An *in silico* approach to find potential SARS-CoV-2 NSP13 inhibitors

Ali Nassiri

Gaia Gastoldi

Mahsa Babaie

Negar Esfahanizadeh Hosseinpour

Team MILANO 2 (paired with Warsaw 3 and Sorbonne 4)

1. Introduction

SARS-CoV-2 (severe acute respiratory syndrome coronavirus-2) is the etiological agent of a respiratory disease whose first outbreak was reported in December 2019 in Wuhan, China. Even if the WHO announced the end of COVID-19 public health emergency during May 2023, the research to find new active molecules against SARS-CoV-2 remains crucial.

Among SARS-CoV-2 proteins, the structural proteins (spike, membrane, envelope and nucleocapsid proteins) are obvious target to prevent viral entry into the host cells and viral assembly. However, also the NSPs (non-structural proteins) implied in viral replication are very interesting target to act against SARS-CoV-2 infection. NSP13 is the helicase of SARS-CoV-2 and it works in coordination with NSP12 (RNA-dependent RNA polymerase), NSP8 and NSP7 (NSP12 cofactors) in the replication-transcription complex.

NSP13 belongs to the helicase superfamily 1B and it is highly conserved among coronaviruses. It is composed by five domains: two ReacA-like domains (1A and 2A domains), the zinc-binding domain, the stalk domain and the 1B domain. The 1A, 1B and 2A domains form the helicase core. NSP13 unwinds double-stranded DNA or RNA (with a preference for DNA in in vitro assays) and has triphosphatase activity on all standard nucleotides and ribonucleotides. Its enzymatic activity is Mg2+-dependent and the ion is involved in ATP binding. NSP13 translocation has a 5'-3' polarity.

For its role in viral replication and its conservation in pathogenic coronaviruses, several efforts have been made to find SARS-CoV-2 helicase inhibitors, both *in vitro* and *in silico*.

The published computational approaches to find new NSP13 inhibitors mostly relied on molecular docking and pocket prediction. Molecular dynamics simulations have also been performed, especially to investigate the binding mode of experimental inhibitors or the relative movements of NSP13 domains during its translocation. Initially, the computational studies used homology modelling to generate the 3D structure of NSP13, then the publication of several SARS-CoV-2

helicase crystal structures provide a more solid starting point. Similarly, our strategy was a structure-based approach based on known inhibitors retrieved from published *in vitro* assays and on the resolved crystal structure of NSP13.

We selected known inhibitors from literature and we completed the training set with decoys that were retrieved from DUD-E and ZINC. All the molecules were ionized at pH 7.4 and minimised with AMMP. The decoys were filtered according to physical-chemical properties of the inhibitors. An unbalanced dataset of 9 active (1%) and 891 (99%) inactive molecules was obtained.

The docking was performed with PLANTS and the generated poses were then rescored with the RESCORE+ tool as implemented in the VEGA ZZ suite of programs. The scores were used to build a model using the Enrichment Factor Optimisation algorithm. The bests linear equations were selected. The virtual screening was performed on the Openscreen ECBL Pilot Library of 5016 molecules. The trained model was used to predict which molecules could be active against SARS-CoV-2 NSP13. Figure 1.1 provides an outline of the basic workflow used in this project.

