

Virtual screening on the web for drug repurposing: a primer

Yu Wai Chen^{1*}, Chin-Pang Benu Yiu², Kwok-Yin Wong¹

¹Department of Applied Biology and Chemical Technology and State Key Laboratory of Chemical Biology and Drug Discovery, The Hong Kong Polytechnic University, Hunghom, Hong Kong, China

²Independent Researcher, La Costa, Ma On Shan, Hong Kong, China

*Corresponding authors: Yu Wai Chen, Email: yu-wai.chen@polyu.edu.hk

Competing interests: The authors have declared that no competing interests exist.

Abbreviations used: ADT, AutoDockTools; CoV, coronavirus; M^{pro}, main protease; PDB, Protein Data Bank; PDB ID, Protein Data Bank accession number; PL^{pro}, papain-like protease; RdRp, RNA-dependent RNA polymerase; SARS, severe acute respiratory syndrome; VS: virtual screening; 3CL^{pro}: 3-chymotrypsin-like protease

Received November 15, 2020; Revision received January 20, 2021; Accepted January 25, 2021; Published May 26, 2021

ABSTRACT

We describe a procedure of performing *in silico* (virtual) screening using a web-based service, the MTiOpenScreen, which is freely accessible to non-commercial users. We shall use the SARS-CoV-2 main protease as an example. Starting from a structure downloaded from the Protein Data Bank, we discuss how to prepare the coordinates file, taking into account the known biochemical background information of the target protein. The reader will find that this preparation step takes up most of the effort before the target is ready for screening. The steps for uploading the target structure and defining the search volume by critical residues, and the main parameters to use, are outlined. When this protocol is followed, the user will expect to obtain a ranked list of small approved drug compounds docked into the target structure. The results can be readily examined graphically on the web site or downloaded for studying in a local molecular graphics program such as PyMOL.

Keywords: AutoDock Vina, main protease, SARS-CoV-2, virtual screening, approved drug

BACKGROUND

At the time of writing (late 2020 to early 2021), mankind is facing one of the worst global health threats. Beginning one year ago, in the winter of 2019, the coronaviral disease known as COVID-19 has spread to more than 200 countries and infected 100-million people with fatality exceeding two millions. In various countries, the number of cases rises and falls like waves. At the moment, no one can predict when or whether it will ever end.

Scientific research has been adversely affected because of laboratory closures and research resources manufacturing and shipping problems. Naturally, some researchers who are confined to work from home turn to *in silico* methods. There has been an unprecedented surge of work deposited into preprint servers, yet the quality of these tide-riding articles varied greatly and many did not advance into peer-reviewed journals.

The purpose of this article is to outline a protocol of performing virtual screening (VS) using web services and tools which are free for non-profit users. We shall illustrate the procedures using a relevant protein as an example of drug development. The goal is to describe the steps to identify potential compounds from approved drug libraries which

can be repurposed as inhibitors targeting this protein. This protocol is based on an earlier article that we published [1]. The reader may refer to this paper for more details of the results and their interpretations.

The genome of the causative agent, the SARS-CoV-2 coronavirus, encodes 29 proteins. Among these, the viral enzymes are considered to be the most promising drug targets because there is less opportunity for the virus to develop resistance once inhibitors are discovered. The scope to escape inhibition by mutation is limited by the pressure to preserve catalytic functions. The viral enzymes include the papain-like protease (PL^{pro}), main protease (M^{pro}; aka 3CL^{pro}), RNA-dependent RNA polymerase (RdRp) and helicase. There are other targets, for example, the viral spike protein which is crucial for the initial contact of the host cells. But some experimental information, *e.g.*, which parts of it bind to the host receptors, is needed before it can be exploited for drug screening. The M^{pro} is an essential enzyme for the virus. Having a 96% identity in protein sequence, its SARS-CoV orthologue was extensively studied for drug development during the SARS outbreak in 2003. The M^{pro} of SARS-CoV-2 was, therefore, selected as a target for drug repurposing, hoping to get rapid results.

