A Generative Deep Learning Approach for the Discovery of SARS CoV2 Protease Inhibitors

Noor Shaker,*[a] Mohamed Abou-Zleikha[a] Mubarak A. Alamri,[b] and Youcef Mehellou*[c]

[a] Dr. Noor Shaker and Dr. Mohamed Abou-Zleikha
Glamorous AI Ltd
29-31 Colville Road, London W11 2BT, U.K.
E-mail: noor@glamorous.ai
[b] Dr. Mubarak A. Alamri
Department of Pharmaceutical Chemistry
College of Pharmacy, Prince Sattam Bin Abdulaziz University
Alkarj, Saudi Arabia
[c] Dr. Youcef Mehellou
Cardiff School of Pharmacy and Pharmaceutical Sciences
Cardiff University
Redwood Building, Cardiff CF10 3NB, U.K.
E-mail: MehellouY1@cardiff.ac.uk

Supporting information for this article is given via a link at the end of the document.((Please delete this text if not appropriate))

Abstract: COVID19 has caused thousands of deaths worldwide within a few months. The rapid spread of this virus that causes COVID19, termed SARS CoV2, has been facilitated by the lack of effective vaccines and treatments against this virus. In recent months, our team has developed a novel deep learning platform, Rosalind, for drug design and optimisation, and it enables rapid in silico discovery and evaluation of novel chemical designs. In the current work, we applied Rosalind for the rapid discovery of SARS CoV2 replication inhibitors that target the virus main protease M^{pro}. Through a series of training and optimisation rounds based on reported SARS CoV2 M^{pro} inhibitors helped by docking into the recently reported crystal structures of SARS CoV2 M^{pro} and medicinal chemistry input, we identified the a series of promising SARS CoV2 M^{pro} inhibitors. These compounds are presented in this work so they scientific community could pursue them while we continue our deep learning-based work in a collaborative manner to identify lead SARS CoV2 M^{pro} compounds with excellent drug-like properties that could be developed in a timely manner to address the urgent need for new and effective COVID19 treatments.

Coronaviruses (CoVs) are distributed in mammals and birds and have been known since the mid-1960s.[1] Most of the pathogenic CoVs are known to cause nonfatal illnesses.[2] However, in the 21st century, there have been three CoVs outbreaks, which caused serious fatal respiratory infections in humans.[3] The first of these three was the severe acute respiratory syndrome coronavirus (SARS-CoV), which emerged in China in 2003 and spread to five continents affecting over 8,000 individuals with an overall fatality rate of 10%.[4] The second was the Middle East Respiratory Syndrome Coronavirus (MERS-CoV), which appeared firstly in Saudi Arabia in 2012 and spread to many places causing a global mortality rate of 35%.[4] The third CoV was first reported in December 2019 in the city of Wuhan in China, and it was noted as a novel coronavirus that causes severe pneumonia.[5] As the RNA genome of this coronavirus was found to be about 82% identical to that of SARS-CoV and both viruses belong to clade b of the genus Betacoronavirus,[5] the World Health Organization (WHO) named this novel CoV virus SARS-CoV-2 and its associated disease the 2019-coronavirus disease (commonly referred to as COVID19). Due to the rapid spread and deadly characteristics of SARS-CoV2, the WHO declared COVID19 as a pandemic.

To date, there is no effective vaccine or treatment for COVID19, and this has partly played a role in its rapid spread and high mortality rate. In order to address this, and given the urgency of discovering new and effective treatments for this infection, we sought to apply a deep learning approach in the discovery of new treatments for COVID19. The application of deep learning strategies in drug design has enjoyed an increasing amount of interest over the past few years but it was only until recently when this approach started bearing fruit in identifying small molecule therapeutics.[6] Notably, deep learning approaches are more fruitful when there are large series of hit compounds against the drug targets as this facilitates training and optimization of the deep learning model.

