequences showed a $\triangle\triangle G_{binding}$ value superior to 0.50 kcal/mol, 2683 (12.27%) had a $\triangle\triangle G_{binding}$ inferior to -0.50 kcal/mol and 18579 (84.96%) had $\triangle\triangle G_{binding}$ values between -0.50 and 0.50 kcal/mol. From these, 55.53% represented mutations from one non-polar to other non-polar residues ($\triangle\triangle G_{binding}$ = 0.14 ± 0.49 kcal/mol), 41.68% from a non-polar to a polar residue ($\triangle\triangle G_{binding}$ = -0.42 ± 0.36 kcal/mol), 2.68% from a polar to another polar residue ($\triangle\triangle G_{binding}$ = 0.65 ± 1.07 kcal/mol), and 0.11% from a polar to a non-polar residue, with $\triangle\triangle G_{binding}$ = 1.14 ± 0.48 kcal/mol (Supplementary Figure 5). For the same 91 SNPs, 90.01% represented mutations from a non-aromatic to another non-aromatic residue ($\triangle\triangle G_{binding}$ = 0.04 ± 0.77 kcal/mol), 7.27% from a non-aromatic to an aromatic residue ($\triangle\triangle G_{binding}$ = 0.09 ± 0.29 kcal/mol), 2.69% from an aromatic to a non-aromatic residue ($\triangle\triangle G_{binding}$ = -0.11 ± 0.30 kcal/mol), and 0.03% from an aromatic to another aromatic residue, ($\triangle\triangle G_{binding}$ = -0.20 ± 0.15 kcal/mol) (Supplementary Figure 5). SNP I82T, located at the TMH3 domain, was the most common SNP detected. This mutation led to the residue's polarity modification from a non-polar residue into a polar one and occurred in 6316 (28.88%) sequences from our dataset. The second most frequent SNP was V70L, at the end of the TMH2 domain. This mutation did not change the type of polarity at that specific position and was detected in 6303 (28.82%) sequences. These were by far the most common SNPs, with the third most common one occurring in only 1455 sequences (more details in Supplementary Table 1).

We also analyzed the type of mutation found in each known clade (Supplementary Figure 6, Supplementary Table 1 - single mutations and Supplementary Table 2 - co-occurring mutations). The most common mutated clade was GRY, where VOCs can be found, with 36.69% of all dimeric detected SNPs. The most frequent mutation found in these homodimeric interfacial residues was V70L, representing 73.30% of all mutations detected in sequences of this clade, with a $\triangle\triangle G_{binding}$ value of -0.02 ± 0.22 kcal/mol. This mutation co-occurred in GRY with M109L (8 cases), A104V (2 cases), A69F (1 case) without any major identifiable energetic advantage ($\triangle\triangle G_{binding}$ around 0 kcal/mol). The second

most frequent mutated clade, where VOCs are also located, was GH, with 21.25%. The most frequent mutation in this clade was I82T, representing 47.23% of all GH clade mutations and a ($\triangle\triangle G_{binding}$ value of -0.49 ± 0.38 kcal/mol). A few mutations also co-occurred with I82T but in low frequency. From these, A85S induced a higher stabilization of the dimer interface ($\triangle\triangle G_{binding}$ value of -1.47 ± 0.47 kcal/mol). G clade mutated sequences constituted 19.06% of the mutated sequences, and the most frequent one was I82T, 71.26%, with a $\triangle\triangle G_{binding}$ value of -0.49 ± 0.38 kcal/mol. A few double mutations of interfacial residues were also found, in particular I82T-R107L (4 cases), I82T-V70F (2 cases), I82T-M109I (2 cases), I82T-V66M (2 cases), I82T-A85S (2 cases), I82T-R107H (2 cases) but none led to higher changes in the binding free energy. Mutated sequences contained in GR clade represent 17.27% of all mutated sequences. The most common mutation in this clade was V70F, 26.32%, with a $\triangle\triangle G_{binding}$ value of 0.17 ± 0.47 kcal/mol. A few mutations were found in association, such as A85S (3 cases, $\triangle\triangle G_{binding}$ = -0.72 ± 0.64 kcal/mol) and A104V (1 case, $\triangle\triangle G_{binding}$ = 0.10 ± 0.54 kcal/mol). The remaining clades were much less populated with mutated sequences: 4.36% in clade GV, 0.90% in clade S, 0.38% in clade O and 0.05% in clades L and V.

In total there were 8951 (40.93%) mutated sequences that were found in VOC and 2757 (12.61%) that were found in VOI. Out of VOC identified sequences, 8474 (94.67%) were contained in pango lineage B.1.1.7 and the most common mutation in this variant was V70L, represented in 6136 sequences (72.41%). In sequences identified as VOI, the most represented pango lineage was B.1.525 (72.59%) and the most frequent mutation for this variant was I82T, present in 2139 sequences (72.48%) (Figure 7).

Solvent occlusion has already been demonstrated as a key aspect of PPIs, as main interfacial residues SASA values are considerably more diminished upon complex formation compared to other interfacial residues[36–41]. The most mutated residues such as I82 (mean value for both monomers: $\triangle$SASA = 54.42 ± 13.27 Å$^2$, relSASA = 0.58 ± 0.12), V70 ($\triangle$SASA = 85.63 ± 11.01 Å$^2$, relSASA = 0.84 ± 0.10), A69 ($\triangle$SASA = 15.37 ± 4.26 Å$^2$, relSASA = 0.90 ± 0.12 Å$^2$) showed higher $\triangle$SASA and relSASA values, which indicates occlusion of these residues upon complex formation, with $SASA_{complex}$ values tending to zero (higher $\triangle$SASA and relSASA closer to 1). Other frequently mutated residues loose accessibility to the solvent but still remained attainable in the complex form: e.g., M109 ($\triangle$SASA = 87.94 ± 14.46 Å$^2$, relSASA = 0.52 ± 0.07), A104 ($\triangle$SASA = 13.69 ± 8.89 Å$^2$, relSASA = 0.21 ± 0.13), R107 ($\triangle$SASA= 62.60 ± 21.03 Å$^2$, relSASA = 0.32 ± 0.10), and W75 ($\triangle$SASA = 49.28 ± 20.33 Å$^2$, relSASA = 0.27 ± 0.11). By preventing bulk water to approximate these interfacial residues, the number and force of interaction established increases and the PPI is strengthened. Residues V70, M109, and I82 established a high number of dimer interactions: 6, 5 and 4, respectively. On the other hand, residues A69, R107, W75 established two interactions each and residue A104 established only one interaction.

As some of these mutations may impact protein's stability, we also look into the identification of their presence in VOI and VOC strains since it can lead to future drug discovery concerning the M protein. The mutations leading to $\triangle\triangle G_{binding}$ below -0.50 kcal/mol or over 0.50 kcal/mol are indicative of such cases. Mutations A69P, R107C, R107H, R107L, and R107S, all have $\triangle\triangle G_{binding}$ values over 0.50 kcal/mol.

Despite the R107H relatively low mutation frequency, it appears in several VOCs as B1.1.7, B.1.351, P.1, and VOI B.1.617.1. On the other hand, mutations I82T, I82S, A69S, A104S, A69T, and A104T have $\triangle\triangle G_{binding}$ values below -0.50 kcal/mol meaning that they have a favorable impact on the mutated protein stability. Mutation I82T has been detected in several VOCs as B.1.617.2 and B.1.1.7, in higher frequency, but also in P.1.1 and B.1.351, and in VOI B.1.525. Mutation I82S has been detected in VOCs B1.1.7 and B.1.351 sparingly and in VOI B.1.617.1 more frequently. Mutation A69S has been detected in VOC B.1.1.7 more frequently than in VOC B.1.351 and in VOI B.1.526 much more infrequently, and in VOI P.2 just once. Mutation A69T is much less frequent than A69S but has also been detected in VOC B.1.1.7. Finally, mutations A104S and A104T have both been identified in VOC B.1.1.7 twice and three times, respectively.

