Predicting Atomization Energies of Molecules: A Machine Learning Approach

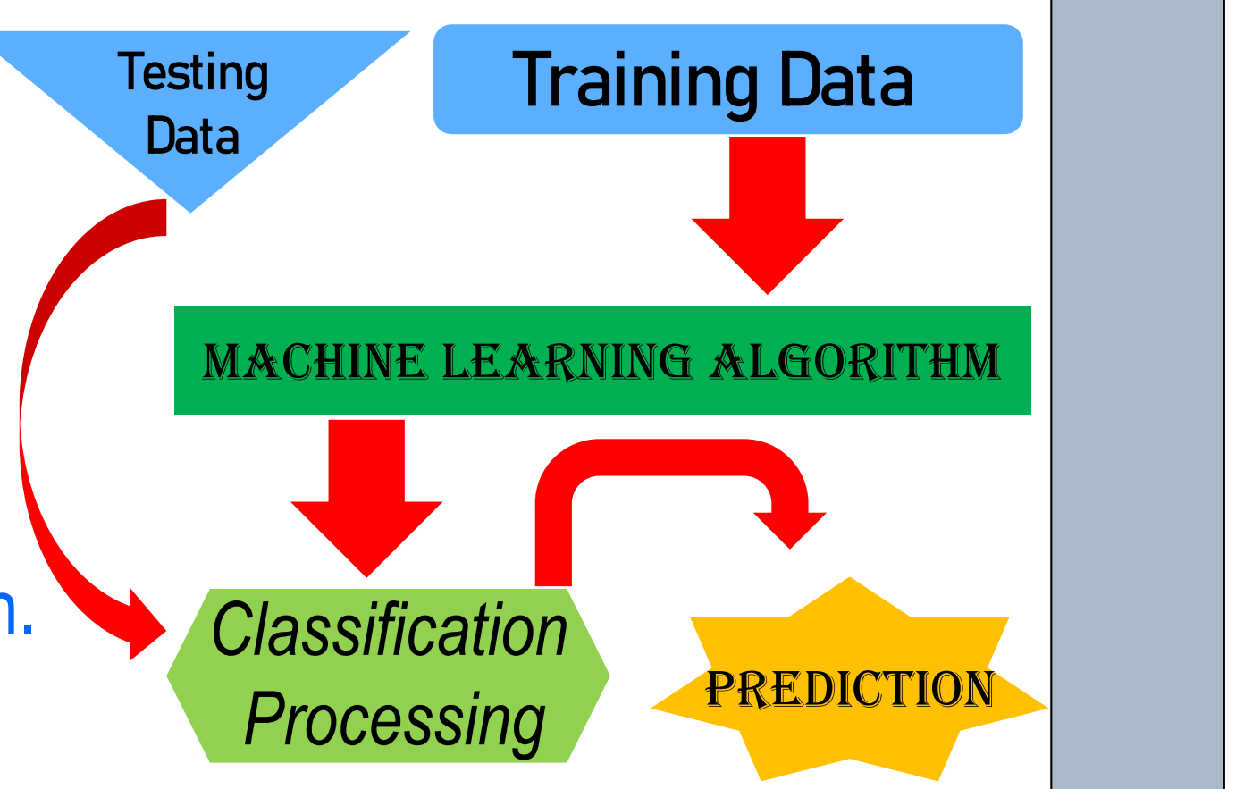
Saroj Upreti(Undergraduate Researcher) and Dr. Ben Dribus (Mentor).
William Carey University, Hattiesburg, MS.

CONVENTIONAL APPROACH

- Atomization Energies are normally predicted using Quantum Espresso Package.
- This package is based on Density Functional Theory.
- Simulations using this method are computationally expensive.
- Using machine learning method can be more effective.

