# A comprehensive SARS-CoV-2 genomic analysis identifies potential targets for drug repurposing

**Nithishwer Mouroug Anand[1], Devang Haresh Liya[1], Arpit Kumar Pradhan[2,3*], Nitish Tayal[4], Abhinav Bansal[5], Sainitin Donakonda[6], Ashwin K. Jainarayanan[7*]**

**[1]Department of Physical Sciences, Indian Institute of Science Education and Research, Mohali, India**
**[2]Graduate School of Systemic Neuroscience, Ludwig Maximilian University of Munich, Germany**
**[3]Klinikum rechts der Isar, Technische Universität München, Germany**
**[4]The Unique Tutorials, Kharghar, Navi Mumbai, India**
**[5]Department of Chemical Sciences, Indian Institute of Science Education and Research, Mohali, India**
**[6]Institute of Molecular Immunology and Experimental Oncology, Klinikum rechts der Isar, Technische Universität München, Germany**
**[7]Interdisciplinary Bioscience DTP, University of Oxford, Oxford, UK**

***Correspondence: Arpit.Pradhan@campus.lmu.de, ashwin.jainarayanan@dtc.ox.ac.uk**

## Abstract

**Background:** The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) which is a novel human coronavirus strain (HCoV) initially reported in December 2019 in Wuhan City, China causing pneumonia-like symptoms and other respiratory tract illness. It's higher transmission and infection rate has successfully enabled it to have a global spread over a matter of small time. With 6,529,240 cases and about 385,264 deaths, this pandemic has become a global concern with certain drugs and vaccines failing at later clinical trials.

**Materials and Methods:** Phylogenetic Analysis, Haplotype Network, Analysis of conserved genes and population-level variants, Using conserved genes as targets for drug designing, Docking studies and Molecular Dynamics (MD) simulations to predict the stability of Drug-Ligand Complex.

**Results:** We identified the most common haplotypes from the haplotype network and at least seven different clusters were found signifying seven different viral lineages across the globe. We studied the mutation frequency across the SARS-CoV-2 viral genome. The conserved genes and population level variants were analyzed and NSP10, Nucleoprotein, Plpro and 3CLpro which were conserved at the highest threshold were used as drug targets for molecular dynamics simulations. Darifenacin, Nebivolol, Bictegravir, Alvimopan and Irbesartan are among the potential drugs which are suggested for further pre-clinical and clinical trials.

**Significance:** This particular study provides a comprehensive targeting of the conserved genes as a novel approach for drug targeting. The conserved gene approach could also be of a big use while designing vaccines and cure. Mutations in the viral genome make the designing of the drugs a challenging task which has a higher risk of failure at later clinical trials. This approach of targeting the stable genes for drug discovery would provide a better therapeutic approach and confidence in the successive clinical trials. We also identified the global level spread of SARS-CoV-2 and mutation frequencies across the viral genome. Our study gives insights of the origin and global spread of the SARS-CoV-2. The data provided in this study can further be used by other groups to understand and combat Covid 19.

**Keywords:** Covid 19, Coronavirus, SARS-CoV, SARS-CoV-2, Haplotype network, Phylogenetic Analysis, MD simulation, Drug-Ligand interactions, Mutation frequency, Docking, Population level variants

## Introduction

The 2019 novel coronavirus strain (2019-nCoV, later officially named SARS-CoV-2) which was initially reported in Wuhan, Hubei Province, People's Republic of China (PRC) belongs to the coronaviridae family of viruses that possess a positive-sense single-stranded RNA genome [1] [2]. Compared to the previous outbreaks of severe acute respiratory syndrome coronavirus (SARS-CoV) in 2003 and Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012, 2019-nCoV has higher transmission and infection rate with an increasing mortality rate [3]. The SARS-CoV-2 genome like other members of the betacoronavirus family has a long ORF1ab polyprotein at the 5' end, which is followed by a set of four major structural proteins, including the spike surface glycoprotein, small envelope protein, matrix protein, and nucleocapsid protein (Figure 1) [4] The 2019-nCoV strain and SARS-CoV share a genome sequence homology of about 79%. The 2019-nCoV has a greater similarity to the SARS-like bat CoVs (MG772933) than the SARS-CoV [1]. The high similarity of receptor-binding domain (RBD) in Spike-protein and several other analyses reveals that SARS-CoV-2 uses angiotensin-converting enzyme 2 (ACE2) as receptor, just like SARS-CoV. Coronavirus via the S protein on the surface identifies the corresponding receptor on the target cell thereby making its entry into the host cell [5]. The higher transmissibility and infection rate of 2019-nCoV as compared to SARS-CoV is attributed to the higher binding affinity of SARS-CoV-2 to the ACE2 receptors [6] [7]. In one of the structure model analysis, SARS-CoV-2 showed a 10-fold higher binding affinity for ACE-2 as compared to that of SARS-CoV [7]. The similarity of sequences between SARS-CoV-2 and SARS-CoV allows utilization of the known protein structures to quickly build a model for drug discovery on this new SARS-CoV-2. In addition, genomic studies of COVID-19 samples have also proven

instrumental in preventing a second wave of the epidemic in countries like New Zealand and The United Kingdom where lockdown restrictions have eased [8]. A comprehensive genomic study could identify the start of community spread immediately and could help in imposing restrictions that could prevent a second wave [9].

As of June 4, 2020, a total of 6,529,240 cases of COVID-19 occurring in at least 213 countries and territories were reported, with approximately 5.9% of fatality rate (385,264/6,529,240). The coronavirus similar to other RNA viruses is characterized by significant genetic variability and high recombination rate which boosts them to be easily distributed among humans and animals in different geographic locations [10]. Numerous coronavirus strains exist within the human and animal populations without causing life threatening diseases [11] However in certain rare cases there is genetic recombination of viruses which produces infectious strains which are pathogenic to humans [12]. What makes SARS-CoV-2 more powerful is the mutation events that allow structural changes in the virus. One of the recent studies suggests the existence of three central variants of SARS-CoV-2 distinguished by amino acid changes [13]. We urgently need therapeutic options to combat this virus infection.

In this study we performed wide array analysis to systematically identify drug targets. Firstly, we performed phylogeny and haplotype analysis which approximately identified seven different clusters based on the haplotype network suggesting the presence of seven different variants of SARS-CoV-2. We also found the genes that are conserved and the population level variants. In this study, we also highlight the mutation frequencies across the viral genome. We then identified the stable genes which have stretches of conserved regions and thereby can be used as efficient Drug-targets. Using this as our base, we identified 4 genes which are stable and conserved in all the strains. We used them as our targets in in-silico drug designing, molecular docking and molecular dynamics simulations. Given the fast mutation rate of these viruses our approach of targeting the stable genes through small molecules would provide a better therapeutic approach and confidence in the successive clinical trials. This study provides new insights into the evolution of COVID-19, identifies the divergence pattern and spread of the virus at the population level and utilises a unique and efficient method of targeting the stable genes for the drug discovery approach.

## Materials and Methods

The complete high throughput FASTA file for 363 nCOV2 viral genomes were downloaded from GISAID on March 17th (Global Initiative on Sharing All Influenza Data; https://www.gisaid.org/) with acknowledgment (Supplementary File 1). While the number of sequences has increased drastically over time, this set of 363 genomes represent the genome sequenced from diverse regions. Sequences and annotations of the reference genome of SARS-CoV-2 (NC_045512) and

other related viruses were downloaded from GenBank and GISAID. Among these 363 sequences, 358 were from humans, four were from pangolin and one from bat.

**Phylogeny Analysis**

Four reference sequences (KF294457.1, AY278489.2, MG772933.1, and MG772934.1) were added to the 358 sequences from humans and the alignment was done with Mafft [14] and MEGA software [15] (version 7.450). The phylogenetic tree was computed using the Neighbour-Joining(NJ) method. A bootstrap value of 1,000 replicates was applied to yield a robust phylogeny. The tree was rooted at KF294457.1 and visualized using the Interactive Tree of Life [16]. The branch lengths of the tree are not proportional to the phylogenetic distance.

**Haplotype Network**

DnaSP v6.12.03 was used to define sequence sets and generate multi-sequence aligned haplotype data in nexus file format [17]. The trait segment was included in the nexus data file for visualisation and drawing of haplotype networks based on the haplotypes generated by the DnaSP. PopART v1.7 was used to draw the haplotype network based on the haplotype generated by DnaSP [18].

**Conserved Gene and Population level Variants**

The 358 sequences from humans were aligned using online MAFFT's closely related viral genome alignment tool [14] with the reference sequence NC_045512. FFT-NS-fragment method was used for alignment with the parameters --reorder --adjustdirection --keeplength --mapout --anysymbol. Default gap penalty of 1.53 and offset value of 0.0 was used. The number of mutations were counted for each nucleotide position using NC_045512 as reference. The ambiguous bases, Ns and gaps were not treated as mutations.

A sliding frame method was used to identify the conserved genes across all the sequences. The programs used for this analysis are publicly shared on GitHub (https://github.com/DevangLiya/CRAM). A master sequence consisting of 1 for the nucleotide position that is conserved across all the genomes and 0 for the nucleotide position that is not conserved was produced for the given alignment. A frame of size 100 was moved across the entire length of this master sequence and each instance of frame was given a score between 0 and 100 to represent the level of conservation by counting the number of 1s in that frame. Starting position of every frame between the given threshold is reported. The nucleotide

sequence corresponding to the conserved frames were reconstructed by adding 100 nucleotides (equal to the frame length) to the reported positions and the sequence was then BLASTed to get the corresponding genes [19]. A few more nucleotides were added on the both ends of the sequence when BLAST did not yield any satisfactory match. The dataset of 358 sequences was divided into the eight population level datasets consisting of China, Japan, Asia (India, Singapore, Cambodia, Nepal, Vietnam, Taiwan, Hong Kong, Thailand, South Korea), Europe (France, Finland, Netherlands, Czech Republic, Switzerland, Italy, Portugal, Germany, Luxembourg, Sweden, Belgium), UK (England, Wales, Ireland), North America (USA and Canada), Oceania (Australia and New Zealand), and Rest of America (Mexico, Chile, Brazil). These sequences were then aligned and visualized in MEGA.

**Protein Structure Modelling**

Complete 3D structures for the four stable proteins (NSP10, Nucleoprotein, Plpro and 3CLpro) of SARS-CoV-2 selected from the above analysis are currently not available. We therefore tried to construct the 3D structures by homology modeling. The sequence of the proteins were blasted against the NCBI database and the suitable templates were selected for the modeling. The 3D structure models for the proteins screened were modeled by comparative protein modeling methods using the SWISS-MODEL server (http://swissmodel.expasy.org) [20]. The structure-based alignment obtained were used and SWISS-MODEL was used in the optimized mode to minimize energy. Models are made according to the target template alignment and the per-residue and the global model quality was accessed using the QMEAN and Global Model Quality Estimate(GMQE) scoring functions. The GMQE score gives an estimate of accuracy of tertiary structure of the protein models. The Q-Mean on the other hand gives an impression of the quality of the submitted model based on its physicochemical properties and then generates a value referring to the overall quality of the structure.

**Validation of Models**

RAMPAGE Ramachandran plot analysis was used for verification of 3D structures. It provides the number of residues in the favored, allowed, and outlier region [21]. If a good proportion of residues lie in the favored and allowed region, then the model is predicted to be good. The quality of the models were also accessed using PROSA, PROCHECK and Verify 3D [22] [23] [24]. Both PROCHECK and RAMPAGE analyze the stereochemical quality of the submitted models based on its phi/psi angle arrangement and then generates Ramachandran plots which highlights the percentage of residues in the favored, allowed or in outlier regions. If a greater proportion of the residues lie in the favored and allowed region then the model is considered to be good. ProSa on the other hand does a comparative analysis by calculating the potential

energy of the protein models and comparing them to the experimental structures deposited in the PDB. The Z-Scores obtained from each model suggest that the structures are comparable to the NMR structures of similar size. Verify3D evaluates the local quality of the protein model on the basis of structure-sequence compatibility to generate a compatibility value for each residue of the protein. A model with 80% of their residues with a 3D-1D score equal to or higher than 0.2 is considered to be a high quality structure.

**Virtual Screening and Molecular Docking**

In order to perform a structure-based virtual screening e-LEA3D,( http://chemoinfo.ipmc.cnrs.fr/) which uses PLANTS (Protein-Ligand ANT System) algorithm was used. In order to find the binding site around a residue metaPocket 2.0 software was used [25]. The virtual screening was done on the basis of docking with the list of FDA approved drugs. The drug-like property and the lipinski's rule of 5 was further used to filter the most suitable drugs. The docking score provided by the e-LEA3D was used to screen the drugs for further analysis.

The ADME (Absorption, Distribution, Metabolism, and Excretion) analysis was performed using the SwissADME(http://www.swissadme.ch/) [26] web server to compute the physicochemical descriptors as well as to predict ADME parameters, pharmacokinetic properties. The toxicity analysis for each drug was performed using the ProTox software. After the ADMET analysis the drugs were then moved to the next phase for MD simulations [27].

