

Screening of World Approved Drugs against Highly Dynamical Spike Glycoprotein SARS-CoV-2 using CaverDock and Machine Learning

Gaspar P. Pinto,^{a,b,†} Ondrej Vavra,^{a,b,†} Sergio M. Marques,^{a,b} Jiri Filipovic,^c David Bednar^{a,b,*}, Jiri Damborsky^{a,b,*}

^a Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic; ^b International Clinical Research Centre, St. Ann's Hospital, Brno, Czech Republic; ^c Institute of Computer Science, Masaryk University, Brno, Czech Republic

[†] Authors contributed equally to this work; ^{*} Authors for correspondence: 222755@mail.muni.cz and jiri@chemi.muni.cz.

Abstract

The new severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causes pathological pulmonary symptoms. Most efforts to develop vaccines and drugs against this virus target the spike glycoprotein, particularly its S1 subunit, which is recognised by angiotensin-converting enzyme 2. Here we use the *in-house* developed tool CaverDock to perform virtual screening against spike glycoprotein using a cryogenic electron microscopy structure (PDB-ID: 6VXX) and the representative structures of five most populated clusters from a previously published molecular dynamics simulations. The dataset of ligands was obtained from the ZINC database and consists of drugs approved for clinical use worldwide. Trajectories for the passage of individual drugs through the tunnel of the spike glycoprotein homotrimer, their binding energies within the tunnel, and the duration of their contacts with the trimer's three subunits were computed for the full dataset. Multivariate statistical methods were then used to establish structure-activity relationships and select top candidate molecules. This new protocol for rapid screening of globally approved drugs (4359 ligands) in a multi-state protein structure (6 states) required a total of 26,148 calculations and showed high robustness. The protocol is universal and can be applied to any target protein with an experimental tertiary structure containing protein tunnels or channels. The protocol will be implemented in the next version of CaverWeb (<https://loschmidt.chemi.muni.cz/caverweb/>) to make it accessible to the wider scientific community.

Introduction

A new coronavirus (SARS-CoV-2) outbreak began in Wuhan in the province of Hubei at the end of 2019. Despite many similarities to the 2002 outbreak of SARS-CoV, the new SARS-CoV-2 outbreak had higher morbidity and mortality. Most infected individuals show no or mild symptoms, but some present general complications such as acute respiratory distress syndrome, pneumonia, and septic shock, potentially leading to the patient's death.^{1–4} Drawing on established knowledge about the original virus, research groups worldwide have focused their efforts on two viral proteins: i) the spike (s)-glycoprotein, with the aim of disrupting its recognition of the membrane-bound angiotensin-converting enzyme 2 (ACE-2); and ii) the main viral protease (Mpro, 3CLpro),^{5,6} with the aim of disrupting viral replication by hindering the processing of several polyproteins that are translated from the viral RNA. Another approach for tackling the spread of the new virus builds on work on the original SARS virus, which resulted in the development of a vaccine designed to induce the production of antibodies against the viral s-glycoprotein,^{7,8} preventing it from recognising and binding to ACE-2. Unfortunately, work on this vaccine was discontinued because it had side effects in animal models that prevented its testing in humans.^{9,10}

Currently, there are over 300 therapies^{11–13} in development that are intended to prevent the spread of the virus and end the pandemic (<https://covid-19tracker.milkeninstitute.org/>). These efforts to create a vaccine or a potent inhibitor that can be used as an a posteriori medical treatment with acceptable side-effects are being undertaken by both private companies and academic institutions. Both viral and host proteins are being targeted. While most efforts are focused on disrupting the viral protease or viral polymerase, the viral genome is also being targeted with the aim of disrupting its replication. In particular, the host enzymes involved in nucleotide synthesis are being studied with the aim of halting the final step in viral genome replication. However, most therapies in development target proteins acting upstream of replication; there are almost 40 preclinical and over 30 clinical trials targeting viral surface proteins including the s-glycoprotein. Several host cell membrane proteins are also being targeted, including CD147 and TMPRSS2¹⁴ and, most importantly, ACE-2.^{15,16}

When the SARS-CoV-2 enters the body, s-glycoprotein units on the surface of the virus act as “hooks”, triggering attachment to a host cell.^{17–19} The s-glycoprotein is homo-trimer with three domains—the cytoplasmic tail, the transmembrane region, and most importantly, the ectodomain.²⁰ The ectodomain is further divided into three areas: the proximal membrane region, the S2 subunit, and at the top, the S1 subunit. The receptor-binding domain is located in the S1 subunit. ACE-2 recognises the S1 subunit, and between 1 and 3 s-glycoprotein monomers can bind to ACE-2 by opening and moving upwards. Before the s-glycoprotein/ACE-2 binding event, the covalent bond between subunits S1 and S2

is primed for cleavage to permit the displacement of the S1 subunit. The viral membrane then fuses with that of the host cell via a series of substantial conformational changes.

Several conformations of the viral s-glycoprotein have been observed by electron microscopy, including both semi-open (PDB ID 6VYB) and closed (PDB ID 6VXX) conformations.²¹ The existence of visibly different conformations demonstrates that the viral s-glycoprotein can undergo conformational changes affecting not just its surface but also the gorge within the S1 subunit and the S2 subunit. Because most studies have focused on localised sites such as the active site of the viral Mpro protease or the receptor-binding domain of the s-glycoprotein, we felt that there was a gap in our knowledge about the virus and that several steps along the pathway from infection to propagation remain to be explored. In particular, the long putative tunnel created by the formation of the s-glycoprotein trimer has, to our knowledge, received little study. Studying drug interactions in such long tunnels would be laborious and computationally expensive if using alchemical^{22,23} or ligand migration methods.^{24,25} However, a long tunnel is a perfect target for study with CaverDock.^{26–28}

CaverDock is an *in-house* tool that uses Caver,²⁹ to identify tunnels in protein structures, and an optimised version of the well-established algorithm from AutoDock Vina to calculate possible ligand trajectories along those tunnels and the corresponding binding energies.³⁰ CaverDock discretises each identified tunnel into a series of discs and models a ligand's passage through the tunnel by constraining one ligand atom to lie within a disc a time, sequentially. The ligand's conformation and binding energies are then calculated using Autodock Vina, with the ligand (aside from the constrained atom) being free to explore the conformational space; the protein is treated as a rigid body. Once the conformation and binding energy have been calculated, the constrained atom is shifted to the next disc and the process is repeated until the ligand has moved through the full length of the tunnel. The tool is continuously maintained and is freely available as both a stand-alone program and a webtool named CaverWeb.^{31,32}

Since the start of the pandemic, the scientific community has recognized the need for collaboration and sharing of results by pledging to make data publicly available as soon as possible. In this work, we used data from a 10 μ s molecular dynamics (MD) simulation of the s-glycoprotein trimer conducted at the D.E. Shaw Institute,³³ from which we extracted the main representative conformations. We also used the original closed structure of the s-glycoprotein retrieved from the Protein Data Bank, giving a total of six structures to study.³⁴ Each structure was subjected to virtual screening using every drug in the globally approved drugs subset of the ZINC15 database.³⁵ This subset contains 4359 unique drugs approved by the US Food and Drug Administration, European Medicines Agency, and other significant authorities.