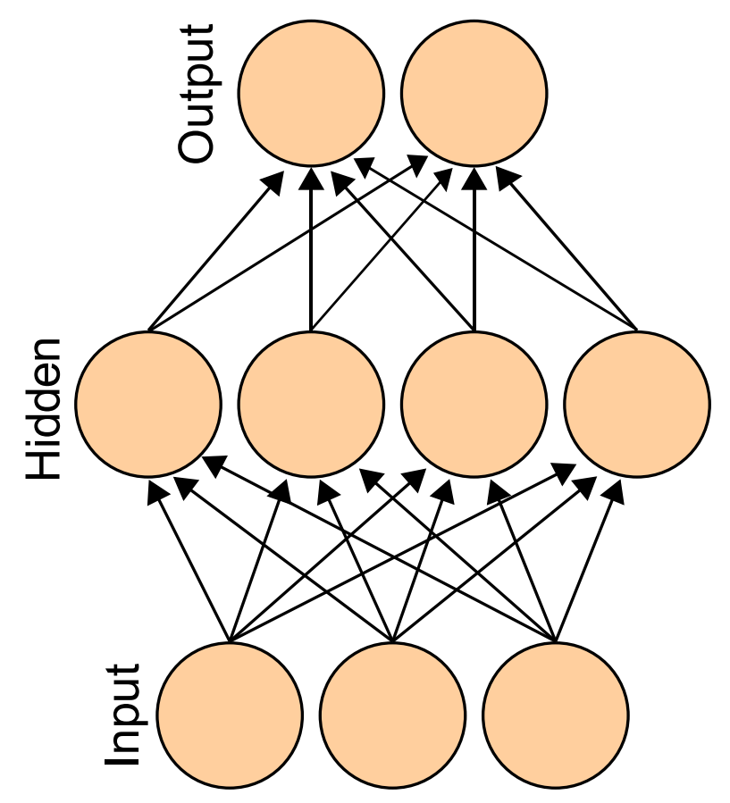
INTRODUCTION TO MACHINE LEARNING

- Data analysis method based on learning from data, identifying patterns and making decisions with minimal human intervention.



Diagrammatic Representation of Machine Learning

- In our case, we use a model of Machine Learning inspired by biological neural networks(of human brain) called Artificial Neural Network(ANN).



Diagrammatic Representation of ANN

PROJECT OUTLINE

- Taking atomic level simulations of 16242 molecules, we attempt to predict the value of Atomization Energy of those molecules.
- These simulations help us build 1275 molecular features (process explained later in the poster).
- Using those features, multiple trainings are carried out in 3 different kinds of neural network architecture, and the predicted results are compared among each other.
- Also, the predicted result is compared to the numbers provided by PubChem, a registered trademark of the National Library of Medicine.

DATA DESCRIPTION AND VISUALIZATION

- The molecules used are composed of subsets of elements C, N, O, H, P and S.
- Each molecule has at least 2 and at most 50 elements.
- The molecular details can be viewed with the help of their PubChem Id.

