MACHINE LEARNING ENABLED SCREENING OF SINGLE ATOM ALLOYS : PREDICTING REACTIVITY TREND

Purab Gupta

2022chb1055

Under guidance of
Dr Arghya Banerjee

Context

The need to calculate **binding energy** here arises in **catalytic hydrogenation** in **pyrolysis** when conversion of **biomass** to **bio oil** occurs which ultimately has to be used as **fuel**

$$\mathrm{RCHO} + 2\mathrm{H_2} \xrightarrow{\text{Catalyst, } T,P} \mathrm{RCH_3} + \mathrm{H_2O}$$

Catalytic hydrogenation of RCHO compound

Introduction

- In transition metals with vacant d band orbitals take parts in catalytic reaction by donating /filling d-orbital
- Generally d band center is a clear descriptor of catalysis rate
- Deviations observed in reactivity in coinage metals (Cu,Au,Ag) where d orbitals are completely filled

Alternate -use available periodic properties of elements to build ML model for binding energy calculations

- Density Function Theory (VASP)

$$E_{ads} = E_{s+a} - (E_s + E_a)$$

- **Machine Learning model (Python , Colab)**

  Available periodic properties

- Microkinetic modelling (CatMap)

  Inputs from DFT

Problem Statement

- Which fundamental properties of **transition metals** (e.g., atomic radius, electron affinity, valence electron count) affect binding energy most significantly?
- How do the properties of the **adsorbates** themselves (e.g., ionic radius, electron affinity) contribute to adsorption strength?
- Can machine learning models provide **accurate** and efficient predictions of binding energy based on these fundamental properties?

ML Setup (approach)

- **Exploratory Data Analysis**
- **Feature Selection**
- **Training models and validation**

- 80:20 train test split
- 6 models applied
- Optimised parameter using grid search

NOTE

This presentation is a comprehensive analysis of 2 areas

- Part 1- Validating the literature (primary report)
- Part 2- Extension of its analysis to broader chemical framework

- Observed on single atom alloys of Cu, Au ,Ag (doped)
- 3 sites observed (3 fold , bridge,atop)
- Site with higher E_o and E_c used
- 20 features used
- 19 dopant metals

Part 1

Note : paper did not contain data so were extracted and researched

Features

Features of dopant metals were taken and were used for separate binding energy

- **Element**
- **Atomic Number**
- **Atomic Mass (u)**
- **Group**
- **Period**
- **Radius (pm)**
- **Electronegativity (Pauling)**
- **Melting Point (°C)**
- **Boiling Point (°C)**
- **Heat of Fusion (kJ/mol)**
- **Ionization Energy (kJ/mol)**
- **Surface Energy (J/m^2)**
- **Density (g/cm^3)**
- **d-band Center (eV)**
- **d-band Filling**
- **Number of d Electrons**
- **Wigner-Seitz Radius (pm)**

