Computational Drug Discovery for COVID-19 (COVID19CADD)

Tristan Camilleri

Supervised by Dr Jean Paul Ebejer

Centre for Molecular Medicine and Biobanking

University of Malta

October, 2022

Dissertation submitted in partial fulfilment of the requirements for the degree of M. Sc. in Bioinformatics

To my wife, Sarah

whose strength and support were instrumental in me pursuing my academic dreams

Acknowledgements

An immense “Thank You” go to my supervisor, colleague and friend, Dr. Jean Paul Ebejer, for his talent and dedication in sharing his knowledge with his students, for his relentless support and for simply being there each time I needed that push.

Deep gratitude and love also go to my wife, Sarah and son, Sven, for patiently putting up with me. I could not have done this without their support through the difficult times.

I am also grateful to the Centre for Molecular Medicine and Biobanking for offering this Master of Science in Bioinformatics. The lecturers and the subject matter covered were just what I was looking for in my quest for knowledge on biosciences and bioinformatics.

Abstract

The COVID-19 pandemic, caused by the novel virus SARS-CoV-2, is most likely here to stay with us and notwithstanding the rapid deployment of vaccines, there is a need to develop antivirals against this virus. This is because new variants of the virus will emerge against which current vaccines might be less effective, there will be people that cannot be vaccinated or areas of the world where deployment of vaccination programs are not as efficient as in other more developed countries. For this reason there is a significant effort to develop antivirals effective against this virus.

Virtual screening is a set of computational tools within the Computer-Aided Drug Design toolbox that is available in order to perform initial filtering of small molecules in order to identify hits that show potential and merit being studied further as part of a drug development pipeline. In this study we make use of two Ligand-Based Virtual Screening (LBVS) techniques – Molecular Fingerprint Similarity Searches and Ultrafast Shape Recognition with CREDO Atom Types (USRCAT) – to search for small molecules that are similar to a set of query molecules that have been identified as having an inhibitory effect against the Main Protease (M^{pro}) of SARS-CoV-2.

Our experiments have resulted in a list of 42 and 195 hits identified from Fingerprint and USRCAT searches, respectively. These results were validated by the calculation of the Enrichment Factor, which resulted in scores (well) above a value of 1, with mean $EF_{1\%}$ values of 7.56 and 2.57 for Fingerprint and USRCAT searches, respectively. These values can be studied using other *in silico* tools such as molecular docking or *in vitro* studies.

Keywords: SARS-CoV-2, COVID-19, CADD, Ligand-based Virtual Screening (LBVS), Main Protease (M^{pro}), Molecular Fingerprint, USRCAT.

Table of Contents

List of Figures

List of Tables

1

Introduction

One of the most memorable quotes from Albert Camus's novel "The Plague" is that *"Everybody knows that pestilences have a way of recurring in the world, yet somehow we find it hard to believe in ones that crash down on our heads from a blue sky. There have been as many plagues as wars in history, yet always plagues and wars take people equally by surprise."* This cannot be truer than the case of the COVID-19 pandemic, where there were two previous outbreaks caused by a Coronavirus, the first one being the Severe Acute Respiratory Syndrome (SARS) outbreak in 2003 and the second one being the Middle East Respiratory Syndrome (MERS) outbreak in 2011 (Franco-Paredes, 2020).

Severe acute respiratory syndrome coronavirus two (SARS-CoV-2), the viral agent giving rise to COVID-19, is a novel coronavirus. This virus belongs to the Coronavirus family, the same family as SARS-CoV and MERS coronavirus. The rapid global spread of the virus has prompted the World Health Organization to declare COVID-19 a pandemic in 2020 (Lan et al, 2020). As of 15 August 2021, there have been 588 million confirmed cases of infection and 6.43 million deaths by SARS-CoV-2 at a Global level, and 114,000 confirmed cases and 797 deaths in Malta. Apart from the direct health impact of the pandemic, it has also had a severe impact on the global economy and educational systems (Nicola, et al., 2020).

1.1 Virtual Screening

The typical cost for a typical drug discovery cycle, covering the entire process from lead identification to clinical trials is US$ 500-800 million and takes 10-15 years (Pandey and Singh, 2017). Other estimates for a drug discovery cycle put a value of US$2 billion over 10 to 15 years (Berdigaliyev and Aljofan, 2020).

Computer-Aided Drug Design (CADD) is one of the methods used in drug development, especially in the earlier stages, that results in improved efficiency and reduced costs (Macalino et al., 2015). Virtual Screening (VS) is the automation of processes related to the screening of large databases of small molecules to identify compounds that elicit the desired biological effect and reduce the chemical library to a more manageable size (Walters, Stahl, and Murcko, 1998). Virtual high-throughput screening (vHTS) is used to identify initial hit compounds from a large dataset for further studies, by assessing the binding affinities of small-molecules to the target protein (Subramaniam, Mehrotra, and Gupta, 2008). Ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS) methods have produced positive results (Seifert, Kraus and Kramer, 2007). One limitation of virtual screening is the predictive range of the mathematical relationships that can reliably differentiate between true hits from inactive non-binding molecules (Westermaier, Baier and Scapozza, 2015).

LBVS involves the use of structure-activity data from a group of compounds that are known binders to a target protein to search for other active compounds (Berenger, Vu and Meiler, 2017). LBVS methods are usually chosen when there is no available 3D-structure of the target protein, nevertheless, similarity searches provide an efficient method of increasing the number of potentially active compounds that could be further studied in a drug discovery project (Macalino et al, 2015). The compound search can be based on similarity, substructure, pharmacophore (abstraction of molecular features resulting in the molecule to be active (Wermuth et al., 1998)), or shape matching.

There is a multitude of LBVS methods, such as Quantitative Structure-Activity Relationship (QSAR) methods, pharmacophore modelling, fingerprints and Ultrafast Shape Recognition (USR), and ligand shape (volume)-based methods, to name a few (Lavecchia and Di Giovanni, 2013; and Hamza, Wei and Zhan, 2012).

The use of virtual screening methodologies has been proven to be a cost-effective and time-saving methodology in the identification and development of lead compounds. The efficiency of these methodologies makes them especially important in the development of drugs effective against COVID-19 (Hou and Zu, 2004).

1.2 Motivation

By 5 June 2021, at least 24 different vaccines have been made available globally to target SARS-CoV-2. Nevertheless, to effectively defeat the COVID-19 pandemic, there will be the need for an arsenal of treatments that includes effective antivirals. New Coronavirus strains will continue emerging, against which the efficacy of vaccines will be challenged; there will always be a cohort of the general population that will refuse vaccination, cannot receive a vaccination, or will still be infected by the virus regardless of vaccination status; and there are countries in which vaccination campaigns will be slow to implement (Wang et al., 2021). These are further justifications for the need to identify suitable SARS-CoV-2 inhibitors to be used in the development of effective drugs against COVID-19 infection.

Modern drug design includes several computational methods as part of the tools used to develop new drugs, these are categorized as Computer-aided drug design (CADD). These methodologies have been and are continuously being refined to increase efficiencies (time and economical resources) in a drug discovery pipeline (Prieto-Martinez, et al., 2019). One of the tools in the CADD portfolio is Virtual Screening (VS) and the use of virtual screening experiments akin to the ones being undertaken by this study, constitute the initial steps of a modern drug discovery pipeline. Such experiments generally improve the efficiency of a drug discovery pipeline by providing a (reduced/targeted) number of possible hits, which are molecules that can be converted into leads that could be further investigated for safety and effectiveness.

The COVID-19 pandemic has been a driver for the scientific community to rapidly develop therapeutics that could be used against the COVID-19 infection. The relative rapidity of CADD, and in particular virtual screening, has driven a significant amount of research aimed toward identifying molecules showing desirable bioactivity against COVID-19 infection. Liu et al (2022) reviewed 181 CADD peer-reviewed publications that identified 779 compounds of interest that were effective against 16 target proteins relevant to the biology of SARS-CoV-2. This shows that CADD methodologies, such as virtual screening, provide a very useful toolset in the development of drugs effective against SARS-CoV-2. The development of drugs that are effective against COVID-19 infection is especially important in the management of endemic COVID-19. An endemic infection is an infection in which the proportion of the population that can get sick is balanced by the viral basic reproduction number (Katzourakis, 2022). It is important to highlight that a viral infection becoming endemic is not directly

proportional to the infection's mortality and therefore we must remain vigilant in the fight against COVID-19.

1.3 Aims and Objectives

The primary aim of this study is to identify promising hits that can be further studied as part of a drug discovery pipeline against infection by SARS-CoV-2. The identification of potential inhibitors shall be performed computationally through VS experiments to identify inhibitors that are active against a SARS-CoV-2 target protein.

The two main scientific research questions that need to be answered in order to achieve the aim of this study are:

1. What is the most suitable SARS-CoV-2 relevant protein target?
2. Out of a defined set of molecules, which molecules are (most) similar to known inhibitors of the chosen protein target?

Ligand-based virtual screening methodologies rely on the use of reference compounds that have been shown to be active against a (protein) target of interest. Therefore, in order to identify relevant ligands, it is important to identify a protein target of interest. As a second step, we need to proceed to identify compounds that are similar to the ligands of interest because we rely on the assumption that compounds with a similar structure will share similar properties and functions (Leach and Gillet, 2007).

1.3.1 Objectives

The objectives of this study are the following:

i. To select a suitable SARS-CoV-2 target protein in order to identify suitable ligands to be used in the LBVS experiments (similarity search);

ii. To prepare a database of inhibitors active against the selected protein target, that are identified from the literature, to be used in the similarity searches;

iii. To prepare a small-molecule database based on publicly available information. These will be filtered for drug likeness, clustered and used for the similarity searches to identify compounds that are similar to the ligands;

iv. To use LBVS tools to identify small molecule inhibitors that are potentially active against the selected SARS-CoV-2 target protein to take forward to lead optimization;

v. To validate the results that were obtained from the LBVS protocols;

vi. To compile a list of putative inhibitors based on LBVS protocols that can be further assessed in more advanced drug development pipelines.

1.4 Approach

The study will be divided into four main tasks. The first task revolves around the identification of the most relevant protein target. A cursory search through literature identifies at least four protein targets (M^{pro}, S-protein, $3CL^{pro}$, and RdRp) relevant to the fight against SARS-CoV-2, however, the Spike protein (S-protein) and Main protease (M^{pro}) enzyme are the two protein targets with the highest number of published studies that tackle the identification of inhibitors against these targets. Several authors (Coelho et al. 2020, Dai et al. 2020, Huang et al., 2020, Xiu et al., 2020b, Xu et al. 2020, Yang et al., 2020, and Zhu et al. 2020) identify the S-protein and M^{pro} as promising candidates for the development of potential drugs that could be used in the fight against COVID-19. The critical roles played by these two proteins in cellular entry and viral replication make them valid targets for drug discovery. High resolution (3Å and less) 3D structures of both these two targets are freely available for download from the Protein Data Bank[1] (Berman et al, 2000), such as 6LU7 for M^{pro} and 6M0J for the S-protein.

Since the start of the COVID-19 pandemic, an enormous global effort has been undertaken to either repurpose already approved drugs (for other diseases) or to identify promising new molecules to be used in the fight against SARS-CoV-2 (Beck et al., 2020; Chenthamarakshan et al., 2020; Fischer et al., 2020; Stebbing et al., 2020; Zhavoronkov et al., 2020). An extensive literature review shall be undertaken to identify known inhibitors of the selected target protein. These shall be compiled in a database and shall be used as the basis for the LBVS experiments. Similarity searches will be carried out using the identified inhibitors against the clustered small molecules.

Secondly, a publicly accessible compound library, such as ZINC15, shall be selected in order to download a tranche of commercially available small molecules. This will then be used to

[1] https://www.rcsb.org/

develop a collection of small molecules for this study. In order to filter out molecules that are not considered to be bioavailable via the oral route, different filtering strategies, such as those based on the Rule of Five (Lipinski et al., 1997), shall be used to create a database of drug-like small molecules. This database shall serve as the basis for the compound search experiments carried out through LBVS.

An initial molecular washing protocol shall be undertaken, to remove any duplicate molecules and to standardize the molecules. The standardization process is necessary to ensure that the correct tautomer of the specific molecule is properly represented and can be used within the experiment. A commonly used filtering strategy is based on the rule of five (Lipinski et al., 1997; Pollastri, 2010), which predicts that poor (oral) absorption or permeation is more likely when there are more than five H-bond donors, ten H-bond acceptors, the molecular weight is greater than 500, and the calculated octanol-water partition coefficient (C Log P) is greater than five. An effectively absorbed molecule, ingested via the oral route, does not violate more than one of these conditions.

Due to the computational resources required for performing VS experiments on large chemical datasets, clustering of molecules shall also be undertaken to reduce the number of molecules used for the selected VS techniques.

To conduct the LBVS experiments, techniques based on Ultrafast Shape Recognition (such as those based on ElectroShape 5D or USRCAT) or similarity searches against Extended Connectivity Fingerprints shall be undertaken to find molecules that are similar to the known inhibitors identified from the literature.

1.4.1 Evaluation

The VS experiments rely on well-established VS methods which have been used widely for drug development, in many instances, including to identify small molecule inhibitors of the S- and M^{pro} proteins. Moreover, as aforementioned, the literature review will identify a number of small molecules that have been putatively identified as SARS-CoV-2 inhibitors.

A two-pronged approach is being proposed to evaluate the validity of the results obtained from the LBVS experiments:

1. The compounds identified from the LBVS experiments will be compared with other inhibitors that are listed in other relevant COVID-19 related repositories of potential

inhibitors such as the COVID Moonshot project[2], which is an open-source repository of inhibitors of M^{pro}.

2. The second approach will involve the calculation of enrichment factors. This shall be carried out through the seeding of the chemical lists with the inhibitors identified from the literature review into the VS experiments. The known inhibitors (positive controls) will need to be identified as such in order for the experiment to be considered valid. It is important to note that for the LBVS experiments, only a selection of the known inhibitors shall be used as a positive control, since LBVS methodologies require by their own nature, known active compounds to be used implicitly for the execution of the experiments.

1.5 Document Structure

This dissertation will be structured as follows.

A comprehensive overview of the existing literature available on the possible SARS-CoV-2 protein targets, including the viral biology and the different computational techniques used in virtual screening shall be provided in the **Background and Literature Overview** chapter.

The **Methodology** chapter will provide a detailed overview of the techniques used in order to achieve the set aims and objectives. This overview shall cover the methods used to clean and standardize the molecular database, the clustering methodology used, and the selected VS models. Arguments shall be presented to back the choices in methods made, coupled with the respective resource requirements.

The **Results and Discussion** chapter will provide a detailed overview of the undertaken experiments and the results that were obtained. The evaluation of the results shall also be provided, including a comparison of the results obtained from the different techniques. The results will be analyzed in the context of the available literature.

In the final chapter, the **Conclusion**, there will be a final overview of the aspects covered by this study, the selected methodologies used to meet the aims and objectives, the

[2] https://postera.ai/moonshot

results that were obtained, the limitations of the study, and any possible future work that might be desirable in the driving the research on the subject forward.

2

Background and Literature Review

The COVID-19 pandemic has driven the research community into a spectacular response to, among others, find suitable therapeutics and develop vaccines. According to the Milken Institute, as of the 5th November 2021, there are 272 vaccines in development, 104 vaccines in clinical testing, and 24 vaccines in use (Shroti et al, 2021). Nevertheless, to effectively defeat the COVID-19 pandemic, there will be the need for an arsenal of treatments that includes effective antivirals. This need for therapeutics arises from the fact that new viral strains will likely emerge which will render current vaccines ineffective. Moreover, there will still be a cohort of the general population that will still be infected by the virus, irrespective of vaccination status. Ultimately, there will always be the need for an effective treatment, irrespective of the existence of vaccines.

2.1 Coronaviruses

Coronaviruses (CoVs) are a large family of single-stranded positive-sense RNA viruses found in various animal species, such as bats, camels, pigs and pangolins (Ye et al, 2020). The Coronavirus family consists of four genera, these being Alpha-, Beta-, Gamma- and Delta-. The viral RNA is coated with a membrane envelope covered with spike-shaped peplomers (glycoproteins), this characteristic crown shape is the reason behind the name of the family. Almeida and Tyrrell (1967) were the first to describe the characteristic feature of coronaviruses, with the surface of the viral particles being "The surface of the particles is covered with a distinct layer of projections roughly 200 Å, long. These projections seem to have a narrow stalk just within the limit of resolution of the microscope and a 'head' roughly 100 Å across." The use of a dictionary led the authors to the Latin term corona (Barlas et al, 2021), which was used to name the new family of viruses (Tyrrell and Fielder, 2002).

CoVs cause a multitude of diseases in humans and animals, some of which have a significant impact, for example Transmissible Gastroenteritis Virus (TGEV) and Porcine Epidemic Diarrhea Virus (PEDV) cause severe gastroenteritis in young piglets with a high

level of morbidity and mortality (Maier, Bickerton and Britton, 2015). Up to now, there are seven known disease-causing coronaviruses in humans, three highly pathogenic CoVs (HPCoVs) (SARS-CoV-2, MERS-CoV, and SARS-CoV) and four normal human CoVs (HCoVs) (HCoV229E, HCoV-OC43, HCoV-NL63, and HCoV-HKU1) (Jiang et al., 2020). The symptoms of these diseases include effects on the respiratory, hepatic, nervous and gastrointestinal systems, which can also be fatal (Harapan et al, 2020). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the viral agent giving rise to COVID-19 is a novel coronavirus belonging to the genus *Betacoronavirus*. This virus has spread globally, prompting the World Health Organization to declare COVID-19 a pandemic in March 2020 (Lan et al, 2020).

The viral genetic material of CoVs is composed of a non-segmented, positive-sense RNA genome ranging between 26 to 32kb (Luk et al., 2019). The strand contains a 5′ cap structure along with a 3′ poly (A) tail, allowing it to act as an mRNA for translation of the replicase polyproteins. The replicase gene encoding the non-structural proteins (nsps) occupies about 20 kb, as opposed to the structural and accessory proteins (around 10 kb). A leader sequence and an untranslated region (UTR) is found at the 5′ end of the RNA. Multiple stem loop structures that are needed for RNA replication and transcription can also be found in this region. Each structural and accessory gene have transcriptional regulatory sequences (TRSs) at their beginning, these are deemed to be required for gene expression (Fehr and Perlman, 2015).

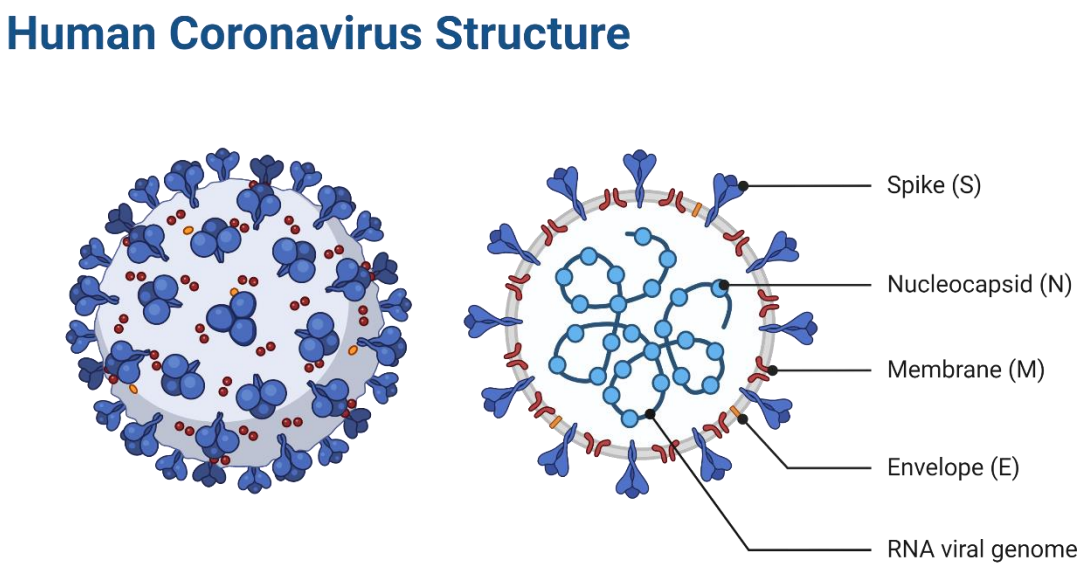


Figure 2.1. Structure of a typical human Coronavirus. Adapted from “Human Coronavirus Structure”, by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates

The four main structural proteins found in coronavirus particles are the spike (S), membrane (M), envelope (E), and nucleocapsid (N) proteins (as shown in Figure 2.1). These are encoded within the 3′ end of the viral RNA. The M protein has three transmembrane domains and it is the most abundant structural protein in coronaviruses, giving the viral particles their shape (Nal et al., 2005). Homotrimers of S protein create the distinctive spike structure on the surface of the virus (Beniac et al., 2006). The S protein mediates binding to the host receptor, which in the case of SARS-CoV-2 it binds to the epithelial peptidase domain of the Angiotensin-Converting Enzyme 2 (ACE2) (Zhang, Li and Niu, 2020).

Successful viral replication is critically dependent on a suite of host factors and host cellular pathways. Therefore, a deep knowledge of the host-virus interactome is a good tool for identifying suitable targets for the development of potential drugs (Luan et al, 2020). This also provides the opportunity for repurposing any approved drugs presenting a good potential for different viruses that rely on the same or similar host factors or pathways (Baggen et al, 2021). Figure 2.2 presents the viral life cycle, which includes the virus-host interactions.

The viral genome of SARS-CoV-2 is around 30kb in length and contains 14 open reading frames – ORFs. The first and second ORF encode for two polyproteins, pp1a and pp1ab

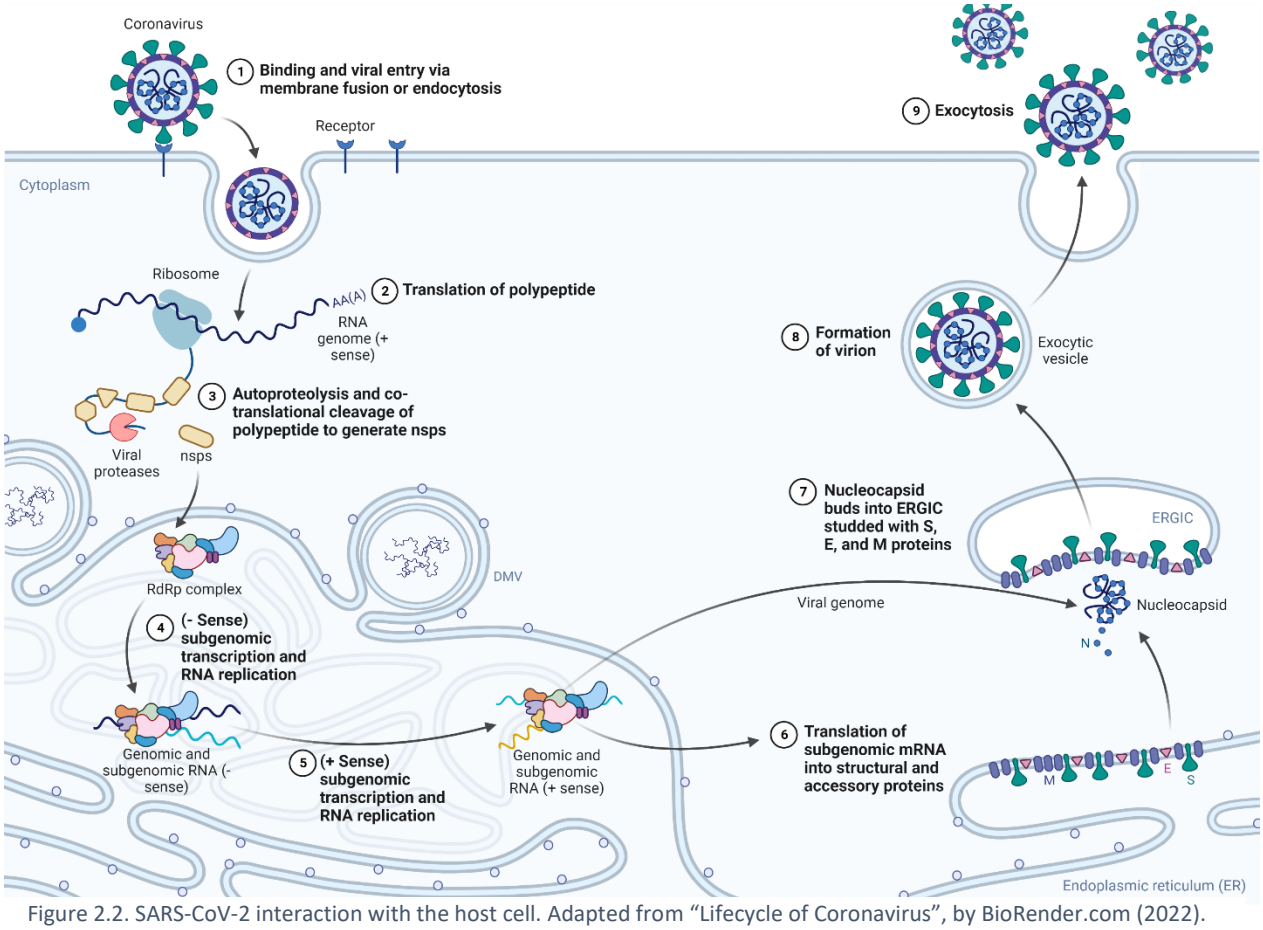


Figure 2.2. SARS-CoV-2 interaction with the host cell. Adapted from "Lifecycle of Coronavirus", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates

(Luk et al., 2019, and Mohamadian et al., 2021). The remaining 12 ORFs give rise to the four structural proteins that are essential to the virus's survival and proliferation, and nine accessory proteins (Rohaim et al, 2020 and Yadav et al. 2021).

Most global efforts to identify new therapeutics revolve around drug repurposing, virtual screening, and *de novo* drug design, with different levels of success. Whilst acknowledging the fact that all viral proteins and enzymes involved in viral replication and processes aimed at controlling the interaction of the virus with the cell are potential targets (Gil et al, 2020), the research efforts broadly revolve around four targets. These are:

- Main protease (M^{pro})
- Spike protein (S-protein)
- Papain-like protease (PL^{pro})
- RNA-dependent RNA polymerase (RdRp)

2.2 Spike Protein

The SARS-CoV-2 transmembrane spike glycoprotein (S-protein) plays an important role in receptor recognition and cellular membrane fusion. The virion has homotrimers of S-proteins spread across its surface, giving it the appearance of a crown, from which the name coronavirus is derived from. This homotrimer binds to the ACE2 receptor (located on the cell surface membrane of cells located in numerous organs across the human body). ACE2 is increasingly expressed in the respiratory and bronchial epithelium which explains why SARS-CoV-2 infection presents itself as a respiratory infection. The S-protein (1,273 amino acids long) is composed of a short (13 amino acid) signal peptide and two subunits (S1 and S2), with S1 containing the Receptor-Binding Domain (RBD) and S2 being responsible for cellular membrane fusion. The S1 unit also contributes to stabilizing the fused state of the S2 subunit to the host cellular membrane (Huang et al, 2020). The S2 subunit is mainly composed of the fusion peptide (FP), heptapeptide repeat sequence 1 (HR1), heptapeptide repeat sequence 2 (HR2), transmembrane anchor (TA) domain, and intracellular tail (IT). HR1 and HR2 together form the six-helical bundle (6-HB), as shown in Figure 2.3.

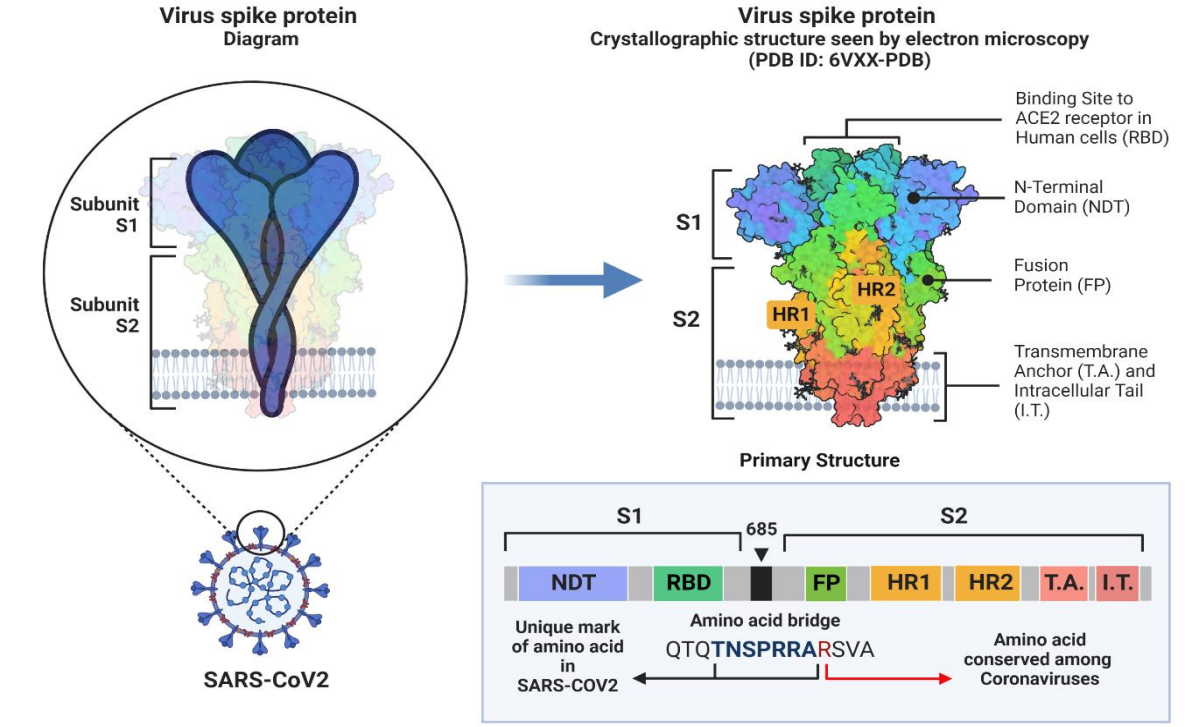


Figure 2.3. S-protein structure, including main sub-units that are essential for receptor binding and cellular fusion. Adapted from "In-depth look into the structure of the SARS-CoV-2 spike protein", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates

When S1 binds to ACE2 the S-protein must then be cleaved by a protease, such as Transmembrane Serine Protease 2 (TMPRSS2), furin or cathepsin L. This results in the cleavage of the S-protein. This process exposes the fusogenic core of S2, which is a hydrophobic high energy metastable state. S2 then undergoes a conformational change resulting in an extended HR1 and HR2 domain, promoting the injection of the FP into the host cellular membrane, forming the pre-hairpin intermediate, which then folds back into a six-helix bundle (6-HB), separating the host membrane. The final stage is the fusion of the viral and host cell membrane, with HR1 and HR2 forming a pore (Figure 2.4) (Krumm et al, 2021).

A key difference of SARS-CoV-2 from other beta-*Coronaviridae*, such as SARS-CoV, is that the S-protein contains an arginine residue rich (multibasic), furin mediated, cleavage site (681–PRRAR↓SV-688). This major difference is purported to contribute to the greater infectivity of SARS-CoV-2 when compared to other related viruses (Hoffmann, Kleine-Weber, and Pöhlmann, 2020, and Huang et al, 2020). Moreover, the SARS-CoV-2 S-protein contains multiple furin cleavage sites, thus facilitating cleavage by proteases similar to furin, thereby increasing the viral infectivity.

The reported drug development strategies targeting the S-protein, in one way or another, revolve around the blockage or modulation of the RBD-ACE2 binding or the inhibition of the

viral membrane fusion. The greatest concern with S-protein targeting therapies is the relatively high mutation rate and genetic variability revolving around the S-protein and the uncertainty this brings on the efficacy of any treatments that are developed. An alternative approach is the

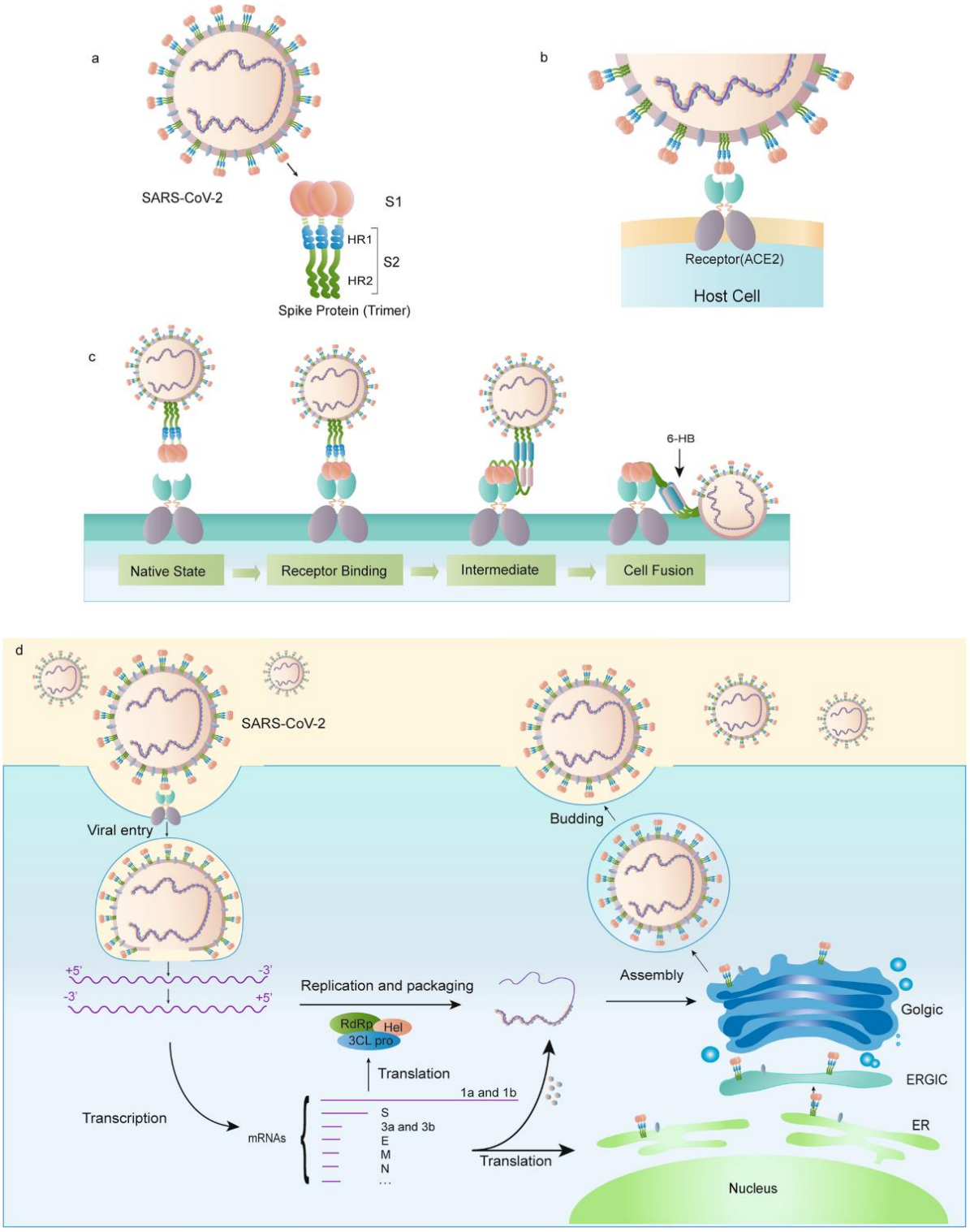


Figure 2.4. a. Schematic structure of the S-protein. b. View of the binding of the spike protein to the ACE2 receptor. c. A mechanistic view of S-protein binding and cellular membrane fusion. d. Representation of the viral life cycle of SARS-CoV-2. Reproduced from Huang et al (2020).

use/development of ACE2 expression modulators or inhibitors of key host cell proteases, such as TMPRSS2, cathepsin, or furin. Hoffmann et al (2020b) have reported that the combined use of cathepsin B and L, and TMPRSS2 protease inhibitors, E-64d, and camostat mesylate, respectively, in Caco-2 and Calu-3 (colon and lung epithelial, respectively) cell lines resulted in the blockage of viral entry via the inhibition of the cleavage of the S1 and S2 sub-units by TMPRSS2.

The currently developed vaccines targeting SARS-CoV-2 all target the viral S-protein. However, the rapid emergence of viral variants of concern has resulted in a significant concern that various mutations might emerge in the near future that will result in conformational changes of significance within S-protein, presenting challenges to the currently available vaccines (Singh et al, 2020). This is also a concern in the realm of drug development, presenting a significant risk in the long-term efficacy of any drug that is developed, which makes it an important consideration when choosing a suitable protein target to study.

2.3 Main Protease

Following viral membrane fusion with the host cellular membrane, the viral RNA that is released into the host cell is translated into two polyproteins (pp1a and pp1ab) by the host ribosomes. The SARS-CoV-2 main protease (M^{pro}), also known as 3-chymotrypsin like protease ($3CL^{pro}$), is responsible for the cleavage of pp1a and pp1b into the non-structural proteins (nsp) 4 to 16, which are critical to viral replication. This makes M^{pro} one of the major targets in the discovery of drugs against SARS-CoV-2 (Mody et al, 2021).

M^{pro} is a homodimeric cysteine protease that is made up of three domains, with domains I and II being mainly composed of antiparallel β-barrels and domain III being mainly made up of five α-helices forming an antiparallel globular structure (Figure. 2.5). M^{pro} has a His41–Cys145 catalytic dyad located in domains I and II with further subsites located in the active site (S1 and S2 being deeply buried in the active site and S3-S5 being shallow subsites). Cys145 and His145 act as a nucleophile and proton acceptor, respectively. The subsites also play a role in the catalytic activity of the enzyme with S1 and S2 providing hydrophobic and electrostatic interactions with the substrates, and the other subsites providing more flexible functionalities (Forrestall et al 2021). The final shape of the pocket of the substrate-binding site is dependent on the dimerization of the protomers, and the active site also contains a number of water molecules that are considered to be critical for enzymatic activity.

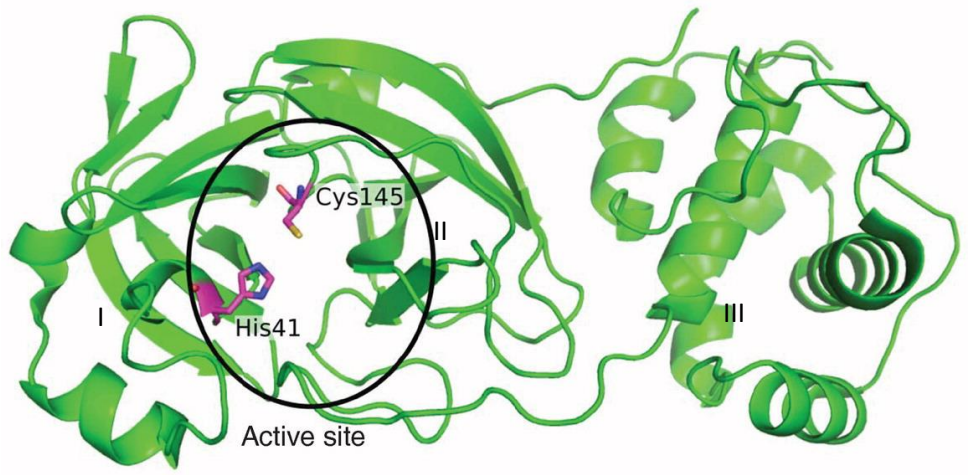


Figure 2.5. 3D structure of M^{pro} protomer from SARS-CoV-2, His41, and Cys145 in magenta, adapted from Bung et al (2021). I, II and III represent the three domains of the enzyme.

Jin et al (2020) report that the substrate-binding pocket is highly conserved among the main protease in all coronaviruses, with the main variability being found in the helical domain III and surface loops. Bung et al (2021) report a root mean square deviation of 0.56 Å between M^{pro} from SARS-CoV and SARS-CoV-2. This implies that any compound inhibiting this substate pocket is also expected to be effective in all coronaviruses sharing the similarity of the protease's active site.

An added advantage of targeting M^{pro} in the drug development strategy is that this enzyme has a unique cleavage specificity, the Leu-Gln↓(Ser, Ala, Gly) sequence (↓ indicates the cleavage site, and the parenthesis indicates that the amino acids are interchangeable with each other), that is not found in other human proteases. Therefore, M^{pro} inhibitors, depending on their specificity, are not expected to present a toxicological concern in relation to other human proteases (Forrestall et al 2021).

The highly unique amino acid cleavage sequence targeted by M^{pro}, and the important role played by this enzyme in viral replication, make M^{pro} a very interesting protein target. However, a number of studies have identified high flexibility and plasticity in the binding site of SARS-CoV-2 M^{pro} when compared to that of SARS-CoV (Citarella et al, 2021 and Shitrit et al, 2021). It was also noted that there are large adjustments in the cavity shape as a result of binding with an inhibitor (Gossen et al, 2021). SARS-CoV M^{pro} and SARS-CoV-2 M^{pro} differ only by 12 residues that are located far from the binding site. This indicates that a limited number of mutations, far from the binding site, can have a significant allosteric effect on binding site plasticity and ligand binding. Notwithstanding the above, recent studies have

indicated the retention of binding affinities for certain ligands, for example GC373 has been found to inhibit both proteases (Vuong et al, 2020 and Vuong et al, 2021).

2.4 Papain-like Protease

The second protease encoded by SARS-CoV-2 is Papain-like protease (PL^{pro}). This protease is responsible for the cleavage of nsp1, nsp2 and nsp3, with PL^{pro} being itself encoded within nsp3. The nsps form the viral replicase complex on the host cellular membrane which serve the purpose of initiating viral replication and transcription. PL^{pro} recognizes the cleavage sequence Leu-X-Gly-Gly↓X-X, which is the same sequence recognized by cellular deubiquitinating enzymes (Osipiuk et al, 2021). Apart from processing the two polyproteins, PL^{pro} also disrupts the host's immune response, through deubiquitination of ubiquitin (Ub) and ubiquitin-like proteins (UbL) from host proteins that are involved in antiviral immunity signalling pathways. One of these is interferon-stimulated gene product 15 (ISG15), which is upregulated upon viral infection (Klemm et al, 2020). When the virus infects human macrophages, PL^{pro} leads to the extracellular release of ISG15, this acts in a cytokine-like manner to exaggerate the subsequent secretion of multiple proinflammatory cytokines and chemokines causing a cytokine storm (Ramasamy and Subbian, 2021). This is believed to be the mechanism behind the super-inflammatory response seen in COVID-19 patients (Cao 2021). Thus, inhibiting PL^{pro} functionality is expected to have a dual effect of decreasing viral replication and reducing the inflammatory response caused by SARS-CoV-2.

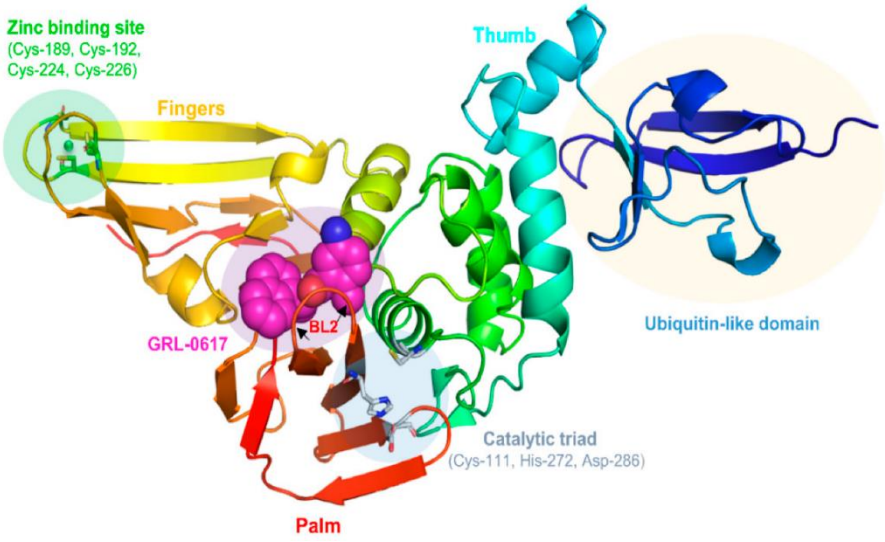


Figure 2.6. Structure of PL^{pro}. Reproduced from Hajbabaie, Harper and Rahman (2021)

PL^{pro} from SARS-CoV-2 shares an 83% sequence similarity with that from PL^{pro} of SARS-CoV, moreover, it shares a high similarity with the sequence of PL^{pro} from MERS-CoV,

Swine Acute Diarrhea Syndrome, Murine Hepatitis Virus, Avian Infectious Bronchitis Virus and Transmissable Gastroenteritis Virus, amongst other coronaviruses (Kandeel et al., 2021). However, it is worth pointing out that the enzyme shares low sequence similarity with any human enzymes, possibly indicating that any PL^{pro} inhibitors would have limited binding to other human enzymes (Osipiuk et al, 2021). PL^{pro} has four domains, a terminal UBl domain, and a 'thumb' catalytic domain, a 'fingers' (zinc-binding) domain and the 'palm' domain, forming a catalytic triad (Figure 2.6) (Hajbabaie, Harper and Rahman, 2021).

Shen et al (2021) report that the highly restricted binding pocket at the substrate-binding sites (Gly-Gly recognition) makes the identification of potential inhibitors difficult, in fact, there are very few PL^{pro} inhibitors with experimentally validated efficacy.

2.5 RNA-dependent RNA Polymerase

RNA-dependent RNA polymerase (RdRp) is an essential enzyme encoded by the SARS-CoV-2 genome, which in association with other transcription factors, is responsible for the replication and transcription of the viral RNA. RdRp is formed by a heterodimer (nsp7 and nsp8), an additional nsp8 subunit and a catalytic subunit (nsp12) (Gao et al, 2020). The RdRp polymerase domain can be subdivided into a right-handed 'fingers-palm-thumb' conformation (Figure 2.7)

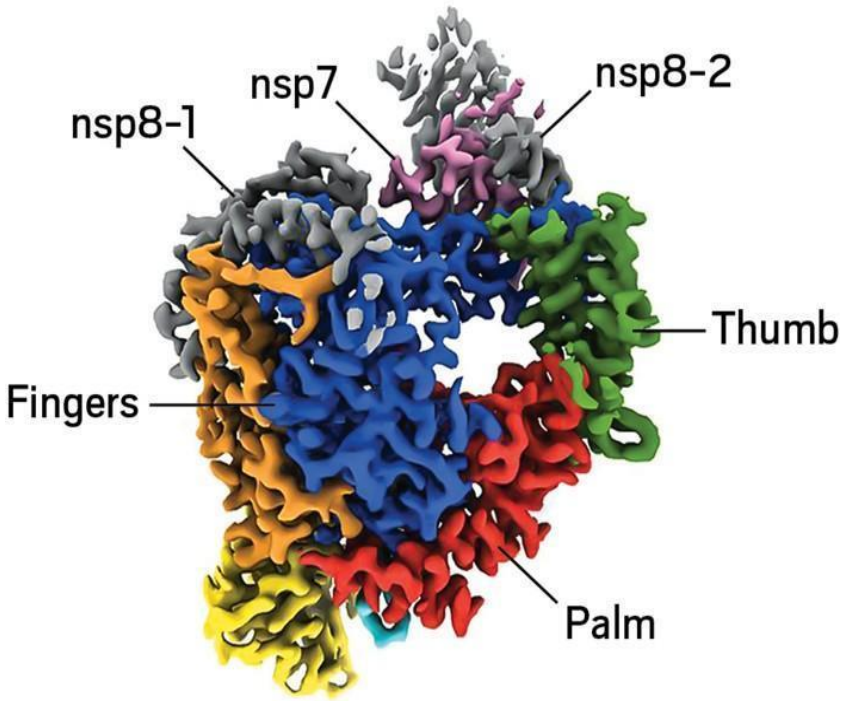


Figure 2.7. Structure of RdRp showing the Fingers, Palm and Thumb domains of the enzyme. Reproduced from Gao et al. (2020)

With the exception of retroviruses, RNA viruses depend on RdRp for the replication and transcription of their RNA. RdRp is considered to be highly conserved within coronaviruses (Aftab et al, 2020), making drug discovery efforts for the inhibition of RdRp possibly useful in the treatment of other diseases caused by other coronaviruses.

One of the first antivirals against COVID-19 that has been approved by the Food and Drugs Agency (FDA) in the USA is Molnupiravir. This broad-spectrum antiviral is metabolized into a ribonucleoside analogue (5′-triphosphate NHC) that resembles cytidine and binds preferentially to the viral RdRp. RdRP then incorporates 5′-triphosphate NHC into the viral RNA which creates a number of mutations that result in lethal mutagenesis or viral error catastrophe (i.e. the mutation frequency is too high, resulting in a sharp drop in viral viability) (Wahl et al., 2021). This is the mechanism of action of Ribavirin in the treatment of Hepatitis C (Anderson, Daifuku, and Loeb, 2004). Another compound (AT-511) with a similar mode of action is in the advanced clinical trials (Good et al., 2021). Nevertheless, only a limited number of studies identifying possible RdRp inhibitors have been published, making it harder to identify ligands that could be used in virtual screening experiments.

2.6 Virtual Screening

The chemical space for molecules below a molecular weight of 500 Da is estimated to be 10^{200}, with the number of drug-like compounds being estimated in excess of 10^{60} (Bohacek, McMartin and Guida, 1996). The 500 Da value is a commonly used cut-off for oral absorption of compounds (Bickerton et al., 2012), combined with other factors that will be described further below. A common assumption is that molecules that share similar structures and properties, tend to have similar biological activities, moreover, molecules lying in close proximity within the chemical space tend to have similar functionalities (Westermaier, Baier and Scapozza, 2015).

Virtual Screening (VS) is the use of computational tools for the screening of large databases of small molecules to identify compounds that elicit the desired biological effect. vHTS has become an essential process in a drug discovery pipeline (Sliwoski et al., 2014). Virtual Screening (VS) can be categorized into two main computational approaches, Ligand-Based Virtual Screening (LBVS) and Structure-Based Virtual Screening (SBVS) (Kucera, 2016) (Figure 2.8).

