**Biola\_gt.csv**

This is the csv file which holds information about each ligand:

Columns:

SMILES – The SMILES string representation of the ligand

Docking\_score\_0-9 – The AutoDock Vina score for the top 10 docking poses

Fusion\_score\_0-9 – The Coherent Fusion score for the top 10 docking poses

Compound\_id – A name for the ligand

Activity – The –log(pKi or pKd or pK) score for the ligand. This is the binding affinity score.

Note: In a classification context, we are using a 4.5 cutoff in the activity column as (positive > 4.5, negative < 4.5)

**Rdkit\_molecular\_descriptors.csv**  
This csv holds calculated molecular descriptors (features) for the ligands. There are 208 features.

Compound\_id – Tells you which ligand the row of features corresponds to.

All other columns – A computed (rdkit) molecular descriptor feature

Note: Some useful preprocessing I’ve seen done on these features is a normalization of the values

Note: I’ve also seen running a Principal Component Analysis on these features be useful

**RDkit\_molecular\_fingerprints\_base64\_encoded.csv**

This csv holds calculated molecular fingerprints for the ligands.

Compound\_id – Again tells you what ligand the row is for

Fingerprint – This is a long base64 encoded string representing the ligand’s fingerprint.

Note: The fingerprint can be useful at a high level for doing tanimoto similarity between ligands. See <https://www.rdkit.org/docs/GettingStartedInPython.html>

Note: Some papers also use a combination of the descriptors / fingerprint representations of the ligands to do prediction.

**Mpro\_pocket\_8\_angstrom\_cut.pdb**

This is a pdb file for the receptor (Mpro) pocket we are predicting against.

This file format can be read by many different python libraries, one great way to look at this file is with PyMol (<https://pymol.org/2/>) which is a visualization tool.

PDB files have atoms and their coordinates and can also be opened by a text editor. This file is informational. The pocket itself will be put into the voxelized format for you along with a ligand at a given pose.

**Pptx files:**

These are some informational powerpoints (previous talks) that describe what we had students do over the summer and detail some of our previous sars-cov-2 efforts

**Data folder:**

First level of directories is compound\_id:

Second level of directories is pose number 0-9 (pose\_[0-9)

Inside each pose folder there are 4 files:

**Raw\_coordinates:**

X, Y, Z locations of every atom in the protein-ligand complex (Number of atoms x 3)

**Pairwise\_distances:**

A squareform pairwise distance matrix from every atom to every other atom (dense matrix)

(Number of atoms x Number of atoms)

**Node\_features:**

The 19 features we compute for each atom + the van\_der\_waals distances (20 features total)

Note: The rows of each of the three aforementioned matrices coorespond to one another

(Number of atoms x 20)

**Voxelized:**

A 19x48x48 matrix of the complex voxelized into the 19-dimensional feature space (19 channels)

Note: Not every ligand had 10 poses that docking found which were valid (so some ligands may have less than 10 poses)