UNIVERSITAT ROVIRA i VIRGILI

Targeting SARS-CoV-2 main protease (M-pro): repositioning of seven approved drugs in a consensus docking-based virtual screening

Júlia Mestres Truyol

TREBALL FINAL DE GRAU BIOTECNOLOGIA

Tutor acadèmic: Dr. Gerard Pujadas Anguiano, Departament de Bioquímica i Biotecnologia, URV (gerard.pujadas@urv.cat)

En cooperació amb: Grup de recerca en Quimioinformàtica i Nutrició (QiN), Departament de Bioquímica i Biotecnologia, URV

Supervisors: Dr. Gerard Pujadas Anguiano i Dr. Santi Garcia-Vallvé, Departament de Bioquímica i Biotecnologia, URV (gerard.pujadas@urv.cat, santi.garcia-vallve@urv.cat)

Juny 2021

Jo, Júlia Mestres Truyol, amb DNI 49535479-B, sóc coneixedora de la guia de prevenció del plagi a la URV Prevenció, detecció i tractament del plagi en la docència: guia per a estudiants (aprovada el juliol 2017) (http://www.urv.cat/ca/vidacampus/serveis/crai/que-us-oferim/formacio-competencies-nuclears/plagi/) i afirmo que aquest TFG no constitueix cap de les conductes considerades com a plagi per la URV.

Tarragona, 7 de juny de 2021

JÚLIA MESTRES TRUYOL - DNI 49535479B

Firmado digitalmente por JÚLIA MESTRES TRUYOL - DNI 49535479B
Fecha: 2021.06.07 09:09:25 +02'00'

(signatura)

Table of contents

Abstract 7

1. Introduction 9

1.1. SARS-CoV-2 is the causative agent of the COVID-19 9

1.2. SARS-CoV-2 M-Pro is a good target to inhibit the virus replication 10

1.3. Drug repurposing is a shortcut strategy in drug development 13

1.4. Protein–ligand docking is popular among virtual screening strategies 14

1.5. The COVID Moonshot initiative aims to identify antiviral drugs targeting SARS-CoV-2 M-pro 17

2. Hypothesis and objective 18

3. Materials and methods 19

3.1. Compound libraries description and preparation 20

3.2. M-Pro structure preparation, grid generation and protein-ligand docking setup 20

3.3. Identification of equivalent and high-affinity docked poses 21

3.4. Analysis of the intermolecular interactions between M-pro and its inhibitors 22

3.5. Virtual screening workflow validation 23

4. Results and discussion 24

4.1. Virtual screening of approved drugs 24

4.2. Selectivity of the virtual screening workflow 31

5. Conclusions 33

6. Acknowledgements 34

7. Reference list 35

8. Self-assessment 37

The present work has been developed during an extracurricular internship at the Cheminformatics and Nutrition (QiN) research group in the Biochemistry and Biotechnology Department of the Rovira i Virgili University (URV), under the supervision of Dr. Gerard Pujadas Anguiano and Dr. Santi Garcia-Vallvé. The results herein presented have been previously published in the International Journal of Molecular Sciences (IJMS) 21(11), 3793, on date 27 May 2020 under the title “Prediction of Novel Inhibitors of the Main Protease (M-pro) of SARS-CoV-2 through Consensus Docking and Drug Reposition”.

Abstract

SARS-CoV-2 is the virus responsible for the COVID-19, a severe respiratory disease that emerged in December 2019 and rapidly spread worldwide. The consequent pandemic has pressed the scientific community to make an unprecedented effort to find an effective vaccine or cure. Currently, despite the implementation of effective vaccines, we are far from herd immunity and thus an alternative therapeutic option is necessary. This work aims to reposition approved drugs as putative inhibitors of a target with a pivotal role in the replication of the virus: the main protease (M-pro). An original consensus docking-based virtual screening (VS) strategy was applied with the intention of focusing on the sampling algorithms of different programs without relying on their scoring functions. Thus, docking was performed with three programs –Glide, FRED and AutoDock Vina– and only the equivalent high affinity binding modes predicted by all of them were considered to correspond to bioactive poses. Seven possible SARS-CoV-2 M-pro inhibitors were predicted with this approach: Perampanel, Carprofen, Celecoxib, Alprazolam, Trovafloxacin, Serafloxacin and Ethyl biscoumacetate. Carprofen, Celecoxib and Sarafloxacin have been selected by the COVID Moonshot initiative for *in vitro* testing, showing a 3.97, 11.90, 20.00% M-pro inhibition at 50 μM, respectively, and a percentage of inhibition at 10 μM of 43 % has been reported for Perampanel elsewhere. These compounds could serve as a starting point for the development of more potent derivatives in a lead optimization process. This VS strategy could further be applied to other databases to predict more putative SARS-CoV-2 M-pro inhibitors which would add up to the list of compounds available for *in vitro* bioactivity assays against COVID-19 as well as to other targets of interest.

Keywords: COVID-19; SARS-CoV-2; M-Pro; 3CL-pro; virtual screening; protein-ligand docking

1. Introduction

1.1. SARS-CoV-2 is the causative agent of the COVID-19

In late 2019, a severe respiratory disease emerged in Wuhan, China, and rapidly spread in many countries around the world culminating in a still ongoing global pandemic. A new coronavirus (CoV) showing 80% of genomic sequence identity compared to the SARS-CoV-2 was soon identified as the ethological cause of the disease. Hence, the virus has been called SARS-CoV-2 and the disease is referred as COVID-19 (COronaVIrus Disease 2019) [1]. As of 6 June 2021, more than 172 million of cases and roughly 3.7 million of deaths have been reported to the World Health Organization (https://covid19.who.int/) since the onset of the pandemic. COVID-19 pathogenesis encompasses mild common symptoms such as fever, fatigue, dry cough and dyspnea as well as gastrointestinal illness which can lead to severe pneumonia culminating in multiorgan failure and death in critical cases. After binding of SARS-CoV-2 to epithelial cells it starts replicating and migrates down the respiratory tract towards alveolar cells in the lungs. Its rapid replication seems to trigger a strong immune response which later progresses to a general pro-inflammatory state, accounting for the acute respiratory response and for long-term ramifications of the pathology [2]. In the absence of any vaccine or cure, the most effective way to "flatten the curve" has been the implementation of measures such as border shutdowns, travel restrictions, lock-down, curfew, social distancing or the use of masks. This, in turn, has repercussed on people's life, health systems and economy [3]. As for now, there is no cure for SARS-CoV-2 and, although vaccination campaigns seem to progress readily, we have not reached herd immunity yet.

1.1.1. The Coronavirus family

Coronaviruses (CoVs) are single-stranded positive-sense RNA viruses (ss-(+)-RNA) that belong to the subfamily *Coronavirinae*, family *Coronavirdiae*, order *Nidovirales*. The subfamily can be further clustered in four genera: *Alphacoronavirus* (α-CoVs), *Betacoronavirus* (β-CoVs), *Gammacoronavirus* (γ-CoVs) and *Deltacoronavirus* (δ-CoVs). Mammals are the main hosts for α- and β-CoVs, while γ- and δ-CoVs mainly infect avian species [4]. On some occasions, CoVs have crossed the interspecies barrier and emerged as human pathogens. Some human CoVs (*i.e.*, HCoV- NL63, HCoV-229E, HCoV- OC43 and HKU1) usually cause common cold disease. However, here we focus on a more pathogenic β-CoVs, SARS-CoV-2, similar to the Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) and the Middle East Respiratory Syndrome coronavirus (MERS- CoV), which emerged on 2002 and 2012, respectively. The closest relative to SARS-CoV-2 is bat CoV, SL-CoV-RaTG13, with 96% of sequence identity [1], indicating that bats could have been a natural reservoir for the virus. However, the divergence between them points at a still unknown intermediate host as the direct progenitor of SARS-CoV-2, and bats would be regarded as an evolutionary precursor [5].

1.1.2. SARS-CoV-2 genome structure and lifecycle

In a SARS-CoV-2 virion, the ss-(+)-RNA genome is protected by complexing into a helical ribonucleocapsid with the nucleocapsid (N) protein. The viral envelope contains the envelope (E), membrane (M) and spike (S) proteins. The S glycoprotein –which gives a crown-like appearance in electron microscope imaging– mediates the virus attachment to the host cell interacting with the angiotensin-converting enzyme 2 (hACE2) [6]. These structural proteins are encoded in the 3' one third of the coding genome. Besides, the two other thirds at the 5' end comprise two overlapping open reading frames, ORF1a and ORF1b, that encode for polyproteins pp1a and pp1b. Their proteolytical processing by the Main-protease (M-pro) and Papain-Like protease (PL-pro) yields 16 nonstructural proteins (nsps1-16) that conform the replicase/transcriptase complex (RTC). The RTC includes, among others, the two proteases themselves, an RNA-dependent RNA polymerase (RdRp), a helicase (Hel) and an exoribonuclease (ExoN) (Figure 1). The coding region of the genome lies within 5'- and 3'- untranslated regions (UTRs) that are important in RNA replication and transcription and a 5'-cap structure with a leader sequence and a 3'-poly(A) tail that allow it to act as a mRNA. All in all, SARS-CoV-2 genome is composed of 29.9 kb and encodes more >9000 amino acids. CoVs genomes are, in fact, the largest among RNA virus [7].

Upon recognition of the cellular receptor ACE2 by the spike protein, viral uptake takes place. Following entry, nucleocapsid degradation allows the release into the cytoplasm of the viral RNA, which immediately undergoes translation of ORF1a and ORF1b. The polyproteins pp1a and pp1b are then processed into nonstructural proteins by M-pro and PL-pro. The transcription of genomic RNA by the RTC proceeds through (-)-RNA intermediates that work as templates for both genomic and a nested set of subgenomic mRNAs. The latter are translated into the four accessory and structural viral proteins. Newly synthesized virions are eventually assembled on intracellular membranes and released from the infected cell by endocytosis [6]. Different steps of the lifecycle of SARS-CoV-2 (Figure 1) have been proposed as targets to inhibit the virus replication [8]. Among them, major attention has been drawn to the inhibition of M-pro as an anti-CoV strategy since the first SARS-CoV outbreak.

