Identification of Potential Inhibitors of the Nsp-13 Protein of the SARS-CoV-2 Virus

Warsaw Team 1 - Sorbonne Team 1

JAKUB JÓZEFOWICZ[1], JAKUB SKRAJNY[2], MICHAŁ KRUTUL[3],
ZOFIA KOCHAŃSKA[4], JULIA SMOLIK[5], MIESZKO SABO[6]

[1] j.jozefowicz2@student.uw.edu.pl, [2] j.skrajny@student.uw.edu.pl
[3] m.krutul@student.uw.edu.pl, [4] z.kochanska@student.uw.edu.pl
[5] j.smolik@student.uw.edu.pl , [6] m.sabo@student.uw.edu.pl

February 8, 2023

1. Introduction

1.1. Coronaviruses

Coronaviruses (CoVs) can be a major cause of acute intestinal and systemic infections. They are also known to attack the respiratory system in humans and other mammals [1]. Coronaviruses are enveloped, single-stranded RNA viruses that belong to the *Coronaviridae* family. SARS-CoV (Severe Acute Respiratory Syndrome Coronavirus) and MERS-CoV (Middle East Respiratory Syndrome CoV) are among the zoonotic pathogens that contributed to outbreaks in 2002 and 2012, respectively.

In December 2019, an outbreak of new pneumonia was identified in Wuhan, Hubei Province, China, which appeared to be caused by a new type of coronavirus, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [2]. The emergence of the new SARS-CoV-2 virus and its high infectivity, which caused the pandemic to spread rapidly around the world, prompted the development of global collaborations, building on the lessons learned from previous outbreaks of SARS-CoV and MERS-CoV [3]. SARS-CoV-2 has been shown to share similarities with SARS-CoV in genome structure, tissue tropism and viral pathogenesis. The genomic similarity and past experience in an earlier outbreak helped scientists to respond quickly to the new pandemic. SARS-CoV-2, however, appears to be more infectious. The two previous outbreaks have led to the identification of key targets, such as the virus protein *spike*. It promotes the attachment of virus particles to host cells via the angiotensin-converting enzyme 2 (ACE2) receptor, which is found on the surface of many eukaryotic cell types.

A new pneumonia has been named Coronavirus Disease 2019 (COVID-19) by the World Health Organization (WHO). According to the latest WHO report, there have been 664,873,023 cases worldwide to date, 6,724,248 of which have been fatal (as of January 25, 2023). Vaccines have made good progress in preventing the coronavirus disease (COVID-19) pandemic by significantly reducing the number of cases (particularly fatal ones) and hospitalization [4]. However, the emergence of numerous variants has brought significant challenges to human health. Therefore, developing effective therapeutics, such as identifying effective inhibitors and designing drugs against COVID-19, could help manage the pandemic more effectively.

1.2. SARS-CoV-2 Virus Genome

The SARS-CoV-2 genome consists of the first two open reading frames (ORFs): ORF1a and ORF1b [5, 6] (Fig. 1). ORF1a/b encodes 16 nonstructural proteins (Nsps). The remaining ORFs encode several structural proteins, such as viral envelope proteins and accessory proteins. Nsp-1 serves the virus to evade the host's immune system and inhibit its gene expression. It is a target protein for vaccine development. Nsp-2 is not well understood, but is known not to be used in viral replication. RNA-binding domains can be distinguished in the structure of Nsp-3. This unstructured protein interacts with Nsp-4 and other cofactors. Loss of the Nsp-3 complex with Nsp-4 eliminates viral replication. Other Nsps play different roles in the viral life cycle. For example, Nsp-12 in complex with Nsp-7 and Nsp-8 enables virus replication. Nsp-9 in complex with Nsp-8 is involved in RNA replication and

virulence, and the Nsp-10 complex with Nsp-16 is essential for capping viral mRNA transcripts. The non-structural protein Nsp-13 functions as a helicase, as it cleaves deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) in a nucleoside triphosphate (NTP)-dependent manner. Nsp-13 is highly conserved within coronavirus species. Genome sequencing has led to the identification of mutations in structural proteins, namely the spike protein, and non-structural proteins. Nsp-13 is an attractive target protein for inhibitors of this virus because the mutations present in it do not affect its interaction with some known compounds.

