UDK 621.3.016.4:536:519.673
Original scientific article/Izvirni znanstveni članek

ISSN 1580-2949
MTAEC9, 55(2)263(2021)

PREDICTION OF FORMATION ENERGY USING TWO-STAGE MACHINE LEARNING BASED ON CLUSTERING

NAPOVED TVORBENE ENERGIJE Z UPORABO DVOSTOPENJSKEGA STROJNEGA UČENJA NA OSNOVI ZDRUŽEVANJA

Xingyue Fan

School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China

Prejem rokopisa – received: 2020-09-07; sprejem za objavo – accepted for publication: 2021-01-15

doi:10.17222/mit.2020.174

The formation energy (H_f) is one of the important properties associated with the thermodynamic stability of ABO3-type perovskite. In this work, two-stage machine learning based on hierarchical clustering and regression was designed for improving the prediction values of the density-functional theory (DFT) H_f of ABO3-type perovskites. A global dataset was clustered into Cluster 1 and Cluster 2 using the CHI (the Calinski-Harabasz index). To compare the prediction performances of H_f, DTR (decision tree regression), GBRT (gradient boosted regression trees), RFR (random forest regression) and ETR (extra tree regression) were applied to build models of Cluster 1, Cluster 2 and the global dataset, respectively. The results showed that all four different regression models of Cluster 1 had a higher R^2, and lower *MSE* and *MAE* than those of the global dataset, while the models of Cluster 2 were poorer. Meanwhile, the GBRT model of Cluster 1 achieved a higher R^2 of 0.917, and lower *MSE* and *MAE* of 0.033 eV/atom and 0.125 eV/atom. We further validated and compared the generalization ability of the models by predicting the H_f of ABO3-type perovskite previously unseen in the training set. The two-stage machine-learning models proposed here can provide useful guidance for accelerating the exploration of materials with desired properties.

Keywords: ABO3-type perovskites, formation energy, hierarchical clustering, regression model

Tvorbena energija (H_f) je ena od pomembnih lastnosti povezanih s termodinamsko stabilnostjo perovskitov tipa ABO3 ($CaTiO_3$). V članku avtorji opisujejo uporabo dvostopenjskega strojnega učenja na hierarhičnem združevanju in nato oblikovanje regresijskih modelov za izboljšanje napovedi vrednosti funkcionalne teorije gostote (DFT) H_f perovskitov tipa ABO3. Globalni set podatkov je avtor združil v dva razreda (razred 1 in razred 2) s CHI (Calinski–Harabasz) indeksom. Za primerjavo vrednosti napovedi H_f je avtor uporabil regresijske metode: DTR, GBRT, RFR in ETR, za izdelavo modelov razreda 1, razreda 2 in globalnem setu podatkov. Rezultati so pokazali, da imajo vsi štirje regresijski modeli zgrajeni na osnovi razreda 1 visok koeficient determinacije R^2, manjši srednji kvadrat napake (*MSE*) in sredje absolutne napake (*MAE*) kot globalni set podatkov, medtem ko so bili modeli izdelani na osnovi razreda 2 slabši. GBRT model izdelan na osnovi razreda 1 je dosegel višji R^2 (0.917) ter manjša MSE (0,033 eV/atom) in *MAE* (0.125 eV/atom). Nadalje je avtor ovrednotil in primerjal sposobnost drugih modelov za napoved H_f perovskitov tipa ABO3, ki zaenkrat še niso bili objavljeni. Predlagani modeli dvostopenjskega strojnega učenja zagotavljajo uporabno metodo za pospešitev raziskav na področju materialov z želenimi lastnostmi.

Ključne besede: ABO3-tip perovskitov, tvorbena energja, hierarhično združevanje, regresijski model

1 INTRODUCTION

Predicting the thermodynamic stability of perovskite-type oxides is critical in materials science.[1–3] And the thermodynamic phase stability is broadly assessed using the convex-hull analysis, where the energy above the convex hull (E_{hull}) of a compound provides a direct measure of its stability.[1–2,4] However, E_{hull} is a poor target metric for a machine learning model.[2] The formation energy (H_f), the key metric of a crystal stability and synthesizability, is typically defined with respect to the stable line combination of the competing phases in E_{hull}.[2–4] Once H_f is predicted, E_{hull} then can be extracted by applying the OQMD database. Generally, a more negative value of H_f indicates a more stable compound.[4,5] Besides, H_f is required to calculate reaction enthalpies and voltages and determine many other properties of materials.[4]

Early measurements of H_f are based on density functional theory (DFT) calculations,[3,4] however, a comparatively expensive cost limits the use of DFT-screening large numbers of possible compounds. Recently, machine learning has been widely used in predicting the quantitative structure-property relationship (QSPR) of perovskite-type compounds,[2,6–9] including H_f. F. Faber et al.[10] investigated kernel ridge regression with its different feature vector representation as the input to predict H_f of solids. W. K. Ye et al.[2] used deep neural networks to predict the H_f of a crystal. J. Im et al.[5] utilized gradient-boosted regression trees to predict the H_f and bandgaps for searching lead-free perovskites used in solar cells.

