Course Project:
Deep Learning to Predict COVID19 Antiviral Drugs

SYSC 4906 - ASSIGNMENT 2

Kevin Dick
Systems & Computer Engineering
Carleton University
Otttawa, Canada
kevin.dick@carleton.ca

James R. Green
Systems & Computer Engineering
Carleton University
Otttawa, Canada
jrgreen@sce.carleton.ca

November 10, 2020

Abstract

You (and your ***partner****) are the lead machine learning researcher(s) among an interdisciplinary team of scientists and virologists during the COVID-19 pandemic. Your team has decided to dedicate their time and resources to identifying candidate anti-viral drugs that may reduce severity of COVID-19 symptoms and the viruses' mortality rate. Unfortunately, your team only has the resources to test a handful of candidate anti-viral drugs among the millions that exist. They turn to you and your team of machine learning engineers and data scientists to generate a deep learning model that can predict which of the millions of drugs is most likely to cure COVID-19 and provide them with a short list of candidate drugs to test in the lab. Time is ticking! Lives are at stake! Can you help bring an end to the pandemic?!*

Assignments 2 (A2) and 3 (A3) will be focused on developing deep machine learning models for the task of **Drug-Target Interaction** (DTI) prediction. A2&A3 can be completed *independently* or in *teams of two (2)*. A2 has two deliverables: a **Notebook with questions** related to the DeepPurpose toolkit that will be used for A3, and a **written proposal** for the machine learning solution you plan to implement for A3 (note: the model and architecture proposed does not need to perfectly align with the model that will ultimately be implemented). The DeepPurpose framework makes easily accessible feature representations for both protein sequences and drug sequences.

In A3, the final model you develop will be compared to the models developed by your peers and others in the scientific literature on an independent dataset in a **mini-competition**. Specifications of A3 will follow.

As part of a **project reach goal**, the most performant models *may* be combined into an *ensemble predictor* and used to generate predictions for the **49,437** drugs listed in the CAS COVID-19 Antiviral Candidate Compounds Dataset and results *may* be prepared as part of an academic research paper.

***K*eywords** Drug-Target Interaction Prediction · COVID19 · Deep Learning

A2 Due Date: Friday, November 20th, 2020 by 23:59 EST

Project Overview

Drug target interaction (DTI) prediction is critical to drug discovery, which is costly and time-consuming given the need of experimentally search for candidate drugs over a tremendously large drug-compound space. To develop a new drug takes **more than 10 years** and costs more than **$2.6 billion**. One strategy is to determine whether an existing drug can be effectively used to treat other diseases, known as *drug repurposing*. However, experimentally determining whether a large number of candidate drugs will *bind* to the intended target protein is also prohibitively expensive and time-consuming. Consequently, in recent years, there has been increased interest and promising progress for deep learning in DTI predictions.

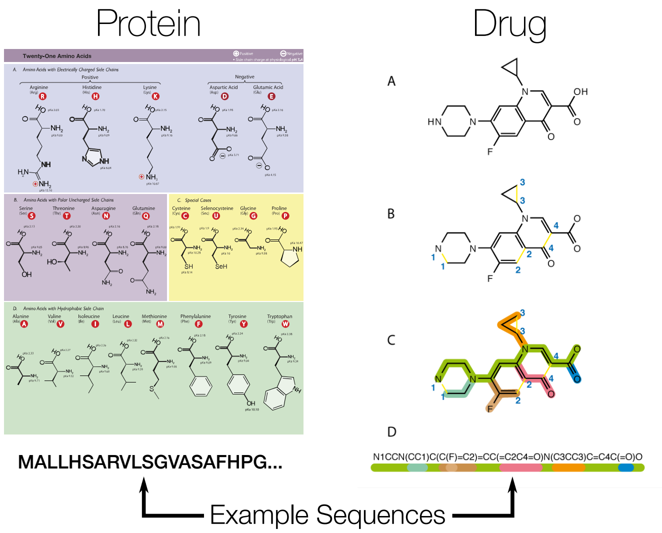

Figure 1: Sequence-Based Alphabets. Partially adapted from: Dancojocari, CC BY-SA 3.0

A majority of diseases can usually be attributed to a set of *target proteins* in a disease pathway or foreign proteins from pathogens or viruses. Typically, a drug (a small molecule) is discovered to modulate one or multiple of these target proteins. Through this modulation of the disease-causing protein, the symptoms of a disease can be greatly reduced, or the the disease altogether cured.

One of the major paradigms of the drug action mechanism is the 'Lock-And-Key' theory [1]: we conceptualise a protein as a "lock" and the discovered drug is a correctly fitting "key" (*i.e.* the right drug to modulate the protein). The "fitness" of a given key is known as it's *binding affinity*.

