Hoai-Huong Emilie DOAN
Brenda ENRIQUEZ
Birsu GUZEL
Isabelle WU

MEET-U – Report

Proposing potential inhibitors for the Sars-CoV-2 Helicase (Nsp13) by Molzip approach

Supervised by :
Juliana Silva Bernardes
Elodie Laine
Vaitea Opuu

2023-2024 Academic Year

SUMMARY

Abstract

Since the SARS-COV-2 pandemic in 2019, several research has been conducted to treat coronavirus infections. A lot of attention has been devoted to non-structural proteins (nsp) as potential targets, particularly NSP13 due to its high sequence conservation and its essential role in viral replication. NSP13 encodes for the virus helicase protein which has an important role in replication, transcription, and translation. Defects in helicase can lead to many genetic disorders, making NSP13 an excellent antiviral target (Spratt et al., 2021). Moreover, many studies demonstrated that NSP13 contains multiple druggable pockets that are known for perturbing the helicase activity (Marecki et al., 2021). The field of ligand prediction for pockets is a growing field. Prediction methods are increasingly used in pharmaceutical research to identify potential inhibitors or to improve known inhibitors. Different computer methods exist and provide important assistance that speeds up research and is efficient.

Introduction

In order to disrupt the SARS-COV-2, we want to identify NSP13 potential inhibitors capable of binding to target pockets. Different approaches exist to predict the binding of a ligand to a pocket, either by a docking method or by machine learning.

Predicting which ligands bind to an active site is a complex problem that can be solved through optimization of structure and thermodynamic equilibrium. These algorithms already exist but require a lot of calculation time to predict and require high-quality data, as well as many features to be measured, which makes this process slow and expensive. Indeed, they are based on complex databases to make their predictions, which may contain erroneous structures. Also, with a classic docking approach, performance can be sensitive to the parameters of the algorithm used.

Our goal is to test Molzip, a new prediction approach that is unconventional as it does not take into account three-dimensional structures to predict (Probst, 2023). Molzip is a classification method based on information theory and uses probabilities to quantify the average information content of a set of messages, including computer coding. And promising results on the prediction of physicochemical variables, but also the prediction of ligand-protein affinity were obtained with this algorithm with only the sequence in the input, which is little information to predict.

Also, we used this notion of compression that we entered into machine learning methods known as Random Forest, Neural Network, and others to see if using the same idea as Molzip would give us interesting results in different algorithms with the same data.

Our approach

Molzip

We chose to study the technique “Molzip” developed and presented by Daniel Probst in a recent article published in 2023 and called “Parameter-Free Molecular Classification and Regression with Gzip” (Probst, 2023). Firstly, we want to test the reliability of the prediction of this tool via different tests and then use it to predict potential inhibitory ligands.

First, we tested the prediction of variables such as molecular weight with a database of over 5,000 ligands. We applied the algorithm while knowing the molecular weight value of the ligands we are trying to predict and we were able to display on a plot the predicted values versus the true values. By tracing the diagonal, we see that the predictions are very close to the latter, thus showing the potential of this tool (Figure 1a). We also applied PCA to visualize the distribution of variables. Thanks to this, we see clusters forming, the low values group together, and the same for the high values (Figure 1b). This graph allows us to confirm that Molzip succeeds in differentiating and predicting ligand variables from only their SMILES sequences.

