Data preparation and structures

For the approach I am going to take to predict molecular compound properties using 1D, 2D, and 3D multimodal data and deep learning algorithms, **I realized that my algorithms need to be compared to previous works**, and for the selection of datasets and comparison of algorithms, I learned about several options:

1) Use independent data different from that used in the previous works (training and test data) and the algorithm I designed to model and compare with the previous works.

2) Use data(**Public dataset**) from previous work for training(**Of course, my data processing methods and training algorithms are unique**), test on the same test set and do a comparison.

3) Combine the above two options.

Since not every work uses the exact same training and test sets, I decided on the third option. That is, I will use the training data((**Public dataset**)) from literature [Prediction of Drug-Induced Liver Toxicity Using SVM and Optimal Descriptor Sets--Jaganathan] as the raw data for my project and compare the results with that literature on the same test set and also with the results of other works that uses a different test set.

Questions:

1, Is the scenario I described above feasible?

2, Is there a problem if my algorithm does not perform as well as the previous works?

3, Does the literature used to make the comparison need to be the latest and greatest currently available in the industry?

If my scenario is feasible, my goal will be to **predict Drug-Induced Liver Toxicity using multimodal data (1D, 2D, 3D) as well as deep learning**.

My raw dataset will be as follows:

SMILES type data (1D) of 1253 compounds(samples) as raw data for training set and SMILES type data of 208 compounds as raw data for test set.

Where the training set consists of 636 hepatotoxic and 617 non-hepatotoxic compounds and the test set consists of 94 hepatotoxic and 114 non-hepatotoxic compounds.and the labels of the samples are “Negative” and “Positive”

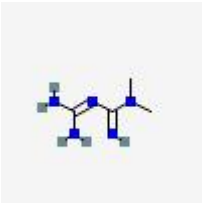
The data can be viewed on the Github:

https://github.com/Charless9/Xiancheng_Predict-the-properties-of-molecular-compounds-base-AI/blob/main/Data/Raw_Dateset.xlsx

For example,for Metformin:

Its SMILES(1D) format is CN(C)C(=N)N=C(N)N

Its 2D format is: Its 3D format is:



And its label is “Negative”.

A more detailed project framework is given below: