Comparison of pharmacophore-based and ML approach to finding inhibitors of Sars-Cov-2 Nsp-13 helicase

Maria Bochenek, Tomasz Cheda, Mateusz Janduła,
Agnieszka Kowalewska, Joanna Krawczyk

November 2022

1 Introduction

The project's task was to identify potential Sars-Cov-2 Nsp-13 helicase inhibitors using virtual screening, docking simulations, and/or ML-based approaches. At the beginning of the course, we were given 5 resolved protein structures and an example of a database of small molecules to perform virtual screening on ([5]). Our pipeline is based on two alternative approaches of virtual screening - pharmacophore-based and DiffDock-based [2].

We built the first approach based on article [1]. We initially explored both ligand-based and receptor-based methods of constructing a pharmacophore, however, we chose the second one. Then, we screened the bioactives subdatabase from [5]. For this purpose (pharmacophore model construction and virtual screening), we used Schrödinger Maestro Phase software [4].

The second approach is based on a recently proposed [2] machine learning method which combines advances in diffusion models and geometric deep learning to reconstruct ligand pose, boasting state-of-the-art performance in redocking. The generated poses are refined and scored with GNINA, a fork of smina (which is a fork of AutoDock Vina) supporting scoring with convolutional neural networks.

Our project workflow is presented in Fig. 1. We describe it in more detail in the section Materials & Methods.

In the project, we were paired with the **second team from Sorbonne University** (S2).

