Journal of Institute of Science and Technology, 29(2), 109-115 (2024)
ISSN: 2467-9062 (print), e-ISSN: 2467-9240
https://doi.org/10.3126/jist.v29i2.70481

Short Communication

MACHINE LEARNING MODEL TO PREDICT THE FORMATION ENERGY OF COPPER-BASED TERNARY ALLOYS

Subash Dahal, Devendra Adhikari, Shashit K. Yadav*
Department of Physics, Mahendra Morang Adarsh Multiple Campus, Tribhuvan University, Biratnagar, Nepal
**Correspondence: yadavshashit@yahoo.com*

(Received:** October 04, 2024; **Final Revision:** December 21, 2024**; Accepted:** December 23, 2024**)

ABSTRACT

Formation energy plays a crucial role in material development, serving as a key metric for understanding material stability and behavior. In this study, machine learning algorithms, namely Random Forest Regressor (RFR) and Gradient Boosting Regressor (GBR), were employed to predict the formation energy of copper-based ternary alloys. The models were implemented using the Scikit-Learn library within the Anaconda distribution. A composition-based featurizer, Magpie elemental properties, from the Matminer toolkit, was utilized to represent the alloy's features. The results demonstrate that the composition-based featurizer effectively captures the relationship between alloy composition and formation energy. Among the models, GBR outperformed RFR, explaining 94% of the variance in formation energy using only five features, compared to 92.5% explained by RFR, which required ten features. These findings highlight the efficiency and accuracy of GBR in predicting formation energy with fewer input features. This work underscores the potential of machine learning models, particularly the GBR, as powerful tools for accelerating material discovery and design. By enabling reliable and efficient predictions, these models provide a pathway to streamline material development processes.

Keywords: Cu-based alloys, featurizer, formation energy, hyperparameters, machine learning

INTRODUCTION

The search for novel materials with desired properties is the foundation of material science and engineering. The alloying phenomena has served as robotic tool to develop new materials with desired properties. The materials characteristic can be enhanced or manipulated by varying the thermal processing routes such as micro-granules, atmospheric conditions, cooling rates, etc. (Dhungana *et al.,* 2023). Among numerous material systems, ternary copper-based (Cu-based) alloys have their own diverse range of applications due to their high mechanical strength, electrical conductivity, and thermal conductivity (Inoue *et al.,* 2001; Kosec & Milosev, 2007). Cu-based ternary alloys such as Cu-Al-Mn, Cu-Al-Ni, Cu-Al-Be, etc. are known as shape memory alloys (SMAs), can restore their shape when heated over a specific temperature (Jani *et al.,* 2014; Mazzer et al., 2022). This phenomenon is caused by a martensitic phase transformation, which occurs when a martensitic phase nucleates and grows from an austenitic phase under shear-dominant diffusion less solid-state conditions. Among them Cu–Zn, Cu–Al, and Cu–Sn alloys, both with and without ternary additives, have demonstrated promise because of their exceptional thermal and electrical conductivity, ease of production, and good shape recovery (Dasgupta, 2014). Moreover, Cu-based ternary alloys like Cu-Sn-Ag, Sn-Cu-Bi, Sn-Cu-La, Sn-Cu-Y, etc. are used as lead-free solders in electronic industries to replace Pb containing Sn-Pb solders (Huang *et al.,* 2022; Islam *et al.,* 2005; Ohnuma *et al.,* 2000; Xia *et al.,* 2006).

Moreover, the equilibrium energies, such as formation energy, cohesive energy and Gibbs free energy determine the stability and spontaneity of different inter-metallic complexes. Being more specific, due to the complex interplay of constituent elements and their interactions, the formation energy of the alloy is influenced thereby obstructing the design and development of the alloy. Formation energy is an important thermodynamic parameter that indicates the binding energy of the condensed state of the alloy. For the development and utilization of Cu-based alloys, the prediction of formation energy plays an important role (Zhou *et al.,* 2022). With the development of computer technology, researchers mostly use density functional theory (DFT) for the prediction of formation energy. However, there are still some difficulties due to the time-consuming and intensive computation costs of DFT. Recently, machine-learning models (MLM) have emerged as robust tools for the prediction of material properties. Therefore, the MLMs have been employed to predict the formation energies of Cu-based ternary alloys have been MLMs in present work.

