HW5 report

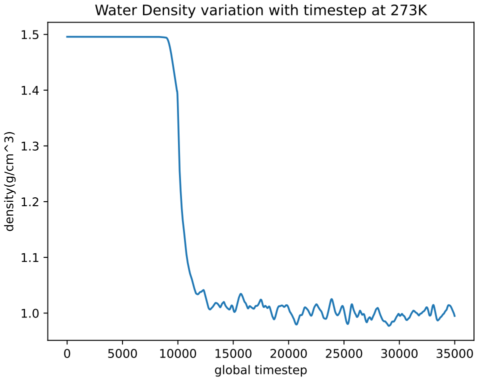
Han Xie

1 MOLECULAR DYNAMIC SIMULATIONS

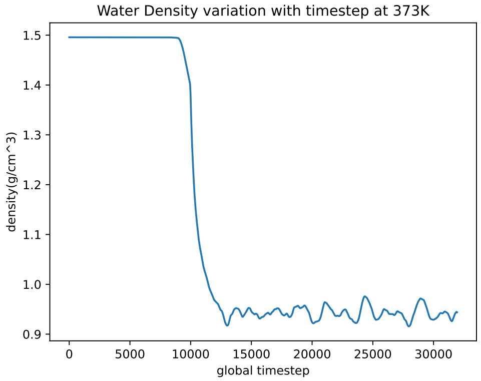
1.1 water density

a. The average water density from the NPT simulation is 1.001155. The value is roughly 0.42% larger than the density from the experiment. The difference is super small. The potential reason behind this is that the simulation assumes an NPT environment. However, in real life the pressure is not always constant. In addition, the SPC/E method might also be the reason why the value is a little bit larger since SPC/E includes a self-polarization energy correction to better produce the value for the enthalpy, but this might leads to an increase in density.

b. The plot of the water density over timestep is as below. From the plot, we could find that the density of water reach the equilibrium roughly after 20000 timesteps.

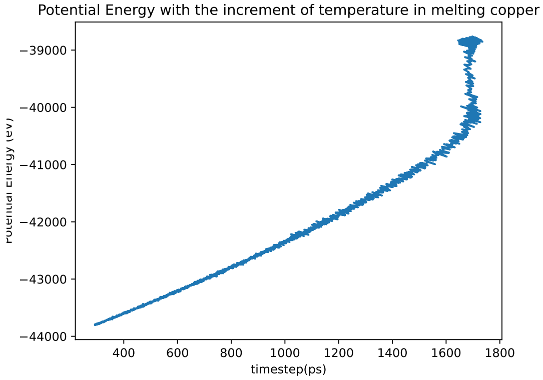


c. To compare 297 K with the 373 K case, we calculate the density of the 373 K case. The average density under the 373 K case is around 0.941937. The value is smaller than the 297 K case since the water vaporize at the temperature of 373K. Thus, gas will spread out more than the liquid. As the increment of volume, the density would decrease. However, the equilibrium time seems faster than the 273K case, this might also due to the greater distance between gas molecules would reduce the molecular interactions among them.

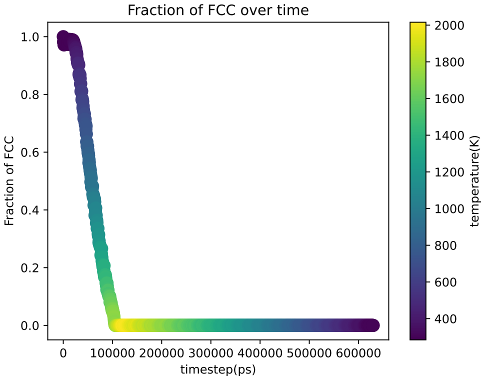


1.2 copper

a. From log.lammps, we plot potential energy (PE) as a function of temperature to identify the melting point (discontinuity in PE) is at around 1700K. Compare the simulated melting point with the experimental value, the value 1700K is much larger than the 1358K. This might due to the fact that the simulation is under the variation of NPT and NVT cases which might not be extremely constant under the experimental case.



