A Comparison of Explainable Artificial Intelligence Methods in the Phase Classification of Multi-Principal Element Alloys

Kyungtae Lee
University of Virginia
Mukil V. Ayyasamy
University of Virginia
Yangfeng Ji
University of Virginia
Prasanna V. Balachandran (pvb5e@virginia.edu)
University of Virginia

Research Article

Keywords:

Posted Date: February 24th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1389922/v1

A Comparison of Explainable Artificial Intelligence Methods in the Phase Classification of Multi-Principal Element Alloys

Kyungtae Lee[1], Mukil V. Ayyasamy[1], Yangfeng Ji[2], and Prasanna V. Balachandran[1,3,*]

[1]Department of Materials Science and Engineering, University of Virginia, Charlottesville, VA 22904, USA
[2]Department of Computer Science, University of Virginia, Charlottesville, VA 22904, USA
[3]Department of Mechanical and Aerospace Engineering, University of Virginia, Charlottesville, VA 22904, USA
[*]pvb5e@virginia.edu

ABSTRACT

We demonstrate the capabilities of two model-agnostic local post hoc model interpretability methods, namely breakDown (BD) and shapley (SHAP), to explain the predictions of a black-box classification learning model that establishes a quantitative relationship between chemical composition and multi-principal element alloys phase formation. We trained an ensemble of support vector machines using a dataset with 1,821 instances, 12 features with low pair-wise correlation, and seven phase labels. Feature contributions to the model prediction are computed by BD and SHAP for each composition. The resulting BD and SHAP transformed data are then used as inputs to identify similar composition groups using k-means clustering. Explanation-of-clusters by features reveal that the results from SHAP agree more closely with the literature. Visualization of compositions within a cluster using Ceteris-Paribus (CP) profile plots show the functional dependencies between the feature values and predicted response. Despite the differences between BD and SHAP in variable attribution, only minor changes were observed in the CP profile plots. Explanation-of-clusters by examples show that the clusters that share a common phase label contain similar compositions, which clarifies the similar-looking CP profile trends. In the limits of a dataset with independent and non-interacting features, SHAP and BD appear to capture similar pattern.

Introduction

Machine learning (ML) techniques have been actively utilized for various applications in recent years due to their superior prediction performance. However, as the applications expand vastly, they also have been required to meet the demand of high level of transparency and accountability due to their black-box framework[1–9]. As a result, explainable ML methods have attracted a great deal of attention in order to further advance the reliability and applicability of ML-based approaches. Model explainability is also becoming increasingly important in the materials science domain as ML and artificial intelligence (AI)-driven algorithms are beginning to show success in simplifying various workflows[10–22]. However, the idea of incorporating explainable ML methods into the current data-driven materials design and discovery workflow is still in its infancy. Unlike many other disciplines, materials science suffers from limited and sparse data. Incorporating interpretablility into the black-box ML models make it possible to explain the ML predictions with more confidence, especially when the models are trained using small and heterogeneous datasets. Explainable ML methods improve the trustworthiness of the trained models by uncovering the biases and errors inside of the models, which would not be identified otherwise using standard metrics such as goodness-of-fit (R^2, mean squared error, mean absolute error, precision, and recall to name a few). Overall, the interpretable and explainable ML framework makes the ML paradigm more accessible to domain experts by the virtue of making the learning process understandable.

One of the representative and popular methods for model explanations is SHapley Additive exPlanations (SHAP) developed by Lundberg and Lee[23]. SHAP is a local-agnostic explanation method based on the SHAP values that were introduced in cooperative game theory to calculate the contributions of players to the total payout[24]. In predictive models, the contribution of each variable can be calculated by averaging over every possible ordering of variables using SHAP, allowing to locally analyze the importance of each input feature for a given instance prediction. Thus, the SHAP method calculates each feature contribution by evaluating a change in the expected model prediction when conditioned on a given feature. In the SHAP algorithm, each feature is assigned an importance value that represents each individual effect on the model prediction as a result of their inclusion in the model. To calculate the importance values, a model is trained in the presence of each corresponding feature, while another model is also trained with the feature withheld. Finally, the predictions from those two models are compared and the SHAP values are computed by averaging the differences for all possible subsets as the effect of withholding a

feature is influenced by other features in the model[24]. This SHAP analysis is particularly useful when one needs to work on specific prediction. In addition, the global understanding of a model can be enhanced by exploiting collective SHAP values. For this reason, the SHAP method has been the predominant method used to analyze machine learning results in various materials science publications on alloys, catalysts, photovoltaics, metal-organic frameworks and oxide glasses to name a few[13,25–29].

In this paper, in addition to SHAP, we will focus on two other instance-level model interpretability approaches, namely breakDown (BD) analysis and Ceteris-Paribus (CP) profiles, that have received little attention in the materials science community. The BD method, similar to the SHAP method, is also based on the variable attribution principle that decomposes the prediction of each individual observation into particular variable contributions[30,31]. Unlike the SHAP values, the BD values provide order specific explanations of variables' contributions in a greedy way[5]. One of the assumptions of BD method is that the input features are independent and non-interacting[32]. There are two algorithms for BD analysis: (1) step-down and (2) step-up. The step-down method starts from a full set of input features. Then, each individual feature contribution is calculated by sequentially removing a single feature from a set followed by variable relaxation in order that the distance to the model prediction is minimized. The step-up method starts from a null set and follows the opposite direction of the step-down method. Both methods have been shown to provide consistent outcomes in feature contributions. At a cursory glance, these assumptions and definitions appear to suggest that the BD method for post hoc model interpretability may find a natural home in explaining hierarchical learning algorithms that preserve a taxonomy. However, Gosiewska and Biecek[33] argue that complex predictive models are usually non-additive and model interpretation should depend on the order in which the explanation is read. Therefore, setting a proper visit order may lead to a better and intuitive understanding of the model prediction. There are papers in the published literature, where the BD method is used for post hoc model interpretability in a non-hierarchical learning setting[5,34–36]. In this study, the BD results are discussed by referring to the BD data from our recent published work[37], where the step-down method was used.

On the other hand, the CP profiles, also referred to as individual conditional expectations (ICE) plots, evaluate the influence of a variable from a trained ML model under the assumption that the values of all other variables are fixed (what-if analysis)[38]. The dependence between the predicted response and a feature is visualized by CP profiles, where one can observe how the input and responses are related at a glance (e.g. in a linear, non-linear or more complex pattern). In this way, the CP analysis helps to quantify the impact of a given variable on the predictions of a black-box model and provide a cursory explanation of the functional form connecting an input with the output. Insights about such a functional dependence is not readily ascertained from the SHAP and BD analysis. As a result, there is immense value in complementing SHAP and BD analyses with CP profile plots.

Although the BD, SHAP and CP profile plots are representative methods for post hoc model interpretation, the comparison of these methods have been rarely reported in ML literature. Lorentzen and Mayer analyzed the policies of motor third-party liability and compared the SHAP decomposition with the BD method[34]. However, the differences were not discussed in sufficient detail. Rinzivillo et al. also reported both SHAP and BD decompositions, but made a simple comparison on the feature importance plots from these methods without a further analysis about the differences[35]. Gosiewska and Biecek examined the dataset about the sinking of the Titanic using BD and SHAP. They presented a different aspect of feature importance from each method, mentioning that it was not clear to determine which one is more reliable. Their local interpretability was modified by including the interaction factors between descriptors, but no comparison between BD and SHAP was made thereafter[33]. Thus, a comparative study on BD and SHAP is necessitated to better understand the usage and effectiveness of these methods.