How to cite this article: Chen YW, Yiu CB, Wong K. Virtual screening on the web for drug repurposing: a primer. *J Biol Methods* 2021;8(2) COVID 19 Special Issue:e148. DOI: 10.14440/jbm.2021.351

MATERIALS

Hardware

This basic hardware required is just a regular modern-day desktop computer running Linux of any favours (e.g., Ubuntu 18.04 LTS) and equipped with a good graphics card. Most of the software are ported to Windows and MacOS/OSX platforms too but these ports are less preferred to the Linux platform, for community support, system management and maintenance.

Web service

MTiOpenScreen [2] is the main service where VS is carried out. It is hosted by the Ressource Parisienne en BioInformatique Structurale (RPBS), Université de Paris. (<https://bioserv.rpbs.univ-paris-diderot.fr/services/MTiOpenScreen>)

We were only interested in clinically approved compounds on the market, for this reason, we restricted our search only on purchasable drugs. The “Drugs-lib” library [3] implemented on the MTiOpenScreen server suits our purpose perfectly. It contains 7173 purchasable drugs with 4574 unique compounds and their stereoisomers. Each library entry in the results is identified with the name of the compound as well as a ZINC15 ID.

Software

The following software should be installed on the local Linux machine.

PyMOL [4]: after the docking, one needs to view the results on molecular graphics. The results are in the pdbqt format which can be readily opened with PyMOL (https://pymolwiki.org/index.php/Linux_Install).

Coot [5]: we need a molecular graphics software which has the capability of editing and manipulation of atomic coordinates in the pdb format. Coot is a protein crystallography utility which serves these functions well (<https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot>).

AutoDockTools (ADT) and AutoDock Vina [6]: both of these are not strictly required if one confines oneself to do a web-based screening. However, if one intends to follow up the results and perform further docking with refined parameters on a local computer, one needs ADT for preparation of the structure files, and AutoDock Vina for docking (<http://autodock.scripps.edu/resources/adt>; <http://vina.scripps.edu>). Smina, a fork of AutoDock Vina, is highly recommended for its extended capabilities including additional scoring functions (<http://smina.sf.net>).

CCP4 suite [7]: this is a complete package for protein crystallography. It contains several useful utilities for manipulation of atomic-coordinate files (<https://www.ccp4.ac.uk>).

TIPS/HINTS: The installation of software on a Linux system may deter some beginners. The only indispensable software are PyMOL and a text editor. The rest is needed if one wants to perform follow-up docking or advanced coordinates manipulation on a local machine.

PROCEDURE

Please refer to **Figure 1** for an overall workflow.

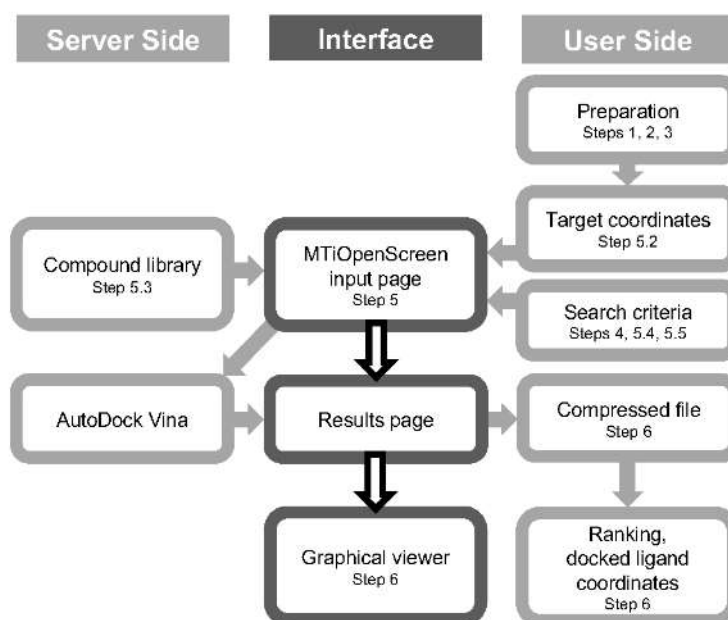


Figure 1. The workflow of the virtual screening procedure using MTiOpenScreen. The middle dark grey boxes are the webpage interfaces where the user interacts with the server. The server side hosts several compound libraries, as well as the screening program AutoDock Vina. On the user side, various activities from supplying input to retrieving output are indicated. The step(s) of this protocol corresponding to each box are indicated.

