Multitarget virtual screening for drug repurposing in COVID19

C.O.S. Sorzano[1], E. Crisman[2], J.M. Carazo[1], R. León[2]

July 14, 2020

[(1)]Biocomputing Unit, National Center for Biotechnology (CSIC), c/Darwin, 3, Campus Universidad Aut'onoma, 28049 Cantoblanco, Madrid, Spain.
[(2)]Hospital de La Princesa, c/Diego de León, 62, 28006, Madrid, Spain

Abstract

Therapeutic or preventive research for coronavirus SARS-CoV2 is an extremely active topic of research since its outbreak in January 2020. In this paper we report the results from a virtual drug screening analysis that, to the best of our knowledge, is the widest work in terms of target proteins and compound library. Our study was focused on the repurposing of currently commercialized drugs, and especially those that can interact with multiple viral proteins and several binding sites within each protein. Additionally, we performed a second virtual screening analysis in which we compared our results to the predicted binding affinities for the drugs currently in clinical trials. We show that the best molecules in our screening compares favourably to those in clinical trials, suggesting their suitability for therapeutic or preventive applications.

1 Introduction

Coronavirus SARS-CoV-2 has largely affected our society since Jan. 2020, and many scientific efforts have been quickly reoriented to find therapeutic and preventive treatments to diminish its impact. Structural biology and computational drug screening can accelerate the search for drugs by elucidating the three-dimensional structure of the viral proteins (Structural Biology) and predicting possible compounds or compound fragments that could interact with these proteins (computational drug screening). Drug repurposing is especially interesting in this endeavour as their safety and secondary effects are well characterized for other applications.

Virtual screening is typically performed by analyzing the catalytic site of a protein of interest and looking for drug fragments that interact with it. These drug fragments can then be appropriately linked and their ADME properties

optimized. This process is long, although rewarding as it may ultimately lead to a commercial product from which the company may benefit. It is more difficult to predict the effectiveness of a binding in an allosteric site. Moreover, because it is difficult to determine the relationship between the allosteric site and the catalytic site.

Most virtual screening works so far has followed an approach similar to the one described above. Only that instead of looking for new molecules, they search for the interaction of existing molecules. However, they still concentrate in one or two proteins, and one or two sites within these proteins (Chen *et al.*, 2020; Contini, 2020; Elmezayen *et al.*, 2020; Jin *et al.*, 2020; Kandeel and Al-Nazawi, 2020; Mittal *et al.*, 2020; Muralidharan *et al.*, 2020; Sekhar, 2020; Smith and Smith, 2020; Ton *et al.*, 2020; Wang, 2020; Xu *et al.*, 2020). Additionally, to speed up the process, they have constructed homology models for the target viral proteins instead of having to wait for the experimental determination of these targets (Arya *et al.*, 2020; Chen *et al.*, 2020; Contini, 2020; Elfiky, 2020; Elmezayen *et al.*, 2020; Sekhar, 2020; Smith and Smith, 2020; Wang, 2020; Xu *et al.*, 2020). The 3CL protease has been the main target of previous works. This is a very important protein for the virus replication. However, hindering the virus cycle at any of its steps would reduce the infection speed, and eventually even stop it by allowing the immune system to have time to fight the virus.

At present there are over 100 atomic models of the virus proteins at the Protein Data Bank (Rose *et al.*, 2011). Unfortunately, there is not an atomic model for all viral proteins and complexes. And, on the other side, many of these 100 models are partially redundant as they describe the same protein or complex as solved by several structural groups.

One of the main problems with a treatment targeting a single viral protein is that the virus may mutate and become immune to the treatment. If a new treatment is found and the virus mutates again, a multirresistant virus could be selected. An alternative approach is to find a multidrug treatment that blocks the virus functions at several points along its cycle, making less likely that the virus successfully mutates at several places in order to escape the treatment. Analogously, a single compound that could interact with the virus at several places would be at an advantageous position as it also makes the resistance raising problem more difficult. Additionally, the prediction problem of allosteric sites is of less importance in this virus blocking application: if we find a molecule that interacts at several places of a target protein, we can foresee that it will hinder its normal function even if we are not able to accurately predict its effect on the catalytic site.

One of the main criticisms to virtual screening is that it does not accurately reproduce reality as experimentally measured binding affinities are normally different from those computationally predicted. Notwithstanding, computational means have been found to be rather meritorious is some tasks (an accuracy of 86% in the identification of binding sites (Halgren, 2009), or hit success between 1 and 40% for molecules at the top of the list (Slater and Kontoyianni, 2019)). This "bad quality" results are also found in other scientific domains. Specifically, in data analysis, many weak classifiers can be gathered into a strong

classifier (Polikar, 2007). Assume that we have to classify objects into class A or B (for the sake of argument, let us assume that both kinds of objects are equally likely). We know how to construct classifiers whose classification accuracy is only 51%, they are just 1% better than a random classification. However, if we construct 20,000 of these weak classifiers, we may accurately determine the class of the object simply by evaluating the difference in the number of votes for classes A and B. This is at the basis of some of the state-of-art classifiers as random forests (Breiman *et al.*, 1993; Breiman, 2001) or other ensemble classifiers (Bishop, 2006). Note that this extreme case of such a weak classifier has been set only as an example, and that virtual screening is in a much better situation.

Joining the specific application of viral cycle blocking (in which as many bindings as possible are preferred) and following an approach similar to the one of ensemble clustering (in which algorithmic errors are compensated by multiple predictions), we have devised a drug repurposing strategy in which, for all viral proteins whose 3D structure is experimentally determined, we computationally find up to 10 possible binding sites. Then, we look for small molecules that are predicted to bind these sites. Finally, we sort the molecules according to these two criteria: binding affinity and number of binding sites predicted by molecule.