In this paper, we present a rigorous comparative study between SHAP and BD on the phase classification problem of multi-principal element alloys (MPEAs). The MPEAs make a good template for this study because the dataset is high-dimensional and has 1,821 compositions in seven class labels (additional details given in the Methods Section). The MPEA phase classification problem is well-studied in the literature[39–44], which provides us with sufficient domain knowledge to compare the outcomes from SHAP and BD results. Previously, we reported a ML study result about BD analysis based an ensemble of support vector machine (eSVM) along with the *k*-means clustering method, followed by CP analysis[37]. This work focuses on providing a detailed comparison between BD and SHAP from the perspective of local post hoc model interpretability of black-box models. Specifically, the local feature importance weights for each instance are separately computed by the BD and SHAP methods. The resulting two different sets of variable attribution data are then independently clustered by the *k*-means clustering algorithm. Both BD and SHAP approaches turn out to be successful in capturing the similarities between the compositions that represent the body-centered cubic (BCC), face-centered cubic (FCC), and amorphous (AM) phases. Explanation-of-clusters by features reveal commonalities and differences between BD and SHAP values. Although the SHAP results are found to be marginally more consistent with the literature, both BD and SHAP values capture well the known findings in the published literature. Despite approaching local feature importance from different perspectives and assumptions, explanation-of-clusters by examples reveal grouping of similar compositions within the clusters. A detailed analysis of the CP profile plots for the compositions within the clusters show similar functional dependencies between the feature values and predicted responses.

Results

In Fig. 1, we discuss the explanation-by-example using the NbTaTiV composition as a template, which is predicted to form in BCC by our eSVM model. Both BD contributions and SHAP values are expressed as bar graphs, where positive and negative values indicate the contribution of each variable to the overall prediction. Both plots show the importance of mean_meltingT, maxdiff_NUnfilled, and maxdiff_AtomicWeight due to the large weights, and thus are identified as dominant for predicting the BCC phase. Some differences in the BD and SHAP variable attributions can also be seen. The Dev_NdValence and mean_ConvalentRadius are identified as important variables by SHAP, whereas the BD analysis identifies mean_NValence and mean_NsValence as important. Our key point from Fig. 1 is the following: even in this simple case, it is important to note that the SHAP and BD results do not completely agree with each other. This is not entirely surprising because in SHAP the contribution of a feature is averaged over all possible conditional expectations, whereas in the greedy BD method only a single order (Fig. 1a) of conditional expectation is performed[5]. In a similar vein, we have calculated the BD and SHAP values for every alloy composition in the training data.

In order to better understand this difference between the BD and SHAP methods, we now transition to model explainability at an intermediate level, where the instances are clustered based on similarities of their attributes as represented by the BD and SHAP values. Although the variable attribution from SHAP and BD for an individual instance may appear differently, does this mean they capture vastly different pattern in the data? The main purpose is to understand and explain similar compositions within a cluster. From the materials science standpoint, clusters are expected to shed insights into the key factors that govern the formation of a particular phase as a function of chemical composition. As a result, there is value in such analysis.

In our earlier work[37], we introduced a novel algorithm (see Supplementary Algorithm S1) that integrated BD values, *k*-means clustering and CP profile plots based on the eSVM framework. We briefly discuss the significance of this algorithm. We have a high-dimensional input feature space and when we visualized it using the t-distributed stochastic neighbor embedding (t-SNE) algorithm[45] we found that the MPEA phases are non-separable (Supplementary Figure S1a). The features in the input space have low pair-wise correlation coefficient (see Methods), thus the independence assumption is weakly satisfied. We also do not explicitly code feature interactions in the input space. The supervised eSVM method with radial-basis kernel function generates a non-linear decision boundary that led to a classification accuracy of 86% on the test data (see Methods). The next step involves performing post hoc model interpretability analysis of the trained eSVM model based on the BD method. We interpret that the BD analysis reformulates the non-linear eSVM mapping into an approximated linear space by using locally linear approximation, where we have a variable importance or weight associated with each input feature. We then redefine every composition (or instance) in our dataset in terms of the variable importance as calculated from the BD analysis. The redefined dataset serves as an input for *k*-means clustering, which in turn identifies similar composition groups. We then visualized the clusters by CP profile plots to explain the similarities and differences. Thus our algorithm provides a unique way to understand the outcomes from the post hoc interpretability methods.

In this work, we performed the SHAP analysis (instead of the BD analysis) that will serve as the input for *k*-means clustering algorithm. The detailed steps are shown in Algorithm 1. The output consists of groups of compositions (or clusters) with similar SHAP values. Once the clusters are identified, we then revisit the individual compositions within each cluster and take an average of the calculated SHAP values and CP profiles. The optimal numbers of clusters from the *k*-means clustering algorithm were identified by plotting the total within sum of square versus the number of clusters (Supplementary Figure S2). The elbow point is (approximately) located at 10 clusters. To analyze the results of *k*-means clustering, we plotted the frequency of occurrence of the number of components in the alloy composition for each individual cluster in Fig. 2. This analysis helps to exclude some clusters that are representative of the binary alloys (our primary interest lies in the high entropy alloys, which normally comprise of more than four components). For this reason, clusters 1, 2, 3, 8, and 10 are down-selected from the SHAP results. Our previous BD work[37] focused on clusters that are representative of the BCC, AM and FCC phases. We now compare the patterns recognized from the SHAP and BD clustering analyses to explain the similarities and differences.

The local variable attribution analysis based on the *k*-means clustering results are given in Figs. 3, 4, and 5. These clusters can be labeled as containing composition groups that form in BCC, AM, and FCC phases. With respect to the BCC phase (Fig. 3), both the BD and SHAP methods identify mean_MeltingT and maxdiff_NUnfilled variables as the significant variables. The mean_DeltaHf, dev_NdValence, and mean_CovalentRadius are found to be unique to SHAP, whereas the mean_NValence, mean_NsValence, and maxdiff_AtomicWeight appear dominant only in the BD results. According to the literature data[46–48], atomic size mismatch (maxdiff_AtomicWeight and mean_CovalentRadius) and mixing enthalpy (represented using mean_DeltaHf) are identified as important for BCC phase formation. Thus, both SHAP and BD results agree with previous reports.

In Figs. 4a and b, we compare the BD and SHAP results that are representative of the compositions that are grouped in the AM clusters. Both BD and SHAP methods identify the importance of MixingEntropy and maxdiff_AtomicWeight variables. The importance of mean_CovalentRadius variable is more apparent in the averaged SHAP values, compared to the averaged BD contributions. Since the AM phase formation is also affected by atomic size mismatch, the SHAP result

appears to agree marginally better with previous reports than the BD result. The SHAP result also stresses the importance of mean_DeltaHf and dev_NdValence. On the other hand, the BD analysis identifies mean_NValence, mean_NsValence, and maxdiff_Electronegativity as important.

