Molecular Dynamics Simulations Indicate the COVID-19 Mpro Is Not a Viable Target for Small-Molecule Inhibitors Design

Maria Bzówka[1#], Karolina Mitusińska[1#], Agata Raczyńska[1], Aleksandra Samol[1], Jack Tuszyński[2,3], Artur Góra[1*]

1) Tunneling Group. Biotechnology Centre, ul. Krzywoustego 8, Silesian University of Technology, Gliwice, 44-100, Poland
2) Department of Physics, University of Alberta, Edmonton, AB, T6G 2E1, Canada
3) DIMEAS, Politecnico di Torino, Corso Duca degli Abruzzi, 24, Turin, 10129, Italy

*Corresponding author: a.gora@tunnelinggroup.pl, phone +48323271659
[#]These authors contributed equally to this work

Abstract

The novel coronavirus whose outbreak took place in December 2019 continues to spread at a rapid rate worldwide. In the absence of an effective vaccine, inhibitor repurposing or *de novo* design may offer a longer-term strategy to combat this and future infections due to similar viruses. Here, we report on detailed molecular dynamics simulations of the main protease (Mpro). We compared and contrasted the Mpro for COVID-19 with a highly similar SARS protein. In spite of a high level of sequence similarity, the active sites in both proteins show major differences in both shape and size indicating that repurposing SARS drugs for COVID-19 may be futile. Furthermore, analysis of the pocket's time-dependence indicates its flexibility and plasticity, which dashes hopes for rapid and reliable drug design. Conversely, structural stability of the protein with respect to flexible loop mutations indicates that the virus' mutability will pose a further challenge to the rational design of small-molecule inhibitors.

Introduction

In early December 2019, the first atypical pneumonia outbreak associated with the novel coronavirus of zoonotic origin (COVID-19) appeared in Wuhan City, Hubei Province, China[1,2]. As of 23 February 2020, COVID-19 has been reported in 26 countries and nearly 79,000 infection cases (both laboratory-confirmed and reported as clinically diagnosed), including more than 2400 fatal ones, have been confirmed[3]. According to the World Health Organization, a precise estimate of the infection fatality rate is therefore impossible at present. However, the National Health Commission of China, at the press conference on February 4 evaluated that the virus mortality rate stood at 2.1% nationwide (4.9% in Wuhan).

In general, coronaviruses (CoVs) are classified into four major genera: *Alphacoronavirus*, *Betacoronavirus* (which primarily infect mammals), *Gammacoronavirus*, and *Deltacoronavirus* (which primarily infect birds)[4–6]. In humans, coronaviruses usually cause mild to moderate upper-respiratory tract illnesses, e.g., the common cold, however, the rarer forms of CoVs can be lethal. By the end of 2019, six kinds of human CoV have been identified: HCoV-NL63, HCoV-229E, belonging to *Alphacoronavirus* genera, HCoV-OC43, HCoV-HKU1, severe acute respiratory syndrome SARS-CoV, and Middle East respiratory syndrome MERS-CoV, belonging to *Betacoronavirus* genera[5]. Of the aforementioned CoVs, the last two are the most dangerous and they were associated with the outbreak of two epidemics at the beginning of the 21st century[7]. On January 7, the COVID-19 was isolated and announced as a

new, seventh, type of human coronavirus (the name was officially given by WHO on February 11). It was classified as *Betacoronavirus*[2]. Investigations to determine the origins of the infection are still ongoing, however, increasing evidence demonstrates a link between the COVID-19 and other similar known coronaviruses circulating in bats. Based on the phylogenetic analysis of the genomic data of COVID-19, Zhang et al. indicated that the COVID-19 is most closely related to two SARS-CoV sequences isolated from bats in 2015 and 2017. This is suggestive that the bat's CoV and COVID-19 share a common ancestor, and the new virus can be considered as a SARS-like virus[8]. Notwithstanding, the transmission route to humans remains unclear. Bats are rather rare in food markets in China but they might be hunted and sold directly to restaurants. The most likely hypothesis is that an intermediary host animal has played a role in the transmission. Since the intermediate hosts are generally mammals, there is also a possibility that living mammals, which are often sold in Chinese food markets, could have caused an outbreak of human infection.

The genome of coronaviruses typically contains a positive-sense, single-stranded RNA but it differs in size ranging between ~26 and ~32 kb. It also includes a variable number of open reading frames (ORFs) – from 6 to 11. The first ORF is the largest, encoding nearly 70% of the entire genome and 16 non-structural proteins (nsps)[4,9]. Of the nsps, the main protease (Mpro, also known as a chymotrypsin-like cysteine protease 3CLpro), encoded by nsp5, has been found to play a fundamental role in viral gene expression and replication, thus it is an attractive target for anti-CoV drug design[10]. The remaining ORFs encode accessory and structural proteins, including spike surface glycoprotein (S), small envelope protein (E), matrix protein (M), and nucleocapsid protein (N).

Based on the three sequenced genomes of COVID-19 (Wuhan/IVDC-HB-01/2019, Wuhan/IVDC-HB-04/2019, and Wuhan/IVDC-HB-05/2019, provided by the National Institute for Viral Disease Control and Prevention, CDC, China), Wu et al., performed a detailed genome annotation. The results were further compared to related coronaviruses – 1,008 human SARS-CoV, 338 bat SARS-like CoV, and 3,131 human METS-CoV indicating that the three strains of COVID-19 have almost identical genomes with 14 ORFs, encoding 27 proteins including 15 non-structural proteins (nsp1-10 and nsp12-16), 4 structural proteins (S, E, M, N), and 8 accessory proteins (3a, 3b, p6, 7a, 7b, 8b, 9b, and orf14). The only identified difference in the genome consisting of ~29.8 kb nucleotides consisted of five nucleotides. The genome annotation revealed that COVID-19 is fairly similar to SARS-CoV at the amino acid level, however, there are some differences in the occurrence of accessory proteins, e.g., the 8a accessory protein, present in SARS-CoV, is absent in COVID and the lengths of 8b and 3b proteins do not match. The phylogenetic analysis of COVID-19 showed it to be most closely related to SARS-like bat viruses, but no strain of SARS-like bat virus was found to cover all equivalent proteins of COVID-19[11].

As previously mentioned, the main protease is one of the key enzymes in the viral life cycle. Together with other non-structural proteins (papain-like protease, helicase, RNA-dependent RNA polymerase) and the spike glycoprotein structural protein, it is essential for interactions between the virus and host cell receptor during viral entry[12]. Initial analyses of genomic sequences of the four nsps mentioned above indicate that those enzymes are highly conserved sharing more than 90% sequence similarity with the corresponding SARS-CoV enzymes[13].