**Molecular Dynamic simulations**

The unbound proteins and Protein-drug complexes were subjected to MD simulation for 25ns to mimic the physiological state of protein molecules. The simulation was performed with GROMACS 2019 (M.J. Abraham) utilizing the GROMOS96 43a1 force field parameters [28]. The topologies of the drug molecules were modeled using the PRODRG web server [29]. The system was made electrostatically neutral by adding counter ions and the complexes were solvated within 10 SPC/E water cube [28] [30]. The whole system was then energy minimized in multiple steps using the steepest descent method. The temperature of the entire system was raised up to 300 K for a time scale of 100ps. Two different phases of equilibration were performed-first with constant pressure and temperature (NPT) and the other with steady volume and temperature (NVT) [31]. The trajectory file of simulated system was then used for calculation of various structural parameters like the Root Mean Square Deviation (RMSD), Root Mean Square Fluctuations (RMSF), Radius of Gyration (Rg), Intermolecular Hydrogen Bonding (H-bonding) and Solvent-Accessible Surface Area (SASA) to understand the structural behavior of the protein-drug complexes [32].

## Results and Discussions

### Phylogeny Analysis

This viral network is a snapshot of the early stages of an epidemic before the phylogeny becomes obscured by subsequent migration and mutation. The question may be asked whether the rooting of the viral evolution can be achieved at this early stage by using the oldest available sampled genome as a root. The described core mutations are considered reliable as they have been verified and validated by a variety of contributing laboratories and sequencing platforms. The phylogeographic patterns in the network are potentially affected by distinctive migratory histories, founder events, and sample size. Nevertheless, it would be prudent to consider the possibility that mutational variants might modulate the clinical presentation and spread of the disease. The phylogenetic classification provided here may be used to rule out or confirm such effects when evaluating clinical and epidemiological outcomes of SARS-CoV-2 infection, and when designing treatment and, eventually, vaccines. We took the phylogenetic data submitted to GISAID and found out that Bat SL-CoV KF294457.1, MG772933.1 and MG772934.1 are the closest relative to 2019-nCoV and used them for rooting the phylogenetic tree. AY278489.2, a human SARS COV was taken as a distant relative [33]. Interestingly, although the differences are very small, the closest strain to the ancestor strain is not the one from Wuhan but from Germany and Finland (Figure 2). Shenzhen and Guangdong were the next before Wuhan could come. This can happen in 3 scenarios: the first scenario is that the virus from Wuhan mutated in such a manner that it came closer to the ancestral strain while traveling to Guangdong, Shenzhen, Finland and Germany. The next in the series is a strain from the Cruise ship Diamond Princess that went to the USA indicating that the strain in Wuhan and the one in the Cruise Ship of USA are closely linked. Further mutations resulted in the strains in Asia and rest of Europe along with Brazil, Mexico and Chile. The viruses in South Korea and Japan have even more mutations. The second scenario is very unlikely and it suggests that the virus originated in Germany and then mutated as it went to other places. In the second scenario, the virus would have infected Finland, Shenzhen, Guangdong and Wuhan in the starting and then it spread everywhere. This does not match with the historically reported cases and thus is unlikely. The third scenario is that all the sequences from Wuhan are not documented missing the most ancestral genome.

### Seven Viral clusters identified via haplotype network

In order to understand the population level divergence of SARS-CoV-2 we tried to map the haplotype network and establish the relationship among the SARS-CoV-2 haplotypes from the genome data collected all over the globe. A total of 194 haplotypes were identified from 358 SARS-CoV-2 genomes. Haplotype 1 had the highest prevalence and was present in diverse

geographical locations (Figure 3). These haplotypes could be roughly classified into 7 different clusters signifying at least 7 distinct strains of the virus present over the globe. The main central hub consists of around 40-45% contribution from China followed by USA and Europe. However, the haplotype from the USA remains mostly inside the USA. It didn't spread much indicating that the infected population in the USA are majorly isolated from the rest of the world. However, the haplotype from China and Europe spread everywhere indicating more connectivity of these 2 regions with the rest of the world. The haplotype in the USA came from China possibly through the cruise ship Diamond Princess via Japan and then spread in the USA giving rise to all other haplotypes in the USA (Figure 3). Almost all the ancestral haplotypes in the USA correspond to cruise ship Diamond Princess.

**Identification of Conserved Genes and mutation frequency across viral genome**

To determine conserved regions we performed systematic sequence analysis which identified the conserved genes with different threshold conservation levels (Table 1). The population variant genes were also identified and highlighted on the basis of their geographic distribution (Table 2). Nucleotide positions 240, 3036, 8781, 11082, 14407, 23402, 28143, with reference to NC_045512 sequence, had mutation frequencies greater than 40 (Figure 4). This represents the highly mutating positions in the genome which we call the "Hotspot Zones". These hotspot zones were distributed over the viral genome. Some of these zones lie in the NSP1, NSP3, NSP4, NSP6, NSP12, spike protein (S-protein) and ORF8 genes. For our further analysis we chose the proteins with the highest conservation thresholds. NSP10, Nucleoprotein, PLpro, and 3CLpro were conserved targets which were chosen for drug targeting. Interestingly, Japan had the least number of variant genes whereas in Asia the population carried a diverse set of SNPs throughout the viral genome (Table 2). Similarly, China, Rest of America (Mexico, Chile, Brazil) and Europe had more number of variant genes as compared to other populations in the UK and North America. Orf1a polyprotein was found to be a variant in all the population (Table 2).

**Homology modeling of Stable targets and virtual screening of small molecules**

The three dimensional structure generated by SWISS-MODEL was checked for its quality based on several parameters (Table 3). For each of the proteins, the models were arranged with respect to the GMQE scoring functions and were checked for the local quality estimate and Z-scores. The protein models which were best fit in all these parameters were accessed further for their quality (Supplementary Figure 1-4).

All the four protein models had a greater proportion of residues in the favoured and allowed region in the Ramachandran Plot. Prose Analysis revealed that structures are in the X-Ray/NMR structure fold and also have a greater stereochemical quality (Supplementary Figure 1-4).

We used a Structure-Based drug designing and docking approach. We carried out the virtual screening of the drugs from the list of FDA approved drugs. In order to define the binding sites around the residue MetaPocket 2.0 metaserver was used to identify the ligand binding site on the protein surface. From the ligand binding site a binding site radius of 10 Å was defined and the docking was performed. The drugs which docked to the proteins with a higher docking score were considered for the further analysis. For each protein two drugs with highest scores were selected and were analysed further for the ADMET analysis and the MD simulations.

The SWISSADME server was employed to predict the pharmacokinetic effects and to predict the likeliness of the drugs. All the seven drugs employed had good drug like characteristics. The toxicity of the drugs was predicted by ProTox and was deemed to be non-toxic. We were particularly interested to look at the interaction of the drugs with the surrounding residues in the protein. The interaction of the drugs with the protein residues can be visualised in Figure 5-6. Taken together, our structure based approach identified good quality models of stable proteins in SARS-CoV-2 and potential small molecules against them.

**Molecular dynamics (MD) simulation**

Molecular Dynamics simulations are employed to study the strength and properties of the protein-drug complexes and their conformational changes on an atomic level. Various parameters such as RMSD, RMSF, Radius of Gyration, Intermolecular H-bonds, and SASA were calculated throughout the simulation trajectory to give insights on the structure of the proteins. To illustrate the dynamics, and conformational stability of the protein-drug complexes, the protein-drug complexes were subjected to MD simulations for a period of 25ns. The binding of the drugs Cilostazol and Elvitegravir destabilized the PLpro complex. Thereby Plpro was not short-listed for further downstream analysis. There were several interactions of Bictegravir and Nebivolol with the Nucleoprotein complex (Nucleoprotein-Bictegravir: Arg68, Gly124, Asn126; Nucleoprotein-Nebivolol: Pro67, Arg68, Tyr123, Ile131, Val133, Ala134). Alvimopan interacted with NSP10 at residues Asp82, His83, Phe89, Cys90, and Lys93 whereas Irbesartan had interactions with NSP10 at Cys74, His83, Pro84, Cys90, Leu92 and Leu112. While Darifenacin has some contacts with 3CLpro at Asn142, Asn214, Val303, Phe305, Nebivolol interacted with the 3CLpro at Lys751 and Thr763 residues (Figure 5, 6). The results of the MD simulations are summarised in table 4 provided below (Table 4). A superimposition of the protein-ligand complexes before and after the simulation has been provided below (Figure 7).

**An overview of the proteins chosen for MD simulation**

The proteins that were found to be conserved from the previous analyses were studied in detail. The interaction map of these SARS-CoV-2 proteins from the study by Gordon et al., 2020 reveals targets for drug repurposing [34].

Nucleoprotein:

The nucleoprotein (N-Protein) is a highly charged, multifunctional, basic protein of 422 amino acids which binds to the viral RNA during the virion assembly and leads to formation of the helical nucleocapsid [35]. The N protein and spike protein (S-protein) are encoded by all coronaviruses. The nucleocapsid (N) protein of COVID-19 has nearly 90% amino acid sequence identity with SARS-CoV [36]. However, we observed that the spike protein is not conserved in different variants of SARS-CoV-2 above 90% threshold. The N protein forms complexes with genomic RNA and creates a capsid around the enclosed nucleic acid [35]. It also assists in RNA synthesis and affects the host cell responses such as cell cycle and translation [37]. It plays an important role in virion assembly and enhances the efficiency of the virus transcription and assembly [37]. The interaction map of N-protein reveals that the N-protein interacts with human protein that are responsible for RNA processing and Stress Granule Regulation [34]. This indicates that similar to the N-protein of SARS-CoV, The N-protein of SARS-CoV-2 also plays an important role in suppressing the RNA interference (RNAi) to overcome the host defense. Previous studies have shown that 15 human proteins interact with the N-protein of SARS-CoV-2 [34]. Out of the 15 human proteins interacting with the N-protein, CSNK2B, CSNK2A2 and LARP1 might be plausible drug targets.

3CLpro

3-chymotrypsin-like cysteine protease (3CLpro) or the NSP5 is also a non structural protein encoded by ORF1a/1b. The SARS-CoV2 replication process involves a series of proteolytic cleavage of the polypeptide to generate various proteins [7]. Among these serial cleavages, the 3CL protease plays a critical role at 11 distinct cleavage sites, and is vital for the replication of virus particles [4]. Further, 3CLpro's location at the 3' end causes the protein to express excessive variability which in turn makes it a potential target for COVID-19 drugs. The interaction map of 3CLpro reveals only one human protein-HDAC2, an enzyme responsible for removing the acetyl groups from lysine residues of core histones [34].HDAC2 plays an important role in regulating the epigenetic features and gene expression patterns in human cells. All of the above make 3CLpro a suitable target for anti-coronavirus drugs.

NSP10

NSP10 is one of the 16 non-structural proteins (NSP1–16) encoded by ORF1a/1b that comprise the RNA-synthesizing machinery of SARS-CoV2. The NSP10 subunit contains two zinc fingers and is known to interact with the NSP14 and NSP16 subunits to increase their 3'-5' exoribonuclease and 2'-O-methyltransferase activities respectively [38]. Existing literature indicates that this interaction between NSP10 and NSP14 subunits is crucial for the viral replication process as mutations in NSP10 that abolished the interaction are known to have yielded replication-negative virus [38]. The network map for NSP10 reveals that the protein interacts with several proteins responsible for endomembrane compartments and vesicle trafficking pathways [34]. Among these human-proteins are the AP2 (AP2A2 and AP2M1) proteins that are associated with clathrin-mediated endocytosis [34]. Interaction of NSP10 with these human-proteins are hypothesized to modify endomembrane compartments to favor coronavirus replication [34].

**Root Mean Square Deviation (RMSD)**

The Root Mean Square Deviation (RMSD) analysis is an important step towards measuring the stability of the protein-ligand complex. A stable RMSD indicates that the binding of the protein-drug complex does not cause any significant changes in the structure of the protein.

Nucleoprotein:

It is evident that the RMSD of the Free Nucleoprotein, Nucleoprotein-Bictegravir, and Nucleoprotein-Nebivolol has remained mostly stable throughout the simulation. The Free Nucleoprotein stabilized at around 6 ns and remained stable throughout the simulation except for a spike at around 17 ns and another at around 19 ns. The RMSD of the Nucleoprotein-Nebivolol complex on the other hand stabilized at around 8 ns and maintained stability throughout except for minor peaks between 15ns and 20ns. The RMSD of the Nucleoprotein-Bictegravir complex starts to stabilize a little later at around 10ns and remains stabilized except for a small dip at around 22ns (Figure 8 A).

3CLpro:

From the RMSD plot of 3CLpro, we can see that the free form of the protein has stabilized earlier in the simulation at around 2ns and remains stabilized till the end. The 3CLpro-Darifenacin and the 3CLpro-Nebivolol complex get stabilized at around 4ns and stay stabilized to the end except for a minor instability in the 3CLpro-Nebivolol complex at around 6ns (Figure 9 A).

NSP10:

Figure 10 A reveals that the RMSD of the Free NSP10 protein, NSP10-Alvimopan, and NSP10-Irbesartan complexes are stabilized. The RMSD of the Free NSP10 protein stabilizes at around 7ns and maintains stability until 25ns. The NSP10-Alvimopan complex attains stability at around 5ns and remains stable throughout the simulation barring a small dip at around 20ns. The NSP10-Irbesartan complex, on the other hand, reaches stability comparatively later at around 8ns and remains stabilized throughout. These results indicate that the drugs did not significantly influence the structural stability of the NSP10 protein. In particular, the NSP10-Alvimopan complex has an average RMSD that is very close to the RMSD of the drug-free form of NSP10 (Figure 10 A).

