Deep learning on protein-ligand pairs

Identification of potential inhibitors of the Nsp-13 protein of the SARS-CoV-2 virus

Warsaw Team 1- Sorbonne Team 2

JAKUB JÓZEFOWICZ, JAKUB SKRAJNY, MICHAŁ KRUTUL,
ZOFIA KOCHAŃSKA, JULIA SMOLIK, MIESZKO SABO

1. Materials & Methods

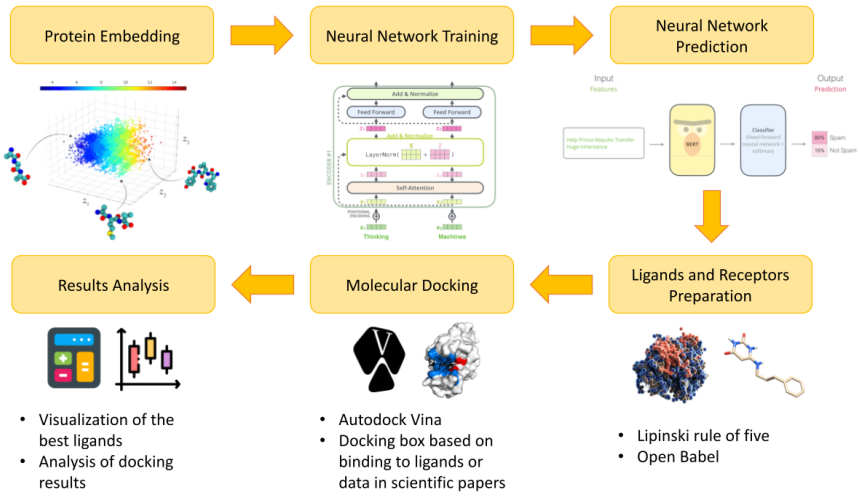

Figure 1: Diagram of the analyses described in detail in the paper

1.1. Neural network

The neural network model was created on the basis of Transformer [1]. The input for training the neural network were protein-ligand pairs with knowledge about their IC50 value. The proteins were represented by the amino acid sequence, and the ligands were represented in SMILES format. The protein and ligand before proceeding to the next stages of model training were first encoded using ChemBERTa [2] and UniRep [3, 4] models respectively.

The training set consisted of ∼500,000 protein-ligand pairs. The SMILES codes of the ligands were retrieved from the BindingDB database [5]. The amino acid sequences of the proteins to which the selected ligands were bound were retrieved from the PDB database [6].

The output of our trained neural network is a probability of a ligand-protein pair having an IC50 value $< 10\mu M$.

The trained network, ligands as well as the test dataset, i.e. resolved Nsp-13 structures 6zsl, 7nio, 7nn0, 5rm2, 5rme (prepared and optimized based on structures from PDB: 6ZSL, 7NIO, 7NNO, 7RDX, 7RDZ) were used to identify potential inhibitors of the Nsp-13 protein [7].

1.2. Ligands & receptors preparation and docking

We limited the available set of ligands, retaining only those that met each of the rules described in Lipinski's rule. The set of 4 rules described by Christopher Lipinski allows to identify those molecules that can potentially act as an orally administered drug. Molecules that were selected as potential inhibitors were then used to build their 3D structure using Open Babel software.

All structures of the selected receptors (Nsp-13 protein structures) contained several protein chains, which were removed from each structure, except for chain A. Ligands were also removed from structures that were not apoproteins. The edited receptors were then exported to mol2 format using Chimera [8], and then using Open Babel converted to pdbqt format.

After obtaining the 3D structures of the ligands and the edited receptor structures, we proceeded to molecular docking used to predict the conformation of bound molecules and their binding affinities. Compounds that the neural network classified as potential inhibitors of the Nsp-13 protein were docked in the AutoDock Vina (Vina) program [9, 10].

Docking of the selected ligands was carried out for all of the selected structures of the Nsp-13 protein. For all the receptors, the size and position of the docking box were identified by analyzing the protein's binding to other ligands (if any). Amino acids that were within 5 Å of the bound ligand were considered to belong to the active site of the protein. If the receptor was an apoprotein, we determined its active site based on other scientific publications [11].

Each docking result is described by a score value (affinity). The smaller the value, the better the docking result.