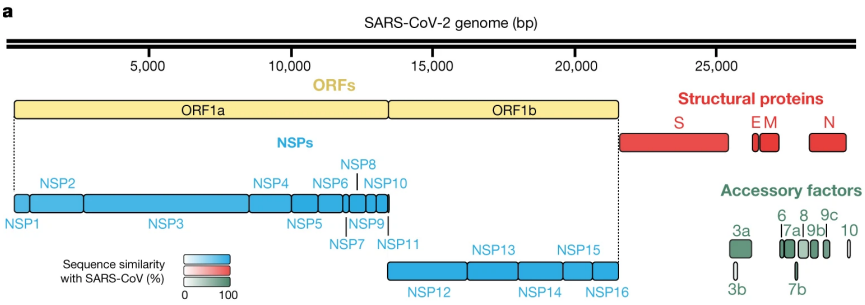


Figure 1: SARS-CoV-2 genome annotation [7]. The first two ORFs are marked in yellow. The intensity of blue (non-structural proteins), red (structural proteins) and green (accessory factors) is proportional to the protein sequence similarity with SARS-CoV homologs (if homologs exist). In the genome of the SARS-CoV-2 virus, 4 structural proteins, 16 non-structural proteins and 9 accessory factors can be distinguished.

1.3. Neural Networks

Developments in computational biology and other sciences have led to faster drug discovery and formulation [8]. Artificial intelligence (AI) is widely used around the world, including the academia. Machine learning (ML), one of AI's offshoots, is connected in many fields, including data generation and analytics. Machine learning algorithms build a mathematical model from sample data, called a training dataset, to make predictions or decisions. Deep learning (DL) is a subcategory of machine learning. The learning process is deep because the structure of artificial neural networks consists of multiple layers. Each layer consists of units that transform input data into information that subsequent layers can use to perform some predictive task. ML and DL have become attractive approaches to drug discovery.

A neural network (NN) is a method in artificial intelligence that enables computers to learn about data in a way that is inspired by the human brain. It is a type of machine learning process (called deep learning), the structure of which resembles the human brain. It is based on a collection of connected computational units called neurons. Each neuron is assigned a weight, called bias, and an activation function. Connections between neurons, also known as edges, as well as neurons, are assigned weights that adjust as learning progresses. Typically, neurons are aggregated into layers. Signals travel from the first layer (input layer) to the last layer (output layer). There are many types of networks that differ, for example, in architecture, the way signals are processed by neurons or the way input data is transferred to the network (all at once, in fragments, recursively, etc.).

1.4. Problem Statement

The aim of our work is to identify potential inhibitors of the Nsp-13 protein of the SARS-CoV-2 virus using an approach based on deep learning (neural networks). Inhibition of the viral protein may result in stopping the replication process and thus inhibit the multiplication of the virus in the attacked organism. Finding the right inhibitor, i.e. a molecule that would be able to fight the SARS-CoV-2 virus, could slow down the development of the global pandemic and even potentially end it.

2. Materials & Methods

2.1. Workflow

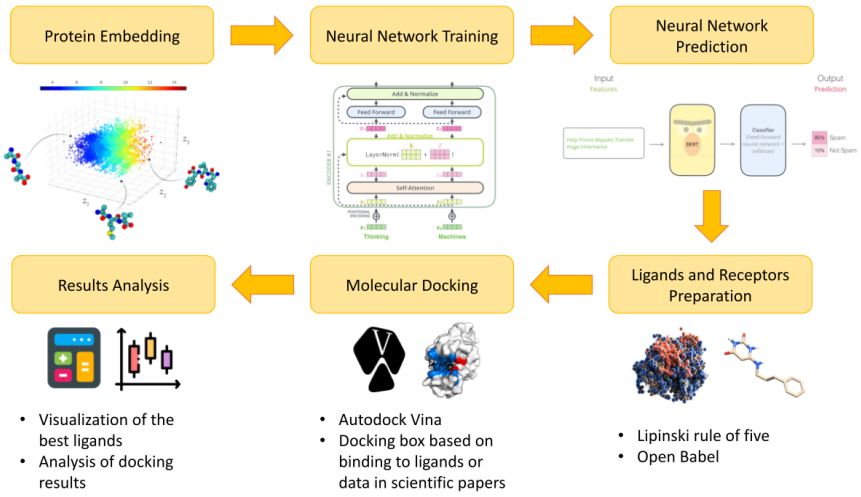


Figure 2: Diagram of the analyses described in detail in the paper

2.2. Training Dataset: Selection of Receptors and Ligands

In an approach based on machine learning, an important element is to provide a large, varied training dataset, which is why in this approach it was crucial to provide information about "good" ligands and "bad" ligands. Inhibitors can be divided into 3 groups depending on the IC50 value: high (IC50 $<$ 1 μM), moderate (1 μM $<$ IC50 $<$ 10 μM), and low (IC50 $>$ 10 μM) potential for drug-drug interaction [9]. In our approach, we considered "good" ligands as those with IC50 values below 10 μM. Many particles can bind to one protein. The binding of one compound could change the conformation of the entire protein, allowing for the binding of another molecule, which could subsequently prove to be a potential inhibitor. For the training dataset, we decided to use only one-to-one protein-ligand pairs.

The training set consisted of ∼500,000 protein-ligand pairs. The SMILES codes of the ligands were retrieved from the BindingDB database [10]. BindingDB is a public, web-based database of measured binding affinities focusing primarily on interactions of proteins considered to be potential pharmacological targets for ligands, which are small, drug-like molecules. It currently contains over 2,500,000 experimentally determined binding affinities of protein-ligand complexes for nearly 9,000 protein targets, including isoforms and mutation variants, and over 1,000,000 small molecule ligands.

The amino acid sequences of the proteins to which the selected ligands were bound were retrieved from the PDB database [11]. The PDB database contains, among other things, data on the 3D structure of large biological molecules (proteins, DNA and RNA). Knowledge of the 3D structure of a biological macromolecule is essential to the understanding of its function.

2.3. Neural Network Training

The input for training the neural network was a set of protein-ligand pairs. A protein was represented by its amino acid sequence, and the ligand was represented in SMILES format. The output is a probability of a ligand-protein pair having an IC50 value $< 10\mu M$. We divided our architecture into 3 parts:

- Ligand encoder
- Protein encoder
- Binding predictor

Ligand Encoder Transformer networks have repeatedly proven effective at text summarisation tasks in multiple contexts [12]. This quality of the network architecture, combined with the emergent GNN-like behaviour of the attention mechanism over SMILES strings, makes the BERT family of models a convincing choice for this part of the architecture.

Because of the prohibitive computational cost of pretraining large transformer-based encoders, we decided to use a pretrained model from the HuggingFace library [13]. ChemBERTa [14], a RoBERTa model trained on the ZINC database [15], was chosen for this task.

The network was trained alongside the binding attention mechanism, with the SMILES strings tokenized by a BPE encoder trained for SMILES strings.

Protein Encoder We initially envisioned the model to include a similar attention-based encoder for the protein sequences, albeit with the use of a transformer adapted for longer sequences [16]. The most appealing choice from a research standpoint would have been the BigBird model [17], the use of which has not yet been explored in the context of protein sequences. However, similarly to the binding predictor, we weren't able to secure the computational resources to train a model from scratch and so decided on the UniRep [18, 19] model for embedding. The protein embeddings were precomputed for use with the rest of the network.

Binding Predictor The binding predictor architecture is created on the basis of Transformer [20]. Transformer was originally designed to solve the Natural Language Processing problems, but it was so successful and innovative that researchers started to apply this network to different types of problems. Before Transformer was introduced, Recurrent Neural Networks (RNNs) [21] were used to deal with sequential data. One of the reasons why Transformer usually works better than RNN is attention mechanism that is able to catch both local and global dependencies, while RNNs struggles with understanding global ones.

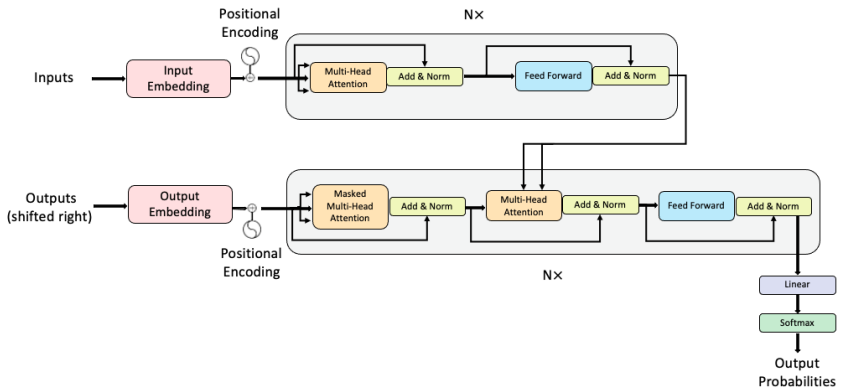


Figure 3: Transformer

Binding predictor's inputs are outputs of Ligand encoder and Protein encoder - one dimensional sequences. In the beginning they are computed independently.

1. [linear -> positional encoder -> linear -> relu -> stack]

First Linear increases number of dimensions. It is necessary for positional encoder to properly encode the position of an element in a sequence. Once information about the position in a sequence is encoded within the sequence, a second Linear is applied to decrease length, which is necessary due to the limitations of our machine. Activation functions like relu are used to enable to model non-linear dependencies. After these computations both elements are stacked together.

2. [norm -> multi-head-attention -> norm -> feed forward] x N

The second part consists of N sublayers that have the same architecture. Normalization Layers are used to stabilize the training. Feed forward is a combination of linear, relu and linear layers. The most important element is a multi-head-attention. This is where most of the dependencies are recognized.