Binding affinity is typically measured and reported by the *equilibrium dissociation constant*, K_d, which is used to evaluate and sort into rank-order the various interaction strength between a drug and it's target: the smaller the K_d value, the greater the binding affinity of the drug for its target. For the purposes of this assignment, the K_d is a float value suitable for a regression-type model.

Both a protein and drug (small molecule) can be represented as **sequences** from two different alphabets (Fig. 1). A protein is represented as a sequence of *amino acids* while a small molecule can be represented as a **s**implified **m**olecular-**i**nput **l**ine-**e**ntry system (SMILE) as seen in Fig. 2.

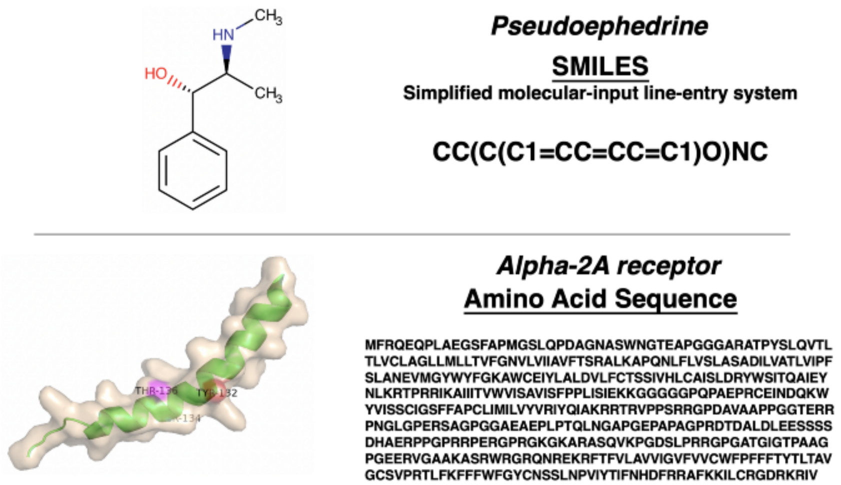
CH3
HN
HO
CH3
Pseudoephedrine
SMILES
Simplified molecular-input line-entry system
CC(C(C1=CC=CC=C1)O)NC
THR-136
TYR-132
-134
Alpha-2A receptor
Amino Acid Sequence
MFRQEQPLAEGSFAPMGSLQPDAGNASWNGTEAPGGGARATPYSLQVTL
TLVCLAGLLMLLTVFGNVLVIIAVFTSRALKAPQNLFLVSLASADILVATLVIPF
SLANEVMGYWYFGKAWCEIYLALDVLFCTSSIVHLCAISLDRYWSITQAIEY
NLKRTPRRIKAIIITVWVISAVISFPPLISIEKKGGGGGPQPAEPRCEINDQKW
YVISSCIGSFFAPCLIMILVYVRIYQIAKRRTRVPPSRRGPDAVAAPPGGTERR
PNGLGPERSAGPGGAEAEPLPTQLNGAPGEPAPAGPRDTDALDLEESSSS
DHAERPPGPRRPERGPRGKGKARASQVKPGDSLPRRGPGATGIGTPAAG
PGEERVGAAKASRWRGRQNREKRFTFVLAVVIGVFVVCWFPFFFTYTLTAV
GCSVPRTLFKFFFWFGYCNSSLNPVIYTIFNHDFRRAFKKILCRGDRKRIV

Figure 2: Example Sequence-Based Representation of a Molecule and Protein.

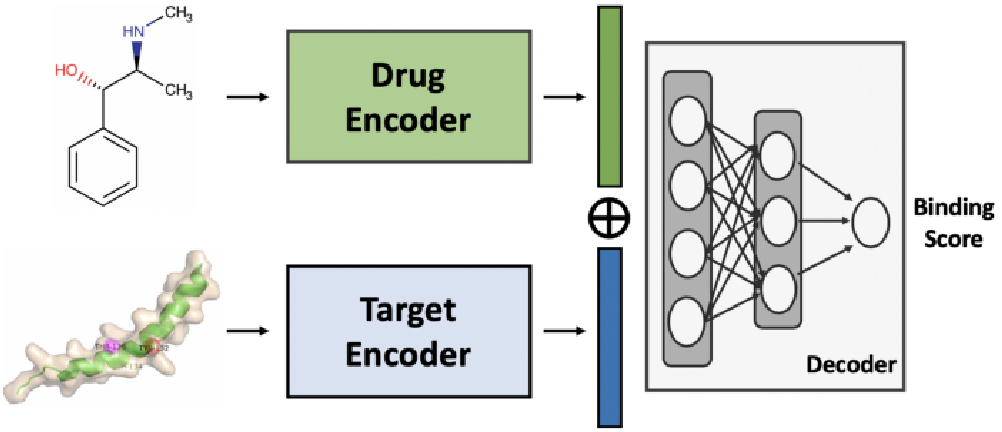

Figure 3: Conceptual Overview of a DeepPurpose DTI Machine Learning Pipeline.