1.1 Project workflow description

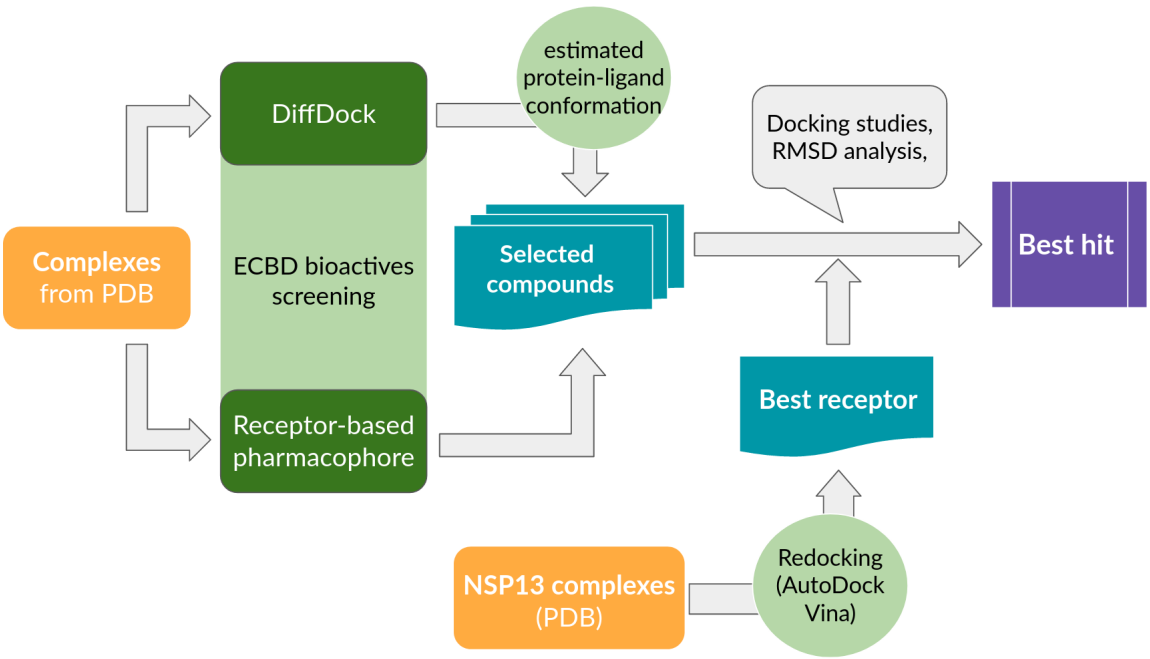

Figure 1: Workflow figure

2 Materials & Methods

2.1 Pharmacophore-based approach

In order to create a pharmacophore model, we have selected and downloaded 55 Nsp-13 complexes with ligands other than simple ions or ADP from the PDB database. Then, we selected one Nsp-13 complex with ADP (PDB 7KRN, chain E) as a reference structure. Consecutively, using PyMol, we have superposed the reference and query structures to check, if the ligand is located in the receptor's ADP binding site. After the visual analysis, we have preselected 15 complexes (5RL7, 5RL9, 5RLI, 5RLJ, 5RM2, 5RM7, 5RLN, 5RLO, 5RLR, 5RLS, 5RLW, 5RLV, 5RLY, 7NN0, 7NNG) to perform redocking using MGLtools [6] (also used for ligand and docking box preparation) and Autodock Vina [7, 8]. As redocking we define removing a ligand (native ligand) from a complex and docking it again in the same complex. Redocking can help us select a receptor for further docking. We define the ADP binding site of Nsp-13 after [1]: ASP374, GLU375, SER377, ASP401, GLN404, ARG443, LYS288, SER289, ARG567, and GLY538.

Based on the redocking score (binding affinity), we have selected a reference structure for docking - 7NN0 (based on binding affinity and visual analysis in PyMol and MGltools). The comparison between the position of the native ligand and the predicted conformation is presented in Fig. 2. The best hits are presented in Table 1. It is worth mentioning that all the ligands except 7NN0 were docked far away from the native ligand (the two molecules were not overlapping).

PDB code	Binding affinity (kcal/mol)
7NN0	-10.7
5RL7	-7.2
5RLS	-7.2
7NNG	-7.1
5RLN	-6.9
5RM7	-6.9
5RM2	-6.7

Table 1: Best AutoDock Vina affinity binding scores for redocked ligands.

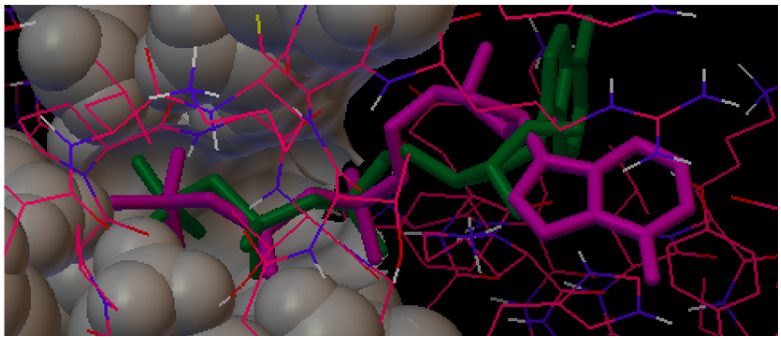

Figure 2: Redocking for complex 7NN0 visualised in MGLTools [6]. The native ligand is marked in magenta. The redocking output - in green. The binding site surface is marked in gray.

We have tried multiple approaches to creating a pharmacophore model. Initially, we wanted to construct a ligand-based pharmacophore using Python libraries - RDKit and Cmapper. Unfortunately, we were not able to succeed on this path, since the Ligand Expo database, which we wanted to use, was out of order for the entire month (January 2023). On the other hand, while pursuing this approach, we have learned about the similarities and dissimilarities of the ligands.

Then, since we could not find another free and open-source tool for pharmacophore construction, we decided to use the trial version of Schrödinger Maestro Phase software [4].

First, we tried developing ligand-based pharmacophore hypothesis using a subset of 15 ligands natively found in ATP binding pocket of nsp13 complexes from PDB database (Table 2).