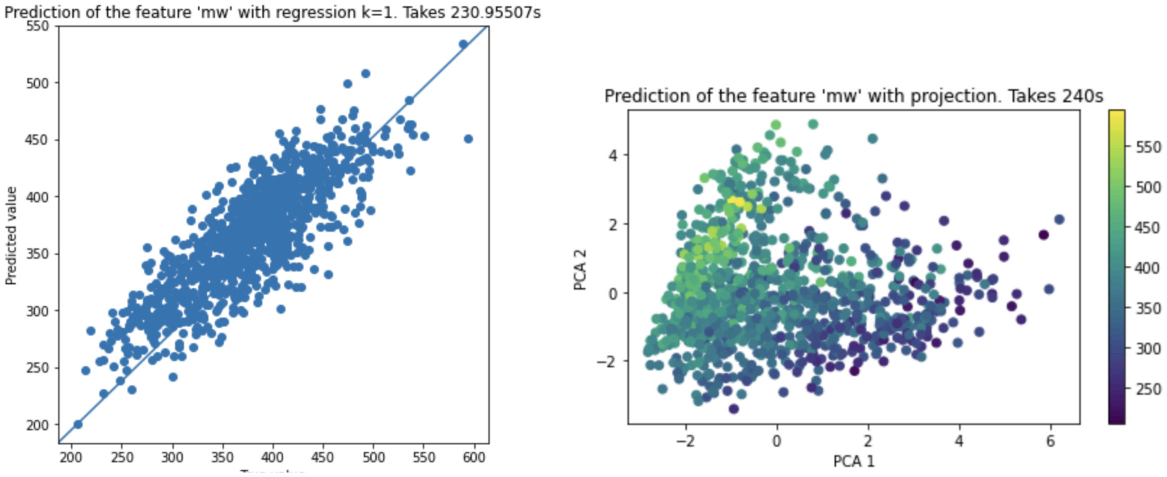


Figure 1. Prediction of the molecular weight with Molzip (Figure.1a in the left) and projection with PCA (Figure.1b in the right)

Second, we used a database (Li et al., 2023) containing ligands and pockets with an affinity score. The database contains ligands that bind to different pockets and we did a cross-validation to predict the affinity scores of the ligands, and to check the performance of the model. Each split gave us similar plots showing the consistency of the model (Figure 2).

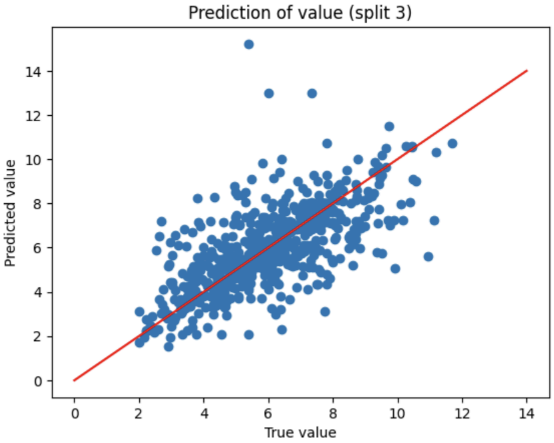


Figure 2. One of the plots given by the cross-validation with the dataset of PDB Bind

Therefore, in order to predict ligands for the ARN and ATP pockets of NSP13, we used the previous dataset (Li et al., 2023) containing the protein-ligand binding affinity scores as a training set. Then, we predicted the affinity of potential ligands (dataset of 5000 potential ligands). We considered that ligands with the highest affinity value are the best-predicted ligands.

Method with known Machine Learning

We wanted to test different supervised machine learning techniques on the PDB Bind dataset in order to compare our results with the Molzip method. To achieve this, we used the same idea of "zip" the ligand with the possible pocket which consists of taking the length of the compression of these sequences. Then, we tested a Neural Network, Random Forest, AdaBoost, KNeighbors, GradientBoosting, and SVM. Moreover, according to their performance on the PDB Bind dataset, we used some of these methods on the dataset with 5000 possible ligands to select the best ligands found by supervised machine learning algorithms.

Materials and Methods

Dataset

We use two differents datasets:

1. Dataset with 5000 possible ligands,
2. PDB Bind Dataset. As well as NSP13 pocket sequences.

Dataset of 5000 possible ligands

We will use a dataset of 5000 potential ligands in SMILES format provided to us by our professors. The file also contains other variables describing the ligands, such as their molecular weight, the hydrophobicity indicator, or the ligand identifier on PubChem. For our project, we will only use the SMILES and we will predict if they are potential inhibitors of NSP13.

PDB Bind dataset

The PDB Bind file was missing in the GitHub of Daniel Probst. Therefore, we found the file "LP_PDBBind.csv" in order to use this database as a training set for our predictions and validate the Molzip method (Li et al. 2023). This database was originally from the PDBBind website but was modified and prepared by Li et al.. We contacted Jie Li to have more information for the columns "kd/ki" and "value", and he told us that the column "value" – which is the affinity of the ligand for the given pocket – corresponds to the negative log of the Ki, Kd or IC50 values; he didn't differentiate the Ki, Kd or IC50 when calculating the value. We are interested in the ligand-protein interaction affinity, so we don't mind if they are not differentiated but we decided to remove the rows which have an IC50 value, leaving us with 13.000 rows. We also preprocessed the data by converting all Ki and Kd values to the same unit, in nanomolar. Kd is a dissociation constant and Ki an inhibitor constant, these constants reflect the ligand-protein binding affinity. Then, we realized a cross-validation to confirm the relevance of the dataset. We used the columns "smiles" and "seq" to representing respectively the SMILES of the ligands and the sequence of the binding pocket. We also used the column "values"

NSP13 pockets sequences

Given that we train our model with SMILES of ligands and the sequences of their binding sites together, we had to find the sequences of pockets of interest on NSP13 to concatenate them with the SMILES of possible ligands. Several studies looking for potentially druggable pockets on NSP13 were made and two pockets were found, one of them is an ATP pocket and the other one is an ARN pocket (Austin et al., 2021). In order to find the amino acid sequence of these pockets we used Pymol with the crystal structure of the SARS-CoV-2 helicase in complex with AMP-PNP (PDB:7NN0) and the SARS-CoV-2 replication-transcription complex bound to nsp13 helicase – nsp13(2)-RTC – apo class (PDB:7RDX).

Molzip algorithm

Molzip is a method based on the theory of information. The method proposed by Daniel Probst compresses the strings defining different ligands and the main idea is that if the strings are similar, the difference between the length of the compressions would be null.

When a new ligand is added to the database, Molzip algorithm compares the length of the compression of the new ligand with other ligands of the database which have a known score. It uses the kNN classification technique and it chooses the "k" closest neighbors between all the ligands of the database and returns the barycentre of the scores of "k" chosen ligands. For k=1, the algorithm will take into account only a single ligand which is the closest and it will allow us to classify the new ligand. With k=10, we can choose 10 ligands closest to the query ligand and we can apply a PCA to cluster and visualize the results.

In order to apply the Molzip algorithm, we concatenated the SMILES of ligands with the sequences of the binding pockets. Afterward, we gave these concatenated SMILES and pocket sequences, along with the true labels we wanted to predict as input. In our case, we predicted the affinity values for each ligand.

Supervised Machine Learning Algorithms

First, we used the PDB Bind dataset to train our models. To do this, we concatenate the SMILES with the sequence and use the "zip" algorithm to compress them. Then, we created a new dataset where we have three columns: 1. The length of the zip, 2. "kd/ki", 3: "values". We used the length of the zip to predict 2 and 3 columns independently, to know which column allowed us to achieve the best result.

We performed a cross validation for each proposed method to know which performs better, and also once we have trained our neural network. We tested with the Dataset of 5000 possible ligands to know which ligands could have a better fixation to our 2 pockets: ARN and ATP by using some of the supervised machine learning methods.

<u>Neural Network:</u> We used the "keras" library in python for the architecture of the neural networks. In order to predict "the value" column of PDB Bind files, we used the activation function "linear" at the last layer because we want to predict continuous values.

To predict the value kd/ki, we divided the values into different categories. Therefore, we used the activation function “softmax” at the last layer of the neural network to predict the different classes.

Supervised Algorithms: We used the algorithms from Scikit Learn. We tested different numbers of groups and according to references, we set the ranges so we let group 0 as the one that will select the ligands that will fix better with the pocket. We proposed the following groups:

Number of groups	kd/ki	value
2	[0-600] (600,inf)	[0-14] (14-16]
3	[0-500] (500-900] (900-inf)	[0-2] (2-10] (10-16]
4	[0-20] (20-500] (500-900] (900-inf)	[0-3] (3-5] (5-7] (7-16]

Results

Identified sequences of NSP13 pockets

In order to identify the sequences of the two pockets, we used two different SARS-CoV-2 crystal structures complexed with either AMP (PDB:7NN0) or the RNA template (PDB:7RDX) (Figure 3). With Pymol we selected amino acids from the NSP13 structure which are at a distance of 5Å or less from the ligand. Then to get the sequence of the pocket, we took the sequence of amino acids between the first and the last amino acid at a distance of 5Å (Figure S1), as we do not take in consideration the three-dimensional structure of the pocket, we have in the sequence a lot of amino acids which are actually not close to the ligand pocket. For the sequence of both pockets, we identified close amino acids – shown in the figures – that were also found in the literature (Austin et al., 2021)(Newman et al., 2021).

