Team name	SAFAN_Cov
Team member(s) (firstname lastname; …)	Luisa Pugliese
Affiliation	S.A.F.AN. BIOINFORMATICS
Contact email	Luisa.pugliese@safan-bioinformatics.it
Contact phone number (optional)	+393336130644
Protein targets (for example: 3CLPro/Nsp5, BoAT1, Fc Receptor, Furin, IL6R, M protein, Nspx, OrfXx, N, E, etc…) | 3 required	3CLPro/Nsp5, Papain-like protease, RNA-dependent RNA polimerase, AP2-associated protein kinase 1

Methods:

Compounds were first selected using SAFAN-ISP proprietary technology.

It is a ligand based method calculating the binding affinities between each ligand and more than 4500 targets from 15 different protein classes.

As many ligand based methods SAFAN-ISP is based on the molecular similarity evaluation between the submitted molecule(s) and those in an active compound database containing experimental data concerning the interactions between molecules and protein targets.

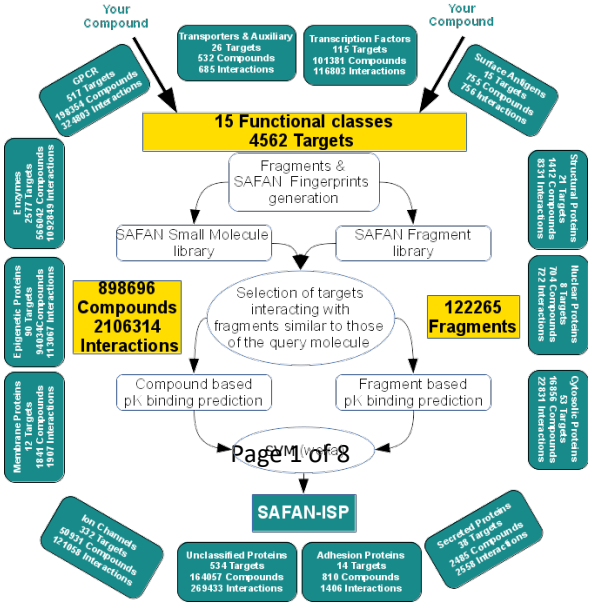
SAFAN-ISP calculates similarities using newly derived fingerprints based on small substructures. The similarity (Tanimoto) is computed matching atoms to substructures and evaluating common atoms.

For the experimental data we use a refactored bioactivity database derived from the CHEMBL25 database.

The technology involves three steps:

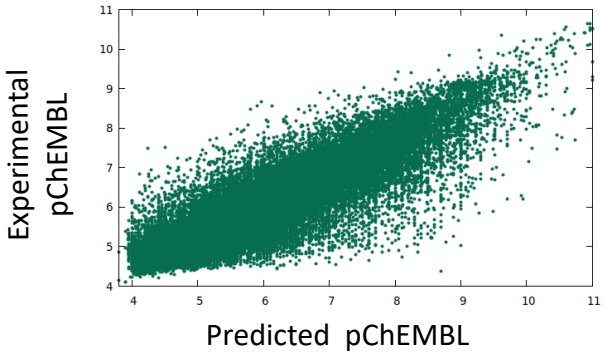
1. Molecule fragmentation : the input molecule is splitted into fragments following an newly derived scheme. Each fragment is compared to SAFAN-ISP database of fragments, derived from the compounds included in our bioactivities database.
   Fragments are used to :
   - select targets sharing similar fragments with the input compounds,
   - calculating affinities combining concerning different compounds binding the same target
2. Affinity calculation based on compound similarity : the similarity of the input compound is derived with all the compounds present in SAFAN-ISP database. Next the binding constant on all targets ouputted from step 1. is computed combining similarities and experimental data using four different schemes, two including chiral fingerprints and two excluding them.
3. Weka Machine Learning approach : the REPTree algorithm, available from the WEKA open source package, is used to combine all binding constants derived in steps 1. and 2. in a single value that will be the final output. RepTree uses the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees.

SAFAN-ISP technology is summarized by the following scheme.