In order to choose the SARS CoV2 molecular target for our deep learning-driven discovery of anti-SARS CoV2 agents, we considered the validated protein targets of its closely-related and previously known coronavirus SARS CoV. Indeed, analysis of these targets and the available molecules that target them highlighted the protease enzymes that play key roles in the proteolytic processing of the virus as attractive molecular targets.[7] Interestingly, these proteases are similar in both SARS CoV and SARS CoV2, and studies have already shown that, akin to SARS CoV, the maturation of SARS CoV2 is mediated by two cysteine protease enzymes; 3-chymotrypsin-like protease ($3CL^{pro}$, commonly referred to as M^{pro}) and papain-like protease (PL^{pro}).[5, 8] Encouragingly, humans lack proteases with M^{pro} and PL^{pro} cleavage specificity, and this suggests that inhibitors of these enzymes would have limited toxicity. In addition, two crystal structures of the SARS CoV2 main protease M^{pro} have recently been reported along with some initial hit compounds,[8] Together, this made the SARS CoV2 protease M^{pro} an attractive target for drug discovery endeavours that are aimed at developing new and effective treatments for COVID19.

To rapidly discover SARS CoV2 M^{pro} protease inhibitors, we used our deep learning platform, Rosalind, which is being developed specifically to deal with small datasets; mainly cases where only a limited number of known inhibitors are known. In contrast to other approaches, Rosalind's capabilities are optimised to efficiently explore a diverse search space around a specific core structure. This is a key feature of Rosalind, which allows it to incorporate medicinal chemistry input while retaining the power of novel, diverse and rapid chemical design.

Our pursuit of discovering novel SARS CoV2 M^{pro} inhibitors started by examining the literature to generate a database of molecules that have been reported as promising SARS CoV2 M^{pro} inhibitors. In particular, we combined the compounds reported in

three different studies,[9] which after removing duplicates, gave a library of 535 compounds (Supporting Information, **Table S1**). Analysis of these structures led to two large clusters (small molecules heterocycles and peptidomimetics [examples given in **Figure 1A-B**]) as well as few nucleosides/nucleotides. Notably, this combined library of 535 compounds contained two molecules, namely remdesivir and niclosamide, which have already been shown to have potent anti-SARS CoV2 activities.[10] Out of the two major clusters of compounds, we decided to focus on the small heterocyclic molecules. Analysis of the compounds among the library of 535 compounds led us to select 42 compounds (Supporting Information, **Table S2**), and identify a key pharmacophore with possible variations (**Figure 1C**).

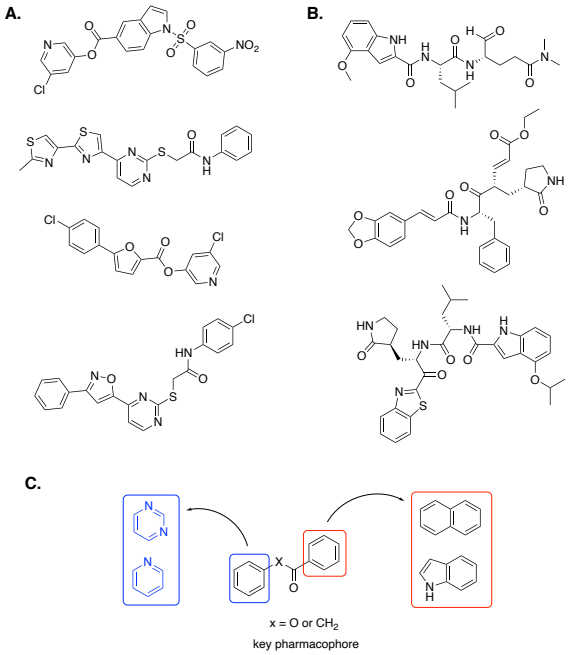

Figure 1. Examples of compounds representing the two major clusters of molecules within the 535 library of compounds gathered from the literature. **A.** Small heterocyclic molecules. **B.** Peptidomimetics. **C.** The key pharmacophore identified from the small heterocyclic compounds that were in the principal database.

