SARS-CoV-2 3D database: Understanding the Coronavirus Proteome and Evaluating Possible Drug Targets.

Ali F. Alsulami[†], Sherine E. Thomas[†], Arian R. Jamasb[†], Christopher A. Beaudoin[†], Ismail Moghul[4], Bridget Bannerman, Liviu Copoiu, Sundeep Chaitanya Vedithi[†], Pedro Torres[†] and Tom L. Blundell

Corresponding author: Tom Blundell, Department of Biochemistry, University of Cambridge, Cambridge, CB2 1GA, UK. E-mail: tlb20@cam.ac.uk
[†] Contributed equally to this work.

Ali F. Alsulami, Arian R. Jamasb, Chris Beaudoin, and Liviu Copoiu are PhD candidates in the Department of Biochemistry, at the University of Cambridge. Their research areas are drug discovery, computational biology, and bioinformatics.

Sherine Thomas is a postdoc in the Department of Biochemistry, University of Cambridge, Cambridge. Her research focuses on drug discovery for infectious diseases.

Ismail Moghul is a PhD candidate at UCL Cancer Institute, University College London. His research areas focus on bioinformatics.

Bridget Bannerman is a postdoc at the Molecular Immunity Unit, Department of Medicine University of Cambridge, MRC Laboratory of Molecular Biology. Her research focuses on developing predictive models for various pathogenic micro-organisms, reviewing treatment management strategies for SARS-CoV-2 and designing tools and strategies for surveillance of antimicrobial resistance.

Sundeep Chaitanya Vedithi is Research Director of the American Leprosy Mission and leads a group of postdoc in the Department of Biochemistry, University of Cambridge, focusing on bioinformatics and drug discovery for *Mycobacterium leprae*.

Pedro Torres is a Professor at the Laboratório de Modelagem e Dinâmica Molecular, Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil. His research focuses on bioinformatics tools for proteomic databases and virtual screening and docking for early drug discovery.

Tom Blundell is a Professor at the Department of Biochemistry, University of Cambridge. His research focuses on structural biology, bioinformatics and drug discovery for cancer and mycobacterial infections.

Abstract

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a rapidly growing infectious disease, widely spread with high mortality rates. Since the release of the SARS-CoV-2 genome sequence in March 2020, there has been an international focus on developing target-based drug discovery, which also requires knowledge of the 3D structure of the proteome. Where there are no experimentally solved structures, our group has created 3D models with coverage of 97.5% and characterised them using state-of-the-art computational approaches. Models of protomers and oligomers, together with predictions of substrate and allosteric binding sites, protein-ligand docking, SARS-CoV-2 protein interactions with human proteins, impacts of mutations, and mapped solved experimental structures are freely available for download. These are implemented in SARS CoV-2 3D, a comprehensive and user-friendly database, available at https://sars3d.com/. This provides essential information for drug discovery, both to evaluate targets and design new potential therapeutics.

Keywords; SARS-CoV-2 proteome modeling, SARS-CoV-2 3D database, SARS-CoV-2 drug targets, proteome analysis, drug discovery.

Introduction:

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first detected in late 2019 in Wuhan (Hubei, China). Since then it has spread dramatically, infecting over 40 million people, with over 1 million deaths reported to date and causing major health and economic challenges globally[1]. The virus belongs to the coronavirus family that includes SARS-CoV-1 and MERS-CoV, and is characterised by a positive-sense single-stranded RNA genome. SARS-CoV-2 has approximately 79% sequence similarity to SARS-CoV-1 and 50% similarity to MERS-CoV[2]. The involvement of bats in transmission to humans is still not clear, although the sequence similarity of human SARS-CoV-2 to that of the bat coronavirus RaTG13 is ~96%[3]. High genetic variability and recombination are believed to enable widespread adaptive evolution of SARS-CoV-2 in humans around the world[4][5].

The full genome of SARS-CoV-2 RNA was released in March 2020 (GenBank: MN908947.3) (https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3) and was shown to be a single strand of 29.9kb, with a 3' poly (A) region and 5' -methylguanosine cap. The entry of the virus into the host cell is facilitated by the viral Spike protein, which includes a receptor-binding domain that recognizes the ACE-2 receptor on the surface of the human respiratory epithelial cell. Upon entry of the SARS-CoV-2, a single positive strand of RNA is released into the cytoplasm of the host cell[6]. The synthesis of viral RNA involves two steps: the first involves genome replication during which the RNA is transcribed via replication-transcription complexes (RTC), producing negative RNA that replicates to positive sense RNA for repackaging into the virion, and the second occurs when the negative sense RNA is transcribed via discontinuous transcription into different mRNA lengths that in turn are translated into various virion proteins. These proteins are essential for viral particle formation and for the viral replication cycle to continue[6, 7].

The SARS-CoV-2 genome can be considered as comprising three regions. The first comprises the RNA used directly as template to translate the non-structural polyproteins pp1a and pp1ab through ribosome frameshifting. The polyproteins are autoproteolytically processed into 16 non-structural proteins (Nsp1-16); these proteins also assemble to form the replication-transcription complex RTC[8]. The second region codes for the four structural proteins that occur in all coronaviruses: Spike (S), envelope (E), membrane protein (M), and nucleoprotein (N), and the third for the accessory proteins such as ORF (3a, 6, 7a, 7b, 8, 9b, 10)[9].

The SARS-CoV-2 virus, like other coronaviruses, is an obligate pathogen that uses the translation machinery of host cells for viral gene expression. It has been established that the Nsp1 viral protein of SARS-CoV-2 disrupts gene expression by inhibiting translation and promoting mRNA degradation in the host cells[10]. Recent studies[11] demonstrate interactions between Nsp1 of the SARS-CoV-2 virus and the host's cellular proteins, such as DNA polymerase and primase subunits, thereby affecting processes such as DNA replication and damage repair. In addition, the SARS-CoV-2 viral protein Nsp2 interacts with the cap-binding protein eIF4E, an enzyme that catalyses a key regulatory step in mRNA translation in the host. The main protease is involved in a number of cellular processes, including translation, replication, cell death and protein modification and regulation. The interactions of the viral proteins with the host's cellular proteins have many roles including ensuring overall efficiency in the synthesis and replication of the viral proteins within the host[11]. These interactions also inhibit the host cellular mRNA processes and accelerate viral disease progression in the host[11][12].

