Journal Report 3
9/16/19-9/20/19
Sohom Paul
Computer Systems Research Lab
Period 1, White

Daily Log

Monday September 9

Successfully wrote script to parse file and run Psi4 energy calculation.
Have not yet found how to programmatically extract features; could let Psi4 print to file and perform file I/O, but that would be slow.
Diverted myself by learning about the differences between UHF and RFH references.

Tuesday September 10

Finished parser to read in data in .xyz format
Started writing script to extract variables from output file.
Continued to research ways to access variables programmatically.

Thursday September 12

Finished writing script to extract variables (rotational constants, dipole moment, homo, lumo) from output file.
Modified script to execute batch jobs and output results to a .csv.
Refactored so it's not totally spaghetti code.

Timeline

Date	Goal	Met
Sep 9	Build first graph convolutional network and install relevant libraries	Partially; No backpropagation.
Sep 16	Run relevant DFT calculations; find best parameters	Yes.
Sep 23	Write Python script to run DFT calculations and pull relevant features from our dataset.	Yes
Sep 30	Build toy networks with multitask learning	
Oct 7	Integrate multitask learning into first GCNs	

Reflection

Writing the parser to read in data from or data set and extract the relevant features was pretty easy, taking only 60 lines of Python in total. What really boggled me was that it looks like the Psi4 API does not offer any better way to actually extract features except by parsing the output file of the calculation. I spent a lot of time trying to research ways around the extra file I/O, but I couldn't find any solutions. That being said, the time Python takes on file I/O should be relatively short compared to the time spent actually processing our molecules, so the speed decrease isn't the worst thing in the world. That being said, because my code is using spaghettied string searching to look for the output variable, the code will fail if the output file ever looks different from my test cases.

Reproduced below is the .csv I generated from running my script on the first 3 molecules in the dataset (methane, ammonia, and water), outputing only dipole moment (in Debye) and HOMO and LUMO energies (in hartrees).

```
Index,Dipole,HOMO,LUMO
1,0.0,0.028443021428518133,0.10485482872576515
2,1.6098,0.014617421847011052,0.09121647966708474
3,1.9148,0.009643355437129025,0.07859165810157169
```

The calculation of above quantities for three simple molecules took 13.3 s to perform. I did not time the length of the calculation once you include thermochemical calculations, but I estimate that it's around 45 seconds total. This is looking good for us; any conceivable neural network is going to beat out DFT's speed.

Now that we know that we can get a baseline for how good our results should be, I need to join my partner in coding up the machine learning side of this project. I don't know much about multitask learning aside from the fact that it can be done in several different ways, so my goals for the coming weeks are learning about these different implementations (while simultaneously learning Keras and Spektral) and then working multitask learning into a graph convolutional network.