1.2. SARS-CoV-2 M-Pro is a good target to inhibit the virus replication

The SARS-CoV-2 M-Pro is a cysteine protease responsible for the processing of the polyproteins at a minimum of 11 sites, starting by the autolytic cleavage from pp1a. The release of most nonstructural proteins, thus, relies on M-pro activity. This central role in the viral life cycle, along with the absence of a human homologue in human cells, make it an inviting target for the development of antiviral drugs [9]. Moreover, no less than 300 crystal structures of M-pro have been uploaded –either in complex with inhibitors or in their apo forms– to the Protein Data bank (https://www.rcsb.org/) hitherto, shedding light onto the tridimensional structure, assembly and catalytic mechanism of the protein.

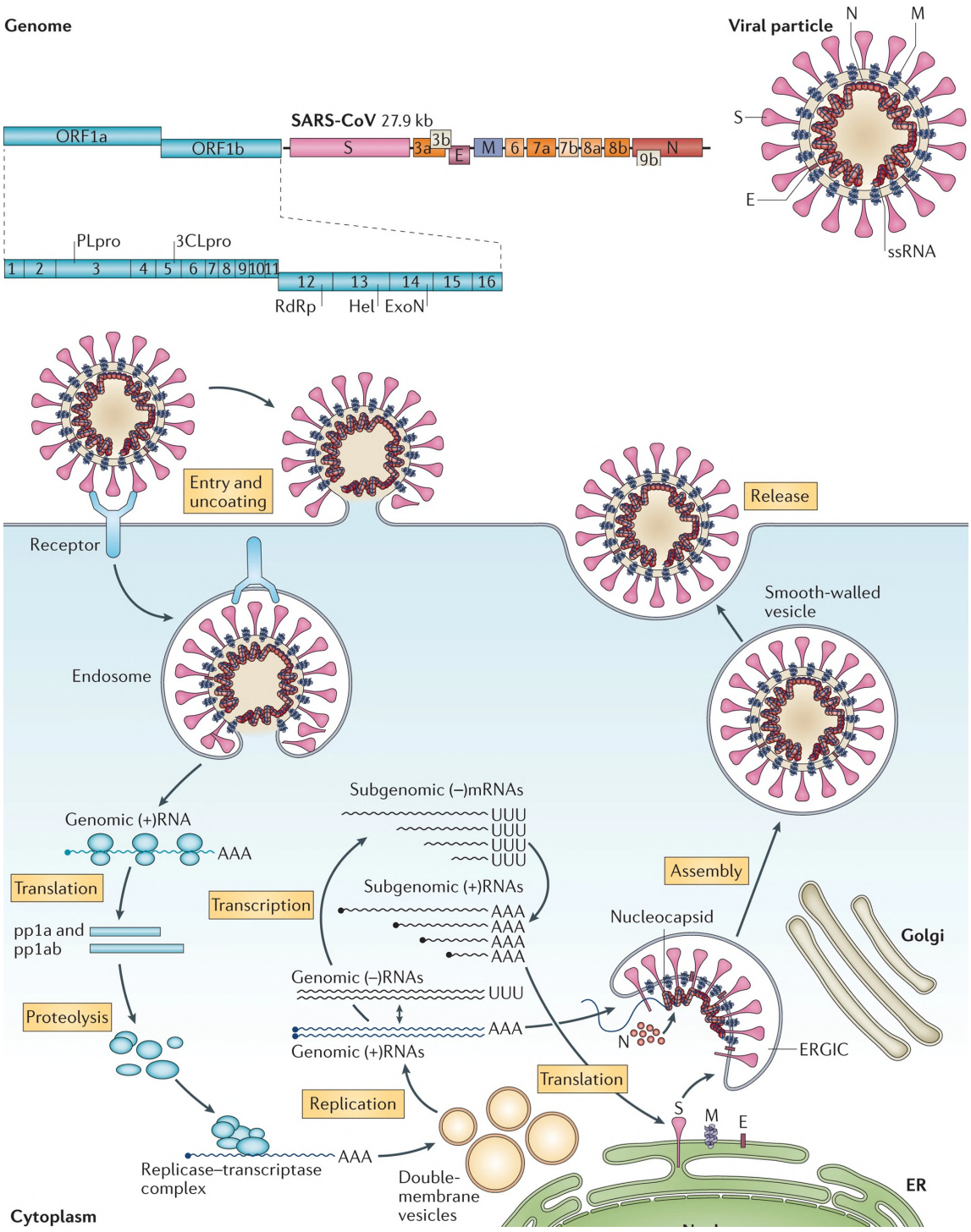


Figure 1 | SARS-CoV-2 genome structure and life cycle. Adapted from [10].

The tridimensional structures of SARS-CoV-2 and SARS-CoV M-pro reveal a high degree of analogy, in consonance with the 96% of sequence identity and the similar catalytic mechanism. As far as the latter is concerned, although proteolysis in other chymotrypsin-like enzymes depends on three catalytic residues, M-pro has a Cys145-His41 catalytic dyad in CoVs active site instead. A buried water molecule takes place in a network of interactions nearby the catalytic site. It has a stabilizing effect on His41 through a strong H-bond and is in turn stabilized through H-bonds with Asp187 and His64 [11]. This water

molecule (H_2O_{cat}) is thus believed to replace the function of the third catalytic residue, completing a non-canonical catalytic triad (Figure 2c) [12]. The proteolytic mechanism follows a common nucleophilic-type reaction. In a first step –acylation– Cys145 covalently binds to the carbon of the P1 residue of the substrate peptide, forming an acyl-enzyme complex. In a second step –deacylation– the acyl-enzyme is hydrolyzed, releasing the substrate and restoring the active site [13]. In this second step the water molecule is postulated to act as a deacetylating nucleophile.

M-pro has been reported to be catalytically inactive as a monomer. Thus, the active form of M-pro is a dimer with each of the protomers composed of three domains (Figure 2a). Domains I (residues 8-101) and II (residues 102-184) share a common fold of ß-barrel structures and have a catalytic role while domain III (residues 201–303) is conformed of 5 α-helices arranged into a largely antiparallel globular cluster and is involved dimerization, primarily through a salt-bridge between Glu290 of one protomer and Arg4 of the other. Dimerization via interactions between domain II of one protomer –specially with Glu166– and the N-finger of the other protomer (residues 1-7) shape the S1 pocket of the binding site. Thus, albeit domain III is non-catalytic, its role in dimerization makes it essential for M-pro activity. Also, it is worth mentioning the structural flexibility of the binding site which can be determined by comparing its structure upon binding of a ligand with the apo form of the protein. This plasticity constitutes an induced fit owing to binding of the ligand [11].

The substrate binding cleft of each protomer is hosted between domains I and II and includes the catalytic dyad. Cys145 is part of the oxyanion loop –an S-shaped loop in domain II composed by Gly138-Gly146– and Cys145 and Gly143 delimit the oxyanion hole, where the carbonyl group of the scissile peptide bond of the susbtrate binds [9]. Five subsites can be identified in the binding site, S1-S1'-S2-S3-S4, where the amino acids P1↓P1'-P2-P3-P4 (where ↓ marks the cleavage site) of the substrate fit, respectively (Figure 2b). The substrate recognition motif Leu-Gln↓Ser-Ala-Gly appears to be shared among most CoV, although the enzyme exhibits sequence promiscuity [11].

The S1 subsite is made up of the side chains of Phe140, Asn142, His163, Glu166, and His172 and the main chains of Phe140 and Leu141. At the bottom of the pocket, the imidazole of His163 is located so it can donate a H-bond to the side chain carbonyl of the substrate Gln P1. The S1' subsite hosts several threonine residues (Thr24, Thr25, Thr26) which can interact with the group at P1' by forming either hydrogen bonds or lipophilic interactions, as well as Gly143, Ser144 from the oxyanion loop and the catalytic dyad. The S2 subsite is a hydrophobic cleft where alkyl/aryl substituents, such as the side chain of substrate Leu P2 can fit. It is covered by a lid consisting of residues 46-51, laterally defined by the main chains of residues 186-188, the side chains of His41, Asp187 and Gln189 and hosts a Met 164 and Met165 in the bottom. Subsites S3 and S4 are delimited by the flexible loops linking residues 165–168 and 189–192. In contrast to S1 and S2, the remaining subsites are less buried and more exposed to the solvent, meaning that they are likely to admit groups of a varied size and nature [9].