The recently released crystal structure of the Mpro of COVID-19 (PDB ID: 6lu7) was obtained by Prof. Yang's group from ShanghaiTech by co-crystallisation with a peptide-like inhibitor N-[(5-methylisoxazol-3-yl)carbonyl]alanyl-L-valyl-N~1-((1R,2Z)-4-(benzyloxy)-4-oxo-1-{[(3R)-2-oxopyrrolidin-3-yl]methyl}but-2-enyl)-L-leucinamide (N3 or PRD_002214). The

same inhibitor was co-crystallised with other human coronaviruses, e.g., HCoV-NL63 (PDB ID: 5gwy), HCoV-KU1 (PDB ID: 3d23), or SARS-CoV (PDB ID: 2amq). Currently, Mpro is the only crystallised COVID-19 protein and it has a 96% sequence identity with Mpro from SARS-CoV. This enzyme naturally forms a dimer whose each monomer consists of the N-terminal catalytic region and a C-terminal region[15]. While 12 residues differ between both CoVs, only one, namely S46 in COVID-19 (A46 in SARS), is located in the proximity of the entrance to the active site. However, such a small structural change would typically be not expected to substantially affect the binding of small molecules[13]. Such an assumption would routinely involve the generation of a library of derivatives and analogous based on the scaffold of a drug that inhibits the corresponding protein in the SARS case. As shown in the present paper, regrettably, this strategy is not likely to succeed with COVID-19 for Mpro as a molecular target. Below, we detail the results that lead to this conclusion.

In this study, we investigate how only 12 different residues, located mostly on the protein's surface, may affect the behaviour of the active site pocket of the COVID-19 Mpro structure. To this end, we performed classical molecular dynamics simulations (cMD) of both SARS and COVID-19 Mpros as well as mixed-solvents MD simulations (MixMD) combined with small molecules' tracking approach to analyse the conformational changes in the binding site. In spite of the structural differences in the active sites of both Mpro proteins, major issues involving plasticity and flexibility of the binding site could result in significant difficulties in inhibitor design for this molecular target. Indeed, an *in silico* attempt has already been made involving a massive virtual screening for Mpro inhibitors of COVID-19 using Deep Docking[15]. Other recent attempts used virtual screening searching searches for putative inhibitors of the same main protease of COVID-19 based on the clinically approved drugs [16–18]. However, none of such attempts is likely to lead to clinical advances in the fight against COVID-19 for reasons we elaborate below.

Results and Discussion

Crystal structures comparison, and location of the replaced amino acids distal to the active site

The COVID-19 main protease's crystallographic structure was recently made publicly available through the Protein Data Bank (PDB)[19] as a complex with an N3 inhibitor (PDB ID: 6lu7). We refer to this structure as COVID-19 CoV Mpro. We used two structures of the SARS-CoV main protease: one, referred to as SARS-CoV Mpro (PDB ID: 2amq), was crystallised with the same inhibitor to compare the structural information, and the other without an inhibitor (PDB ID: 1q2w), which we refer to as SARS-CoV Mpro-f. The COVID-19 Mpro and SARS-CoV Mpro structures differ by only 12 amino acids located mostly on the proteins' surface (Figure 1A, Supplementary Table S1). Both enzymes share the same structural composition; they comprise three domains: domains I (residues 1-101) and II (residues 102-184) consist of an antiparallel β-barrel, and the α-helical domain III (residues 201-301) is required for the enzymatic activity[20]. Both enzymes resemble the structure of cysteine proteases, although their active site is lacking the third catalytic residue[21]; their active site comprises a catalytic dyad, namely H41 and C145, and a particularly stable water molecule forms at least three hydrogen bond interactions with surrounding residues, including the catalytic histidine, which corresponds to the position of a third catalytic member (Figure 1B). It should be also noted that one of the differing amino acids in COVID-19 Mpro, namely S46, is located on a C44-P52 loop which is flanking the active site cavity.

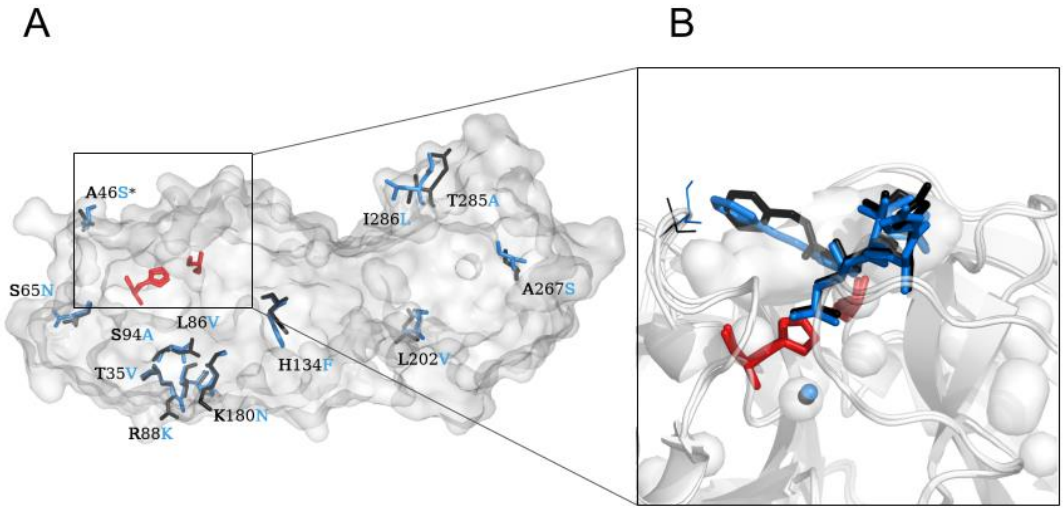


Figure 1. The differences between the SARS-CoV Mpro and COVID-19 Mpro structures. (A) The overall structure of both SARS-CoV and COVID-19 Mpros with differing amino acids marked as black (SARS-CoV Mpro) and blue (COVID-19 Mpro). (B) Close-up of the active site cavity and bound N3 inhibitor into SARS-CoV (black sticks) and COVID-19 (blue sticks) Mpros. The catalytic water molecule that resembles the position of the third member of the catalytic triad adopted from the cysteine proteases is shown for both SARS-CoV (black sphere) and COVID-19 (blue sphere) Mpros. The active site residues are shown as red sticks and the proteins' structures are shown in surface representation. The differing residues in position 46 located near the entrance to the active site are marked with an asterisk (*) on the (A) and as blue and black lines on the (B) panel.

Plasticity of the binding cavities

We performed 50 ns MD simulations of both SARS-CoV Mpros, and COVID-19 Mpro to gain insight into the plasticity of the binding cavity with a classical MD approach with water molecules used as molecular probes. Such a strategy is assumed to provide a highly detailed picture of protein's interior dynamics[22]. The small molecules tracking approach was used to determine the accessibility of the active site pocket in both SARS-CoV Mpros and COVID-19 Mpro, and a local distribution approach was used to provide information about an overall distribution of solvent in the protein's interior. To properly examine the flexibility of both active site cavities, we used the *time-window* mode of the AQUA-DUCT (AQ) software[23] to analyse the water molecules' flow through the cavity in a 10 ns time step and combined that with the *outer* pocket calculations to examine the plasticity and maximal accessible volume of the binding cavity.

