Data-Driven Prediction of Formation Energies of Binary Compounds Using Machine Learning

By Tetiksha Jain

Introduction

A precise estimation of the thermodynamic stability of materials remains central to the successful formation and synthesis of new compounds in materials science, especially in discovering materials with optimal properties. Formation energy, which can often be computed via the quantum mechanical method of density functional theory (DFT), is an important measure of stability (Kirklin et al., 2015). It is also known that DFT is accurate, but also very expensive computationally. This research, adopt the principles of materials informatics and seek to determine if formation energies can be defined from simple atomic characteristics instead of full information captured at the crystal structure level using machine learning. This will facilitate rapid assessment of materials. The work relies on data in the publication from Wang et al.

Mao, Y., Yang, H., Sheng, Y., Wang, J., Ouyang, R., Ye, C., Yang, J., & Zhang, W. (2021). Prediction and Classification of Formation Energies of Binary Compounds by Machine Learning: An Approach without Crystal Structure Information. *ACS Omega*, *6*(22), 14533–14541. https://doi.org/10.1021/acsomega.1c01517

The dataset is used from the above work and combined with atomic level descriptors such as ionization potential, electron affinity and bond dissociation energy regarding the constituents of the binary compounds.

Methods

I used a dataset of 183 binary compounds from the supplemental information provided by Wang et al. (2021), which included experimentally measured formation enthalpies per atom and over 90 atomic and compositional features. From these, 23 features with clear physical

relevance such as bond dissociation energy (BDE), electron affinity (EA), electronegativity (EN), atomic radii, estimated volume, and basic unit cell parameters were selected. These features provided a meaningful description of the chemical and structural behavior of compounds without requiring detailed crystal structure information.

To model the data, I implemented two tree-based machine learning algorithms: Random Forest Regressor and Gradient Boosting Regressor. Random Forest Regressor averages the predictions of many independent decision trees, providing robustness to overfitting. In Gradient Boosting Regressor, each tree is built sequentially, improving upon the accuracy of the previous one. The dataset was divided in a stratified manner, whereby 80% constituted the training subset and 20% served the testing subset. Mean Absolute Error (MAE) and R^2 score were used to evaluate the model's performance. To ensure reliability, I also used 5-fold cross-validation, averaging performance across multiple train-test splits.

In addition to full-model training, a simplified model was also built using only the five most important features, as identified through feature importance scores, to assess whether a smaller model could still perform well. Secondly, I applied Principal Component Analysis (PCA) to reduce the feature set to five uncorrelated components and trained a model on those.

Results

Random Forest and Gradient Boosting Algorithms were trained to see which one performed best with the full set of 23 features. The best performing model was the Gradient Boosting Regressor which had a R^2 score of 0.86 and an MAE of 28,000 J/mol-atom, outperforming the Random Forest model (R^2 = 0.78, MAE = 30,973 J/mol-atom). Cross-validation score further confirmed the model's stability, with Gradient Boosting achieving a cross-validated R^2 of 0.79 and an average MAE of 29,400 J/mol-atom. These conclusions imply that Gradient Boosting is excellent for capturing intricate relationships between compositional features and formation energy-based phenomena.

To help interpret the model's behavior, I graphed the predicted versus actual values of the formation energies (Figure 1). Several predictions are near the ideal line, demonstrating that

the model is capable of picking up some of the energy ranges which are significant, these ranges include stable or (more negative) and less stable compounds.