We have chosen the subset of ligands with relatively similar scaffolds and features that consist of the following PDB structures: 5RLI, 5RLJ, 5RLW, 5RLR, 5RL9. This combination of ligands from PDB complexes was almost the same as the one presented in [1], difference being that our subset additionally included ligands from 5RLI and 5RLR and did not include ligands from 5RLN

PDB structure	Chain with ligand	Ligand
5RL7	A	VVD
5RL9	B	UR7
5RLJ	B	VW4
5RLN	A	NZG
5RLO	B	UQS
5RLW	B	S9S
5RLI	B	JFM
5RLR	B	VWD
5RLS	A	VWG
5RM7	B	N0E
5RLY	A	K34
5RM2	B	UXG
5RLV	A	VWJ
7NNG	A	UJK
7NN0	A	ANP

Table 2: PDB structures with ligands found in ATP-binding pocket.

and 5RLO. Structure alignment of the ligands brought a lot of questions about the validity of this approach as functional groups that interacted with protein were grouped with functional groups that had no interactions in complexes.

After analyzing receptor-ligand interactions, we discovered that the common ligand features are located in different sites of the receptor's binding pocket and hence create different interactions.

Moreover, all of our ligand-based pharmacophores were not specific enough and we did not get any valid results from virtual screening of [5].

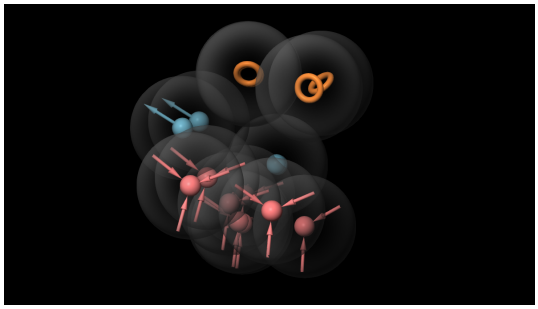

Figure 3: Ligand-based merged Pharmacophore hypothesis.

Another important thing to note was the fact that ligand-based pharmacophore modeling is usually used when trying to infer hypotheses based on known inhibitors. Therefore this approach didn't seem like the best fit considering given data and the possibility that it might not be specific enough for our purpose. We wanted to build the pharmacophore hypothesis based on the receptor-ligand interactions and not only on the common aspects of ligands' structure.

Besides, it is crucial to remember that the pharmacophore model in [1] was receptor-based, not ligand-based. Considering the mentioned arguments, we have decided to create a receptor-based pharmacophore.

2.2 DiffDock-based approach

2.2.1 Protein preparation

Initially, we were planning to build an ensemble model and, in preparation, selected 3 PDB structures: 6ZSL, 7NN0 and 7NIO. We chose them mainly based on resolution. The 6ZSL PDB structure defines two biological assemblies, corresponding to two copies of the Nsp-13 protein. Each assembly was downloaded as a separate .pdb file. Similarily, 7NN0 yielded 4 structures. 7NIO does not define assemblies, but similarily contains two copies. Each copy was extracted into a separate file.

This step yielded 8 distinct structures. Heterogenous atoms were deleted using pdbtools [13]. The structures were then aligned as rigid bodies, pairwise to 6zsl assembly 1, using the alignment

tool at https://www.rcsb.org/alignment with the jFATCAT-rigid method, selecting residues 259-440, corresponding roughly to the ATP binding domain.

The resulting .cif files were transformed back to .pdb using openbabel [14].

Basing our preprocessing on that of DiffDock, the proteins' hydrogens were then fixed with the 'reduce' tool.[15]

This process resulted in 8 prepared structures. However, during initial experimentation we realized that we would not have enough computational resources to fully dock all ligands to all structures. This is in part due to diffusion models being notoriously computationally expensive. For the next steps, we chose 6zsl assembly 1 and 7nn0 assembly 1, to represent respectively the apo and holo forms of Nsp13.

2.2.2 Pose generation with DiffDock

DiffDock is a graph neural network trained to 'reverse the diffusion', transforming a random initial ligand pose by translation, rotation and torsion conditioned on a protein structure. The final pose is assigned a confidence score by a similar machine learning model. This process is repeated (20 or 40 times) and the poses and confidence scores are returned. Authors report achieving sub-2ÅRMSD for the Top-1 pose on 38 percent of the test set, which outperforms traditional docking methods by a large margin. The model was trained and tested on the PDBBind dataset [16]. We use a pretrained checkpoint and code provided by the authors and available at https://github.com/gcorso/DiffDock . We followed the installation and running instructions provided, choosing to substitute the number of samples with 20. Running on Nvidia Titan V and RTX3070 GPUs, the inference took approximately 20 seconds per ligand,protein pair.