The 42 compounds and the identified key pharmacophore were used in the training of the SARS CoV2 Rosalind model. At this early stage of employing the deep learning model, we applied a number of physicochemical properties filters including molecular weight < 500 g/mol and a LogP = 2-3 (see Experimental for full details). These were applied at this stage to ensure that from the onset of this work, we generate molecules that have key drug-like physicochemical properties. The generated compounds were then docked using Smina,[11] a fork of Autodock Vina into the crystal structure of SARS CoV2 M^{pro} protease (6YB7)[12] and that of ist closely related SARS CoV (3V3M)[13]. The compounds that gave an energy binding score of < -8 Kcal/mol on both structures were selected (Supporting Information, **Table S3**), and this gave 49 compounds, which were subsequently analysed according to their drug-likeness, ease of synthesis and presence of known toxic groups. The top 20 compounds were identified as promising SARS CoV2 M^{pro} protease inhibitors (**Figure 2A**, Supporting Information, **Table S4**).

While conducting this work, the chemical structures of a series of covalent and non-covalent binding fragments co-crystalised with SARS CoV2 M^{pro} protease were reported.[14] To capitalise on this valuable information, we subsequently analysed these structures and selected 21 non-covalent binding fragments (Supporting Information, **Table S5**) and used them in our Rosiland model to generate new SARS CoV2 structures based on these fragments. In this process, we again applied a number of physicochemical properties filters that include molecular weight < 500 g/mol and LogP = 2-3 (see Experimental for full details) to ensure that the newly designed compounds confine to the most desirable physicochemical properties of known drugs. The model generated 500 structures, which upon docking into the SARS CoV2 (6YB7) M^{pro} protease, the top 20 compounds in terms the energy of binding were chosen (**Figure 2B,** Supporting Information, **Table S6**).

Subsequently, we analysed the final 40 compounds shown in **Figure 2**, which represent the combined deep learning-designed compounds based on the reported SARS CoV2 M^{pro} inhibitors and the non-covalent binders of SARS CoV2 M^{pro}, and chose five final SARS CoV2 M^{pro} inhibitors (**12**, **15**, **16**, **2´** and **11´**, **Figure 3**). Although these five compounds did not necessarily have the best binding free energy, they were chosen based on their ease of synthesis and lack of functional groups that are known in the medicinal chemistry field to be associated with some toxicity (e.g., phenols, and anilines).

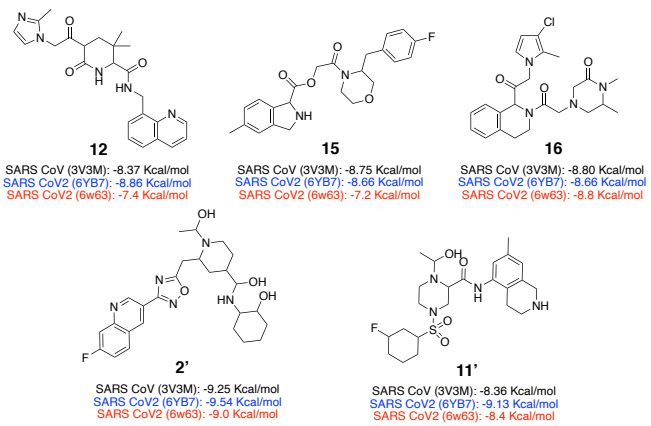

Figure 3. Chemical structures of the top five compounds selected from the deep learning-driven SARS CoV2 M^{pro} inhibitors.