Plot of E_o and E_c vs dopant group

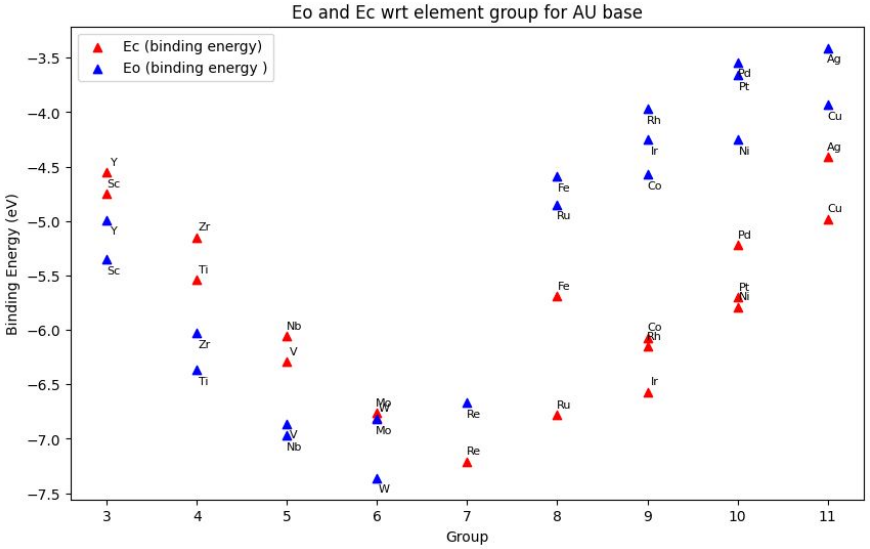


figure 1

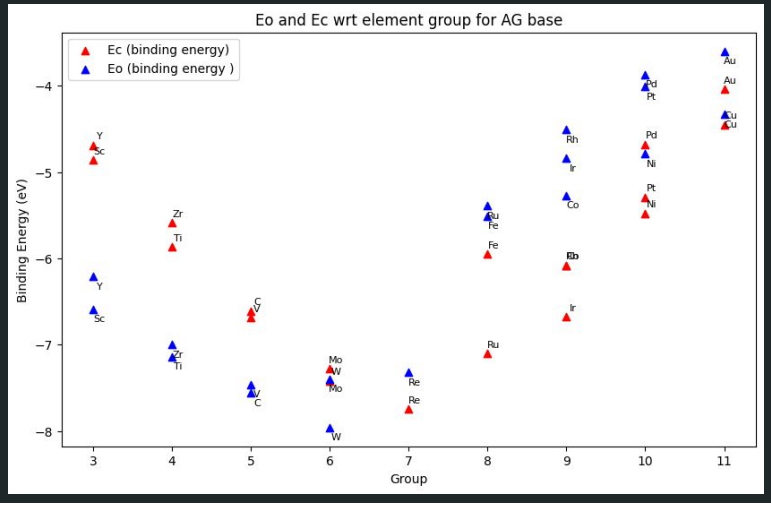


figure 2

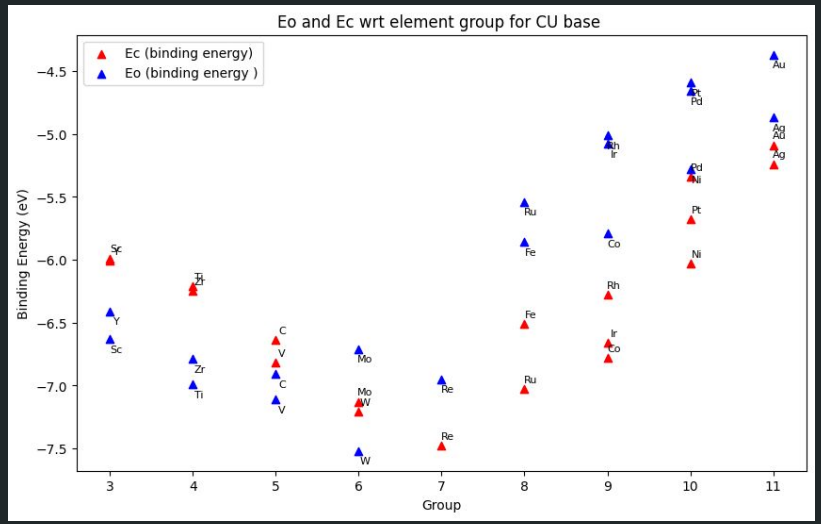
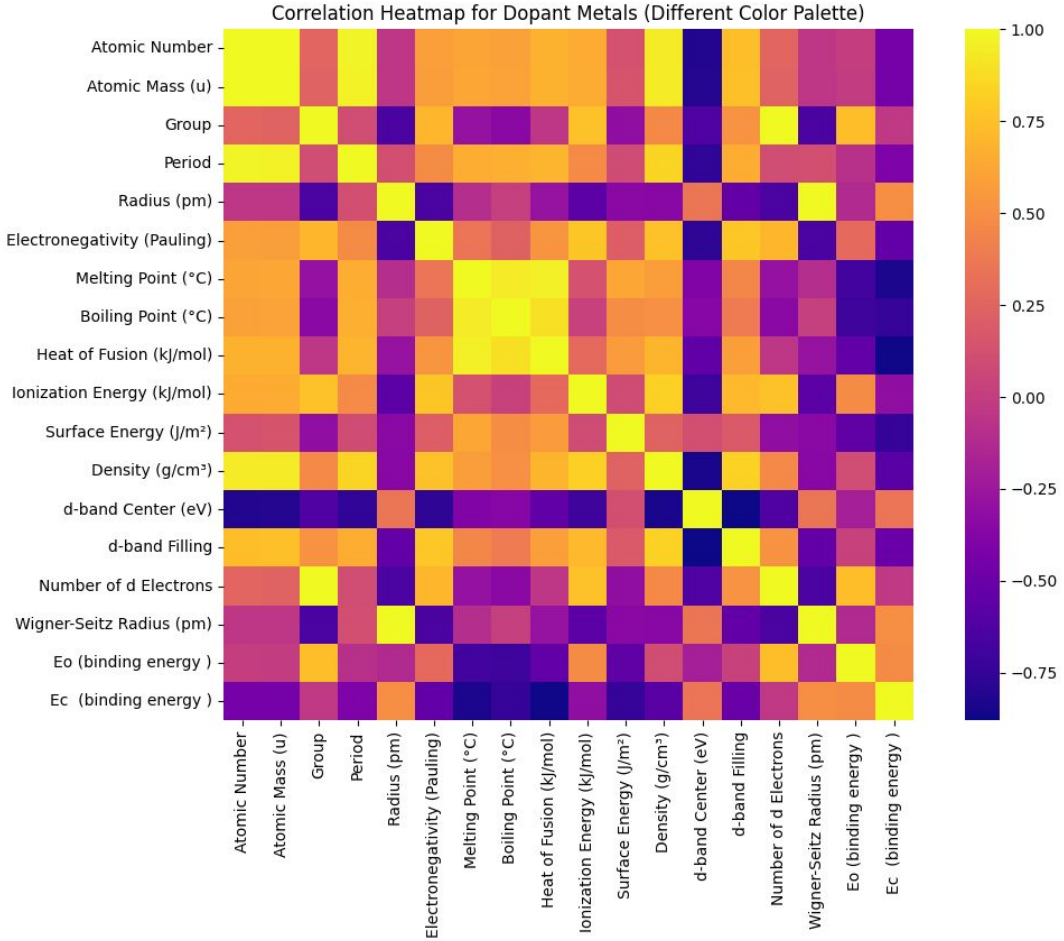


