UPDATE: Test Dataset section changed.

Project due date: Week 10, Monday, 23:59pm

Background

Molecular recognition between proteins and ligands plays an important role in many biological processes, such as membrane receptor signaling and enzyme catalysis. Predicting the structures of protein-ligand complexes and finding ligands by virtual screening of small molecule databases are two long-standing goals in molecular biophysics and medicinal chemistry [1, 2]. Knowledge-based statistical potentials have been developed for modeling protein-ligand interactions. They are based on distributions of intermolecular features in large databases of protein-ligand complexes.

Over the past decade, deep learning has achieved remarkable success in various artificial intelligence research areas. Evolved from the previous research on artificial neural networks, this technology has shown superior performance to other machine learning algorithms in areas such as image and voice recognition, natural language processing, among others. The first wave of applications of deep learning in pharmaceutical research has emerged in recent years, and its utility has gone beyond bioactivity predictions and has shown promise in addressing diverse problems in drug discovery [3].

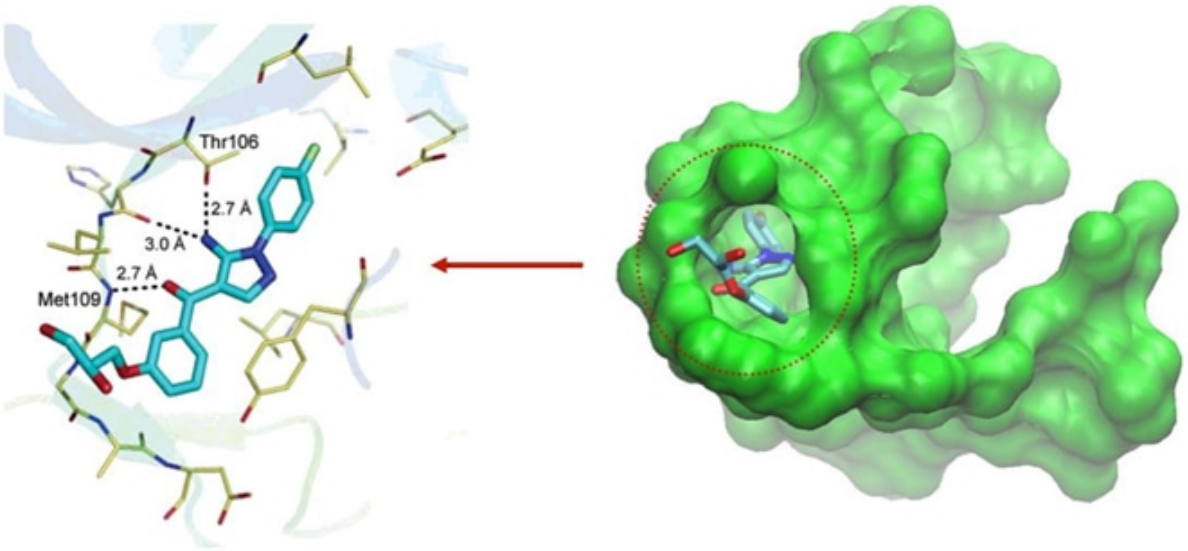


Figure 1: Illustration of protein-ligand complex. Source

Project Task

In this project, the task is to find the most possible ligand which binds a given protein at a specific position. This is determined by (1) A *SMILES formula* (an unambiguous sequence representing a chemical structure) of ligand, (2) Structural information of a protein, including

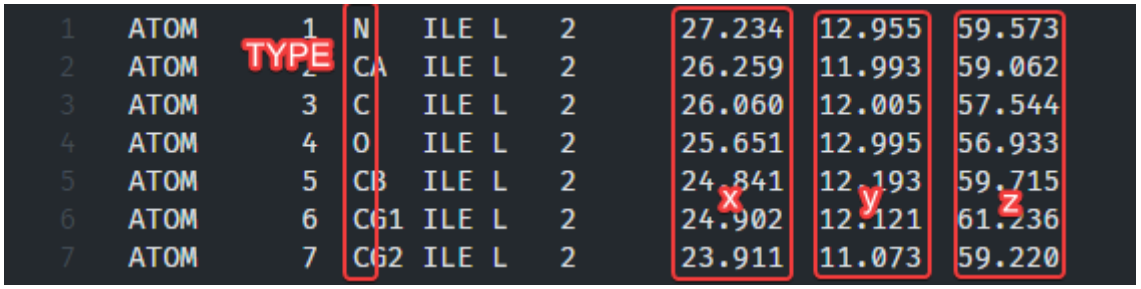
thousands of atoms and their properties, (3) The (x,y,z) location on the protein where some ligand will dock. The above data can serve as input to a neural network to predict a possibility score of protein-ligand binding.