Since in the deep learning-driven design of these compounds we docked them in the SARS CoV2 M^{pro} (6YB7) structure, we docked the top five compounds in a different SARS CoV2 M^{pro} (6w63)[15] structure to get a better insight into their binding modes and whether they retain good binding energies in their binding to SARS CoV2 M^{pro} (6w63), which is different from the SARS CoV2 M^{pro} (6YB7) structure that was used in their design. Notably, the SARS CoV2 M^{pro} (6w63) was obtained in complex with the broad spectrum non-covalent inhibitor X77.[15] Thus, we initially performed unbias docking of ligand X77 in the SARS CoV2 M^{pro} (6w63) structutre and X77 was docked in the same pocket where the co-crystallised X77 was bound (**a**, **Figure 4B**) [-8.2 Kcal/mol binding energy]. Critically, the docking pose of this ligand matched that of the cocrystallised ligand [calculated docking RMSD: 0.933Å) (**a**, **Figure 4B**) and this gave us confidence in our docking method. Subsequently, we docked our top five compounds shown in **Figure 3** in the SARS CoV2 M^{pro} (6w63) structutre, and these compounds were docked in the same pocket where the original co-crystallised ligand (X77) was bound (**Figure 4A**). Encouragingly, all of our top five compounds engaged the catalytic His41 residue in their binding akin to ligand X77, while compounds **2´** and **12** (**b** and **f**, respectively, **Figure 4B**) formed further interactions with the catalytic Cys145 of the SARS CoV2 M^{pro}. Additionally, out of all the five docked compounds and the ligand X77, compounds **2´** formed the most interactions with the SARS CoV2 M^{pro} (**Figure S1**) and had the best binding energy to SARS CoV M^{pro} (3V3M), SARS CoV2 M^{pro} (6YB7) and SARS CoV2 M^{pro} (6w63). Together, this makes compounds **2´** a promising candidate to investigated for its ability to inhibit SARS CoV2 M^{pro}.

In conclusion, this work presents the strategy we adopted in developing a deep learning approach that will be a powerful tool in the rapid discovery of SARS CoV2 M^{pro} inhibitors. Indeed, starting from 535 known SARS CoV2 M^{pro} predicted inhibitors and reported structures of 21 non-covalent binders to SARS CoV2 M^{pro} protease, we performed a series of cycles where a human chemistry knowledge informed a deep learning model optimisation approach towards the identification of forty promising SARS CoV2 M^{pro} inhibitors, which the scientific community could now pursue through the various COVID19 open and collaborative