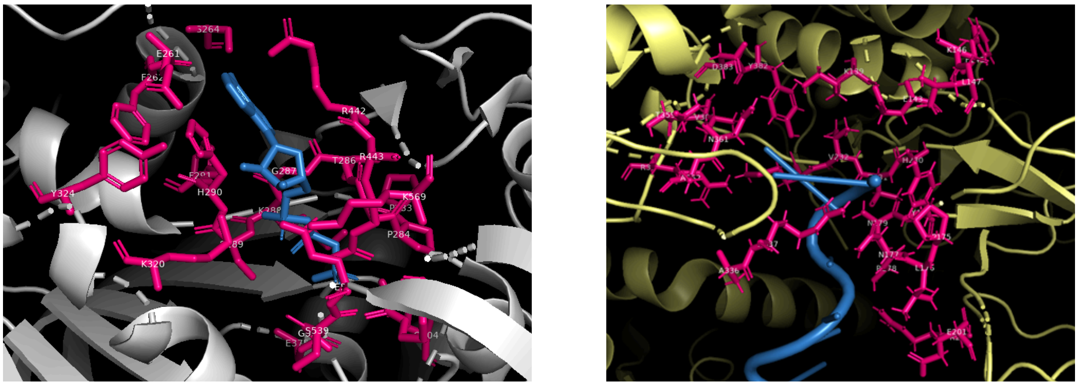


Figure 3. Crystal structure of the SARS-CoV-2 helicase in complex with AMP-PNP (PDB:7NN0) or the RNA PDB:7RDX). Focus on the ATP pocket of Helicase nsp13 (on the left) and on the RNA pocket of nsp13 (on the right). The AMP (left) or RNA structure (right) are in blue, the amino acids of NSP13 around 5Å distance from the AMP in magenta.

Molzip

To validate the performance of the Molzip technique, we tested the method on different predictions like we said earlier in the approach. For each pocket, we predicted the potential ligand's affinity value and selected those with the highest affinity scores. To do so, we used "LP_PDBBind.csv" as a training set and we only took as entries lines which have a K_D value so that our training base has an affinity score which is based on the same constant which is here the equilibrium dissociation constant (K_D). Therefore, we used around 6000 samples and took "value" as a prediction value. Molzip has no problem in predicting affinity scores as it can manage continuous values. The highest values represent the best affinities and we took the ranking of the 20 best potential ligands (Figure S2). These ligands can be found with their EOS numbers on the data file "pilot_library.csv" and also on the internet. It should be noted that certain ligands have been predicted to bind with both pockets, which is not normal. Our results with Molzip are therefore not relevant.

Different Machine Learning Methods

	Neural Network	Random Forest	AdaBoost	Gradient Boosting	SVM
kd/ki	0.54 (±0.02)	0.540873	0.534415	0.540873	0.542290
values	0.84 (±0.01)	0.977004	0.977319	0.977477	0.977477

Table 1. Accuracy for K_D *and values (4 categories)*

When we predicted the "kd/ki" and "values" columns of the PDB Bind file, the supervised machine learning algorithms didn't show a good performance so we decided to make categories according to literature. After categorizing the values of the "kd/ki" columns, we observed accuracy scores around 0.5. This indicates that the classifiers obtained by the tested methods perform similarly to a random classifier, achieving approximately 50% accuracy. These results suggest that the classifiers tend to give incorrect results around half of the time.

According to our methodology, we got better results when we categorized the “values” column than when using the “kd/ki” feature (categorized) to classify the fixation index (Table 1). However, while testing with the Dataset 5000, we realized that it was hard to predict and classify them as this value is in logarithm scale and so a very small variance in the dataset caused a lot of divergence in the results. On the other hand, “kd/ki” proved a lower accuracy level but since we have information about the performance of this parameter in the fixation, we decided to use this feature to classify the possible fixation ligands on the dataset of 5000 ligands, and we set a very small range for class 0 so we are very confident that we’re proposing highly good performing ligands.

Comparaison

By comparing the results obtained with Molzip and supervised machine learning algorithms, we observe that Molzip predicts better the “values”, meaning affinities of different ligands, when we test the “LP_PDBBind.csv” data. However, Molzip performs as a random classifier when we want to predict the categories of kd/ki, like the supervised machine learning algorithms.