The predictive accuracy of H_f is critical in a QSPR model. Traditionally, a global QSPR model built on an entire diverse dataset with a wide range is not always sat-

*Corresponding author's e-mail:
annpion@shu.edu.cn (Xingyue Fan)

isfactory for it is difficult to capture the detailed structure-property relationship of each group.[11–13] H. Yuan et al.[11] indicated that the prediction performance of the baseline toxicity of local models based on sub-clusters was much superior to that of a global model based on a global dataset. E. Stevens et al.[12] found that linear relationships between treatment hours and mastery of learning objectives were strong within sub-groups. Y. Liu et al.[13] clustered a creep dataset into eight clusters after trial and error to accelerate the prediction of the creep-rupture life. Thus, it is reasonable to hypothesize that a model built on sub-clusters can have a better prediction performance of H_f than a model built on a global dataset.

In this paper, we designed a two-stage machine-learning strategy based on a hierarchical-clustering method and then regression to better predict H_f. Firstly, the clustering method was used for nature grouping, and then different models of sub-clusters and the global dataset were built, seeking the best model. As there was little or no prior knowledge, the clustering method allowed the optimal natural grouping of compounds based on the similarity of structure descriptors evaluated by cluster internal indicators. Then, the H_f of sub-clusters and the global dataset were predicted by four commonly used regression models: DTR, GBRT, RFR and ETR. Finally, we validated the models on completely new test samples.

2 MATERIALS AND METHODS

2.1 Dataset and pre-processing

The study started with perovskite-type oxides, and the used dataset was from the work by P. Balachandran et al.[14] Removing the compounds with missing values, we got 386 ABO3 compounds and 9 corresponding descriptors. These descriptors include r_A (the Shannon ionic radii for A), r_B (the Shannon ionic radii for B), M_A (the Mendeleev numbers for A), M_B (the Mendeleev numbers for B), d_{AO} (the A-O bond length), d_{BO} (the B-O bond length), r_A/r_O (the radius ratio of A to O), r_B/r_O (the radius ratio of B to O), t (the tolerance factor). Since r_A and r_A/r_O, r_B and r_B/r_O are completely linearly related, we only keep r_A/r_O and r_B/r_O. Thus, the retained dataset includes 386 ABO3 compounds and 7 corresponding descriptors. Considering the model application on previously unseen samples, 15 samples were randomly reserved for the validation of the performance of our model, and the remaining 371 samples were used for building the model.

The descriptors were pre-processed with normalization and PCA before applying hierarchical clustering. In this work, the Euclidean distance was used to measure the distance or similarity between the objects. However, for the Euclidean distance, if descriptors are measured on different scales, descriptors with large values contribute more to the distance measure than variables with small ones. To overcome this problem, min-max normalization is adopted so that all these input descriptors are dimensionless and have the same range of (0–1).[15] PCA is used to transform a set of possible correlation variables into a set of linearly uncorrelated variables and to extract the most variations of the dataset. The variables from the transformed set are called the principal components (PCs). The orthogonal transformation was written as column vectors $W = (\omega_1, \omega_2, \ldots\omega_N)$ and the weights of the PCs were defined as ω_l, determined with an eigenproblem.[16]

$$X^T X\omega_1 = \lambda_1 \omega_1 \tag{1}$$

The eigenvalues are arranged in descending order, that is, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$. The variance contribution rate of each eigenvalue is:

$$\tilde{\lambda}_1 = \frac{\lambda_1}{\sum_{l=1}^{N} \lambda_1} \tag{2}$$

It is termed as the explained variance ratio in PCA.

2.2 Hierarchical clustering

Hierarchical clustering attempts to divide the samples in a dataset into several disjoint subsets. The divided samples are called clusters. There are two methods of hierarchical clustering – agglomerative and divisive. The process of these two methods of hierarchical clustering can be visualized through a dendrogram, which can demonstrate the clustering results. In this work, the agglomerative method of hierarchical clustering was performed, based on Ward's method,[17] to measure the similarity between the clusters and the Euclidean distance to measure the distances between the samples.

