Improving Molecular Modeling through Advances in Force Field Development

Submitted by: Daniel Cabrera
Donald Bren School of Information and Computer Sciences
Software Engineering
University of California, Irvine

UROP Proposal

Supervisors:
Dr. David Mobley
Vice Chair, Professor of Chemistry and Pharmaceutical Sciences
School of Physical Sciences
University of California, Irvine

Jessica Maat
Graduate Researcher
University of California, Irvine

Proposal for Research in Fulfillment of Guidelines for UROP Proposal

1. Introduction

Computational chemistry uses molecular modeling, which often relies on force fields, to aid in drug development. The force field utilizes potential energy functions that describe bonded and non-bonded interactions of a molecule. Molecular models that rely on force fields are capable of accurate predictions of molecular behavior, which can prove extremely useful in aiding the development of pharmaceutical drugs.

The four most historical force fields currently are, AMBER, CHARM, GROMOS, and OPLS, all rely on atom typing in their simulations (Guvench, O., & MacKerell, A. D). Atom typing uses preset "atom types" to classify atoms and assign parameters in a given simulation, as opposed to using the actual atom or molecule itself. This layer of ambiguity results in simulations and molecular models that are not as accurate as they can be. Mobley Lab at UCI has addressed this issue through the creation and use of smirnoff99Frosst, a force field that acts directly onto the atoms and molecules being used in a simulation. This allows for more accurate parameterization in a simulation, which results in an improved molecular model and increasingly accurate predictions about molecular behavior (Mobley et al., 2018).

For this project, I will be working within UCI's Mobley Lab with Dr. Mobley and graduate researcher Jessica Maat on the open source force field development project Open Force Field (openforcefield.org). Ultimately, we will be working to use computer science techniques to improve molecular modeling and force fields, specifically smirnoff99Frosst.

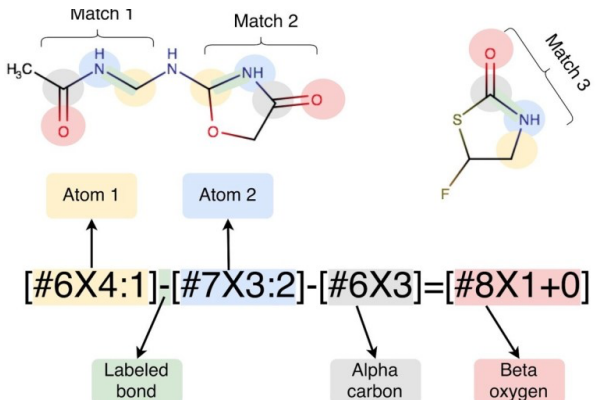

Figure 1: Direct chemical perception of a molecule. SOURCE: mobleylab.org

2. Objective

Through my own research, I aim to continue improvement on molecular modeling through contribution to the DANCE pipeline. DANCE is a program that acts as a filter for databases such as eMolecules, a popular molecular database of all commercially available compounds. DANCE works to generate diverse datasets of molecules with fragment based fingerprints. This allows computational chemists to filter for molecules with specific molecular substructures and properties to improve the force fields (dance.readthedocs.io). My contributions to DANCE will include developing a customizable molecular fingerprinting method, curating a diverse dataset for molecular bonds, and performing chemical diversity analysis on these datasets.

