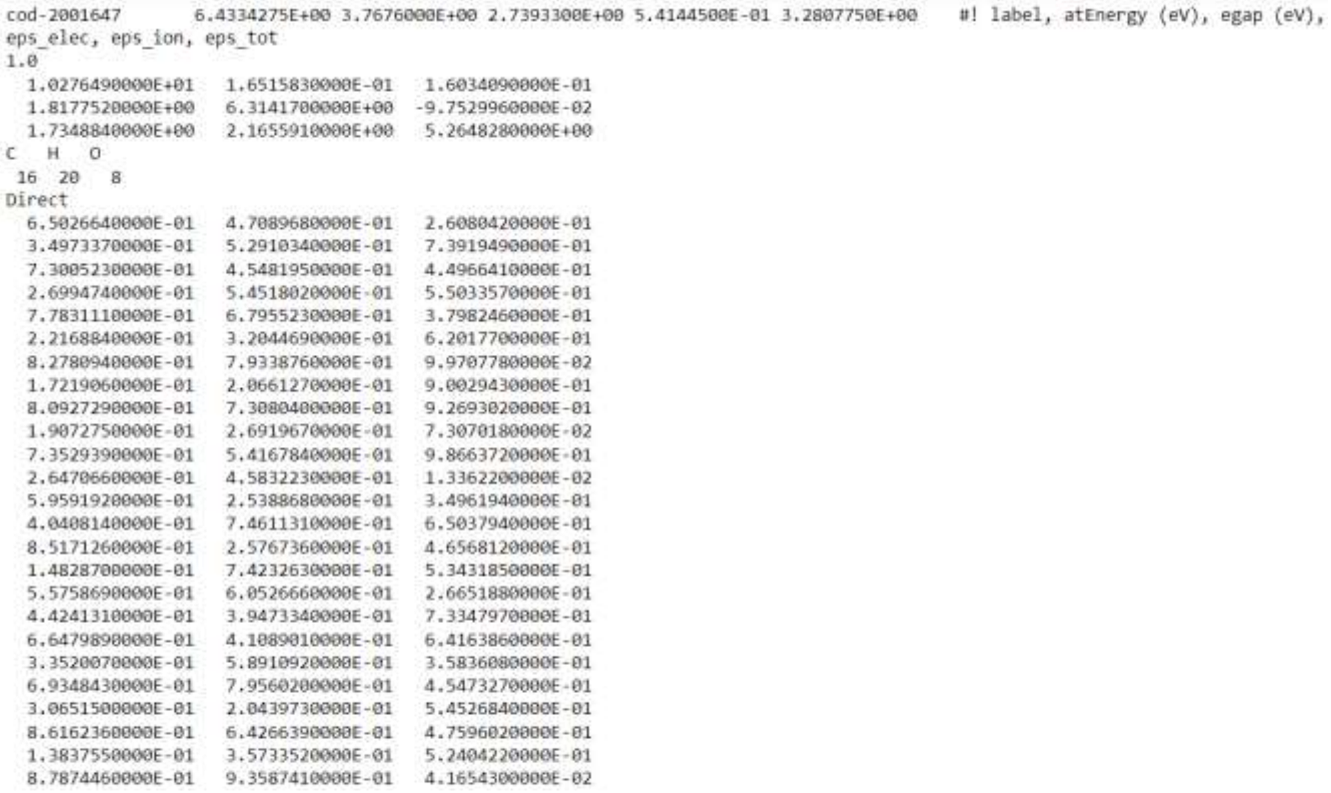
CH5650: Molecular Data Science and Informatics

Assignment-1 Report

Shrivarshan K, MM20B058

Data:

A crystal dataset of 215 optimized structures are provided. All of them are given in the POSCAR format. An example of the data of a structure is shown below.



```
cod-2001647       6.4334275E+00 3.7676000E+00 2.7393300E+00 5.4144500E-01 3.2807750E+00    #! label, atEnergy (eV), egap (eV),
eps_elec, eps_ion, eps_tot
1.0
  1.0276490000E+01   1.6515830000E-01   1.6034090000E-01
  1.8177520000E+00   6.3141700000E+00  -9.7529960000E-02
  1.7348840000E+00   2.1655910000E+00   5.2648280000E+00
C   H   O
 16  20   8
Direct
  6.5026640000E-01   4.7089680000E-01   2.6080420000E-01
  3.4973370000E-01   5.2910340000E-01   7.3919490000E-01
  7.3005230000E-01   4.5481950000E-01   4.4966410000E-01
  2.6994740000E-01   5.4518020000E-01   5.5033570000E-01
  7.7831110000E-01   6.7955230000E-01   3.7982460000E-01
  2.2168840000E-01   3.2044690000E-01   6.2017700000E-01
  8.2780940000E-01   7.9338760000E-01   9.9707780000E-02
  1.7219060000E-01   2.0661270000E-01   9.0029430000E-01
  8.0927290000E-01   7.3080400000E-01   9.2693020000E-01
  1.9072750000E-01   2.6919670000E-01   7.3070180000E-02
  7.3529390000E-01   5.4167840000E-01   9.8663720000E-01
  2.6470660000E-01   4.5832230000E-01   1.3362200000E-02
  5.9591920000E-01   2.5388680000E-01   3.4961940000E-01
  4.0408140000E-01   7.4611310000E-01   6.5037940000E-01
  8.5171260000E-01   2.5767360000E-01   4.6568120000E-01
  1.4828700000E-01   7.4232630000E-01   5.3431850000E-01
  5.5758690000E-01   6.0526660000E-01   2.6651880000E-01
  4.4241310000E-01   3.9473340000E-01   7.3347970000E-01
  6.6479890000E-01   4.1089010000E-01   6.4163860000E-01
  3.3520070000E-01   5.8910920000E-01   3.5836080000E-01
  6.9348430000E-01   7.9560200000E-01   4.5473270000E-01
  3.0651500000E-01   2.0439730000E-01   5.4526840000E-01
  8.6162360000E-01   6.4266390000E-01   4.7596020000E-01
  1.3837550000E-01   3.5733520000E-01   5.2404220000E-01
  8.7874460000E-01   9.3587410000E-01   4.1654300000E-02
```