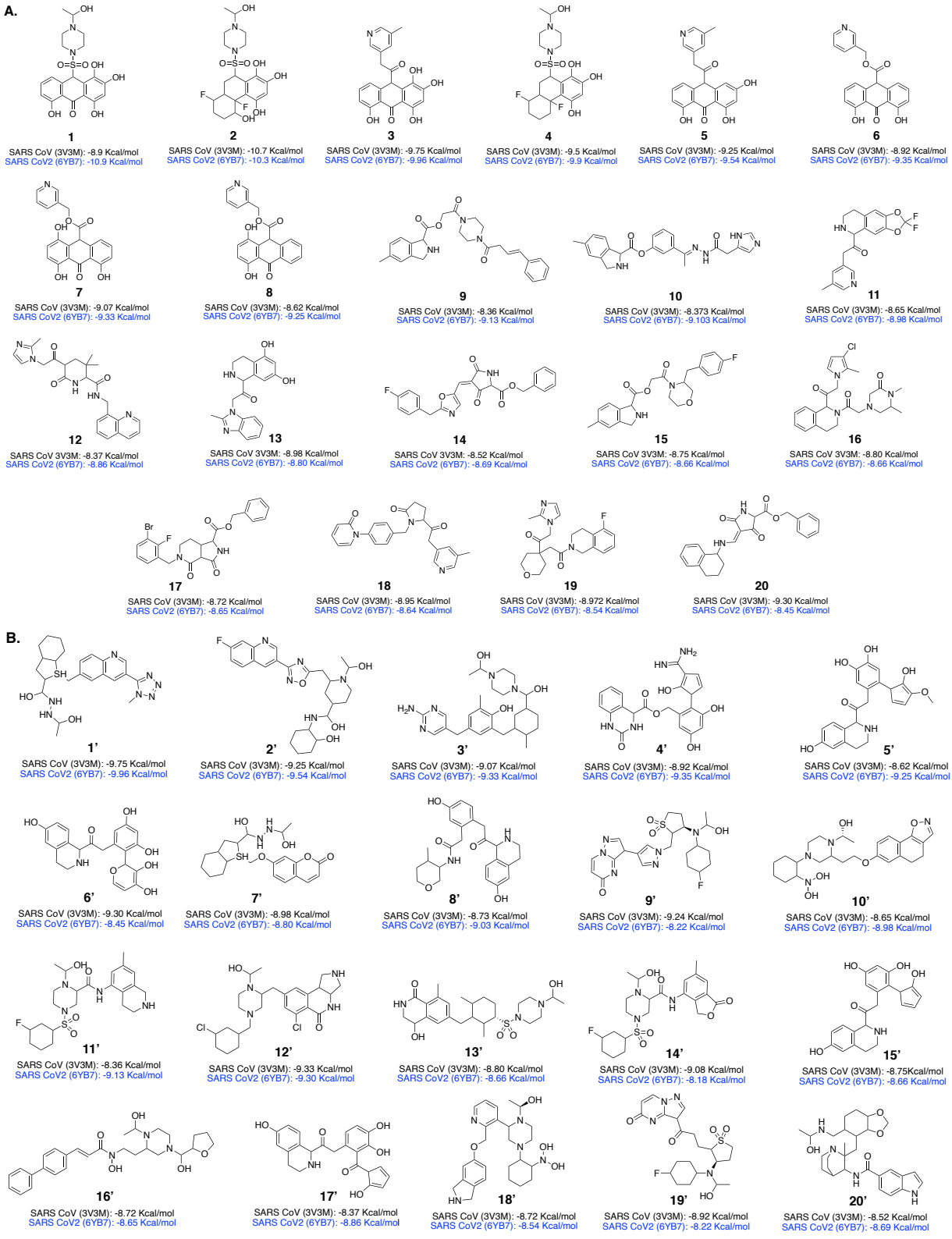

Figure 2. Chemical structures of the top promising SARS CoV2 M^{pro} inhibitors derived form reported inhibitors (**A**) and those based on the non-covalent binders of the SARS CoV2 (**B**). SMILES of these compounds are provided in the Supporting Information (**Tables S4** and **S6**).

initiatives, e.g., the COVID MoonShot project. In particular, we highlighted five molecules that could be investigated first for their ability to inhibit SARS CoV2 M^{pro}. Critically, we are continuing the training and optimization of our deep learning model in a collaborative manner to identify lead SARS CoV2 M^{pro} compounds with excellent drug-like properties that could be developed in a timely manner to address the urgent need for new and effective COVID19 treatments.