For each pair, the pose with the highest confidence score was selected. The resulting .sdf files were coalesced into a single .sdf file per protein structure using rdkit [17] for easier batch processing.

2.2.3 Scoring with GNINA

GNINA is a molecular docking program, based on a fork of AutoDock Vina, which supports scoring using convolutional neural networks.

For each prepared structure and pose set, two scorings were performed: one with the –score-only options, which preserves the supplied pose exactly, and one with the –minimize option, which performs refinement steps approximating local energy minimization, using default settings for other options.

The output contained, among others:

- an AutoDock Vina affinity score expressed in kcal/mol (less is better)
- a CNN computed quality score, expressing probability that the pose is ¡ 2 RMSD
- a CNN computer affinity score, predicting affinity in pK units

2.2.4 Analysis

The output of GNINA was loaded with rdkit into pandas. A subset of ligands for which all steps succeeded was chosen, yielding 2136 rows out of the 2464 starting compounds.

EOS ids were recovered based on the initial bioactives.csv file.

As the minimized poses scored consistently higher, best hits were chosen based on the following criteria:

- top 10 ranked by AutoDock Vina score, docked to 7NN0, minimized pose
- top 10 ranked by AutoDock Vina score, docked to 6ZSL, minimized pose
- top 10 ranked by CNN score, docked to 7NN0, minimized pose
- top 10 ranked by CNN score, docked to 6ZSL, minimized pose

3 Results

3.1 Pharmacophore-based approach

We have created a pharmacophore hypothesis for each of the 15 complexes separately and analyzed all of them visually. Based on ligand interaction diagrams (see example in Fig. 4) and the amount of information and specificity that a single hypothesis brings into the entire model, we have selected a subset of hypotheses to merge. For example, we have entirely excluded the 5RLW, 5RLJ, 5RLI,

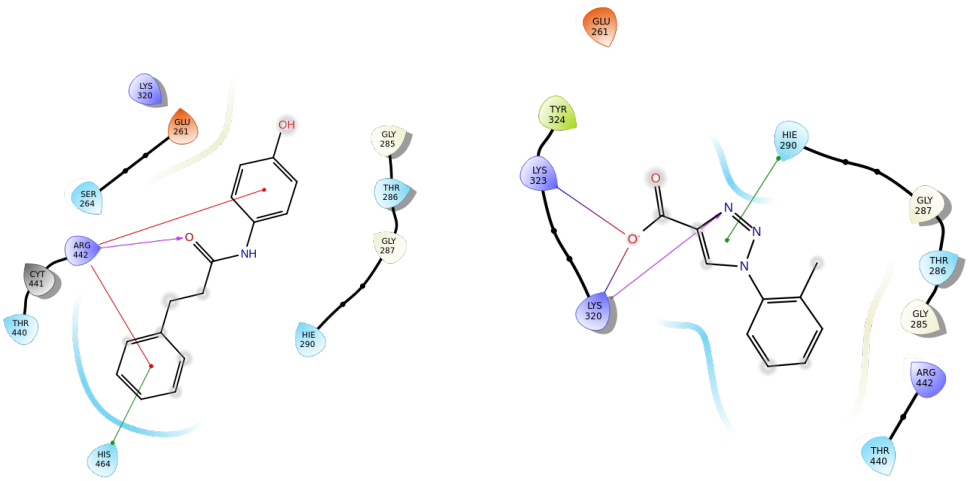

Figure 4: Ligand-receptor interactions diagrams (left: 5rm7, right: 7nng).

and 5RL7 since their hypotheses seemed ambiguous (a ligand's interaction diagram did not support its pharmacophore). From the remaining hypotheses, we were trying to identify the most common features. By merging such features, we have developed multiple hypotheses based on different subsets of the chosen 15 ligand-receptor complexes and screened against bioactives from ECBD database. We validated the hypotheses by analyzing their results of screening a database composed of active molecules (Table 3) and decoys generated for this set of actives using DUD-E database generate tool [3].