b. From dump.cu.lammpstrj, I find that most of the atoms are in FCC structure, however, some of them are remaining disordered during the runing process in the beginning, with the melting and cooling process, there are more disordered atoms exist. There are also other types of structures such as type 3 and 4 (BCC or ICO). By plotting the FCC atom fraction as a function of temperature. We are able to observe structural loss during heating during cooling. And then the BCC and ICO structures appear during the cooling process, with the cooling ICO structures is more and more apparent. The reason that structures does not get back to FCC structure might due to some force setting issues.



c. By watching the visualization in the Ovito, I am getting roughly similar result. The three pictures above shows the structure at different steps of the simulation process. The black is atom at the structure FCC, white is atom at the structure of unknown (disordered), and the red is ICO structures. There are some BCC embeded in there but is unable to see directlly due to the small number and they are embeded in the box.

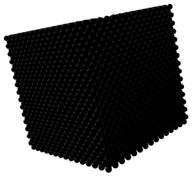


Figure 1: original

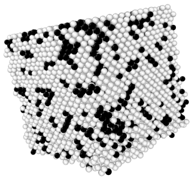


Figure 2: melting

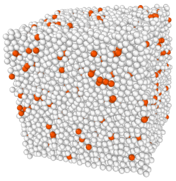


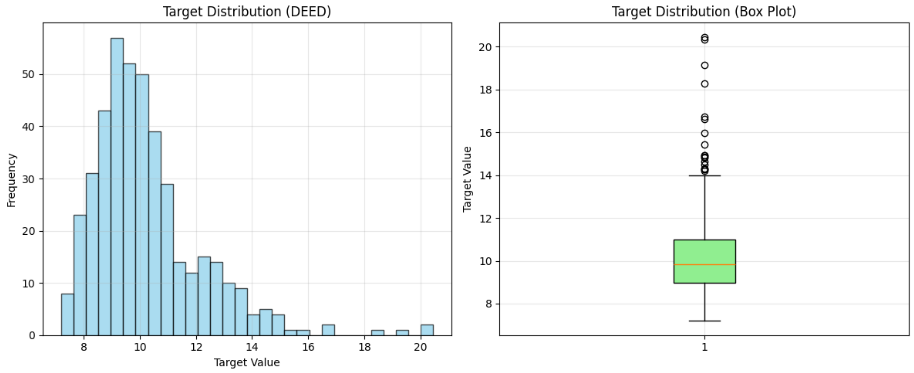
Figure 3: cooling

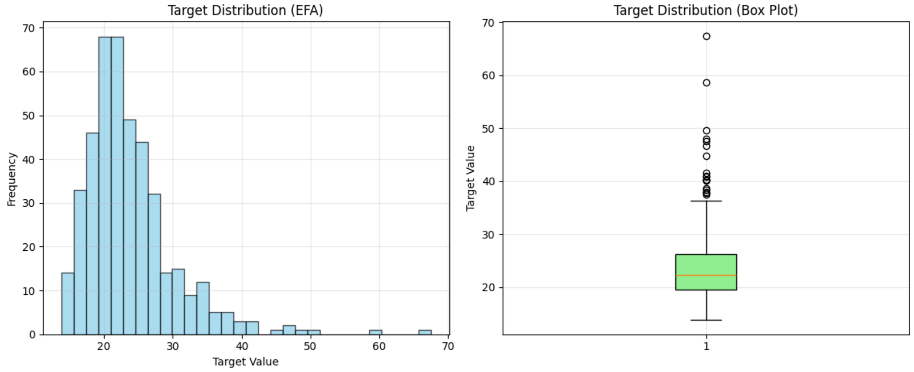
d. During the process, the system does not fully crystallize to FCC, but the number of ICO structure increases a lot. This might due to the cooling rate is still too fast (even I have already set to be 10 K every 10000 steps of cooling with 10000 step of holding at the temperature before keep on cooling).

2 Machine Learning

2.1 Learn the catalyst performance descriptor using data from the High-Entropy Alloy (HEA) database

The code have been modified accordingly to train the machine learning models for both descriptors, EFA and DEED. By comparing the performace then and model accuracy of both descriptors and analyze the difference from the perspective of data distribution and descriptor characteristics.





