UNIVERSITAT ROVIRA i VIRGILI

FiBeFTa: Finding the best fingerprint for discerning between actives and decoys: application to the SARS-CoV-2 main protease

Oriol Villaró Serrano

BIOTECHNOLOGY FINAL DEGREE PROJECT

Academic tutor: Santiago Garcia-Vallve
Biochemistry and Biotechnology Department, URV
santi.garcia-vallve@urv.cat

Supervisor: Gerard Pujadas Anguiano
Biochemistry and Biotechnology Department, URV
gerard.pujadas@urv.cat

Table of contents

Abstract

Molecular fingerprints have been used regularly in virtual screening and drug discovery. Molecular fingerprints combine the results of more complex techniques, paired with the efficiency that comes with binary data structures. However, there are different types with a huge variety within them, that each results in a different fingerprint for the same molecule. This project aims to develop a tool to compare 10 different fingerprints, and use it to rank how each of them performs according to 4 different metrics. In ordero to achieve this, we will use this tool on one of the common benchmarks that exist in cheminformatics, the Directory of Useful Decoys.

Key words

Molecular fingerpint, Virtual screening, Bioinformatics, Directory of Useful Decoys

Introduction

Virtual screening

Finding new drugs in chemical databases is an essential step in any project that aims to discover and develop new drugs. There are two ways for achieving this goal: experimentally testing compound libraries to find molecules that show the desired bioactivity against specific targets (a process known as high throughput screening; HTS); and computationally predicting the bioactivity of interest in files containing molecular databases (known as virtual screening; VS). HTS uses robotic arms, liquid handling peripherics and control software that allows for rapid testing of chemical libraries for biological activity. Technological hardware like robotic plate handling is critical for the procedure and reagents like antibodies and recombinant proteins can also make the whole procedure very expensive [1]. On the other hand, VS provides a more affordable approach for the research of new bioactive compounds by using a set of computational techniques that are able to predict new bioactive compounds in databases of small chemical compounds [2]. These small molecules are predicted to complement the binding site of a particular target molecule in terms of shape, electrostatic surface, hydrophobicity and spatial location of hidrogen-bond donors/acceptors pairs [3]. As the result of the formation of this complex it is expected that the bioactivity of the target becomes either inhibited or stimulated by the bound ligand [2].

The specific VS technique applied depends on the amount of information available for the particular target, typically a protein. If the three-dimensional structure of the target is available, structure-based virtual screening (SBVS) techniques (typically, protein-ligand docking) are usually applied. For example, protein-ligand docking algorithms (one of the most applied applied VS methods) aim to predict the 3D structure of a protein-ligand complex given the individual structure of the protein and the ligand. Thus, docking methods fit the ligand into the target binding site by combining and optimizing variables like steric, hydrophobic and electrostatic complementarity and estimate the binding free energy by means of the so-called docking score [4]. Usually, protein-ligand docking algorithms have two parts (*i.e.*, an initial sampling algorithm that explores the possible conformations of the ligand at the target binding

site; and a second part that ranks these ligand conformation by decreasing predicted affinity for the target) [4].

When the 3D structure of the target is not known, then ligand-based virtual screening (LBVS) techniques can be applied. In LBVS, a database of molecules is searched looking for those that most closely resemble to an active molecule that is acting as a template during the search [5]. For instance, using a set of active molecules against the same target, it is possible to develop a pharmacophore (*i.e.*, a spatial arrangement of functional groups that allow the bioactive conformation of each ligand to interact with the binding site of a target protein) by: **(1)** sampling the conformational space of the active molecules; **(2)** looking for sets of conformations (one conformer per active molecule [6]) that share the 3D location of different functional groups that are expected to be used for the intermolecular interaction with the target. Then this pharmacophore can be used to look in databases of molecular conformers for similar spatial arrangement of pharmacophoric features [7]. Apart from using ensembles of ligand conformers to obtain pharmacophores, they can also be directly obtained from known 3D structures of drug/target complexes [8].

Other computational strategies that can be used indistinctly in SBVS or in LBVS are shape similarity (SS) or electrostatic similarity (ES) methods. In the case of SBVS, both methodologies use experimental conformers (the so-called experimental *poses*) as the query molecules. In SS, the conformations from a database of chemical compounds is compared with the known (or expected) bioactive conformation of an active in order to find which of the molecules from the database have a global molecular shape that is similar to the one from the query [9]. ES, that is usually used together with SS to improve the VS results, compares the electrostatic surface from the query active with the ones from the conformers at the molecular database in order to identify those compounds with the greatest electrostatic similarity with the known active [10].

Molecular Fingerprints

Molecular fingerprints represent structural information of a molecule as binary arrays that facilitates the comparison between molecular pairs. Thus, in two-dimensional fingerprints, each molecule is encoded as a binary vector characterizing the absence or presence of

specific properties of its 2D structure [11]. Depending on the information that these bits represent, different types of molecular fingerpints can be differentiated. For instance, key-based substructural fingerprints are based on the occurrences of predefined chemical groups —“keys”, and are encoded as a bit string that can be easily analyzed. The number of bits is determined by the number of structural keys [12]. On the other hand, path-based, or topological, fingerprints index all the fragments of the molecule following a path up to a set number of bonds and then, these indexes are hashed to create a fingerprint. Unlike in structural fingerprints, the presence or absence of a bit cannot be tracked to a single feature of the molecule [13]. Finally, circular fingerprints are another category of hashed topological fingerprints that use the environment of each atom up to a set radius to generate the corresponding fingerprint. Fingerprints can also be used in VS workflows by comparing the ones from known actives with those from the molecules at the database that is being screened.

Measuring molecular similarity with fingerprints

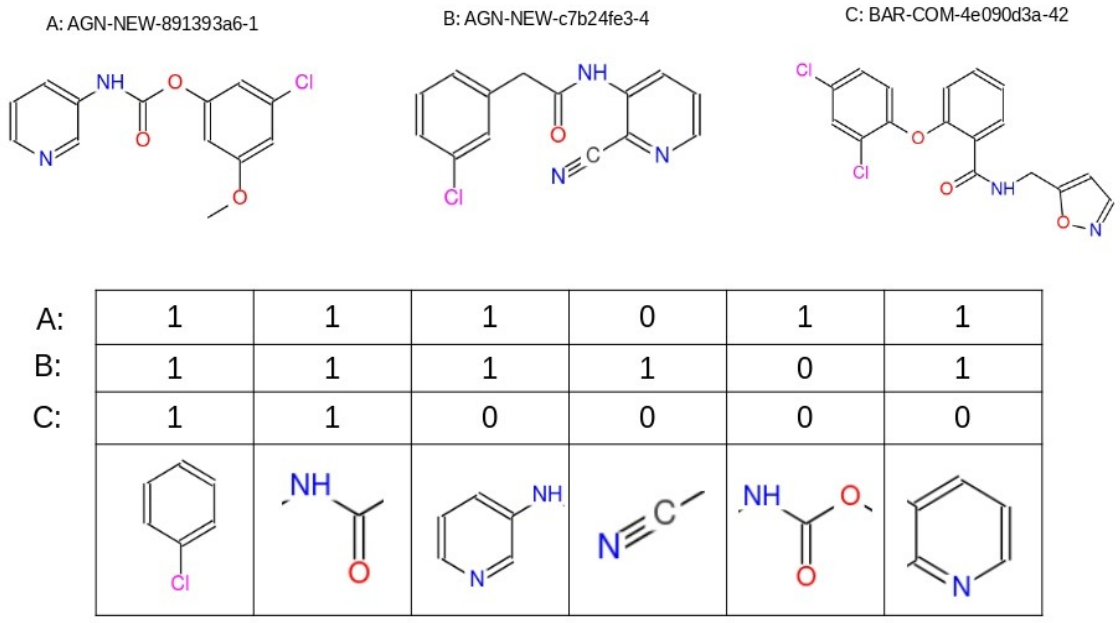
The Tanimoto similarity coefficient is a statistic used to measure the similarity and diversity of sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. By design, the Tanimoto coefficient has a value between 0 and 1, both included. The Tanimoto coefficient is represented by the following mathematical equation:

$$T(a,b)=\frac{N_c}{N_a+N_b+N_c}$$

In this equation, **N** represents the number of distinct attributes in each object (**a**, **b**). In this case, **c** corresponds to the intersection set.