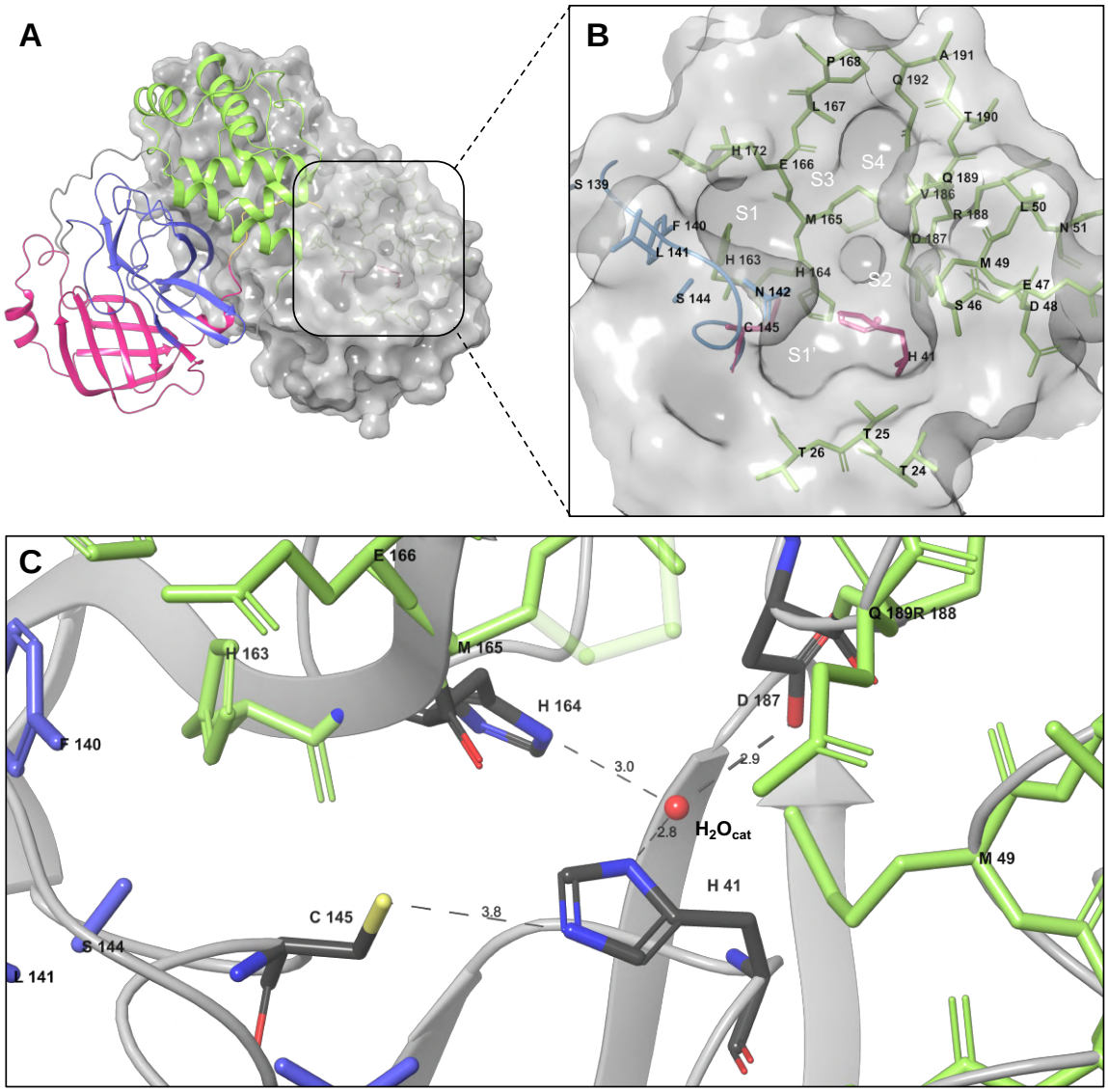


Figure 2 | In panel **A** one protomer of the functional dimer of SARS-CoV-2 M-Pro is shown in cartoon with domains I, II and III and the N-finger colored in pink, bule, green and yellow, respectively. The other protomer is shown as a surface with the binding site residues highlighted. Panel **B** provides a closeup look at the binding site in which the residues of the binding site (green), the catalytic dyad and the H_2O_{cat} (pink) and the oxyanion loop (blue) are highlighted. Panel **C** is a detail of the catalytic site of SARS-CoV-2 M-pro with key residues colored in CPK. Distances between interacting atoms are given in Ångstroms. All figures were generated with Shcrödinger's Maestro [14], using the crystal structure of the apo protein (PDBid 6WQF).

1.3. Drug repurposing is a shortcut strategy in drug development

Drug repurposing (also referred as drug repositioning) is a widely used strategy for identifying alternative medical uses for already approved or investigational drugs. This approach provides numerous advantages in comparison to developing a new drug entirely from scratch. As safety has already been assessed in early-stage trials with the repurposed drug, the risk of failure is greatly reduced. Also, time-consuming steps in drug development (*e.g.*, preclinical testing, safety assessment or formulation development) are shortened, thereby minimizing the time frame for the development of the drug. These advantages, together with a reduction of the associated costs, allow a faster and less risky return on investment in the development of repurposed drugs [15].

In contexts where time presses, such as the currently ongoing COVID-19 pandemic, drug repositioning turns out to be a convenient approach.

The first step in a drug repurposing strategy is the identification of a candidate molecule for a target of interest. Given the high throughput of computational approaches, they tend to be a mainstay in this step, often synergistically with experimental approaches. Among the most popular structure-based *in silico* approaches are virtual screening strategies based on protein-ligand molecular docking, which have been extensively used in both drug discovery and drug repurposing [16,17].

1.4. Protein–ligand docking is popular among virtual screening strategies

1.4.1. Virtual screening

Virtual screening (VS) is a computational approach employed to select compounds that can bind to a specific target –generally a receptor or an enzyme– among a large library of initial compounds. Different methods are executed in a stepwise manner in a VS procedure, which act as filters that keep only a part of the compounds from the previous step. This leads to a funnel-like hierarchical workflow, where only the "hit" compounds make it through all the filters [18]. With the VS, the candidates list can be dramatically reduced, although the hit compounds' biological activity needs to be experimentally tested to consider them active compounds.

Thus, the main advantage of VS is that it allows processing of thousands of compounds in only several hours and avoids synthesizing, buying and experimentally testing all the compounds, thus reducing the costs. It is worth bearing in mind that not always will a VS procedure give hit compounds with actual high activity [15]. However, obtaining structurally divergent lead compounds that can further be optimized in subsequent steps is yet another objective of VS. VS results can also be useful to understand the molecular basis behind active compounds, specifically if receptor-based methods are being used, and this information can further be considered during the optimization process [16].

One can distinguish between two kinds of VS methodologies: ligand-based VS (*e.g.*, fingerprint-based methods, quantitative structure activity relationship (QSAR), or supervised machine learning) and receptor-based VS (*e.g.*, protein-ligand docking or structure-based pharmacophores). The former is based on the similarity of the library of compounds with active compounds, while the latter focus on the complementarity of the compounds with the binding site of the protein [18]. Although ligand-based methods can work when no structural information of the protein in available, they require at least one reported active molecule. Also, ligand-based VS methods assume that similar compounds will have similar activities, but on some occasions small changes in the structure of a compound cause drastic variations of activity [19]. These cases, often referred as activity cliffs, can be bypassed if the structure of the receptor is considered, like in receptor-based VS methods.

1.4.2. Protein-ligand docking

Protein-ligand molecular docking is a receptor-based VS method that uses the tridimensional structure of a target protein to predict how ligands from the initial library (*e.g.,* libraries of approved drugs) bind to the binding site of the target (*e.g.*, SARS-CoV-2 M-pro). For each of the compounds, possible conformations of the compounds placed in a certain orientation so that they may fit in the binding site (*i.e.*, docked poses) are generated [16]. Its capacity to interrogate thousands of drugs against one protein target at a relatively low computational cost make protein-ligand docking a popular tool in computational and structural biology. Protein-ligand docking is preceded by the following steps [16]:

1. Library preparation: the initial library refers to the set of compounds to which the docking will be applied to. It can be obtained from databases, commercial suppliers or directly from an in-house collection of compounds. If the structures of this compounds are collected in 2D format, they need to go through a process called conformational sampling in order to predict 3D conformations. Also, different tautomers and possible protonation states at a pH of interest have to be generated for each molecule.
2. Protein preparation: X-ray crystallographic structures are experimental and present problems that must be corrected before being used in the docking. For instance, they can have missing hydrogen atoms, missing residues, incomplete side-chains, undefined protonation states or there can be crystallization products that are not found *in vivo.*
3. Binding site definition: the space that will be occupied by the docked poses is restricted by defining the limits of the pocket of the proteins where the compounds should be docked.

Molecular docking methods use two independent steps [20]:

1. Conformational sampling: different docked poses within the binding site of the protein are generated by a search algorithm.
2. Scoring: the protein-ligand affinity of each docked pose is estimated by a scoring function that predicts the strength of the interactions. This score is then used to rank the different docked poses for each compound, so that the first docked pose will be the best one and most likely represent the real binding mode.

Glide, FRED and AutoDock Vina are extensively used docking software, that follow different approaches to generate docked poses and score the results. Glide and FRED use different exhaustive algorithms to obtain docked poses, while AutoDock Vina uses an iterated local search global optimizer [21,22]. Regarding their scoring functions, while the Glide and FRED scoring functions are fully empirical [23,24], the scoring function of AutoDock Vina is a hybrid scoring function that incorporates empirical and knowledge-based elements [25].

1.4.3. Protein-ligand docking limitations

The main limitation of docking programs is that their scoring functions often have a low success rate at predicting the binding affinity of compounds. A better docking score is not a reliable criteria for selecting the best docking pose, as it does not necessarily correlate with biological activity. Therefore, protein-ligand docking should not be used to predict compound activity, but to enrich the initial library in active compounds by keeping those that are more likely to have a good binding affinity, according to the scoring function, and ruling out those that do not fit in the binding site. The former can be solved by setting a docking score threshold [16].

Because of the differences between these three protein-ligand docking programs –both in their search algorithms and scoring functions– and the abovementioned unreliability of their scoring functions one possible solution to improve the performance of docking-based VS is implementing a consensus docking. Integrating the results of different docking programs and focusing on their intersection compensates for their individual weaknesses [26].

Here, for instance, a consensus docking strategy was followed, where docking simulations were performed parallelly with Glide, FRED and AutoDock Vina and docked poses predicted by the three of them were selected. After that, a docking score threshold was applied so as to select probable high-affinity docked poses [27].

Another shortcoming of docking programs is the lack of motion. While the ligand is treated as a flexible molecule, the protein conformation is considered rigid and its atoms coordinates are not allowed to change [20,28]. Also, docked poses are unexpectedly dependent on the ligand input structure [29], and the ability of docking programs to discern actives from inactives is largely dependent on the protein structure and the similarity between screened ligand and the ligands co-crystallized with the employed structure [17]. A last limitation is that docking programs neglect the role of water, which has an important role in the binding process [20]. It is also worth noting that docking scores obtained by different programs are not comparable, as they are dependent on the force fields and algorithms used.

1.4.4. Protein-ligand docking validation

For all the limitations exposed above, it is of major importance to validate a protein-ligand docking methodology, any VS step or a VS protocol in general. The ideal validation is to experimentally assess the biological activity of the hit compounds. However, if that is not possible, or if one wants to validate the methodology before testing the hit compounds *in vitro*, it must be subject to an *in silico* validation. A VS strategy can be computationally validated by being applied to a set of active compounds and to a set of inactive or decoy compounds [16]. The use of inactives or decoys is key as they are a negative control. Naturally, a validation of this kind is only applicable when there is available information about the target protein.

Decoys are putative inactive compounds. Briefly, they are compounds which physically resemble active compounds, but are chemically different and since are presumed to be inactive. In validation processes, decoys are commonly used rather than inactives due to the lack of data on inactive compounds that can be found in the literature. Decoys can either be obtained from databases, such as DUD·E [30] or generated with tools, such as DecoyFinder [31].