1,3-Dimethoxybutane

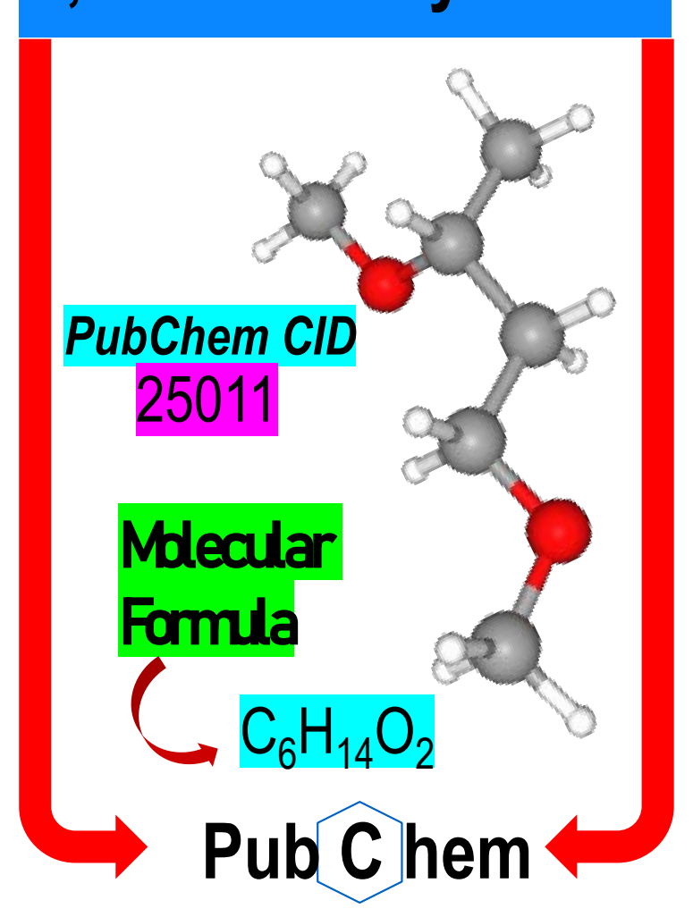


FIG: Sample 3D Structure of a Molecule Obtained with PubChem Id.

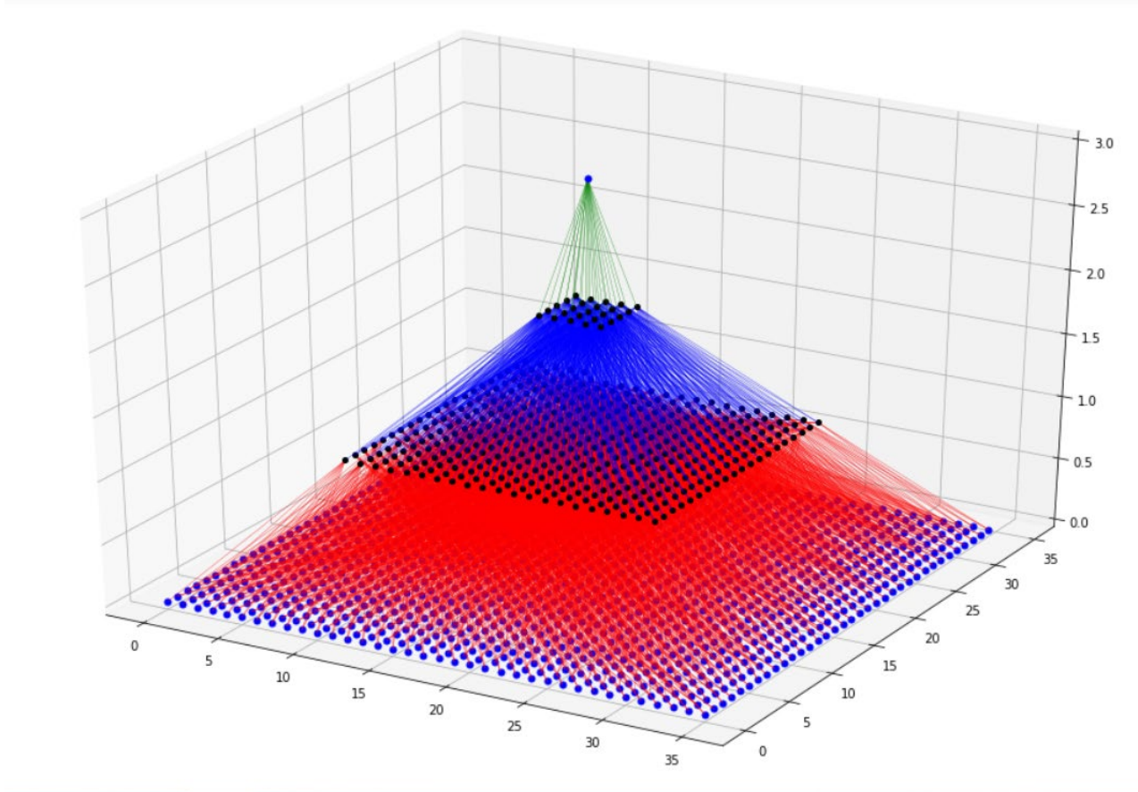
How are molecular features obtained?