After testing our models, we found the following top of sequences to be analyzed further to confirm if they might work as inhibitors. It is important to mention that while using AdaBoost as well as SVM, the models didn’t find any ligand that might fix our ATP Pocket. Also, with the proposed methods it was harder to find inhibitors to our ARN Pocket. While comparing the results given by the different supervised learning methods, we didn’t find a common set of proposed inhibitors, as we expected, there were only three common inhibitors found for ATP Pocket (Highlighted in Tables 2 and 3). However, we don’t observe any ligands in common between the ligands found with Molzip and the ligands found with supervised learning algorithms.

Discussion

First of all, the article written by Daniel Probst in order to represent the Molzip method is a preprint version and it hasn’t undergone peer review. Therefore, the number of documentation about the Molzip algorithm is really limited. Moreover, the code isn’t well commented and there are some missing files such as “meta.csv”. The results obtained with the code are not significant, in fact it displays scores out of nowhere which in no way prove the robustness of the algorithm because it does not explain it clearly. We tried to run the code ourselves but it was not possible. As a consequence, we had difficulties understanding and running the code with the documents Daniel Probst proposes in the github for implementing Molzip. We tried contacting Daniel Probst to ask about the missing files but we didn’t get any response. Therefore, we had to carry out tests to confirm the relevance of the algorithm.

Second, even though if we do not use complex data containing structure and features, the chosen database implies a selection bias which means that we learn in particular affinities of ligands and that this is not suitable for perfectly predicting all the ligands protein binding. And another problem is that unlike other machine learning models which can fit a training set, Molzip does not have this aspect to fit (that the model learns) and needs to compare with all the training sets every time that we want to make a prediction.

The execution time of the Molzip algorithm was long when we executed it on large databases, such as the dataset containing 19000 sequences (PDB Bind dataset). However, the results that were interesting for us took less calculation time, it took only 1 hour to predict the affinities for the two pockets and that was faster than using classical prediction algorithms to predict all the 5000 ligands but that does not mean better prediction.

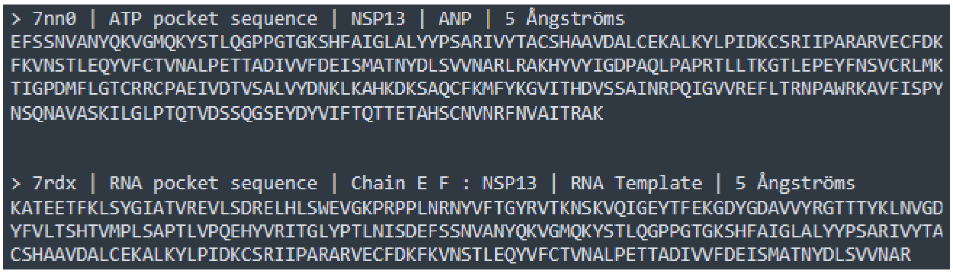
We were expecting to find similar outputs while evaluating our 5000 ligands with different methods but we found that different methods gave us very divergent results even using the same input coming from “zip”. We think that the different results may be due to the fact that the tested methods function differently in order to give predictions. Therefore, the ligands that were found by the algorithms should be biologically studied to see if the predicted ligands fix really into the pocket, which can also give an idea about which prediction tool works better.

Conclusion

After our experiments, we can prove that by using “zip” we reduce significantly the size of the database and this can be used to give a first approach of which ligands can be interesting. However, it might be used with other methods or with complete sequences once made the first filter using the proposed methodology.

We don’t have any results that are relevant at the moment with Molzip. Through our predictions, we can notice that this approach can surprisingly achieve some good results but it is not completely precise and fair. This may depend on the database. It would therefore be interesting to continue with it but by applying several other methods in addition to improve the results and reliability. This method can also make it possible to make an initial selection of potential ligands in order to reduce the number of ligands that could be docked with more traditional methods.