This multitarget approach is, in a way, similar to that followed by Zhou *et al.* (2020). In both approaches, we use experimentally determined structures for the viral proteins and several atomic models for each one of them. However, there are some important differences: 1) (Zhou *et al.*, 2020) analyzed 7 viral proteins and 1 binding pocket per target, this binding pocket was manually selected; we analyzed 10 proteins, and almost 300 binding sites predicted by Schrödinger `sitemap` (one of the most established computer programs for this task); 2) (Zhou *et al.*, 2020) used Rosetta GAligandDock (an unpublished docking protocol in Rosetta) for the ligand docking, while we used Schrödinger `glide`, that has over 3,000 citations (this change in the program may explain the differences in the predicted binding affinities); 3) they explored a wide range of compounds from DrugBank, while we have focused in compounds approved by the FDA and other world drug agencies, and we have given especial attention to already cheap and commercialized drugs that could immediately be used for a clinical trial; 4) we have explicitly looked for drugs that bind several proteins simultaneously (rather than they bind to the different atomic models of the same protein as in Zhou *et al.* (2020)).

From the previous analysis several molecules outstand from the rest of compounds. We then compared these molecules in a second virtual screening analysis with those currently in clinical trials for COVID19. We show that the compounds found in the first analysis compare favourably to the ones in clinical trials. We also draw some interesting results applicable to the compounds in clinical trials. Finally, we report a list of compounds that are widely available, cheap and have the potential of binding with high affinity to many of the viral proteins.

2 Methods

The following process has been conducted in our workflow engine Scipion (de la Rosa-Trevín *et al.*, 2016), through an extension that we have recently performed. The whole process required more than 2,000 Scipion protocols, i.e., the largest Scipion project ever performed, and some of the sets involved more than 9 million entries. Without the help of this workflow engine, it would have been very difficult to carry out this massive analysis.

Target reparation

We analyzed all viral proteins for which an experimentally determined three-dimensional structure was known. For each of the target molecules, we analyzed several of its atomic models, with a total of 30 atomic structures being analyzed. Table 1 shows the list of PDB codes for each one of the studied viral proteins. S2 is a protein of the virus spike. The spike contains the Receptor Binding Domain (RBD), which interacts with the Angiotensin Converter Enzyme 2 during cell infection. Non-structural proteins 10 and 16 make the viral RNA to become unnoticed by the cell foreign RNA recognition system, NSP3 is responsible for cutting and untagging viral proteins, NSP9 helps to construct pores in the cell nucleus membrane, NSP15 is an endoribonuclease in charge of removing viral RNA, NSP12 is an RNA polymerase, and the complex NSP12&7&8 is the RNA replication complex. The virus cycle and possible points to interfere with drugs are reviewed in Guy *et al.* (2020).

Viral protein	**PDB codes**
Spike	6VSB, 6VXX, 6VYB
S2	6LXT
RBD/ACE2	6M17, 6M0J, 6LZG
RBD	6W41
NSP10 & 16	6W75, 6WJT
NSP3	6W6Y, 6VXS, 6W02, 6W01
NSP9	6W9Q, 6W4B
Papain-like protease	6LU7, 6W9C, 6YB7, 7BQY, 6M03, 6M2Q, 6Y84
NSP15	6VWW
Nucleocapsid	6VYO, 6M3M
NSP12	7BTF, 6M71
NSP12 & 7 & 8	7BV1, 7BV2

Table 1: Viral proteins and PDB codes used for virtual screening. RBD: Receptor Binding Domain. ACE2: Angiotensin Converter Enzyme 2. NSP: Non-structural protein

For each target molecule, we called Schrödinger `prepwizard` for a pH of 7 and a pH range of 2. We asked the program to fill side chains and loops, create bonds between proximal sulfurs, convert the selenomethionines to methionines,

delete and re-add hydrogens, add the cap termini, remove water molecules, and organic molecules used for co-crystallization. We computed the network of hydrogen bonds and the protonation states with `propka` (Olsson *et al.*, 2011). Finally, we performed a restrained minimization using the OPLS force field 3 with an RMSD convergence threshold of 0.3 Å.

For each one of the prepared targets, we have predicted at most 10 binding sites using Schrödinger `sitemap` program. This program has an accuracy of 86% in the identification of binding sites (Halgren, 2009). We removed from the analysis those binding sites whose binding score was smaller than 0.8. Around each binding site, we prepared a search grid with an inner box of 10Å and an outer box of 30Å.

We repeated the same process for each of the domains of these 30 proteins. In total, we analyzed 297 binding pockets whose binding score was above 0.8. Fig. 1 shows the numerical description made by Schrödinger `sitemap` for this set of binding pockets.

Compound library preparation

For the compound library in the first screening we have taken all compounds approved by the FDA and other world drug agencies as listed by the ZINC15 database (FDA and world-non-FDA) (Sterling and Irwin, 2015). This library comprises 5,905 compounds. We then prepared these compounds using Schrödinger `ligprep`. We used Epik for the generation of the ligand ionization (pH of 7 and an pH range of 2) and tautomeric states. We used OPLS3e force field for the ligand preparation. This process may duplicate some of the compounds depending on possible ionization states along the pH range. This step expanded the 5,905 input compounds to 9,723 derived compounds.

