Contributors: Pranav Punuru, Adithya Chandrasekar, Aditi Anand, Amrutha Vaidyam, Pranesh Monda, Ongshu Dutta, Shreya Krishnan, Alyssa Wan, Akshitha Kartigueyan, Praneel Bhandari, Twinkal Barai

Introduction

Drug Discovery is a useful yet very arduous process in pharmaceutical sciences that allows for candidate medications to be identified and possibly implemented to solve healthcare problems. On average, the whole process takes 10-15 years [1] and the result of the study may not even be fully implemented into the market. The process has two phases: early discovery and development, with development taking significantly longer [2]. However, the power of technology allows for less false positives and drugs that are better candidates for actual implementation in the healthcare system. Specifically, Machine Learning and Artificial Intelligence pre-screening approaches provide promising potential to identify highly potent drug candidates. The quantitative measure that will be used in this project for the machine learning algorithm will be docking scores. Docking scores provide the binding affinity between two molecules after they have been docked and are useful because they can be used for virtual screening as a starting point in drug discovery. This project develops an artificial neural network model to accurately predict the docking scores of candidate drug molecules on SARS-CoV-2 protein targets. The results of the model provide a good system for pre-screening drug candidates and allow for the drug discovery process to be more efficient and ultimately yield more effective drug products that solve a vast array of healthcare problems.

[1] https://phrma.org/policy-issues/Research-and-Development-Policy-Framework
[2] https://doctortarget.com/wp-content/uploads/2018/02/Diapositiva2.png

Methods

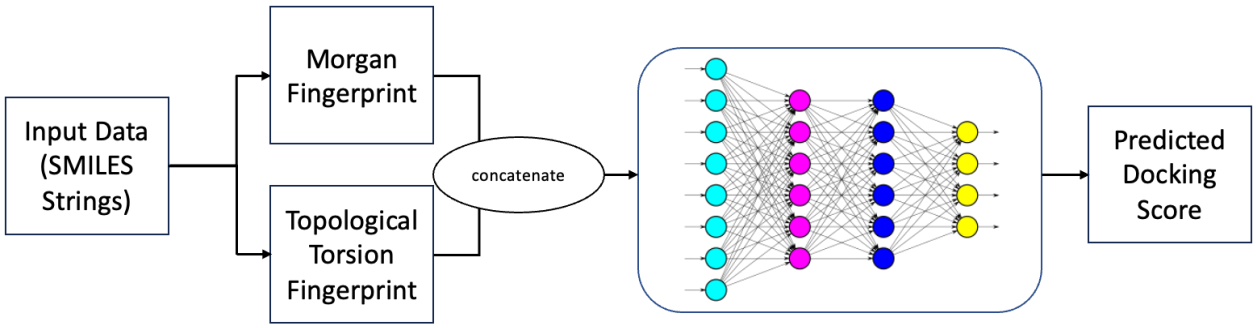

Figure 1. Overview of proposed method for predicting docking scores on SARS-CoV-2 targets

Figure 1 presents an overview of the proposed method, which consists of representation extraction followed by classification with a neural network.

1. Each entry in the data set comprises a SMILES string of a drug molecule, followed by the docking scores which are the binding affinities of the molecule to various COVID protein targets.

2. To train a neural network on this data we must transform the SMILES string into a representation that faithfully captures the essence of a molecule.
    a. We consider two representations - Morgan Fingerprint and a Topological Torsion Fingerprint, as well as a concatenation of both.
    b. Morgan Fingerprint: This fingerprinting method captures the local chemical environment around each atom in the molecule. For each atom we look at, its neighbors up to a specified radius are considered, resulting in a bit vector that indicates the presence or absence of a particular substructure in the molecule.
    c. Topological Torsion Fingerprint: This technique emphasizes the environment around bonds or torsions. It looks at sequences of four bonded atoms, encoding information about their types and connectivity into a bit vector.
    d. For each molecule, the two fingerprints are concatenated which creates a good feature set for training.

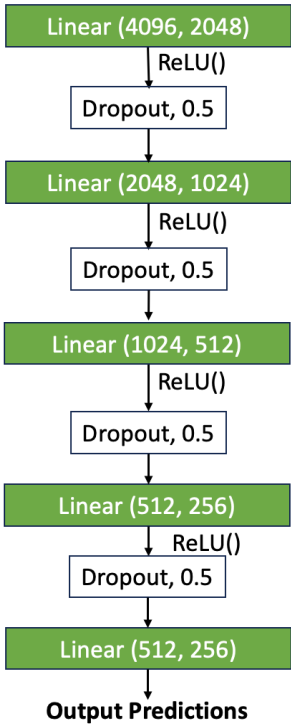

Figure 2. Topology of the neural network used for docking score prediction

Results

We implemented the method described in Figure 1 using RDKit, a collection of cheminformatics and machine learning software written in C++ and Python [cite-rdkit]. The neural network topology shown in Figure 2 was implemented using the Pytorch machine learning framework

[cite-pytorch]. In our experiments, we considered three different representations for the input molecule - Morgan Fingerprint, Topological Torsion Fingerprint, and a concatenation of both.

The provided data set of 270,000 samples is randomly split into a training set and validation set in the ratio of 80:20. The training hyperparameters used consisted of a learning rate of 0.0005, 150 epochs and a batch size of 32.

The results of our experiments are presented in Table 1 below. We see that using both the Morgan Fingerprint and Topological Torsion Fingerprint results in the best Mean Absolute Error (MAE) of 0.388 on the validation set. Using the Morgan Fingerprint alone and Topological Torsion Fingerprint alone result in MAEs of 0.414 and 0.412, respectively.

Table 1: Mean absolute error of predictive model using each combination of training data

Training Data Combination	**Mean Absolute Error**
Morgan Fingerprint Only	0.4127280606582977
Topological Torsion Fingerprint Only	0.41452315418230806
Both Fingerprints (concatenated input)	0.3191453876294856