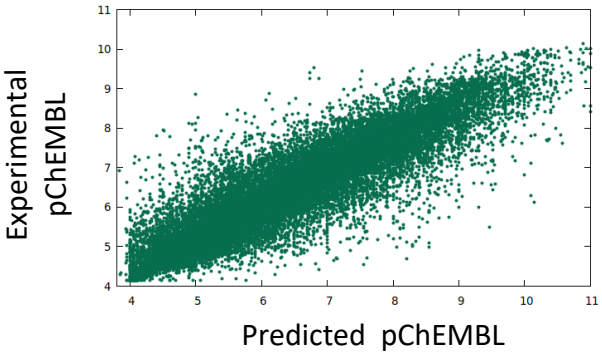


The output of SAFAN-ISP is a pChEMBL defined as: -Log(molar IC50, XC50, EC50, AC50, Ki, Kd or Potency).
SAFAN-ISP results were validated on 21000 compounds by the leave-one-out method, resulting in a Pearson correlation with experimental data of 0.89 and 0.91 for non chiral and chiral compounds respectively.

NON CHIRAL



CHIRAL



SAFAN-ISP was further validated on the selected targets:

Target	Pearson Correlation with experimental data	Mean Difference from the experimental data
3CLPro/Nsp5	0.82	0.004
Papain-like protease	0.27	0.03
RNA-dependent RNA polimerase	0.91	0.14
AP2-associated protein kinase 1	0.84	0.15

The poor performance, in term of Pearson correlation, for Papain-like protease compounds is mainly due to the fact that the dispersion of the experimental data is very narrow. Indeed the average difference of the predicted and the experimental pChEMBL is 0.03.

Since SAFAN-ISP is a profiling algorithm for each compound we get information on many different targets. To select interesting compounds it is thus possible combine the predicted pChEMBL with the relative position in the profiling. Compound ranking higher in the profiling are likely to be more selective.
The top 2000 compounds were thus selected and used for docking experiments with autodock VINA to check if they could fit into the relative binding sites.
VINA scoring was also checked for correlation with experimental data on the same data sets used to validate SAFAN-ISP predictions on the four targets.
Since no correlation was found, VINA results were used to evaluate how many residues within the binding site were in contact with the docked molecule in the best pose.
In the last column of the JEDI_Stage1_template_v1.csv the pChEMBL value is reported with the fraction of binding site residues in contact with the molecule.

Compounds for which the last column is empty, are sorted by similarity to compounds binding the selected target within SAFAN-ISP database.

Targets:

The first three targets, 3CLPro/Nsp5, Papain-like protease, RNA-dependent RNA polimerase, were selected because they are part of those that will be experimentally validated within the challenge. The fourth, AP2-associated protein kinase 1 is part of the CAS COVID 19 Protein Target Thesaurus and at the moment no approved drug exists having AP2-associated protein kinase 1 as main target. Since the goal of docking was to check if the selected compounds were able to reach the binding site, we looked for x-ray structures in complex with inhibitor and we restrained then the docking to the region where the ligand was positioned in the crystal structure.

The following structures were actually used:

Target	**PDB code**
RNA-dependent RNA polimerase	4o4r
Papain-like protease	6wx4,6w9c
3CLPro/Nsp5	6lu7
AP2-associated protein kinase 1	5te0

We VINA run through the macro available from YASARA software package. All side chains were free to move with the exception of Cysteines involved in disulphide bridges.

Libraries*:*

The following libraries were tested:

Library	Link	Number of compounds
CHEMBL selection (includes compounds with known clinical phase of development)	https://www.ebi.ac.uk/chembl/	1000000
MilliporeSigma AMS Data	Obtained from Thomas Herman	5049312
Molport	www.molport.com	7615837
Cas antiviral	www.cas.org/covid-19-antiviral-compounds-dataset	49438
SWEETLEAD	simtk.org/projects/sweetlead	4317
Foodb	Foodb.ca	26485
PCIDB	www.genome.jp/db/pcidb	36783
pepbank	pepbank.mgh.harvard.edu	15491
Total number of compounds		13797663

In order to get rid of duplicates in our output list, we sorted the compounds as described in the method section (by decreasing pChEMBL and by increasing position in the target list and by

decreasing % of binding site residues in contact) and then we added new compounds up to 10000 only if the corresponding InChI was not already found.
We screened in total 13797663 compounds, much less than the billion originally required.
However the compounds were not screened against one target but profiled against 4500 targets and this was very helpful to filter more selective compounds.
As Thomas Herman knows we had huge problems in getting the required computational time.