The target distribution refers to the value distribution of the target properties in the dataset. From the result, we find that the distribution of DEED have more narrow distribution compare to EFA. With DEED obtain the standard deviation of 1.940, and EFA's standard deviation is 6.574. A smaller standard deviation in a target distribution indicates that the data points are tightly clustered around the mean, signifying low variability and high precision. In machine learning, this means a model is likely to make more consistent and predictable predictions, as the target values are less spread out.

From K-means clustering analysis, we have confirm the statement mentioned above, since DEED have greater accuracy in K-mean clustering analysis comparing with EFA descriptor. With the test accuracy for DEED is 0.605 and for EFA is 0.504; the train accuracy is 0.574 for DEED and 0.530 for EFA. From Confusion Matrix analysis, we also see the similar trend. The accuracy for DEED is 0.6047 and for EFA is 0.5039, indicating that the narrower distribution of the targets would provide better result in the methods we were choosing.

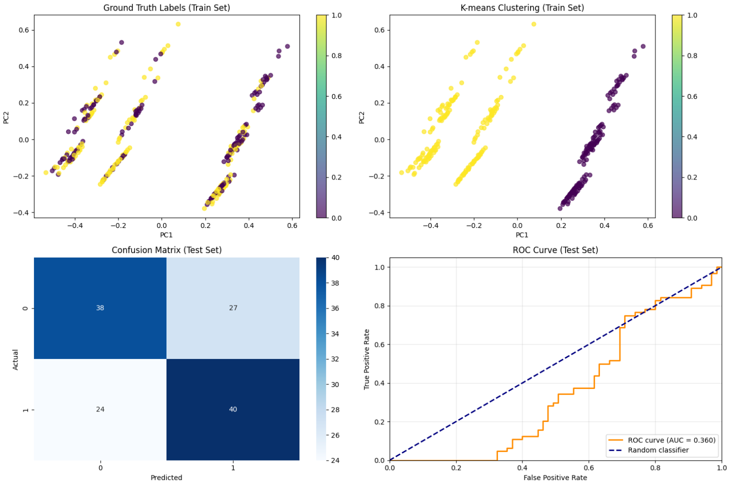


Figure 4: The K-mean clustering analysis for DEED

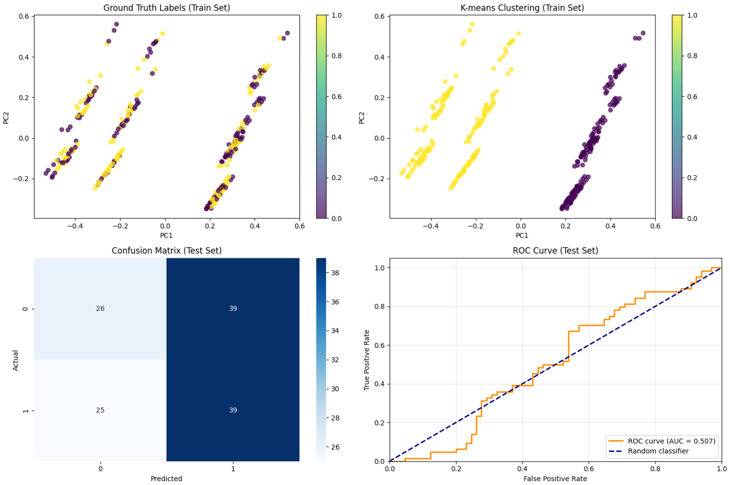


Figure 5: The K-mean clustering analysis for EFA

Since the value for r square is lower, R square provide a better result (closer to 0) in both random forest and gradient boosting models for DEED, we are using MAE metrics for DEED analysis. And all the other methods have the negative difference which means that there are potential possibility of underfitting. Compare to different regression method, random forest regression have better performance for DEED by using R square matrice.

