Implementation of Machine Learning Models to Predict Chemical Properties

Ruoshan Dong, ddong1@hwemail.com
Harvard-Westlake School, Class of 2021
USC Viterbi Department of Chemical Engineering, SHINE 2018

Introduction

Machine learning application for chemical properties prediction

- Density Functional Theory (DFT) is widely used in the field of quantum chemistry because of its high accuracy and reasonable computational cost. However, in recent years, machine learning models have been proved to reduce computational cost while maintaining an acceptable level of accuracy.

Bioinspired Dicopper Complexes Screening by DFT for Methane Hydroxylation:

- Dicopper Complexes inspired by pMMO is promising to catalyze methane activation at low temperature and pressure.

Computational Methods

Computational Methods:

- **Dataset**: **QM9 Google Database** (rotational constant(GHz), dipole moment(Debye), isotropic polarizability(Bohr^3) ,energy of HOMO(Ha), etc.)
- **Model**: **Linear Regression,** Logistic Regression, Support Vector Machine (SVM), Regression Tree, etc.)

$$\hat{y}(w, x) = w_0 + w_1 x_1 + ... + w_p x_p$$
$$\min_w ||Xw - y||_2^2$$

- **Representation**: **Fingerprint Molecule**

Skills Learned

Machine Learning:

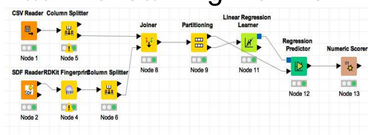
- Basic ideas and principles
- Different models (ex: linear regression, , logistic regression, etc.)
- Optimization (regularization, cross validation)

Programming in Python:

- Use Spyder program in Anaconda

KNIME

- Build machine learning workflow



KNIME Workflow, screenshot

Next Steps and Advice for Future SHINE Students

Advice for Future Students:

- I was worried about being one of the only sophomores in the program, but quickly grew to realize that age doesn't even matter, because we are all here for the same purpose.
- Don't be afraid of not knowing enough, you are here to learn!
- Keep in touch with the SHINE cohort and your mentors.

Future Steps:

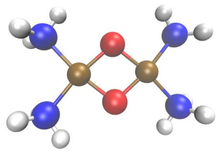
- Continue programming and learning more about machine learning algorithms
- Enter project into science fair
- Pursue more research programs

Objective & Impact of Professor's Research

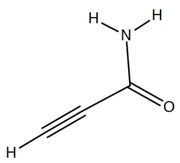
Professor Sharada is implementing state-of-the-art machine learning techniques in combination with chemical research. Her main goals are to:

- Predict chemical properties more efficiently in a sensible, low-cost method
- Predict the connectivity between atoms based on previously trained datasets
- Comprehend the mechanisms of catalyzed reactions, thus enabling the design of new and high-performance catalysts.

The overall impact of this work is to allow chemical research to be done more efficiently, and design selective catalysts in order to save energy in the chemical industry, therefore contributing to the reduction of greenhouse gas levels in the environment.



Dicopper catalyst with NH_3 ligand



Molecule from QM9

Results and Discussion

Table 1: Linear Regression Mean Absolute Error (MAE) of Chemical Properties

Properties	C_v (cal/mol K^{-1})	μ (Debye)	α ($Bohr^3$)	HOMO (eV)	LUMO (eV)
LR MAE	1.6	0.84	3.65	0.27	0.54
Properties	ZPVE (eV)	U_0 (Ha)	U_{298} (Ha)	H_{298} (Ha)	G_{298} (Ha)
LR MAE	0.27	16.04	16.04	16.04	16.04

Table 2: Benchmark MAE[1]

Properties	C_v (cal/mol K^{-1})	μ (Debye)	α ($Bohr^3$)	HOMO (eV)	LUMO (eV)
DFT MAE	0.34	0.1	0.4	2.0	2.6
KRR MAE	1.84	0.49	4.17	0.124	0.133

- **Linear Regression vs DFT**
  Linear regression has greater accuracy when predicting HOMO and LUMO
- **Linear Regression vs Kernel Ridge Regression (KRR)**
  Linear regression favored over KRR when predicting C_v and α

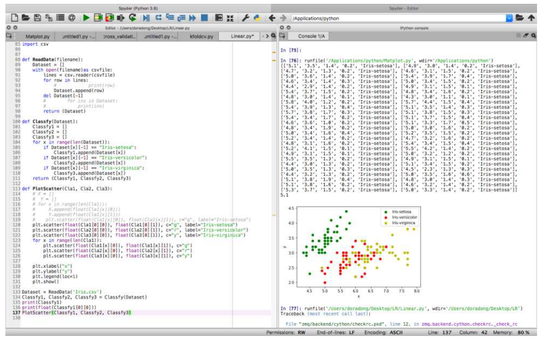
How This Relates to My STEM Coursework

Computer Science:

- I am currently learning Java at school, and the similarities between Python and Java will help me be more prepared to take AP Computer Science A

Chemistry:

- Knowledge of different chemical attributes (heat capacity, HOMO/LUMO) will help me with my future science courses



kNN algorithm coded in Python using Spyder, screenshot

Reference

1. Faber, F. A., Hutchison, L., Huang, B., Gilmer, J., Schoenholz, S. S., Dahl, G. E., ... & von Lilienfeld, O. A. (2017). Machine learning prediction errors better than DFT accuracy. *arXiv preprint arXiv:1702.05532*.

Acknowledgements

Thank you Professor Shaama Sharada, Zhenzhuo Lan, Dr.Katie Mills, Dr. Megan Herrold, and the SHINE cohort for making this opportunity possible, and for making my summer so amazing!