figure 3



Correlation plot

figure 4

Methodology (Machine learning models)

The following ML models were introduced

1. K Nearest Neighbors (KNN)
2. Support Vector Regression (SVR)
3. Random Forest Regression (RFR)
4. Extra Tree regression (ETR)
5. Gradient Boost Regression (GBR)
6. Artificial Neural Network (ANN)

- **KNN**
  - Hyperparameter K = 4
  - RMSE E_o = 0.5089 eV
  - RMSE E_c = 0.5451 eV
- **SVR**
  - Hyperparameters
    - Kernel = RBF
    - kernel coefficient (gamma)= 0.02
    - regularization parameter (C) = 1000
  - RMSE E_o = 0.4651 eV
  - RMSE E_c = 0.3567 eV

- **RFR**
  - Hyperparameters
    - number of decision trees = 500
    - maximum depth of the tree = 6
  - RMSE E_o = 0.29574eV
  - RMSE E_c = 0.37300eV
- **ETR**
  - Hyperparameters
    - number of decision trees= 500
    - maximum depth of the tree =5
  - RMSE E_o = 0.2589 eV
  - RMSE E_c = 0.47855 eV

- **GBR**
  - Hyperparameters
    - number of decision trees = 500
    - learning rate = 0.3
    - maximum depth of the tree = 3
  - RMSE E_o =0.3485eV
  - RMSE E_c = 0.4435eV
- **ANN**
  - Hyperparameters
    - ANN activation function = rectified linear unit
    - loss function = mean square error
    - Optimizer = stochastic gradient descent
  - RMSE E_o = 0.5902eV
  - RMSE E_c = 1.50790V