Binding energies along the s-glycoprotein tunnel were calculated for every drug and all six structures. We then compared the results obtained to identify the best ligands for each tunnel position in each conformation. We also analysed each drug to identify the contacts made with each monomeric unit of the s-glycoprotein trimer. This allowed us to select drugs that were predicted to interact with all three monomers and are thus likely to suppress opening of the S1 subunits and thereby prevent the binding of the s-glycoprotein to ACE-2. Finally, we performed quantitative structure-activity analysis (QSAR) to correlate the binding energies of the drugs with their physicochemical properties and used multivariate statistical methods to select top candidates. We are currently implementing this virtual screening methodology into CaverWeb³¹ to allow the community to perform automated calculations against other target proteins using the globally approved drug dataset (Figure 1).

Methods

Construction of the S-glycoprotein ensemble

The cryo-EM structure of the trimeric SARS-CoV-2 spike glycoprotein was obtained from the RCSB Protein Data Bank.³⁶ The selected structure (PDB ID: 6VXX) corresponds to the closed state of this protein. To obtain sufficient conformational diversity for our analysis of the s-glycoprotein trimer, we used the results of a 10 μ s MD simulation conducted by the D. E. Shaw group, which started from the same cryo-EM structure of s-glycoprotein. This trajectory was clustered using the ctptraj³⁷ module of AmberTools 16³⁸ and a distance-based metric defined by the mass-weighted root-mean-square deviation (RMSD) of the backbone atoms of the residues surrounding the gorge of the S1 domain. The RMSD was calculated relative to the starting structure. All residues located within 20 Å of the centreline of the tunnel in the initial s-glycoprotein structure (calculated as described below; 565 in total) were included when calculating this metric. The hierarchical agglomerative clustering algorithm was used with average-linkage, a minimum distance between clusters (epsilon) cut-off of 2.5, sieve 5, and a minimum of 5 clusters.

Tunnel analysis

Before the tunnel analysis, three residue segments were removed from the MD snapshots (residues 365 to 372, 1333 to 1340, and 2301 to 2308). These segments were loose during the simulation and bind to the mouth of the s-glycoprotein tunnel. The tunnel extending through the s-glycoprotein trimer was characterized using HOLE v2.2.005.³⁹ The vector for the HOLE calculation was defined by the

centre points between the C-alpha atoms of the following residues: LYS 1034 and PRO 986 in all three subunits of the s-glycoprotein structure, and LYS 858, 1826, 2794 and PRO 810, 1778, 2746 in the MD snapshots. A sample rate of 0.9 Å was used, and the end radius was set to 10 Å. We analysed the tunnel radii and cut the segment going through the S1 domain until the first extreme tunnel bottleneck was reached; the distance at which this bottleneck was encountered varied between 60 and 80 Å depending on the structure or snapshot under consideration. The output of the HOLE was converted into the CAVER 3 PDB file format²⁹ to enable discretization for CaverDock calculations. However, the tunnel predicted by HOLE for the s-gp structure contained disconnections that made it undiscretisable. Therefore, we remodelled this tunnel using CAVER 3.02, starting from C-alpha of Thr A 1009. The probe radius, shell radius, and shell depth were set to 0.7, 20, and 20, respectively. Finally, the selected tunnel parts were discretized into a series of discs using the discretiser tool with default settings.²⁷

Ligand dataset

The globally approved drug dataset was downloaded from the ZINC database³⁵ on the 26th of May 2020 in mol2 format. Only the first protonation state of each drug molecule was saved. The SMILES codes for all ligands were collected and stored in CSV files, which were then uploaded to the Mordred⁴⁰ web server to obtain the molecular descriptor values needed for the QSAR calculations.

CaverDock calculations

Only the part of the tunnel in the S1 domain was considered in the CaverDock calculations. We discretised the tunnel into a set of discs using the program's default settings.²⁷ The ligand and receptor files were prepared using MGLtools 1.5.7.⁴¹ The grid box was generated around the relevant part of the tunnel using a script from the CaverDock package. The default drag atom (i.e. the atom closest to the centroid of the molecule) was used. Calculations were run in the inward direction only, in the lower-bound trajectory mode.

Principal Components Analysis

Principal Component Analysis (PCA)⁴² was used to facilitate understanding of the data resulting from the CaverDock calculations. The data matrix consisted of 4358 ligands (objects) docked into six different protein states obtained from the CaverDock trajectories. The data for each ligand consisted of its minimum binding energy along the CaverDock trajectory and three percentage values representing the

proportion of the trajectory during which the ligand was in contact with one, two, or all three individual units of the s-glycoprotein trimer. The data were autoscaled to unit variance and centred before analysis.

Partial Least Squares Analysis

Partial Least Squares (PLS) analysis⁴³ was used to explore the relationships between the minimal binding energies of 4358 ligands (objects) docked to six different protein states (dependent variables Y) and 1326 molecular descriptors of individual ligands (independent variables X). 2D and 3D molecular descriptors were calculated using the software tool Mordred⁴⁰, which is particularly suitable for our purpose because it can calculate descriptors even for large molecules. PLS reveals the correlation structure among variables X and Y by reweighting variables X with PLS weights and projecting them to a smaller number of new latent variables. Autoscaled and centred data were used in the PLS analysis. The importance of every molecular descriptor in the model was assessed using the variable importance in the projection (VIP) parameter⁴⁴ and plots of the PLS variable weights.⁴⁴ Internal validation was performed to assess the quality of the developed PLS models⁴⁵ by cross-validation and permutation testing. During cross-validation,⁴³ a portion of the Y data are excluded during model development, and the resulting model is used to predict the missing data. The predictions are then compared to the original data to obtain a Q^2 value. Q^2 provides a more realistic estimate of a model's predictive power than the squared multiple regression coefficient R^2 . In this study, 1/7 of the compounds were deleted during each cross-validation round. During permutation testing, the model was recalculated 999 times by randomly re-ordering the dependent variable y. The statistical package SIMCA-P version 12 (Umetrics, Umeå, Sweden) was used to perform all statistical analyses.