In Figure 1 we can see an example of how a substructure fingerprint works, and how to calculate the Tanimoto coefficient with it by using 3 different molecules. To calculate the Tanimoto coefficient between the molecules we need to know the number of bits set in both (N_c), as well as the number of bits set in one but not in the other (N_a and N_b). In this case:

- T(A, B) = 4 / (1 + 1 + 4) = 4 / 6 = 0.67
- T(A, C) = 2 / (3 + 0 + 2) = 2 / 5 = 0.40



A:	1	1	1	0	1	1
B:	1	1	1	1	0	1
C:	1	1	0	0	0	0

Figure 1: Example 6-substructure key fingerprint visualization with Tanimoto coefficient example

Validation of Virtual Screening

Evaluating the performance of VS methods is a necessary practice for both the method developers and the end-users. For the developers, it is done to parametrize and validate the methods, while for the end-users it's a way to select which method performs best in a given situation. The validation of a VS is based on its capacity for discerning between known actives and decoys (compounds that are presumably inactive against the target of the known actives).

Different evaluation metrics exist to measure the performance of a VS method. The results are reported as a single number that estimates the ability of a VS method to retrieve active compounds out of a mixed set of active compounds and decoys. In the case of similarity searching, predicting the bioactivity of compounds is more difficult than quantifying other properties, such as solubility or log P [14,15]. The subtle geometrical or functional differences

between two very similar compounds can reduce or enhance binding affinity and bioactivity in a meaningful manner [16]. The most frequently used metrics for evaluating the performance of a VS are the Enrichment Factor (EF), the Area under the Curve (AUC) and the Boltzmann-Enhanced Discrimination of ROC (BEDROC).

The Enrichment Factor (EF) denotes the quocient of true actives among a subset of predicted actives and the overall fraction of actives. In other words, the EF is the concentration of active molecules among the top-scoring hits compared to their concentration throughout the entire database [17]. EF with a factor of 1 or a factor of 10 are the most commonly used to evaluate VS performance. This means surveying the number of top-scoring hits caused by active molecules in the top 1% and 10% of the list of molecules respectively.

$$EF(X) = \frac{\frac{actives \text{ in } top-ranked\ X\ \%}{number\ of\ molecules \text{ in } the\ X\ \%}}{\frac{total\ actives}{total\ molecules}}$$

Receiver operating characteristic (ROC) curve is a tool for characterizing and comparing the diagnostic accuracy of binary classifier systems. The ROC curve is generated by plotting its true positive rate against its false positive rate at various thresholds [18]. The area under the ROC curve analysis metric (AUC) provides a quantitative measure for the discrimination ability of a VS method. AUC value range from 0 to 1, with 1 being a perfect classification where all the active compounds are ranked before the decoys, and anything lower than 0.5 is treated as a bad prediction and similar to a random classification. In Figure 2 we can see an example of these ROC curves, and the AUC that each of them determines. The red dashed line represents a classifier that decides the label of the molecule at random. The closer the ROC curve gets to the top-left part of the chart, the better classifier it its. As such, the closer the AUC gets to 1 the better that VS scores. In the case of the Figure 2, the best-performing VS would be represented by the blue line.

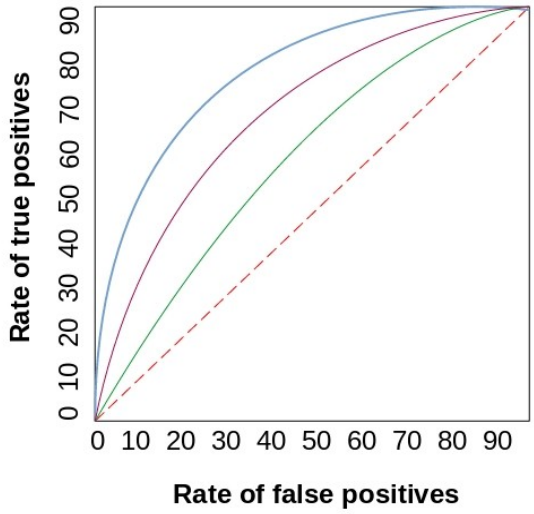


Figure 2: Comparison of different ROC curves

The Boltzmann-Enhanced Discrimination of ROC (BEDROC) score [19] is a metric that assigns more weight to early ranked compounds than late ranked compounds. The active compounds are classified depending on their position using an exponential factor, ranging from 1 for the top ranked compound to close to 0 for the lowest ranked compound. BEDROC is recognized as a proper and robust evaluation measure for early discovery, making up for some of the limitations that the AUC presents, like late-recognition [20]. BEDROC values can be interpreted as the probability that an active molecule being ranked better than a molecule selected at random from an exponential probability distribution function of parameter (α).

Hypothesis

A lot of research has been done in the subject of VS in order to identify promising compounds in drug discovery. VS is useful since it saves time and resources that would be otherwise wasted by testing molecules with a low chance of presenting the bioactivity desired in the target molecule. For VS to be possible, molecular information must be converted to a format that a computer can interpret and process. One of such methods are molecular fingerprints, but the same molecule is differently encoded depending on the fingerprint that is being used and, in the same way that not all fingerprints

have not the same performance when trying to discern between a set of active and decoys molecules that are mixed. Here, **we hypothesize that developing a tool for calculating the best fingerprint for one specific target and using such fingerprint during the VS can improve its performance for that target.**

Objectives

The overall objectives of this project are the following:

1. Develop a tool (so called FiBeFTa; ***Fi****nding the* ***Be****st* ***Fi****ngerprint for* ***Ta****rget*) that takes as input a file with the structures of active molecules and another file with the structures of the decoys for a given target and process that information to calculate which of the fingerprints analyzed allows a better separation of both groups.
2. Use FiBeFTa to evaluate the targets in the Directory of Useful Decoys (DUD) [17].
3. Use the SARS-CoV-2 main protease (M-pro) as a relevant target to observe the functionality of the algorithm.

The long-term objectives of he current work are the following:

1. Convert FiBeFTa into a web-service, in order to make it available and contribute to other VS projects.

Methodology

Targets

In this project, the performance of the tool that has been developed was analyzed with 41 targets [40 from the Directory of Useful Decoys (DUD) and the SARS-CoV-2 M-pro].

DUD contains a total of 2.950 active compounds for the 40 targets and 36 decoys per active molecule. The decoys for each target have similar physical properties (*e.g.*, molecular weight, calculated LogP) but dissimilar molecular topology than the corresponding actives. The files with the sets of active/decoy molecules can be found in the DUD website organized by target (http://dud.docking.org/).

Table A: Number of ligands and decoys for the forty targets in DUD. When the number of actives or decoys analyzed in this work is lower than the total number available in DUD, the original number is reported between brackets.

Protein	PDB code	No. of ligands	No. of decoys
Nuclear Hormone Receptors			
AR – Androgen receptor	1xq2	79	2854
$ER_{agonist}$ – Estrogen receptor	1l2i	67	2570
$Er_{antagonist}$ – Estrogen receptor alpha	3ert	39	1448
GR – Glucocorticoid receptor	1m2z	78	2947
MR – Mineralocorticoid receptor	2aa2	15	636
PPArg – Peroxisome proliferator activated receptor gamma	1fm9	85	3127
PR – Progesterone receptor	1sr7	27	1041
RXRa – Retinoic X receptor alpha	1mvc	20	750
Kinases			
CDK2 – Cyclin-dependent protein kinase 2	1ckp	72	2073 (2074)
EGFr – Epidermal growth factor receptor	1m17	475	15996
FGFr1 – Fibroblast growth factor receptor 1	1agw	120	4550
HSP90 – Human heat shock protein 90	1uy6	37	979
P38 MAP – p38 mitogen activated protein	1kv2	454	9141
PDGFrb – Platelet derived growth factor receptor kinase	model	170	5980
SRC – Tyrosine kinase SRC	2src	152 (159)	6319
TK – Thymidine kinase	1kim	22	891
VEGFr2 – Vascular enothelial growth factor receptor	1vr2	88	2906
Serine Proteases			
Fxa – Coagulation factor Xa	1f0r	146	5743 (5745)

Protein	PDB code	No. of ligands	No. of decoys
Thrombin	1ba8	72	2456
Trypsin	1bju	49	1664
Metalloenzymes			
ACE – Angiotensin-converting enzyme	1o86	49	1797
ADA – adenosine deaminase	1ndw	39	927
COMT – Catechol-O-methyltransferase	1h1d	11	468
PDE5 – Phosphodiesterase 5	1xp0	88	1978
Folate Enzymes			
DHFR – Dihydrofolate reductase	3dfr	410	8364 (8367)
GART – Glycinamide ribonucleotide transformylase	1c2t	40	879
Other Enzymes			
AChE – Acetylcholinesterase	1eve	107	3892
ALR2 – Aldose reductase	1ah3	26	995
AmpC – AmpC Beta-lactamase	1xgj	21	786
COX-1 – Cyclooxygenase-1	1q4g	25	911
COX-2 – Cyclooxygenase-2	1cx2	426	13289
GPB – Glycogen phosphorylase beta	1a8i	52	2134 (2140)
HIVPR – HIV protease	1hpx	62	2038
HIVRT – HIV reverse transcriptase	1rt1	43	1519
HMGA – Hydroxymethylglutaryl-CoA reductase	1hw8	35	1480
InhA – Enoyl ACP reductase	1p44	79 (86)	3266
NA – Neuraminidase	1a4g	49	1873 (1874)
PARP – Poly (ADP-ribose) polymerase	1efy	35	1351
PNP – Purine nucleoside phosphorylase	1b80	50	1036
SAHH – S-adenosyl-homocysteine hydrolase	1a7a	33	1344 (1346)

The forty first target analyzed in this work is the SARS-CoV-2 M-pro (see Figure 3). M-pro is a key enzime of coronaviruses and has a pivotal role in mediating viral replication and transcription, making it an attractive drug target for COVID-19 treatment [21].