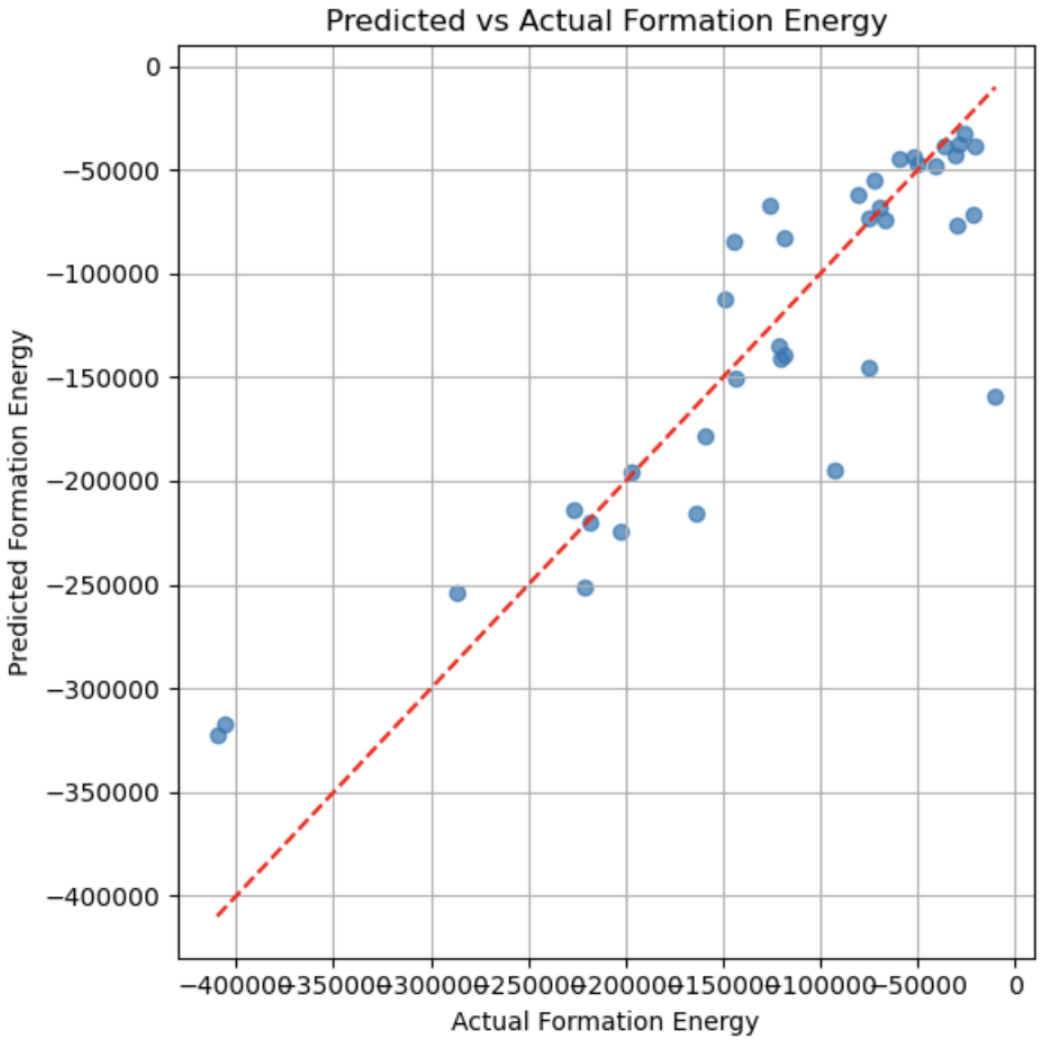

Figure 1: *Scatter plot of actual vs. predicted formation energies using Gradient Boosting. Most values fall near the ideal line, indicating good predictive performance.*

The residual plot (Figure 2), shows that the prediction errors are not only small but also approximately evenly distributed. It also depicts that the model is slightly underpredicting for highly negative formation energy compounds. The residuals are balanced around 0, meaning the predictions made are unbiased and balanced.

