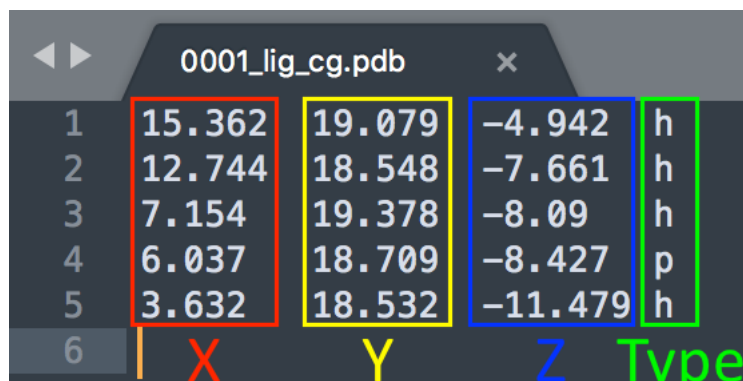


## Testing Data

The testing dataset consists of a set of protein .pdb files and ligand .pdb files. Each file is named as rrrr\_pro.cg.pdb and tttt\_lig.cg.pdb where “rrrr” and “tttt” are randomized indices. **There are 824 protein and 824 ligand files and they are named from 0001 to 0824; however, matching protein and ligand file ids (e.g. “rrrr”=0001 and “tttt”=0001) do not imply they are binding! This is your task to find matching pairs.**

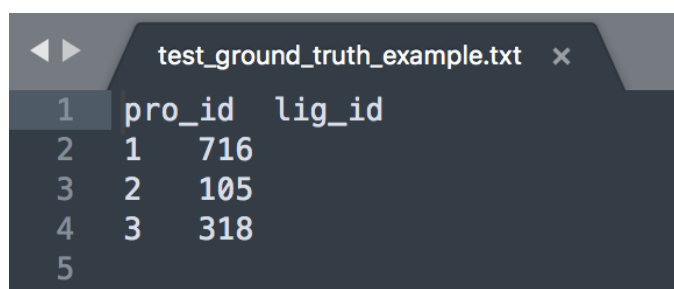
Each protein and ligand file contains four tab-separated columns storing X, Y, Z coordinates data and atom type data as shown in Figure 1. You can read these files with “read\_testing\_pdb\_file.py”.



	X	Y	Z	Type
1	15.362	19.079	-4.942	h
2	12.744	18.548	-7.661	h
3	7.154	19.378	-8.09	h
4	6.037	18.709	-8.427	p
5	3.632	18.532	-11.479	h
6	X	Y	Z	Type

Figure 1: Data structure in protein and ligand files

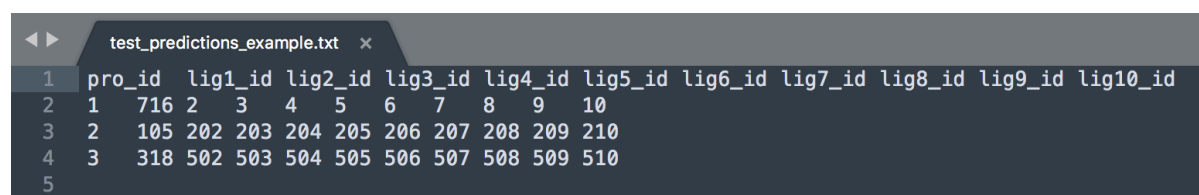
For the metric of grading, for each protein in the test data set, you are required to identify 10 ligands that are predicted to bind the protein. Prediction is considered to be correct if true ligand, which will bind to protein, is among the 10 ligands predicted to bind. For each protein, only one ligand will bind to it. You are allowed to suggest 10 candidates and hope that among your 10 candidates, you identify the correct one. Final score of the project is the number of proteins with a correct prediction for binding. We provide you with a script to calculate the top 10 accuracy: “calculate\_top\_10\_acc.py”. An example ground truth file is shown in Figure 2 in case you want to prepare a validation set and test your model.



	pro_id	lig_id
2	1	716
3	2	105
4	3	318
5		

Figure 2: Example structure for ground truth file

For your project output, you need to provide “test\_predictions.txt” file consisting of 11 tab-separated columns. First column contains protein id (scalar from 1 to 824) and the other 10 columns contain candidate ligand ids (scalars from 1 to 824). Example file is shown in Figure 3. **Please do not forget the header of the file.**



	pro_id	lig1_id	lig2_id	lig3_id	lig4_id	lig5_id	lig6_id	lig7_id	lig8_id	lig9_id	lig10_id
2	1	716	2	3	4	5	6	7	8	9	10
3	2	105	202	203	204	205	206	207	208	209	210
4	3	318	502	503	504	505	506	507	508	509	510
5											

Figure 3: Example predictions file

Structure of whole testing data folder is shown in Figure 4.