# Discussion

In this work, our starting point was the AlphaFold's M protein monomer for which we predicted its membrane orientation using six different membrane orientation software's. After minimization and MD equilibration, we chose TMHMM M protein monomer membrane orientation prediction for the following studies since it showed a higher stability, with low RMSD values upon comparison with the initial AlphaFold's structure, and without any major conformational change. SARS-CoV M protein monomer domains were previously predicted in an experimental research that elucidated M protein dimer interactions[10]. In that experiment, residues 15-37 were shown to belong to TMH1, residues 50-72 to TMH2 and residues 77-99 to TMH3[10]. For the first time, a reliable SARS-CoV2 M protein membrane orientation was proposed by this work that showed that residues 20-38 belong to TMH1, residues 46-70 to TMH2 and residues 76-100 to TMH3, results in agreement to the above mentioned SARS-CoV experimental results.

Despite the M protein dimer being crucial for various biological functions such as SARS-CoV-2 virion assembly and shape formation, the type of interactions established in its homodimer form are still poorly understood. Experimental SARS-CoV M protein dimer data demonstrated that residues W19, W57, P58, W91, L92, Y94, F95 and C158 were relevant, suggesting that homologous residues W20 (TMH1 domain), W58, P59 (TMH2 domain), W92, L93 Y95, F96 (TMH3 domain) and C159 (endo-domain) of SARS-CoV-2 may also be important for M dimer interaction and stabilization[10]. Authors also hypothesized that SARS-CoV residues C63 and C85 mutations did not interfered with M dimer formation, suggesting that homologous SARS-CoV-2 M protein residues C64, C86 and C159 may also not be involved in M dimer interface[10]. This information was used as cue for various docking experiments as already detailed in the Results section. A high confidence docking decoy based on the TMHMM monomer was subjected to further studies due to its proper membrane orientation regarding previous analysis. In particular, it was subjected to 1.5 μs MD, which showed that overall conformational stability for monomer A and monomer B was slightly different (e.g., dissimilar RMSDs), whereas RMSF results were alike, especially in TMH domains. TMHs showed low fluctuations, which allowed the establishment of highly prevalent and meaningful interactions between the two monomers. We identified 34 main interactions responsible for

the M protein dimer 3D structure stabilization, between 17 residues from monomer A and 21 residues from monomer B. From these interactions, 73.53% occurred between transmembrane residues, which was expected as the M protein is a transmembrane dimeric system. From these interactions, 12 were conserved throughout the entire MD simulation time, including interactions between W55-L93, W92-W92, L93-P59 and F96-F96, homologous residues from the ones detected to SARS-CoV[10]. This suggests that these four interactions are pivotal towards M protein dimer stabilization. Other interacting residues were present in lasting interactions throughout the MDs simulations, and thus important residues to further study and validate were W55, V66, A69, V70, Y71, W75, I82, F103 and M109.

Regarding mutation analysis, from the 1271550 genomes analyzed, 21868 sequences carried SNPs at M protein dimer predicted interaction residues. This represents only 1.7% of all retrieved genomes suggesting that the predicted interfacial region is extremely conserved[42]. We identified 91 unique SNPs in this predicted interface. From these, 2.77% had a $\triangle\triangle G_{binding}$ higher than 0.50 kcal/mol, which means that these mutations can have a negative impact in the M protein dimer stability and 12.27% had a $\triangle\triangle G_{binding}$ lower than -0.50 kcal/mol hence, could have a favorable impact in M protein dimer stability. The majority of mutations did not appear to influence M protein dimer interfacial stabilization, since about 85% had $\triangle\triangle G_{binding}$ values between -0.50 kcal/mol and 0.50 kcal/mol. The ones that seem to lead to a gain of stabilization were I82T, I97T, I82S, W92Q, L62S, A104T, I97S, L93S, F100S, P59Q, Y71H, A104S, A69T, A85S, L67H and A69S with $\triangle\triangle G_{binding}$ 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. We included here I82T as it is very closed to our stablished threshold and is the most prevalent detected mutation. Most SNPs remained as non-polar residues (55.53%) or transitioned from non-polar to polar residues (41.68%) and most continued as non-aromatic residues. Since the M protein is a membrane protein, many non-polar residues were found within the membrane region, and, as such, most predicted interactions involved non-polar residues. However, mutations from non-polar to polar residues may confer a gain in conformation stability as they may establish hydrogen bonds. In our work, 99.36% of non-polar to polar SNPs had $\triangle\triangle G_{binding}$ negative values, which endorses the maintenance or increase in stability as proposed. Mutations in homologous SARS-CoV experimentally interacting residues P59, W92, L93 and F96 were sparse and showed $\triangle\triangle G_{binding}$ values close to zero. Three exceptions were exposed: L93S and W92Q with $\triangle\triangle G_{binding}$ values lower than -0.5 kcal/mol, suggesting that these residues were also extremely important for M protein dimer interaction; and L93P ($\triangle\triangle G_{binding}$ = 2.29 kcal/mol) value, the second highest, probably due to the destabilization caused by Proline in the TMH3 α-helix.