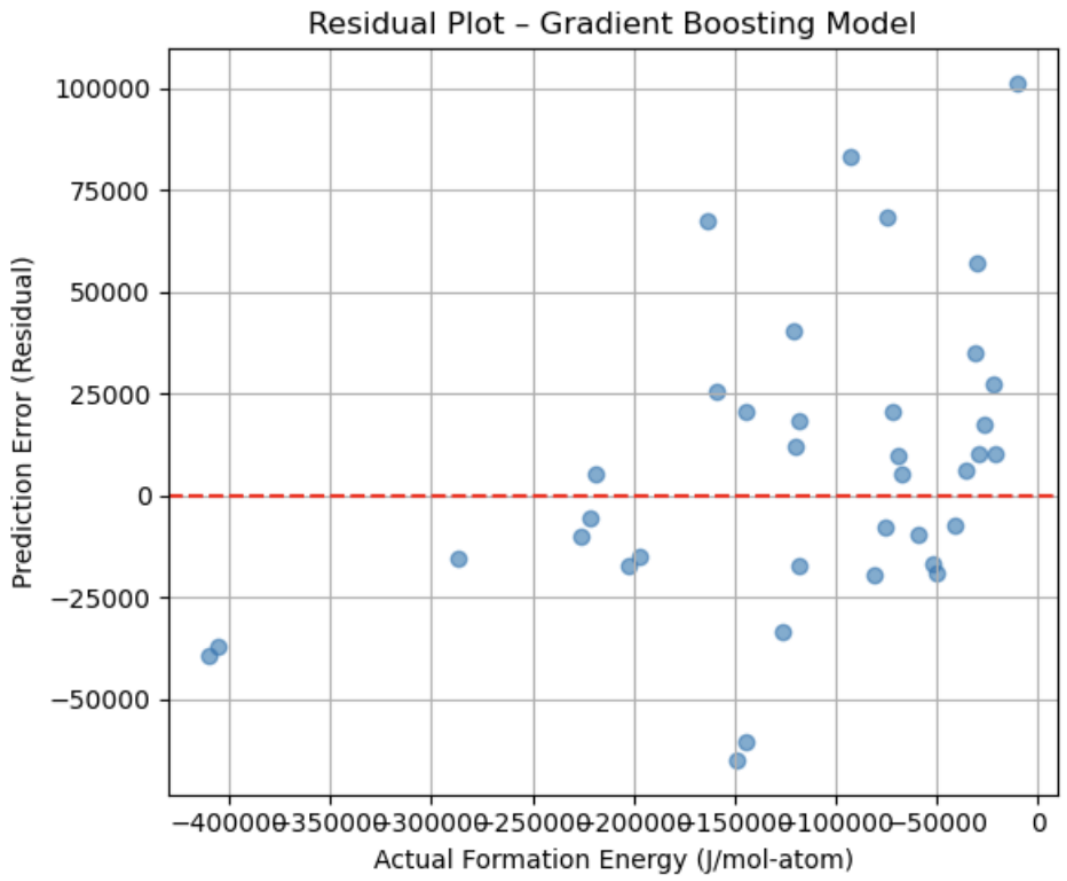

Figure 2: *Residual plot showing prediction errors. Most residuals are small and centered around zero, with some underprediction for highly stable compounds.*

The Random Forest model allowed us to determine feature importance as depicted in Figure 3. Bond Dissociation Energy (BDE) as well as Electron Affinity and Electronegativity were amongst the most important features, which correspond to chemical expectations, hence verifying that the model is learning useful patterns. The importance ranking aligns very closely with Wang et al 2021's results, except Bond Dissociation Energy, which strengthens the argument that both our model and the previous model identify similar key driver of formation energy.

