In silico prediction of NSP13 inhibition and virtual screening campaign

Team members:

- Asma Alimolaei
- Elnaz Vojoudi Yazdi
- Fateme Sarhandi
- Riham Ibrahim
- Sara Shirvani
- Trishang Udhwani
- Zeinab Salehian

Partner teams:

- Sorbonne 3
- Heidelberg 1

1. Introduction [1 2 3 4 5 6]

The global COVID-19 pandemic continues to wreak economic and social problems globally. Effective vaccines and drug therapies are essential to bring the pandemic to an end. This global challenge has seen a unique and intense focus on coronavirus research, resulting in the development of vaccines in impressively short times. Similarly, empirical and limited rational selection of drugs, such as remdesivir, provided early drug treatments that limited morbidity and mortality. However, more effective drugs are still required to treat COVID-19 and other coronavirus diseases, such as SARS and MERS, as well as new viruses that may emerge in the future.

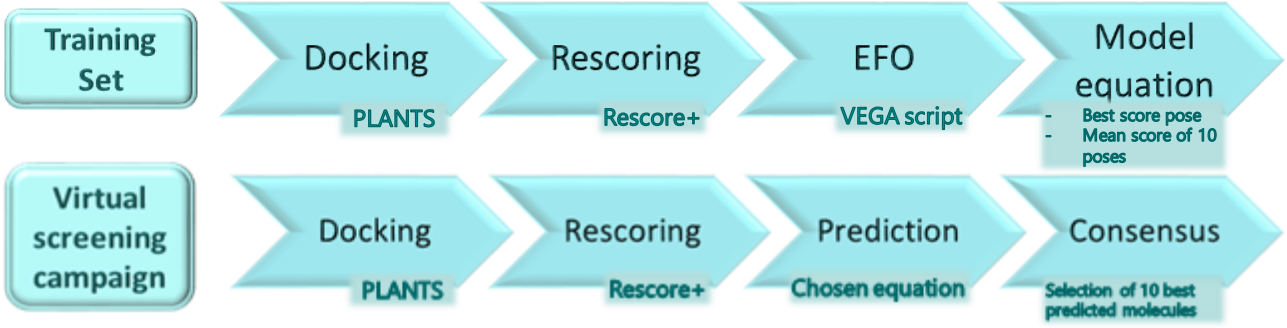
In addition to structural proteins, the SARS-CoV-2 non-structural proteins are of interest and these include the helicase NSP13.
NSP13 is a 67 kDa protein that belongs to the helicase superfamily 1B, it utilises the energy of nucleotide triphosphate hydrolysis to catalyse the unwinding of double-stranded DNA or RNA in a 5′ to 3′ direction. Although NSP13 is believed to act on RNA *in vivo* enzymatic characterization shows a significantly more robust activity on DNA in *in vitro* assays with relatively weak non-processive helicase activity when compared to other superfamily 1B enzymes.
NSP13 has been shown to interact with the viral RNA-dependent RNA polymerase NSP12, and acts in concert with the replication-transcription complex (NSP7/NSP8/NSP12). This interaction has been found to significantly stimulate the helicase activity of NSP13 possibly by means of mechano-regulation. In addition to its helicase activity, NSP13 also possesses RNA 5′ triphosphatase activity within the same active site, suggesting a further essential role for NSP13 in the formation of the viral 5′ mRNA cap.

NSP13 contains 5 domains, a N-terminal Zinc binding domain (ZBD) that coordinates 3 structural Zinc ions, a helical “stalk” domain, a beta-barrel 1B domain and two “RecA like” helicase subdomains 1 A and 2 A that contain the residues responsible for nucleotide binding and hydrolysis.

The SARS-CoV-2 Non-structural protein 13 (NSP13) has been identified as a target for anti-virals due to its high sequence conservation and essential role in viral replication. Structural analysis reveals two druggable pockets on NSP13 that are among the most conserved sites in the entire SARS-CoV-2 proteome. NSP13 has been suggested as a good target for the development of new antiviral drugs.