Figure 2 presents the differences in sizes and shapes of the *outer* pockets detected in the two systems. Surprisingly, the volume of the *outer* pockets of both SARS main proteases structures is on average at least 2-fold larger than those of COVID-19 Mpro (Supplementary Table S2). Since both structures are highly similar, it might be expected that their binding pocket would also be very similar. This observation suggests that there can be large differences between the accessibility to the binding cavity and/or the accommodation of the shape of the cavity in response to an inhibitor that can be bound. There are also differences in the *outer* pockets' volumes between the two structures of SARS main proteases; the inhibitor-free SARS-CoV Mpro-f structure used as a starting point of MD simulations has shown the largest *outer* pocket of all the analysed systems. These results suggest that the SARS main proteases' binding cavity is highly flexible and changes both in volume and shape significantly after ligand binding. This

finding indicates a serious obstacle for a classical virtual screening approach and drug design in general. Numerous novel compounds exist that are considered as potential inhibitors of SARS-CoV, although they have not reached the stage of clinical trials. The lack of success might be related to the above-mentioned plasticity of the binding cavity. Some of these compounds have been used for docking and virtual screening research, aimed not only at SARS-CoV[24,25], but also at the novel CoV[15,26]. Such an approach focuses mostly on the structural similarity between the binding pockets but ignores the fact that the actual available binding space differs significantly. In general, a rational drug design can be a very successful tool in the identification of possible inhibitors in cases where the atomic resolution structure of the target protein or complex is known. This approach is referred to as Structure-Based Drug Design (SBDD)[27]. For a new target, when a highly homologous structure is available with a co-crystalised inhibitor exists, then a very logical strategy can be used by seeking chemically similar compounds or creating derivatives of this inhibitor, and finding those that are predicted to have a higher affinity for the new target structure than the original one. This would be expected to work for COVID-19 proteins (such as Mpro) using SARS proteins as a template. However, our in-depth analysis indicates a very different situation taking place, with major shape and size differences emerging due to the binding site flexibility. Although discouraging, such important results should be taken into consideration in future research.

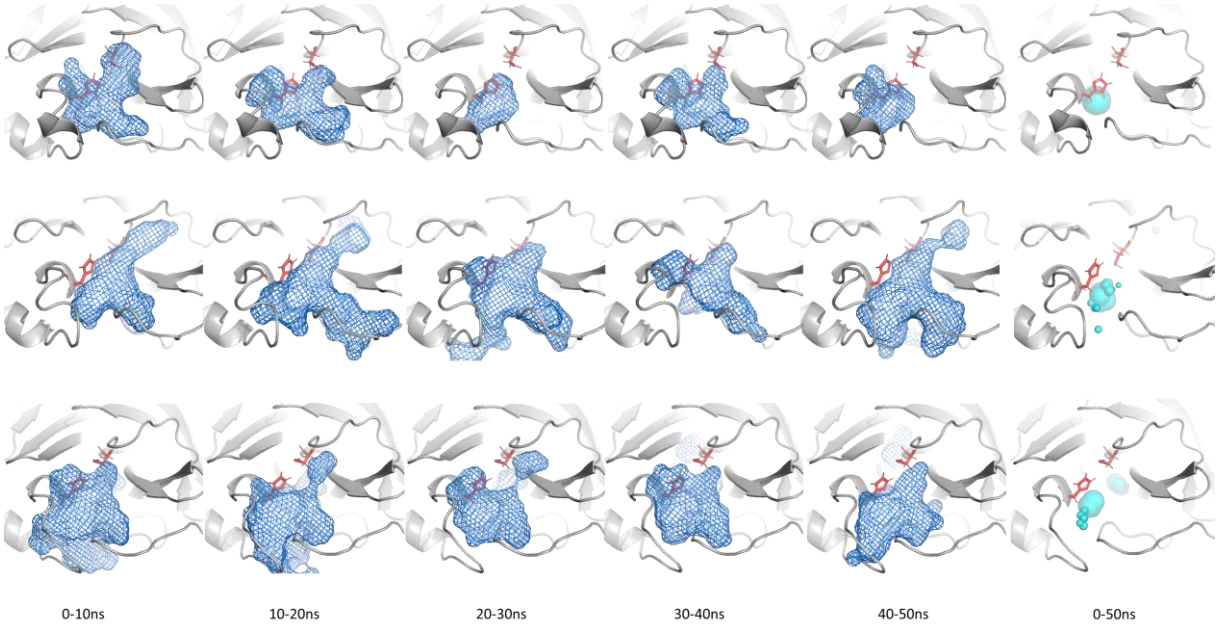


Figure 2. The *outer* pockets of COVID-19 Mpro (upper row), SARS-CoV Mpro (middle row), and SARS-CoV Mpro-f (bottom row) structures in 5 *time-windows* (10 ns each). The *outer* pocket (blue mesh) represents the maximal possible space that could be explored by water molecules. The catalytic dyad is shown as red sticks. Note that the *outer* pocket calculated for both SARS-CoV Mpros is larger than the *outer* pocket of COVID-19 Mpro which shows a higher level of plasticity and flexibility of the SARS-CoV Mpros binding cavity (see also Supplementary Table 2). The last column shows the average location of water hot-spots (cyan spheres) during the simulation time. The position of the biggest hot-spot in each row reflects the position of the catalytic water molecule.

As we have shown in previous research, tracking of water molecules in the binding cavity combined with the local distribution approach can identify catalytic water positions[28]. Despite differences in the size and dynamics of the binding cavities of SARS-CoV and COVID-19 Mpros, the main identified water hot-spot was always found in a position next to the H41 residue, and this location is assumed to indicate catalytic water of Mpro replacing the missing

third catalytic site amino acid[21]. The remaining water hot-spots correspond to a much lower water density level and are on the borders of the binding cavity, which suggests rather hydrophobic or neutral interior of the binding cavity. Therefore, we applied MixMD simulations with various cosolvents to examine in detail the plasticity of the binding site cavity in response to molecular probes with various physico-chemical properties.

Cosolvent hot-spots analysis

The mixed-solvent MD simulations were run with the following cosolvents: acetonitrile (ACN), benzene (BNZ), dimethylsulfoxide (DMSO), methanol (MEO), phenol (PHN), and urea (URE). Cosolvents were used as specific molecular probes, representing different chemical properties and functional groups that would complement the different regions of the binding site and the protein itself. Using small molecules tracking approach we analysed the flow through the Mpros structures and identified the regions in which those molecules are being trapped and/or caged, located within the protein itself (global hot-spots; Supplementary Figure S1) and inside the binding cavity (local hot-spots; Supplementary Figure S2). The size and location of both types of hot-spots differ and provide complementary information. The global hot-spots identify potential binding/interacting sites in the whole protein structure and additionally provide information about regions attracting particular types of molecules, whereas local hot-spots describe the actual available binding space of a specific cavity.

