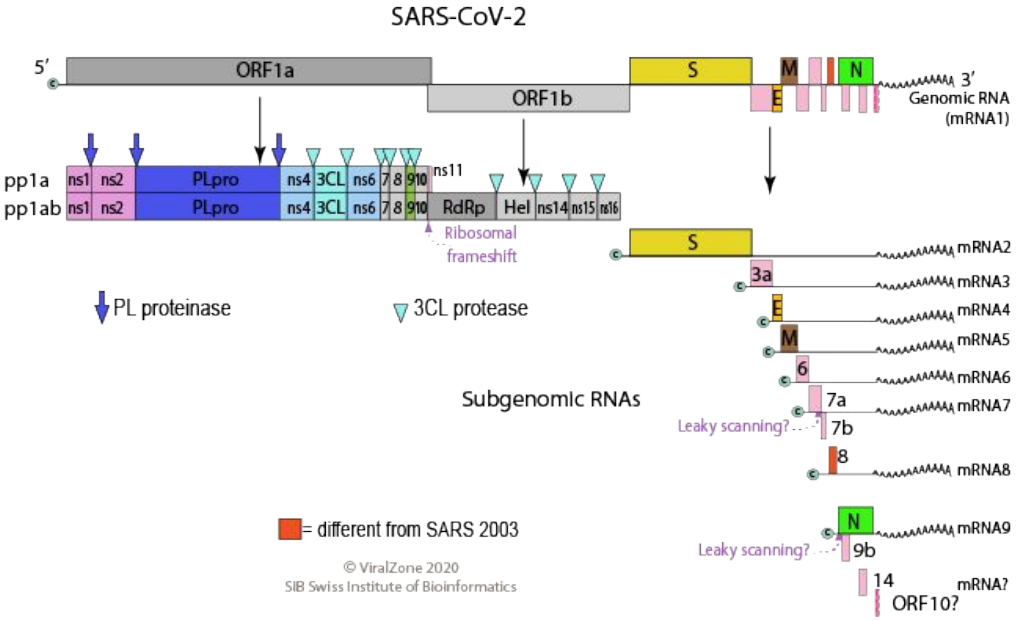
Activity prediction of SARS-CoV-2 Mpro inhibitors based on ensemble docking and machine learning

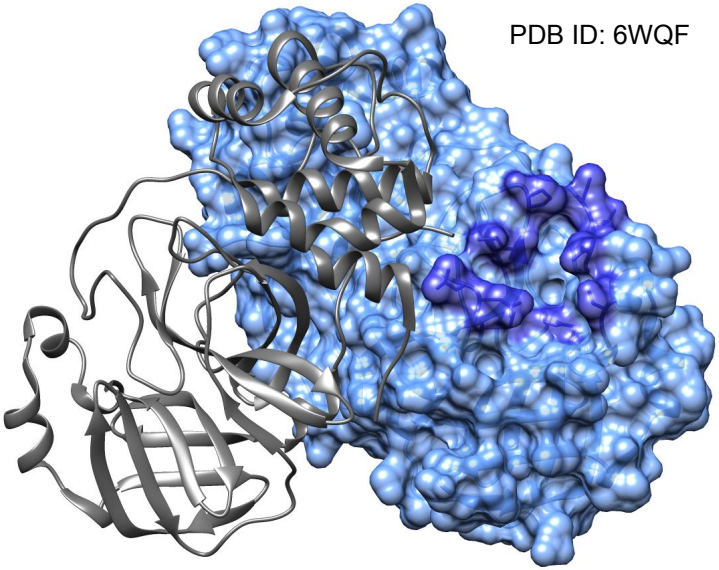
Anastasia D. Fomina[1,2], Dmitry I. Osolodkin[2]

[1]Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia
[2]FSASI “Chumakov FSC R&D IBP RAS”, Moscow, Russia

anastasiia.d.fomina@gmail.com

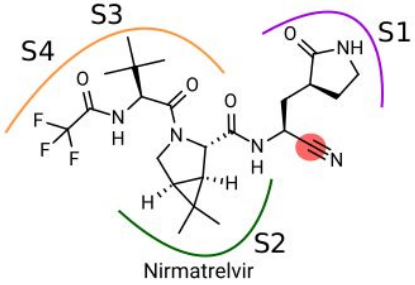
SARS-CoV-2 Main Protease

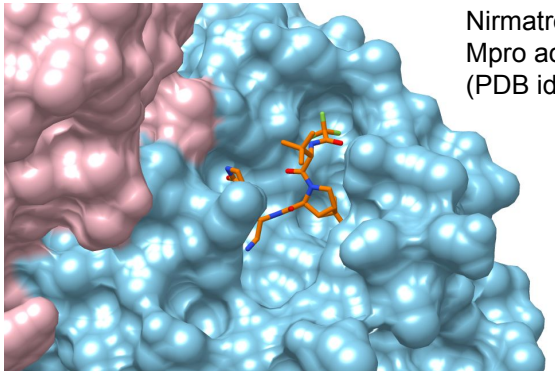




PDB ID: 6WQF

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



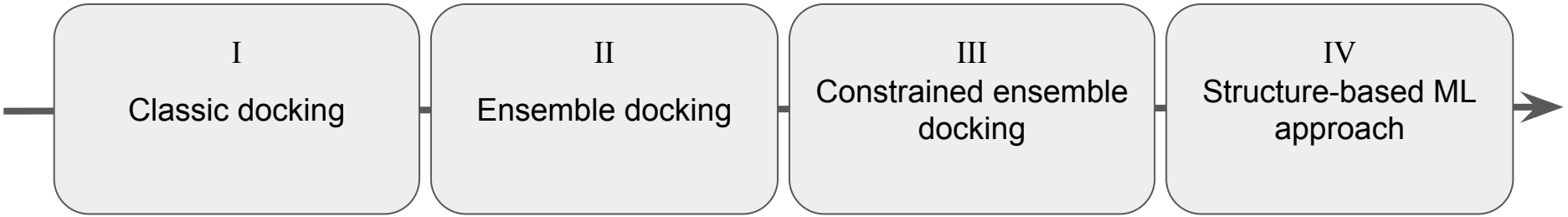


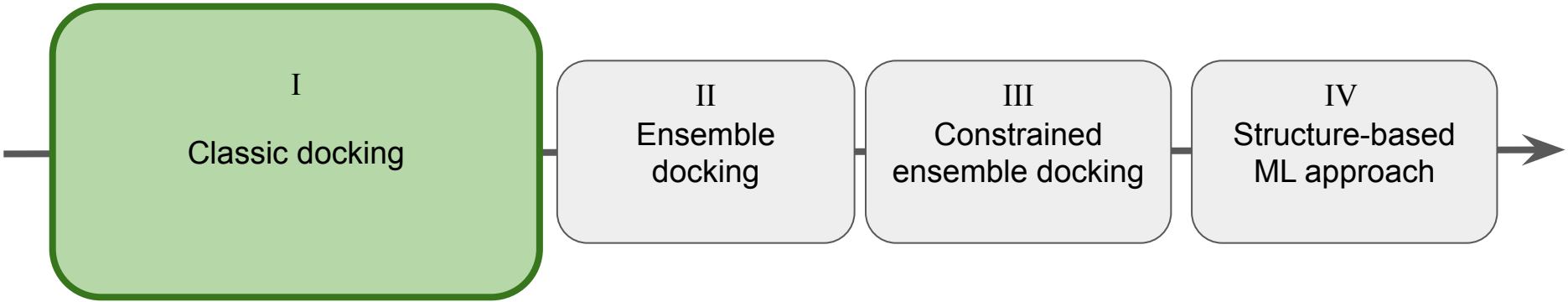
Nirmatrelvir in Mpro active site (PDB id: 4TLL)

Objective

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.

