Activity prediction of SARS-CoV-2 Mpro inhibitors based on ensemble docking and machine learning

Anastasia D. Fomina\textsuperscript{1,2}, Dmitry I. Osolodkin\textsuperscript{2}

\textsuperscript{1}Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia
\textsuperscript{2}FSASI “Chumakov FSC R&D IBP RAS”, Moscow, Russia

anastasiia.d.fomina@gmail.com
SARS-CoV-2 Main Protease

- Crucial for viral replication - processes polyprotein in 11 sites
- Highly conserved between different coronaviruses
- Substrate specificity to viral polyprotein
- A lot of available data
Objective

Our aim was to develop an approach for predicting the activity of SARS-CoV-2 3CLpro protease inhibitors based on ensemble docking and machine learning.

1st structure - 05.02.2020
To date - more than 400 structures in PDB!

Our aim was to develop an approach for predicting the activity of SARS-CoV-2 3CLpro protease inhibitors based on ensemble docking and machine learning.
I  Classic docking

II  Ensemble docking

III  Constrained ensemble docking

IV  Structure-based ML approach
Docking a library of compounds proposed by generative topographic mapping

PDB ID: 6wqf

473 GTM hits 39 docking hits with grid Score <-50 10 compounds tested

Experimental evaluation

10 hits from the virtual screening were bought and tested on the inhibitory activity against Mpro SARS-CoV-2. Based on the docking of the two most active hits’ analogs, the new extended series of compounds was selected, bought and experimentally tested.

Ensemble docking

Classic docking

Constrained ensemble docking

Structure-based ML approach
Library of compounds with known activity against coronaviral Mpro

Sources: articles and preprints published between July 2020 and March 2021

This library was used for docking validation and comparison in further experiments.

Viruses

- IBV 1.9%
- PEDV 6.1%
- FIPV 8.3%
- HCoV-229E 8.3%
- HCoV-NL63 3.2%
- MERS-CoV 21.1%
- SARS-CoV 44.1%
- SARS-CoV-2 7.0%

July 2020

2105 compounds

- 751 Active against Mpro
- 370 Inactive against Mpro
- 984 Activity against Mpro unknown
- 1106 Activity against virus unknown
- 717 Active against virus
- 282 Inactive against virus

March 2021
The First Ensemble

- structures from the same series with minimal differences

PDB - 168 structures

- selection by mean RMSD and conformational diversity of active site residues (visual analysis)

57 structures

August-September 2020

22 structures
Docking results analysis - first ensemble

No big difference between docking into different structures + too many structures.

We needed a stricter approach
The Second Ensemble

- mutants
- immature forms
- oxidized forms
- unresolved parts

- selection by the diversity of active site residues conformations

PDB - 168 structures
57 structures
22 structures
- mutants
- immature forms
- oxidized forms
- unresolved parts

March 2021

Pairwise RMSD between active site residues (222x222)

Download → Filter → 3D-align Cα chain A → Calculate active site RMSDs → RMSD distribution → Pick the most diverse pairs

New data + automatisation

Download Filter 3D-align Cα chain A Calculate active site RMSDs RMSD distribution Pick the most diverse pairs

283 PDB - 168 structures
222 57 structures
7 22 structures

222
The second ensemble

Download → Filter → 3D-align Cα chain A → Calculate active site RMSDs → RMSD distribution → Pick the most diverse pairs

99% percentile - RMSD > 1.4046
75% percentile - RMSD < 0.9657

✅ Scalable
✅ Semi-automated
✅ Not too many structures
The Third Ensemble: the same algorithm - more data

Download → Filter → 3D-align Ca chain A → Calculate active site RMSDs → RMSD distribution → Pick the most diverse pairs

- PDB - 283 structures (416 - 306 = 110)
- mutants, immature forms, oxidized forms, unresolved parts
- selection by the diversity of active site residues conformations

January 2022

222 structures (283 - 61 = 222)
I
Classic docking

II
Ensemble docking

III
Constrained ensemble docking

IV
Structure-based ML approach
Interaction Fingerprints

Data: holo-structures from PDB

Interaction Fingerprints were calculated using Flare 4.0.2 python interpreter. Three types of interactions were considered: h-bonds, C-C interactions (good), C-heteroatom interactions (bad).