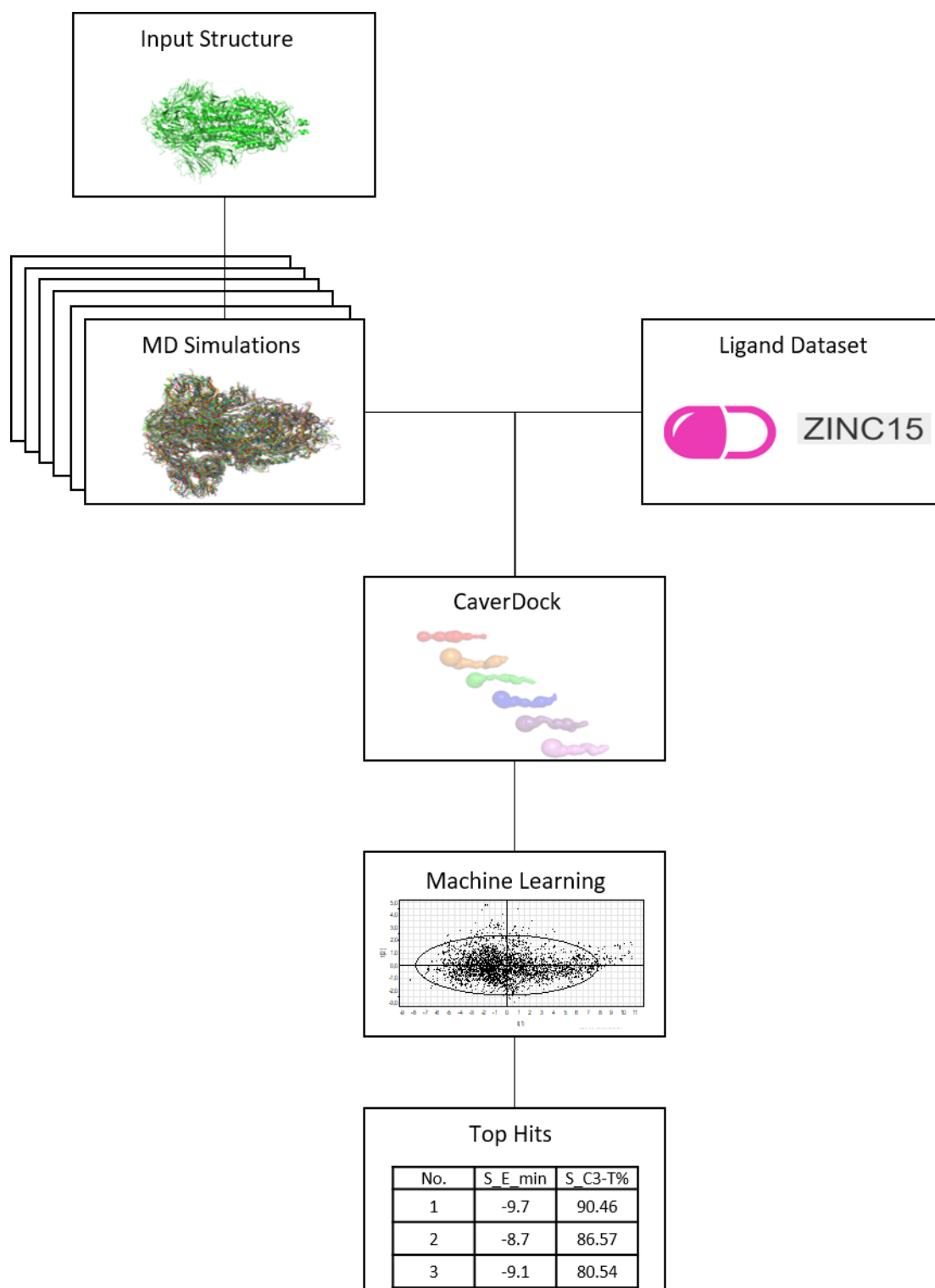


Figure 1 Workflow showing the steps performed during the virtual screening with CaverDock using the full globally approved drug dataset and six protein states, along with the subsequent analytical steps. This workflow is currently being implemented on the CaverWeb²⁹ web server to allow the wider community to easily perform such virtual screens.

Results and discussion

The cryo-EM structure of the spike glycoprotein

We initially analysed the cryogenic electron microscopy (cryo-EM) structure in the closed conformation (PDB ID: 6XVV). This choice was made because our objective was to block the viral infection mechanism by over-stabilizing the closed conformation to suppress the protein's biological activity. Despite missing some loops on the surface, the cryo-EM structure had a sufficiently high resolution and structural integrity inside the tunnel for virtual screening with CaverDock. Because the goal was to block large conformational changes of the s-glycoprotein trimer, we ranked the best binding drugs based on both their overall binding energies and the extent of their contacts with all three monomeric units. Three distinct clusters of drugs with binding profiles showing clear energy minima were identified, each binding to a different region of the tunnel (Figure 2). The first cluster consisted of drugs binding in the region immediately behind the first bottleneck of the subunit S1 gorge, between 12 Å and 21 Å from the trimer's surface. Since this region is immediately behind the tunnel's second tightest bottleneck, we hypothesise that drugs in this cluster are flexible enough to cross that narrow part of the tunnel and then undergo a conformational change to adopt an optimal binding conformation.

The second and smallest cluster of drugs binds in the middle of the tunnel. Although we consider this group to be a cluster, the binding positions of the drugs at the extremes of the cluster differ by 10 Å: ZINC000004099004 binds 26 Å from the surface, while ZINC000008214470 binds at 36 Å. The final region of the tunnel is also the most populated; 99.5% of the drugs tested in the virtual screen bind most strongly in its deepest third, between 45 Å and 65 Å from the surface. All of the top ten drugs identified in this study (Figure 2) belong to this final cluster (Electronic Supporting Information - ESI) and have consistently lower binding energies than any drug binding preferentially in the other two regions. In addition, most of the drugs with the lowest binding energies belong to the cluster binding at position 3 (ESI). The profile of the tunnel in this region is narrower than in the other tunnel regions.

The S-glycoprotein dynamical ensemble

The D. E. Shaw research institute studied the dynamical ensemble of the s-glycoprotein by performing a 10 μ s MD simulation starting from the closed cryo-EM structure mentioned above (PDB ID: 6VXX). This simulation became stable after 6 μ s, as shown by the root-mean-square deviation (RMSD) plot (SI-Figure 1). Due to the s-glycoprotein's high flexibility, the cryo-EM structure lacks several parts of its sequence, causing several segments to be seemingly disconnected from the rest of the structure. Unfortunately, during the MD simulation, two fragments corresponding to residues 446-454 and 461-468 detached themselves from their correct positions and drifted to different locations within the structure. These events are responsible for the two spikes seen in the RMSD plots at around 2.1 and 5.2 μ s (SI-Figure 2). These unrealistically dynamical fragments, which were originally located on the outer surface of the s-glycoprotein, were excluded from all subsequent analyses in this work.

We clustered the MD snapshots based on the RMSD of the gorge residues to obtain diverse but biologically relevant conformations of the s-glycoprotein. The obtained clusters are ranked in terms of their populations. The most populated cluster, s1, dominated almost the entire second half of the trajectory (SI-Figure 1). The mean RMSD of the gorge residues in this cluster was 3.46 ± 0.13 Å, which is close to the average value for the entire simulation (3.66 ± 0.38 Å) (SI-Figure 3). Conversely, the least populated cluster (s7) had RMSD values indicating that it remained close to its starting structure (1.62 ± 0.69 Å). Representative structures of the clusters (SI-Figure 3) and their tunnels (SI-Figures 5 and 6) were also obtained, enabling further analysis (SI-Figure 3).

CaverDock calculations were performed using representative structures of the 5 most populated clusters in the same way as described for the cryo-EM structure (Figure 2). Each tunnel had a unique profile, but in all cases, the narrowest section was in the deepest region of the tunnel, close to the S2 subunit. The vast majority of the ligands have their lowest binding energies in this region (Figure 3). The sole exception is the most populated state, s1, for which the majority of the ligands have their lowest binding energies in the middle of the tunnel (Figure 2). The tunnel in this state is slightly wider than in the other states, making it difficult for ligands to form contacts with all three monomers. The tendency for the binding energies of drugs to be lowest immediately before or after a bottleneck was seen for all states.