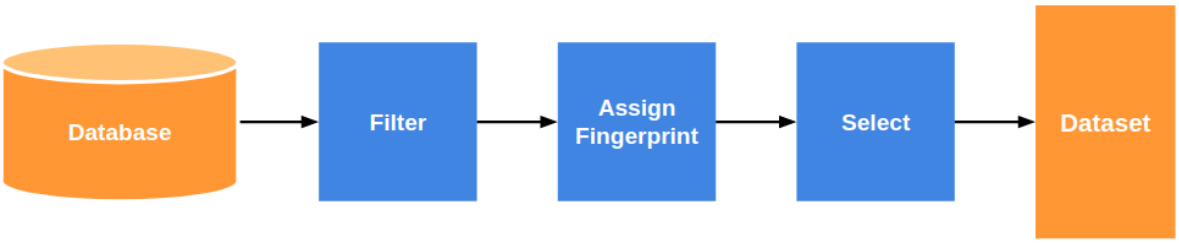

Figure 2: Diagram of the DANCE pipeline steps. SOURCE: dance.readthedocs.io

3. Aims

Aim 1: Develop a customizable molecular fingerprinting method

Fingerprinting is a common method of defining molecules to extract relevant properties to select for diversity or similarity, cluster databases, or molecular characterization. For example, MACCS, LINGO, Circular, Path, and Tree are some of the most popular fingerprint types that encode the 2D graph features of a molecule (docs.eyesopen.com). Although these techniques are valuable for characterization, they take into account the entire molecule. There is currently no existing fingerprinting method that encodes for targeted molecular fragments. In force field development, it is extremely important to select training data that contains not only targeted molecular substructures, but also specified molecular properties. DANCE seeks to fill this need for customizable fragment based fingerprints.

The first aim of my UROP proposal is to develop a customized fingerprint for molecular fragments. DANCE currently implements fragment based fingerprint for impropers and select torsion parameters. The aim of my project is to extend this method to bonds, in particular nitrogen-nitrogen bonds. In order to accomplish this, I will be examining two specific traits of these bonds.

The first trait I will be examining is the neighboring atoms of the nitrogen-nitrogen bond. To acquire the neighboring atoms of the bond, I will develop scripts that utilize the OpenEye toolkit to isolate the nitrogen-nitrogen bond in a molecule. From this point, I will iterate through the rest of the molecule, with the knowledge of the nitrogen-nitrogen bond's location and create a small data set of all the neighboring atoms, identified by their atomic numbers. Importantly, I will set a maximum number of four neighbors that will be located. This allows for standardized fingerprinting data for every molecule and ensures the most relevant neighbors to the bond are being examined, leading to more accurate fingerprint comparison and more diverse datasets. This also displays the ability for customization in the fingerprint as the amount of neighbors taken into account can be adjusted.

The next property I will be examining in nitrogen-nitrogen bonds is their Wiberg bond order. In order to obtain this Wiberg bond order, I will make use of various tools from the Openeye toolkit

to create a function capable of using the Wiberg bond order equation, $W_{AB} = \sum_{\mu \in A} \sum_{v \in B} P_{\mu v}^2$, to calculate this value. The first step of the function will conform molecules to an OpenEye molecule. Molecules that were successfully converted will then be input into the Wiberg bond order equation. The returned value will be added to the fingerprint values for the molecule. This process also highlights the ability for this fingerprint to be customized using different values or traits from a bond or molecule as the Wiber bond order is a chemical bond property.

The neighboring atoms and Wiberg bond order are both relevant properties to a bond. This fingerprint utilizes these properties to highlight the customizability of the fingerprint, as the values of the fingerprint themselves can be expanded or changed for different results. At the same time, these values also work to create fingerprints with diverse values that will be used to generate diverse datasets.

