BS6207, Nov2022 project description:
Predicting protein ligand interactions

Understanding and predicting protein-ligand interactions is very important in in-silico drug discovery. Effective predictions can save billions of dollars within the whole pipeline of drug development.

The objective of this project is to develop AI models to predict protein-ligand interaction. The framework in which we can achieve this is through the fully supervised learning method. You are given a long list of protein and ligands as training set. In the testing set, identical data format is given except that your AI task is to predict actives and decoys.

There are 3 columns in your training data,

1. The first column gives the protein databank identification number (PDB-ID). You can go to protein databank to search for the structure of the protein.
2. The second column give the SMILES sequence of the ligand
3. The third column tells if this ligand will bind to the corresponding protein

Some useful links:

You can download the training data from: https://web.bii.a-star.edu.sg/~leehk/bs6207/bs6207.html
https://www.rcsb.org/structure/{PDB_ID}
https://www.rcsb.org/fasta/entry/{PDB_ID}
https://files.rcsb.org/download/{PDB_ID}.pdb
https://www.rdkit.org

Some references:

a. DrugVQA: https://www.nature.com/articles/s42256-020-0152-y (2D CNN + LSTM)
b. BridgeDPI: https://academic.oup.com/bioinformatics/article-abstract/38/9/2571/6547049?redirectedFrom=fulltext (CNN + FNN + meta-GNN)
c. PotentialNet: https://pubs.acs.org/doi/10.1021/acscentsci.8b00507(GNN).
d. AtomNet: https://arxiv.org/pdf/1510.02855.pdf (3D CNN). I don't see this published in a journal, just arxiv.