Co-crystallized ligands form the most
- H-bonds with Glu166 and His163
- good contacts with His41, Met49 and Leu141
- bad contacts with Phe140, Asn 142 and Ser 144
Constrained docking

Constraints:
- Glu166 (H-bond donor),
- His163 (H-bond donor)

Method:
Dock6.9 chemical matching

Result:
No difference from unconstrained docking
Interaction Fingerprints

**Custom Flare-based IFPs**

0 - no contact  
-1 - C-hetero contact  
1 - C-C contact  
2 - H-bond

```
resid 1  resid 2  resid 3  resid 4  resid 5  resid N
0  -1  2  0  1  ...  0
```

**Vanilla ODDT-based IFPs**

Hydrophobic  
Arom. face-to-face  
Arom. edge to face  
H-bonds donator  
H-bonds acceptor  
Salt bridge +  
Salt bridge -  
Salt bridge metall

```
0  0  1  0  0  0  0  0  0  ...  
```

Docking

```
Interaction Fingerprints  
Docking Scores  
Morgan Fingerprints
```

Model

Training data

Assay Report Card

Basic Information

<table>
<thead>
<tr>
<th>Assay ID</th>
<th>CHMPB4495552</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Functional</td>
</tr>
<tr>
<td>Description</td>
<td>SARS-CoV-2 3CLpro protease inhibition percentage at 20μM by FRET-based assay of substrate cleavage</td>
</tr>
<tr>
<td>Format</td>
<td>ISO_0900019</td>
</tr>
<tr>
<td>Journal</td>
<td>[2021] -</td>
</tr>
<tr>
<td>Organism</td>
<td>Severe acute respiratory syndrome coronavirus 2</td>
</tr>
<tr>
<td>Method</td>
<td>---</td>
</tr>
<tr>
<td>Tissue</td>
<td>---</td>
</tr>
<tr>
<td>Cell Type</td>
<td>---</td>
</tr>
<tr>
<td>Substratar/Pronection</td>
<td>---</td>
</tr>
<tr>
<td>Target</td>
<td>CHMPB4495552</td>
</tr>
<tr>
<td>Document</td>
<td>CHMPB4495554</td>
</tr>
<tr>
<td>Cell</td>
<td>---</td>
</tr>
<tr>
<td>Tissue</td>
<td>---</td>
</tr>
</tbody>
</table>

Distribution of inhibition percent

- ≥ 60% - active
- < 60% - inactive

ABSTRACT: Compound repurposing is an important strategy for the identification of effective treatment options against SARS-CoV-2 infection and COVID-19 disease. In this regard, SARS-CoV-2 main protease (3CL-Pro), also termed M-Pro, is an attractive drug target as it plays a central role in viral replication by processing the viral polyproteins pp1a and pp1ab at multiple distinct cleavage sites. We here report the results of a repurposing program involving 8,7 K compounds containing marketed drugs, clinical and preclinical candidates, and small molecules regarded as safe in humans. We confirmed previously reported inhibitors of 3CL-Pro and have identified 62 additional compounds with IC_{50} values below 1 μM and profiled their selectivity toward chymotrypsin and 3CL-Pro from the Middle East respiratory syndrome virus. A subset of eight inhibitors showed antiviral activity in a Vero-E6 cell line, and the compounds thiopurine and MG-132 were analysed for their predicted binding characteristics to SARS-CoV-2 3CL-Pro. The X-ray crystal structure of the complex of myricetin and SARS-CoV-2 3CL-Pro was solved at a resolution of 1.77 Å, showing that myricetin is covalently bound to the catalytic Cys45 and therefore inhibits its enzymatic activity.