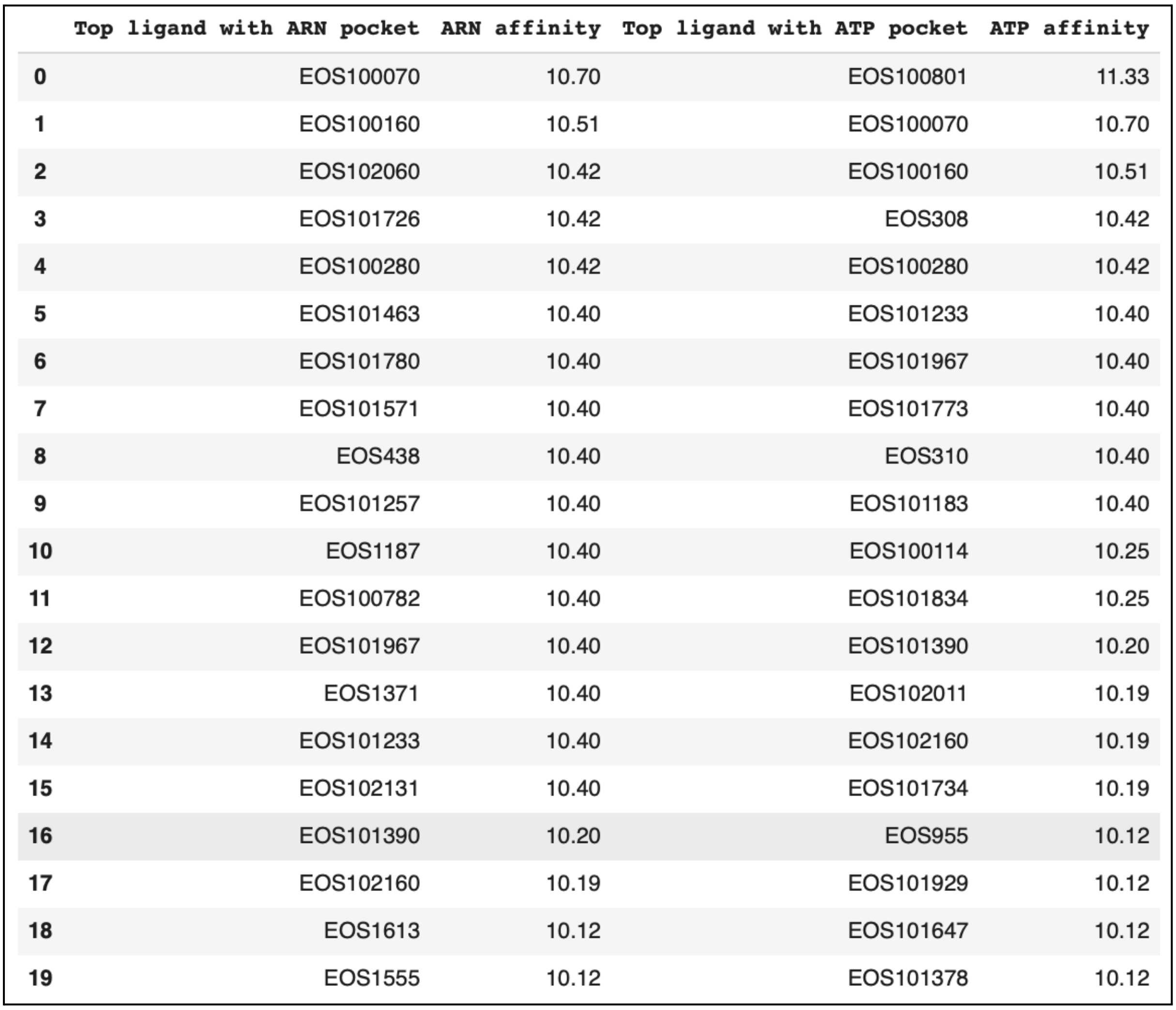
Annexes



```
> 7nn0 | ATP pocket sequence | NSP13 | ANP | 5 Ångströms
EFSSNVANYQKVGMQKYSTLQGPPGTGKSHFAIGLALYYPSARIVYTACSHAAVDALCEKALKYLPIDKCSRIIPARARVECFDK
FKVNSTLEQYVFCTVNALPETTADIVVFDEISMATNYDLSVVNARLRAKHYVYIGDPAQLPAPRTLLTKGTLEPEYFNSVCRLMK
TIGPDMFLGTCRRCPAEIVDTVSALVYDNKLKAHKDKSAQCFKMFYKGVITHDVSSAINRPQIGVVREFLTRNPAWRKAVFISPY
NSQNAVASKILGLPTQTVDSSQGSEYDYVIFTQTTETAHSCNVNRFNVAITRAK


> 7rdx | RNA pocket sequence | Chain E F : NSP13 | RNA Template | 5 Ångströms
KATEETFKLSYGIATVREVLSDRELHLSWEVGKPRPPLNRNYVFTGYRVTKNSKVQIGEYTFEKGDYGDAVVYRGTTTYKLNVGD
YFVLTSHTVMPLSAPTLVPQEHYVRITGLYPTLNISDEFSSNVANYQKVGMQKYSTLQGPPGTGKSHFAIGLALYYPSARIVYTA
CSHAAVDALCEKALKYLPIDKCSRIIPARARVECFDKFKVNSTLEQYVFCTVNALPETTADIVVFDEISMATNYDLSVVNAR
```