**Radius of Gyration(Rg)**

The radius of gyration is a key parameter of the Protein-Drug complex that is used to study the folding properties and conformations of the protein-drug complexes. A comparatively high radius of gyration value indicates that a protein molecule is packed loosely while a lower radius of gyration value indicates a protein structure that is more compact. A more compact protein indicates that the drug molecule has not significantly interfered with the folding mechanism of the protein.

NucleoProtein:
In the case of Nucleoprotein, the radius of gyration of Nucleoprotein-Bictegravir complex and the Nucleoprotein-Nebivolol complex is found to be close to that of the unbound protein. The average Rg value of Unbound Nucleoprotein, Nucleoprotein-Bictegravir complex and Nucleoprotein-Nebivolol complex is found to be 1.45 nm, 1.44 nm, and 1.46 nm respectively. However, this difference in the mean radius of gyration between drugs is not significant as they are well within the standard deviation of the respective complexes. The minor variations in radius of gyration can be attributed to the conformational changes that the protein-drug complex undergoes (Figure 8 B).

3CLpro Protein:
In the case of 3CLpro, the compactness of the protein is found to be unaffected by the binding of the drugs as they have similar radius of gyration. The average Radius of gyration value of Unbound 3CLpro Protein, 3CLPro-Nebivolol complex, and 3CLPro-Darifenacin complex is found to be 2.45 nm, 2.44 nm, and 2.46 nm respectively. The differences in the radius of gyration are well within the standard deviation of the respective proteins. We also observe a gradual decrease in Radius of gyration value of the protein-ligand complexes. This indicates that the

secondary structure of the protein is not significantly affected by the binding of the drugs (Figure 9 B).

NSP10:
The mean radius of gyration for the Free-NSP10, NSP10-Alvimopan complex, and NSP10-Irbesartan complex is found to be 1.37, 1.39, and 1.40 respectively. Although the mean radius of gyration indicates that the NSP10-Irbesartan and NSP10-Alvimopan complexes are not as compact as the Free-NSP10 complex. The radius of gyration plot (Figure 10 B) reveals that the final conformations of the Free-NSP10 and NSP10-Alvimopan complex have a very similar radius of gyrations. This indicates that the binding of Alvimopan has not affected the foldability of the protein. The binding of Irbesartan on the other hand slightly affects the foldability of the protein.

**Intermolecular Hydrogen Bonding**

The number of intermolecular hydrogen bonds is an important parameter that can be used to quantify the binding affinity between the protein and the drug molecule. The presence of a large number of H-bonds between protein and drug molecules signifies a strong binding between the molecules.

Nucleoprotein:
We observed the maximum number of 7 hydrogen bonds between the protein and drug in both the Nucleoprotein-Bictegravir complex and the Nucleoprotein-Nebivolol complex. The average value of intermolecular H-bonds is 3 for Nucleoprotein-Nebivolol complex while 2 for Nucleoprotein-Bictegravir complex. The significant number of hydrogen bonds shows that drug molecules have a high affinity towards the active site of Nucleoprotein (Figure 8 C).

3CLpro:
In the case of 3CLpro, the maximum number of intermolecular hydrogen bonds in the 3CLpro-Nebivolol complex and the 3CLpro-Darifenacin complex is found to be 6 and 3 respectively. The average number of intermolecular H-bonds for both 3CLpro-Nebivolol complex and 3CLpro-Darifenacin complex was found to be 1. Unlike the 3CLpro-Darifenacin complex where Hydrogen bonds can be observed since the start of the simulation, the hydrogen bonds in 3CLpro-Nebivolol complex start appearing only after 13ns. (Figure 9 C)

NSP10:
In the case of NSP10, both NSP10-Alvimopan and NSP10-Irbesartan complexes have a maximum of 3 Hydrogen bonds between the protein and the drug. The average number of intermolecular hydrogen bonds is 1 in the case of Irbesartan and less than one in the case of Alvimopan. This

indicates that the Drug-protein affinity is higher in the case of Irbesartan than in the case of Alvimopan (Figure 10 C).

**Root Mean Square Fluctuations (RMSF)**

Root Mean Square Fluctuations (RMSF) is a vital structural parameter that is used to quantify the flexibility and rigidity of the protein-drug complexes. Since the RMSF measures the deviations of residue from its initial position, it is also highly useful in exploring the conformational flexibility of the protein-drug complexes. In all three proteins, the RMSF at the binding sites was below 0.2nm. This indicates that the drugs kept close contact with their binding pockets during the MD simulations.

Nucleoprotein:
In the case of Nucleoprotein, we observed the highest fluctuations between 400-700 atoms stretch. The average RMSF values of Nucleoprotein, Nucleoprotein-Bictegravir complex and Nucleoprotein-Nebivolol were found to be 0.203nm, 0.219nm and 0.216nm respectively. Further, the RMSF of most residues of the protein is found to be stable below 0.3 nm thereby preserving the flexibility of the protein (Figure 8 D).

3CLpro:
In the case of 3CLpro, we observe high fluctuations throughout the protein chain in both free protein and protein-drug complexes. No major differences are observed in the RMSF profiles of the free protein and protein-drug complexes. The average RMSF values of 3CLpro, 3CLpro-Darifenacin complex, and 3CLpro-Nebivolol were found to be 0.163nm, 0.169nm and 0.162nm respectively (Figure 9 D). These mean RMSF value indicates that the binding of Darifenacin and Nebivolol preserve the flexibility of the protein.

NSP10:
The RMSF profile of NSP10 and its complexes reveal that the protein has high fluctuations between the 600 and 1000 atom loop. The overall RMSF profile of free NSP10 is found to be similar to that of the drug-complexes. The average RMSF of free NSP10, NSP10-Alvimopan and NSP10-Irbesartan was found to be 0.209, 0.194, and 0.176 respectively. This indicates that there might be a slight loss of flexibility from the binding of the drug molecules (Figure 10 D).

**Solvent Accessible Surface Area analysis (SASA)**

To better understand the solvent Hydrophobic and Hydrophilic behaviour of the protein-drug complexes, solvent accessible surface area analysis (SASA) was performed. These results

indicated that all the proteins-ligand complexes are well solvated after the binding of drug molecules.

The Solvent Accessible Surface Area analysis revealed that no major differences are observed in the SASA profiles of Nucleoprotein and its protein-drug complexes. The mean SASA values for the free Nucleoprotein, Nucleoprotein-Bictegravir complex and Nucleoprotein-Nebivolol complex were 73.959 $nm^2$, 74.005 $nm^2$, and 77.085 $nm^2$ respectively (Figure 8 E). The average SASA values of 3CLpro, 3CLpro-Darifenacin complex and 3CLpro-Nebivolol complex are found to be 232.278 $nm^2$, 237.846 $nm^{2,}$ and 234.738 $nm^2$, respectively (Figure 9 E). In the case of NSP10, the Free NSP10 Protein, NSP10-Alvimopan complex, and NSP10-Irbesartan complex are found to have average SASA values of 66.666 $nm^2$, 67.3879 $nm^2$, and 70.351 $nm^2$ respectively (Figure 10 E). No major differences are observed in the SASA profiles of these complexes.

**Side-Effects of the drugs chosen for targeting**

The drugs selected for repurposing are Alvimopan, Nebivolol, Darifenacin, Irbesartan and Bictegravir. 3CLpro is targeted by Darifenacin and Nebivolol, NSP10 is targeted by Irbesartan and Alvimopan and Nucleoprotein is targeted by Nebivolol and Bictegravir. Alvimopan a selective peripherally acting mu-opioid receptor antagonist is used for accelerating upper and lower gastrointestinal tract recovery after a bowel resection [39]. Nebivolol is a beta blocker that is used to treat hypertension and heart failure [40]. Bictegravir is an integrase inhibitor class virial drug that is used to treat HIV and other retroviral diseases [41]. Irbesartan is an angiotensin receptor blocker that is used to treat hypertension and to protect the kidneys from damage due to diabetes [42]. Irbesartan is also used to prevent heart attacks by lowering blood pressure [43]. Darifenacin is a medication to treat urinary incontinence [44]. Darifenacin interacts with the M3 muscarinic acetylcholine receptors, which mediate bladder muscle contractions [45]. The side effects of these drugs were analysed from the SIDER database of drugs and side effects (http://sideeffects.embl.de/about/). This revealed that the major side effects of these FDA approved drugs are Headache, Dizziness, Diarrhoea and Constipation. In addition to these, back pain and dry mouth were also observed in the case of Darifenacin.

## Conclusion

The genomic organisation of SARS-CoV-2 is similar to other beta-coronaviruses which consists of a 5'-untranslated region (UTR), a replicase complex (orf1ab) encoding non-structural proteins (NSPs), a spike protein (S) gene, envelope protein (E) gene, a membrane protein (M) gene, a nucleocapsid protein (N) gene, 3'-UTR, and several unidentified non-structural open reading frames [4]. The rapid transmission and infection rate of SARS-CoV-2 can be attributed to the high mutation rates. Nucleotide substitution has often been one of the most studied mechanisms in

the viral evolution process [46]. The mutation events in the viral genome contribute to structural changes in the proteins thereby making it a difficult therapeutic target. This is one of the essential parameters which needs to be reconsidered in the drug-development process for a successful and effective design of therapeutics.

Understanding the phylogeny and the analysis of the mutations was thereby an important step to design effective methods to identify the list of conserved genes and the population level divergence of the genes. Report suggests a closely related bat coronavirus strain with 96.2% similarity to that of the SARS-CoV-2 [47]. The three scenarios predicted in this paper needs to be investigated further in order to know about proper ancestry and origins of these lineages.

There have been several studies to find out the variant strains of the COVID-19 on a population level. In one of the recent studies the SARS-CoV-2 was found to have 3 central variants distinguished by amino acid changes [13]. In our study we tried to construct the haplotype network and qualitatively analyze the haplotypes to understand the population level divergence of the SARS-CoV-2. We observed that one of the haplotypes which correlates to the previous sequences obtained from Wuhan was the most divergent and was present across the globe at a higher frequency of occurrence. From the haplotype network it was evident that there were at least seven different lineages of SARS-CoV-2.

This warranted an investigation into detailed perspective at the conserved genes and the population level variant genes. Based on different threshold levels of conservation, we divided them into different groups. NSP10, Nucleoprotein, PLpro and 3CLpro were some of the most conserved genes. These genes were chosen for drug screening, docking and molecular dynamic simulation. However, PLpro was not stable throughout the simulations. So, it was not used in further analysis. This particular approach gave a better edge in the drug development process and serves as a better therapeutic method. We also studied the mutations in the SARS-CoV-2 viral genome and analyzed the population level variant genes which can be found in Table 2. Nucleotide positions 240, 3036, 8781, 11082, 14407, 23402, 28143 were the hotspot zones with mutation frequencies greater than 40. Some of these residues are in the NSP1, NSP3, NSP4, NSP6, NSP12, spike protein (S-protein) and ORF8 genes.

Chymotrypsin-like (3C-like protease, 3CLpro) and papain-like protease (PLP) are the non-structural proteins which are key players in the proteolytic processing of viral polyproteins essential in the viral replication and can inhibit the host innate immune responses [48]. Both Darifenacin and Nebivolol docked to the 3CLpro complex with a higher docking score and were further analyzed for their interaction with the 3CLpro protein. The average RMSD of the Drug-3CLpro complex is similar to that of the free protein. Both 3CLpro-Darifenacin and 3CLpro-Nebivolol get stabilized around 4ns which shows the structural stability of the complex. Both the

intermolecular H-bond analysis and the SASA analysis suggested that the proteins-drug complexes were stabilised after the binding of drug molecules to their active sites. Nucleoprotein was also conserved and its role in the virion assembly makes it a great therapeutic target. The RMSD values of Nucleoprotein-Drug complexes indicated that the complex remained stable throughout the run. A smaller Rg value for the Nucleoprotein-Bictegravir complex as compared to the Nucleoprotein-Nebivolol complex suggested a more compact binding of the protein with the Bictegravir. The higher number of the intermolecular hydrogen bonds of both the drugs with their respective protein complexes shows the high affinity of the drug to the active site of the Nucleoprotein. A similar SASA profile of the Protein-Drug complex to the free protein suggests structural stability after binding of the drugs. NSP10 has been shown to interact with NSP14 and this interaction most likely enables coronaviruses to reliably replicate their long RNA genome [38]. The drugs Alvimopan and Irbesartan which had higher docking scores during virtual screening were selected for the MD simulations. Both Alvimopan and Irbesartan make the NSP10 complex stabilised as evident from their MD analysis. The binding of cilostazol and elvitegravir did not stabilize the PLpro complex. We thereby suggest Darifenacin, Nebivolol, Bictegravir, Alvimopan and Irbesartan as potential drugs for the clinical trials against SARS-CoV-2. Further, a BLAST search of human proteins with the selected SARS-CoV-2 proteins indicates that there are no human proteins that are similar to shortlisted viral proteins minimising the off target binding of the drugs.