KEYWORDS: SARS-CoV-2, main protease, screening, FRET, repurposing

<table>
<thead>
<tr>
<th>8702 (inorganic)</th>
<th>6897 (organometallics)</th>
</tr>
</thead>
<tbody>
<tr>
<td>221 active</td>
<td></td>
</tr>
<tr>
<td>6676 inactive</td>
<td></td>
</tr>
</tbody>
</table>
Datasets

ODDT IFP

512 Morgan FP

Set6 Set7 Set8

Grid Scores, Mw, logP

1024 Morgan FP

Set9

Flare IFP

Set2

Ligand Efficiencies, Mw, logP

Set1 Set3 Set4
ROC-curves

**ROC-curve Logistic Regression**

**ROC-curve Random Forest**

**ROC-curve Catboost**

**ROC-curve SVM**

**ROC-curve Tabnet**

**Best models by test AUC**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>AUC train</th>
<th>AUC test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set9</td>
<td>LR</td>
<td>0.98±0.00</td>
<td>0.71±0.03</td>
</tr>
<tr>
<td>Set8</td>
<td>LR</td>
<td>0.98±0.00</td>
<td>0.71±0.03</td>
</tr>
<tr>
<td>Set8</td>
<td>RF</td>
<td>0.97±0.01</td>
<td>0.71±0.06</td>
</tr>
<tr>
<td>Set8</td>
<td>Catboost</td>
<td>1.00±0.00</td>
<td>0.70±0.06</td>
</tr>
<tr>
<td>Set9</td>
<td>RF</td>
<td>0.97±0.01</td>
<td>0.70±0.06</td>
</tr>
</tbody>
</table>
### Best models

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>ROCAUC train</th>
<th>ROCAUC test</th>
<th>F1 train</th>
<th>F1 test</th>
<th>Precision1 test</th>
<th>Recall1 test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set7</td>
<td>ODDT IFP</td>
<td>LR</td>
<td>0.71±0.01</td>
<td>0.63±0.05</td>
<td>0.06±0.00</td>
<td>0.06±0.00</td>
<td>1.00±0.00</td>
</tr>
<tr>
<td>Set4</td>
<td>Morgan 512, Flare IFP</td>
<td>SVM</td>
<td>0.80±0.01</td>
<td>0.68±0.04</td>
<td>0.10±0.01</td>
<td>0.08±0.01</td>
<td>0.75±0.07</td>
</tr>
<tr>
<td>Set8</td>
<td>Morgan 1024</td>
<td>SVM</td>
<td>0.89±0.01</td>
<td>0.71±0.02</td>
<td>0.19±0.01</td>
<td>0.13±0.01</td>
<td>0.57±0.07</td>
</tr>
<tr>
<td>Set8</td>
<td>Morgan 1024</td>
<td>LR</td>
<td>0.86±0.01</td>
<td>0.70±0.04</td>
<td>0.18±0.01</td>
<td>0.13±0.01</td>
<td>0.56±0.06</td>
</tr>
<tr>
<td>Set9</td>
<td>Morgan 1024, ODDT IFP</td>
<td>LR</td>
<td>0.86±0.01</td>
<td>0.70±0.04</td>
<td>0.18±0.01</td>
<td>0.13±0.01</td>
<td>0.56±0.06</td>
</tr>
</tbody>
</table>
Conclusion

● A simple docking model has shown the best results yet
● We developed a consensus docking approach and use it in routine research
● Constrained docking with DOCK6 chemical matching shows the same results, as unconstrained
● Classification ML approach didn’t work out - to be continued…

Acknowledgements

This study was supported by the Non-commercial Foundation for the Advancement of Science and Education INTELLECT and State Research Funding № FNZG-2022-0002.

I thank Cresset team for Flare academic license.

I thank my colleagues for their help and for making this work possible
Thank you for your attention!