The number of actives for M-pro consist on 81 active molecules obtained from the COVID Moonshot initiative whereas the decoy file was created by using DecoyFinder [22]. DecoyFinder is a graphical tool able to find sets of decoy molecules for a given set of active ligands. It can do so with two different methods:

- By finding molecules with a molecular weight similar to the actives.

- By finding molecules which have a similar number of rotational bonds, hydrogen bond acceptors, hydrogen bond donors, logP value and molecular weight, but are chemically different (molecular descriptor based decoys).

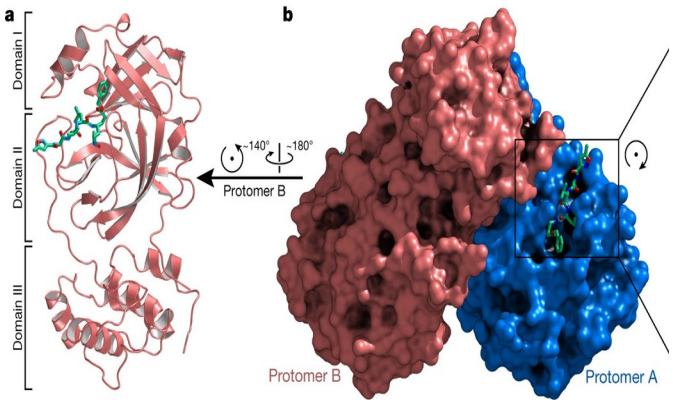


*Figure 3: Crystal structure of SARS-CoV2 M-pro in a covalent complex with N3 [21]. Panel **a** shows the cartoon representation of one protomer of the dimeric M-pro whereas panel b shows the surface representation of the M-pro homodimer (with subunit A in blue and subunit B in salmon). The covalently bound N3 ligand is shown in green.*

In this work we use the molecular-weight based decoys method to find 2.916 decoy molecules (36 decoys per active molecule) from the ZINC database [23,24].

Fingerprint generation

Fingerprint generation and similarity search was perfomed with Chemfp [25]. Chemfp is a set of command-line tools and a Python library for fingerprint generation and high-performance search. Apart from this high-performance, that enables for fast searches and fingerprint generation, Chemfp also implements a standard file format (.fps) and performs high-speed Tanimoto similarity searches.

Chemfp supports the fingerprints available for the toolkits from Open Babel[1], OpenEye[2] and RDKit[3], of which only Open Babel and RDKit will be discussed here, since they both have a permanent open source license for academic purposes. In order to use both toolkits some active/decoy molecules from

1 **Open Babel: The Open Source Chemistry Toolbox -** http://openbabel.org/wiki/Main_Page
2 **OpenEye Scientific -** https://www.eyesopen.com/
3 **RDKit: Open-Source Cheminformatics Software -** https://rdkit.org/

the original DUD sets had to be removed for some of the targets because some nitrogen atoms in the structure of these molecule had too many bonds and a valid Lewis dot structure could not be drawn. In that case, RDKit could not interpret the structural information of these molecules without correcting the valences of these atoms. Therefore, considering that this problem affects very few molecules, we decided to remove them from the sample (see Table 1). As we can see in Table 1, this issue affected almost exclusively the decoys files, with the exception of 7 actives for the protein Enoyl ACP reductase (**Inha**) and for the Tyrosine kinase SRC (**SRC**). In order to avoid differences between the RDKit and the Open Babel results, these molecules were also removed from the set that was processed by the later toolkit.

The 10 fingerprints used and compared in this work, together with their category are shown in Table 2. In total, 5 *substructure key*, 4 *topological* and 1 *circular* fingerprints were employed.

Table B: Fingerprints used and type

Name	Type
Open Babel	
FP2	Topological
FP3	Substructure key
ChemFP-Substruct	Substructure key
RDKit	
AtomPair	Topological
Avalon	Substructure key
Fingerprint	Topological
MACCS166	Substructure key
Morgan	Circular
Pattern	Substructure key
Torsion	Topological

FiBeFTa implementation details

FiBeFTa is structured as a Python 2 script. Python 2 stopped being updated on April 2020 with the release of version 2.7, which signified the cease of development for Python 2 in favor of Python 3. Even though development of new tools in Python 2 is discouraged, it is used in FiBeFTa for 2 reasons:

1. The free open-source version of Chemfp only supports Python 2. Free academic licenses for Python 3 are also offered, but that would mean having to renew the license every year in order for FiBeFTa to work.

2. To maintain compatibility between other projects in the Cheminformatics & Nutrition Research Group.

All things considered, an effort has been made to develop FiBeFTa in such a way that works either in both Python 2 and 3, to facilitate the eventual transition to the new Python update.

Results and discussion

FiBeFTa Algorithm

The FiBeFTa algorithm uses two structural files as input, one for the active molecules of the target of interest and one for the corresponding decoys. Accepted formats for the molecules in these two input files are SDF and SMILES. The first step is converting independently both files into lists of molecules that are labeled as actives (*i.e.*, ***Active list***, see Figure 4) or decoys (*i.e.*, ***Decoy list***, see Figure 4). Once labeled, both lists are joined in a single file (*i.e.*, ***Total list***, see Figure 4). In order for the algorithm to be more efficient, all the fingerprints will be calculated now for the ***Active*** and the ***Total*** lists. Using these fingerprints, the program will iterate over the ***Total list*** calculating the Tanimoto coefficient with all the molecules at the ***Active list*** in order to find which is its closest active compound (during this process, each active in the ***Total list*** is not paired with itself at the ***Active list***). When there are no more molecules to pair, the list will be sorted by descending Tanimoto coefficient (if one decoy and one active have the same Tanimoto value, then the decoy will rank first in order to not favor false metrics). Once finished this sorting, the different validation metrics will be calculated and a **Molecule ranking** file for the corresponding target/fingerprint pair will be generated. These are *.csv* files where each molecule at the ***Total list*** is paired with its most similar active at the ***Active list*** and where the rows are sorted by default by decreasing Tanimoto coefficent. These files are named by the specific fingerprint they use and are useful for allowing the traceability of the FiBeFTa final output and, therefore, for each molecule, they include the next fields:

1. **Molecule ID:** the name for the molecule provided in the corresponding input file (*e.g.*, the ZINC code).
2. **Molecule SMILES:** the SMILES for the molecule.
3. **Tanimoto score:** the Tanimoto score between the molecule and its most similar active

4. **It's Active?:** a binary indicator (1/0) to show whether the molecule is an active (*i.e.*, 1) or a decoy (*i.e.*, 0). This data is used when calculating the metrics to see how well the corresponding fingerprint is able to distinguish between active/decoys.
5. **Closest Active ID:** the name for the most similar active molecule in the corresponding input file (*e.g.*, the ChEMBL ID).
6. **Closest Active SMILES:** the SMILES code for the most similar active molecule.

Once all the metrics for each of the fingerprints have been calculated a new file summarizing the results for each metric (by default EF1%, EF10%, AUC and BEDROC) is generated. These files acts as a summary of the execution, and concludes which fingerprint performs best for the given metric and target.