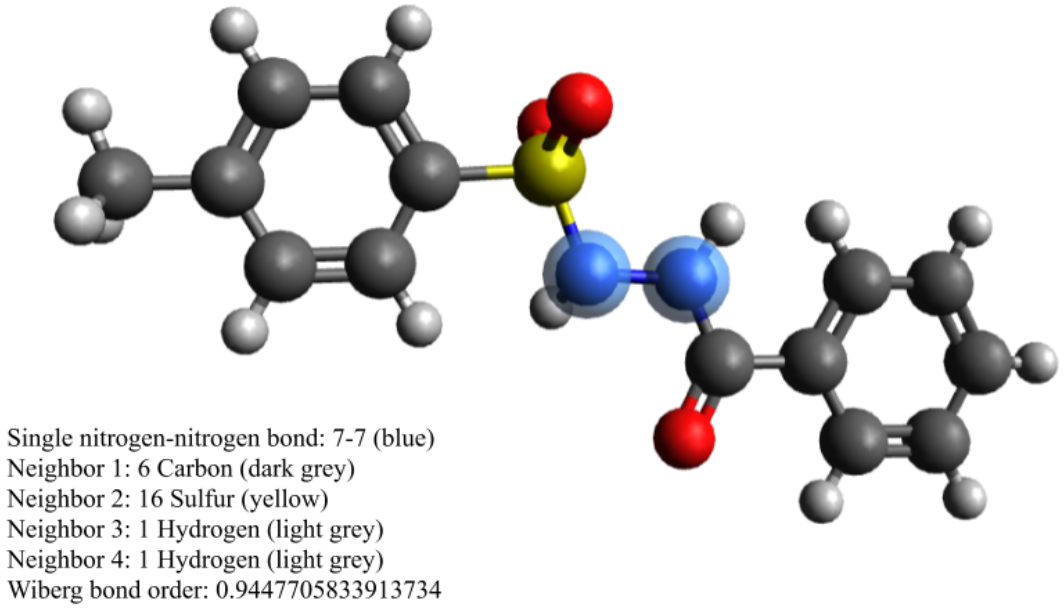

Figure 3: Fingerprinted Cc1ccc(cc1)S(=O)(=O)NNC(=O)c2ccccc2 molecule SOURCE: avogadro.cc

Aim 2: Curate a diverse dataset for molecular bonds

After the development of the nitrogen-nitrogen fingerprinting method, my goal is to use these fingerprints to curate a diverse dataset for molecular bonds. The first step for this aim is to select a starting dataset. I will select a random 60,000 molecules from the eMolecules database of 5.9 million commercially available molecules (eMolecules.com). Using a random set of molecules helps to add diversity to the dataset by erasing the chance for groupings of similar molecules in the eMolecules database to end up in the dataset. Following this, the molecules will be run through the DANCE pipeline.

As seen in the figure above, the DANCE pipeline begins by marking molecules passed through it

with a relevance function I will create. This relevance function will utilize the Openeye Scientific toolkit to search the structure of the eMolecule for instances of a single nitrogen-nitrogen bond. Additionally, this function will check how many nitrogen-nitrogen bonds are found in a given molecule and will only mark molecules as relevant if they have specifically one single nitrogen-nitrogen bond. This step especially demonstrates the value of DANCE as molecules can be filtered for very specific traits, even before being fingerprinted.

These filtered molecules will then be passed to the fingerprinting function discussed earlier where they will be assigned fingerprint values based on the atomic number of their neighboring atoms and their Wiberg bond order. The assignment of fingerprint values operates in the way that each molecule is assigned their values based on the return value of the fingerprint function. In this case, the molecules will be fingerprinted with five values, the first being the Wiberg bond order value and the last four being the neighboring atoms of the molecule. Using these values in my fingerprint will add to the diversity of the dataset as the neighboring atoms and Wiberg bond order of every bond will be unique. In both cases, the properties of the entire molecule are taken into account to create the fingerprint. Every molecule will be composed of different atoms, meaning that there will undoubtedly be different neighbors to each molecule's nitrogen-nitrogen bond. This can also be seen through the Wiberg bond order value as OpenEye Scientific defines the Wiberg bond order as, "A measure of electron population overlap between two atoms" (OpenEye). As this value is influenced by the other atoms in the molecule, each fingerprint will contain unique Wiberg bond order values. In my experiment, this will be displayed using a nitrogen-nitrogen bond, however this can be done using any bond, again showing the possibility for this fingerprint to be customizable to any molecular attribute. Also, this highlights the ability of DANCE in targeting molecular substructures and examining their unique traits to create equally unique grounds to generate a dataset upon.