In this regard, Zhuo *et al.* employed a support vector model to predict the band gap of an inorganic compound using only the composition descriptors (Zhuo et al., 2018). Olsthoorn *et al.* used Organic Materials Database (OMDB) data for 12500 crystal structures and predicted the band gap of 260092 materials in the Crystallography Open

Database (COD) by using the kernel ridge regression model and the deep learning model SchNet (Olsthoorn *et al.,* 2019). Similarly, different researches so far have been conducted to predict properties, such as heat capacity (Alade *et al.,* 2020; Aldosari *et al.,* 2021; Kauwe *et al.,* 2018) and Gibbs free energy (Bitencourt-Ferreira and de Azevedo 2018; Desgranges and Delhommelle 2018) with high accuracy. Meanwhile, Faber *et al.* used a MLM to calculate the formation energy of 2 million elapsolite crystals (Faber *et al.,* 2016). A machine learning algorithm (MLA) has been used to predict the formation energy of different materials due to their flexibility and reliability (Faber *et al.,* 2015; Zhou *et al.,* 2022).

Therefore, the goal of this work is to use MLM to predict the formation energy of ternary Cu-based alloys. For this purpose, Random Forest Regressor (RFR) and Gradient Boosting Regressor (GBR) MLMs, have been employed to predict the formation energy of the alloys. In machine learning, specific characteristics of data can be described by descriptors or features. In this work, only a composition-based descriptor of material has been used to predict the formation energy.

MATERIALS AND METHODS

Data collection and preprocessing

The input dataset for the formation energies of Cu-based ternary alloys have been taken from material project database which uses DFT approach for the respective purpose (Jain *et al.,* 2013). The dataset initially contained information about the material ID, structure, band gap, composition, chemical formula, and stability of the material. The obtained data was preprocessed by selecting only the stable materials, removing the duplicates, and sorting the data on the basis of their formation energies. Then after, our dataset contained 973 rows of data for different Cu-based ternary alloys. Before performing machine learning analysis, we need to extract the relevant feature from the raw data. For this, we implemented the Magpie elemental property compositional-based featurizer from Matminer to composition column of the dataset (Ward *et al.,* 2018). This featurizer transforms the chemical composition into a set of numerical features that contain information about elemental composition and properties. We have then dropped the irrelevant column, which contains information about material ID, structure, composition, chemical formula, and stability. Finally, our dataset contains 973 rows and 134 columns, and then we have trained the MLM on this dataset.

Model selection

Different factors must be considered while selecting MLM for particular task, such as nature of the problem, complexity of the model, robustness to noise and outliers, interpretability, and performance metrics. While considering all these factors, we selected two models namely RFR and GBR, both of them follow the ensemble learning methods.

Random Forest Regressor (RFR)

It was first introduced by Tin Kam Ho in 1995. It is an ensemble method which constructs a forest of decision trees on a random subset of the training data and subset of features. This property of RFR aided in reducing the over fitting and improving the performance of the model. By combining the prediction made from multiple trees, RFR provides robust prediction by averaging the prediction obtained from each tree in the forest. Additionally, the versatility of RFR offers robustness to noisy data, high performance and provides depth insight to the feature importance by measuring the impurity (Breiman, 2001; Ho, 1995).

Gradient Boosting Regressor (GBR)

GBR is another powerful ensemble learning method and was first introduced by Leo Breiman in 1997. Unlike Random Forest, which builds random trees independently, Gradient Boosting builds trees sequentially, with each tree focusing to minimize the errors made by the previous one. This process overall improves the performance by learning from the mistakes made by the preceding tree. At each iteration, it fits the new decision tree to the residuals of the current prediction and minimizes the error. This property makes Gradient Boosting the powerful regression algorithm that offers superior performance and interpretability (Breiman, 1997).