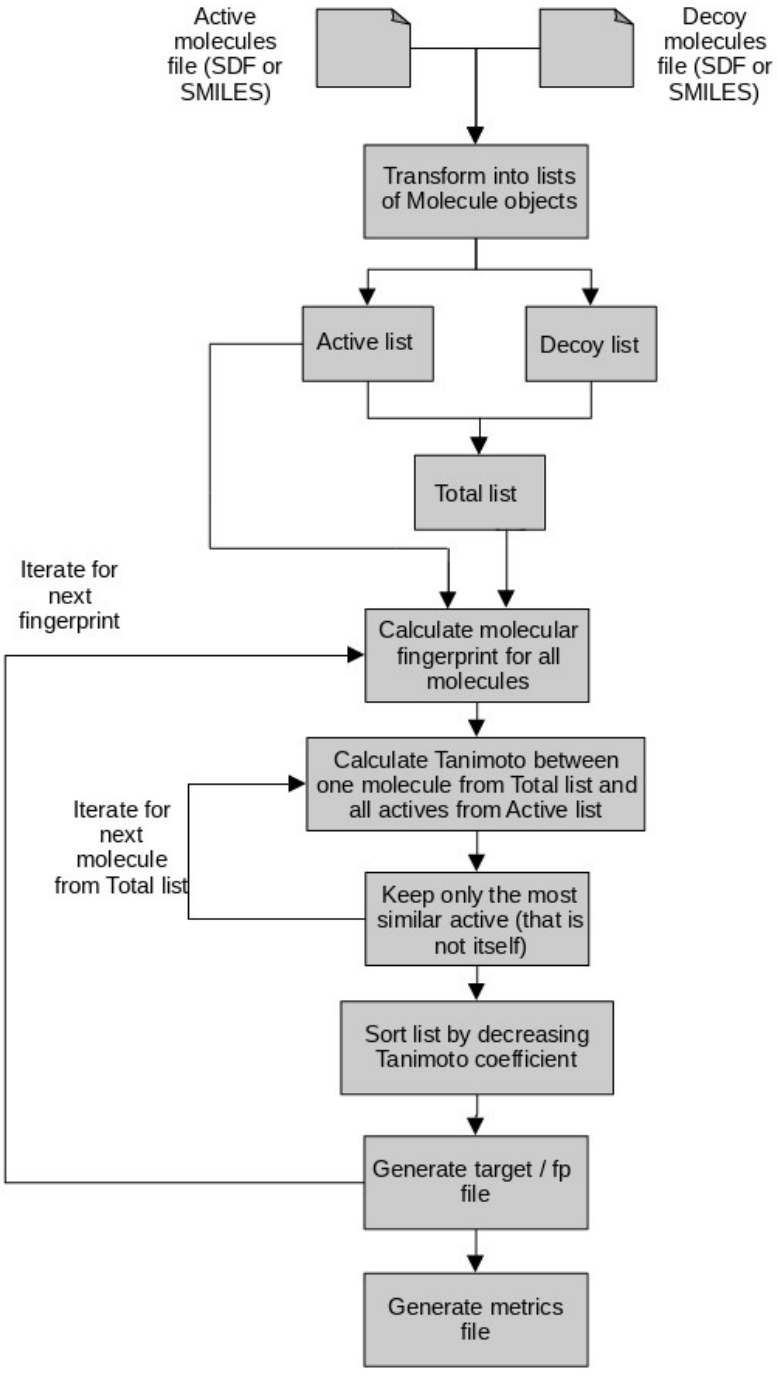


Figure 4: Pipeline of the algorithm

Analysis of DUD targets

Table 3 shows the data obtained with the 40 targets present at (DUD) [17]. Each of the fingerprints shows two values per metric: the first one is the number of targets for which the corresponding fingerprint obtained the best score, and the second is the number of times the same fingerprint fall above the 75th percentile results for a target. Since the results obtained for the different fingerprints are generally high (see Tables from S1 to S4), this double evaluation of the fingerprint performance allows for a better characterization of their robustness in VS.

Table 3 shows that the values for “Best” column (irrespective of the metrics) do not add to 40 (*i.e.*, the number of targets available at DUD). This is due to the fact that when two or more fingerprints shared the highest score for a particular target all of them are counted at the corresponding cell. As it can be seen in the results, this happens more frequently with the EF, especially with a factor of 1, since the result are determined by the outcome of fewer molecules.

Table C: Fingerprint performance Metric value classification by fingerprint

Fingerprint	EF1%		EF10%		AUC		BEDROC	
	Best	75th PCTL	Best	75th PCTL	Best	75th PCTL	Best	75th PCTL
Open Babel								
FP2	36	36	18	26	9	22	4	18
FP3	0	0	0	0	0	0	0	0
ChemFP-Substruct	34	35	8	14	2	19	2	5
RDKit								
AtomPair	32	33	20	26	11	25	8	20
Avalon	36	36	9	15	4	15	5	9
Fingerprint	36	36	11	15	6	18	7	15
MACCS166	32	33	14	19	4	21	2	8
Morgan	35	35	29	32	18	29	14	28
Pattern	38	39	11	15	4	16	3	6
Torsion	35	35	20	29	8	28	7	18
	314		140		66		52	

In Figure 5 and 6 we can see the previous data in a graphical format and separated by fingerprint type (Substructure fingerprints for Figure 5 and hashed fingerprints for Figure 6). Regarding the EF with a factor of 1 (that is, how more likely it is to statistically find an active in the top 1% of the list relative to a random order), most fingerprints perform fairly well (except FP3), with almost all fingerprints

performing best in, at least, 32 of the 40 targets. This is most likely due to the presence of stereoisomers, which will get paired regardless of the fingerprint used, and will occupy some of the top positions in the 1st top percent of the list of molecules. The best performing fingerprint by this metric is RDKit-Pattern (that is the best for 38 targets), but apart from this particular fingerprint both substructural and topological fingerprints perform similarly (see Table 3 and Figures 5 and 6). As mentioned, the only exception to these general good results for the EF1% is the fingerprint OpenBabel-FP3.

Looking at the results with the EF with a factor of 10, we start to see some differences in the performance of the fingerprints. RDKit-Morgan outperforms all the rest, giving the best results in 29 out of the 40 targets (9 more than the second-best options: RDKit-Atompair and Torsion). OpenBabel-FP2, while not scoring the best in most targets, 18 out of 40, matches AtomPair when comparing results in the 75th percentile, still falling behind Torsion (*i.e.*, 29).

Analyzing results for the AUC metric, we observe again how RDKit-Morgan performs better, having the best results in 18 of the targets, 7 more than RDKit-Atompair. This doesn't change when taking into account results in the 75th percentile, where these 2 fingerprints obtain results that are above than 75% of all results in over half the targets, together with FP2, MACCS166 and Torsion.

For the last metric taken into account, BEDROC, RDKit-Morgan is the best-perforiming fingerprint with achieving the best result in 14 targets, almost doubling the second-best, RDKit-Atompair with a value of 8. Observing the 75th percentile category we can see that only RDKit-Morgan and AtomPair score in more than half of the targets.

As we can see in the figures below, the hashed fingerprints score better according to almost all validation metrics, with the exception of EF1, where both perform similarly. Of the different hashed fingerprints RDKit-Morgan is the one that excels the most. From the category of hashed fingerprints, RDKit-Morgan is the only circular fingerprint evaluated.

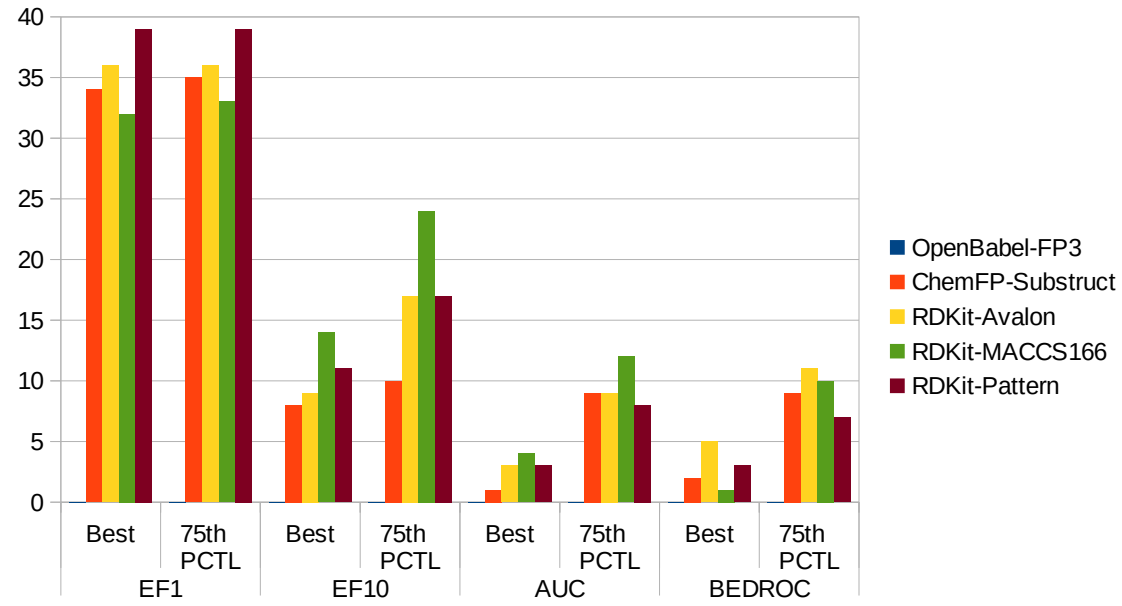


Figure 5: Number of DUD targets for which a substructure key fingerprint scores in the top positions