Various viral proteins are being studied as potential drug-targets and candidates for vaccine development. Attempts to therapeutically modulate the S glycoprotein include developing antibodies against ACE2 and S protein subunits, and identifying small molecules to abolish the S-ACE2 interaction[13, 14]. The crucial role played by this interaction in viral entry requires priming of S protein by host proteases, such as the serine protease TMPRSS2, subtilisin-like furin and cathepsin L, inhibition of which is being explored[15–18]. Yet another viral structural protein of considerable therapeutic interest is the viral nucleoprotein protein that packages the viral genomic RNA. Due to its immunogenic potential, this protein is being studied as a candidate for vaccines and diagnostics[14, 19].

The viral proteases (3CL^{pro} and PL^{pro}) on the other hand are non-structural proteins, which cleave the viral polyproteins pp1a and pp1b, resulting in the formation of 16 non-structural proteins that perform a range of important functions[20]. Nsp12-16 are believed to be involved in the crucial replication-transcription complexes of the virus along with a number of viral cofactors such as Nsp7, 8 and 10[21, 22]. Various drug candidates that target the proteases, such as Lopinavir, Ritonavir and Disulfiram, as well as those that target the viral RNA-dependent RNA polymerase (Nsp12), including Remdesivir, Favipiravir and Ribavirin, are currently in clinical trials along with others that modulate inhibition of virus entry and control of immune response[14, 23]. The coronavirus NTPase / helicase (Nsp13), a multifunctional protein that forms part of the core replicative complex of the virus, along with Nsp12 and Nsp14 (exoribonuclease), is also considered an important target for anti-viral drug discovery. In addition to disrupting RNA secondary structures during replication, this protein is thought to catalyse the first step in 5'- cap formation of the viral mRNA[26][27]. The subsequent steps in the process are understood to be mediated by the Nsp14 (N7-MTase domain) and Nsp16 (2' O MTase)[22]. In addition, a number of viral proteins have been identified to play a role in evading the host innate immune response. These proteins, including non-structural proteins (Nsp1, Nsp9 and Nsp15), viral proteases (Nsp3 and Nsp5) and accessory proteins (ORF3b, ORF6), are thus being further investigated for the development of novel therapeutics[11, 14, 28].

Definition of the structural proteome is essential for understanding the molecular basis of any disease and for structure-guided design of new drugs. Therefore, databases act not only as a hub for experimental structures, most importantly the Protein Data Bank (PDB)[29], but also as an archive for the homology models generated by different labs, for example Genome3D[30]. Our group in the past has successfully generated bacterial structural proteome databases, for example CHOPIN for *Mycobacterium tuberculosis*[31] and Mabellini for *Mycobacterium abscessus*[32], which include experimentally defined and modelled protomers and ligand interactions.

Here we describe a novel, extensively annotated SARS-CoV-2 3D database. We focus on structures of all gene products and their higher-order assemblies, i.e. homo- and hetero-oligomers, and trans-membrane regions, as well as ligand and metal-ion interactions, with acceptable assessment score. The SARS-CoV-2 3D database provides a web interface that is user-friendly and easily accessible, so that end-users can navigate, inspect and download the 3D structural proteome data, visualise modelled oligomeric complexes, analyse pockets of modelled structures and investigate SARS-CoV-2 human protein interactions, mutations, and protein-ligand docking.

Methods

Proteome Modelling

All SARS-CoV-2 modelled structures built using MODELLER[33] version(2.24). All sequences for SARS-CoV-2, obtained from GenBank: MN908947.3, were compared to structures in the Protein Data Bank (PDB) [29] using PSI-BLAST[34], which relies on Position-Specific Scoring Matrix (PSSM) profile-profile alignment, in conjunction with several multiple-sequence alignment methods, including FUGUE[35], which recognizes distant homologs using combined information from both sequence and structure, and HHsearch[36], which uses hidden Markov models (HHM). Templates for models were selected based on percentage sequence identity, amino acid sequence coverage, and structure resolutions. The selected templates were re-aligned to the target sequence using Clustal Omega software[37], and the produced alignment was used in MODELLER to build the final modelled structure. All the hetero-atoms and cofactors, such as bound ligands and metal ions, were obtained from selected templates. The SARS-CoV-2 genome has nine genes annotated as trans-membrane proteins (Nsp3, Nsp4, Nsp6, ORF3a, ORF7a, ORF7b, S, E, and M) and the Orientations of Proteins in Membranes (OPM) database[38] was used to annotate these transmembrane regions.

The homo-oligomeric models were built manually and automatically via the ProtCHOIR pipeline (P. Torres and T. L. Blundell, manuscript in preparation https://doi.org/10.5281/zenodo.3384945). For example, Nucleoprotein (N) has been solved experimentally in separate PDB entries; PDB ID: 6M3M is a monomer covering residues 49-174, and PDB ID: 6WZQ is a homodimer covering residues 252-364. Initially, we have built the unsolved missing regions between positions 1-49 using (PDB ID: 5NP3) as a template and regions between the 174-252 using structures from PDB IDs: 6K12, 1F15 as templates. All the regions were assembled to obtain a full protomer, followed by aligning the modelled protomer to the experimentally defined homodimer to obtain the full homodimeric modelled structure. The homodimeric S protein has been modelled in both open and the closed conformational states (Supplementary S9, S10).

Protein models of homo-oligomers such as E protein were generated using our novel modelling tool: ProtCHOIR. The ProtCHOIR pipeline relies on a homo-oligomeric-protein database and uses well-established tools such as MolProbity[39], PISA[40], GESAMT (PSI-BLAST and MODELLER to search for homologous templates, assemble the model homo-oligomer and finally assess its accuracy. This pipeline also allows for the generation of protomeric/monomeric structures to cope with the need for high-throughput comparative modelling of entire proteomes. ProtCHOIR creates oligomeric structures by performing a search on the locally created databases using PSI-BLAST before comparative modelling using MODELLER and assembling the oligomeric structure, for which a detailed report is output, documenting all the analyses performed. All generated models are assessed using MolProbity, GESAMT and PISA (in the case of oligomeric models).