Drug target structure preparation

1. Oligomerization state

This step deserves a lot of attention. A crystal structure is the usual starting point. However, a crystal structure downloaded from the Protein Data Bank (PDB) [8] may need substantial editing before it is submitted for VS. At this stage, one must exhaust one's efforts in understanding the basic biology and chemistry of the drug target. In particular, attention must be paid to the physiological oligomerization state, the active and inactive conformations, any missing or disordered residues, the essential residues relating to the drug target's functions (and the catalytic mechanism, if applicable) and the nature of the substrates.

One usually chooses a structure at the highest resolution if there are several available. There are debates on whether the structure of the apo-enzyme (with no ligand bound) or that of the complexed enzyme should be used. If both are available, both should be tried.

Before the crystal structure of SARS-CoV-2 M^{Pro} was solved, we looked for a crystal structure in the Protein Data Bank for modelling. Previously, work on the near-identical SARS-CoV orthologue established that its functional form is dimeric [9]. If homology modelling is used, the user needs to ensure that they use a dual-functional dimer as the template. Some dimeric crystal structures of SARS-CoV M^{Pro} are known to have one inactive protomer with its active site collapsed.

It was not long before many crystal structures, both apo- and complexed with ligands, of SARS-CoV-2 M^{Pro} were reported and released in the PDB. The highest-resolution crystal structure of the apo-enzyme is PDB ID 6YB7, which is to 1.25Å, and its asymmetric unit contains one molecule of the active state. We used pdbset of the CCP4 suite, with the help of Coot, to generate the crystallographic symmetry-related dimer (consists of "A" and "B" chains).

CAUTION: If the monomer is used for screening, the ligands may be docked to residues that are otherwise buried in the interface and false positives may be identified.

NOTE: If a crystal structure of the target is not available, one may use homology modelling to build a starting structure. There are web services for this purpose, for example, the Swiss-Model (<https://swissmodel.expasy.org/interactive>). Modeller is another popular tool, but it needs to be installed on a local machine (<https://salilab.org/modeller/>). In our earlier work, we used a homology model of SARS-CoV-2 M^{Pro}, built with Modeller, for virtual screening because the crystal structure was not yet available at that time [1].

2. Protonation states

AutoDock Vina does not take atomic charges into account for docking. It does use polar hydrogens. It is necessary to check the protonation of side chains at the active site residues, especially histidines. This information may be available in the literature if homologous structures have been published. The protonation states can sometimes be inferred from the nature of the neighbouring atoms (electron donating or accepting atoms) and interatomic bond distances and angles in the crystal structure. In this case, note that H41 is N δ -protonated, H163 is N ϵ -protonated, H172 is protonated on both N δ and N ϵ . These hydrogen atoms can be defined in ADT (add hydrogens).

3. Alternative conformations

At 1.25 Å resolution, the protein structure contains a substantial number of residues which show alternative side-chain conformations. The user needs to choose one in each residue. For this purpose, the user may refer to lower resolution crystal structures to select a major conformer for each of these residues. This is done by editing the PDB file using a text editor (*e.g.*, *nedit* in Linux). For example, in 6YB7 residue valine 104 (V104), there are two alternative conformations. The user needs to choose either all the "A" set atoms or all those of the "B" set. The other set of atoms need to be deleted, and the "A" or "B" in front of the residue name "VAL A 104" of the remaining atoms replaced by a space character.

NOTE: In the PDB file, the alternative conformation indicator is the character ("A", "B", "C"...) immediately in front of the residue name. Do not mix it up with the character after the residue name, which is the chain identifier (representing "A" chain, "B" chain, *etc.*). *E.g.*, The A and B conformers of V104 of "A" chain are noted as "... AVAL A 104" and "... BVAL A 104", respectively.

4. Important residues

The active-site residues as well as those responsible for substrate binding need to be identified. Again, these knowledge are available from the literature of this protein and its orthologues. Keep a list of these residues which will be used to define the search box in VS in MTiOpenScreen (Step 5.4).

TIPS/HINTS: In the follow-up work on the local machine, AutoDock Vina can be instructed to treat a subset of residues to have flexible side chains for optimal docking. The active site residues should be used for this purpose (see Discussion).