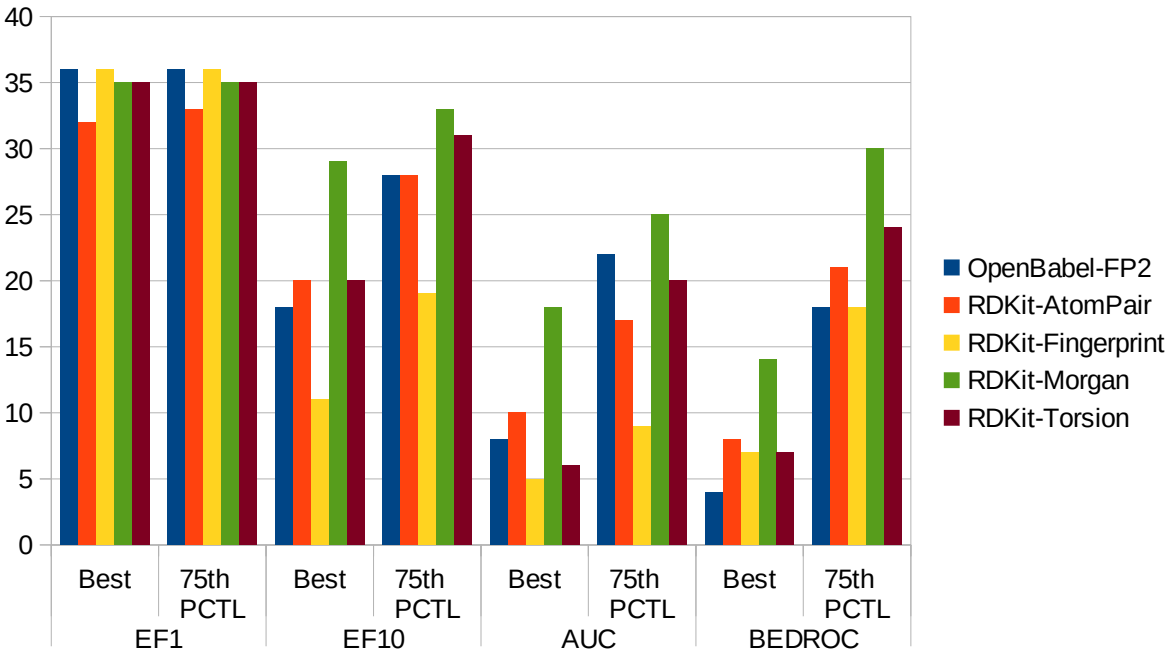


Figure 6: Number of DUD targets for which a hashed fingerprint scores in the top positions

Analysis of M-pro

In order to see a more every-day use of this tool we will analyze the output generated for the SARS-CoV2 M-pro. In Table 4 we can see the values of the metrics for each of the fingerprints.

Table D: Validation of VS for M-pro

Fingerprint	EF1%	EF10%	AUC	BEDROC
Open Babel				
FP2	35.19	8.89	0.97	0.87
FP3	0.00	0.00	0.89	0.04
ChemFP-Substruct	35.19	8.52	0.96	0.86
RDKit				
AtomPair	35.19	9.38	0.98	0.9
Avalon	35.19	8.03	0.94	0.8
Fingerprint	35.19	8.03	0.93	0.83
MACCS166	35.19	9.01	0.97	0.88
Morgan	35.19	9.75	0.99	0.97
Pattern	35.19	9.14	0.97	0.87
Torsion	35.19	9.26	0.97	0.9

As we can see in Figure 7 and Table 4, the ranges of the Enrichment Factor 10 go from 8.03 to 9.75 (excluding Open Babel FP3 that, again, performs very bad), and the fingerprint that has the best score is Morgan. FP3 uses SMARTS patterns for a variety of functional groups, but it completely ignores atoms that do not form part of these patterns, for example Bromine, and assigns a Tanimoto score of 1 to molecules that are not similar because of this. Since the FiBeFTa algorithm prioritizes decoys over active molecules for equal Tanimoto scores, OpenBabel-FP3 does not perform well in regardless of the metric analyzed. It should be noted that OpenBabel-FP3 is not usually used, falling behind FP2, which is considered the default fingerprint for substructure searching inside its package, OpenBabel, which agrees with the results obtained for the OpenBabel fingerprints.

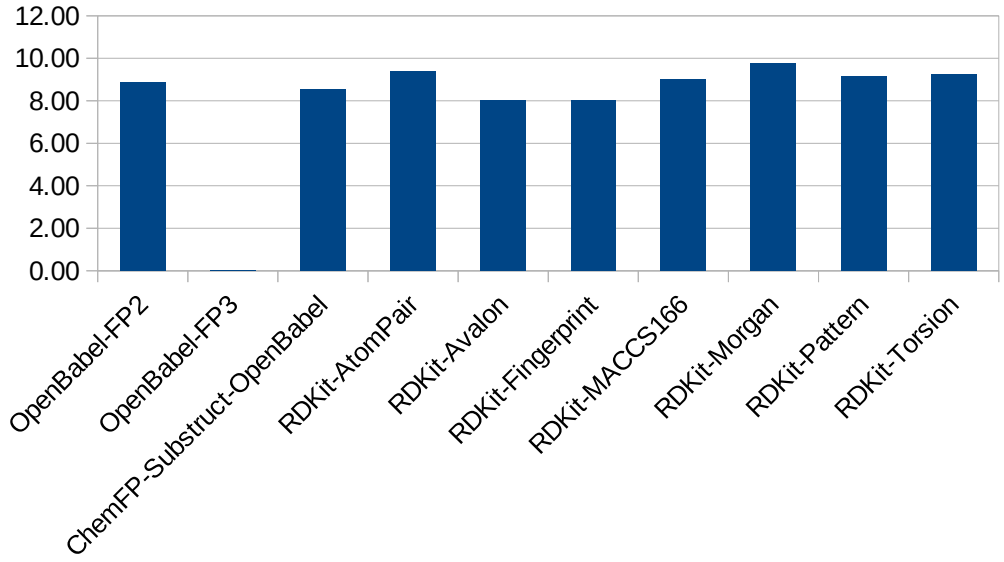


Figure 7: EF 10 of each fingerprint for M-pro

Figure 8 shows the values for both AUC and BEDROC scores. This Figure also confirms that for M-pro the most appropiate fingerprint to use is the circular fingerprint Morgan, since it obtains the best scores by both metric. In this figure we can also see an example of the "Early Recognition Problem" [19] for FP3, in which BEDROC assigns a poor score, since it gives priority to the decoys, and this metric gives more value to the first molecules identified and their order.

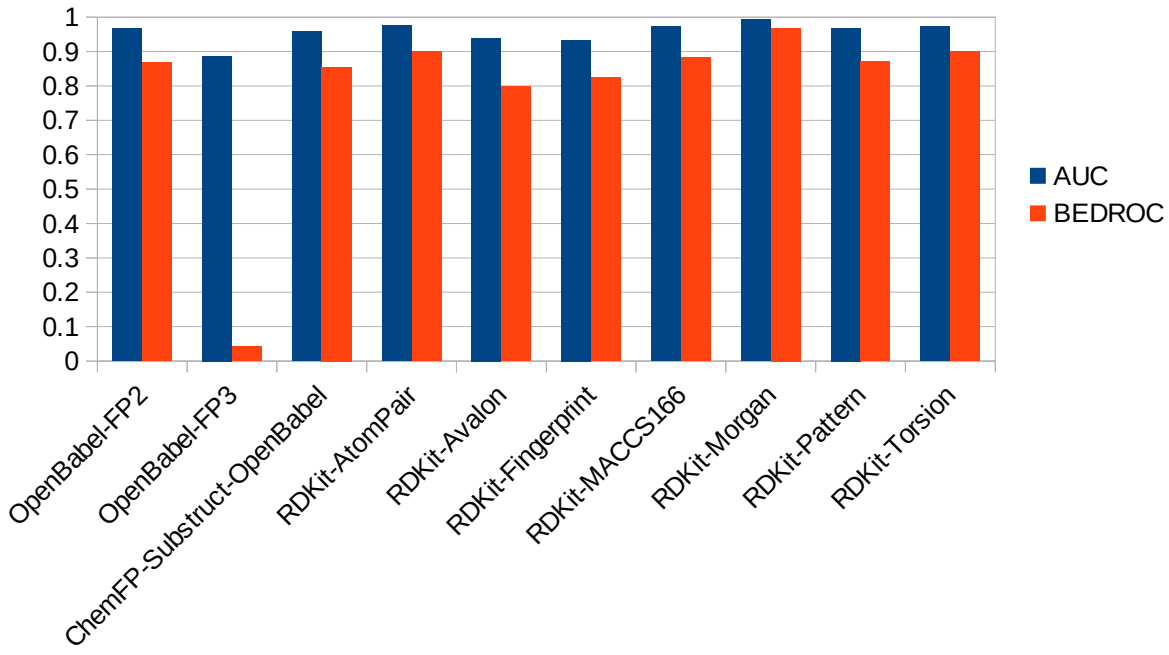


Figure 8: AUC and BEDROC of each fingerprint for M-pro

Conclusion

After reviewing the data generated we are in a position to answer some of the questions we have asked during this project. As hypothesized, using the algorithm before any actual study would help choose the best one for the specific molecule, which in turn would help separate potentially useful molecules *in silico.*

Looking at the results from Figures 5 and 6 we observe that for three out of the four metrics hashed fingerprints outperform substructure key fingerprints, and perform similarly in the fourth one. Out of all the hashed fingerprints, it's RDKit – Morgan, the only circular fingerprint evaluated that performs consistently well, irregardless of the metric.