Computational-aided drug discovery (CADD) methods can increase the odds of identifying compounds with desirable features, speed up the hit-to-lead process and improve the chances of getting the compound pass preclinical testing.

In order to give a contribute to NSP13 studies, the main aim of this work is to develop prediction models able to discriminate which ligands could have inhibition activity towards the target, through virtual screening campaigns.



2. Materials and Methods

2.1 Protein preparation

Among the resolved NSP13 structures, this study involved the complex between the protein and the ligand 1-(diphenylmethyl)azetidin-3-ol (PDB ID: 5RM2), chosen due to its high resolution (1.82Å). In VEGAZZ suite [7,8], after deleting water molecules, ions and crystallographic additives, hydrogen atoms were added to the amino acid residues. In order to carry out studies on physiological pH value, the protein was ionised considering pH 7.4. The force field CHARMM22 and the Gasteiger's charges were assigned. Then, a general check of the protein was performed, by which the absence of cis peptide bonds, R- amino acids, rings intersections and altered bond lengths were verified. The protein was then refined by 2 minimisation procedures using NAMD. The first procedure was carried out with all the protein atoms fixed except for hydrogens; the second one was performed with the backbone atoms under constraints to preserve the resolved folding. The so prepared protein structure underwent the following docking simulations.

2.2 Training set preparation

The training set was collected including 1% of active molecules and 99% of inactive molecules.

The active molecules were chosen among known inhibitors from literature. The 10 known NSP13 inhibitors were generated starting from their 2D structure, they were drawn by Ketcher (in VEGAZZ suite). The obtained structures were then optimised, using AMMP as implemented in VEGAZZ environment. Gasteiger-Marsili charges and CHARMM22 force field were assigned, and an energy minimisation was carried out using the conjugate gradient algorithm with 3000 iterations and a tolerance value of 0.01.

In the same way the inactive dataset was prepared, choosing 990 decoys downloaded by ZINC database.

In order to reduce the bias as much as possible, the ZINC database was filtered by considering the physico-chemical properties of the active molecules, such as angles, atoms, bonds, charge, chiral atoms, dipole, flexible torsions, H-bond acceptors and donors, heavy atoms, lipole, mass, PSA, rings, SAS, SAV, surface, torsions, virtual LogP, volume.

2.3 Training set docking

Docking simulations were carried out using PLANTS (Protein-Ligand ANT System)[9], which generates reliable ligand poses using the ant colonization algorithm (ACO). In detail, the search was focused on a 10 Å radius sphere around the bound ligand, thus including the entire binding cavity; for each ligand 10 poses were generated and scored by using the ChemPLP function[10]. Then a rescore analysis was performed on these poses, using Rescore+[11].

Two sets of molecular docking studies were conducted, considering both binding pockets of NSP13. The active sites were chosen considering a neighbourhood of the co-crystallised ligands. The first binding pocket centre was set at the following coordinates -12.95 39.96

-18.27 with a radius of 12.87Å, while the second one was set at -23.61 17.80 -11.30 with a radius of 13.90Å.

2.4 Generation of predictive models

For the generation of the predictive models, the interaction scores (the primary docking score and the ones obtained from rescoring) were used as descriptors. In detail, for each binding pocket, two models were created: the former considering as descriptors the scores of the best pose as defined by the primary score, the latter considering as descriptors the mean scores of the 10 poses. For the last, an implemented script in VEGAZZ environment was used.

Thanks to another script in VEGAZZ, the enrichment factor optimisation (EFO)[12 13] was applied, and the predictive ability of the model was evaluated. Beside the enrichment factors, the AUC of the ROC curve and the confusion matrix were computed as well.

2.5 OpenScreen virtual screening campaign

To predict unknown molecules activity, the best model chosen in the previous step was used.