In the above data, the first line contains the information about atomization energy, energy band gap, electronic part of dielectric constant, ionic part of dielectric constant, total dielectric constant of a structure. The 8th line contains the information of number of carbon, hydrogen and oxygen atoms. The remaining information are predominantly the coordinates of the atomic positions in each of the molecular crystal.

The aim of the assignment is to develop one model for each physical property, which could use the information about number of C, H, O bonds and predict each of the physical properties listed above for all the 215 structures.

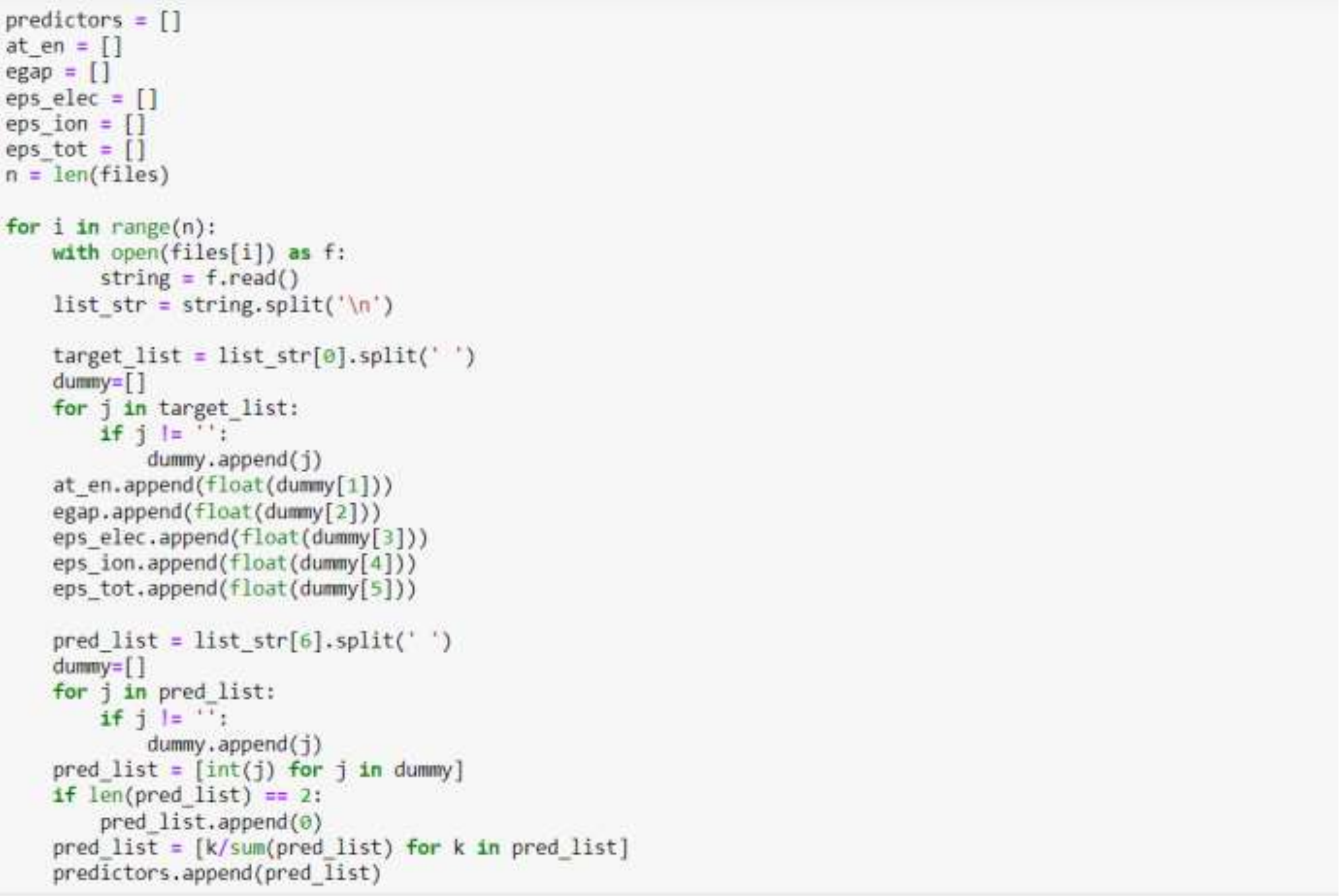
Data preprocessing:

The preprocessing of the data is very vital because the data given is in POSCAR format, which cannot be used as input for a deep learning model. So the numerical values should be separated out from the text file and can then be used to train the model. As discussed in the previous section, the first line of the data contains the information about the physical properties and the 8th line contains information about the number of C, H, O bonds.

So the data of each structure is loaded as string as the first and 8th line are taken separately for preprocessing. The numerical values present in those lines are taken separately and array of such values are created for each structure. This array which contains the number of C, H, O bonds is normalized using Standard Scaler function from sklearn.preprocessing module and a dataset of such arrays is created.

The dataset which contains the information of normalized number of C, H, O bonds is called the predictors dataset. The dimension of the dataset is 215 x 3 and each of the columns of this dataset is used as input values of the model to predict the physical properties.

The following is the code used for preprocessing,



```
predictors = []
at_en = []
egap = []
eps_elec = []
eps_ion = []
eps_tot = []
n = len(files)

for i in range(n):
    with open(files[i]) as f:
        string = f.read()
    list_str = string.split('\n')

    target_list = list_str[0].split(' ')
    dummy=[]
    for j in target_list:
        if j != '':
            dummy.append(j)
    at_en.append(float(dummy[1]))
    egap.append(float(dummy[2]))
    eps_elec.append(float(dummy[3]))
    eps_ion.append(float(dummy[4]))
    eps_tot.append(float(dummy[5]))

    pred_list = list_str[6].split(' ')
    dummy=[]
    for j in pred_list:
        if j != '':
            dummy.append(j)
    pred_list = [int(j) for j in dummy]
    if len(pred_list) == 2:
        pred_list.append(0)
    pred_list = [k/sum(pred_list) for k in pred_list]
    predictors.append(pred_list)
```

In the above code, the predictors list contains the normalized number of C, H, O bonds and at_en list contains the atomization energy value, egap list contains the energy gap value, eps_elec contains the electronic part of dielectric constant, eps_ion contains the ionic part of dielectric constant and eps_tot contains the total dielectric constant value of all the 215 structures.

Fingerprinting:

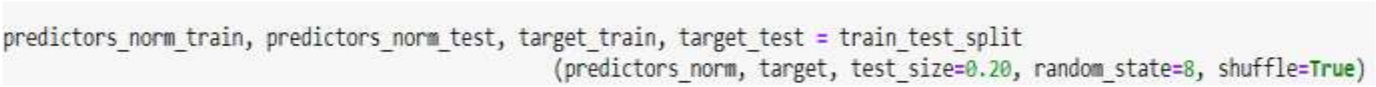
The fingerprinting technique used here is called motif-based fingerprinting. Motif-based fingerprinting is a technique used in bioinformatics to identify and compare patterns of short DNA or protein sequences, called motifs, within different biological sequences. The technique involves searching for specific motifs in a database of sequences and then generating a fingerprint, or summary, of the motif occurrences within each sequence.

Since only the number of C, H, O atoms data is available for each of the molecules, 0th order motif-based fingerprinting is used here. This technique necessarily uses the number of individual atoms as input values to predict the physical properties of the molecules.

Model building:

The deep leaning framework is used to predict the 5 physical properties of the molecules. There are two ways of building the model. One way is to build a single model which takes the number of C, H, O as input values and predict the 5 physical properties. But this is not a good approach because the model takes 3 input values and predicts 5 values for each molecule. A different way is adopted in which for each physical property a model is built, such that it takes 3 input values and predicts the physical property of all 215 molecules.

The KERAS platform is used to build the deep learning models. To predict each of the physical properties, 80% of the data is used for training the model and tested its performance on remaining 20% of the data. To split the data train_test_split function from sklearn.model_selection module is used. This function split the predictor dataset and the target data into train and test data.



```
predictors_norm_train, predictors_norm_test, target_train, target_test = train_test_split
                                                (predictors_norm, target, test_size=0.20, random_state=8, shuffle=True)
```

The optimizer used for all the 5 models is the popular Adam optimizer and loss function is mean-squared error. Each of the models is also validated during training on certain amount of data which varies within the models.

Each of the models are qualified based on a metric called R2 score which is also called as regression score. This score tells how close is the predicted target value to the actual target value. Each model has two R2 scores, train R2 score and test R2 score.

This metric depends a lot on the hyperparameters of the model. The hyperparameters of our models are, the number of hidden layers, number of neurons in each layer, activation function learning rate of the optimization function. These hyperparameters are turned such that a good R2 score is obtained for each of the models.

