Molecular Charge **PREDICTION**

Machine Learning

GROUP **MEMBERS**

Kanchi Mohan Krishna	20CS10030
Lav Jharwal	20CS30031
Gangaram Sudewad	20CS30017

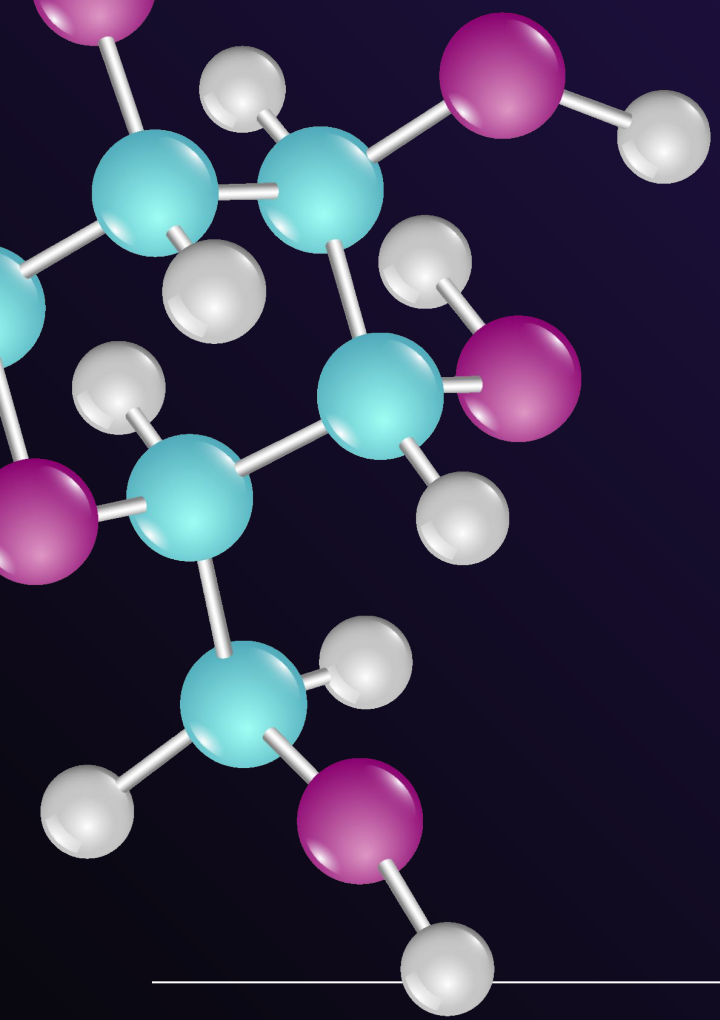
Problem Statement

Determining the Partial Charges of the Molecule from the 3D Structure using any Computational Method

What is an Atom Pair Fingerprint ?

An atom pair fingerprint is used to encode the structural information of a molecule. It is based on the frequency of occurrences of atom pairs, where an atom pair is defined as a unique combination of two atoms and the bond that connects them.

We encode this information in a binary vector format, it is possible to compare and classify different molecules based on their structural similarity.



Atom Pair Fingerprint

Descriptors are generated using atom pair fingerprint which involve counting the occurrences of atom pairs in a molecule and applying a hashing function to generate a fixed-size vector that represents the molecule.

In order to use the information generated by the fingerprints as an input to a Machine Learning model , we need to convert the bit vector representation of the fingerprint into a numerical representation. We have converted this information into an Array.

Random Forest Regressor

In a random forest regressor, multiple decision trees are constructed independently, each using a random subset of the available features and training data. The randomness introduced by using a subset of features and data is intended to reduce overfitting and improve the generalization performance of the model.

The attributes in the decision trees of the RFR are based on the descriptors provided by the atom pair fingerprint function . The parameter tuning was done on the model to increase the accuracy.

The models are generated for the atoms: (H, C, N, O, F, P, S, Cl, Br, I)

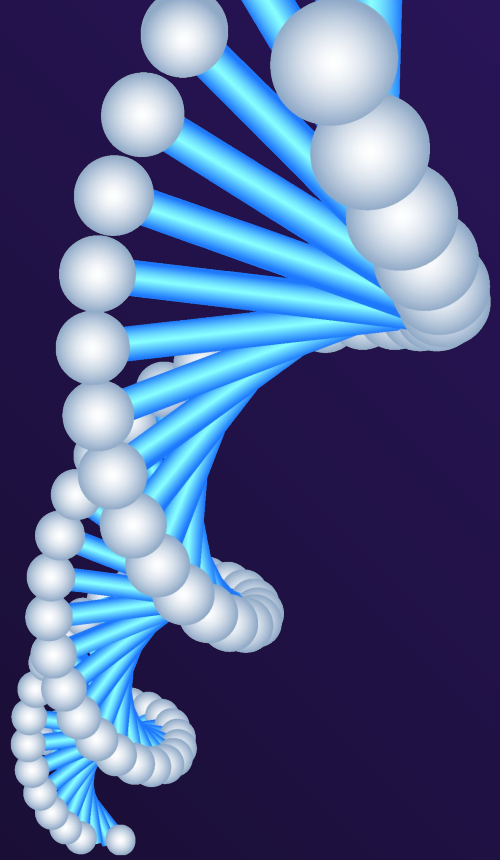
Neural Network

At a high level, a neural network consists of a large number of interconnected processing nodes, called neurons, which are organized into layers. The input layer receives the input data, which is then passed through one or more hidden layers of neurons before reaching the output layer.