The most common mutations were I82T (28.88%) and V70L (28.82%), key residues for M monomers interaction as I82 and V70 interaction was conserved throughout the entire MDs simulation with a mean distance of 8.62 ± 0.65 Å for I82-V70 and 9.08 ± 0.66 Å for V70-I82 interactions (monomer A - monomer B). Both these residues (V70 and I82) had low RMSF values and were occluded from solvent upon complex formation (ΔSASA values between 45-88 $Å^2$), which protects the established interactions. I82T and V70L, showed $\triangle\triangle G_{binding}$ values of -0.49 ± 0.38 kcal/mol and -0.02 ± 0.22 kcal/mol, suggesting that I82T is the most favorable, high-prevalent mutation and should be further studied.

Overall, most represented clades in our mutation study were GRY (36.69%), containing VOC and GH (21.25%), G (19.06%) and GR (17.27%), containing VOC and VOI. This could mean that SNPs in the interface region may impact SARS-CoV-2 life cycle, specifically regarding the M protein functions. Furthermore, these mutations are intrinsically related to known VOC and VOIs. For instance, V70L and I82T mutations appeared in 99.5% and 97.64% of clades sequences that contain VOC and VOI. The most common mutation in VOC was V70L, detected in 6137 VOC genomes, and 97.35% of the time this mutation was detected, it appeared in pango lineage B.1.1.7, a VOC in clade GRY.

There were 25 co-occurrent mutations on the GISAID data, 12 of which on interfacial residues involved in PPIs present throughout the entire MDs simulation. Even though SNP V70L only co-occurred with other mutations in 9 cases, these sequences were from clade GRY, which contains several VOC. Overall, clades G (27.45%), GRY (23.53%), GH (23.53%) and GR (19.61%) were the most represented in our co-occurrence results, all containing VOC. V70L does not seem to be by itself relevant for homodimer formation but seems to be a catalyzer if co-occurring with other interfacial mutations as found in various VOCs. Clades GV (3.92%) and S (1.96%) also contained sequences with co-occurring mutations, and the remaining ones did not show any co-occurring mutations. It is possible to conclude that the majority of co-occurring mutations were indeed in VOC and VOI containing clades.

As M protein dimer has several important functions during SARS-CoV-2 life cycle, it is fundamental to understand its structure-function relationship. Herein, upon establishing a comprehensive and well detailed computational pipeline, we were able not only to assess mutation effects at this interface but also to understand the dynamic behavior of the region and establish the consequences for dimer stability for the first time. This was the first time that SARs-CoV-2 M protein dimer structure and interactions were proposed and thoroughly studied either computationally or experimentally. As confirmed in this and other studies, M protein is very well conserved, and thus a good candidate for new therapeutic solutions regarding SARS-CoV-2.

## Methods

This work can be split into three main steps: M protein monomer membrane orientation prediction, M protein dimer 3D structure prediction and mutation effect assessment in the homodimer interface. The overall workflow to accomplish these goals is illustrated in Figure 1.