Keywords: COVID19 • SARS CoV2 • M^{pro} Protease • Inhibitor • Deep Learning

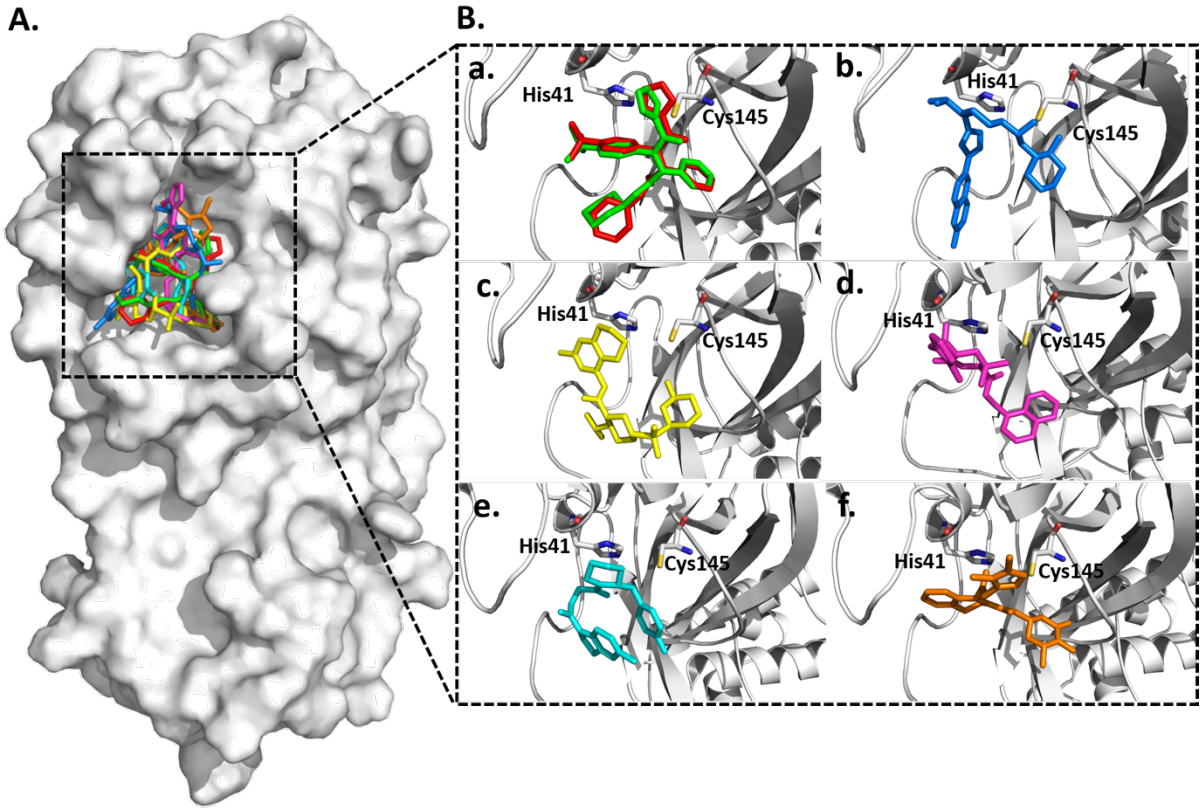

Figure 4. **A**. Molecular surface representation of SARS SRAS-CoV-2 M^{pro} (6w63) with the re-docked-X77 (green), **2'** (blue), **11'** (yellow), **12** (pink), **15** (cyan) and **16** (orange). The original co-crystallized ligand (X77) is shown in red sticks. **B**. Ribbon representation of the binding mode of the docked compounds. The catalytic His41 and Cys145 residues are showing in sticks.

EXPERIMENTAL

Data Preparation

Initially, Rosalind's deep learning generative models are trained on a large set of known active small-molecule drugs. The molecules were represented in Simplified Molecular Input Line Entry System (SMILES) format.[16] The SMILES dataset was preprocessed by applying sequential filters to remove stereochemistry, salts, undesirable atoms or groups and to maintain a canonical representation throughout the training and validation process. The RDKit library in python was used for dataset preprocessing.

De Novo Training Procedure

The generative model is initially trained to produce valid SMILES by training it on a set of 1.5 million active molecules curated from ChEMBL (compounds with a pChEMBL score greater than 7 (calculated from IC_{50}).[17]

To produce SARS CoV inhibitors, we focused on reported inhibitors on SARS CoV main protease M^{pro}. There is a reasonable number of compounds known to inhibit the SARS CoV main protease M^{pro} and two crystal structures of the SARS CoV2 main protease M^{pro} have recently been reported along with some initial hit compounds.[8, 12] This allowed us to curate a set of 535 inhibitors reported SARS CoV main protease M^{pro}. (Supporting Information, **Table S1**).[9] This set was visually inspected to identify promising core structures. A small number of 42 prioritised compounds were then further analysed (Supporting Information, **Table S2**) and a handful were manually docked using AutoDock Vina software[18] with a known protease crystal structure (PDB ID: 3V3M)[13]. After this process, a key pharmacophore was identified (**Figure 1C**) and selected to seed the de novo design procedure.