The database was downloaded from the OpenScreen drug database and it was filtered by the same mass range and heavy atom number range as the training set. Then it was cleaned and standardised, removing molecules with rare element and salts, resulting in 4167 molecules. Hydrogen atoms were added and an AMMP optimisation was done.

The dataset underwent to molecular docking, as done with the training set, keeping the same setting – binding pocket, scoring function, number of poses – as before. Then, Rescore+ script was used.

In order to select the best predicted as active molecules, a consensus score docking was performed. Each equation was used to predict which molecules could be considered as active from this dataset. Then, all these results were cross-checked, in order to select which molecules were predicted as active from all these models simultaneously. There were 71.

3. Results

3.1 Models selection

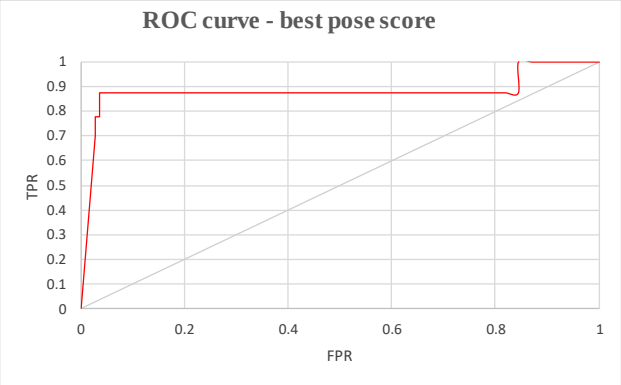
For the validation, an equation for each setting (*i.e.* two binding pockets and best pose score evaluation and mean score of 10 poses evaluation) was chosen according to the best Enrichment Factor (EF), absence of redundant descriptors and reduced number of variables.

The first step was to calculate the confusion matrix, which allows an evaluation of the performance of the algorithm, through the calculation of the Matthews correlation coefficient (MCC), AUC, Precision, Recall and, mostly important, ROC curves.

In the following tables, there is a summary of the results for each case.

For the first binding pocket:

BEST POSE SCORE			
Model ID	**Model equation**	**Actives in 1%**	**EF**
2	1.00 PLANTS_CHEMPLP_NORM_WEIGHT	5	50

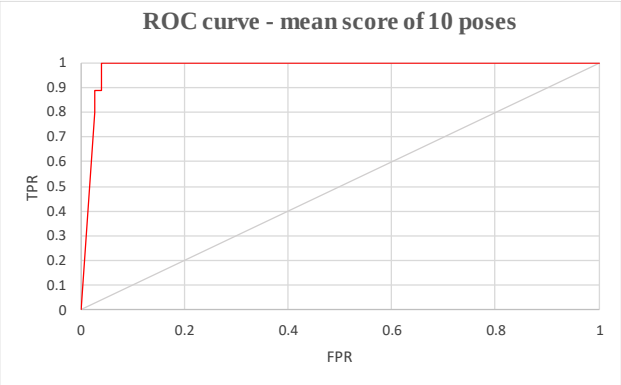


Confusion matrix

		predicted classes	
		0	**1**
Experimental classes	**0**	963	27
	1	3	7

F1 score	0.318182
MCC	0.369342
AUC	0.869865
Precision	0.205882
Recall	0.7

MEAN SCORE OF 10 POSES			
Model ID	**Model equation**	**Actives in 1%**	**EF**
3	1.00 PLANTS_CHEMPLP_NORM_WEIGHT	5	50