The hetero-oligomer models were built based on the selected templates, where there is evidence of equivalent interactions for homologues in the literature. For example, the Nsp14-Nsp10 complex with functional ligands was built based on the experimental structures of the SARS-CoV-1 heterodimer (PDB IDs: 5C8S, 5C8T).

After obtaining the final model, side-chain energy minimization implemented in Foldit[41] was used to remove side-chain clashes. We have utilised jsPISA[42] to calculate the surface interface assemblies for only homo- and hetero-oligomeric models. The interfacial regions are characterised by parameters such as interface area (Å), solvation energy (kcal/mol), total binding energy (kcal/mol), hydrophobic logP-value, hydrogen bonds, salt bridges, and disulphide bonds. Furthermore, the MolProbity was used as quality assessment to validate the quality of modelled structures. The MolProbity log-weighted single value is a score derived from combination of multiple features such as percentage Ramachandran not-favoured, clashscore, and percentage bad side-chain rotamers.

We have not modelled proteins, such as Nsp5 and Nsp16, which have full structural coverage define by X-ray, cryo-EM, and NMR experimental approaches. Structures were downloaded from the RCSB PDB, and saved as biological assemblies. We have implemented the Fpocket into our final models in order to identify potential ligands, and allosteric binding sites. The pocket ranking is based on the possibility of binding small drug-like molecules.

Methods for Predicting impacts of Mutations

In order to understand the impacts of mutations on overall protein stability, where there are experimentally-solved protein structures, Nsp3, Nsp5, Nsp9, Nsp12, Nsp13, Nsp15, and S, we utilized mCSM-Stability[43], mCSM-PPI[43], DeepDDG[44], PROVEAN[45], MAESTRO[46], and I-Mutant[47]. These tools either categorize the mutation as stabilizing or destabilizing or quantify the change in predicted protein folding values (ΔΔG) between wild type and mutant forms (ΔΔG = ΔGwildtype - ΔGmutant). PROVEAN[45] and I-Mutant2.0[48] are sequence-based tools, which take into account the evolutionary conservation of amino acid motifs. mCSM[43], DeepDDG[44], and MAESTRO[46] use the 3D structure of the protein to predict the thermodynamic stability changes between mutant and wild type forms based on residue/atom distances, residue conservation, energy calculations, and solvent accessibility. mCSM-PPI[43] was used to investigate impacts on protein-protein interactions and oligomeric interfaces. The mCSM-based tools use graph signatures to encode the atomic environment and train a predictive model, DeepDDG[44] utilizes a neural network with shared network parameters for each target residue-neighbour residue pair, and MAESTRO[46] implements a multi-agent machine learning system. The most destabilizing mutations, i.e. those reported to have the lowest average ΔΔG values, are displayed in the tables corresponding to each protein on the website.

Protein-Protein Docking

A list of high-confidence viral-human protein-protein interactions determined through tandem affinity purification mass spectrometry was obtained from Gordon *et al*.[11]. UNIPROT[49] identifiers for host proteins were mapped to corresponding PDB entries using the genome-scale models where available with protein structures (using the GEM-PRO pipeline implemented in SSBIO[50]), resulting in a list of 39 unique human structures for docking after quality assessment. Quality assessment involved performing pairwise sequence alignments between the UniProt sequence and the PDB sequence for each polypeptide chain in the candidate structures, candidate structures retrieved using the PDBe best structures API. Quality assessment thresholds ensured that representative structures have >90% coverage of the UniProt sequence, >95% identity to the sequence, excluding a 5% on the sequence termini, and resolution <5 Å. PDB files were cleaned to remove solvents and ligands, and the highest scoring chain in the alignment selected as a representative structure. Sequence alignments and quality assessment were performed using EMBOSS Needle [51] via the SSBIO python library. Protein-protein docking for known-interaction pairs was performed using ClusPro[52], and the docked structures were minimized using CHARMM22[53]. The top 4 docked poses with lowest CHARMM22 energy score were selected and presented into the SAR CoV-2 & Human Proteins interaction table. However, all other poses can be downloaded from the website help page.

Ligand Virtual Screening

The following non-structural proteins and their corresponding Protein Data Bank (PDB) files were screened using the FDA approved drug library of 1930 compounds extracted from the eDrug 3D web-resource[54].

1. Nsp3 (PL Proteinase)- PDB Id 6XAA
2. Nsp5 (3CLPro/Main Protease)- PDB Id 6XMK
3. Nsp12 (RNA dependent RNA polymerase in complex with nsp7 and nsp8 cofactors) – PDB Id 7C2K
4. Nsp14 (Guanine N7 methyltransferase). At the time of virtual screening, the crystal structure of nsp14 of SARS-CoV-2 was not solved experimentally. Hence, the crystal structure of nsp14 of SARS-CoV-1 virus with PDB Id 5C8S has been used in the virtual screening experiments as this protein has 95% amino acid sequence identity to that of SARS-CoV-2.
5. Nsp15 (Uridylate specific endoribonuclease) - PDB Id: 6XDH
6. Nsp16 (2'-O-methyltransferase) - PDB Id: 6WKQ
7. S protein subunit 1 PDB Id: 6VSB

The above-mentioned non-structural proteins were selected for virtual screening as most of them were identified as potential drug targets in SARS-CoV-2. The proteins and the ligands were prepared using protein preparation wizard[55] and Ligprep[56] modules in Schrodinger Suite 2020-2. For 1930 drug molecules, 24,992 conformers were generated and submitted to the virtual screening workflow[56] To facilitate molecular docking with Glide[57], grid boxes were generated[58] using specific atoms of existing native/reference ligand in the active sites. Later the reference ligands were extracted from the binding sites and re-docked into the Glide-specified grid boxes to determine the differences in binding patterns between the structurally solved native pose and the docked pose. This process was repeated with each atom in the reference ligand until the lowest root mean square deviation (RMSD) was obtained between the two poses described above. Once this was achieved, the prepared drug library was docked into the active site noted by the grid with the lowest RMSD to reference. Three levels of docking were performed: high throughput virtual screening (HTVS) was first used to select the top 10% of all the docks based on the Glide docking scores, secondly these were repeated using Glide Standard precision docking and thirdly this was followed by the top 10% with Glide extra-precision docking. This workflow substantially reduces the number of ligand conformers to be docked by extra-precision docking, eventually reducing time and providing higher quality docking scores. The MM/GBSA[59] method was also employed to determine the ligand-binding affinities.