# M protein monomer structure and membrane orientation

As there are no experimentally resolved structures for SARS-CoV-2 M protein dimer or monomer, and protein homology to other known 3D structures is reduced, we used AlphaFold´s[17] team proposed monomeric structure. AlphaFold is a state-of-the-art Neural Network (NN)-based algorithm that predicts protein 3D structures from their sequence with a mean accuracy of 2.1 Å[43]. From 223 amino acids

present in M protein, AlphaFold was able to confidently predict a structure encompassing residues 11 to 203, which were the ones studied henceforth. Six different web-based resources for membrane orientation prediction were used: OPM[18], TMpred[19], TMHMc[20,21], PSIPRED[22,23], CCTOP[24,25] and SACSMEMSAT[26]. OPM database is able to predict protein structure within the lipid bilayer and it optimizes position taking into account protein-membrane interactions[18]. TMpred predicts membrane-spanning regions and orientations from naturally occurring membrane proteins[19]. TMHMM correctly predicts membrane proteins' α-helices positions with an accuracy of 77%, differentiating between soluble and membrane proteins[20,21]. PSIPRED predicts membrane protein secondary structure based on position-specific scoring matrices[22,23]. CCTOP predicts transmembrane topology using known experimental and computational membrane topologies[24,25]. SACSMEMSAT is able to predict protein secondary structure and membrane protein topology from well-defined membrane protein data[26].

We used MD simulations for the M monomer initial minimization considering each membrane orientation obtained via OPM, TMpred and TMHMM, PSIPRED, CCTOP and SACSMEMSAT. MDs were performed using GROMACS[28,29] and the CHARMM36 force field[44]. Each system was built with CHARMM-GUI[27] membrane builder with TIP3 waters, 0.9 M $Na^+$ and $Cl^-$ ions and a bilayer membrane with POPC:POPE:PI:POPS:PSM:Cholesterol, in order to replicate human ER membrane[45], as M protein is translated and virus is assembled in this organelle. System size, water molecules, ion numbers and lipid composition are described in Supplementary Table 3. Systems initial minimization was performed in order to remove bad contacts using the steepest descent algorithm. In this step, systems were heated with a Berendsen-thermostat at 310 K in the canonical ensemble (NVT) over 7 ns, and pressure was kept constant at one bar with isothermal–isobaric ensemble (NPT) for 20 ns with a semi-isotropic pressure coupling algorithm[46]. Long-range electrostatic interactions were treated by the fast smooth Particle-Mesh Ewald (PME) method[47]. RMSD analysis was conducted in Pymol, version 1.2r3pre with protein and transmembrane Cα residues in order to establish structural differences between AlphaFold M protein prediction and membrane orientation equilibrated results.

# M protein dimer and interface prediction

OPM, TMpred and TMHMM protein monomers were selected from system equilibration results and subjected to M protein dimer prediction. To guide the protein-protein docking we used known information on SARS-CoV M protein that has a 90.5% sequence identity and 90% homology with SARS-CoV-2 M protein[48]. Two equilibrated M protein monomers from each membrane orientation were used for dimer prediction using the docking tool HADDOCK[30], version 2.4, a protein quaternary structure predictor based on experimental data. Since M protein is a membrane protein and most homodimers are symmetric[49], water docking results were not considered and docking results with TMH2 and TMH3 non-crystallographic symmetry restraints were generated. To determine M protein monomer’s active residues,

CPORT[50], a protein-protein residue interaction predictor at an atomic-level, was used and only transmembrane residues predicted by this tool were considered for downstream steps. For each membrane predictor, 5000 dimer structures were generated in rigid body docking phase (it0) and 1000 structures for the semi-flexible refinement phase (it1). Dimer results were examined, according to each monomer membrane orientation prediction through an in-house Python script. Upon the selection of the most 20 promising HADDOCK dimers 3D structures, we extended our work towards interface interacting residues prediction. Protein Interfaces, Surfaces and Assemblies (PISA)[51], a web-based tool that resorts to chemical-physical principles for analyzing and modeling of macromolecular interactions, was used as a first predictor for dimer interface residues on all twenty dimer structures. Two dimers were chosen based on PISA results and their comparison with SARS-CoV's M protein dimer experimental results, highlighted homologous SARS-CoV-2 residues W20, W58, P59, W92, Y95, F96 and C159 as important residues for dimer stabilization. Selected structures were further subjected to PRODIGY[52,53]. PRODIGY not only predicts dimer interacting residues, but also helps to determine if a protein interface is crystallographic or biological, the latter meaning that the predicted dimer is biologically relevant.

