COMP 562 Final Report

Matthew Hart   Austin Snyder

12/7/2021

1 Introduction and motivation

Transition-Metal Complexes (TMCs) are molecules consisting of one or more transition-metal centers surrounded by organic molecules (known as ligands), bound to the metal atom by dative bonds (fig 1). These molecules are of interest to researchers because of their potential to mimic the efficiency and specificity of catalytic biological molecules, such as enzymes. However, engineering these molecules for use in catalysis has proven to be difficult, owing to the large combinatorial space of of transition metals and ligands, as well as lack of design rules relating molecular structure to physical mechanisms of catalysts. Simulation methods (such as DFT calculations) are oftentimes used as a means of finding the electronic properties of these catalysts, which are important for determining the physical mechanisms of catalysis. However, these methods are computationally expensive and can take upwards of one month running on large clusters for complicated systems. Machine learning has emerged in the chemical sciences as a method for producing models of complicated systems with large amounts of variables. While these methods have been extensively applied to areas such as drug discovery, there have been little applications to the discovery and modelling of TMCs. Here, we use simple machine learning methods to predict the metal-center charges of TMCs based off structural features of the molecules. The use of these methods could be easily expanded to other electronic properties with the goal of creating models capable of predicting complete electronic profiles of TMCs.

1.1 Application of Machine Learning

Machine learning methods have been consistently shown to be useful for problems meeting a few main criteria. The problems must exist in a very large combinatorial space of variables, the problem should be large and difficult for humans to understand (black-box modelling should be avoided if possible), and there must be a large amount of data associated with said problem. The problem of predicting electronic properties of transition metal complexes is suitable for the application of machine learning because there is a large corpus of structural and DFT data associated with TMCs, the relationships between TMC structure and electronic structure is difficult to grasp within a few terms of explanation, and there is a very large amount of possible combinations of transition metals and ligands (the number of small molecules is estimated to be over 10^{60}). The prediction of electronic properties of TMCs is therefore an appropriate problem for machine learning. Furthermore, the molecules in question are able to be featurized into large-dimensional spaces via popular cheminformatics tool kits, such as RDKit, which give our models a larger space to explore and learn from.

2 Preprocessing and Curation

The TMqm dataset is a large (80,000) data set of crystallographic TMC structures collected from the Cambridge Structural Database(source) . Balcells et al(Ref.1) complied this large set of structural data and performed the DFT calculations necessary to compute the electronic properties of the TMCs (dispersion energy, molecule charge, HOMO, LUMO, etc). For initial curation, we downloaded the dataset and converted the molecular representations from their native .xyz (Cartesian coordinate) formats into Simplified Molecular-Input Line-Entry System (SMILES) format using the OpenBabel cheminformatics tookit. This step was necessary for a number of reasons. First, the .xyz files are much larger than SMILES representations, and smaller molecular representations formats will be more useful is speeding up computational efficiency. Second, .xyz file formats do not represent molecular bonds explicitly; this is a problem when trying to curate data about the electronic structure of bonds within molecules. Upon conversion from .xyz to SMILES, several thousand molecules of the original data set were fragmented or otherwise erroneous due to a bug in the OpenBabel package and were eliminated from our modelling data set. SMILES strings containing "@" (indicating the chirality of a molecule) were eliminated so that there would not be problems with enantiomers in our data set. Duplicate analysis was performed to balance the dataset. RDKit boasts a known bug of failing to import SMILES strings as chemical objects if the resulting chemical graph is highly interconnected, which TMCs generally are. Therefore, as a last step in curation, we eliminated all molecules which resulted in highly connected chemical graphs. Our final modelling data set consisted of around 14,000 molecules.

2.1 Feature Processing

For our chemical features, we employed a widely used set of cheminformatics descriptors. These features were calculated through the RDKit package and included: the exact molecular weight of the molecule, Morgan fingerprints, the heavy-atom molecular weight, the total molecular weight, the number of radical electrons, the number of valence electrons, and a set of 16 mathematical descriptors of the chemical graph of the molecule. These features were selected because they did not include any of our modelling endpoints (electronic properties) and were based on the structure and stoichiometry of the molecule alone. A total of 23 structural descriptors were calculated for 14,000 molecules. The features and modelling endpoints were structured into a final modelling matrix and used for training, test, and validation sets.