SARS-CoV-2 3D web interface

The front-end of the web interface is implemented in HTML5, CSS, Bootstrap 4.5, jQuery, and Font Awesome library to add icon functionality. All the tables are stored in PostgreSQL server and Express.js, a web application framework for Node.js used in the back-end to query the stored tables in PostgreSQL. We used the Embedded JavaScript template (EJS) as a template engine that dynamically generates the final HTML. The dynamic sunburst chart was created using the Plotly.js library, and the network graph interaction was created in Data Driven Documents (D3.js) package. To facilitate programmatic access to the SARS-CoV-2 3D database we have constructed a stored model, protein-protein, and PDB tables that have data available through RESTFUL APIs. The data are returned in a JSON object that is easily parsable by other software.

Results

SARS-CoV-2 proteome analysis

We have built oligomeric models with almost full sequence coverage for the 21 proteins that's partially solved or do not have structures deposited in the PDB (Figure 1). The longest modelled protein is papain-like proteinase (PL^{pro}) Nsp3, with 1,945 residues, whereas the smallest is Nsp11 with 13 residues. The modelled structures exist as monomers, or as homo- or hetero-oligomeric complexes. The ligands and cofactors bound to the structures are retrieved from selected experimental structures of homologues. There are 15 proteins where experimentally determined structures have a mean sequence coverage is 81.41%, whereas the coverage for the modelled structures is 97.5%. The mean average of the quality assessment MolProbity score for our models is 2.52 with standard deviation of 0.65, the lowest and highest values are 0.85 and 3.47 respectively.

Examples of SARS-CoV-2 modelled proteins

In this section we illustrate some of the modelled structures listed in Table 1 produced using the approaches described in the Methods section. The remaining models are presented in the Supplementary Data.

Complex of non-structural proteins Nsp14 and Nsp10

The non-structural protein 14 (Nsp14), conserved throughout the CoV family, plays fundamental roles in the viral replication/transcription complex. Structurally it has two domains, the N-terminal exoribonuclease (ExoN) and the C-terminal N7 methyltransferase domain. The Nsp14-Nsp10 complex is essential for viral replication and transcription, and disruption of the complex decreases replication fidelity[60]. The full model structure of the Nsp14-Nsp10 hetero-dimer was built based on (PDB ID: 5C8T, 5C8S) (Figure 2A). The model has a MolProbity score of 3.16, TM-score: 0.99, RMSD: 0.54 Å. Both the s-adenosyl-l-homocysteine (SAH) and guanosine-p3-adenosine-5',5'-triphosphate (G3A) ligands, as well as the three zinc ions, are modelled from selected templates (Figure 2A). Since the selected templates are very close homologues and the TM-score between the model and the template is high, this model could serve as a reliable target in molecular docking studies as well as mutation analysis.

Envelope protein

The E protein, 75 amino acids long, is one of the smallest transmembrane proteins in SARS-CoV-2. The lack of E protein not only reduces viral loads but also budding of vesicles from the plasma membrane. Structurally, the E protein consists of three regions: N-terminal, transmembrane and C-terminal. The template selected for modelling the homo-pentameric E protein is PDB ID: 5X29, the NMR structure of the SARS-CoV-1 E protein[61]. The model has a MolProbity score of 3.22, TM-score of 0.99, RMSD of 0.25 Å. The transmembrane region is annotated using the Orientations of Proteins in Membranes (OPM) database. (Figure 2B).

Membrane protein

The M protein, one of the most abundant type III glycoproteins in coronavirus particles, is located in the viral envelope between the S proteins and facilitates the virus budding[62]. Knowledge of the structure of this viral membrane protein is important for developing new therapeutics that stop the virus budding inside host cells. The M protein exists as a dimer with each protomer structurally comprising a short N-terminal region, followed by the N-domain and three transmembrane helices. It was modelled using four templates (PDB IDs: 3A7K_A, 5UTT_A, 6SPB_V, 6XDC), yielding a model with 3.04 MolProbity score. The transmembrane region was annotated using OPM. (Figure 2C)

Non-structural protein 1

The non-structural protein 1 (Nsp1) plays an essential role in suppressing gene expression of the host cell via association with the ribosome. Nsp1 in SARS-CoV-1 inhibits cellular anti-viral defence mechanisms via partial shutdown of the innate immune system, so facilitating viral replication[10]·[63]. Nsp1 in SARS-CoV-2 has 84% sequence identity to Nsp1 SARS-CoV-1, indicating similarity of biological function[64]. The C-terminal of Nsp1, covering residues 148-180, has recently been published as a complex with human 40S ribosomal subunit PDB ID: 6ZLW, covering residues 148-180. The modelled Nsp1 structure was generated using multiple templates PDB ID 2GDT_A, 5C5S_A, 6ZLW_i. We were able to model the full-length protein in complex with the human 40S ribosomal subunit to produce a hetero 35-mer complex. The modelled structure could plausibly serve as a target to study the effect of missense mutations on the interaction between Nsp1 and 40S ribosomal subunit. (Figure 2D)

Protein-protein docking

We modelled structures for 308 experimentally confirmed viral-human protein-protein interactions. Viral infections can, at one level, be viewed as a perturbation of host protein interaction networks. Structural modelling of these complexes should lead to improved understanding of how SARS-CoV-2 manipulates and disrupts cellular processes. Furthermore, most antiviral development programs focus on inhibition of viral proteins[65]**,** resulting in a small pool of targets. Inhibition of vital viral protein-human-host protein interactions provides other avenues for small-molecule development and drug repurposing screens. We have mapped the viral human protein-protein interactions using D3.js Force-Directed Graph, which provides an interactive data visualization for web browsers. The viral human protein-protein interaction data are stored in JSON format and loaded from an Application Program Interface (APIs) to produce the 2D graph. The node in each 2D graph always represented the SARS-CoV-2 protein, whereas the edges represented the human proteins. In SARS-CoV-2 3D database the human drug target proteins are highlighted as a black arrow where human non-drug target proteins are highlighted in grey arrow.