For the metric of grading, for each protein in the test data set, you are required to predict 10 ligands that are most possible to bind the protein. For each protein, there is only one matching ligand. Prediction is considered correct if the matching ligand is among your 10 candidate ligands for it. Final score of the project is the number of proteins with correct prediction for binding.

Training Dataset

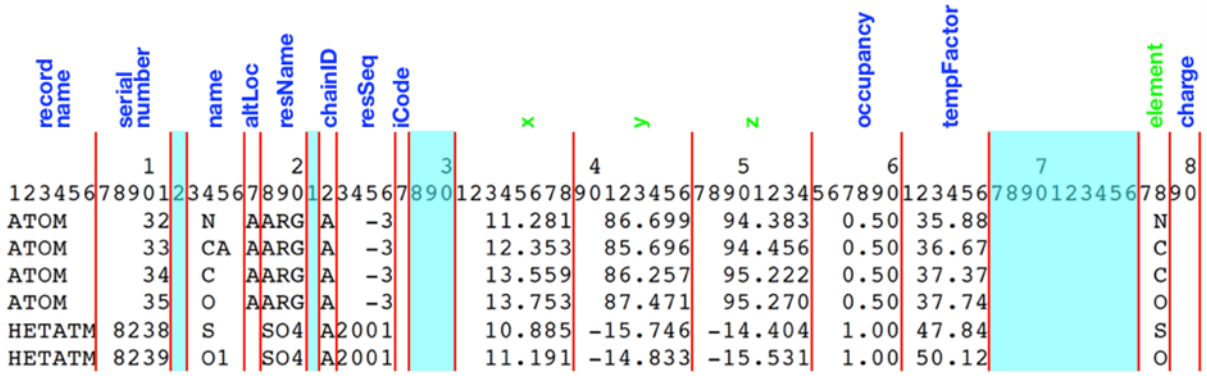
In our dataset, there are 3000 protein-ligand complexes that were determined experimentally with 3D structures available. Proteins are archived in **proteins.zip**. After extracting, you can get a folder **pdbs/** storing all proteins. Each file is for one protein and contains its atom information, and the filename is the protein's ID. For example, A file with path **pdbs/1A0Q.pdb** stores atom information about the protein with ID *IA0Q*.

Proteins consist of atoms and these atoms form the structure of them (characterize them) by coming together. Protein data files contain (x, y, z) coordinates of each atom of the proteins. In Figure 2, data structure of these protein files are shown. Note that atom types are given as *the first character of the name field*, with meaning 'C - Carbon', 'O – Oxygen', 'N – Nitrogen', etc. For instance, the first atom in Figure 2 is Nitrogen. Figure 3 shows all data fields in a protein data file. For the detailed description of each field, please refer to pp.180-182 and pp.187-188 of [4].



1	ATOM	1	N	ILE	L	2	27.234	12.955	59.573		
2	ATOM	2	CA	ILE	L	2	26.259	11.993	59.062		
3	ATOM	3	C	ILE	L	2	26.060	12.005	57.544		
4	ATOM	4	O	ILE	L	2	25.651	12.995	56.933		
5	ATOM	5	CB	ILE	L	2	24.841	12.193	59.715		
6	ATOM	6	CG1	ILE	L	2	24.902	12.121	61.236		
7	ATOM	7	CG2	ILE	L	2	23.911	11.073	59.220		

Figure 2: Data structure in protein data files



record name	serial number	name	altLoc	resName	chainID	resSeq	iCode	x	y	z	occupancy	tempFactor	element	charge
	1			2			3	4		5	6		7	8
12345678901	23456	7890	1	23456	7	890	1	2345678	90123456	7890	1234567	8901234	5678	90
ATOM	32	N	AARG	A	-3			11.281	86.699	94.383	0.50	35.88	N	
ATOM	33	CA	AARG	A	-3			12.353	85.696	94.456	0.50	36.67	C	
ATOM	34	C	AARG	A	-3			13.559	86.257	95.222	0.50	37.37	C	
ATOM	35	O	AARG	A	-3			13.753	87.471	95.270	0.50	37.74	O	
HETATM	8238	S	SO4	A	2001			10.885	-15.746	-14.404	1.00	47.84	S	
HETATM	8239	O1	SO4	A	2001			11.191	-14.833	-15.531	1.00	50.12	O	

Figure 3: Data fields in protein data files

In this project, we are solely interested in coordinates and types of the atoms (which are indicated in Figure 2 and Figure 3) constructing the structure of the proteins. In this project, we will not use atom types directly; instead we will treat them either as hydrophobic or polar. 'C' is interpreted as hydrophobic, while 'O' and 'N' are interpreted as polar. The reason why we haven't changed them to hydrophobic ('h') and polar ('p') in our dataset is that you may visualize the structure of proteins and ligands by using some open-source software like PyMOL, which requires actual atom types [5, 6], such as in Figure 4. An example python script (read_pdb_file.py) is also provided in our course website for reading pdb files to extract atom coordinates and atom types.