Virtual screening on the webserver

The workflow of the MTiOpenScreen server is described on its landing page (<https://bioserv.rpbs.univ-paris-diderot.fr/services/MTiOpenScreen/>).

5. Submitting VS job

At the MTiOpenScreen homepage, click the “Run MTiOpenScreen” button.

NOTE: One can use the services without registering, as “guest”, but running with a registered account will have the job logging and results better organised. Note that the service required is not MTiAutoDock.

- 5.1. In the “Demonstration Mode” area, choose “No”.
- 5.2. In the “Protein receptor” area; under “Receptor file”, choose the “upload” tab. Then click the “Choose file” button; then browse in the local computer and choose the pdb-format structure file as prepared in the above procedures. If the file is read successfully, its contents will appear in the text window of the “upload” tab of the webserver.
- 5.3. In the “Compounds database” area; “Select library...” and choose “Drugs-lib” from the dropdown menu. In the “Compound library filters” area, leave all fields empty to use default values.

CRITICAL STEP: This step is concerning how to define the search box. If one wants to define this by known active-site and substrate-binding residues, do Step 5.4. Otherwise, if one wants to define with a search box, then do Step 5.5:

- 5.4. In the “Grid calculation” area, choose “List of residues” from the dropdown menu corresponding to “Mode”. Skip the “Calculate grid from coordinates (custom mode)” areas. Then in the “Calculate grid from a list of residues (list of residues mode)” area, fill in the list of residues into the input box. Here is the list we used for screening SARS-CoV-2 M^{pro} [1], in the format MTiOpenScreen specified:
`_A_HIS_41_,_A_MET_49_,_A_GLY_143_,_A_SER_144_,_A_CYS_145_,_A_HIS_163_,_A_HIS_164_,_A_MET_165_,_A_GLU_166_,_A_LEU_167_,_A_ASP_187_,_A_ARG_188_,_A_GLN_189_,_A_THR_190_,_A_ALA_191_,_A_GLN_192_`

HINT: In this specific case, only the “A” chain active site needs to be screened as that of the “B” chain is identical by crystallographic symmetry. If the two protomers are not identical (*e.g.*, from a structure which has “A” and “B” chains in the asymmetric unit), then each set of residues corresponding to the “A” or “B” chain needs to be screened in separate runs. *I.e.*, first run, for the “A” chain active site (as above); second run, for the “B” chain active site (use “_B_HIS_41_,_B_MET_49_, ...”). Note that the trailing underscores in the residue format is two characters long. If after job submission, the “job progress report” window at the top shows the text “1/5 input processing” only and never proceed to stage 2/5 and beyond, and printed a line in red “Your job finished with an unusual status code (1), check your results carefully”, then it is likely that there were errors in the input, either in the PDB file or the parameters (*e.g.*, the residue format if one uses a list to specify the search volume).

- 5.5. Alternatively, if one prefers to define the search box by the AutoDock Vina convention, one can use a coordinate as the centre and specify the three box dimensions. In the “Grid calculation” area, Choose “Custom parameters” from the dropdown menu corresponding to “Mode”. In the “Calculate grid from coordinates (custom mode)” area, fill in the X, Y, Z coordinates (in Å) in the “Grid center coordinates” area; and in the “Sizes of the search space” area, fill in the X, Y, Z box lengths (in Å).
- 5.6. Press the Run button.

Downloading docking results

6. After the run is completed, the results are presented on the webpage. Press the blue button on the top of the MTiOpenScreen job page to download a single compressed (“.zip”) file containing all results.

ANTICIPATED RESULTS

This demonstration job took approximately an hour to complete but the actual run time will depend on the actual service loading at the host side.

On the web page of the job, the user will see a list of the best binders, ranked by Vina binding energies (in kcal/mol). The web service tactfully offers a graphics viewing tool which facilitates the immediate examination of the top 100 scorers (Fig. 2), with limited graphics capabilities. The complete docking job output can be downloaded as a compressed (zipped) file for local analysis. Upon decompression (unzipped), several files will be available. The file “output.vis.html” is supposed to be openable by a web browser and contains a utility to display the results graphically. But this does not work under our hands and only an empty graphics window was shown with some control radio buttons and a sliding bar, irrespective of browsers: Chrome, Brave Internet Explorer or Edge, on a Windows desktop computer.