Feature importance according to GBR on Au Base

E_o

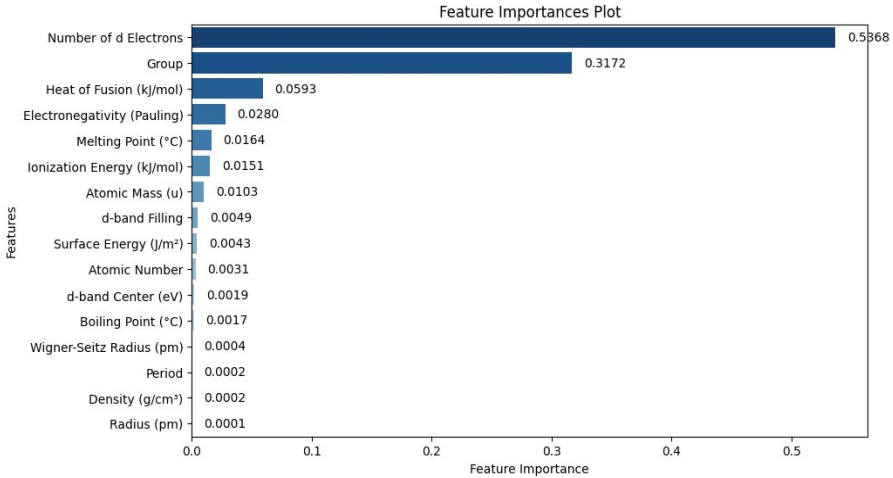


figure 5

E_c

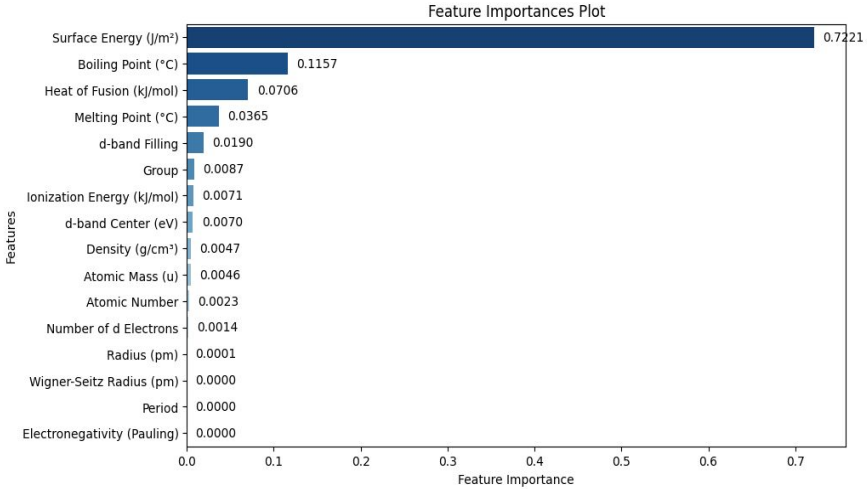


figure 6

Comparing Research Paper Results

Model	RMSE (Ec)	RMSE (Eo)	Validation RMSE (Ec)	Validation RMSE (Eo)
KNN	**0.50**	**0.47**	0.5451	0.5089
SVR	**0.21**	**0.26**	0.3567	0.4651
RFR	**0.34**	**0.31**	0.3567	0.29574
ETR	**0.28**	**0.29**	0.47855	0.2589
GBR	**0.19**	**0.22**	0.4435	0.3485
ANN	**0.1**	**0.19**	1.5079	0.5902

Reasons for result variation

- **System variation** - Both systems may have different computer as well as software powers . The version and language also makes a difference
- **Absence of dataset** - As not direct reference to all the data used was given and contained complex measurements through advance softwares , those feature values were extracted from not so reliable sources
- **Difference in data size** - Different models perform differently depending on data size
- **Feature engineering** - different preprocessing and methodologies may have been used

- Does not have separate target variable
- Features of both adsorbate and adsorbent were taken
- Unimolecular non-doped metal surface considered
- 23 features used
- 9 surface metals (9x3)

Part 2

Binding energy targeted for carbon(C),oxygen(O) , hydrogen(H)

Features (both adsorbent and adsorbate)

- Atomic Number
- Atomic Radius (Å)
- Ionic Radius (Å)
- Electrons in Outer Shell
- Number of Shells
- Valence Electrons
- Electron Affinity (kJ/mol)
- Electronegativity (Pauling Scale)
- Melting Point (°C)
- Boiling Point (°C)
- Heat of Fusion (kJ/mol)
- Heat of Vaporization (kJ/mol)
- Adsorbate Atomic Radius (Å)
- Adsorbate Atomic Number
- Adsorbate Ionic Radius (Å)
- Adsorbate Electrons in Outer Shell
- Adsorbate Electron Affinity (kJ/mol)
- Adsorbate Electronegativity (Pauling Scale)
- Adsorbate Boiling Point (°C)
- Adsorbate Melting Point (°C)
- Adsorbate Heat of Fusion (kJ/mol)
- Adsorbate Heat of Vaporization (kJ/mol)

Binding energy variation with Metal (hue = adsorbate)