Principal Component Analysis (PCA)

Multivariate statistical analyses were used to: (i) comprehend the large data sets obtained from the CaverDock calculations, (ii) establish structure-activity relationships, and (iii) select the best potential drug candidates. Two statistically significant models were generated by PCA using the CaverDock results obtained using the set of 4358 ligands and six protein states. The data used in the PCA were the minimum binding energies for each drug along the trajectory and the proportions of the trajectory during which the docked ligand was in contact with one, two, and all three individual subunits of the s-glycoprotein trimer, expressed as percentages.

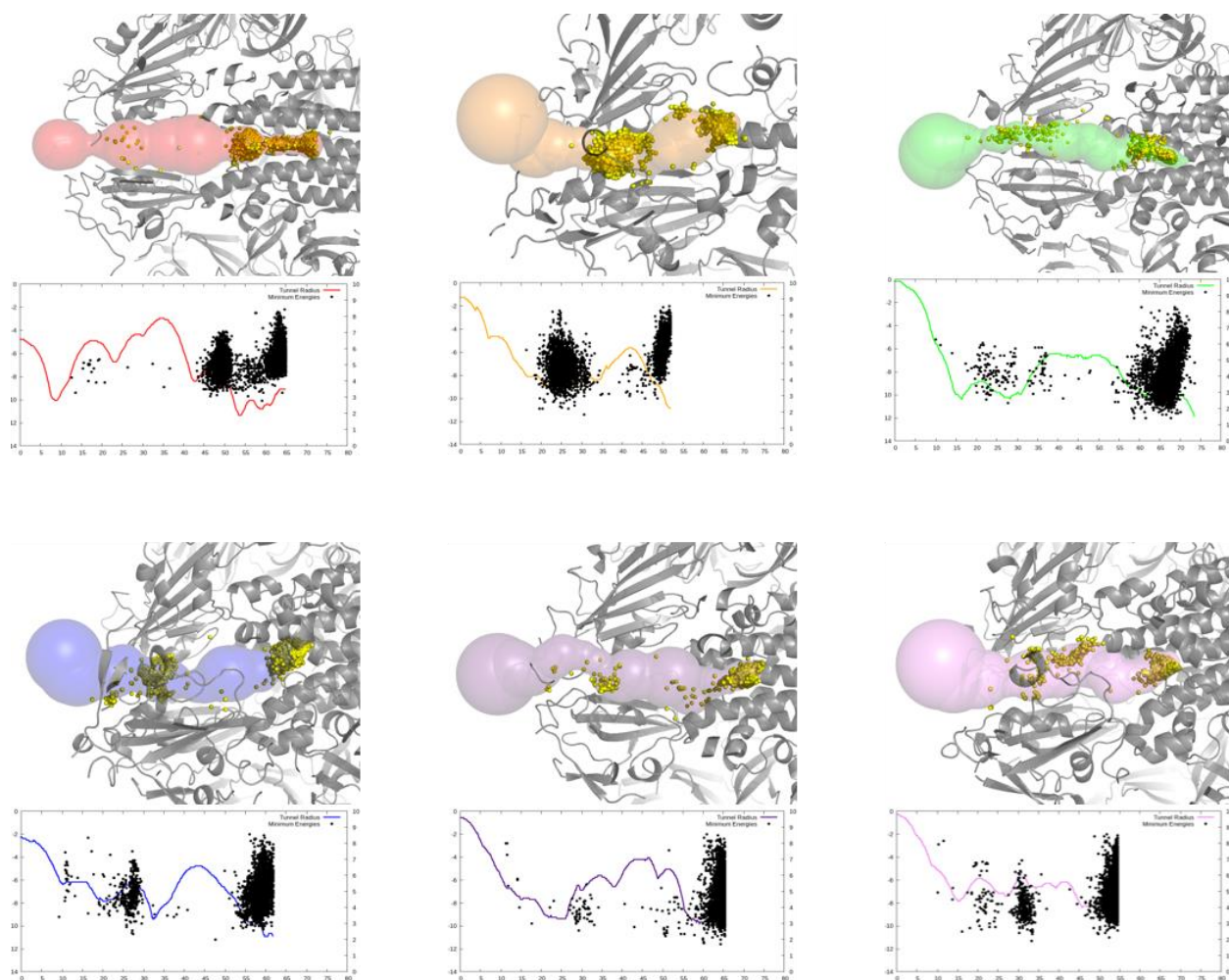


Figure 2 Tunnels in the six protein states showing the regions where the drugs bind with the lowest binding energy. Top: Visualization of the tunnel used for virtual screening in the six protein states analysed with CaverDock. These states are the cryo-EM structure (red) and 5 representative structures (s1 in orange, s2 in green, s3 in blue, s4 in purple and s5 in pink) obtained by clustering the results of an MD simulation. Yellow spheres in the tunnels indicate

the centre of mass of each drug when bound at the location where it binds most strongly. The plots below each structure show the corresponding tunnel profiles (in Å) using solid lines. Each black dot indicates the position where one drug binds most strongly together with the corresponding binding energy in kcal/mol.

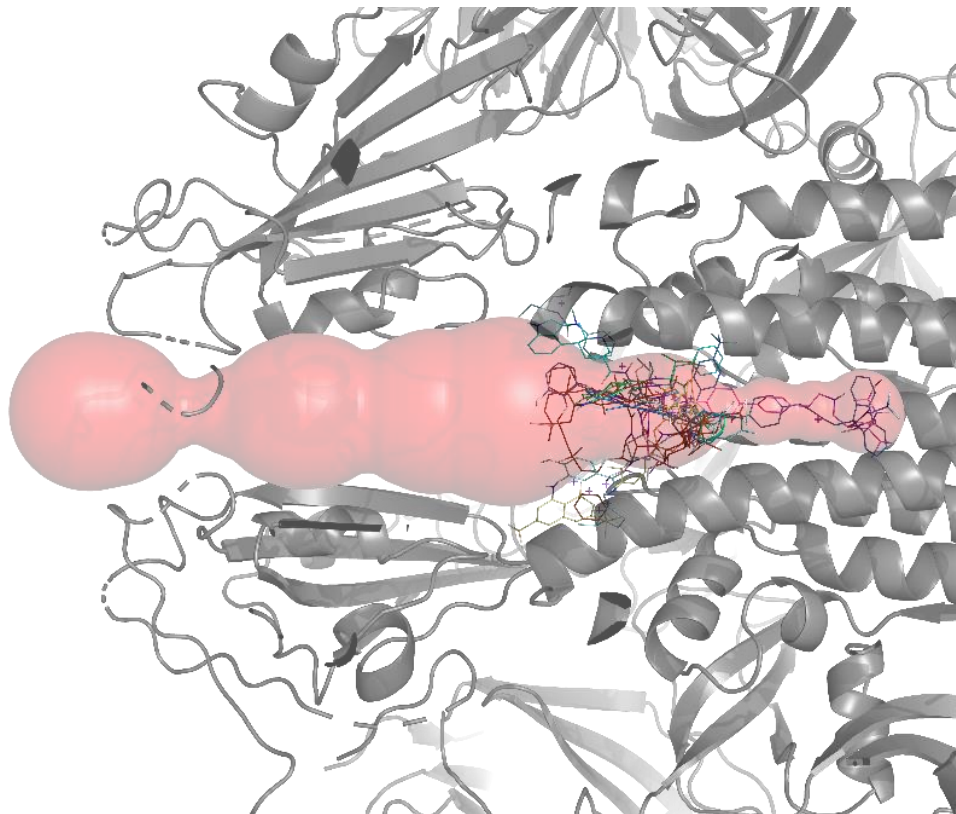


Figure 3 Visualization of the tunnel in the cryo-EM structure with the top ten inhibitors bound to the positions corresponding to their lowest binding energy. Drugs were ranked by multivariate analysis. The protein structure (PDB ID: 6VXX) is shown as a grey ribbon, while the tunnel predicted by CaverDock is indicated by the red surface. Inhibitors are shown using all-atom models, coloured by atom type.