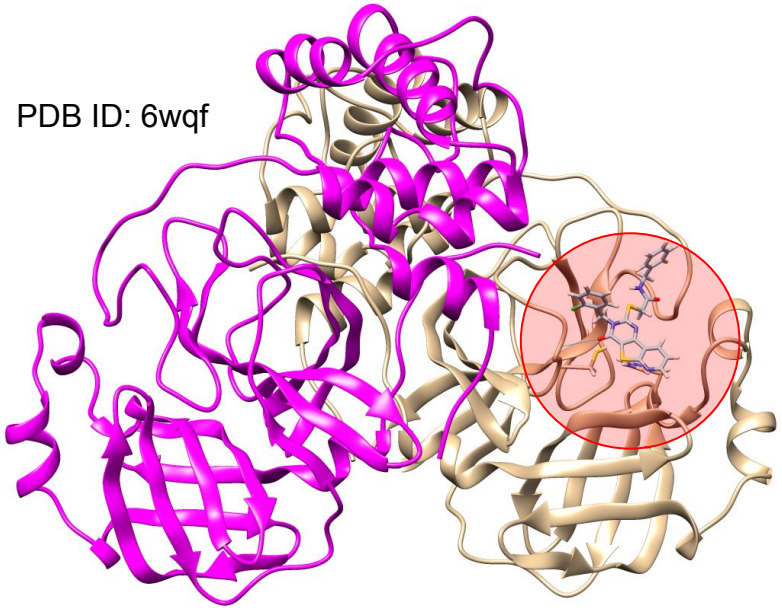


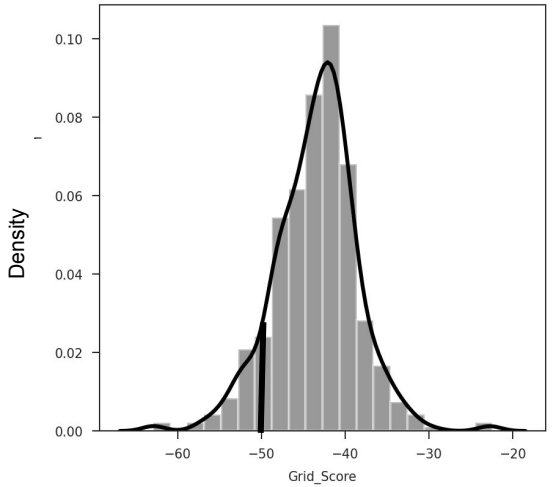

I
Classic docking
II
Ensemble docking
III
Constrained ensemble docking
IV
Structure-based ML approach

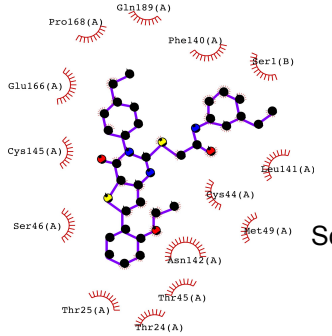
Docking a library of compounds proposed by generative topographic mapping

I

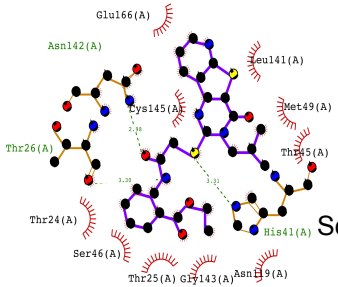
PDB ID: 6wqf



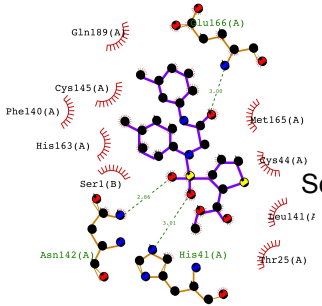




Scaffold 1



Scaffold 2



Scaffold 3

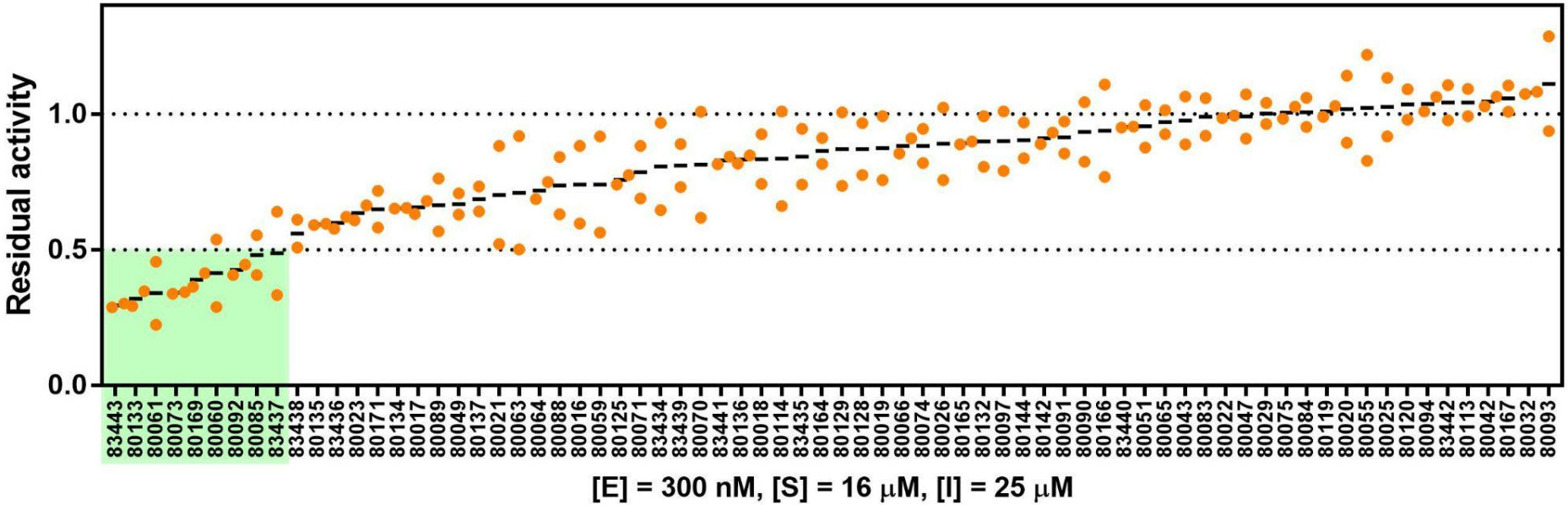
473 GTM hits → 39 docking hits with grid Score <-50 → 10 compounds tested

D. Horvath, A. Orlov, D. I. Osolodkin, A. A. Ishmukhametov, G. Marcou, A. Varnek, *Mol. Inf.* **2020**, *39*, 2000080.

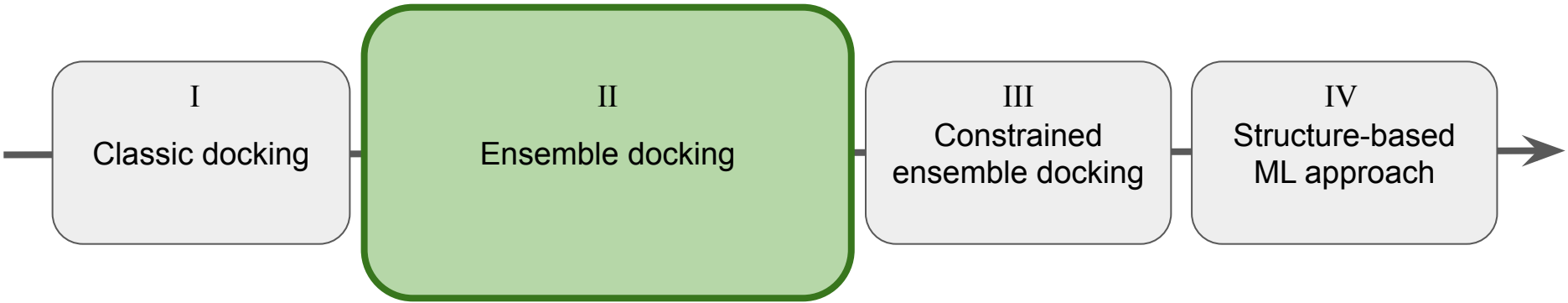
Experimental evaluation

I

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.