Small molecule ligand docking against target proteins

Top hits, scored by Glide XP[57] and MM/GBSA for receptors Nsp 3, Nsp 5, Nsp 12, Nsp 14, Nsp15, Nsp 16, and S protein are included in the database as a table. Scores for all receptors were noted to be above the docking score of the reference/native ligand. Users can download the docked poses as PDB files from the web-resource and also view the docked structures on the MolStar viewer. In the viewer, docked ligands can be selected using the corresponding name/id of the ligand, located at the top of the sequence viewer. By focusing on the ligand, the user can recognize the interatomic interactions that the ligand forms with the surrounding residue environment. All possible interactions can be noted by toggling options on the “Representations” menu on the right-hand panel. As noted by other groups[66][67], we have identified Mitoxantrone, a drug that was used to treat acute myeloid leukemia and multiple sclerosis, as a potential hit against Nsp3 in our virtual screening experiments. Protokylol, a β-adrenergic receptor agonist, was also noted among the top hits for Nsp15 and Nsp16. These screens provide initial information on the potential of repositioning of FDA-approved drugs to act on specific target proteins in SARS-CoV-2. All used anti-viral FDA drugs are mapped to DrugBank (www.drugbank.ca) with a link on the web interface.

Mutation Analysis

Single nucleotide mutations leading to amino acid changes can have a significant impact on protein structure and function. In order to understand the structural impact of mutations on the viral proteins, we used several tools, such as mCSM-Stability, mCSM-PPI, Provean, Maestro, I-Mutant, and DeepDDG, to estimate the change in folding energy between mutant and native forms. We performed a saturation mutagenesis substituting each amino acid in the protein sequence on experimentally-validated viral protein structures to gain a comprehensive view of the most stabilizing and destabilizing amino acid substitutions. The mutations that had, on average, the most destabilizing effect on local and global protein stability are reported. Destabilizing mutations in the vicinity of enzyme active sites or protein-ligand/protein interfaces can negatively impact necessary interactions for the viral life cycle. For example, in our analysis, several amino acid positions in different proteins were largely destabilizing regardless of the substituted amino acid. These residues were primarily located internally inside hydrophobic-rich protein domains. The most destabilizing mutated residues for S protein were found buried within the subunit 1 C-terminal and N-terminal receptor binding domains (e.g. T54, L303, G431, N439). The most destabilizing mutations for the main protease were predicted near drug binding sites (e.g. V20, V148) (supplementary S14), while the most destabilizing mutations for Nsp12 were located near cofactor binding residues (e.g. A418, K508)[68–70]. The frequently reported D614G Spike mutant was predicted to have a small destabilizing effect using mCSM (-0.210 kJ/mol) and DeepDDG (-0.129 kJ/mol), a small stabilizing effect with Maestro (0.214 kJ/mol), and a larger stabilizing effect with Provean (0.903 kJ/mol). Of note, most of the protein stability prediction tools require protein models to be in their apo forms; thus, protein stability related to glycation and cofactor-binding was not evaluated in this analysis. These data can provide valuable insights into the functionality and structural robustness of different protein domains through understanding their capacity to retain stability after mutation.

Front-end pages

The website can be accessed from HTTPS URL: https://sars3d.com/. The website has a ‘Navbar’ including the help page, which shows in detail how the database can be accessed programmatically through the RESTFUL API. The first page with a jumbotron identifies four features of the SARS-CoV-2 3D database: oligomeric modelling, binding-site prediction of modelled structures, mutation analysis, and protein-protein and protein-ligand docking. For simplicity, the database can be queried in two ways: through a table on the right that contains the gene identifier, or through a sunburst viewer on the left that contains the gene identifier. (Figure-3A)

The results page gives a brief description of the queried gene and links to different modelling pipelines, such as SWISS Model[71], I-TASSER[72] and AlphaFold[73], to compare the different modelling approaches. The model page integrates multiple tools, such as the MolStar (Mol*) viewer that is used to view the model, PDB structures, pocket predictions, and potential ligands obtained through virtual screening. The Mol* viewer shows the desired structural sequence at the top of the viewer. Calculating interactions, for example hydrogen bonding or π–π stacking, of any selected residue or ligands, can suggest interactions that will be important to consider in designing new ligands to satisfy binding-site residue interactions. The Fpocket prediction is located in a table presented under Mol* viewer; the pockets can be displayed for a target structure in the Model/PDB table in order to explore ligand binding sites. In addition, the experimentally solved structures have been annotated from RCSB and can be loaded into the Mol* viewer. A UniProt viewer[49] integrates data curated by UniProt teams, including domain and binding-site prediction, topology. This viewer has the advantage of automatically incorporating information updates from UniProt as they become available in the future.

The Models/PDB table contains a target model structure that can be viewed in the Mol* viewer and downloaded locally or programmatically through a RESTFUL API. Also, it contains other information such as templates used to build the model, MolProbity score, PDB coverage, model coverage, sequence length, oligomeric states, oligomeric interfaces, and model information.

Discussion

There are now multiple online databases specific for genomic data, including GenBank[74] containing all publicly available DNA sequences, Ensembl[75] containing annotated genomes mainly of vertebrates, and GISAID[76] containing annotated viral genomes. Most of these genome databases have manually curated data and display sequences of the target gene. In addition, there are multiple proteomic databases such as RCSB PDB for experimentally solved structures. The

CATH[77] and SCOP[78] databases focus on protein domains. The UniProt database merges the sequence and structural annotations. COSMIC (the Catalogue of Somatic Mutation in Cancer)[79] is concerned with the impact of mutations in human cancer. Recently developed databases, such as Genome3D comprising collaborative work between different groups based on comparative modelling, are beginning to narrow the huge gap between three-dimensional structures and sequence annotation.

Two 3-D structural databases, developed by our group using in-house pipeline tools such as Vivace[32], are Chopin, the *M. tuberculosis* database originally established in 2015, and Mabellini for *M. abscessus* established in 2019. Since building our models we have noted the databases such as Coronavirus3D[80], which are focused on mapping mutations to experimentally solved structures. I-TASSER, entitled Genome Wide structure and function modelling of SARS-CoV-2, provides a set of monomeric models. Swiss Model provides a set of monomeric and oligomeric models. AlphaFold provides a set of monomeric models. D3Targets-2019-nCoV a molecular docking database for SARS CoV-2 drug targets. We have provided links into our database to allow comparisons with all these major contributions.