The values for all targets of the DUD were also analysed (see Annex) in order to see if there was any target that ranked low on all fingrprints. Some targets have higher scores overall, and there are targets that rank lower for a specific metric. For example, the target thymidine kinase (tk) is the only target to have a maximum BEDROC of less than 0.8000 for any fingerprint (0.7852), however, it scores normally for the other metrics, showing no correlation between poor performance of one metric affecting the rest, at first sight.

Bibliography

1. Shukla, A.A. High Throughput Screening of Small Molecule Library: Procedure, Challenges and Future. *J. Cancer Prev. Curr. Res.* **2016**, *Volume 5*, doi:10.15406/JCPCR.2016.05.00154.

2. Gimeno, A.; Ojeda-Montes, M.; Tomás-Hernández, S.; Cereto-Massagué, A.; Beltrán-Debón, R.; Mulero, M.; Pujadas, G.; Garcia-Vallvé, S. The Light and Dark Sides of Virtual Screening: What Is There to Know? *Int. J. Mol. Sci.* **2019**, *20*, 1375, doi:10.3390/ijms20061375.

3. Barakat, K.H.; Mane, J.Y.; Tuszynski, J.A. Virtual screening: An overview on methods and applications. In *Handbook of Research on Computational and Systems Biology: Interdisciplinary Applications*; IGI Global, 2011; pp. 28–60 ISBN 9781609604912.

4. Sethi, A.; Joshi, K.; Sasikala, K.; Alvala, M. Molecular Docking in Modern Drug Discovery: Principles and Recent Applications. *Drug Discov. Dev. - New Adv.* **2019**, doi:10.5772/INTECHOPEN.85991.

5. Ballester, P.J.; Westwood, I.; Laurieri, N.; Sim, E.; Richards, W.G. Prospective virtual screening with ultrafast shape recognition: The identification of novel inhibitors of arylamine N-acetyltransferases. *J. R. Soc. Interface* **2010**, *7*, 335–342, doi:10.1098/rsif.2009.0170.

6. Schaller, D.; Šribar, D.; Noonan, T.; Deng, L.; Trung, |; Nguyen, N.; Pach, S.; David Machalz, |; Bermudez, M.; Wolber, G. Next generation 3D pharmacophore modeling Programming Molecular and Statistical Mechanics > Molecular Interactions. **2020**, doi:10.1002/wcms.1468.

7. Seidel, T.; Wieder, O.; Garon, A.; Langer, T. Applications of the Pharmacophore Concept in Natural Product inspired Drug Design. *Mol. Inform.* **2020**, *39*, 2000059, doi:10.1002/MINF.202000059.

8. Kirchmair, J.; Ristic, S.; Eder, K.; Markt, P.; Wolber, G.; Laggner, C.; Langer, T. Fast and efficient in silico 3D screening: toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *J. Chem. Inf. Model.* **2007**, *47*, 2182–96, doi:10.1021/ci700024q.

9. Puertas-Martín, S.; Redondo, J.L.; Ortigosa, P.M.; Pérez-Sánchez, H. OptiPharm: An evolutionary algorithm to compare shape similarity. *Sci. Reports 2019 91* **2019**, *9*, 1–24, doi:10.1038/s41598-018-37908-6.

10. Puertas-Martín, S.; L. Redondo, J.; Pérez-Sánchez, H.; M. Ortigosa, P. Optimizing Electrostatic Similarity for Virtual Screening: A New Methodology. *Informatica* **2020**, *31*, 821–839, doi:10.15388/20-INFOR424.

11. Kuwahara, H.; Gao, X. Analysis of the effects of related fingerprints on molecular similarity using an eigenvalue entropy approach. *J. Cheminformatics 2021 131* **2021**, *13*, 1–12, doi:10.1186/S13321-021-00506-2.

12. K, R.; W, C.; S, P.; A, P.; AJ, B. Substructural Connectivity Fingerprint and Extreme Entropy Machines-A New Method of Compound Representation and Analysis. *Molecules* **2018**, *23*, doi:10.3390/MOLECULES23061242.

13. Cereto-Massagué, A.; Ojeda, M.J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63, doi:10.1016/j.ymeth.2014.08.005.

14. Bender, A.; Glen, R.C. Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem.* 2004, *2*, 3204–3218.

15. Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity - a Review. *QSAR Comb. Sci.* **2003**, *22*, 1006–1026, doi:10.1002/qsar.200330831.

16. Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035–1042, doi:10.1021/jm0003992.

17. Huang, N.; Shoichet, B.K.; Irwin, J.J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801, doi:10.1021/jm0608356.

18. Song, L.; Liu, A.; Shi, J. Genetics and population analysis SummaryAUC: a tool for evaluating the performance of polygenic risk prediction models in validation datasets with only summary level statistics. *Bioinformatics* **2019**, *4038–4044*, 20, doi:10.1093/bioinformatics/btz176.

19. Truchon, J.-F.; Bayly, C.I. Evaluating virtual screening methods: good and bad metrics for the “early recognition” problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508, doi:10.1021/ci600426e.

20. Zakeri, P.; Simm, J.; Arany, A.; Elshal, S.; Moreau, Y. Gene prioritization using Bayesian matrix factorization with genomic and phenotypic side information. In Proceedings of the Bioinformatics; Oxford University Press, 2018; Vol. 34, pp. i447–i456.

21. Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; et al. Structure of M pro from SARS-CoV-2 and discovery of its inhibitors. *Nat. 2020 5827811* **2020**, *582*, 289–293, doi:10.1038/s41586-020-2223-y.

22. Cereto-Massagué, A.; Guasch, L.; Valls, C.; Mulero, M.; Pujadas, G.; Garcia-Vallvé, S. DecoyFinder: An easy-to-use python GUI application for building target-specific decoy sets. *Bioinformatics* **2012**, *28*, doi:10.1093/bioinformatics/bts249.

23. Sterling, T.; Irwin, J.J. ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337, doi:10.1021/ACS.JCIM.5B00559.

24. Irwin, J.J.; Sterling, T.; Mysinger, M.M.; Bolstad, E.S.; Coleman, R.G. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.* 2012, *52*, 1757–1768.

