CH5650: Molecular Data Science and Informatics
Assignment#1
Maximum Marks – 15

Timeline: March 12, 2023

Data: A crystal dataset of 215 optimized structures are provided. All of them are given in the POSCAR format. The origin of the structures and calculated properties are given in the first line of each structure files as #! label, atEnergy (eV), egap (eV), eps_elec, eps_ion, eps_tot. The entries in the first line are as mentioned in the following.

"Label" is used to indicate the origin of the structures (A reference number)
"atEnergy": atomization energy, in eV
"egap" : energy band gap, in eV
"eps_elec": electronic part of the dielectric constant
"eps_ion" : ionic part of the dielectric constant
"eps_tot" : total dielectric constant, which is the sum of the electronic and the ionic parts

Problem Statement

Build machine learning models using the given data to predict atomization energy, energy band gap, electronic part of the dielectric constant, ionic part of the dielectric constant and total dielectric constant of a molecular crystal. Use 80% of the data for model building, and test the models' performance for the remaining 20% of the data. You can build 5 models each for a single property. Alternatively, you can build one single model that predict 5 properties. Write a details description of your method, fingerprinting and model building. Use motif-based fingerprints. Define a loss function or cost function for model building and draw the function during the training of models. Compare the actual property and predicted property for both the training and test set in graphs. Compare the performance for different orders of fingerprints. Submit your python script along with the report. For all cases, report the parity plots.