The final dimer system was built in a similar way as above-mentioned for M protein monomer MD simulations[45] (Supplementary Table 3). Three independent dimer system replicas of 0.5 μs MD simulations were produced with GROMACS. M protein dimer equilibration was performed as described in the previous section. MD simulations were performed with an isothermal–isobaric ensemble. Temperature coupling was done using a Nose-Hoover thermostat with a time constant of 1 ps. In order to maintain a constant pressure, a semi-isotropic Parrinello–Rahman barostat was used with a time constant of 5 ps and compressibility of 4.5 × 10-5 bar-1. Electrostatic interactions were performed with fast smooth Particle-Mesh Ewald, with a cutoff of 1.2 nm and Hydrogen bonds were constrained using the linear constraint solver.

Dimer system RMSD and RMSF calculations were performed using Cα atoms with GROMACS package. CCA, which tracks the movements of two or more sets of time series data relative to one another, was performed using the Bio3D R package[54] based on the Cα atoms. SASA analysis for each residue was performed with the GROMACS package. SASA analyses were performed for the dimer complex ($SASA_{complex}$) and each monomer separately ($SASA_{monomerA}$ and $SASA_{monomerB}$), and ΔSASA was calculated for each residue as $SASA_{complex} – (SASA_{monomerA} + SASA_{monomerB})$. ΔSASA values provide another quantitative measure of conformational change upon protein coupling. To further understand the behavior upon complex formation, we also calculated relSASA for each residue that comes from the quotient between ΔSASA and $SASA_{monomer}$. To detect possible interacting residues, a structure was retrieved every 2 ns, totaling 100 structures from 300 ns until 500 ns, for each replica. These structures were then submitted to an *in-house* script that detected residues for which side chains were within 5 Å of each other, using a 90% prevalence time as a cut-off.

# M protein mutation analysis

Genome and protein sequences for this study were obtained from the GISAID[35] database (Accession Numbers were listed at Supplementary Information) and are available upon request on https://www.gisaid.org. MicroGMT[31], a python package, was developed, optimized, and used for SARS-CoV-2 M gene mutation analysis, to track indels and SNPs. This software requires raw or assembled genome sequences and works through database comparison to detect genomic mutations. Only non-synonymous SNPs at the M gene region for predicted interacting residues were considered for further studies. For M protein sequence mutation analysis, we used the Rahman *et al.* approach that works through pairwise analysis and comparison[32]. This method uses Multiple Sequence Alignment (MSA) and pairwise alignments to detect mutations in large datasets in a fast and accurate manner and has also been used in other studies regarding different SARS-CoV-2 proteins. Both of these tools were used with default parameters and all available sequences were compared against a reference, the first SARS-CoV-2 genome sequenced (NC_045512.2).

To determine the impact of mutations in M protein dimer stability, Gibbs energy difference was calculated using FoldX[33], an empirical force field. This approach evaluates the impact of mutations in protein stability through free energy variation ($\triangle\triangle G_{binding} = \triangle G_{mutant} - \triangle G_{WT}$) between mutant protein and reference protein, taking into account contributions from hydrophobic, polar, Van der Waals, hydrogen bonds and electrostatic interactions[33]. In order to avoid considering mean $\triangle\triangle G_{binding}$ values close to zero as relevant for protein stability, we established a low (below -0.5 kcal/mol) and high cut-off off (above 0.5 kcal/mol). Results for this step were analyzed taking into account residues polarities, both for the WT and mutated proteins, as well as splitting residues by aromaticity, as both these characteristics have a major impact on protein-protein interactions. 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 F, W and Y were considered as aromatic.

All presented structure images were produced with Protein Imager[55], ggplot2 R package[56]

and Bio3D R package[54].