Table 1. Hyperparameter used for tuning, and the best parameter obtained for two models

Model	Hyperparameter used for tuning	Best parameter
RFR	Bootstrap = [True, False] max_depth = [1,2,3,4,5,6,7] min_samples_leaf = [1,2,3,4,5,6,7] max_leaf_nodes = [None, 5, 10, 20, 50] max_samples = [0.5, 0.75, 0.85] n_estimators = [10,20,30,40,50,100,150,200]	Bootstrap = True max_depth = 7 max_leaf_nodes = 50 max_samples = 0.85 min_samples_leaf = 2 n_estimators = 40
GBR	n_estimators = [10,30,50,70,100,150] max_depth = [1,3, 5, 7] min_samples_split = [2, 5, 7,10] min_samples_leaf = [1, 2, 3,4]	max_depth = 3 min_samples_leaf = 2 min_samples_split = 2 n_estimators = 150

Feature Selection and Hyperparameter

At first, we divided the dataset into training and testing set with test size of 0.2. After that we train the MLM with its default hyperparameter on training set using scikit learn (Buitinck *et al.,* 2013; Pedregosa *et al.,* 2011) in Anaconda distribution (Anaconda Inc., 2020). We then used the inherent characteristics of RFR and GBR algorithms to identify the best features for the model. For RFR and GBRs, we have identified the top 10 and 5 features, respectively. The important features of the models are shown in Figures 1 and 2.

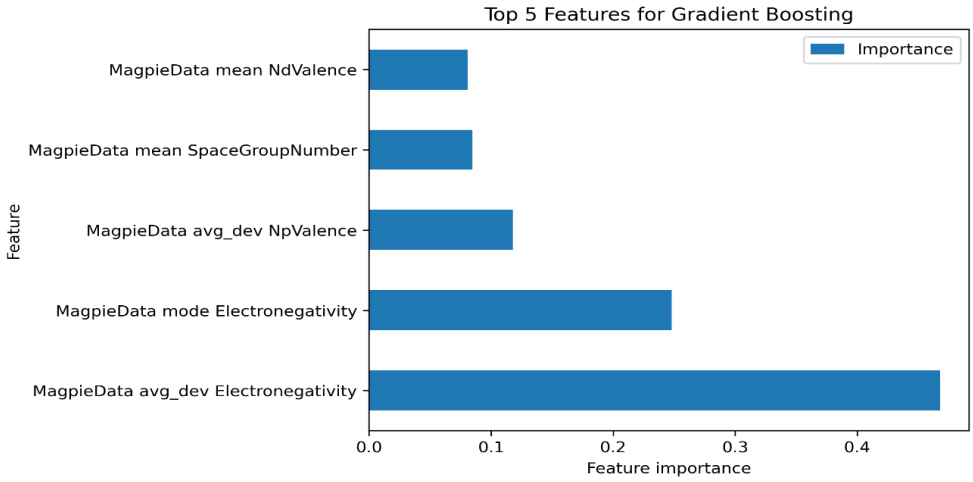

Figure 1. Feature importance for GBR.

The selected features have the potential to maintain a balance between computational complexity and dimensionality reduction, hence improving model predictions and interpretability. Furthermore, hyperparameters are the parameters that improve the performance of the model. In ML, a crucial step is to identify the parameters that best suit an accurate prediction of desired properties of materials. We have implemented GridSearchCV to tune the hyperparameters for both the models. The details of the hyperparameters used to tune and the best parameters obtained for both models are presented in Table 1. We then evaluated the performance of the models using the 10-fold cross-validation technique and metrics, such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), R^2, and adjusted R^2 which are listed in Table 2.