The overfitting analysis for random forest(DEED):

```
R² difference (Train - Test): 0.2448
```

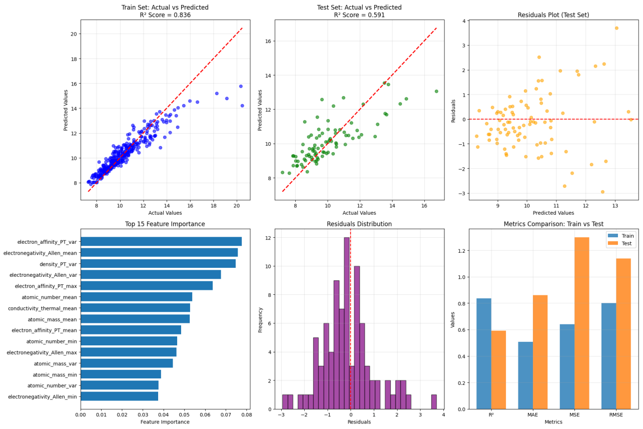


Figure 6: The Random Forest Regression Implementation for DEED

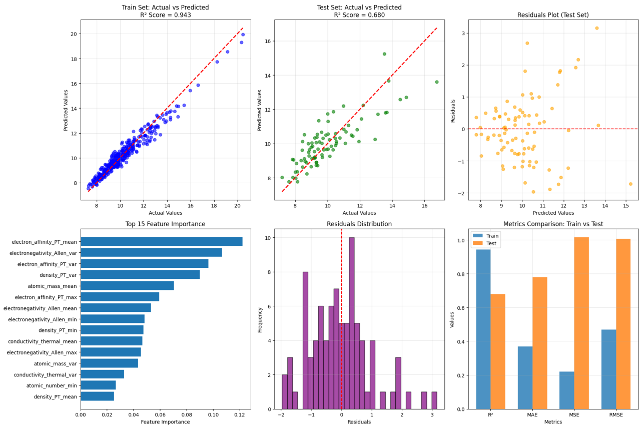


Figure 7: The Gradient Boosting Regression Implementation for DEED

```
MAE difference (Train - Test): -0.3547
MSE difference (Train - Test): -0.6579
RMSE difference (Train - Test): -0.3392
MAPE difference (Train - Test): -3.8637
```

The overfitting analysis for Gradient Boosting Regression(DEED):

```
R² difference (Train - Test): 0.2716
MAE difference (Train - Test): -0.4190
MSE difference (Train - Test): -0.8210
RMSE difference (Train - Test): -0.5511
MAPE difference (Train - Test): -4.1085
```

For the EFA analysis, the R square analysis is also the one that mostly close to zero. All the rest are much smaller than zero which potentially leads to the underfitting problem. Compare the R square value, we also find that the Gradient Boosting Regression have smaller R square difference in overfitting analysis.