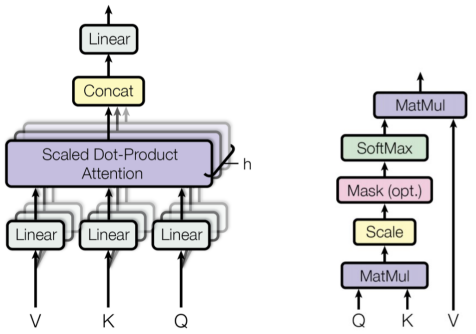


Figure 4: Multi-Head-Attention and Scaled Dot-Product Attention

$$\text{Attention(Q, K, V)} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$
$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\text{head}_1, \ldots, \text{head}_h]\,\mathbf{W}_0$$
$$\text{where} \quad \text{head}_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right)$$

3. [linear, relu, linear, relu, linear]

The last part consists of linear and relu layers. Model predicts which compound-protein pairs can bind with each other. During training model's predictions are compared to ground truth, and based on that model's weights are changing.

2.4. Prediction: Identification of Potential Nsp-13 Protein Inhibitors

The trained network, ligands used for training, 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) [22] were used to identify potential inhibitors of the Nsp-13 protein.

2.5. Ligand and Receptor Preparation

Ligands that were predicted as potential inhibitors of the Nsp-13 protein were filtered out. 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. It checks such characteristics of molecules as their molecular weight, octanol-water partition coefficient and the number of hydrogen bond acceptors and donors. It is assumed that a molecule should not have more than 5 hydrogen bond donors and 10 hydrogen bond acceptors, the molecular weight should not exceed 500 Da, and the octanol-water partition coefficient is expected not to exceed the value of 5. Originally, the rule defined by Christopher Lipinski assumes that to further consider a molecule as a potential oral drug, a maximum of one of the 4 rules can be broken. During our analyses, we decided to apply stricter criteria and assumed that each of the described rules must be met for a molecule to be considered in subsequent analysis steps.

Molecules that were selected as potential inhibitors in the previous step were then used to build their 3D structure using Open Babel software. This is a chemical toolkit designed for a wide range of chemical data. It is an open, collaborative project that allows anyone to search, convert, analyze and store data from molecular modeling, chemistry, semiconductor materials, biochemistry and other related fields. The ligand structure was generated using the following command:

```
obabel ligand.smi –O ligand.pdbqt —gen3d
```

where

```
ligand.smi  - ligand in SMILES format
ligand.pdbqt  - the resulting 3D structure of the ligand
```

All structures of the selected receptors (Nsp-13 protein structures) contained several protein chains. All chains 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 [23], and then using Open Babel converted to pdbqt format with the command:

```
obabel receptor.mol2 –O receptor.pdbqt –xc –xr
```

where

```
receptor.mol2  - receptor in mol2 format
receptor.pdbqt  - the resulting structure of the receptor
```

After obtaining the 3D structures of the ligands and the edited receptor structures, we proceeded to molecular docking.

2.6. Molecular Docking

Molecular docking is a computational procedure that is used to efficiently predict the non-covalent binding of a macromolecule (receptor) and a small molecule (ligand), starting from their unbound structures [24]. Docking is used to predict the conformation of bound molecules and their binding affinities. The prediction of ligand binding to proteins is particularly important as it is used in practice to screen virtual libraries of molecules with drug-like properties in order to identify potential inhibitors.

One of the elements of docking is the reproduction of chemical potentials that determine bound conformation preferences and free energy of the binding. Docking programs typically use a scoring function, which can be seen as an attempt to approximate the standard chemical potentials of a system.

Compounds that the neural network classified as potential inhibitors of the Nsp-13 protein were docked in the AutoDock Vina (Vina) program [24, 25]. It is a molecular docking and virtual screening program. AutoDock Vina automatically calculates the map grid and clusters the results transparently to the user.

Docking for each protein structure was performed separately in the Vina program using the following command:

```
vina --receptor receptor.pdbqt --ligand ligand.pdbqt
--size_x X --size_y Y --size_z Z --center_x CX --center_y CY
--center_z CZ --seed 123 --log ligand.txt --out ligand_out.pdbqt
```

where

```
receptor.pdbqt  - receptor structure in pdbqt format
ligand.pdbqt  - structure of individual ligands in pdbqt format
size  - docking box size
center  - coordinates of the center of the docking box
ligand.txt  - file into which the logs of the program's operation were saved
ligand_out.pdbqt  docking output file
```

Docking of the selected ligands was carried out for all of the selected structures of the Nsp-13 protein. For all of 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 [26]. When determining these values, it is crucial to keep the box space as small as possible, allowing for a more exhaustive search of the space when trying to find the most optimal site for ligand attachment to the protein. The use of too large dimensions or incorrect determination of the position of the docking box results in unsatisfactory docking results, including, for example, binding of the ligand to the protein in the wrong place.