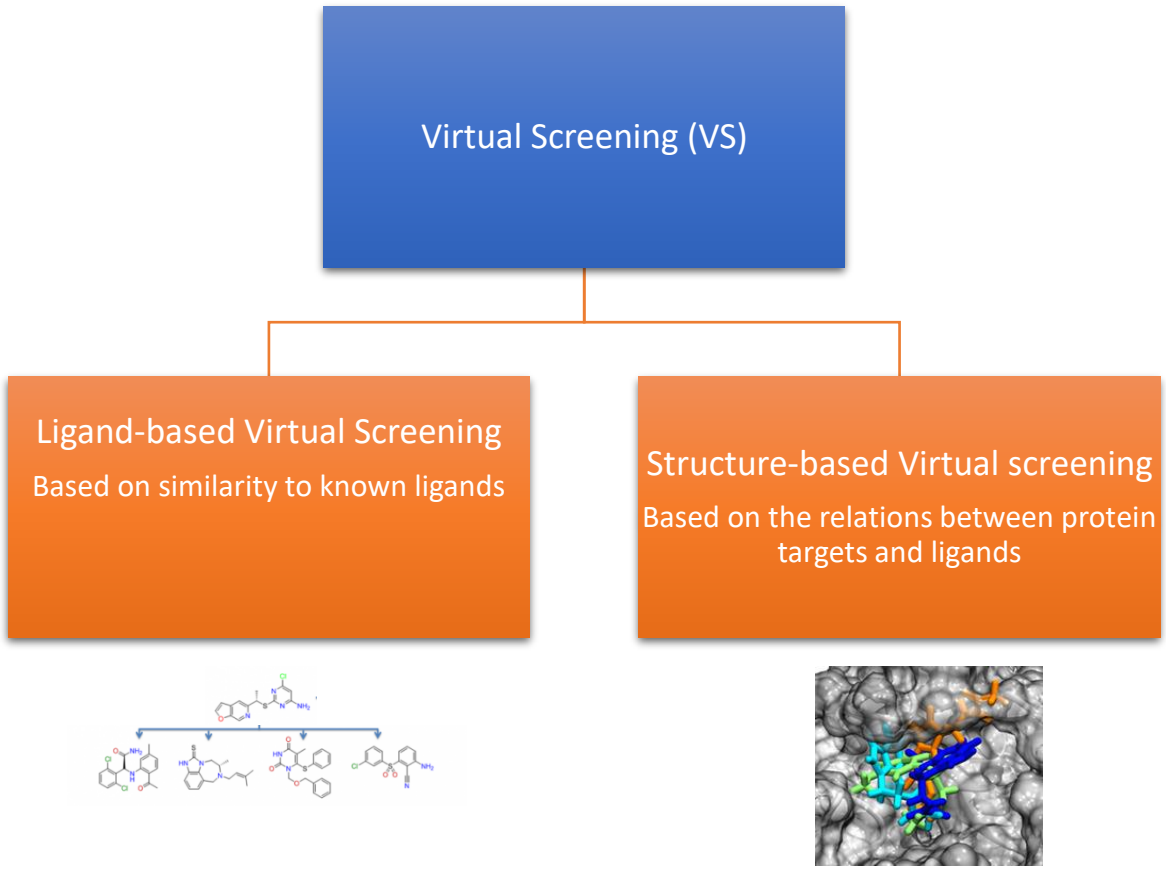


Figure 2.8. Illustration showing the main two branches of Virtual Screening techniques - ligand-based and structure-based.

2.6.1 Ligand-Based Virtual Screening

LBVS involves the use of molecular descriptors from a group of compounds that are known actives to search for other similar active compounds (Berenger, Vu and Meiler, 2017). LBVS methods are particularly useful when the precise 3D structure of the target protein is unavailable. This approach requires a lower computational effort (when compared to SBVS), nevertheless, methodologies such as similarity searches provide an efficient approach to provide potential leads that could be considered further studied in a drug discovery project (Macalino et al, 2015). An important assumption made by LBVS methods is that structurally similar compounds share similar biological interactions with the target protein, which is known as the Similar-Structure, Similar-Property Principle (Johnson and Maggiora, 1990).

There is a multitude of LBVS methods, including Quantitative Structure-Activity Relationship (QSAR) methods, pharmacophore modelling, fingerprints and Ultrafast Shape Recognition (USR), and ligand shape (volume)-based methods (Lavecchia and Di Giovanni, 2013; and Hamza, Wei and Zhan, 2012). For the purpose of this study, the two most relevant LBVS methods are the following:

2.6.1.1 Ultrafast Shape Recognition with CREDO Atom Types

Ultrafast shape recognition (USR) is an interesting LBVS method because it provides the advantage of having information about the shape of a molecule condensed into a 12-element decimal vector descriptor, making it very efficient to search for similarity among compounds based on this descriptor (Ballester, Finn and Richards, 2009). It is important to note that USR is an alignment-free LBVS method, which means that the method does not need to perform the computationally resource-intensive step of superposing the molecules with each other, (both target and query) since the shape information is compressed into rotationally and translationally invariant numerical form (Ballester and Richards, 2007).

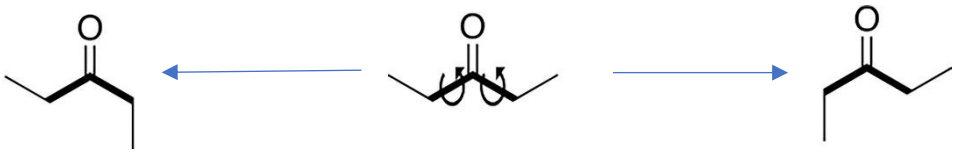


Figure 2.9. Different conformers of pentan-3-one.

As a result of the rotatability of single bonds, molecules can exist in different conformations, as shown in Figure 2.9, where three conformers of the same molecule are illustrated. Each conformer has different free energy, with the conformer holding the lowest free energy being the most dimensionally/structurally favourable molecule. However, the desirable binding of a ligand to the binding pocket of a target protein is affected by the shape of the conformer, with only a limited number of conformers being able to bind effectively. Since conformers with high free energy are less likely to occur in nature, the lowest energy free energy conformer is selected for USR. This lowest-energy (conformer) is parameterized using Force Field methods, such as the Molecular Mechanics Force Field (Ballester, Finn and Richards, 2009). While the actual similarity search in USR methods is computationally inexpensive, the derivation of the lowest energy conformer for each molecule in a dataset is an intensive process.

Since its inception, several USR-like methods have been developed, which brought improvements to USR. Such methods include USR-VS, Ultra-Fast Shape Recognition with Atom Types (UFSRAT), and Ultrafast Shape Recognition with CREDO Atom Types (USRCAT) (Banegas-Luna, Cerón-Carrasco, and Pérez-Sánchez, 2018).

One major disadvantage of USR is that this algorithm does not discriminate between pharmacophoric features, since it is indifferent to atom types. USRCAT is an algorithm that was developed to address this limitation. In order to include pharmacophore information in the

USR descriptor, USRCAT calculates additional USR moments which represent atomic subsets that are classified as either hydrophobic, aromatic, hydrogen bond donor or acceptor, a methodology used for atomic classification in the CREDO database. This results in a vector containing an additional 48 moments to the original 12 USR moments (Schreyer and Blundell, 2012).

2.6.1.2 Fingerprints

Fingerprints are another popular category of molecular descriptors that can be used for LBVS. These can be divided into two large groups, the first one takes only into consideration the 2D structure (topology) of a molecule, whilst the second group also includes information on the 3D structure of the ligand in conjunction with 2D information. The efficiency of this methodology is derived from the computationally efficient abstract representation, where structural features are represented by either bits in a bit string or counts in a count vector, that facilitate molecular comparisons (Riniker and Landrum, 2013).

The first fingerprints were based on the Morgan algorithm, which is a process whereby numerical identifiers are assigned to each atom iteratively. Initially by encoding an initial atom identifier, which is the numbering invariant atom information, and then iterating through the identifiers until all the atom identifiers are as close to unique as possible. The intermediate results are discarded, leaving canonical identifiers for the atoms (Morgan, 1965).

Extended-Connectivity Fingerprints (ECFP) is an improvement over Morgan fingerprints, and like Morgan fingerprinting, they provide a method that captures molecular features relevant to a molecular activity by using a circular fingerprint. The main advantages of this method are that the ECFPs are easily interpretable, can be calculated rapidly, can represent a large number of different molecular features, and are flexible (Rogers and Hahn, 2010). The major differences between ECFP and the Morgan algorithm are:

- the number of iterations in the ECFP generation process is limited;
- all identifier information, including the intermediate atom identifiers, is collected, with the identifiers generated following each iteration being collected into a set;

- in ECFP the identifier generation process includes a fast-hashing scheme, making it more computationally efficient.

The ECFP Generation Process (represented in Figure 2.10.) consists of four stages:

1. Initial stage – all atoms are assigned a unique integer identifier generated by using the Daylight atomic invariants rule – i.e. a set of seven properties related to the respective atom, which are i. number of non-hydrogen immediate neighbours, ii. Bond order without the inclusion of hydrogen bonds, iii. Atomic number, iv. Atomic mass, v. atomic charge, vi. number of bonded hydrogens, and vii. Ring membership;
2. Update stage – For each atom, the identifier of the atom itself together with the identifiers of its immediate neighbours, are collected into an array and a hash function is applied to the array to condense it into a single-integer identifier. Essentially, this step condenses the information related to the atom's neighbourhood. This process is repeated iteratively to account for the number of atomic neighbours (i.e. atomic diameter) that is user defined;
3. Deduplication stage – multiple occurrences of the same feature are removed so that only a single representative identifier is left in the final feature list;
4. Bit vector generation stage – the final feature list is converted into a bit vector.

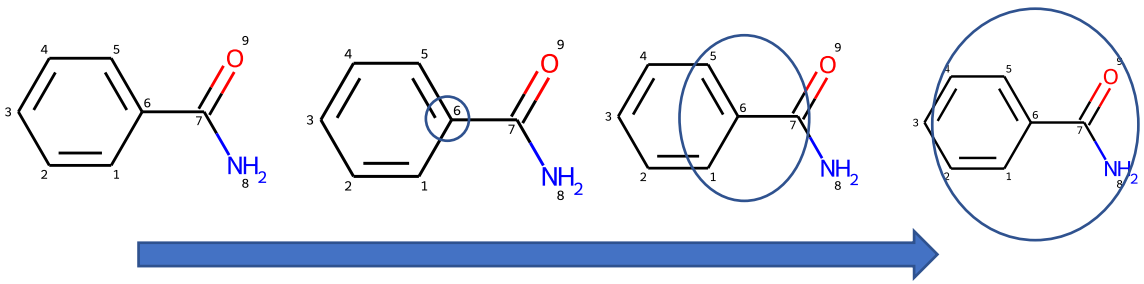


Figure 2.10. Iterative update of the information for an atom identifier, atom 6 in benzamide. At each iteration, the neighbourhood being considered is increased until the user-defined diameter, in this case, 2, is reached. The identifiers are then hashed into a bit vector.

2.6.2 Structure-Based Virtual Screening

In contrast to LBVS, SBVS methodologies rely on information of both protein target and ligand in order to evaluate the interactions between these two molecular structures. The basis for such assessments is the assumption that the interaction of a molecule with a protein target site results in a (desirable) biological effect (Sliwoski et al., 2014).

Molecular Docking is the most commonly applied method of SBVS (Cheng et al., 2012). Leach and Gillet (2007) explain that the "aim of a docking experiment is to predict the 3D structure (or structures) formed when one or more molecules form an intermolecular complex".

Molecular docking experiments follow a two-step process, the first one being the generation of a number of molecular poses of the ligand molecule into the binding site of the protein target, and the second being the actual scoring of the poses (Torres et al., 2019). The outcome of a molecular docking study is a score or estimate of the binding affinity between the different conformations of small molecule with the binding pocket of the target protein (Cheng et al., 2012; Macalino et al., 2015). The most crucial aspect of such experiments is the accurate scoring of the binding and ranking of docked compounds.

The scoring functions may be based on the estimation of the binding energy, free energy, or interaction energies and fall into three types – force field calculations (summation of energy terms), empirical (based on quantitative structure-activity relationships), and knowledge-based functions (based on the experimentally observed interactions of atom pairs) (Torres et al., 2019).

2.7 Molecular Databases

Computational lead generation efforts can be broadly categorized into two approaches, the *de novo* design of candidate molecules or the virtual screening of molecular databases. Whilst the first approach provides the flexibility of designing 'ideal' molecules, it presents two key challenges. There is no guarantee that the designed compound can be readily synthesized and being a completely new molecule, there would not be any information available on the molecule, including its ADMET (Absorption, Digestion, Metabolism, Excretion, and Toxicity) properties, albeit this is also true for many virtual screening hits. The virtual screening of (curated) molecular databases is by far the most common approach used in lead generation, however, it also presents its own challenges when compared to *de novo* design. One of the main challenges is that a number of widely used databases run into the millions of molecules, which would necessitate a proportional amount of computational resources should all molecules be screened, akin to finding a needle in a haystack (Raymond et al., 2010).

The effectiveness of a virtual screening pipeline is highly dependent on the quality of the compound database that is used for the computational experiments. Table 2.1 contains a

list of commonly used public chemical databases, including for virtual screening experiments. These repositories contain a vast number of small molecules, however, for the purpose of increasing the efficiency of screening experiments, such databases are usually pre-processed through the application of a number of filters. The pre-processing can involve the removal of duplicates, removal of compounds containing chemically reactive groups, removal of compounds containing undesirable atoms, the use of property filters, etc (Saxena and Prathipati, 2006).

Table 2.1. Public databases of chemical compounds

Database	No of compounds	Link
Pubchem	110 million	https://pubchem.ncbi.nlm.nih.gov/
ChEMBL	2.1 million	https://www.ebi.ac.uk/chembl/
ChemSpider	103 million	http://www.chemspider.com/
ZINC	230 million	http://zinc.docking.org/

The most commonly used filtering strategies are the following:

Lipinski's Rule of Five – This is the most common filtering strategy reported in the literature. The filtering strategy is based on a number of numerical filters that are multiples of five, hence the name Rule of Five (Lipinski et al., 1997; Pollastri, 2010). This filtering strategy predicts that poor (oral) absorption or permeation is more likely when there are more than five H-bond donors, ten H-bond acceptors, the molecular weight is greater than 500, and the calculated octanol-water partition coefficient (CLogP) is greater than five. An effectively absorbed molecule, ingested via the oral route, does not typically violate more than one of these conditions. Nevertheless, there are well-known exceptions to this rule, such as many natural products, whereby many bioactive molecules have 'evolved' to have low hydrophobicity and intermolecular H-bond donating potential in order for them to be absorbable (Ganesan, 2008).

Veber Filter – This filtering strategy assumes that molecular flexibility, in terms of the number of rotatable bonds, and polar surface area/number of hydrogen bonds are good predictors of bioavailability, independent of molecular weight (Veber et al, 2002). This filtering strategy filters out compounds that have more than ten rotatable bonds and a polar surface area of more than 140 $Å^2$ (or more than 12 H-bond donors and acceptors).

Ghose Filter - Ghose, Viswanadhan and Wendoloski (1999) analysed the Comprehensive Medicinal Chemistry (CMC) database (v. 97.1) and seven subsets from the CMC database and developed a consensus definition of a drug-like molecule from a physicochemical and structural perspective; and based on a comparison between the property distribution in different (drug) subsets, developed a filtering strategy based on the following physicochemical properties:

- calculated log P within the range of -0.4 to 5.6;
- molecular weight within the range of 160 and 480;
- molecular refractivity between 40 and 130; and
- atom count between 20 and 70.

Rapid Elimination of Swill filter - Rapid Elimination Of Swill (REOS) is a filtering strategy that deploys an initial filtering that is based on physicochemical properties and is then followed by the filtering out of compounds containing chemical functional groups known to be problematic (such as long aliphatic chains). The filtering strategy allows the flexibility of selecting which parameters can be accepted to fail. The following physicochemical parameters are the initial filtering properties:

- Molecular weight between 200 and 500
- Calculated LogP between -5.0 and +5.0
- H-bond donor count between 0 and 5
- H-bond acceptor count between 0 and 10
- Formal charge between -2 and +2
- Rotatable bond count between 0 and 8
- Heavy atom count between 15 and 50

Quantitative Estimate of Drug-likeness – Unlike the other drug filters, which rely on a number of pass or fail limits related to physicochemical properties, the Quantitative Estimate of Drug-likeness (QED) approach relies on the estimate of the physicochemical properties in terms of a desirability function reflecting how desirable that property is in a drug-like compound. The individual property estimates are then integrated into a single value, ranging from 0 to 1, with one being the most drug-like (Bickerton et al, 2012). The authors indicated that when applying QED to ChEMBL (Gaulton et al, 2017), the best cluster of human drug targets was obtained when applying a QED cut-off of 0.693, and 0.766 for the best cluster of

oral drugs. The user can set an arbitrary cut-off to filter out compounds based on empirical values or the user's experience.

Drug-like filter – A further filter for drug-likeness is implemented in ChemAxon products and implemented in the Squonk Computational Notebook[3] that is loosely based on the Murcko drug filter (Walters, Murcko and Murcko, 1999) and the Oprea drug filter (Oprea, 2000). The following are the rules used by this filter:

- molecular mass < 400
- ring count > 0
- rotatable bond count < 5
- hydrogen bond donor count ≤ 5
- hydrogen bond acceptor count ≤ 10
- c log P < 5

2.8 Molecular clustering

Clustering is the process of grouping objects (such as molecules) into clusters, whereby the data in each cluster shares similarities with other members of the cluster while being dissimilar from the members of other clusters. One key point of clustering data into clusters is the minimization of the intercluster similarity while maximizing the intracluster similarity (Young, 2009). When members of a cluster are allowed to be members of another cluster, this is known as soft clustering, while hard clustering is when cluster members can only be assigned to one cluster.

Clustering algorithms can be grouped into two broad categories, hierarchical and non-hierarchical/ partition-based algorithms. Hierarchical clustering involves the successive partitioning of data into clusters that are further divided into a larger number of clusters from one step to the other. In agglomerative hierarchical clustering, each element of a dataset is a separate cluster, which are merged successively into larger clusters, hence a bottom-up approach, on the other hand, in divisive hierarchical clustering the entire dataset is grouped as a single cluster which is further divided into smaller clusters.

[3] Squonk Computational Notebook. https://squonk.it/docs/cells/Drug-like%20Filter%20(CXN)/ accessed on 5 August 2021

Non-hierarchical clustering involves the generation of new clusters by merging or splitting clusters without following a particular order. Non-hierarchical clustering is sub-divided into four categories of methods (Cassar, 2019):

1. Density-based algorithms – clustering based on the density distribution of a particular property, with data points falling within a pre-defined density cut-off being clustered together (e.g. DBSCAN);
2. Relocation algorithms – data points are moved from one cluster to the other in order to increase the intragroup similarity for a particular property (e.g. *k*-means);
3. Nearest neighbour algorithms – the distance of each datapoint from one another is calculated and datapoints are clustered on the basis of this distance, usually in accordance with a pre-defined cut-off (e.g. Jarvis-Patrick);
4. Single pass algorithms – data points are assigned to a particular cluster on the basis of a particular threshold, with the ultimate aim of assigning membership of each datapoint in a single step (e.g. Leader/Taylor-Butina and Sphere Exclusion).

The Jarvis–Patrick clustering algorithm (Jarvis and Patrick, 1973) creates a set of non-overlapping clusters by defining a number of neighbours in common and a distance metric. Two points are considered to form part of the same cluster when they share the defined number of neighbours and therefore fall within the defined distance. This is a deterministic algorithm and therefore will result in the same cluster members for the same starting set.

In the k-means (Loyd, 1982) clustering algorithm, k number of clusters are generated, where k is a pre-defined value. Randomly selected k number of points are generated and serve as cluster centroids, with surrounding data points being allocated to the cluster. This process is iterated until centroids do not shift and therefore this method is non-deterministic.

Taylor-Butina clustering (Butina, 1999, and Taylor, 1995) and Sphere exclusion (Hudson, et al., 1996) clustering are examples of single pass algorithms. In sphere exclusion, three sets are used to cluster the data points. The first set contains singletons (elements that could not be clustered), which initially consists of all elements of the dataset, the second set collects all the cluster centroids and is therefore initially empty, and the third set contains the clusters and its assigned cluster members. Clustering through this algorithm is heavily influenced by the initial choice of centroids, therefore the cluster centroid identification is a non-deterministic process, and to try and introduce order in the centroid selection process, Butina included a pre-processing step in the overall algorithm.

2.9 Related Work - Inhibitor Identification from Literature

A Google Scholar search was conducted between 21st and 29th July 2021 using the keywords "COVID-19 *in silico* inhibitors", "SARS-CoV-2 *in silico* inhibitors", and interchanging the term "inhibitors" with "virtual screening", as well as omitting the term "*in silico*". A total of 32 publications (Appendix II) were identified. The publications were used to:

- identify strategies used for identifying new leads in drug development against SARS-CoV-2;
- protein targets used;
- develop a chemical library of potential leads identified via *in silico* methods;
- develop a library of *in vitro* tested compounds.

A total of 19 publications were deemed to be relevant in terms of *in silico* identification of hits against SARS-CoV-2. A summary of the findings is presented in Table 2.2. Table 2.3 provides a summary of the reviewed studies.

Table 2.2 Reviewed studies

Target	Number of studies	Number of studies containing *in vitro* data (ELISA-based methodologies or competitive inhibition methods)	Number of hits	Number of compounds tested *in vitro*
M^{pro}	13	4	194	19
PL^{pro}	3	1	15	3
S-protein	5	1	44	1
RdRp	2	1	13	3
TMPRSS2	1	-	10	-

Forrestall *et al.* (2021) performed a molecular docking experiment to predict the binding energies and inhibition constants (K_i values) of a number of 2-pyridone containing natural compounds against M^{pro} (PDB: 6WQF). The ADMET properties of the compounds showing the strongest binding affinities were then evaluated. However, it is relevant to mention that the results that were obtained were not validated using any *in vitro* methodology. This was a frequent pattern observed throughout the studies that were evaluated. This was a similar

finding to that published by Macip *et al.* (2021). The authors reviewed 61 peer-reviewed publications that reported at least one docking step to identify SARS-CoV-2 M^{pro} inhibitors. Of these 61 studies, only two tested their *in silico* predictions using *in vitro* methodologies, and therefore lacked a robust validation of the findings. Moreover, when the authors assessed the use of re-docking (removing the ligand from a protein-ligand complex to re-dock it and assess whether the initial binding pose can be re-obtained) or self-docking as a verification step, they noticed out of the 18 studies that used re-docking, only two studies modified the ligand in order to lose the binding coordinates prior to performing the re-docking. An alternative validation methodology in molecular docking experiments can be the utilization of sets of active and inactive compounds as controls.

The majority (~ 58%) of studies reviewed involved an initial molecular docking step, followed by a second *in silico* validation step, most often using Molecular Dynamics. This was the approach employed by Ogedjo et al (2021), Kumar et al (2020), Tejera et al (2020), Gurung et al (2021), Puttaswamy et al (2020), Cuesta, Mora and Marquez (2021), Pitsillou et al (2020), Pitsillou et al (2021), Reyaz et al (2021), Gangadevi et al (2021) and Basu et al (2020).

Of the reviewed manuscripts, only four studies made use of computational techniques based on artificial intelligence approaches. Gaudêncio and Pereira (2020) used a random forest machine learning approach in order to develop a QSAR classification model that was subsequently used to select potential hits from three chemical libraries of marine natural products (MNP) that resulted in 494 MNPs being selected. The best performing fifteen MNPs were then assessed through molecular docking (Autodock Vina (Trott and Olson (2010)) against M^{pro} (PDB: 6LU7) followed by an assessment of ADMET properties. Using a dataset of ~ 1.6 million drug-like small molecules obtained from the ChEMBL database, Bung et al (2021) trained a deep neural network-based generative model. This generative model was then trained to develop novel drug-like small molecules. Using transfer learning, followed by reinforced learning, the authors used the optimized model to generate a number of molecules that could inhibit M^{pro}. Molecular docking was then used to rank and validate the generated molecules.

Tejera et al (2020) developed a QSAR model based on machine learning, using graph convolutional networks (molecules were converted to molecular graphs). The authors created a dataset of molecules from ChEMBL, with reported activity against M^{pro} and polyprotein 1ab, they then used an IC_{50} cut-off of 10μM to distinguish between inhibitors and non-inhibitors.

Table 2.3 Summary of reviewed studies

Study	Target protein	Methods used	AI methodologies	PDB reference
Forrestall et al (2021)	M^{pro}	i. Molecular docking ii. ADMET	N	6WQF
Gaudêncio and Pereira (2020)	M^{pro}	i. QSAR – machine learning ii. Molecular docking	Y - QSAR	6LU7
Ogedjo et al (2021)	RBD-ACE2	i. Molecular docking ii. Molecular dynamics	N	6M17
Bung et al (2021)	M^{pro}	i. De novo design of small molecules based on deep neural network-based generative and predictive model ii. Molecular docking	Y	6LU7
Kumar et al (2020)	M^{pro}	i. Molecular docking ii. Molecular dynamics + Structure-activity relationship	N	6LU7
Manandhar et al (2021)	M^{pro}	i. In vitro screening ii. Molecular dynamics	N	7B3E
Tejera et al (2020)	M^{pro}	i. QSAR – machine learning ii. Molecular docking + Molecular dynamics	Y - QSAR	6Y2G
Hajbabaie, Harper and Rahman (2021)	PL^{pro}	i. LBVS (scaffold hopping) ii. Molecular docking	N	7JRN
Gurung et al (2021)	RBD-ACE2	i. Molecular docking ii. Molecular dynamics	N	6M0J
Liu and Wang (2020)	M^{pro}	i. Molecular docking	N	6LU7
Jin et al (2020)	M^{pro}	i. Molecular docking ii. In vitro screening	N	6LU7
Puttaswamy et al (2020)	TMPRSS2 RBD-ACE2 M^{pro} RdRp	i. Molecular docking ii. QSAR	N	- 6M0J 6LU7 -
Cuesta, Mora and Marquez (2021)	M^{pro} PL^{pro} RdRp	i. QSAR based on experimental data (pIC_{50}) ii. Molecular docking + Molecular dynamics	Y – K-means clustering	6LU7 6WUU 7BV2
Pitsillou et al (2020)	M^{pro}	i. Molecular docking ii. Molecular dynamics iii. In vitro	N	6LU7 6Y2G 6M03
Pitsillou et al (2021)	PL^{pro}	i. Molecular docking ii. Molecular dynamics iii. In vitro	N	6XAA 6W9C 6WUU 6WX4 7JRN
Reyaz et al (2021)	M^{pro}	i. QSAR ii. Molecular docking + Molecular dynamics	N	6YB7
Gangadevi et al (2021)	RBD-ACE2	i. In vitro ii. Molecular docking + Molecular dynamics	N	6M0J
Basu et al (2020)	M^{pro}	i. Molecular docking ii. Molecular dynamics	N	6LU7
Day et al (2021)	RBD-ACE2	i. Molecular docking ii. Surface plasmon resonance iii. In vitro	N	6VW1

The dataset was then used to train the graph convolutional network-type classifier model. The model was then used to screen molecules found on the DrugBank database (Wishart et al,

2018) as an initial virtual screening step to obtain a list of 20 top-performing hits which were then assessed using molecular docking and molecular dynamics.

Cuesta, Mora and Marquez (2021) used experimentally verified pIC50 data from a dataset of structurally diverse molecules to develop a QSAR model. The QSAR model was developed using K-means cluster analysis, using a training set of 35 molecules and a test set of 9 molecules. The model was then used to predict the pIC50 values of compounds found in the DrugBank database (using four major datasets). The best performers were then used in molecular docking and molecular dynamics simulations against M^{pro}, PL^{pro} and RdRP, in order to assess the model performance.

2.10 Evaluation Criteria

The calculation of the Enrichment Factor (EF) is one of the methods that are available to assess the quality of the results obtained from the LBVS experiments (Mishra and Basu, 2013). In order to obtain the enrichment factor, the lists of molecules obtained from the similarity searches are ranked by order of similarity and are seeded with known active molecules. A user-defined percentage cut-off is selected, and the number of true actives recovered from this user-defined sub-set is used to estimate the EF as shown in Equation 2.1. Essentially, the enrichment factor is the ratio of the fraction of actives in the (user-defined) subset of molecules relative to the fraction of actives in the original dataset.

$$EF_{x\%} = \left(S_{act}/S_{all}\right) \Big/ \left(P_{act}/P_{all}\right)$$

Equation 2.1 $EF_{x\%}$ is the Enrichment Factor at the percentage cut-off x%, P_{all} is the total number of molecules in all of the list, S_{all} is the number of molecules in the subset chosen with the x% cut-off, P_{act} is the number of active molecules in all of the list, and S_{act} is the number of actives found in the subset.

The EFs obtained for all the ligands used for the similarity searches are used to obtain an average enrichment factor.

2.11 Summary

Through this chapter, we have provided a comprehensive overview of the biology of Coronaviruses, in particular SARS-CoV-2, and protein targets that are relevant in the context of Virtual Screening for the discovery of SARS-CoV-2 inhibitors. We have then provided an

overview of Virtual Screening methodologies and the major steps undertaken in performing this ligand-based virtual screening study, which are:

- Selection of suitable ligands;
- Preparation of lists of molecules (including different filtering strategies) to use in similarity searches;
- Clustering of molecules;
- LBVS methods;
- Use of Enrichment Factors to assess the results of this study.

3

Methodology

This chapter details the methods and approaches to be able to run ligand-based virtual screening experiments using filtered lists of small molecules, using a number of putatively active inhibitors identified from the literature as queries.

3.1 Approach Overview

The overall approach used in this study is outlined in Figure 3.1 and it is subdivided into a number of functional parts, which will be further elucidated below.

1. Protein target identification. A list of protein targets relevant to SARS-CoV-2, that are of interest for drug discovery purposes is generated through a literature review. The protein target that is most suitable for the purpose of conducting the virtual screening experiments shall be chosen on the basis of an evaluation of the desirable properties of the respective protein target.
2. Inhibitor identification. Through the literature review process referred to previously, a list of small molecules identified as inhibiting the function of the selected protein target is compiled.

NOTE: The first two steps are considered to both form part of the literature review process.

3. Curation of small molecule list. The drug-like portion of ZINC15 (Sterling and Irwin, 2015) was downloaded[4] locally using the "2D" and "in-stock" commercial availability subsets. The small molecules were then standardized and any duplicates were removed.
4. Database creation. In order to have a single repository of information that can be easily retrieved and used in further experiments, the lists of small molecules, list of protein targets, and list of inhibitors were then loaded into a database using POSTGRESQL. This provides the opportunity

[4] https://zinc15.docking.org/ last accessed on 21 July 2021

of running the same experiments using the identified inhibitors for the remaining protein targets as part of future work.

5. Small molecule filters. For the purpose of removing small molecules possessing undesirable properties with respect to their drug-like properties, a number of filtering strategies were applied to the table of small molecules, and to the table of inhibitors. This also ensured that subsets of molecules and inhibitors that were subjected to the same filters were used for the purpose of the virtual screening experiments. This, for example, avoids the possibility of using inhibitors with a molecular weight of 500, in the case of Lipinski's Rule of Five, which is especially important when considering that the molecules in the dataset downloaded from ZINC15 have a molecular weight lower than 500.

   The filtered molecules were then clustered in order to have a representative and diverse more manageable subset of small molecules.

6. Similarity searches were performed to identify molecules sharing similarities with the inhibitor molecules used as queries. These molecules can then be further evaluated in terms of their bioactivity against the protein target of interest and their ADMET properties. Two distinct methodologies were utilized for performing the similarity searches in order to assess the results obtained from two different methodologies. Moreover, having the same molecules obtain a high ranking in both methods, would have served as an internal validation of the results.
   a. Molecular Fingerprint similarity search. The Morgan fingerprint was obtained for each cluster centroid and filtered inhibitor and a similarity search was performed between the two sets to derive the respective Tanimoto score and Dice score.
   b. USRCAT search. The lowest energy conformers for the cluster centroids and filtered inhibitors were obtained, and similarity based on the USRCAT score was performed.
7. Creation of (two) lists of candidates for further evaluation, through for example *in vitro* studies, derived from the above-mentioned similarity searches.

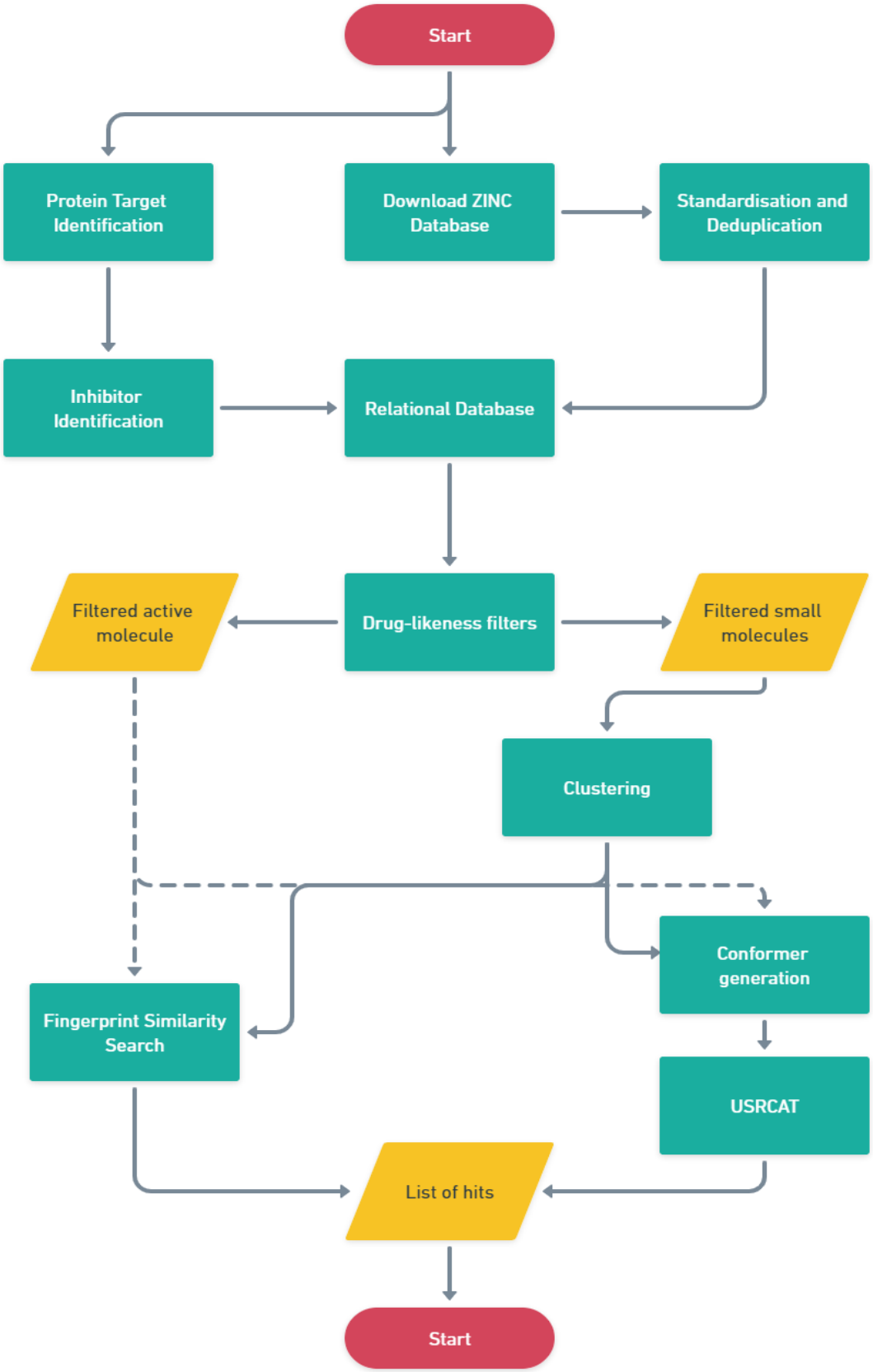


Figure 3.1 Flowchart showing the steps involved in the study

3.2 Protein Target and Inhibitor Identification

The literature review was undertaken by searching Google Scholar, between the 21st and 29th July 2021. The following were the terms used for the search:

- "COVID-19 *in silico* inhibitors";
- "SARS-CoV-2 *in silico* inhibitors";
- "COVID-19 *in silico* virtual screening"; and
- "SARS-CoV-2 *in silico* virtual screening".

The searches was also done without the term "*in silico*".

A total of 39 publications (Appendix 1) were identified as being relevant to this study. The publications were used to:

- identify strategies used for identifying new leads in drug development against SARS-CoV-2;
- protein targets used;
- develop a chemical library of potential actives identified via in silico methods; and
- develop a library of in vitro tested compounds.

The identified studies were reviewed in order to catalogue the following information:

1. protein targets addressed within the study;
2. hits that were identified;
3. *in silico* method used for hit identification;
4. *in vitro* validation of results;
5. use of AI-based methodologies; and
6. PDB reference of the protein target

A summary of the above-mentioned information is presented in Table 2.2. M^{pro} was selected as the protein target of interest for this virtual screening study for the following reasons:

- the amino acid sequence specificity for its binding, that is not found in other known human proteases;
- the larger volume of studies that have studied M^{pro};
- the larger number of identified hits that have been shown to inhibit this protease; and
- the ready availability of a good protein structure on PDB.

3.3 Development of a List of Commercially Available Small Molecules

ZINC is a publicly available database and toolset that was first developed to provide a freely available source of compounds that can be used in virtual screening experiments. In its latest iteration, ZINC15, the creators of this database have implemented an annotation system, that provides information on the

biological association, target, and commercial availability of the compounds found in the database (Sterling and Irwin, 2015). On the date of publication (2015), ZINC15 contained around 120 million compounds. As of 21 July 2021, it has grown to a database of over 997 million compounds (750 million of these are searchable analogs). For the purpose of this study, the portion of 2D compounds that were categorized as in-stock and as drug-like, was downloaded locally (on 21 July 2021) from ZINC15[5]. This was done by selecting the respective tranches in ZINC15, selecting the SMILES file format for the output files, and the download method, which in this case was an executable script. This resulted in 6,877,387 molecules being downloaded, divided into 720 .smi files, from ZINC15. Each file contains a header with each line containing the SMILES formula and ZINCID of the molecule.

3.3.1 Standardization

The downloaded molecules were standardized so that the molecules are represented in a 'standard form' that allows a harmonized treatment of molecules and prevents issues when processing molecules for virtual screening through RDKit[6]. This standardization was performed through RDKit and involves the following processes (Landrum, 2021):

1. The clean-up (normalisation) of substructures to their standardized form (e.g. -[S+2]([O-])[O-] to -S(=O)=O).
2. The calculation of implicit and explicit valence of all atoms to check that valence rules are obeyed, such as carbon with a valency higher than four.
3. Neutralization of charges followed by reionization.
4. Identification of parent fragment.
5. Breaking of bonds to metal atoms.

The above step was performed on each separate .smi file and the resulting standardized molecules were saved into a file with extension .stnd.smi, with each line containing the SMILES formula and ZINCID of the molecule.

3.3.2 Deduplication

As a follow-up to standardization, a deduplication step was performed in order to remove molecules that are listed more than once. This step is required due to the fact that the standardization step might have resulted in duplicates of molecules being generated. For example, if the original pre-standardization list contained different salt forms of the parent molecule (e.g. Na^+ and K^+ salt of the same organic parent molecule), standardization would result in the same SMILES formula being generated by the standardization step. Deduplication would preserve only one copy of the molecule.

[5] https://zinc15.docking.org/tranches/home/ accessed on 21 July 2021
[6] RDKit: Open-source cheminformatics; http://www.rdkit.org accessed on 15 July 2021

Deduplication is a step that is required to make the following computational steps more efficient by preventing processing the same molecule more than once, as well as reducing the memory requirements for the database and subsequent processing steps.

InChI (Heller, et al., 2015) is the International Chemical Identifier developed under the auspices of the International Union of Pure and Applied Chemistry. The InChI identifier is a machine-readable text string abstraction of the molecule obtained through a three-stage process (normalization, canonicalization, and serialization). The InChIKey is a compact hashed code that is calculated from the InChIKey (Heller, et al., 2015) and each key is unique to a molecule. The InChIKey is derived by condensing the InChI identifier into a 27-character key (the first 14 characters represent information about the core molecular skeleton, including the molecular formula; the next eight characters represent molecular features, such as tautomerism, stereochemistry, and metal ligation; the next two characters encode for the type of InChI and the software version used for the generation of the InChI; and the last character represents the protonation state).

The InChIKey was obtained for each molecule in the SMILES files, using RDKit, and the result was saved in a .inc file, with each line containing the SMILES formula and InChIKey of each molecule. In hindsight, an improvement to this process would be the inclusion of the ZINC ID for each molecule since this would ensure easy retrieval of commercial information when acquiring the compounds for testing. The individual files were combined together and entries having the same InChIKey were removed. This process resulted in 6,065,672 unique molecules and 811,715 molecules being removed as duplicates.

3.4 Database Creation – DisCO

A relational database, named DisCO, was created using the PostgreSQL cartridge within the RDKit toolkit. The database was meant to contain all the information generated from the literature review and small molecule list generation. The information can be split into three categories:

- protein targets and related studies;
- small molecule inhibitors; and
- sanitized and deduplicated molecules downloaded from ZINC15.

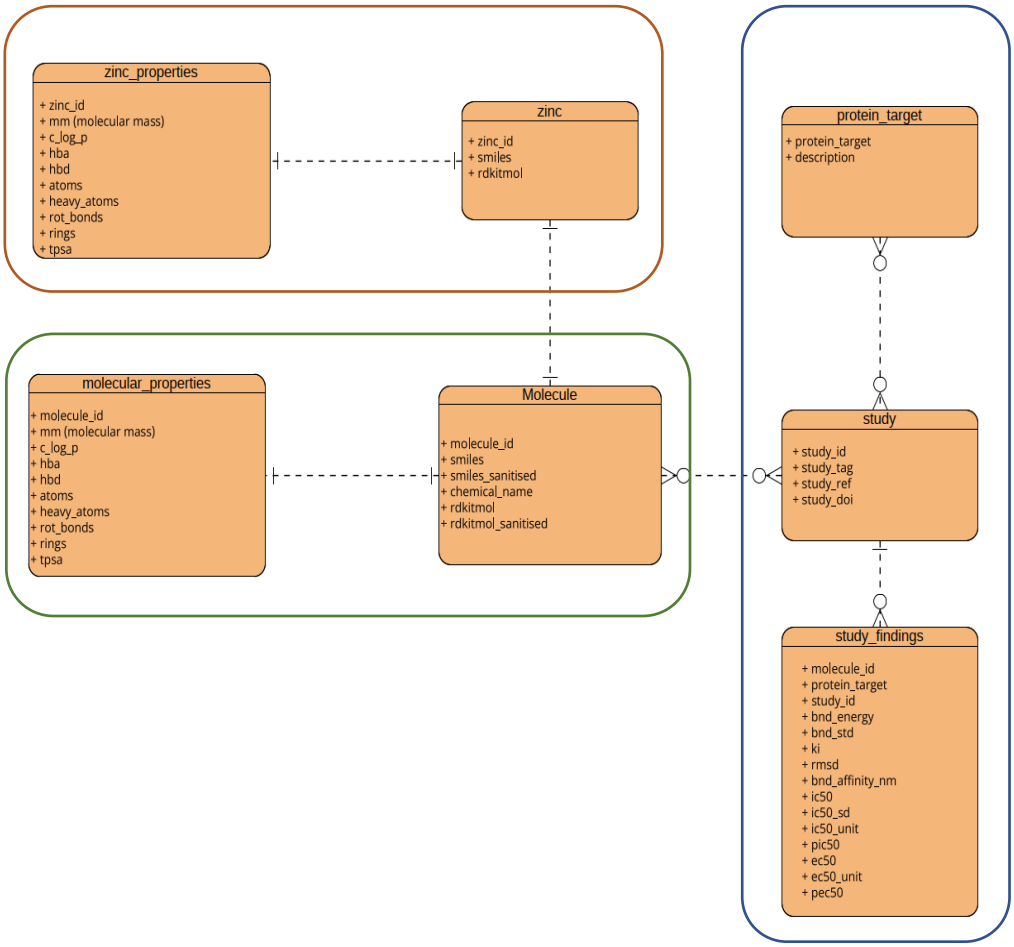


Figure 3.2 Logical Entity Relationship Diagram for the DisCO database. The orange box is the data related to the small molecules downloaded from ZINC15, the green box is the data related to the inhibitors and the blue box is the data pertaining to the protein targets and respective studies

3.4.1. Database Schema

Figure 3.2 illustrates the Entity Relationship Diagram (ERD) for the database and shows how the database is split into entities and attributes, including their relationship. Tables 3.1 to 3.7 present the physical implementation details of the database, including the relevant constraints. The database provides a large volume of information, including binding affinities and results from *in vitro* testing (IC_{50}) of the inhibitors where these were available from the studies that were assessed. The table also contains information about the molecular properties of each molecule, that were generated and used in the drug filtering step. Even though the information related to binding energies and information obtained from *in vitro* testing was not used directly in this study, this information (as well as the entire database) can be used in further work, such as in *in vitro* testing and molecular docking experiments of the hits identified from the virtual screening experiments with M^{pro}.

Table 3.1 molecule database table – This table contains information related to the inhibitors identified from the different studies. The information includes a molecular ID, the Chemical name (where available), the SMILES code, the sanitised SMILES code, and their respective mol object generated using RDKit

Attribute	Data type	Length	Constraint	Description	Example
molecule_id	[7]citext		Primary key (Unique and Not Null)	Custom six letter identifier for the molecule.	DOCCSE
smiles	varchar	1000	Not Null	A text-based canonical molecular representation.	C=CO/C=C/CSC(=S)N(CC)CC
smiles_sanitised	varchar	1000	Not Null	A text-based canonical molecular representation.	C=CO/C=C/CSC(=S)N(CC)CC
chemical_name	varchar	100		Molecule name, if available	
rdkitmol	mol[8]			Molecule instance	[bytes]
rdkitmol_sanitised	mol			Molecule instance	[bytes]

Table 3.2 zinc database table – This table contains information related to the small molecules downloaded from ZINC15. The information consists of the molecular ID, SMILES code, and its respective mol object

Attribute	Data type	Length	Constraint	Description	Example
zinc_id	citext		Primary key (Unique and Not Null)	Custom six letter identifier for the molecule.	DOCCSE
smiles	varchar	1000	Not Null	A text-based canonical molecular representation.	C=CO/C=C/CSC(=S)N(CC)CC
rdkitmol	mol			Molecule instance	[bytes]

[7] The citext datatype allows a primary key to be case-insensitive.
[8] RDKit data type representing the RDKit molecule

Table 3.3 molecular_properties database table – This table contains information related to molecular properties of the inhibitors.

Attribute	Data type	Length	Constraint	Description	Example
molecule_id	citext_mol		Primary Key (Unique and Not Null) Foreign Key (references entity molecule on update set null, on delete cascade)	Custom six letter identifier for the molecule.	DOCCSE
name	varchar	300		Molecule name, if available.	hydroxychloroquine
mm	real		Not Null	Molecular Mass.	179.91
c_log_p	real		Not Null	Log P	1.301
hba	integer		Not Null	Number of Hydrogen Bond Acceptors	3
hbd	integer		Not Null	Number of Hydrogen Bond Donors	2
atoms	integer		Not Null	Number of atoms	15
heavy_atoms	integer		Not Null	Number of heavy atoms	7
rot_bonds	integer		Not Null	Number of rotational bonds	3
rings	integer		Not Null	Number of rings	1
tpsa	real		Not Null	Topological polar surface area	66.130

Table 3.4 zinc_properties database table - This table contains information related to molecular properties of the small molecules downloaded from ZINC15.

Attribute	Data type	Length	Constraint	Description	Example
zinc_id	citext_mol		Primary Key (Unique and Not Null) Foreign Key (references entity zinc on update set null, on delete cascade)	Custom six letter identifier for the molecule.	DOCCSE
mm	real		Not Null	Molecular Mass.	179.91
c_log_p	real		Not Null	Log P	1.301
hba	integer		Not Null	Number of Hydrogen Bond Acceptors	3
hbd	integer		Not Null	Number of Hydrogen Bond Donors	2
atoms	integer		Not Null	Number of atoms	15
heavy_atoms	integer		Not Null	Number of heavy atoms	7
rot_bonds	integer		Not Null	Number of rotational bonds	3
rings	integer		Not Null	Number of rings	1
tpsa	real		Not Null	Topological polar surface area	66.130

Table 3.5 protein_target database table – This table lists all protein targets that were assessed.

Attribute	Data type	Length	Constraint	Description	Example
protein_target	char	3	Unique and Not Null	Three letter identifier for the protein target.	RBD
description	varchar	250	Not Null	A description of the protein target.	Receptor binding domain of the s-protein

Table 3.6 study database table – This table contains information related to the individual studies that were used.

Attribute	Data type	Length	Constraint	Description	Example
study_id	char	3	Primary key (Unique and Not Null)	Custom three-letter identifier for the study.	AAA
study_tag	varchar	100	Not Null	Tags aimed at categorising the studies.	MPRO, AI, DOCK, MDYN
study_ref	varchar	500	Not Null	Reference to the study	Forrestall, K.L., Burley, D.E., Cash, M.K., Pottie, I.R. and Darvesh, S., 2021. 2-Pyridone natural products as inhibitors of SARS-CoV-2 main protease. Chemico-biological interactions, 335, p.109348.
study_doi	varchar	500	Not Null	DOI for the study	https://doi.org/10.1016/j.cbi.2020.109348

Table 3.7 zinc_properties database table – This table contains the information related to the inhibitors, that was extracted from the studies that were identified.