25. Dalke, A. The chemfp project. *J. Cheminform.* **2019**, *11*, 76, doi:10.1186/s13321-019-0398-8.

Annex

DUD Results

Table S1: Enrichment Factor 1

Targets	OpenBabel			RDKit						
	FP2	FP3	Substruct	AtomPair	Avalon	Fingerprint	MACCS166	Morgan	Pattern	Torsion
ace	37.67	0	37.67	37.67	35.58	35.58	37.67	35.58	37.67	37.67
ache	37.37	0	35.46	37.37	37.37	36.42	34.5	37.37	37.37	37.37
ada	24.77	0	24.77	24.77	24.77	24.77	24.77	24.77	24.77	24.77
alr2	39.27	0	39.27	39.27	39.27	39.27	35.34	39.27	39.27	39.27
ampc	38.43	0	38.43	38.43	38.43	38.43	38.43	38.43	38.43	33.63
ar	35.85	0	37.13	33.29	37.13	37.13	29.45	37.13	37.13	35.85
cdk2	29.79	0	29.79	29.79	29.79	29.79	29.79	29.79	29.79	29.79
comt	43.55	0	43.55	43.55	43.55	43.55	43.55	43.55	43.55	43.55
cox1	37.44	0	37.44	37.44	37.44	37.44	33.28	37.44	37.44	37.44
cox2	32.19	0	32.19	32.19	32.19	32.19	31.96	31.96	32.19	32.19
dhfr	21.4	0	21.15	21.4	21.4	21.4	21.4	21.15	21.4	21.4
egfr	34.68	0	34.68	34.68	34.68	34.68	34.68	34.68	34.68	34.68
er_agonist	39.36	0	36.33	39.36	39.36	39.36	39.36	37.84	39.36	37.84
er_antagonist	38.13	0	35.4	38.13	35.4	32.68	38.13	38.13	38.13	38.13
fgfr1	38.92	0	38.92	38.92	38.92	38.92	38.92	38.92	38.92	38.92
fxa	40.34	0	40.34	39.64	40.34	40.34	40.34	40.34	40.34	40.34
gart	20.42	0	22.97	22.97	22.97	22.97	17.87	22.97	22.97	22.97
gpb	40.04	0	42.04	38.03	40.04	42.04	30.03	42.04	42.04	40.04
gr	38.78	0	38.78	38.78	38.78	38.78	38.78	38.78	38.78	38.78
hivpr	33.87	0	33.87	33.87	33.87	33.87	33.87	33.87	33.87	33.87
hivrt	36.33	0	36.33	36.33	36.33	36.33	36.33	36.33	36.33	36.33
hmga	43.29	0	43.29	43.29	43.29	43.29	43.29	43.29	43.29	43.29
hsp90	27.46	0	27.46	27.46	27.46	27.46	27.46	27.46	27.46	27.46
inha	42.34	0	42.34	42.34	42.34	42.34	42.34	42.34	42.34	42.34
mr	43.4	0	43.4	43.4	36.17	43.4	43.4	43.4	43.4	43.4
na	39.22	0	39.22	35.1	39.22	39.22	39.22	39.22	39.22	39.22
p38	21.13	0	21.13	21.13	21.13	21.13	21.13	21.13	21.13	21.13
parp	39.6	0	39.6	39.6	39.6	39.6	39.6	39.6	39.6	39.6
pde5	23.48	0	23.48	23.48	23.48	23.48	23.48	23.48	19.96	23.48
pdgfrb	36.18	0	36.18	36.18	36.18	36.18	36.18	36.18	36.18	36.18
pnp	21.72	0	21.72	21.72	21.72	21.72	21.72	21.72	21.72	21.72
ppar_gamma	37.79	0	37.79	36.61	37.79	37.79	37.79	37.79	37.79	37.79
pr	39.56	0	39.56	39.56	39.56	39.56	39.56	39.56	39.56	39.56
rxr_alpha	38.5	0	38.5	38.5	38.5	38.5	38.5	38.5	38.5	38.5
sahh	41.73	0	41.73	38.52	41.73	41.73	41.73	41.73	41.73	41.73
src	40.74	0	40.74	40.74	40.74	40.74	40.74	40.74	40.74	40.74
thrombin	35.11	0	35.11	35.11	35.11	35.11	35.11	35.11	35.11	35.11
tk	27.67	0	32.28	32.28	36.89	27.67	32.28	9.22	18.44	13.83
trypsin	34.96	0	34.96	34.96	34.96	34.96	34.96	34.96	34.96	34.96
vegfr2	34.02	0	32.85	31.68	34.02	34.02	34.02	34.02	34.02	34.02

Table S2: Enrichment Factor 10

Target	OpenBabel			RDKit						
	FP2	FP3	Substruct	AtomPair	Avalon	Fingerprint	MACCS166	Morgan	Pattern	Torsion
ace	9.01	0	8.39	9.83	9.42	9.01	9.21	9.83	7.58	8.8
ache	9.18	0	8.71	8.99	9.18	9.27	8.8	8.99	8.71	8.99
ada	9.55	9.29	9.29	10.06	9.55	8.77	10.06	10.06	9.55	10.06
alr2	8.08	0	7.7	8.08	7.7	8.47	8.08	8.08	8.85	8.08
ampc	8.65	0	9.61	7.69	9.13	8.65	8.17	9.13	8.17	9.13
ar	9.88	6.34	9.76	9.88	9.88	9.76	9.88	9.88	9.88	9.88
cdk2	9.05	0	8.21	8.63	8.63	8.91	8.35	9.47	9.19	9.05
comt	7.41	5.56	7.41	8.34	8.34	6.49	8.34	7.41	6.49	8.34
cox1	7.25	0	7.25	5.64	6.44	6.44	6.44	7.25	7.25	7.25
cox2	9.82	0	9.63	9.79	9.44	9.37	9.35	9.82	9.6	9.79
dhfr	10	1.66	10	10	10	10	10	10	10	10
egfr	9.98	0	9.96	9.98	9.87	9.98	9.98	9.98	9.98	9.98
er_agonist	9.58	8.98	10.03	9.88	9.13	9.28	10.03	10.03	9.13	9.73
er_antagonist	10.05	9.53	10.05	10.05	9.53	9.79	10.05	10.05	9.53	9.79
fgfr1	10	0	10	10	10	10	10	10	10	10
fxa	9.88	0	9.6	9.6	9.33	9.4	9.4	9.95	9.05	9.6
gart	9.09	0	9.09	8.33	9.59	9.09	9.34	8.84	9.34	9.34
gpb	9.06	3.28	9.64	9.83	9.26	9.06	9.64	9.26	9.45	9.26
gr	10.02	0	9.89	9.89	9.89	10.02	9.89	10.02	10.02	10.02
hivpr	9.84	7.26	9.68	9.52	9.52	9.68	8.71	9.84	9.68	9.84
hivrt	8.62	0	8.62	7.68	7.92	8.38	5.82	7.68	6.52	7.45
hmga	9.46	9.17	9.75	10.03	8.03	8.03	9.75	10.03	9.46	10.03
hsp90	9.79	9.24	9.79	10.06	9.79	9.79	10.06	10.06	9.79	10.06
inha	9.89	0	9.63	9.63	9.63	9.89	9.89	9.89	9.63	9.89
mr	8.01	7.34	8.01	8.68	8.68	8.01	8.01	8.68	7.34	7.34
na	10.01	0	9.81	10.01	9.81	9.81	9.81	9.6	8.99	9.19
p38	10.01	0	9.96	10.01	10.01	10.01	9.87	10.01	10.01	10.01
parp	8.9	0	9.18	8.61	8.32	8.32	9.47	8.61	8.32	9.47
pde5	9.8	0	9.35	9.57	8.78	8.78	8.89	9.92	8.09	9.92
pdgfrb	9.94	0	9.94	10	9.88	9.94	9.76	9.94	9.94	9.94
pnp	9.65	0	9.25	9.05	9.65	8.85	9.65	10.06	9.85	9.25
ppar_gamma	9.77	0	9.65	9.77	9.54	9.77	9.54	9.77	9.3	9.77
pr	10.08	0	9.7	10.08	10.08	10.08	10.08	10.08	10.08	10.08
rxr_alpha	10	0	10	10	10	10	10	10	10	10
sahh	10.05	9.75	9.75	10.05	9.44	10.05	10.05	10.05	10.05	10.05
src	9.95	0	9.95	9.89	9.89	9.82	9.7	9.89	9.82	10.01
thrombin	8.64	0	8.5	9.06	8.78	8.64	8.64	9.34	8.5	8.92
tk	8.21	0	8.66	9.58	9.12	8.66	8.66	9.58	8.66	8.66
trypsin	9.4	0	9.4	9.61	9.4	9.4	8.59	9.81	8.79	9.61
vegfr2	8.76	0	7.62	8.19	6.94	7.28	7.85	8.76	7.74	7.74