Lastly, the clustering results that are reflective of the FCC phase is compared in Figs. 5a and b based on the averaged BD and SHAP data, respectively. Three common important descriptors identified by both BD and SHAP include: (1) mean_NValence, (2) frac_pValence, and (3) maxdiff_AtomicWeight. While the importance of mean_MeltingT and mean_CovalentRadius variables are emphasized by the SHAP values, the BD contributions identify mean_NsValence and maxdiff_Electronegativity as important. According to our previous work[37], the constituent elements of MPEA compositions that form in the FCC phase span the first and second rows of the *d*-block elements from the periodic table. These elements have a wide range of valence electron number. In contrast, the BCC phase mostly forms in MPEA compositions with early transition-series metals (tighter range of valence electron number). Thus, the mean_NValance, mean_MeltingT and mean_CovalentRadius descriptors are likely to have larger variations in the FCC cluster than in the BCC cluster, providing a better explanation about the formation of FCC phase. From this context, both BD and SHAP results are relevant in capturing the insights that govern the FCC phase formation.

The results from clustering analysis of BD and SHAP data were further analyzed through the lens of the averaged CP profile plots. In Fig. 6, we show the averaged CP profile plots for the compositions in the FCC cluster. The black dots in each panel indicate the feature values for each data point within the cluster. There were a total of 152 or 125 compositions that were labeled as FCC in the cluster depending on whether the inputs were from the BD or SHAP data, respectively. When we compared the specific MPEA compositions in the two clusters, we found an overwhelming 87% overlap between them. The pie charts shown in Figs. 7a and b reveal the similar elemental distributions in the two FCC clusters. This is an intriguing result. Despite the differences in the calculation of variable attribution (shown in Fig 5a and b), we find that similar alloy compositions are grouped in each cluster. This consistency is also reflected in the CP profile plots shown in Figs. 6a and b, where the functional dependencies between each variable and the phases follow a similar pattern. Some minor differences can be found between the BD and SHAP data in the CP profile plots. For example, consider the maxdiff_AtomicWeight variable. In Fig. 6a, the intersection (or the decision boundary) between the cyan (representing FCC phase) and blue (representing presence of Mixed Phases in the microstructure) curves occur at about 0.5 for the normalized maxdiff_AtomicWeight variable. However, the same two curves intersect at $\sim$0.25 for the same variable in Fig. 6b. Nonetheless, the overarching functional dependency looks similar. When we analyzed the averaged CP profile plots for the compositions within the BCC (Supplementary Figure S3) and AM clusters (Supplementary Figure S4), a similar behavior was found. The BCC and AM clusters also show consistent elemental distributions between the two methods (Supplementary Figures S5). In fact, the AM clusters had identical alloy compositions. Data visualization based on t-SNE algorithm show that the BD (Supplementary Figure S1b) and SHAP datasets (Supplementary Figure S1c) are able to separate the clusters well and have similar patterns. This result further reinforces the similarity between BD and SHAP methods for our chosen dataset.

Discussion

The capabilities of BD and SHAP as post hoc model interpretability methods were demonstrated on the local feature analysis of the MPEA phase classification problem using the black-box eSVM method, *k*-means clustering, and CP profile plots. The BD and SHAP methods did not produce identical variable attribution. There were some agreements and disagreements, which can be directly ascribed to the way conditional estimates are calculated between the two methods. The SHAP analysis is more robust (and hence expensive), whereas the BD method is greedy. In-depth analysis via clustering provided the meaningful data for substantive discussion. Both BD and SHAP results consistently identify maxdiff_AtomicWeight and mean_CovalentRadius as key factors in impacting the formation of BCC and AM phases. Explanation-of-clusters by features revealed that the SHAP results are relatively more consistent with the existing understanding of BCC and AM phase formation. Both BD and SHAP methods capture the key descriptors that govern the formation of FCC phase. Intriguingly, explanation-of-clusters by examples revealed that both methods identify similar alloy composition groupings. Visualization based on CP profile plots show more or less identical pattern in the learned feature trends. Despite the substantive difference in the instance-level variable attribution, *k*-means clustering uncovered the common trends between the two methods. We attribute the similarity in the BD and SHAP results to the presence of independent and non-interacting input features. Under this special setting, as Molnar also noted[7], the differences between the greedy BD method and the exhaustive SHAP method appear to be negligible.

Methods

Dataset, Feature Selection, and Machine Learning

The dataset used in this work contains 1,821 compositions, which was created by collecting and preprocessing data from published literature. Each composition is labeled as BCC, FCC, BCC+FCC, hexagonal-closed packed (HCP), AM, Intermetallics (IM), or Mixed-phases (MP) depending on the experimentally determined X-ray diffraction data. A total of 125 input features

were added to the dataset using the Magpie program[49] and down-selected by pair-wise Pearson correlation coefficient (PCC)[50] within the RSTUDIO environment[51]. Two different thresholds of PCC < 0.4 and PCC < 0.6 were considered for the feature selection step. As there is no standard approach to decide the threshold, we referred to a report from Pei *et al.*[42] and started from a PCC criterion of 0.6 and then examined a more stricter PCC value of 0.4 to look for further simplification. The correlation analysis resulted in two pre-processed datasets containing 12 and 20 variables for PCC < 0.4 and PCC < 0.6, respectively.

Each dataset was randomly divided into two portions, with 75% for training and 25% for testing. For multi-class classification learning, we constructed the eSVM model where multiple SVM models were generated by the bootstrap sampling method[52,53] and the nonlinear Gaussian radial basis function kernel was used using the e1071 package[54]. The eSVM hyperparameters were determined using the out-of-bag evaluation based on grid search. The eSVM model performed with a classification accuracy of 86% on the independent test set. The performance metrics are given in Supplementary Table S1. As there was no difference in prediction performance between 12 and 20 feature sets, we selected the simpler 12 feature set which is listed in Table 1 and used for post hoc local feature analysis to compare SHAP with BD. The details of the dataset and ML model building have been described in our previous paper[37].

breakDown, SHAP and Ceteris-Paribus Profile Plots

Post hoc model interpretability analysis of black-box models can be broadly classified into two types. One is the global variable importance analysis, where the feature importance is evaluated based on an entire dataset. The other is the local variable importance analysis such as BD and SHAP, where feature importance is evaluated for each individual instance. Unlike BD and SHAP, the CP profile plots help in visualizing the relationship between each feature and the predicted response for each instance or compositions in our dataset. The BD, SHAP, and CP methods from the DALEX package[55] were used for post-hoc local model interpretation, where the contributions of each descriptor to ML prediction were calculated. We performed clustering analysis using the result from BD or SHAP data. We used the *k*-mean clustering algorithm as implemented in the factoextra package[56]. For the local feature importance analysis of each cluster, we averaged the BD or SHAP data of those instances and computed the local feature contributions and CP profiles. In Supplementary Table S2, the cluster numbers from *k*-means clustering are linked to the appropriate phase labels for the ease of interpretation. To visualize the high-dimensional data in two-dimensions, we used the t-distributed stochastic neighbor embedding (t-SNE) method from the Rtsne package[57] with a perplexity value of 350 and a learning rate of 200.

Acknowledgments

Research was sponsored by the Defense Advanced Research Project Agency (DARPA) and The Army Research Office and was accomplished under Grant Number W911NF-20-1-0289. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of DARPA, the Army Research Office, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Competing Interests

The authors declare that there are no competing interests.

