Journal Report 20
3/2/20-3/6/20
Sohom Paul
Computer Systems Research Lab
Period 1, White

Daily Log

Monday, March 2

Installed RDKit library.
Read API for fingerprinting and drawing functions.

Wednesday, March 4

Research comparison of molecular fingerprinting schemes. I have decided to build my first fingerprints using the circular fingerprinting scheme (e.g. consider subgraphs of fixed breadth, hash the results and OR into a bit vector).

Thursday, March 5

Encountered and investigated possible bug in RDKit draw function. (Methanenitrile, or HCN, is being displayed as if it were H3CNH3, which is just not chemically valid.)
Wrote code to convert between our network's matrix representation of molecules to the RDKit representation.

Timeline

Date	Goal	Met
Feb 24	Build first example MPNN.	No; still in process of setting up testbed to use sample code.
Mar 2	Build first example MPNN (possibly in Python 3).	Partial; need to finish building and testing.
Mar 9	Finish building example MPNN. Start implementing neural fingerprinting.	Implemented nonneural circular fingerprinting instead.
Mar 16	Modify network structure to accept fingerprints.	
Mar 23	Examine existing technologies for drawing molecules.	

Final Goal Contract	
A	Our integrated multitask-GCN model produces results that are on-par with other models.* Our project is well-documented and works out-of-the-box with clear installation instructions available on our project GitHub. The user is given a visual interface to interactively draw molecules and predict chemical properties.
B	Our integrated multitask-GCN model produces results that are on-par with other models. Our project documentation exists and the project works with some minor hassles when installing from GitHub. The user is capable of viewing and selecting molecules from a pre-set list in order to predict chemical properties.
C	Our integrated multitask-GCN model produces results that are significantly worse than other methods. Our project is poorly documented and installation from GitHub is very difficult. The user can select molecules to test from a list, without any visuals showing molecular structure.

* I have adjusted this goal statement because we really have no way of ensuring that our innovations will actually improve accuracy. For this reason, our goal is merely to match the performance of other models / the models we produced earlier this year.

Reflection

This past week, I have realized that I've reached the point where, even though my prototype MPNN is not yet functional, I am falling behind schedule so I ought to carry forward. I have read all the code, so I ought to theoretically be able to quickly produce results now that I'm starting molecular fingerprinting. Also, I had initially planned to implement neural fingerprinting (which I still will do eventually), but I realized it might be more prudent to start with nonneural methods so I can modify my network pipeline and get baseline results to compare the neural methods to. I am currently exploring the API for RDKit, which can generate various types of molecular fingerprints from SMILES strings.

I have also noted this above, but while I was initially very excited at seeing the RDKit draw function (which we could potentially use for our final UI), I have already found a case where it fails. That being said, I trust the rest of the library because of its high usage in chemical computation. I have also figured out to interconvert between RDKit and Spektral representations of molecules. For next week, I will try to adjust my networks to accept molecular fingerprints.