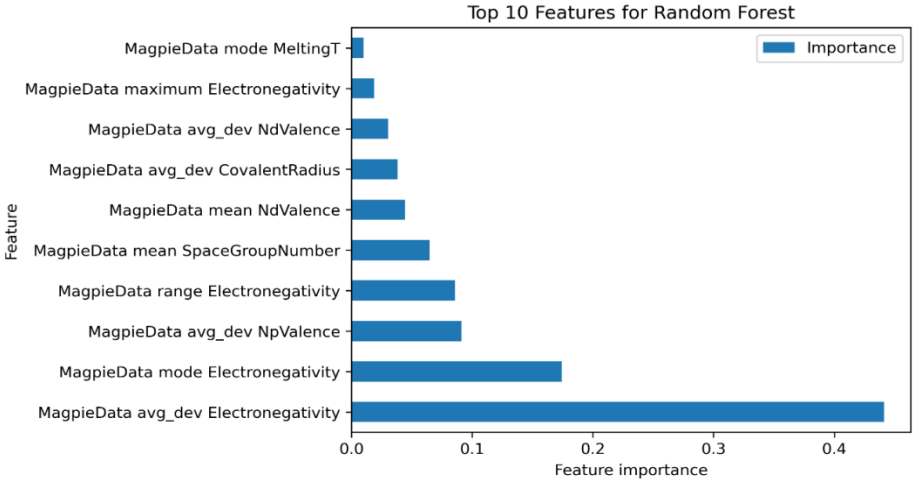

Figure 2. Feature importance for RFR.

RESULTS AND DISCUSSION

One important method for obtaining the electronic structure and material properties using density functional theory is the use of traditional computational approaches, such as first-principles calculation, molecular dynamics simulations, finite element method simulations, etc., which treat electrons as the primary object of study. This process eliminates the need for empirical and semi-empirical parameters and enables precise material calculations (Zhang *et al.,* 2024). Because of their simulation nature or certain presumptions, the aforementioned computational simulation tools rely on theoretical models that deviate from actual experimental research to varying degrees. Thankfully, the emergence of machine learning methods can make up for the limitations of theoretical models. Using the training set-a collection of data with specific attributes chosen at random from the acquired dataset-machine learning approaches typically build some models and train the models in accordance with the related algorithms (Zhang *et al.,* 2024).

Therefore, we have divided the entire above mentioned input dataset into two parts, including training dataset (80%) and testing dataset (20%). The model we implemented to predict the formation energies of Cu-based ternary alloys show remarkable performances by using only the Magpie Elemental Property Featurizer. The RFR, after performing hyperparameter tuning, shows robust capabilities for predicting formation energy. The 10-fold cross-validation model achieved 0.91 in the training set and 0.85 in the testing set. The lower values of MAE, MSE, and RMSE show that the predicted values have a minimal deviation from actual values. Figure 2 shows the comparison between actual values and predicted values of formation energy. Furthermore, the value of R^2 and adjusted R^2 are found to be 0.925 in the testing set, which shows that the model has the ability to explain approximately 92.5% of the variance of formation energy. Due to the involvement of large dataset, the results obtained from the work could not be tabulated and presented in the work. In this regard, only the accuracy levels between the training dataset and predicted dataset are displayed in Table 2.

Table 2. Comparison between input and predicted values of formation energies for Cu-based ternary alloys

Metric	RFR		GBR	
	Training	Testing	Training	Testing
10-fold CV Score	0.911520373	0.85095688	0.901155199	0.866228323
Score	0.965968633	0.928966294	0.967320891	0.942604141
MAE	0.076778491	0.096609038	0.074965362	0.093351844
MSE	0.012152652	0.020475518	0.011669758	0.016544399
RMSE	0.110239066	0.14309269	0.108026655	0.128625031
R^2	0.965968633	0.928966294	0.967320891	0.942604141
Adjusted R^2	0.965524939	0.925105766	0.967109239	0.941085732