The first PCA model (PCA-1) used 24 variables: 3 related to the minimum binding energies for each protein state, and 3 quantifying the percentages of the trajectory during which the drug was in contact with 1, 2, and 3 units of the trimeric s-glycoprotein. Ten statistically significant principal components were obtained, collectively explaining 98% of the variation in the data. The second model, PCA-2, was generated using 12 variables representing the energy minima and the percentages of each trajectory during which the drug was in contact with all three monomeric units of the s-glycoprotein trimer for each of the six studied protein states. This model yielded only two principal components that explained 85% and 8% of the variation in the data, respectively. Because it had only two principal components, this model was

easier to interpret than the first. The top hits predicted by the two models were very similar, so only the results obtained with the simpler model 2 will be discussed further. By inspecting the distribution of the docked compounds in the 2D space spanned by the first two principal components (Figure 4), the compounds interacting most strongly with all three subunits of the spike protein were identified (ESI). Such compounds are most likely to modify the conformational behaviour of the s-glycoprotein and thus affect its biological function. The distribution of the 12 variables used to cluster the ligands is shown at the bottom of Figure 4.

Partial Least Squares Analysis (PLS)

A PLS analysis was performed to correlate the minimum binding energies for each ligand from the CaverDock calculations with the molecular descriptors of the docked ligands. Binding energies calculated for all six states of the s-glycoprotein were considered simultaneously using a single PLS model. The initial model, PLS-1, used 1326 independent variables and consisted of four principal components collectively explaining 87% of the variation in the data. The correlation coefficient ($R^2 = 0.87$) and cross-validated correlation coefficient ($Q^2 = 0.87$) of this model are identical, suggesting excellent predictive power. To simplify the model, the variable selection was performed. Specifically, independent variables were selected based on their position in the loadings plot and variable importance in the projection (VIP) plot. In this way, the number of variables was reduced from 1326 to 56. A new model generated with these variables, PLS-2, had three principal components, with an R^2 of 0.84 and a Q^2 of 0.84. Validation by permutation testing - scrambling the Y variables while keeping the X-matrix unchanged – indicated that this correlation would be very unlikely to be observed by chance, as expected given the large number of observations on which the model is based.

The observed minimal binding energies were plotted against the corresponding predicted values for the starting structure 6VXX and state s4, for which the worst and best fits were obtained, respectively (SI-Figure 7). VIP values were computed to quantify the relative importance of the chosen molecular descriptors in explaining the differences in the minimum binding energies for all six states (SI-Figure 8). The most influential variables were FMF (a molecular framework ratio descriptor of the shape of the molecule), BalabanJ (Balaban's J graph index, which describes the molecular structure of small molecules), piPC (a path count descriptor of molecular topology), MWC and SRW04 (walk count descriptors, the latter of which relates to self-returning walks), VR (a normalised Randic-like eigenvector-based index derived from the Barysz matrix, weighted by atomic number), and VE (the average coefficients of the last eigenvectors of the Barysz matrix, weighted by van der Waals' volume). Detailed information about all

molecular descriptors computed using MORDRED is available at <https://mordred-descriptor.github.io/documentation/master/descriptors.html> and in the book 3D QSAR in Drug Design.³⁹