Results:

Among four targets selected, for RNA-dependent RNA polimerase were available a lot of experimental data and SAFAN-ISP easily selected more than 10000 compounds.
For AP2-associated protein kinase 1 and 3CLPro/Nsp5 much less data were available. SAFAN-ISP was able to retrieve between 2000 and 3000 compounds and the other compounds to reach 10000, were selected by similarity with respect to the compounds associated to AP2-associated protein kinase 1 or 3CLPro/Nsp5 in SAFAN-ISP database.
Concerning the papain-like protease no experimental data were available. However the papain-like protease share a high folding homology with the Ubiquitin carboxyl-terminal hydrolase 2 and for this enzyme many experimental data are available.
Thus we decided to select the compounds looking for interactions with Ubiquitin carboxyl-terminal hydrolase 2 and then check if the selected compounds were able to bind into papain-like protease active site. We consider results on this target less reliable that those obtained for the other targets.

Next we describe results for the first 5 compounds obtained for each target.

RNA-dependent RNA polimerase

1) CHEMBL38955
This compound has no name other than the IUPAC name and has no clinical studies on-going.
On CHEMBL is described as a potent and selective inhibitor of human renal renin.
RNA-dependent RNA polimerase is the 5th target in the profiling list with a predicted pChEMBL values of 8.7. JAK1, another COVID19 target is predicted to bind to CHEMBL38955 in the nanomolar range.
In the complex obtained from the docking experiment, 80% of the residues present in the binding site interact with CHEMBL38955

2) CHEMBL364574
This compound has no name other than the IUPAC name and has no clinical studies on-going.
On CHEMBL is described as a micromolar range inhibitor of Matrix Metallo Proteases.
RNA-dependent RNA polimerase is the 1st target in the profiling list and the predicted pChEMBL is 8.9. Within the profiling many metalloproteases appear but also histone deacetylases that might be considered targets for COVID19.
In the complex obtained from the docking experiment, 60% of the residues present in the binding site interact with CHEMBL364574

3) CHEMBL1940827
This compound has no name other than the IUPAC name and has no clinical studies on-going.
On CHEMBL is described as agonist of human beta2-adrenoceptor. No other data ar available.

RNA-dependent RNA polimerase is the 4th target in the profiling list and the predicted pChEMBL is 8.1.
In the complex obtained from the docking experiment, 95% of the residues present in the binding site interact with CHEMBL1940827

4) CHEMBL3957152
This compound has no name other than the IUPAC name and has no clinical studies on-going.
On CHEMBL is described an experiment yielding a pChEMBL 8.5 to Dual specificity protein kinase TTK.
RNA-dependent RNA polimerase is the 1st in the profiling list and the predicted pChEMBL is 8.7.
In the complex obtained from the docking experiment, 55% of the residues present in the binding site interact with CHEMBL3957152.

5) CHEMBL398150
This compound has no clinical studies on-going.
(R)-1-(4-(1-(3,5-bis(trifluoromethyl)benzylamino)-1-oxo-4-(4-phenylpiperidin-1-yl)butan-2-yl)thiazol-2-yl)-3-methylurea on CHEMBL is described as a C-C chemokine receptor type 2 antagonist.
RNA-dependent RNA polimerase is the 2nd in the profiling list and the predicted pChEMBL is 8.3.
JAK1, another COVID19 target is predicted to bind to CHEMBL398150 with a pChEMBL of 6.8.
In the complex obtained from the docking experiment, 65% of the residues present in the binding site interact with CHEMBL398150.