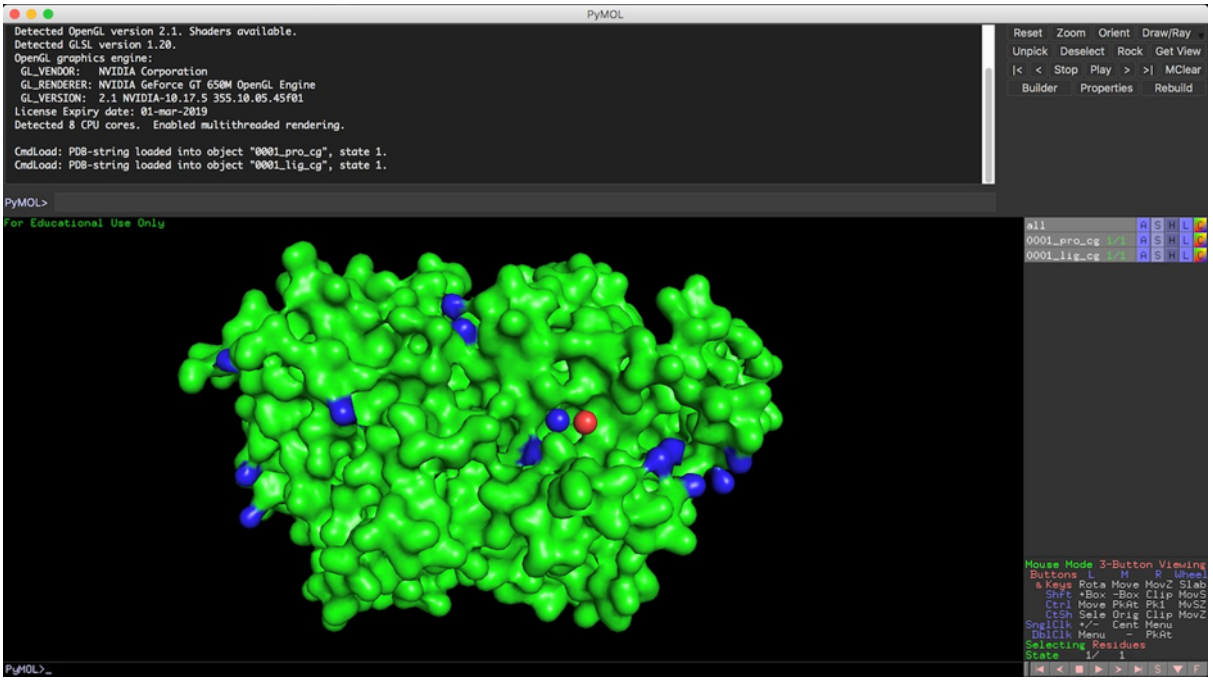


Figure 4: Visualizing structure in PyMOL

Apart from protein files, three other csv files are given to represent ligand and protein-ligand binding information.

File **centroids.csv** gives (x, y, z) coordinates of protein-ligand binding locations. For example, the 3rd line

112M,34.8922,7.174,12.4984

indicates that "protein with ID *112M* can bind a ligand at 3D position *(x=34.8922, y=7.174, z=12.4984)*".

File **ligand.csv** assigns each ligand with one ID (Column *LID*), and gives its structure (Column *Smiles*). For example, the 732nd line

732,C=NCCC

indicates that "structure of the ligand with ID *732* is *C=NCCC*".

File **pair.csv** shows the final one-to-one correspondence between ligand and protein with Columns (*PID, LID*). For example, the 3rd line

112M,732

indicates that "protein with ID *112M* and the ligand with ID *732* could bind".

Combining all information above, we can reach the fact that “protein with ID *112M* and ligand with structure *C=NCCC* binds at 3D position *(x=34.8922, y=7.174, z=12.4984)*”.

Test Dataset

We will not provide any testing data for you. However, the test dataset will consist of *exactly the same files and the same formats*, with different protein data. You are required to submit your code, and we will run your code with our test dataset to get predictions and do grading based on them.

We are still making final decisions for the code submission specification (e.g. how we retrieve your predictions). For now, you have been given enough information to start the project.

References

1. https://en.wikipedia.org/wiki/Protein
2. https://en.wikipedia.org/wiki/Protein_structure
3. https://en.wikipedia.org/wiki/Drug_design
4. http://ftp.wwpdb.org/pub/pdb/doc/format_descriptions/Format_v33_A4.pdf
5. https://pymol.org/2/#download
6. http://pymol.sourceforge.net/newman/userman.pdf