The SARS-CoV-2 pandemic was declared a Public Health Emergency of International Concern (PHEIC) by the World Health Organisation (WHO) on 30th January, 2020. Since then there have been several studies regarding drug designing and appropriate pre-clinical and clinical trials for drugs and vaccines. This particular study finds its significance in utilising the conserved genes as stable targets for drug designing which gives a greater confidence while testing the drugs in the clinical trials. The drugs Darifenacin, Nebivolol, Bictegravir, Alvimopan and Irbesartan targeted the stable genes 3CLpro, Nucleoprotein and NSP10 and were shown to stabilize the Drug-Protein complex in MD simulations. We also find the mutation frequency across the viral genome, the conserved genes and the population level variant genes which would greatly benefit the designing of vaccines and cure for SARS-CoV-2. Our haplotype network gives an impression of seven different viral strains spread across the globe with different frequencies and phylogenetic tree raises concerns about its origin. The drugs reported in this paper can be further analysed and used as an antiviral drug against SARS-CoV-2 upon further downstream analysis and appropriate clinical trials.

## Conflict of interest

The authors declare no conflict of interest.

## Funding

The authors did not receive funding from any source for this work.

## Acknowledgement

We would like to thank GISAID (Acknowledgement table in the supplementary file) for the database of SARS-CoV-2 genome sequences. The authors would like to thank Shubham Kumar Sinha and Mridula for their help with global conservation analysis.

## Authors contribution

AKJ and AKP conceptualized and designed the project. AKP and DHL performed the phylogeny analysis. AKP and NT constructed the haplotype network. DHL performed mutation frequency and conserved gene analysis. DHL and AB performed population wise variant gene analysis. AKP performed the docking analysis and the drug interaction study. NMA worked on MD simulation and its analysis. AKP, NMA, DHL, AKJ and NT wrote the manuscript. NT and SD provided intellectual support in interpreting the results and editing the manuscript.

## Reference

[1] L. Wang, Y. Wang, D. Ye, and Q. Liu, "Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence," *Int. J. Antimicrob. Agents*, no. xxxx, p. 105948, 2020.

[2] M. S. Nadeem *et al.*, "Origin, potential therapeutic targets and treatment for coronavirus disease (COVID-19)," *Pathogens*, vol. 9, no. 4, pp. 1–13, 2020.

[3] D. Raoult, A. Zumla, F. Locatelli, G. Ippolito, and G. Kroemer, "Coronavirus infections: Epidemiological, clinical and immunological features and hypotheses," *Cell Stress*, vol. 4, no. 4, pp. 66–75, 2020.

[4] M. Tahir ul Qamar, S. M. Alqahtani, M. A. Alamri, and L. L. Chen, "Structural basis of SARS-CoV-2 3CLpro and anti-COVID-19 drug discovery from medicinal plants," *J. Pharm. Anal.*, no. xxxx, pp. 1–7, 2020.

[5] W. Tai *et al.*, "Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine," *Cell. Mol. Immunol.*, no. March, 2020.

[6] M. Hoffmann *et al.*, "SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor," *Cell*, vol. 181, no. 2, pp. 271-280.e8, 2020.

[7] S. Xia *et al.*, "Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion," *Cell Res.*, vol. 30, no. 4, pp. 343–355, 2020.

[8] Y. Z. Zhang and E. C. Holmes, "A Genomic Perspective on the Origin and Emergence of

SARS-CoV-2," *Cell*, vol. 181, no. 2, pp. 223–227, 2020.
[9] L. Caly *et al.*, "Isolation and rapid sharing of the 2019 novel coronavirus (SAR-CoV-2) from the first patient diagnosed with COVID-19 in Australia," *Med. J. Aust.*, no. March, pp. 459–462, 2020.
[10] Y. F. Tu *et al.*, "A review of sars-cov-2 and the ongoing clinical trials," *Int. J. Mol. Sci.*, vol. 21, no. 7, 2020.
[11] Z. W. Ye, S. Yuan, K. S. Yuen, S. Y. Fung, C. P. Chan, and D. Y. Jin, "Zoonotic origins of human coronaviruses," *Int. J. Biol. Sci.*, vol. 16, no. 10, pp. 1686–1697, 2020.
[12] M. Pérez-Losada, M. Arenas, J. C. Galán, F. Palero, and F. González-Candelas, "Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences," *Infect. Genet. Evol.*, vol. 30, pp. 296–307, 2015.
[13] P. Forster, L. Forster, C. Renfrew, and M. Forster, "Phylogenetic network analysis of SARS-CoV-2 genomes," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 17, pp. 9241–9243, 2020.
[14] K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7: Improvements in performance and usability," *Mol. Biol. Evol.*, vol. 30, no. 4, pp. 772–780, 2013.
[15] S. Kumar, G. Stecher, and K. Tamura, "MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets," *Mol. Biol. Evol.*, vol. 33, no. 7, pp. 1870–1874, 2016.
[16] I. Letunic and P. Bork, "Interactive Tree of Life (iTOL) v4: Recent updates and new developments," *Nucleic Acids Res.*, vol. 47, no. W1, pp. 256–259, 2019.
[17] J. Rozas *et al.*, "DnaSP 6: DNA sequence polymorphism analysis of large data sets," *Mol. Biol. Evol.*, vol. 34, no. 12, pp. 3299–3302, 2017.
[18] J. W. Leigh and D. Bryant, "POPART: Full-feature software for haplotype network construction," *Methods Ecol. Evol.*, vol. 6, no. 9, pp. 1110–1116, 2015.
[19] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, 1990.
[20] A. Waterhouse *et al.*, "SWISS-MODEL: Homology modelling of protein structures and complexes," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W296–W303, 2018.
[21] S. C. Lovell *et al.*, "Structure validation by C alpha geomF. Altschul, S., Gish, W., Miller, W., W. Myers, E., & J. Lipman, D. (1990). Basic Local Alignment Search Tool. Journal of Molecular Biology.etry: phi,psi and C beta deviation," *Proteins-Structure Funct. Genet.*, vol. 50, no. August 2002, pp. 437–450, 2003.
[22] M. Wiederstein and M. J. Sippl, "ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins," *Nucleic Acids Res.*, vol. 35, no. SUPPL.2, pp. 407–410, 2007.
[23] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, "PROCHECK: a program to check the stereochemical quality of protein structures," *J. Appl. Crystallogr.*, vol. 26, no. 2, pp. 283–291, 1993.
[24] J. U. Bowie, R. Luthy, and D. Eisenberg, "A method to identify protein sequences that fold into a known three-dimensional structure," *Science (80-. ).*, vol. 253, no. 5016, pp. 164–170, 1991.
[25] D. Douguet, H. Munier-Lehmann, G. Labesse, and S. Pochet, "LEA3D: a computer-aided ligand design for structure-based drug design," *J. Med. Chem.*, vol. 48, no. 7, pp. 2457–2468, 2005.
[26] A. Daina, O. Michielin, and V. Zoete, "SwissADME: A free web tool to evaluate

pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules," *Sci. Rep.*, vol. 7, no. October 2016, pp. 1–13, 2017.
[27] P. Banerjee, A. O. Eckert, A. K. Schrey, and R. Preissner, "ProTox-II: A webserver for the prediction of toxicity of chemicals," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W257–W263, 2018.
[28] S. W. Chiu, S. A. Pandit, H. L. Scott, and E. Jakobsson, "An improved united atom force field for simulation of mixed lipid bilayers," *J. Phys. Chem. B*, vol. 113, no. 9, pp. 2748–2763, 2009.
[29] A. W. Schüttelkopf and D. M. F. Van Aalten, "PRODRG: A tool for high-throughput crystallography of protein-ligand complexes," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 60, no. 8, pp. 1355–1363, 2004.
[30] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, "The missing term in effective pair potentials," *J. Phys. Chem.*, vol. 91, no. 24, pp. 6269–6271, 1987.
[31] A. D. Elmezayen, A. Al-Obaidi, A. T. Şahin, and K. Yelekçi, "Drug repurposing for coronavirus (COVID-19): in silico screening of known drugs against coronavirus 3CL hydrolase and protease enzymes," *J. Biomol. Struct. Dyn.*, pp. 1–12, 2020.
[32] R. J. Khan *et al.*, "Targeting SARS-CoV-2: a systematic drug repurposing approach to identify promising inhibitors against 3C-like proteinase and 2'-O-ribose methyltransferase," *J. Biomol. Struct. Dyn.*, vol. 0, no. 0, pp. 1–14, 2020.
[33] R. Lu *et al.*, "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding," *Lancet*, vol. 395, no. 10224, pp. 565–574, 2020.
[34] D. E. Gordon *et al.*, "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing," *Nature*, 2020.
[35] C. K. Chang, M. H. Hou, C. F. Chang, C. D. Hsiao, and T. H. Huang, "The SARS coronavirus nucleocapsid protein - Forms and functions," *Antiviral Res.*, vol. 103, no. 1, pp. 39–50, 2014.
[36] S. Kannan, P. Shaik Syed Ali, A. Sheeza, and K. Hemalatha, "COVID-19 (Novel Coronavirus 2019) - recent trends," *Eur. Rev. Med. Pharmacol. Sci.*, vol. 24, no. 4, pp. 2006–2011, 2020.
[37] I. Sola, F. Almazan, S. Zuniga, and L. Enjuanes, "Continuous and discontinuous RNA synthesis in coronaviruses," *Annu. Rev. Virol.*, vol. 2, pp. 265–288, 2015.
[38] M. Bouvet *et al.*, "Coronavirus Nsp10, a critical co-factor for activation of multiple replicative enzymes," *J. Biol. Chem.*, vol. 289, no. 37, pp. 25783–25796, 2014.
[39] J. B. Leslie, "Alvimopan: a peripherally acting mu-opioid receptor antagonist," *Drugs Today (Barc)*, vol. 43, no. 9, pp. 611–625, 2007.
[40] O. Hilas and D. Ezzo, "Nebivolol (bystolic), a novel beta blocker for hypertension," *P T*, vol. 34, no. 4, pp. 188–192, 2009.
[41] E. D. Deeks, "Bictegravir/Emtricitabine/Tenofovir Alafenamide: A Review in HIV-1 Infection," *Drugs*, vol. 78, no. 17, pp. 1817–1828, 2018.
[42] E. J. Lewis and J. B. Lewis, "Treatment of diabetic nephropathy with angiotensin II receptor antagonist," *Clin. Exp. Nephrol.*, vol. 7, no. 1, pp. 1–8, 2003.
[43] M. Burnier, V. Forni, G. Wuerzner, and M. Pruijm, "Long-term use and tolerability of irbesartan for control of hypertension," *Integr. Blood Press. Control*, p. 17, 2011.
[44] N. Zinner *et al.*, "Darifenacin treatment for overactive bladder in patients who expressed

dissatisfaction with prior extended-release antimuscarinic therapy," *Int. J. Clin. Pract.*, vol. 62, no. 11, pp. 1664–1674, 2008.
[45] S. S. Hegde, "Muscarinic receptors in the bladder: From basic research to therapeutics," *Br. J. Pharmacol.*, vol. 147, no. SUPPL. 2, pp. 80–87, 2006.
[46] S. Duffy, L. A. Shackelton, and E. C. Holmes, "Rates of evolutionary change in viruses: Patterns and determinants," *Nat. Rev. Genet.*, vol. 9, no. 4, pp. 267–276, 2008.
[47] P. Zhou *et al.*, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, no. 7798, pp. 270–273, 2020.
[48] Y. M. Báez-santos, S. E. S. John, and A. D. Mesecar, "The SARS-coronavirus papain-like protease: Structure, functionand inhibition by designed antiviral compounds COVID-19 resource centre is hosted on Elsevier Connect , the company ' s public news and information website . Elsevier hereby grants permission t," *Antiviral Res.*, vol. 115, no. January, pp. 21–38, 2015.

## FIGURES

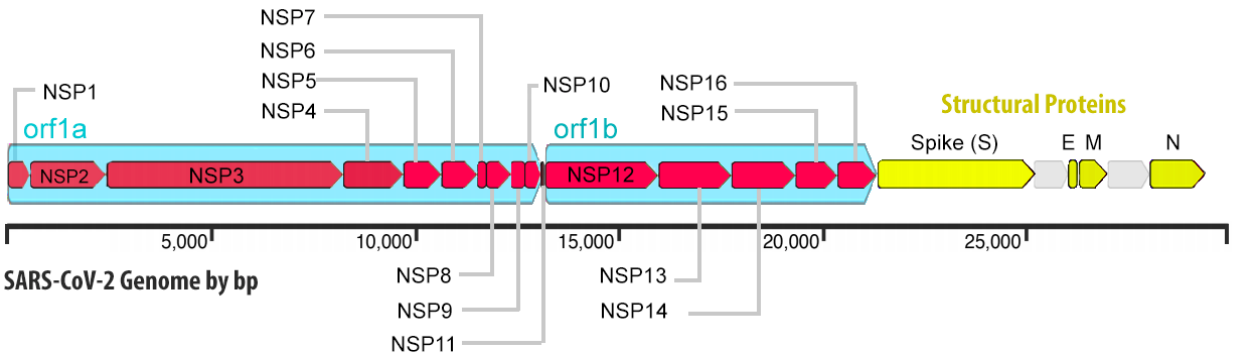



**Figure 1:** A detailed schematic representation of the SARS-CoV-2 viral genome. The figure represents the detailed view of structural and non-structural proteins (NSPs).