2. Results

2.1. Neural network

Applying the criteria of Lipinski's rule on the set of potential inhibitors allowed us to limit the set of searched molecules to 333736 compounds that have a chance after oral administration to penetrate the cell barrier and reach their target. The ligands were sorted by probability value, which was the resulting value from the neural network prediction. We selected the first 500 ligands (most likely to be a good inhibitor based on the IC50 value) that additionally met all the Lipinski rule requirements and further analyzed them. Additionally, we have chosen 100 ligands with the lowest score to serve as a control group to check if our neural network can serve as a useful tool for ligand selection.

2.2. Molecular docking

Each docking resulted in 9 structures of a given ligand docked to the receptor. For each docked ligand, we selected its best conformation based on binding affinity value. In the next step, for each receptor, we chose 100 ligands that had the lowest binding affinity value. Then, we looked for common ligands amongst all of the top particles chosen for each of the Nsp-13 structures. This resulted in ten particles from PubChem which are shown in Fig. 2.

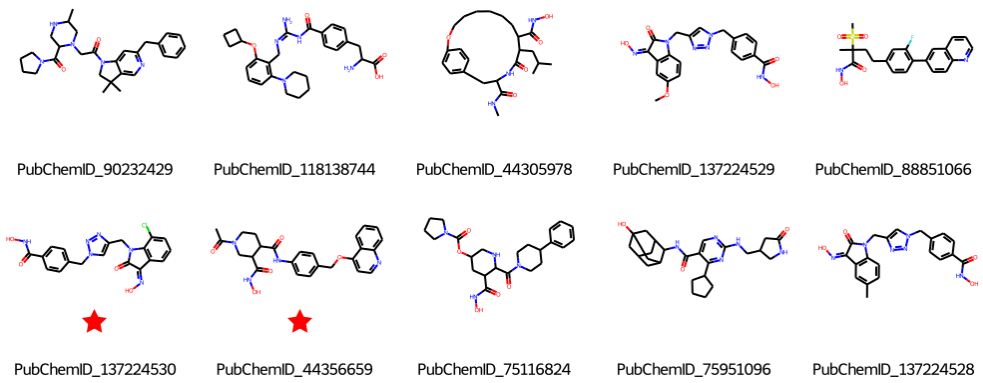

Figure 2: Common ligands found in 100 particles with the lowest vina affinity scores chosen for every protein, where two ligands marked with asterisks had the best scores.

Furthermore, after choosing only the 10 best ligands, based on their probability of low IC50 score and looking for ones that were present in all of the top choices for each Nsp-13 structure we identified two particles from PubChem: 44356659 and 137224530.

To check if our neural network can serve as a tool for screening particles we have compared the affinity scores of 500 ligands that were designated as best candidates and 100 ligands that were predicted to have a low probability of having an IC50 score below 10μM. Four out of five analyzed proteins showed a statistically significant difference between the results of docking ligands labeled as promising and those from the bottom of the list returned by our neural network. Only for Nsp-13 structure 5rm2 did we obtain the results which differences were not statistically significant.

References

[1] Vaswani, A. *et al.* Attention is all you need. *CoRR* **abs/1706.03762** (2017). URL `http://arxiv.org/abs/1706.03762`. `1706.03762`.

[2] Chithrananda, S., Grand, G. & Ramsundar, B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885* (2020).

[3] Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods* **16**, 1315–1322 (2019).

[4] Ma, E. J. & Kummer, A. Reimplementing unirep in jax. *bioRxiv* 2020–05 (2020).

[5] Gilson, M. K. *et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research* **44**, D1045–D1053 (2015).

[6] Burley, S. K. *et al.* RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Research* **51**, D488–D508 (2022).

[7] `http://bioinfo4eu.unimi.it/meet-eu-materials/`.

[8] Pettersen, E. F. *et al.* Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605–1612 (2004).

[9] Trott, O. & Olson, A. J. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **31**, 455–461 (2010).

[10] Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling* **61**, 3891–3898 (2021).

[11] Halder, U. C. Predicted antiviral drugs darunavir, amprenavir, rimantadine and saquinavir can potentially bind to neutralize sars-cov-2 conserved proteins. *Journal of Biological Research-Thessaloniki* **28**, 18 (2021).