Atomization energy:

1. Hyperparameters:

Number of hidden layers: 1

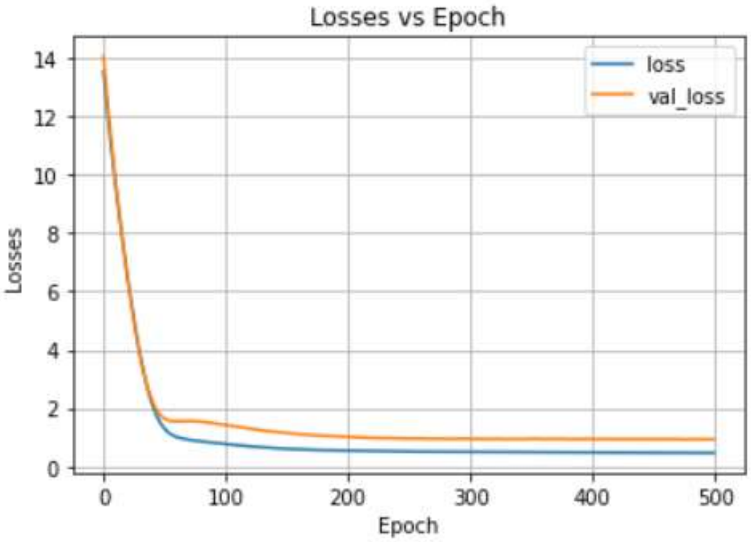
Number of neurons: 500

Fraction of dropout: 0.2

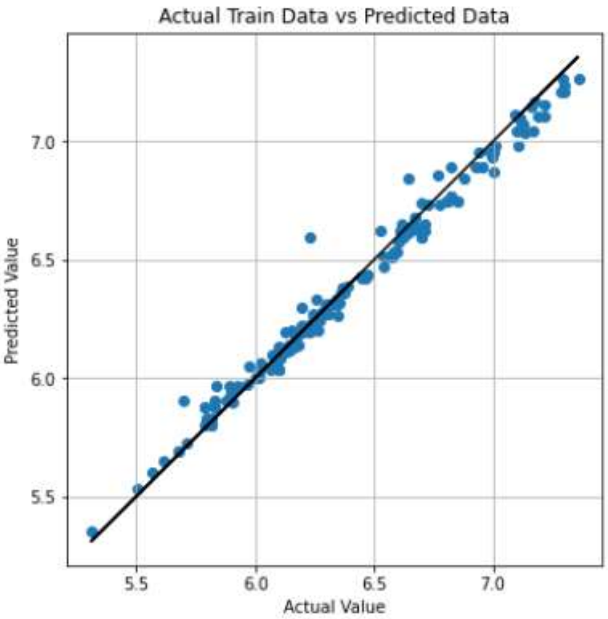
Activation function: ReLU

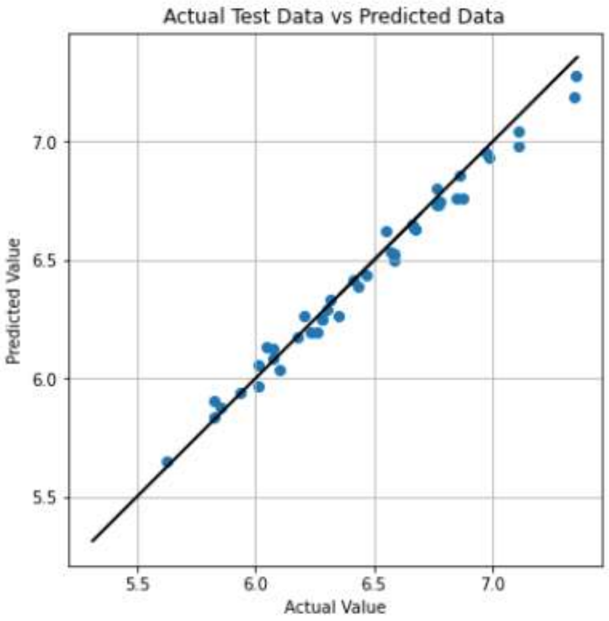
Learning rate: 1e-2

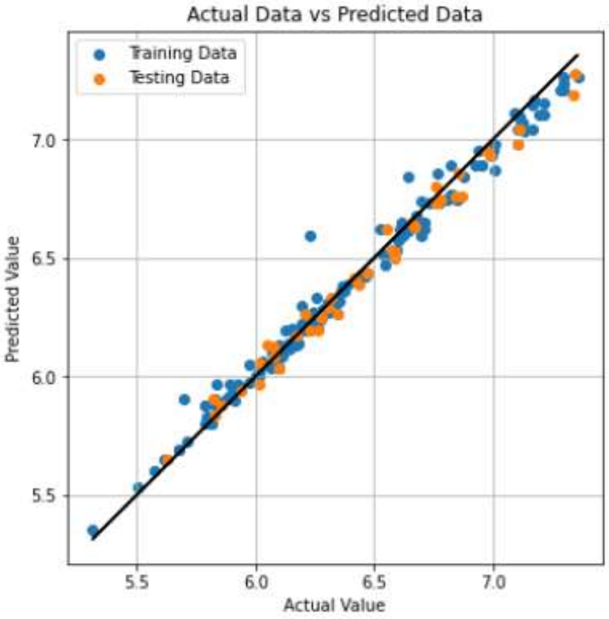
2. Loss vs Epoch plot:



3. Parity plots:







4. <u>R2 score</u>:

R2 score of train data = 0.9821081252552027
R2 score of test data = 0.9811872084186273

Energy gap:

1. **Hyperparameters**:

Number of hidden layers: 1

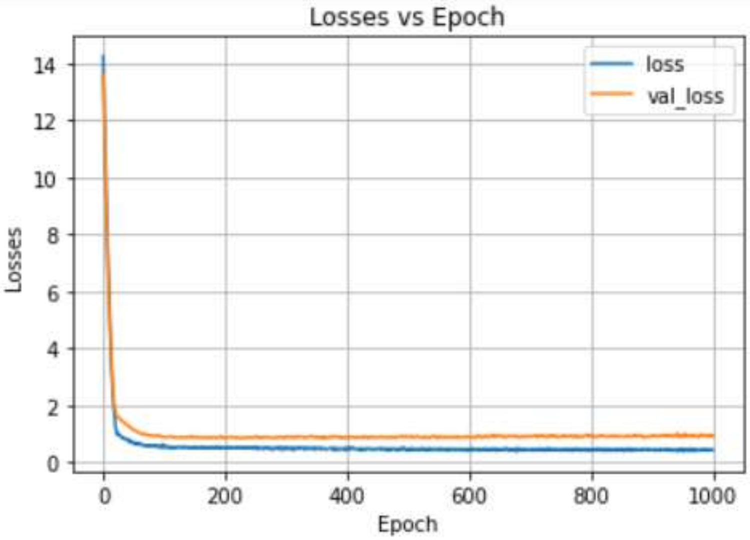
Number of neurons: 200

Fraction of dropout: 0.1

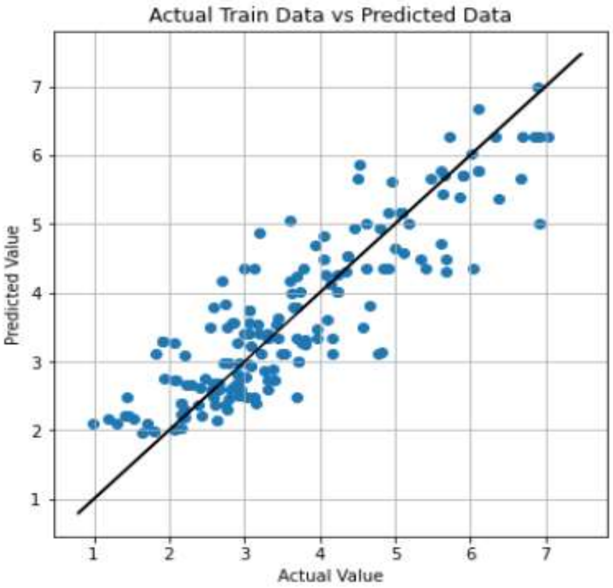
Activation function: ReLU

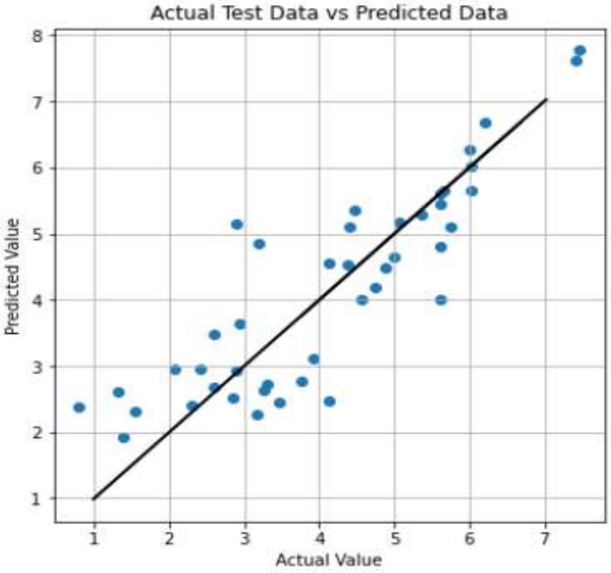
Learning rate: 1e-3

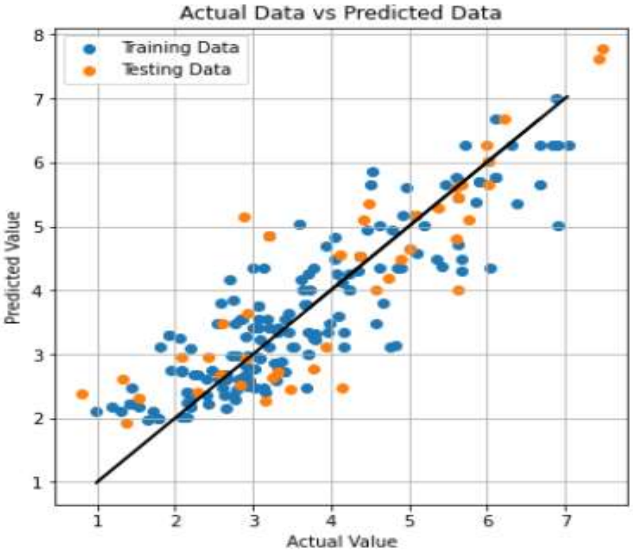
2. **Loss vs Epoch plot**:



3. **Parity plots**:







4. <u>R2 score</u>:

R2 score of train data = 0.7704008696949497
R2 score of test data = 0.745529827676009

Electronic part of dielectric constant:

1. Hyperparameters:

Number of hidden layers: 2

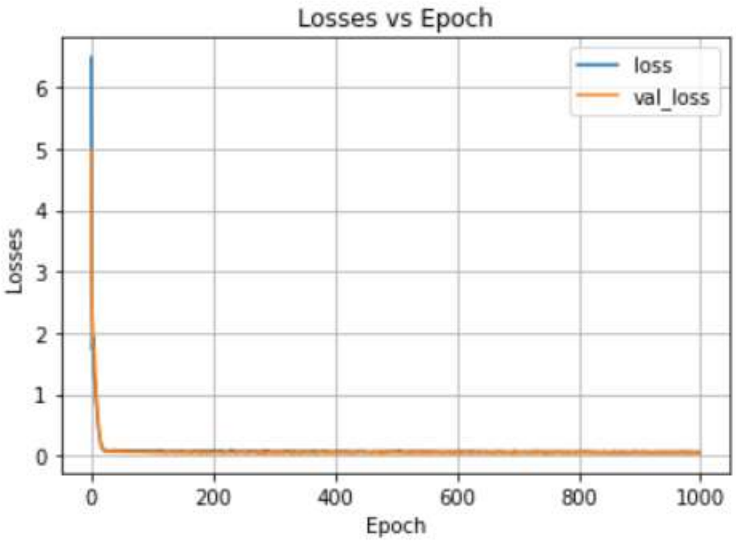
Number of neurons: 300

Fraction of dropout: 0.1

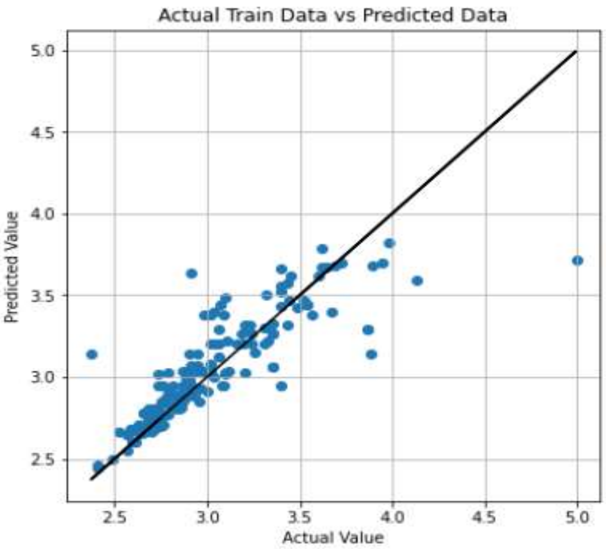
Activation function: ReLU

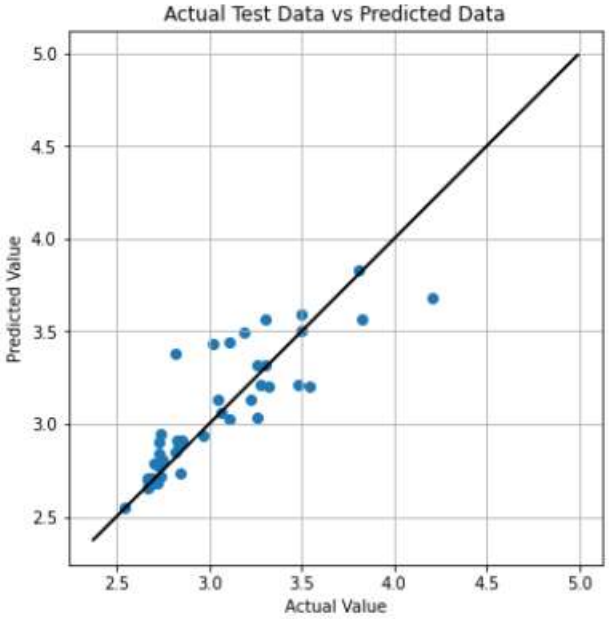
Learning rate: 1e-3

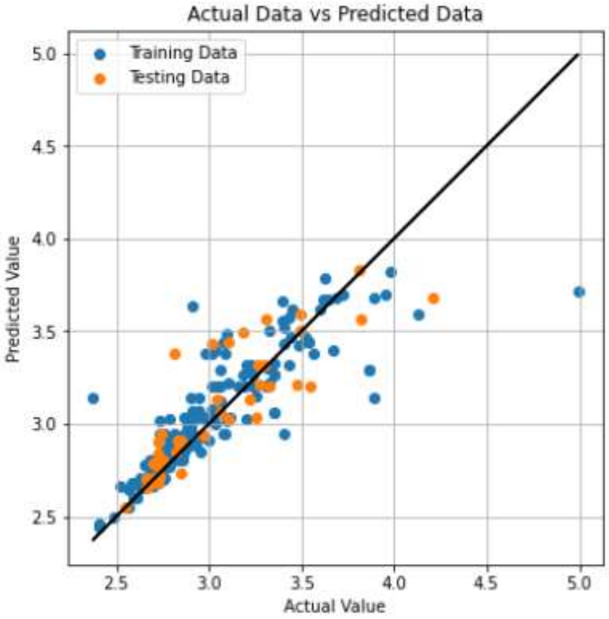
2. Loss vs Epoch plot:



3. Parity plots:







4. R2 score:

R2 score of train data = 0.7330840862259485
R2 score of test data = 0.7430867016195075

Ionic part of dielectric constant:

1. Hyperparameters:

Number of hidden layers: 5

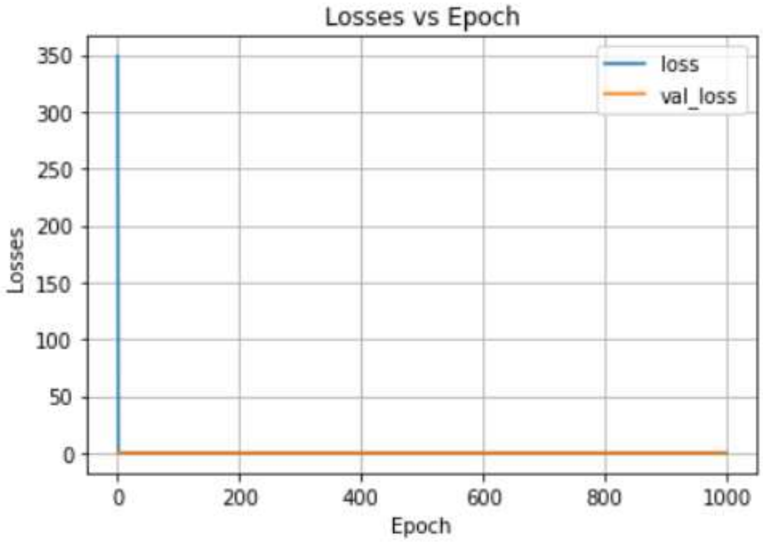
Number of neurons: 500

Fraction of dropout: 0.2

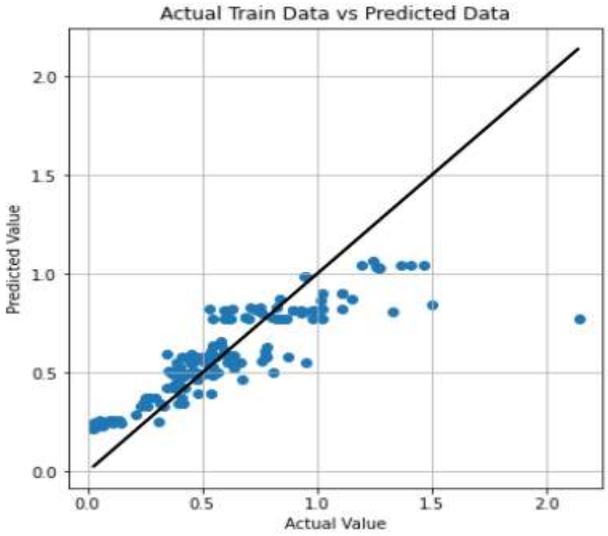
Activation function: ReLU

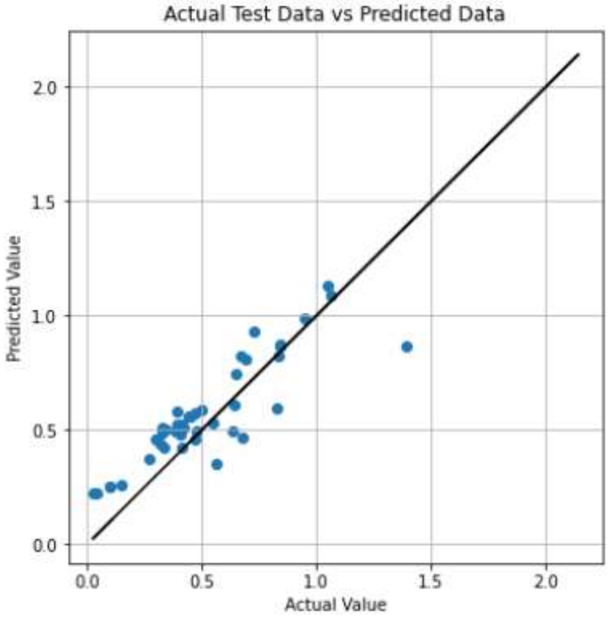
Learning rate: 1e-2

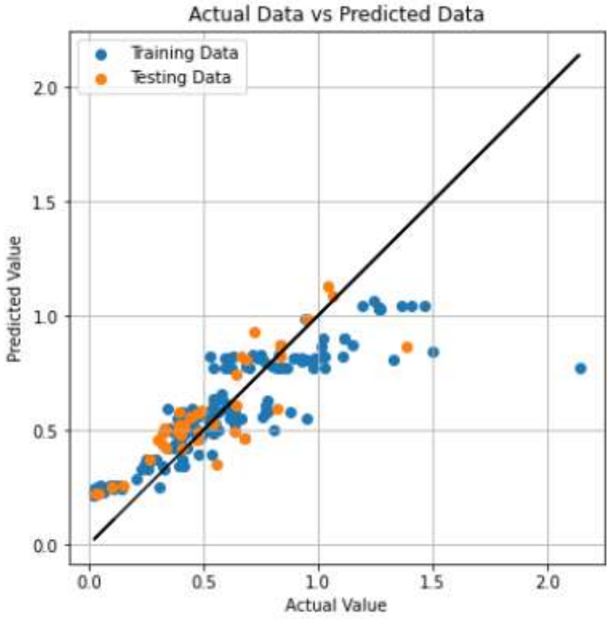
2. Loss vs Epoch plot:



3. Parity plots:







4. <u>R2 score</u>:

R2 score of train data = 0.7302927937050213
R2 score of test data = 0.7207677949173429