- First of all, we construct 50x50 matrix C_{IJ} using intermolecular Coulomb repulsion operators which are defined as follows:

$$C_{IJ} = \begin{cases} 0.5\, Z_I^{2.4} & I = J \\ \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} & I \neq J \end{cases}$$

where Z_I are atomic numbers, R_I are atomic coordinates, and indices I, J run over the atoms in a given molecule.

- The off-diagonal terms refer to ionic repulsion between atoms I and J and the diagonal terms are obtained from a fit of the atomic numbers to the energies of isolated atoms.



The diagram in the left is a general visualization of a ANN architecture for our project. Later, we introduce randomness and weight sharing to make new architectures. The dataset we are dealing predicts a single value using 1275 input features, deep layers with 125 and 25 nodes, and a predicted value of output.

- All of these processes used to obtain features are based on Density Functional Theory, and hence, the data extraction part of our project relies on the theory.
- Molecules with less than 50 atoms have their Coulomb matrices appended by columns and rows of 0 to complete them to have dimensions of 50 × 50.
- From this 50 x 50 matrix for each molecule, the upper triangular part, or the lower triangular part(since, they are like mirror images to each other) provides a total of 1275 features.

All of these 1275 molecular features obtained are used as inputs, which goes into further processing. We have a total of 16242 molecules a fraction of which is used as training data. The remaining is used for testing purpose once the data is trained.

Processing here is carried out in various architectures which are as follows:

LOCAL ARCHITECTURE

- ❖ No randomness is introduced.
- ❖ Predicts least accurate results among 3.
- ❖ Average of minima error percent = 4.53%

LOCAL ARCHITECTURE WITH WEIGHT SHARING

- ❖ Weight sharing is introduced.
- ❖ Predicts better than Local Architecture.
- ❖ Average of minima error percent = 2.94%

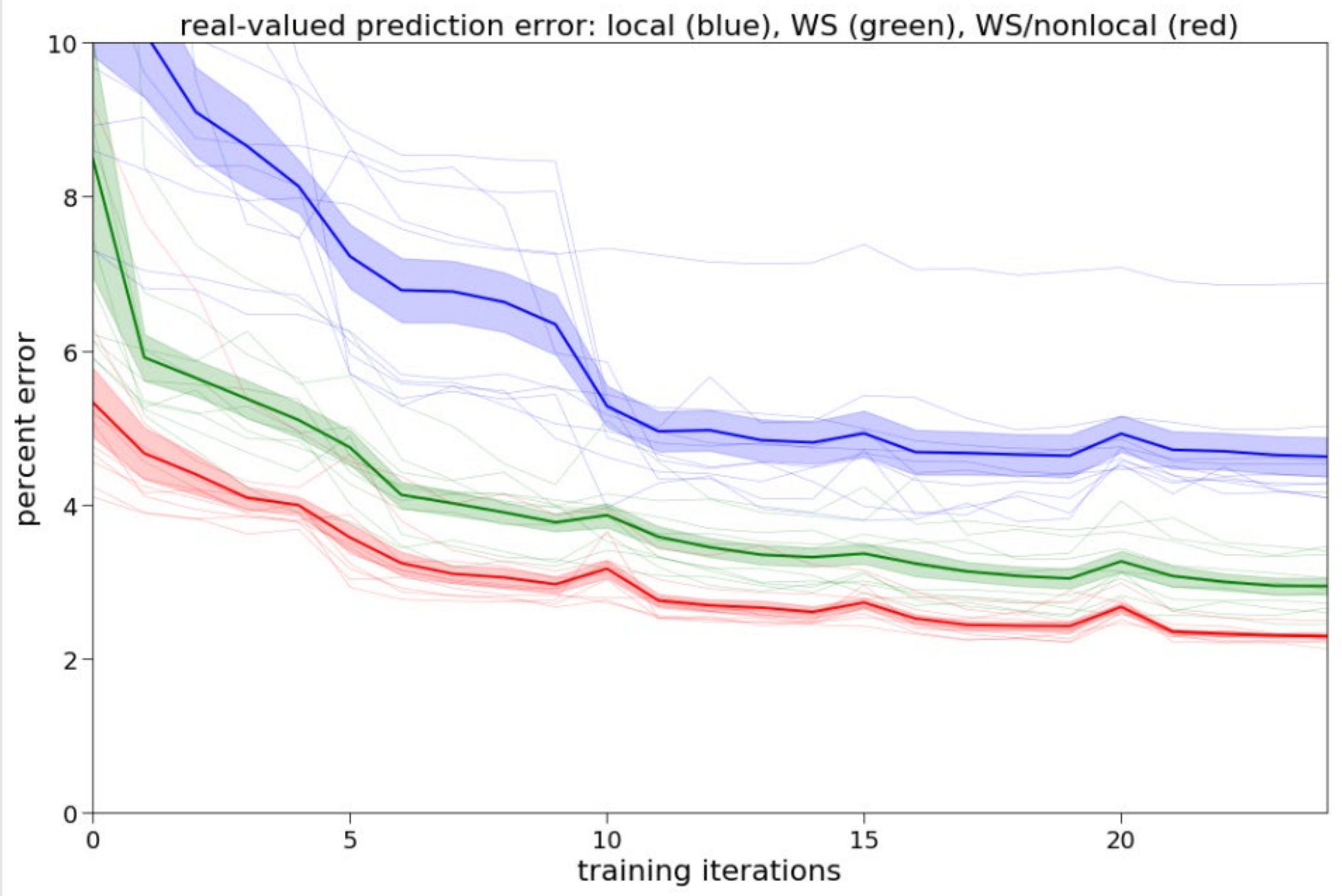


Fig: Percent Error Comparison Between 3 Architectures

NON LOCAL ARCHITECTURE WITH WEIGHT SHARING

- ❖ Weight Sharing and Randomness, both are introduced.
- ❖ Performs best of all of our architectures.
- ❖ Average of Minima Error Percent= 2.27%

For each architecture: A total of 25 iterations are run.
(5 epochs of 5 iterations each)
A total of 10 trials is averaged out.

CONCLUSION AND FUTURE RESEARCH

- Using non local architecture with weight sharing, an error of around 2% is obtained.
- Machine learning method of predicting Atomization Energy is computationally inexpensive and less time consuming than the conventional method.
- Our computing limitation did not allow us create more advanced architectures. With more powerful computing capability, we might get up to more than 99% accurate in the predictions.
- In the near future, our first goal would be reduce this percent error to as low as possible.
- In the recent years, quantum chemistry simulations have proven to be great alternatives in solving complex chemistry. Deep learning method might solve the Schrödinger equation for the electrons around atoms by finding their wave functions almost exactly.
- With our architecture, and the theories/codes behind them, a wide range of machine learning applications in the field of chemistry can be explored in the future.

REFERENCES

Ameya Prabhu, Girish Varma, Anoop Namboodiri. Deep Expander Networks: Efficient Deep Networks from Graph Theory. Preprint, 2018.

Benjamin F. Dribus et al. Network Horizon Dynamics II: 3-Generation Case. Preprint, to appear.

Matthias Rupp et al. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. Preprint,2012

Charu C. Aggarwal. Neural Networks and Deep Learning. Springer, 2018.

CONTACT INFORMATION

Email : Supreti380337@student.wmcarey.edu
Phone: +1 (601) 447 7029