There exists numerous ways of encoding these sequences into numerical representation for inclusion within a machine learning pipeline. An example conceptual overview of this process is seen in Fig. 3 exemplifying the DeepDTA model [2]. Examples of other (deep) learning mode to predict DTI includes the use of LASSO-DNN [3], logistic regression [4, 5], GNN-CPI [6], DeepDTI [7], and DeepConv-DTI [8]:

- **LR** [4, 5]: applies a logistic regression model on the concatenated drug and protein feature vectors
- **GNN-CPI** [6]: uses graph neural network to encode drugs and use CNN to encode proteins. The latent vectors are then concatenated into a neural network for compound-protein interaction prediction.
- **DeepDTI** [7]: models DTI using Deep Belief Network, which is a stack of Restricted Boltzmann Machines. It uses the concatenation of Extended-Connectivity Fingerprints (ECFP2, ECFP4, ECFP6) as the drug feature and uses PSC for protein features.
- **DeepDTA** [2]: applies CNN on both raw SMILES string and protein sequence to extract local residue patterns. The task is to predict binding affinity values.
- **DeepConv-DTI** [8]: uses CNN and global max pooling layer to extract various length local pattern in protein sequence and applies fully connected layer on drug fingerprint ECFP4.

Promisingly, by varying the sequence-based representations and the deep learning architecture, it is possible to produce improved DTI predictors. A robust model could then be used to screen for candidate drugs among the millions of candidate drugs that may target the proteins of the coronavirus (SARS-CoV-2) that causes COVID-19. Machine learning practitioners can also leverage the existing datasets and feature representations too develop their own models.

To this end, throughout A2 and A3, the DeepPurpose toolkit will be used to make easily accessible the datasets, representations, and state-of-the-art models for the task of DTI prediction. High performming models could then be used to screen for candidate drugs that may bind to SARS-CoV-2 proteins with high confidence.

The DeepPurpose ToolKit

The course project will leverage the DeepPurpose toolkit as it facilitates the building of deep machine learning solutions for drug-target interaction prediction. Many of the chemical and biological facets of DTI prediction are abstracted or wrapped up in a scikit-learn-like API. Consequently, one can focus on model development without worrying about feature extraction or feature representation.

DeepPurpose is a deep learning-based toolkit for **drug-target interaction prediction**, drug repurposing, and virtual screening. Recently, Kexin Huang developed DeepPurpose that works specifically for drug-target interaction prediction, a fundamental task for the drug discovery process designed in a scikit-learn-fashion framework that wraps 50+ models in less than 10 lines. See this Towards Data Science blog post for an overview: Drug Discovery with Deep Learning. Under 10 Lines of Codes.

From DeepPurpose's GitHub Page:

> *"15+ powerful encodings for drugs and proteins, ranging from deep neural network on classic cheminformatics fingerprints, CNN, transformers to message passing graph neural network, with 50+ combined models! Most of the combinations of the encodings are not yet in existing works. All of these under 10 lines but with lots of flexibility! Switching encoding is as simple as changing the encoding names!"*

To help guide writing the A2 proposal and in anticipation of implementing A3, the following section describes some of the data and experimental designs that could be used to evaluate individual models.

Data & Experimental Design

Two benchmark datasets may be used to evaluate the predictive performance of a model: DAVIS and BindingDB. DAVIS consists of wet lab assay K_d values among 68 drugs and 379 proteins [9] and BindingDB consists of K_d values among 10,665 drugs and 1,413 proteins [10]. DTI pairs that have Kd values less than 30 units are typically considered positive.

Optionally, balanced training can be achieved by sub-sampling the same number of negative DTI pairs as the positive samples for training set. For validation and test sets, the datasets may or many not be artificially balanced.

Models will often be evaluated on **cold target/cold drug** where none of the protein or drug sequences in the test set exist in the training or validation sets. This generates more robust and models and evaluations that reflect expected performance in reality.

Dataset should be divided in into training, validation, and testing sets according to a given split ratio. Experiments should be independently repeated a number of times with different random splits of dataset and select the best performing model.

Binary classification metrics can be used if treated as a classification problem (*e.g.* ROC-AUC, area under the receiver operating characteristic curve; PR-AUC, area under the precision-recall curve). Again, a $K_d < 30$ can be classified as a positive interaction (class 1) whereas $K_d \geq 30$ are negative (class 0). However, the raw predicted K_d value can be evaluated with regression-type metrics (*e.g.* RMSE, root mean squared error).

Assignment 2 Deliverables

Creating Models with the DeepPurpose Framework

A DeepPurpose tutorial notebook (covered in Tutorial 7) is provided to demonstration the various usages of the DeepPurpose framework for DTI prediction.