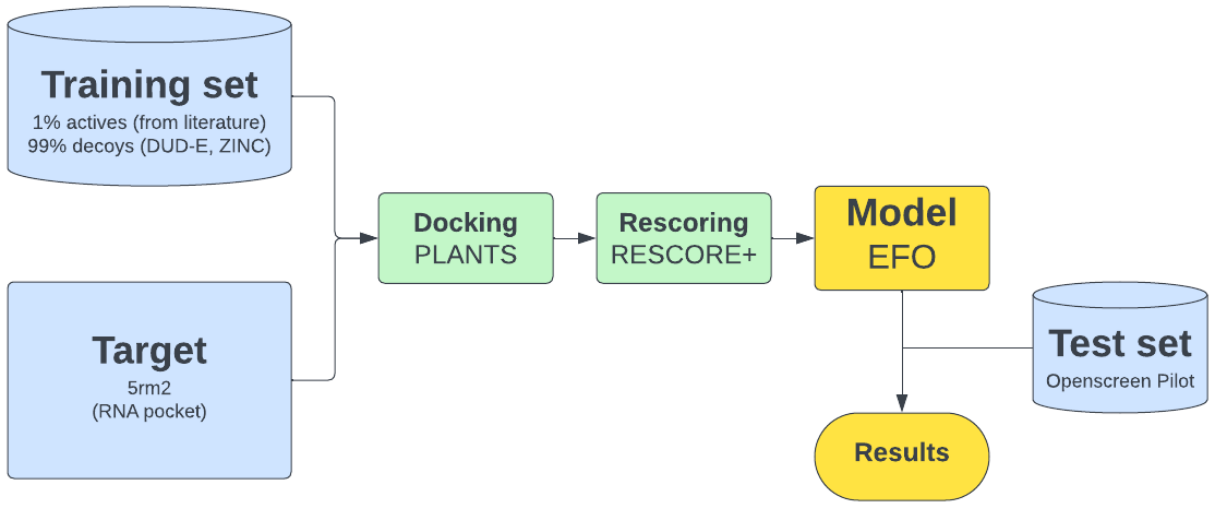

Figure 1.1 Virtual screening workflow followed in this project.

2. Material & Methods

Training set preparation

9 inhibitors were selected from literature: baicalein, flavanone, flavanone-7-O-glucoside, kaempferol, licoflavone, myricetin, quercetin, SSYA10-001, zafirlukast. Their SMILES were retrieved from PubChem and used to build the 2D molecules in VEGA ZZ.

VEGA ZZ served as the primary platform for all the key steps of our project. It is a suite of programs that are useful for different operations in computer-aided drug design, such as molecules and proteins 3D visualisation, manipulation and minimisation, molecular docking and rescore, database generation and conversion.

The 2D structures of the inhibitors were then ionised at pH 7.4 and the minimisation was performed with AMMP (conjugate gradient). VEGA ZZ also calculated several physical-chemical properties of the active molecules that were later used to filter the decoys and obtain a homogeneous training set. We operated this filtration to avoid biases in the training set because of the inclusion of outlier molecules that could be easily classified as inactive by the model.

The decoys were generated in DUD-E using the *Generate* functions that allows to obtain decoys starting from the SMILES structure of the active molecules. To complete the training set, we obtained other decoys from the ZINC database. The inactive molecules were prepared and minimised in VEGA ZZ in the same way of the inhibitors. As said, all the decoys were then filtered according to physical-chemical properties of the 9 active molecules.

Pocket definition

Once the training set was completed, we considered the protein structure. Starting from literature, we identified the ATP-binding pocket and the RNA-binding pocket as the most interesting for inhibition of the helicase. To diversify from the project of the Milano 2 Team, we chose to work on the 5'-RNA pocket. Among the already-prepared structures available on MEET-EU 2023 website, we chose 5RM2 because it was the only structure with a small molecule docked at the RNA-binding site. Using the co-crystallised small molecule as the centre of the binding pocket in VEGA ZZ, we define the coordinates of the binding pocket (-24.6800; 17.7100; -10.6000) using an 8 Å radius around the ligand.

Redocking

We decided to use PLANTS to perform the docking as it is already implemented in VEGA ZZ. **PLANTS** (Protein-Ligand ANT System) is a docking algorithm based on ACO (ant colony optimisation). ACO algorithms are stochastic algorithms (optimisation algorithms that make use of randomness to optimise the target function) that fall within the swarm intelligence techniques, which are an artificial intelligence approach that recalls the natural behaviour of some animals to solve optimisation problems. Swarm intelligence is based on a population of simple agents that interact

locally with one another and with the environment and that can self-organize as a functioning system. ACO is inspired by the foraging behaviour of real ants: to find the shortest path between the nest and the food source they deposit pheromone trails on their way and identify the path with the highest pheromone concentration as the shortest, because it's the one that has been taken most times by previous ants. Similarly, ant colony optimisation algorithms employ a colony of virtual ants to find the best function to solve a problem. After each iteration of the algorithm, the goodness of each solution is determined based on the quantity of pheromones. The ligand is treated as flexible and a pheromone vector is calculated for each of its degrees of freedom, which are three translational and three rotational degrees of freedom for each ligand plus as many torsional degrees of freedom as the single bonds that are not part of a ring system and thus have freedom of rotation. To rank the poses PLANTS uses three empirical scoring functions: PLP (piecewise linear potential), PLP95, and ChemPLP. Empirical scoring functions calculate the binding affinity through an equation that considers a set of non-related individual

contributions or interaction descriptors. PLP and PLP95 consider steric interactions, torsional potential, a clash term, intramolecular interactions; ChemPLP in addition considers H-bonds and metal-acceptor interactions. Constraints from experimental data can be added to improve the docking poses. In this case, the final score is multiplied for or summed to a certain factor if the

pose matches the constraints in shape, H-bonds, distance from the protein surface or other features.