Author Contribution Statements

The study was planned by K.L and P.V.B. The manuscript was prepared by K.L, M.V.A, Y.J and P.V.B. The data set construction was done by K.L. The machine learning studies were performed by K.L and M.V.A. The interpretation of the results were performed by K.L, Y.J and P.V.B. All authors discussed the results, wrote, and commented on the manuscript.

Data Availability

The dataset used for the ML study is freely available in our Web App (https://adaptivedesign.shinyapps.io/AIRHEAD/) and on Figshare[58].

References

1. Edwards, L. & Veale, M. Enslaving the algorithm: From a "right to an explanation" to a "right to better decisions"? *IEEE Secur. Priv.* **16**, 46–54, DOI: 10.1109/MSP.2018.2701152 (2018).

2. Patrick Hall, N. G. *An Introduction to Machine Learning Interpretability* (O'Reilly Media, Inc., 2018).

3. Bryce Goodman, S. F. European union regulations on algorithmic decision-making and a "right to explanation. *arXiv:1606.08813v3* (2016).
4. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci.* **116**, 22071–22080 (2019).
5. Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **23**, DOI: 10.3390/e23010018 (2021).
6. Chen, H. *et al.* Explaining neural network predictions on sentence pairs via learning word-group masks. *arXiv preprint arXiv:2104.04488* (2021).
7. Molnar, C. *Interpretable machine learning* (Lulu.com, 2020).
8. Molnar, C., Casalicchio, G. & Bischl, B. Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 417–431 (Springer, 2020).
9. Staniak, M. & Biecek, P. Explanations of Model Predictions with live and breakDown Packages. *The R Journal* **10**, 395–409, DOI: 10.32614/RJ-2018-072 (2018).
10. Stach, E. *et al.* Autonomous experimentation systems for materials development: A community perspective. *Matter* **4**, 2702–2726 (2021).
11. Kusne, A. G. *et al.* On-the-fly closed-loop materials discovery via Bayesian active learning. *Nat. communications* **11**, 5966 (2020).
12. Kim, Y., Kim, E., Antono, E., Meredig, B. & Ling, J. Machine-learned metrics for predicting the likelihood of success in materials discovery. *npj Comput. Mater.* **6**, 131 (2020).
13. Gurnani, R., Yu, Z., Kim, C., Sholl, D. S. & Ramprasad, R. Interpretable Machine Learning-Based Predictions of Methane Uptake Isotherms in Metal-Organic Frameworks. *Chem. Mater.* **33**, 3543–3552, DOI: 10.1021/acs.chemmater.0c04729 (2021).
14. Talapatra, A. *et al.* Autonomous efficient experiment design for materials discovery with Bayesian model averaging. *Phys. Rev. Mater.* **2**, 113803 (2018).
15. Ament, S. *et al.* Autonomous materials synthesis via hierarchical active learning of nonequilibrium phase diagrams. *Sci. advances* **7**, eabg4930.
16. Hart, G. L., Mueller, T., Toher, C. & Curtarolo, S. Machine learning for alloys. *Nat. Rev. Mater.* **6**, 730–755 (2021).
17. Balachandran, P. V. Adaptive machine learning for efficient materials design. *MRS Bull.* **45**, 579–586 (2020).
18. Pilania, G. Machine learning in materials science: From explainable predictions to autonomous design. *Comput. Mater. Sci.* **193**, 110360 (2021).
19. Vasudevan, R. K. *et al.* Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics. *MRS communications* **9**, 821–838 (2019).
20. Iwasaki, Y. *et al.* Identification of advanced spin-driven thermoelectric materials via interpretable machine learning. *npj Comput. Mater.* **5**, 103, DOI: 10.1038/s41524-019-0241-9 (2019).
21. Lu, Z. *et al.* Interpretable machine-learning strategy for soft-magnetic property and thermal stability in Fe-based metallic glasses. *npj Comput. Mater.* **6**, 187, DOI: 10.1038/s41524-020-00460-x (2020).
22. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 83, DOI: 10.1038/s41524-019-0221-0 (2019).
23. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777 (2017).
24. Shapley, L. S. *A value for n-person games* (Princeton University Press, 2016).
25. Kim, G. *et al.* First-principles and machine learning predictions of elasticity in severely lattice-distorted high-entropy alloys with experimental validation. *Acta Materialia* **181**, 124–138 (2019).
26. Witman, M. *et al.* Data-driven discovery and synthesis of high entropy alloy hydrides with targeted thermodynamic stability. *Chem. Mater.* **33**, 4067–4076 (2021).
27. Mine, S. *et al.* Analysis of updated literature data up to 2019 on the oxidative coupling of methane using an extrapolative machine-learning method to identify novel catalysts. *ChemCatChem* **13**, 3636–3655 (2021).

28. Hartono, N. T. P. *et al.* How machine learning can help select capping layers to suppress perovskite degradation. *Nat. communications* **11**, 1–9 (2020).

29. Zaki, M. *et al.* Interpreting the optical properties of oxide glasses with machine learning and shapely additive explanations. *J. Am. Ceram. Soc.* .

30. Staniak, M. & Biecek, P. Explanations of Model Predictions with live and breakDown Packages. *The R J.* **10**, 395, DOI: 10.32614/rj-2018-072 (2019).

31. Sykes, A. L. *et al.* Interpretable machine learning applied to on-farm biosecurity and porcine reproductive and respiratory syndrome virus. *Transboundary Emerg. Dis.* DOI: https://doi.org/10.1111/tbed.14369. https://onlinelibrary.wiley.com/doi/pdf/10.1111/tbed.14369.

32. Biecek, P. & Burzykowski, T. *Explanatory model analysis: Explore, explain and examine predictive models* (Chapman and Hall/CRC, 2021).

33. Gosiewska, A. & Biecek, P. Do not trust additive explanations. *arXiv preprint arXiv:1903.11420* (2019).

34. Lorentzen, C. & Mayer, M. Peeking into the black box: An actuarial case study for interpretable machine learning. *Available at SSRN 3595944* (2020).

35. Bodria, F. *et al.* Benchmarking and survey of explanation methods for black box models. *arXiv preprint arXiv:2102.13076* (2021).

36. Molnar, C., Bischl, B. & Casalicchio, G. iml: An R package for Interpretable Machine Learning. *J. Open Source Softw.* **3**, 786, DOI: 10.21105/joss.00786 (2018).

37. Lee, K., Ayyasamy, M., Delsa, P., Hartnett, T. Q. & Balachandran, P. V. Phase classification of multi-principal element alloys via interpretable machine learning. *npj Comput. Mater.* **8**, 25, DOI: https://doi.org/10.1038/s41524-022-00704-y (2022).

38. Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.* **24**, 44–65, DOI: 10.1080/10618600.2014.907095 (2015). https://doi.org/10.1080/10618600.2014.907095.

39. Huang, W., Martin, P. & Zhuang, H. L. Machine-learning phase prediction of high-entropy alloys. *Acta Materialia* **169**, 225–236, DOI: https://doi.org/10.1016/j.actamat.2019.03.012 (2019).

40. Zhou, Z. *et al.* Machine learning guided appraisal and exploration of phase design for high entropy alloys. *npj Comput. Mater.* **5**, 128, DOI: 10.1038/s41524-019-0265-1 (2019).