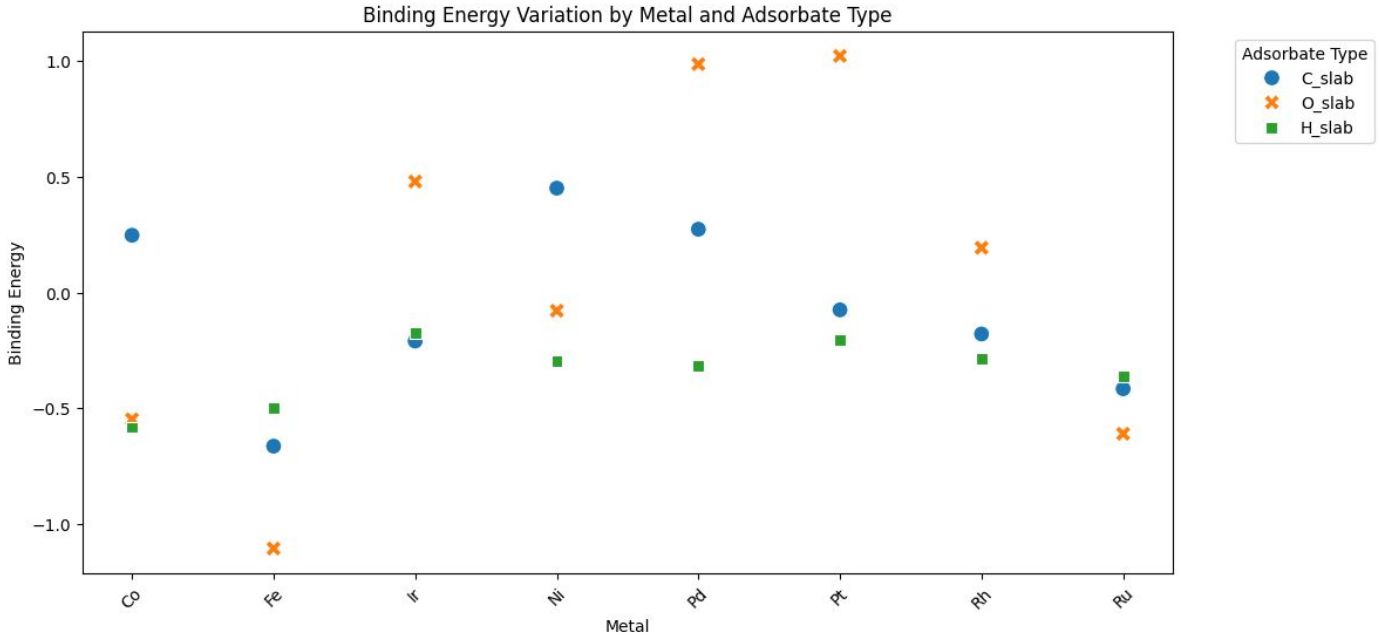


figure 7

Correlation plot

(binding energy highlighted)