Finally, we have chosen a pharmacophore model (Fig. 5) consisting of:

- 2 donors,
- 3 aromatic rings,
- 2 negative charges.

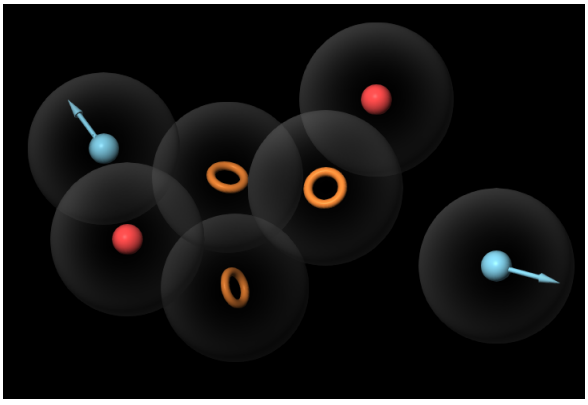

Figure 5: Pharmacophore hypothesis.

This hypothesis was the best in terms of validation performance. After screening it against the database [5], we obtained 76 hits. We then docked it into 7NN0 receptor (the reference structure chosen before). While analyzing the docking results, we encountered a problem. We wanted -10kcal/mol to be the maximal binding affinity threshold, but we get only 2 hits that fulfilled this

Compound name	Compound CID
A-385358	11556440
ABT-737	11228183
Adapalene	60164
Adomeglivant	91933867
Avasimibe	166558
CARM1-IN-1	24827559
Cintirorgon	124126348
Diphenyl Blue	6296
Elaidic acid	37517
Evans Blue	9409
Fenretinide	5288209
Gossypol	3503
GW7647	3392731
Idasanutlin	53358942
Linifanib/ABT-869	11485656
Navitoclax	24978538
NF 023	6093160
Oleic acid	445639
PDK1/Akt/Flt Dual Pathway Inhibitor	5113385
PPNDS	5311367
RO8994	53238217
Suramin	5361
TCID	2729042
TW-37	11455910
Venetoclax/ABT-199	49846579
Zafirlukast	5717

Table 3: Actives used for Pharmacophore hypothesis validation [12].

criterion (UTP and ATP - common in human cells). Due to that, we were forced to lower the threshold - we set it to -9kcal/mol. The best binding affinities are presented in Table 4. Apart from molecules common in the human body, we found 4 molecules (ECBD codes EOS101850, EOS101674, EOS102024, EOS101092). Two of them have their trade names defined in ChEMBL database [9]: Folotyn and Tomudex. The rest can be found in ZINC databases and is available for purchase.

ECBD code	Ligand name (if defined)	Binding affinity (kcal/mol)
EOS100357	Uridine 5'-triphosphate	-10.5
EOS100983	Adenosine 5'-triphosphate (sodium salt)	-10.1
EOS102340	Guanosine-5'-triphosphate	-9.8
EOS101850	-	-9.7
EOS101674	Pralatrexate (trade name Folotyn)	-9.1
EOS102024	-	-9.1
EOS100865	Adenosine-5'-diphosphate	-9.0
EOS101092	Raltitrexed (trade name Tomudex)	-9.0

Table 4: Top AutoDock Vina affinity binding scores for docked ligands.

We visualized docking results in PyMol. The visualizations for EOS100357 and EOS101674 are presented in Fig. 6. A visual summary of the pharmacophore-based approach is presented in Fig. 7.