Figure S1. Amino acids sequence of targets pockets



	Top ligand with ARN pocket	ARN affinity	Top ligand with ATP pocket	ATP affinity
0	EOS100070	10.70	EOS100801	11.33
1	EOS100160	10.51	EOS100070	10.70
2	EOS102060	10.42	EOS100160	10.51
3	EOS101726	10.42	EOS308	10.42
4	EOS100280	10.42	EOS100280	10.42
5	EOS101463	10.40	EOS101233	10.40
6	EOS101780	10.40	EOS101967	10.40
7	EOS101571	10.40	EOS101773	10.40
8	EOS438	10.40	EOS310	10.40
9	EOS101257	10.40	EOS101183	10.40
10	EOS1187	10.40	EOS100114	10.25
11	EOS100782	10.40	EOS101834	10.25
12	EOS101967	10.40	EOS101390	10.20
13	EOS1371	10.40	EOS102011	10.19
14	EOS101233	10.40	EOS102160	10.19
15	EOS102131	10.40	EOS101734	10.19
16	EOS101390	10.20	EOS955	10.12
17	EOS102160	10.19	EOS101929	10.12
18	EOS1613	10.12	EOS101647	10.12
19	EOS1555	10.12	EOS101378	10.12

Figure S2. Top 20 potentials ligands found with Molzip

Neural Network	Random Forest	AdaBoost	Gradient Boosting	SVM
None	EOS1801 EOS2344 EOS2347 EOS100612 EOS100745 EOS101359 EOS101497 EOS101962 EOS101985 EOS101825 EOS101859 EOS102084 EOS102142 EOS102212 EOS102354	None	EOS933 EOS1768 EOS1776 EOS1801 EOS2263 EOS2344 EOS2347 EOS100170 EOS100134 EOS100163 EOS100378 EOS100390 EOS100567 EOS100672 EOS100612 EOS100745 EOS100756 EOS100890 EOS100894 EOS101020	None

Table 2. Proposed inhibitors for ATP Pocket (4 categories)

Neural Network	Random Forest	AdaBoost	Gradient Boosting	SVM
None	EOS107 EOS110 EOS123 EOS159 EOS308 EOS472 EOS735 EOS736 EOS652 EOS656 EOS482 EOS717 EOS511 EOS989 EOS1097 EOS1298 EOS1379 EOS1388 EOS1550 EOS1562	None	None	None

Table 3. Proposed inhibitors for ARN Pocket (4 categories)

References

1. Probst D. Parameter-Free Molecular Classification and Regression with Gzip. ChemRxiv. 2023; doi:10.26434/chemrxiv-2023-v1s2s.
2. A.N. Spratt, F. Gallazzi, T.P. Quinn, C.L. Lorso, A. Sönnerborg, K. Singh. Coronavirus helicases: attractive and unique targets of antiviral drug-development and therapeutic patents. Expert Opin Ther Pat. 2021;31(4):339-350. doi:10.1080/13543776.2021.1884224
3. J.C. Marecki, B. Belachew, J.Gao, K.D. Raney. Chapter Ten - RNA helicases required for viral propagation in humans. The Enzymes. 2021; 50: 335-367. https://doi.org/10.1016/bs.enz.2021.09.005
4. Li J, Guan X, Zhang O, Sun K, Wang Y, Bagni D, Head-Gordon T. Leak Proof PDBBind: A Reorganized Dataset of Protein-Ligand Complexes for More Generalizable Binding Affinity Prediction. Preprint. 2023.
5. Austin N. Spratt, Fabio Gallazzi, Thomas P. Quinn, Christian L. Lorson, Anders Sönnerborg & Kamal Singh. Coronavirus helicases: attractive and unique targets of antiviral drug-development and therapeutic patents, Expert Opinion on Therapeutic Patents. 2021; 31:4, 339-350, doi: 10.1080/13543776.2021.1884224
6. 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). https://doi.org/10.1038/s41467-021-25166-6