Binding energy centric correlation plot

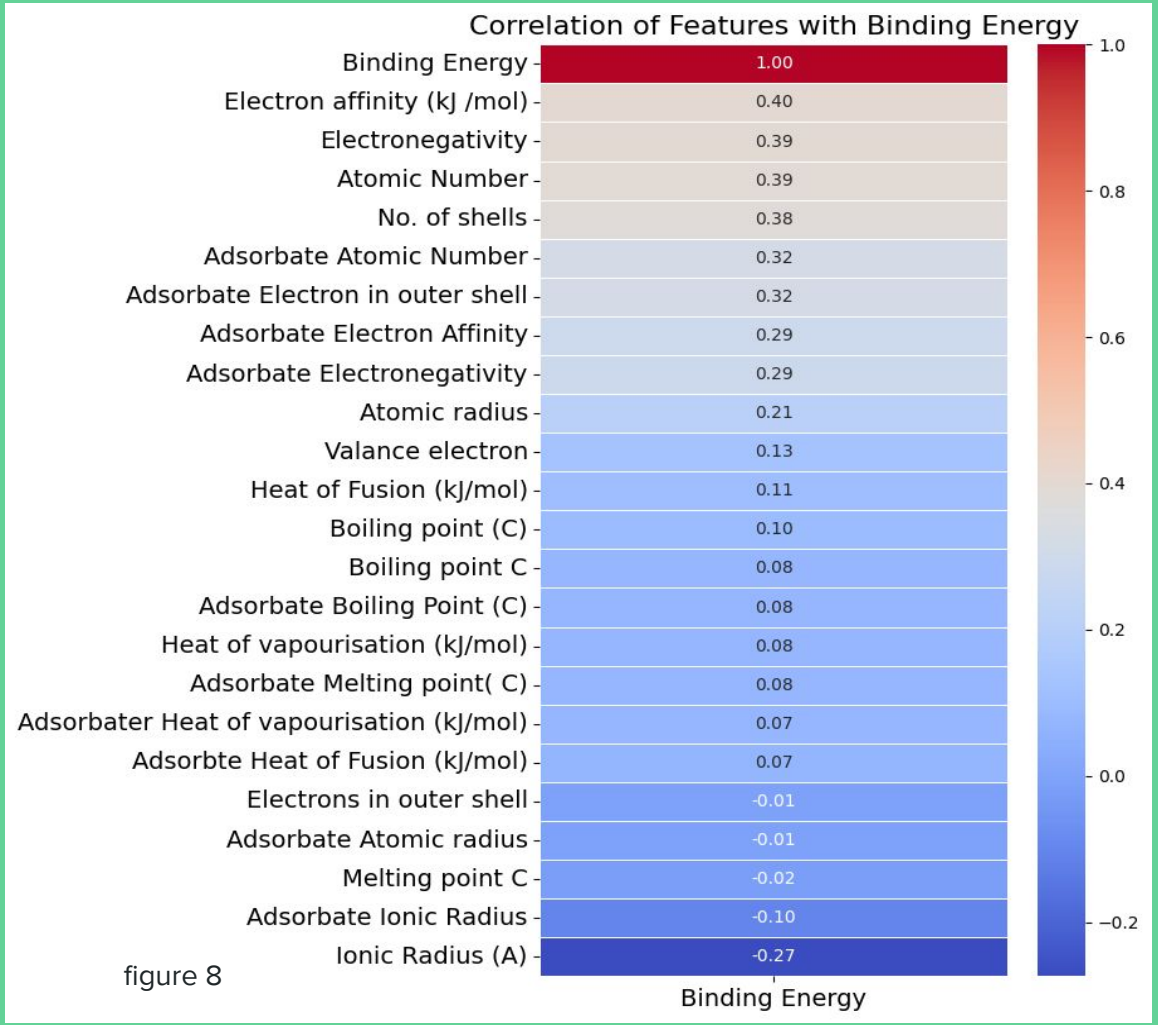


figure 8

ML model hyperparameter and results

- **KNN**
  - Hyperparameter K = 6
  - RMSE E = 0.3418 eV

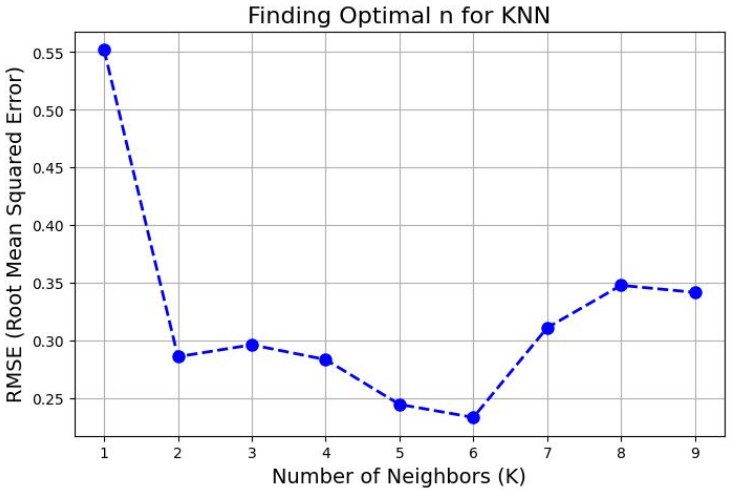


figure 9

- **SVR**
  - Hyperparameters
    - Kernel = RBF
    - Tolerance (ϵ) = 0.01
    - regularization parameter (C) = 1
  - RMSE E = 0.247 eV