Attribute	Data type	Length	Constraint	Description	Example
molecule_id	citext_mol	6	Primary key (Not Null) Foreign Key (references entity molecule on update set null, on delete cascade)	Custom six letter identifier for the molecule.	DOCCSE
protein_target	char	3	Primary key (Not Null) Foreign Key (references entity protein_target on update set null, on delete cascade)	Three letter identifier for the protein target.	RBD
study_id	char	3	Primary key (Not Null) Foreign Key (references entity study on update set null, on delete cascade)	Custom three-letter identifier for the study.	AAA
bnd_energy_kj_mol	real			Binding energy	1.45
bnd_std	real			Binding energy standard deviation	0.45
ki	real			Inhibition constant	0.15
rmsd	real			Root mean square deviation	0.11
ic50	real			IC50	
ic50_sd	real			IC50 standard deviation	
ic50_unit	char	2	CHECK ((ic50_unit IS NOT NULL AND ic50 IS NOT NULL) OR (ic50_unit IS NULL AND ic50 IS NULL))	Unit for the IC50 value	
pic50	real			Logarithmic value of IC50	
ec50	real			EC50	
ec50_unit	char	2	CHECK ((ic50_unit IS NOT NULL AND ic50 IS NOT NULL) OR (ic50_unit IS NULL AND ic50 IS NULL))	Unit for the EC50 value	
pec50	real			Logarithmic value of EC50	

3.5 Small Molecule Filtering

Even though the ZINC15 portion of molecules that we downloaded corresponds to the drug-like portion of the database, six small molecule filters were applied in order to harmonize the molecules that were selected and to ensure that the assessed molecules are bioavailable through the oral route. These filters were used in order to compare the results obtained from the different filtering strategies, as well as to attempt to identify a filtering strategy that could result in an

increase in the efficient use of computational resources whilst not sacrificing the effectiveness of the similarity searches.

The following filtering strategies were applied separately on the dataset of molecules that were downloaded from ZINC15 and the inhibitors identified from open literature, hereinafter referred to as actives.

Lipinski's Rule of Five – This is the most common filtering strategy reported in the literature. This filtering strategy selects molecules based on a number of rules whose values are based on multiples of five (Lipinski et al., 1997; Pollastri, 2010). The rules are the following:

- ≤ 5 H-bond donors,
- ≤ 10 H-bond acceptors,
- molecular weight ≤ 500, and
- calculated octanol-water partition coefficient (C Log P) ≤ 5.

An effectively absorbed molecule, ingested via the oral route, does not typically violate more than one of these conditions. For the purpose of this study, we applied this filter in a strict manner, i.e. no rule violation, as well as accepting one rule violation, referred to as Lipinski-relaxed.

Veber Filter – Veber et al. (2002) postulated that molecular flexibility, in terms of the number of rotatable bonds, and polar surface area/number of hydrogen bonds give a good indication of bioavailability, independent of molecular weight. The following rules were applied for this filtering strategy:

- ≤ 10 rotatable bonds, and
- Topological polar surface area ≤ 140$Å^2$

Ghose Filter - Ghose, Viswanadhan and Wendoloski (1999) developed the following filtering rules based on the analysis of the Comprehensive Medicinal Chemistry (CMC) database (v. 97.1) and seven subsets from the CMC database:

- C log P within the range of -0.4 to 5.6;
- molecular weight within the range of 160 and 480;
- molecular refractivity between 40 and 130; and
- atom count between 20 and 70.

Rapid Elimination of Swill filter – This filtering strategy relies on the initial filtering based on physicochemical parameters, followed by the filtering out of compounds containing chemical functional groups known to be problematic (such as long aliphatic chains). The filtering strategy allows the flexibility of selecting which parameters can be accepted to fail. For the purpose of this study, only the initial filtering based on physicochemical parameters was implemented:

- Molecular weight between 200 and 500
- C Log P between -5.0 and +5.0
- hydrogen bond donor count between 0 and 5
- hydrogen bond acceptor count between 0 and 10
- formal charge between -2 and +2
- rotatable bond count between 0 and 8
- heavy atom count between 15 and 50

Quantitative Estimate of Drug-likeness – This filter relies on a scoring function based on how the individual physicochemical properties affect molecular behaviour *in vivo*, which are then integrated into a single value, ranging from 0 to 1, with one being the most drug-like (Bickerton et al, 2012). The authors indicated that when applying QED to ChEMBL (Gaulton et al, 2017), the best cluster of human drug targets was obtained when applying a QED cut-off of 0.693, and 0.766 for the best cluster of oral drugs. The authors also indicated that a QED is a good value to use in the construction of a library, with a value of 0.9 being the value to use to have a high cut-off for drug-likeness. Based on these considerations, we decided to apply this filter using a QED cut-off of 0.6 and 0.9.

Drug-like filter – A further filter for drug-likeness is implemented in ChemAxon products and implemented in the Squonk Computational Notebook[9] that is loosely based on the Murcko drug filter (Walters, Murcko and Murcko, 1999) and the Oprea drug filter (Oprea, 2000). The following are the rules used by this filter:

- molecular mass < 400
- ring count > 0

[9] Squonk Computational Notebook. https://squonk.it/docs/cells/Drug-like%20Filter%20(CXN)/ accessed on 5 August 2021

- rotatable bond count < 5
- hydrogen bond donor count ≤ 5
- hydrogen bond acceptor count ≤ 10
- c log P < 5

3.6 Small Molecule Clustering

Virtual screening of large multi-million molecule libraries is computationally intensive, and this is especially true for docking experiments and similarity searches that are based on the computation of distance matrices (Gentile et al., 2022). Willett, Barnard, and Downs (1998) defined clustering as "*the process of subdividing a group of objects (chemical molecules in the present context) into groups, or clusters, of objects that exhibit a high degree of both intracluster similarity and intercluster dissimilarity*."

Clustering permits to have an overview of the range of properties within a dataset by selecting cluster representatives from each cluster. This approach increases the efficiency of virtual screening exercises by reducing the number of molecules used for searches, and redundancy. If a molecule (cluster representative) is identified as being a promising inhibitor, then further studies/assessments can be performed on the other members of the cluster, conversely, inactive molecules will lead to the rejection of the members of that particular cluster (Ebejer, 2014).

In order to decrease the computational resource requirements for the execution of the virtual screening experiments, we used an implementation of the sphere exclusion clustering algorithm, using RDKit to cluster the different lists of filtered ZINC15 molecules. A distance threshold of 0.65 was selected to obtain clusters having a high level of intracluster similarity. Please note that due to the small number of actives, these were not clustered.

3.7 Generation of Morgan Fingerprints

Molecular fingerprints are molecular descriptors that are used extensively in virtual screening because they offer a number of advantages, including their ease of use, and can be used efficiently to compare molecular structures). In order to reduce the computational intensity of the comparison of molecules in a similarity search, molecules can be abstracted into molecular fingerprints through their conversion into a sequence of bits (Cereto-Massagué et al., 2015). The size and complexity of a molecule heavily influence the success of a similarity search (Riniker and Landrum, 2013). Morgan Fingerprints are the most popular fingerprints used in small molecule virtual screening.

For the purpose of this study, we calculated the Morgan Fingerprint of each molecule (cluster representative and filtered active molecule) as 2048-bit vector of radius 2. Morgan fingerprints are a type of circular fingerprint based on the Morgan algorithm, that consist of a hashed topological fingerprint that records the neighbourhood environment of each atom to a pre-set radius.

3.8 Similarity Search Using Morgan Fingerprints

The similarity between two vectors can be quantified and assessed in many ways, with the Tanimoto coefficient being one of the most commonly used metrics to estimate this similarity (Willett, 2006). The Dice coefficient is another commonly used similarity coefficient (Bajusz, Rácz, and Héberger, 2015). The equations used to calculate these two coefficients are provided in Table 3.8. Irrespective of the length of fingerprints, the Tanimoto and Dice coefficients will always range between 0 and 1.

Table 3.8 Similarity coefficients

Similarity/Distance metric	Equation
Dice	$\frac{2c}{a+b}$
Tanimoto	$\frac{c}{a+b-c}$
where, given two fingerprints A and B, a is the number of bits set to 1 in fingerprint A, b is the number of bits set to 1 in fingerprint B, and c is the number of bits set to 1 that are common in both fingerprints.	

In order to compare the Tanimoto and Dice metrics, two similarity searches were performed for each active molecule within the respective filtered group, against the cluster representatives. The two searches were performed through the calculation of the Tanimoto and the Dice coefficients for each active-cluster representative pair of molecular fingerprints, using the appropriate module within RDKit. For each filtered group of molecules, two pairs of lists were generated, one pair consisting of the Tanimoto coefficients for all cluster representatives against each active molecule in descending order and the Tanimoto coefficients for the fifteen top-scoring (to have a general overview of the top scoring results) cluster representatives for each active molecule, and a similar pair based on the Dice coefficients.

Landrum[10] (2021) suggests the use of a Tanimoto similarity threshold of 0.3 and better still 0.4 at a noise level of 0.95, with the 0.4 cut-off being the most relevant cut-off value. Molecules scoring a

[10] Fingerprint similarity thresholds for database searches. https://greglandrum.github.io/rdkit-blog/similarity/reference/2021/05/21/similarity-search-thresholds.html accessed on 21 July 2022.

Tanimoto similarity threshold higher than 0.4 were then filtered out into a final list of hits (Appendix I).

3.9 Ultrafast Shape Recognition with Pharmacophoric Constraints

In addition to performing similarity searches based on fingerprint searches, a similarity search based on molecular shape comparison was performed. Based on the fact that the effect of a small molecules on a macromolecule, such as a protein/enzyme, frequently relies on the binding of the small molecule to a specific pocket on the macromolecule, a molecular shape comparison is an additional effective tool for ligand based virtual screening. Ultrafast Shape Recognition (USR) is a method to perform such molecular shape comparisons in an efficient manner (Ballester and Richards, 2007). USR is molecular shape comparison tool that is agnostic to the atom types of the molecules it is comparing and therefore it cannot discriminate on the basis of pharmacophoric features of the respective molecules. An improvement over the USR method is Ultrafast Shape Recognition with CREDO atom types (USRCAT) developed by Schreyer and Blundell (2012), whereby in addition to the unique vector of geometrical descriptors, which spans a 12-dimensional molecular shape space produced by the USR method, the vector now consists of 60 elements.

An attempt was made to generate the lowest energy conformer of each molecule (cluster representative and filtered active molecule) within the set of molecules filtered by using the drug-like filter, using an implementation of the Experimental-Torsion Basic Knowledge Distance Geometry (ETKDG) algorithm within RDKit (Riniker and Landrum, 2015). It is worth noting that for 878 molecules, it was not possible to generate a conformer using the tools provided by RDKit.

The lowest energy conformer was then used to generate the USRCAT vector for each molecule and then these vectors were then compared to the vectors for the active molecules. The comparison of the different molecular shapes results in a USR score, with a higher score indicating a higher degree of similarity. Like the previous similarity searches, two lists were developed, one containing all the scores for all the comparisons with the respective active molecules and a list containing the best performing 15 molecules when compared to the respective active molecule.

The generation of the lowest energy conformers, derivation of the USRCAT vectors and the molecular shape similarity search was performed by running the python script *4_usrcat_conf_gen.py* on each list of molecules.

3.10 Enrichment Factor Calculation

The calculation of the Enrichment factor gives an indication of the number of active compounds identified within a pre-defined cut-off by using a virtual screening method (Bender and Glen, 2005). The enrichment factor is a frequently used metric to assess the success rate or quality of a particular virtual screening protocol. Equation 3.2 represents the formula used to calculate the enrichment factor (Fechner and Schneider, 2004) for each active and then the average for the respective filtered set of actives was calculated.

$$EF_{x\%} = \left({}^{S_{act}}/_{S_{all}}\right) \Big/ \left({}^{P_{act}}/_{P_{all}}\right)$$

Equation 3.2 $EF_{x\%}$ is the Enrichment Factor at the percentage cutoff x%, P_{all} is the total number of molecules in all of the list, S_{all} is the number of molecules in the subset chosen with the x% cutoff, P_{act} is the number of active molecules in all of the list, and S_{act} is the number of actives found in the subset.

3.11 Technology stack

All development was done in Python 3.9, using RDKit 2022.03.1. The processes, with the exception of the clustering of molecules, were performed in Ubuntu 20.04. The molecular clustering was performed using Ubuntu 18.04 Server with 1 TB of RAM.

The extraction of the molecules of interest from the database and filtering was performed by running the python script *2_file_importer_filters.py*

The molecular clustering of the filtered molecules was performed by running the python script *3_excl_clustering_tc.py* on the lists of filtered molecules.

The molecular fingerprints for the molecules and respective similarity searches (using Tanimoto and Dice coefficients) were performed by executing the python script *5_lbvs_tanimoto.py* on each filtered list of molecules.

The enrichment factors were calculated for all of the generated lists by running the python script *6_stat_enr_factor.py* on each of the generated lists.

All of the code that has been developed for the purpose of all the experiments referred to above have been made available at the following repository:

https://github.com/gandalfdgrey/COVID19CADD.git

4

Results and Discussion

In accordance with the aim and objectives of this study, a desk review coupled with a series of *in silico* experiments were implemented. The following processes were undertaken, as per the methodology described in Chapter 3:

- Identification of a suitable protein target relevant to inhibitor identification against SARS-CoV-2 – Protein target identification has an important role in drug development (Schenone, et al., 2013). This process is also important in order to identify suitable active ligands that are known to be active against the protein target of interest.
- Virtual screening relies on the screening of lists or databases of small molecules and therefore the creation of a suitable database or list of small molecules is an important step in many virtual screening studies that are based on similarity searches.
- Ligand-Based Virtual Screening – Development of a list of hits that can be further evaluated as part of further *in vitro* or *in silico* studies through two different similarity search experiments, the first being the Tanimoto similarity search using Morgan Fingerprints and the second being a similarity search using the Ultrafast Shape Recognition with CREDO Atom Types (USRCAT) algorithm.

4.1 Protein Target Identification

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the virus responsible for the COVID-19 pandemic. Notwithstanding the development of more than effective vaccines against this virus (strains), it is manifestly evident that SARS-CoV-2 will remain a problematic disease-causing virus for the foreseeable future through the mutation of the virus in novel strains. By 25 July 2022, there were 40 approved vaccines, 11 WHO EUL (Emergency Use Listing) vaccines, and 220 vaccine candidates (Shroti et al, 2021). Moreover, at a global level, 62.03% of the population has received the last dose of the primary vaccination series and 26.06% of the population has received a primary booster

dose of a COVID-19 vaccine (WHO COVID-19 Dashboard, accessed on 27 July 2022[11]). By this date, there were an astounding 574,940,327 confirmed cases of COVID-19 infection, globally. This means that in view of the number of unvaccinated people, gaps in efficacy, and new viral strains, there will be a sustained need for the development of new therapeutic agents that are effective against SARS-CoV-2 and possibly also effective against related coronaviruses (van de Leemput and Han, 2021).

As reported in Table 2.1, the literature review identified five major protein targets that are being studied for the identification of hits that can be further studied as part of a therapeutic development strategy. These are the Main protease (M^{pro}), Papain-like protease (PL^{pro}), Spike protein (S-protein), RNA-dependent RNA polymerase (RdRp), and Transmembrane serine protease 2 (TMPRSS2). Unlike the other protein targets, TMPRSS2 is a protease that is expressed from the human genome. Of the 19 studies that were reviewed, the most studied protein target is M^{pro} with 13 studies evaluating this protein target. It is also the protein target with most studies that included a validation study for the results obtained from the *in silico* experiment, through an *in vitro* study (four studies) and the highest number of identified hits (194). Kulkarni and Ingale (2022) identified around 52 *in silico* drug repurposing studies that were performed on M^{pro} with the protein structure PDB ID: 6LU7 (co-crystallized with N3) being the most frequently used in these studies. This was also observed in the reviewed studies.

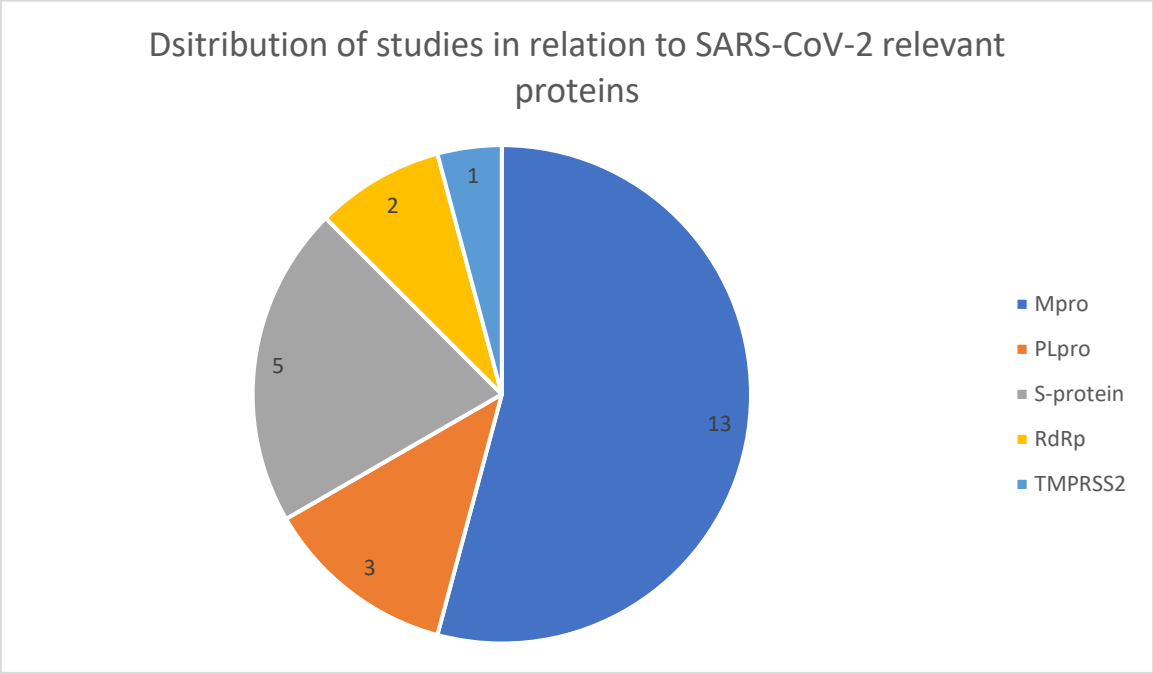


Figure 4.1 Summary of reviewed studies and identified SARS-C-V-2 relevant protein targets

[11] https://covid19.who.int/

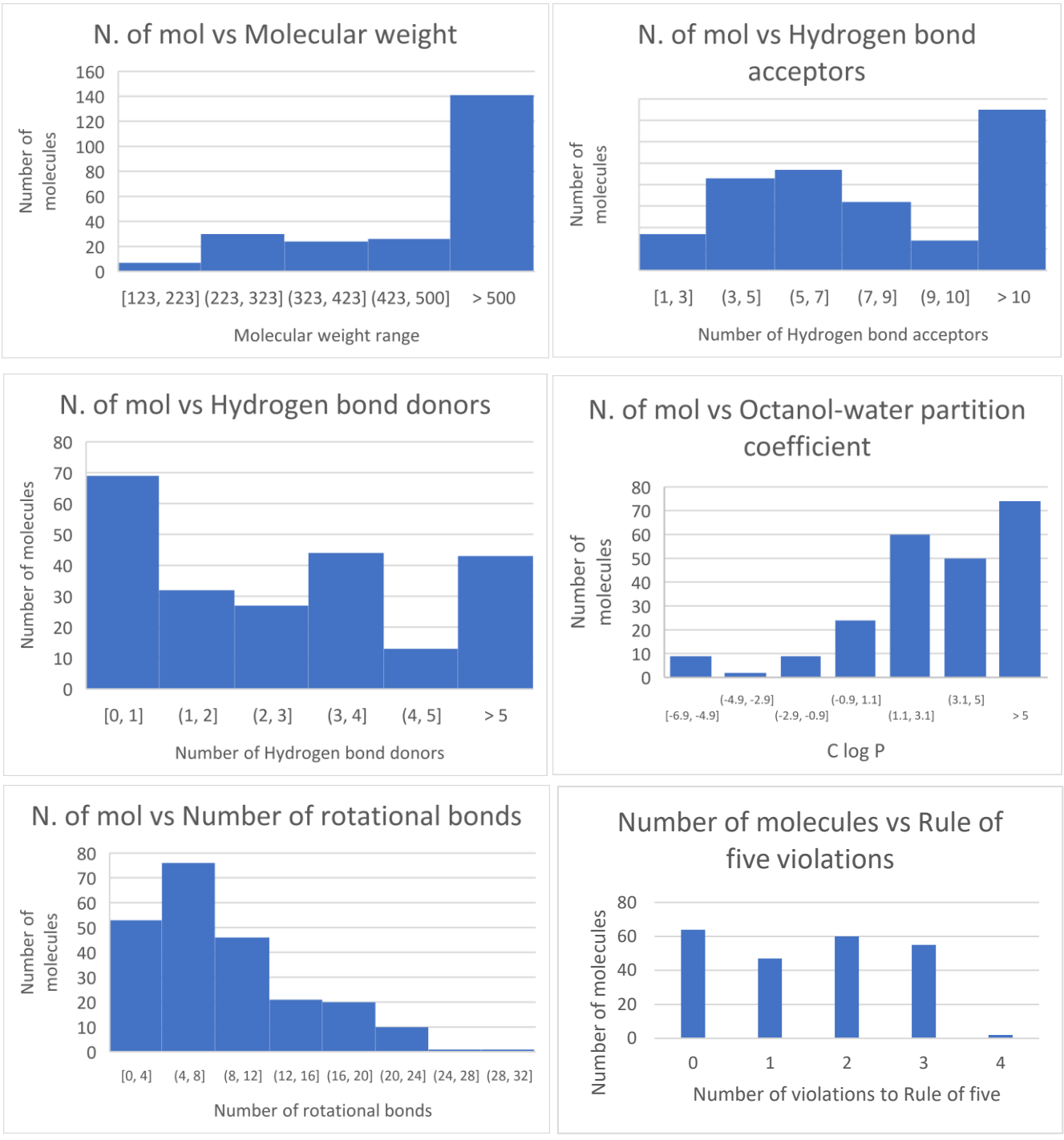


Figure 4.2 Charts indicating the different molecular properties relevant to drug-likeness versus the number of molecules identified from the literature

A summary of the molecular properties of hits effective against M^{pro} that were identified from the literature review is provided in Table 4.2. As explained further below, the implementation of Lipinski's rule of five (which is based on cut-off points for the molecular weight, number of hydrogen bond donors and acceptors, and octanol-water partition coefficient) allows for a maximum violation of one of its cut-offs. The number of hits effective against M^{pro} that violated more than one of these cut-offs was around 51%, notwithstanding the fact that more than 59% of the hits are either approved drugs administered via the oral route or are in some stage of drug-development.

SARS-CoV-2 is closely related, from a phylogenetic point of view, to SARS-CoV (sharing around 80% of their genome), moreover, their proteases are highly conserved (Wu et al, 2020). M^{pro} from these two viruses cleaves the peptide bond following a glutamine amino acid, more specifically the Leu-Gln↓(Ser, Ala, Gly) sequence (Anand et al., 2003, and Zhang et al., 2020). The processing of the polyprotein 1ab (pp1ab) encoded by the viral RNA of SARS-CoV-2, by cleavage at 11 sites of this polyprotein by M^{pro} (Hilgenfield, 2014 and Jin et al., 2020) is considered to be critical to the survival/replication of the virus and therefore the inhibition of the activity of this enzyme is expected to lead to an effective therapeutic against infection by this virus. The fact that no human protease that possesses a similar cleavage specificity is known, is also very desirable since such inhibitors are unlikely to have undesirable effects on humans.

In view of the large number of studies concentrating on M^{pro}, the large number of hits that have been identified in the literature as being possibly effective against this protease, the highly conserved nature of the enzyme, the criticality of this enzyme to the viral lifecycle, and the cleavage site that lacks closely related homologues in humans were considered as justification to focus this study's efforts on this protein target (Pillaiyar et al., 2016).

4.2 Molecular database creation

ZINC (ZINC Is Not Commercial) is a publicly accessible database that provides access to a repository of molecules that can be used for a multitude of studies, including virtual screening. In its current version, ZINC15 associates the chemical repository data with biological activity by gene product, drugs, and natural products, as well as commercial availability.

The tranche of 2D drug-like in-stock compounds of ZINC15 was downloaded on 21 July 2021, which consisted of 6,877,387 molecules, divided into 720 .smi files. Each line in the .smi files contains the SMILES formula and ZINCID of the molecule, as well as a header line at the top of the file. The molecules were then sanitized and deduplicated, which resulted in a list of 6,065,672 unique molecules. Drug filtering strategies were applied separately to this list of unique molecules to develop filtered lists of molecules that are considered drug-like and orally active in accordance with the respective drug-like filter. The filtering parameters for each filtering strategy that was used are summarised in Table 4.1. The results obtained from the application of the different filters are summarised in Table 4.2.

Lipinski's rule of five is the most prevalently used filtering strategy applied in virtual screening pipelines and is considered to give a good indication of oral absorption (Linpinski et al., 1997). Veber studied the relation between the number of rotatable bonds and topological polar surface area, and the permeability of the cell membranes within the gastrointestinal tract, with highly flexible and highly polar molecules being less permeable (Veber et al., 2002). In another study, Ghose concluded that the

total number of atoms, molar refractivity, an increased range of log P and a smaller molecular weight range are more indicative of oral bioavailability than what was proposed by Lipinski. Moreover, Ghose pointed out that an abundance of phenyl rings in relation to heterocyclic rings resulted in a higher hydrophobicity (Ghose, Viswanadhan and Wendoloski, 1999). Oprea (2000) and Murcko (Walters, Murcko and Murcko, 1999), respectively, continued working on Lipinski's rule of five to make sure that a more diversified pool of compounds is achieved as part of the filtering strategy (Protti et al., 2021). Saxena and Prathipati (2006) evaluated Lipinski's rule of five, Veber's filter and the Oprea filtering strategies and concluded that they all have their strengths and weaknesses, moreover, drug-likeness is influenced by various molecular properties and structural features. To this effect a drug-like filter, refining Lipinski's rule of five was implemented in ChemAxon products and implemented in the Squonk Computational Notebook[12].

Table 4.1 Summary of filtering strategies.

Parameter	Lipinski (pass all parameters)	Lipinski[13]	Veber	Ghose	REOS	Drug-like
H-bond donors	≤ 5	≤ 5			0 to 5	≤ 5
H-bond acceptors	≤ 10	≤ 10			0 to 10	≤ 10
Molecular weight	≤ 500	≤ 500		160 to 480	200 to 500	< 400
C Log P	≤ 5	≤ 5		-0.4 to 5.6	-5.0 to 5.0	< 5
Rotatable bonds			≤ 10		0 to 8	< 5
Atom count			≤ 140Å^2	20 to 70		
Topological Polar Surface Area						
Molecular refractivity				40 to 130		
Formal charge					-2 to 2	
Heavy atom count					15 to 50	
Ring count						> 0

Table 4.2 Results from the application of the different filters

Filter	ZINC15 list		Actives		Percentage of filtered ZINC15 molecules vs filtered hits (actives)
	Number of molecules	% Filtered	Number of molecules	% Filtered	
Lipinski (All parameters)	3,834,942	63.22%	21	9.33%	14.76%
Lipinski	6,045,537	99.67%	41	18.22%	18.28%
Veber	5,966,447	98.36%	63	28.00%	28.47%
Ghose	4,923,340	81.17%	18	8.00%	9.86%
REOS	5,646,742	93.09%	27	12.00%	12.89%
QED – 0.6	4,376,791	72.16%	17	7.56%	10.47%
QED – 0.9	329,391	5.43%	0	0.00%	0.00%
Drug-like	2,360,459	38.92%	14	6.22%	15.99%

[12] Squonk Computational Notebook. https://squonk.it/docs/cells/Drug-like%20Filter%20(CXN)/ accessed on 5 August 2021

[13] The Lipinski rule of five filter allows one violation of the parameters.

The quantitative estimate of drug-likeness, QED, developed by Bickerton and his team (Bickerton, et al., 2012), integrates (geometric mean of) desirability functions for eight molecular properties relevant to drug-likeness (molecular weight, octanol-water partition coefficient number of hydrogen bond donors and acceptors, molecular polar surface area, number of rotatable bonds, number of aromatic rings and number of structural alerts) into an index ranging from 0 to 1, with 1 being the highest score for drug-likeness.

From the results obtained, the application of Lipinski's rule of five resulted in a 99.67% recovery of small molecules from the ZINC15 list, which shows that the drug-like tranche of ZINC15 is filtered using a filtering strategy very similar (if not identical) to Lipinski's rule of five. However, a lower number of hits identified from the literature was filtered (18.22%) when applying the rule of five, with the best result being observed when using the Veber filter (28.00%). Other filters that gave good recovery rates from the ZINC15 listed were the QED at a 0.6 cut-off, the Ghose filter, and REOS filters with recovery rates of 72.16%, 81.17%, and 98.36%, respectively. However, the same was not observed when applying these filtering strategies to the list of actives, with a filtering of 7.56%, 8.00%, and 12.00% for the respective filter. When comparing the percentage recovery of both lists together, the best performing filter is the Veber filter. The list generated by the Lipinski filter (i.e. allowing one violation of the rules) was selected for the rest of the study, since the same molecules captured by the strict implementation of Lipinski's rule of five would be included in the relaxed rule of five implementation.

Based on the fact that no active molecule identified from the literature was filtered by the Quantitative Estimate of Drug-Likeness (at a cut-off of 0.9), this was considered as the worst-performing filtering strategy. In view of the poor recovery of small molecules and complete rejection of actives when using the QED filter cut-off of 0.9, the lists generated using this filter were not considered any further.

Table 4.3 SMILES formula for the actives filtered using the drug-like filter

SMILES	Reference
CCC(C)SSc1ncc[nH]1	AAAABU
Oc1ccc(C=Cc2cc(O)cc(O)c2)cc1	AAAAHO
CC=CC1Oc2c(-c3ccccc3)cn(O)c(=O)c2C2C(C)CCCC12	AAAAKO
COn1ccc2c(c1=O)C1C(C)CC(C)CC1(C)C(C)O2	AAAAME
CC1CC(C)(O)C(C)C(c2c(O)[nH]ccc2=O)C1C	AAAANE
C=CC1CC(C)CC(C)C1C1CN(O)C=CC1=O	AAAANO
C=CC1CC(C)CC(C)C1c1c(O)ccn(O)c1=O	AAAAOA
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOI
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOU
COc1ccn(C)c(=O)c1C#N	AAAAQA
CCN(CC)C(=O)c1cccnc1	AAABLA
CCC1(CCC(C)C)C(=O)NC(=O)NC1=O	AAABLU
CC1CS(=O)(=O)CCN1/N=C/c1ccc([N+](=O)[O-])o1	AAABNE
O=c1c2ccccc2[se]n1-c1ccccc1	AAACBA

Table 4.4 SMILES formula for the actives filtered using the Ghose filter

SMILES	Reference
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAACU
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAAEU
CC(CC(C)C(Cl)CCl)C(=O)c1c(O)[nH]c(CO)c(CO)c1=O	AAAAPE
CC=CC1Oc2c(-c3ccccc3)cn(O)c(=O)c2C2C(C)CCCC12	AAAAKO
CC1CC[C@@H]2C(C(=O)c3c(O)[nH]cc(C4(O)CC[C@@H](O)[C@H]5O[C@H]54)c3=O)C(C)C=C[C@H]2C1	AAAALA
CCC(CC)NC(=O)C(Cc1ccccc1)NC(=O)c1cc(Cl)cc(Cl)c1	AAAAYA
CC1CCC2C(C=CC(C)C2C(=O)c2c(O)[nH]cc(C3(O)CCC(O)(O)C4OC43)c2=O)C1	AAAALU
COn1ccc2c(c1=O)C1C(C)CC(C)CC1(C)C(C)O2	AAAAME
CCC(/C=C/C=C/C=C/C(=O)c1c(O)c(C2(O)CCC(O)CC2)cn(O)c1=O)CO	AAAAPA
Cc1ccc(NC(=O)c2cc(C#N)ncc2F)cc1-n1ccn2nc(-c3cccnc3)cc12	AAAAUU
CCOC(=O)C1=CC(OC(CC)CC)C(NC(C)=O)C(N)C1	AAABAE
O=C(CCc1ccc(F)cc1F)N[C@@H](Cc1ccccc1)[C@@H](O)CNc1ccc(F)cc1	AAABBO
CNC(=O)C(NC(=O)C(CC(C)C)C(O)C(O)=NO)C(C)(C)C	AAABLI
CCS(=O)(=O)N1CC(CC#N)(n2cc(-c3ncnc4[nH]ccc34)cn2)C1	AAACAI
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAABWI
CC1OC(Oc2c(-c3ccc(O)c(O)c3)oc3cc(O)cc(O)c3c2=O)C(O)C(O)C1O	AAACCE
OCC1OC(Oc2cc3c(O)cc(O)cc3[o+]c2-c2ccc(O)c(O)c2)C(O)C(O)C1O	AAACEE
CCc1cc2c(cc1CC)CC(NCC(O)c1ccc(O)c3[nH]c(=O)ccc13)C2	AAACHA

Table 4.5 SMILES formula for the actives filtered using the relaxed Lipinski filter

SMILES	Reference
CCN(CC)C(=S)SSC(=S)N(CC)CC	AAAABA
CCCCCCNC(=O)n1cc(F)c(=O)[nH]c1=O	AAAABI
CCC(C)SSc1ncc[nH]1	AAAABU
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAACU
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAAEU
Oc1ccc(C=Cc2cc(O)cc(O)c2)cc1	AAAAHO
CC(CC(C)C(Cl)CCl)C(=O)c1c(O)[nH]c(CO)c(CO)c1=O	AAAAPE
CC(C)CC(NC(=O)OCc1ccccc1)C(=O)N[C@@H](CC1CCNC1=O)C(O)S(=O)(=O)O	AAAAAA
CC=CC1Oc2c(-c3ccccc3)cn(O)c(=O)c2C2C(C)CCCC12	AAAAKO
CC1CC[C@@H]2C(C(=O)c3c(O)[nH]cc(C4(O)CC[C@@H](O)[C@H]5O[C@H]54)c3=O)C(C)C=C[C@H]2C1	AAAALA
CCC(CC)NC(=O)C(Cc1ccccc1)NC(=O)c1cc(Cl)cc(Cl)c1	AAAAYA
CC1CCC2C(C=CC(C)C2C(=O)c2c(O)[nH]cc(C3(O)CCC(O)(O)C4OC43)c2=O)C1	AAAALU
COn1ccc2c(c1=O)C1C(C)CC(C)CC1(C)C(C)O2	AAAAME
CCC(/C=C/C=C/C=C/C(=O)c1c(O)c(C2(O)CCC(O)CC2)cn(O)c1=O)CO	AAAAPA
CC1CC(C)(O)C(C)C(c2c(O)[nH]ccc2=O)C1C	AAAANE
C=CC1CC(C)CC(C)C1C1CN(O)C=CC1=O	AAAANO
C=CC1CC(C)CC(C)C1c1c(O)ccn(O)c1=O	AAAAOA
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOI
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOU
COc1ccn(C)c(=O)c1C#N	AAAAQA
FC(F)(F)c1ccc(Nc2ncnc3cc(-c4ncccc4C(F)(F)F)ccc23)cc1	AAAAQI
COc1ccc2c(OC3CC4C(=O)NC5(CCCCCCCCCC(=O)N(C)C5)C(=O)N4C3)cc(-n3cccc3)nc2c1	AAAASO
O=C(NC(Cc1ccccc1)C(=O)N1CCOCC1)C(c1ccccc1)c1ccc2ccccc2c1	AAAATA
O=S(=O)(Nc1ccc2c(C=Cc3ccc4ncccc4c3)cccc2c1)C(F)(F)F	AAAATU
COc1ccc2c(OC3CC4C(=O)NC5(C(=O)O)OCC5COCCCC(=O)N4C3)cc(-c3cccs3)nc2c1	AAAAUE
O=C(O)CC(O)(CC(=O)O)C(=O)O	AAABMI
Cc1ccc(NC(=O)c2cc(C#N)ncc2F)cc1-n1ccn2nc(-c3cccnc3)cc12	AAAAUU
CCOC(=O)C1=CC(OC(CC)CC)C(NC(C)=O)C(N)C1	AAABAE
CC(C)(C)C(O)C(Cc1ccccc1)NC(=O)C(Cc1ccccc1)NC(=O)c1ccc(Cl)cc1	AAAAYE
Nc1ncnc2c(C3NC(CO)C(O)C3O)c[nH]c12	AAABAU
O=C(CCc1ccc(F)cc1F)N[C@@H](Cc1ccccc1)[C@@H](O)CNc1ccc(F)cc1	AAABBO
CCN(CC)C(=O)c1cccnc1	AAABLA
CNC(=O)C(NC(=O)C(CC(C)C)C(O)C(O)=NO)C(C)(C)C	AAABLI
CCC1(CCC(C)C)C(=O)NC(=O)NC1=O	AAABLU
Cc1cccc(-c2ccc3c(c2)c(C(=O)C(=O)O)cn3Cc2ccc(C(C)(C)C)cc2)c1	AAABMU
CC1CS(=O)(=O)CCN1/N=C/c1ccc([N+](=O)[O-])o1	AAABNE
CCS(=O)(=O)N1CC(CC#N)(n2cc(-c3ncnc4[nH]ccc34)cn2)C1	AAACAI
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAABWI
O=c1c2ccccc2[se]n1-c1ccccc1	AAACBA
OCC1OC(Oc2cc3c(O)cc(O)cc3[o+]c2-c2ccc(O)c(O)c2)C(O)C(O)C1O	AAACEE
CCc1cc2c(cc1CC)CC(NCC(O)c1ccc(O)c3[nH]c(=O)ccc13)C2	AAACHA

Table 4.6 SMILES formula for the actives filtered using the QED (0.6) filter

SMILES	Reference
CCCCCCNC(=O)n1cc(F)c(=O)[nH]c1=O	AAAABI
CCC(C)SSc1ncc[nH]1	AAAABU
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAACU
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAAEU
Oc1ccc(C=Cc2cc(O)cc(O)c2)cc1	AAAAHO
CC=CC1Oc2c(-c3ccccc3)cn(O)c(=O)c2C2C(C)CCCC12	AAAAKO
CCC(CC)NC(=O)C(Cc1ccccc1)NC(=O)c1cc(Cl)cc(Cl)c1	AAAAYA
COn1ccc2c(c1=O)C1C(C)CC(C)CC1(C)C(C)O2	AAAAME
CC1CC(C)(O)C(C)C(c2c(O)[nH]ccc2=O)C1C	AAAANE
C=CC1CC(C)CC(C)C1C1CN(O)C=CC1=O	AAAANO
C=CC1CC(C)CC(C)C1c1c(O)ccn(O)c1=O	AAAAOA
CCOC(=O)C1=CC(OC(CC)CC)C(NC(C)=O)C(N)C1	AAABAE
CCN(CC)C(=O)c1cccnc1	AAABLA
CCC1(CCC(C)C)C(=O)NC(=O)NC1=O	AAABLU
CCS(=O)(=O)N1CC(CC#N)(n2cc(-c3ncnc4[nH]ccc34)cn2)C1	AAACAI
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAABWI
O=c1c2ccccc2[se]n1-c1ccccc1	AAACBA

Table 4.7 SMILES formula for the actives filtered using the REOS filter

SMILES	Reference
CCN(CC)C(=S)SSC(=S)N(CC)CC	AAAABA
CCCCCCNC(=O)n1cc(F)c(=O)[nH]c1=O	AAAABI
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAACU
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAAEU
Oc1ccc(C=Cc2cc(O)cc(O)c2)cc1	AAAAHO
CC(CC(C)C(Cl)CCl)C(=O)c1c(O)[nH]c(CO)c(CO)c1=O	AAAAPE
CC=CC1Oc2c(-c3ccccc3)cn(O)c(=O)c2C2C(C)CCCC12	AAAAKO
CC1CC[C@@H]2C(C(=O)c3c(O)[nH]cc(C4(O)CC[C@@H](O)[C@H]5O[C@H]54)c3=O)C(C)C=C[C@H]2C1	AAAALA
CCC(CC)NC(=O)C(Cc1ccccc1)NC(=O)c1cc(Cl)cc(Cl)c1	AAAAYA
CC1CCC2C(C=CC(C)C2C(=O)c2c(O)[nH]cc(C3(O)CCC(O)(O)C4OC43)c2=O)C1	AAAALU
COn1ccc2c(c1=O)C1C(C)CC(C)CC1(C)C(C)O2	AAAAME
CCC(/C=C/C=C/C=C/C(=O)c1c(O)c(C2(O)CCC(O)CC2)cn(O)c1=O)CO	AAAAPA
CC1CC(C)(O)C(C)C(c2c(O)[nH]ccc2=O)C1C	AAAANE
C=CC1CC(C)CC(C)C1C1CN(O)C=CC1=O	AAAANO
C=CC1CC(C)CC(C)C1c1c(O)ccn(O)c1=O	AAAAOA
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOI
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOU
O=C(NC(Cc1ccccc1)C(=O)N1CCOCC1)C(c1ccccc1)c1ccc2ccccc2c1	AAAATA
Cc1ccc(NC(=O)c2cc(C#N)ncc2F)cc1-n1ccn2nc(-c3cccnc3)cc12	AAAAUU
CCOC(=O)C1=CC(OC(CC)CC)C(NC(C)=O)C(N)C1	AAABAE
CNC(=O)C(NC(=O)C(CC(C)C)C(O)C(O)=NO)C(C)(C)C	AAABLI
CCC1(CCC(C)C)C(=O)NC(=O)NC1=O	AAABLU
CC1CS(=O)(=O)CCN1/N=C/c1ccc([N+](=O)[O-])o1	AAABNE
CCS(=O)(=O)N1CC(CC#N)(n2cc(-c3ncnc4[nH]ccc34)cn2)C1	AAACAI
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAABWI
O=c1c2ccccc2[se]n1-c1ccccc1	AAACBA
CCc1cc2c(cc1CC)CC(NCC(O)c1ccc(O)c3[nH]c(=O)ccc13)C2	AAACHA

Table 4.8 SMILES formula for the actives filtered using the Veber filter

SMILES	Reference
CCN(CC)C(=S)SSC(=S)N(CC)CC	AAAABA
CCCCCCNC(=O)n1cc(F)c(=O)[nH]c1=O	AAAABI
CCC(C)SSc1ncc[nH]1	AAAABU
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAACU
COc1ccc(C2=C(c3ccc(OC)cc3)N3C(=NC4c5c(ccc6ccccc56)OCC4C3(C)C)S2)cc1	AAAADE
COc1ccc(-c2sc3n(c(NC(C)(C)C)c4[n+]3C3CC(=O)C=CC3(OC)CC4)c2-c2ccc(OC)cc2)cc1	AAAADO
CC(C)(C)Nc1c2[n+](c3sc(-c4ccccc4)c(-c4ccccc4)n13)C13CCCC1(C=CC(=O)C3)OC2	AAAAEA
COc1ccc(-c2sc3n(c(NC(C)(C)C)c4[n+]3[C@H]3CC(=O)C=C5CCCC[C@@]53OC4)c2-c2ccc(OC)cc2)cc1	AAAAEI
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAAEU
Cc1nc(-c2cccnc2)sc1C(=O)Nc1ccccc1-c1cn2c(CN3CCOCC3)csc2n1	AAAAHE
Oc1ccc(C=Cc2cc(O)cc(O)c2)cc1	AAAAHO
CC(CC(C)C(Cl)CCl)C(=O)c1c(O)[nH]c(CO)c(CO)c1=O	AAAAPE
CCC(C)CC(C)/C=C(\C)C1(C)OC(C)(c2c3c(cn(C)c2=O)C2(OC)CCC(=O)CC2O3)CCC1C	AAAAJO
CCC(C)CC(C)/C=C(\C)C1(C)OC(C)(c2c3c(cn(C)c2=O)C2(OC)CCC(OC)(OC)CC2O3)CCC1C	AAAAKA
CC=CC1Oc2c(-c3ccccc3)cn(O)c(=O)c2C2C(C)CCCC12	AAAAKO
CC1CC[C@@H]2C(C(=O)c3c(O)[nH]cc(C4(O)CC[C@@H](O)[C@H]5O[C@H]54)c3=O)C(C)C=C[C@H]2C1	AAAALA
CCC(CC)NC(=O)C(Cc1ccccc1)NC(=O)c1cc(Cl)cc(Cl)c1	AAAAYA
COn1ccc2c(c1=O)C1C(C)CC(C)CC1(C)C(C)O2	AAAAME
CCC(C)CC(C)/C=C(\C)C1(C)OC(C)(c2c3c(cn(C)c2=O)C2(OC)CCC(=O)CC2O3)CCC1C	AAAAMO
CC1CC(C)(O)C(C)C(c2c(O)[nH]ccc2=O)C1C	AAAANE
C=CC1CC(C)CC(C)C1C1CN(O)C=CC1=O	AAAANO
C=CC1CC(C)CC(C)C1c1c(O)ccn(O)c1=O	AAAAOA
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOI
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOU
COc1ccn(C)c(=O)c1C#N	AAAAQA
FC(F)(F)c1ccc(Nc2ncnc3cc(-c4ncccc4C(F)(F)F)ccc23)cc1	AAAAQI
COc1cc(N2CCN(C(C)=O)CC2)ccc1Nc1ncc(Cl)c(Nc2ccc(N3CCN(C(C)=O)CC3)cc2OC)n1	AAAAQU
CC(C)(C)c1cc(NC(=O)Nc2ccc(-c3cn4c(n3)sc3cc(OCCN5CCOCC5)ccc34)cc2)no1	AAAARE
CC(C)(C)NC(=O)C1CC2CCCCC2CN1CC(OCc1ccccn1)C(=O)NCc1ccc2c(c1)OCO2	AAAASE
COc1ccc2c(OC3CC4C(=O)NC5(CCCCCCCCCC(=O)N(C)C5)C(=O)N4C3)cc(-n3cccc3)nc2c1	AAAASO
O=C(NC(Cc1ccccc1)C(=O)N1CCOCC1)C(c1ccccc1)c1ccc2ccccc2c1	AAAATA
O=C(O)c1ccc(O)c(C(Cc2cccc(Cl)c2)c2ccc(OC3CCOC3)c(Cl)c2)c1	AAAATI
O=S(=O)(Nc1ccc2c(C=Cc3ccc4ncccc4c3)cccc2c1)C(F)(F)F	AAAATU
COc1ccc2c(OC3CC4C(=O)NC5(C(=O)O)OCC5COCCCC(=O)N4C3)cc(-c3cccs3)nc2c1	AAAAUE
O=C(O)CC(O)(CC(=O)O)C(=O)O	AAABMI
Cc1ccc(NC(=O)c2cc(C#N)ncc2F)cc1-n1ccn2nc(-c3cccnc3)cc12	AAAAUU
Cc1ncc(C2=C(C(=O)Nc3cc(Br)cs3)C(C(=O)NS(=O)(=O)C3CC3)N=C(C3CC4CCN4C3)N2C)s1	AAAAVE
CC(C)(C)NC(=O)C1CC2CCCCC2CN1CC(OCC1CCCCC1)C(=O)NC1c2ccccc2CC1O	AAAAVO
CCOC(=O)C1=CC(OC(CC)CC)C(NC(C)=O)C(N)C1	AAABAE
CC(C)(C)NC(=O)C1CC2CCCCC2CN1CC(O)C(Cc1ccccc1)NC(=O)c1cccc(O)c1	AAAAWE
CC(C)(C)OC(=O)NC(Cc1ccccc1)C(Cc1ccccc1)C(=O)NC1c2ccccc2CC1O	AAAAXE
CC(C)(C)NC(=O)C1CCCCN1CC(O)C(Cc1ccccc1)NC(=O)c1ccc2ccccc2n1	AAAAXO
CC(C)(C)C(O)C(Cc1ccccc1)NC(=O)C(Cc1ccccc1)NC(=O)c1ccc(Cl)cc1	AAAAYE
Cc1c(O)cccc1C(=O)NC(CSc1ccccc1)C(O)CN1CC2CCCCC2CC1C(=O)NC(C)(C)C	AAABAA
O=C(CCc1ccc(F)cc1F)N[C@@H](Cc1ccccc1)[C@@H](O)CNc1ccc(F)cc1	AAABBO
CC(Cc1ccccc1)C(=O)N1CCN(C[C@H](O)[C@H](Cc2ccccc2)NC(=O)c2c3ccccc3cc3ccccc23)CC1	AAABDE

Table 4.9 SMILES formula for the actives filtered using the Veber filter (continued)

CC(C)C(NC(=O)OC(C)(C)C)C(=O)N1CCN(C[C@H](O)[C@H](Cc2ccccc2)NC(=O)c2c3ccccc3cc3ccccc23)CC1	AAABFI
CC1(C)[C@H]2COc3ccc4ccccc4c3[C@H]2N=C2SC=C(c3ccc(S(=O)(=O)N4CCOCC4)cc3)N21	AAABIE
CC(OC1OCCN(Cc2n[nH]c(=O)[nH]2)C1c1ccc(F)cc1)c1cc(C(F)(F)F)cc(C(F)(F)F)c1	AAABHE
COc1ccc(-c2sc3n(c(NC(C)(C)C)c4[n+]3C3CC(=O)C=CC3(OC)CC4)c2-c2ccc(OC)cc2)cc1	AAABII
CC(C)(C)Nc1c2[n+](c3sc(-c4ccccc4)c(-c4ccccc4)n13)C13CCCC1(C=CC(=O)C3)OC2	AAABIU
COc1ccc(-c2sc3n(c(NC(C)(C)C)c4[n+]3[C@H]3CC(=O)C=C5CCCC[C@@]53OC4)c2-c2ccc(OC)cc2)cc1	AAABJE
CCN(CC)C(=O)c1cccnc1	AAABLA
CNC(=O)C(NC(=O)C(CC(C)C)C(O)C(O)=NO)C(C)(C)C	AAABLI
CCC1(CCC(C)C)C(=O)NC(=O)NC1=O	AAABLU
Cc1cccc(-c2ccc3c(c2)c(C(=O)C(=O)O)cn3Cc2ccc(C(C)(C)C)cc2)c1	AAABMU
CC1CS(=O)(=O)CCN1/N=C/c1ccc([N+](=O)[O-])o1	AAABNE
CCS(=O)(=O)N1CC(CC#N)(n2cc(-c3ncnc4[nH]ccc34)cn2)C1	AAACAI
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAABWI
Cc1c(O)cccc1C(=O)NC(CSc1ccccc1)C(O)CN1CC2CCCCC2CC1C(=O)NC(C)(C)C	AAABZE
O=c1c2ccccc2[se]n1-c1ccccc1	AAACBA
Cc1nc(-c2cccnc2)sc1C(=O)Nc1ccccc1-c1cn2c(CN3CCOCC3)csc2n1	AAACBI
CCc1cc2c(cc1CC)CC(NCC(O)c1ccc(O)c3[nH]c(=O)ccc13)C2	AAACHA

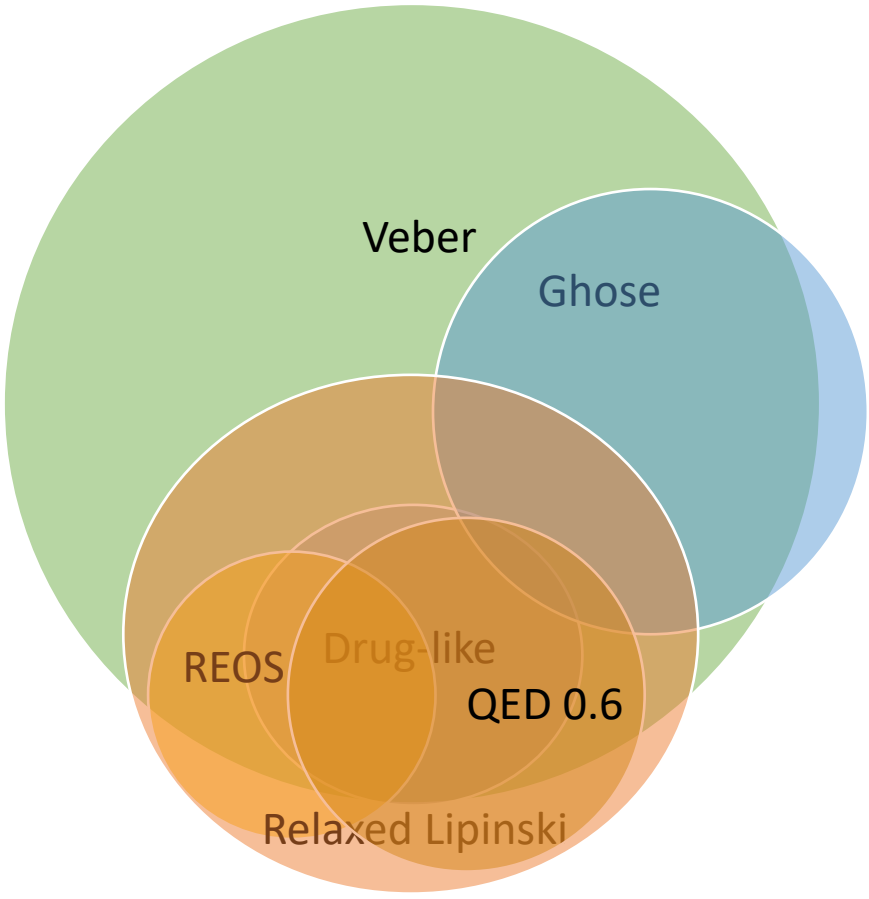


Figure 4.3 Venn diagram showing the relationship between the actives filtered using the different filters

The SMILES formulas of all the active molecules that were filtered using the respective filtering rules are represented in Tables 4.3 to 4.9 (structural representations of the actives can be found in Appendix I). It was noted that there is a significant overlap between the actives filtered by the relaxed Lipinski filter and the Veber filter. All actives filtered through the drug-filter can be found in both the

actives filtered using Lipinski's rule of five and the Veber filtering rules, this gives an indication that the drug-like filter is a good representation of the different filtering rules, since it gives good coverage from the different filtering strategies. Moreover, the vast majority of filtered molecules contain at least a single ring in their structure (as shown in Table 4.10), which is one of the rules found in the drug-like filter. It is also worth pointing out that all actives filtered by the REOS filter and the QED at a cut-off of 0.6, were filtered also by the relaxed Lipinski filter. A visual representation of the relationship/overlap between the filtered actives is provided in Figure 4.3. Therefore, as a result of this comparative assessment of filtering rules, it appears that the Drug-like filter and the relaxed Lipinski filters are adequate filters that offer a good representation of the variety within the filtered actives.