Zakharova M. Yu. et al. Pre-Steady-State Kinetics of the SARS-CoV-2 Main Protease as a Powerful Tool for Antiviral Drug Discovery // Frontiers in Pharmacology. 2021. (12). C. 773198.

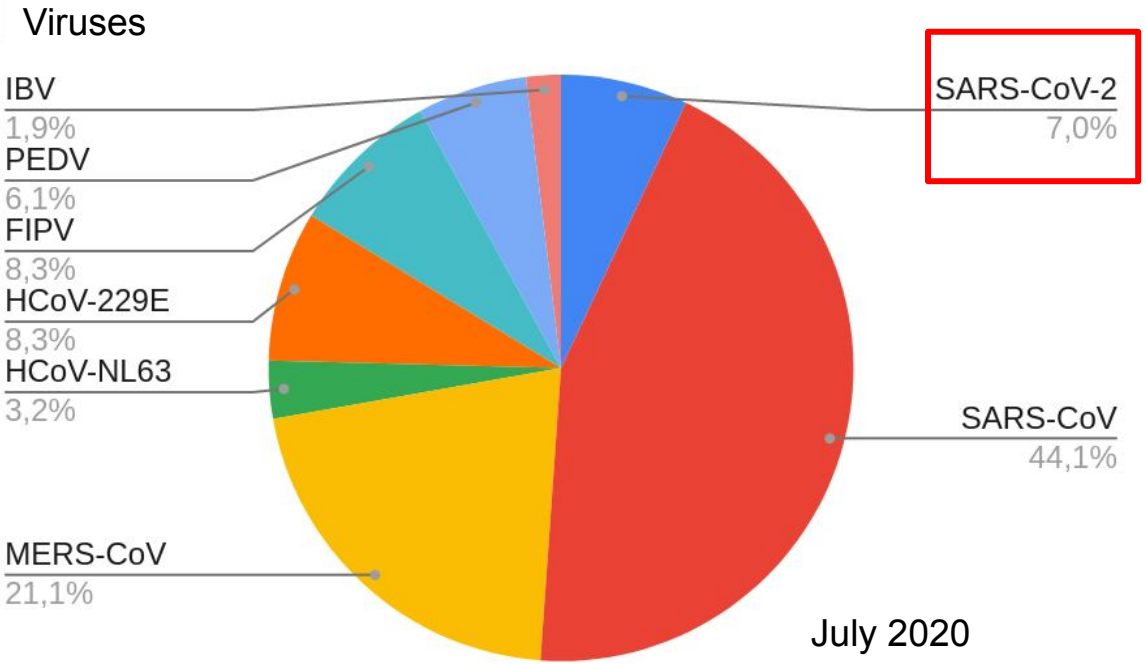

I
Classic docking
II
Ensemble docking
III
Constrained ensemble docking
IV
Structure-based ML approach

Library of compounds with known activity against coronaviral Mpro

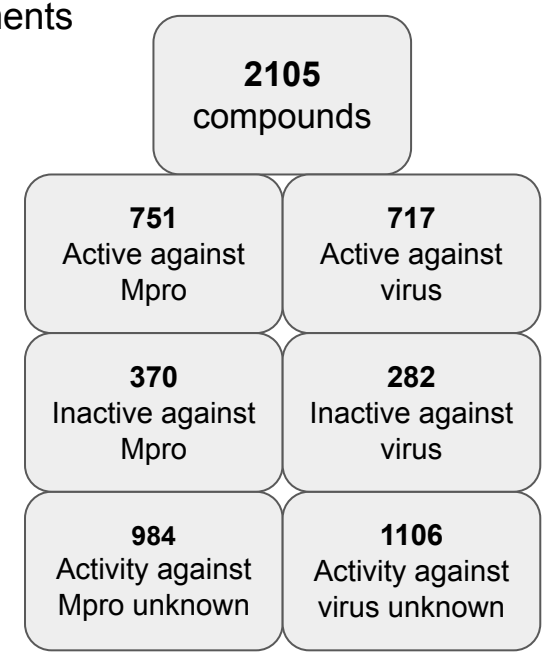
II

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

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



July 2020



March 2021

The First Ensemble

August-September 2020

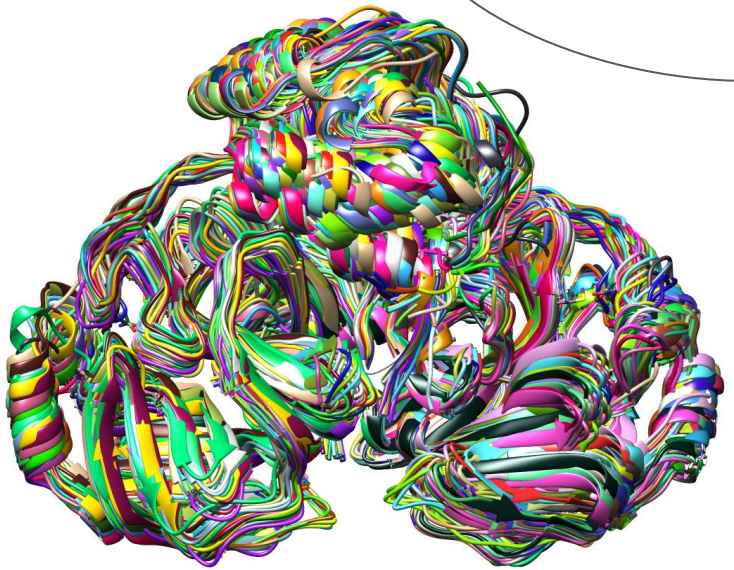
PDB - 168 structures

- structures from the same series with minimal differences

57 structures

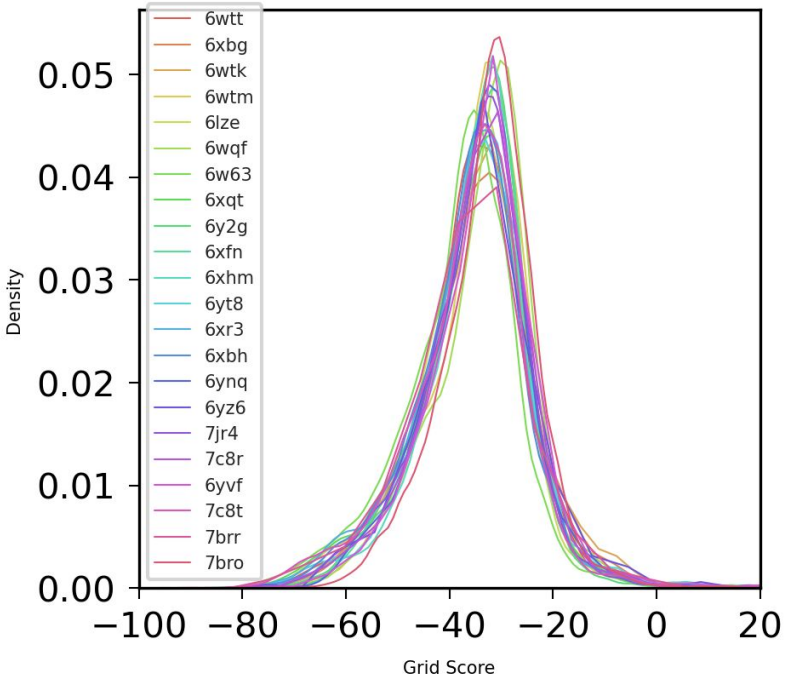
- selection by mean RMSD and conformational diversity of active site residues (visual analysis)