We Implemented total of 3 layers (Input, Hidden, Output) , The number of nodes for each layer were decided during parameter tuning, For each layer a corresponding activation function was assigned.

During training, the network learns to adjust the weights of the connections between neurons in order to minimize a specific objective function, such as the mean squared error.

The models are generated for the atoms: (H, C, N, O, F, P, S, Cl, Br, I)



Suggestions and Improvements

Circular Fingerprint Method gives the better description of an atom for better prediction of charges. But the complexity of the Algorithm is high , We implemented a function using circular fingerprints but the algorithm takes more time to train the molecules. So there is a tradeoff between complexity and accuracy. So we decided to use the function from RDkit.

Neural Networks models gives better accuracy on charge prediction with relatively less amount of data on molecules. Whereas Random Forest models give better accuracy when data is abundant.

So , To predict better one way is to use Ensemble Learning methods to include both RFR and NN models to predict charges.

CONCLUSION

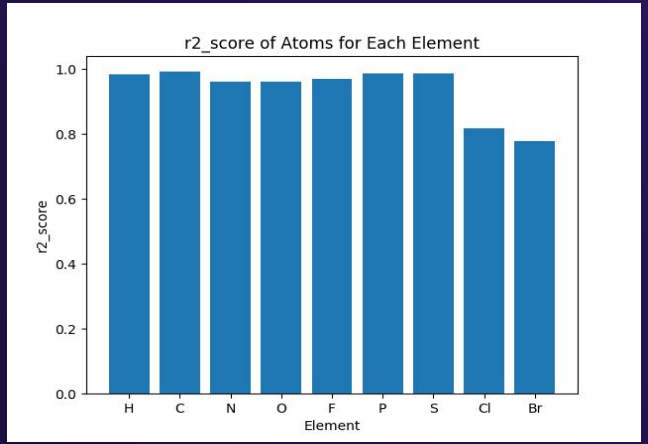
We Tried the Random Forest Regression & Neural Networks Separately.

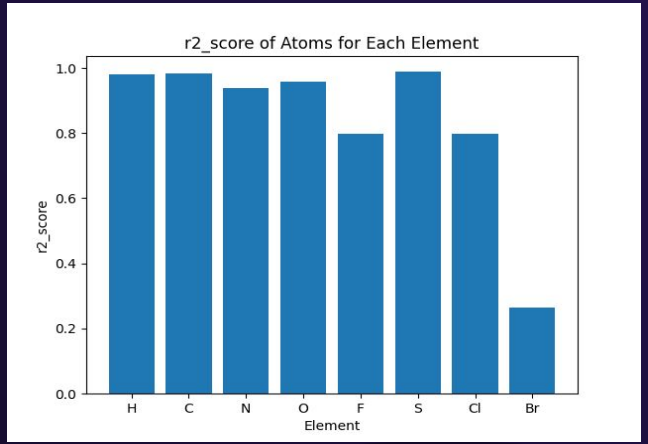
After experimenting we came to the conclusion that Neural Network gave better accuracy results.

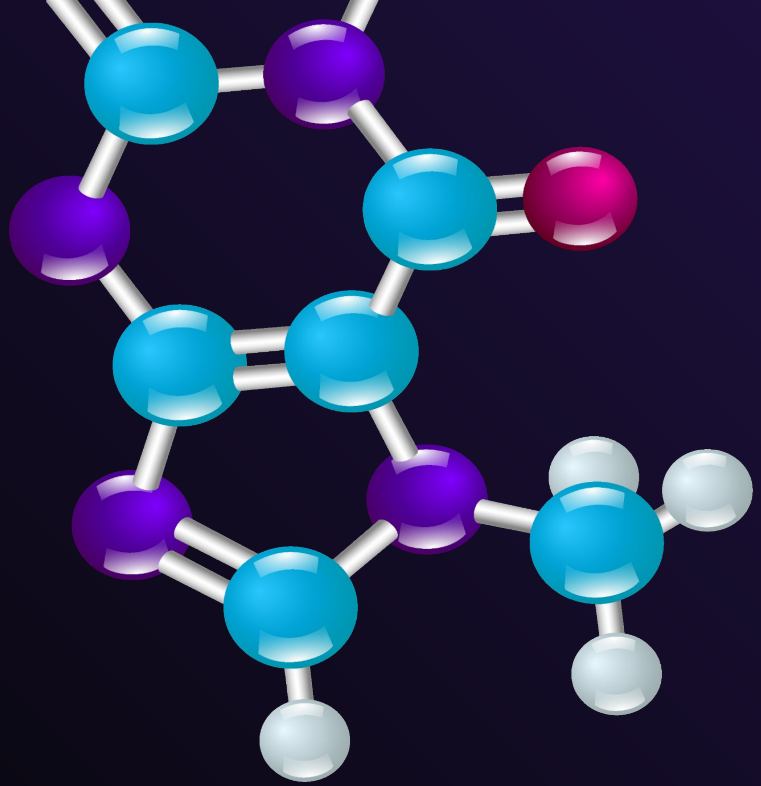
Here the 2 graphs shows us the R2 scores of each model of an atom after parameter tuning. (Dataset have been divided into 80% and 20% for training and testing)

We got an accuracy of nearly 97% when trained on a large dataset.

We also have created a program which takes sdf file as an input and predicts the charges of molecules based on the 3D coordinates using the models we trained earlier.







Thank You

References :
Dataset - https://www.research-collection.ethz.ch/handle/20.500.11850/230799