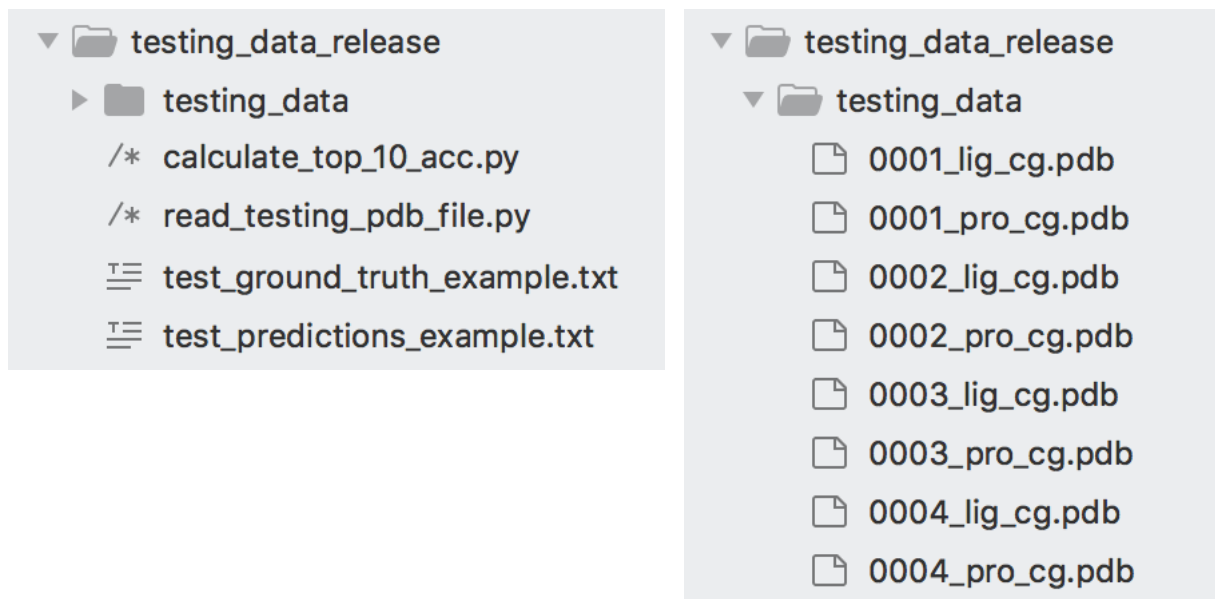


Figure 4: Folder structure of testing data

### Submission Format

You will submit a “.zip” file in IVLE under ‘Files/Student Submission/Project/’. The name of the file must be the id of the students in the group connected with ‘\_’ (e.g. e0123456\_e9876543.zip for groups and e0123456.zip for individuals). **Your folder must contain the followings:**

**1) ‘test\_predictions.txt’**

Structure of “test\_predictions.txt” file must be as described above.

**2) ‘report.pdf’**

Report format: **Strictly limited to 6 pages** (figures and text font size >10) + 1 page reference, no appendix. Any longer report will be rejected.

Report template:

- Problem definition
- Highlights (new algorithms, insights from the experiments)
- Dataset pre-processing description
- Training and testing procedure
- Experimental study

Clarity, model understanding, and highlights are important for the assessment.

**3) ‘code’ folder**

You will submit your code inside ‘code’ folder. Your code should be clean, well structured and documented (includes comments).

**4) ‘README.txt’**

You should write a proper README file defining how to prepare the data and run your code so that we will be able to reproduce your results.

Folder structure of the compressed file must be as shown in Figure 5.

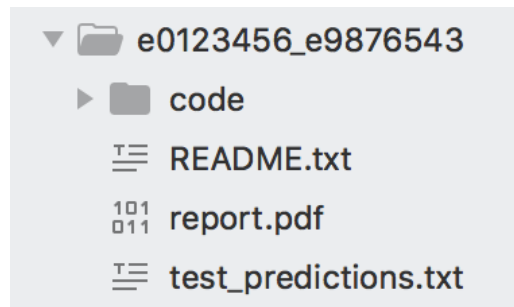


Figure 5: Folder structure of compressed submission file