Table 4.10 Number of ring-containing filtered actives

	Veber	Lipinski	REOS	Ghose	QED 0.6	Drug-like
Ring containing molecules	60	39	25	17	17	14
Total filtered molecules	63	41	27	18	17	14

4.3 Molecular Clustering

In the context of virtual screening, clustering is the process of grouping small molecules into smaller clusters. The intracluster similarity between the members is higher than intercluster similarity (Young, 2009). Clustering provides several advantages. Using cluster centroids or representatives reduces the computational resources needed to perform a similarity search, moreover, clustering of similar molecules together reduces possible over-representation of molecules sharing similar properties. This is achieved by maximising intercluster diversity. In view of the limited availability of computational resources, as well as to minimise any possible bias in the results as a result of over-representation of molecules sharing similar properties, we decided to implement a clustering algorithm prior to performing the similarity searches.

The selection of an appropriate clustering algorithm that could handle the different datasets using the available computational resources proved to be the hardest challenge in this study. The implementation of the Taylor-Butina clustering algorithm, using RDKit, involved the computation of a distance matrix for all the members of each respective small molecule list. Due to the time and space complexity of $O(N^2)$ of this algorithm, it was not possible to calculate the distance matrix. This was a similar challenge for the implementation of the *k*-means clustering algorithm. Finally, it was decided to use the RDKit implementation of the Sphere exclusion clustering algorithm by Roger Sayle as part of the 2019.09 release of RDKit. The selection of cluster centroids is a relatively quick process, which for the largest set of small molecules took around 24 hours. However, the allocation of the different molecules into their respective clusters took 30 days and 6 hours for the Lipinski filtered list. The clustering for the different filtered lists was run in parallel since each instance is run on a single

processing thread. A similarity threshold of 0.65 was used as a similarity cut-off for the centroid selection, based on the Fingerprint thresholds developed by Greg Landrum for Morgan fingerprints with atom radius of two and having a 99% coverage[14].

Table 4.11 Results from the clustering processing

Filtered list	Number of molecules - initial set	Number of centroids	Percentage recovery
Lipinski	6,045,537	70,457	1.17%
Veber	5,966,447	69,428	1.16%
Ghose	4,923,340	54,050	1.10%
REOS	5,646,742	66,049	1.17%
QED 0.6	4,376,791	55,533	1.27%
Drug-like	2,360,459	49,046	2.08%

The number of clusters obtained from the clustering of the filtered small molecules is shown Table 4.11. The largest number of cluster centroids, which can be seen as a measure of diversity, was obtained from the Lipinski small molecule set. Notwithstanding the fact that the drug-like filter started with a smaller dataset, the number of centroids that were identified is higher when compared to the other filtering methods, as a percentage recovery. Moreover, when comparing the number of cluster centroids obtained from the drug-like small molecule set with the number of centroids obtained from the other small molecule sets, there is an overlap (in terms of numbers) of ~70% of centroids identified from the Lipinski filter and ~91% of the Ghose filter. The above further indicates that the use of the Lipinski and Drug-like filters is a suitable choice of drug filters.

4.4 Ligand-Based Virtual Screening Experiments - Fingerprint Similarity Search

Separate fingerprint similarity searches were performed using filtered lists of molecules (from ZINC15) and their respective filtered list of actives, using Tanimoto and Dice coefficients. For the purpose of this evaluation, only the results obtained from the Drug-like, Veber, and Lipinski filtered lists shall be discussed further. The Veber list has the largest number of filtered actives, the list obtained from applying the Lipinski filter resulted in the largest number of ZINC15 cluster centroids and the drug-like list provides good coverage, both in terms of actives and ZINC15 molecules.

The searches undertaken using the Tanimoto and Dice ceofficients resulted in the same ranking of molecules for the filtered molecules between the two coefficients. It was observed that the use of the Dice coefficient always resulted in higher scores than the Tanimoto coefficient. This is also in line with the findings of Bajusz, Rácz, and Héberger (2015), where they observed that certain metrics result in

[14] Fingerprint Thresholds. http://rdkit.blogspot.com/2013/10/fingerprint-thresholds.html accessed on 5 June 2022.

similar behaviour, with identical ranking of results. This similarity stems from the mathematical definition of the different coefficients. They also observed that the Tanimoto and Dice coefficients are monotonic with each other. The study also identified the Dice and Tanimoto similarities as the best among a number of similarity metrics, moreover the Tanimoto coefficient is identified as the most popular coefficient metric (Gimeno, et al, 2019). When comparing the top 15 Tanimoto coefficients with the top 15 Dice coefficients obtained from the similarity search of each active (refer to Figures 4.4 to 4.6), apart from the fact that we obtain identical rankings, we find that there is a very strong linear correlation between the two coefficients, which is expected.

When analysing the results of a similarity search there are a number of considerations, namely:

- A high similarity coefficient value does not automatically result in compounds having similar activity. Differences in protein interaction can result from minor structural/functional group differences – activity cliffs. Activity cliffs are defined as pairs of compounds with very high similarity yet highly different activity (Cereto-Massagué, et al. 2015);
- The selection of an appropriate cut-off value to be applied to the ranking of compounds;
- Identification of false positives and false negatives as a result of the fact that similarity metrics do not differentiate between the importance of different fingerprint bits, in relation to activity.

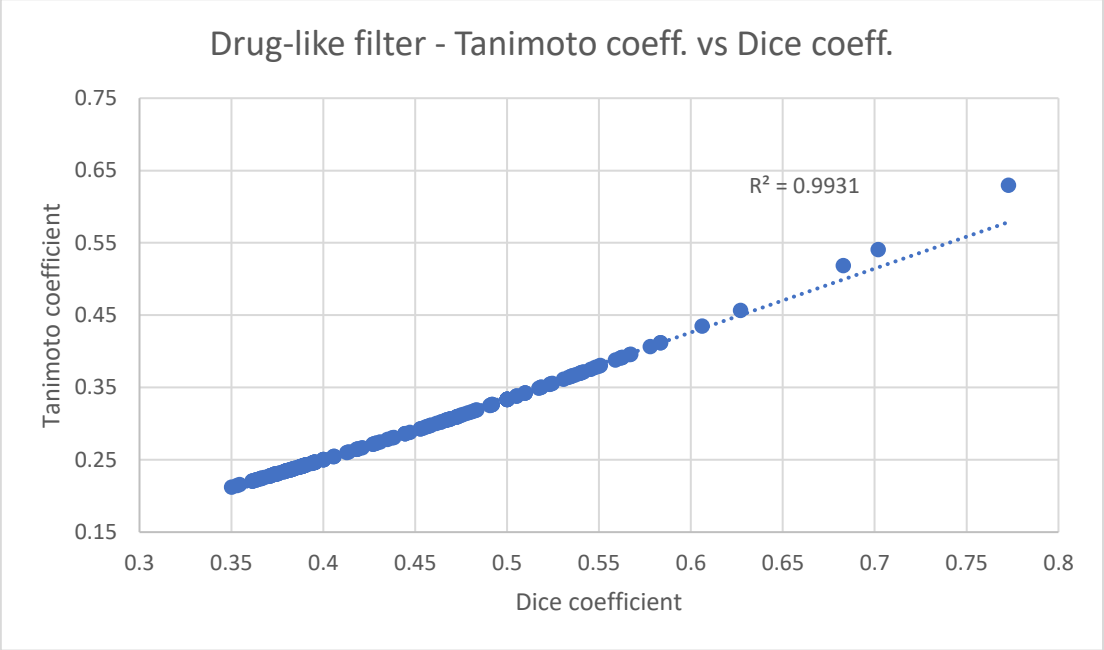


Figure 4.4 Scatter plot of the top 15 Dice and Tanimoto coefficients each active filtered using the Drug-like filter

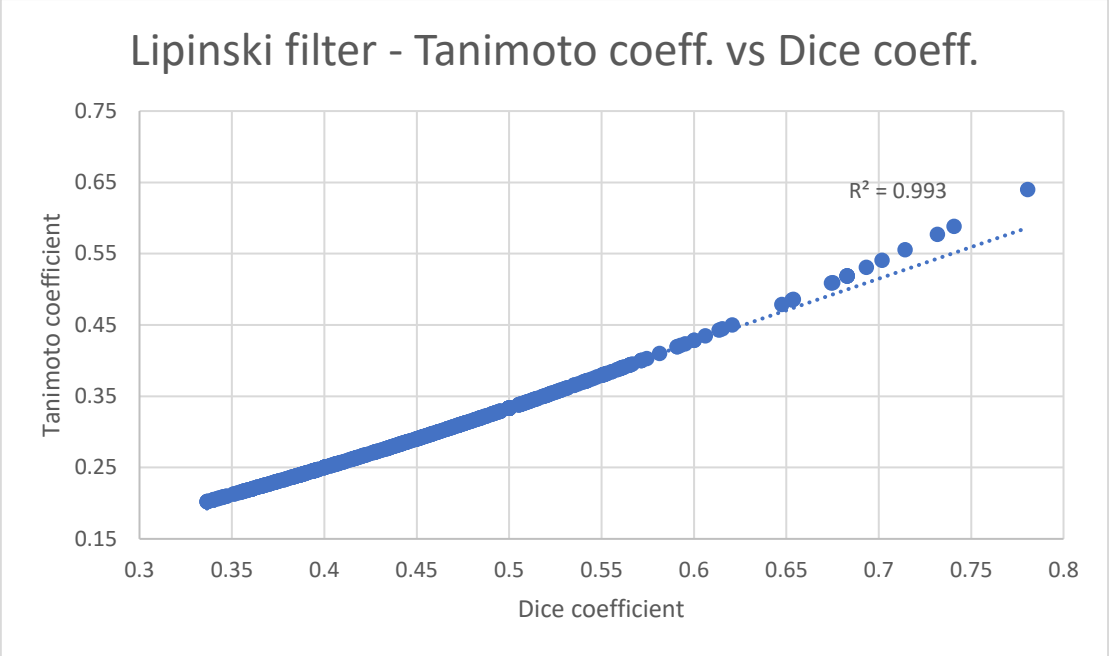


Figure 4.5 Scatter plot of the top 15 Dice and Tanimoto coefficients each active filtered using the Lipinski filter

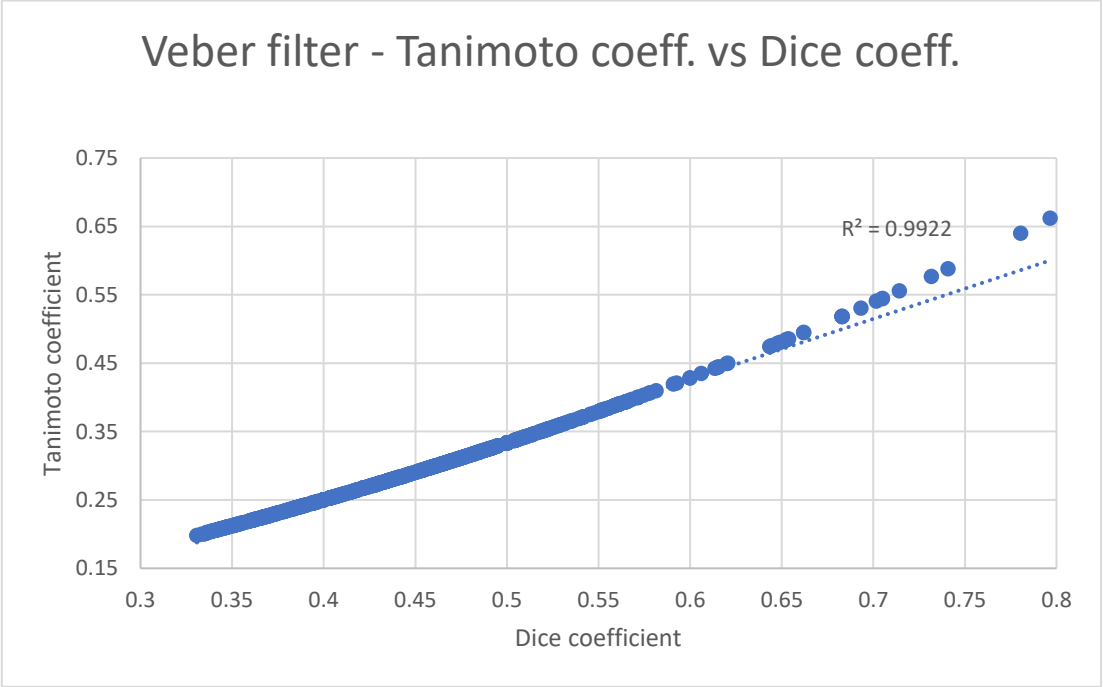


Figure 4.6 Scatter plot of the top 15 Dice and Tanimoto coefficients each active filtered using the Veber filter

Landrum[15] (2021) performed a statistical analysis of the similarity searches using the Tanimoto coefficient on a number of different fingerprints. At a 0.95 noise level, the threshold for a bit vector Morgan fingerprint of radius two is 0.27, and therefore it is recommended that a similarity threshold of

[15] Fingerprint similarity thresholds for database searches. https://greglandrum.github.io/rdkit-blog/similarity/reference/2021/05/21/similarity-search-thresholds.html accessed on 21 July 2022.

0.3 and better still 0.4 be used, with the 0.4 cut-off being the most relevant cut-off value. The noise level refers to the desired user-defined recovery rate, i.e. a noise level of 0.95 would result in a recovery of 95% of the related compounds. In view of the fact that the we have shown a strong correlation between the Dice and Tanimoto coefficients, with the same ranking of molecules being obtained from the similarity search using the Tanimoto, it was decided to only use the results obtained from the Tanimoto similarity searches in order to arrive to a list of hits to the filtered hits.

Table 4.12 Summary of identified hits for the Lipinski, Veber, and Drug-like filters. We are listing the number of actives being searched, the number of molecules obtained using a Tanimoto coefficient cut-off of 0.3 and 0.4. For each cut-off we are listing the number of unique molecules that were identified, i.e. after the removal of duplicate molecules.

	Number of hits		
Tanimoto similarity search	Lipinski	Veber	Drug-like
Actives	41	63	14
Cut-off – 0.3	260	352	75
Unique small molecules	220	282	75
Cut-off – 0.4	31	42	8
Unique small molecules	28	31	8

Table 4.13 Hits identified from Tanimoto similarity search using a 0.4 cut-off (molecules in common shaded in green). SMILES formula are provided in Appendix I

Molecule IDs		
Drug-like	Lipinski	Veber
AACUGO	AISOOA	AISOOA
DAQOZU	IIRONO	IIRONO
BUKUUU	FABKIU	FABKIU
BITVFO	AAFUBO	AAFUBO
AOBGSU	AAIWRI	AAIWRI
	BADVLU	BADVLU
	AAENMO	AAENMO
	EELHGO	EELHGO
	GEERHU	GEERHU
	AEVSIE	AEVSIE
	AOQEZE	AOQEZE
	FIULEI	FIULEI
AABQNE	AABQNE	AABQNE
	BOJEME	BOJEME
AAIAJO	AAIAJO	AAIAJO
	AAAJWE	AAAJWE
	AUDQFU	AUDQFU
	AAMTFE	AAMTFE
	CAPEUA	CAPEUA
AAOPOA	AAOPOA	AAOPOA
	AUEZQE	AUEZQE
	AADZUU	AADZUU
	EAFIQO	LARPRE
	FOOUTU	AONUIA
	FEGYWI	COUEQU
	JUDSCI	FUAVOI
	EUSOWU	BOBDUI
	BUQHJA	EAOMMI
		DEFEHA
		MOMMQU
		AUIRFU

Table 4.14 Overlap of actives from the Lipinski, Veber and Drug-like filters (molecules in common shaded in green). SMILES formula are provided in Appendix I

Molecule IDs			
Lipinski	Veber		Drug-like
AAAABA	AAAABA	AAAAXO	AAAABU
AAAABI	AAAABI	AAAAYE	AAAAHO
AAAABU	AAAABU	AAABAA	AAAAKO
AAAACU	AAAACU	AAABBO	AAAAME
AAAAEU	AAAADE	AAABDE	AAAANE
AAAAHO	AAAADO	AAABFI	AAAANO
AAAAPE	AAAAEA	AAABIE	AAAAOA
AAAAAA	AAAAEI	AAABHE	AAAAOI
AAAAKO	AAAAEU	AAABII	AAAAOU
AAAALA	AAAAHE	AAABIU	AAAAQA
AAAAYA	AAAAHO	AAABJE	AAABLA
AAAALU	AAAAPE	AAABLA	AAABLU
AAAAME	AAAAJO	AAABLI	AAABNE
AAAAPA	AAAAKA	AAABLU	AAACBA
AAAANE	AAAAKO	AAABMU	
AAAANO	AAAALA	AAABNE	
AAAAOA	AAAAYA	AAACAI	
AAAAOI	AAAAME	AAABWI	
AAAAOU	AAAAMO	AAABZE	
AAAAQA	AAAANE	AAACBA	
AAAAQI	AAAANO	AAACBI	
AAAASO	AAAAOA	AAACHA	
AAAATA	AAAAOI		
AAAATU	AAAAOU		
AAAAUE	AAAAQA		
AAABMI	AAAAQI		
AAAAUU	AAAAQU		
AAABAE	AAAARE		
AAAAYE	AAAASE		
AAABAU	AAAASO		
AAABBO	AAAATA		
AAABLA	AAAATI		
AAABLI	AAAATU		
AAABLU	AAAAUE		
AAABMU	AAABMI		
AAABNE	AAAAUU		
AAACAI	AAAAVE		
AAABWI	AAAAVO		
AAACBA	AAABAE		
AAACEE	AAAAWE		
AAACHA	AAAAXE		

When using a Tanimoto coefficient cut-off of 0.3, as summarised in Table 4.12, 260, 352, and 75 hits were identified for the Lipinski, Veber and drug-like filtered actives, respectively. From these, for the Lipinski and Veber filters there were 40 and 70 (Lipinski and Veber, respectively) molecules that were identified as a possible hit for more than one active. One reason for this identification of duplicates is due to the actives list for the Veber and Lipinski filters containing duplicates, such as chloroquine. Another reason is the result of some filtered actives being very similar, differing only in one or a few atoms. As can be seen from Figure 4.7, five actives that were filtered using the Veber filter resulted in the same hit being identified with Tanimoto similarity scores ranging from 0.449 to 0.662.

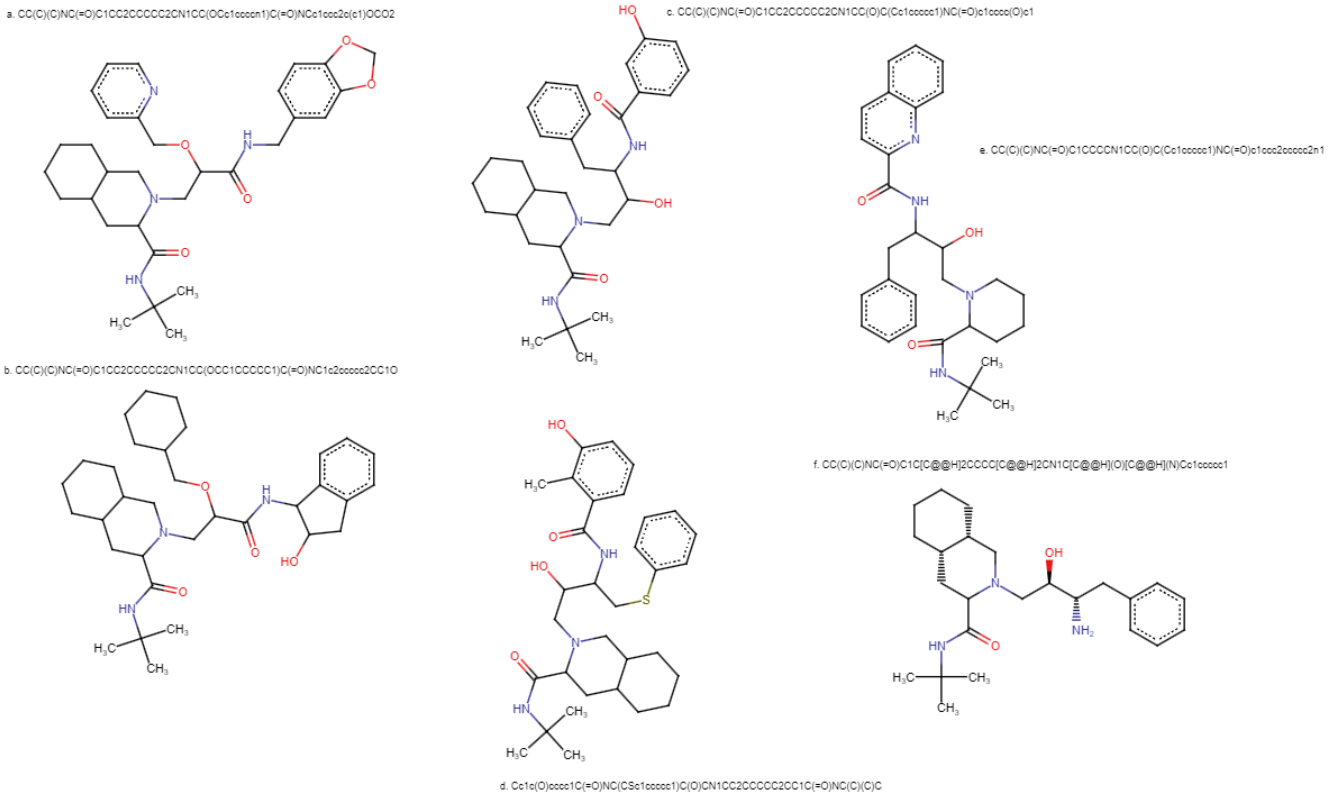


Figure 4.7 *a* to *e* molecular structure of actives which resulted in the identification of the same hit; *f* hit obtained from similarity search of molecules *a* to *e*

The Lipinski actives list has 36 actives in common with the Veber filter. On the other hand, the Lipinski filter has five actives not shared with the Veber filter and the Veber filter has 27 unique actives (see Table 4.15). The actives filtered using the Drug-like filter are all shared with the Veber and Lipinski filters. The Lipinski filter resulted in the identification of 22 hits in common with the Veber filter, and six distinct hits, while the Veber filter resulted in nine distinct/unique hits (see table 4.14). The drug-like filter resulted in three hits in common with the Veber and Lipinski filters and five distinct hits, notwithstanding a complete overlap of actives with the other two filters. These differences can be attributed to the different actives and ZINC15 molecules being filtered with the respective filter. The consolidation of the three lists of hits resulted in a list of 42 hits which is represented in Table 4.15. As

further described in Table 4.16, enrichment factors of 11.55, 6.25 and 4.88 were obtained for the drug-like, Veber and Lipinski drug filters, respectively at a 1% cut-off. This indicates that a good recovery of seeded actives was achieved and therefore the similarity search using the Tanimoto coefficient was effective.

Table 4.15 Consolidated list of hits using Veber, Lipinski and Drug-like filters, using the Tanimoto similarity score cut-off of 0.4

Index	SMILES
AISOOA	CCN(CC)C(=S)S[Sn](c1ccccc1)(c1ccccc1)c1ccccc1
IIRONO	CC(C)(C)N(CCNC(=O)n1cc(F)c(=O)[nH]c1=O)C(=O)O
FABKIU	CCCCCCNC(=O)n1cc(Cl)c(=O)n(C(=O)OCC(C)C)c1=O
AAFUBO	CN(C)CCCNc1ccnc2cc(Cl)ccc12
AAIWRI	Oc1cc(O)cc(C=Cc2ccc(O)c(O)c2)c1
BADVLU	OC[C@H]1O[C@@H](Cc2cc(O)cc(C=Cc3ccc(O)cc3)c2)[C@H](O)[C@@H](O)[C@@H]1O
AAENMO	CC(C)NC(=O)c1cc(Cl)cc(Cl)c1
EELHGO	O=C(Nc1nncs1)C(Cc1ccccc1)NC(=O)N1CCOCC1
GEERHU	CC(N)C(=O)NC(Cc1ccccc1)C(=O)N1CCCC1C(=O)Nc1ccc2ccccc2c1
AEVSIE	O=C(O)CC(O)(CC(=O)O)CC(=O)O
AOQEZE	CCOC(=O)C1=CC(OC(CC)CC)[C@@H]2N[C@@H]2C1
FIULEI	CCC(CC)OC1=CC(C(=O)O)C[C@H](N)[C@H]1NC(C)=O
AABQNE	CCN(CC)C(=O)c1cncc(C(N)=O)c1
BOJEME	CCN(CC)CCN(CC(=O)N(C1CCCCC1)C1CCCCC1)C(=O)c1cccnc1
AAIAJO	CCN(CC)C(=O)c1ccncc1B1OC(C)(C)C(C)(C)O1
AAAJWE	CCN(CC)C(=O)c1ccc(N(C)S(=O)(=O)c2ccccc2)cc1
AUDQFU	CC(C)CC(C(=O)NC(C(=O)OC1CCCC1)c1ccccc1)C(O)C(O)=NO
AAMTFE	CCCC1(C[C@H](C)CC)C(=O)NC(=O)NC1=O
CAPEUA	CC[C@]1(CCC(C)C)C(=O)NC(=O)N(C(=O)c2ccccc2I)C1=O
AAOPOA	O=C1OCCN1N=Cc1ccc([N+](=O)[O-])o1
AUEZQE	N#CC[C](C1CCCC1)n1cc(-c2ncnc3[nH]ccc23)cn1
AADZUU	COc1ccc(C[C@@H](C)NC[C@H](O)c2ccc(O)c3[nH]c(=O)ccc23)cc1
EAFIQO	CCCCCCCCSCCNC(=O)CCn1cc(F)c(=O)[nH]c1=O
FOOUTU	CC(C=O)NC(=O)C(Cc1ccc[nH]1)NC(=O)c1ccccc1
FEGYWI	Nc1ncnc2c(-c3[nH]c(CO)c(O)c3O)c[nH]c12
JUDSCI	O=C(NCCO)NC(=S)c1cccnc1
EUSOWU	Oc1cc(O)c2cc(O)c(-c3ccc(O)c(O)c3)[o+]c2c1
BUQHJA	C[C@@H]1O[C@@H](Oc2cc3c(O)cc(O)cc3cc2-c2cc(O)c(O)c(O)c2)[C@H](O)[C@H](O)[C@H]1O
LARPRE	COc1cc(F)cc(C(=O)Nc2ccccc2-c2cn3c(CN4CCNCC4)csc3n2)c1
AONUIA	CC[C@H](C)C[C@H](C)/C=C(\C)[C@@H]1O[C@H](c2c3c(cn(C)c2=O)[C@]2(O)CCC(=O)C[C@@H]2O3)CC[C@@H]1C
COUEQU	CC(=O)CNC(=O)C(Cc1ccc[nH]1)NC(=O)c1ccccc1
FUAVOI	COc1cc(N2CCC(N)CC2)ccc1Nc1ncc(Cl)c(-c2c[nH]c3ccccc23)n1
BOBDUI	O=C(c1ccc2ccccc2c1)C([C@H](c1ccccc1)N1CCOCC1)N1CCOCC1
EAOMMI	C[N+](C)(C)CCOC(=O)C(O)(CC(=O)O)CC(=O)O
DEFEHA	CC(C)(C)NC(=O)C1C[C@@H]2CCCC[C@@H]2CN1C[C@@H](O)[C@@H](N)Cc1ccccc1
MOMMQU	O=C(N[C@@H]1c2ccccc2C[C@@H]1O)C1COCCN1CC1CCC1
AUIRFU	CC(C)C(=NC(=O)C(Cc1ccc(Cl)cc1)NC(=O)c1ccc(Br)cc1)C(=O)O
AACUGO	COc1cc(C=Cc2ccc(O)cc2)cc(OC)c1
DAQOZU	O=c1cc(O)cc(/C=C/c2ccccc2)o1
BUKUUU	C=C[C@@H](C=Cc1ccc(O)cc1)c1ccc(O)cc1
BITVFO	COc1cc(C(C)=O)ccc1Oc1c(C#N)c(=O)n(C)c(=O)n1C
AOBGSU	CN(C(=O)c1cccnc1)c1nc2ccccc2o1

4.5 Ultrafast Shape Recognition with CREDO Atom Types

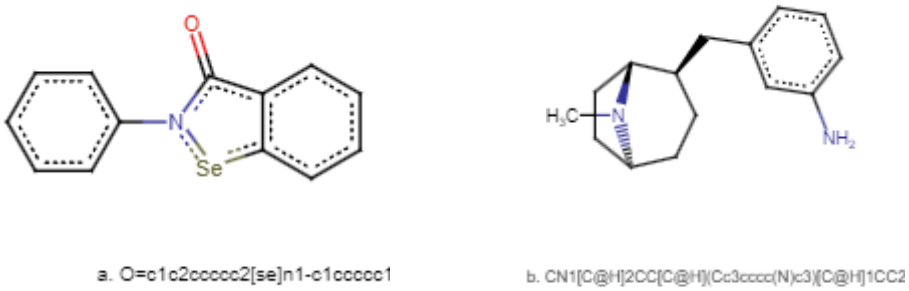
The Ultrafast Shape Recognition (USR) algorithm was developed for the purpose of screening millions of molecules in a relatively short period of time, through the use of 3-dimensional information about the shape of molecules (Ballester, 2011). However, one significant limitation of this algorithm is its inability to discriminate between different pharmacophoric features. Ultrafast Shape Recognition with CREDO Atom Types (USRCAT) is an improvement over USR whereby, the vector describing the molecule includes pharmacophoric information through the inclusion of 48 additional elements to the original 12 USR moments contained in the USR vector (Schreyer and Blundell, 2012).

4.5.1 Conformer generation

The first step in performing a USRCAT similarity search was the generation of the lowest energy conformer for each molecule. This step proved to be the most challenging part of the experiment, in view of the computational time required to derive the lowest energy conformer for the filtered cluster centroids and due to the inability of RDKit to generate the lowest energy conformer for certain molecules. In view of the above constraints, it was decided to proceed only with the cluster centroids obtained from the application of the drug-like filter to the ZINC15 and actives lists. With respect to the ZINC15 list, it was not possible to derive the lowest energy conformer for 878 molecules (out of an initial list of 49,046 molecules), whilst it was not possible to derive the lowest energy conformer for one of the drug-like actives.

In the case of the actives list, it was not possible to generate the lowest energy conformer for Ebselen (O=c1c2ccccc2[se]n1-c1ccccc1). This is most likely due to the unavailability of parameters for force field optimization (UFF) of the sp2 hybridized Se atom[16] in Ebselen (refer to Figure 4.8a). For the other molecules from the ZINC15 filtered list, the most common reason for the failure to generate conformers was the result of the molecules having a fused ring system with a highly constrained stereochemical arrangement (an example of such a molecule is provided in Figure 4.8 and the algorithm being unable to treat these molecules[17].

[16] https://github.com/rdkit/rdkit/issues/2892 last accessed: 30 July 2022
[17] Reply provided by Greg Landrum - https://github.com/rdkit/rdkit/issues/5250 last accessed: 30 July 2022



a. O=c1c2ccccc2[se]n1-c1ccccc1 b. CN1[C@H]2CC[C@H](Cc3cccc(N)c3)[C@H]1CC2

Figure 4.8 Molecular structure of (a.) Ebselen and (b.) one of the molecules with a highly constrained fused ring system

4.5.2 USRCAT

Once the lowest energy conformers for all molecules were generated, a similarity search was performed through the RDKit implementation of USRCAT, by obtaining the USRScore for the pairwise comparison between the respective active molecule and each centroid molecule of the list of filtered molecules. For each active molecule, the 15 molecules with the highest similarity score (using a cut-off of 0.3) were compiled into a list. The molecular structures of these molecules are presented in Appendix I.

When comparing the top 15 similarity scores per active obtained for the USRCAT similarity search with the top 15 similarity scores obtained for the Tanimoto similarity search for the Drug-like filtered actives, the two lists shared only one hit – Index AISYXU – which was identified for different actives. It was also noted that eight molecules scored highly for two different actives, i.e. featured in the top 15 molecules twice (molecules AASNQO, AACKSO, AAUHYO, AAIBTO, AIWBPA, BOODIE, JEZYEO, HOYEWO). The consolidated list of top 15 hits per active is provided in Appendix I.

4.6 Evaluation

In view of the fact that it was not possible to validate the results using *in vitro* methods (being outside the scope of this project), it was decided to use the enrichment factor as a method of evaluating the results. In order to determine the enrichment factor, the lists of molecules filtered by using the different filtering strategies were seeded with their respective list of active molecules. This process had a dual purpose, firstly to use this seeded list for internal validation purposes and secondly to calculate the enrichment factor for each filtered list.

For internal validation purposes, when the lists of small molecules are seeded with actives, the algorithms that were used for the similarity searches should match perfectly the active molecule being queried with the respective molecule in the seeded set, i.e. obtaining a similarity score of 1.0 (i.e. identical molecule) to the respective active being assessed.

The enrichment factor was used to assess the quality of the screening methods that were used to identify the number of seeded actives (Mishra and Basu, 2013). The enrichment factor is the ratio of the fraction of actives in the (user-defined) subset of molecules relative to the fraction of actives in the original dataset, as shown in Equation 4.1. The enrichment factor was calculated for each active molecule within each filtered set of actives and then the average of the enrichment factors was calculated. Table 4.16 presents the average enrichment factors for the respective filtering strategy and at different subset cut-offs.

$$EF_{x\%} = \left(S_{act}/S_{all}\right) \Big/ \left(P_{act}/P_{all}\right)$$

Equation 4.1 $EF_{x\%}$ is the Enrichment Factor at the percentage cut-off x%, P_{all} is the total number of molecules in all of the list, S_{all} is the number of molecules in the subset chosen with the x% cut-off, P_{act} is the number of active molecules in all of the list, and S_{act} is the number of actives found in the subset.

An enrichment factor larger than one indicates that the method outperforms a random selection of compounds, with the bigger the value the better performing the method is (Fechner and Schneider, 2004). As indicated in Table 4.16, the best performing method was the Tanimoto similarity search using the drug-like filtered actives and ZINC15 molecules, this is most likely due to the lower number of ZINC15 (centroid) molecules and relatively high representativeness of the filtered active molecules. This result also confirms that the use of the drug-like filter coupled with the Tanimoto similarity search results in a cost-effective (in terms of computational resources) initial virtual screening method in a drug-development pipeline.

Table 4.16 Average enrichment factors calculated at different percentage cut-offs for the Tanimoto similarity search and USRCAT search for the different filters.

Filter and Similarity search	$EF_{1\%}$	$EF_{2.5\%}$	$EF_{5\%}$	$EF_{10\%}$
Drug-like Tanimoto search	11.55	6.82	3.96	2.47
Veber Tanimoto search	6.25	3.18	2.18	1.64
Lipinski Tanimoto search	4.88	2.88	2.05	1.57
Drug-like USRCAT	2.57	2.31	1.8	1.54

4.7 Discussion

As outlined in Section 1.3 the two main aims of this study are to firstly identify a suitable SARS-CoV-2 protein target and secondly to develop a list of hits that can be used in further studies to further refine the compound selection for a drug discovery pipeline. The two aims have been achieved through the identification of M^{pro} as a suitable protein target for the inhibition of SARS-CoV-2, the identification of possible leads from literature, and the use of the leads to perform further Ligand-Based Virtual

Screening (LBVS) experiments to arrive at a list of hits that can be further studied and refined as part of a drug development exercise.

The exploration-exploitation trade-off refers to a common decision-making problem faced in situations where a decision is taken based on the assumption that it is the optimal one in view of the observed data (exploitation) or not taking the apparently optimal decision, in view of the fact that it is assumed that the available data is not sufficient to truly choose the best option (Angus, 2021). In the context of this study, one is easily led to believe that in view of the amino acid sequence specificity of M^{pro}'s binding (that is not found in other known human proteases), the larger volume of studies that have worked on M^{pro}, the larger number of identified hits and the ready availability of a good protein structure for this protease, make M^{pro} the preferred target for a virtual screening study. After all, these are all pros when conducting such a study. These being:

i. The amino acid sequence specificity is expected to lead to inhibitors that have smaller chances of interacting with other human proteins/enzymes;
ii. A larger volume of published literature (a larger volume of available data) increases the likelihood of a good target selection, model creation, and hit generation;
iii. A larger number of available hits from other studies will facilitate LBVS experiments since these methodologies rely on the availability of identified and active ligands; and
iv. An accurate protein structure is at the basis of SBVS studies.

Nevertheless, we are faced with the dilemma of abandoning the selection of another protein target or virtual screening methodology on the simple basis that it was not 'the obvious/optimal' choice with the data that is available.

The list of 42 hits obtained from the fingerprint similarity search, outlined in Table 4.15, together with the list of hits obtained from USRCAT search, listed in Appendix I, presents a list of compounds identified from 2D-fingerprint search and a (3D) shape-based similarity search. It is to be noted that the very low overlap between the identified hits from the two methods gives a good molecular overview of possible hits, giving a good scope for further work. This minimal overlap was expected in view of the fact that different virtual screening methodologies result in different hits being identified, as a result of different molecular features being used as the basis of the respective virtual screening method. This was also the basis for choosing to perform two different methods.

The enrichment factors that were obtained indicate that the methodology that was used was an effective one, with the Tanimoto similarity search using the Drug-like filter being the most effective method in terms of enrichment factor ($EF_{1\%}$ of 11.55). The list of computational hits should be seen as an initial step in lead identification that can be further worked on through the following suggested processes:

1. Validation step

Whilst noting the computational efficiency of LBVS methods in screening millions of molecules, these methods rely heavily on the activity of the initial ligand used for the similarity search. Macip et al. (2022) report that out of 61 peer-reviewed studies to develop hits against M^{pro} using molecular docking studies, only two studies validated their results *in vitro*, and 18 studies used re-docking as a validation strategy. This indicates that there might be issues in relation to the bioactivity of molecules identified in the literature, especially for those compounds identified from studies that lacked an *in vitro* validation step or other suitable validation steps. Moreover, this study revolved entirely around LBVS experiments and did not include an *in vitro* validation step, which would be recommended to be performed as a further improvement of this study. It is worth pointing out that out of the 194 M^{pro} relevant actives that were identified from the literature review, only 19 compounds were tested in vitro for their bioactivity against M^{pro}.

2. Complementary use of Structure-Based Virtual Screening

The list of computational hits that was developed as part of this study can be used in SBVS experiments for the purpose of further refining the list into a smaller list of compounds of interest. Molecular docking is a suitable SBVS technique that can give a good indication of affinity between the hit and protein of interest. Moreover, since the hits that were identified are cluster centroids, the individual clusters can be explored further to identify compounds that might have a better binding affinity to the binding pocket or possess side groups that present more desirable ADMET properties.

Overall, this study presents a good opportunity for further work that can be used as part of a drug discovery pipeline in the fight against SARS-CoV-2 and provides a good starting point that can be used in a multitude of studies (*in vitro* and *in silico*).

4.8 Summary

The main aim of the study was to develop a list of small molecule hits that can be used to inhibit the protein function of a suitably selected protein target that leads to the inhibition of infection by SARS-CoV-2, which is the virus behind the COVID-19 pandemic. As clearly described in this chapter, these objectives were all achieved through the following:

- **Identification of a suitable protein target (relevant to SARS-CoV-2)** – A literature review was undertaken in order to identify the main protein targets being studied in the fight against SARS-CoV-2. This resulted in the identification of the Main protease (M^{pro}) as the main target protein for this study. This protein was selected on the basis of the fact that the largest number of studies that were identified concentrated on M^{pro}, a large number of hits thus providing a good basis for LBVS experiments, the amino acid sequence of the enzyme is highly conserved,

the criticality of this enzyme to the viral lifecycle, and the cleavage site that lacks closely related homologs in humans (Pillaiyar et al., 2016).

- **Molecular database creation** – Through the identification of suitable actives from the literature review and by selecting the most relevant tranche of the ZINC15 database, a database of small molecules was developed. The actives and small molecules downloaded from ZINC15 were filtered using different drug-filtering strategies. The filtered ZINC15 compounds were then clustered in order to reduce the extent of computational resources that were required to conduct the similarity searches.
- **Ligand-Based Virtual Screening** – The filtered actives were used as the ligands for similarity searches using two different methods:
    1. Morgan fingerprint similarity search using the Tanimoto coefficient; and
    2. Ultrafast Shape Recognition with CREDO Atom Types.

  The results of the two LBVS experiments were assessed through the calculation of enrichment factors and the small molecules that were identified were condensed into a list of hits that can be used/further refined through additional studies (such as Molecular Docking), as part of a drug discovery pipeline.

5

Conclusion

The aim of this dissertation is to identify small molecules that can be used to develop drugs that are effective against infection by SARS-CoV-2, which is the causative virus of COVID-19 infection. In order to achieve this main aim, a protein target (Main protease M^{pro}) was identified as being the most suitable target to use for the purpose of running virtual screening experiments. Two main Ligand-based virtual screening (LBVS) methods, molecular fingerprint similarity search and USRCAT, are used to search for small molecules that have structural similarities to the active ligands that are identified as inhibiting M^{pro}.

In Chapter 2 we give an overview to the biology of SARS-CoV-2, the virus responsible for the COVID-19 pandemic. In particular, we present the main protein targets that are being studied for the purpose of developing drugs against SARS-CoV-2 and the background justifying the selection of M^{pro} as the selected protein target. This choice was based on the basis of the large number of studies concentrating on M^{pro}, the large number of hits that have been identified in the literature as being possibly effective against this protease, the highly conserved nature of the enzyme, the criticality of this enzyme to the viral lifecycle, and the cleavage site that lacks closely related homologues in humans.

We also provide details related to creation of small molecule databases, and methodologies that are used in Ligand-Based Virtual Screening. We also provide detailed information on the use of Morgan Fingerprints and Ultrafast Shape Recognition with CREDO Atom Types (USRCAT) in LBVS. The choice of these two LBVS methods revolved on the fact that:

i. Similarity searches based on Morgan Fingerprints is one of the most used LBVS methodologies in virtual screening experiments in view of their relatively low requirement for computational resources;
ii. USRCAT is very computationally efficient shape-based LBVS method.

Moreover, it was decided to use two different LBVS methods in view of the fact that different virtual screening methodologies result in different hits being generated from each method (due to the different molecular features used as a basis of the respective method). Thus ending in a varied list of hits that can be further refined as further studies.

In Chapter 3 we provide the implementation details used for the purpose of the studies. During the project we identified M^{pro} as a suitable protein target, identified active ligands that can be used in LBVS experiments, developed a small molecule database that is used in the LBVS experiments, implemented algorithms to filter the small molecules and actives on the basis of drug-likeness, clustered these small molecules to ensure good molecular representation (and optimise the use of available resources), and algorithms for similarity searches. In the case of USRCAT, we also generated lowest-energy conformers for each respective molecule.

The results of the experiments are presented in Chapter 4. We provide information on the M^{pro} and related active ligands. We also provide results related to the filters for drug-likeness and from the clustering algorithms. A list of hits based on the similarity searches against the filtered active ligands is provided in this Chapter and in Appendix I. The main outcomes of this study are the following:

- The use of the drug-like filter and Lipinski's Rule of Five is a good filtering approach, on the basis that they provide good molecular coverage, thus ensuring representativeness of selected compounds;
- Molecular clustering based on the sphere exclusion algorithm is a computationally efficient clustering method;
- There is a linear correlation between the results obtained from the fingerprint searches using the Tanimoto and Dice coefficients;
- Morgan fingerprint similarity searching using the Tanimoto coefficient provides a good approach for identifying hits. Using a Tanimoto coefficient cut-off of 0.4, we identified 42 hits that can be further studied;
- A search based on USRCAT generated 179 hits that can be further studied using different approaches, such as in *in vitro* studies or molecular docking;
- The results obtained were adequately validated using enrichment factors.

5.1 Review of Aims and Objectives

The two main aims of this study are to firstly identify a suitable SARS-CoV-2 protein target and secondly to develop a list of hits that can be used in further studies to further refine the compound selection for a drug discovery pipeline.

The two aims have been achieved through:

- the identification of M^{pro} as a suitable protein target for the inhibition of SARS-CoV-2,
- the identification of possible leads from literature,
- the creation of a relational database of small molecules,

- the use of a number of filters for drug-likeness to refine the small molecules and actives,
- the use of molecular clustering to ensure optimal representativeness and computational efficiency, and
- the use of the filtered leads to perform further Ligand-Based Virtual Screening (LBVS) experiments to arrive at a list of hits that can be further studied and refined as part of a drug development exercise.

The main validation of the results was carried out through the calculation of enrichment factors, which indicated that the experiment was an effective one since all the estimated factors were much higher than one (the highest being an $EF_{1\%}$ of 11.55 for the Morgan fingerprint similarity search using the molecules generated from the use of the drug-like filter).

The similarity searches resulted in a list consisting of 42 hits obtained from the fingerprint similarity search, and a list of 179 hits obtained from USRCAT search. It is to be noted that the very low overlap between the identified hits from the two methods gives a good molecular overview of possible hits, giving a good scope for further work. Two LBVS methods were used in view of the fact that different computational LBVS methods result in different hits being identified, and therefore provide higher representation of the chemical space of the identified hits.

5.2 Critique and Limitations

The use of the drug-like filters for selecting actives and ZINC15 molecules provides a good compromise between computational resources that were required for performing the similarity searches and the coverage of molecular diversity. However, this does not involve a full search among all of the ZINC15 tranche that was downloaded or all of the actives that were selected. This might result in the omission of good molecular candidates being assessed for molecular similarity. The fact that apart from filtering for drug-likeness, the molecular sets were clustered further increased the possibility of missing a good molecular candidate. In the case of the USRCAT similarity search, this fact was further exacerbated by the fact that only the drug-like actives were screened. The right balance between efficient use of computational resources and ensuring adequate representativeness of the selected molecular sets served as the justification for the selected approach. It is also pertinent to point out that due to time limitations, it was not possible to refine the results obtained from the similarity searches by exploring the respective cluster of the top scoring small molecules in terms of similarity coefficient. Moreover, since it was out of the scope of this study, it was also not possible to refine the results using *in vitro* methods.

Another limitation of this study relates to the validation of the results, whereby it was not possible to validate the identified hits through *in vitro* or different *in silico* studies. In a review of studies targeting molecular docking experiments to identify M^{pro} inhibitors, Macip et al. (2022) highlight that only a limited number of the reviewed studies performed a validation step that is either based on in vitro testing or a proper re-docking validation step.