Table S3: AUC

Targets	OpenBabel-FP2			RDKit-AtomPair						
	FP2	FP3	Substruct	AtomPair	Avalon	Fingerprint	MACCS166	Morgan	Pattern	Torsion
ace	0.9663	0.7972	0.9405	0.989	0.98	0.9584	0.9641	0.987	0.9201	0.9558
ache	0.9583	0.7042	0.9556	0.9113	0.9376	0.9542	0.9066	0.9463	0.9161	0.9364
ada	0.9912	0.9774	0.9919	0.9997	0.9898	0.9861	0.9991	0.9989	0.9942	0.9978
alr2	0.8726	0.8177	0.9127	0.8714	0.8747	0.8932	0.9081	0.9233	0.8952	0.8697
ampc	0.9687	0.6623	0.9831	0.9621	0.9783	0.9609	0.9251	0.9832	0.9587	0.9663
ar	0.9901	0.9347	0.9892	0.9902	0.99	0.9903	0.9846	0.9875	0.9893	0.9868
cdk2	0.9498	0.6923	0.941	0.9412	0.955	0.9536	0.9202	0.9838	0.9526	0.959
comt	0.8831	0.8278	0.8904	0.9172	0.8619	0.8353	0.9194	0.8856	0.859	0.9113
cox1	0.9232	0.6399	0.8325	0.7542	0.7896	0.8243	0.8218	0.8467	0.8707	0.9102
cox2	0.9916	0.7815	0.983	0.9917	0.9711	0.9721	0.9753	0.9935	0.9834	0.9916
dhfr	1	0.9516	0.9998	1	0.9998	1	1	0.9998	1	1
egfr	0.9995	0.908	0.9981	0.999	0.9967	0.9993	0.9984	0.9996	0.9993	0.9996
er_agonist	0.9884	0.9371	0.9936	0.996	0.9679	0.971	0.9953	0.9957	0.9704	0.9941
er_antagonist	0.9968	0.9626	0.9971	0.9971	0.9873	0.9896	0.9965	0.9985	0.9931	0.9949
fgfr1	1	0.9101	1	1	1	1	0.9995	1	1	1
fxa	0.9953	0.8866	0.9836	0.9813	0.9765	0.9763	0.9646	0.9939	0.9705	0.987
gart	0.9846	0.8385	0.9805	0.9662	0.9741	0.9832	0.9873	0.9845	0.9878	0.9824
gpb	0.9802	0.9002	0.9858	0.9869	0.984	0.9617	0.9708	0.9832	0.9837	0.9716
gr	1	0.8509	0.9929	0.9987	0.9977	0.9991	0.9905	1	0.9998	0.9999
hivpr	0.9941	0.9258	0.9762	0.9802	0.9718	0.9734	0.9408	0.988	0.9895	0.9974
hivrt	0.9339	0.7356	0.9327	0.8522	0.8829	0.978	0.8391	0.8504	0.8459	0.8988
hmga	0.9435	0.9416	0.9943	1	0.9225	0.9149	0.9862	1	0.9671	0.9996
hsp90	0.997	0.9567	0.9981	0.9994	0.9941	0.9958	0.9998	0.9999	0.997	0.9988
inha	0.9873	0.7082	0.9825	0.9713	0.9682	0.9873	0.9841	0.9895	0.9818	0.9911
mr	0.9016	0.8307	0.8924	0.9123	0.9435	0.8817	0.9193	0.9033	0.9401	0.9072
na	0.9941	0.9044	0.9933	0.9947	0.9861	0.9933	0.9935	0.9929	0.9806	0.9642
p38	0.9999	0.7559	0.999	0.9998	0.9998	0.9999	0.9974	1	0.9998	0.9999
parp	0.9726	0.7605	0.9725	0.8928	0.9028	0.8702	0.9773	0.9602	0.8762	0.9714
pde5	0.9775	0.7207	0.9862	0.9773	0.9471	0.954	0.9538	0.9989	0.9073	0.9915
pdgfrb	0.9979	0.859	0.9992	0.9996	0.9951	0.9961	0.9923	0.9977	0.9993	0.9994
pnp	0.9946	0.8327	0.9779	0.9716	0.9947	0.9759	0.993	0.9983	0.9971	0.9929
ppar_gamma	0.983	0.8379	0.973	0.9759	0.971	0.9766	0.9715	0.9858	0.9601	0.9805
pr	0.9998	0.8089	0.9824	0.9979	0.9998	0.9998	0.9985	0.9996	0.9989	0.9995
rxr_alpha	0.9998	0.8192	0.9991	1	0.9997	0.9999	0.9981	1	0.9997	0.9998
sahh	0.9954	0.9749	0.9932	0.9944	0.9905	0.9938	0.9977	0.995	0.9951	0.9952
src	0.9961	0.8918	0.9975	0.9982	0.9967	0.9942	0.9926	0.9987	0.995	0.9992
thrombin	0.9406	0.912	0.9441	0.9487	0.9388	0.9184	0.9237	0.9747	0.9281	0.9702
tk	0.9703	0.8548	0.963	0.9584	0.9716	0.9627	0.9628	0.9751	0.9665	0.9553
trypsin	0.9769	0.8857	0.9523	0.9618	0.952	0.9567	0.9497	0.9929	0.9582	0.9735
vegfr2	0.95	0.5895	0.9182	0.9272	0.8994	0.9022	0.9292	0.9365	0.8899	0.9124

Table S4: BEDROC

Target	OpenBabel-FP2			RDKit						
	FP2	FP3	Substruct	AtomPair	Avalon	Fingerprint	MACCS166	Morgan	Pattern	Torsion
ace	0.8415	0.0013	0.7933	0.9131	0.8235	0.8545	0.8888	0.9042	0.751	0.8444
ache	0.9092	0.0001	0.8636	0.8947	0.9186	0.9203	0.8669	0.912	0.8763	0.9157
ada	0.9312	0.497	0.9384	0.9961	0.9203	0.9099	0.9857	0.9838	0.9587	0.9756
alr2	0.7577	0.0073	0.7857	0.8419	0.7922	0.8425	0.8337	0.8337	0.8333	0.8292
ampc	0.8637	0.0005	0.8667	0.8014	0.911	0.8843	0.7631	0.9086	0.7879	0.8713
ar	0.9428	0.1819	0.9495	0.9416	0.9677	0.9495	0.8862	0.9413	0.9174	0.9167
cdk2	0.9223	0.0019	0.7883	0.8675	0.8893	0.9043	0.8282	0.9474	0.8881	0.9044
comt	0.6955	0.3723	0.688	0.8364	0.8001	0.6301	0.7722	0.7598	0.5956	0.7336
cox1	0.7125	0.0002	0.6717	0.5435	0.6574	0.6823	0.6563	0.7018	0.6597	0.7247
cox2	0.9784	0.0008	0.9565	0.9795	0.9463	0.9473	0.9276	0.9784	0.9468	0.9759
dhfr	1	0.1582	0.9964	1	0.9969	1	0.9997	0.9961	1	1
egfr	0.9952	0.0489	0.9918	0.9956	0.9849	0.9952	0.9843	0.998	0.9936	0.9974
er_agonist	0.9332	0.3282	0.9111	0.9577	0.8975	0.9077	0.9413	0.9536	0.9003	0.944
er_antagonist	0.9527	0.2691	0.9537	0.9573	0.9183	0.9329	0.9497	0.9779	0.9408	0.9595
fgfr1	1	0.0455	1	0.9997	1	1	0.9939	1	1	1
fxa	0.9626	0.0534	0.9303	0.9477	0.9338	0.9334	0.9215	0.9734	0.8848	0.9457
gart	0.8565	0.004	0.8686	0.8137	0.8901	0.8758	0.8809	0.8621	0.9175	0.8796
gpb	0.8724	0.1275	0.8987	0.9186	0.8565	0.8622	0.8255	0.9005	0.8485	0.8856
gr	0.9995	0.0102	0.9899	0.9915	0.9905	0.9921	0.9856	0.9999	0.9966	0.9988
hivpr	0.9721	0.5266	0.9649	0.9507	0.9531	0.9503	0.8952	0.9753	0.9513	0.9854
hivrt	0.856	0.0308	0.8363	0.7872	0.7953	0.8671	0.6476	0.8078	0.6569	0.7656
hmga	0.9352	0.7427	0.9715	0.9997	0.8443	0.8419	0.9471	1	0.9106	0.9938
hsp90	0.9837	0.4645	0.9862	0.9932	0.9807	0.982	0.9979	0.9982	0.9824	0.989
inha	0.9847	0.0001	0.9704	0.95	0.9644	0.9894	0.9496	0.9893	0.9587	0.9787
mr	0.7779	0.5838	0.7846	0.7985	0.8073	0.7993	0.7894	0.7605	0.7425	0.7309
na	0.9359	0.1131	0.9567	0.9339	0.932	0.9458	0.9323	0.9389	0.8865	0.9189
p38	0.9992	0.0002	0.9939	0.9978	0.9976	0.9984	0.9845	0.9996	0.9979	0.9983
parp	0.8973	0.0034	0.8823	0.8838	0.8648	0.8644	0.9428	0.8708	0.8648	0.9146
pde5	0.9514	0.0044	0.9263	0.9454	0.8743	0.8798	0.8739	0.9907	0.7579	0.9819
pdgfrb	0.995	0.013	0.9938	0.9959	0.9896	0.9948	0.9757	0.9955	0.9956	0.9962
pnp	0.9713	0.0392	0.943	0.919	0.9652	0.9136	0.9374	0.9824	0.9784	0.9607
ppar_gamma	0.9694	0.0102	0.9534	0.9722	0.943	0.9739	0.9498	0.9749	0.9335	0.9758
pr	0.9968	0.0419	0.9616	0.978	0.9967	0.9968	0.9815	0.9946	0.9813	0.9929
rxr_alpha	0.9969	0.0085	0.9863	1	0.9951	0.999	0.9699	1	0.9951	0.9971
sahh	0.9425	0.4158	0.9288	0.9213	0.9112	0.9363	0.9652	0.9289	0.9418	0.9318
src	0.9881	0.0362	0.9883	0.9901	0.9856	0.9773	0.9722	0.9894	0.9816	0.9919
thrombin	0.8747	0.1127	0.8641	0.8926	0.8828	0.8599	0.8688	0.8975	0.8243	0.8956
tk	0.7774	0.0164	0.7597	0.745	0.7852	0.774	0.6598	0.7124	0.7621	0.6017
trypsin	0.9421	0.0244	0.9333	0.946	0.9367	0.9416	0.8739	0.9589	0.8562	0.958
vegfr2	0.8625	0.0004	0.7542	0.8396	0.7063	0.7361	0.7363	0.8579	0.7251	0.8006