Each docking result was described by a score value (affinity). The smaller the value, the better the docking result. The general form of the conformation-dependent part of the scoring function that Vina uses is:

$$c = \sum_{i<j} f_{t_i t_j}(r_{ij}), \tag{1}$$

where the sum is for all pairs of atoms that can move relative to each other, usually excluding interactions 1-4, i.e. atoms separated by three consecutive covalent bonds. Here, each atom i is assigned the type t_i, and a symmetric set of interaction functions $f_{t_i t_j}$ of the interatomic distance r_{ij} should be defined.

3. Results

3.1. Neural Network Training

The neural network was trained with the full BindingDB dataset of known IC50 values ($n \approx 500,000$). The network was tasked with predicting whether a given protein-ligand pair had an IC50 $< 10\ \mu M$. This value was chosen first and foremost because, below this threshold, drugs are typically in low enough concentrations to be able to be admitted orally, without disrupting enzyme function. This makes the given candidate qualifiable for downstream analysis in programs such as NIH's NCI60. Additionally, the value splits the dataset in approximately half, which made the dataset unbiased between the two classes of predictions, maximizing the utility of the dataset for training. The network was trained for 40 epochs, on a cluster equivalent to about 150 hours on a GTX2080ti. Further training would have allowed for only marginally better prediction results. In the end the prediction resulted in an accuracy of 86%, precision of 95% and recall of 88%. The parameters used during training can be found in the project's repository, along with additional guidance for adjusting parameters for different training scenarios.

3.2. Neural Network Prediction

The network was then tasked with predicting which substances fall below the IC50 threshold when given the Nsp-13 encoding in combination with every substance in the BindingDB dataset, this constituted an additional 2 hours of computation. The results of the prediction were sorted with regard to the probability of the required concentration falling below the threshold and passed on to be further filtered based on other criteria before molecular docking.

3.3. Ligand and Receptor Preparation

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 prediction. We selected the first 500 ligands (most likely to be a good inhibitor based on the IC50 value) that additionally met all of the requirements of the Lipinski rule 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.

3.4. Molecular Docking

Each docking resulted in 9 structures of a given ligand docked to the receptor. Every model was described using the scoring function (binding affinity). Docked particles should have a low score, which takes negative values. For each docked ligand, we selected its best conformation and best binding affinity value. On this basis, 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. 5.

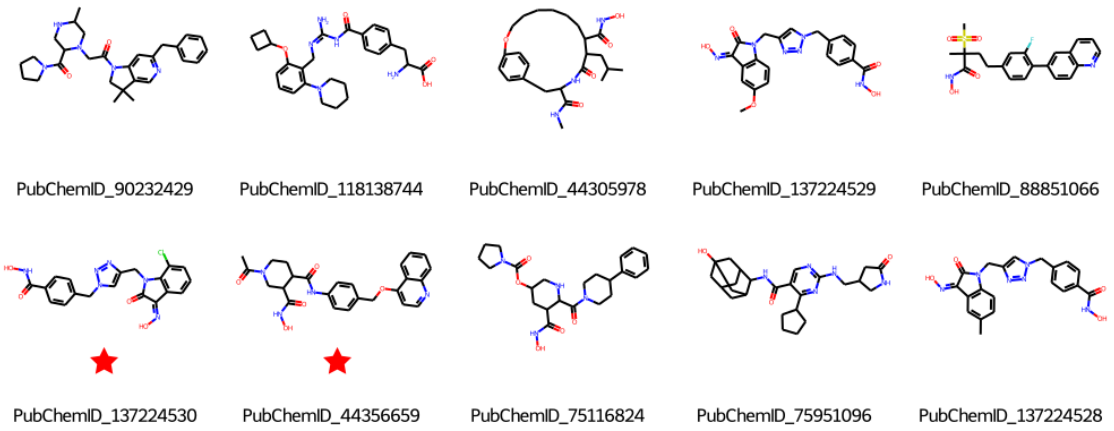
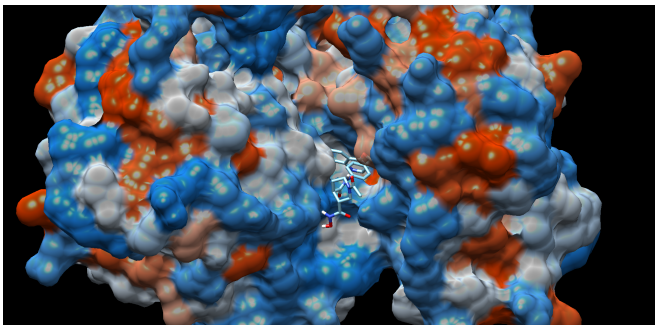
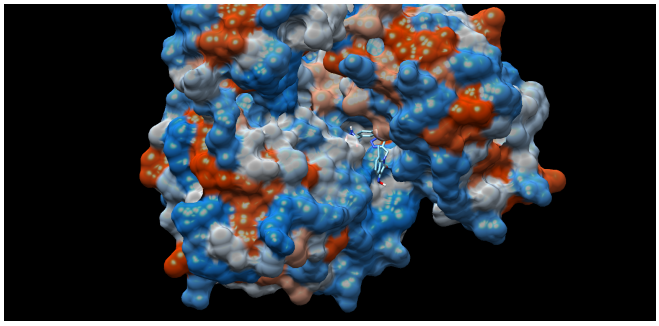


Figure 5: Common ligands found in 100 particles with the lowest vina affinity scores chosen for every protein, where two ligands marked with asterisks had the best scores.