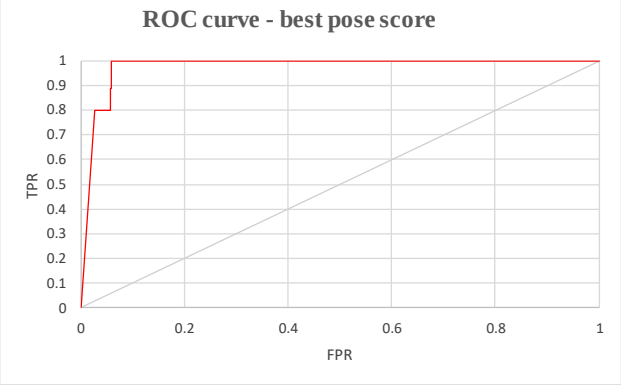
Confusion matrix

		predicted classes	
		0	**1**
Experimental classes	**0**	964	26
	1	2	8

F1 score	0.363636
MCC	0.424799
AUC	0.972298
Precision	0.235294
Recall	0.8

For the second binding pocket:

BEST POSE SCORE			
Model ID	**Model equation**	**Actives in 1%**	**EF**
1	1.00000000 PLANTS_PLP95_NORM_WEIGHT + 0.00268536 ELECTDD	6	60

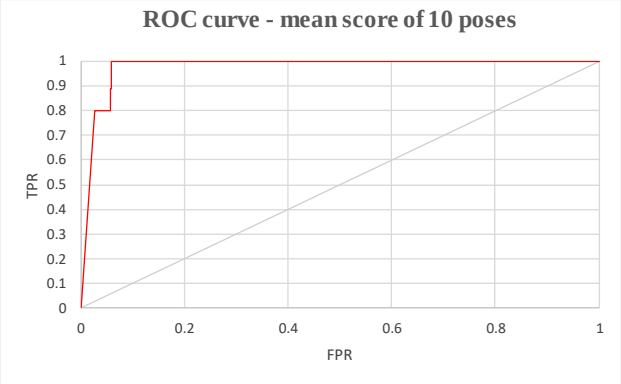


Confusion matrix

		predicted classes	
		0	**1**
Experimental classes	**0**	964	26
	1	2	8

F1 score	0.363636
MCC	0.424799
AUC	0.967488
Precision	0.235294
Recall	0.8

MEAN SCORE OF 10 POSES			
Model ID	**Model equation**	**Actives in 1%**	**EF**
1	1.00000000 PLANTS_PLP95_NORM_WEIGHT + 0.51208049 Contacts_NORM_HEVATMS	5	50



Confusion matrix

		predicted classes	
		0	**1**
Experimental classes	**0**	962	28
	1	4	6

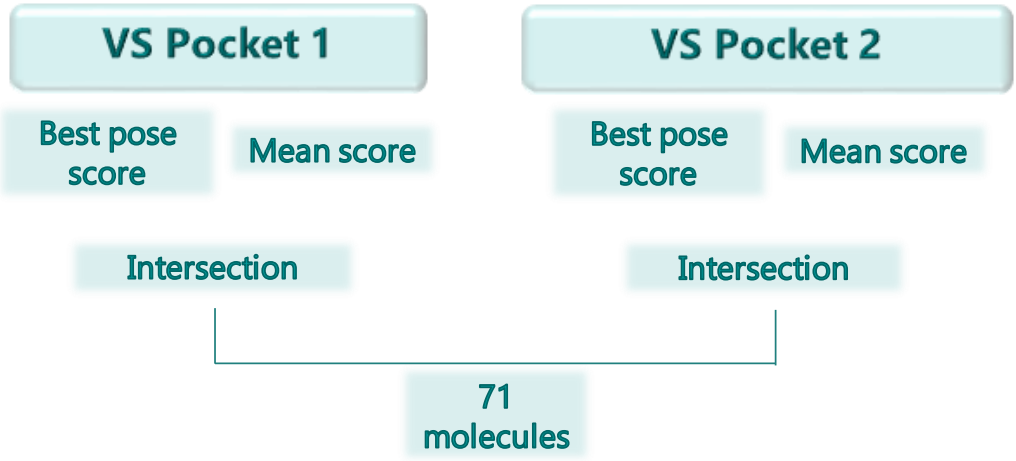
F1 score	0.272727
MCC	0.313885
AUC	0.937271
Precision	0.176471
Recall	0.6

It is worth noticing that the enrichment factor reaches high values in each model, as well as the AUC value (Area Under the Curve, that is a measure of accuracy) which is always up to 0.86. For this reason, all models can be considered robust.