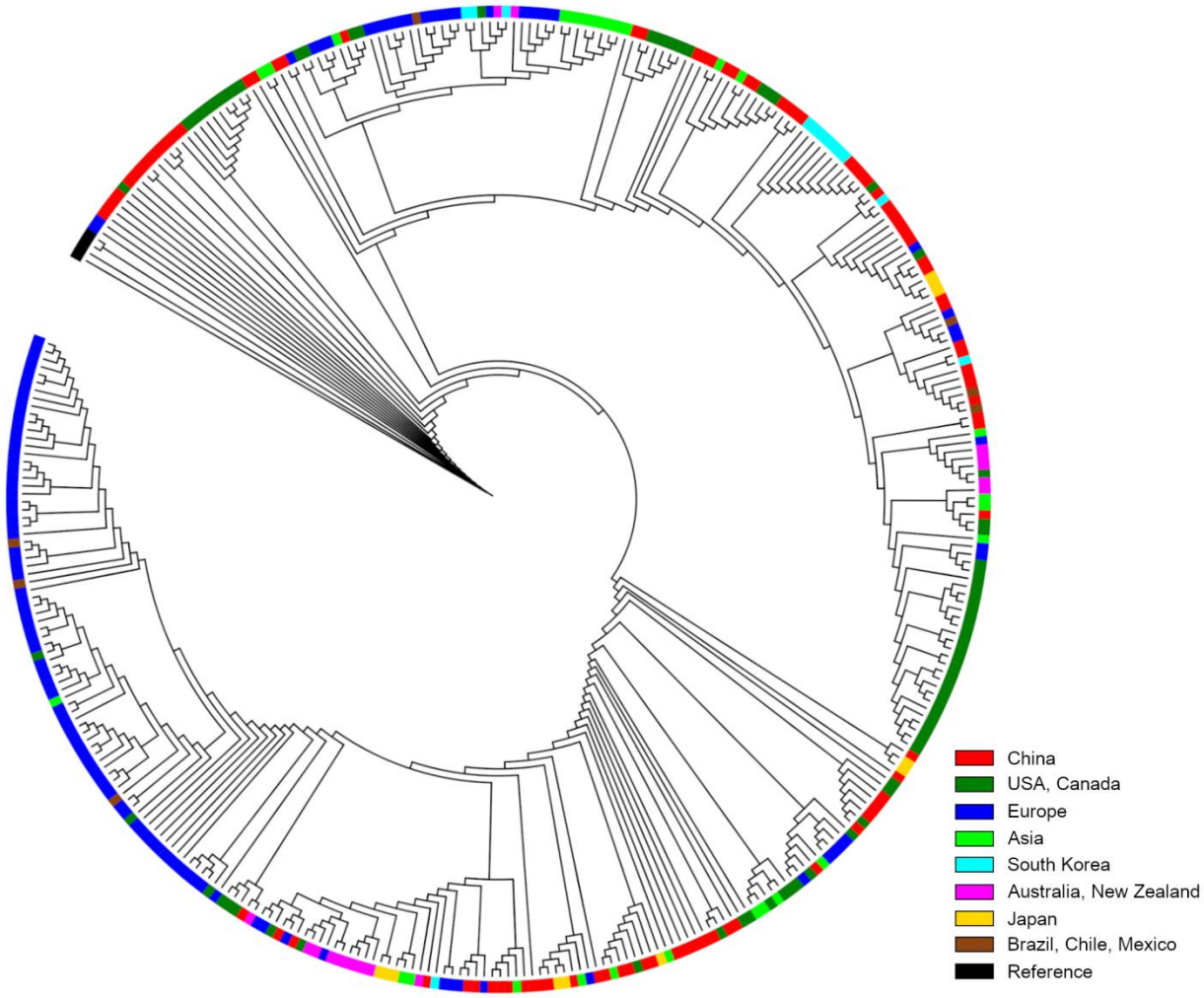



**Figure 2:** Phylogenetic tree representing 363 complete throughput SARS-CoV-2 genomes.
Europe: France, Finland, Netherlands, Czech Republic, Ireland, Switzerland, England, Italy, Portugal, Germany, Luxembourg, Wales, Sweden, Belgium
Asia: India, Singapore, Cambodia, Nepal, Vietnam, Taiwan, Hong Kong, Thailand
Reference: MG772933.1, MG772934.1, KF294457.1, AY278489.2

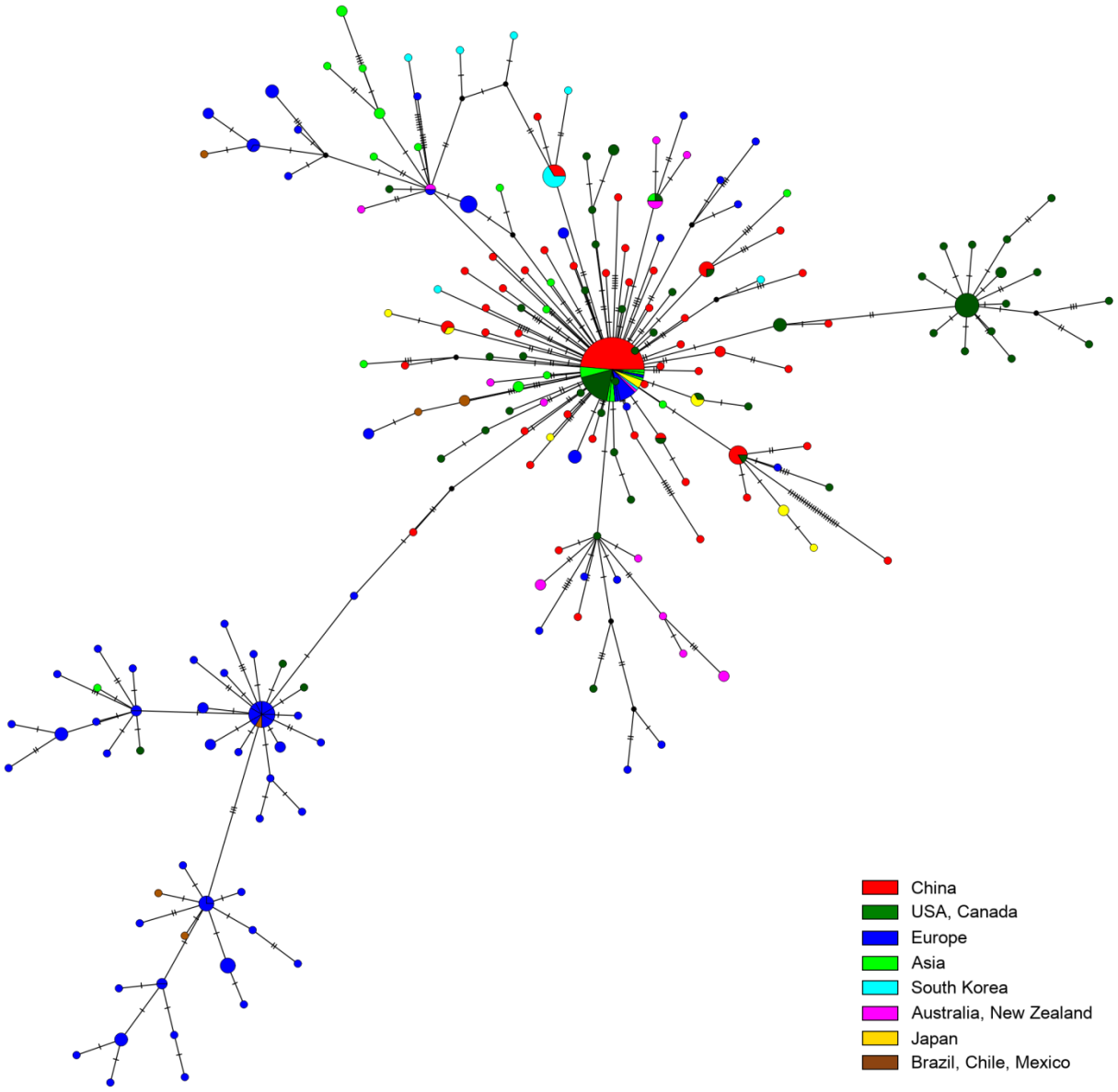



**Figure 3:** Haplotype analysis of SARS-CoV-2 viruses. Haplotype Network of 358 SARS-CoV-2 viral genomes. The distribution of haplotypes over geographical areas were inserted as a part of the traits section in the Nexus file. The color code and its respective geographical distribution is marked on the bottom right corner.

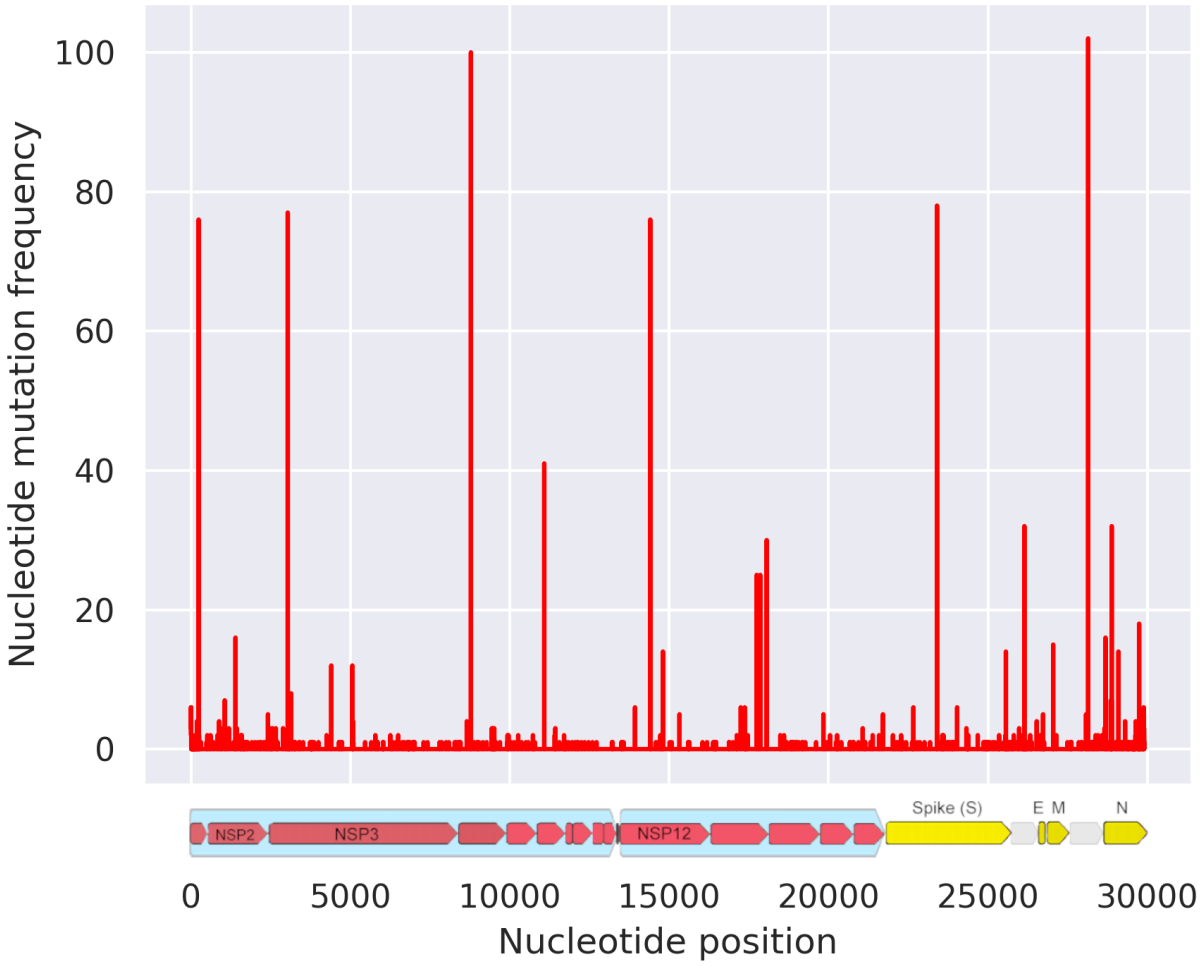



**Figure 4:** Mutation Frequency across the SARS-CoV-2 viral genome. The red lines represent the number of mutations at a particular nucleotide position. On the abscissa is the nucleotide numbered from 0 to 30,000. To better understand the mutations across the viral genome, the genomic representation of SARS-CoV-2 is provided in the bottom panel. The red ones in the bottom panel represent the non-structural proteins while the yellow ones represent Spike, E-proteins and the N-proteins.