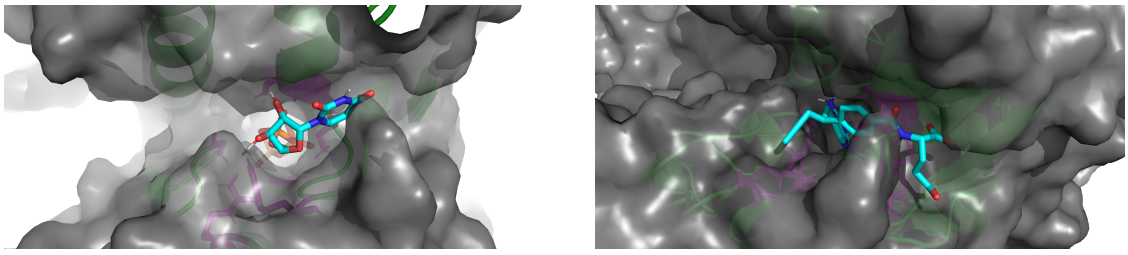

Figure 6: Docking results visualised in PyMol (left: EOS100357, right: EOS101674). The amino acids from ADP binding site are colored magenta and the surface of the receptor is colored gray.

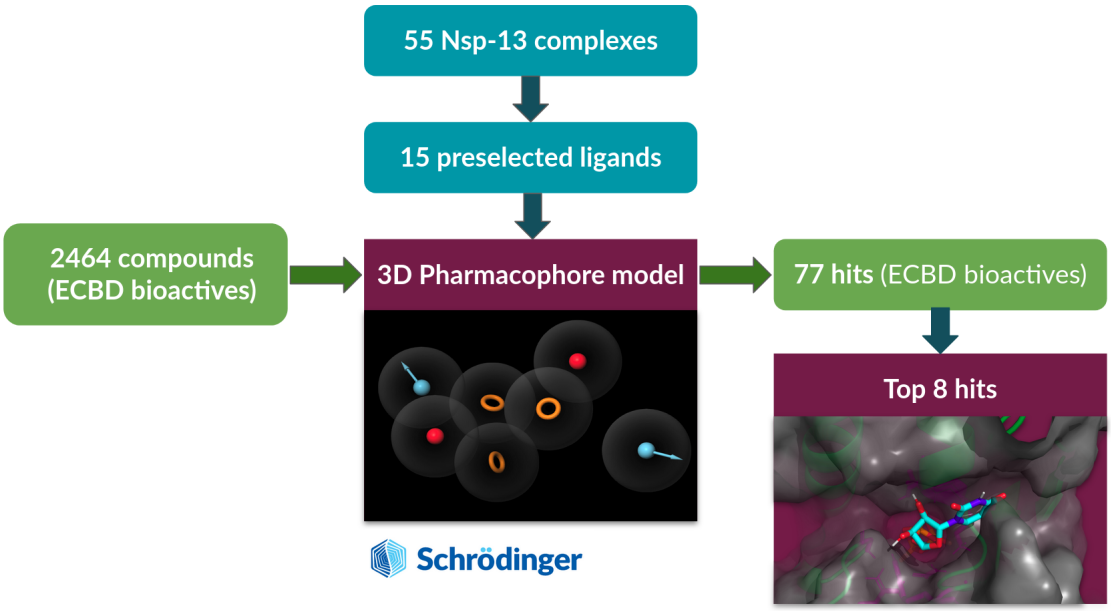

Figure 7: Result figure for pharmacophore-based approach.

3.2 DiffDock-based approach

3.2.1 Top 10 - vina score - 7NN0, minimized

Idarubicin (hydrochloride) was detected in the EOS300008 assay as active.

ECBD code	Ligand name (if defined)	Score
EOS100329	CYCLIC AMP	-9.53
EOS100212	KML29	-9.37
EOS100596	Tadalafil	-9.32
EOS100983	Adenosine 5'-triphosphate (sodium salt)	-9.12
EOS100422	avacopan	-9.01
EOS100681	Teneligliptin	-8.93
EOS101097	irinotecan	-8.92
EOS101383	Idarubicin (hydrochloride)	-8.83
EOS100910	Apigenin 7-glucoside	-8.82
EOS100026	Maslinic Acid	-8.78

Table 5: Top vina scores for 7NN0.

3.2.2 Top 10 - vina score - 6ZSL, minimized

Idarubicin (hydrochloride) was detected in the EOS300008 assay as active.