Now tagged with their own fingerprint values, the DANCE pipeline will sort the molecules based on the values within their fingerprint. Specifically, the DANCE pipeline sorts in a lexicographical fashion from least to greatest, where the molecules are sorted based on the first value, then the second value should the first values be equal, then the third value should the second values be equal, and so on. In simpler terms, the values will be sorted similarly to the alphabet. In the alphabet, words beginning with "a" will always come before words beginning with "b" and words with "aa" before words with "ab". In this case, the fingerprint values of one molecule will be compared against that of another one by one until there is a difference. When there is a difference, the molecule with the smaller value will be placed earlier in the sort.

With molecules both fingerprinted and sorted, the selection frequency parameter of the DANCE configuration will be taken into account. Before running the DANCE pipeline, users have the ability to specify a selection frequency which works to further filter the molecules. For example, a selection frequency of three would mean that every third molecule in the sorted by fingerprint list of molecules will be selected and added to the final dataset. The advantage to this is that it allows for different scales of diversity to be taken into the final dataset, allowing for even more customizability through the DANCE process.

To put it simply, this fingerprint aims to achieve diversity by focusing the natural diversity of molecules into the small scope of one, specific molecular trait. In this instance, I plan to achieve

this using single nitrogen-nitrogen bonds, but the main takeaway of this process is its ability to be adapted to accommodate any molecular trait. This can be seen as the final dataset curated through the DANCE pipeline will include a set of molecules as diverse as their fingerprint allows them to be. As diversity is the goal for these molecules, it will be important to analyze the chemical diversity of the final dataset.

Aim 3: Perform chemical diversity analysis on datasets

In order to analyze the results of aims 1 and 2, it is important to use a chemical diversity measure for the datasets. To do so, we must come up with diversity measures. For aim 3 we plan to measure the chemical diversity of the datasets using 3 methods: (1) MACCS key fingerprint, (2) number of atoms, and (3) bond length analysis. In the DANCE pipeline I will implement a diversity analysis after the selection step called analysis. I will perform the analysis step on the results from aim 2 and generate plots of the chemical diversity of the fingerprinted and filtered molecules. I will also perform analysis on randomly selected molecules from the eMolecules database that contain nitrogen-nitrogen bonds. I will compare the two datasets produced by DANCE and random selection. We predict that DANCE's fingerprinting method will produce a more diverse dataset, and this experiment will prove the robustness of the fragment based fingerprinting method DANCE uses.

The three analysis methods I plan to implement in DANCE focus on different measures of diversity. The first method, MACCS key fingerprint, focuses on measuring the diversity of an entire molecule. This fingerprinting technique considers the entire chemical structure of the molecule. The next measure is the bond length analysis which focuses on the diversity of the bond length of the nitrogen-nitrogen bond. Because the fingerprinting of the nitrogen-nitrogen fragment selects for Wiberg bond order diversity, the bond length should also have diverse values. The bond length analysis will give insight into the chemical diversity of the specific fingerprinted fragment. The final diversity measure is the number of atoms. This measure may result in a similar distribution for the number of atoms for the outlined experiment in the previous paragraph. Overall these measures will focus on the diversity of the entire molecule and the specific fingerprinted fragment, and will give insight into the robustness of DANCE.

4. Responsibilities

While conducting my research under UROP, I will be responsible to complete the aims I have outlined above within the timeline I have provided in order to contribute to a Mobley Lab published manscript. In order to accomplish this, I will be consistently updating my graduate mentor, Jessica Maat, on Slack. Additionally, I will be meeting with Jessica at least once a week over Zoom to discuss the work I completed during the week and to discuss the next steps for the week ahead. During these meetings, I will be taking notes to keep track of what I have accomplished on a weekly basis and intend to accomplish in the following week. I will also be documenting all of my work on a Mobley Lab GitHub repository, following professional formatting to ensure readability and understanding of my code by others. I will also be holding myself responsible to complete a poster outlining the most important findings of my research. Building on this, I plan to present this poster at least once, hopefully at the UCI Undergraduate Research Symposium, but if that is not possible at least in a Mobley Lab meeting with other

members.

5. Timeline

This timeline outlines the following school year and the steps I will take in order to accomplish the aims of this project.