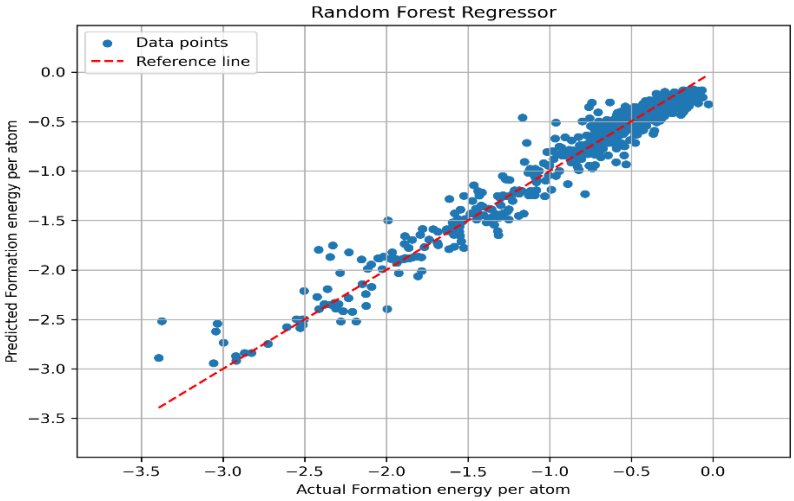

Figure 3. Predicted versus actual formation energy using RFR model.

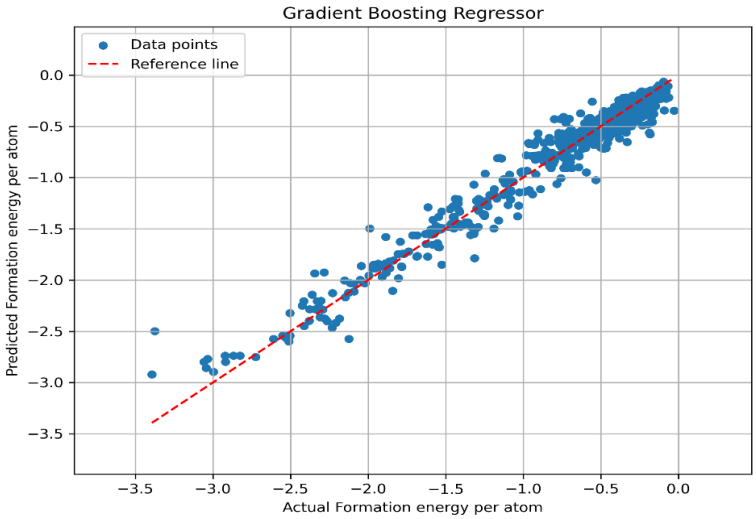

Figure 4. Predicted versus actual formation energy using GBR model.

On the other hand, GBR shows remarkable predictive accuracy. With the optimal hyperparameters mentioned in Table 1, this model achieves lower values of MAE, MSE, and RMSE than RFR. The higher value in R^2 and adjusted R^2 show that the model is performing slightly better than the RFR. Figure 4 shows the comparison between the predicted and actual values of formation energy per atom performed in GBA. Additionally, the adjusted value of R^2 is 0.94 revealing approximately 94% of the variance in formation energy is explained by the model.

Further, the performance differences between RFR and GBR can be attributed to their inherent algorithmic strengths. While RFR is known for its robustness and resistance to overfitting, GBR leverages an ensemble of weak learners to iteratively minimize errors, leading to slightly superior predictions. Both models, however, benefit significantly from the use of the Magpie Elemental Property Featurizer, which effectively captures the compositional features of the alloys. These findings underscore the potential of machine learning models to accurately and efficiently predict complex material properties such as formation energy. The ability of GBR to achieve high accuracy with fewer features and lower error metrics highlights its suitability for predictive tasks in materials science. By providing reliable predictions, these models can significantly accelerate the process of material discovery and optimization.

CONCLUSIONS

Both models, Gradient Boosting Regressor and Random Forest Regressor, demonstrated exceptional accuracy in predicting the formation energy of Cu-based ternary alloys, as evidenced by the close alignment of data points around the reference line. This strong correlation indicates that the models effectively capture the underlying patterns in the dataset. Furthermore, the low error metrics reinforce the reliability and robustness of the developed models. The results highlight the potential of machine learning in materials science, particularly in predicting complex material properties. Among the tested models, the Gradient Boosting Regressor emerged as a standout performer, explaining 94% of the variance using only five features. In comparison, the Random Forest Regressor accounted for 92.5% of the variance, requiring ten features. The efficiency and precision of the Gradient Boosting Regressor make it a promising tool for material scientists aiming to understand and predict material behavior more effectively. By enabling faster and more accurate predictions, these machine learning models could significantly accelerate the process of material discovery and design. The Gradient Boosting Regressor, in particular, has the potential to streamline research workflows and inspire innovative approaches in the study of Cu-based alloys and beyond. This study underscores the transformative role of machine learning in addressing complex challenges in materials science, paving the way for data-driven advancements in the field.