AP2-associated protein kinase 1

1) CHEMBL562594
This compound has no name and other than the IUPAC name has no clinical studies on-going.
On CHEMBL a very low affinity is described to Aldo-keto reductase family 1 member B1 (pChEMBL 4.07).
AP2-associated protein kinase 1 is the 1st in the profiling list and the predicted pChEMBL is 8.8.
Dipeptidyl peptidase IV, another COVID19 target is predicted to bind to CHEMBL562594 with a pChEMBL of 6.9.
In the complex obtained from the docking experiment, 75% of the residues present in the binding site interact with CHEMBL562594.

2) CHEMBL1469075
This compound has no clinical studies on-going.
For 2-(2,6-dimethylmorpholino)-3-phenylquinoline on CHEMBL a micromolar interaction is described to Rap guanine nucleotide exchange factor 4.
AP2-associated protein kinase 1 is the 1st in the profiling list and the predicted pChEMBL is 8.7.
Several other kinases are listed in the profiling output. Angiotensin AT1 receptors, another COVID19 potential target is predicted to bind CHEMBL1469075 with a pChEMBL of 6.254.
In the complex obtained from the docking experiment, 66% of the residues present in the binding site interact with CHEMBL1469075.

3) CHEMBL4076880
This compound has no name other than the IUPAC name and has no clinical studies on-going.

On CHEMBL activities against different member of the Adenosine receptor family in the submicromolar range are described.
AP2-associated protein kinase 1 is the 1st in the profiling list and the predicted pChEMBL is 8.4.
In the complex obtained from the docking experiment, 80% of the residues present in the binding site interact with CHEMBL4076880.

4) CHEMBL1581758
This compound has no clinical studies on-going.
2,6-dimethyl-4-(4-phenyl-2-quinolinyl)morpholine on CHEMBL is described as slightly active against ATP-dependent Clp protease proteolytic subunit.
AP2-associated protein kinase 1 is the 1st in the profiling list and the predicted pChEMBL is 8.8.
In the complex obtained from the docking experiment, 58% of the residues present in the binding site interact with CHEMBL1581758.

5) CHEMBL4094007
This compound has no name other than the IUPAC name and has no clinical studies on-going.
On CHEMBL is described its binding to adenosine A3 receptor with a pChEMBL of 6.9.
AP2-associated protein kinase 1 is the 1st in the profiling list and the predicted pChEMBL is 8.4.
In the complex obtained from the docking experiment, 75% of the residues present in the binding site interact with CHEMBL4094007.

3CLPro/Nsp5

1) MolPort-001-014-336
4-bromophenyl furan-2-carboxylate has no clinical studies on-going and no activities reported in CHEMBL.
RNA-dependent RNA polimerase and Sodium/hydrogen exchanger 1, that could be a potential target because COVID-19 may affect the endocrine pancreas by activating Na+/H+ exchanger 2 and increasing lactate levels, appear in the profiling output in the micromolal range.
This compound could be particularly interesting because it could work following different mechanisms.
3CLPro/Nsp5 is the 1st in the profiling list and the predicted pChEMBL is 7.4.
A Sea prediction reported on ZINC agrees with our prediction.
In the complex obtained from the docking experiment, 63% of the residues present in the binding site interact with MolPort-001-014-336.

2) CHEMBL1939869
5-Pyridin-3-yl-1H-indole-2-carboxylic acid ethyl ester has no clinical studies on-going.
On CHEMBL a submillimolar interaction with matrix metallo proteases is described.
3CLPro/Nsp5 is the 11th in the profiling list and the predicted pChEMBL is 7.3
In the complex obtained from the docking experiment, 72% of the residues present in the binding site interact with CHEMBL1939869.

3) CHEMBL1939870
5-Pyridin-4-yl-1H-indole-2-carboxylic acid ethyl ester differs from CHEMBL1939869 in the position of the nitrogen atom in the pyridin ring. Cannot be distinguish from it in term of profiling or docking.