For inhibition of SARS CoV2, our starting set constitutes 21 non-covalent binding fragments co-crystallized[14] with SARS CoV2 M^{pro} (Supporting Information, **Table S5**). Rosalind's generative models are then seeded with the selected fragments and the trained models are run for approximately 24 hours rapidly producing millions of design ideas and pruning them according to a predefined drug-likeness scoring measure. For this study, the fitness score used constitutes drug-likeness and medicinal chemistry measures. Drug-likeness reward molecules with properties representative for protease inhibition; LogP: 2-3; molecular weight (MW: < 500 g/mol). Additional drug-likeness filters used include: number of hydrogen bond donors (HBD): 0-7; Number of hydrogen bond acceptors (HBA): 4-11; and topological polar surface area (tPSA): 60–200 Å. Medicinal chemistry filters were applied to incorporate expert insights filtering out structures containing rings bigger than six atoms and polypeptides (n ≥ 4).

RDKit predictive models were used to predict LogP and molecular weights (MW). A deep learning predictive model based on Ramsundar B *et al.*[19] was used for toxicity prediction.

Given that the system is designed to produce only valid SMILES, and because of the use of the above restrictions, Rosalind automatically filtered unfit designs and produced a valid set of 500 compounds for each strategy followed after 24 hours.

Docking Procedure

SARS CoV M^{pro} (PDB ID: 3V3M),[13] SARS CoV2 M^{pro} (PDB ID: 6YB7)[12] and SARS CoV2 M^{pro} (PDB ID: 6w63)[15] structures were obtained from PDB (https://www.rcsb.org/), in .pdb format. The files were prepared for docking using AutoDockTools by removing the crystallised inhibitors, removing water, adding polar hydrogens to the protein structures, defining the dimension and the center of the grid box for docking simulation and converting the PDB files into PDBQT. Candidate compounds in both sets were docked against the selected structures. For automatic docking, we used Smina,[11] a fork of Autodock Vina that focuses on improving scoring and minimization. Docking was performed at exhaustiveness of 10, and a random seed was selected. The docking pipeline was distributed and ran for about five hours on a computing cluster with 16 CPUs.

The generated compound sets were then ordered according to the binding energies obtained from the docking. Hits with binding free energy higher than -7.5 kcal/mol on either structure were discarded. The remaining set contains 100 compounds scoring below -7.5 kcal/mol on both structures. A final score is given to the remaining set of hits that is the sum of both binding energies. This score is used to order the set and a final list of top 20 hits (all scoring below -8.5 kcal/mol on both proteases) is selected for the final manual docking

and visual inspection (Supporting Information, **Table S4** and **Table S6**).