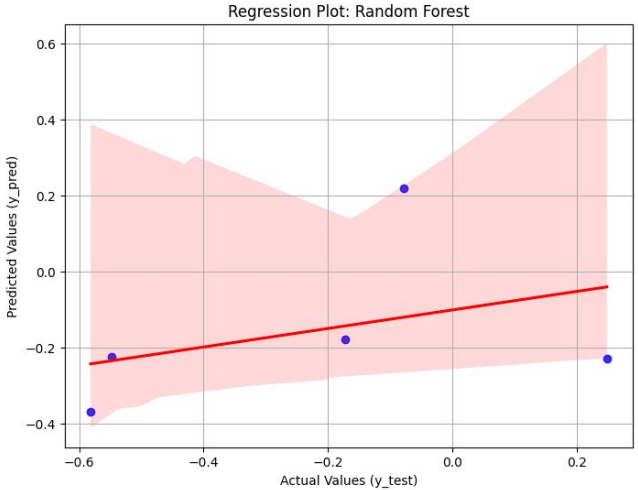


figure 10

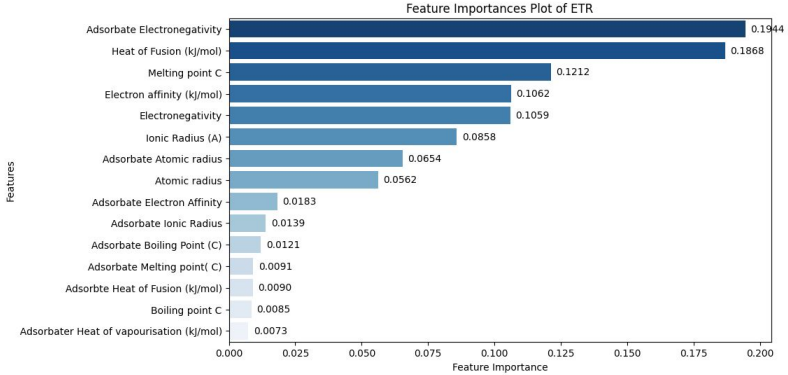


figure 11

- **RFR**
  - Hyperparameters
    - number of decision trees = 500
    - maximum depth of the tree 6
  - RMSE E = 0.306 eV

- **ETR**
  - Hyperparameters
    - number of decision trees= 500
    - maximum depth of the tree 20
  - RMSE E = 0.432 eV

- **GBR**
  - Hyperparameters
    - number of decision trees = 100
    - learning rate = 0.1
    - maximum depth of the tree 3
  - RMSE E = 0.437 eV
- **ANN**
  - Hyperparameters
    - ANN activation function = ReLu
    - loss function mean square error
    - optimizer = Adam optimizer
    - Number of hidden layer = 3
    - epochs=100
  - RMSE E = 0.087 eV

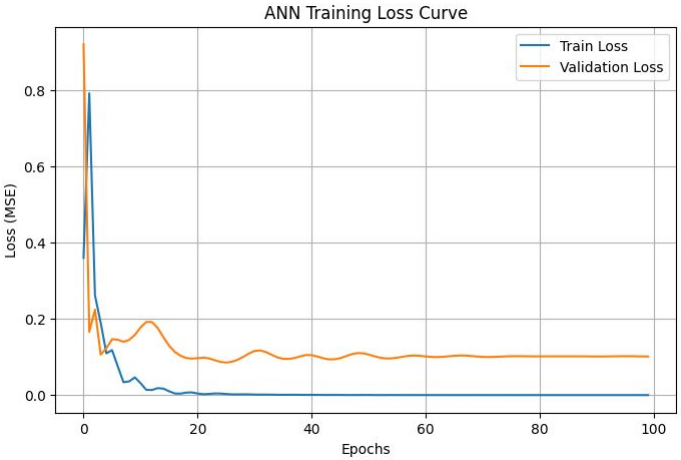


figure 11

Analyzing ANN

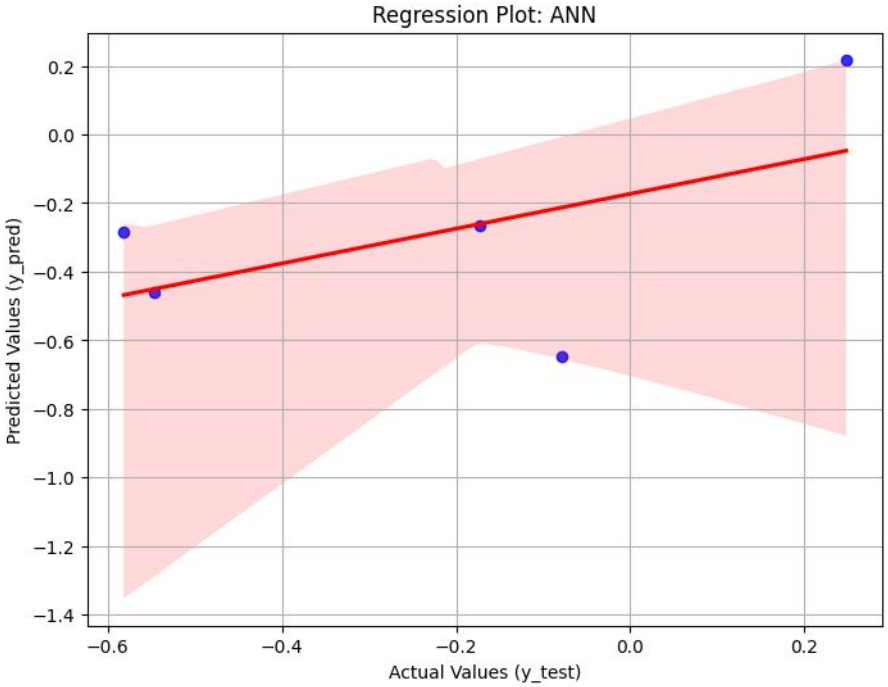


figure 12 Actual vs Predicted plot

	Actual (y_test)	Predicted (y_pred)
0	-0.54722	-0.460582
1	-0.58157	-0.285437
2	0.24781	0.218869
3	-0.17317	-0.267375
4	-0.07809	-0.645439