41. Kaufmann, K. & Vecchio, K. S. Searching for high entropy alloys: A machine learning approach. *Acta Materialia* **198**, 178–222, DOI: https://doi.org/10.1016/j.actamat.2020.07.065 (2020).

42. Pei, Z., Yin, J., Hawk, J. A., Alman, D. E. & Gao, M. C. Machine-learning informed prediction of high-entropy solid solution formation: Beyond the Hume-Rothery rules. *npj Comput. Mater.* **6**, 50, DOI: 10.1038/s41524-020-0308-7 (2020).

43. Zhang, Y. *et al.* Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models. *Acta Materialia* **185**, 528–539, DOI: https://doi.org/10.1016/j.actamat.2019.11.067 (2020).

44. Feng, S. *et al.* A general and transferable deep learning framework for predicting phase formation in materials. *npj Comput. Mater.* **7**, 1–10 (2021).

45. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. machine learning research* **9** (2008).

46. Takeuchi, A. & Inoue, A. Calculations of mixing enthalpy and mismatch entropy for ternary amorphous alloys. *Mater. Transactions, JIM* **41**, 1372–1378, DOI: 10.2320/matertrans1989.41.1372 (2000).

47. Takeuchi, A. & Inoue, A. Classification of bulk metallic glasses by atomic size difference, heat of mixing and period of constituent elements and its application to characterization of the main alloying element. *MATERIALS TRANSACTIONS* **46**, 2817–2829, DOI: 10.2320/matertrans.46.2817 (2005).

48. Zhang, Y., Zhou, Y., Lin, J., Chen, G. & Liaw, P. Solid-solution phase formation rules for multi-component alloys. *Adv. Eng. Mater.* **10**, 534–538, DOI: https://doi.org/10.1002/adem.200700240 (2008). https://onlinelibrary.wiley.com/doi/pdf/10.1002/adem.200700240.

49. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 16028, DOI: 10.1038/npjcompumats.2016.28 (2016).

50. John D. Kelleher, A. D., Brian Mac Namee. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies* (The MIT Press, 2020).

51. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2012). ISBN 3-900051-07-0.

52. Vapnik, V. *The Nature of Statistical Learning Theory* (Springer-Verlag New York, 2000).

53. MacKinnon, D. P., Lockwood, C. M. & Williams, J. Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. *Multivar. Behav. Res.* **39**, 99–128, DOI: 10.1207/s15327906mbr3901_4 (2004).

54. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien* (2015). R package version 1.6-7.

55. Biecek, P., Maksymiuk, S. & Baniecki, H. *moDel Agnostic Language for Exploration and eXplanation* (2021). R package version 2.2.0.

56. Kassambara, A. & Mundt, F. *Extract and Visualize the Results of Multivariate Data Analyses* (2020). R package version 1.0.7.

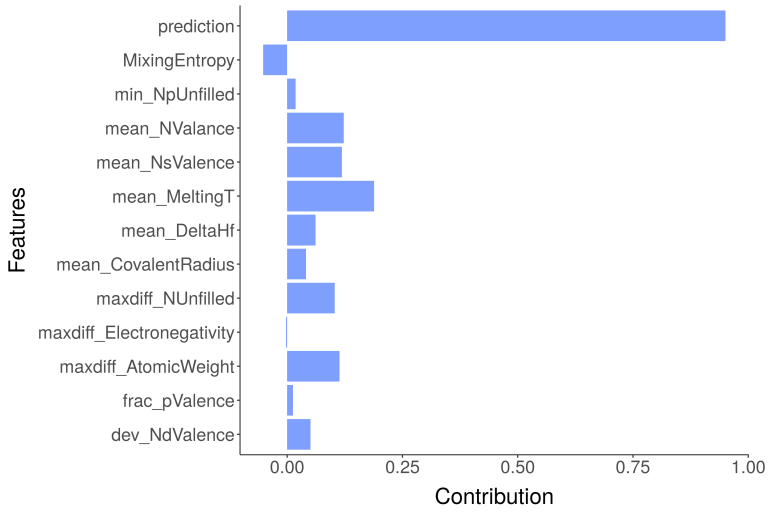
57. Krijthe, J., van der Maaten, L. & Krijthe, M. J. Package 'rtsne' (2018).

58. Lee, K., Ayyasamy, M. V., Delsa, P., Hartnett, T. Q. & Balachandran, P. V. Phase classification of multi-principal element alloys via interpretable machine learning. *figshare* DOI: https://doi.org/10.6084/m9.figshare.15098094.v1 (2021).

Table 1. List of the 12 descriptors selected from 125 descriptors by PCC > 0.4

Notation	Description
maxdiff_NUnfilled	Difference between minimum and maximum numbers of unfilled valence orbitals
min_NpUnfilled	Minimum number of unfilled p valence orbitals
maxdiff_AtomicWeight	Difference between minimum and maximum atomic weights
mean_NValance	Average number of filled valence electrons
mean_MeltingT	Average melting temperature
mean_NsValence	Average number of filled s valence electrons
dev_NdValence	Standard deviation of the number of filled d valence electrons
frac_pValence	Fraction of filled p valence electrons
MixingEntropy	Mixing entropy
mean_DeltaHf	Average mixing enthalpy
maxdiff_Electronegativity	Difference between minimum and maximum electronegativity values
mean_CovalentRadius	Average covalent radius of constituent elements

(a)



(b)

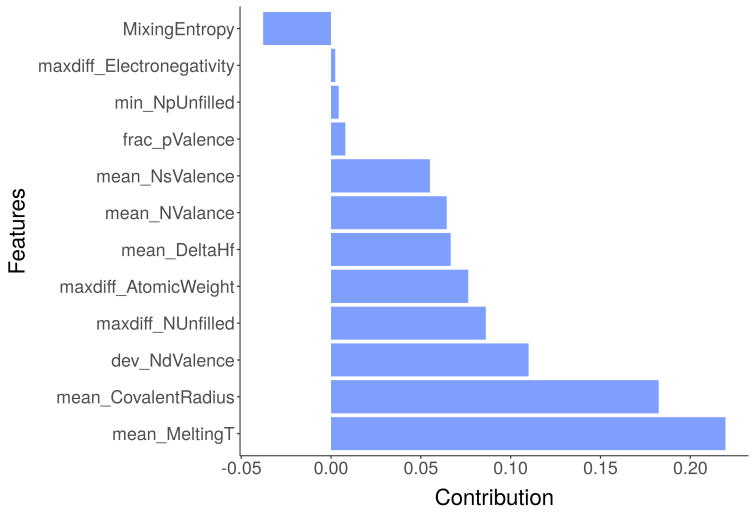


Figure 1. The (a) BD and (b) SHAP decomposition are plotted for the NbTaTiV composition, which is predicted to have BCC phase by the eSVM model. The sizes of each bar indicate the averaged contributions of the respective variables towards the overall prediction for a given instance.