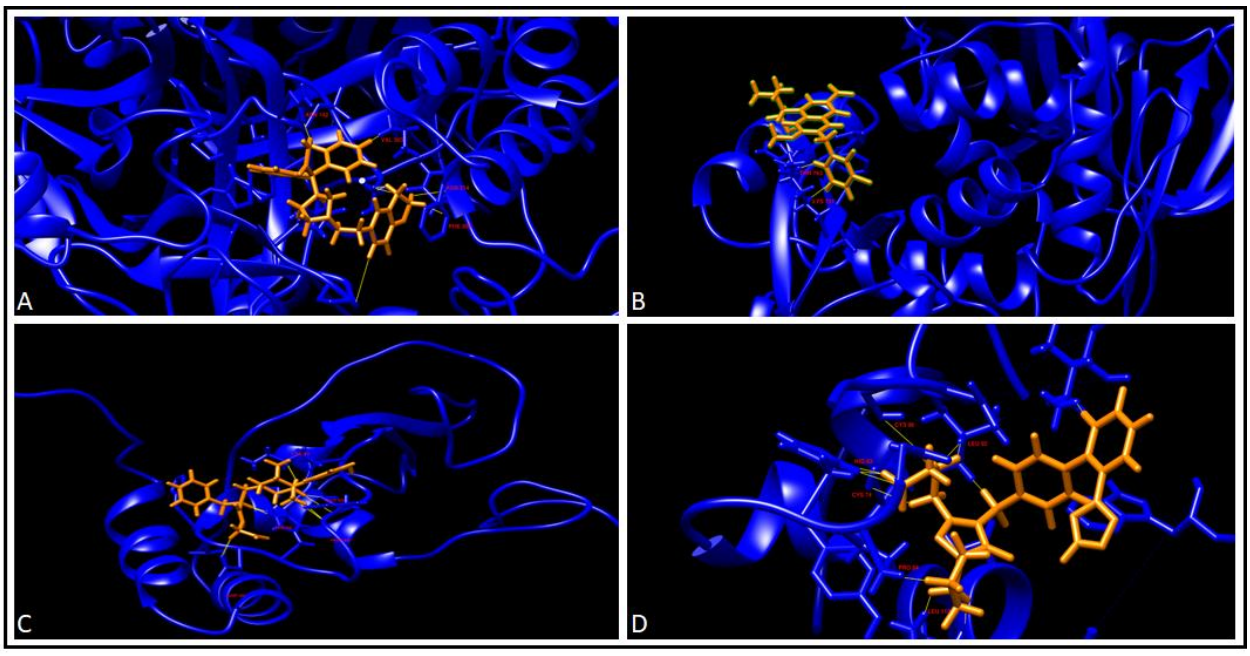



**Figure 5:** Drug-Protein interaction after docking. A. 3CLPro-Darifenacin interaction, B. 3CLPro-Nebivolol interaction, C. NSP10-Alvimopan interaction, and D. NSP10-Isbesartan interaction. Drugs are in orange while the proteins are labelled in blue and the residues interacting with the drugs are highlighted in red. The contacts are shown in yellow.

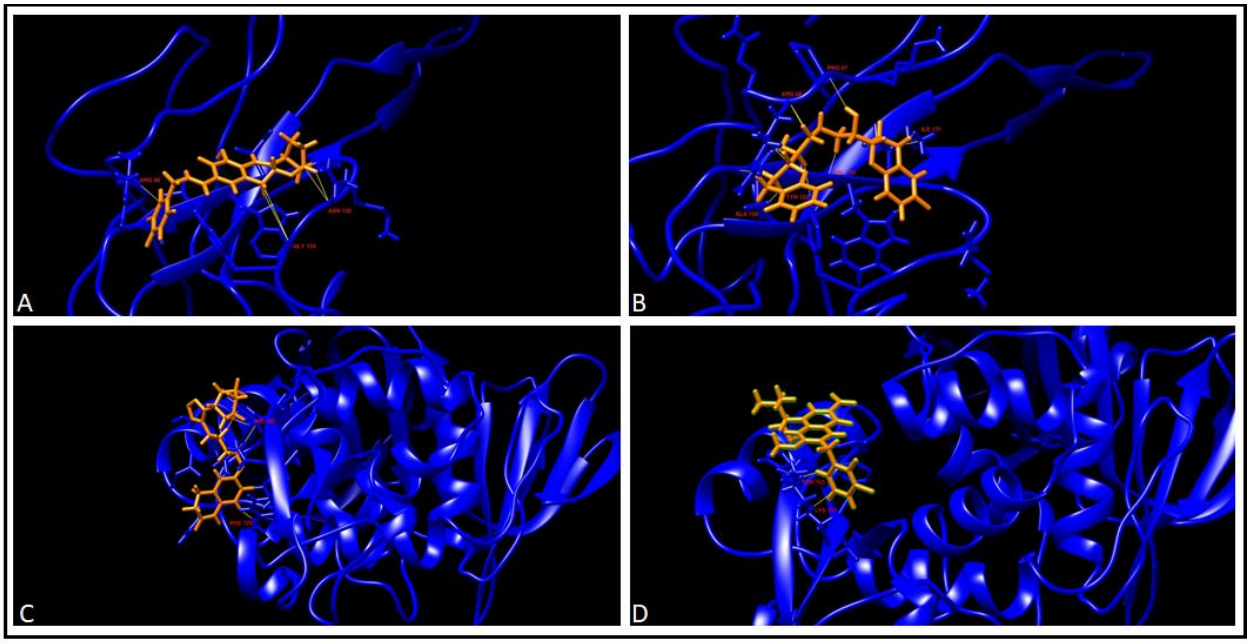



**Figure 6:** Drug-Protein interaction after docking. A. Nucleoprotein-Bictegravir interaction, B. Nucleoprotein-Nebivolol interaction, C. PL Pro-Cilostazol interaction, and D. PL Pro-Elvitegravir interaction. Drugs are in orange while the proteins are labelled in blue and the residues interacting with the drugs are highlighted in red. The contacts are shown in yellow.

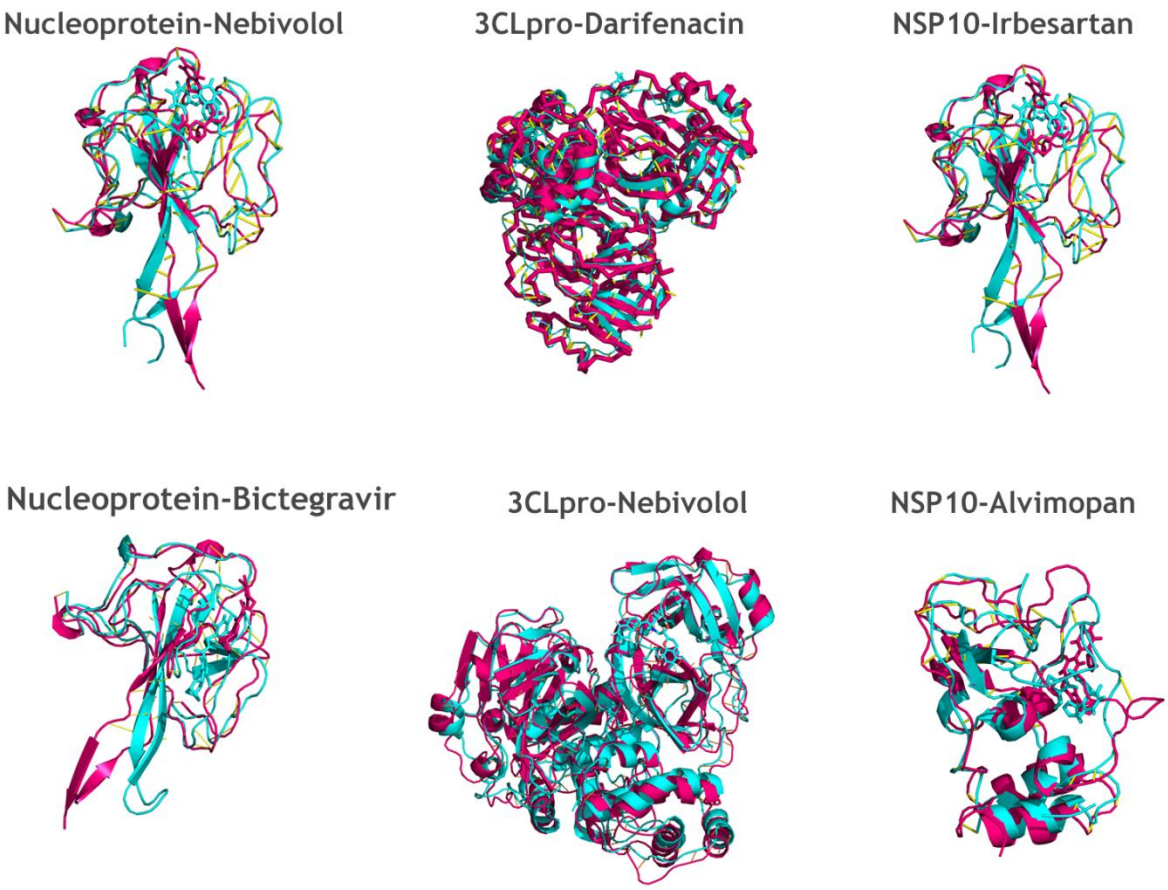
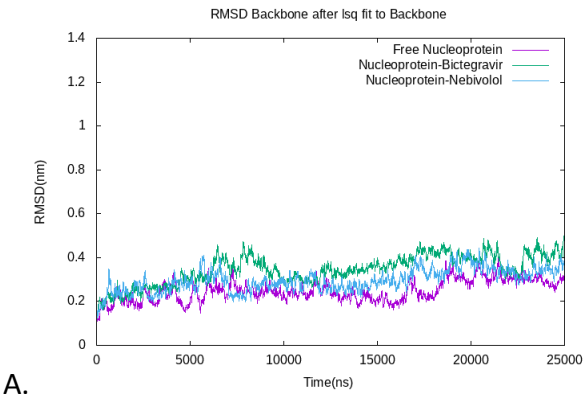
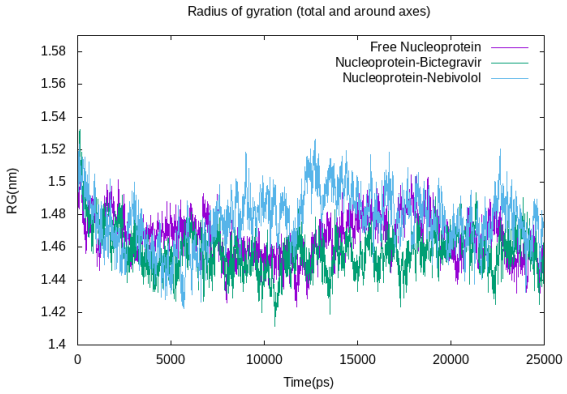


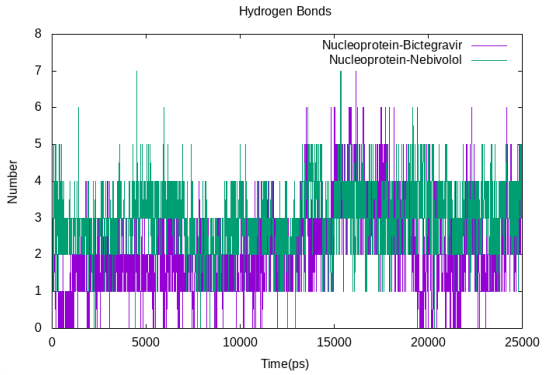

**Figure 7:** A superimposition of the protein-ligand complexes before and after the MD simulation. The Protein-ligand complex before the MD simulation is shown in magenta while the complex after the simulation is shown in cyan.



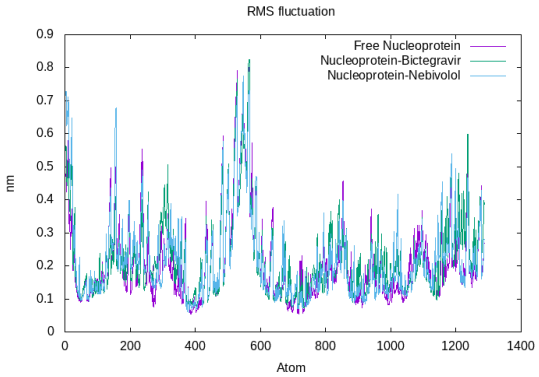

A.



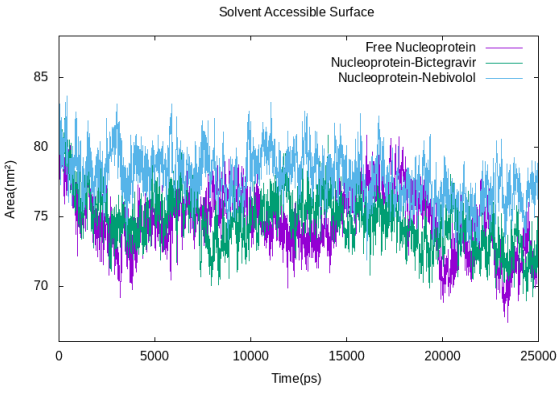

B.




C.




D.




E.

**Figure 8:** Analysis of RMSD, Radius of Gyration, Hydrogen Bonding, RMSF and SASA of Nucleoprotein and drugs Bictegravir and Nebivolol. A. Root-mean-square deviation of the Cα atoms, B. Radius of gyration (Rg) over the entire simulation, where the ordinate is Rg (nm) and the abscissa is time (ps), C. Total number of H-bond count throughout the simulation, D. RMSF values over the entire simulation, where the ordinate is RMSF (nm) and the abscissa is residue, and E. Solvent accessible surface area (SASA), where the ordinate is SASA ($nm^2$) and the abscissa is time (ps).