The assignment 2 notebook contains a number of questions and is the first A2 deliverable:

1. **Step 0:** Ensure that the DeepPurpose library was properly loaded.
2. **Step 1:** Load the DAVIS Dataset where the target value is in binary based on a threshold value for the binding affinity scores of 30. Print out the total number of elements and first value of X_{drug}, X_{target}, and y. Plot the distribution of y (should be a histogram for class 0 vs. class 1 with 2 bins and a title with the percentage of positives).
3. **Step 2:** Encode both the SMILES and targets using the CNN-based encodings for feature representations. Split the dataset into a training:validation:test sets by fractions 60:20:20 and ensuring none of the SMILES or Protein sequences in the test set appear in the training or validation sets. Print out the first representation of the first training SMILE and Protein sequences.
4. **Step 3:** In two subplots, plot a distribution of sequence lengths for all SMILE sequences (subplot 1) and all protein sequences (subplot 2) in the training set.
5. **Step 4:** Generate the configuration that implements the DeepDTA model (details in notebook). Train and Test the model.
6. **Step 5:** Report what the final test AUROC, AUPRC, and F1 values are. Briefly describe what each of these metrics represents.
7. **Step 6:** With reference to the configuration from Step 4, modify the configuration in such a way that might improve test performance. Briefly describe what changes you made and briefly justify those choices. Experimentally determine whether your model modifications did or did not improve performance by training the model and comparing performance of the model above.

8. **Step 7:** Report the *difference* in your new model performance compared to the original model performance. Lesser performance should be a negative value; greater performance should be a positive value.

Written Project Proposal

In anticipation of assignment 3, the second A2 deliverable is a written proposal of the model you plan to implement to generate predictions between candidate drugs and protein targets.

Prepare a written proposal describing the implementation of the machine learning solution you will create in **Assignment 3** using DeepPurpose. You are strongly recommended to familiaariize yourself with DeepPurpose while preparing the proposal and considering the information in the previous section. Assume that you are preparing this proposal for your interdisciplinary team of scientists as well as other machine learning engineers.

The proposal **must** be written in LaTeXaccording to the IEEE conference style (PDF format); you are strongly recommended to use Overleaf (sign up for a free account), and you can access the template here: Overleaf IEEE Conference Template.

The proposal should be between **3-5 pages** and may include figures, tables, equations, (sub)sections, in-text citations, and reference. Outlined below are the expected sections and approximate word count and topics for each:

1. **Abstract** (<150 words): A brief summary of the contents of the proposal
2. **Project Overview** (200-300 words, with references):
   (a) Briefly describe *Drug-Target Interaction* (DTI) prediction
   (b) Briefly describe how DTI can be *leveraged for COVID-19*
   (c) Briefly describe a few *recent DTI methods*
3. **Proposed Data & Methods** (300-500 words, with figures & references):
   (a) Describe the *datasets* proposed to be used
   (b) Describe any *data transformations and feature representation* you plan to use
   (c) Describe the *model architecture* you plan to use and any experiments you will conduct to *improve the model*
   (d) Describe strategies you will attempt to *promote model generalization*
   (e) Include a figure with a conceptual overview of your machine learning solution
4. **Expected Results** (100-200 words): Describe your anticipated results and how the experiments you expect to perform will influence that performance

Project Resources:

- Drug Discovery with Deep Learning. Under 10 Lines of Codes.
- DeepPurpose GitHub Repository
- DeepPurpose Documentation
- Video on Locally Setting up DeepPurpose
- Loading Datasets
- DeepDTI Github
- Info on the DAVIS & KIBA Datasets

References

[1] Daniel A Gschwend, Andrew C Good, and Irwin D Kuntz. Molecular docking towards drug discovery. *Journal of Molecular Recognition: An Interdisciplinary Journal*, 9(2):175–186, 1996.

[2] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[3] Jiaying You, Robert D McLeod, and Pingzhao Hu. Predicting drug-target interaction network using deep learning model. *Computational Biology and Chemistry*, 80:90–101, 2019.

[4] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[5] Dong-Sheng Cao, Qing-Song Xu, and Yi-Zeng Liang. propy: a tool to generate various modes of chou's pseaac. *Bioinformatics*, 29(7):960–962, 2013.

[6] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2):309–318, 2019.

[7] Ming Wen, Zhimin Zhang, Shaoyu Niu, Haozhi Sha, Ruihan Yang, Yonghuan Yun, and Hongmei Lu. Deep-learning-based drug–target interaction prediction. *Journal of proteome research*, 16(4):1401–1409, 2017.

[8] Ingoo Lee, Jongsoo Keum, and Hojung Nam. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15(6):e1007129, 2019.

[9] Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

[10] Michael K Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, 44(D1):D1045–D1053, 2016.