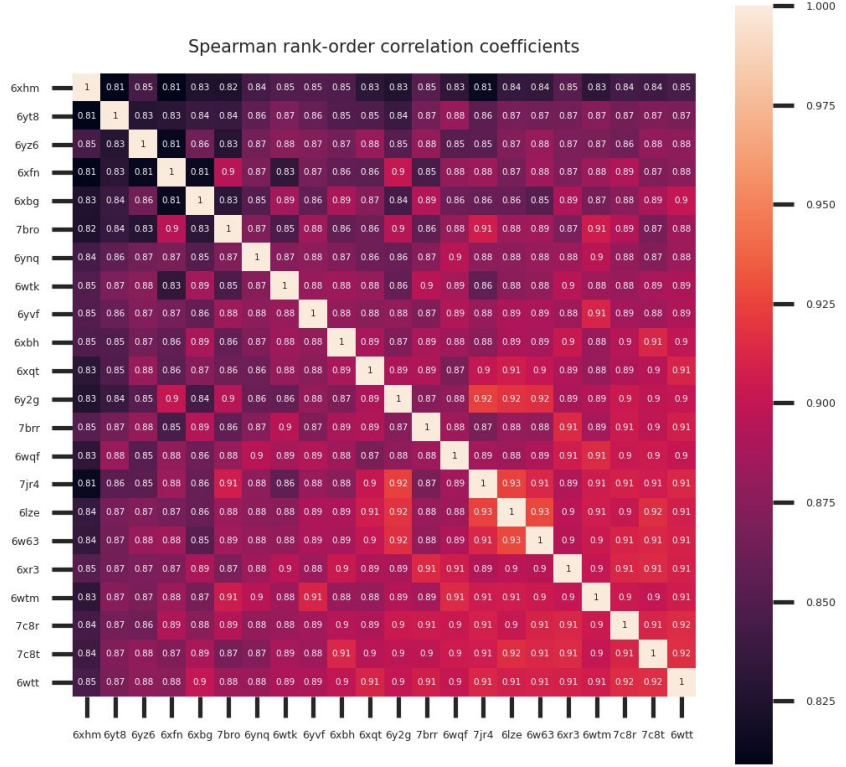
22 structures



Docking results analysis - first ensemble

II





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

We needed a stricter approach

The Second Ensemble

New data + automatisation

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

PDB - ~~168~~ 283 structures

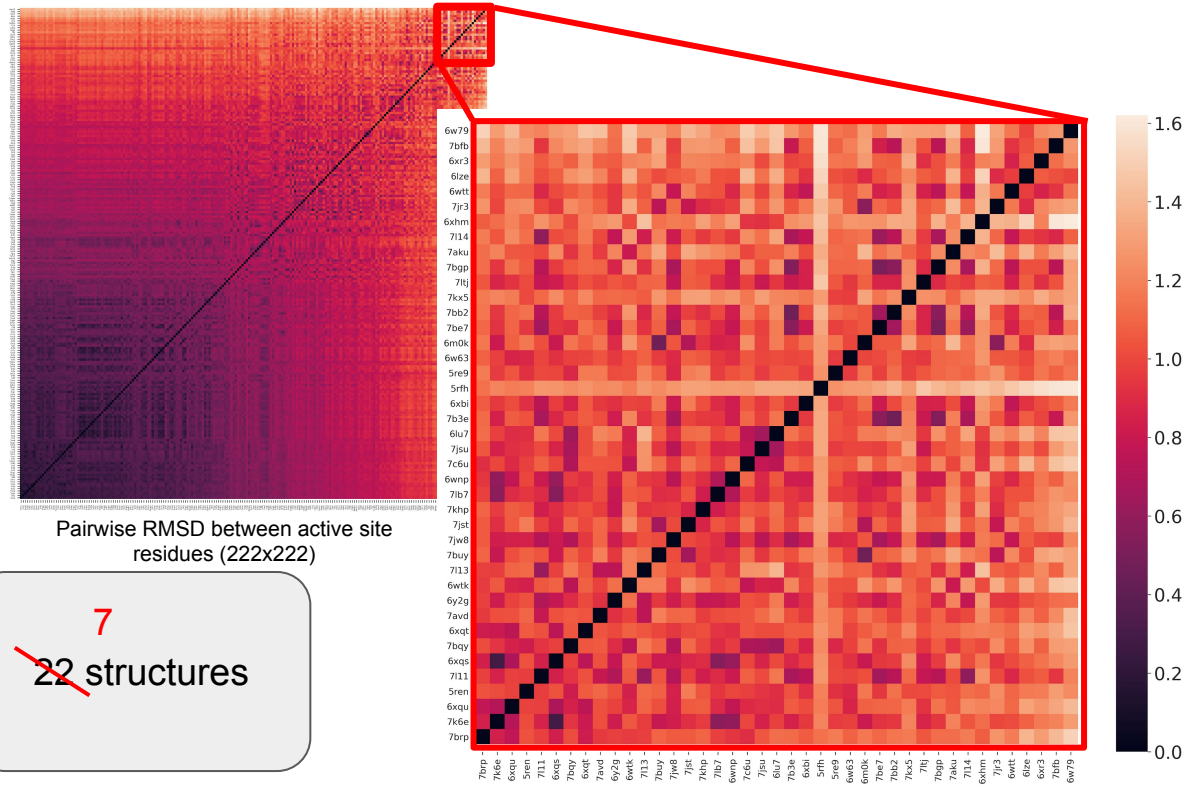
March 2021

- mutants
- immature forms
- oxidized forms
- unresolved parts

~~57~~ 222 structures

- selection by the diversity of active site residues conformations

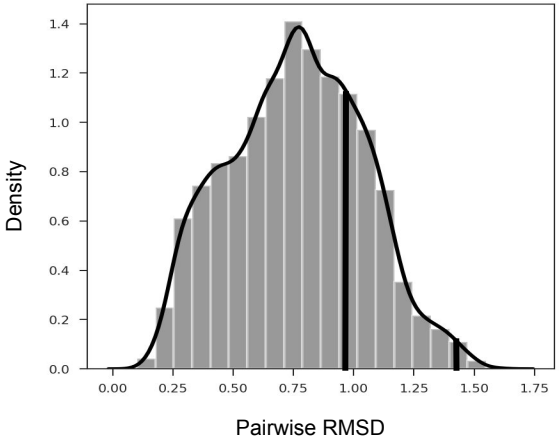
~~22~~ 7 structures



Pairwise RMSD between active site residues (222x222)

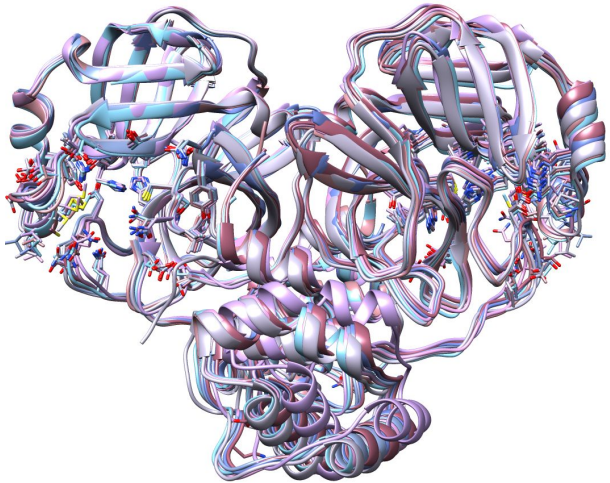
The second ensemble

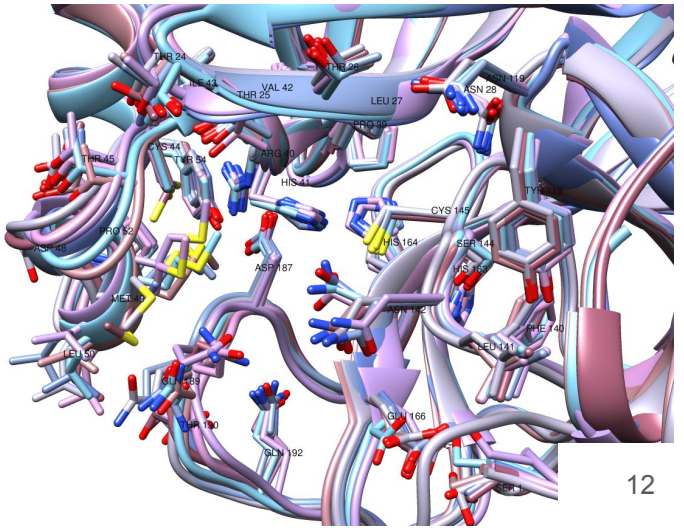
Download → Filter → 3D-align Cα chain A → Calculate active site RMSDs → RMSD distribution → Pick the most diverse pairs