Nevertheless, all 1500 hits were included in the file “output.table.csv”. In the file “output.pdbqt”, A total of 4500 models comprising 3 docking modes (with different binding energies, with the top one used for ranking) for each of the 1500 top-binding compounds.

To study the results on the local machine in molecular graphics, the user needs to open the target molecule’s coordinates (the PDB file which was uploaded for screening) in PyMOL first, and then open the “output.pdbqt” file of the docked results of the ligands. Conveniently, the ZINC reference of each docked compound was also listed so the user can refer back to the ZINC15 web site (<http://zinc.docking.org/>) to find out about the chemical properties of each compound of interest.

TIPS/HINTS: One can edit the “output.pdbqt” file using a text editor and extract the top scorers (*e.g.*, the top 20) and save them into a new file for examination in PyMOL. Each solution was enclosed in a pair of “MODEL *n*” (*n* is the rank, an integer) and “ENDMDL” lines. In PyMOL, each solution can be viewed by using the “video-player-like” control buttons on the bottom left, advancing one at a time.

NOTE: A “.csv” file is a text file openable by Microsoft Excel, it can be opened with any text editor too. A “.pdbqt” file is a structure coordinates file (modified “.pdb” format) which can be opened with PyMOL or ADT.



Figure 2. The graphical visualizer on the MTiOpenScreen web site. The user can quickly examine the results of the 100 best-docked compounds here. A simple click on the bottom list, which includes the ZINC identifier and the Vina binding energy, will switch to the next ligand. The button corresponding to the active ligand shown is highlighted in green.

DISCUSSION

The purpose of this article is to show how easy it is to perform VS nowadays. Before web services like MTiOpenScreen are available, the learning curve of molecular docking in silico was steep. Some of the

major hurdles lie in the setting up of the workstation and installation of software. With a web service, these issues are avoided. The trade-off is that, understandably, some features of AutoDock Vina are not offered (*e.g.*, flexible side chains on the target). Some parameters which are used to fine-tune the program are left out. MTiOpenScreen offers sev-

eral useful compound libraries, which are adequate for a first attempt of drug repurposing. In addition to Drugs-lib, two other libraries may be useful: the “FOOD-lib”, with 10997 food-constituent compounds, and the “NP-lib”, with 1228 natural-product compounds [3]. All compounds in these three libraries are purchasable. MTiOpenScreen allows the user to upload a customized library too. However, it is advisable to perform this on a local setup of AutoDock Vina, or better still, Smina. The preparation of the compound structures and library is beyond the scope of this tutorial. The interested readers can refer to this excellent article discussing the extended use of AutoDock Vina [10]. Finally, the server itself needs to be improved on stability and reliability. The authors have encountered some occasions when it was out of service.

We have to issue a word of warning: since it is so convenient to generate numerical results, it can be easily abused. We cannot stress more how important it is to interpret the results carefully and to follow up with further analyses and *in vitro* studies. AutoDock Vina presents its results in “Vina binding energies” which is an analogous value for scoring. It is not a bona fide free energy of binding and should not be used for quantitative comparison between different systems. After the initial top-scoring compounds were obtained from MTiOpenScreen, it is advisable to repeat the results on a local computer, using AutoDock Vina with different scoring functions and checking for consistency. There are at least two other scoring models which should be tried: Vinardo [11] and Smina. This article by the group who hosts the MTiOpenScreen and Drugs-lib has an excellent discussion of the pitfalls of VS and some attempts to overcome or minimize the various intrinsic biases [12]. In addition, one should define flexible side chains for the active-site residues and those which are directly involved in substrate binding for optimal docking (Step 4).

If there are ligands known to bind well with the target, then one should validate the VS procedure and parameters using the “custom compound library” option. Alternatively, a local docking set up of AutoDock Vina can be used. In the case we are using as an example here, the target is 96% identical in amino-acid sequence with its SARS coronaviral (SARS-CoV) M^{pro} orthologue with identical active sites [1]. There were reports of drug leads against this protein, with crystal structures solved (*e.g.*, [13]). These inhibitors may serve as good candidates for validation. Further, the first crystal structure of the SARS-CoV-2 M^{pro} is solved with an inhibitor, N3 [14], which can be used to validate the VS protocol.