4) CHEMBL1444763 6 weka: 7.271 dock: 0.681818181818182 ave: 6.8083
5-(4-chlorophenyl)-N-(pyridin-3-yl)furan-2-carboxamide has no clinical studies on-going.
On CHEMBL a very low activity against Thioredoxin glutathione reductase and Luciferin 4-monooxygenase is reported.
3CLPro/Nsp5 is the 6th in the profiling list and the predicted pChEMBL is 7.3.
Neuronal acetylcholine receptor, reported as a potential COVID19 target is the 1st of the profiling output with a predicted pChEMBL of 8.1.
In the complex obtained from the docking experiment, 68% of the residues present in the binding site interact with CHEMBL1444763.

5) CHEMBL3824099 1 weka: 7.208 dock: 0.681818181818182 ave: 6.74930909090909
[5-(2-chlorophenyl)furan-2-yl]-pyrazol-1-ylmethanone has no clinical studies on-going.
On CHEMBL a very low activity against cAMP-specific 3',5'-cyclic phosphodiesterase 4B is described.
3CLPro/Nsp5 is the 1st in the profiling list and the predicted pChEMBL is 7.2.
In the profiling output appear several histone deacetylases, that might be potential targets for COVID19, with predicted pChEMBL in the micromolar range.
In the complex obtained from the docking experiment, 68% of the residues present in the binding site interact with CHEMBL3824099.

Papain-like protease

In our view predictions for this target are less reliable than those obtained for the other targets. The ranking in the profiling list is much lower that the other three targets reported above and the percentage of binding site residues in contact with the predicted ligands are very low.
We will still report them in case they could be useful to add information to results obtained from other groups.

1) CHEMBL193810
1-[2-(4-Cyclohexyl-piperazin-1-yl)-2-(4-trifluoromethyl-phenyl)-ethyl]-3-(4-methoxy-phenyl)-urea has no clinical studies on-going.
On CHEMBL an interaction with C-C chemokine receptor type is reported having pChEMBL 5.7.
Papain-like protease is the 50th in the profiling list and the predicted pChEMBL is 7.0.
In the complex obtained from the docking experiment, 14% of the residues present in the binding site interact with CHEMBL193810.

2) CHEMBL3823073 31 weka: 6.832 dock: 0.148148148148148 ave: 5.66802962962963
N-[3-(2-amino-2-oxoethyl)phenyl]-1H-indazole-3-carboxamide has no clinical studies on-going.
On CHEMBL a low affinity for Casein kinase II subunit alpha is described
Papain-like protease is the 31th in the profiling list and the predicted pChEMBL is 6.8.
In the complex obtained from the docking experiment, 14% of the residues present in the binding site interact with CHEMBL3823073

3) PCIDB C00002593 6 weka: 6.76 dock: 0.148148148148148 ave: 5.6082962962963
5'-Demethoxydeoxypodophyllotoxin : no activity is described. It is a natural compound present in Anthriscus sylvestris and Bursera morelensis
Papain-like protease is the 6th in the profiling list and the predicted pChEMBL is 6.8.

In the complex obtained from the docking experiment, 14% of the residues present in the binding site interact with C00002593.

4) CHEMBL1568754 15 weka: 6.515 dock: 0.148148148148148 ave: 5.40503703703704
N-[3-(5-chloro-1,3-benzoxazol-2-yl)-2-methylphenyl]-2-[(4-chlorophenyl)thio]acetamide has no clinical studies on-going.
On CHEMBL a low activity against major cysteine proteinase is reported.
Papain-like protease is the 15th in the profiling list and the predicted pChEMBL is 6.5.
In the complex obtained from the docking experiment, 14% of the residues present in the binding site interact with CHEMBL1568754.

5) CHEMBL2180408 37 weka: 6.477 dock: 0.148148148148148 ave: 5.37351111111111
JNJ-42396302 completed phase I clinical trial.
On CHEMBL inhibition of PDE10A2 with a pChEMBL of 7.2 is reported.
Papain-like protease is the 37th in the profiling list and the predicted pChEMBL is 6.5.
Several kinases are reported in the profiling output. Among them also JAK1 (pChEMBL 6.6) that is listed as a potential COVID19 target.
In the complex obtained from the docking experiment, 14% of the residues present in the binding site interact with CHEMBL2180408.