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

II

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

416
~~283~~ PDB structures

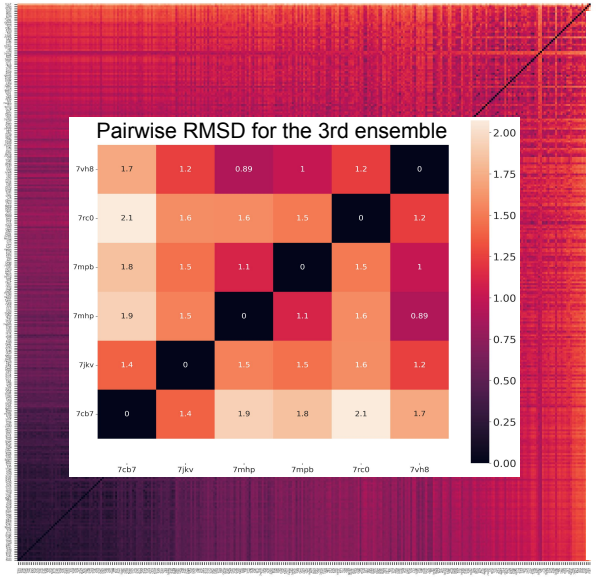
January 2022

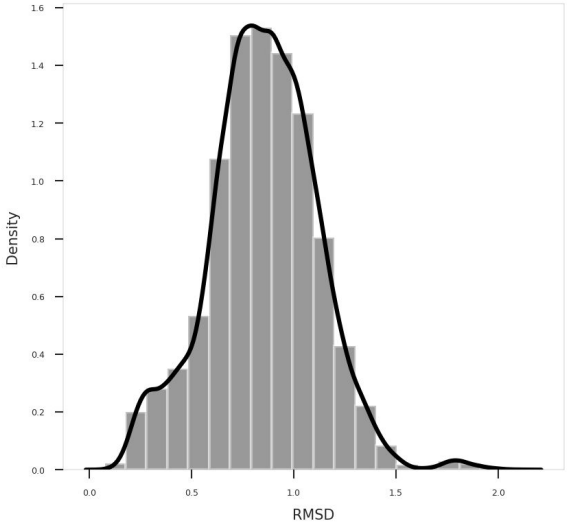
- mutants
- immature forms
- oxidized forms
- unresolved parts

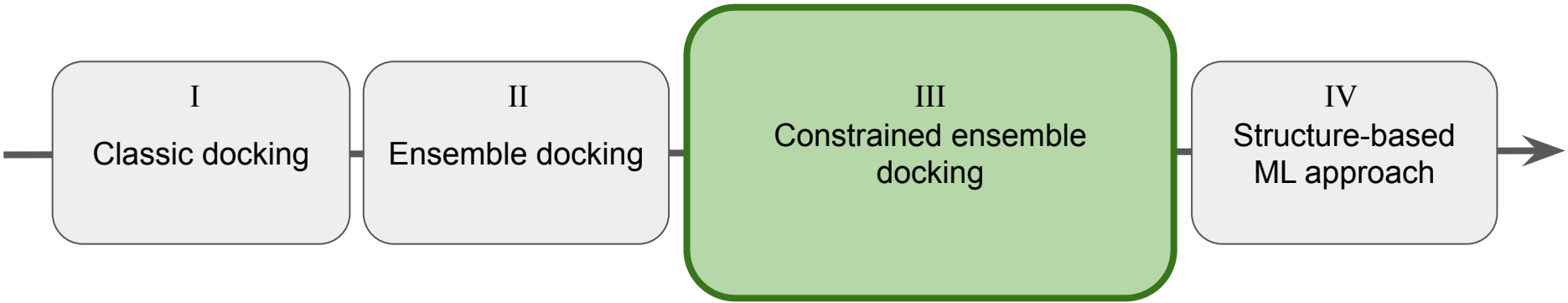
306
~~222~~ structures

- selection by the diversity of active site residues conformations

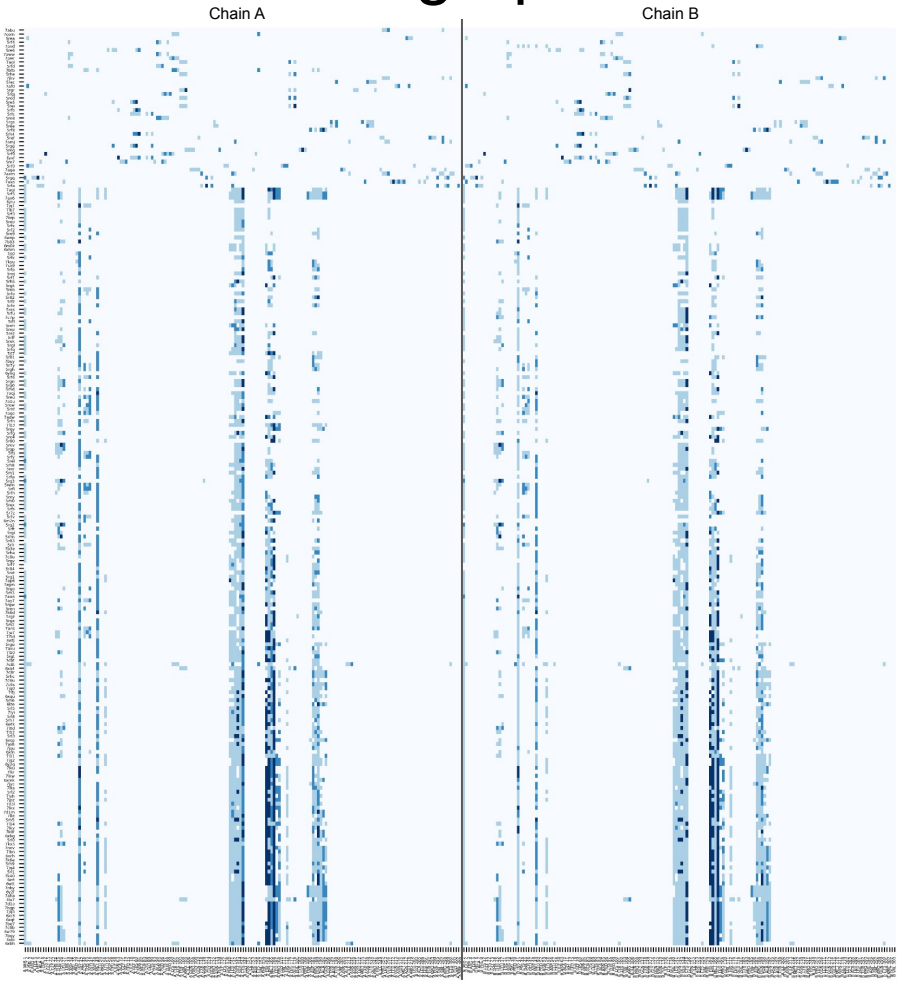
6
~~7~~ structures

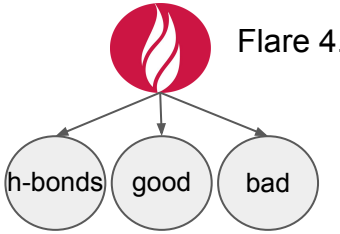





I
Classic docking
II
Ensemble docking
III
Constrained ensemble docking
IV
Structure-based ML approach