# Data Availability

The genomic datasets analyzed during the current study are freely available in the GISAID repository, https://www.gisaid.org/, and Accessions Numbers are available at Supplementary Information. GISAID has an application procedure for obtaining access to the data, which should be followed for any researcher that wants to use it. Detailed data analysis results are also available at Supplementary Information. Any material requests should be addressed to ISM: irina.moreira@cnc.uc.pt.

## Declarations

# Acknowledgements

We gratefully acknowledge the Authors from all the Originating laboratories responsible for obtaining the specimens and the Submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based (listed at Supplementary Information). All submitters of data may be contacted directly via www.gisaid.org).

# Competing interests

The authors declare that they have no competing interests.

# Author contributions

ISM conceived the presented idea. CM-P and MNP performed necessary computations and carried out the main experiments. NNP contributed to docking analysis and RPG performed MD calculations. CM-P and MNP wrote the manuscript, with the help of NNP and ABC, and under NR-F and ISM supervision. All authors discussed the results and contributed to the final manuscript. All authors have read and agreed to the published version of the manuscript.

# Funding

This work was funded by COMPETE 2020 - Operational Programme for Competitiveness and Internationalization and Portuguese national funds via FCT - Fundação para a Ciência e a Tecnologia, under projects POCI-01-0145-FEDER-031356, UIDB/04539/2020, and DSAIPA/DS/0118/2020. NR-F and CM-P were also supported by FCT through Ph.D. scholarships PD/BD/135179/2017 and 2020.07766.BD (DOCTORATES 4 COVID-19), respectively. ABC and RPG were supported by scholarships PTDC/QUI-OUT/32243/2017 and PTDC/QUI-NUC/30147/2017, respectively. Authors also acknowledge FCT, Advanced Computing Project DSAIPA/DS/0118/2020 and LCA (Laboratório de Computação Avançada da Universidade de Coimbra).

## References

1. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. (2021).

2. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).

3. Wang, M.-Y. *et al.* SARS-CoV-2: Structure, Biology, and Structure-Based Therapeutics Development. *Front. Cell. Infect. Microbiol.* **10**, 587269 (2020).

4. GISAID - Clade and lineage nomenclature aids in genomic epidemiology of active hCoV-19 viruses. (2021).

5. SeyedAlinaghi, S. *et al.* Characterization of SARS-CoV-2 different variants and related morbidity and mortality: a systematic review. *Eur. J. Med. Res.* **26**, 51 (2021).

6. Hamed, S. M., Elkhatib, W. F., Khairalla, A. S. & Noreddin, A. M. Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology. *Sci. Rep.* **11**, 8435 (2021).

7. Khailany, R. A., Safdar, M. & Ozaslan, M. Genomic characterization of a novel SARS-CoV-2. *Gene Rep* **19**, 100682 (2020).

8. Bianchi, M. *et al.* Sars-CoV-2 Envelope and Membrane Proteins: Structural Differences Linked to Virus Characteristics? *Biomed Res. Int.* **2020**, 4389089 (2020).

9. 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).

10. 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).

11. Satarker, S. & Nampoothiri, M. Structural Proteins in Severe Acute Respiratory Syndrome Coronavirus-2. *Arch. Med. Res.* **51**, 482–491 (2020).

12. Neuman, B. W. *et al.* A structural analysis of M protein in coronavirus assembly and morphology. *J. Struct. Biol.* **174**, 11–22 (2011).

13. 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).

14. Mariano, G., Farthing, R. J., Lale-Farjat, S. L. M. & Bergeron, J. R. C. Structural Characterization of SARS-CoV-2: Where We Are, and Where We Need to Be. *Front Mol Biosci* **7**, 605236 (2020).

15. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **20**, 681–697 (2019).

16. Aslam, B., Basit, M., Nisar, M. A., Khurshid, M. & Rasool, M. H. Proteomics: Technologies and Their Applications. *J. Chromatogr. Sci.* **55**, 182–196 (2017).

17. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).

18. 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).

19. K. Hofmann, W. S. TMbase-a database of membrane spanning proteins segments. *Biol. Chem. Hoppe Seyler* **374**, 166 (1993).

20. 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).

21. 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).

22. Buchan, D. W. A. & Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407 (2019).

23. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).