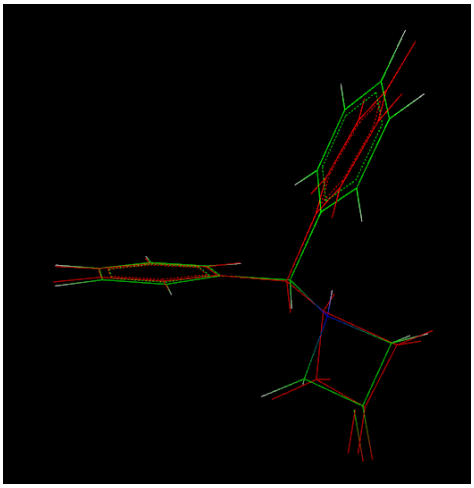

Figure 2.1 Overlapping structures of the original co-crystallised ligand and the minimized-and-docked ligand after the applying of the shape constraints

To test PLANTS ability of docking a small molecule in the chosen binding site, we performed the redocking using the minimised co-crystallised ligand. We considered a redocking satisfying when

the RMSD (root-mean square deviation) between the best docking pose and the original co-crystallised ligand was less than 2 Å. Since the result wasn't satisfying, we decided to use the original pose of the ligand as a shape constraint in the docking procedure: for an optimal atom-overlap of a ligand and a shape atom –3.0 value is added to the score. In this way we obtained an RMSD of 0.3783 Å (Figure 2.1).

Docking

The docking of the training set was performed with PLANTS in the VEGA ZZ environment. The binding site was the one defined in the *Pocket definition* paragraph. We used ChemPLP scoring function. Ten poses were generated for each molecule (9000 poses were obtained in total).

Rescoring

The rescoring of the generated poses was performed with **RESCORE+** as implemented in the VEGA ZZ suite. Rescoring is a useful tool to calculate new scores after molecular docking, to refine the already-obtained scores and generate new others to have more features to be used in the model generation. The selected scoring functions were: CHARMM, Elect, ElectDD, MLPInS, MLPIns_2, MLPInS_3, MLPInS_F, Contacts, ChemPlp, Plp, Plp95, X-Score.

Model design

The model was developed using the **Enrichment Factor Optimisation algorithm** as implemented in VEGA ZZ. EFO is a classification algorithm that allows to calculate a new scoring function by linearly combining different descriptors (in this case, the docking scores). The result of the calculation is a first-degree equation whose coefficients are calculated by maximising the enrichment factor (EF), calculated as below:

$$EF_{\%} = \frac{\frac{Actives_{\%}}{N_{\%}}}{\frac{Actives_{tot}}{N_{tot}}}$$

The EF quantifies the capacity of the model to rank the correct compounds in the top-scoring molecules and the bigger the EF is, the more robust the model is.

EFO maximises the EF using the gradient-free Hooke-Jeeves algorithm (pattern search) and random sampling to avoid local maxima. The number of variables is set by the user. We chose EFO to develop our model because it was specifically developed to perform virtual screening campaigns on unbalanced data sets (as in our case, with 1% of active and 99% of inactive compounds).

We generated 3 models (using 1 variable, 2 variables and 3 variables respectively) considering only the best pose for each molecule and 3 models (using 1 variable, 2 variables and 3 variables respectively) considering the mean of the ten poses of each molecule. Therefore, we obtained 6 models in total and we decided to keep only one model for the best poses and one model for the mean of the poses.

In the output generated by the EFO script in VEGA ZZ, the models are already ranked according to the EF. We chose the best model (the one with the higher EF) discarding the models whose linear equations were redundant (with more than one score of the same type). We thus selected these two linear equations:

- For the best poses:

PLANTS_CHEMPLP_NORM_HEVATMS + 0,09523810 MLPINS_2 + 0,09523810 ELECTDD

Enrichment factor (at 1%) = 44.44

- For the mean of 10 poses:

PLANTS_CHEMPLP_NORM_WEIGHT + 0,00948905 ELECTDD

Enrichment factor (at 1%) = 55.55

These equations allow to calculate a score to rank the molecules. The lower the score it is, the most active is predicted to be based on his docking and rescoring performance.