The approach described here in the SARS-CoV-2 3D database includes not only domain or protomer structures but also multi-domain structures, homo- and hetero-oligomers, transmembrane proteins, and ligand and cofactor binding for the full SARS-CoV-2 proteome. We generated protomer structures with good sequence coverage and possible oligomeric states from selected templates. All the models in the SARS-CoV-2 3D database have good quality scores and exhibit known folds, therefore providing a reliable basis for multiple purposes, such as protein-ligand docking, protein-protein docking and mutation analysis. We have reviewed all solved experimental structures from PDB, we have mapped protein-protein interactions between SARS-CoV-2 and human proteins, considered protein-ligand docking of proved anti-viral FDA drugs, and predicted the impacts of mutations using a diverse set of tools.

Modelling the relative orientations of individual domains in multi-domain gene products remains a challenge in the field of comparative modelling. However, over the past decade the resolution revolution in cryo-EM has proved able to provide reliable data for multi-domain and multi-component systems. Modelling the interactions of the proteins of SARS-CoV-2 with human proteins through identification of conserved amino acids located on interfacial regions is a possible solution for improving docking protein-protein interaction. We have included several methods to assess the impacts of mutations in order to indicate uncertainties and help identify false-positive predictions. We intend to update new experimentally solved structures from PDB as they become available, and address all aspects of new data in future updates of SARS-CoV-2 3D.

Conclusion

SARS-CoV-2 3D database, a comprehensive resource for the SARS-CoV-2 proteome, is based on three-dimensional structures, using either experimentally solved structures or through comparative computational modelling based on structures of close homologues. Since the SARS-CoV-2 proteome is relatively small with 25 gene products, we have extended the methodology, used to construct our previously published databases such as Mabellini, and Chopin, to include models of transmembrane proteins, multi-domain proteins, and homo- and hetero-oligomers manually as well as using our recently-developed software ProtCHOIR that is designed to build homo-oligomeric assemblies. While structural models for SARS-CoV-2 have been developed by multiple tools such as SWISS MODEL, I-TASSER, and AlphaFold, our SAR-CoV-2 3D database is the first resource to our knowledge integrating not only the SARS-CoV-2 proteome model but also protein-protein interactions, mutation analysis, and protein-ligand docking, all in a user-friendly manner.

We have built an entirely new website based on Node.js with new functionality such as protein-protein interactions 2D viewer to visualize human SARS-CoV-2 interactions, UniProt viewer for sequence annotations, sunburst to navigate and query the proteome, MolStar viewer to render the 3D structure, and a RESTFUL API.

The goals of the SARS-CoV-2 3D database are to bring together the most important data that help the drug discovery process by modelling unsolved proteins structures, often as oligomers, exploiting protein-protein interaction data that have recently been published. Furthermore, Fpocket is used to predict the potential ligand binding sites and potential allosteric sites. The SARS-CoV-2 3D database is regularly updated with new mutation data and a range of sequence and structure-based methods are exploited to analyse the impacts of mutations. With the likely future availability of new templates, the model structures will be updated to increase the model accuracy. Efforts are in process to add new annotations such as pathway analysis and to model further protein-protein interactions.

Key points

- SARS-CoV-2 3D is a comprehensive database for COVID-19 bringing together computational and experimental 3D structures on one platform that provides essential information for drug discovery.
- SARS-CoV-2 3D database includes protomers, homo-oligomers, hetero-oligomers, and transmembrane modelled proteins, including with ligands and cofactors.
- SARS-CoV-2 3D includes predictions of allosteric and small ligand binding sites prediction.
- SARS-CoV-2 3D includes protein-protein interactions between SARS-CoV-2 and human proteins.
- SARS-CoV-2 3D incorporates protein-ligand docking and saturation mutagenesis.

Acknowledgements

This work is supported and funded by King Abdullah scholarship (Saudi Arabia research council), and American Leprosy Missions grants (G88726), SET is funded by the Cystic Fibrosis Trust (RG 70975) and Fondation Botnar (RG91317). A.R.J is funded by the Biotechnology and Biological Sciences Research Council (BBSRC) DTP studentship (BB/M011194/1). B.B. is funded by the Cystic Fibrosis Trust and L.C. on a studentship from Ipsen. T.L.B. is funded by a the Wellcome Trust Investigator Award, PHZJ/489 RG83114 (2016-2021)

Competing interests

The authors declare no competing interest.

Additional information

Supplementary information is available for this paper.

References

1. World Health Organization. Weekly Operational Update on COVID-19 september 27, 2020. World Heal Organ. 2020; October:1–10.

2. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020;395:565–74.

3. Zhou P, Yang X Lou, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579:270–3. doi:10.1038/s41586-020-2012-7.

4. Sironi M, Hasnain SE, Rosenthal B, Phan T, Luciani F, Shaw MA, et al. SARS-CoV-2 and COVID-19: A genetic, epidemiological, and evolutionary perspective. Infect Genet Evol. 2020;84 May.

5. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, et al. Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients. Clin Infect Dis. 2020;2019 Xx Xxxx:3–10.

6. Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, et al. Cell entry mechanisms of SARS-CoV-2. Proc Natl Acad Sci U S A. 2020;117.

7. Romano M, Ruggiero A, Squeglia F, Maga G, Berisio R. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping. Cells. 2020;9.

8. Sawicki SG, Sawicki DL, Siddell SG. A Contemporary View of Coronavirus Transcription. J Virol. 2007;81:20–9.

9. Liu DX, Fung TS, Chong KKL, Shukla A, Hilgenfeld R. Accessory proteins of SARS-CoV and other coronaviruses. Antiviral Res. 2014;109:97–109.

10. Narayanan K, Huang C, Lokugamage K, Kamitani W, Ikegami T, Tseng C-TK, et al. Severe Acute Respiratory Syndrome Coronavirus nsp1 Suppresses Host Gene Expression, Including That of Type I Interferon, in Infected Cells. J Virol. 2008;82:4471–9.

11. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. 2020;583 March.

12. Cencic R, Desforges M, Hall DR, Kozakov D, Du Y, Min J, et al. Blocking eIF4E-eIF4G Interaction as a Strategy To Impair Coronavirus Replication. J Virol. 2011;85:6381–9.

