Deep Learning in Drug Discovery

Three areas of focus

1. Drug properties: Predict drug properties, could be framed as a supervised learning problem
2. Drug Design: Design a compound that has certain properties (hard problem to solve, could use autoencoders or GANs)
3. Drug Targe Interaction (DTI): predicting whether a particular drug can bind to particular proteins or not.

Few Things to keep in mind

Drug compounds are complex molecular structures, that require some handling before they can be fed into the model

The project proposal pdf that Dr Sun provided mentions one such technique called **SMILES**. It is the way of converting graph structure data to textual content and using the text (encoded string) in the machine learning pipeline. This then can be used to conduct NLP tasks like predict the properties, side effects and chemical-chemical interaction.

Problem with SMILES is the context loos, the conversion from graph structure to textual content leads to lossy compression.

Link: <http://opensmiles.org/opensmiles.html>

Representing the molecular structure as a Graph, a compound can be considered as a graph, where vertices are atoms, and chemical bonds between atoms are edges.

Links: <https://tkipf.github.io/graph-convolutional-networks/>

Graph libraries: <https://github.com/rusty1s/pytorch_geometric>

<https://github.com/facebookresearch/PyTorch-BigGraph>

Of all the three areas listed above, DTI probably is the one we should focus on. Normally existing traditional machine learning methods for predicting DTI can be roughly divided into similarity-based and feature-based approaches, and most of them formulate the problem as a classification problem, even DNN addresses this as a classification problem, which can again cause binding context loss.

PADME)([Feng et. al.](https://arxiv.org/pdf/1807.09741.pdf)).: Using RNN and SMILES, framed as a regression problem with NLP

Protein And Drug Molecule interaction prEdiction

(PADME) overcomes the limitations of the existing methods by predicting real-valued interaction strengths instead of binary class labels, and, to address the cold-start problems (drugs or targets that are absent from the training set but appear in the test set).

Diagram

Description automatically generated

Here SMILES is used to create text representations, they us a RNN as the NN. As a rule of thumb, you can use RNNs based architectures (GRU, LSTM, …) and transformers in case of text-based representation for inputs, and Convolutional neural networks in case of image or 3D structure.

DTI can be framed binary classification (whether a compound bind to the target) or regression (prediction the strength of affinity between compound and proteins).

**FRnet-DTI**[**(Rayhan et. al.)**](https://www.sciencedirect.com/science/article/pii/S2405844020302899)Two convolutional neural networks are proposed: FRnet-Encode and FRnet-Predict. Here, one model is used for feature manipulation and the other one for classification. Using the first method FRnet-Encode, we generate 4096 features for each of the instances in each of the datasets and use the second method, FRnet-Predict, to identify interaction probability employing those features.

Diagram

Description automatically generated

Diagram

Description automatically generated

Graph Neural Networks using 3D molecular structure([Lim et. al.](https://arxiv.org/pdf/1904.08144.pdf)): CNN

In their work, the atomic coordinates of a protein-ligand complex around the binding site are represented on a 3D rectangular grid. Then, a 3D convolutional neural network is applied to the rectangular grid to classify whether the complex is active or not. This allows not only classify active and inactive compounds but also distinguish active and inactive binding poses.

Methodology: 1. embedding the 3D structure of a protein-ligand complex in an adjacency matrix, 2. devising attention algorithm considering 3D structures and 3. introducing a variant of graph neural networks that is suitable for learning protein-ligand interaction.

GRAPHDTA([Nguyen et. al.)](https://www.biorxiv.org/content/10.1101/684662v7.full) Using CNN

Graphical user interface, application

Description automatically generated

Data sets:

DrugBank database (<http://www.drugbank.ca/>). KEGG DRUG captures abundant approved drugs in Japan, USA and Europe based on the chemical structure and molecular interaction network information, of which most drugs reported corresponding the information of target proteins

SNAP: https://snap.stanford.edu/biodata/datasets/10002/10002-ChG-Miner.html