The enrichment factor was used to assess the quality of the screening methods that were used to identify the number of seeded actives (Mishra and Basu, 2013). The enrichment factor is the ratio of the fraction of actives in the (user-defined) subset of molecules relative to the fraction of actives in the original dataset. The enrichment factors that were estimated for the respective searches of filtered actives and small molecules provide comfort in view of the fact that all were above the value of 1, which indicates that the similarity searches that were undertaken are better than a random selection of compounds. The best performing method was the Tanimoto similarity search using the drug-like filtered actives and ZINC15 molecules, with an $EF_{1\%}$ of 11.55. This result also confirms that the use of the drug-like filter coupled with the Tanimoto similarity search results in a cost-effective (in terms of computational resources) initial virtual screening method in a drug-development pipeline.

5.3 Future Work

Through the identification of the hits presented in Chapter 4 and listed in Appendix I, we have achieved the two main aims of this study. Nevertheless, a number of limitations and areas for further study have been identified. These can be addressed as part of future work. The following suggested approach can be considered as a further refinement of the results obtained:

i. For the small molecules that scored highest, and especially for those that scored higher than the 0.4 cut-off for the Tanimoto coefficient, further similarity searches would be performed between the respective active and cluster members to which the cluster centroids belong. One of the benefits of performing similarity searches on clustered data is that through the minimisation of the distance between intracluster members and the maximisation of the intercluster distance, there is a reduction in the risk of having an over-representation of inactive compounds (Pasupa and Kudisthalert, 2018). However, this does not mean that the cluster centroid is the best performing molecule from within the cluster, thus meriting further refinement by evaluating the members of the cluster.

ii. The protein structure of M^{pro} is well represented within the PDB database and has been used extensively in molecular docking studies for the identification of inhibitors. Molecular docking is more demanding in terms of computational resources (Ma, Chan, and Leung, 2011) when compared to fingerprint similarity searches or USRCAT. Therefore, the list of hits can be used as a starting point for a Molecular Docking experiment and the molecular docking experiment would itself serve as an *in silico* validation method for the results that have been obtained through this dissertation. Moreover, different cluster members can also be used to identify the best fitting molecule.

iii. The list of hits, once refined through molecular docking as well as including those small molecules that would have scored highly from performing similarity searches within the clusters referred to in *i* above, should be tested *in vitro* in order to assess the enzyme inhibition of any of the identified hits. This would also provide a dose-dependent indication of the activity of the compounds at an *in vitro* level. A comparison between the highest scoring hits identified from the different filtering strategies could also be considered in order to assess the impact of the different filters.

iv. The best performing compounds should be further assessed in terms of their Absorption, Distribution, Metabolism, Excretion, and Toxicology (ADMET) related properties to remove any compounds that present any issues in terms of their ADMET properties. This assessment can be performed computationally, using for example ADMETlab (Dong, et al., 2018) and admetSAR (Yang, et al., 2019).

The greatest challenge that we faced in undertaking this study was related to the molecular clustering of the filtered small molecule sets. It was not possible to immediately identify a clustering computational tool or algorithm that could cluster efficiently the relatively large, small molecule sets. This indicates that such methods are either not readily available or not widely disseminated. Whilst the RDKit implementation of the sphere exclusion algorithm is computationally efficient in identifying the cluster centroids, the population of the individual clusters is highly time consuming. This provides a good opportunity for further work in this field.

5.4 Final Remarks

On a final note, whilst acknowledging that the main aims of this dissertation has been achieved, the results provide significant opportunity to continue developing the science behind the fight against COVID-19. Moreover, these additional studies should be seen in the context of the lead identification process within the drug discovery process.

We firmly believe that the results obtained are a valid contribution to the fight against the SARS-CoV-2 infection, especially when one considers the high likelihood that the virus will persist like seasonal flu, there are people that will require drugs against COVID-19 infection due to them being unable to get vaccinated, and the adaptive resistance of the virus to vaccines through the development of new strains.

References

Abraham, S., Kienzle, T.E., Lapps, W. and Brian, D.A., 1990. Deduced sequence of the bovine coronavirus spike protein and identification of the internal proteolytic cleavage site. Virology, 176(1), pp.296-301.

Aftab, S.O., Ghouri, M.Z., Masood, M.U., Haider, Z., Khan, Z., Ahmad, A. and Munawar, N., 2020. Analysis of SARS-CoV-2 RNA-dependent RNA polymerase as a potential therapeutic drug target using a computational approach. Journal of translational medicine, 18(1), pp.1-15.

Aghaee, E., Ghodrati, M. and Ghasemi, J.B., 2021. In silico exploration of novel protease inhibitors against coronavirus 2019 (COVID-19). Informatics in Medicine Unlocked, 23, p.100516.

Almeida, J.D. and Tyrrell, D.A., 1967. The morphology of three previously uncharacterized human respiratory viruses that grow in organ culture. Journal of General Virology, 1(2), pp.175-178.

Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J.R. and Hilgenfeld, R., 2003. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. Science, 300(5626), pp.1763-1767.

Anderson, J.P., Daifuku, R. and Loeb, L.A., 2004. Viral error catastrophe by mutagenic nucleosides. Annual review of microbiology, 58(1), pp.183-205.

Angus, D.C., 2020. Optimizing the trade-off between learning and doing in a pandemic. Jama, 323(19), pp.1895-1896.

Baggen, J., Vanstreels, E., Jansen, S. and Daelemans, D., 2021. Cellular host factors for SARS-CoV-2 infection. Nature Microbiology, 6(10), pp.1219-1232.

Bajusz, D., Rácz, A. and Héberger, K., 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. Journal of cheminformatics, 7(1), pp.1-13.

Ballester, P.J., 2011. Ultrafast shape recognition: method and applications. Future Medicinal Chemistry, 3(1), pp.65-78.

Ballester, P.J., Finn, P.W. and Richards, W.G., 2009. Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology. Journal of Molecular Graphics and Modelling, 27(7), pp.836-845.

Ballester, P.J. and Richards, W.G., 2007. Ultrafast shape recognition to search compound databases for similar molecular shapes. Journal of computational chemistry, 28(10), pp.1711-1723.

Banegas-Luna, A.J., Cerón-Carrasco, J.P. and Pérez-Sánchez, H., 2018. A review of ligand-based virtual screening web tools and screening algorithms in large molecular databases in the age of big data. Future medicinal chemistry, 10(22), pp.2641-2658.

Barlas, S.B., Adalier, N., Dasdag, O. and Dasdag, S., 2021. Evaluation of SARS-CoV-2 with a biophysical perspective. Biotechnology & Biotechnological Equipment, 35(1), pp.392-406.

Basu, S., Veeraraghavan, B., Ramaiah, S. and Anbarasu, A., 2020. Novel cyclohexanone compound as a potential ligand against SARS-CoV-2 main-protease. Microbial Pathogenesis, 149, p.104546.

Beck, B.R., Shin, B., Choi, Y., Park, S. and Kang, K., 2020. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. Computational and structural biotechnology journal, 18, pp.784-790.

Bender, A. and Glen, R.C., 2005. A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. Journal of chemical information and modeling, 45(5), pp.1369-1375.

Beniac, D.R., Andonov, A., Grudeski, E. and Booth, T.F., 2006. Architecture of the SARS coronavirus prefusion spike. Nature structural & molecular biology, 13(8), pp.751-752.

Berdigaliyev, N. and Aljofan, M., 2020. An overview of drug discovery and development. Future medicinal chemistry, 12(10), pp.939-947.

Berenger, F., Vu, O. and Meiler, J., 2017. Consensus queries in ligand-based virtual screening experiments. Journal of cheminformatics, 9(1), pp.1-13.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., 2000. The protein data bank. Nucleic acids research, 28(1), pp.235-242.

Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S. and Hopkins, A.L., 2012. Quantifying the chemical beauty of drugs. Nature chemistry, 4(2), pp.90-98.

Bohacek, R.S., McMartin, C. and Guida, W.C., 1996. The art and practice of structure-based drug design: a molecular modeling perspective. Medicinal research reviews, 16(1), pp.3-50.

Bonanno, E. and Ebejer, J.P., 2020. Applying Machine Learning to Ultrafast Shape Recognition in Ligand-Based Virtual Screening. Frontiers in pharmacology, 10, p.1675.

Bung, N., Krishnan, S.R., Bulusu, G. and Roy, A., 2021. De novo design of new chemical entities for SARS-CoV-2 using artificial intelligence. Future medicinal chemistry, 13(06), pp.575-585.

Butina, D., 1999. Unsupervised database clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. Journal of Chemical Information and Computer Sciences, 39(4), pp.747-750.

Calligari, P., Bobone, S., Ricci, G. and Bocedi, A., 2020. Molecular investigation of SARS–CoV-2 proteins and their interactions with antiviral drugs. Viruses, 12(4), p.445.

Camus, A., 2000. The Plague:[Excerpts]. Academic Medicine, 75(9), p.945.

Cao, X., 2021. ISG15 secretion exacerbates inflammation in SARS-CoV-2 infection. Nature immunology, 22(11), pp.1360-1362.

Caruso, F.P., Scala, G., Cerulo, L. and Ceccarelli, M., 2021. A review of COVID-19 biomarkers and drug targets: resources and tools. Briefings in bioinformatics, 22(2), pp.701-713.

Cassar, J., 2019. A big data approach for clustering large chemical datasets (Master's thesis, University of Malta).

Cereto-Massagué, A., Ojeda, M.J., Valls, C., Mulero, M., Garcia-Vallvé, S. and Pujadas, G., 2015. Molecular fingerprint similarity search in virtual screening. Methods, 71, pp.58-63.

Cheng, T., Li, Q., Zhou, Z., Wang, Y. and Bryant, S.H., 2012. Structure-based virtual screening for drug discovery: a problem-centric review. The AAPS journal, 14(1), pp.133-141.

Chenthamarakshan, V., Das, P., Hoffman, S.C., Strobelt, H., Padhi, I., Lim, K.W., Hoover, B., Manica, M., Born, J., Laino, T. and Mojsilovic, A., 2020. Cogmol: Target-specific and selective drug design for covid-19 using deep generative models. arXiv preprint arXiv:2004.01215.

Choudhary, S., Malik, Y.S. and Tomar, S., 2020. Identification of SARS-CoV-2 cell entry inhibitors by drug repurposing using in silico structure-based virtual screening approach. Frontiers in immunology, 11, p.1664.

Citarella, A., Scala, A., Piperno, A. and Micale, N., 2021. SARS-CoV-2 Mpro: A Potential Target for Peptidomimetics and Small-Molecule Inhibitors. Biomolecules, 11(4), p.607.

Coelho, C., Gallo, G., Campos, C.B., Hardy, L. and Würtele, M., 2020. Biochemical screening for SARS-CoV-2 main protease inhibitors. PloS one, 15(10), p.e0240079.

Congreve, M., Carr, R., Murray, C. and Jhoti, H., 2003. A 'rule of three' for fragment-based lead discovery?. Drug discovery today, 8(19), pp.876-877

Cuesta, S.A., Mora, J.R. and Márquez, E.A., 2021. In silico screening of the DrugBank database to search for possible drugs against SARS-CoV-2. Molecules, 26(4), p.1100.

Dai, W., Zhang, B., Jiang, X.M., Su, H., Li, J., Zhao, Y., Xie, X., Jin, Z., Peng, J., Liu, F. and Li, C., 2020. Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. Science, 368(6497), pp.1331-1335.

Day, C.J., Bailly, B., Guillon, P., Dirr, L., Jen, F.E.C., Spillings, B.L., Mak, J., von Itzstein, M., Haselhorst, T. and Jennings, M.P., 2021. Multidisciplinary approaches identify compounds that bind to human ACE2 or SARS-CoV-2 spike protein as candidates to block SARS-CoV-2–ACE2 receptor interactions. MBio, 12(2), pp.e03681-20.

Ebejer, J.P., 2014. Data driven approaches to improve the drug discovery process: a virtual screening quest in drug discovery (Doctoral dissertation, University of Oxford).

Ebejer, J P. (2021) Standardizing a molecule using RDKit. Available at: https://bitsilla.com/blog/2021/06/standardizing-a-molecule-using-rdkit/ (Accessed: 21 August 2021)

Eweas, A.F., Maghrabi, I.A. and Namarneh, A.I., 2014. Advances in molecular modeling and docking as a tool for modern drug discovery. Der Pharma Chemica, 6(6), pp.211-228.

Fechner, U. and Schneider, G., 2004. Evaluation of distance metrics for ligand-based similarity searching. ChemBioChem, 5(4), pp.538-540.

Fehr, A.R. and Perlman, S., 2015. Coronaviruses: an overview of their replication and pathogenesis. Coronaviruses, pp.1-23.

Fischer, A., Sellner, M., Neranjan, S., Smieško, M. and Lill, M.A., 2020. Potential inhibitors for novel coronavirus protease identified by virtual screening of 606 million compounds. International journal of molecular sciences, 21(10), p.3626.

Forrestall, K.L., Burley, D.E., Cash, M.K., Pottie, I.R. and Darvesh, S., 2021. 2-Pyridone natural products as inhibitors of SARS-CoV-2 main protease. Chemico-biological interactions, 335, p.109348.

Franco-Paredes, C., 2020. Albert Camus' the Plague revisited for COVID-19. Clinical Infectious Diseases, 71(15), pp.898-899.

Ganesan, A., 2008. The impact of natural products upon modern drug discovery. Current opinion in chemical biology, 12(3), pp.306-317.

Gangadevi, S., Badavath, V.N., Thakur, A., Yin, N., De Jonghe, S., Acevedo, O., Jochmans, D., Leyssen, P., Wang, K., Neyts, J. and Yujie, T., 2021. Kobophenol A inhibits binding of host ACE2 receptor with spike RBD domain of SARS-CoV-2, a lead compound for blocking COVID-19. The journal of physical chemistry letters, 12(7), pp.1793-1802.

Gao, Y., Yan, L., Huang, Y., Liu, F., Zhao, Y., Cao, L., Wang, T., Sun, Q., Ming, Z., Zhang, L. and Ge, J., 2020. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science, 368(6492), pp.779-782.

Gaudêncio, S.P. and Pereira, F., 2020. A computer-aided drug design approach to predict marine drug-like leads for SARS-CoV-2 main protease inhibition. Marine drugs, 18(12), p.633.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E. and Davies, M., 2017. The ChEMBL database in 2017. Nucleic acids research, 45(D1), pp.D945-D954.

Gentile, F., Yaacoub, J.C., Gleave, J., Fernandez, M., Ton, A.T., Ban, F., Stern, A. and Cherkasov, A., 2022. Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. Nature Protocols, 17(3), pp.672-697.

Ghose, A.K., Viswanadhan, V.N. and Wendoloski, J.J., 1999. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. Journal of combinatorial chemistry, 1(1), pp.55-68.

Gil, C., Ginex, T., Maestro, I., Nozal, V., Barrado-Gil, L., Cuesta-Geijo, M.Á., Urquiza, J., Ramírez, D., Alonso, C., Campillo, N.E. and Martinez, A., 2020. COVID-19: drug targets and potential treatments. Journal of medicinal chemistry, 63(21), pp.12359-12386.

Gimeno, A., Ojeda-Montes, M.J., Tomás-Hernández, S., Cereto-Massagué, A., Beltrán-Debón, R., Mulero, M., Pujadas, G. and Garcia-Vallvé, S., 2019. The light and dark sides of virtual screening: what is there to know?. International journal of molecular sciences, 20(6), p.1375.

Good, S.S., Westover, J., Jung, K.H., Zhou, X.J., Moussa, A., La Colla, P., Collu, G., Canard, B. and Sommadossi, J.P., 2021. AT-527, a double prodrug of a guanosine nucleotide analog, is a potent inhibitor of SARS-CoV-2 in vitro and a promising oral antiviral for treatment of COVID-19. Antimicrobial Agents and Chemotherapy, 65(4), pp.e02479-20.

Gossen, J., Albani, S., Hanke, A., Joseph, B.P., Bergh, C., Kuzikov, M., Costanzi, E., Manelfi, C., Storici, P., Gribbon, P. and Beccari, A.R., 2021. A blueprint for high affinity SARS-CoV-

2 Mpro inhibitors from activity-based compound library screening guided by analysis of protein dynamics. ACS pharmacology & translational science, 4(3), pp.1079-1095.

Grosdidier, A., Zoete, V. and Michielin, O., 2011. SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic acids research, 39(suppl_2), pp.W270-W277.

Gurung, A.B., Ali, M.A., Lee, J., Farah, M.A. and Al-Anazi, K.M., 2021. Identification of potential SARS-CoV-2 entry inhibitors by targeting the interface region between the spike RBD and human ACE2. Journal of infection and public health, 14(2), pp.227-237.

Hajbabaie, R., Harper, M.T. and Rahman, T., 2021. Establishing an Analogue Based In Silico Pipeline in the Pursuit of Novel Inhibitory Scaffolds against the SARS Coronavirus 2 Papain-Like Protease. Molecules, 26(4), p.1134.

Hamza, A., Wei, N.N. and Zhan, C.G., 2012. Ligand-based virtual screening approach using a new scoring function. Journal of chemical information and modeling, 52(4), pp.963-974.

Harapan, H., Itoh, N., Yufika, A., Winardi, W., Keam, S., Te, H., Megawati, D., Hayati, Z., Wagner, A.L. and Mudatsir, M., 2020. Coronavirus disease 2019 (COVID-19): A literature review. Journal of infection and public health.

Heller, S.R., McNaught, A., Pletnev, I., Stein, S. and Tchekhovskoi, D., 2015. InChI, the IUPAC international chemical identifier. Journal of cheminformatics, 7(1), pp.1-34.

Hilgenfeld, R., 2014. From SARS to MERS: crystallographic studies on coronaviral proteases enable antiviral drug design. The FEBS journal, 281(18), pp.4085-4096.

Hoffmann, M., Kleine-Weber, H. and Pöhlmann, S., 2020. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. Molecular cell, 78(4), pp.779-784.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.H., Nitsche, A. and Müller, M.A., 2020. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. cell, 181(2), pp.271-280.

Hou, T. and Xu, X., 2004. Recent development and application of virtual screening in drug discovery: an overview. Current pharmaceutical design, 10(9), pp.1011-1033.

Huang, Y., Yang, C., Xu, X.F., Xu, W. and Liu, S.W., 2020. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacologica Sinica, 41(9), pp.1141-1149.

Hudson, B.D., Hyde, R.M., Rahr, E., Wood, J. and Osman, J., 1996. Parameter based methods for compound selection from chemical databases. Quantitative Structure-Activity Relationships, 15(4), pp.285-289.

Jarvis, R.A. and Patrick, E.A., 1973. Clustering using a similarity measure based on shared near neighbors. IEEE Transactions on computers, 100(11), pp.1025-1034.

Jiang, C., Yao, X., Zhao, Y., Wu, J., Huang, P., Pan, C., Liu, S. and Pan, C., 2020. Comparative review of respiratory diseases caused by coronaviruses and influenza A viruses during epidemic season. Microbes and infection, 22(6-7), pp.236-244.

Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C. and Duan, Y., 2020. Structure of M pro from SARS-CoV-2 and discovery of its inhibitors. Nature, 582(7811), pp.289-293.

Johnson, M.A. and Maggiora, G.M., 1990. Concepts and applications of molecular similarity. Wiley.

Kandeel, M., Kitade, Y., Fayez, M., Venugopala, K.N. and Ibrahim, A., 2021. The emerging SARS-CoV-2 papain-like protease: Its relationship with recent coronavirus epidemics. Journal of Medical Virology, 93(3), pp.1581-1588.

Katzourakis, A., 2022. COVID-19: endemic doesn't mean harmless. Nature, pp.485-485.

Keretsu, S., Bhujbal, S.P. and Cho, S.J., 2020. Rational approach toward COVID-19 main protease inhibitors via molecular docking, molecular dynamics simulation and free energy calculation. Scientific reports, 10(1), pp.1-14.

Klemm, T., Ebert, G., Calleja, D.J., Allison, C.C., Richardson, L.W., Bernardini, J.P., Lu, B.G., Kuchel, N.W., Grohmann, C., Shibata, Y. and Gan, Z.Y., 2020. Mechanism and inhibition of the papain-like protease, PLpro, of SARS-CoV-2. The EMBO journal, 39(18), p.e106275.

Krumm, Z.A., Lloyd, G.M., Francis, C.P., Nasif, L.H., Mitchell, D.A., Golde, T.E., Giasson, B.I. and Xia, Y., 2021. Precision therapeutic targets for COVID-19. Virology journal, 18(1), pp.1-22.

Kucera, T., 2016. Virtual screening in drug design–overview of most frequent techniques. Military Medical Science Letters, 85(2), pp.75-79

Kulkarni, S.A. and Ingale, K., 2022. In Silico Approaches for Drug Repurposing for SARS-CoV-2 Infection.

Kumar, S., Sharma, P.P., Shankar, U., Kumar, D., Joshi, S.K., Pena, L., Durvasula, R., Kumar, A., Kempaiah, P., Poonam and Rathi, B., 2020. Discovery of new hydroxyethylamine analogs

against 3CLpro protein target of SARS-CoV-2: Molecular docking, molecular dynamics simulation, and structure–activity relationship studies. Journal of chemical information and modeling, 60(12), pp.5754-5770.

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L. and Wang, X., 2020. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature, 581(7807), pp.215-220.

Landrum, G. (2021) RSC OpenScience Standardization 202104. Available at: https://github.com/greglandrum/RSC_OpenScience_Standardization_202104/blob/main/MolStandardize%20pieces.ipynb (Accessed: 21 August 2021)

Lavecchia, A. and Di Giovanni, C., 2013. Virtual screening strategies in drug discovery: a critical review. Current medicinal chemistry, 20(23), pp.2839-2860.

Leach, A.R. and Gillet, V.J., 2007. An introduction to chemoinformatics. Springer.

Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J., 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced drug delivery reviews, 23(1-3), pp.3-25.

Liu, Y., Gan, J., Wang, R., Yang, X., Xiao, Z. and Cao, Y., 2022. DrugDevCovid19: An atlas of anti-COVID-19 compounds derived by computer-aided drug design. Molecules, 27(3), p.683.

Liu, X. and Wang, X.J., 2020. Potential inhibitors against 2019-nCoV coronavirus M protease from clinically approved medicines. Journal of Genetics and Genomics, 47(2), p.119.

Lloyd, S., 1982. Least squares quantization in PCM. IEEE transactions on information theory, 28(2), pp.129-137.

Luan, B., Huynh, T., Cheng, X., Lan, G. and Wang, H.R., 2020. Targeting proteases for treating COVID-19. Journal of proteome research, 19(11), pp.4316-4326.

Luk, H.K., Li, X., Fung, J., Lau, S.K. and Woo, P.C., 2019. Molecular epidemiology, evolution and phylogeny of SARS coronavirus. Infection, Genetics and Evolution, 71, pp.21-30.

Macalino, S.J.Y., Gosu, V., Hong, S. and Choi, S., 2015. Role of computer-aided drug design in modern drug discovery. Archives of pharmacal research, 38(9), pp.1686-1701.

Macip, G., Garcia-Segura, P., Mestres-Truyol, J., Saldivar-Espinoza, B., Ojeda-Montes, M.J., Gimeno, A., Cereto-Massagué, A., Garcia-Vallvé, S. and Pujadas, G., 2021. Haste makes waste: A critical review of docking-based virtual screening in drug repurposing for SARS-CoV-2 main protease (M-pro) inhibition. Medicinal research reviews.

Maier, H.J., Bickerton, E. and Britton, P., 2015. Coronaviruses. Methods in molecular biology.

Manandhar, A., Srinivasulu, V., Hamad, M., Tarazi, H., Omar, H., Colussi, D.J., Gordon, J., Childers, W., Klein, M.L., Al-Tel, T.H. and Abou-Gharbia, M., 2021. Discovery of Novel Small-Molecule Inhibitors of SARS-CoV-2 Main Protease as Potential Leads for COVID-19 Treatment. Journal of Chemical Information and Modeling, 61(9), pp.4745-4757.

Mishra, N. and Basu, A., 2013. Exploring different virtual screening strategies for acetylcholinesterase inhibitors. BioMed research international, 2013.

Mody, V., Ho, J., Wills, S., Mawri, A., Lawson, L., Ebert, M.C., Fortin, G.M., Rayalam, S. and Taval, S., 2021. Identification of 3-chymotrypsin like protease (3CLPro) inhibitors as potential anti-SARS-CoV-2 agents. Communications biology, 4(1), pp.1-10.

Mohamadian, M., Chiti, H., Shoghli, A., Biglari, S., Parsamanesh, N. and Esmaeilzadeh, A., 2021. COVID-19: Virology, biology and novel laboratory diagnosis. The journal of gene medicine, 23(2), p.e3303.

Morgan, H.L., 1965. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. Journal of chemical documentation, 5(2), pp.107-113.

Murugan, N.A., Kumar, S., Jeyakanthan, J. and Srivastava, V., 2020. Searching for target-specific and multi-targeting organics for Covid-19 in the Drugbank database with a double scoring approach. Scientific reports, 10(1), pp.1-16.

Nal, B., Chan, C., Kien, F., Siu, L., Tse, J., Chu, K., Kam, J., Staropoli, I., Crescenzo-Chaigne, B., Escriou, N. and van der Werf, S., 2005. Differential maturation and subcellular localization of severe acute respiratory syndrome coronavirus surface proteins S, M and E. Journal of general virology, 86(5), pp.1423-1434.

Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M. and Agha, R., 2020. The socio-economic implications of the coronavirus pandemic (COVID-19): A review. International journal of surgery, 78, pp.185-193.

Ogedjo, M., Onoka, I., Sahini, M. and Shadrack, D.M., 2021. Accommodating receptor flexibility and free energy calculation to reduce false positive binders in the discovery of natural products blockers of SARS-COV-2 spike RBD-ACE2 interface. Biochemistry and biophysics reports, 27, p.101024.

Oprea, T.I., 2000. Property distribution of drug-related chemical databases. Journal of computer-aided molecular design, 14(3), pp.251-264.

Osipiuk, J., Azizi, S.A., Dvorkin, S., Endres, M., Jedrzejczak, R., Jones, K.A., Kang, S., Kathayat, R.S., Kim, Y., Lisnyak, V.G. and Maki, S.L., 2021. Structure of papain-like protease from SARS-CoV-2 and its complexes with non-covalent inhibitors. Nature communications, 12(1), pp.1-9.

Pandey, S. and Singh, B.K., 2017. Current advances and new mindset in computer-aided drug design: A review. The Pharma Innovation, 6(8, Part B), p.72.

Pillaiyar, T., Manickam, M., Namasivayam, V., Hayashi, Y. and Jung, S.H., 2016. An overview of severe acute respiratory syndrome–coronavirus (SARS-CoV) 3CL protease inhibitors: peptidomimetics and small molecule chemotherapy. Journal of medicinal chemistry, 59(14), pp.6595-6628.

Pitsillou, E., Liang, J., Karagiannis, C., Ververis, K., Darmawan, K.K., Ng, K., Hung, A. and Karagiannis, T.C., 2020. Interaction of small molecules with the SARS-CoV-2 main protease in silico and in vitro validation of potential lead compounds using an enzyme-linked immunosorbent assay. Computational biology and chemistry, 89, p.107408.

Pitsillou, E., Liang, J., Ververis, K., Hung, A. and Karagiannis, T.C., 2021. Interaction of small molecules with the SARS-CoV-2 papain-like protease: In silico studies and in vitro validation of protease activity inhibition using an enzymatic inhibition assay. Journal of Molecular Graphics and Modelling, 104, p.107851.

Pollastri, M.P., 2010. Overview on the Rule of Five. Current protocols in pharmacology, 49(1), pp.9-12.

Prieto-Martínez, F.D., López-López, E., Juárez-Mercado, K.E. and Medina-Franco, J.L., 2019. Computational drug design methods—current and future perspectives. In silico drug design, pp.19-44.

Protti, Í.F., Rodrigues, D.R., Fonseca, S.K., Alves, R.J., de Oliveira, R.B. and Maltarollo, V.G., 2021. Do Drug-likeness Rules Apply to Oral Prodrugs?. ChemMedChem, 16(9), pp.1446-1456.

Puttaswamy, H., Gowtham, H.G., Ojha, M.D., Yadav, A., Choudhir, G., Raguraman, V., Kongkham, B., Selvaraju, K., Shareef, S., Gehlot, P. and Ahamed, F., 2020. In silico studies evidenced the role of structurally diverse plant secondary metabolites in reducing SARS-CoV-2 pathogenesis. Scientific reports, 10(1), pp.1-24.

Ramasamy, S. and Subbian, S., 2021. Critical determinants of cytokine storm and type I interferon response in COVID-19 pathogenesis. Clinical microbiology reviews, 34(3), pp.e00299-20.

RDKit: Open-source cheminformatics; http://www.rdkit.org

RDKit: Fingerprint Thresholds, 2018. Available online: http://rdkit.blogspot.com/2013/10/fingerprint-thresholds.html (last accessed: 5 June 2022)

Reyaz, S., Tasneem, A., Rai, G.P. and Bairagya, H.R., 2021. Investigation of structural analogs of hydroxychloroquine for SARS-CoV-2 main protease (Mpro): A computational drug discovery study. Journal of Molecular Graphics and Modelling, 109, p.108021.

Reymond, J.L., Van Deursen, R., Blum, L.C. and Ruddigkeit, L., 2010. Chemical space as a source for new drugs. MedChemComm, 1(1), pp.30-38.

Riniker, S. and Landrum, G.A., 2013. Open-source platform to benchmark fingerprints for ligand-based virtual screening. Journal of cheminformatics, 5(1), pp.1-17.

Riniker, S. and Landrum, G.A., 2015. Better informed distance geometry: using what we know to improve conformation generation. Journal of chemical information and modeling, 55(12), pp.2562-2574.

Rogers, D. and Hahn, M., 2010. Extended-connectivity fingerprints. Journal of chemical information and modeling, 50(5), pp.742-754.

Rohaim, M.A., El Naggar, R.F., Clayton, E. and Munir, M., 2020. Structural and functional insights into non-structural proteins of coronaviruses. Microbial pathogenesis, p.104641.

Romeo, A., Iacovelli, F. and Falconi, M., 2020. Targeting the SARS-CoV-2 spike glycoprotein prefusion conformation: virtual screening and molecular dynamics simulations applied to the identification of potential fusion inhibitors. Virus research, 286, p.198068.

Saxena, A., 2020. Drug targets for COVID-19 therapeutics: Ongoing global efforts. Journal of biosciences, 45(1), pp.1-24.

Saxena, A.K. and Prathipati, P., 2006. Collection and preparation of molecular databases for virtual screening. SAR and QSAR in Environmental Research, 17(4), pp.371-392.

Schenone, M., Dančík, V., Wagner, B.K. and Clemons, P.A., 2013. Target identification and mechanism of action in chemical biology and drug discovery. Nature chemical biology, 9(4), pp.232-240.

Schreyer, A.M. and Blundell, T., 2012. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. Journal of cheminformatics, 4(1), pp.1-12.

Seifert, M.H., Kraus, J. and Kramer, B., 2007. Virtual high-throughput screening of molecular databases. Current opinion in drug discovery & development, 10(3), pp.298-307.

Shen, Z., Ratia, K., Cooper, L., Kong, D., Lee, H., Kwon, Y., Li, Y., Alqarni, S., Huang, F., Dubrovskyi, O. and Rong, L., 2021. Design of SARS-CoV-2 PLpro Inhibitors for COVID-19 Antiviral Therapy Leveraging Binding Cooperativity. Journal of medicinal chemistry.

Shitrit, A., Zaidman, D., Kalid, O., Bloch, I., Doron, D., Yarnizky, T., Buch, I., Segev, I., Ben-Zeev, E., Segev, E. and Kobiler, O., 2020. Conserved interactions required for inhibition of the main protease of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Scientific reports, 10(1), pp.1-11.

Shrotri, Swinnen, Kampmann, Parker (2021). An interactive website tracking COVID-19 vaccine development. Lancet Glob Health; 9(5):e590-e592

Singh, P.K., Kulsum, U., Rufai, S.B., Mudliar, S.R. and Singh, S., 2020. Mutations in SARS-CoV-2 leading to antigenic variations in spike protein: a challenge in vaccine development. Journal of laboratory physicians, 12(02), pp.154-160.

Sliwoski, G., Kothiwale, S., Meiler, J. and Lowe, E.W., 2014. Computational methods in drug discovery. Pharmacological reviews, 66(1), pp.334-395.

Squonk Computational Notebook. https://squonk.it/docs/cells/Drug-like%20Filter%20(CXN)/ accessed on 5 August 2021

Stebbing, J., Phelan, A., Griffin, I., Tucker, C., Oechsle, O., Smith, D. and Richardson, P., 2020. COVID-19: combining antiviral and anti-inflammatory treatments. The Lancet Infectious Diseases, 20(4), pp.400-402.

Sterling, T. and Irwin, J.J., 2015. ZINC 15–ligand discovery for everyone. Journal of chemical information and modeling, 55(11), pp.2324-2337.

Subramaniam, S., Mehrotra, M. and Gupta, D., 2008. Virtual high throughput screening (vHTS)-A perspective. Bioinformation, 3(1), p.14.

Tari, L.W. ed., 2012. Structure-based drug discovery. Humana Press.

Taylor, R., 1995. Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals. Journal of Chemical Information and Computer Sciences, 35(1), pp.59-67.

Tejera, E., Munteanu, C.R., López-Cortés, A., Cabrera-Andrade, A. and Pérez-Castillo, Y., 2020. Drugs repurposing using QSAR, docking and molecular dynamics for possible inhibitors of the SARS-CoV-2 Mpro protease. Molecules, 25(21), p.5172.

Torres, P.H., Sodero, A.C., Jofily, P. and Silva-Jr, F.P., 2019. Key topics in molecular docking for drug design. International journal of molecular sciences, 20(18), p.4574.

Trott, O. and Olson, A.J., 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry, 31(2), pp.455-461.

Tyrrell, D.A.J. and Fielder, M., 2002. Cold wars: the fight against the common cold. Oxford University Press, USA.

van de Leemput, J. and Han, Z., 2021. Understanding individual SARS-CoV-2 proteins for targeted drug development against COVID-19. Molecular and Cellular Biology, 41(9), pp.e00185-21.

Vázquez, J., López, M., Gibert, E., Herrero, E. and Luque, F.J., 2020. Merging ligand-based and structure-based methods in drug discovery: An overview of combined virtual screening approaches. Molecules, 25(20), p.4723.

Veber, D.F., Johnson, S.R., Cheng, H.Y., Smith, B.R., Ward, K.W. and Kopple, K.D., 2002. Molecular properties that influence the oral bioavailability of drug candidates. Journal of medicinal chemistry, 45(12), pp.2615-2623.

Vuong, W., Khan, M.B., Fischer, C., Arutyunova, E., Lamer, T., Shields, J., Saffran, H.A., McKay, R.T., van Belkum, M.J., Joyce, M.A. and Young, H.S., 2020. Feline coronavirus drug inhibits the main protease of SARS-CoV-2 and blocks virus replication. Nature communications, 11(1), pp.1-8.

Vuong, W., Fischer, C., Khan, M.B., van Belkum, M.J., Lamer, T., Willoughby, K.D., Lu, J., Arutyunova, E., Joyce, M.A., Saffran, H.A. and Shields, J.A., 2021. IMproved SARS-CoV-2 Mpro inhibitors based on feline antiviral drug GC376: structural enhancements, increased solubility, and micellar studies. European journal of medicinal chemistry, p.113584.

Wahl, A., Gralinski, L.E., Johnson, C.E., Yao, W., Kovarova, M., Dinnon, K.H., Liu, H., Madden, V.J., Krzystek, H.M., De, C. and White, K.K., 2021. SARS-CoV-2 infection is effectively treated and prevented by EIDD-2801. Nature, 591(7850), pp.451-457.

Walters, W.P. and Namchuk, M., 2003. Designing screens: how to make your hits a hit. Nature reviews Drug discovery, 2(4), pp.259-266.

Walters, W.P., Murcko, A.A. and Murcko, M.A., 1999. Recognizing molecules with drug-like properties. Current opinion in chemical biology, 3(4), pp.384-387.

Walters, W.P., Stahl, M.T. and Murcko, M.A., 1998. Virtual screening—an overview. Drug discovery today, 3(4), pp.160-178.

Wang, Q., Yang, L., Jin, H. and Lin, L., 2021. Vaccination against COVID-19: A systematic review and meta-analysis of acceptability and its predictors. Preventive medicine, 150, p.106694.

Wermuth, C.G., Ganellin, C.R., Lindberg, P. and Mitscher, L.A., 1998. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). Pure and Applied Chemistry, 70(5), pp.1129-1143.

Westermaier, Y., Barril, X. and Scapozza, L., 2015. Virtual screening: an in silico tool for interlacing the chemical universe with the proteome. Methods, 71, pp.44-57.

WHO COVID-19 Dashboard. Geneva: World Health Organization, 2020. Available online: https://covid19.who.int/ (last accessed: 27 July 2022).

Willett, P., 2006. Similarity-based virtual screening using 2D fingerprints. Drug discovery today, 11(23-24), pp.1046-1053.

Willett, P., Barnard, J.M. and Downs, G.M., 1998. Chemical similarity searching. Journal of chemical information and computer sciences, 38(6), pp.983-996.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z. and Assempour, N., 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research, 46(D1), pp.D1074-D1082.

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z., Zhang, Z., Wang, J. and Sheng, J., 2020. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell host & microbe, 27(3), pp.325-328.

Xia, S., Lan, Q., Su, S., Wang, X., Xu, W., Liu, Z., Zhu, Y., Wang, Q., Lu, L. and Jiang, S., 2020a. The role of furin cleavage site in SARS-CoV-2 spike protein-mediated membrane fusion in the presence or absence of trypsin. Signal transduction and targeted therapy, 5(1), pp.1-3.

Xia, S., Liu, M., Wang, C., Xu, W., Lan, Q., Feng, S., Qi, F., Bao, L., Du, L., Liu, S. and Qin, C., 2020b. Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. Cell research, 30(4), pp.343-355.

Xu, C., Ke, Z., Liu, C., Wang, Z., Liu, D., Zhang, L., Wang, J., He, W., Xu, Z., Li, Y. and Yang, Y., 2020. Systemic in silico screening in drug discovery for Coronavirus Disease

(COVID-19) with an online interactive web server. Journal of chemical information and modeling, 60(12), pp.5735-5745.

Yadav, R., Chaudhary, J.K., Jain, N., Chaudhary, P.K., Khanra, S., Dhamija, P., Sharma, A., Kumar, A. and Handu, S., 2021. Role of structural and non-structural proteins and therapeutic targets of SARS-CoV-2 for COVID-19. Cells, 10(4), p.821.

Yan, S. and Wu, G., 2021. Spatial and temporal roles of SARS-CoV PLpro—A snapshot. The FASEB Journal, 35(1), p.e21197.

Yang, L., Pei, R.J., Li, H., Ma, X.N., Zhou, Y., Zhu, F.H., He, P.L., Tang, W., Zhang, Y.C., Xiong, J. and Xiao, S.Q., 2020. Identification of SARS-CoV-2 entry inhibitors among already approved drugs. Acta Pharmacologica Sinica, pp.1-7.

Ye, Z.W., Yuan, S., Yuen, K.S., Fung, S.Y., Chan, C.P. and Jin, D.Y., 2020. Zoonotic origins of human coronaviruses. International journal of biological sciences, 16(10), p.1686.

Young, D.C., 2009. Computational drug design: a guide for computational and medicinal chemists. John Wiley & Sons.

Zeng, F., Hon, C.C., Yip, C.W., Law, K.M., Yeung, Y.S., Chan, K.H., Peiris, J.S.M. and Leung, F.C.C., 2006. Quantitative comparison of the efficiency of antibodies against S1 and S2 subunit of SARS coronavirus spike protein in virus neutralization and blocking of receptor binding: implications for the functional roles of S2 subunit. FEBS letters, 580(24), pp.5612-5620.

Zhang, X., Li, S. and Niu, S., 2020. ACE2 and COVID-19 and the resulting ARDS. Postgraduate medical journal, 96(1137), pp.403-407.

Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K. and Hilgenfeld, R., 2020. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science, 368(6489), pp.409-412.

Zhavoronkov, A., Aladinskiy, V., Zhebrak, A., Zagribelnyy, B., Terentiev, V., Bezrukov, D.S., Polykovskiy, D., Shayakhmetov, R., Filimonov, A., Orekhov, P. and Yan, Y., 2020. Potential 2019-nCoV 3C-like protease inhibitors designed using generative deep learning approaches. ChemRxiv. Preprint. https://doi. org/10.26434/chemrxiv, 1182(9102), p.v1.

Zhu, W., Xu, M., Chen, C.Z., Guo, H., Shen, M., Hu, X., Shinn, P., Klumpp-Thomas, C., Michael, S.G. and Zheng, W., 2020. Identification of SARS-CoV-2 3CL protease inhibitors by a quantitative high-throughput screening. ACS pharmacology & translational science, 3(5), pp.1008-1016.

Appendix I

1. Molecular structures of filtered actives

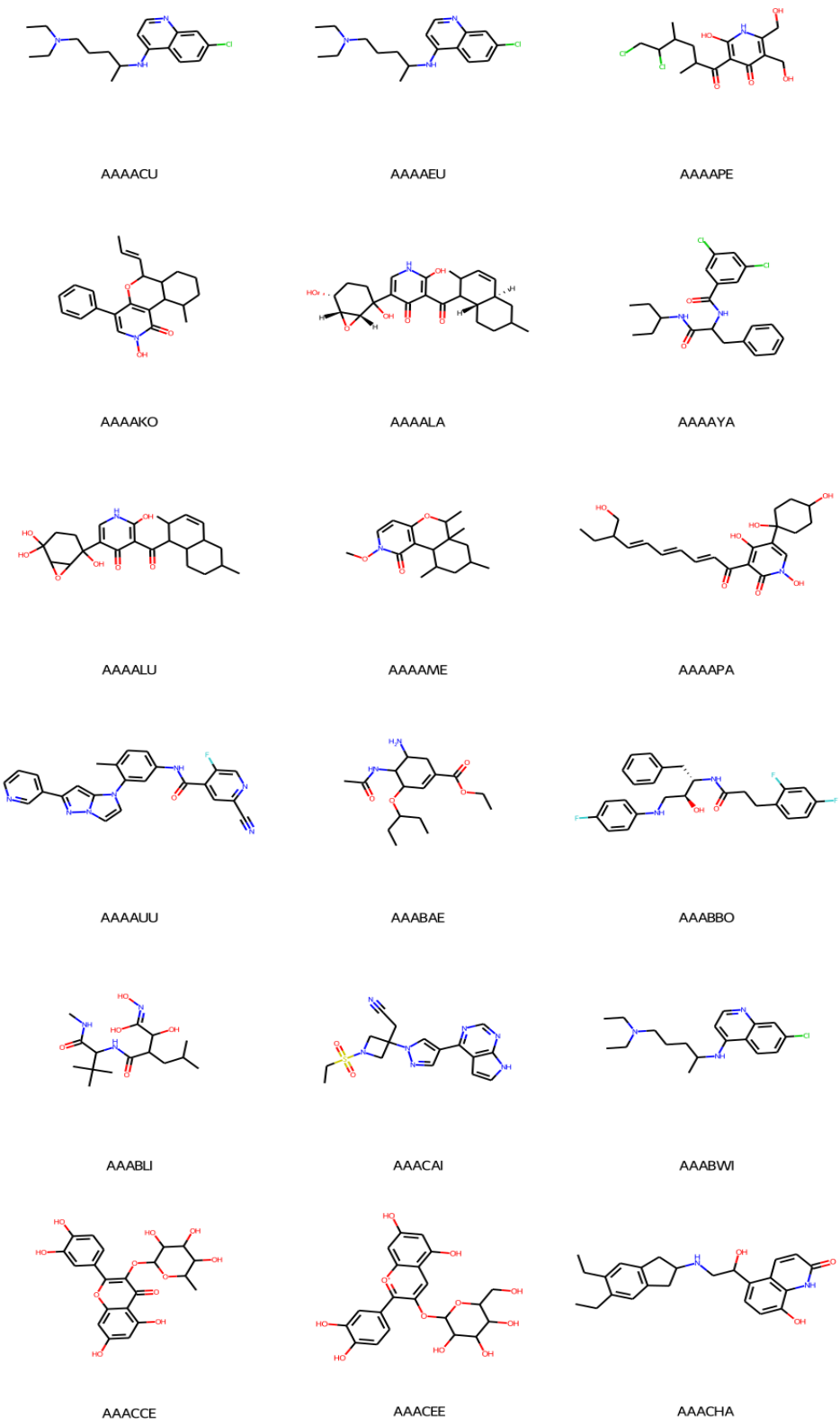


Figure A.1. Molecular structure of actives filtered using the Ghose filter

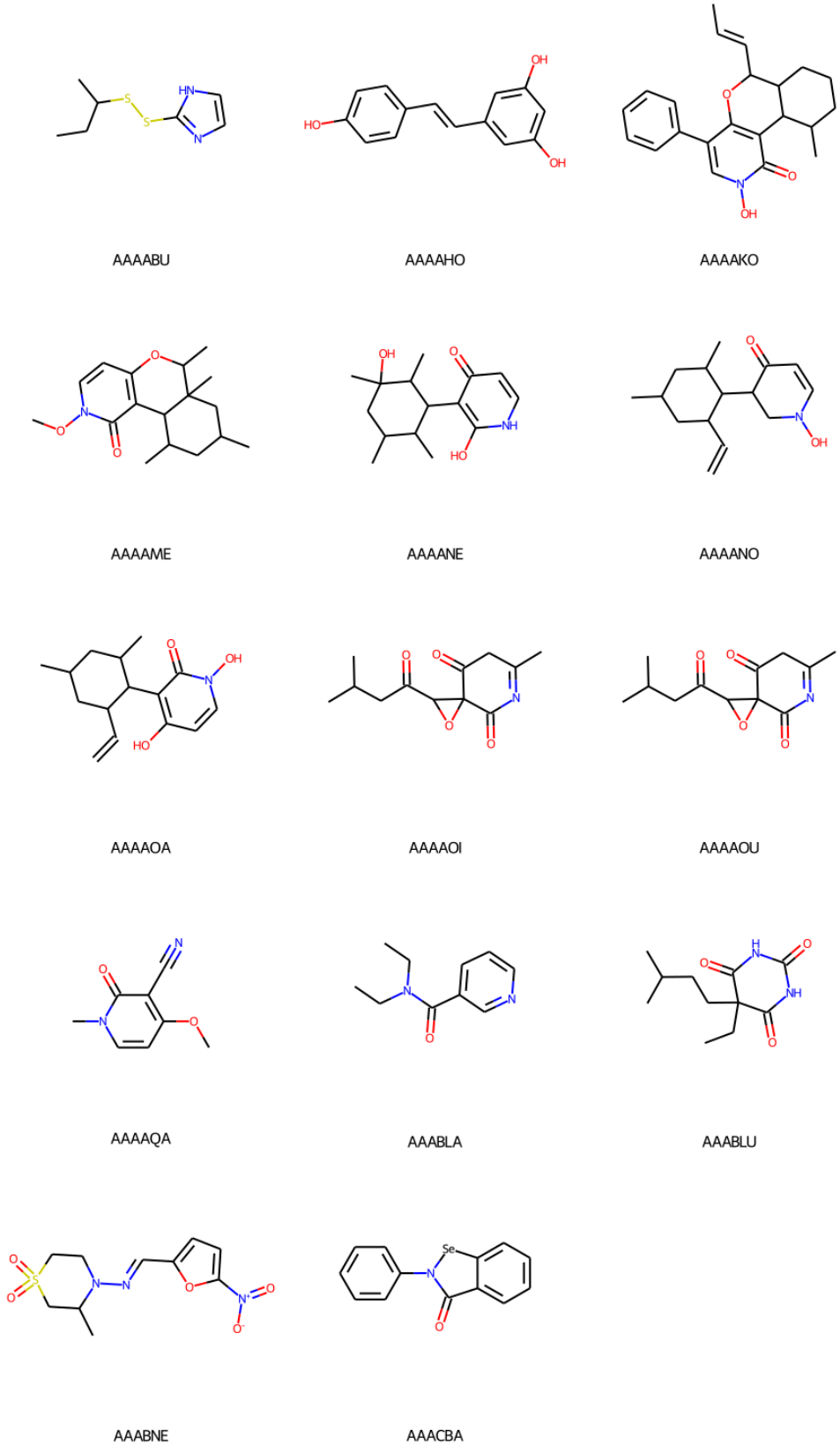


Figure A.2. Molecular structures of actives filtered using the drug-like filter

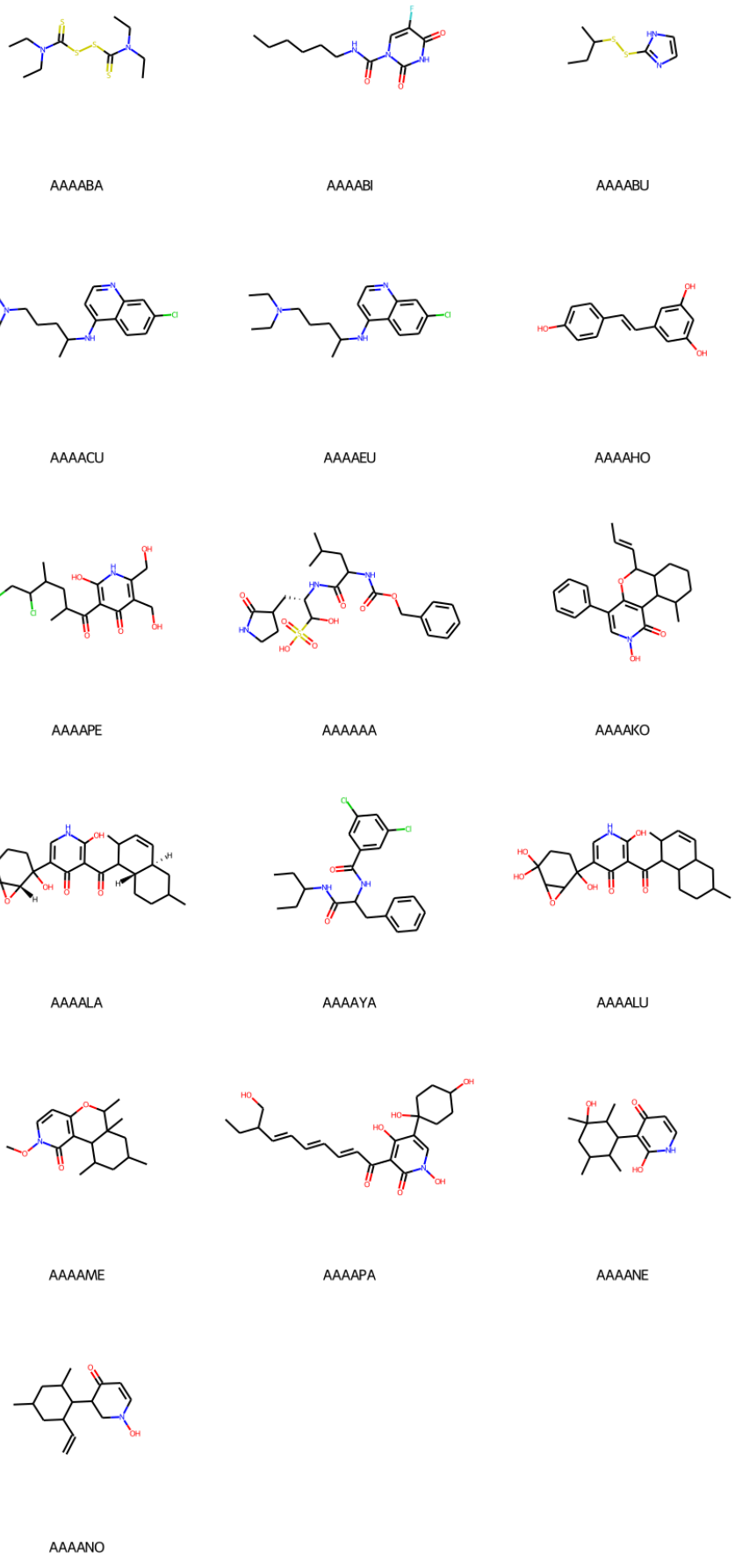


Figure A.3. Molecular structures of actives filtered using Lipinski filter

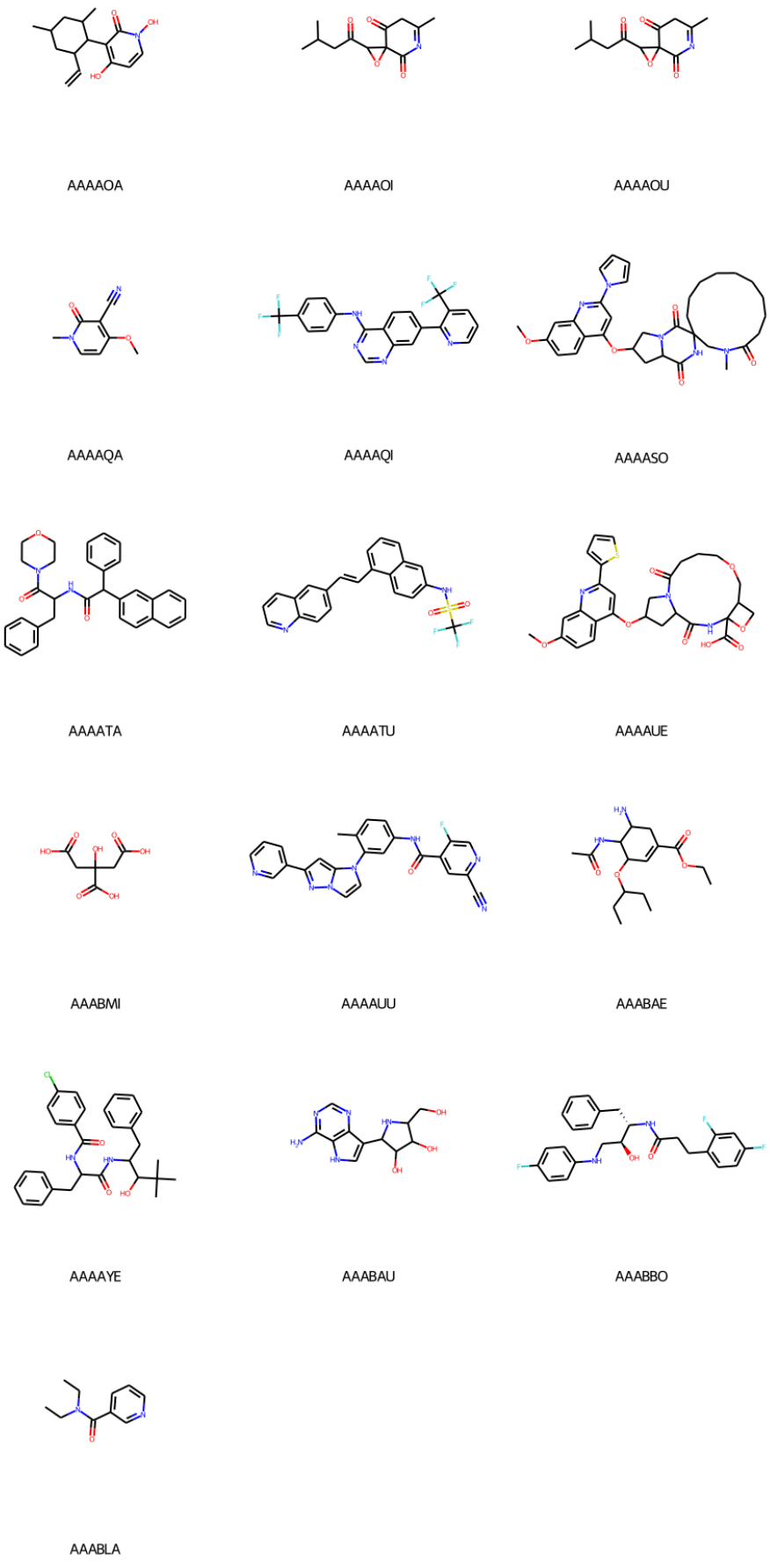
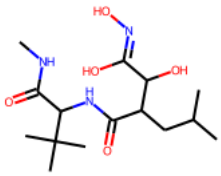
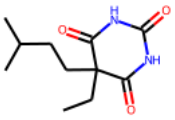