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),

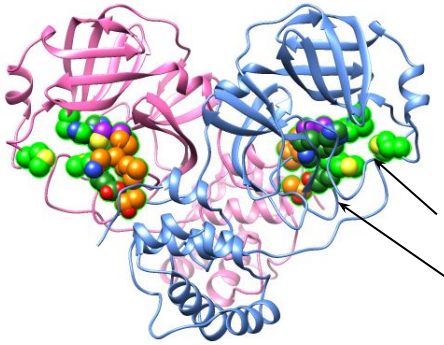
His163 (H-bond donor)

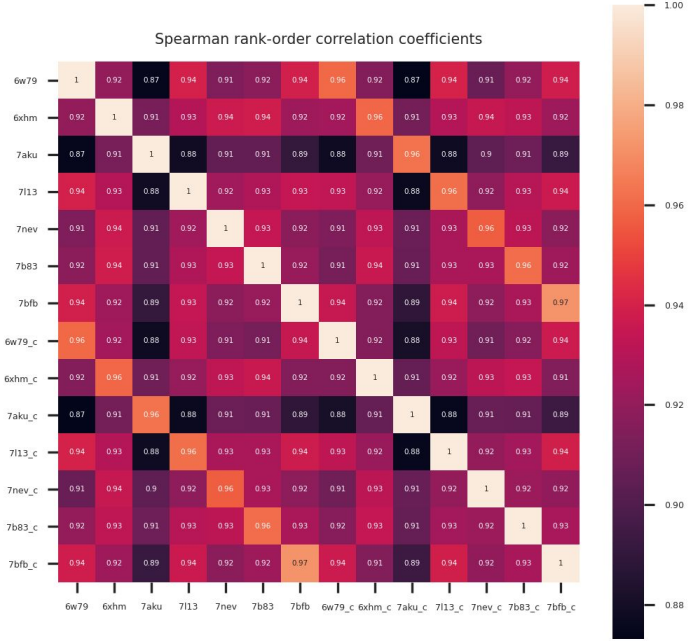
Method:

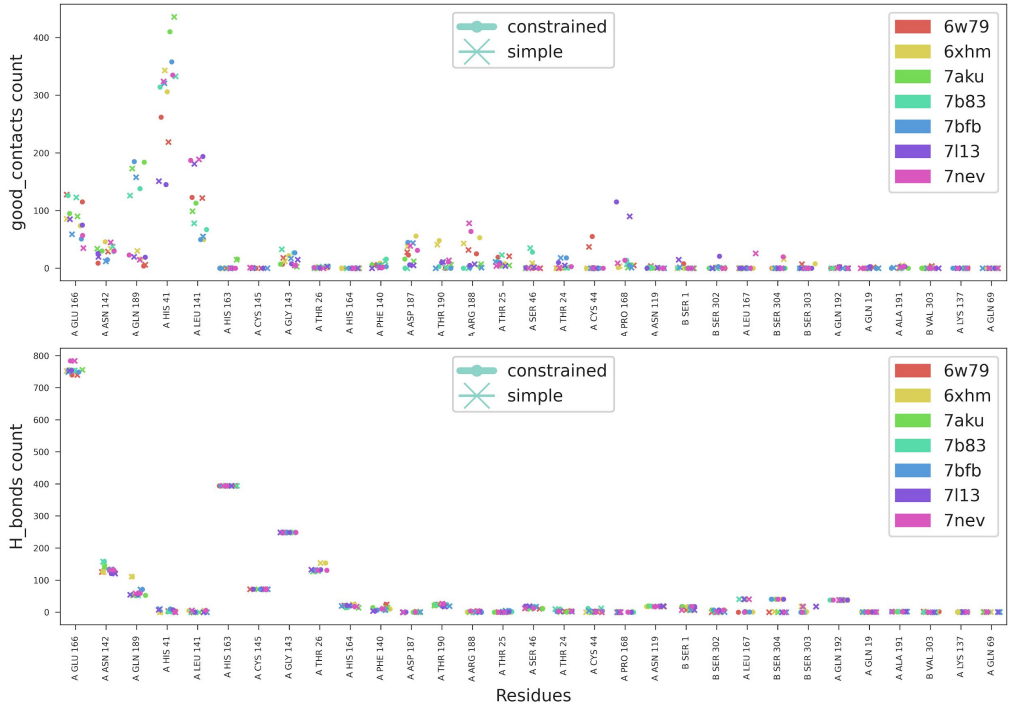
Dock6.9 chemical matching

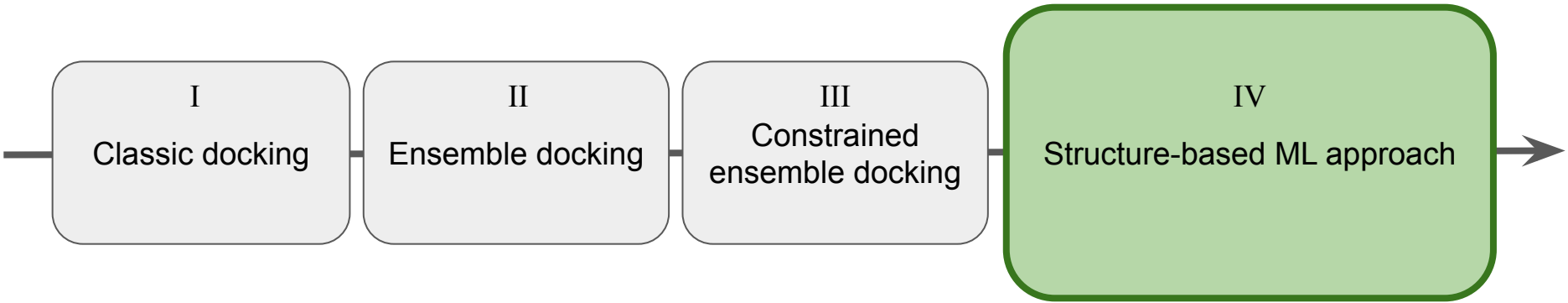
Result:

No difference from unconstrained docking





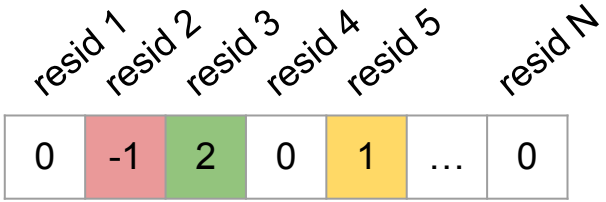



I
Classic docking
II
Ensemble docking
III
Constrained ensemble docking
IV
Structure-based ML approach