For the compound library in the second screening, we have taken all compounds currently in clinical trials as listed in ClinicalTrials.gov, the set of the top scoring compounds in our first screening, and a random subset of the ZINC FDA and world-non-FDA catalogs of the same size as the other two subsets (clinical trials plus top scoring in the first screening).

1st virtual screening

For each binding site of each target, we have performed a two-level docking prediction using Schrödinger `glide`. In both stages we have used a flexible docking (`confgen`, Watts *et al.* (2010)). In the first stage we used a low docking precision (HTVS). For each compound and site comparison, we allowed 5,000 initial poses per ligand and kept up to 5 of the best poses. We chose 2% of the poses coming out of the first docking analysis and removed the redundancy (several poses) keeping only 1 pose per ligand. With the surviving set of ligands, we performed a second docking stage with a more precise docking model (SP). The number of initial poses and kept poses were the same as for the first docking stage. We kept the best 10% of the poses sorted by binding affinity. Then, for each protein target we gathered the results of all the second docking rounds

in each one of its binding sites. From this gathering, we kept 20% of the best poses.

2nd virtual screening

The goal of the 2nd virtual screening was to compare the results of the compounds found in the first virtual screening with respect to the results predicted for compounds already in clinical trials against COVID19. The compound library of this virtual screening was much smaller and made of three subsets: 1) top compounds found in our first screening, 2) compounds in clinical trials, 3) random subset of the ZINC FDA and world-non-FDA catalogs with the same size as subsets 1 and 2. This third subset will give a background distribution for compounds that, in general, are not expected to especially interact with the viral proteins.

For the second screening we used the more precise docking model (SP) and there was no filtering of poses by docking score so that we can see the full distribution of predicted values for all compounds in the compound library. The rest of parameters remained equal to those used in the first screening.

3 Results

1st virtual screening

Table 2 shows a summary of the results coming out the massive virtual screening with multiple viral targets and multiple structures per target. We report on the number of unique compounds found to interact with the different proteins, the average number of binding sites per structure, the average number of compounds per site and the range of affinity bindings and ligand efficiencies. We remind that for each target protein we also predict binding sites for each of its individual domains so that, even if we ask the program a maximum of 10 possible binding sites, the average number per structure can be a bit higher.

From this analysis, we can already learn some interesting results:

- There is a wide range of possible binding sites among the viral proteins (average number of binding sites per structure). The ones more difficult to attack are the one forming nucleus pores (NSP9), the nucleocapsid, and the two proteases (NSP3 and papain-like protease). On the other side, the ones with more possible binding sites are the endoribonuclease (NSP12 & 7 & 8), the methyltransferase complex (NSP10 & 16), and the spike.

- The two proteases have very few possible binding sites (2-3), but very promiscuous, capable of interacting with many different partners (average number of unique compounds per site above 100). This is, probably, a consequence of their physiological function of having to interact with many different protein substrates.

It is remarkable that most previous studies on drug repurposing has focused on the papain-like protease, instead of the more accessible spike or receptor binding proteins. This massive study may suggest new research directions for therapeutic and preventive treatments.

Viral protein	**#Unique compounds**	**#Binding sites/Structure** (Volume [Å^3])	**#Unique compounds /site**	**Affinity binding [kcal/mol]**	**Ligand efficiency [kcal/mol]**
Spike	786	10.3 (15,300)	76.3	[-11.34, -7.57]	[-0.88,-0.11]
S2	97	3.4 (1,100)	28.5	[-10.78, -8.56]	[-1.27, -0.12]
RBD/ACE2	409	7.2 (7,500)	56.8	[-10.75, -7.54]	[-1.11, -0.12]
RBD	211	6.8 (2,000)	31.0	[-9.12, -6.93]	[-1.02, -0.10]
NSP10 & 16	423	10.4 (2,500)	40.7	[-9.97, -7.03]	[-1.23, -0.11]
NSP3	350	2.7 (1,100)	129.6	[-12.96, -7.18]	[-1.01, -0.11]
NSP9	98	1.5 (200)	65.3	[-8.63, -6.29]	[-0.80, -0.09]
Papain-like protease	413	2.1 (3,700)	196.7	[-10.88, -6.41]	[-0.81, -0.09]
NSP15	213	6.0 (1,700)	35.5	[-10.66, -7.27]	[-0.92, -0.11]
Nucleocapsid	147	2.0 (2,300)	73.5	[-10.91, -8.49]	[-1.04, -0.13]
NSP12	475	11.2 (3,400)	42.4	[-9.55, -7.00]	[-0.91, -0.11]
NSP12 & 7 & 8	510	13.9 (2,700)	36.7	[-11.14, -7.00]	[-0.99, -0.10]

Table 2: Summary of the virtual screening performed on several structures of the viral proteins. We report the number of unique compounds found to interact with the macromolecule, the average number of binding sites per PDB structure (and the total volume, on average, for all sites in the target), the average number of unique compounds per binding site, and the range of affinity bindings and ligand efficiencies.

The first screening resulted in 12,503 poses of 2,031 compounds, distributed along the 297 possible binding sites of the 30 target structures. Suppl. Fig. 2 shows the docking score (binding affinity) and ligand efficiency of these poses. We also analyzed the number of poses for each one of the compounds (remind that each compound may occupy up to 5 different poses in the same binding site depending on its binding affinity). The most frequently observed value is 1 pose, the median is 3, and 95% of the compounds occupy less than 22 poses. However, 5% of the compounds have more than 22 poses, and 1% more than 63. As argued in the introduction, this set of molecules with many binding sites are the most interesting ones.