The more active compounds that get through the VS step (*i.e.*, true positives) and inactive compounds that do not get through the VS step (*i.e.*, true negatives), the more accurate the methodology is. The extent to which a VS can be enriched with active compounds can be quantified with the area under a ROC curve or with the Enrichment Factor (EF) [16]:

$$EF = \frac{\frac{a_2}{a_2+d_2}}{\frac{a_1}{a_1+d_1}}$$

a_1 = actives before the VS step
a_2 = actives after the VS step
d_1 = decoys before the VS step
d_2 = decoys after the VS step

1.5. The COVID Moonshot initiative aims to identify antiviral drugs targeting SARS-CoV-2 M-pro

COVID Moonshot is a crowdsourced initiative that aims contribute to pandemic preparedness without relying on commercial interests. It is an international consortium of scientists that work in a non-profit open-science drug-discovery model, with the purpose of developing antiviral drugs targeting the SARS-CoV-2 M-pro. Drug-designers around the world have submitted roughly 10000 molecule designs relying on different hypotheses and methods (*e.g.,* docking, approaches, by-eye structure-based designs, machine learning approaches, SARS and MERS literature reviewing, etc.). The activity against M-pro of over 1000 out of these compounds has been assessed after either ordering or synthesizing them. Two complementary biochemical assays have been used: a fluorescence-based assay and a RapidFire Mass Spectometry assay. This activity data, together with over 200 crystal structures of M-pro complexes with some of these compounds, is made publicly available and frequently updated for other researchers who may find it useful [32].

2. Hypothesis and objective

The dreadful situation triggered by the COVID-19 pandemic demands an efficient strategy to urgently find drugs for the treatment of COVID-19. SARS-CoV-2 M-Pro is a key target for the development of such therapies not only for its essential role in the replication of the virus, but also because this enzyme is highly conserved among related viruses and does not exist in humans. Also, the structure of M-pro is well-known, as it has been obtained by X-ray crystallography.

Computational approaches, such as protein-ligand docking, are useful to speed up drug discovery and reduce costs, which makes them a promising choice in a context as such. Moreover, drug repurposing, rather than traditional drug discovery, allows the identification of drugs whose safety profiles have already been proved, thus reducing costs, time and risk even more. Finally, alternative approaches that take advantage of its potential and outweigh their shortcomings, such as consensus docking, can be implemented to deal with the main limitation of protein-ligand docking programs: the unreliability of their scoring functions.

With this idea in mind, the main objective of the current work is to identify putative SARS-CoV-2 M-Pro inhibitors among approved drugs by implementing a VS strategy based on consensus protein-ligand docking. At the same time, this work aims to prove the efficiency of this VS strategy.

3. Materials and methods

In order to predict inhibitors that could bind to the active site of M-pro, a receptor-based VS strategy relying mainly on protein-ligand docking was applied to two libraries of approved drugs (*i.e.,* e-Drug3D and Reaxys-marketed). The strategy was built on the premise that different docking programs use different algorithms to search binding modes and to give them a score, and that the calculation of the latter as a binding free energy solely based on docking poses tends to be insufficiently reliable, as stated above. Considering the differences in the approaches taken by different software, focusing on their intersections should compensate for the weaknesses of each [26,33]. Thus, the results of three different docking programs were compared in order to identify docked poses that were predicted simultaneously by the three of them.

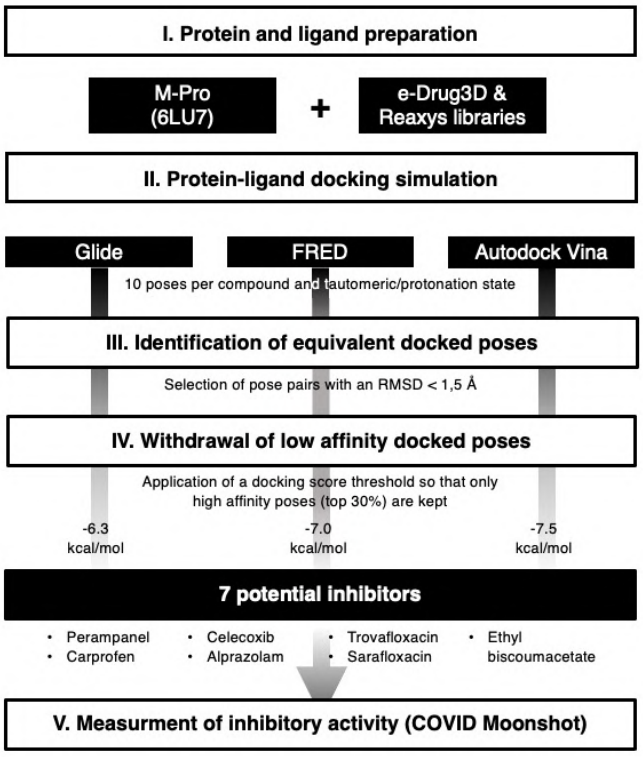


Figure 3 | Overview of the main steps of the VS strategy.

First, protein-ligand docking simulations were performed with the M-pro structure and the two libraries of approved drugs using Glide, FRED and AutoDock Vina, independently. Then, equivalent docked poses (referred as 'triplets' for simplicity) were identified among the results of the three programs. The resemblance was determined by calculating the root mean square deviation (RMSD) between docked poses, which is a measure of the average distance between the atoms of two molecules. Only those triplets of poses with an RMSD low enough to be considered as the same binding mode were kept. Also, a docking score threshold was applied to consider hits of the VS only the compounds with high affinity for M-pro. The Glide pose of each VS hit triplet was submitted to an energy minimization with the binding site of M-pro by using the MM-GBSA minimization

The validity of this protocol was assessed by checking how it performed on a sample comprising 28 experimentally known M-pro inhibitors and 1600 calculated decoys (*i.e.*, compounds that resemble actives but present low or null bioactivity for the target of interest). Finally, the hit compounds were submitted to the COVID Moonshot initiative. Thanks to this, the M-pro inhibitory activity of some of them could be validated *in vitro*.

This VS strategy is summarized in Figure 3 and explained in detail in the following sections.

3.1. Compound libraries description and preparation

The following libraries of approved compounds were used for drug repositioning purposes:

- e-Drug3D library: library of 1930 drugs and active metabolites approved by the FDA between 1939 and 2019 with a molecular weight ≤2000 Da from the database e-Drug3D [34].
- Reaxys-marketed library: library of 4536 drugs labeled as “marketed” in the field “Highest clinical phase” from the Reaxys database [35].

The following library was used as a reference to establish docking score thresholds:

- OTAVA-ML-SARS: library from OTAVA which contains 1577 compounds with predicted activity against SARS-CoV-2 based on machine learning approaches [36].

Compounds to be docked using Glide [37] and Autodock Vina [38] were prepared using the following instructions: **(1)** generate one 3D conformation per compound and discard the compounds with unspecified chiralities with Omega [39]; and **(2)** prepare the compounds for docking with LigPrep [40] by generating all the possible protonation states for each compound in the pH range 7.2±1.0 and the default number of tautomers while respecting the chiralities from the input geometry of each compound. Compounds to be docked using Fred [41] were prepared using the following instructions: **(1)** set the ionization states of the compounds with *fixpka* [42]; **(2)** enumerate tautomeric forms with *tautomer* [42]; **(3)** assign atomic partial charges with *molcharge* [42]; and **(4)** generate one conformation per compound and discard compounds with unspecified chiralities with Omega [39].

3.2. M-Pro structure preparation, grid generation and protein-ligand docking setup

The structure of M-pro in complex with the inhibitor N3 was obtained from the Protein Data Bank (PDBid 6LU7). Before its preparation for each docking software, the covalent bond with N3 was removed. Preparation was performed with different tools depending on the docking software used. During docking, only the best docked pose was generated for the reference libraries, whereas 10 poses were generated for each tautomer and protonation state of the compounds in the libraries of approved compounds used for drug repositioning purposes. In the case of dockings with Glide, the M-pro structure was prepared with Maestro [14] using Protein Preparation Wizard [43–46] with the following settings: **(a)** hydrogens were added after removal of original hydrogens; **(b)** N and C termini were capped; **(c)** disulphide bonds were created between sulphur atoms within 3.2 Å; **(d)** Epik [45] was used to generate probable tautomers and protonation states at a neutral pH; **(e)** H-bond assignment was further optimized using PROPKA [47] at a

default pH value; (f) all water molecules were removed from the structure; and (g) the structure was minimized with the default force field.

Glide was used to generate the grid around the cavity of the protein where the compounds were supposed to bind using the N3 inhibitor bound to 6LU7 as a reference. The center coordinates corresponded to the centroid of N3 (-10.36, 12.46, 68.7) and the box size was set to 35 x 35 x 35 Å. Glide was used to dock the different compound libraries to the M-pro structure by: (a) using standard-precision (SP) mode, and (b) generating 1 or 10 binding poses per compound depending on the library. In the case of dockings with FRED, MakeReceptor [42] was used to set up the receptor for docking by: (a) defining a box that enclosed the active site with its center coordinates and dimensions established based on the grid previously defined with Glide; (b) setting a shape potential; and (c) defining the inner and outer contours of the receptor with the default options. FRED 3.3.0.1. was used to dock the different compound libraries to the M-pro structure using the default settings to generate 1 or 10 binding poses per compound depending on the library. In the case of dockings with AutoDock Vina, AutoDockTools [48] from MGL Tools 1.5.6. was used to prepare the protein structure by: (a) removing all waters, and (b) adding polar hydrogens. The grid was defined as a box of the same size and center coordinates as in the other two docking programs and AutoDock Vina 1.1.2 was used to dock the different compound libraries to the M-pro structure using the default settings to generate 1 or 10 binding poses per compound depending on the library.

3.3. Identification of equivalent and high-affinity docked poses

Equivalent docked poses among the three protein-ligand docking programs were identified by comparing the root mean square deviation (RMSD) between the heavy atoms of all docked poses obtained for the same tautomeric and protonation state of each molecule. If the RMSD between all possible pairs of docked poses in each triplet was less than or equal to 1.5 Å, the docked poses obtained with the three programs were considered equivalent. The threshold value of 1.5 Å was chosen on the basis of visual inspection and general agreement in the literature [49]. The Schrödinger script *rmsd.py* [50] was used to calculate the RMSD values between poses.

The docking score thresholds chosen to select only the highest affinity equivalent poses were selected by docking the OTAVA-ML-SARS library with the same running conditions that were previously described for the approved compound libraries (with the only exception that for the OTAVA compounds only one docked pose for each tautomeric and protonation state was kept). Then, the docking score that kept only the top 30% of these compounds in each docking software was used as a threshold value to determine the minimum docking score that was regarded as indicating docked poses binding to M-pro with *high affinity*. These thresholds are -6.3, -7.0 and -7.5 Kcal/mol for Glide, FRED and AutoDock Vina, respectively (Figure 4).