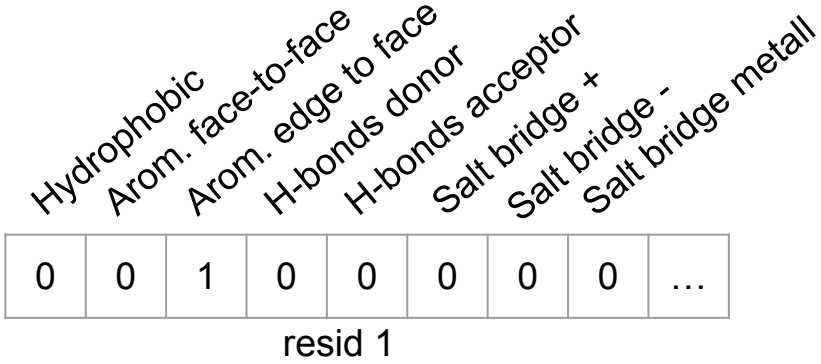
Interaction Fingerprints

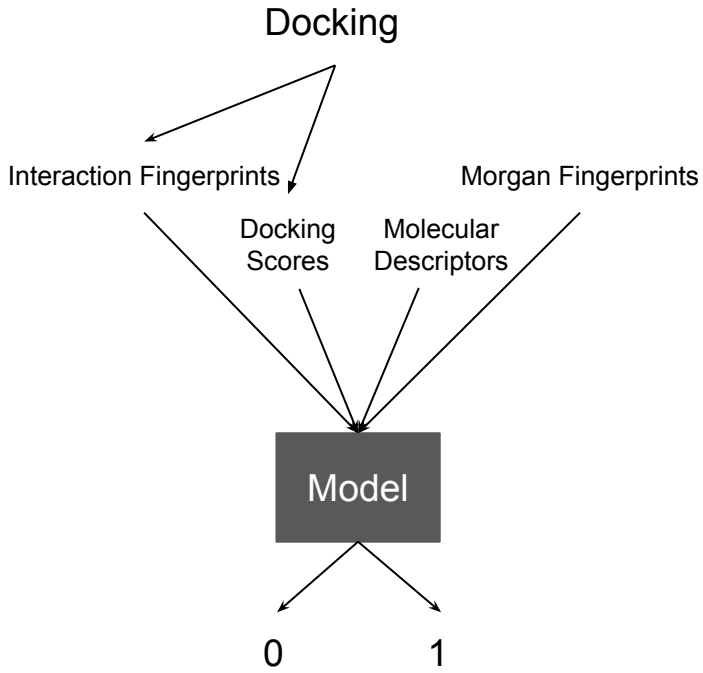
Custom Flare-based IFPs

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



***Vanilla* ODDT-based IFPs**





Wójcikowski, M.; Zielenkiewicz, P.; Siedlecki, P. Open Drug Discovery Toolkit (ODDT): A New Open-Source Player in the Drug Discovery Field. *J Cheminform* **2015**, *7* (1), 26.

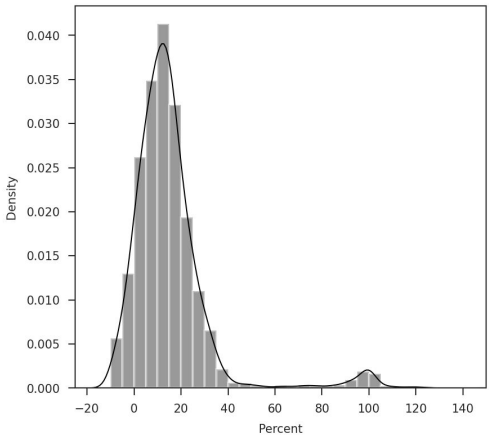
Training data

IV

Assay Report Card

Basic Information

Assay ID:	CHEMBL4495582
Type:	Functional
Description:	SARS-CoV-2 3CL-Pro protease inhibition percentage at 20μM by FRET kind of response from peptide substrate
Format:	BAO_0000019
Journal:	(2020) -
Organism:	Severe acute respiratory syndrome coronavirus 2
Strain:	---
Tissue:	---
Cell Type:	---
Subcellular Fraction:	---
Target:	CHEMBL4523582
Document:	CHEMBL4495564
Cell:	
Tissue:	



Distribution of inhibition percent

≥ 60% - active
< 60% - inactive

ACS Pharmacology & Translational Science

pubs.acs.org/ptsci Article

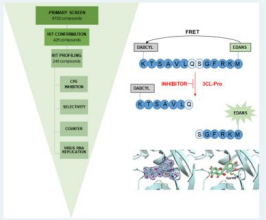
Identification of Inhibitors of SARS-CoV-2 3CL-Pro Enzymatic Activity Using a Small Molecule in Vitro Repurposing Screen

Maria Kuzikov,* Elisa Costanzi, Jeanette Reinshagen, Francesca Esposito, Laura Vangeel, Markus Wolf, Bernhard Ellinger, Carsten Claussen, Gerd Geisslinger, Angela Corona, Daniela Iaconis, Carmine Talarico, Candida Manelfi, Rolando Cannalire, Giulia Rossetti, Jonas Gossen, Simone Albani, Francesco Musiani, Katja Herzog, Yang Ye, Barbara Giabbai, Nicola Demitri, Dirk Jochmans, Steven De Jonghe, Jasper Rymenants, Vincenzo Summa, Enzo Tramontano, Andrea R. Beccari, Pieter Leyssen, Paola Storici, Johan Neyts, Philip Gribbon, and Andrea Zaliani

Cite This: *ACS Pharmacol. Transl. Sci.* 2021, 4, 1096−1110 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

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 anticytopathic effect in a Vero-E6 cell line, and the compounds thioguanosine and MG-132 were analyzed 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 Cys145 and therefore inhibiting its enzymatic activity.