This virtual screening produces a formidable amount of data (the raw results are available as Suppl. Material). In order to combine the two requirements (high affinity and high number of binding sites) we summarize the output data by taking the best binding affinity for each pair compound-site. Then, for each pair compound and target protein, we added all the binding affinities. Table 3 shows these aggregated values as well as the number of sites occupied by each compound. In this table we show a cephalosporin and a quinolone. However, these families of molecules systematically appeared interfering with the NSP10

& 16 or NSP12 & 7 & 8 complexes (the full results are available as Suppl. Material).

2nd virtual screening

The second virtual screening aims at comparing the top drugs coming out from the first analysis (14 compounds) and the ones currently in clinical trials (159 compounds; both together they are called the treatment group and it is formed by 173 compounds) with respect to a random subset of compounds approved by the FDA or world-non-FDA agencies (we refer to this random subset as the control group and it is also made of 173 compounds). For each compound we kept the best 5 poses in each one of the possible binding sites. Overall, the analysis resulted in 418,605 poses (that are the best 5 poses of all the ionized and tautomerized versions of the compounds in all studied binding sites). Fig. 3 shows the binding affinity distribution for both groups. As can be seen, both distributions significantly overlap. However, a two-sample Kolmogorov-Smirnov test gives a p-value of 0.006671. The median values for both distributions are -8.11 (treatment) and -7.70 (control). Actually, the most interesting part, from a therapeutic perspective, of this distribution is the region below -8 kcal/mol. This small difference between both distributions is enlarged if the bottom 1% poses are considered (-11.44 for the treatment group, -10.33 for the control group).

Only 39 pairs of treatment compound-binding site had a better binding affinity than the 1% percentile of the control group. Among these pairs, we found (in parentheses we have given the structure it fits in and the binding pocket number assigned during our analysis):

- From our first virtual screening: FAD (14 poses, 1 binding site of NSP3 ($6W02_1$), 1 of S2 ($6LXT_2$)), NADH (8 poses, 1 binding site of NSP3 ($6W02_1$)), Coenzyme A (5 poses, 1 binding site of NSP3 ($6W02_1$)), Cangrelor (7 poses, 1 binding site of NSP3 ($6W02_1$)), Isovuconazole (3 poses, 1 binding site of NSP3 ($6W02_1$)), Integrilin (2 poses, 1 binding sites of NSP3 ($6W02_1$) and the Spike ($6VSB_1$)).

- From the compounds in clinical trials: None

The most prominent family among the compounds in clinical trials are those derived from heparin, and these have already proved to be effective in the treatment of COVID19 (Thachil, 2020). It could be because 1) the drug interacts with the virus, or 2) because the drug ameliorates or prevents some of the lesions caused by the virus. In this work we have shown that the first hypothesis is true (which does not mean that the second is not also true). It is interesting that one of the outstanding molecules in the clinical trial pipeline is a cephalosporine (ceftaroline) confirming the systematic appearance of this family of molecules in our first virtual screening.

It is also very interesting that all the clinical trials listed above are addressing just 2 binding sites of a single complex (the RNA polymerase).

Remdesivir has already been shown to shorten the recovery time in groups of COVID19 patients (Grein *et al.*, 2020). In our study it appears as one of the drugs with a medium potency in its iteraction with the virus (best binding affinity -8.68 kcal/mol). Hydroxychloroquine has already been reported to have a positive effect in-vitro although its clinical application has been abandoned due to its seemingly increased risk of death. Its highest binding affinity in all possible target sites is -7.39 kcal/mol. For the sake of comparison, from now one we will only count those interactions with a binding affinity better than -8 kcal/mol. Remdesivir has 3 binding sites above this level on 3 different proteins.

Above the level of remdesivir in the set of clinical trials, we have (best binding affinity [kcal/mol], BS=binding site): Umifenovir (-8.01, 1 BS), Clavulanic acid (-8.09, 1 BS), Captopril (-8.20, 1 BS), Amoxicillin (-8.51, 2 BS), Piperacillin (-8.56, 2 BS), Atorvastatin (-8.7, 1 BS), Linagliptin (-8.73, 2 BS), Nintedanib (-8.77, 1 BS), Omeprazole (-8.80, 4 BS), Tinzaparin (-8.92, 1 BS), Heparin (-8.92, 1 BS), Curcumin (-9.12, 3 BS), Anakinra (-8.55, 1 BS), Folic acid (-9.30, 3 BS), VazegepantBHV3500 (-9.79, 7 BS), Telmisartan (-9.85, 6 BS), Metenkefalin (-9.90, 12 BS), Deferoxamine (-9.90, 2 BS), Famotidine (-9.90, 3 BS), Isoprinosine (-9.97, 3 BS), AT001 (-10.14, 2 BS), Cobiscitat (-10.17, 11 BS), Methotrexate (-10.27, 7 BS), Defibrotide (-10.29, 2 BS), IMU838 (-10.44, 3 BS), Danoprevir (-10.46, 3 BS), Doxycycline (-11.44, 2 BS), Ceftaroline (-11.45, 6 BS). These are 28 out of the 159 clinical trials (18%).

Above the level of remdesivir in our first virtual screening: Plazomycin (-9.23, 14 BS), Neomycin (-10.06, 14 BS), Atosiban (-10.22, 24 BS), Capreomycin (-10.15, 15 BS), Isovuconazole (-10.45, 23 BS), Cefoporazone (-10.88, 6 BS), Ceftolozane (-10.93, 6 BS), Cangrelor (-10.94, 22 BS), Integrilin (-11.02, 8 BS), Coenzyme A (-11.36, 13), NADH (-12.44, 8), FAD (-12.96, 12).