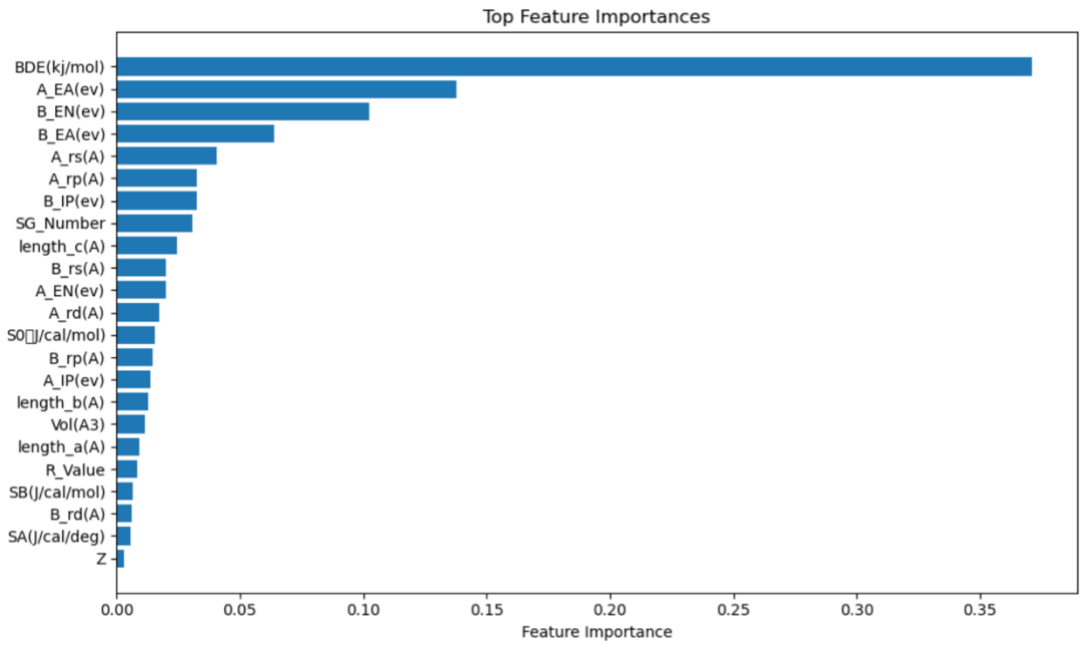

Figure 3: *Feature importance plot from the Random Forest model. BDE, electron affinity, and electronegativity contribute the most to prediction accuracy.*

Further, I retrained the model to only include the top five features and this derived model exhibited R^2 of 0.83, which closely aligns with the full-feature model. The simplified model outperformed the PCA-based model that utilized five principal components R^2 = 0.65 and MAE = 42,237 J/mol-atom. This implies that while the PCA based model had preserved a large amount of variance, it had also ignored many signals.

Discussion and Summary

In this project I aimed to determine if it is possible to predict the formation energy of binary compounds using only the atomic and compositional features of the structure without needing the full crystal structure via machine learning methods. After training two tree based models, Random Forest and Gradient Boosting, it can be confirmed that reasonable predictions can indeed be made. The Gradient Boosting Regressor performed the best with an R^2 score of 0.86 and a mean absolute error of approximately 28,000 J/mol-atom. These results add credence with the standpoint that formation energy can relatively be captured using elemental

descriptors at the level of the constituent atoms. A notable outcome from the analysis was that the model which relied solely on the topmost five features performed nearly as well as the complete model (R^2=0.83).

In combination with well-chosen features based on chemistry and physics, this project has shown that machine learning, when used with a specific approach, can reliably and efficiently predict material properties machine learning is an effective tool. The achievement of a basic, structureless model suggests that materials could be screened more easily within larger chemical frameworks. Future developments may include the use of SHAP for explanation on specific predictions, evaluation of model generalization on ternary or quaternary compounds, and the application of uncertainty quantification to enhance strategic decision-making in materials design.

References

Mao, Y., Yang, H., Sheng, Y., Wang, J., Ouyang, R., Ye, C., Yang, J., & Zhang, W. (2021). Prediction and Classification of Formation Energies of Binary Compounds by Machine Learning: An Approach without Crystal Structure Information. *ACS Omega*, *6*(22), 14533–14541. https://doi.org/10.1021/acsomega.1c01517

Kirklin, S., Saal, J. E., Meredig, B., Thompson, A., Doak, J. W., Aykol, M., Rühl, S., & Wolverton, C. (2015). The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *Npj Computational Materials*, *1*(1). https://doi.org/10.1038/npjcompumats.2015.10