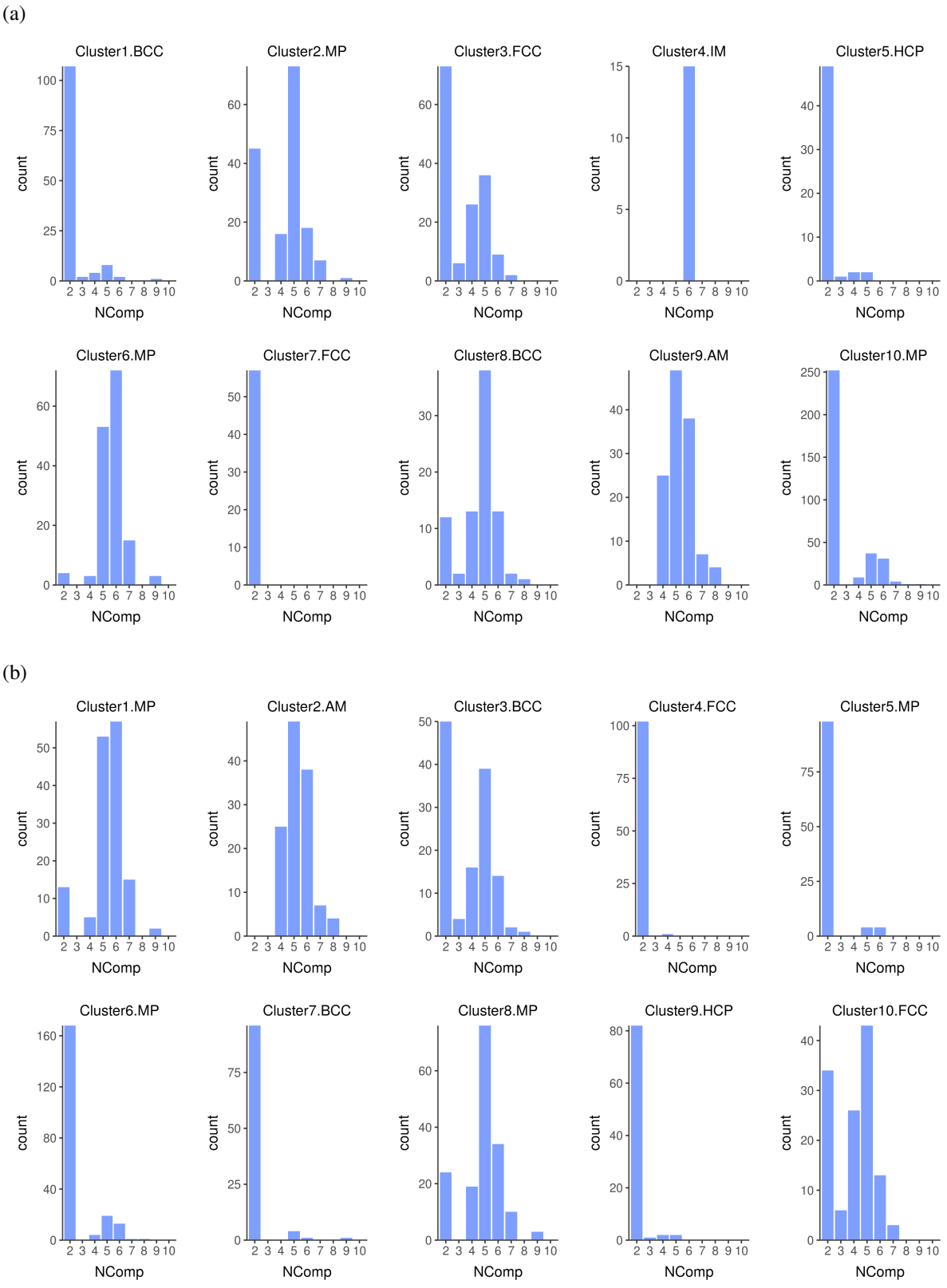
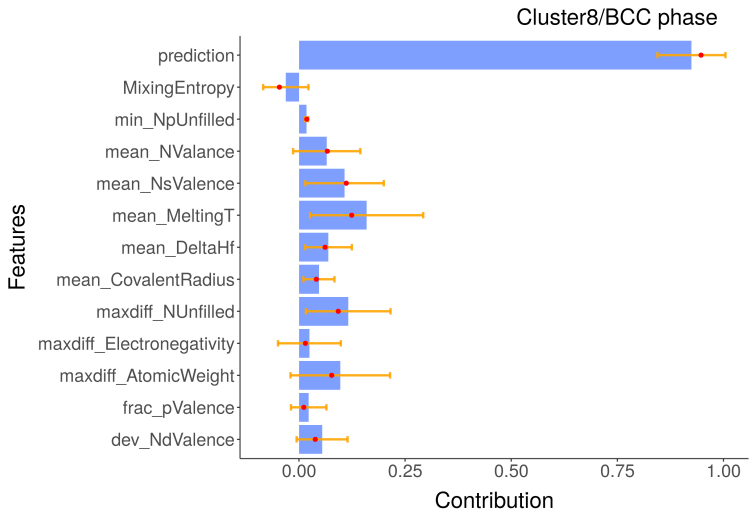


Figure 2. The distribution of the number of components (represented as NComp) for each cluster generated by (a) the BD-based and (b) SHAP-based data using the *k*-means clustering algorithm, respectively.

(a)



(b)

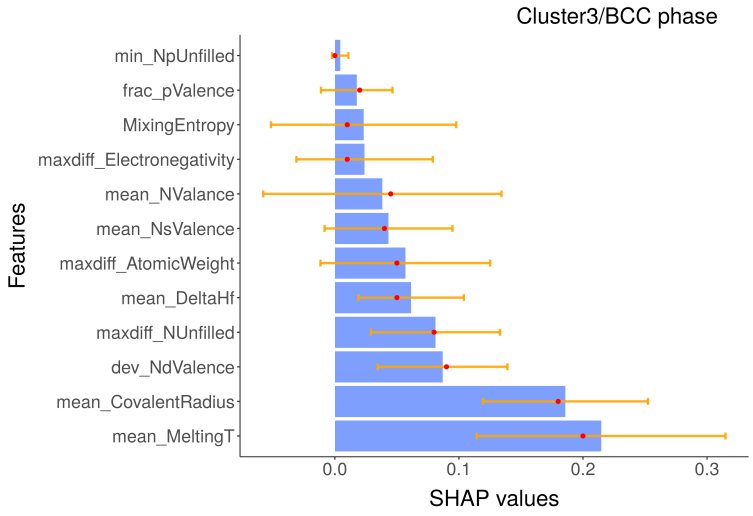
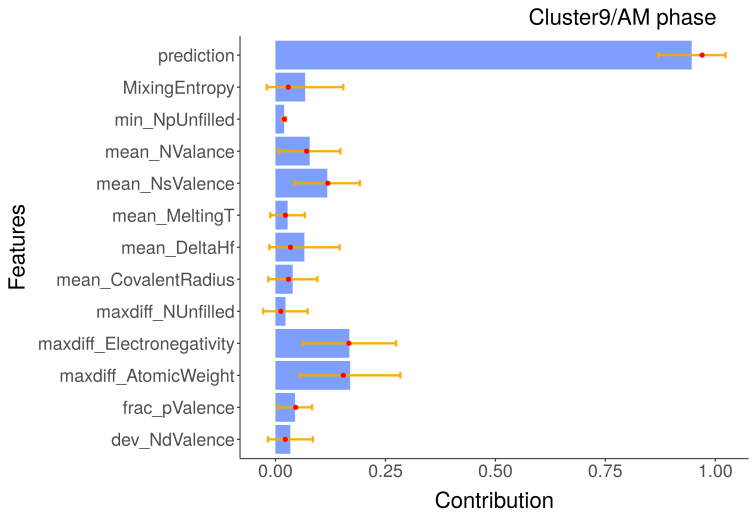


Figure 3. The (a) BD and (b) SHAP decomposition are plotted for the representative clusters (clusters 3 and 8, respectively), which are predicted to have BCC phase by the eSVM model. The sizes of each bar indicate the averaged contributions of the respective variables towards the overall prediction for a given instance. The first row indicates the sum of the overall mean prediction value and the changes. Red dots and yellow lines denote median values and error bars, respectively.