From the top compounds (both in number of binding sites and binding affinities) of the second virtual screening, we narrowed the list to those commercial compounds whose price in the public market is smaller than 15 euros and they interact with at least 3 binding sites (we removed heparin as it caused Schrödinger `glide` to fail). For these compounds, we run another round of virtual screening, this time with the XP accuracy (the most accurate analysis of `glide`). Table 4 shows the aggragated binding affinities. We find a few antibiotics (neomycin, amoxicillin+clavulanic acid), non-steroid antiiflamatories (indomethacine), antihypertensives (telmisartan, ramipril, and atorvastatin), and a proton-pump inhibitor (omeprazole). There are also some rather common compounds (folic acid, coenzyme A, NADH, and FAD) that are considered nutritional supplements (FAD is not commercialized as a supplement). Coenzyme A, NADH and FAD can be derived from their vitamin precursors, B_5, B_3, and B_2, respectively.

4 Conclusions

After several rounds of screening with different accuracy models for the binding affinity (from least to most accurate), a number of compounds has consistently outperformed the rest. All the compounds in the first virtual screening

were among the best matching compounds in the second round. The fact that many of the compounds currently in clinical trials were also among the top compounds showed the capacity of the virtual screening to identify interesting compounds (compounds currently in clinical trials have been selected because they have already shown some positive features either in *in vitro* studies or with patients). Very interestingly, some of the compounds in clinical trials (Amoxicillin+Clavulanic acid, Heparin, Indomethacin, Telmisartan, Ramipril, and Atorvastatin) have high binding affinities (¡-8 kcal/mol) to several binding sites and several target proteins. This fact casts a positive forecast on these clinical trials. On the other side, some of the other clinical trials seem to be targeting a single protein (Levofloxacin, Naproxen, and Omeprazole).

Among the most prominent compounds, we have detected Neomycin that is a very cheap antibiotic widely available and produced by many independent companies, which guarantees its supply. The fact that it is an antibiotic may help the acute pneumonia caused by the virus. If we need to tackle hypertension or inflammatory response, we have also identified that some of the compounds currently in clinical trials (telmisartan, ramipril, and atorvastatin, for the hypertension; and naproxene and indomethacine, for the inflammation) seem to interact with the viral proteins, which might be a help in order to reduce, not only the virus effects, but also its infection.

We have also identified FAD, NADH+NA and Coenzyme A, and to a smaller extent, folic acid. These compounds largely outperform the drugs currently in clinical trials in terms of binding affinity to viral proteins and variety and potency of binding sites. They are also found in the normal metabolism of cells. NADH, Coenzyme A, and folic acid are sold as nutritional supplements. If not in their pure form, their precursors are Vitamins B2 (FAD), B3 (NADH+NA) and B5 (Coenzyme A), which are also normally sold as nutritional complements. We are conscious that these compounds can be taken up by any other cell in the body. However, globally raising their concentration, and given their high affinity for the viral proteins, could be used as a preventive treatment in healthy patients. Their use in COVID19 affected patients may also slow down the growth rate of the virus.

Some of the compounds coming out from the first screening are difficult to handle in a widely adopted manner. For instance, Cangrelor has the highest number of binding sites and high affinity. However, Cangrelor is a rather expensive drug, that is continually perfused in a hospital environment (in part because its pharmacokinetic half-life is between 3 and 6 minutes). Similar comments on the price, ease of wide application, pharmaceutical form, or diversity of supply could be made on other drugs in the list of drugs in clinical trials or drugs in the outcome of the first screening.

In this paper we have taken the position of looking for compounds that are cheap, widely available, with low toxicity (neomycin is widely used at the moment), and binding to many sites and many viral proteins. Although, there is no guarantee of success, this fact is very suggestive as a predictor of a possible success in a clinical trial.

Funding

We thank financial support from the Spanish Ministry of Economy and Competitiveness through Grants BIO2016-76400-R(AEI/FEDER, UE), the "Comunidad Autónoma de Madrid" through Grant: S2017/BMD-3817, Instituto de Salud Carlos III, PT17/0009/0010 (ISCIII-SGEFI/ERDF), European Union (EU) and Horizon 2020 through grants: CORBEL (INFRADEV-1-2014-1, Proposal: 654248), INSTRUCT-ULTRA (INFRADEV-03-2016-2017, Proposal: 731005), EOSC Life (INFRAEOSC-04-2018, Proposal: 824087), HighResCells (ERC-2018-SyG, Proposal: 810057), IMpaCT (WIDESPREAD-03-2018 - Proposal: 857203), EOSC-Synergy (EINFRA-EOSC-5, Proposal: 857647), and iNEXT-Discovery (Proposal: 871037). The authors acknowledge the support and the use of resources of Instruct, a Landmark ESFRI project.

References

Arya, R. *et al.* (2020). Potential inhibitors against papain-like protease of novel coronavirus (covid-19) from fda approved drugs.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

Breiman, L. *et al.* (1993). *Classification and regression trees*. Chapman & Hall/CRC.

Chen, Y. W. *et al.* (2020). Prediction of the sars-cov-2 (2019-ncov) 3c-like protease (3cl pro) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Research*, **9**.

Contini, A. (2020). Virtual screening of an fda approved drugs database on two covid-19 coronavirus proteins. *ChemRxiv*.

de la Rosa-Trevín, J. M. *et al.* (2016). Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. *J. Structural Biology*, **195**, 93–99.