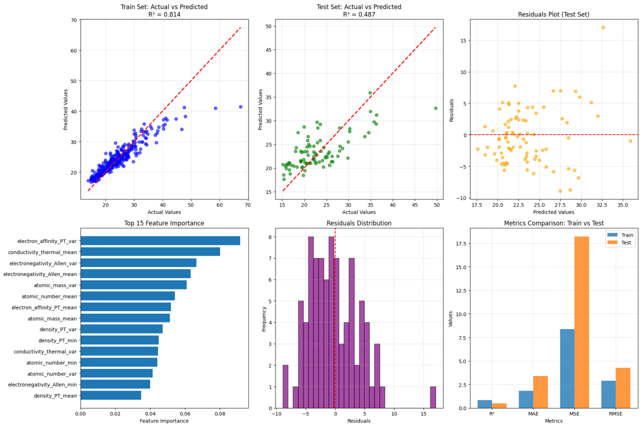


Figure 8: The Random Forest Regression Implementation for EFA

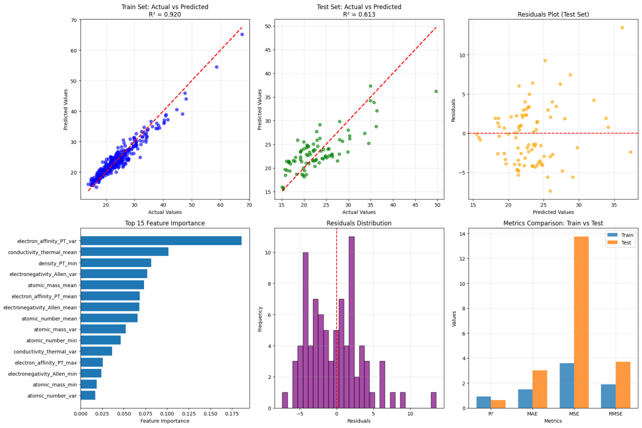


Figure 9: The Gradient Boosting Regression Implementation for EFA

The overfitting analysis for random forest(EFA):

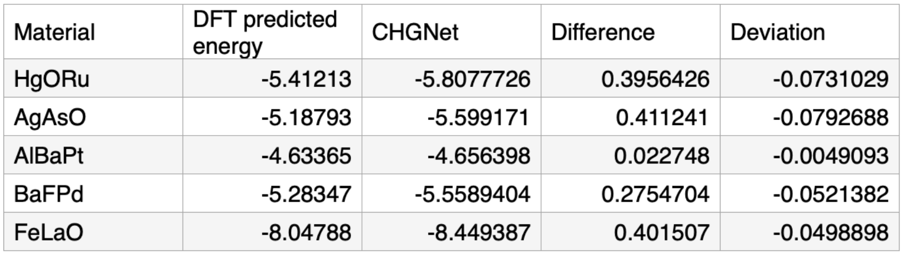
```
R² difference (Train - Test): 0.3272
MAE difference (Train - Test): -1.5775
MSE difference (Train - Test): -9.8401
RMSE difference (Train - Test): -1.3740
MAPE difference (Train - Test): -7.6042
```

The overfitting analysis for Gradient Boosting Regression(EFA):

```
R² difference (Train - Test): 0.3075
MAE difference (Train - Test): -1.5068
MSE difference (Train - Test): -10.1511
RMSE difference (Train - Test): -1.8117
MAPE difference (Train - Test): -6.8248
```

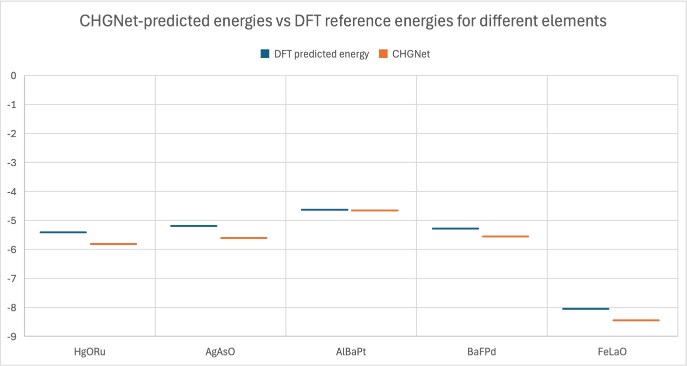
2.2 MLFF

The energies calculate by CHGNet model have been calculated by the jupyter notebook. The energy difference and deviation from DFT calculation are listed as below:



Material	DFT predicted energy	CHGNet	Difference	Deviation
HgORu	-5.41213	-5.8077726	0.3956426	-0.0731029
AgAsO	-5.18793	-5.599171	0.411241	-0.0792688
AlBaPt	-4.63365	-4.656398	0.022748	-0.0049093
BaFPd	-5.28347	-5.5589404	0.2754704	-0.0521382
FeLaO	-8.04788	-8.449387	0.401507	-0.0498898

Figure 10: Table 1. the table of energy comparison from CHGNet calculated energies and DFT-predicted energies



From the result, the AlBaPt shows the smallest deviation compare with other materials (0.49%). The largest deviation comes from AgAsO (7.92%). This might due to the fact that those are some non-common materials with poor representation in the training set, and the improvement of energies might be able to reach with better dataset.

3 contribution

The group contribute equally. We listen and help each other out through the process.