The ROC curve is the graphical representation of the binary classification. In the chart, on the *x axis* there is the proportion of false positives, which is also defined as specificity, and on the *y axis* there is the proportion of true positive, also called sensitivity. While the EF focuses only on the correct compounds at the top of the ranking without considering what the rest of it, ROC curve evaluates the robustness of the entire ranking

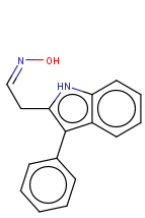
3.2 Virtual Screening results

In the following diagram, there is the summary of the selection made:

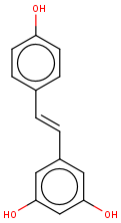


The first 10 molecules are listed below:

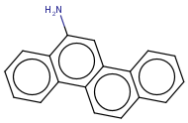
EOS98620
EOS102407
EOS100899
EOS100433
EOS2253
EOS100710
EOS101630
EOS100134
EOS102462
EOS102211



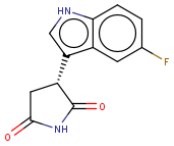
EOS2253



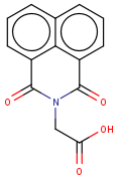
EOS98620



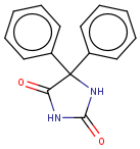
EOS100134



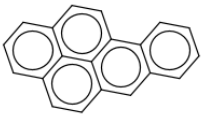
EOS100433



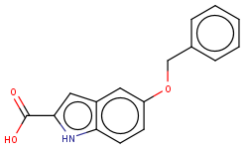
EOS100710



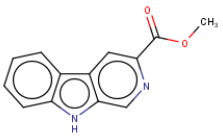
EOS100899



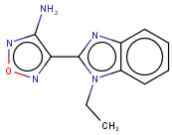
EOS101630



EOS102211



EOS102407



EOS102462

3.3 Partner's results

The two other Teams worked on the search of potential binding pockets of NSP13. Both of them focused their studies on PDB ID 6ZSL.

Heidelberg group 1 screened a dataset of 5000 molecules from Zinc database (randomly selected from the sub-set of drug-like molecules) optimized with AutoGrow4, while Sorbonne group 3 screened OpenScreen database without any filtering steps.

Heidelberg group 1 performed the binding pockets prediction with fpocket and then they performed molecular docking studies using AutoDock Vina, with Vinardo scoring function. They validated the results with molecular dynamic simulations with GROMACS.

Sorbonne group 3 used p2rank to predict potential binding pockets on the NSP13 and then they performed molecular docking studies with Autodock Vina and finally scored the interactions using Convex-PL.

The first 10 best predicted molecules are listed below:

Heidelberg 1
ZINC000514436632
ZINC000104309836
ZINC000620748806
ZINC000019015192
ZINC000104277568
ZINC001164872284
ZINC001164979321
ZINC000101042701
ZINC000095523345
ZINC000257281912

Sorbonne 3
EOS100851
EOS101596
EOS100853
EOS100609
EOS101513
EOS117
EOS101596
EOS101170
EOS1709
EOS101554

Unfortunately, no molecule from Sorbonne group 1 prediction is in common with our consensus prediction.

4. Discussion

It is well known that the assessment of druggability of specific regions of protein surfaces may offer great opportunity to identify small molecule inhibitors, capable of developing promising therapeutic candidates. Distinct enzymes involved in the SARS-CoV-2 replication are characterised by shallow surfaces, thus, resulting in challenges for the discovery process, which employs computational approaches, aimed at searching druggable binding sites. To detect the presence of druggable areas suitable for the binding of inhibitors, it is crucial to exploit the binding hot spots at the viral protein surfaces.