ECBD code	Ligand name (if defined)	Score
EOS100245	Amentoflavone	-11.56
EOS100687	Mangiferin	-10.07
EOS100870	Bemcentinib	-9.36
EOS100851	Grazoprevir	-9.16
EOS100882	NADP (sodium salt)	-8.76
EOS101383	Idarubicin (hydrochloride)	-8.70
EOS101251	PLX8394	-8.55
EOS101777	itacitinib	-8.42
EOS101357	lifirafenib	-8.38
EOS101260	SCH772984	-8.36

Table 6: Top vina scores for 6ZSL.

3.2.3 Top 10 - CNN score - 7NN0, minimized

ECBD code	Ligand name (if defined)	Score
EOS101269	Venetoclax	7.27
EOS100509	LY2955303	7.02
EOS100971	UNII-Q8MI0X869M	7.01
EOS100781	Bosentan Hydrate	6.95
EOS100979	AP-III-a4 (ENOblock)	6.91
EOS100679	FIIN-3	6.87
EOS101362	CS-1295	6.87
EOS100280	RO8994	6.84
EOS101607	Telaprevir	6.79
EOS101814	Maraviroc	6.78

Table 7: Top CNN scores for 7NN0.

3.2.4 Top 10 - CNN score - 6ZSL, minimized

ECBD code	Ligand name (if defined)	Score
EOS101827	Semapimod 4HCl	7.46
EOS100552	GSK2837808A	7.26
EOS101495	birinapant	7.19
EOS100421	A-1155463	7.17
EOS100971	UNII-Q8MI0X869M	6.97
EOS101549	Aurora A inhibitor I	6.97
EOS100260	AM679	6.81
EOS100895	A-1210477	6.67
EOS100531	UNC0631	6.59
EOS101592	RN486	6.56

Table 8: Top CNN scores for 6ZSL.

4 Discussion

Our work in the project can be summarized not only by the results presented but also by the biological knowledge and practical insights into real-life bioinformatics problems, such as identifying potential inhibitors of different proteins. Due to differences in our academic paths and experience (mathematics, bioinformatics, machine learning, and computer science), first, we had to create a pipeline that would allow using our skills and experiences. Next, we needed to divide the tasks among the members of the team so that we could use the knowledge we already have as well as cooperate and learn from each other. Of course, we also gained many entirely new skills and intuitions - only one of us had heard of a pharmacophore hypothesis before.

While preparing the project, we encountered several problems and we still have some doubts considering the choices we made throughout the project. One of the doubts concerns selecting the ligands building the pharmacophore. As in [1], we chose the ligands from the ones found in complexes with Sars-Cov-2 Nsp-13 helicase in PDB database. The problem is that almost all of them are really shallow in the ADP binding pocket, however, redocked ligands are located deep inside the pocket. We consider such redocking outcomes invalid. Moreover, as in [1], screening the database with the receptor-based pharmacophore based on (not necessarily only) native ligands lying shallow in the pocket, returns ligands, that lie deep in the pocket. We were not sure whether this behavior of results is normal, however we accepted it, as was done in [1].

Considering our results from the pharmacophore-based approach, we hoped to get more ligands with scores higher than the threshold. However, after examining our results, we found out that the ligand (EOS101674) with the trade name Folotyn has been previously identified as a potential Sars-Cov-2 inhibitor [10]. The scientists at the Chinese Academy of Sciences Shenzhen Institutes of Advanced Technology (SIAT) showed that Folotyn (pralatrexate), a chemotherapy drug could be a potential remedy against SARS-CoV-2 - "They found that pralatrexate more strongly inhibited SARS-CoV-2 replication than did Gilead Sciences' remdesivir under the same experimental conditions" [11]

The results for DiffDock need further validation. Unfortunately, the current frequency of 2 active compounds out of 40 suspected hits does not rise significantly above random chance. Please note that the results for DiffDock were obtained by full protein docking and were not filtered to exclude ligands docked outside the ATP pocket. Filtering and re-ranking is an interesting next step - the proteins are already aligned to facilitate filtering. The low rankings of ATP call into question the results obtained by CNN scoring - it ranks around 1200 for 7NN0 and 1400 for 6ZSL. The difference in ranking of adenosine phosphates between 6ZSL and 7NN0 highlights the importance of choosing the right protein conformation. EOS100280 narrowly missed the threshold for being considered active in assay EOS300008. Similarily, EOS100260 demonstrated noticeable activity.