Elfiky, A. A. (2020). Anti-hcv, nucleotide inhibitors, repurposing against covid-19. *Life sciences*, page 117477.

Elmezayen, A. D. *et al.* (2020). Drug repurposing for coronavirus (covid-19): in silico screening of known drugs against coronavirus 3cl hydrolase and protease enzymes. *J. Biomolecular Structure and Dynamics*, pages 1–12.

Grein, J. *et al.* (2020). Compassionate use of remdesivir for patients with severe covid-19. *New England Journal of Medicine*.

Guy, R. K. *et al.* (2020). Rapid repurposing of drugs for covid-19. *Science*, **368**(6493), 829–830.

Halgren, T. A. (2009). Identifying and characterizing binding sites and assessing druggability. *J. Chemical Information and Modeling*, **49**(2), 377–389.

Jin, Z. *et al.* (2020). Structure-based drug design, virtual screening and high-throughput screening rapidly identify antiviral leads targeting covid-19. *BioRxiv*.

Kandeel, M. and Al-Nazawi, M. (2020). Virtual screening and repurposing of fda approved drugs against covid-19 main protease. *Life sciences*, page 117627.

Mittal, L. *et al.* (2020). Identification of potential molecules against covid-19 main protease through structure-guided virtual screening approach. *J. Biomolecular Structure and Dynamics*, pages 1–26.

Muralidharan, N. *et al.* (2020). Computational studies of drug repurposing and synergism of lopinavir, oseltamivir and ritonavir binding with sars-cov-2 protease against covid-19. *J. Biomolecular Structure and Dynamics*, pages 1–7.

Olsson, M. H. M. *et al.* (2011). Propka3: consistent treatment of internal and surface residues in empirical p k a predictions. *J. Chemical Theory and Computation*, **7**, 525–537.

Polikar, R. (2007). Bootstrap-inspired techniques in computational intelligence. *IEEE Signal Processing Magazine*, **24**(4), 59–72.

Rose, P. W. *et al.* (2011). The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res*, **39**(Database issue), D392–D401.

Sekhar, T. (2020). Virtual screening based prediction of potential drugs for covid-19.

Slater, O. and Kontoyianni, M. (2019). The compromise of virtual screening and its impact on drug discovery. *Expert Opinion on Drug Discovery*, **14**, 619–637.

Smith, N. and Smith, J. C. (2020). Repurposing therapeutics for covid-19: supercomputer-based docking to the sars-cov-2 viral spike protein and viral spike protein-human ace2 interface. *ChemRxiv*.

Sterling, T. and Irwin, J. J. (2015). Zinc 15–ligand discovery for everyone. *J. Chemical Information and Modeling*, **55**(11), 2324–2337.

Thachil, J. (2020). The versatile heparin in covid-19. *J. Thrombosis and Haemostasis*, **18**, 1020–1022.

Ton, A. T. *et al.* (2020). Rapid identification of potential inhibitors of sars-cov-2 main protease by deep docking of 1.3 billion compounds. *Molecular Informatics*.

Wang, J. (2020). Fast identification of possible drug treatment of coronavirus disease-19 (covid-19) through computational drug repurposing study. *J. Chemical Information and Modeling*.

Watts, K. S. *et al.* (2010). Confgen: a conformational search method for efficient generation of bioactive conformers. *J. Chemical Information and Modeling*, **50**(4), 534–546.

Xu, Z. *et al.* (2020). Nelfinavir was predicted to be a potential inhibitor of 2019-ncov main protease by an integrative approach combining homology modelling, molecular docking and binding free energy calculation. *BioRxiv*.

Zhou, G. *et al.* (2020). Computational drug repurposing studies on sars-cov-2 protein targets. *ChemRxiv*.