Therefore, any of the equivalent docked poses of the approved compounds with docking scores more positive than their corresponding threshold were discarded (*i.e.*, poses

obtained with Glide, FRED or AutoDock Vina were compared, respectively with the Glide, FRED or AutoDock Vina thresholds). Then, only those approved compounds with a triplet of equivalent *high affinity* docked poses were considered VS hits (if more than one triplet of poses was found for the same hit, only the one that presented the highest mean docking score was chosen).

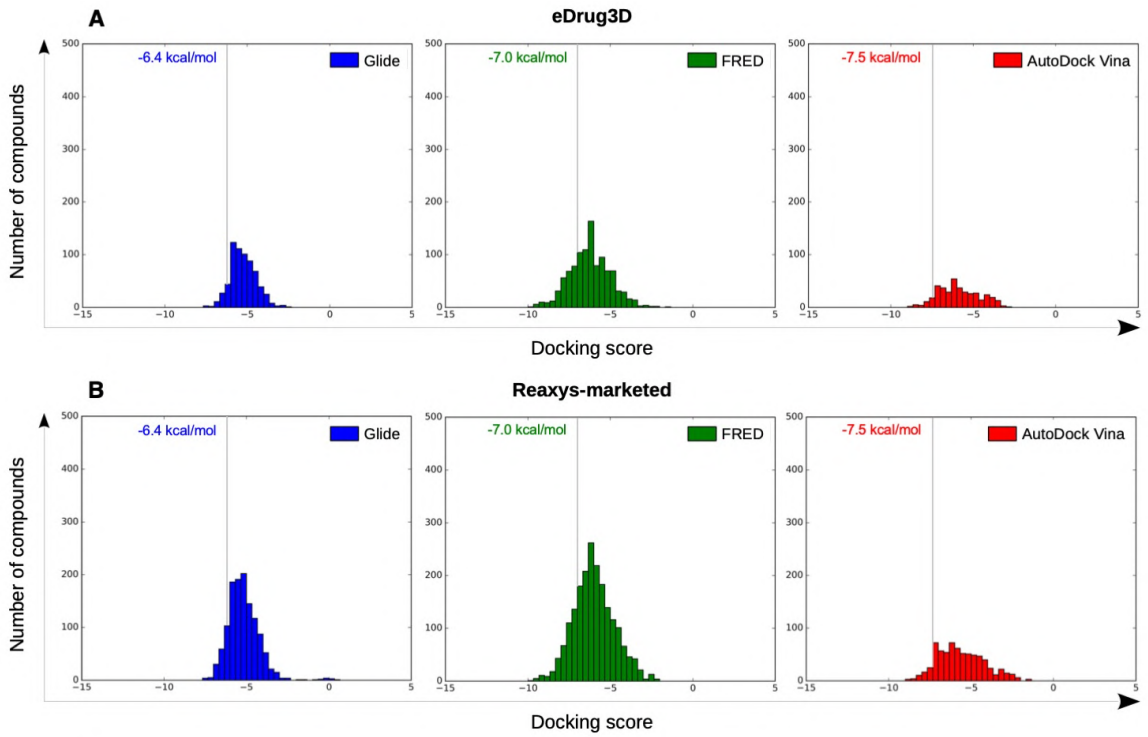


Figure 4 | Histograms corresponding to the docking scores of the two screened libraries: eDrug3D (panel **A**) and Reaxys-marketed (panel **B**). Docking score thresholds for selecting the ligands with the highest affinity are shown.

3.4. Analysis of the intermolecular interactions between M-pro and its inhibitors

The Glide pose from each VS hit triplet was submitted to the MM-GBSA methodology available in Prime [51]. This methodology calculates the binding free energies (ΔG_{bind}) from the predicted complexes obtained from ligand-docking simulations. During each MM-GBSA run, energy minimization was carried out to keep all protein residues frozen, except for a flexible region of 6 Å around the ligand. Otherwise, the remaining parameters used were the default values.

The coordinates of the protein-ligand complexes obtained after the MM-GBSA calculation were analyzed with the *poseviewer_interactions.py* script [14] to obtain the intermolecular interactions between the docked poses and the M-pro binding site. The following interactions were analyzed: hydrogen bonds (HAccep, HDonor and Ar-Hbond), halogen bonds (XBond), salt-bridge interactions (Salt), π-cation interactions (PiCat), π-π interactions (PiFace, PiEdge) and hydrophobic interactions (HPhob).

3.5. Virtual screening workflow validation

The ability of the designed VS workflow to discern between active and non-active molecules was evaluated by: **(a)** collecting from the literature all molecules with known *in vitro* activity as SARS-CoV-2 M-pro inhibitors; and **(b)** using this set of known M-pro inhibitors to obtain a set of 1600 decoys with the Generate DUD·E Decoys tool (http://dude.docking.org/generate) [30]. The docking of these two sets of molecules was performed with the same conditions that were previously described at section 3.1.2 and 10 docked poses for each program were kept per tautomeric and protonation state.

4. Results and discussion

4.1. Virtual screening of approved drugs

After performing docking simulations to the two libraries of approved drugs (*i.e.*, eDrug3D and Reaxys-marketed) and the M-pro structure in parallel with Glide, Fred and AutoDock Vina, the docked poses obtained by different programs for each compound were compared. Only those triplets where the RMSD for all comparisons was lower than 1.5 Å were kept. From these triplets of equivalent docked poses for each compound, only those with the docking scores for the three programs below the threshold were selected. Also, if more than one triplet was found for the same hit, the one with the highest docking score mean was discarded. As a result, the VS led to the identification of seven potential M-pro inhibitors: Perampanel, Carprofen, Celecoxib, Alprazolam, Trovafloxacin, Sarafloxacin and Ethyl biscoumacetate. The three equivalent docked poses and the result of the MM-GBSA minimization of the corresponding Glide pose at the M-pro binding site are shown in Figure 5. A description of the seven drugs and their predicted intermolecular interactions with M-pro is provided below and summarized in Tables 1 and 2, respectively.

Perampanel is an AMPA glutamate receptor antagonist used as an anticonvulsant to treat partial-onset seizures (Table 1). The FDA warns of severe side effects such as serious or life-threatening behavioral and psychiatric reactions [52]. Considering these adverse effects, a risk-benefit assessment should be conducted for COVID-19 patients at different stages of the disease after having confirmed its in vitro activity against M-pro.

Carprofen is a selective cyclooxygenase-2 (COX-2) inhibitor that was previously used as a pain reliever in the treatment of joint and post-surgical pain (Table 1). According to the DrugBank database [52], this drug was withdrawn from the market in 1995 on commercial grounds. Carprofen was previously used in human medicine for over 10 years, and its use in dogs is approved. Regarding its adverse effects, Carprofen was generally well tolerated with mild adverse effects, such as gastro-intestinal pain and nausea, similar to those of aspirin and other nonsteroidal anti-inflammatory drugs (NSAIDs). However, NSAIDs increase the risk of cardiotoxicity and may worsen the course of community-acquired pneumonia [53].

Celecoxib is a selective COX-2 inhibitor indicated for arthritis pain and to reduce precancerous polyps in the colon in familial adenomatous polyposis (Table 1). Although significant concerns regarding the safety of COX-2 selective NSAIDs emerged in the early 2000s, in 2005 the FDA concluded that the benefits of Celecoxib treatment outweighed the potential risks in properly selected and informed patients. However, it is not advisable to administer Celecoxib or other NSAIDs to patients with previous cardiovascular events including acute myocardial infarction, coronary revascularization, or coronary stent insertion [52], and NSAIDs may worsen the course of community-acquired pneumonia [53].

Alprazolam acts on the receptors BNZ-1 and BNZ-2 and it is used for the treatment of anxiety and panic disorders (Table 1). According to the DrugBank database [53], Alprazolam is mainly metabolized by CYP3A and, thus, its administration together with CYP3A inhibitors like ketoconazole and itraconazole is contraindicated. Its adverse effects are usually linked to the sedation it may cause.

Trovafloxacin is a broad spectrum antibiotic that inhibits DNA gyrase and topoisomerase IV (Table 1). However, according to the DrugBank database [53], this drug was withdrawn from the market due to its hepatotoxic potential.

Sarafloxacin is a quinolone antibiotic (Table 1). According to the DrugBank database [35], Sarafloxacin was discontinued by its manufacturer before receiving approval for its use in the US and Canada. Therefore, even if its M-pro inhibitory activity is confirmed, more data about its putative adverse effects would be necessary before being considered as a candidate for the treatment of COVID-19.

Ethyl biscoumacetate is a vitamin K antagonist used as an anticoagulant (Table 1). However, since according to the DrugBank database [53], Ethyl biscoumacetate was withdrawn from the market and it can produce prolonged bleeding and severe hemorrhage.

All the VS hits establish similar interactions in the binding site of M-pro (Table 2). Firstly, most of them contain a hydrophobic substructure buried mainly in the S2 subsite involving lipophilic interactions with His41, Met49, Met165, Gln189 and Asp187. In this sense, Celecoxib only interacts with Met49, probably because of the smaller size of its substituent compared to the other compounds which establish a higher number of interactions of this type. Secondly, H-bond interactions with residues from the oxyanion loop and the catalytic dyad are common in all the compounds. Perampanel, Alprazolam, Sarafloxacin and Ethyl biscoumacetate interact with Gly143 and Carprofen and Trovafloxacin interact with Ser144 and the catalytic Cys145. The other catalytic residue, His41, establishes either π-stacking interactions or H-bonds with all the compounds but Trovafloxacin. The main oxygen of His164, which is coordinated with the H_2O_{cat}, establishes an H-bond with Perampanel and Carprofen and an aromatic H-bond with Alprazolam and Sarafloxacin. Finally, Celecoxib and Perampanel establish an H-bond with Leu141, which has been reported to be a mutational “coldspot” (*i.e.*, a residue which has undergone no mutations) with implications in forming the binding site [54]. While some of these interactions are common in other crystallized complexes (*e.g.*, 13a, 13b, N3 and X77), the hydrogen bond interaction with the main chain oxygen from Leu141 in the case of Perampanel and Celecoxib or a halogen bond interaction with Cys44 in the case of Carprofen are not [27]. The fact that these interactions are shared with known crystallized inhibitors of M-pro and that they involve residues with relevant roles in the binding site structure and catalytic mechanism sustain that herein predicted compounds could be possible high-affinity effective inhibitors.