For each model we had to define the final-score value that we considered as the cut-off for distinguish active and inactive compounds. Since we defined the EF at 1%, we took the score of the 9^{th} molecule (out of 900) in the ordered rank as the cut-off. They result

Virtual screening

The database used to search for new inhibitors of NSP13 in the virtual screening was the Openscreen ECBL (European Chemical Biology Library) Pilot Library of 5016 molecules. The database was retrieved from the Openscreen website and the molecules were prepared with the same procedures used for the training set. Salts were removed and the remaining molecules were filtered according to the physical-chemical properties of the inhibitors (and of the whole training set), whose ranges

we considered as the applicability domain of the model. Duplicates with the training set were removed. We obtained 859 molecules.

We performed the docking with PLANTS and the rescore with RESCORE+ with the same settings used to generate the model. Then we applied the selected equations on the best poses and on the mean of the poses. The molecules with higher score than the cut-off value were the ones predicted as actives.

3. Results

In Table 3.1 the molecules predicted as active are listed, both for the model on the best poses and for the model on the mean of the poses.

BEST	MEAN
1,00000000 PLANTS_CHEMPLP_NORM_HEVATMS + 0,09523810 MLPINS_2 + 0,09523810 ELECTDD	1,00000000 PLANTS_CHEMPLP_NORM_WEIGHT + 0,00948905 ELECTDD
ModID_1	ModID_2
11 molecules predicted as active	15 molecules predicted as active
cut-off: -3,7314	cut-off: -0,22935
EOS100027	EOS98616
EOS100104	EOS100653
EOS100653	EOS100677
EOS100947	EOS593
EOS98616	EOS100072
EOS100072	EOS102169
EOS1251	EOS1341
EOS102267	EOS100027
EOS1601	EOS1763
EOS2290	EOS100804
EOS1341	EOS102409
	EOS101604
	EOS1240
	EOS100470
	EOS102410

Table 3.1 The table shows the molecules identified as active according to the two selected models. The BEST column refers to the model built on the best poses, while the MEAN column refers to the model built on the mean of ten poses. The respective equations, model names, number of molecules predicted as active and cut-off values are reported in each column.

As can be seen, 5 molecules are recognised as active by both models: EOS100027, EOS100653, EOS98616, EOS100072, EOS1341. Their structures can be seen in Figure 3.1. 3 of the top-scoring hits are flavonoids. Among flavonoids, the *in vitro* activity as NS13 inhibitors has been proven for several molecules and in fact, among the inhibitors selected for the training set, seven were flavonoids.

In conclusion, our model allowed us to search for inhibitors of NSP13, using docking, rescoring and the EFO algorithm. For the construction of the training set, we relied on the available experimental data about inhibitors whose activity has been confirmed in vitro. The docking and rescoring procedure can be time-consuming, but VEGA ZZ made it easy to perform the calculations and develop the model even without coding.