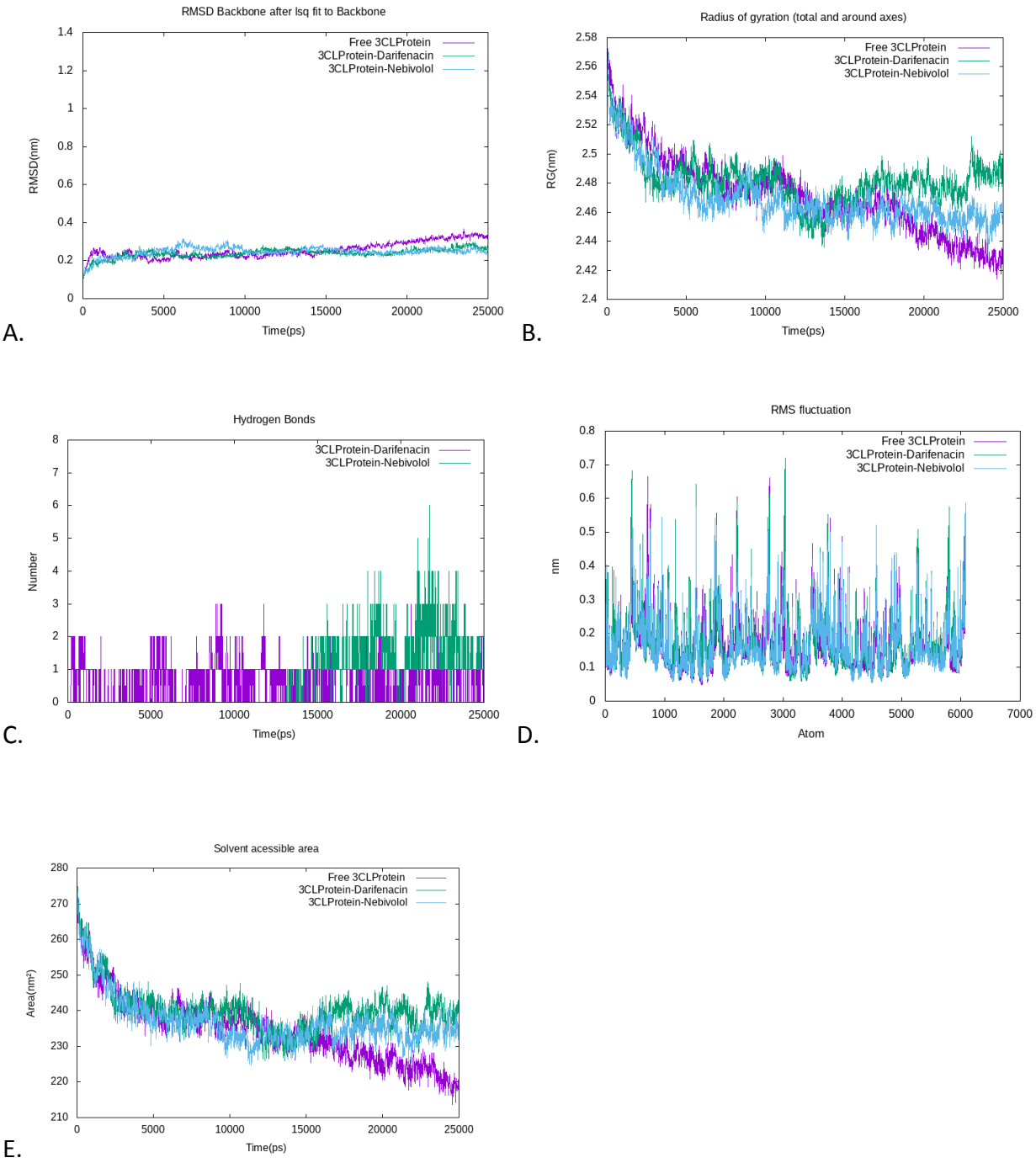



**Figure 9:** Analysis of RMSD, Radius of Gyration, Hydrogen Bonding, RMSF and SASA of 3CLpro Protein and drugs Darifenacin and Nebivolol. A. Root-mean-square deviation of the Cα atoms, B. Radius of gyration (Rg) over the entire simulation, where the ordinate is Rg (nm) and the abscissa is time (ps), C. Total number of H-bond count throughout the simulation, D. RMSF values over the entire simulation, where the ordinate is RMSF (nm) and the abscissa is residue, and E. Solvent accessible surface area (SASA), where the ordinate is SASA ($nm^2$) and the abscissa is time (ps).

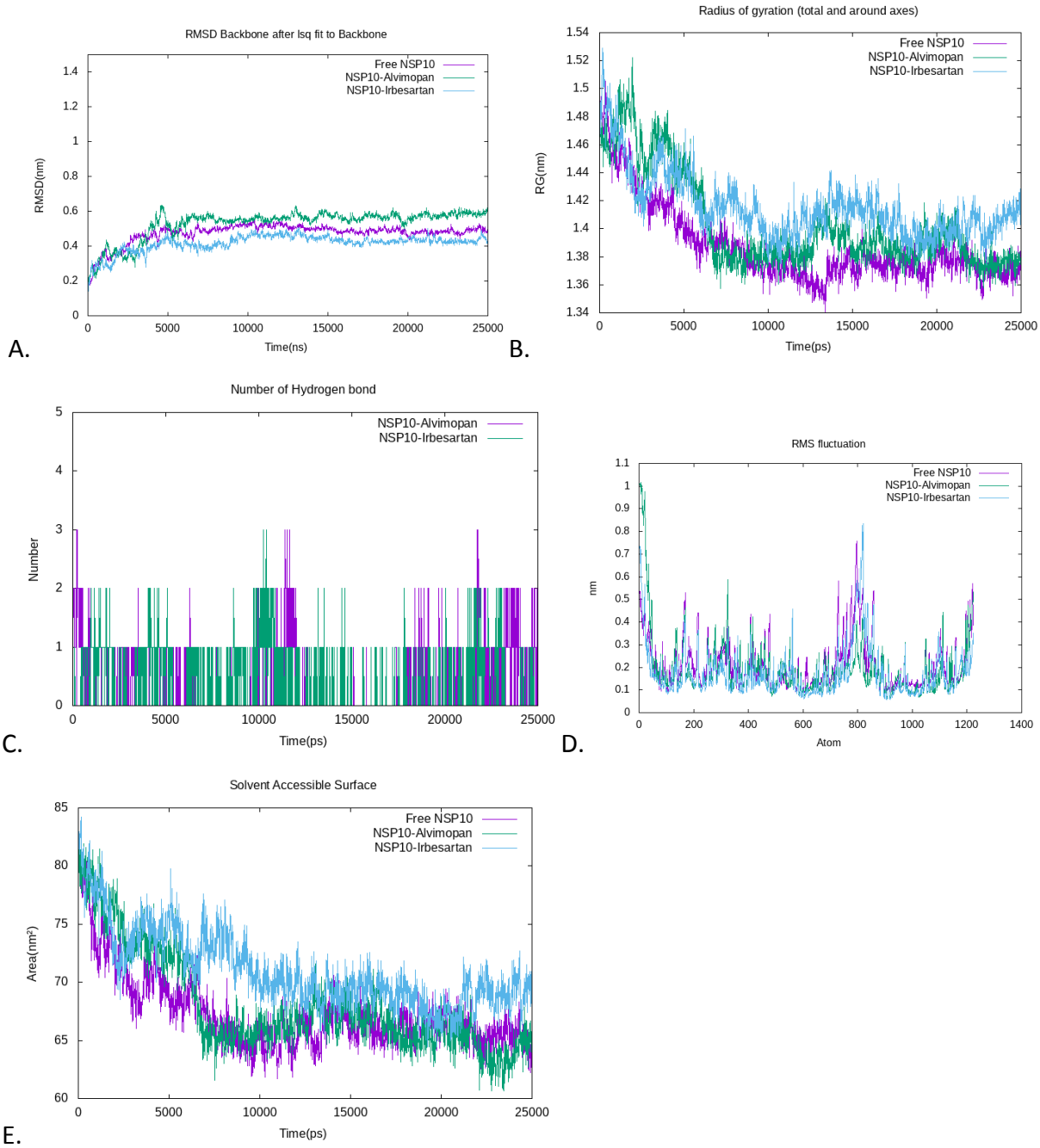



**Figure 10:** Analysis of RMSD, Radius of Gyration, Hydrogen Bonding, RMSF and SASA of NSP10 Protein and drugs Alvimopan and Irbesartan . A. Root-mean-square deviation of the Cα atoms, B. Radius of gyration (Rg) over the entire simulation, where the ordinate is Rg (nm) and the abscissa is time (ps), C. Total number of H-bond count throughout the simulation, D. RMSF values over the entire simulation, where the ordinate is RMSF (nm) and the abscissa is residue, and E. Solvent accessible surface area (SASA), where the ordinate is SASA ($nm^2$) and the abscissa is time (ps).

## Tables

**Table 1:** Detailed list of conserved genes arranged into their respective thresholds of conservation

| THRESHOLD 100-95 | THRESHOLD 95-90 | THRESHOLD 90-85 | THRESHOLD 85-80 | THRESHOLD 80-75 |
|---|---|---|---|---|
| Chain B, NSP10 | Chain A, Nucleocapsid protein | ORF1a polyprotein | ORF1ab polyprotein | ORF1ab polyprotein, partial |
| NSP10 | Chain A, Papain-like proteinase | Nucleocapsid phosphoprotein | ORF1a polyprotein, partial | Surface glycoprotein |
| Membrane glycoprotein | 3C-like proteinase | NSP2 | NSP3 | |
| Nucleocapsid phosphoprotein, partial | | ORF1ab polyprotein | NSP3 (residues 207-377) | |
| RNA binding domain of nucleocapsid protein | | | ADRP | |

**Table 2:** Population wise variant genes arranged in reference to their geographical locations

| Oceania | China | Rest of America | UK | Japan | North America | Europe |
|---|---|---|---|---|---|---|
| NSP3 | ORF1a polyprotein, partial | ORF1a polyprotein, partial | ORF1ab polyprotein, partial | ORF1a polyprotein , partial | ORF1a polyprotein, partial | NSP13-pp1ab |
| ORF1a polyprotein, partial | ORF1a polyprotein | ORF1ab polyprotein | ORF1ab polyprotein | ORF1a polyprotein | NSP3 | Chain A, Uridylate-specific endoribonuclease |
| ORF1ab polyprotein | Chain A, Uridylate-specific endoribonuclease | ORF1a polyprotein | NSP2 | ORF10 protein | Chain A, Papain-like proteinase | NSP15-pp1ab (endoRNAse) |
| ORF1a polyprotein | Chain A, Replicase polyprotein 1ab | ORF1ab polyprotein, partial | ORF1a polyprotein | ORF10 protein, partial | Chain A, Peptidase C16 | ORF3a protein |

| | | | | | | |
|---|---|---|---|---|---|---|
| NSP4 | Chain A, Non-structural Protein 3 | ORF1ab polyprotein | Surface glycoprotein | | ORF1a polyprotein | Membrane glycoprotein, partial |
| Spike glycoprotein | Chain A, NSP3 macrodomain | ORF3a, partial | Nucleocapsid phosphoprotein, partial | | ORF1ab polyprotein partial | Membrane glycoprotein |
| ORF3a protein | NSP3 | ORF3a protein | Chain A, Nucleoprotein | | ORF10 protein | ORF8 protein |
| Surface glycoprotein | ORF10 protein | Nucleocapsid phosphoprotein, partial | Chain A, SARS-CoV-2 nucleocapsid protein | | ORF10 protein, partial | ORF1ab polyprotein |
| Surface glycoprotein, partial | ORF10 protein, partial | Nucleocapsid phosphoprotein | NSP2 | | Chain A, SARS-CoV-2 NSP16 | NSP2 |
| ORF8 protein, partial | NSP14 | | | | Chain A, 2'-O-methyltransferase | ORF1a polyprotein, partial |
| ORF8 protein | | | | | | |
| ORF10 protein | | | | | | |

**Table 3:** Parameters for the validation of the homology modeled protein

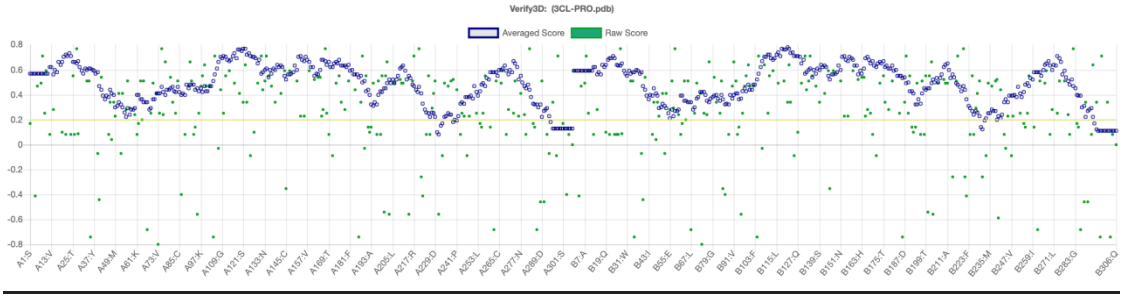
| Proteins | GMQE Score | Q-Mean | Z-Score |
|---|---|---|---|
| PL-PRO | 0.11 | -0.28 | -8.87 |
| Nucleoprotein | 0.24 | 0.03 | -5.03 |
| NSP10 | 0.86 | -0.93 | -3.58 |
| 3CLPro | 0.99 | 0.45 | -7.2 |

**Table 4:** A table illustrating the mean of various structural parameters for the simulated proteins and protein-ligand complexes

| Complex | RMSD (nm) | RMSF (nm) | Radius of Gyration (nm) | SASA (nm^2) | H-bonds |
|---|---|---|---|---|---|
| Free NSP-10 | 0.470361 | 0.20984 | 1.37233 | 66.6657 | - |
| NSP10-Alvimopan | 0.525528 | 0.19471 | 1.38811 | 67.3879 | 0 |
| NSP10-Irbesartan | 0.413957 | 0.17612 | 1.40021 | 70.351 | 1 |
| Free Nucleoprotein | 0.24749 | 0.20316 | 1.45041 | 73.9592 | - |
| Nucleoprotein-Nebivolol | 0.293992 | 0.21579 | 1.46075 | 77.0852 | 3 |
| Nucleoprotein-Bictegravir | 0.342362 | 0.21871 | 1.44189 | 74.0045 | 2 |
| Free 3CL Protein | 0.252472 | 0.16319 | 2.44544 | 232.278 | - |
| 3CL pro-Darifenacin | 0.237117 | 0.16895 | 2.45577 | 237.846 | 1 |
| 3CL pro-Nebivolol | 0.244245 | 0.16176 | 2.44327 | 234.738 | 1 |

## Supplementary Figures

A.