(a)



(b)

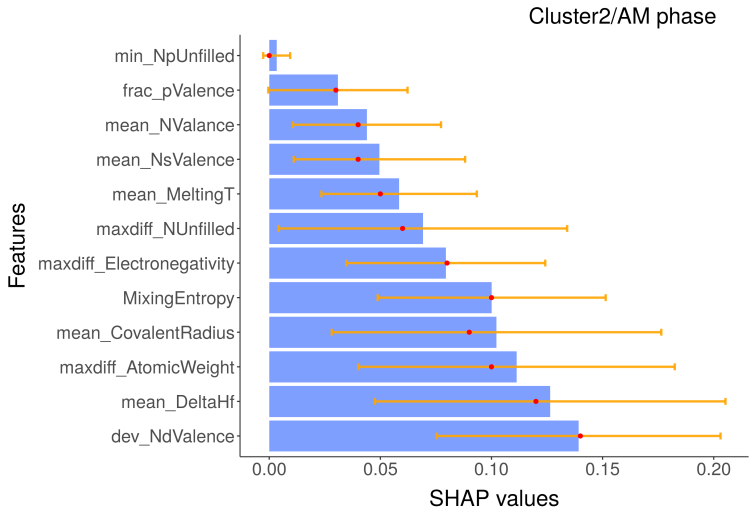
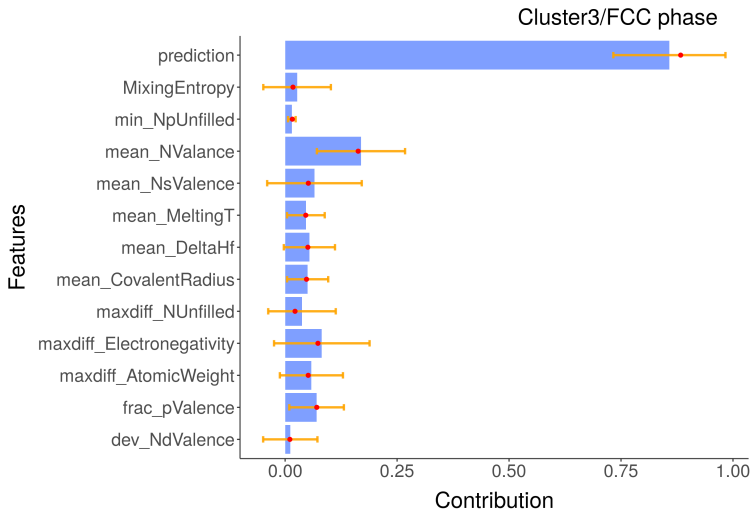


Figure 4. The (a) BD and (b) SHAP decomposition are plotted for the representative clusters (clusters 9 and 2, respectively), which are predicted to have AM phase by the eSVM model. The sizes of each bar indicate the averaged contributions of the respective variables towards the overall prediction for a given instance. The first row indicates the sum of the overall mean prediction value and the changes. Red dots and yellow lines denote median values and error bars, respectively.

(a)



(b)

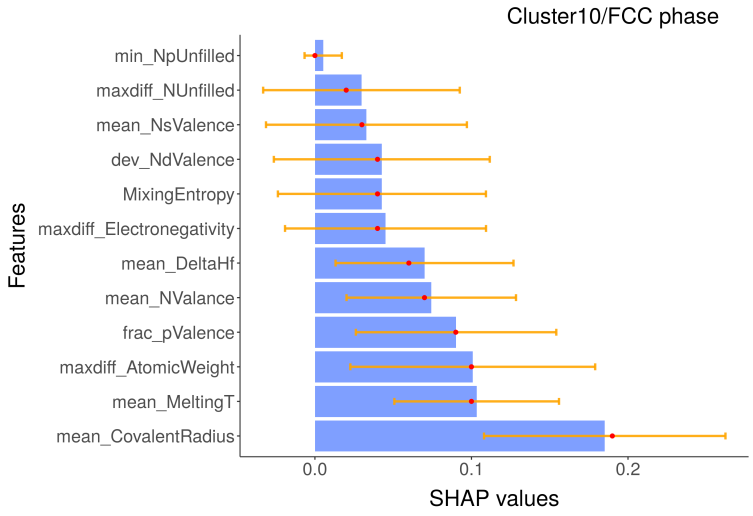


Figure 5. The (a) BD and (b) SHAP decomposition are plotted for the representative clusters (clusters 3 and 10, respectively), which are predicted to have FCC phase by the eSVM model. The sizes of each bar indicate the averaged contributions of the respective variables towards the overall prediction for a given instance. The first row indicates the sum of the overall mean prediction value and the changes. Red dots and yellow lines denote median values and error bars, respectively.