2.2 Assumptions and Shortcomings

The key assumptions of this data set are that initial data collected from the TMqm data set is composed of unbiased and accurate electronic feature calculations. We have also assumed that the creation of different data sets for each kind of metal center involved in a TMC is unnecessary. Another assumption we made is that the structure of the molecules tested would be the key determinate of their electronic properties. One of the shortcomings of this study was our reliance on cheminformatics tool kits for both featurization and molecular representation formats, which severely cut down on the size of our modelling data set

3 Models & Results

In order to generate models, 80% of the 14,000-molecule dataset was randomly selected for training, while the other 20% was left for eventually testing the models. This means that the training set contained 8,456 molecules with 23 structural descriptors for each, while the testing dataset was left with 2,819 molecules. Additionally, 25% of the training set was randomly selected for inclusion in the validation set, resulting in 2,819 molecules. Models were then generated using a variety of standard methodologies, including ridge regression, AdaBoost regression, gradient boosting regression, random forest regression, extra trees regression, K-Nearest Neighbor regression, support vector regression (SVR), and linear support vector regression (LSVR).

The results of each of these regression analyses are included in Table 1 below.

Regression Methodology	**R^2**	**Fit Time (s)**
Ridge	0.75	0.020
AdaBoost	0.64	2.558
Gradient Boosting	0.77	7.433
Random Forest	0.83	21.138
Extra Trees	0.84	6.897
SVR	0.37	6.070
LSVR	0.63	1.460

Ridge regression also performed relatively well due to the high correlation across considered independent variables. Generic gradient boosting performed relatively well, but the special case of AdaBoost was not able to perform similarly. Support vector regression and linear support vector regression both performed poorly due to the training data being significantly larger than the number of features.

Extra trees and random forest regression were found to be the best-performing regression methodologies in this instance, but extra trees was found to be significantly faster. This is likely due to how each regression model chooses their split points during decision tree creation. While random forest regression spends a majority of its computation time identifying the optimal split point, extra trees chooses it randomly. Because a key goal of this report is to identify a time-saving methodology for accurate identification of molecular metal charges, extra trees regression is the best-performing methodology between the two.

Additionally, like other models used, extra trees had a reasonable time complexity with comparison to current, non-machine learning methods of calculation. The training data was then fit to this best-performing model and evaluated for accuracy, mean absolute error, and root mean square error.

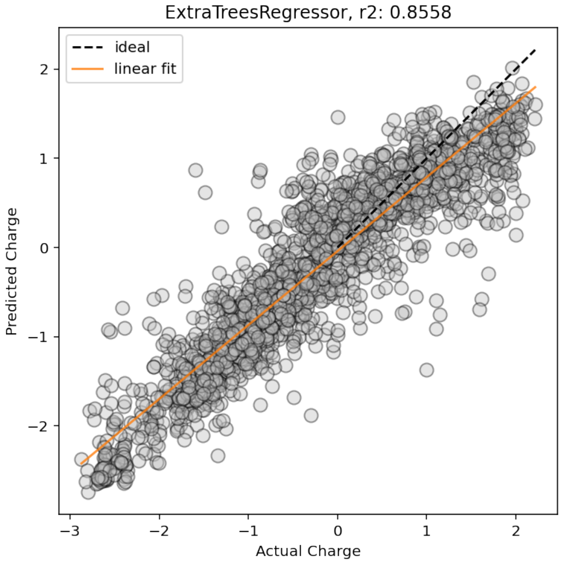

Figure 1: Regression plot of predicted metal charge as a function of actual metal charge, with prediction data sourced from the extra trees regression methodology. The R^2 is .86, indicating a good fit.

4 Conclusions

In this work, we have presented a simple machine learning model capable of predicting a single electronic property, metal charge, with relative accuracy. We employed simple structural chemical feature along with common machine learning models to predict an electronic property in seconds, compared to several hours for explicit calculations. Due to the models assumptions and shortcomings, further work is needed to produce a model that can complete the electronic profile of a TMC. Improvements to this model will include the prediction of the HOMO/LUMO gap and the conformational energy of the molecule. We will also utilize the entirety of the original data set by including the molecules that could not be read by RDKit. Furthermore, we will search for other molecular representation platforms that make up for the shortcomings of SMILES when applied to TMCs. Finally, we will use experimental data around TMCs and catalysis to relate electronic properties to catalytic mechanisms.

5 References

1. tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes David Balcells and Bastian Bjerkem Skjelstad Journal of Chemical Information and Modeling 2020 60 (12), 6135-6146 DOI: 10.1021/acs.jcim.0c01041