B.

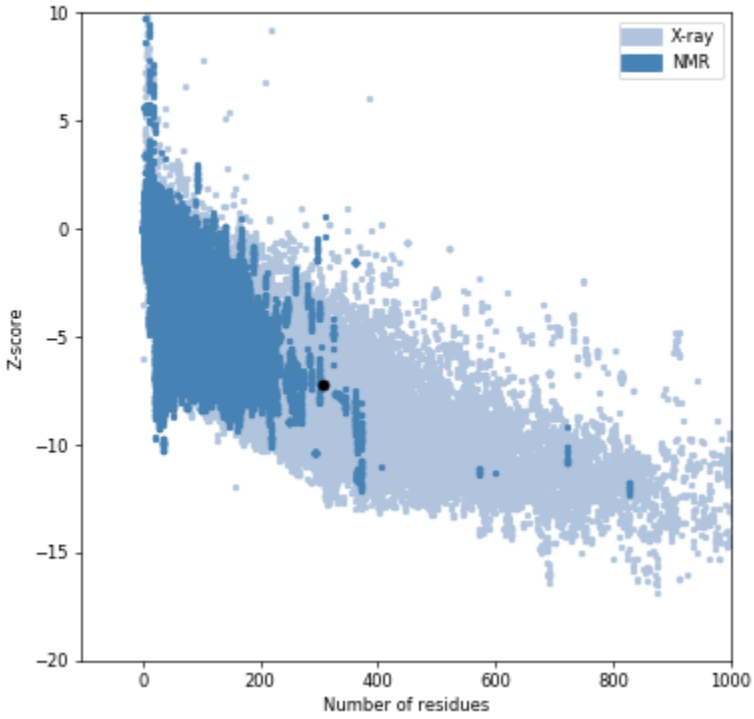
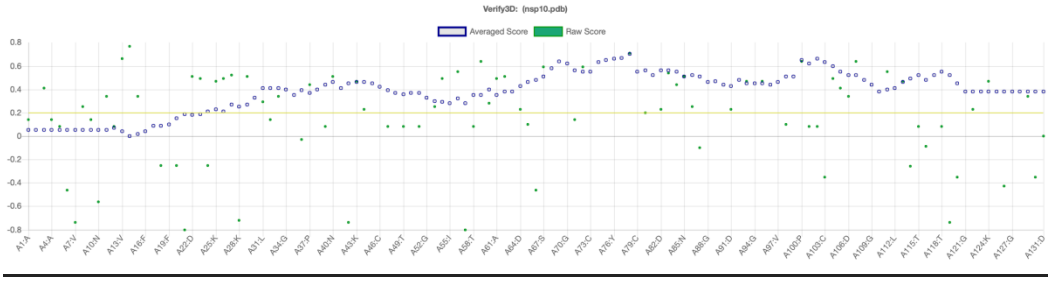
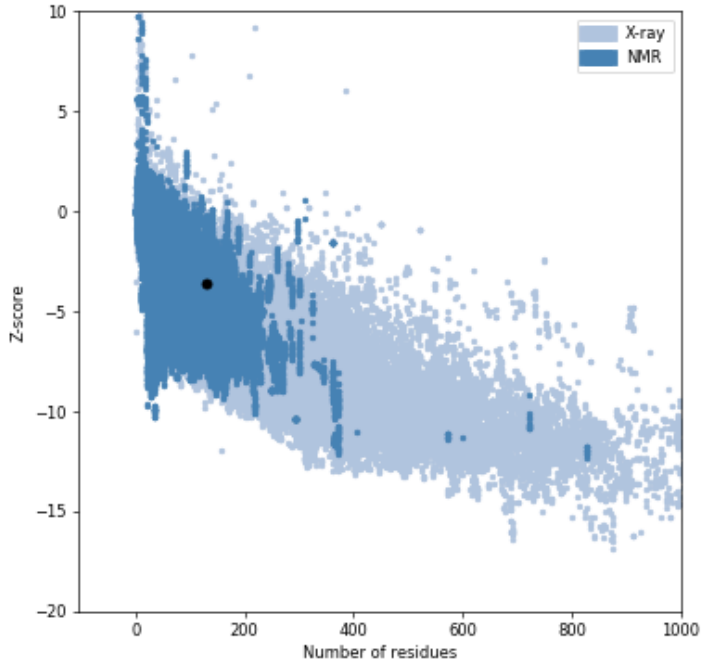



**Figure 1**: Validation of the in silico predicted model of 3CLpro by ProSa and Verify3D. A. Structure validation by ProSa, which shows the Z-score of the predicted model of 3CLpro (black dot), when compared to a non-redundant set of crystallographic structures (light blue dots) and NMR structures (dark blue dots). B. Structure validation by Verify3D, which shows the 3D-1D score for each atom of the predicted model of 3CLpro. The graphic shows that 94.28% of the residues of the in silico structure of 3CLpro presented a compatibility score of 0.2 or higher, which indicates that the structure is a high-quality model according to Verify3D.

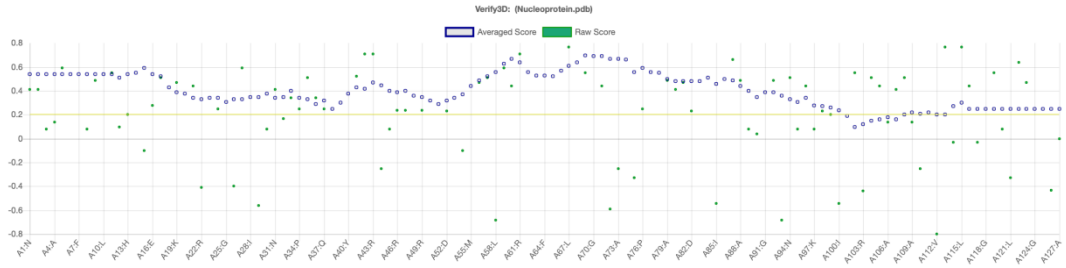



**A.**

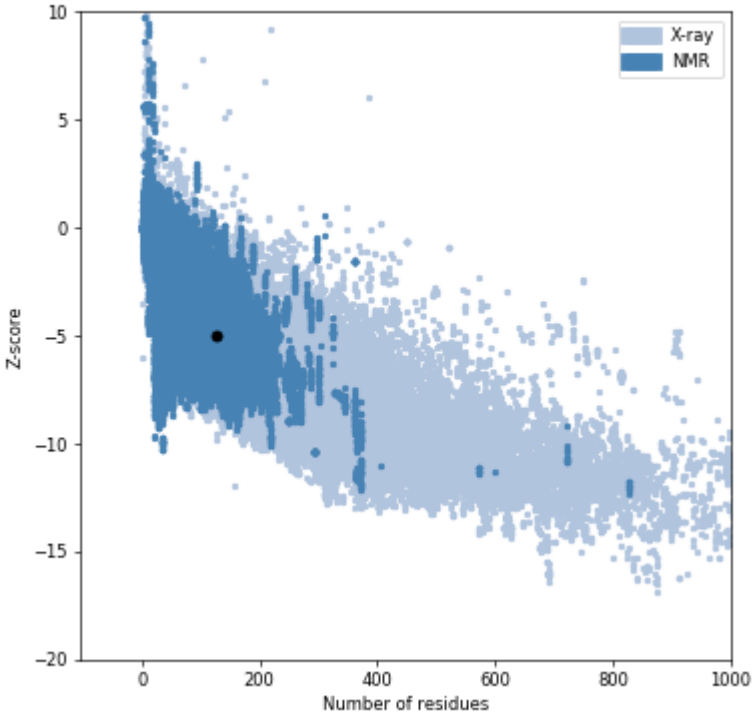



**B.**

**Figure 2**: Validation of the in silico predicted model of NSP10 by ProSa and Verify3D. A. Structure validation by ProSa, which shows the Z-score of the predicted model of NSP10 (black dot), when compared to a non-redundant set of crystallographic structures (light blue dots) and NMR structures (dark blue dots). B. Structure validation by Verify3D, which shows the 3D-1D score for each atom of the predicted model of NSP10. The graphic shows that 82.44% of the residues of the in silico structure of NSP10 presented a compatibility score of 0.2 or higher, which indicates that the structure is a high-quality model according to Verify3D.



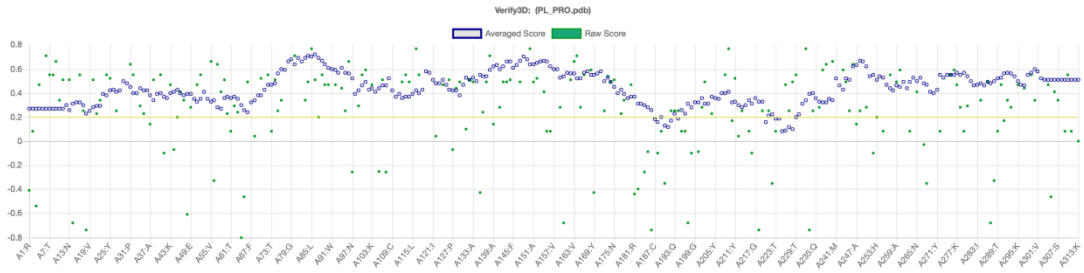

**A.**




**B.**

**Figure 3**: Validation of the in silico predicted model of Nucleoprotein by ProSa and Verify3D. A. Structure validation by ProSa, which shows the Z-score of the predicted model of Nucleoprotein (black dot), when compared to a non-redundant set of crystallographic structures (light blue dots) and NMR structures (dark blue dots). B. Structure validation by Verify3D, which shows the 3D-1D score for each atom of the predicted model of Nucleoprotein. The graphic shows that 94.49% of the residues of the in silico structure of Nucleoprotein presented a compatibility score of 0.2 or higher, which indicates that the structure is a high-quality model according to Verify3D.

A.




B.

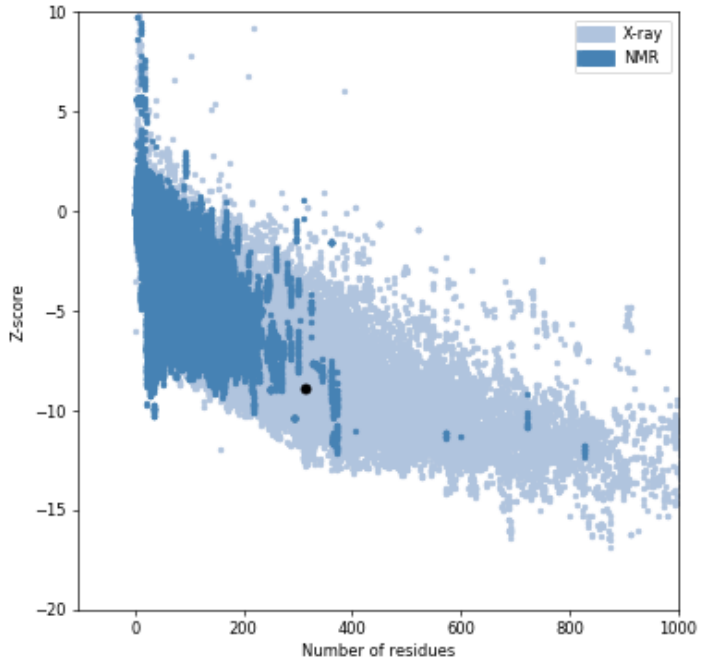



**Figure 4**: Validation of the in silico predicted model of PLpro by ProSa and Verify3D. A. Structure validation by ProSa, which shows the Z-score of the predicted model of PLpro (black dot), when compared to a non-redundant set of crystallographic structures (light blue dots) and NMR structures (dark blue dots). B. Structure validation by Verify3D, which shows the 3D-1D score for each atom of the predicted model of PLpro. The graphic shows that 95.85% of the residues of the in silico structure of PLpro presented a compatibility score of 0.2 or higher, which indicates that the structure is a high-quality model according to Verify3D.