Figure 3 shows the location of global hot-spots for COVID-19 Mpro structure. For clarity, for each cosolvent, only the most important hot-spots are shown. Figure 3 also presents amino acids that differ between the SARS-CoV Mpros and COVID-19 Mpro structures. The largest number and the densest hot-spots are located within the catalytic dyad and the binding cavity. The binding cavity is particularly occupied by urea and phenol hot-spots, which is especially interesting, due to the fact that these solvents exhibit different chemical properties. Such an observation applies also to both SARS-CoV Mpros structures (Supplementary Figure S3). The general distribution of the hot-spots from particular cosolvents is quite similar and verifies specific interactions with the particular regions of the analysed proteins. It is worth mentioning that around the amino acids that vary between the structures of COVID-19 Mpro and SARS-CoV Mpros, there is also a notable number of hot-spots. Hot-spots for urea and phenol also stand out in these places, however, hot-spots for other cosolvents also appear, though they exhibit a lower density.

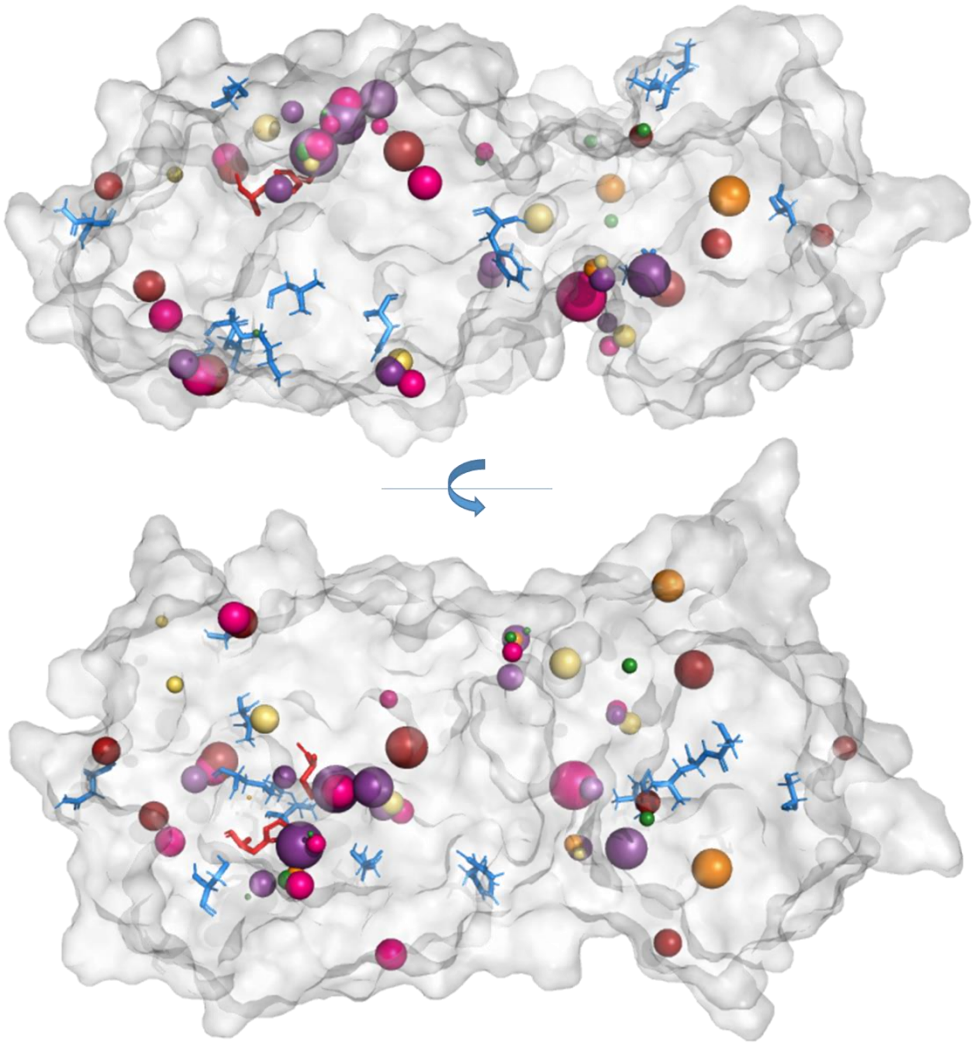


Figure 3. Localisation of the most important hot-spots identified in COVID-19 Mpro. Hot-spots for individual cosolvents are represented by spheres, and their size reflects the hot-spots density. The colour coding is as follows: purple - urea, green - DMSO, yellow - methanol, orange - acetonitrile, pink - phenol, red - benzene. The active site residues are shown as red sticks, the unique residues of COVID-19 Mpro as blue sticks, and the proteins' structures are shown in surface representation.

Figure 4 presents a close-up of the binding pockets in Mpros. In the first row global hot-spots are shown, whereas the second row presents the local hot-spots. In the case of both COVID-19 Mpro and SARS-CoV Mpros structures, hot-spots are located near the catalytic dyad and in the places corresponding to the locations of functional groups of the N3 inhibitor. However, the chemical properties of hot-spots clearly differ between both structures. The active site cavity of the COVID-19 Mpro structure is occupied mostly by urea and phenol hot-spots, while SARS-CoV Mpro features mostly benzene hot-spots. Such findings could suggest that potential COVID-19 Mpro inhibitors may exhibit diverse chemical characteristics. The hot-spots distribution of the SARS-CoV Mpro-f structure differs from that of Mpros. Both global and local hot-spots of the SARS-CoV Mpro-f structure are located in the proximity of the C44-P52 loop, which potentially regulates the access to the active site, whereas both the COVID-19 and SARS-CoV Mpros are accessible to cosolvent molecules. It is worth noting that the binding cavity in the SARS-CoV Mpro-f structure is less occupied in comparison with two other Mpros. A caveat to this analysis must be added that accounts for the differences between COVID-19 and SARS-CoV situations discussed above. While for SARS-CoV a ligand is included in the

pocket and its presence or absence can be compared, for COVID-19 we do not have an 'empty' (apo) structure available and its presence could in principle explain the lack of stability and flexibility of the analysed loop. Therefore, our conclusions here are still somewhat tentative.