Validation with DUD-E or a similar decoy dataset would be interesting. So far, DiffDock has only demonstrated that it can reconstruct the pose of ligands, but it has not been shown to be able to discriminate from duds. Perhaps it can come up with poses that are convincing for scoring algorithms and even humans, but ultimately false.

References

[1] El Hassab MA, Eldehna WM, Al-Rashood ST, Alharbi A, Eskandrani RO, Alkahtani HM, Elkaeed EB, Abou-Seri SM. Multi-stage structure-based virtual screening approach towards identification of potential SARS-CoV-2 Nsp-13 helicase inhibitors. J Enzyme Inhib Med Chem. 2022 Dec;37(1):563-572. doi: 10.1080/14756366.2021.2022659. PMID: 35012384; PMCID: PMC8757614.

[2] G. Corso, H. Stärk, B. Jing, R. Barzilay, T. Jaakkola - DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking (2022) - https://arxiv.org/abs/2210.01776 - doi 10.48550/ARXIV.2210.01776

[3] Mysinger MM, Carchia M, Irwin JJ, Shoichet BK J. Med. Chem., 2012, Jul 5. doi 10.1021/jm300687e

[4] https://www.schrodinger.com/products/maestro

[5] https://ecbd.eu/

[6] https://ccsb.scripps.edu/mgltools/

[7] J. Eberhardt, D. Santos-Martins, A. F. Tillack, and S. Forli. (2021). AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. Journal of Chemical Information and Modeling.

[8] O. Trott, A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, Journal of Computational Chemistry 31 (2010) 455-461

[9] https://www.ebi.ac.uk/chembl/

[10] Senthil Srinivasan, R. A.; Meenakshi, R. Computational Screening of Folate Analogues against the Sars-Cov-2 Corona Virus by Molecular Docking. Journal of Advanced Scientific Research, v. 11, p. 176–180, 2020.

[11] https://www.fiercebiotech.com/research/ai-spots-lymphoma-chemotherapy-as-stronger-covid-19-drug-than-gilead-s-remdesivir

[12] Zeng J, Weissmann F, Bertolin AP, Posse V, Canal B, Ulferts R, Wu M, Harvey R, Hussain S, Milligan JC, Roustan C, Borg A, McCoy L, Drury LS, Kjaer S, McCauley J, Howell M, Beale R, Diffley JFX. Identifying SARS-CoV-2 antiviral compounds by screening for small molecule inhibitors of nsp13 helicase. Biochem J. 2021 Jul 16;478(13):2405-2423. doi: 10.1042/BCJ20210201. PMID: 34198322; PMCID: PMC8286831.

[13] Rodrigues JPGLM, Teixeira JMC, Trellet M and Bonvin AMJJ. pdb-tools: a swiss army knife for molecular structures. F1000Research 2018, 7:1961 (https://doi.org/10.12688/f1000research.17456.1)

[14] O'Boyle, N.M., Banck, M., James, C.A. et al. Open Babel: An open chemical toolbox. J Cheminform 3, 33 (2011). https://doi.org/10.1186/1758-2946-3-33

[15] Word, et al.(1999) "Asparagine and glutamine: using hydrogen atom contacts in the choice of sidechain amide orientation" J. Mol. Biol. 285, 1735-1747. (doi.org/10.1006/jmbsecondi.1998.2401)

[16] Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li, and Renxiao Wang. Forging the basis for developing protein–ligand interaction scoring functions. Accounts of Chemical Research, 50 (2):302–309, 2017.

[17] RDKit: Open-source cheminformatics. https://www.rdkit.org