This is the Dataset construction protocol with output format description.

INPUT: The raw datafiles should contain the protein sequences, compound smiles, and target interaction.

OUTPUT: input for training NN models (4 components):

adjacencies.npy

compounds.npy

proteins.npy

pssms.npy

energy.npy

rdkitfeatures.npy

fingerprint\_dict.pickle

word\_dict.pickle

Step1: process raw experimental data.

The process is to read the experimental data and select the regression or multi-class classification model according to the aim to the model.

Generate the raw\_data.txt, which contains these columns:

Item number

Sequence of the protein

Smiles of the ligand

Interaction values

Step2: generate PSSM files.

1. extract all the possible FASTA files from the raw data.

Create the path “./fasta/”, and then run the code "find\_uni\_sequence.py" to prepare the .fasta files for sequence alignment. The code will extract all the FASTA files in the folder and generate the “pre\_dataset.txt”file.

1. generate PSSM files with the FASTA files by PSI-BLAST

PSI-BLAST is a software on National Center for Biotechnology Information (NCBI) website (<http://www.ncbi.nlm.nih.gov/BLAST/>).

The code for running PSI-BLAST:

psiblast \

-query fasta\_file.fasta \

-db /data/database/blast\_data/nr \

-num\_iterations 3 \

-out output\_file.out \

-out\_ascii\_pssm pssm\_file.pssm)

1. process PSSM original results

Create the path “./pssm/”, then create the folder “originresults”and “slidingwindow”

Put the pssm files from PSI-BLAST into “./pssm/originresults/”.

Input “sliding window size”when running the “pssm\_process.py”. The processing pssm results will be saved in “./pssm/slidingwindow/”.

Step3: generate RdKit features files.

Run the code “rdkitfeatures.py”

Note: the “inputdir”in the code is the path of the “pre\_dataset.txt”.

This code will generate the RDKit descriptors for each item and select the features based on the feature importance in tree base models (Random Forest).

Step4: generate Energy terms features.

Using RosettaCM to generate the energy terms of the variants.

Run the code “weighted\_energy\_file.py” to generate the out files.

Step5: extract the Input files for training the model.

Run “preprocess\_data.sh” with paths of the sliding-window folder and the folder contains the “dataset.txt” and “rdkit\_final.csv”. (modified the the preprocess\_data.sh with the path of the folders)

Finally, all the input files are generated and saved in the input path.