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

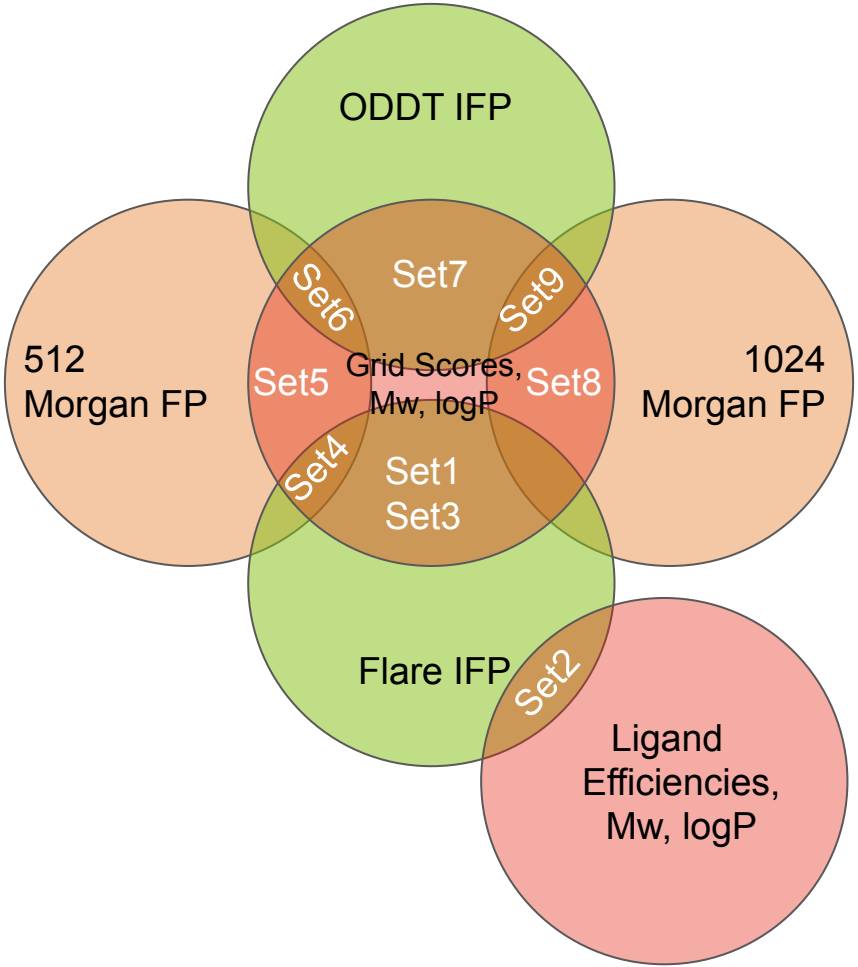
8702 → 6897

- inorganic
- organometallics

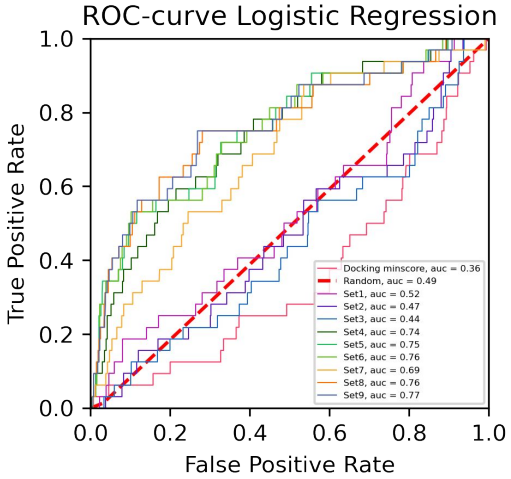
→ 221 active

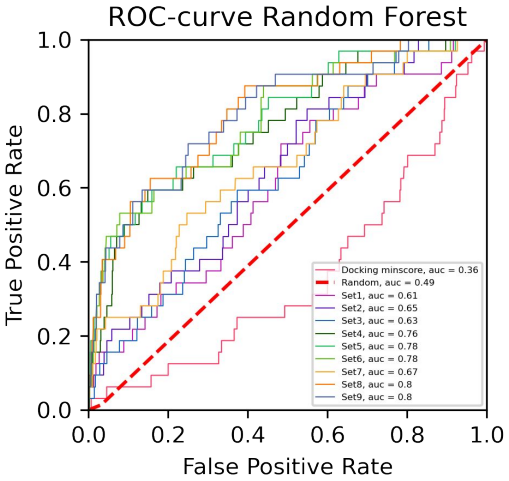
→ 6676 inactive

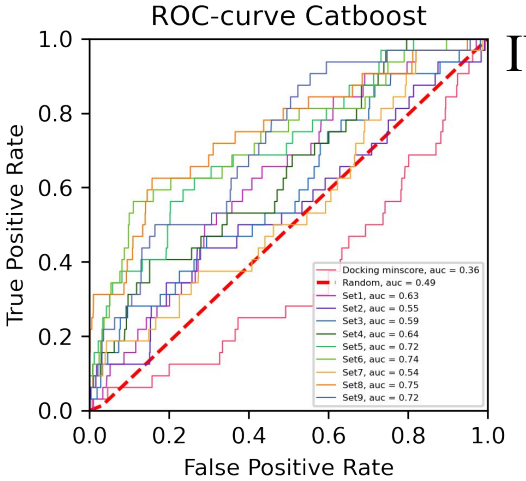
Datasets


ODDT IFP
512
Morgan FP
1024
Morgan FP
Flare IFP
Ligand
Efficiencies,
Mw, logP
Set7
Set6
Set9
Set5
Grid Scores,
Mw, logP
Set8
Set4
Set1
Set3
Set2

ROC-curves



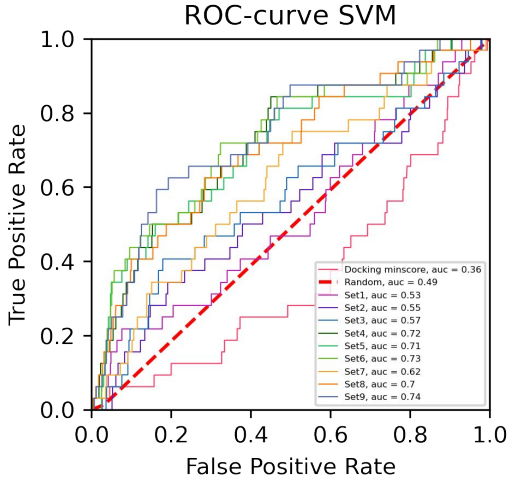


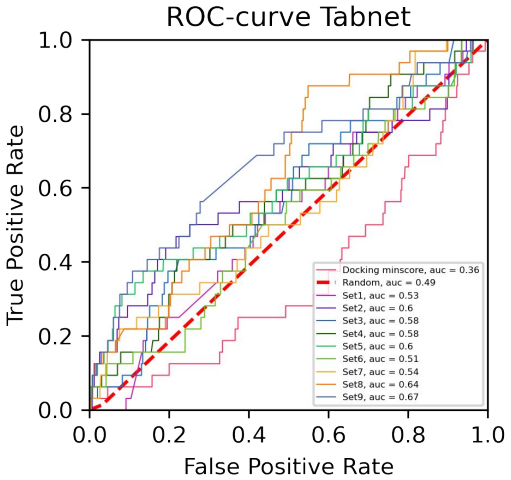


IV

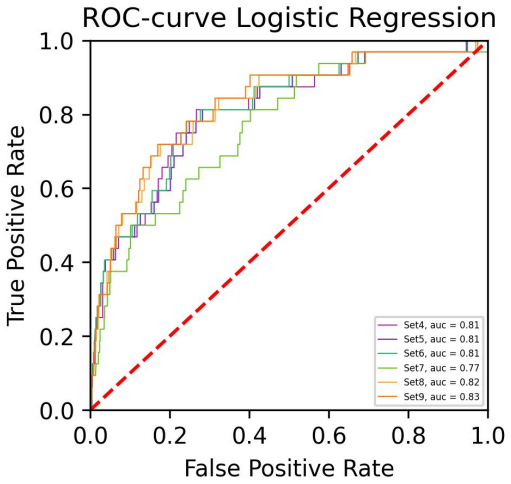
Best models by test AUC

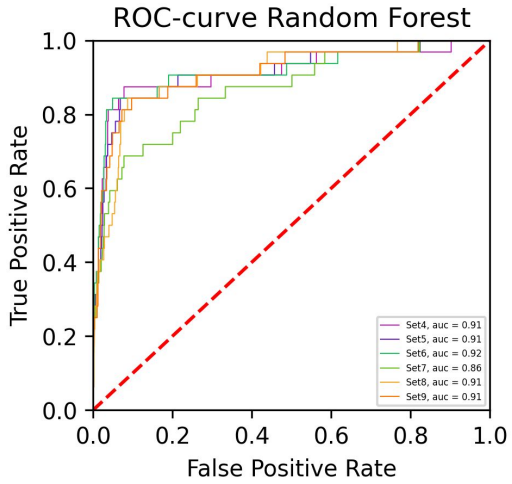
Dataset	Model	AUC train	AUC test
Set9	LR	0.98±0.00	0.71±0.03
Set8	LR	0.98±0.00	0.71±0.03
Set8	RF	0.97±0.01	0.71±0.06
Set8	Catboost	1.00±0.00	0.70±0.06
Set9	RF	0.97±0.01	0.70±0.06

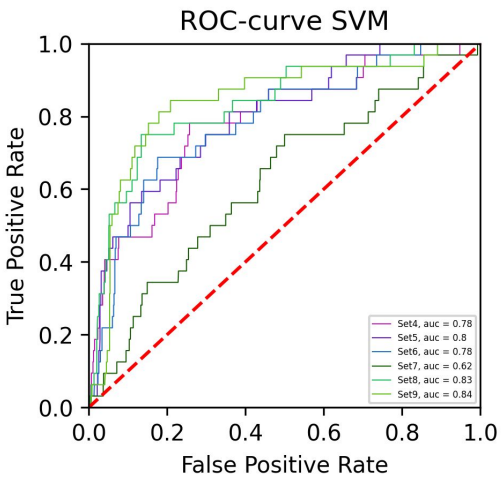




Best models







Dataset		Model	ROCAUC train	ROCAUC test	F1 train	F1 test	Precision1 test	Recall1 test
Set7	ODDT IFP	LR	0.71±0.01	0.63±0.05	0.06±0.00	0.06±0.00	0.03±0.00	1.00±0.00
Set4	Morgan 512, Flare IFP	SVM	0.80±0.01	0.68±0.04	0.10±0.01	0.08±0.01	0.04±0.01	0.75±0.07
Set8	Morgan 1024	SVM	0.89±0.01	0.71±0.02	0.19±0.01	0.13±0.01	0.07±0.01	0.57±0.07
Set8	Morgan 1024	LR	0.86±0.01	0.70±0.04	0.18±0.01	0.13±0.01	0.07±0.01	0.56±0.06
Set9	Morgan 1024, ODDT IFP	LR	0.86±0.01	0.70±0.04	0.18±0.01	0.13±0.01	0.07±0.01	0.56±0.06

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!