Here, we employed distinct computational approaches to explore protein surfaces that are major contributors for the binding of ligands, capable of inhibiting the activity of helicase NSP13 of SARS-CoV-2, thus, reducing the virus replication and infectious capacity. To gain insights into the protein hot spots involved in the binding process, the two potential druggable sites of the NSP13 protein were explored.

Four predictive models were generated and were used to predict potential inhibitors of NSP13 from a known drug database. Since the prediction is highly influenced by the training set composition, all datasets were accurately prepared.

The final consensus study represents a fruitful strategy to cross-check diverse predictions and to find the best overall results.

The virtual screening has led to the identification of different compounds. Waiting for an experimental validation of these results, this model could represent a good starting point for the rational design of new potential inhibitors of NSP13, in order to offer a new way to fight COVID-19.

References

1. Ricci, Federico, et al. "In Silico Insights towards the Identification of SARS-CoV-2 NSP13 Helicase Druggable Pockets." Biomolecules 12.4 (2022): 482.
2. Tolbatov, Iogann, Loriano Storchi, and Alessandro Marrone. "Structural Reshaping of the Zinc-Finger Domain of the SARS-CoV-2 nsp13 Protein Using Bismuth (III) Ions: A Multilevel Computational Study." Inorganic Chemistry 61.39 (2022): 15664-15677.
3. Berta, Dénes, et al. "Modelling the active SARS-CoV-2 helicase complex as a basis for structure-based inhibitor design." Chemical Science 12.40 (2021): 13492-13505.
4. Raubenolt, Bryan A., et al. "Molecular dynamics simulations of the flexibility and inhibition of SARS-CoV-2 NSP 13 helicase." Journal of Molecular Graphics and Modelling 112 (2022): 108122.
5. El Hassab, Mahmoud A., et al. "Multi-stage structure-based virtual screening approach towards identification of potential SARS-CoV-2 NSP13 helicase inhibitors." Journal of enzyme inhibition and medicinal chemistry 37.1 (2022): 563-572.
6. Newman, Joseph A., et al. "Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase." Nature Communications 12.1 (2021): 4848.
7. Pedretti, Alessandro, Luigi Villa, and Giulio Vistoli. "VEGA: a versatile program to convert, handle and visualize molecular structure on Windows-based PCs." Journal of Molecular Graphics and Modelling 21.1 (2002): 47-49.
8. Pedretti, Alessandro, Luigi Villa, and Giulio Vistoli. "VEGA–an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming." Journal of computer-aided molecular design 18 (2004): 167-173.
9. Korb, Oliver, Thomas Stützle, and Thomas E. Exner. "PLANTS: Application of ant colony optimization to structure-based drug design." Ant Colony Optimization and Swarm Intelligence: 5th International Workshop, ANTS 2006, Brussels, Belgium, September 4-7, 2006. Proceedings 5. Springer Berlin Heidelberg, 2006.
10. Korb, Oliver, Thomas Stutzle, and Thomas E. Exner. "Empirical scoring functions for advanced protein− ligand docking with PLANTS." Journal of chemical information and modeling 49.1 (2009): 84-96.
11. Pedretti, Alessandro, et al. "Structural Effects of Some Relevant Missense Mutations on the MECP2-DNA Binding: A MD Study Analyzed by Rescore+, a Versatile Rescoring Tool of the VEGA ZZ Program." Molecular Informatics 35.8-9 (2016): 424-433.
12. Mazzolari, Angelica, et al. "Prediction of the formation of reactive metabolites by a novel classifier approach based on enrichment factor optimization (EFO) as implemented in the VEGA program." Molecules 23.11 (2018): 2955.
13. Pedretti, Alessandro, et al. "Rescoring and linearly combining: A highly effective consensus strategy for virtual screening campaigns." International Journal of Molecular Sciences 20.9 (2019): 2060.