13. Du L, He Y, Zhou Y, Liu S, Zheng BJ, Jiang S. The spike protein of SARS-CoV - A target for vaccine and therapeutic development. Nat Rev Microbiol. 2009;7:226–36.

14. Gil C, Ginex T, Maestro I, Nozal V, Barrado-Gil L, Cuesta-Geijo MÁ, et al. COVID-19: Drug Targets and Potential Treatments. J Med Chem. 2020.

15. Shen LW, Mao HJ, Wu YL, Tanaka Y, Zhang W. TMPRSS2: A potential target for treatment of influenza virus and coronavirus infections. Biochimie. 2017;142:1–10. doi:10.1016/j.biochi.2017.07.016.

16. Ivanova T, Hardes K, Kallis S, Dahms SO, Than ME, Künzel S, et al. Optimization of Substrate-Analogue Furin Inhibitors. ChemMedChem. 2017;12:1953–68.

17. Zhou Y, Vedantham P, Lu K, Agudelo J, Carrion R, Nunneley JW, et al. Protease inhibitors targeting coronavirus and filovirus entry. Antiviral Res. 2015;116:76–84.

18. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020;181:271-280.e8.

19. Liu W, Liu L, Kou G, Zheng Y, Ding Y, Ni W, et al. Evaluation of nucleocapsid and spike protein-based enzyme-linked immunosorbent assays for detecting antibodies against SARS-CoV-2. J Clin Microbiol. 2020;58:1–7.

20. Thiel V, Ivanov KA, Putics Á, Hertzig T, Schelle B, Bayer S, et al. Mechanisms and enzymes involved in SARS coronavirus genome expression. J Gen Virol. 2003;84:2305–15.

21. Hilgenfeld R. From SARS to MERS: crystallographic studies on coronaviral proteases enable antiviral drug design. FEBS J. 2014;281:4085–96.

22. Subissi L, Imbert I, Ferron F, Collet A, Coutard B, Decroly E, et al. SARS-CoV ORF1b-encoded nonstructural proteins 12-16: Replicative enzymes as antiviral targets. Antiviral Res. 2014;101:122–30. doi:10.1016/j.antiviral.2013.11.006.

23. Kupferschmidt K, Cohen J. Race to find COVID-19 treatments accelerates. Science (80-). 2020;367:1412–3.

24. Yin, W., Luan, X., Li, Z., Zhang, L., Zhou, Z., Gao, M., Wang, X., Zhou, F., Wang, Q., Wang, Q., Jiang, Y., Jiang, H., Xiao, G., Yu, X., Zhang, S., Xu H. Structure of COVID-19 RNA-dependent RNA polymerase bound to suramin. 2020.

25. Dey SK, Saini M DC. Penciclovir and Anidulafungin bind nsp12, which governs the RNA-dependent-RNA polymerase activity of SARS-CoV-2, with higher interaction energy than Remdesivir, indicating potential in the treatment of Covid-19. 2020.

26. Ivanov KA, Ziebuhr J. Human Coronavirus 229E Nonstructural Protein 13: Characterization of Duplex-Unwinding, Nucleoside Triphosphatase, and RNA 5′-Triphosphatase Activities. J Virol. 2004;78:7833–8.

27. M.-H. L, D.C. M, C.-H. H, S.-C. C, Y.-H. C, C.-Y. S, et al. Disulfiram can inhibit MERS and SARS coronavirus papain-like proteases via different modes. Antiviral Res. 2018;150 November:155–63. http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L620021275%0Ahttp://dx.doi.org/10.1016/j.antiviral.2017.12.015.

28. Kopecky-Bromberg SA, Martínez-Sobrido L, Frieman M, Baric RA, Palese P. Severe Acute Respiratory Syndrome Coronavirus Open Reading Frame (ORF) 3b, ORF 6, and Nucleocapsid Proteins Function as Interferon Antagonists. J Virol. 2007;81:548–57.

29. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. The protein data bank. Acta Crystallogr Sect D Biol Crystallogr. 2002;58 6 I:899–907.

30. Sillitoe I, Andreeva A, Blundell TL, Buchan DWA, Finn RD, Gough J, et al. Genome3D: Integrating a collaborative data pipeline to expand the depth and breadth of consensus protein structure annotation. Nucleic Acids Res. 2020;48:D314–9.

31. Ochoa-Montaño B, Mohan N, Blundell TL. Chopin: A web resource for the structural and functional proteome of Mycobacterium tuberculosis. Database. 2015;2015:1–10.

32. Skwark MJ, Torres PHM, Copoiu L, Bannerman B, Floto RA, Blundell TL. Mabellini: A genome-wide database for understanding the structural proteome and evaluating prospective antimicrobial targets of the emerging pathogen Mycobacterium abscessus. Database. 2019;2019:1–16.

33. Sali A, Blundell T. Sali, A. & Blundell, T. L. Comparative modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815. Journal of molecular biology. 1994;234:779–815.

34. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.

35. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol. 2001;310:243–57.

36. Fidler DR, Murphy SE, Courtis K, Antonoudiou P, El-Tohamy R, Ient J, et al. Using HHsearch to tackle proteins of unknown function: A pilot study with PH domains. Traffic. 2016;17:1214–26.

37. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7.

38. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI. OPM: Orientations of proteins in membranes database. Bioinformatics. 2006;22:623–5.

39. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: All-atom structure validation for macromolecular crystallography. Acta Crystallogr Sect D Biol Crystallogr. 2010;66:12–21.

40. Krissinel E, Henrick K. Inference of Macromolecular Assemblies from Crystalline State. J Mol Biol. 2007;372:774–97.

41. Kleffner R, Flatten J, Leaver-Fay A, Baker D, Siegel JB, Khatib F, et al. Foldit Standalone: a video game-derived protein structure manipulation interface using Rosetta. Bioinformatics. 2017;33:2765–7.

42. Krissinel E. Stock-based detection of protein oligomeric states in jsPISA. Nucleic Acids Res. 2015;43:W314–9.

43. Pires DEV, Ascher DB, Blundell TL. MCSM: Predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 2014;30:335–42.

44. Cao H, Wang J, He L, Qi Y, Zhang JZ. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. J Chem Inf Model. 2019;59:1508–14.

45. Choi Y, Chan AP. PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31:2745–7.

46. Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P. MAESTRO - multi agent stability prediction upon point mutations. BMC Bioinformatics. 2015;16:1–13.

