Using Machine Learning Algorithms to Predict the Effects of Substituents on Molecular Properties and Structures

Shane Leonard†,*
† Lakehead University

Keywords: Computational Chemistry, Substituent effects, Machine Learning, QTAIM

1. Background

Chemistry is the study of atoms and molecules, the behavior and properties of molecules from the identity of the composite atoms. To understand the properties of molecules, chemists divide molecules into large fragments called substituents. This is essential in many fields from research to industrial applications. There are currently many rigorous studies being conducted attempting to better understand what effects a molecules behavior, structure and properties [1]. The focus of my research is the effects of substituents. Another important component of understanding chemical reaction outcomes is the prediction of stereochemistry [2]. It is known that substituents can influence the outcome of reactions, but the specific mechanism is still being elucidated. One property that can be used is optical rotation, the rotation of plane polarized light by a chiral molecule.

Substituents can be an atom or group of atoms such as rings or large chains. Depending on the identity of the substituent the properties, structure and reactivity of a molecule can be affected. These effects are electronic in nature. Two prominent examples are the inductive effect and the resonance effect. The inductive effect is an electrostatic transfer of energy through sigma bonds (single bonds) while the resonance effect is the transfer of energy through π (double) bonds [3]. These effects are often further delineated into whether the donor or withdraw electrons from the system. Substituent properties have traditionally been assessed using a substituent-backbone probe approach [1]. By changing the substituent and measuring the property of the probe (most commonly the pKa) with respect to hydrogen, the effect of the substituent can be assessed. Through the use of different backbones, chemists separate the transmission of the effect based on the number and types of bonds.

The Quantum Theory of Atoms In Molecules (QTAIM) was developed by Bader to describe the structure and properties of molecules [4]. It provides a mechanism for identifying molecular fragments and assessing their properties. QTAIM describes the conventional three-dimensional structures of molecules through critical points in the electron density. There are four different types of critical points, Nuclear Attractor Critical Points (NACP), Bond Critical Points, Ring Critical Points (RCP), and Cage Critical Points (CCP). The electron density gradient vector field from which the critical points are assessed provide a well-defined surface separating each and every atom within a molecule [5]. This surface can be used to define an atomic volume element over which any property of that atom can be assessed. The properties are additive, so their sum captures the overall property in any molecule. As a result, one can also examine the summed properties of a molecular fragment (a substituent).

2. Research Goals

In 2020 Vis et. al. showed the many different structures silicate polyols can adapt in aqueous solutions [6]. The existence of these structures was proven using nuclear magnetic
resonance spectra. Many different complexes were shown with 4-6 bond coordination. The reason for the specific structures these silicate polyols adapt is not completely understood. Using different substituents with known properties a better understanding about which structures are more likely to form during reactions.

Lefrancois-Gagnon et al. developed a QTAIM feature set of 116 different substituents and various properties including the dipole moment and quadrupole moment [7]. Using this data two sub projects will be done. The first project is to test the inductive effect in three different backbones, bicyclopentane, bicyclooctane and bicycloundecane. The inductive effect will be observed through an isodesmic reaction, which is a chemical reaction in which the bonds broken in the reaction are the same type of bonds formed in the product. The acid dissociation energy using each of the 116 substituents will be determined. This data will comprise the feature set to test the machine learning algorithm and compare the results to previous studies [8, 1]. The second sub project will be to assess the effects of substituents on optical rotation. The data will be generated using QTAIM and the machine learning algorithm will be further tested and calibrated. These first two steps will hopefully lead to determining the effect of substituents on polyol silicate complexes.

3. Machine Learning

The main goal after the relevant training and testing data has been generated, is to select the most appropriate machine learning method to be used for this data set. To begin building the algorithm a matrix and vectors need to be defined. For this research the matrix will be composed of the substituents and the properties determined by Lefrancois-Gagnon [6]. For investigating the isodesmic reactions acid dissociation energy the problem seems linear in nature. An input based on a property of the substituent and the acid dissociation energy being the output. If it turns out to be completely linear a simple linear regression model may be used. This would not require a large amount of computational power. A Naïve Bayesian model is also being considered using a “tree” structure with a parent node splitting to different child nodes. The child nodes could potentially help processing more than a single input variable and make specific predictions, it would also not require too much computational power though may turn out to be too rigid of a structure.

When starting the second sub project looking at chirality these more linear models may not be sufficient. The study of substituent effects on the chirality of tetramethyl structures has not been done before, meaning we can not be sure the input and output will be as linear as the study of isodesmic reactions. Simple linear regression may be inefficient and even non-linear regression may not fit the data accurately enough.

The reason for doing the two sub projects is to build a base machine learning algorithm to begin testing with the polyol silicate complexes. Given the complexity of these structures and the large amount of possibilities linearity cannot be relied upon. Therefore, neural networks are being looked at as the possible end goal method. Neural network methods are being considered because of their ability to process substantial amounts of data input and recognize patterns or interpret the data, distributing it into different classes or categories. The algorithm may be able to recognize similarities and eventually predict outcomes based off the past training data fed into the program. The input and output data can also be clustered based on chosen properties and characteristics. With enough of these clusters and training, unlabeled data may be used as inputs to generate predictions and even identify unknowns. This research is still at an early stage, therefore the specific method that will be used is still undetermined.
Acknowledgments
I wish to acknowledge my supervisor, Dr. R. Mawhinney for their support and guidance as well as K. Lefracois-Gagnon, A. Smith and A. Degrazia for their previous research that made this project possible. I would also like to acknowledge Compute Canada for computational resources and NSERC for funding.

References
[1] A. Smith. Assessing and characterizing the inductive effect through silicon containing backbones and on silicon reactivity Lakehead University, Knowledge Commons 2014
[8] A. Degrazia Do QTAIM substituent dipole moment descriptors account for the inductive effect in larger bicyclo-type systems? Lakehead University 2021