Identifying Potential Treatments for COVID-19 with a Deep Learning Model

Alperen Bağ[1], Berk Atıl[1], Rıza Özçelik[1], Elif Ozkirimli[2], Arzucan Özgür[1]
1 Department. of Computer Engineering, Boğaziçi University, Istanbul, Turkey.
2 Data and Analytics Chapter, Pharma International Informatics, F. Hoffmann-La Roche AG, Switzerland

Main result

Tezacaftor and **Lurasidone** are found to be good candidates related to Covid-19 thanks to high interaction possibility with **ACE2** and **GAK** respectively.

Introduction

SARS-CoV-2 infected millions of people around the world and had severe implications. However, an effective treatment is yet to be found. To accelerate drug discovery and to identify lead molecules, drug repurposing is an attractive approach. In this study, we trained the DeepDTA model(1), a state-of-the-art deep drug-target affinity prediction model that represents ligands by their Simplified Molecular Input Line Entry System (SMILES) strings and proteins by their amino-acid sequences, to virtually scan the existing drugs to suggest drug candidates for the treatment of coronavirus disease (COVID-19).

Data

- We used the BindingDB(4) data set by retrieving only interactions with either pK_i or pK_d values. We classified the interactions as 1 (binds) or 0 (not binds) based on thresholds 7.6 for pK_i and 7.1 for pKd.
- We splitted the data into 1 train set, 2 validation sets and 2 test sets. The first validation/test set is a set with similar number of positive and negative examples for each ligand and the other one is a random set. The reason for this is to prevent the fake success when the model always predicts either 1 or 0 for some ligands and these ligands have only either positive or negative interactions in the validation and test set.

Set Name	# of Examples	# of Positive	# of Negative	# of Ligands	# of Proteins
Train	259240	72603	186637	156474	3815
Balanced Test	28442	14629	13813	12176	1668
Random Test	61008	18905	42103	48802	2784
Balanced Validation	27000	13922	13078	11599	1617
Random Validation	61008	18670	42338	48898	2790

Target Proteins

As targets, we chose 13 human proteins associated with Covid19 based on (5)(6). Our targets are as follows; ACE2, RBC Band3, JUN, NPM1, XPO1, ROA1, TMPS2, Furin, Cathepsin L, AAK1, GAK, TPC2, and FYV1.

Model

We used DeepDTA(1), a state-of-the-art deep learning model that predicts binding affinity between a ligand and a protein. It takes a FASTA sequence of a protein and a SMILES representation of a ligand.

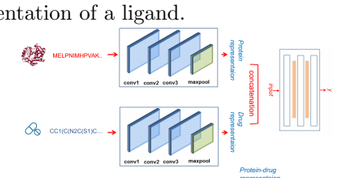


Figure 1: DeepDTA architecture, taken from the DeepDTA(1)

Bias Problem

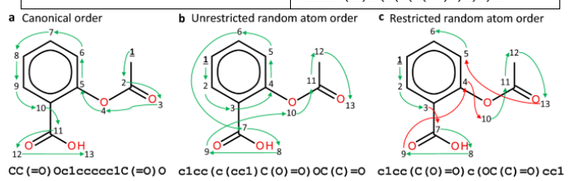
- Upon detailed analysis of results, we found that the model is likely to give low affinity scores, which is mainly caused by imbalanced training data, since it has much more negative interactions than the positive ones.
- In order to solve this problem, we tested **2** different approaches which are described in the following section.

Data Balancing Methods

We trained 2 models with the methods below, then both models were used together as ensemble for inference.

- Firstly, we tried to utilize the class weights technique to reduce the effect of imbalanced data. This technique increases the weight of the minority class in the loss function, thereby giving the out-numbered class higher importance, which is positive class (1) in our case. That is, when the model makes a wrong prediction for a positive class (1), it is penalized more.
- The other approach that we used is making the training set more balanced. To do this, we used SMILES enumeration as an offline augmentation. We make use of the fact that SMILES are not unique. The difference stems from the starting position or the visiting order of the atoms. We increased the number of positive examples by taking equivalent SMILES representations and input them into the model as different interactions. Below, some enumerated examples are shown(2).

Canonical SMILES	Enumerated SMILES
C[C@](N)(Cc1ccc(O)cc1)C(=O)O	C(O)(=O)[C@@](C)(N)Cc1ccc(O)cc1, c1(C[C@](N)(C(=O)O)O)ccc(O)cc1
CC(c1cc2ccccc2s1)N(O)C(N)=O	c1c2sc(C(C)N(O)C(=O)N)cc2ccc1, c1ccc2c(c1)sc(C(N(C(=O)N)O)C)c2



Training

- We trained the models using the Adam optimizer with a learning rate of 0.001. The batch size was 32 and binary cross-entropy loss function was utilized.
- Different than the original DeepDTA paper(1), we applied online augmentation to the SMILES sequences to reduce the over-fitting and ligand-specific bias. We used SMILES enumeration for this purpose, however this time it is used as an online augmentation technique during the training instead of balancing the train data before the training. In other words, each SMILES in a batch was changed to another equivalent SMILES representation that also represents the same molecule before being processed by the model. While this technique does not affect the general performance of the model, it increased the precision when the classification threshold is 0.9, which is important for confident proposals. Specifically, the precision increased from **0.7117** to **0.7576** in the balanced validation set. Furthermore, when SMILES enumeration used, the precision consistently improves with an increasing classification threshold.