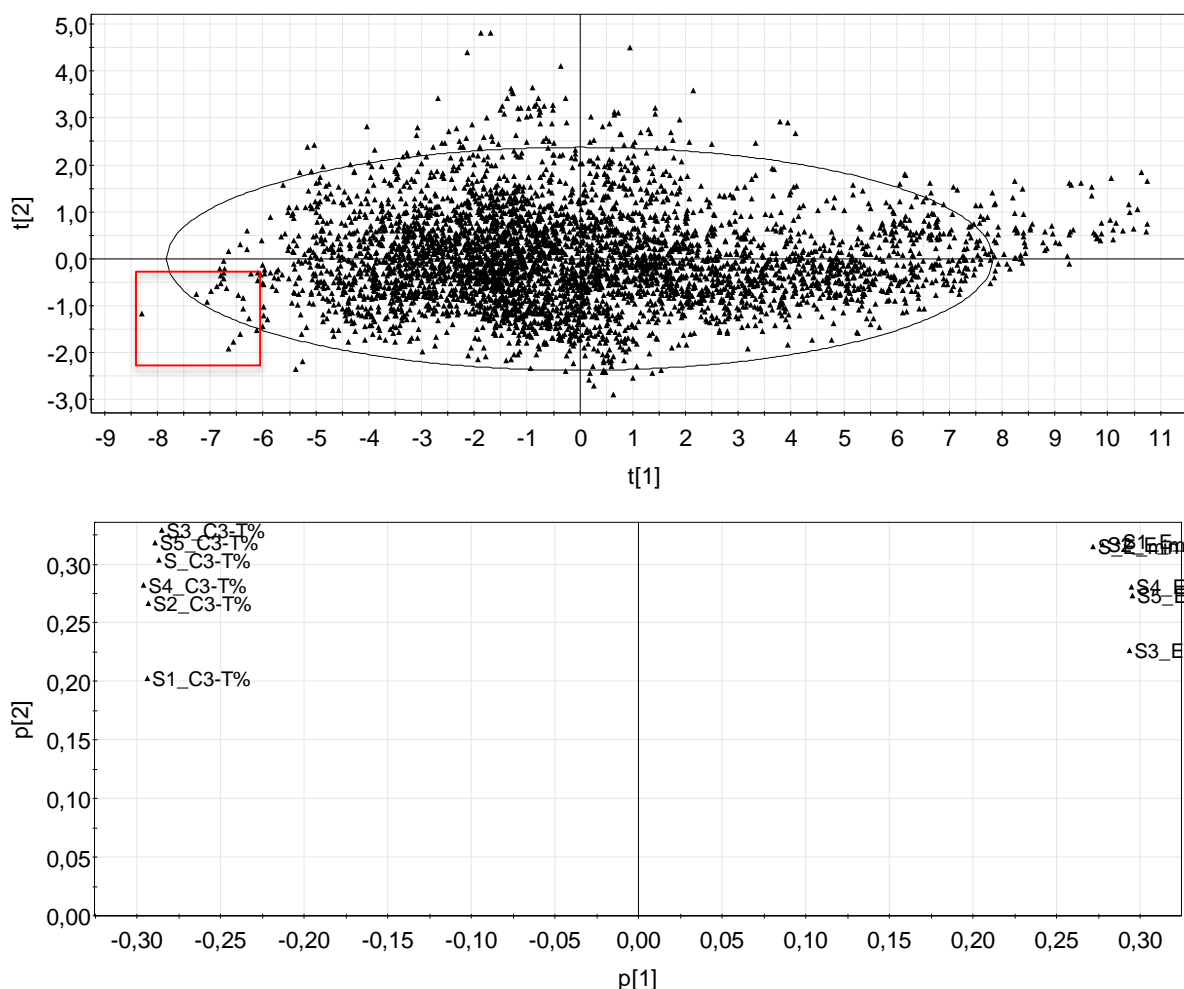


Figure 4 Scores and loadings plot of the first two principal components of the second PCA model. Top: Scores plot of the first two principal components showing the distributions of all studied compounds based on their minimal binding energies and number of contacts with the three subunits of the spike glycoprotein. The top hits were selected from this plot. The positions of the compounds in the 2D space are determined by the locations of variables in the loadings plot (bottom). Compounds showing the strongest binding to all three units in the different states of the spike protein are located on the left of the plot, inside the red box. Bottom: Loadings plot of the first two principal components showing the distribution of the variables in the 2D space. This plot corresponds to the scores plot presented above. The variables describing the minimal binding energies calculated for the six different s-glycoprotein states are on the right, while those describing the contact percentage with the three individual subunits of the spike protein trimer are located on the left.

Top Ranked Drugs

We obtained a ranking of the best binders from the PCA and selected the top ten for further analysis (Figure 5). These ligands had consistently low binding energies in all of the studied protein structures and exhibited a high percentage of contacts with all three monomeric units of the s-glycoprotein trimer during the CaverDock simulations. Although drugs in clusters S1, S3, and S5 occasionally formed contacts with only one monomer, these cases represented less than 10% of the corresponding trajectory. This ranking reflects our assumption that strong interactions with all three monomers in different states of the trimer will reduce the trimer's capacity for conformational change, which is essential for the biological activity of the spike glycoprotein. We also found that multivariate statistical methods were needed to rank the drugs meaningfully. For example, a simple ranking of the drugs based on their minimum binding energies would not have placed Daclatasvir (Figure 5) in the top 10 because its binding energies for all six conformations are higher than those of some drugs that were not selected. It was thus clear that interaction with all three monomers was weighted strongly in the ranking of the drugs; for three of the studied protein states, Daclatasvir was observed in contact with two and three subunits of the s-glycoprotein trimer, and in the remaining three states (clusters s1, s3 and s5) it was in contact with two or three subunits for at least 96.4% of the trajectory (SI-Table1).

Among the drugs ranked in the top 10 was a dye for cataract surgery (ZINC000169289767), three drugs currently used as antiviral agents against the hepatitis C virus (ZINC000164760756, ZINC000936069565, ZINC000068204830), an antifungal (ZINC000028639340), a microsomal triglyceride transfer protein inhibitor (ZINC000027990463), a hepatoprotective drug for chronic hepatitis (ZINC000096015174), an agent used to treat squamous cell carcinoma of the head and neck (ZINC000003934128), a vasoconstrictor used to treat migraines (ZINC000003978005), and an agent for treating cerebral and peripheral vascular events that are also used in Alzheimer's studies to inhibit γ -secretase (ZINC000003995616).

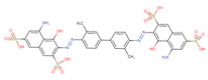
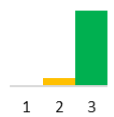


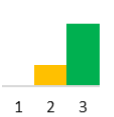

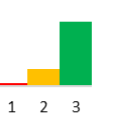
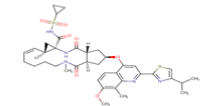
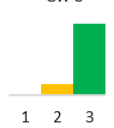
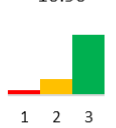
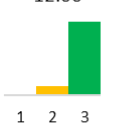
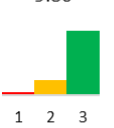
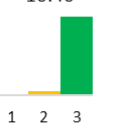
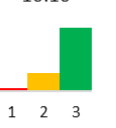
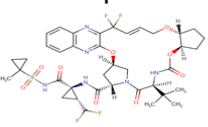
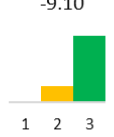
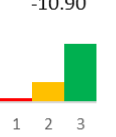
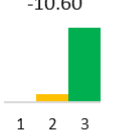
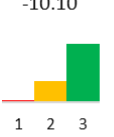
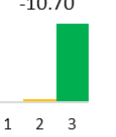

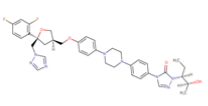
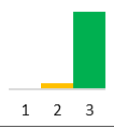



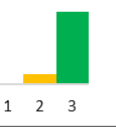

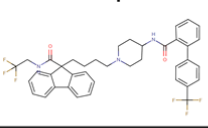
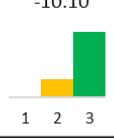




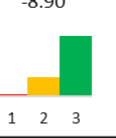
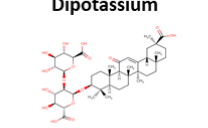






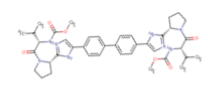
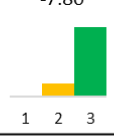


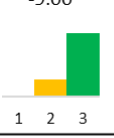

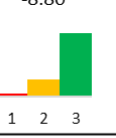
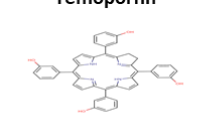
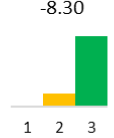
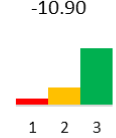
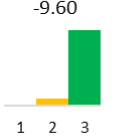
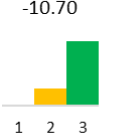
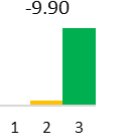
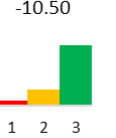
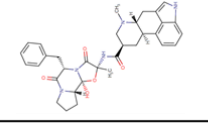
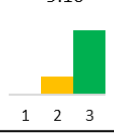
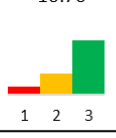


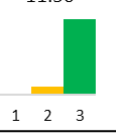
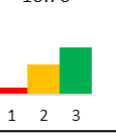
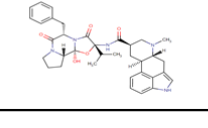
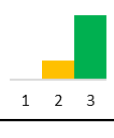
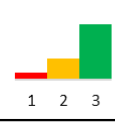
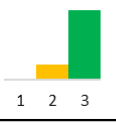

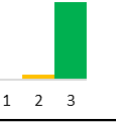
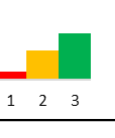
	6VXX	S1	S2	S3	S4	S5
Trypan blue 	-9.70 	-10.60 	-12.00 	-11.20 	-11.40 	-10.30 
Simeprevir 	-8.70 	-10.90 	-12.00 	-9.80 	-10.40 	-10.10 
Glecaprevir 	-9.10 	-10.90 	-10.60 	-10.10 	-10.70 	-10.20 
Posaconazole 	-9.20 	-9.80 	-11.20 	-9.00 	-10.40 	-9.60 
Lomitapide 	-10.10 	-10.00 	-11.50 	-8.90 	-10.10 	-8.90 
Glycyrrhizinate Dipotassium 	-9.00 	-9.70 	-11.30 	-9.00 	-10.30 	-10.30 
Daclatasvir 	-7.80 	-9.69 	-9.80 	-9.00 	-9.40 	-8.80 
Temoporfin 	-8.30 	-10.90 	-9.60 	-10.70 	-9.90 	-10.50 
Dihydroergotamine 	-9.10 	-10.70 	-11.10 	-10.10 	-11.50 	-10.70 
Dihydroergocristine 	-9.20 	-10.40 	-11.00 	-9.90 	-11.50 	-10.60 

Figure 5 Top ten inhibitors predicted using CaverDock simulations and machine learning. Drugs are shown in the first column with their names and chemical structures. Binding energies per drug for each protein state (cryoEM 6VXX and MD states S1-S5) are reported in kcal/mol. The bar plots under each binding energy represent the percentage of the corresponding trajectory during which these compounds formed contacts with one monomer (red), two monomers (yellow), and three monomers (green).