Table 1 | Main characteristics of the seven drugs predicted as SARS-CoV-2 M-pro inhibitors.

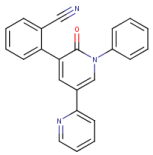
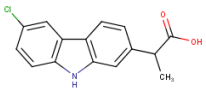
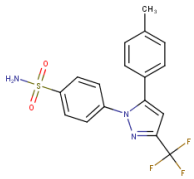
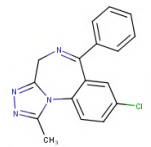
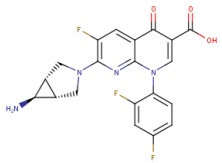
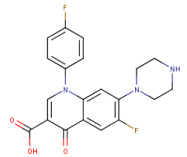
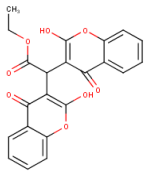
Compound	Drugbank ID, COVID Moonshot ID, % Inhibition at 50 μM (when available)	Status	Mechanism	Indication	Adverse effects
 **Perampanel**	DB08883 GER-UNI-cfb91824-1	Approved	AMPA glutamate receptor antagonist	Anticonvulsant: treatment of partial-onset seizures that may or may not occur with generalized seizures	Serious or life-threatening behavioral and psychiatric reactions
 **Carprofen**	DB00821 GER-UNI-ec786817-1 3.97 %	Approved; Withdrawn[1]	Selective cyclooxygenase-2 (COX-2) inhibitor	Pain reliever in the treatment of joint pain and postsurgical pain	Mild, such as gastro-intestinal pain and nausea, like those recorded with aspirin and other nonsteroidal anti-inflammatory drugs (NSAIDS)
 **Celecoxib**	DB00482 GER-UNI-05c7e912-1 11.89 %	Approved	Selective COX-2 inhibitor	Arthritis pain and in familial adenomatous polyposis (FAP) to reduce precancerous polyps in the colon	Like other NSAIDS it is not advisable to administer it to patients with previous cardiovascular events
 **Alprazolam**	DB00404 GER-UNI-cad0fe83-1	Approved	Acts on benzodiazepine receptors BNZ-1 and BNZ-2	Treatment of anxiety and panic disorders	Generally related to its sedative effects. Mixed with alcohol it may lead to coma and death

Table 1 | Cont.

Compound	Drugbank ID, COVID Moonshot ID, % Inhibition at 50 μM (when available)	Status	Mechanism	Indication	Adverse effects
 **Trovafloxacin**	DB00685 GER-UNI-c2851835-1	Approved; Withdrawn	Inhibition of DNA gyrase and topoisomerase IV	Broad spectrum antibiotic	It was withdrawn in 1999 due to its hepatotoxic potential.
 **Sarafloxacin**	DB11491 GER-UNI-caecb3b0-1 20.00 %	Vet approved; Withdrawn[2]		Antibiotic	
 **Ethyl biscoumacetate**	DB08794 GER-UNI-9e096ee1-1	Withdrawn	Vitamin K antagonist	Anticoagulant	It is contraindicated in conditions like myocardial infarction, liver diseases, postpartum, hypersensitivity, pregnancy, bleeding, kidney disease, breast feeding and duodenal ulcer. It can produce increased blood clotting time, prolonged bleeding and severe hemorrhage.

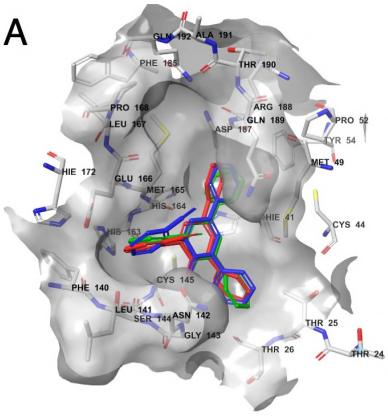
Data was obtained from DrugBank (https://www.drugbank.ca/). [1] It is no longer marketed for human usage, after being withdrawn in 1995 on commercial grounds. [2] It was discontinued in 2001 by its manufacturer, Abbott Laboratories, before receiving approval for use in the US and Canada.

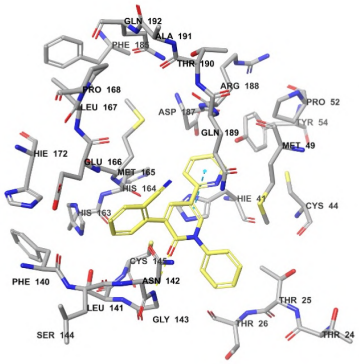
Table 2 | Summary of the intermolecular interactions between M-pro and the seven compounds identified as putative M-pro inhibitors. Interactions were obtained with the poseviewer_interactions.py script after the Glide poses and the M-pro binding site were submitted to a MM-GBSA minimization.

Subsite	Residue	Perampanel	Carprofen	Celecoxib	Alprazolam	Trovafloxacin	Sarafloxacin	Ethyl biscoumacetate
S_3	Met165	CB^{h}			CB^{h}	CB^{h}	CB^{h}	CB^{h}, SD^{h}
	Gln189		CG^{h}		CG^{h}	$NE2^{a}$, CG^{h}	CG^{h}	CG^{h}
	Thr190						O^{d}	
S_2	Met49	CE^{h}, SD^{h}	CB^{h}, CG^{h}, SD^{h}	SD^{h}	CB^{h}, CE^{h}, CG^{h}, SD^{h}	CG^{h}, SD^{h}	CB^{h}, CG^{h}, SD^{h}	SD^{h}
	His164	O^{Ar}	O^{d}		O^{Ar}		O^{Ar}	
	Asp187	CB^{h}	CB^{h}		CB^{h}		CB^{h}	
S_1	Leu141	O^{Ar}		O^{Ar}				
	Asn142			$OD1^{d}$, CB^{h}				CB^{h}
	Glu166	CB^{h}		CB^{h}	O^{Ar}	CB^{h}		
S_1'	Thr26	O^{Ar}						
	His41	CG^{p}, CB^{h}	CG^{p}, CG^{p}, CB^{h}		CG^{p}, CB^{h}		$NE2^{a}$, CB^{h}, CG^{p}	$NE2^{a}$, $CD2^{Ar}$, CB^{h}
	Gly143	N^{a}			N^{a}		N^{a}	N^{a}
	Ser144		N^{a}			N^{a}		
	Cys145		N^{a}			N^{a}		
	Cys44		SG^{x}, CB^{h}					
	Pro52		CG^{h}					

Interactions are indicated with the protein atom that is involved and refer to the role played by the ligand in the intermolecular interaction with that protein atom: ^{a}HAccep, ^{Ar}Ar-Hbond, ^{d}HDonor, ^{h}HPhob, ^{p}PiFace and ^{x}XBond.

A



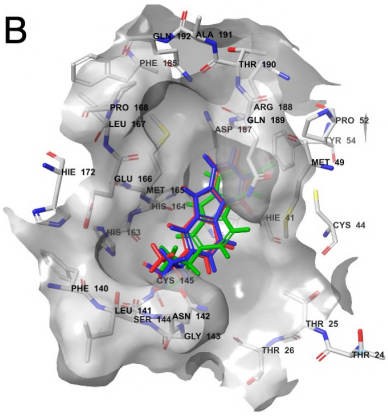


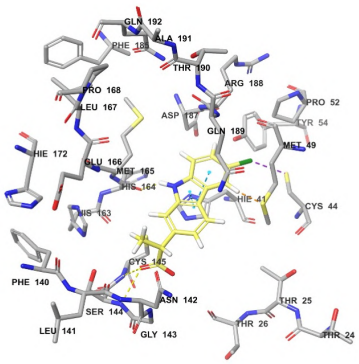
Perampanel DB08883

Docking score:

Glide	-6.9 Kcal/mol
FRED	-9.7 Kcal/mol
AutoDock Vina	-8.4 Kcal/mol
MM-GBSA ΔGbind	-63.9 Kcal/mol

B



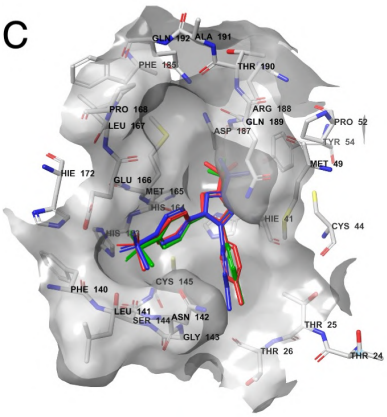


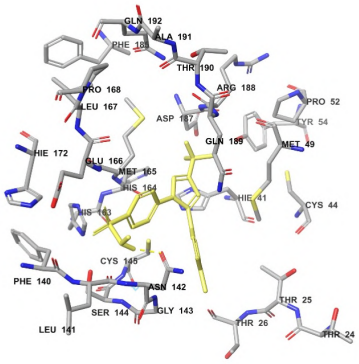
Carprofen DB00821

Docking score:

Glide	-6.7 Kcal/mol
FRED	-9.4 Kcal/mol
AutoDock Vina	-7.6 Kcal/mol
MM-GBSA ΔGbind	-37.6 Kcal/mol

C



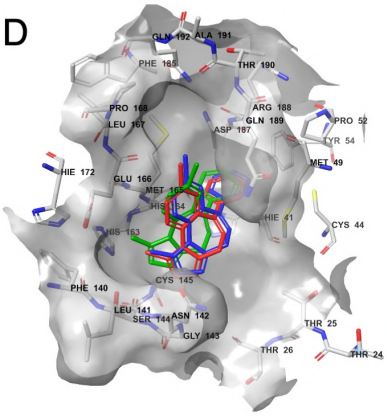


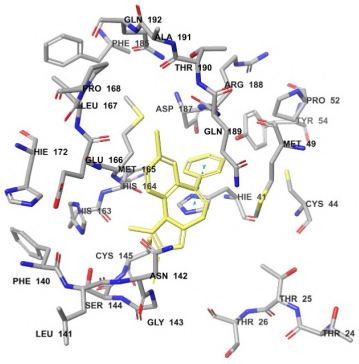
Celecoxib DB00482

Docking score:

Glide	-6.7 Kcal/mol
FRED	-8.6 Kcal/mol
AutoDock Vina	-7.8 Kcal/mol
MM-GBSA ΔGbind	-42.1 Kcal/mol

D





Alprazolam DB00404

Docking score:

Glide	-6.5 Kcal/mol
FRED	-7.8 Kcal/mol
AutoDock Vina	-7.5 Kcal/mol
MM-GBSA ΔGbind	-57.6 Kcal/mol

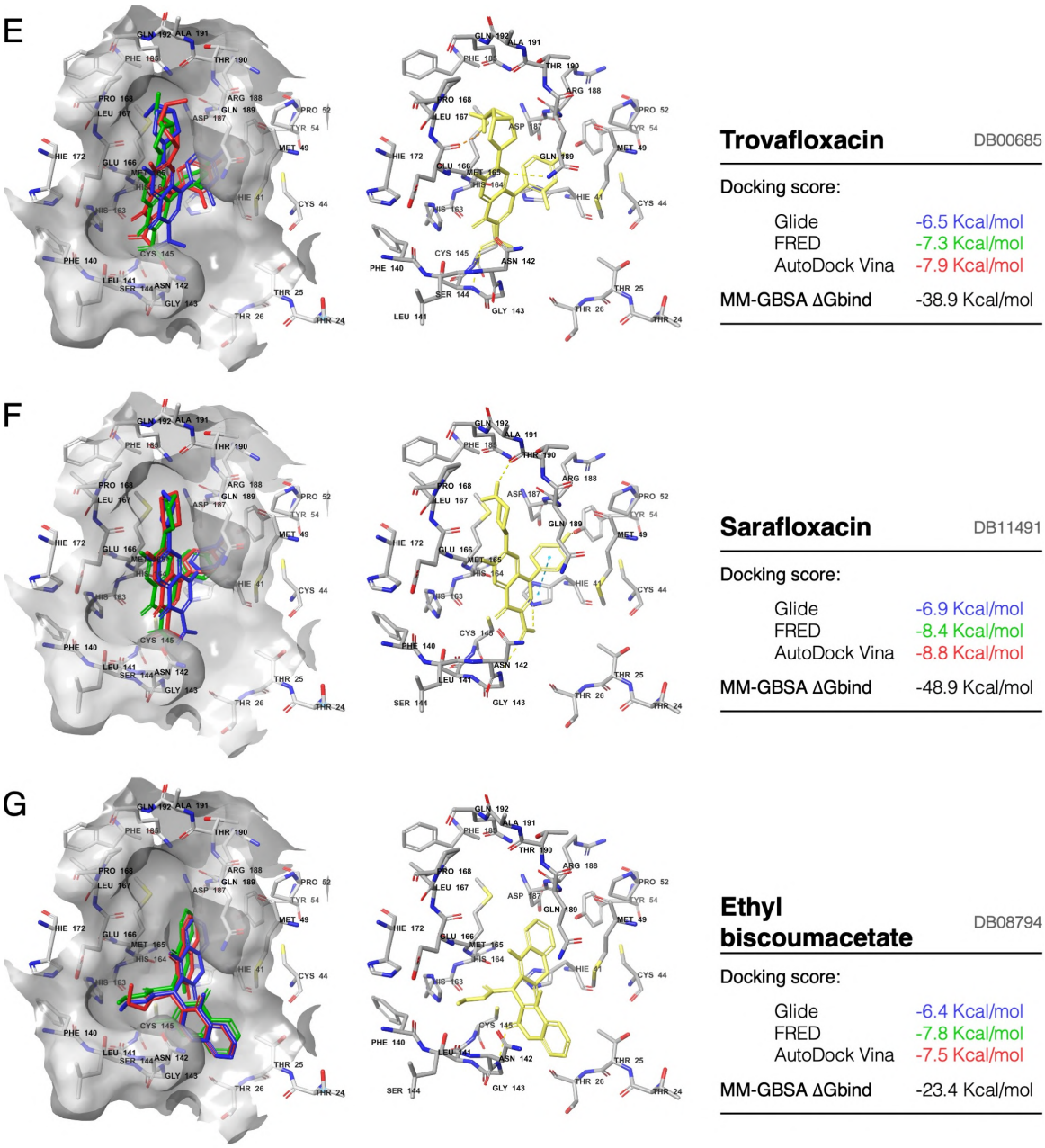


Figure 5 | On the left, panels **A-G** show the superposition of the docking poses obtained with Glide (blue), FRED (green) and AutoDock Vina (red) for each of the inhibitors. In the middle, the corresponding Glide docked pose after MM-GBSA minimization (yellow) is shown. On the right, the docking scores for the three programs and the ΔG_{bind} obtained after MM-GBSA minimization are provided. All figures were generated with Schrödinger's Maestro [14].

Regarding the adverse effects of the drugs, a risk-benefit analysis of each drug must be conducted for COVID-19 patients at different stages of the disease. The drugs with few adverse effects (i.e., Perampanel, Celecoxib and Aprazolam) could be considered for COVID-19 clinical trials as long as their activity was proved. As for the rest of drugs, despite the possible adverse effects that might be a limitation to their fast use, their experimental validation would be a set point to further research aiming at the development of more potent inhibitors taking those as lead compounds. For the moment,

the predicted inhibitory activity against M-pro of three of these drugs has been experimentally proven by the COVID Moonshot initiative. Celecoxib, Carprofen and Sarafloxacin showed a percentage of inhibition at 50 μM of 11.89%, 3.97% and 20.00%, respectively.

Interestingly, the inhibitory activity against M-pro of Perampanel has also been reported elsewhere [55], with a percentage of inhibition at 10 μM of 43 %, and an IC50 of roughly 100-250 μM. According to the authors, the IC50 could only be approximated due to the interference of the compound's fluorescence with the product of the assay. In fact, in a subsequent lead optimization study, they redesigned Perampanel to noncovalent, nonpeptidic derivates and, with the aid of free-energy perturbation calculations (FEP) obtained inhibitors with IC50 in the low nanomolar range [56]. The resemblance of the orientation in the binding site of M-pro of the docked pose obtained herein and the crystallized structure of Perampanel analogues they provide indicates the accuracy of the predictions (Figure 6). Plus, this work demonstrates the potential of predicted inhibitors with weak biological activity as possible lead compounds for the development of much more potent analogues.

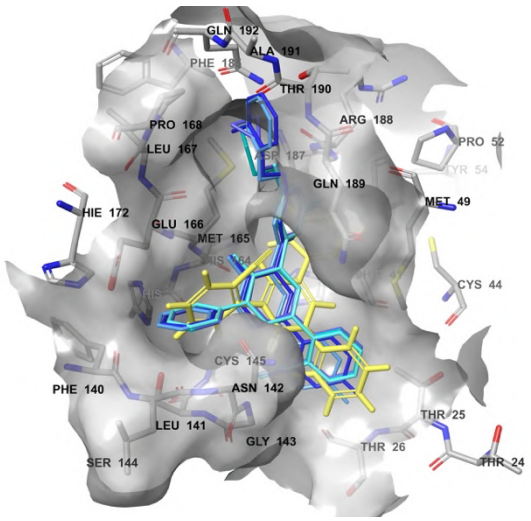


Figure 6 | Superposition of the docked pose of Perampanel (yellow) with the Perampanel analogs designed and assayed by Zhang et al. (PDBids: 7L10, 7L11, 7L12, 7L13, 7L14) (blue) [56] in the binding site of M-pro (6LU7).

4.2. Selectivity of the virtual screening workflow

The validity of the predictions that resulted from this VS workflow was assessed by evaluating it performance on a sample of 28 experimentally known M-pro inhibitors and 1600 calculated decoys. Of note, two molecules could not be docked because of problems with either Omega or Glide. However, none of the M-pro inhibitors was recovered with this set-up. Therefore, the triplets of equivalent docked poses for each inhibitor were determined by visually inspection and their docking scores were not considered. This was done under the assumption that the most reliable part of protein-ligand docking is their capacity to explore the hypothetical binding modes of a ligand, and not their scoring function. 8 out of the 26 known inhibitors (*i.e.*, 11a, Carmofur, Dipyridamole, Oxytetracycline, PX-12, Shikonin, Sulfacetamide and Tideglusib) had an equivalent triplet with upper RMSD at the [1.29-2.64] Å interval and were thus selected. Interestingly, other compounds with similar RMSD values were discarded because the docked poses were significantly different. Parallelly, a RMSD threshold of 2.5 Å was applied to the docked poses of the 1600 calculated decoys, keeping a total of 131

compounds. The enrichment factor (EF) of this VS workflow is thus 3.6. Considering that no visual inspection of the docked poses of the decoys was done, meaning the threshold was somehow permissive, the actual EF could be even higher. This validation not only proves the selectivity of the VS workflow used herein, but also shows that the conditions used were very strict and the rate of false positives will most likely be low. This is in agreement with the fact that four compounds that have been experimentally tested *in vitro* (*i.e.*, Celecoxib, Carprofen, Sarafloxacin and Perampanel) have shown M-pro inhibitory activity.

5. Conclusions

The critical situation caused by the COVID-19 pandemic is ameliorating with widespread vaccination campaigns, but we are still far from herd immunity. This scenario calls for new therapeutic options to tackle with COVID-19. Looking for these drugs among approved drugs is a shortcut to clinical trials and a sooner spread of their application among infected people. Computational approaches can speed up this search by predicting approved drugs with a greater potential to be tested *in vitro*. M-pro is a key enzyme in SARS-CoV-2 replication which is highly conserved among related viruses and has a different cleavage specificity relative to human proteases. For all this, it is being considered one of the main targets to inhibit SARS-CoV-2 replication.

In this work, a VS strategy consisting of finding compounds that are predicted to bind in the same manner and with high affinity to the binding site of M-pro by three docking programs (*i.e.*, Glide, FRED and AutoDock Vina) simultaneously. Seven drugs were predicted to inhibit M-pro: Perampanel, Carprofen, Celecoxib, Alprazolam, Trovafloxacin, Sarafloxacin and Ethyl biscoumacetate.

The performance of the VS was assessed with a set of SARS-CoV-2 inhibitors and calculated decoys, confirming its efficiency and revealing that the criteria employed was very strict and could even be relaxed. Plus, three out of the seven predicted inhibitors were proved to have inhibitory activity against M-pro *in vitro*. For these reasons, the VS workflow developed herein could be applied to other commercial databases of nonapproved drugs to predict more possible SARS-CoV-2 inhibitors, as well as to other targets of future interest.