References

[1] a) S. R. Weiss, J. L. Leibowitz, *Adv. Virus Res.* **2011**, *81*, 85-164; b) G. Lu, Q. Wang, G. F. Gao, *Trends Microbiol.* **2015**, *23*(8), 468-478.
[2] C. M. Coleman, M. B. Frieman, *J. Virol.* **2014**, *88*(10), 5209-5212.
[3] J. Guarner, *Am. J. Clin. Pathol.* **2020**, *153*(4), 420-421.
[4] E. Mahase, *BMJ* **2020**, *368*, m641.
[5] X. Xu, P. Chen, J. Wang, J. Feng, H. Zhou, X. Li, W. Zhong, P. Hao, *Sci. China. Life Sci.* **2020**, *63*(3), 457-460.
[6] Y. Shen, D. Zhou, L. Qiu, X. Lai, M. Simon, L. Shen, Z. Kou, Q. Wang, L. Jiang, J. Estep, R. Hunt, M. Clagett, P. K. Sehgal, Y. Li, X. Zeng, C. T. Morita, M. B. Brenner, N. L. Letvin, Z. W. Chen, *Science* **2002**, *295*(5563), 2255-2258.
[7] a) T. Pillaiyar, M. Manickam, V. Namasivayam, Y. Hayashi, S. H. Jung, *J. Med. Chem.* **2016**, *59*(14), 6595-6628; b) Y. M. Baez-Santos, S. E. St John, A. D. Mesecar, *Antiviral Res.* **2015**, *115*, 21-38; c) R. Ramajayam, K. P. Tan, P. H. Liang, *Biochem. Soc. Trans.* **2011**, *39*(5), 1371-1375; d) H. Wang, S. Xue, H. Yang, C. Chen, *Virol. Sin.* **2016**, *31*(1), 24-30; e) Q. Zhao, E. Weber, H. Yang, *Recent Patents Anti-Infective Drug Discov.* **2013**, *8*(2), 150-156.
[8] a) 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, H. Yang, *Nature* **2020**, DOI: 10.1038/ s41586-020-2223-y; b) L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox, R. Hilgenfeld, *Science* **2020**, DOI: 10.1126/science.abb3405.
[9] a) D. Duy Nguyen, K. Gao, J. Chen, R. Wang, G. Wei, *bioRxiv* **2020**, DOI: 10.1101/2020.01.30.927889; b) B. Tang, F. He, D. Liu, M. Fang, Wu, D. Xu, *bioRxiv* **2020**, DOI: doi.org/10.1101/2020.03.03.972133; c) https://ghddi-ailab.github.io/Targeting2019-nCoV/CoV_Experiment_Data/ (accessed in March **2020**).
[10] a) J. Xu, P. Y. Shi, H. Li, J. Zhou, *ACS Infect. Dis.* **2020**, DOI: 10.1021/acsinfecdis.0c00052; b) B. N. Williamson, F. Feldmann, B. Schwarz, K. Meade-White, D. P. Porter, J. Schulz, N. v. Doremalen, I. Leighton, C. K. Yinda, L. Pérez-Pérez, A. Okumura, J. Lovaglio, P. W. Hanley, G. Saturday, C. M. Bosio, S. Anzick, K. Barbian, T. Cihlar, C. Martens, D. P. Scott, V. J. Munster, E. d. Wit, *bioRxiv* **2020**, DOI: 2020.2004.2015.043166.
[11] D. R. Koes, M. P. Baumgartner, C. J. Camacho, *J. Chem. Inf. Model.* **2013**, *53*(8), 1893-1904.
[12] C. D. Owen, P. Lukacik, C. M. Strain-Damerell, A. Douangamath, A. J. Powell, D. Fearon, J. Brandao-Neto, A. D. Crawshaw, D. Aragao, M. Williams, R. Flaig, D. R. Hall, K. E. McAuley, M. Mazzorana, D. I. Stuart, F. von Delft, M. A. Walsh, SARS-CoV-2 main protease with unliganded active site (2019-nCoV, coronavirus disease 2019, COVID-19), **2020** (https://www.rcsb.org/structure/6YB7).
[13] J. Jacobs, V. Grum-Tokars, Y. Zhou, M. Turlington, S. A. Saldanha, P. Chase, A. Eggler, E. S. Dawson, Y. M. Baez-Santos, S. Tomar, A. M. Mielech, S. C. Baker, C. W. Lindsley, P. Hodder, A. Mesecar, S. R. Stauffer, *J. Med. Chem.* **2013**, *56*(2), 534-546.
[14] https://www.diamond.ac.uk/covid-19/for-scientists/Main-protease-structure-and-XChem/Downloads.html (Accessed March **2020**).
[15] A. D. Mesecar, Structure of COVID-19 main protease bound to potent broad-spectrum non-covalent inhibitor X77, **2020,** https://www.rcsb.org/structure/6W63 (Accessed April **2020**).
[16] D. Weininger, A. Weininger, J. L. Weininger, *J. Chem. Inf. Comput. Sci.* **1989**, *29*(2), 97-101.
[17] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, *Nucleic Acids Res.* **2012**, *40*, D1100-1107.
[18] O. Trott, A. J. Olson, *J. Comput. Chem.* **2010**, *31*(2), 455-461.
[19] B. Ramsundar, P. Eastman, P. Walters, V. Pande, *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*, O'Reilly, **2019**.