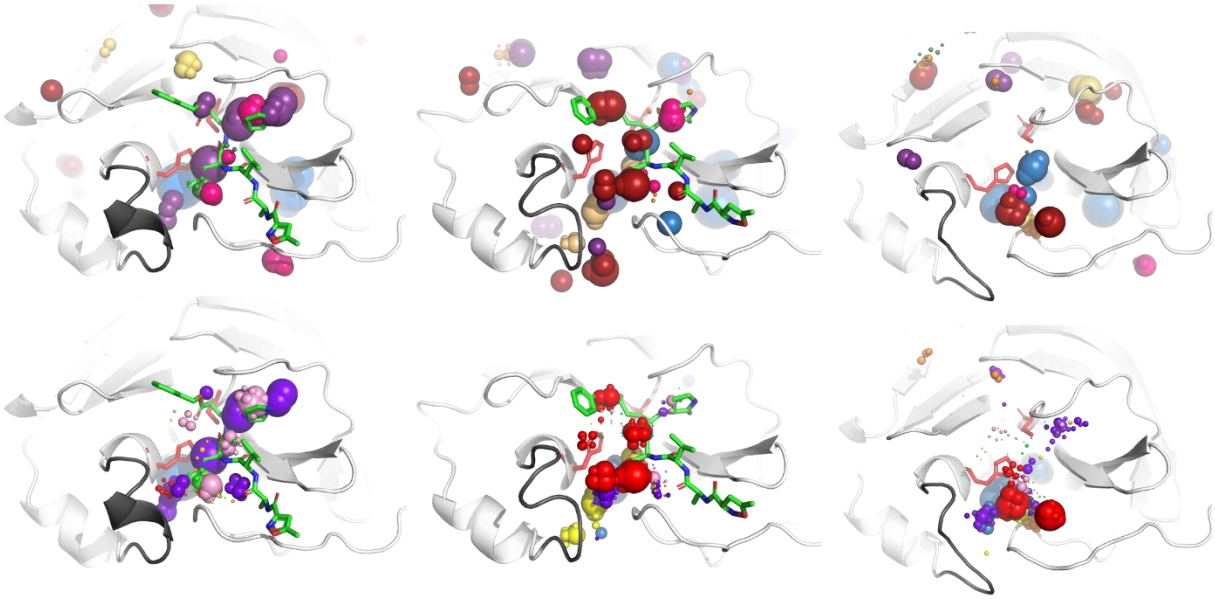
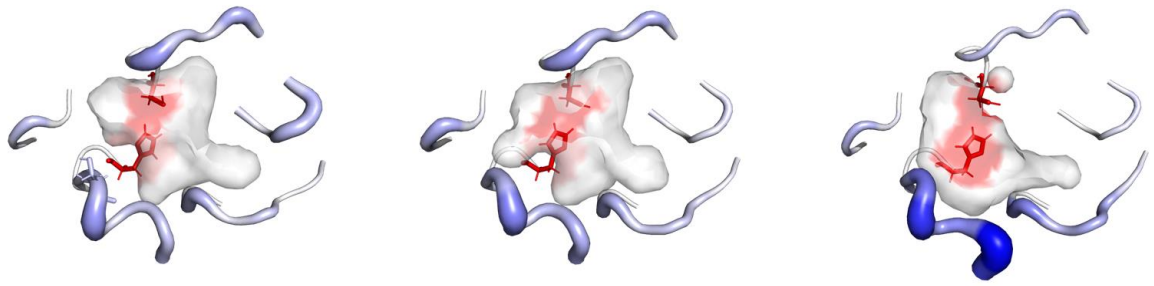


Figure 4. Localisation of the global (upper row) and local (bottom row) hot-spots identified in the binding site cavities in analysed proteins (from left, COVID-19 Mpro, SARS-CoV Mpro, and SARS-CoV Mpro-f). Hot-spots for individual cosolvents are represented by spheres, and their size reflects the hot-spots density. The colour coding is as follows: purple - urea, green - DMSO, yellow - methanol, orange - acetonitrile, pink - phenol, red - benzene. The active site residues are shown as red sticks, the N3 inhibitor structure from the crystal structures as green sticks, and the proteins' structures are shown in cartoon representation, loop 44-52 is grey.

Flexibility of the active site entrance

To further examine the plasticity and flexibility of the main proteases binding cavities, we focused on the movements of loops surrounding their entrances and regulating the active sites' accessibility. We found that one of the analysed loops of the SARS-CoV Mpro-f, namely C44-P52 loop, is more flexible than the corresponding loops of two other Mpros structures, while the adjacent loops are mildly flexible (Figure 5). This could be indirectly assumed from the absence of the C44-P52 loop in the crystallographic structure of SARS-CoV Mpro-f structure. On the other hand, such flexibility could suggest that the presence of an inhibitor might stabilise the loops surrounding the active site. The other Mpros structures with bound N3 inhibitor did not show such loop movements.



A B C

Figure 5. Flexibility of loops surrounding the entrance to the binding cavity of (A) COVID-19 Mpro, (B) SARS-CoV Mpro and (C) SARS-CoV Mpro-f. For the picture clarity, only residues creating loops were shown. The active site residues are shown as red sticks and the A46S replacement between SARS and COVID-19 main proteases is shown as light blue sticks. The width and colour of the shown residues reflect the level of loop flexibility. The wider and darker residues are more flexible.

Potential mutability of COVID-19

In general, all the above-mentioned findings indicate potential difficulties in the identification of specific inhibitors toward Mpro proteins. First, the binding site itself is characterised by huge plasticity and probably even distant to active site mutations modify their properties. Secondly, the C44-P52 loop regulates access to the active site and can contribute to the discrimination of potential inhibitors. Therefore, additional mutations in mentioned regions, which could appear during further COVID-19 evolution, can immediately change the affinity between Mpro and its ligands. To verify potential threat of further mutability of the Mpro protein we performed: i) correlated mutation analyses (CMA) on multiple sequence alignments, ii) the analysis of the contribution of already identified differences between the SARS and COVID-19 Mpro proteins to protein stability, and iii) have predicted further possible mutations caused by the most probable mutations, substitution of single nucleotides in mRNA sequence of Mpro.

Indeed, the analysis performed with Comulator software[29] shows, that within Mpros from the coronavirus family evolutionary-correlated residues are dispersed throughout the structure. This indirectly supports our previous findings that distant amino acids mutation can contribute significantly to binding site plasticity. It is worth to add that among evolutionary-correlated residues we identified also those that differ between COVID-19 and SARS-CoV Mpros, located on the C44-P52 loop (Supplementary Figure S4) and the F185-T201 linker loop. The C44-P52 loop is likely to regulate the access to the active site by enabling entrance of favourable small molecules and blocking the entry of unfavourable ones. Such a conclusion may also imply that a sufficiently potent inhibitor of SARS-CoV and/or COVID-19 Mpros needs to be able to open its way to the active site before it can successfully bind to its cavity. The F185-T201 loop starts in the vicinity of the binding site and links I and II domains with the III domain; it contributes significantly to Mpro dimerization[30]. The CMA analysis indicate that Q189 from the linker loop corelates with residues from the C44-P52 loop, whereas R188, A191, and A194 correlate with selected residues from all domains, but not with the C44-P52 loop (Supplementary Figure S4). As reported in the previous research, the overall plasticity of Mpro is required for proper enzyme functioning[31,32]. In the case of SARS-CoV the truncation of the linker loop (F185-T201) gave rise to a significant reduction in protein activity and

confirmed that the proper orientation of the linker allows the shift between dimeric and monomeric forms[30]. Dimerization of the enzyme is necessary for its catalytic activity and the proper conformation of the seven N-terminal residues (N-finger) is required[33]. In COVID-19 Mpro, the T285 is replaced by alanine, and the I286 by leucine. It has been shown that replacing S284, T285, and I286 by alanine residues in SARS-CoV Mpro leads to a 3.6-fold enhancement of the catalytic activity of the enzyme. This is accompanied by changes of the structural dynamics of the enzyme that transmit the effect of the mutation to the catalytic center. Indeed, the T285A replacement observed in the COVID-19 Mpro allows the two domains III to approach each other a little closer[34].