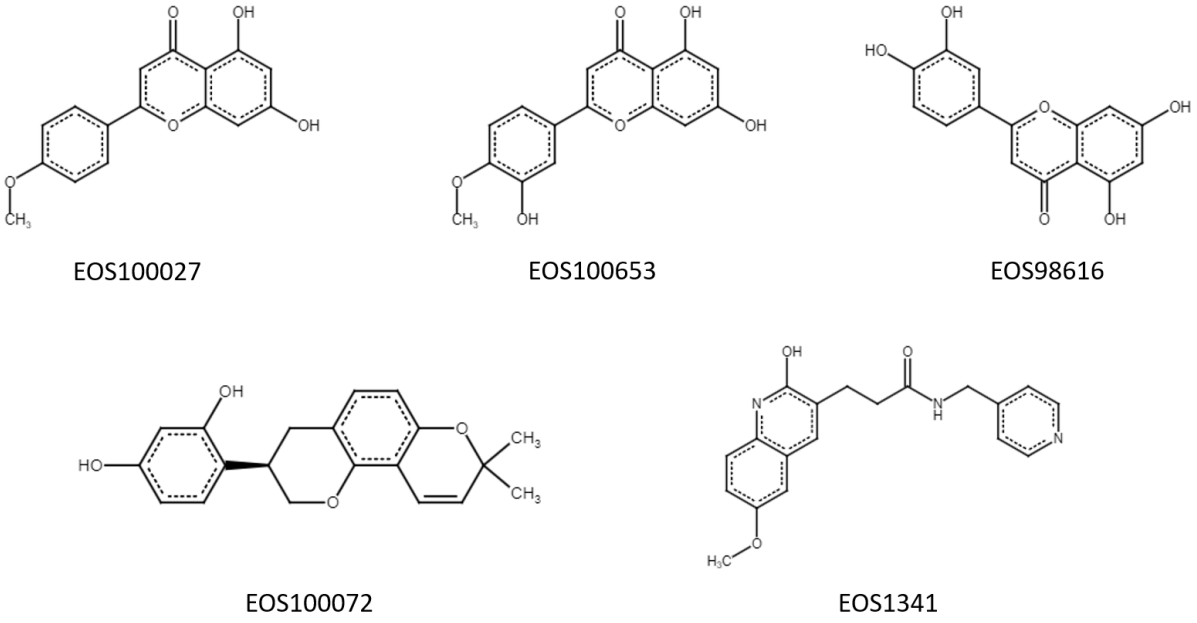

Figure 3.1 2D structures of the 5 molecules that was recognised as active both by the model on the best poses and the model on the mean of the poses.

Comparison with the paired groups

Warsaw3 team explored three NSP13 pockets using peptides (retrieved from literature review, from structural search on PDB or newly generated using ProtGPT2) and among them they considered also 5'-RNA binding pocket. The 5'-RNA binding pocket is quite big and this is probably the reason why we needed to use constraints in our docking procedure, while the peptides are probably big enough to docking in a satisfactory way in the pocket. The analysed peptides often bound this pocket, except in the models generated with AlphaFold2. One common difficulty was probably finding enough

active molecules (small molecule for us and peptides for them) in the available published data. The differences in the ligands we were working with oriented our respective choices about the docking tools (HPepDock and CABS-dock have been specifically developed for protein-peptide docking).
Our approach has probably more in common with the Sorbonne4 team, because they work with small molecules as well. However, there are some differences in our work: they identified the pocket with P2rank and they used AutoDock Vina for docking and scoring. None of the top scoring molecules was in common. In general, their approach is more similar to ours, but we relied more on literature for pocket and training set selection.

4. Conclusions

During our project we learned how to perform a virtual screening starting from an available crystal structure of the target protein and some experimentally-active molecules. In the Openscreen ECBL Pilot Library the molecules identified as active by both of our models were mainly flavonoids and this is probably because the inhibitors that we selected from literature were mainly flavonoids as well. To improve the predictive power of our models it would be useful to expand the training set with more diverse compounds.
In conclusion, our approach allowed us to exploit the potential of different tools for drug-design but always relying on the available experimental data about NSP13 to find new potential inhibitors.

References

Newman, J.A., Douangamath, A., Yadzani, S. et al. Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase. Nat Commun 12, 4848 (2021).

Angela Corona, Krzysztof Wycisk, Carmine Talarico, Candida Manelfi, Jessica Milia, Rolando Cannalire, Francesca Esposito, Philip Gribbon, Andrea Zaliani, Daniela Iaconis, Andrea R. Beccari, Vincenzo Summa, Marcin Nowotny, and Enzo Tramontano, ACS Pharmacology & Translational Science 2022 5 (4), 226-239

N. Mehyar, A. Mashhour, I. Islam, H.A. Alhadrami, A.M. Tolah, B. Alghanem, S. Alkhaldi, B.A. Somaie, M. Al Ghobain, Y. Alobaida, A.S. Alaskar & M. Boudjelal (2021) Discovery of Zafirlukast as a novel SARS-CoV-2 helicase inhibitor using in silico modelling and a FRET-based assay, SAR and QSAR in Environmental Research, 32:12, 963-983

Corona, A., Madia, V. N., De Santis, R., Manelfi, C., Emmolo, R., Ialongo, D., ... & Tramontano, E. (2023). Diketo acid inhibitors of nsp13 of SARS-CoV-2 block viral replication. Antiviral Research, 217, 105697.

Irwin, J. J., & Shoichet, B. K. (2005). ZINC—a free database of commercially available compounds for virtual screening. Journal of Chemical Information and Modeling, 45(1), 177-182.

Mysinger, M. M., Carchia, M., Irwin, J. J., & Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. Journal of Medicinal Chemistry, 55(14), 6582-6594.

Pedretti, A., Villa, L., & Vistoli, G. (2004). VEGA—an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming. Journal of Computer-Aided Molecular Design, 18(3-4), 167-173.

Korb, O., Stützle, T., & Exner, T. E. (2009). PLANTS: Application of ant colony optimization to structure-based drug design. In Ant Colony Optimization and Swarm Intelligence (pp. 247-258). Springer.

Pedretti, A., Mazzolari, A., Gervasoni, S., & Vistoli, G. (2019). Rescoring and linearly combining: a highly effective consensus strategy for virtual screening campaigns. International Journal of Molecular Sciences, 20(9), 2060.