ACKNOWLEDGEMENTS

The authors are thankful to Office of Campus Chief and Department of Physics, Mahendra Morang Adarsh Multiple Campus, Tribhuvan University, Biratnagar, Nepal for providing computational facility.

AUTHOR CONTRIBUTIONS

S. Dahal: Conceptualized the project, did the computational works and drafted the manuscript. D. Adhikari: helped in drawing important conclusions and reviewed the manuscript. S.K. Yadav: edited the written content, assisted with data curation and validation, and supervised the research.

CONFLICT OF INTEREST

Authors declare that there is no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

REFERENCES

Alade, I.O., Abd Rahman, M.A., Abbas, Z., Yaakob, Y., & Saleh, T.A. (2020). Application of support vector regression and artificial neural network for prediction of specific heat capacity of aqueous nanofluids of copper oxide. *Solar Energy, 197*, 485–490. https://doi.org/10.1016/j.solener.2020.01.022.

Aldosari, M.N., Yalamanchi, K.K., Gao, X., & Sarathy, S.M. (2021). Predicting entropy and heat capacity of hydrocarbons using machine learning. *Energy and AI, 4*, 100054. https://doi.org/10.1016/j.egyai.2021.100054.

Anaconda Inc. (2020). *Anaconda software distribution*. Retrieved September 02, 2024, from https://www.anaconda.com.

Bitencourt-Ferreira, G., & de Azevedo, W.F. (2018). Development of a machine-learning model to predict Gibbs free energy of binding for protein-ligand complexes. *Biophysical Chemistry, 240*, 63–69. https://doi.org/10.1016/j.bpc.2018.07.002

Breiman, L. (1997). Arcing the edge. *Technical Report No. 486*. Statistics Department, University of California, Berkeley.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., & et al.. (2013). API design for machine learning software: Experiences from the Scikit-learn project. *arXiv preprint arXiv:1309.0238*.

Dasgupta, R. (2014). A look into Cu-based shape memory alloys: Present scenario and future prospects. *Journal of Materials Research, 29*(16), 1681–1698. https://doi.org/10.1557/jmr.2014.196.

Desgranges, C., & Delhommelle, J. (2018). A new approach for the prediction of partition functions using machine learning techniques. *Journal of Chemical Physics, 149*(4), 044118. https://doi.org/10.1063/1.5031861.

Dhungana, A., Yadav, S.K., Mehta, U., Novakovic, R., & Adhikari, D. (2023). Thermodynamic and surface properties of liquid Al-Cu-Ni alloys. *Journal of Materials Engineering and Performance*, 1–11. https://doi.org/10.1007/s11665-023-07715-y.

Faber, F.A., Lindmaa, A., Von Lilienfeld, O.A., & Armiento, R. (2016). Machine learning energies of 2 million elpasolite (ABC2D6) crystals. *Physical Review Letters, 117*(13), 135502. https://doi.org/10.1103/PhysRevLett.117.135502.

Faber, F.A., Lindmaa, A., Von Lilienfeld, O.A., & Armiento, R. (2015). Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry, 115*(16), 1094–1101. https://doi.org/10.1002/qua.24917.

Ho, T.K. (1995). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278–282). IEEE. https://doi.org/10.1109/ICDAR.1995.598994.

Huang, H., Chen, B., Hu, X., Jiang, X., Li, Q., Che, Y., Zu, S., & Liu, D. (2022). Research on Bi contents addition into Sn-Cu-based lead-free solder alloy. *Journal of Materials Science: Materials in Electronics, 33*(19), 15586–15603. https://doi.org/10.1007/s10854-022-08983-w.