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 receptors we identified two particles from PubChem: 44356659 and 137224530. The best docking conformation of those ligands for each of the Nsp-13 proteins is shown in Fig 6-10.

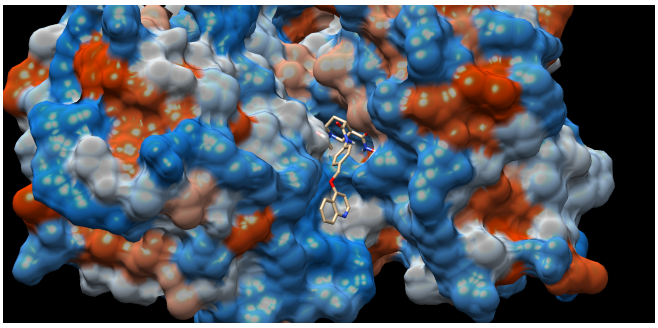


(a) 5rm2-44356659 complex, affinity= -8.8 kcal/mol

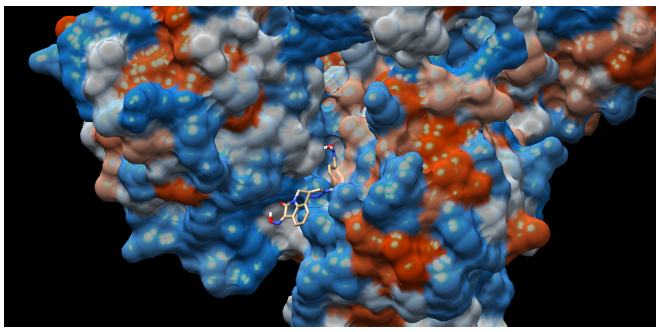


(b) 5rm2-126842746 complex, affinity= -8.9 kcal/mol

Figure 6: Docking results for 5rm2 protein and two best ligands

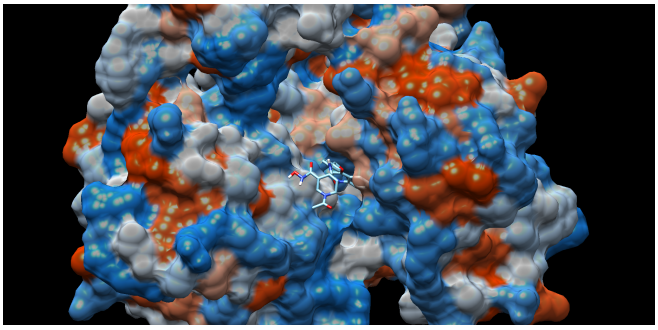


(a) 5rme-44356659 complex, affinity= -9.0 kcal/mol

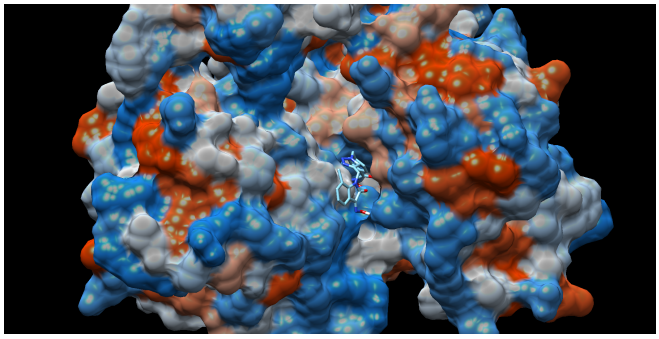


(b) 5rme-126842746 complex, affinity= -8.3 kcal/mol

Figure 7: Docking results for 5rme protein and two best ligands

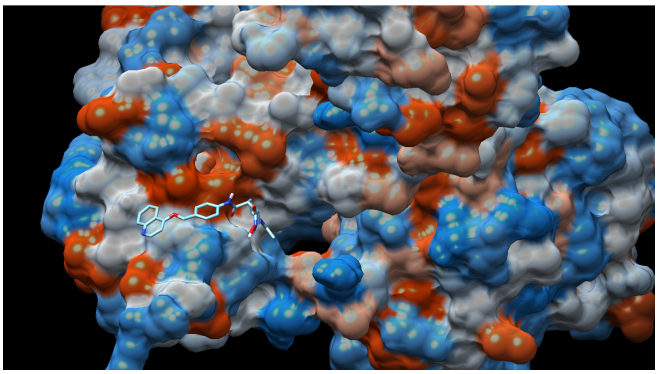


(a) 6zsl-44356659 complex, affinity= -8.0 kcal/mol

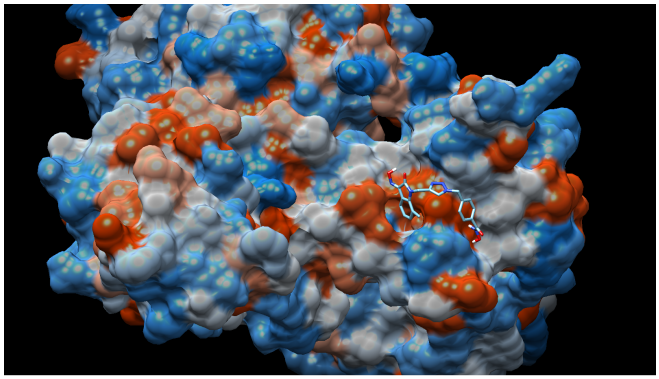


(b) 6zsl-126842746 complex, affinity= -8.1 kcal/mol

Figure 8: Docking results for 6zsl protein and two best ligands