fig 13 ***Actual vs Predicted values***

Overall analysis of all models

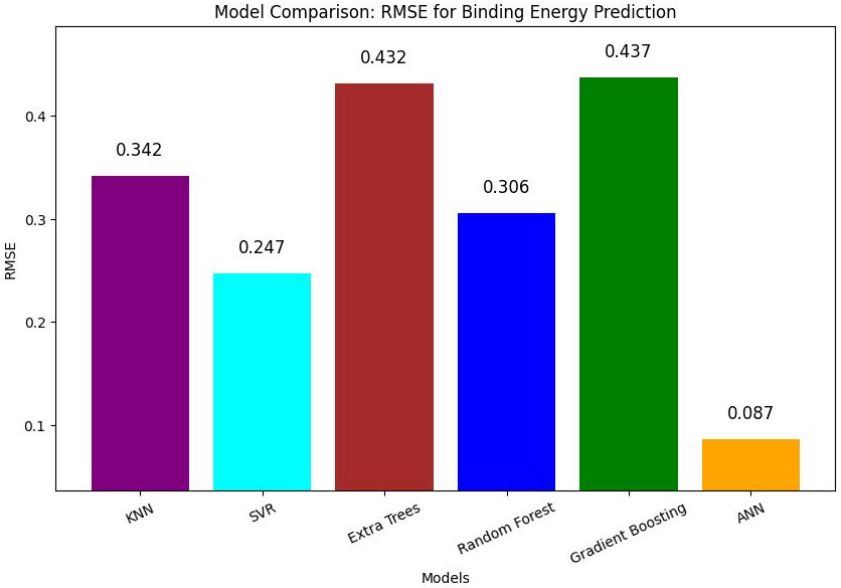


fig14 ***RMSE plot***

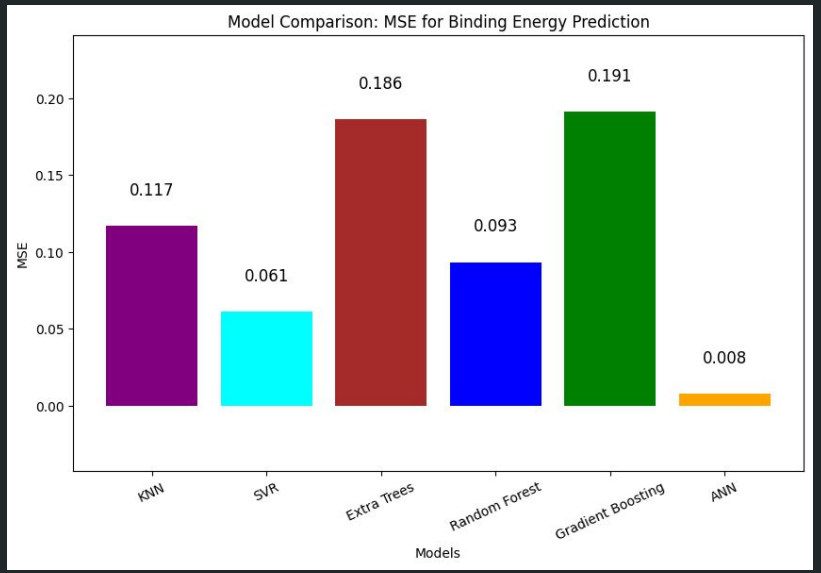


fig 15 ***MSE plot***

Conclusion

- **ANN** and **SVR** came out be the most effective followed by **RFR**
- Threshold value accepted of error is **0.5 eV** , which every predictive model used passes
- Feature importance analysis revealed that **Adsorbate Electronegativity, Heat of Fusion, and Melting Point** had the highest impact
- The trained models showed **good generalizability**

Future work

- With proper dataset DFT values as well as d band properties if incorporated may help in efficiency
- Expanding Feature engineering
- Increasing data generalization capacity
- Implementing Advanced deep learning