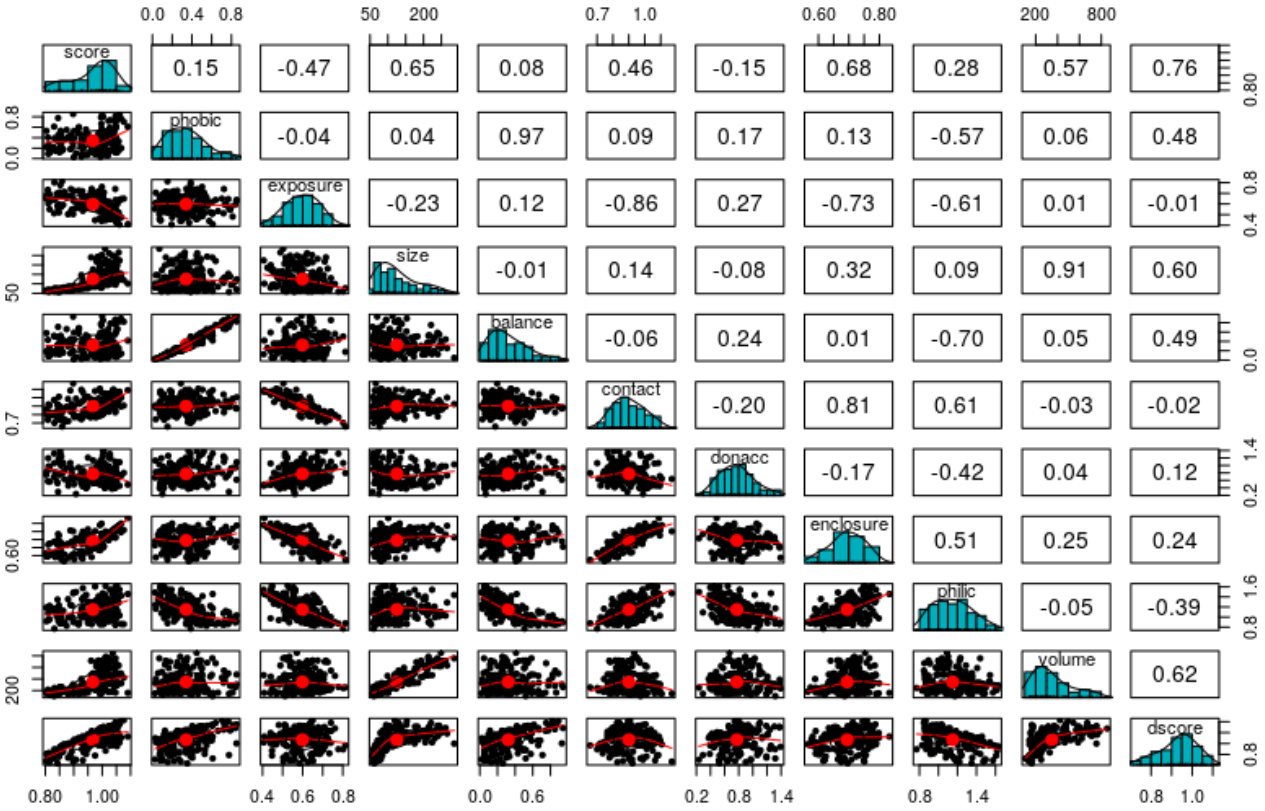

Figure 1: Matrix scatter plot of the variables describing the binding sites after filtration. The numbers in the upper left triangle is the correlation between variables. The parameter description of the binding sites are: score, phobic, exposure, size, balance, contact, donacc, enclosure, philic, volume, dscore. Their meaning is described in Schrödinger `sitemap` user's manual.

Supplementary Material

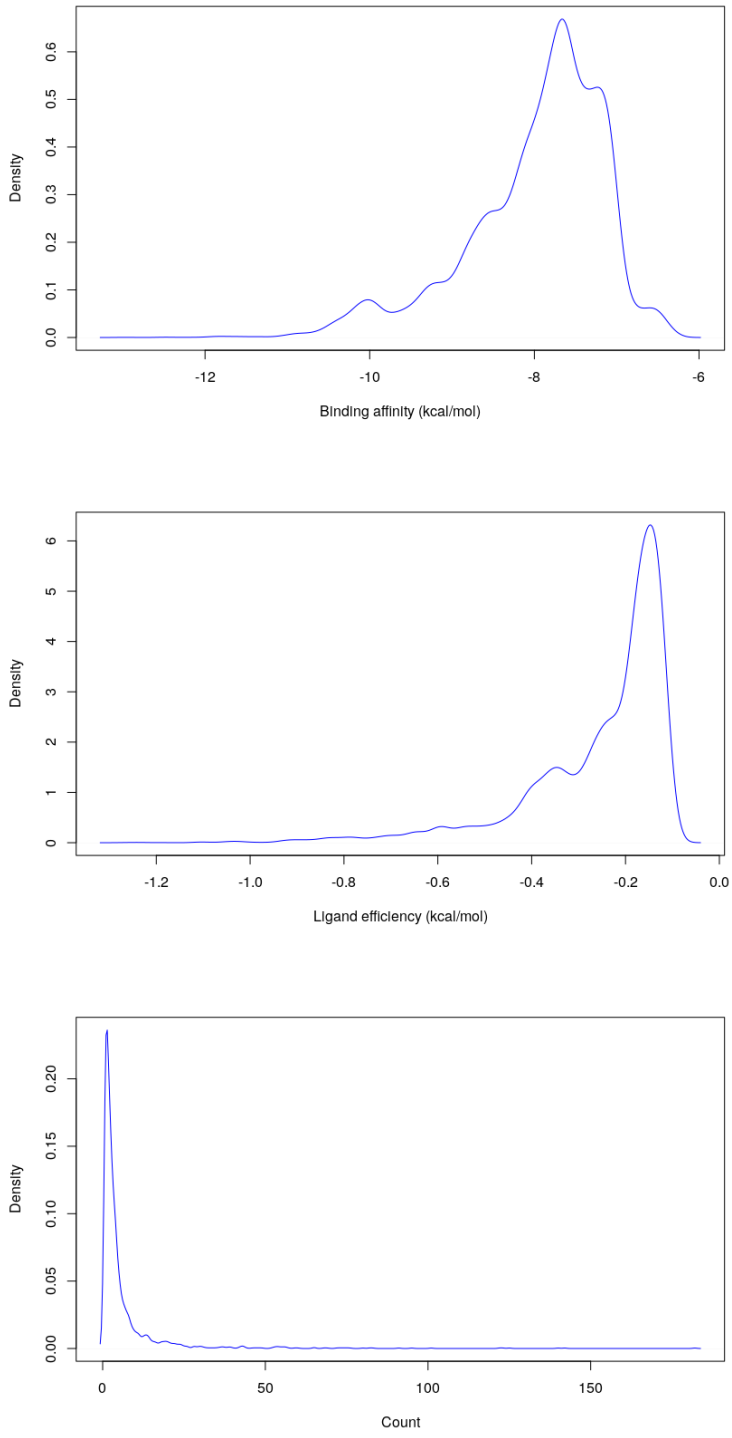

Figure 2: Affinity binding (docking score), ligand efficiency, and pose count distribution for the 12,306 top poses of the first virtual screening.