Conclusions

Here we describe a computational workflow that was used to perform virtual screening based on CaverDock trajectories for 4358 drug molecules and six conformational states of the s-glycoprotein of SARS-CoV-2. This analysis involved a total of 26,148 calculations. Each calculation took a real-time average of 37 minutes to complete on 8 CPUs, making the method sufficiently fast for thorough virtual screening. It should be noted that the length of the tunnel in the studied s-glycoprotein structures ranges between 57 Å and 77 Å, making it several times longer than typical enzyme tunnels. However, this long tunnel can serve as a good representative of the structural features present in transmembrane proteins.

We used machine learning to identify the most promising drug candidates based on their strength of binding inside the tunnel and their likely ability to prevent the s-glycoprotein trimer from undergoing functionally necessary conformational change. Although we only selected 10 inhibitors here for the sake of brevity, this number could easily be increased. CaverDock is fast enough to analyse an even higher number of snapshots to cover the protein's conformational space more comprehensively or to examine a significantly greater number of ligands. Importantly, this workflow is currently being made available on the CaverWeb tool to enable automated virtual screenings of the ZINC globally approved drugs dataset. This will enable researchers around the world to perform virtual screening and data analysis in the same way as reported here, in a user-friendly manner. It will also be possible to export the results as comma separated value (CSV) files and/or Pymol sessions to be opened and processed locally by the user. The procedure will be applicable to any protein with an available tertiary structure containing tunnels or channels and should thus find diverse applications in drug design, protein engineering, and metabolic engineering.

Acknowledgements

The authors would like to express their deep gratitude to Stanislav Mazurenko, Joan Planas-Iglesias, Rayyan Khan, Milos Musil, Jan Stourac and Jan Mican (Masaryk University, Brno, Czech Republic) for help with the project's conceptualization. The research was conducted with financial support from the Czech Ministry of Education (02.1.01/0.0/0.0/18_046/0015975, CZ.02.1.01/0.0/0.0/16_026/0008451), the Grant Agency of the Czech Republic (20-15915Y) and European Union (857560, 720776 and 814418). Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140).

Bibliography

- 1 M. Cascella, M. Rajnik, A. Cuomo, S. C. Dulebohn and R. Di Napoli, in *StatPearls*, StatPearls Publishing, Treasure Island (FL), 2020.
- 2 S. Murthy, C. D. Gomersall and R. A. Fowler, Care for Critically Ill Patients With COVID-19, *JAMA*, 2020, **323**, 1499–1500.
- 3 D. L. Heymann, N. Shindo and WHO Scientific and Technical Advisory Group for Infectious Hazards, COVID-19: what is next for public health?, *Lancet*, 2020, **395**, 542–545.
- 4 D. S. Hui, E. I Azhar, T. A. Madani, F. Ntoumi, R. Kock, O. Dar, G. Ippolito, T. D. Mchugh, Z. A. Memish, C. Drosten, A. Zumla and E. Petersen, The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health — The latest 2019 novel coronavirus outbreak in Wuhan, China, *International Journal of Infectious Diseases*, 2020, **91**, 264–266.
- 5 H. M. Mengist, X. Fan and T. Jin, Designing of improved drugs for COVID-19: Crystal structure of SARS-CoV-2 main protease M pro, *Signal Transduction and Targeted Therapy*, 2020, **5**, 1–2.
- 6 Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, Y. Duan, J. Yu, L. Wang, K. Yang, F. Liu, R. Jiang, X. Yang, T. You, X. Liu, X. Yang, F. Bai, H. Liu, X. Liu, L. W. Guddat, W. Xu, G. Xiao, C. Qin, Z. Shi, H. Jiang, Z. Rao and H. Yang, Structure of M pro from SARS-CoV-2 and discovery of its inhibitors, *Nature*, 2020, **582**, 289–293.
- 7 H. Pearson, Caution raised over SARS vaccine, *Nature*, , DOI:10.1038/news050110-3.
- 8 Integrative illustration for coronavirus outreach, <https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000815>, (accessed 10 September 2020).
- 9 C.-T. Tseng, E. Sbrana, N. Iwata-Yoshikawa, P. C. Newman, T. Garron, R. L. Atmar, C. J. Peters and R. B. Couch, Immunization with SARS Coronavirus Vaccines Leads to Pulmonary Immunopathology on Challenge with the SARS Virus, *PLoS One*, , DOI:10.1371/journal.pone.0035421.
- 10 Z. Yang, H. C. Werner, W. Kong, K. Leung, E. Traggiai, A. Lanzavecchia and G. J. Nabel, Evasion of antibody neutralization in emerging severe acute respiratory syndrome coronaviruses, *PNAS*, 2005, **102**, 797–801.
- 11 A. Chernyshev, Pharmaceutical Targeting the Envelope Protein of SARS-CoV-2: the Screening for Inhibitors in Approved Drugs, , DOI:10.26434/chemrxiv.12286421.v1.
- 12 A. Jiménez-Alberto, R. M. Ribas-Aparicio, G. Aparicio-Ozores and J. A. Castelán-Vega, Virtual screening of approved drugs as potential SARS-CoV-2 main protease inhibitors, *Comput Biol Chem*, 2020, **88**, 107325.