As for the seven putative SARS-CoV-2 M-pro inhibitors, if their activity was assayed both *in vitro* and *in cellulo* and their benefits were considered to outweigh their adverse effects, they could be considered for COVID-19 clinical trials. Alternatively, if they were proved to be weak inhibitors, as it happens with Celecoxib, Carprofen, Sarafloxacin or Perampanel, they could serve as lead compounds and go through an optimization process to lead to more potent derivatives. In fact, other authors have obtained derivatives from Perampanel with IC50 in the nanomolar range, proving the potential of virtual screening hits to be used in lead optimization.

6. Acknowledgements

I would like to first express my gratitude to Dr. Gerard Pujadas and Dr. Santi Garcia-Vallvé not only for their supervision and tutoring of the present work, but also for the opportunity they gave me to join the Cheminformatics and Nutrition research group (QiN). Had not it been for them it would never have crossed my mind the idea of me working in bioinformatics. I will always be hugely grateful for this. Also, my thanks go to Dr. Aleix Gimeno for guiding me in my first steps in this field. And last, but not least, I want to thank all the members of the group for their unconditional help and support along the way. I cannot think of a better team to work with.

7. Reference list

[1] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020;579:265–9.

[2] Harrison AG, Lin T, Wang P. Mechanisms of SARS-CoV-2 Transmission and Pathogenesis. Trends Immunol 2020;41:1100–15.

[3] Nicola M, Alsafi Z, Sohrabi C, Kerwan A, Al-Jabir A, Iosifidis C, et al. The socio-economic implications of the coronavirus pandemic (COVID-19): A review. Int J Surg 2020;78:185–93.

[4] Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 2019;17:181–92.

[5] Hu B, Guo H, Zhou P, Shi ZL. Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol 2021;19:141–54.

[6] V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. Coronavirus biology and replication: implications for SARS-CoV-2. Nat Rev Microbiol 2021;19:155–70.

[7] Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 Transcriptome. Cell 2020;181:914-921.e10.

[8] Iacob S, Iacob DG. SARS-CoV-2 Treatment Approaches: Numerous Options, No Certainty for a Versatile Virus. Front Pharmacol 2020;11:1224.

[9] Cannalire R, Cerchia C, Beccari AR, Di Leva FS, Summa V. Targeting SARS-CoV-2 Proteases and Polymerase for COVID-19 Treatment: State of the Art and Future Opportunities. J Med Chem 2020.

[10] De Wit E, Van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: Recent insights into emerging coronaviruses. Nat Rev Microbiol 2016;14:523–34.

[11] Kneller DW, Phillips G, O'Neill HM, Jedrzejczak R, Stols L, Langan P, et al. Structural plasticity of SARS-CoV-2 3CL Mpro active site cavity revealed by room temperature X-ray crystallography. Nat Commun 2020;11:1–6.

[12] Lee J, Worrall LJ, Vuckovic M, Rosell FI, Gentile F, Ton AT, et al. Crystallographic structure of wild-type SARS-CoV-2 main protease acyl-enzyme intermediate with physiological C-terminal autoprocessing site. Nat Commun 2020;11:1–9.

[13] Ramos-Guzmán CA, Ruiz-Pernía JJ, Tuñón I. Unraveling the SARS-CoV-2 Main Protease Mechanism Using Multiscale Methods. ACS Catal 2020;10:12544–54.

[14] Schrödinger Release 2019-3: Maestro, Schrödinger, LLC, New York, NY, 2019

[15] Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: Progress, challenges and recommendations. Nat Rev Drug Discov 2018;18:41–58.

[16] Gimeno A, Ojeda-Montes M, Tomás-Hernández S, Cereto-Massagué A, Beltrán-Debón R, Mulero M, et al. The Light and Dark Sides of Virtual Screening: What Is There to Know? Int J Mol Sci 2019;20:1375.

[17] Pinzi L, Rastelli G. Molecular docking: Shifting paradigms in drug discovery. Int J Mol Sci 2019;20(18), 4331.

[18] Kumar A, Zhang KYJ. Hierarchical virtual screening approaches in small molecule drug discovery. Methods 2015;71:26–37.

[19] Scior T, Bender A, Tresadern G, Medina-Franco JL, Martínez-Mayorga K, Langer T, et al. Recognizing pitfalls in virtual screening: A critical review. J Chem Inf Model 2012;52:867–81.

[20] Pantsar T, Poso A. Binding affinity via docking: Fact and fiction. Molecules 2018;23(8):1899.

[21] Novič M, Tibaut T, Anderluh M, Borišek J, Tomašič T. The Comparison of Docking Search Algorithms and Scoring Functions. Methods Algorithms Mol. Docking-Based Drug Des. Discov., Hershey, USA: Medical Information Science Reference; 2016, p. 99–127.

[22] Ul-Haq Z, Madura JD. Frontiers in Computational Chemistry. Ul-Haq, Z., Madura, J.D., Eds. Volume 3. Sharjah, UAE: Bentham Science Publishers; 2017.

[23] McGann M. FRED pose prediction and virtual screening accuracy. J Chem Inf Model 2011;51:578–96.

[24] Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, et al. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 2004;47:1750–9.

[25] Torres PHM, Sodero ACR, Jofily P, Silva-Jr FP. Key Topics in Molecular Docking for Drug Design. Int J Mol Sci 2019;20:E4574.

[26] Chaput L, Mouawad L. Efficient conformational sampling and weak scoring in docking programs? Strategy of the wisdom of crowds. J Cheminform 2017;9-37.

[27] Gimeno A, Mestres-Truyol J, Ojeda-Montes MJ, Macip G, Saldivar-Espinoza B, Cereto-Massagué A, et al. Prediction of novel inhibitors of the main protease (M-pro) of SARS-CoV-2 through consensus docking and drug reposition. Int J Mol Sci 2020;21(11):3793

[28] Erickson JA, Jalaie M, Robertson DH, Lewis RA, Vieth M. Lessons in Molecular Recognition: The Effects of Ligand and Protein Flexibility on Molecular Docking Accuracy. J Med Chem 2004;47:45–55.

[29] Feher M, Williams CI. Effect of input differences on the results of docking calculations. J Chem Inf Model 2009;49:1704–14.

[30] Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. J Med Chem 2012;55:6582–94.

[31] Cereto-Massagué A, Guasch L, Valls C, Mulero M, Pujadas G, Garcia-Vallvé S. DecoyFinder: An easy-to-use python GUI application for building target-specific decoy sets. Bioinformatics 2012;28:1661–2.

[32] Achdout H, Aimon A, Bar-David E, Barr H, Ben-Shmuel A, Bennett J, et al. COVID moonshot: Open science discovery of SARS-CoV-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning. BioRxiv 2020:2020.10.29.339317.

[33] Li J, Fu A, Zhang L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. Interdiscip Sci Comput Life Sci 2019;11:320–8.

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

[35] Reaxys Database; Elsevier 2020. Available online: https://www.reaxys.com (accessed on March 16, 2020).

[36] OTAVA Chemicals. Available online: https://otavachemicals.com/targets/sars-cov-2-targeted-libraries (accessed on 16 March 2020).

[37] Schrödinger Release 2019-3: Glide, Schrödinger, LLC, New York, NY, 2019

[38] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 2010;31:455–61.

[39] OMEGA 3.1.1.2: OpenEye Scientific Software, Santa Fe, NM.

[40] Schrödinger Release 2019-3: LigPrep, Schrödinger, LLC, New York, NY, 2019

[41] OEDOCKING 3.4.0.2: OpenEye Scientific Software, Santa Fe, NM.

[42] QUACPAC 2.0.2.2: OpenEye Scientific Software, Santa Fe, NM.

[43] Schrödinger Release 2019-3: Protein Preparation Wizard, Schrödinger, LLC, New York, NY, USA, 2019

[44] Schrödinger Release 2019-3: Impact, Schrödinger, LLC, New York, NY, USA, 2019

[45] Schrödinger Release 2019-3: Epik, Schrödinger, LLC, New York, NY, USA, 2019

[46] Schrödinger Release 2019-3: Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2019

[47] Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. J Chem Theory Comput 2011;7:525–37.

[48] Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem 2009;30:2785–91.

[49] Hevener KE, Zhao W, Ball DM, Babaoglu K, Qi J, White SW, et al. Validation of molecular docking programs for virtual screening against dihydropteroate synthase. J Chem Inf Model 2009;49:444–60.

[50] Schrödinger release 2019-3, Schrödinger, LLC. New York, NY, USA, 2019

[51] Schrödinger Release 2019-3: Prime, Schrödinger, LLC, New York, NY, USA, 2019

[52] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46:D1074–82.

[53] Basille D, Plouvier N, Trouve C, Duhaut P, Andrejak C, Jounieaux V. Non-steroidal Anti-inflammatory Drugs may Worsen the Course of Community-Acquired Pneumonia: A Cohort Study. Lung 2017;195:201–8.

[54] Krishnamoorthy N, Fakhro K. Identification of mutation resistance coldspots for targeting the SARS-CoV2 main protease. IUBMB Life 2021;73:670–5.

[55] Ghahremanpour MM, Tirado-Rives J, Deshmukh M, Ippolito JA, Zhang CH, Cabeza De Vaca I, et al. Identification of 14 Known Drugs as Inhibitors of the Main Protease of SARS-CoV-2. ACS Med Chem Lett 2020;11:2526–33.

[56] Zhang CH, Stone EA, Deshmukh M, Ippolito JA, Ghahremanpour MM, Tirado-Rives J, et al. Potent Noncovalent Inhibitors of the Main Protease of SARS-CoV-2 from Molecular Sculpting of the Drug Perampanel Guided by Free Energy Perturbation Calculations. ACS Cent Sci 2021;7(3):467-75.

8. Self-assessment

When I joined the Cheminformatics and Nutrition group, bioinformatics was a totally new research field for me, and the project I had to work in, a challenge. In fact, the first thing I learnt was to work autonomously and overcome the problems which emerged at each step. Now, after two years of experience in the group, I can say I have learnt from scratch to use multiple in silico tools, such as different docking programs, to manage files and data as well as to program in R. Needless to say, with this work I have managed to improve my critical thinking, methodological organization and scientific writing.

Parallelly, I have had the chance to actively take part in the tasks carried out in a research group; that is, protocol establishment, collaborations with other groups or article publishing. Not to mention, the fact of working as part of a team, making suggestions and contribution, but also being questioned. I believe that these aspects are as important as the bioinformatical knowledge and can be totally extrapolated to any other research field.

All in all, the experience and knowledge I have got from the present work –and, in general, from the time in the group– have been indeed rewarding both in a professional and personal level.