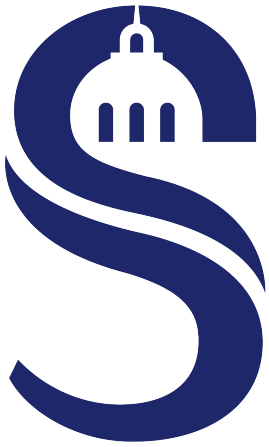

SORBONNE UNIVERSITÉ

Summary MEET-U 2023 / 2024

Henri GUILHON Mathys DELATTRE Lorenzo BALLESTRA
Quentin DESPRETZ

January 31, 2024

Contents

Abstract

The aim of this project is identifying potential therapeutic molecules targeting Sars-Cov2's nsp13 using docking and machine learning tools. We will be focusing on the ATP pocket of nsp13 as an hypothetical therapeutic target.

1 Introduction

We will try to determine whether a Neural Network is capable of predicting the ligands that have the best affinity to the ATP-pocket, thus saving a considerable amount of time by pre-selecting potential inhibitor candidates.

2 Method

2.1 Docking Tool : AutoDock Vina

First, we will be using Autodock-Vina to compute the affinity of several ligands to the ATP pocket. This score informs us about the strength of the bonding between the tested ligand and the ATP pocket.

We chose to focus on the ligands contained in the pre-selected dataset. We targeted the ATP-pocket that was found by searching the amino acids forming hydrogen bonds with ATP made in the literature. Vina will dock up to 20 conformations for each ligand and return the affinity of the ligand and its docked positions on nsp13. Those informations will subsequently be used to train the neural network and to evaluate the potential of the ligand.

2.2 Converting database into graphs

The next step consists in obtaining the graph representation of the ligands from their SMILES formula with the help of RDKIT's chem package. The atoms are represented as nodes and the bonds as edges. The graph will also take into account :

- the atomic number of atoms
- the degree of nodes
- the formal charge
- the presence of an aromatic cycle

2.3 Selecting the conformation

A "pocket occupation" score will finally be introduced as a way to evaluate the degree with which the ligand can substitute ATP and the specificity of the ligands to fit in the ATP pocket.

First, every amino acid is simplified as their alpha carbon in order for every docked molecules to have the same atom references. Then taking the ATP as the reference, the minimal distance between each of its atoms with all alpha carbons were taken. The mean maximal of mean distances of all conformations of ATP in its pocket is around 5 Å : this will subjectively now be the threshold of reference.

From this threshold were computed another research between ATP and nsp13 : all amino acids that are below 5Å of an atom of the ATP will be kept as a reference ATP-pocket amino acid. This search is made among 9 conformations returned by AutoDock-Vina and only the amino acids with more or equal than 10 occurrences were kept as reference. This allows to search for reference amino acids taking in account the plasticity of ATP in its pocket.

By calculating the proportion of amino acids of a given conformation below 5Å to a reference amino acid in the pocket, we hope to be able to establish a new criteria to assess a ligand's pocket occupation. With the multiple conformations given by vina for a single ligand, the one with the highest occupation score will be chosen instead of the first with the highest affinity.

2.4 Training the neural network

Now that the graph representation of the ligands have been successfully computed, we want to train a neural network whose ultimate goal will be to highlight the ligands with the highest affinity scores without having to dock them all. The inputs into the model are the graphs obtained from the ligand's SMILES formula and the output is the predicted affinity of this ligand on the ATP pocket.

3 Results

3.1 Overall Observations

The results are very encouraging, we obtained a model capable of sorting the molecules and extracting the potential best hits within seconds. We found out that the best use for it was to filter the entire dataset and throw out the ligands that are sure to be the ones with the lowest affinity.

3.2 Network observations

The network nevertheless allows to quickly select the best predicted affinities among all the database. Thus fulfilling its purpose as to select the best potential inhibitor candidates. These ones will have to be docked again but docking the entire database will not be needed anymore with the network as it tackles the computational cost of docking.

We tried different input parameters to attest what is best for an accurate prediction. We saw that smaller batches lead to better results overall as well as a much smaller training time.

At the end of the pipeline, we predicted the 16 best candidates of the pilot library that was given to us. These candidates were afterwards compared to the real docking affinity. The results confirmed what we observed beforehand, our prediction contains mostly high scoring molecules which are promising, but also some relatively bad candidates.

Our network is able to predict the affinity scores of the 5000 molecules forming the dataset in a few seconds. While the network can help predict some of the top scoring molecules instantly, it can not replace any docking model. This model should help pre-processing huge datasets by filtering ligands with a high affinity.

4 Discussion

4.1 Network Accuracy

The relative low accuracy of the model flatten the highest scores of affinity of some molecules : lowering them from -10.2 to -8.9 as an example. Thus the best potential inhibitors could be mixed up with ones with lower affinity (-8 which is still a good affinity considering ATP's is -7.4)

4.2 Batch Size and Training Dataset

Smaller batches come with a model more sensitive to local minimums. Also adding the hydrogens gives best performance at the expense of the training time. Furthermore the training dataset was chosen randomly among the docked molecules, however the docked molecules were not.

5 Conclusion

The network is still imperfect with a mean squared error relatively high, but it allows to make a fast filtering of a huge database more consequent than the one used here. The addition of some outliers molecules with very high or very low affinities could make the network differentiate them better. Adding more features to the network could help improving its accuracy but computational training times are wide and don't allow many tries.