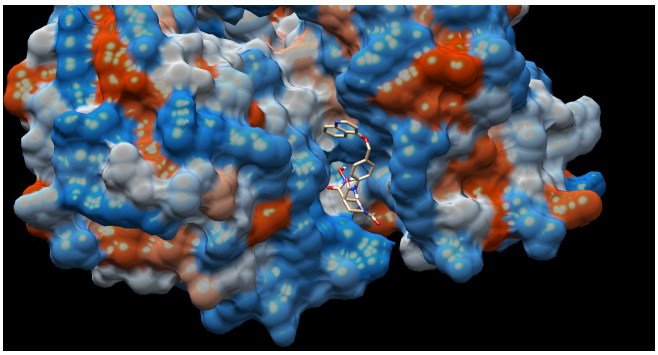


(a) 7nio-44356659 complex, affinity= -6.8 kcal/mol

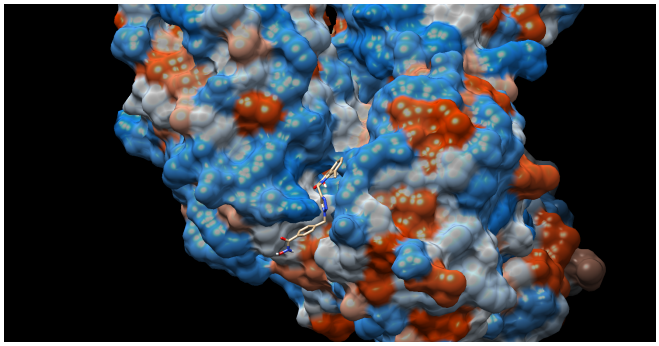


(b) 7nio-126842746 complex, affinity= -7.0 kcal/mol

Figure 9: Docking results for 7nio protein and two best ligands



(a) 7nn0-44356659 complex, affinity= -9.0 kcal/mol



(b) 7nn0-126842746 complex, affinity= -8.9 kcal/mol

Figure 10: Docking results for 7nn0 protein and two best ligands

To check if our neural network can serve as a tool for screening many 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 a low IC50 score. Results of those comparisons are shown in Fig 11, which also contains information about the statistical significance of those results. The samples which were labeled as being statistically different from each other are marked with an asterisk. Four out of five analyzed proteins showed a noticeable 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 protein 5rm2 did we obtain the results which differences were not statistically significant.