Inoue, A., Zhang, W., Zhang, T., & Kurosaka, K. (2001). High-strength Cu-based bulk glassy alloys in Cu-Zr-Ti and Cu-Hf-Ti ternary systems. *Materials Transactions, JIM, 42*(7), 1147–1152. https://doi.org/10.2320/matertrans1989.42.1147.

Islam, M.N., Chan, Y., Rizvi, M.J., & Jillek, W. (2005). Investigations of interfacial reactions of Sn-Zn based and Sn-Ag-Cu lead-free solder alloys as replacement for Sn-Pb solder. *Journal of Alloys and Compounds, 400*(1–2), 136–144. https://doi.org/10.1016/j.jallcom.2005.04.065.

Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., & Persson, K.A. (2013). Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials, 1*(1), 011002. https://doi.org/10.1063/1.4812323.

Jani, J.M., Leary, M., Subic, A., & Gibson, M.A. (2014). A review of shape memory alloy research, applications, and opportunities. *Materials & Design, 56*, 1078–1113. https://doi.org/10.1016/j.matdes.2013.11.084.

Kauwe, S.K., Graser, J., Vazquez, A., & Sparks, T.D. (2018). Machine learning prediction of heat capacity for solid inorganics. *Integrating Materials and Manufacturing Innovation, 7*(2), 43–51. https://doi.org/10.1007/s40192-018-0118-7.

Kosec, T., & Milosev, I. (2007). Comparison of a ternary Cu-18Ni-20Zn alloy and binary Cu-based alloys in alkaline solutions. *Materials Chemistry and Physics, 104*(1), 44–49. https://doi.org/10.1016/j.matchemphys.2007.03.015.

Mazzer, E.M., Da Silva, M.R., & Gargarella, P. (2022). Revisiting Cu-based shape memory alloys: Recent developments and new perspectives. *Journal of Materials Research, 37*(1), 162–182. https://doi.org/10.1557/s43578-021-00409-7.

Ohnuma, I., Miyashita, M., Anzai, K., Liu, X.J., Ohtani, H., Kainuma, R., & Ishida, K. (2000). Phase equilibria and the related properties of Sn-Ag-Cu-based Pb-free solder alloys. *Journal of Electronic Materials, 29*, 1137–1144. https://doi.org/10.1007/s11664-000-0078-2.

Olsthoorn, B., Geilhufe, R.M., Borysov, S.S., & Balatsky, A.V. (2019). Band gap prediction for large organic

crystal structures with machine learning. *Advanced Quantum Technologies, 2*(7–8), 1900023. https://doi.org/10.1002/qute.201900023.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N.E., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., & et al. (2018). Matminer: An open source toolkit for materials data mining. *Computational Materials Science, 152*, 60–69. https://doi.org/10.1016/j.commatsci.2018.05.018.

Xia, Y., Xie, X., Xie, X., & Lu, C. (2006). Intermetallic compounds evolution between lead-free solder and Cu-based lead frame alloys during isothermal aging. *Journal of Materials Science, 41*, 2359–2364. https://doi.org/10.1007/s10853-006-7513-3.

Zhang, Y., Dang, S., Chen, H., Li, H., Chen, J., Fang, X., Shi, T., & Zhu, X. (2024). Advances in machine learning methods in copper alloys: A review. *Journal of Molecular Modeling, 30*(12), 398. https://doi.org/10.1007/s00894-024-05549-3.

Zhou, Y., Jing, G., Yiting, G., Jun, W., Yan, W., Xiaoxiao, H., Jun, C., Quanjin, L., Qiang, W., & Chenlong, W. (2022). Prediction of formation energies of UCr4C4-type compounds from Magpie feature descriptor-based machine learning approaches. *Optical Materials: X, 16*, 100196. https://doi.org/10.1016/j.omx.2022.100196.

Zhuo, Y., Mansouri Tehrani, A., & Brgoch, J. (2018). Predicting the band gaps of inorganic solids by machine learning. *Journal of Physical Chemistry Letters, 9*(7), 1668–1673. https://doi.org/10.1021/acs.jpclett.8b00124.