- 13 K. Miroshnychenko and A. V. Shestopalova, Combined Use of Amentoflavone and Ledipasvir Could Interfere with Binding of Spike Glycoprotein of SARS-CoV-2 to ACE2: The Results of Molecular Docking Study, , DOI:10.26434/chemrxiv.12377870.v1.
- 14 I. Glowacka, S. Bertram, M. A. Müller, P. Allen, E. Soilleux, S. Pfefferle, I. Steffen, T. S. Tsegaye, Y. He, K. Gnirss, D. Niemeyer, H. Schneider, C. Drosten and S. Pöhlmann, Evidence that TMPRSS2 Activates the Severe Acute Respiratory Syndrome Coronavirus Spike Protein for Membrane Fusion and Reduces Viral Control by the Humoral Immune Response, *Journal of Virology*, 2011, **85**, 4122–4134.
- 15 South Andrew M., Brady Tammy M. and Flynn Joseph T., ACE2 (Angiotensin-Converting Enzyme 2), COVID-19, and ACE Inhibitor and Ang II (Angiotensin II) Receptor Blocker Use During the Pandemic, *Hypertension*, 2020, **76**, 16–22.
- 16 R. Chowdhury and C. D. Maranas, Biophysical characterization of the SARS-CoV-2 spike protein binding with the ACE2 receptor and implications for infectivity, *bioRxiv*, 2020, 2020.03.30.015891.
- 17 D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, C.-L. Hsieh, O. Abiona, B. S. Graham and J. S. McLellan, Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation, *Science*, 2020, **367**, 1260–1263.
- 18 S. Satarker and M. Nampoothiri, Structural Proteins in Severe Acute Respiratory Syndrome Coronavirus-2, *Arch Med Res*, 2020, **51**, 482–491.
- 19 Y. L. Siu, K. T. Teoh, J. Lo, C. M. Chan, F. Kien, N. Escriou, S. W. Tsao, J. M. Nicholls, R. Altmeyer, J. S. M. Peiris, R. Bruzzone and B. Nal, The M, E, and N Structural Proteins of the Severe Acute Respiratory Syndrome Coronavirus Are Required for Efficient Assembly, Trafficking, and Release of Virus-Like Particles, *Journal of Virology*, 2008, **82**, 11318–11330.
- 20 X. Ou, Y. Liu, X. Lei, P. Li, D. Mi, L. Ren, L. Guo, R. Guo, T. Chen, J. Hu, Z. Xiang, Z. Mu, X. Chen, J. Chen, K. Hu, Q. Jin, J. Wang and Z. Qian, Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV, *Nature Communications*, 2020, **11**, 1620.
- 21 A. C. Walls, Y.-J. Park, M. A. Tortorici, A. Wall, A. T. McGuire and D. Veasler, Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein, *Cell*, 2020, **181**, 281-292.e6.
- 22 J. D. Chodera, D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson and V. S. Pande, Alchemical free energy methods for drug discovery: Progress and challenges, *Curr Opin Struct Biol*, 2011, **21**, 150–160.
- 23 I. Cabeza de Vaca, R. Zarzuela, J. Tirado-Rives and W. L. Jorgensen, Robust Free Energy Perturbation Protocols for Creating Molecules in Solution, *J. Chem. Theory Comput.*, 2019, **15**, 3941–3948.
- 24 V. Guallar, C. Lu, K. Borrelli, T. Egawa and S.-R. Yeh, Ligand Migration in the Truncated Hemoglobin-II from Mycobacterium tuberculosis THE ROLE OF G8 TRYPTOPHAN, *J. Biol. Chem.*, 2009, **284**, 3106–3116.
- 25 M. F. Lucas and V. Guallar, An Atomistic View on Human Hemoglobin Carbon Monoxide Migration Processes, *Biophysical Journal*, 2012, **102**, 887–896.
- 26 G. P. Pinto, O. Vavra, J. Filipovic, J. Stourac, D. Bednar and J. Damborsky, Fast Screening of Inhibitor Binding/Unbinding Using Novel Software Tool CaverDock, *Front. Chem.*, , DOI:10.3389/fchem.2019.00709.
- 27 J. Filipovic, O. Vávra, J. Plhák, D. Bednar, S. M. Marques, J. Brezovsky, L. Matyska and J. Damborsky, CaverDock: A Novel Method for the Fast Analysis of Ligand Transport, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019, 1–1.
- 28 O. Vavra, J. Filipovic, J. Plhak, D. Bednar, S. M. Marques, J. Brezovsky, J. Stourac, L. Matyska and J. Damborsky, CaverDock: a molecular docking-based tool to analyse ligand transport through protein tunnels and channels, *Bioinformatics*, 2019, **35**, 4986–4993.
- 29 E. Chovancova, A. Pavelka, P. Benes, O. Strnad, J. Brezovsky, B. Kozlikova, A. Gora, V. Sustr, M. Klvana, P. Medek, L. Biedermannova, J. Sochor and J. Damborsky, CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures, *PLOS Computational Biology*, 2012, **8**, e1002708.
- 30 O. Trott and A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, *J Comput Chem*, 2010, **31**, 455–461.

- 31 J. Stourac, O. Vavra, P. Kokkonen, J. Filipovic, G. Pinto, J. Brezovsky, J. Damborsky and D. Bednar, Caver Web 1.0: identification of tunnels and channels in proteins and analysis of ligand transport, *Nucleic Acids Res.*, 2019, **47**, W414–W422.
- 32 J. Stourac, O. Vavra, P. Kokkonen, J. Filipovic, G. Pinto, A. Schenkmyerova, J. Damborsky and D. Bednar, Caver web: identification of tunnels and channels in proteins and analysis of ligand transport, *Journal of Biotechnology*, 2019, **305**, S72.
- 33 D.E. Shaw Research, Molecular Dynamics Simulations Related to SARS-CoV-2, https://www.deshawresearch.com/downloads/download_trajectory_sarscov2.cgi/.
- 34 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, The Protein Data Bank, *Nucleic Acids Res*, 2000, **28**, 235–242.
- 35 T. Sterling and J. J. Irwin, ZINC 15 – Ligand Discovery for Everyone, *J. Chem. Inf. Model.*, 2015, **55**, 2324–2337.
- 36 P. W. Rose, C. Bi, W. F. Bluhm, C. H. Christie, D. Dimitropoulos, S. Dutta, R. K. Green, D. S. Goodsell, A. Prlić, M. Quesada, G. B. Quinn, A. G. Ramos, J. D. Westbrook, J. Young, C. Zardecki, H. M. Berman and P. E. Bourne, The RCSB Protein Data Bank: new resources for research and education, *Nucleic Acids Res*, 2013, **41**, D475–D482.
- 37 PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data | Journal of Chemical Theory and Computation, <https://pubs.acs.org/doi/10.1021/ct400341p>, (accessed 9 September 2020).
- 38 D. Case, R. Betz, D. S. Cerutti, T. Cheatham, T. Darden, R. Duke, T. J. Giese, H. Gohlke, A. Götz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.-S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko and P. Kollman, *Amber 16*, University of California, San Francisco., 2016.
- 39 O. S. Smart, J. G. Neduvellil, X. Wang, B. A. Wallace and M. S. Sansom, HOLE: a program for the analysis of the pore dimensions of ion channel structural models, *J Mol Graph*, 1996, **14**, 354–360, 376.
- 40 H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, Mordred: a molecular descriptor calculator, *Journal of Cheminformatics*, 2018, **10**, 4.
- 41 G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson, AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility, *J Comput Chem*, 2009, **30**, 2785–2791.
- 42 A. Höskuldsson, PLS regression methods, *Journal of Chemometrics*, 1988, **2**, 211–228.
- 43 H. Kubinyi, Ed., *3D QSAR in Drug Design: Volume 1: Theory Methods and Applications*, Springer Netherlands, 1994.
- 44 S. Wold and W. J. Dunn, Multivariate quantitative structure-activity relationships (QSAR): conditions for their applicability, *J. Chem. Inf. Comput. Sci.*, 1983, **23**, 6–13.
- 45 S. Wold, Validation of QSAR's, *Quantitative Structure-Activity Relationships*, 1991, **10**, 191–193.