Total dielectric constant:

1. Hyperparameters:

Number of hidden layers: 2

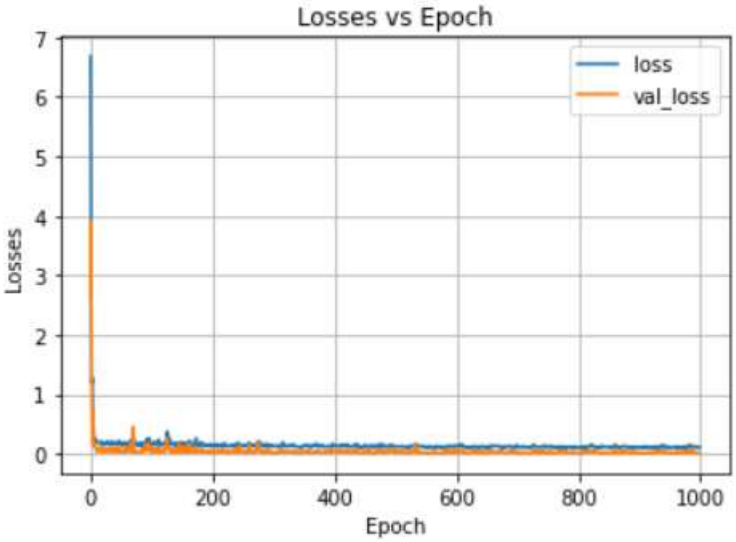
Number of neurons: 600

Fraction of dropout: 0.1

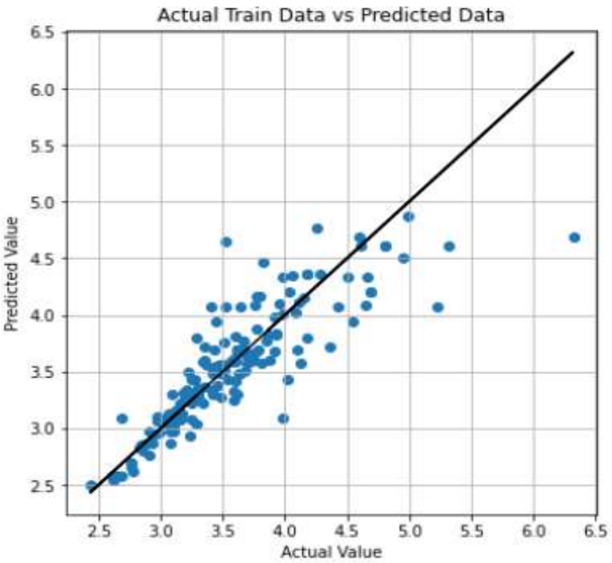
Activation function: ReLU

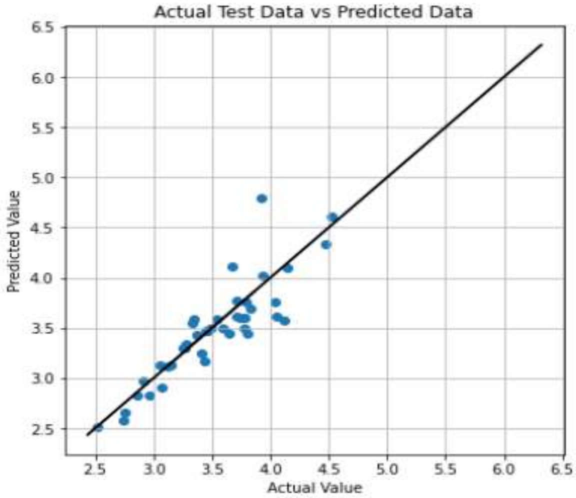
Learning rate: 1e-2

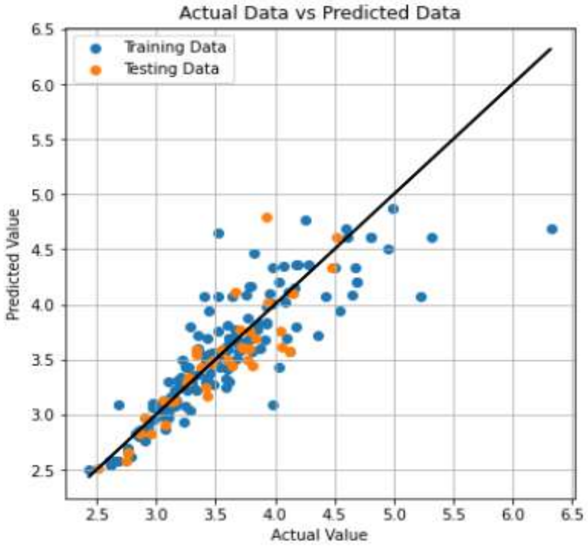
2. Loss vs Epoch plot:



3. Parity plots:







4. R2 score:

R2 score of train data = 0.7490626542723815
R2 score of test data = 0.7388639037259573

Conclusion:

The zeroth order fingerprinting works well for predicting atomization energy as it depends more on the number of atoms. But for predicting the remaining physical properties it seems that the zeroth order fingerprinting is not working well. This can be intuitively understood as these physical properties don't depend only on the number of C,H,O atoms of the molecules but also some other features of the molecules.