47. Capriotti E, Fariselli P, Calabrese R, Casadio R. Predicting protein stability changes from sequences using support vector machines. Bioinformatics. 2005;21 SUPPL. 2:54–8.

48. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005;33 SUPPL. 2:306–10.

49. Bateman A. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47:D506–15.

50. Mih N, Brunk E, Chen K, Catoiu E, Sastry A, Kavvas E, et al. Ssbio: A Python framework for structural systems biology. Bioinformatics. 2018;34:2155–7.

51. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47:W636–41.

52. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, et al. The ClusPro web server for protein-protein docking. Nat Protoc. 2017;12:255–78.

53. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B. 1998;102:3586–616.

54. Douguet D. Data Sets Representative of the Structures and Experimental Properties of FDA-Approved Drugs. ACS Med Chem Lett. 2018;9:204–9.

55. Madhavi Sastry G, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. J Comput Aided Mol Des. 2013;27:221–34.

56. Brooks WH, Daniel KG, Sung SS, Guida WC. Computational validation of the importance of absolute stereochemistry in virtual screening. J Chem Inf Model. 2008;48:639–45.

57. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. J Med Chem. 2004;47:1739–49.

58. Ban T, Ohue M, Akiyama Y. Multiple grid arrangement improves ligand docking with unknown binding sites: Application to the inverse docking problem. Comput Biol Chem. 2018;73:139–46. doi:10.1016/j.compbiolchem.2018.02.008.

59. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov. 2015;10:449–61.

60. Ma Y, Wu L, Shaw N, Gao Y, Wang J, Sun Y, et al. Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. Proc Natl Acad Sci U S A. 2015;112:9436–41.

61. Surya W, Li Y, Torres J. Structural model of the SARS coronavirus E channel in LMPG micelles. Biochim Biophys Acta - Biomembr. 2018;1860:1309–17.

62. J Alsaadi EA, Jones IM. Membrane binding proteins of coronaviruses. Future Virol. 2019;14:275–86.

63. Wathelet MG, Orr M, Frieman MB, Baric RS. Severe Acute Respiratory Syndrome

Coronavirus Evades Antiviral Signaling: Role of nsp1 and Rational Design of an Attenuated Strain. J Virol. 2007;81:11620–33.

64. Thoms M, Buschauer R, Ameismeier M, Koepke L, Denk T, Hirschenberger M, et al. Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2. Science (80-). 2020;8665:eabc8665.

65. de Chassey B, Meyniel-Schicklin L, Vonderscher J, André P, Lotteau V. Virus-host interactomics: New insights and opportunities for antiviral drug discovery. Genome Med. 2014;6:1–14.

66. Lokhande KB, Doiphode S, Vyas R, Swamy KV. Molecular docking and simulation studies on SARS-CoV-2 Mpro reveals Mitoxantrone, Leucovorin, Birinapant, and Dynasore as potent drugs against COVID-19. J Biomol Struct Dyn. 2020;0:1–12. doi:10.1080/07391102.2020.1805019.

67. Mittal L, Kumari A, Srivastava M, Singh M, Asthana S. Identification of potential molecules against COVID-19 main protease through structure-guided virtual screening approach. J Biomol Struct Dyn. 2020;0:1–19. doi:10.1080/07391102.2020.1768151.

68. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. 2020;181:281-292.e6. doi:10.1016/j.cell.2020.02.058.

69. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science (80-). 2020;368:779–82.

70. Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. Nature. 2020;582:289–93. doi:10.1038/s41586-020-2223-y.

71. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. 2003;31:3381–5.

72. Roy A, Kucukural A, Zhang Y. I-TASSER: A unified platform for automated protein structure and function prediction. Nat Protoc. 2010;5:725–38.

73. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577:706–10. doi:10.1038/s41586-019-1923-7.

74. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Res. 2019;47:D94–9.

75. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020;48:D682–8.

76. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance. 2017;22:2–4.

77. Sillitoe I, Dawson N, Lewis TE, Das S, Lees JG, Ashford P, et al. CATH: Expanding the horizons of structure-based functional annotations for genome sequences. Nucleic Acids Res. 2019;47:D280–4.

78. Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: Expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res. 2020;48:D376–82.

79. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: The Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47:D941–7.

80. Sedova M, Jaroszewski L, Alisoltani A, Godzik A. Coronavirus3D: 3D structural visualization of COVID-19 genomic divergence. Bioinformatics. 2020; May:1–3.

Figure 1 Statistical analysis of modelled proteome. (A) The total percentage sequence coverage of experimentally solved structures deposited in RSCB (Research Collaboratory for Structural Bioinformatics) is represented in cyan, whereas the total percentage coverage of each modelled

structure is shown above it in red. (B) MolProbity scores for all modelled SARS-CoV-2 structures deposited in the SARS-CoV-2 3D database.

Figure 2. Four modelled oligomeric targets selected from the SARS-CoV-2 proteome. (A) Nsp14 (white-grey) and Nsp10 (green). The zinc ions are shown as silver spheres, and the magnesium ion as a green sphere. The SAH ligand is represented in magenta pink, and the G3A in green. Binding interactions of both ligands in the Nsp14 binding sites are represented as dashed lines highlighted in black. (B) Homo-pentameric model of Envelope protein E with each protomer indicated in a different colour. The membrane is represented as a red/ blue circular structure. (C) Homodimeric model of the membrane protein M, with protomers coloured in green and white. The membrane is represented as a red / blue circular structure. (D) The structure of the Nsp1 model, coloured dark blue and encircled, complexed with the 40S ribosome (a hetero 35-mer) with the 35 proteins coloured differently.

Figure 3 Website interface front page and result page: (A) Home page: the jumbotron at the top left represents the main ideas of the database. The query options such as a table or sunburst are represented below. (B) Results page. The top represents a brief description of the target genes with external links to other modelling pipelines: MolStar viewer for protein visualizations, Fpocket table, and UniProt viewer for sequence annotation. There are five tables represented in the result pages: models, PDB structures, mutations, protein-protein interactions, and virtual screening of ligands. The SARS-CoV-2 protein interaction with the human protein is annotated at the bottom of the page; the black arrows indicate human proteins annotated as drug targets.