November:

- Understand the differences between bonds, angles, and impropers for parameterization of force fields
- Look through and familiarize myself with the OpenEye Scientific toolkit
- Begin development of a customizable relevance function for a DANCE pipeline to filter for single nitrogen-nitrogen bonds

December:

- Generate a small dataset of eMolecules to test the relevance function
- Test the first implementation of the relevance function through scripts to examine eMolecules
- Develop a script capable of visualizing the molecules from the .oeb files returned through the DANCE pipeline

January:

- Analyze the results of the data returned through the relevance function
- Generate a pull request for the DANCE repository on GitHub
- Begin documentation of my code through GitHub and assign reviewers
- Receive feedback on the relevance function and make fixes as necessary

February:

- Begin development of a fingerprinting function
- Create the function that will return the neighbors of a nitrogen-nitrogen bond
- Test the first implementation of the fingerprint on a small dataset of eMolecules
- Upload the progress of the fingerprint to GitHub
- Receive feedback on the fingerprint and make adjustments as necessary

March:

- Begin development of the function to calculate the Wiberg bond order
- Test the function on a small dataset of eMolecules
- Upload the progress of the fingerprint to GitHub
- Receive feedback on the fingerprint and make adjustments as necessary
- Sit it on a group meeting for Mobley Lab to hear research being presented by others

April:

- Develop a script that generates a file to explicitly outline the fingerprint data for every molecule
- Develop a script that generates a file to outline molecules that failed to conform when calculating the Wiberg bond order
- Write a README.md file which will describe the project and how to run my

experiment
- Update GitHub to include these scripts and README files
- Make contributions to a Mobley Lab research paper

May:
- Create a poster to present at the UCI Undergraduate Research Symposium (assuming it is held)
- Present at the UCI Undergraduate Research Symposium

June:
- Write papers to submit to the UCI Undergraduate Research Journal
- Submit papers to the UCI Undergraduate Research Journal

6. Itemized Budget

Poster- $80
Total- $80

The only funds required for this research are allocated for a poster to present my research. As far as I know currently, posters may be presented in the UCI Undergraduate Research Symposium at the end of the year or pushed to a later year due to Covid-19. Regardless of when posters are presented, the funds I am requesting will be enough to cover a poster that I intend to present. It is possible for the budget of my research to increase if there travel opportunities open up for other research conferences, however I have not included this as there is no confirmation of these events, again due to Covid-19.

References

Avogadro—Free cross-platform molecular editor. (n.d.). Avogadro. Retrieved October 28, 2020, from https://avogadro.cc/

DANCE Documentation—DANCE documentation. (n.d.). Retrieved October 28, 2020, from https://dance.readthedocs.io/en/latest/

eMolecules Database Download—EMolecules. (n.d.). Retrieved October 28, 2020, from https://www.emolecules.com/info/plus/download-database

Guvench, O., & MacKerell, A. D. (2008). Comparison of Protein Force Fields for Molecular Dynamics Simulations. In A. Kukol (Ed.), *Molecular Modeling of Proteins* (pp. 63–88). Humana Press. https://doi.org/10.1007/978-1-59745-177-2_4

Mobley, D. L., Bannan, C. C., Rizzi, A., Bayly, C. I., Chodera, J. D., Lim, V. T., Lim, N. M., Beauchamp, K. A., Slochower, D. R., Shirts, M. R., Gilson, M. K., & Eastman, P. K. (2018). Escaping Atom Types in Force Fields Using Direct Chemical Perception. *Journal of Chemical Theory and Computation*, *14*(11), 6076–6092. https://doi.org/10.1021/acs.jctc.8b00640

Mobley Lab, UCI. (n.d.). Mobley Lab, UCI. Retrieved October 28, 2020, from https://mobleylab.org/

Open Force Field Initiative. (n.d.). Retrieved November 5, 2020, from https://openforcefield.org/

OpenEye Toolkits 2020.1.0—Toolkits—Python. (n.d.). Retrieved October 28, 2020, from https://docs.eyesopen.com/toolkits/python/