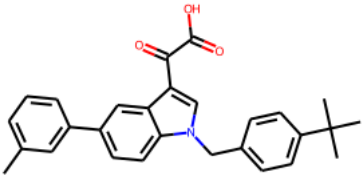
Figure A.4. Molecular structures of actives filtered using Lipinski filter (continued)



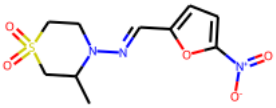
AAABLI



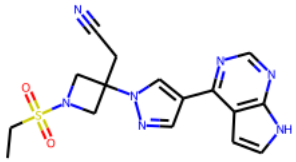
AAABLU



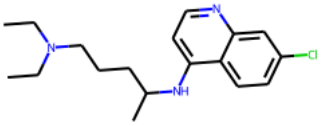
AAABMU



AAABNE

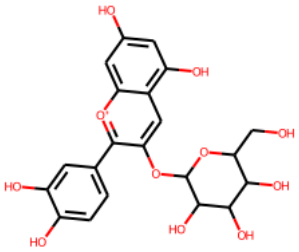


AAACAI

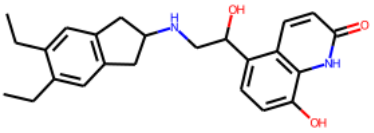


AAABWI

AAACBA



AAACEE



AAACHA

Figure A.5. Molecular structures of actives filtered using Lipinski filter (continued)

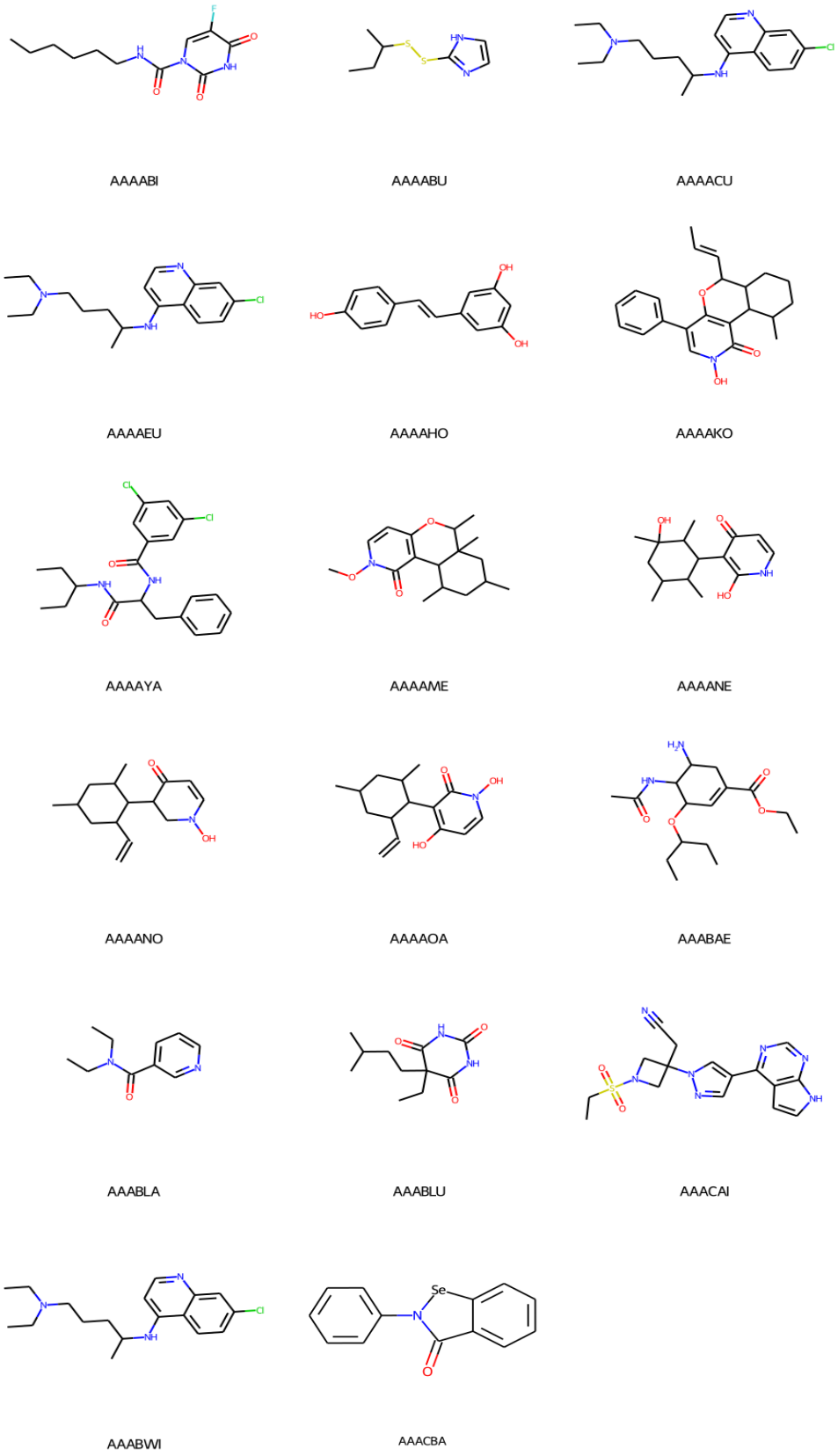


Figure A.6. Molecular structures of actives using the QED 0.6 cut-off filter

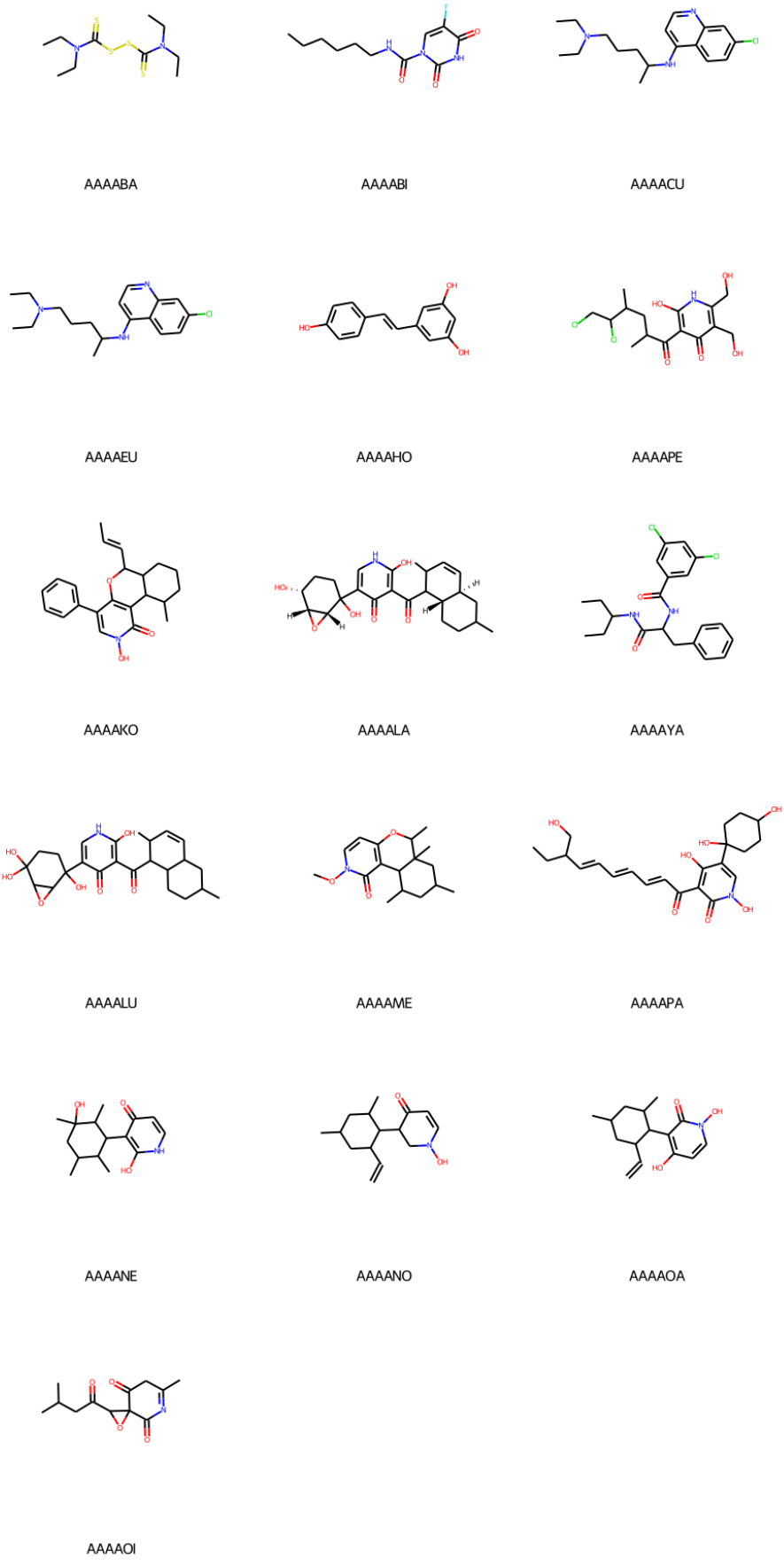
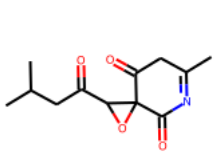
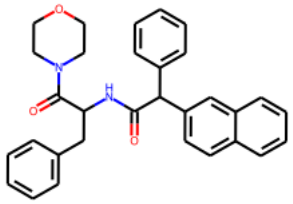


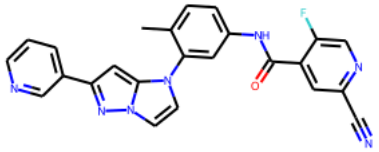
Figure A.7. Molecular structures of actives filtered using the REOS filter



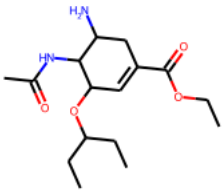
AAAAOU



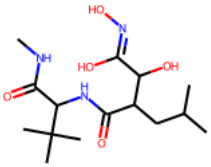
AAAATA



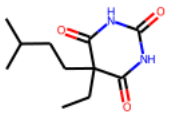
AAAAUU



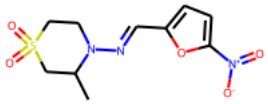
AAABAE



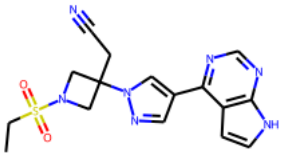
AAABLI



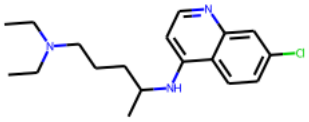
AAABLU



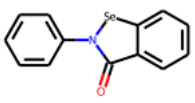
AAABNE



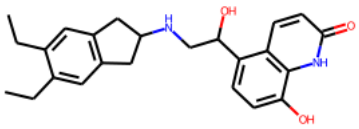
AAACAI



AAABWI



AAACBA



AAACHA

Figure A.8. Molecular structures of actives filtered using the REOS filter (continued)

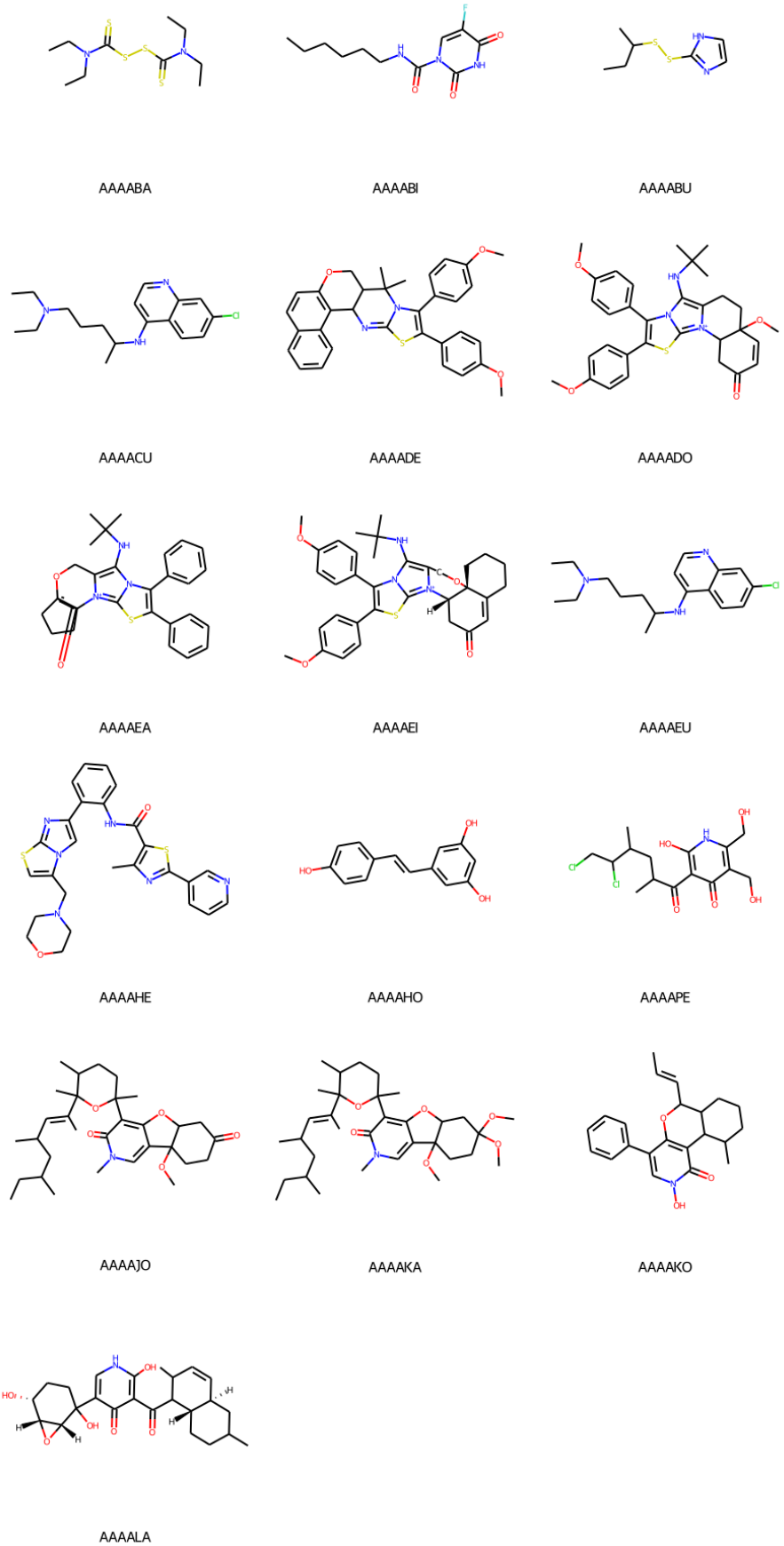


Figure A.9. Molecular structure of actives filtered using the Veber filter

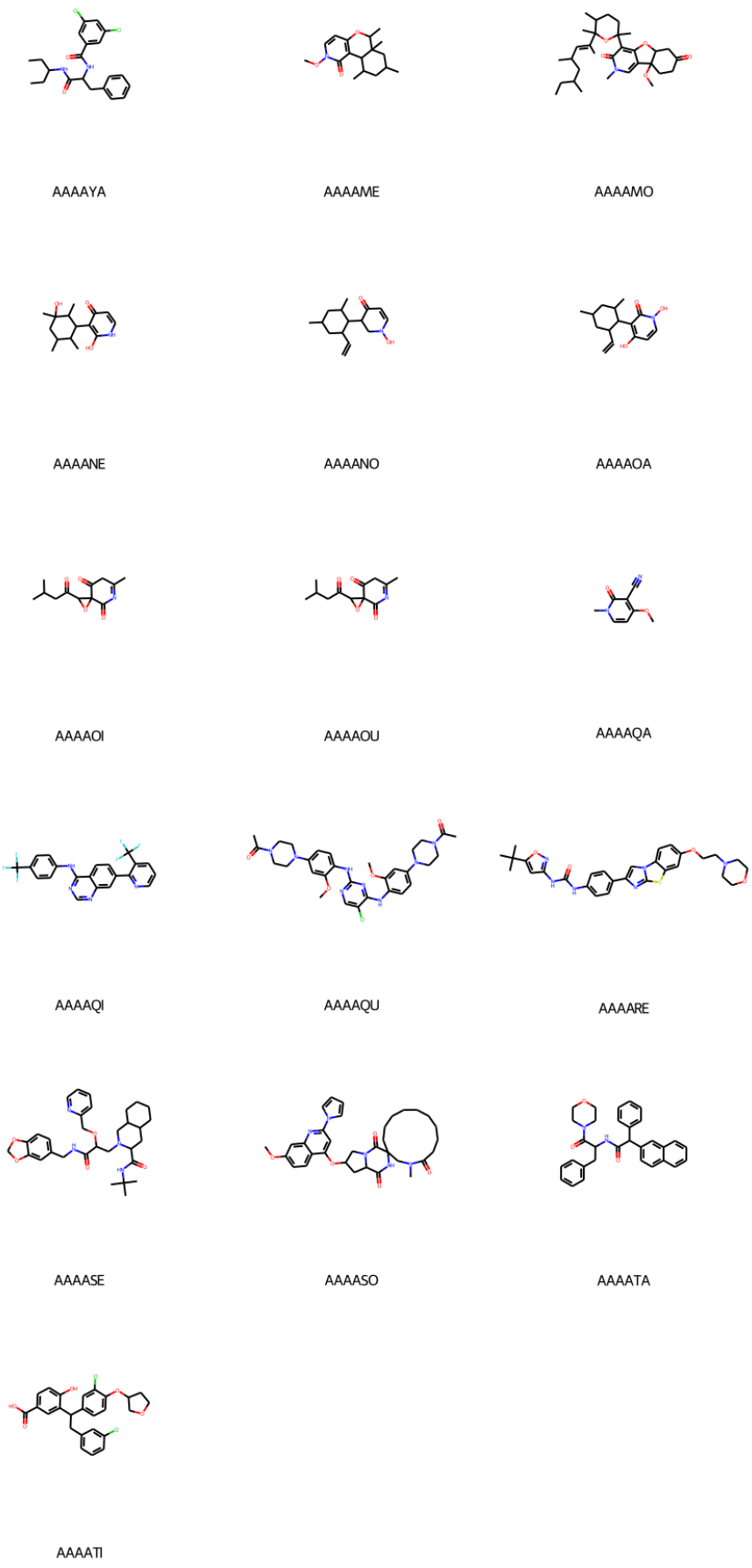


Figure A.10. Molecular structures of actives filtered using the Veber filter (continued)

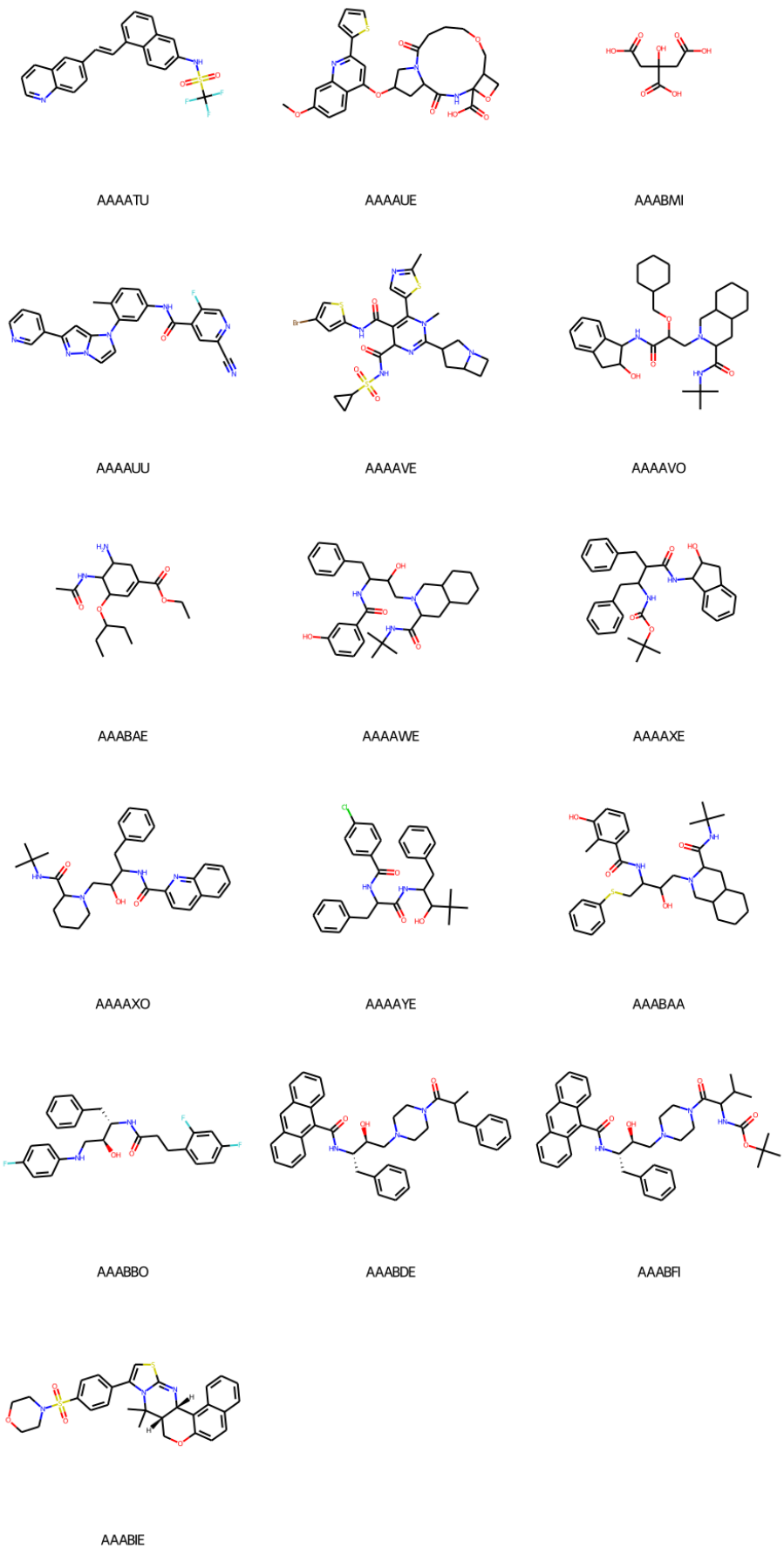


Figure A.11. Molecular structures of actives filtered using the Veber filter (continued)

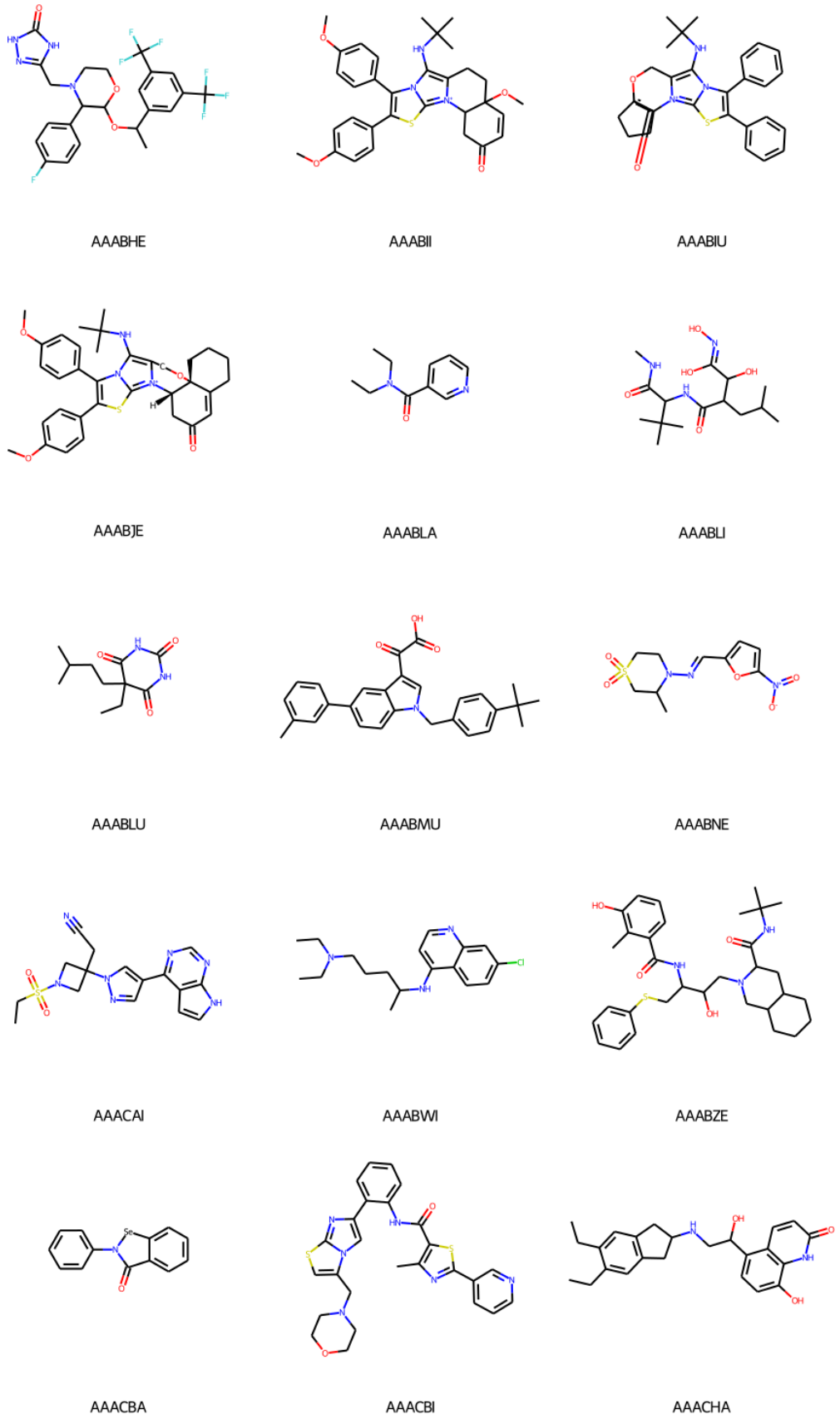


Figure A.12. Molecular structures of actives filtered using the Veber filter (continued)

2. SMILES formula of filtered hits having a Tanimoto coefficient cut-off higher than 0.4

Table A.1 Hits with a Tanimoto coefficient higher than 0.4, obtained from the small molecule list filtered using the drug-like *filter*

SMILES	Molecule ID
COc1cc(C=Cc2ccc(O)cc2)cc(OC)c1	AACUGO
O=c1cc(O)cc(/C=C/c2ccccc2)o1	DAQOZU
C=C[C@@H](C=Cc1ccc(O)cc1)c1ccc(O)cc1	BUKUUU
COc1cc(C(C)=O)ccc1Oc1c(C#N)c(=O)n(C)c(=O)n1C	BITVFO
CCN(CC)C(=O)c1cncc(C(N)=O)c1	AABQNE
CN(C(=O)c1cccnc1)c1nc2ccccc2o1	AOBGSU
CCN(CC)C(=O)c1ccncc1B1OC(C)(C)C(C)(C)O1	AAIAJO
O=C1OCCN1N=Cc1ccc([N+](=O)[O-])o1	AAOPOA

Table A.2 Hits with a Tanimoto coefficient higher than 0.4, obtained from the small molecule list filtered using the Lipinski filter

SMILES	Molecule ID
CCN(CC)C(=S)S[Sn](c1ccccc1)(c1ccccc1)c1ccccc1	AISOOA
CC(C)(C)N(CCNC(=O)n1cc(F)c(=O)[nH]c1=O)C(=O)O	IIRONO
CCCCCCCCSCCNC(=O)CCn1cc(F)c(=O)[nH]c1=O	EAFIQO
CCCCCCNC(=O)n1cc(Cl)c(=O)n(C(=O)OCC(C)C)c1=O	FABKIU
CN(C)CCCNc1ccnc2cc(Cl)ccc12	AAFUBO
Oc1cc(O)cc(C=Cc2ccc(O)c(O)c2)c1	AAIWRI
OC[C@H]1O[C@@H](Cc2cc(O)cc(C=Cc3ccc(O)cc3)c2)[C@H](O)[C@@H](O)[C@@H]1O	BADVLU
CC(C)NC(=O)c1cc(Cl)cc(Cl)c1	AAENMO
CC(C=O)NC(=O)C(Cc1ccc[nH]1)NC(=O)c1ccccc1	FOOUTU
O=C(Nc1nncs1)C(Cc1ccccc1)NC(=O)N1CCOCC1	EELHGO
CC(N)C(=O)NC(Cc1ccccc1)C(=O)N1CCCC1C(=O)Nc1ccc2ccccc2c1	GEERHU
O=C(O)CC(O)(CC(=O)O)CC(=O)O	AEVSIE
CCOC(=O)C1=CC(OC(CC)CC)[C@@H]2N[C@@H]2C1	AOQEZE
CCC(CC)OC1=CC(C(=O)O)C[C@H](N)[C@H]1NC(C)=O	FIULEI
Nc1ncnc2c(-c3[nH]c(CO)c(O)c3O)c[nH]c12	FEGYWI
CCN(CC)C(=O)c1cncc(C(N)=O)c1	AABQNE
CCN(CC)CCN(CC(=O)N(C1CCCCC1)C1CCCCC1)C(=O)c1cccnc1	BOJEME
CCN(CC)C(=O)c1ccncc1B1OC(C)(C)C(C)(C)O1	AAIAJO
CCN(CC)C(=O)c1ccc(N(C)S(=O)(=O)c2ccccc2)cc1	AAAJWE
O=C(NCCO)NC(=S)c1cccnc1	JUDSCI
CC(C)CC(C(=O)NC(C(=O)OC1CCCC1)c1ccccc1)C(O)C(O)=NO	AUDQFU
CCCC1(C[C@H](C)CC)C(=O)NC(=O)NC1=O	AAMTFE
CC[C@]1(CCC(C)C)C(=O)NC(=O)N(C(=O)c2ccccc2I)C1=O	CAPEUA
O=C1OCCN1N=Cc1ccc([N+](=O)[O-])o1	AAOPOA
N#CC[C](C1CCCC1)n1cc(-c2ncnc3[nH]ccc23)cn1	AUEZQE
Oc1cc(O)c2cc(O)c(-c3ccc(O)c(O)c3)[o+]c2c1	EUSOWU
C[C@@H]1O[C@@H](Oc2cc3c(O)cc(O)cc3cc2-c2cc(O)c(O)c(O)c2)[C@H](O)[C@H](O)[C@H]1O	BUQHJA
COc1ccc(C[C@@H](C)NC[C@H](O)c2ccc(O)c3[nH]c(=O)ccc23)cc1	AADZUU

Table A.3 Hits with a Tanimoto coefficient higher than 0.4, obtained from the small molecule list filtered using the Veber filter

SMILES	Molecule ID
CCN(CC)C(=S)S[Sn](c1ccccc1)(c1ccccc1)c1ccccc1	AISOOA
CC(C)(C)N(CCNC(=O)n1cc(F)c(=O)[nH]c1=O)C(=O)O	IIRONO
CCCCCCNC(=O)n1cc(Cl)c(=O)n(C(=O)OCC(C)C)c1=O	FABKIU
CN(C)CCCNc1ccnc2cc(Cl)ccc12	AAFUBO
COc1cc(F)cc(C(=O)Nc2ccccc2-c2cn3c(CN4CCNCC4)csc3n2)c1	LARPRE
Oc1cc(O)cc(C=Cc2ccc(O)c(O)c2)c1	AAIWRI
OC[C@H]1O[C@@H](Cc2cc(O)cc(C=Cc3ccc(O)cc3)c2)[C@H](O)[C@@H](O)[C@@H]1O	BADVLU
CC[C@H](C)C[C@H](C)/C=C(\C)[C@@H]1O[C@H](c2c3c(cn(C)c2=O)[C@]2(O)CCC(=O)C[C@@H]2O3)CC[C@@H]1C	AONUIA
CC(C)NC(=O)c1cc(Cl)cc(Cl)c1	AAENMO
CC(=O)CNC(=O)C(Cc1ccc[nH]1)NC(=O)c1ccccc1	COUEQU
COc1cc(N2CCC(N)CC2)ccc1Nc1ncc(Cl)c(-c2c[nH]c3ccccc23)n1	FUAVOI
CC(C)(C)NC(=O)C1C[C@@H]2CCCC[C@@H]2CN1C[C@@H](O)[C@@H](N)Cc1ccccc1	DEFEHA
O=C(Nc1nncs1)C(Cc1ccccc1)NC(=O)N1CCOCC1	EELHGO
CC(N)C(=O)NC(Cc1ccccc1)C(=O)N1CCCC1C(=O)Nc1ccc2ccccc2c1	GEERHU
O=C(c1ccc2ccccc2c1)C([C@H](c1ccccc1)N1CCOCC1)N1CCOCC1	BOBDUI
O=C(O)CC(O)(CC(=O)O)CC(=O)O	AEVSIE
C[N+](C)(C)CCOC(=O)C(O)(CC(=O)O)CC(=O)O	EAOMMI
O=C(N[C@@H]1c2ccccc2C[C@@H]1O)C1COCCN1CC1CCC1	MOMMQU
CCOC(=O)C1=CC(OC(CC)CC)[C@@H]2N[C@@H]2C1	AOQEZE
CCC(CC)OC1=CC(C(=O)O)C[C@H](N)[C@H]1NC(C)=O	FIULEI
CCN(CC)C(=O)c1cncc(C(N)=O)c1	AABQNE
CCN(CC)CCN(CC(=O)N(C1CCCCC1)C1CCCCC1)C(=O)c1cccnc1	BOJEME
CCN(CC)C(=O)c1ccncc1B1OC(C)(C)C(C)(C)O1	AAIAJO
CCN(CC)C(=O)c1ccc(N(C)S(=O)(=O)c2ccccc2)cc1	AAAJWE
CC(C)CC(C(=O)NC(C(=O)OC1CCCC1)c1ccccc1)C(O)C(O)=NO	AUDQFU
CCCC1(C[C@H](C)CC)C(=O)NC(=O)NC1=O	AAMTFE
CC[C@]1(CCC(C)C)C(=O)NC(=O)N(C(=O)c2ccccc2I)C1=O	CAPEUA
O=C1OCCN1N=Cc1ccc([N+](=O)[O-])o1	AAOPOA
N#CC[C](C1CCCC1)n1cc(-c2ncnc3[nH]ccc23)cn1	AUEZQE
COc1ccc(C[C@@H](C)NC[C@H](O)c2ccc(O)c3[nH]c(=O)ccc23)cc1	AADZUU
CC(C)C(=NC(=O)C(Cc1ccc(Cl)cc1)NC(=O)c1ccc(Br)cc1)C(=O)O	AUIRFU

3. SMILES Formula of Filtered Actives

Table A.4 Actives filtered using the drug-like filter

SMILES	Molecule ID
CCC(C)SSc1ncc[nH]1	AAAABU
Oc1ccc(C=Cc2cc(O)cc(O)c2)cc1	AAAAHO
CC=CC1Oc2c(-c3ccccc3)cn(O)c(=O)c2C2C(C)CCCC12	AAAAKO
COn1ccc2c(c1=O)C1C(C)CC(C)CC1(C)C(C)O2	AAAAME
CC1CC(C)(O)C(C)C(c2c(O)[nH]ccc2=O)C1C	AAAANE
C=CC1CC(C)CC(C)C1C1CN(O)C=CC1=O	AAAANO
C=CC1CC(C)CC(C)C1c1c(O)ccn(O)c1=O	AAAAOA
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOI
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOU
COc1ccn(C)c(=O)c1C#N	AAAAQA
CCN(CC)C(=O)c1cccnc1	AAABLA
CCC1(CCC(C)C)C(=O)NC(=O)NC1=O	AAABLU
CC1CS(=O)(=O)CCN1/N=C/c1ccc([N+](=O)[O-])o1	AAABNE
O=c1c2ccccc2[se]n1-c1ccccc1	AAACBA

Table A.5 Actives filtered using the Ghose filter

SMILES	Molecule ID
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAACU
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAAEU
CC(CC(C)C(Cl)CCl)C(=O)c1c(O)[nH]c(CO)c(CO)c1=O	AAAAPE
CC=CC1Oc2c(-c3ccccc3)cn(O)c(=O)c2C2C(C)CCCC12	AAAAKO
CC1CC[C@@H]2C(C(=O)c3c(O)[nH]cc(C4(O)CC[C@@H](O)[C@H]5O[C@H]54)c3=O)C(C)C=C[C@H]2C1	AAAALA
CCC(CC)NC(=O)C(Cc1ccccc1)NC(=O)c1cc(Cl)cc(Cl)c1	AAAAYA
CC1CCC2C(C=CC(C)C2C(=O)c2c(O)[nH]cc(C3(O)CCC(O)(O)C4OC43)c2=O)C1	AAAALU
COn1ccc2c(c1=O)C1C(C)CC(C)CC1(C)C(C)O2	AAAAME
CCC(/C=C/C=C/C=C/C(=O)c1c(O)c(C2(O)CCC(O)CC2)cn(O)c1=O)CO	AAAAPA
Cc1ccc(NC(=O)c2cc(C#N)ncc2F)cc1-n1ccn2nc(-c3cccnc3)cc12	AAAAUU
CCOC(=O)C1=CC(OC(CC)CC)C(NC(C)=O)C(N)C1	AAABAE
O=C(CCc1ccc(F)cc1F)N[C@@H](Cc1ccccc1)[C@@H](O)CNc1ccc(F)cc1	AAABBO
CNC(=O)C(NC(=O)C(CC(C)C)C(O)C(O)=NO)C(C)(C)C	AAABLI
CCS(=O)(=O)N1CC(CC#N)(n2cc(-c3ncnc4[nH]ccc34)cn2)C1	AAACAI
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAABWI
CC1OC(Oc2c(-c3ccc(O)c(O)c3)oc3cc(O)cc(O)c3c2=O)C(O)C(O)C1O	AAACCE
OCC1OC(Oc2cc3c(O)cc(O)cc3[o+]c2-c2ccc(O)c(O)c2)C(O)C(O)C1O	AAACEE
CCc1cc2c(cc1CC)CC(NCC(O)c1ccc(O)c3[nH]c(=O)ccc13)C2	AAACHA

Table A.6 Actives filtered using the Lipinski filter

SMILES	Molecule ID
CCN(CC)C(=S)SSC(=S)N(CC)CC	AAAABA
CCCCCCNC(=O)n1cc(F)c(=O)[nH]c1=O	AAAABI
CCC(C)SSc1ncc[nH]1	AAAABU
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAACU
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAAEU
Oc1ccc(C=Cc2cc(O)cc(O)c2)cc1	AAAAHO
CC(CC(C)C(Cl)CCl)C(=O)c1c(O)[nH]c(CO)c(CO)c1=O	AAAAPE
CC(C)CC(NC(=O)OCc1ccccc1)C(=O)N[C@@H](CC1CCNC1=O)C(O)S(=O)(=O)O	AAAAAA
CC=CC1Oc2c(-c3ccccc3)cn(O)c(=O)c2C2C(C)CCCC12	AAAAKO
CC1CC[C@@H]2C(C(=O)c3c(O)[nH]cc(C4(O)CC[C@@H](O)[C@H]5O[C@H]54)c3=O)C(C)C=C[C@H]2C1	AAAALA
CCC(CC)NC(=O)C(Cc1ccccc1)NC(=O)c1cc(Cl)cc(Cl)c1	AAAAYA
CC1CCC2C(C=CC(C)C2C(=O)c2c(O)[nH]cc(C3(O)CCC(O)(O)C4OC43)c2=O)C1	AAAALU
COn1ccc2c(c1=O)C1C(C)CC(C)CC1(C)C(C)O2	AAAAME
CCC(/C=C/C=C/C=C/C(=O)c1c(O)c(C2(O)CCC(O)CC2)cn(O)c1=O)CO	AAAAPA
CC1CC(C)(O)C(C)C(c2c(O)[nH]ccc2=O)C1C	AAAANE
C=CC1CC(C)CC(C)C1C1CN(O)C=CC1=O	AAAANO
C=CC1CC(C)CC(C)C1c1c(O)ccn(O)c1=O	AAAAOA
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOI
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOU
COc1ccn(C)c(=O)c1C#N	AAAAQA
FC(F)(F)c1ccc(Nc2ncnc3cc(-c4ncccc4C(F)(F)F)ccc23)cc1	AAAAQI
COc1ccc2c(OC3CC4C(=O)NC5(CCCCCCCCCC(=O)N(C)C5)C(=O)N4C3)cc(-n3cccc3)nc2c1	AAAASO
O=C(NC(Cc1ccccc1)C(=O)N1CCOCC1)C(c1ccccc1)c1ccc2ccccc2c1	AAAATA
O=S(=O)(Nc1ccc2c(C=Cc3ccc4ncccc4c3)cccc2c1)C(F)(F)F	AAAATU
COc1ccc2c(OC3CC4C(=O)NC5(C(=O)O)OCC5COCCCC(=O)N4C3)cc(-c3cccs3)nc2c1	AAAAUE
O=C(O)CC(O)(CC(=O)O)C(=O)O	AAABMI
Cc1ccc(NC(=O)c2cc(C#N)ncc2F)cc1-n1ccn2nc(-c3cccnc3)cc12	AAAAUU
CCOC(=O)C1=CC(OC(CC)CC)C(NC(C)=O)C(N)C1	AAABAE
CC(C)(C)C(O)C(Cc1ccccc1)NC(=O)C(Cc1ccccc1)NC(=O)c1ccc(Cl)cc1	AAAAYE
Nc1ncnc2c(C3NC(CO)C(O)C3O)c[nH]c12	AAABAU
O=C(CCc1ccc(F)cc1F)N[C@@H](Cc1ccccc1)[C@@H](O)CNc1ccc(F)cc1	AAABBO
CCN(CC)C(=O)c1cccnc1	AAABLA
CNC(=O)C(NC(=O)C(CC(C)C)C(O)C(O)=NO)C(C)(C)C	AAABLI
CCC1(CCC(C)C)C(=O)NC(=O)NC1=O	AAABLU
Cc1cccc(-c2ccc3c(c2)c(C(=O)C(=O)O)cn3Cc2ccc(C(C)(C)C)cc2)c1	AAABMU
CC1CS(=O)(=O)CCN1/N=C/c1ccc([N+](=O)[O-])o1	AAABNE
CCS(=O)(=O)N1CC(CC#N)(n2cc(-c3ncnc4[nH]ccc34)cn2)C1	AAACAI
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAABWI
O=c1c2ccccc2[se]n1-c1ccccc1	AAACBA
OCC1OC(Oc2cc3c(O)cc(O)cc3[o+]c2-c2ccc(O)c(O)c2)C(O)C(O)C1O	AAACEE
CCc1cc2c(cc1CC)CC(NCC(O)c1ccc(O)c3[nH]c(=O)ccc13)C2	AAACHA

Table A.7 Actives filtered using a QED filter with a cut-off of 0.6

SMILES	Molecule ID
CCCCCCNC(=O)n1cc(F)c(=O)[nH]c1=O	AAAABI
CCC(C)SSc1ncc[nH]1	AAAABU
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAACU
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAAEU
Oc1ccc(C=Cc2cc(O)cc(O)c2)cc1	AAAAHO
CC=CC1Oc2c(-c3ccccc3)cn(O)c(=O)c2C2C(C)CCCC12	AAAAKO
CCC(CC)NC(=O)C(Cc1ccccc1)NC(=O)c1cc(Cl)cc(Cl)c1	AAAAYA
COn1ccc2c(c1=O)C1C(C)CC(C)CC1(C)C(C)O2	AAAAME
CC1CC(C)(O)C(C)C(c2c(O)[nH]ccc2=O)C1C	AAAANE
C=CC1CC(C)CC(C)C1C1CN(O)C=CC1=O	AAAANO
C=CC1CC(C)CC(C)C1c1c(O)ccn(O)c1=O	AAAAOA
CCOC(=O)C1=CC(OC(CC)CC)C(NC(C)=O)C(N)C1	AAABAE
CCN(CC)C(=O)c1cccnc1	AAABLA
CCC1(CCC(C)C)C(=O)NC(=O)NC1=O	AAABLU
CCS(=O)(=O)N1CC(CC#N)(n2cc(-c3ncnc4[nH]ccc34)cn2)C1	AAACAI
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAABWI
O=c1c2ccccc2[se]n1-c1ccccc1	AAACBA