The simple method outlined here applies to a single protein. But drug targets are not restricted to viral proteins. The host receptor (*e.g.*, the angiotensin-converting enzyme-2, ACE2) and other host proteins (*e.g.*, the transmembrane serine protease 2, TMPRSS2; [11]) could be targeted too. For instance, these researchers mapped the virus-host interactions and identified 66 host proteins which are potential drug targets [15].

Acknowledgments

We acknowledge support from the Innovation and Technology

Commission of Hong Kong, the Hong Kong Polytechnic University and the Life Science Area of Strategic Fund 1-ZVH9.

References

- Chen YW, Yiu CB, Wong K. (2020) Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL pro) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Res* 9: 129. doi: 10.12688/f1000research.22457.2. PMID: 32194944
- Labbé CM, Rey J, Lagorce D, Vavruša M, Becot J, et al. (2015) MTiOpenScreen: a web server for structure-based virtual screening. *Nucleic Acids Res* 43: W448-W454. doi: 10.1093/nar/gkv306. PMID: 25855812
- Lagarde N, Rey J, Gyulkhandanyan A, Tufféry P, Miteva MA, et al. (2018) Online structure-based screening of purchasable approved drugs and natural compounds: retrospective examples of drug repositioning on cancer targets. *Oncotarget* 9: 32346-32361. doi: 10.18632/oncotarget.25966. PMID: 30190791
- The PyMOL molecular graphics system. Schrödinger LLC. Version 1.7 or later. Available from: https://pymolwiki.org/index.php/Linux_Install
- Emsley P, Lohkamp B, Scott WG, Cowtan K. (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66: 486-501. doi: 10.1107/S0907444910007493. PMID: 20383002
- Trott O, Olson AJ. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 31: 455-461. doi: 10.1002/jcc.21334. PMID: 19499576
- Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, et al. (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* 67: 235-242. doi: 10.1107/S0907444910045749. PMID: 21460441
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28: 235-242. doi: 10.1093/nar/28.1.235. PMID: 10592235
- Muramatsu T, Takemoto C, Kim Y, Wang H, Nishii W, et al. (2016) SARS-CoV 3CL protease cleaves its C-terminal autoproteolytic site by novel subsite cooperativity. *Proc Natl Acad Sci U S A* 113: 12997-13002. doi: 10.1073/pnas.1601327113. PMID: 27799534
- Jaghooori MM, Bleijlevens B, Olabarriaga SD. (2016) 1001 Ways to run AutoDock Vina for virtual screening. *J Comput Aided Mol Des* 30: 237-249. doi: 10.1007/s10822-016-9900-9. PMID: 26897747
- Quiroga R, Villarreal MA. (2016) Vinardo: A Scoring Function Based on AutoDock Vina Improves Scoring, Docking, and Virtual Screening. *PLoS One* 11: doi: 10.1371/journal.pone.0155183. PMID: 27171006
- Singh N, Decroly E, Khatib A, Villoutreix BO. (2020) Structure-based drug repositioning over the human TMPRSS2 protease domain: search for chemical probes able to repress SARS-CoV-2 Spike protein cleavages. *Eur J Pharm Sci* 153: 105495. doi: 10.1016/j.ejps.2020.105495. PMID: 32730844
- Yang H, Yang M, Ding Y, Liu Y, Lou Z, et al. (2003) The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proc Natl Acad Sci USA* 100: 13190-13195. doi: 10.1073/pnas.1835675100. PMID: 14585926
- Jin Z, Du X, Xu Y, Deng Y, Liu M, et al. (2020) Structure of M^{pro} from SARS-CoV-2 and discovery of its inhibitors. *Nature* 582: 289-293. doi: 10.1038/s41586-020-2223-y. PMID: 32272481
- Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, et al. (2020) A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583: 459-468. doi: 10.1038/s41586-020-2286-9. PMID: 32353859



This work is licensed under a Creative Commons Attribution-Non-Commercial-ShareAlike 4.0 International License: <http://creativecommons.org/licenses/by-nc-sa/4.0>