Deep learning in drug discovery

Abstract

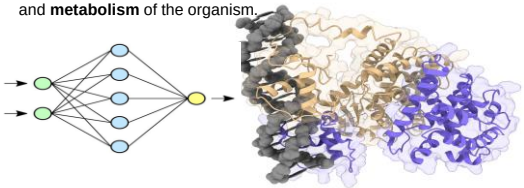
Determination of protein-drug interactions is crucial in drug discovery. The rising of deep learning techniques provides new possibilities to learn the molecular properties directly from chemical data.

Here, we present a framework for predicting drug interactions with nuclear receptors that uses graph convolutional neural network. Nuclear receptors are among the most popular drug targets since they regulate a huge number of processes in the cell.

In our research we made a comparison between different data splitting methods to see how the choice of the splitting algorithm can affect model training efficiency. Different ways of splitting are important to prevent data leakage which leads to overfitting.

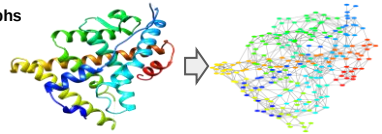
Introduction

A **neural network** is an algorithm that aims to replicate the workings of real life neurons. It can learn by adjusting variables in its functions and comparing its results with the given correct result.
Graph Neural Network is a type of Neural Network which directly operates on the Graph structure. Graphs are a kind of data structure which models a set of objects (nodes) and their relationships (edges). GNNs provide an easy way to do node-level, edge-level, and graph-level prediction tasks.
Nuclear receptors are a class of proteins found within cells that are responsible for sensing **steroid and thyroid hormones** and certain other molecules. Nuclear receptors regulate the expression of specific genes, thereby controlling the **development**, **homeostasis** and **metabolism** of the organism.

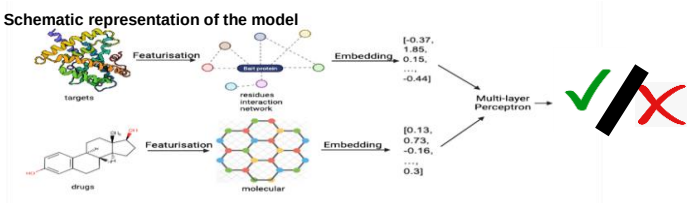


Results

1. Protein graphs



2. Schematic representation of the model

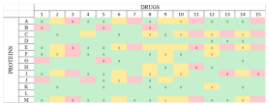


3. Splitting the dataset into train, test and validation subsets facilitates the recognition of "***over-fitting***" of the model. What this means is that the model would otherwise just memorize the right answers and would not be adequate for dealing with data that wasn't originally present in the dataset.

Graphic representation of a splitting method where all interactions of both randomly selected proteins and drugs are kept hidden from the model in the training subset (**cold-protein-cold-drug split, top-left**)

Cold-protein-cold-drug split:

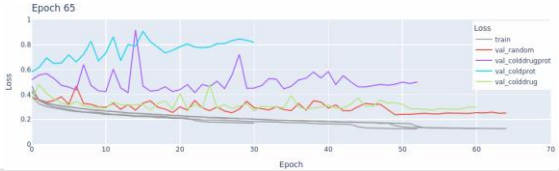
Random split:



Cold-drug split:

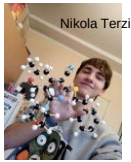
Cold-protein split:

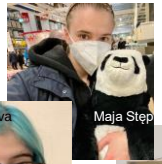
4. Loss of different models: training vs. validation

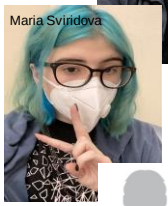


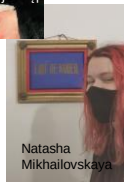
Lab members

Students

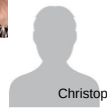

Nikola Terzic


Maja Step


Maria Sviridova


Natasha Mikhailovskaya


Sasha Panina


Christopher Sun

Instructors


Ilya Senatorov


Olga Kalinina


Ilya Mazein

Conclusions

- We have created the first of its kind deep learning model for predicting drug binding to nuclear receptors, the second most important drug target class
- Based on our results, our model needs further improvements before being used with new proteins and drugs
- Possible refinement of the model would be writing a regression model instead of a classification one
- Hopefully, this project will inspire more research in the field