In the interest of examining the energetical effect of the 12 amino acid replacement in the COVID-19 Mpro structure, we performed FoldX[35] calculations for these residues. As expected, the calculated differences in total energies of the SARS-CoV Mpro and variants with introduced mutation from COVID-19 Mpro residue did not represent a significant energy change (Supplementary Table S3). The biggest energy reduction was found for mutation H134F (-0.85 kcal/mol) and mutations R99K, S94A, T285A, I286L only slightly reduced the total energy (Supplementary Table S1).

In order to investigate further possible mutations of COVID-19 Mpro, single nucleotide substitutions were introduced to the COVID-19 main protease gene. If a substitution of a single nucleotide caused translation to a different amino acid than compared to the corresponding residue in the wild-type structure, an appropriate mutation was proposed with FoldX calculations. The most energetically favourable potential mutations were chosen based on -1.5 kcal/mol threshold (Figure 6A, Supplementary Table S3). Most of the energetically favourable potential mutations include amino acids that are solvent-exposed on the protein's surface, according to NetSurfP[36] results. These results show that in general, exposed amino acids are more likely to mutate.

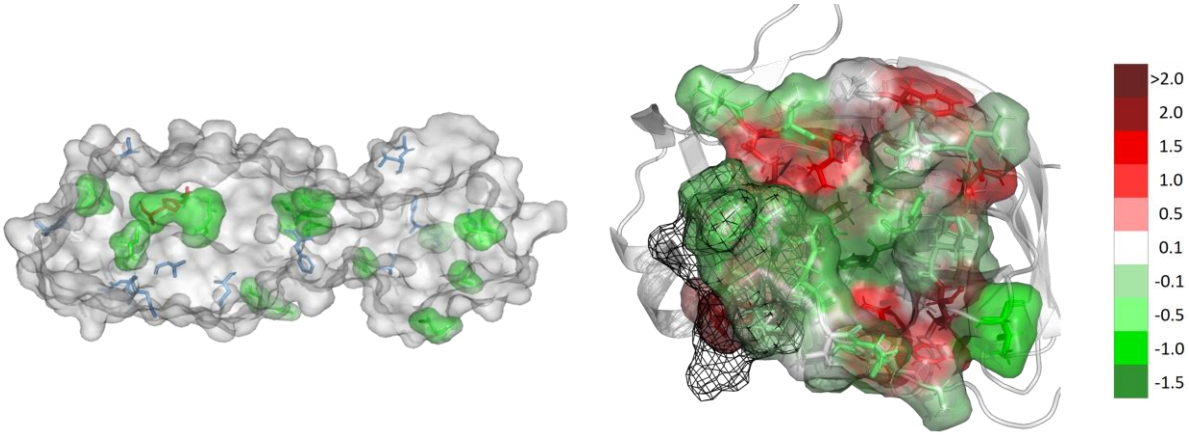


Figure 6. Potential mutability of COVID-19 Mpro. (A) Structure of COVID-19 Mpro with the most energetically favourable potential mutations of amino acids marked as green surface. Positions of amino acids that differ from the ones in SARS-CoV Mpro structure marked as blue sticks. Catalytic dyad marked as red. (B) The catalytic site of COVID-19 Mpro is shown as surface with the most energetically favourable potential mutations shown as green, neutral as white and unfavourable as red. The C44-P52 loop is shown as black mesh.

Additionally, the potential mutability of the binding cavity was investigated. Residues belonging to the binding cavity were found within 7 Å from the N3 inhibitor. FoldX energy calculations of possible mutations were performed for these amino acids (Supplementary Table S4). A heatmap of these residues was then created based on the differences of Gibbs free energy of protein folding compared to the wild-type structure. The most energetically favourable potential mutations are shown as green, neutral as white and unfavourable as red (Figure 6B). Interestingly, residues forming the catalytic dyad, namely H41 and C145, are also prone to mutate. However, probably the most important message comes from the analysis of the potential mutability of the C44-P52 loop. Mutation of four of them has a stabilising effect for the protein and rest near-neutral contribution to the energy. This result indicates that the future evolution of the Mpro protein can significantly reduce the potential use of this protein as a molecular target for coronavirus treatment due to a highly probable development of drug resistance of this virus through mutations.

In this paper, we reported on molecular dynamics simulations of the main protease (Mpro), whose crystal structure has been recently released. We compared and contrasted the Mpro for COVID-19 with a highly similar SARS-CoV protein. In spite of a high level of sequence similarity between these two homologous proteins, their active sites show major differences in both shape and size indicating that repurposing SARS-CoV drugs for COVID-19 may be futile. Furthermore, a detailed analysis of the binding pocket's time-dependence indicates its flexibility and plasticity, which dashes hopes for rapid and reliable drug design. Moreover, our findings show the presence of a flexible loop occluding the entrance to the binding pocket. A successful inhibitor may need to have an ability to move the loop from the entrance in order to bind to the catalytic pocket. However, mutations leading to changes in the amino acid sequence of the loop, while not affecting the folding of the protein, may result in the putative inhibitors' inability to access the binding pocket. We conclude that Mpro is unlikely to represent a fruitful target for drug design against COVID-19. In our opinion, drug development efforts aimed at combatting this virus should focus on other molecular targets.

Methods

Classical MD simulations

The H++ server[37] was used to protonate the COVID-19 and SARS-CoV main proteases' structures (PDB IDs: 6lu7, and 2amq and 1q2w, respectively) using standard parameters and pH 7.4. The missing 4-amino-acids-long loop of the 1q2w model was added using the corresponding loop of the 6lu7 model. Water molecules were placed using the combination of 3D-RISM[38] and the Placevent algorithm[39]. The AMBER 18 LEaP[40] was used to immerse models in a truncated octahedral box of TIP3P water molecules and prepare the systems for simulation using the ff14SB force field[41]. Additionally, 4 and 3 Na + ions were added to the COVID-19 and to the SARS, respectively. AMBER 18 software[40] was used to run 50 ns simulations of both systems. The minimisation procedure consisted of 2000 steps, involving 1000 steepest descent steps followed by 1000 steps of conjugate gradient energy minimisation, with decreasing constraints on the protein backbone (500, 125 and 25 kcal x mol^{-1} x $Å^2$) and a final minimisation with no constraints of conjugate gradient energy minimization. Next, gradual heating was performed from 0 K to 300 K over 20 ps using a Langevin thermostat with a temperature coupling constants of 1.0 ps in a constant volume periodic box. Equilibration and production stages were run using the constant pressure periodic boundary conditions for 1 ns with 1 fs step and 50 ns with a 2 fs time step, respectively. Constant temperature was maintained using the weak-coupling algorithm for 50 ns of the production simulation time,

with a temperature coupling constant of 1.0 ps. Long-range electrostatic interactions were modelled using the Particle Mesh Ewald method with a non-bonded cut-off of 10 Å and the SHAKE algorithm. The coordinates were saved at an interval of 1 ps.