Table A.8 Actives filtered using the REOS filter

SMILES	Molecule ID
CCN(CC)C(=S)SSC(=S)N(CC)CC	AAAABA
CCCCCCNC(=O)n1cc(F)c(=O)[nH]c1=O	AAAABI
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAACU
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAAEU
Oc1ccc(C=Cc2cc(O)cc(O)c2)cc1	AAAAHO
CC(CC(C)C(Cl)CCl)C(=O)c1c(O)[nH]c(CO)c(CO)c1=O	AAAAPE
CC=CC1Oc2c(-c3ccccc3)cn(O)c(=O)c2C2C(C)CCCC12	AAAAKO
CC1CC[C@@H]2C(C(=O)c3c(O)[nH]cc(C4(O)CC[C@@H](O)[C@H]5O[C@H]54)c3=O)C(C)C=C[C@H]2C1	AAAALA
CCC(CC)NC(=O)C(Cc1ccccc1)NC(=O)c1cc(Cl)cc(Cl)c1	AAAAYA
CC1CCC2C(C=CC(C)C2C(=O)c2c(O)[nH]cc(C3(O)CCC(O)(O)C4OC43)c2=O)C1	AAAALU
COn1ccc2c(c1=O)C1C(C)CC(C)CC1(C)C(C)O2	AAAAME
CCC(/C=C/C=C/C=C/C(=O)c1c(O)c(C2(O)CCC(O)CC2)cn(O)c1=O)CO	AAAAPA
CC1CC(C)(O)C(C)C(c2c(O)[nH]ccc2=O)C1C	AAAANE
C=CC1CC(C)CC(C)C1C1CN(O)C=CC1=O	AAAANO
C=CC1CC(C)CC(C)C1c1c(O)ccn(O)c1=O	AAAAOA
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOI
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOU
O=C(NC(Cc1ccccc1)C(=O)N1CCOCC1)C(c1ccccc1)c1ccc2ccccc2c1	AAAATA
Cc1ccc(NC(=O)c2cc(C#N)ncc2F)cc1-n1ccn2nc(-c3cccnc3)cc12	AAAAUU
CCOC(=O)C1=CC(OC(CC)CC)C(NC(C)=O)C(N)C1	AAABAE
CNC(=O)C(NC(=O)C(CC(C)C)C(O)C(O)=NO)C(C)(C)C	AAABLI
CCC1(CCC(C)C)C(=O)NC(=O)NC1=O	AAABLU
CC1CS(=O)(=O)CCN1/N=C/c1ccc([N+](=O)[O-])o1	AAABNE
CCS(=O)(=O)N1CC(CC#N)(n2cc(-c3ncnc4[nH]ccc34)cn2)C1	AAACAI
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAABWI
O=c1c2ccccc2[se]n1-c1ccccc1	AAACBA
CCc1cc2c(cc1CC)CC(NCC(O)c1ccc(O)c3[nH]c(=O)ccc13)C2	AAACHA

Table A.9 Actives filtered using the Veber filter

SMILES	Molecule ID
CCN(CC)C(=S)SSC(=S)N(CC)CC	AAAABA
CCCCCCNC(=O)n1cc(F)c(=O)[nH]c1=O	AAAABI
CCC(C)SSc1ncc[nH]1	AAAABU
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAACU
COc1ccc(C2=C(c3ccc(OC)cc3)N3C(=NC4c5c(ccc6ccccc56)OCC4C3(C)C)S2)cc1	AAAADE
COc1ccc(-c2sc3n(c(NC(C)(C)C)c4[n+]3C3CC(=O)C=CC3(OC)CC4)c2-c2ccc(OC)cc2)cc1	AAAADO
CC(C)(C)Nc1c2[n+](c3sc(-c4ccccc4)c(-c4ccccc4)n13)C13CCCC1(C=CC(=O)C3)OC2	AAAAEA
COc1ccc(-c2sc3n(c(NC(C)(C)C)c4[n+]3[C@H]3CC(=O)C=C5CCCC[C@@]53OC4)c2-c2ccc(OC)cc2)cc1	AAAAEI
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAAAEU
Cc1nc(-c2cccnc2)sc1C(=O)Nc1ccccc1-c1cn2c(CN3CCOCC3)csc2n1	AAAAHE
Oc1ccc(C=Cc2cc(O)cc(O)c2)cc1	AAAAHO
CC(CC(C)C(Cl)CCl)C(=O)c1c(O)[nH]c(CO)c(CO)c1=O	AAAAPE
CCC(C)CC(C)/C=C(\C)C1(C)OC(C)(c2c3c(cn(C)c2=O)C2(OC)CCC(=O)CC2O3)CCC1C	AAAAJO
CCC(C)CC(C)/C=C(\C)C1(C)OC(C)(c2c3c(cn(C)c2=O)C2(OC)CCC(OC)(OC)CC2O3)CCC1C	AAAAKA
CC=CC1Oc2c(-c3ccccc3)cn(O)c(=O)c2C2C(C)CCCC12	AAAAKO
CC1CC[C@@H]2C(C(=O)c3c(O)[nH]cc(C4(O)CC[C@@H](O)[C@H]5O[C@H]54)c3=O)C(C)C=C[C@H]2C1	AAAALA
CCC(CC)NC(=O)C(Cc1ccccc1)NC(=O)c1cc(Cl)cc(Cl)c1	AAAAYA
COn1ccc2c(c1=O)C1C(C)CC(C)CC1(C)C(C)O2	AAAAME
CCC(C)CC(C)/C=C(\C)C1(C)OC(C)(c2c3c(cn(C)c2=O)C2(OC)CCC(=O)CC2O3)CCC1C	AAAAMO
CC1CC(C)(O)C(C)C(c2c(O)[nH]ccc2=O)C1C	AAAANE
C=CC1CC(C)CC(C)C1C1CN(O)C=CC1=O	AAAANO
C=CC1CC(C)CC(C)C1c1c(O)ccn(O)c1=O	AAAAOA
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOI
CC1=NC(=O)C2(OC2C(=O)CC(C)C)C(=O)C1	AAAAOU
COc1ccn(C)c(=O)c1C#N	AAAAQA
FC(F)(F)c1ccc(Nc2ncnc3cc(-c4ncccc4C(F)(F)F)ccc23)cc1	AAAAQI
COc1cc(N2CCN(C(C)=O)CC2)ccc1Nc1ncc(Cl)c(Nc2ccc(N3CCN(C(C)=O)CC3)cc2OC)n1	AAAAQU
CC(C)(C)c1cc(NC(=O)Nc2ccc(-c3cn4c(n3)sc3cc(OCCN5CCOCC5)ccc34)cc2)no1	AAAARE
CC(C)(C)NC(=O)C1CC2CCCCC2CN1CC(OCc1ccccn1)C(=O)NCc1ccc2c(c1)OCO2	AAAASE
COc1ccc2c(OC3CC4C(=O)NC5(CCCCCCCCCC(=O)N(C)C5)C(=O)N4C3)cc(-n3cccc3)nc2c1	AAAASO
O=C(NC(Cc1ccccc1)C(=O)N1CCOCC1)C(c1ccccc1)c1ccc2ccccc2c1	AAAATA
O=C(O)c1ccc(O)c(C(Cc2cccc(Cl)c2)c2ccc(OC3CCOC3)c(Cl)c2)c1	AAAATI
O=S(=O)(Nc1ccc2c(C=Cc3ccc4ncccc4c3)cccc2c1)C(F)(F)F	AAAATU
COc1ccc2c(OC3CC4C(=O)NC5(C(=O)O)OCC5COCCCC(=O)N4C3)cc(-c3cccs3)nc2c1	AAAAUE
O=C(O)CC(O)(CC(=O)O)C(=O)O	AAABMI
Cc1ccc(NC(=O)c2cc(C#N)ncc2F)cc1-n1ccn2nc(-c3cccnc3)cc12	AAAAUU
Cc1ncc(C2=C(C(=O)Nc3cc(Br)cs3)C(C(=O)NS(=O)(=O)C3CC3)N=C(C3CC4CCN4C3)N2C)s1	AAAAVE
CC(C)(C)NC(=O)C1CC2CCCCC2CN1CC(OCC1CCCCC1)C(=O)NC1c2ccccc2CC1O	AAAAVO
CCOC(=O)C1=CC(OC(CC)CC)C(NC(C)=O)C(N)C1	AAABAE
CC(C)(C)NC(=O)C1CC2CCCCC2CN1CC(O)C(Cc1ccccc1)NC(=O)c1cccc(O)c1	AAAAWE
CC(C)(C)OC(=O)NC(Cc1ccccc1)C(Cc1ccccc1)C(=O)NC1c2ccccc2CC1O	AAAAXE
CC(C)(C)NC(=O)C1CCCCN1CC(O)C(Cc1ccccc1)NC(=O)c1ccc2ccccc2n1	AAAAXO
CC(C)(C)C(O)C(Cc1ccccc1)NC(=O)C(Cc1ccccc1)NC(=O)c1ccc(Cl)cc1	AAAAYE
Cc1c(O)cccc1C(=O)NC(CSc1ccccc1)C(O)CN1CC2CCCCC2CC1C(=O)NC(C)(C)C	AAABAA
O=C(CCc1ccc(F)cc1F)N[C@@H](Cc1ccccc1)[C@@H](O)CNc1ccc(F)cc1	AAABBO
CC(Cc1ccccc1)C(=O)N1CCN(C[C@H](O)[C@H](Cc2ccccc2)NC(=O)c2c3ccccc3cc3ccccc23)CC1	AAABDE
CC(C)C(NC(=O)OC(C)(C)C)C(=O)N1CCN(C[C@H](O)[C@H](Cc2ccccc2)NC(=O)c2c3ccccc3cc3ccccc23)CC1	AAABFI
CC1(C)[C@H]2COc3ccc4ccccc4c3[C@H]2N=C2SC=C(c3ccc(S(=O)(=O)N4CCOCC4)cc3)N21	AAABIE
CC(OC1OCCN(Cc2n[nH]c(=O)[nH]2)C1c1ccc(F)cc1)c1cc(C(F)(F)F)cc(C(F)(F)F)c1	AAABHE
COc1ccc(-c2sc3n(c(NC(C)(C)C)c4[n+]3C3CC(=O)C=CC3(OC)CC4)c2-c2ccc(OC)cc2)cc1	AAABII
CC(C)(C)Nc1c2[n+](c3sc(-c4ccccc4)c(-c4ccccc4)n13)C13CCCC1(C=CC(=O)C3)OC2	AAABIU
COc1ccc(-c2sc3n(c(NC(C)(C)C)c4[n+]3[C@H]3CC(=O)C=C5CCCC[C@@]53OC4)c2-c2ccc(OC)cc2)cc1	AAABJE
CCN(CC)C(=O)c1cccnc1	AAABLA
CNC(=O)C(NC(=O)C(CC(C)C)C(O)C(O)=NO)C(C)(C)C	AAABLI
CCC1(CCC(C)C)C(=O)NC(=O)NC1=O	AAABLU
Cc1cccc(-c2ccc3c(c2)c(C(=O)C(=O)O)cn3Cc2ccc(C(C)(C)C)cc2)c1	AAABMU
CC1CS(=O)(=O)CCN1/N=C/c1ccc([N+](=O)[O-])o1	AAABNE
CCS(=O)(=O)N1CC(CC#N)(n2cc(-c3ncnc4[nH]ccc34)cn2)C1	AAACAI
CCN(CC)CCCC(C)Nc1ccnc2cc(Cl)ccc12	AAABWI
Cc1c(O)cccc1C(=O)NC(CSc1ccccc1)C(O)CN1CC2CCCCC2CC1C(=O)NC(C)(C)C	AAABZE
O=c1c2ccccc2[se]n1-c1ccccc1	AAACBA
Cc1nc(-c2cccnc2)sc1C(=O)Nc1ccccc1-c1cn2c(CN3CCOCC3)csc2n1	AAACBI
CCc1cc2c(cc1CC)CC(NCC(O)c1ccc(O)c3[nH]c(=O)ccc13)C2	AAACHA

4. Results obtained from USRCAT similarity search using a cut-off of 0.3 (top 15 molecules per active)

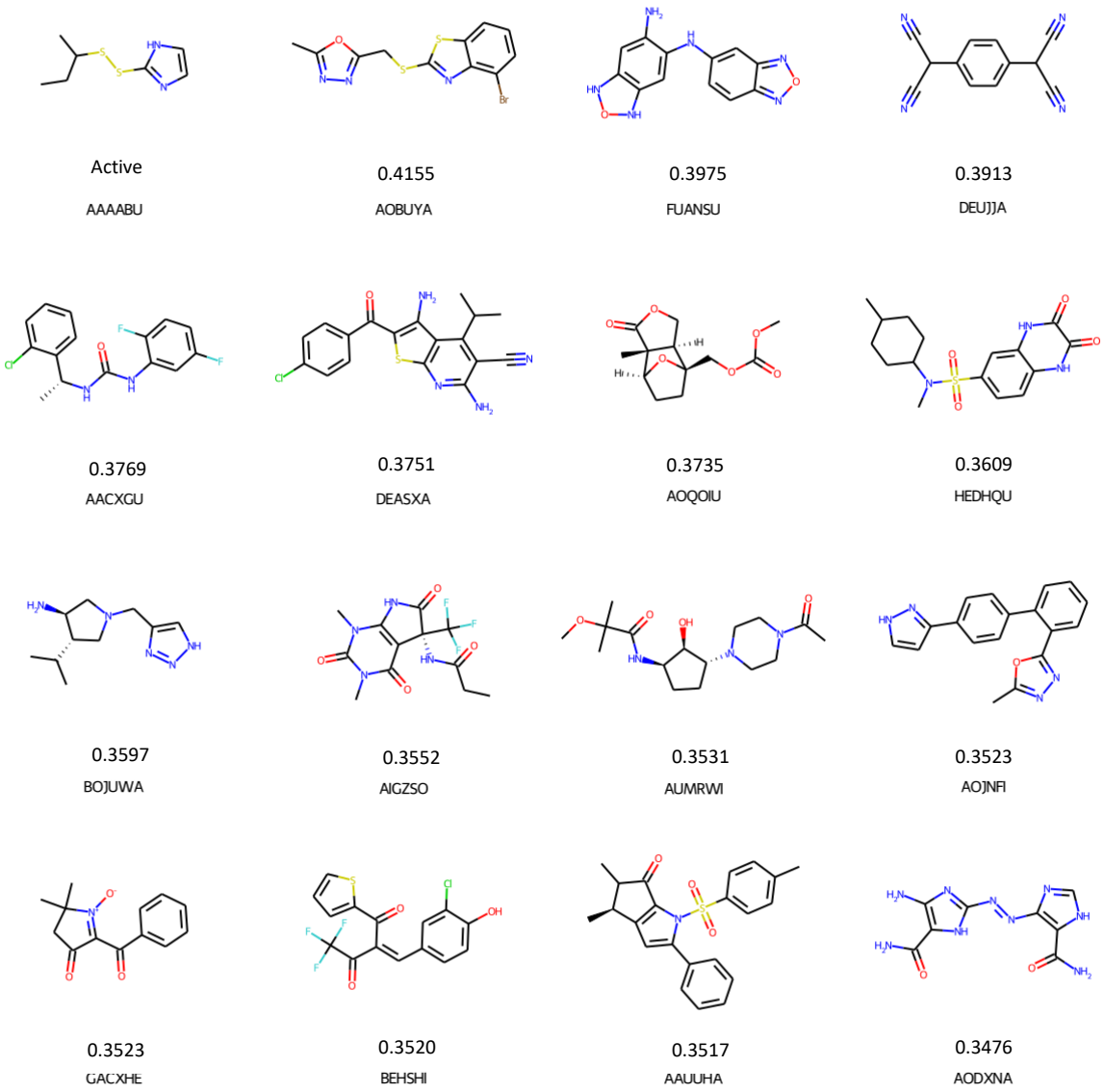


Figure A.13 Top 15 scoring molecules for the active molecule with index AAAABU. The quoted figure is the similarity score.

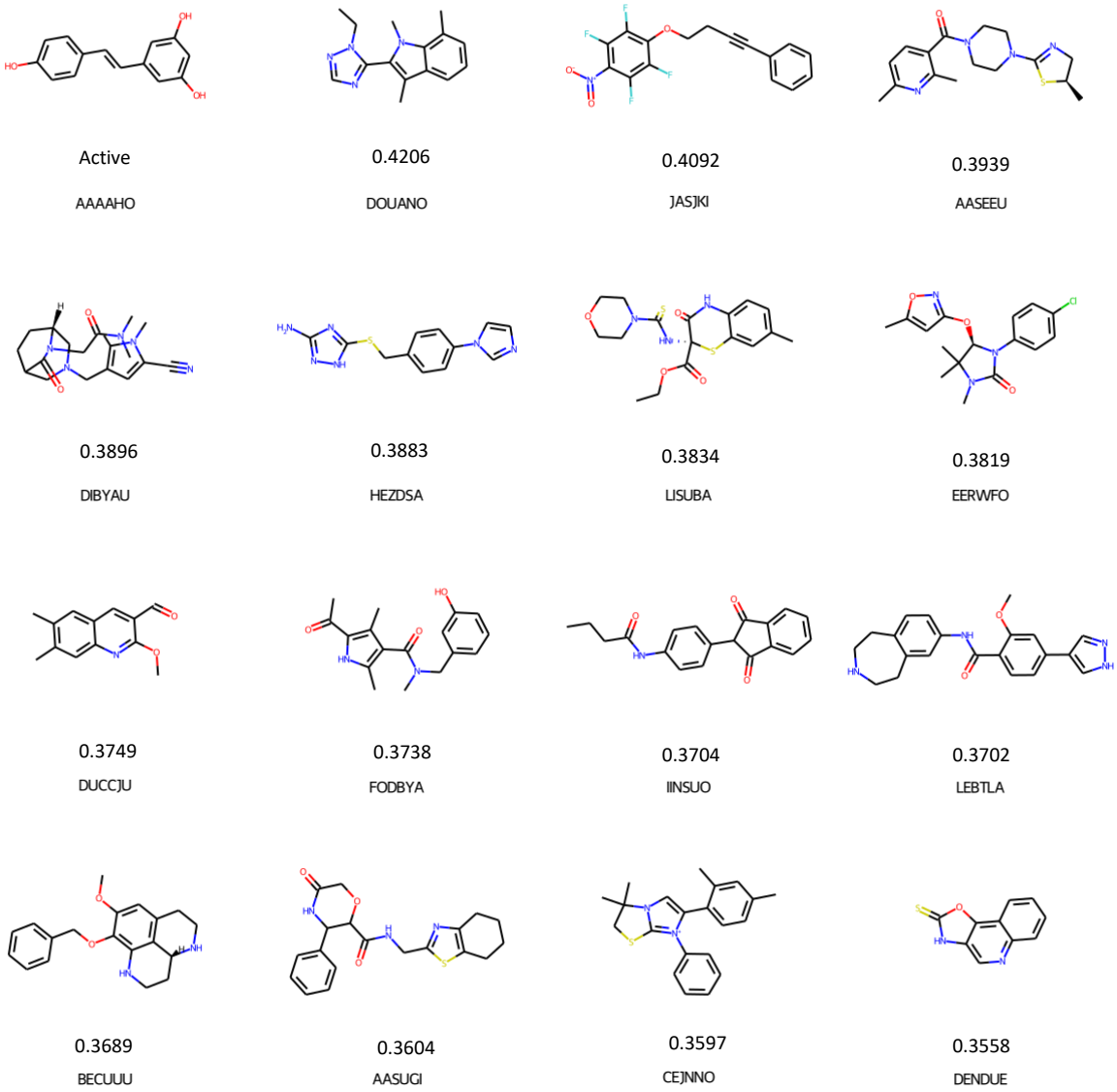


Figure A.14 Top 15 scoring molecules for the active molecule with index AAAAHO. The quoted figure is the similarity score.

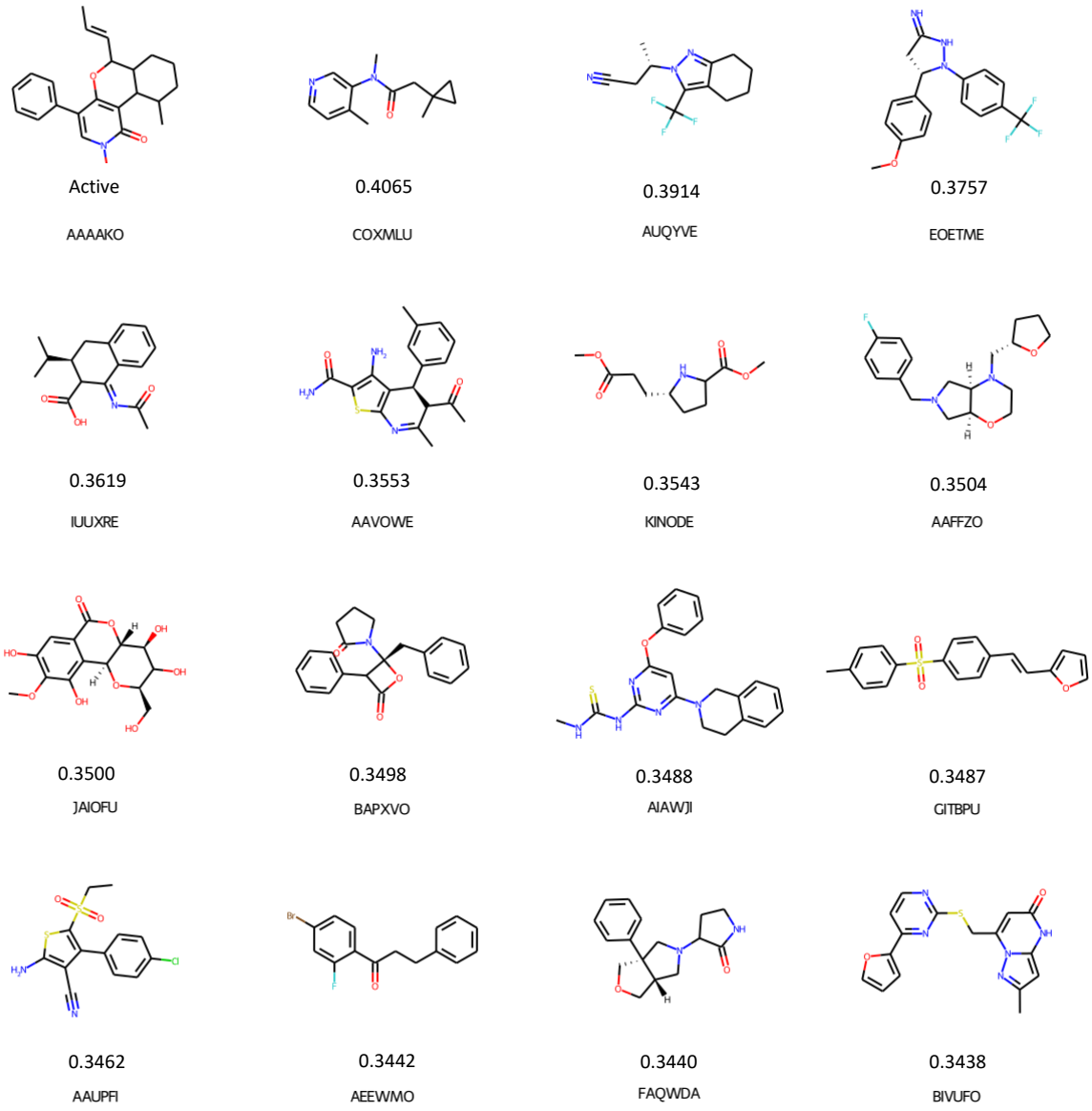


Figure A.15 Top 15 scoring molecules for the active molecule with index AAAAKO. The quoted figure is the similarity score.

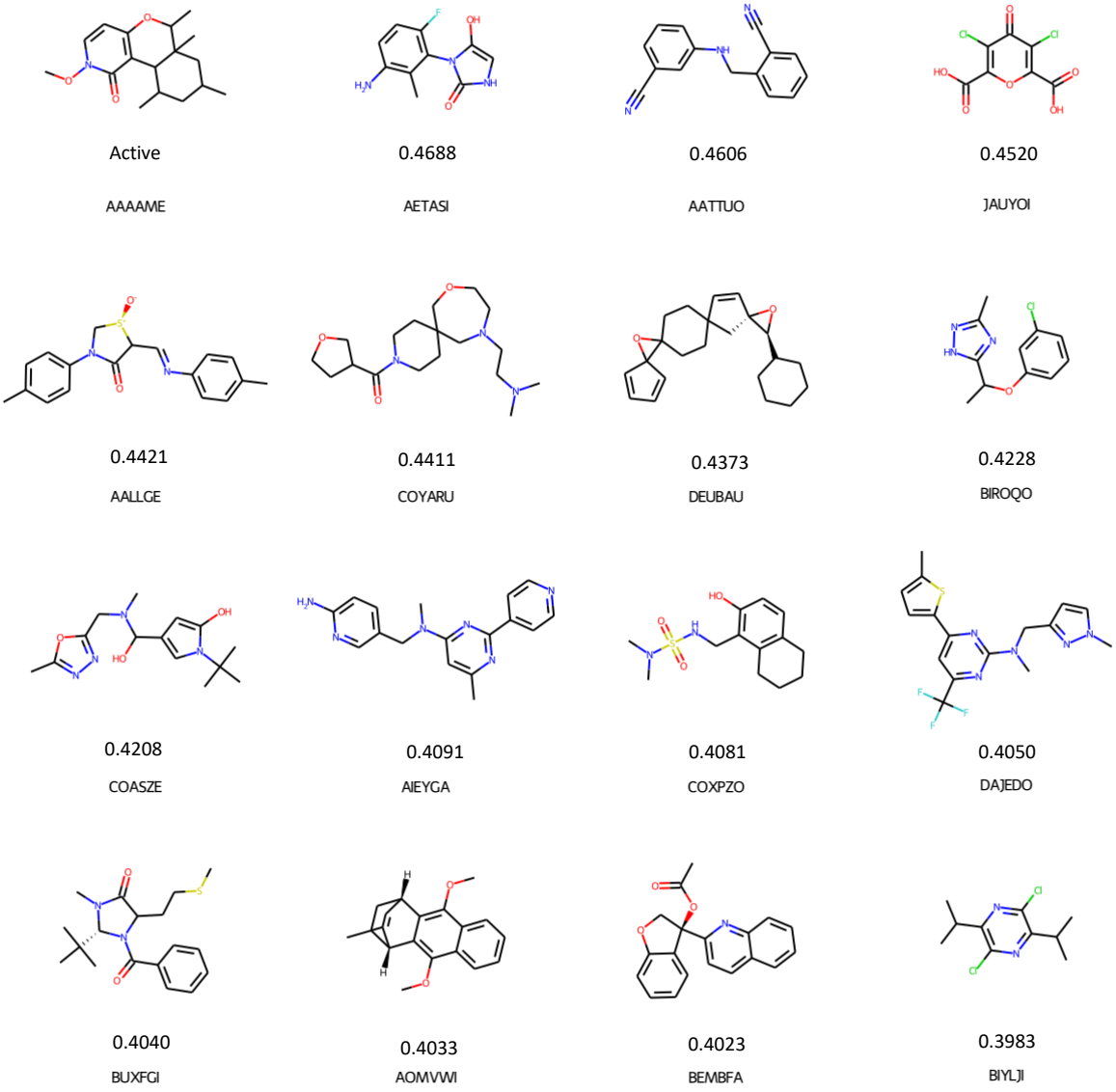


Figure A.16 Top 15 scoring molecules for the active molecule with index AAAAME. The quoted figure is the similarity score.

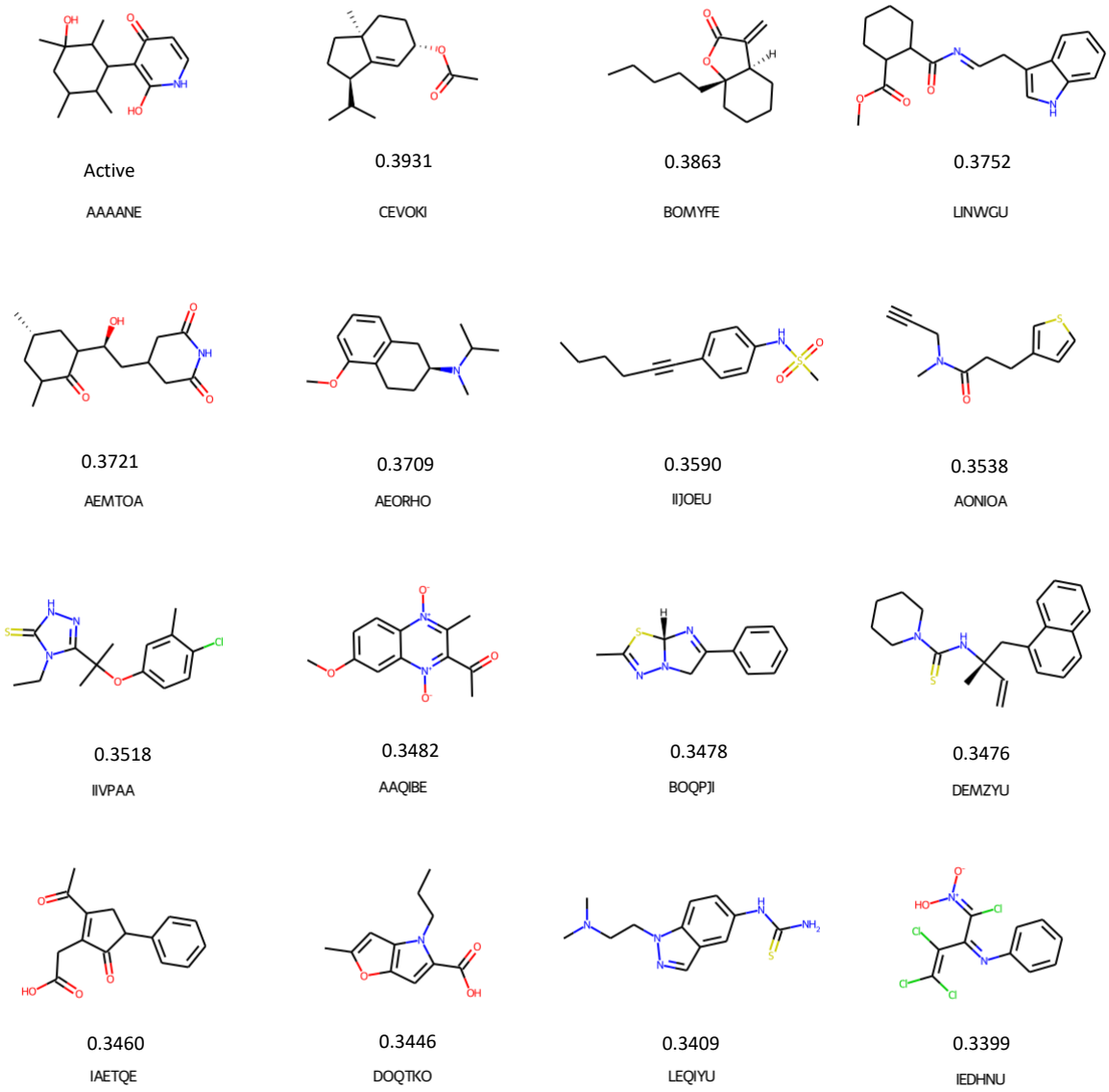


Figure A.17 Top 15 scoring molecules for the active molecule with index AAAANE. The quoted figure is the similarity score.

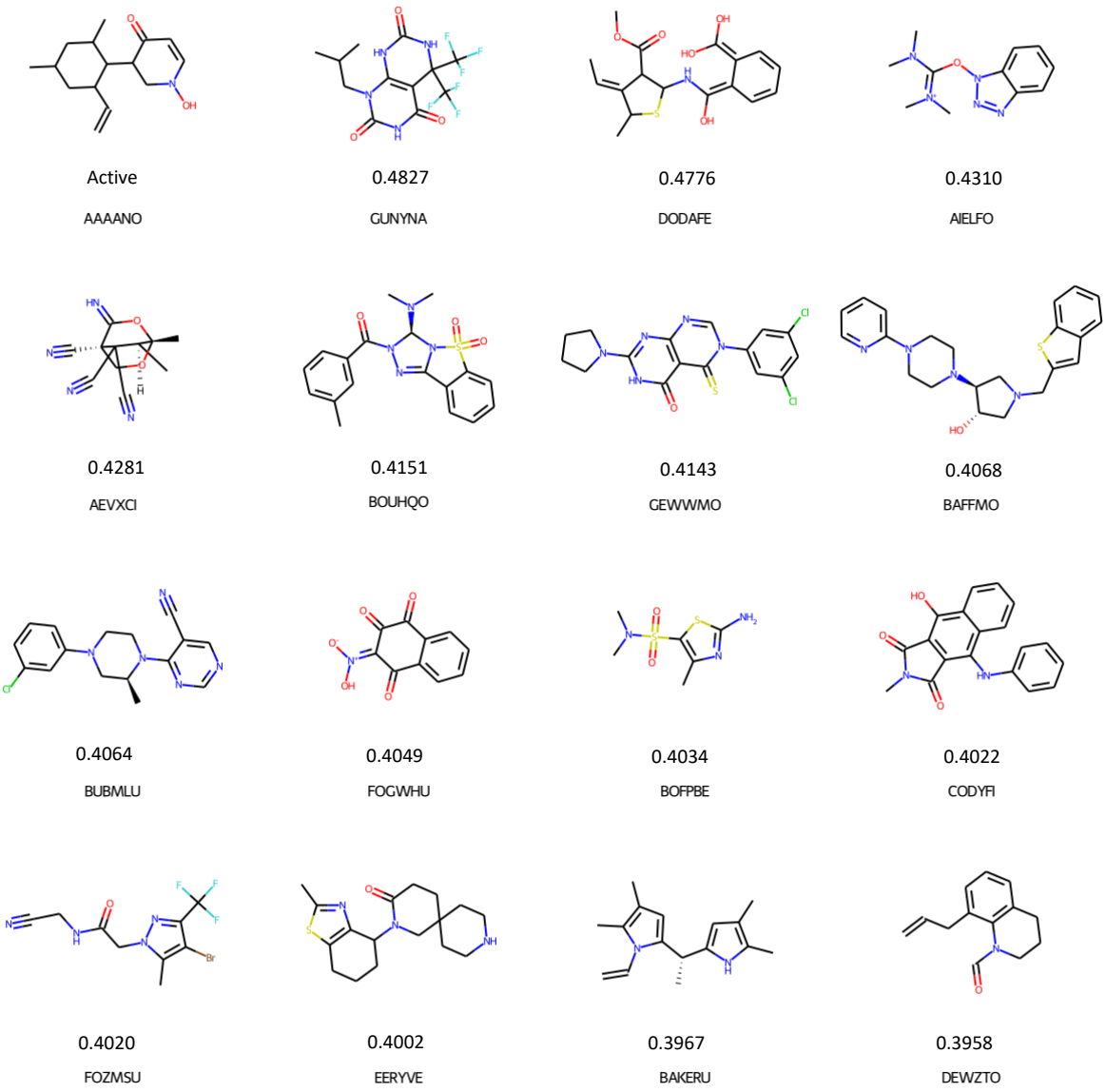


Figure A.18 Top 15 scoring molecules for the active molecule with index AAAANO. The quoted figure is the similarity score.

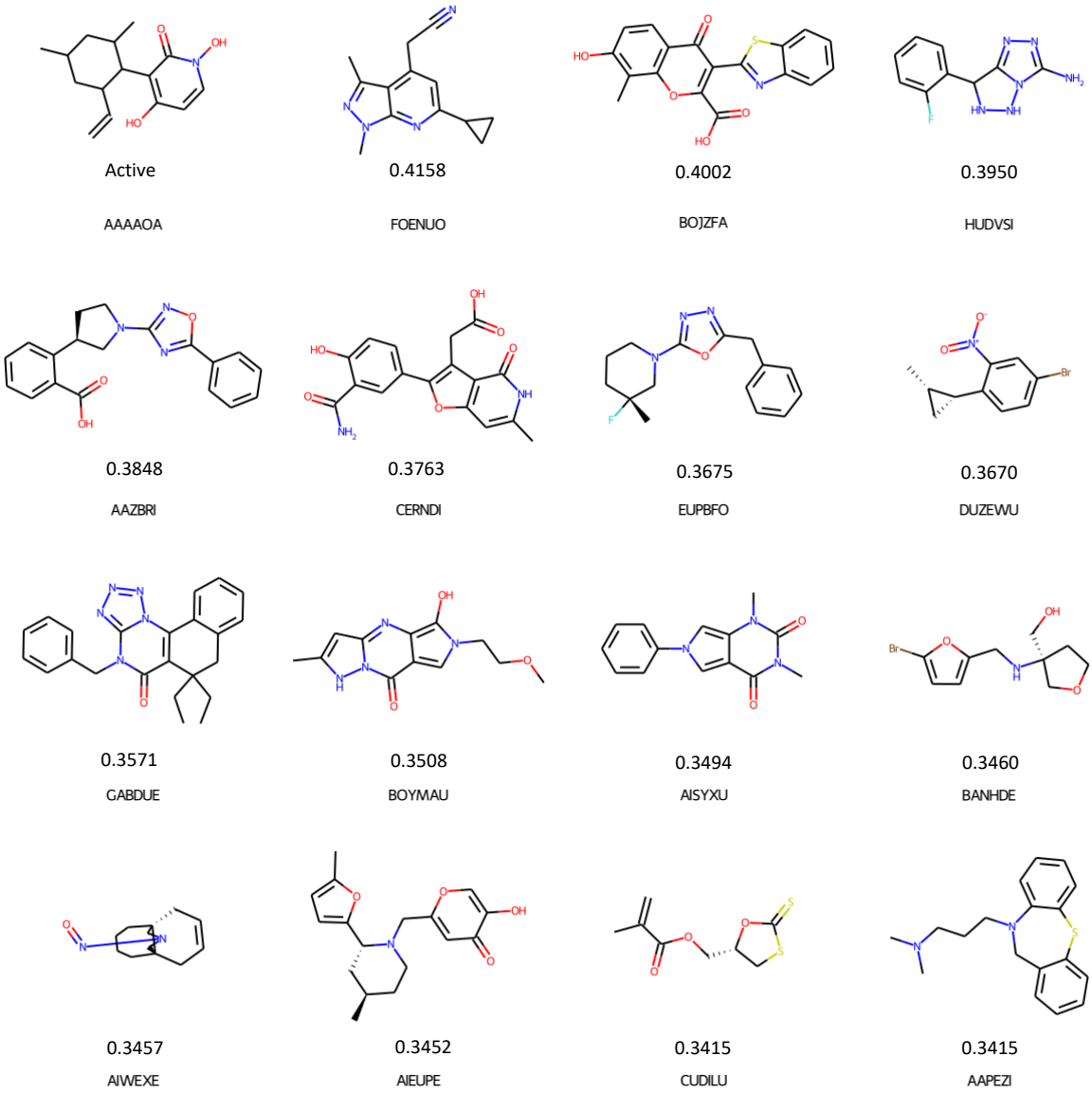


Figure A.19 Top 15 scoring molecules for the active molecule with index AAAAOA. The quoted figure is the similarity score.

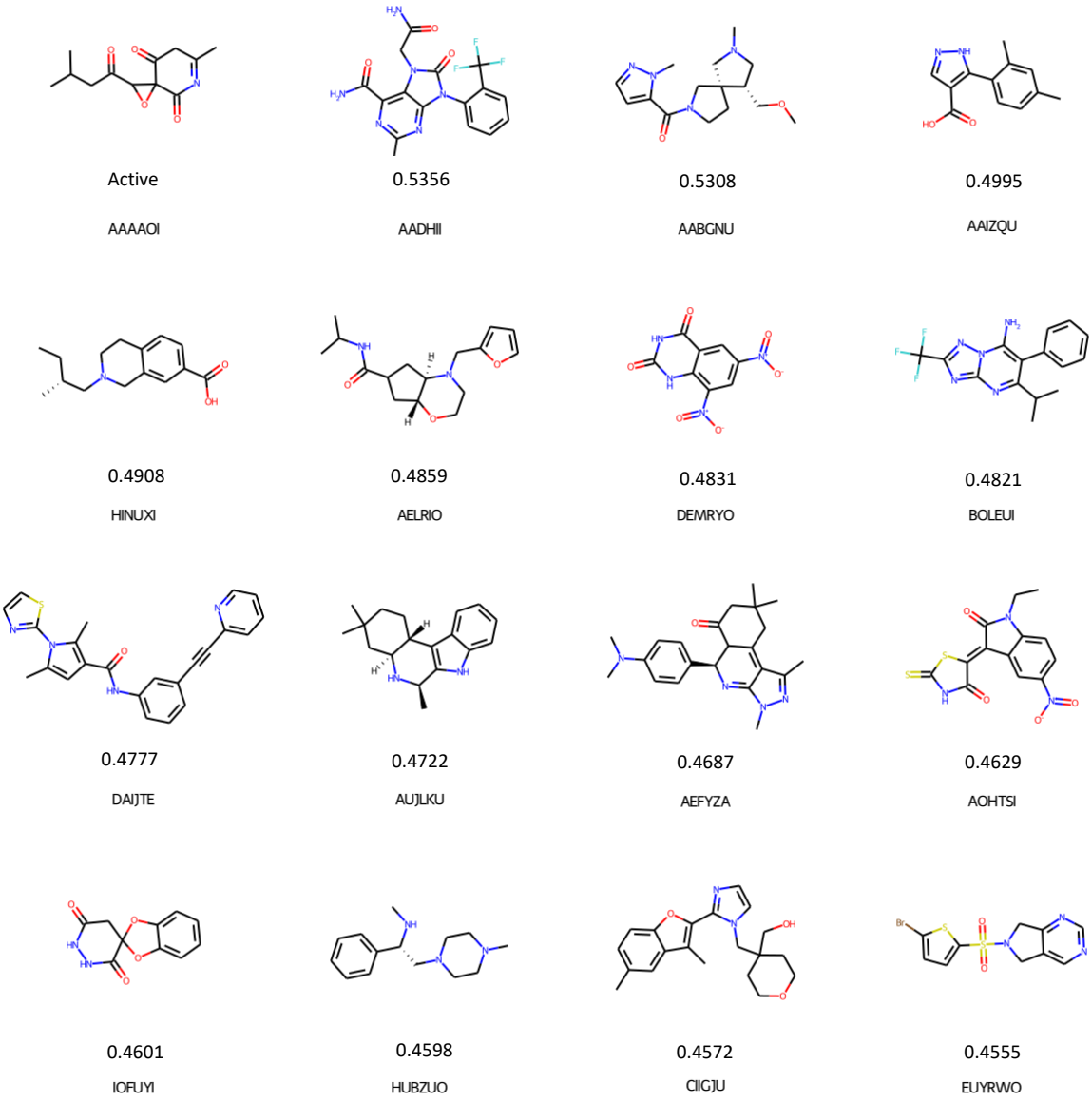


Figure A.20 Top 15 scoring molecules for the active molecule with index AAAAOI. The quoted figure is the similarity score.

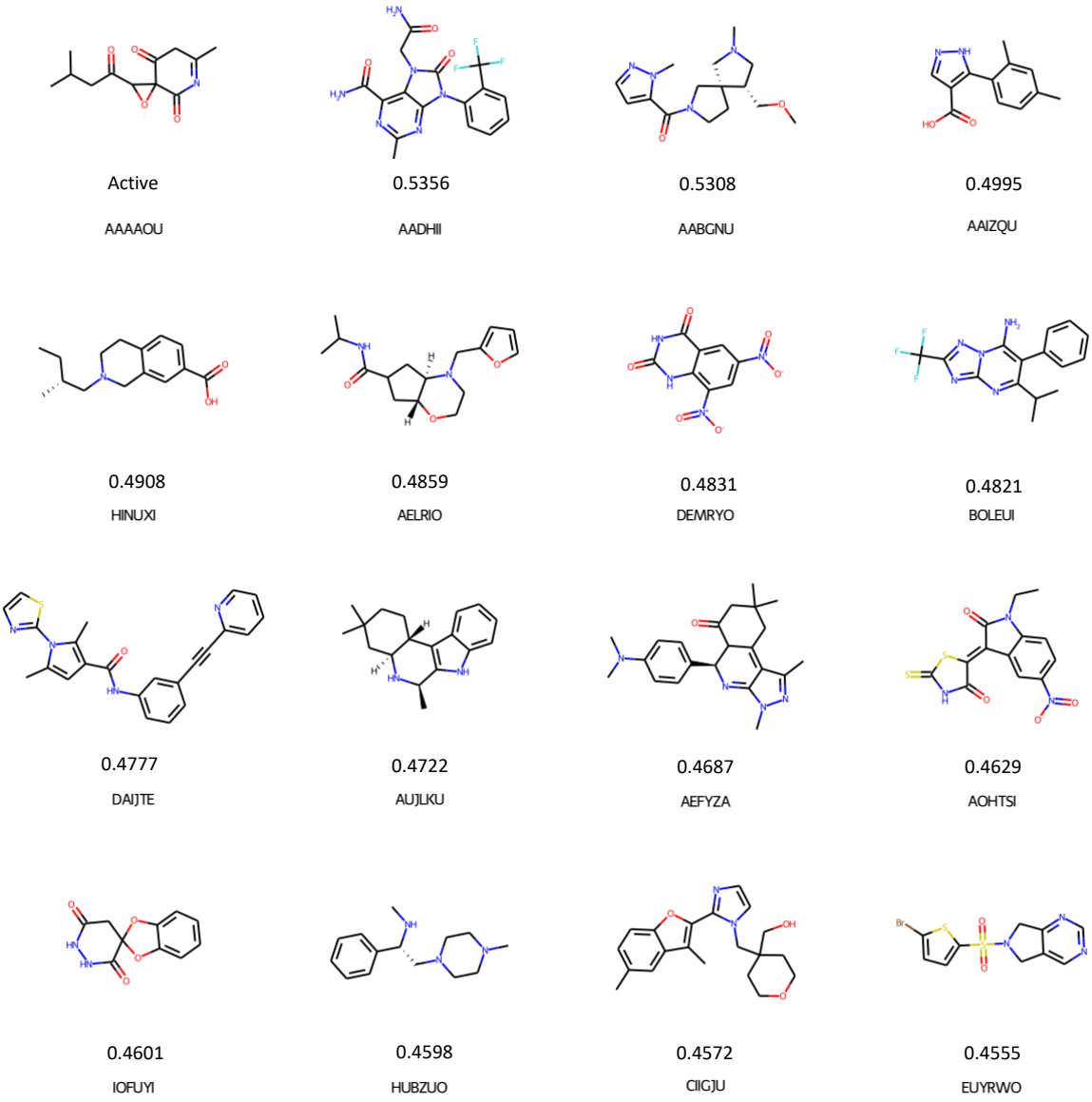


Figure A.21 Top 15 scoring molecules for the active molecule with index AAAAOU. The quoted figure is the similarity score.

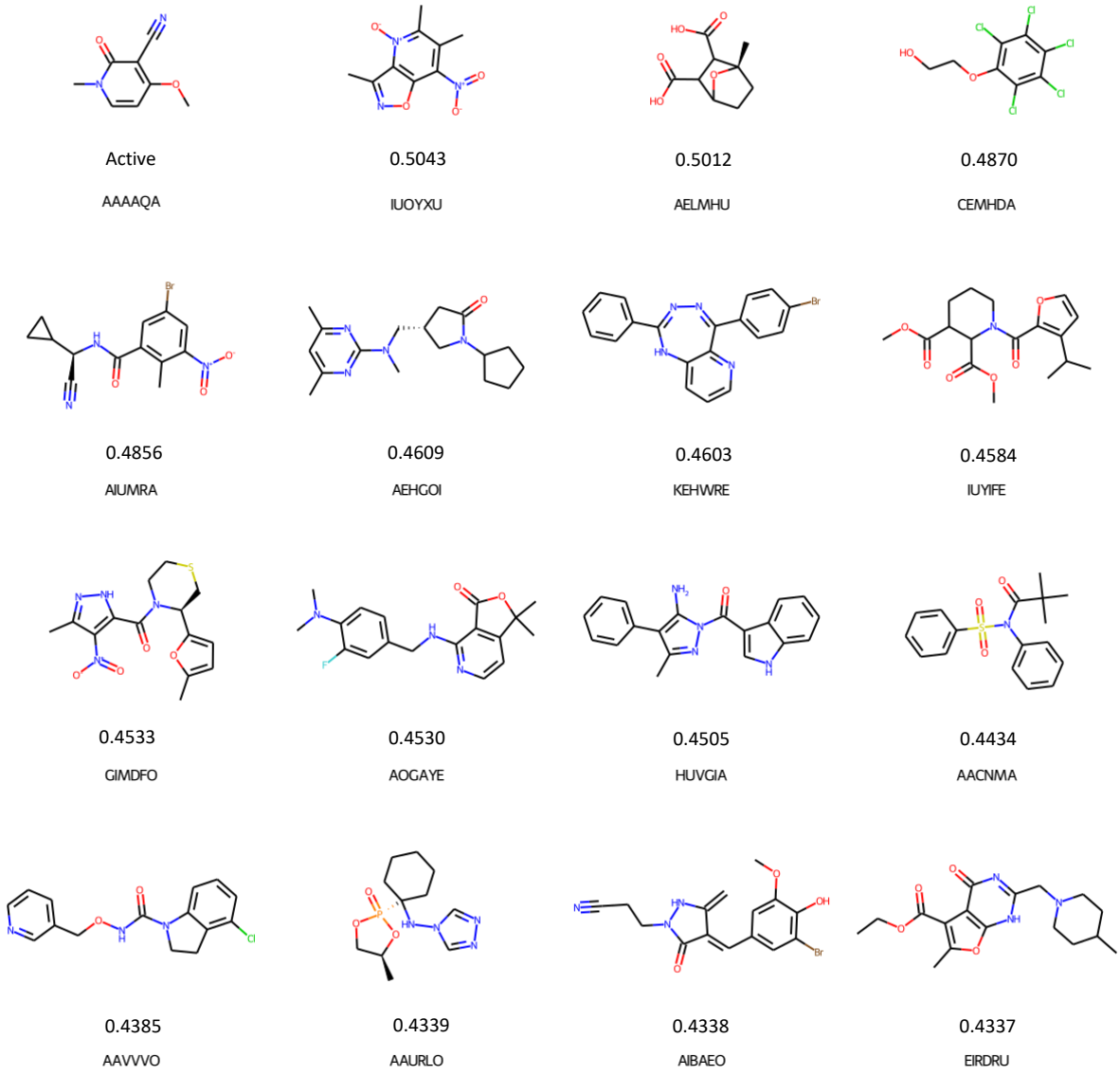


Figure A.22 Top 15 scoring molecules for the active molecule with index AAAAQA. The quoted figure is the similarity score.