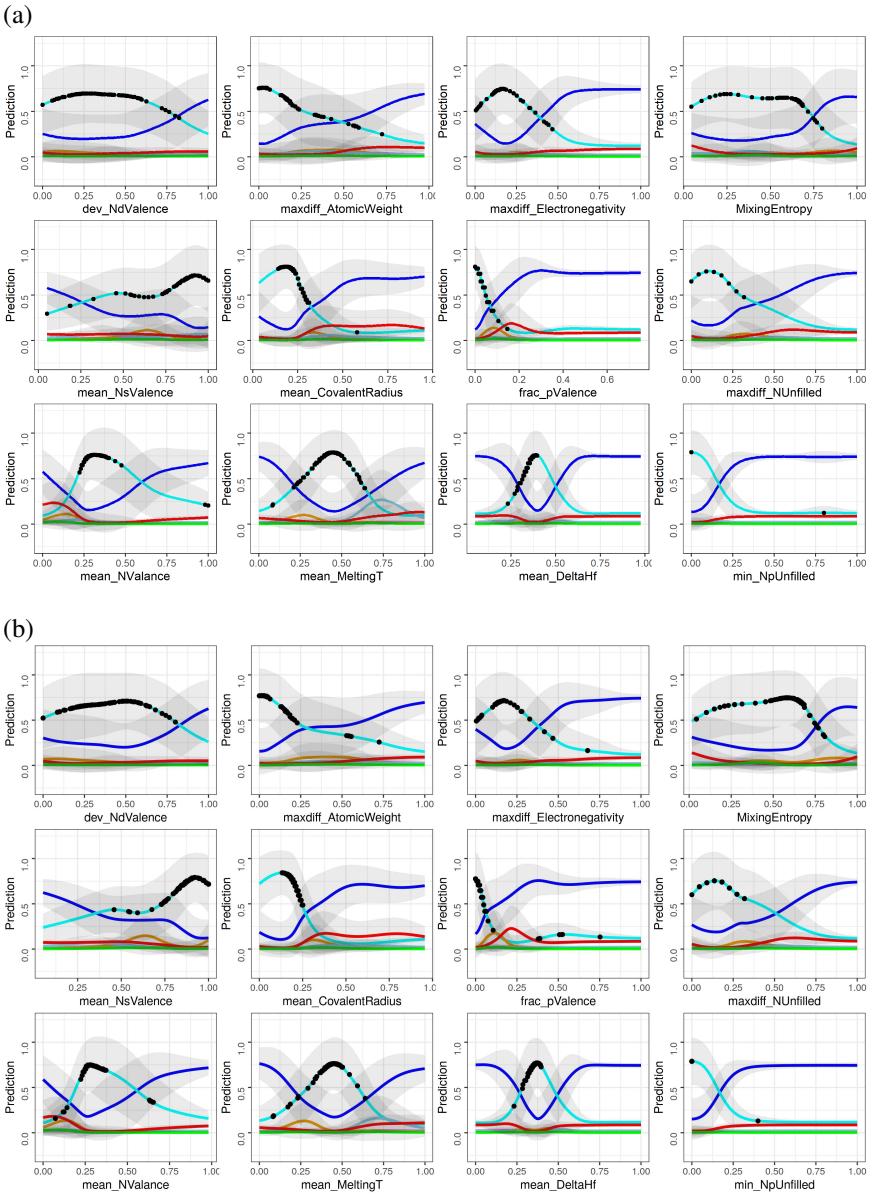


Figure 6. The averaged CP profiles with respect to the 12 input variables based on the (a) BD and (b) SHAP data of the representative clusters (clusters 3 and 10, respectively), which are predicted to have FCC phase by the eSVM model. The black dots indicate the true feature values for all the data points within a given cluster. Line colors denote phase information: blue, MP; violet, AM; cyan, FCC; orange, BCC+FCC; lightblue, HCP; red, BCC; green, IM.

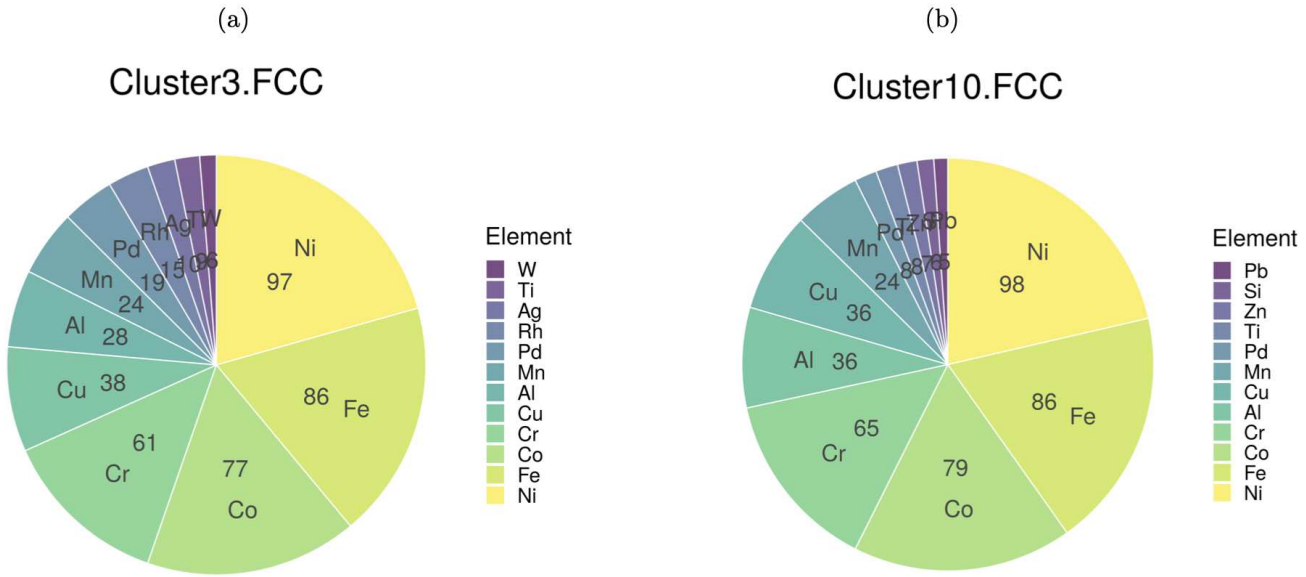


Figure 7. Pie chart showing the distribution of elements in the FCC clusters based on (a) BD and (b) SHAP decomposition.

Algorithm 1. Local interpretable ML algorithm using the SHAP and CP methods along with k-means clustering.

```
procedure SHAP_ANALYSIS(D, eSVM)  ▷ Procedure to construct the SHAP dataframe with the training dataset (D)
    RowLength ← Size(D)  ▷ Total number of instances of D
    for i ← 1 to RowLength do  ▷ Loops through each instance of D
        for j ← 1 to 50 do  ▷ Loops through 50 bootstrap samples
            M ← eSVM[j]
            Exp ← Model_explainer(M, D[j])  ▷ Generates a model explainer for a given bootstrap sample
            SHAP_pred[j] ← Predict_parts(Exp, new_observation=D[i])  ▷ Calculates the variable attributions to the prediction of a given instance
            Merged_SHAP_pred[i] ← Binding(SHAP_pred[j])  ▷ Merges the resulting variable attributions on every loop iteration
        end for
        Avg_Merged_SHAP_pred[i] ← Mean(Merged_SHAP_pred[i])  ▷ Averages the SHAP values of all the bootstrap samples for a given instance
        SHAP_dataframe ← Binding(Avg_Merged_SHAP_pred[i])
    end for
    return SHAP_dataframe
end procedure
procedure k-MEANS CLUSTERING(SHAP_dataframe)  ▷ Procedure for k-means clustering based on SHAP values
    k ← 10  ▷ k: the number of clusters
    Cluster_info ← kmean(SHAP_dataframe, k)  ▷ Implements the k-means clustering algorithm
    return Cluster_info  ▷ Classifies each instance with a specific cluster label
end procedure
procedure CP_ANALYSIS(D, eSVM, Cluster_info)  ▷ Procedure for CP analysis based on cluster information
    idx ← cluster_label  ▷ Choose a cluster label of interest
    for i ← 1 to length(Cluster_info[idx]) do  ▷ Loops through all the instances with the given cluster label
        for j ← 1 to 50 do  ▷ Loops through 50 bootstrap samples
            M ← eSVM[j]
            Exp ← Model_explainer(M, D[j])
            CP_pred[j] ← Predict_profile(Exp, new_observation=D[i])  ▷ Calculates individual CP profiles
            Merged_CPpred ← Binding(CP_pred[j])  ▷ Merges the resulting CP data on every iteration of the inner loop
        end for
        Merged_CPdata ← Binding(Merged_CPpred)  ▷ Merges the resulting CP data on every iteration of the outer loop
    end for
    CP_dataframe ← Mean(Merged_CPdata)  ▷ Averages the CP data across all the instances with the given cluster label
    return CP_dataframe
end procedure
```

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryDocument.pdf