Mixed-solvent MD simulations - cosolvent preparation

Six different cosolvents: acetonitrile (ACN), benzene (BNZ), dimethylsulfoxide (DMSO), methanol (MEO), phenol (PHN), and urea (URE) were selected to perform the mixed-solvent MD simulations. The chemical structures of cosolvents molecules were downloaded from the ChemSpider database[42] and a dedicated set of parameters was prepared. Parameters for ACN were adopted from the work by Nikitin and Lyubartsev[43], and parameters for URE were modified using the 8Mureabox force field to obtain parameters for a single molecule. For the rest of the co-solvent molecules, parameters were prepared using Antechamber[44] with Gasteiger charges[45].

Mixed-solvent MD simulations - initial configuration

The Packmol software[46] was used to build the initial systems consisting of protein (protonated according to the previously described procedure), water, and particular cosolvent molecules. 4 and 3 Na^+ ions were added to the COVID-19 Mpro and to the SARS-CoV Mpros, respectively. It was assumed that the percentage concentration of the cosolvent should not exceed 5% (in the case of ACN, DMSO, MEO, and URE), or should be about 1% in the case of BNZ and PHN phenol (see Supplementary Table S5). The mixed-solvent MD simulation procedures (minimization, equilibration, and production) carried out using the AMBER 18 package were identical as for the classical MD simulations. Only the heating stage differed - it was extended up to 40 ps.

Water and cosolvent molecules tracking

The AQUA-DUCT 1.0 (AQ) software was used to track water and cosolvent molecules. Molecules of interests, which have entered the so-called *Object*, defined as 5Å sphere around the centre of geometry of active site residues, namely H41, C145, H164, and D187, were traced within the *Scope* region, defined as the interior of a convex hull of both COVID-19 Mpro and SARS Mpro Cα atoms. All visualizations were made in PyMol[47].

Outer pocket analysis

AQUA-DUCT defines the pockets as areas of the overall distribution of tracked water molecules; the *outer* pocket represents the maximal possible space that could be explored by tracked molecules.

Hot-spots identification and selection

AQ was used to detect regions occupied by molecules of interests, and identify the densest sites using a local solvent distribution approach. Those so-called hot-spots could be calculated as local and/or global, based on the distribution of tracked molecules which visited the *Object* (local) or just the *Scope* without visiting the *Object* (global); here, they are considered as potential binding sites. For clarity, the size of each sphere representing a particular hot-spot has been changed to reflect its occupation level. The selection of the most significant hot-spots consisted of indicating points showing the highest density in particular regions. From the set of points in the space, small groups of hot-spots were determined. Groups were further defined

by distance (radius) from each other. Any point found within a distance shorter than the determined radius (3Å) from any other point being part of a given group was counted toward the group. For each so designated group of points, one showing the highest density was chosen as representing the place.

Obtaining COVID-19 Mpro gene sequence

COVID-19 Mpro was downloaded from the PDB as a complex with an N3 inhibitor (PDB ID: 6lu7). Tblastn[48] was run based on the protein amino acid sequence. 100% identity with 10055-10972 region of COVID-19 complete genome (Sequence ID: MN985262.1) was obtained. Blastx[49] calculations were run with the selected region, and orf1a polyprotein (NCBI Reference Sequence: YP_009725295.1) amino acid sequence, identical with the previously downloaded COVID-19 Mpro, was received.

FoldX mutations

FoldX software was used to insert substitutions into the structures of SARS and COVID-19 Mpros. In order to analyse the changes in the two structures, 12 single-point mutations were introduced to the SARS structure. Each of the residues in SARS-CoV Mpro was mutated to the respective COVID-19 Mpro residue, and the difference in total energies of the wild-type COVID-19 Mpro and the mutant structures were calculated. Then, in order to investigate further possible mutations of COVID-19 Mpro, single nucleotide substitutions were introduced to the COVID-19 main protease gene. If a substitution of a single nucleotide caused translation to a different amino acid than the corresponding residue in the wild-type structure, an appropriate mutation was proposed with FoldX software.

Comulator calculations of correlation between amino acids

SARS-CoV Mpro was downloaded from the PDB (PDB ID: 1q2w). Blast[50] was run based on the amino acid sequence. As a result, 2643 sequences of viral main proteases similar to chain A SARS-CoV Mpro were obtained. Clustal Omega[51] was used to prepare an alignment of those sequences. Comulator[29] was then employed to calculate the correlation between amino acids and based on the results, groups of positions in SARS-CoV Mpro sequence were selected, whose amino acid occurrences strongly depended on each other.

Acknowledgements

KM, MB, AR, AS and AG work was supported by the National Science Centre, Poland, grant no DEC-2013/10/E/NZ1/00649 and DEC-2015/18/M/NZ1/00427. JT expresses gratitude for research support for this project received from IBM CAS and NSERC (Canada).

Authors Contribution

MB and **KM**: Resources, Calculations, Data analysis, Data curation, Writing- Original draft preparation, Writing - Review & Editing, Visualization. **AR**: Calculations, Data analysis, Writing - Review & Editing, Visualization. **AS**: Calculations, Data analysis. **JT**: Funding acquisition, Writing - Review & Editing. **AG**: Conceptualization, Supervision, Data analysis, Visualization, Funding acquisition, Project administration, Writing - Review & Editing.