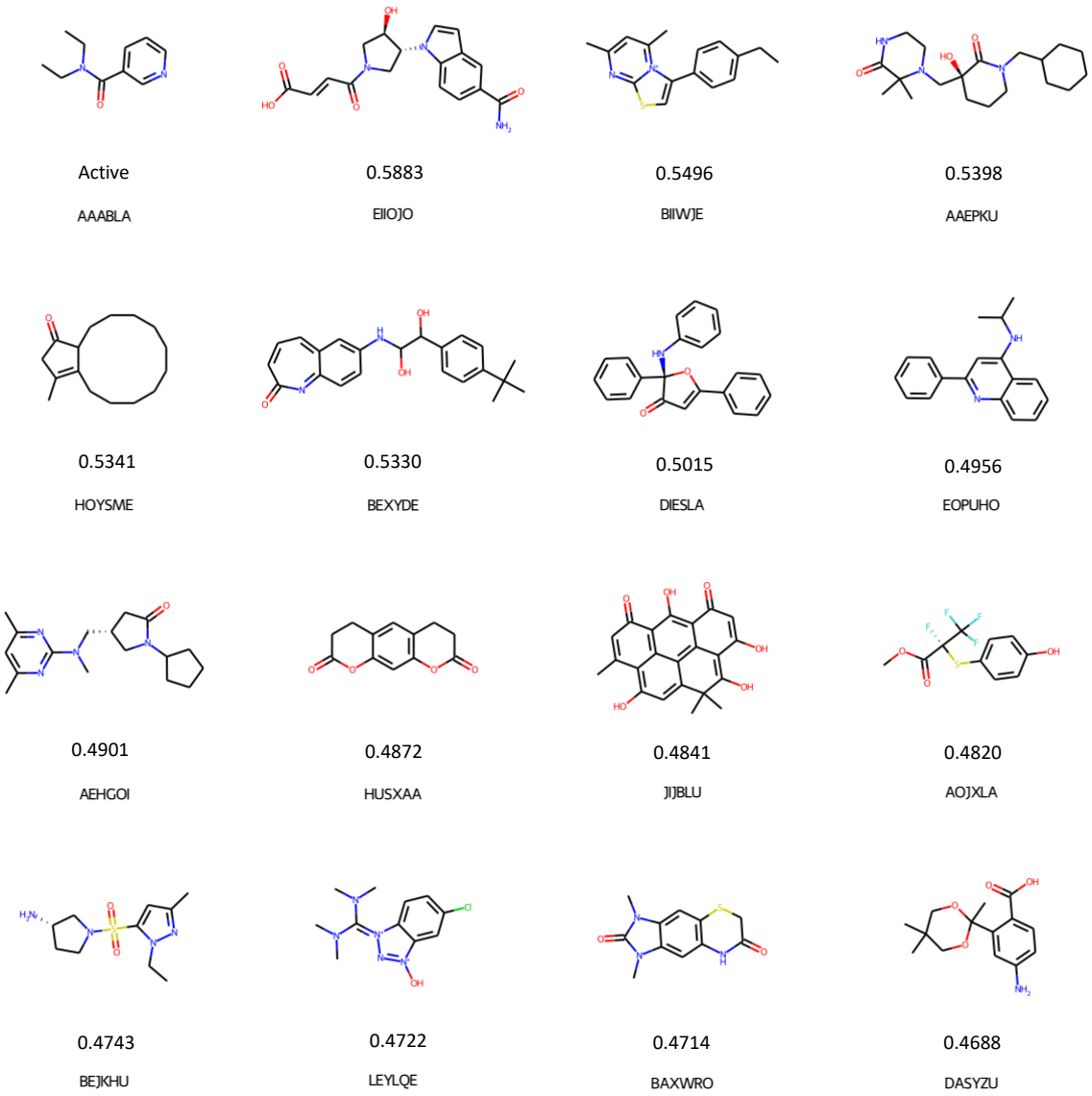


Figure A.23 Top 15 scoring molecules for the active molecule with index AAABLA. The quoted figure is the similarity score.

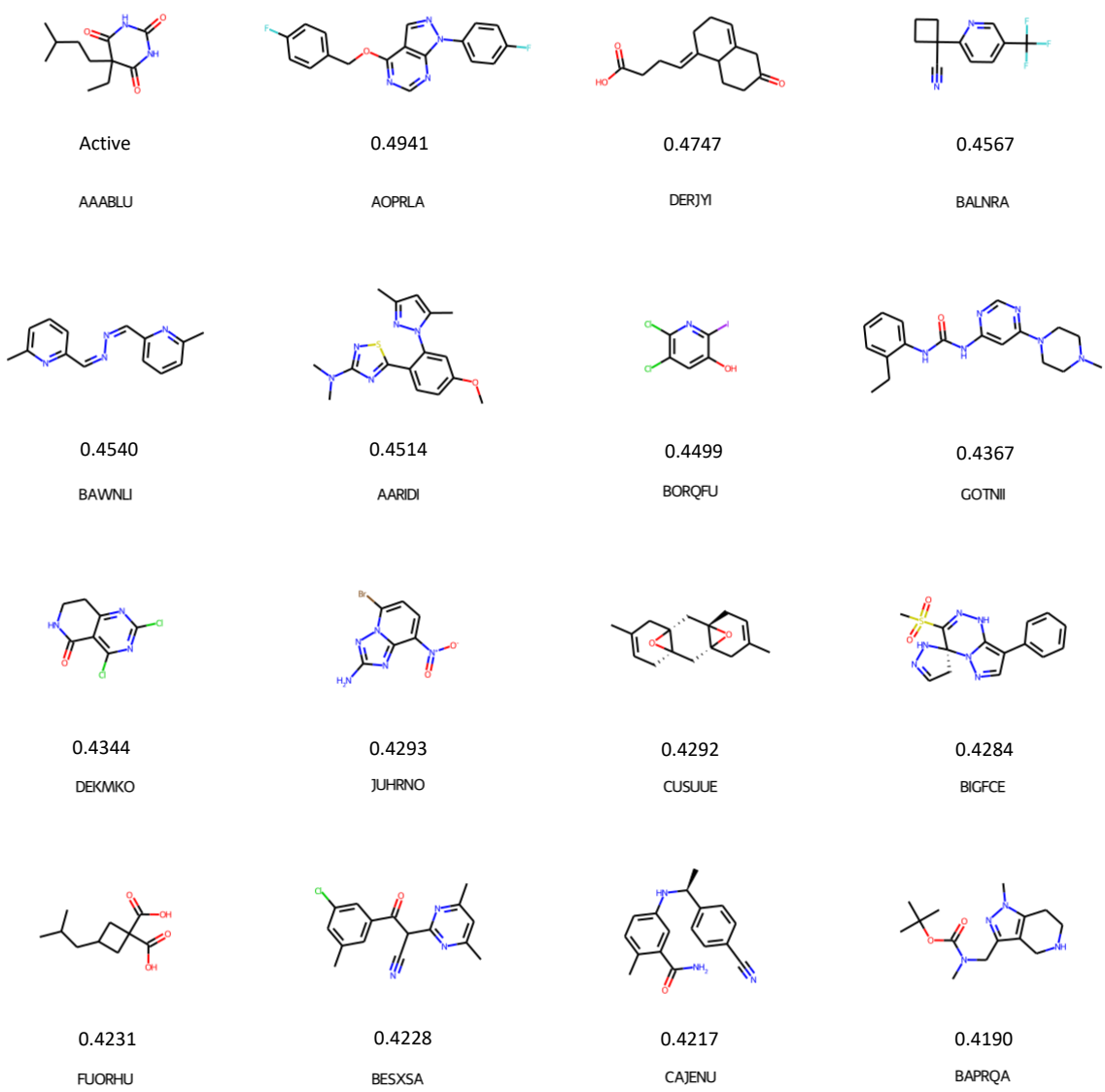


Figure A.24 Top 15 scoring molecules for the active molecule with index AAABLU. The quoted figure is the similarity score.

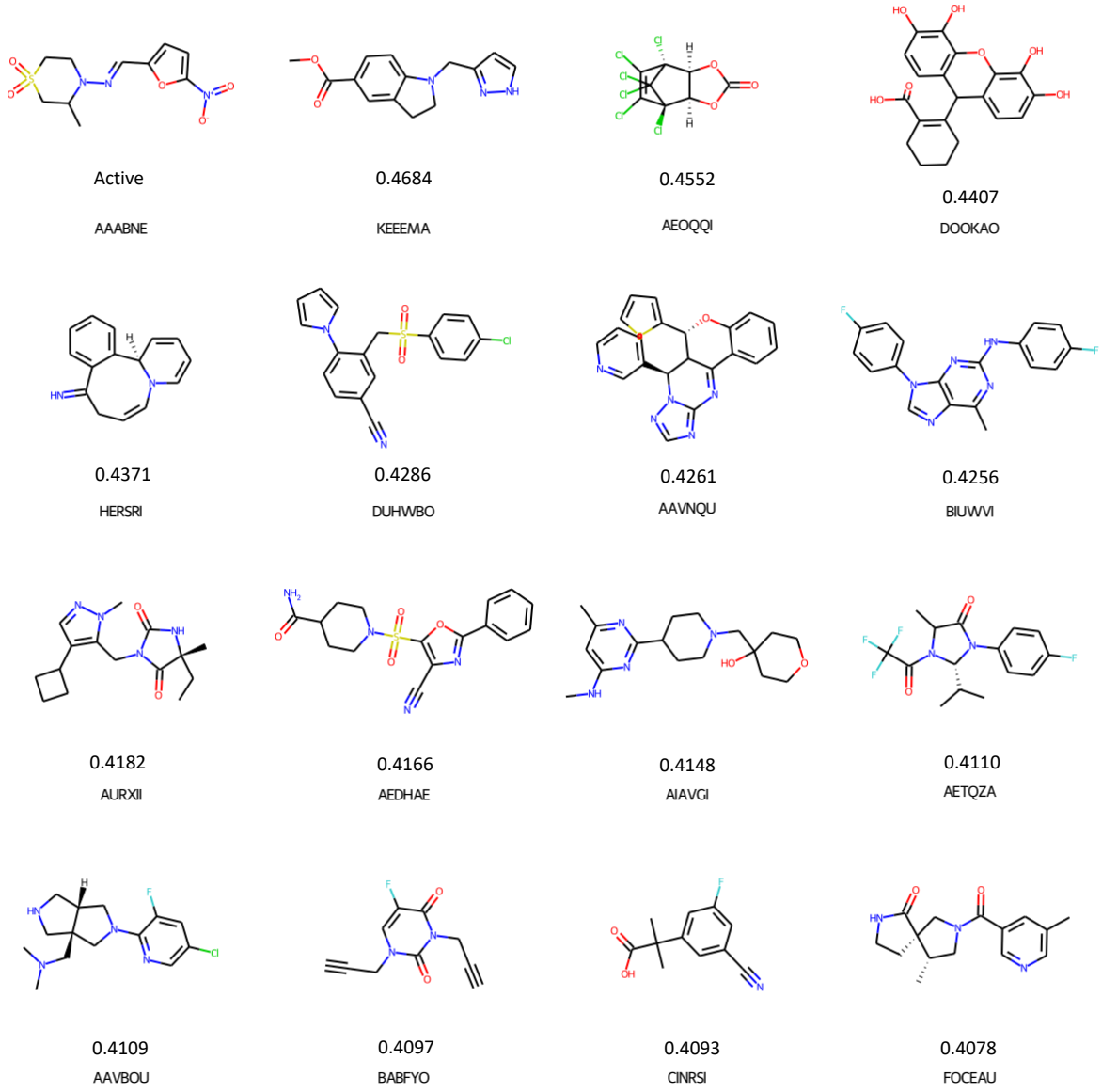


Figure A.25 Top 15 scoring molecules for the active molecule with index AAABNE. The quoted figure is the similarity score.

5. Consolidated list of hits

Table A.10 Consolidated list of hits obtained from fingerprint searches (Tanimoto coefficient - cut-off 0.4)

Molecule ID	SMILES
AISOOA	CCN(CC)C(=S)S[Sn](c1ccccc1)(c1ccccc1)c1ccccc1
IIRONO	CC(C)(C)N(CCNC(=O)n1cc(F)c(=O)[nH]c1=O)C(=O)O
FABKIU	CCCCCCNC(=O)n1cc(Cl)c(=O)n(C(=O)OCC(C)C)c1=O
AAFUBO	CN(C)CCCNc1ccnc2cc(Cl)ccc12
AAIWRI	Oc1cc(O)cc(C=Cc2ccc(O)c(O)c2)c1
BADVLU	OC[C@H]1O[C@@H](Cc2cc(O)cc(C=Cc3ccc(O)cc3)c2)[C@H](O)[C@@H](O)[C@@H]1O
AAENMO	CC(C)NC(=O)c1cc(Cl)cc(Cl)c1
EELHGO	O=C(Nc1nncs1)C(Cc1ccccc1)NC(=O)N1CCOCC1
GEERHU	CC(N)C(=O)NC(Cc1ccccc1)C(=O)N1CCCC1C(=O)Nc1ccc2ccccc2c1
AEVSIE	O=C(O)CC(O)(CC(=O)O)CC(=O)O
AOQEZE	CCOC(=O)C1=CC(OC(CC)CC)[C@@H]2N[C@@H]2C1
FIULEI	CCC(CC)OC1=CC(C(=O)O)C[C@H](N)[C@H]1NC(C)=O
AABQNE	CCN(CC)C(=O)c1cncc(C(N)=O)c1
BOJEME	CCN(CC)CCN(CC(=O)N(C1CCCCC1)C1CCCCC1)C(=O)c1cccnc1
AAIAJO	CCN(CC)C(=O)c1ccncc1B1OC(C)(C)C(C)(C)O1
AAAJWE	CCN(CC)C(=O)c1ccc(N(C)S(=O)(=O)c2ccccc2)cc1
AUDQFU	CC(C)CC(C(=O)NC(C(=O)OC1CCCC1)c1ccccc1)C(O)C(O)=NO
AAMTFE	CCCC1(C[C@H](C)CC)C(=O)NC(=O)NC1=O
CAPEUA	CC[C@]1(CCC(C)C)C(=O)NC(=O)N(C(=O)c2ccccc2I)C1=O
AAOPOA	O=C1OCCN1N=Cc1ccc([N+](=O)[O-])o1
AUEZQE	N#CC[C](C1CCCC1)n1cc(-c2ncnc3[nH]ccc23)cn1
AADZUU	COc1ccc(C[C@@H](C)NC[C@H](O)c2ccc(O)c3[nH]c(=O)ccc23)cc1
EAFIQO	CCCCCCCCSCCNC(=O)CCn1cc(F)c(=O)[nH]c1=O
FOOUTU	CC(C=O)NC(=O)C(Cc1ccc[nH]1)NC(=O)c1ccccc1
FEGYWI	Nc1ncnc2c(-c3[nH]c(CO)c(O)c3O)c[nH]c12
JUDSCI	O=C(NCCO)NC(=S)c1cccnc1
EUSOWU	Oc1cc(O)c2cc(O)c(-c3ccc(O)c(O)c3)[o+]c2c1
BUQHJA	C[C@@H]1O[C@@H](Oc2cc3c(O)cc(O)cc3cc2-c2cc(O)c(O)c(O)c2)[C@H](O)[C@H](O)[C@H]1O
LARPRE	COc1cc(F)cc(C(=O)Nc2ccccc2-c2cn3c(CN4CCNCC4)csc3n2)c1
AONUIA	CC[C@H](C)C[C@H](C)/C=C(\C)[C@@H]1O[C@H](c2c3c(cn(C)c2=O)[C@]2(O)CCC(=O)C[C@@H]2O3)CC[C@@H]1C
COUEQU	CC(=O)CNC(=O)C(Cc1ccc[nH]1)NC(=O)c1ccccc1
FUAVOI	COc1cc(N2CCC(N)CC2)ccc1Nc1ncc(Cl)c(-c2c[nH]c3ccccc23)n1
BOBDUI	O=C(c1ccc2ccccc2c1)C([C@H](c1ccccc1)N1CCOCC1)N1CCOCC1
EAOMMI	C[N+](C)(C)CCOC(=O)C(O)(CC(=O)O)CC(=O)O
DEFEHA	CC(C)(C)NC(=O)C1C[C@@H]2CCCC[C@@H]2CN1C[C@@H](O)[C@@H](N)Cc1ccccc1
MOMMQU	O=C(N[C@@H]1c2ccccc2C[C@@H]1O)C1COCCN1CC1CCC1
AUIRFU	CC(C)C(=NC(=O)C(Cc1ccc(Cl)cc1)NC(=O)c1ccc(Br)cc1)C(=O)O
AACUGO	COc1cc(C=Cc2ccc(O)cc2)cc(OC)c1
DAQOZU	O=c1cc(O)cc(/C=C/c2ccccc2)o1
BUKUUU	C=C[C@@H](C=Cc1ccc(O)cc1)c1ccc(O)cc1
BITVFO	COc1cc(C(C)=O)ccc1Oc1c(C#N)c(=O)n(C)c(=O)n1C
AOBGSU	CN(C(=O)c1cccnc1)c1nc2ccccc2o1

Table A.11 Consolidated list of hits obtained from USRCAT using a cut-off of 0.3 (Top 15 molecules per active)

Molecule ID	SMILES
AOBUYA	Cc1nnc(CSc2nc3c(Br)cccc3s2)o1
FUANSU	Nc1cc2c(cc1Nc1ccc3nonc3c1)NON2
DEUJJA	N#CC(C#N)c1ccc(C(C#N)C#N)cc1
AACXGU	C[C@@H](NC(=O)Nc1cc(F)ccc1F)c1ccccc1Cl
DEASXA	CC(C)c1c(C#N)c(N)nc2sc(C(=O)c3ccc(Cl)cc3)c(N)c12
AOQOIU	COC(=O)OC[C@@]12CC[C@H](O1)[C@]1(C)C(=O)OC[C@@H]21
HEDHQU	CC1CCC(N(C)S(=O)(=O)c2ccc3[nH]c(=O)c(=O)[nH]c3c2)CC1
BOJUWA	CC(C)[C@H]1CN(Cc2c[nH]nn2)C[C@@H]1N
AIGZSO	CCC(=O)N[C@@]1(C(F)(F)F)C(=O)Nc2c1c(=O)n(C)c(=O)n2C
AUMRWI	COC(C)(C)C(=O)N[C@@H]1CC[C@@H](N2CCN(C(C)=O)CC2)[C@@H]1O
AOJNFI	Cc1nnc(-c2ccccc2-c2ccc(-c3cc[nH]n3)cc2)o1
GACXHE	CC1(C)CC(=O)C(C(=O)c2ccccc2)=[N+]1[O-]
BEHSHI	O=C(C(=Cc1ccc(O)c(Cl)c1)C(=O)C(F)(F)F)c1cccs1
AAUUHA	Cc1ccc(S(=O)(=O)n2c(-c3ccccc3)cc3c2C(=O)C(C)[C@@H]3C)cc1
AODXNA	NC(=O)c1[nH]c(N=Nc2nc[nH]c2C(N)=O)nc1N
DOUANO	CCn1ncnc1-c1c(C)c2cccc(C)c2n1C
JASJKI	O=[N+]([O-])c1c(F)c(F)c(OCCC#Cc2ccccc2)c(F)c1F
AASEEU	Cc1ccc(C(=O)N2CCN(C3=NC[C@@H](C)S3)CC2)c(C)n1
DIBYAU	Cc1c(CN2CC3CC[C@H](C2)N(CC(=O)N(C)C)C3=O)cc(C#N)n1C
HEZDSA	Nc1n[nH]c(SCc2ccc(-n3ccnc3)cc2)n1
LISUBA	CCOC(=O)[C@]1(NC(=S)N2CCOCC2)Sc2cc(C)ccc2NC1=O
EERWFO	Cc1cc(O[C@H]2N(c3ccc(Cl)cc3)C(=O)N(C)C2(C)C)no1
DUCCJU	COc1nc2cc(C)c(C)cc2cc1C=O
FODBYA	CC(=O)c1[nH]c(C)c(C(=O)N(C)Cc2cccc(O)c2)c1C
IINSUO	CCCC(=O)Nc1ccc(C2C(=O)c3ccccc3C2=O)cc1
LEBTLA	COc1cc(-c2cn[nH]c2)ccc1C(=O)Nc1ccc2c(c1)CCNCC2
BECUUU	COc1cc2c3c(c1OCc1ccccc1)NCC[C@H]3NCC2
AASUGI	O=C1COC(C(=O)NCc2nc3c(s2)CCCC3)C(c2ccccc2)N1
CEJNNO	Cc1ccc(-c2cn3c([n+]2-c2ccccc2)SCC3(C)C)c(C)c1
DENDUE	S=c1[nH]c2cnc3ccccc3c2o1
COXMLU	Cc1ccncc1N(C)C(=O)CC1(C)CC1
AUQYVE	C[C@@H](CC#N)n1nc2c(c1C(F)(F)F)CCCC2
EOETME	COc1ccc([C@@H]2CC(=N)NN2c2ccc(C(F)(F)F)cc2)cc1
IUUXRE	CC(=O)N=C1c2ccccc2C[C@H](C(C)C)C1C(=O)O
AAVOWE	CC(=O)C1C(C)=Nc2sc(C(N)=O)c(N)c2[C@H]1c1cccc(C)c1
KINODE	COC(=O)CC[C@H]1CCC(C(=O)OC)N1
AAFFZO	Fc1ccc(CN2C[C@@H]3OCCN(C[C@@H]4CCCO4)[C@@H]3C2)cc1
JAIOFU	COc1c(O)cc2c(c1O)[C@@H]1O[C@H](CO)[C](O)[C@H](O)[C@H]1OC2=O
BAPXVO	O=C1O[C@@](Cc2ccccc2)(N2CCCC2=O)C1c1ccccc1
AIAWJI	CNC(=S)Nc1nc(Oc2ccccc2)cc(N2CCc3ccccc3C2)n1
GITBPU	Cc1ccc(S(=O)(=O)c2ccc(/C=C/c3ccco3)cc2)cc1
AAUPFI	CCS(=O)(=O)c1sc(N)c(C#N)c1-c1ccc(Cl)cc1
AEEWMO	O=C(CCc1ccccc1)c1ccc(Br)cc1F
FAQWDA	O=C1NCCC1N1C[C@@H]2COC[C@]2(c2ccccc2)C1
BIVUFO	Cc1cc2[nH]c(=O)cc(CSc3nccc(-c4ccco4)n3)n2n1
AETASI	Cc1c(N)ccc(F)c1-n1c(O)c[nH]c1=O
AATTUO	N#Cc1cccc(NCc2ccccc2C#N)c1
JAUYOI	O=C(O)c1oc(C(=O)O)c(Cl)c(=O)c1Cl
AALLGE	Cc1ccc(N=CC2C(=O)N(c3ccc(C)cc3)C[S@@+]2[O-])cc1
COYARU	CN(C)CCN1CCOCC2(CCN(C(=O)C3CCOC3)CC2)C1
DEUBAU	C1=CC2(C=C1)OC21CCC2(C=C[C@@]3(C2)O[C@H]3C2CCCCC2)CC1
BIROQO	Cc1n[nH]c(C(C)Oc2cccc(Cl)c2)n1
COASZE	Cc1nnc(CN(C)C(O)c2cc(O)n(C(C)(C)C)c2)o1
AIEYGA	Cc1cc(N(C)Cc2ccc(N)nc2)nc(-c2ccncc2)n1
COXPZO	CN(C)S(=O)(=O)NCc1c(O)ccc2c1CCCC2
DAJEDO	Cc1ccc(-c2cc(C(F)(F)F)nc(N(C)Cc3ccn(C)n3)n2)s1
BUXFGI	CSCCC1C(=O)N(C)[C@@H](C(C)(C)C)N1C(=O)c1ccccc1

Table A.12 Consolidated list of hits obtained from USRCAT (continued)

Molecule ID	SMILES
AOMVWI	COc1c2c(c(OC)c3ccccc13)[C@@H]1CC[C@H]2C=C1C
BEMBFA	CC(=O)O[C@]1(c2ccc3ccccc3n2)COc2ccccc21
BIYLJI	CC(C)c1nc(Cl)c(C(C)C)nc1Cl
CEVOKI	CC(=O)O[C@@H]1C=C2[C@@H](C(C)C)CC[C@@]2(C)CC1
BOMYFE	C=C1C(=O)O[C@@]2(CCCCC)CCCC[C@H]12
LINWGU	COC(=O)C1CCCCC1C(=O)N=CCc1c[nH]c2ccccc12
AEMTOA	CC1C[C@H](C)CC([C@@H](O)CC2CC(=O)NC(=O)C2)C1=O
AEORHO	COc1cccc2c1CC[C@H](N(C)C(C)C)C2
IIJOEU	CCCCC#Cc1ccc(NS(C)(=O)=O)cc1
AONIOA	C#CCN(C)C(=O)CCc1ccsc1
IIVPAA	CCn1c(C(C)(C)Oc2ccc(Cl)c(C)c2)n[nH]c1=S
AAQIBE	COc1ccc2c(c1)[n+]([O-])c(C(C)=O)c(C)[n+]2[O-]
BOQPJI	CC1=NN2CC(c3ccccc3)=N[C@H]2S1
DEMZYU	C=C[C@@](C)(Cc1cccc2ccccc12)NC(=S)N1CCCCC1
IAETQE	CC(=O)C1=C(CC(=O)O)C(=O)C(c2ccccc2)C1
DOQTKO	CCCn1c(C(=O)O)cc2oc(C)cc21
LEQIYU	CN(C)CCn1ncc2cc(NC(N)=S)ccc21
IEDHNU	[O-][N+](O)=C(Cl)C(=Nc1ccccc1)C(Cl)=C(Cl)Cl
GUNYNA	CC(C)Cn1c2c(c(=O)[nH]c1=O)C(C(F)(F)F)(C(F)(F)F)NC(=O)N2
DODAFE	CC=C1C(C)SC(NC(O)=c2ccccc2=C(O)O)C1C(=O)OC
AIELFO	CN(C)C(On1nnc2ccccc21)=[N+](C)C
AEVXCI	C[C@H]1C(C#N)(C#N)[C@@]2(C#N)CO[C@@]1(C)OC2=N
BOUHQO	Cc1cccc(C(=O)N2N=C3c4ccccc4S(=O)(=O)N3[C@@H]2N(C)C)c1
GEWWMO	O=c1[nH]c(N2CCCC2)nc2ncn(-c3cc(Cl)cc(Cl)c3)c(=S)c12
BAFFMO	O[C@H]1CN(Cc2cc3ccccc3s2)C[C@@H]1N1CCN(c2ccccn2)CC1
BUBMLU	C[C@H]1CN(c2cccc(Cl)c2)CCN1c1ncncc1C#N
FOGWHU	O=c1c(=[N+]([O-])O)c(=O)c2ccccc2c1=O
BOFPBE	Cc1nc(N)sc1S(=O)(=O)N(C)C
CODYFI	CN1C(=O)c2c(c(Nc3ccccc3)c3ccccc3c2O)C1=O
FOZMSU	Cc1c(Br)c(C(F)(F)F)nn1CC(=O)NCC#N
EERYVE	Cc1nc2c(s1)CCCC2N1CC2(CCNCC2)CCC1=O
BAKERU	C=Cn1c([C@@H](C)c2cc(C)c(C)[nH]2)cc(C)c1C
DEWZTO	C=CCc1cccc2c1N(C=O)CCC2
FOENUO	Cc1nn(C)c2nc(C3CC3)cc(CC#N)c12
BOJZFA	Cc1c(O)ccc2c(=O)c(-c3nc4ccccc4s3)c(C(=O)O)oc12
HUDVSI	Nc1nnc2n1NNC2c1ccccc1F
AAZBRI	O=C(O)c1ccccc1[C@H]1CCN(c2noc(-c3ccccc3)n2)C1
CERNDI	Cc1cc2oc(-c3ccc(O)c(C(N)=O)c3)c(CC(=O)O)c2c(=O)[nH]1
EUPBFO	C[C@@]1(F)CCCN(c2nnc(Cc3ccccc3)o2)C1
DUZEWU	C[C@H]1C[C@H]1c1ccc(Br)cc1[N+](=O)[O-]
GABDUE	CCC1(CC)Cc2ccccc2-c2c1c(=O)n(Cc1ccccc1)c1nnnn21
BOYMAU	COCCn1cc2c(=O)n3[nH]c(C)cc3nc2c1O
AISYXU	Cn1c(=O)c2cn(-c3ccccc3)cc2n(C)c1=O
BANHDE	OC[C@]1(NCc2ccc(Br)o2)CCOC1
AIWEXE	O=NN1C[C@]23CC=CC[C@@]2(CCCC3)C1
AIEUPE	Cc1ccc([C@H]2C[C@H](C)CCN2Cc2cc(=O)c(O)co2)o1
CUDILU	C=C(C)C(=O)OC[C@H]1CSC(=S)O1
AAPEZI	CN(C)CCCN1Cc2ccccc2Sc2ccccc21
AADHII	Cc1nc(C(N)=O)c2c(n1)n(-c1ccccc1C(F)(F)F)c(=O)n2CC(N)=O
AABGNU	COC[C@H]1CN(C)C[C@@]12CCN(C(=O)c1ccnn1C)C2
AAIZQU	Cc1ccc(-c2[nH]ncc2C(=O)O)c(C)c1
HINUXI	CC[C@@H](C)CN1CCc2ccc(C(=O)O)cc2C1
AELRIO	CC(C)NC(=O)C1C[C@H]2OCCN(Cc3ccco3)[C@@H]2C1
DEMRYO	O=c1[nH]c(=O)c2cc([N+](=O)[O-])cc([N+](=O)[O-])c2[nH]1
BOLEUI	CC(C)c1nc2nc(C(F)(F)F)nn2c(N)c1-c1ccccc1
DAIJTE	Cc1cc(C(=O)Nc2cccc(C#Cc3ccccn3)c2)c(C)n1-c1nccs1

Table A.13 Consolidated list of hits obtained from USRCAT (continued)

Molecule ID	SMILES
AUJLKU	C[C@H]1N[C@H]2CC(C)(C)CC[C@@H]2c2c1[nH]c1ccccc21
AEFYZA	Cc1nn(C)c2c1=C1CC(C)(C)CC(=O)C1[C@H](c1ccc(N(C)C)cc1)N=2
AOHTSI	CCN1C(=O)/C(=C2\SC(=S)NC2=O)c2cc([N+](=O)[O-])ccc21
IOFUYI	O=C1CC2(Oc3ccccc3O2)C(=O)NN1
HUBZUO	CN[C@H](CN1CCN(C)CC1)c1ccccc1
CIIGJU	Cc1ccc2oc(-c3nccn3CC3(CO)CCOCC3)c(C)c2c1
EUYRWO	O=S(=O)(c1ccc(Br)s1)N1Cc2cncnc2C1
IUOYXU	Cc1c([N+](=O)[O-])c2onc(C)c2[n+]([O-])c1C
AELMHU	C[C@@]12CCC(O1)C(C(=O)O)C2C(=O)O
CEMHDA	OCCOc1c(Cl)c(Cl)c(Cl)c(Cl)c1Cl
AIUMRA	Cc1c(C(=O)N[C@@H](C#N)C2CC2)cc(Br)cc1[N+](=O)[O-]
KEHWRE	Brc1ccc(C2=NN=C(c3ccccc3)Nc3cccnc32)cc1
IUYIFE	COC(=O)C1CCCN(C(=O)c2occc2C(C)C)C1C(=O)OC
GIMDFO	Cc1ccc([C@@H]2CSCCN2C(=O)c2[nH]nc(C)c2[N+](=O)[O-])o1
AOGAYE	CN(C)c1ccc(CNc2nccc3c2C(=O)OC3(C)C)cc1F
HUVGIA	Cc1nn(C(=O)c2c[nH]c3ccccc23)c(N)c1-c1ccccc1
AACNMA	CC(C)(C)C(=O)N(c1ccccc1)S(=O)(=O)c1ccccc1
AAVVVO	O=C(NOCc1cccnc1)N1CCc2c(Cl)cccc21
AAURLO	C[C@H]1CO[P@](=O)(C2(Nn3cnnc3)CCCCC2)O1
AIBAEO	C=c1[nH]n(CCC#N)c(=O)c1=Cc1cc(Br)c(O)c(OC)c1
EIRDRU	CCOC(=O)c1c(C)oc2[nH]c(CN3CCC(C)CC3)nc(=O)c12
EIIOJO	NC(=O)c1ccc2c(ccn2[C@@H]2CN(C(=O)/C=C/C(=O)O)C[C@H]2O)c1
BIIWJE	CCc1ccc(-c2csc3nc(C)cc(C)[n+]23)cc1
AAEPKU	CC1(C)C(=O)NCCN1C[C@@]1(O)CCCN(CC2CCCCC2)C1=O
HOYSME	CC1=C2CCCCCCCCCCC2C(=O)C1
BEXYDE	CC(C)(C)c1ccc(C(O)C(O)Nc2ccc3nc(=O)cccc3c2)cc1
DIESLA	O=C1C=C(c2ccccc2)O[C@@]1(Nc1ccccc1)c1ccccc1
EOPUHO	CC(C)Nc1cc(-c2ccccc2)nc2ccccc12
AEHGOI	Cc1cc(C)nc(N(C)C[C@@H]2CC(=O)N(C3CCCC3)C2)n1
HUSXAA	O=C1CCc2cc3c(cc2O1)OC(=O)CC3
JIJBLU	Cc1cc(=O)c2c(O)c3c(=O)cc(O)c4c3c3c(cc(O)c1c23)C(C)(C)C=4O
AOJXLA	COC(=O)[C@](F)(Sc1ccc(O)cc1)C(F)(F)F
BEJKHU	CCn1nc(C)cc1S(=O)(=O)N1CC[C@H](N)C1
LEYLQE	CN(C)C(N(C)C)=[N+]1N=[N+](O)c2cc(Cl)ccc21
BAXWRO	Cn1c(=O)n(C)c2cc3c(cc21)NC(=O)CS3
DASYZU	CC1(C)COC(C)(c2cc(N)ccc2C(=O)O)OC1
AOPRLA	Fc1ccc(COc2ncnc3c2cnn3-c2ccc(F)cc2)cc1
DERJYI	O=C(O)CCC=C1CCC=C2CC(=O)CCC12
BALNRA	N#CC1(c2ccc(C(F)(F)F)cn2)CCC1
BAWNLI	Cc1cccc(/C=N\N=C/c2cccc(C)n2)n1
AARIDI	COc1ccc(-c2nc(N(C)C)ns2)c(-n2nc(C)cc2C)c1
BORQFU	Oc1cc(Cl)c(Cl)nc1I
GOTNII	CCc1ccccc1NC(=O)Nc1cc(N2CCN(C)CC2)ncn1
DEKMKO	O=C1NCCc2nc(Cl)nc(Cl)c21
JUHRNO	Nc1nc2c([N+](=O)[O-])ccc(Br)n2n1
CUSUUE	CC1=CC[C@]23C[C@]45CC(C)=CC[C@]4(C[C@@]2(C1)O3)O5
BIGFCE	CS(=O)(=O)C1=NNc2c(-c3ccccc3)cnn2[C@]12CC=NN2
FUORHU	CC(C)CC1CC(C(=O)O)(C(=O)O)C1
BESXSA	Cc1cc(Cl)cc(C(=O)C(C#N)c2nc(C)cc(C)n2)c1
CAJENU	Cc1ccc(N[C@@H](C)c2ccc(C#N)cc2)cc1C(N)=O
BAPRQA	CN(Cc1nn(C)c2c1CNCC2)C(=O)OC(C)(C)C
KEEEMA	COC(=O)c1ccc2c(c1)CCN2Cc1cc[nH]n1
AEOQQI	O=C1O[C@H]2[C@@H](O1)[C@]1(Cl)C(Cl)=C(Cl)[C@]2(Cl)C1(Cl)Cl

Table A.14 Consolidated list of hits obtained from USRCAT (continued)

Molecule ID	SMILES
DOOKAO	O=C(O)C1=C(C2c3ccc(O)c(O)c3Oc3c2ccc(O)c3O)CCCC1
HERSRI	N=C1CC=CN2C=CC=C[C@@H]2c2ccccc21
DUHWBO	N#Cc1ccc(-n2cccc2)c(CS(=O)(=O)c2ccc(Cl)cc2)c1
AAVNQU	c1cncc([C@H]2C3C(=Nc4ncnn42)c2ccccc2O[C@H]3c2cccs2)c1
BIUWVI	Cc1nc(Nc2ccc(F)cc2)nc2c1ncn2-c1ccc(F)cc1
AURXII	CC[C@@]1(C)NC(=O)N(Cc2c(C3CCC3)cnn2C)C1=O
AEDHAE	N#Cc1nc(-c2ccccc2)oc1S(=O)(=O)N1CCC(C(N)=O)CC1
AIAVGI	CNc1cc(C)nc(C2CCN(CC3(O)CCOCC3)CC2)n1
AETQZA	CC(C)[C@H]1N(c2ccc(F)cc2)C(=O)C(C)N1C(=O)C(F)(F)F
AAVBOU	CN(C)C[C@@]12CNC[C@@H]1CN(c1ncc(Cl)cc1F)C2
BABFYO	C#CCn1cc(F)c(=O)n(CC#C)c1=O
CINRSI	CC(C)(C(=O)O)c1cc(F)cc(C#N)c1
FOCEAU	Cc1cncc(C(=O)N2C[C@H](C)[C@]3(CCNC3=O)C2)c1

Appendix II – List of reviewed studies

study_id	study_ref	study_doi
001	Hung, H.C., Ke, Y.Y., Huang, S.Y., Huang, P.N., Kung, Y.A., Chang, T.Y., Yen, K.J., Peng, T.T., Chang, S.E., Huang, C.T. and Tsai, Y.R., 2020. Discovery of M protease inhibitors encoded by SARS-CoV-2. Antimicrobial agents and chemotherapy, 64(9), pp.e00872-20.	https://doi.org/10.1128/AAC.00872-20
002	David, A.B., Diamant, E., Dor, E., Barnea, A., Natan, N., Levin, L., Chapman, S., Mimran, L.C., Epstein, E., Zichel, R. and Torgeman, A., 2021. Identification of SARS-CoV-2 Receptor Binding Inhibitors by In Vitro Screening of Drug Libraries. Molecules, 26(11), p.3213.	https://doi.org/10.3390/molecules26113213
003	David, A.B., Diamant, E., Dor, E., Barnea, A., Natan, N., Levin, L., Chapman, S., Mimran, L.C., Epstein, E., Zichel, R. and Torgeman, A., 2021. Identification of SARS-CoV-2 Receptor Binding Inhibitors by In Vitro Screening of Drug Libraries. Molecules, 26(11), p.3213.	https://doi.org/10.3390/molecules26113213
004	Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C. and Duan, Y., 2020. Structure of M pro from SARS-CoV-2 and discovery of its inhibitors. Nature, 582(7811), pp.289-293.	https://doi.org/10.1038/s41586-020-2223-y
005	Frediansyah, A., Tiwari, R., Sharun, K., Dhama, K. and Harapan, H., 2021. Antivirals for COVID-19: A critical review. Clinical Epidemiology and global health, 9, pp.90-98.	https://doi.org/10.1016/j.cegh.2020.07.006
006	Yamamoto, M., Kiso, M., Sakai-Tagawa, Y., Iwatsuki-Horimoto, K., Imai, M., Takeda, M., Kinoshita, N., Ohmagari, N., Gohda, J., Semba, K. and Matsuda, Z., 2020. The anticoagulant nafamostat potently inhibits SARS-CoV-2 S protein-mediated fusion in a cell fusion assay system and viral infection in vitro in a cell-type-dependent manner. Viruses, 12(6), p.629.	https://doi.org/10.3390/v12060629
007	Yao, X., Ye, F., Zhang, M., Cui, C., Huang, B., Niu, P., Liu, X., Zhao, L., Dong, E., Song, C. and Zhan, S., 2020. In vitro antiviral activity and projection of optimized dosing design of hydroxychloroquine for the treatment of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Clinical infectious diseases, 71(15), pp.732-739.	https://doi.org/10.1093/cid/ciaa237
008	Manandhar, A., Srinivasulu, V., Hamad, M., Tarazi, H., Omar, H., Colussi, D.J., Gordon, J., Childers, W., Klein, M.L., Al-Tel, T.H. and Abou-Gharbia, M., 2021. Discovery of Novel Small-Molecule Inhibitors of SARS-CoV-2 Main Protease as Potential Leads for COVID-19 Treatment. Journal of Chemical Information and Modeling, 61(9), pp.4745-4757.	https://doi.org/10.1021/acs.jcim.1c00684
009	Manandhar, A., Srinivasulu, V., Hamad, M., Tarazi, H., Omar, H., Colussi, D.J., Gordon, J., Childers, W., Klein, M.L., Al-Tel, T.H. and Abou-Gharbia, M., 2021. Discovery of Novel Small-Molecule Inhibitors of SARS-CoV-2 Main Protease as Potential Leads for COVID-19 Treatment. Journal of Chemical Information and Modeling, 61(9), pp.4745-4757.	https://doi.org/10.1021/acs.jcim.1c00684
010	Cuesta, S.A., Mora, J.R. and Márquez, E.A., 2021. In Silico Screening of the DrugBank Database to Search for Possible Drugs against SARS-CoV-2. Molecules, 26(4), p.1100.	https://doi.org/10.3390/molecules26041100
011	Pitsillou, E., Liang, J., Karagiannis, C., Ververis, K., Darmawan, K.K., Ng, K., Hung, A. and Karagiannis, T.C., 2020. Interaction of small molecules with the SARS-CoV-2 main protease in silico and in vitro validation of potential lead compounds using an enzyme-linked immunosorbent assay. Computational biology and chemistry, 89, p.107408.	https://doi.org/10.1016/j.compbiolchem.2020.107408
012	Gangadevi, S., Badavath, V.N., Thakur, A., Yin, N., De Jonghe, S., Acevedo, O., Jochmans, D., Leyssen, P., Wang,	https://doi.org/10.1021/acs.jpclett.0c03119

	K., Neyts, J. and Yujie, T., 2021. Kobophenol a inhibits binding of host ace2 receptor with spike rbd domain of sars-cov-2, a lead compound for blocking covid-19. The journal of physical chemistry letters, 12(7), pp.1793-1802.	
013	Klemm, T., Ebert, G., Calleja, D.J., Allison, C.C., Richardson, L.W., Bernardini, J.P., Lu, B.G., Kuchel, N.W., Grohmann, C., Shibata, Y. and Gan, Z.Y., 2020. Mechanism and inhibition of the papain-like protease, ppro, of SARS-CoV-2. The EMBO journal, 39(18), p.e106275.	https://doi.org/10.15252/embj.2020106275
014	Day, C.J., Bailly, B., Guillon, P., Dirr, L., Jen, F.E.C., Spillings, B.L., Mak, J., von Itzstein, M., Haselhorst, T. and Jennings, M.P., 2021. Multidisciplinary Approaches Identify Compounds that Bind to Human ACE2 or SARS-CoV-2 Spike Protein as Candidates to Block SARS-CoV-2–ACE2 Receptor Interactions. Mbio, 12(2), pp.e03681-20.	https://doi.org/10.1128/mBio.03681-20
015	Forrestall, K.L., Burley, D.E., Cash, M.K., Pottie, I.R. and Darvesh, S., 2021. 2-Pyridone natural products as inhibitors of SARS-CoV-2 main protease. Chemico-biological interactions, 335, p.109348.	https://doi.org/10.1016/j.cbi.2020.109348
016	Gaudêncio, S.P. and Pereira, F., 2020. A computer-aided drug design approach to predict marine drug-like leads for SARS-CoV-2 main protease inhibition. Marine drugs, 18(12), p.633.	https://doi.org/10.3390/md18120633
017	Ogedjo, M., Onoka, I., Sahini, M. and Shadrack, D.M., 2021. Accommodating receptor flexibility and free energy calculation to reduce false positive binders in the discovery of natural products blockers of SARS-COV-2 spike rbac interface. Biochemistry and biophysics reports, p.101024.	https://doi.org/10.1016/j.bbrep.2021.101024
018	Bung, N., Krishnan, S.R., Bulusu, G. and Roy, A., 2021. De novo design of new chemical entities for SARS-CoV-2 using artificial intelligence. Future medicinal chemistry, 13(06), pp.575-585.	https://doi.org/10.4155/fmc-2020-0262
019	Kumar, S., Sharma, P.P., Shankar, U., Kumar, D., Joshi, S.K., Pena, L., Durvasula, R., Kumar, A., Kempaiah, P., Poonam and Rathi, B., 2020. Discovery of new hydroxyethylamine analogs against 3CLpro protein target of SARS-CoV-2: Molecular docking, molecular dynamics simulation, and structure–activity relationship studies. Journal of chemical information and modeling, 60(12), pp.5754-5770.	https://doi.org/10.1021/acs.jcim.0c00326
020	Liu, X. and Wang, X.J., 2020. Potential inhibitors against 2019-nCoV coronavirus M protease from clinically approved medicines. Journal of Genetics and Genomics, 47(2), p.119.	https://dx.doi.org/10.1016%2Fj.jgg.2020.02.001
021	Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C. and Duan, Y., 2020. Structure of M pro from SARS-CoV-2 and discovery of its inhibitors. Nature, 582(7811), pp.289-293.	https://doi.org/10.1038/s41586-020-2223-y
022	Manandhar, A., Srinivasulu, V., Hamad, M., Tarazi, H., Omar, H., Colussi, D.J., Gordon, J., Childers, W., Klein, M.L., Al-Tel, T.H. and Abou-Gharbia, M., 2021. Discovery of Novel Small-Molecule Inhibitors of SARS-CoV-2 Main Protease as Potential Leads for COVID-19 Treatment. Journal of Chemical Information and Modeling, 61(9), pp.4745-4757.	https://doi.org/10.1021/acs.jcim.1c00684
023	Tejera, E., Munteanu, C.R., López-Cortés, A., Cabrera-Andrade, A. and Pérez-Castillo, Y., 2020. Drugs repurposing using QSAR, docking and molecular dynamics for possible inhibitors of the SARS-CoV-2 mpro protease. Molecules, 25(21), p.5172.	https://doi.org/10.3390/molecules25215172
024	Gurung, A.B., Ali, M.A., Lee, J., Farah, M.A. and Al-Anazi, K.M., 2021. Identification of potential SARS-CoV-2 entry inhibitors by targeting the interface region between the spike RBD and human ACE2. Journal of infection and public health, 14(2), pp.227-237.	https://doi.org/10.1016/j.jiph.2020.12.014

025	Puttaswamy, H., Gowtham, H.G., Ojha, M.D., Yadav, A., Choudhir, G., Raguraman, V., Kongkham, B., Selvaraju, K., Shareef, S., Gehlot, P. and Ahamed, F., 2020. In silico studies evidenced the role of structurally diverse plant secondary metabolites in reducing SARS-CoV-2 pathogenesis. Scientific reports, 10(1), pp.1-24.	https://doi.org/10.1038/s41598-020-77602-0
026	Cuesta, S.A., Mora, J.R. and Márquez, E.A., 2021. In Silico Screening of the DrugBank Database to Search for Possible Drugs against SARS-CoV-2. Molecules, 26(4), p.1100.	https://doi.org/10.3390/molecules26041100
027	Pitsillou, E., Liang, J., Karagiannis, C., Ververis, K., Darmawan, K.K., Ng, K., Hung, A. and Karagiannis, T.C., 2020. Interaction of small molecules with the SARS-CoV-2 main protease in silico and in vitro validation of potential lead compounds using an enzyme-linked immunosorbent assay. Computational biology and chemistry, 89, p.107408.	https://doi.org/10.1016/j.compbiolchem.2020.107408
028	Pitsillou, E., Liang, J., Ververis, K., Hung, A. and Karagiannis, T.C., 2021. Interaction of small molecules with the SARS-CoV-2 papain-like protease: In silico studies and in vitro validation of protease activity inhibition using an enzymatic inhibition assay. Journal of Molecular Graphics and Modelling, 104, p.107851.	https://doi.org/10.1016/j.jmgm.2021.107851
029	Reyaz, S., Tasneem, A., Rai, G.P. and Bairagya, H.R., 2021. Investigation of structural analogs of hydroxychloroquine for SARS-CoV-2 main protease (mpro): A computational drug discovery study. Journal of Molecular Graphics and Modelling, 109, p.108021.	https://doi.org/10.1016/j.jmgm.2021.108021
030	Gangadevi, S., Badavath, V.N., Thakur, A., Yin, N., De Jonghe, S., Acevedo, O., Jochmans, D., Leyssen, P., Wang, K., Neyts, J. and Yujie, T., 2021. Kobophenol a inhibits binding of host ace2 receptor with spike rbd domain of sars-cov-2, a lead compound for blocking covid-19. The journal of physical chemistry letters, 12(7), pp.1793-1802.	https://doi.org/10.1021/acs.jpclett.0c03119
031	Basu, S., Veeraraghavan, B., Ramaiah, S. and Anbarasu, A., 2020. Novel cyclohexanone compound as a potential ligand against SARS-CoV-2 main-protease. Microbial Pathogenesis, 149, p.104546.	https://doi.org/10.1016/j.micpath.2020.104546
032	Day, C.J., Bailly, B., Guillon, P., Dirr, L., Jen, F.E.C., Spillings, B.L., Mak, J., von Itzstein, M., Haselhorst, T. and Jennings, M.P., 2021. Multidisciplinary approaches identify compounds that bind to human ACE2 or SARS-CoV-2 spike protein as candidates to block SARS-CoV-2–ACE2 receptor interactions. MBio, 12(2), pp.e03681-20.	https://doi.org/10.1128/mBio.03681-20