Comparison

When we compared the validation performances of the 3 models, the AUC of ROC is better especially with the model with balanced train set. Since we would prefer confident predictions, we decided to use both of them by taking their common proposals. Their separate test performances above threshold 0.9 are shown below.

Threshold	Precision	Recall	Total Positive Predictions	TP
0.9	0.8190	0.1154	2061	1688

(a) Model with Balanced Train

Threshold	Precision	Recall	Total Positive Predictions	TP
0.9	0.7169	0.0507	1035	742

(b) Model with Class Weight

Table 1: Balanced Test Set

Threshold	Precision	Recall	Total Positive Predictions	TP
0.9	0.8762	0.2221	4791	4198

(a) Model with Balanced Train

Threshold	Precision	Recall	Total Positive Predictions	TP
0.9	0.8803	0.1485	3190	2808

(b) Model with Class Weight

Table 2: Random Test Set

Potential Compounds

We trained 2 models with the methods below, then both models were used together as ensemble for inference. Since our main goal is detecting drug-protein pairs with high binding affinity related to Covid-19, we made an inference with target proteins and **2255** approved drugs retrieved from ChEMBL(3). We identified two potential interactions using our ensemble model. The first one is **Tezacaftor** for Angiotensin-converting enzyme 2, ACE2. Tezacaftor is a drug of the cystic fibrosis transmembrane conductance regulator (CFTR) potentiator class(10). The other one is **Lurasidone** for Cyclin G-associated kinase, GAK. Lurasidone is used in the treatment of schizophrenia(8). We used VAPUR(11) and checked to see the relevance of these compounds in COVID-19 literature. Lurasidone was also proposed by docking calculations to be potential(9)

References

[1] Öztürk, Hakime and Özgür, Arzucan and Ozkirimli, Elif *Deepdta: deep drug–target binding affinity prediction*. Bioinformatics, 34(17):i821–i829, 2018
[2] Arús-Pous, J., Johansson, S.V., Prykhodko, O. et al. *Randomized SMILES strings improve the quality of molecular generative models*. J Cheminform 11, 71. https://doi.org/10.1186/s13321-019-0393-0,2019
[3] Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, Bellis L and Overington JP *ChEMBL web services: streamlining access to drug discovery data and utilities*. Nucleic Acids Res 43, W612–W620, 2015
[4] Chen,X., Liu,M. and Gilson,M.K. *BindingDB: A Web-accessible molecular recognition database*. Comb. Chem. HighThroughput Screen,4, 719–725, 2002
[5] Cosic, I., Cosic, D., Loncarevic, I. *RRM prediction of erythrocyte band3 protein as alternative receptor for SARS-CoV-2virus*. Applied Sciences (Switzerland), 10(11). https://doi.org/10.3390/app10114053, 2020
[6] Gil, C., Ginex, T., Maestro, I., Nozal, V., Barrado-Gil, L., Cuesta-Geijo, M. Á., Urquiza, J., Ramírez, D., Alonso, C., Campillo, N. E., Martinez, A. *COVID-19: Drug Targets and Potential Treatments*. Journal of Medicinal Chemistry.https://doi.org/10.1021/acs.jmedchem.0c00606, 2020
[7] Wang, X., Guan, Y. *COVID-19 drug repurposing: A review of computational screening methods, clinical trials, and protein interaction assays*. Medicinal Research Reviews, July, 1–24. https://doi.org/10.1002/med.21728, 2020
[8] Ishibashi, T., Horisawa, T., Tokuda, K., Ishiyama, T., Ogasa, M., Tagashira, R., Matsumoto, K., Nishikawa, H., Ueda, Y., Toma, S., Oki, H., Tanno, N., Saji, I., Ito, A., Ohno, Y.,, Nakamura, M. *Pharmacological Profile of Lurasidone, a Novel Antipsychotic Agent with Potent 5-Hydroxytryptamine 7 (5-HT7) and 5-HT1A Receptor Activity*. Journal of Pharmacology and Experimental Therapeutics, 334(1), 171 LP – 181. https://doi.org/10.1124/jpet.110.167346,2010
[9] Ammar D. Elmezayen, Anas Al-Obaidi, Alp Tegin Şahin, Kemal Yelekçi *Drug repurposing for coronavirus (COVID-19): in silico screening of known drugs against coronavirus 3CL hydrolase and protease enzymes*. Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2020.1758791,2020
[10] Hoy, S.M. *Elezacaftor/Ivacaftor/Tezacaftor: First Approval*. Drugs 79, 2001–2007. https://doi.org/10.1007/s40265-019-01233-7,2019
[11] Köksal, A., Dönmez, H.,Öz "celik, R, Ozkirimli, E., Özgür, A.¨, A., *Vapur: A Search Engine to Find Related Protein - Compound Pairs in COVID-19 Literature*. BioRxiv, 2020.09.05.284224. https://doi.org/10.1101/2020.09.05.284224,2020