	Cangrelor	FAD	Plazomycin	NADH+NA	Isovuconazole	Capreomycin	Coenzyme A	Integrilin	Cephtolozane	Cepho porazone	Neomycin	Atosiban	Hexo prenaline
Spike	-23.61	-31.96	-25.06	-15.65	-46.06	-26.91	-33.13	-11.34		-7.71	-25.98	-36.38	-25.74
S2	-9.09						-9.26	-10.15	-26.2	-18.09	-14.54		
RBD	-14.72	-24.79	-51.5	-31.84	-79.42	-66.51	-16.18		-24.25		-59.72	-16.86	-59.13
NSP10 & 16	-41.68	-45.64	-7.14	-65.27	-16.00	-24.26	-41.99	-8.25	-7.64	-22.83	-7.09	-8.16	
NSP3	-31.22	-33.13		-56.88	-15.70	-8.67	-29.36		-7.92		-7.41		-14.67
NSP9	-7.50			-6.38			-15.77						
Papain-like protease	-29.70	-53.33	-13.67	-65.02	-38.80	-21.49	-58.68			-8.00	-42.36	-22.51	-14.93
NSP15	-15.73		-17.08	-15.55	-8.70		-16.47		-6.98		-15.41		
Nucleocapsid	-9.12			-18.95			-19.72						-8.58
NSP12	-30.67	-24.88	-21.74	-43.59	-32.39	-31.66	-30.37	-15.42			-15.02	-16.21	-23.14
NSP12 & 7 &8	-34.20	-33.65	-31.03	-31.00	-32.09	-32.07	-41.09	-23.84	-7.35			-15.05	-7.79
Sum	-247.24	-247.38	-167.22	-284.86	-269.16	-211.57	-312.02	-69.00	-80.34	-56.63	-187.53	-115.17	-153.98
# Sites	158	140	21	123	36	95	216	17	26	23	17	43	18

Table 3: Summary of the virtual screening performed on several structures of the viral proteins. We report the number of unique compounds found to interact with the macromolecule, the average number of binding sites per PDB structure (and the total volume, on average, for all sites in the target), the average number of unique compounds per binding site, and the range of affinity bindings and ligand efficiencies. FAD: flavin adenine dinucleotide. NADH: Nicotinamide adenine dinucleotide. NA: Nicotinic acid (it has been put together with NADH as NA is a degradation product of NADH).

	Neom.	Amox.	Indom.	Telm.	Ram.	Ator.	Omep.	Folic	Coenz.A	NADH	FAD	Vit	Vit+ Neom
Spike	-127.3	-19.52	-9.10	-16.17	-8.49			-33.94	-74.73	-47.30	-56.76	-88.68	-127.3
S2	-47.64	-17.04				-8.30		-18.61	-27.25	-26.74	-28.80	-29.15	-47.64
RBD/ACE2	-55.72			-8.54				-8.02	-18.52	-35.55	-41.19	-41.19	-55.72
NSP 10&16	-9.58					-9.92		-9.62	-10.35	-17.07	-8.2	-18.70	-19.93
NSP3	-21.61						-8.91	-8.10	-27.27	-32.35	-32.76	-33.13	-34.31
Papain like Protease	-31.08								-16.71	-24.84	-17.95	-26.06	-31.08
NSP15	-11.13							-8.04		-10.06	-8.37	-10.06	-11.13
Nucleocapsid	-9.19									-9.07	-9.24	-9.24	-9.24
NSP12	-9.22								-17.90	-17.19	-18.16	-18.57	-18.57
NSP12&7&8	-10.50								-8.50	-9.19	-8.03	-9.19	-10.50
Sum	-332.9	-36.6	-9.1	-24.7	-8.5	-18.2	-8.9	-86.3	-201.2	-229.4	-229.5	-284.0	-365.4
# Sites	28	4	1	3	1	2	1	10	22	25	24	33	33

Table 4: Summary of the 2nd virtual screening for some proteins and compounds. Neom: Neomycin, Amox.: Amoxicillin+Clavulanic acid, Indom.: Indomethacine, Telm.: Telmisartan, Ram.: Ramipril, Ator.: Atorvastatin, Omep.: Omeprazole, Folic: Folic acid, Coenz.A: Coenzyme A, NADH: Nicotinamide adenine dinucleotide, FAD: Flavin adenine dinucleotide, Vit: Vitamins, Vit+ Neom: Vitamins and Neomycin

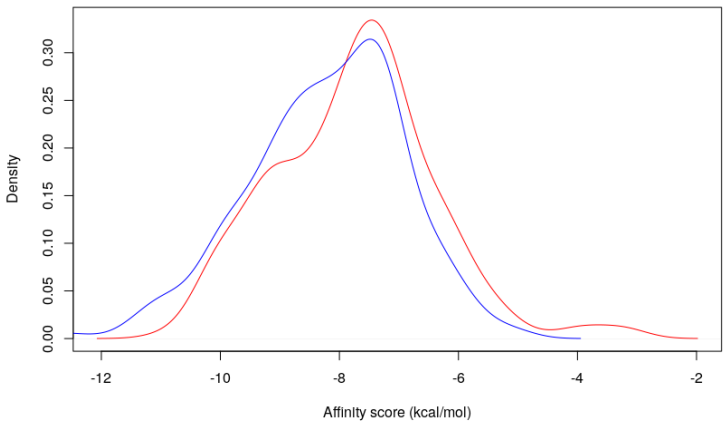

Figure 3: Affinity binding for the treatment group (blue) and control group (red) of the second virtual screening.