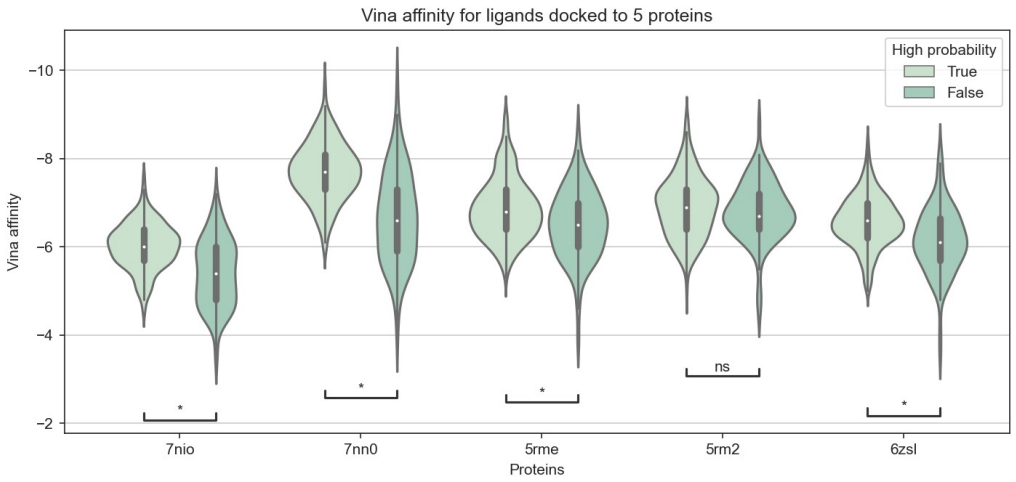


Figure 11: Violinplot which shows the differences between vina affinities for ligands marked as having a high probability of having a low IC50 value and those which have a high probability returned by our neural network. Asterisks indicate groups in which the differences were statistically significant.

4. Discussion

In the current situation in the world, finding an effective inhibitor of the Nsp-13 protein of the SARS-CoV-2 virus is the goal of many scientists.

In the above considerations, we presented steps that can be taken in the process of further analysis and refinement of the search for an inhibitor of the Nsp-13 protein of the SARS-CoV-2 virus. In our work, we were able to identify and present 10 compounds that are potentially good inhibitors of the Nsp-13 protein and which could help start more thorough research into a cure for COVID-19. The results obtained by us can be used as an introduction to further analyzes and clinical trials. We have also marked two particles that we believe to be the best out of our 10 presented ligands. Those structures are not the perfect cure but docking results obtained for them lead us to believe that it might be beneficial to investigate them further.

Our neural network proves to serve its purpose and can be used for the initial screening of the ligands. For four out of five proteins results obtained for docking of the molecules which were identified as promising were statistically different than those which came from docking of the low-probability particles. Of course, there is always room for improvement and we recognize that our network still needs some adjustments to serve as a fully functional tool. Nevertheless, we find that the usage of AI in the task of looking for inhibitors of proteins can be advantageous. If the architecture of the neural network is correctly planned and then trained using accurate descriptors, it may significantly shorten the time needed to identify inhibitors by limiting the number of potential ligands.

References

[1] Yaghoubi, A., Jamehdar, S. A., Movaqar, A., Milani, N. & Soleimanpour, S. An effective drug against covid-19: reality or dream? *Expert Review of Respiratory Medicine* **15**, 505–518 (2021).

[2] Qiu, R. *et al.* The therapeutic effect and safety of the drugs for covid-19: A systematic review and meta-analysis. *Medicine* **100** (2021).

[3] Harrison, A. G., Lin, T. & Wang, P. Mechanisms of sars-cov-2 transmission and pathogenesis. *Trends in Immunology* **41**, 1100–1115 (2020).

[4] Wang, J., Zhang, Y., Nie, W., Luo, Y. & Deng, L. Computational anti-COVID-19 drug design: progress and challenges. *Briefings in Bioinformatics* **23** (2021).

[5] Pitsillou, E., Liang, J., Hung, A. & Karagiannis, T. C. The SARS-CoV-2 helicase as a target for antiviral therapy: Identification of potential small molecule inhibitors by in silico modelling. *J Mol Graph Model* **114**, 108193 (2022).

[6] Raj, R. Analysis of non-structural proteins, nsps of sars-cov-2 as targets for computational drug designing. *Biochemistry and Biophysics Reports* **25**, 100847 (2021).

[7] Gordon, D. E. *et al.* A sars-cov-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020).

[8] Patel, L., Shukla, T., Huang, X., Ussery, D. W. & Wang, S. Machine learning methods in drug discovery. *Molecules* **25** (2020).

[9] Krippendorff, B. F., Lienau, P., Reichel, A. & Huisinga, W. Optimizing classification of drug-drug interaction potential for CYP450 isoenzyme inhibition assays in early drug discovery. *J Biomol Screen* **12**, 92–99 (2007).

[10] 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).

[11] 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).

[12] Miller, D. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165* (2019).

[13] Huggingface transformers library. `https://huggingface.co/`.

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

[15] Irwin, J. J. & Shoichet, B. K. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling* **45**, 177–182 (2005). `https://zinc.docking.org/`.

[16] Tay, Y. *et al.* Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006* (2020).

[17] Zaheer, M. *et al.* Big bird: Transformers for longer sequences. *Advances in neural information processing systems* **33**, 17283–17297 (2020).

[18] 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).

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

[20] Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

[21] Schmidt, R. M. Recurrent neural networks (rnns): A gentle introduction and overview. *CoRR* **abs/1912.05911** (2019). URL `http://arxiv.org/abs/1912.05911. 1912.05911`.

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

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

[24] 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).

[25] 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).

[26] 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).