References

1. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* (2020). doi:10.1016/S0140-6736(20)30183-5
2. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
3. WHO. *Coronavirus disease 2019 (COVID-19) Situation Report – 34*.
4. Woo, P. C. Y., Huang, Y., Lau, S. K. P. & Yuen, K.-Y. Coronavirus Genomics and Bioinformatics Analysis. *Viruses* **2**, 1804–1820 (2010).
5. Tang, Q. *et al.* Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. *Sci. Rep.* **5**, 17155 (2015).
6. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
7. Fehr, A. R. & Perlman, S. Coronaviruses: An Overview of Their Replication and Pathogenesis. *Methods Mol Biol.* **1282**, 1–23 (2015).
8. Zhang, L., Shen, F., Chen, F. & Lin, Z. Origin and evolution of the 2019 novel coronavirus. *Clin. Infect. Dis.* (2020). doi:10.1093/cid/ciaa112
9. Song, Z. *et al.* From SARS to MERS, Thrusting Coronaviruses into the Spotlight. *Viruses* **11**, 59 (2019).
10. Xue, X. *et al.* Structures of Two Coronavirus Main Proteases: Implications for Substrate Binding and Antiviral Drug Design. *J. Virol.* **82**, 2515–2527 (2008).
11. Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* (2020). doi:10.1016/j.chom.2020.02.001
12. Zumla, A., Chan, J. F. W., Azhar, E. I., Hui, D. S. C. & Yuen, K.-Y. Coronaviruses — drug discovery and therapeutic options. *Nat. Rev. Drug Discov.* **15**, 327–347 (2016).
13. Liu, W., Morse, J. S., Lalonde, T. & Xu, S. Learning from the Past: Possible Urgent Prevention and Treatment Options for Severe Acute Respiratory Infections Caused by 2019-nCoV. *ChemBioChem* cbic.202000047 (2020). doi:10.1002/cbic.202000047
14. Lee, T.-W. *et al.* Crystal Structures of the Main Peptidase from the SARS Coronavirus Inhibited by a Substrate-like Aza-peptide Epoxide. *J. Mol. Biol.* **353**, 1137–1151 (2005).
15. Ton, A.-T., Gentile, F., Hsing, M., Ban, F. & Cherkasov, A. Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds. *ChemRxiv* (2020).
16. Xu, Z. *et al.* Nelfinavir was predicted to be a potential inhibitor of 2019-nCov main protease by an integrative approach combining homology modelling, molecular docking and binding free energy calculation. *bioRxiv* (2020). doi:10.1101/2020.01.27.921627
17. Liu, X. & Wang, X.-J. Potential inhibitors for 2019-nCoV coronavirus M protease from clinically approved medicines. *bioRxiv* (2020). doi:10.1101/2020.01.29.924100
18. Li, Y. *et al.* Therapeutic Drugs Targeting 2019-nCoV Main Protease by High-Throughput Screening. *bioRxiv* (2020).
19. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
20. Bacha, U., Barrila, J., Velazquez-Campoy, A., Leavitt, S. A. & Freire, E. Identification of Novel Inhibitors of the SARS Coronavirus Main Protease 3CL pro †. *Biochemistry* **43**, 4906–4912 (2004).
21. Anand, K. Coronavirus Main Proteinase (3CLpro) Structure: Basis for Design of Anti-SARS Drugs. *Science (80-.).* **300**, 1763–1767 (2003).
22. Mitusińska, K., Raczyńska, A., Bzówka, M., Bagrowska, W. & Góra, A. Applications

of water molecules for analysis of macromolecule properties. *Comput. Struct. Biotechnol. J.* **18**, 355–365 (2020).

23. Magdziarz, T. *et al.* AQUA-DUCT 1.0: structural and functional analysis of macromolecules from an intramolecular voids perspective. *Bioinformatics* (2019). doi:10.1093/bioinformatics/btz946
24. Chang, C. *et al.* Structure-based virtual screening and experimental validation of the discovery of inhibitors targeted towards the human coronavirus nucleocapsid protein. *Mol. Biosyst.* **12**, 59–66 (2016).
25. Dayer, M. R., Taleb-Gassabi, S. & Dayer, M. S. Lopinavir; A Potent Drug against Coronavirus Infection: Insight from Molecular Docking Study. *Arch. Clin. Infect. Dis.* **12**, (2017).
26. Chen, Y. W., Yiu, C.-P. & Wong, K.-Y. Prediction of the 2019-nCoV 3C-like Protease (3CLpro) Structure: Virtual Screening Reveals Velpatasvir, Ledipasvir, and Other Drug Repurposing Candidates. *ChemRxiv* (2020). doi:/10.26434/chemrxiv.11831103.v1
27. Anderson, A. C. The Process of Structure-Based Drug Design. *Chem. Biol.* **10**, 787–797 (2003).
28. Mitusińska, K., Magdziarz, T., Bzówka, M., Stańczak, A. & Gora, A. Exploring Solanum tuberosum Epoxide Hydrolase Internal Architecture by Water Molecules Tracking. *Biomolecules* **8**, 143 (2018).
29. Kuipers, R. K. *et al.* 3DM: Systematic analysis of heterogeneous superfamily data to discover protein functionalities. *Proteins Struct. Funct. Bioinforma.* **78**, 2101–2113 (2010).
30. Tsai, M.-Y. *et al.* Essential covalent linkage between the chymotrypsin-like domain and the extra domain of the SARS-CoV main protease. *J. Biochem.* **148**, 349–358 (2010).
31. Needle, D., Lountos, G. T. & Waugh, D. S. Structures of the Middle East respiratory syndrome coronavirus 3C-like protease reveal insights into substrate specificity. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **71**, 1102–1111 (2015).
32. Zhang, L., Lin, D., Sun, X., Rox, K. & Hilgenfeld, R. X-ray Structure of Main Protease of the Novel Coronavirus SARS-CoV-2 Enables Design of α-Ketoamide Inhibitors. *bioRxiv* (2020). doi:10.1101/2020.02.17.952879
33. Anand, K. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *EMBO J.* **21**, 3213–3224 (2002).
34. Lim, L., Shi, J., Mu, Y. & Song, J. Dynamically-Driven Enhancement of the Catalytic Machinery of the SARS 3C-Like Protease by the S284-T285-I286/A Mutations on the Extra Domain. *PLoS One* **9**, e101941 (2014).
35. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382–W388 (2005).
36. Klausen, M. S. *et al.* NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins Struct. Funct. Bioinforma.* **87**, 520–527 (2019).
37. Anandakrishnan, R., Aguilar, B. & Onufriev, A. V. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.* **40**, W537–W541 (2012).
38. Luchko, T. *et al.* Three-Dimensional Molecular Theory of Solvation Coupled with Molecular Dynamics in Amber. *J. Chem. Theory Comput.* **6**, 607–624 (2010).
39. Sindhikara, D. J., Yoshida, N. & Hirata, F. Placevent: An algorithm for prediction of explicit solvent atom distribution-Application to HIV-1 protease and F-ATP synthase. *J. Comput. Chem.* **33**, 1536–1543 (2012).

40. Case, D. A. *et al.* AMBER 2018. (2018).
41. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
42. Pence, H. E. & Williams, A. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.* **87**, 1123–1124 (2010).
43. Nikitin, A. M. & Lyubartsev, A. P. New six-site acetonitrile model for simulations of liquid acetonitrile and its aqueous mixtures. *J. Comput. Chem.* **28**, 2020–2026 (2007).
44. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **25**, 247–260 (2006).
45. Gasteiger, J. & Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **36**, 3219–3228 (1980).
46. Martínez, L., Andrade, R., Birgin, E. G. & Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **30**, 2157–2164 (2009).
47. Delano, W. L. PyMOL: An Open-Source Molecular Graphics Tool. *Ccp4 Newslett Protein Crystallogr* **40**, (2002).
48. Gertz, E. M., Yu, Y.-K., Agarwala, R., Schäffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biol.* **4**, 41 (2006).
49. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
50. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
51. Sievers, F. & Higgins, D. G. Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. in *Multiple Sequence Alignment Methods* 105–116 (2014). doi:10.1007/978-1-62703-646-7_6