24. Dobson, L., Reményi, I. & Tusnády, G. E. The human transmembrane proteome. *Biol. Direct* **10**, 31 (2015).

25. Dobson, L., Reményi, I. & Tusnády, G. E. CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res.* **43**, W408–W412 (2015).

26. 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).

27. 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).

28. Bekker, H. *et al.* Gromacs-a parallel computer for molecular-dynamics simulations. *4th International Conference on Computational Physics (PC 92)* 252–256 (1993).

29. 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).

30. van Zundert, G. C. P. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **428**, 720–725 (2016).

31. 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).

32. 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.

33. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382-8 (2005).

34. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Challenges* **1**, 33–46 (2017).

35. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, (2017).

36. 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).

37. 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).

38. Munteanu, C. R. *et al.* Solvent Accessible Surface Area-Based Hot-Spot Detection Methods for Protein–Protein and Protein–Nucleic Acid Interfaces. *Journal of Chemical Information and Modeling* vol. 55 1077–1086 (2015).

39. 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).

40. 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).

41. Bogan, A. A. & Thorn, K. S. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1–9 (1998).

42. Majumdar, P. & Niyogi, S. SARS-CoV-2 mutations: the biological trackway towards viral fitness. *Epidemiol. Infect.* **149**, e110 (2021).

43. AlQuraishi, M. Machine learning in protein structure prediction. *Curr. Opin. Chem. Biol.* **65**, 1–8 (2021).

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

45. O'Donnell, V. B. *et al.* Potential Role of Oral Rinses Targeting the Viral Lipid Envelope in SARS-CoV-2 Infection. *Function* **1**, (2020).

46. 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).

47. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

48. Thomas, S. The Structure of the Membrane Protein of SARS-CoV-2 Resembles the Sugar Transporter SemiSWEET. *Pathog Immun* **5**, 342–363 (2020).

49. 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).

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

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

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

53. 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).

54. 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).

55. 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).

56. Wilkinson, L. ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. *Biometrics* vol. 67 678–679 (2011).

# Tables

**Table 1:** 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 (Å$^2$) | $_{rel}$SASA A | Monomer B | ΔSASA B (Å$^2$) | $_{rel}$SASA 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 ± 14.7033 ± 12 | 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 |

# Figures

**Figure 1**

Project Pipeline. M protein structure was predicted by AlphaFOLD[17]. Membrane orientation was predicted with Orientations of Proteins in Membranes (OPM)[18], prediction of Transmembrane Helices (TMpred)[19], TransMembrane prediction using cyclic Hidden Markov Model (TMHMM)[20,21], Prediction of secondary structure (PSIPRED)[22,23], Consensus Constrained TOPology prediction (CCTOP)[24,25] and Sequence Analysis & Consulting Service MEMbrane protein Structure And Topology (SACSMEMSAT)[26]. Protein-membrane systems were constructed with Chemistry at HARvard Macromolecular Mechanics Graphical User Interface (CHARMM-GUI)[27] and minimization and equilibration were conducted using GROningen MAchine for Chemical Simulations (GROMACS)[28,29]. M protein dimer was predicted with High Ambiguity Driven protein-protein DOCKing (HADDOCK)[30] and results were compared to SARS-CoV experimental data. Gene and protein mutations were analyzed with Microbial Genomics Mutation Tracker (MicroGMT)[31] and Rahman *et al*[32] programs and energy variation of mutations in dimer interaction residues were calculated with FoldX[33].

## Figure 2

SARS-CoV-2 M protein monomer. **a)** M protein domains predicted by TMHMM[20,21] membrane predictor. **b)** TMHMM[20,21] M protein monomer structure prediction after equilibration in membrane with ER membrane composition. **c)** M protein structure with domains highlighted.

## Figure 3

SARS-CoV-2 M protein dimer HADDOCK[30] prediction using TMHMM[20,21] 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).

## Figure 4

**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.

**Figure 5**

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.

**Figure 6**

$\triangle\triangle G_{binding}$ values of predicted interfacial residues with major impact in protein stability. Color represents the alteration from 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).

**Figure 7**

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

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MarquesPereiraPiresSI.pdf