Cluster-validation techniques were used to evaluate the performance of the cluster results. In general, there are three criteria for a cluster validation: external criteria, internal criteria and relative criteria. To evaluate the results of a clustering method based on the inherent similarity of the datasets, one of the most used internal criteria, i.e., the Calinski-Harabasz index[18] was chosen to determine the optimal number of clusters.

2.3 Regression model

For the regression model, we tested four different machine-learning models: DTR, GBRT, RFR and ETR. They are the extensions of decision trees,[19] gradient boosted trees,[20] random forest[21] and extra trees[22] models of regression problems, respectively. DT is a diagram used to determine the process of an action or result. However, DT is prone to overfitting, while ensemble methods such as boosting and bagging can help prevent overfitting and improve accuracy. GBT adopts a boosting strategy; for regression, it uses the gradient descent algorithm to optimize the loss function iteratively so that the loss function of a sample is as small as possible; for classification, a log-likelihood loss needs to be introduced to help optimize the loss function. RF uses the bagging

strategy on the base learners of decision trees and further introduces the random-attribute selection in the training process. The ET algorithm is a more randomized version of the RF algorithm and it learns much faster.

The coefficient of determination (R^2), mean squared error (*MSE*) and mean absolute error (*MAE)* were chosen as the regression-model evaluation indicators. The R^2 represents the goodness of fit of a regression algorithm, the value of 1.0 indicates a perfect fit. *MSE* is the mean squared error between the predicted and actual values. *MAE* is the average of the absolute difference between the predicted and actual values, n represents the size of the data set, $\hat{y}_i$ is the predicted value of the *i*-th sample and y_i is the corresponding true value of the dataset. Mathematical expressions of the evaluation indicators are given in **Table 1**.

Table 1: Statistical-error measures applied for a method comparison

Measure	Expression
R^2	$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y_i')^2}$
MSE	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2$
MAE	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert(y_i - y_i')\rvert$

3 RESULTS AND DISCUSSION

3.1 Implementation of hierarchical clustering

For data pre-processing, the original dataset was normalized firstly. Then, PCA was applied to the normalized dataset. PCA not only reduces redundancy and noise, but also retains the most essential characteristics of the global dataset. As depicted in **Figure 1**, the bar chart shows the corresponding explained variance ratio of PC, which decreases monotonically. Notably, the contributions of the last two PCs are so small that we only retain the top five PCs used for hierarchical clustering.

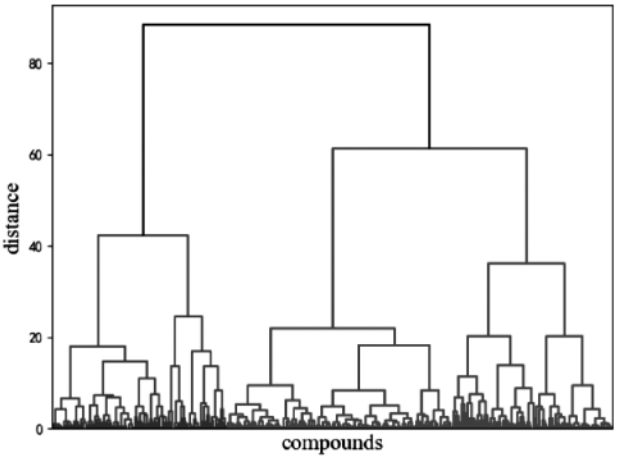

Figure 2: Dendrogram for the ABO3 dataset

Hierarchical clustering based on Ward's method and Euclidean distance was performed, generating a dendrogram to visualize the clustering results for 371 ABO3 compounds as shown in **Figure 2.** In the dendrogram, all the leaf nodes are compounds and the heights are the distances between two clusters, measuring the similarity. In general, a dendrogram is beneficial as it indicates the cluster and sub-cluster relationship, helping us evaluate the similarity of two materials and assess the clustering process,[15] especially when the dataset is small.[23]

The Calinski-Harabasz index (*CH*)[18] was applied to determine the optimal number of clusters. The larger the *CH*, the closer are the clusters, and the more dispersed are the clusters, the better is the clustering result. In **Figure 3**, the optimal number of clusters determined by *CH* is 2, and the suboptimal number is 6. Thus, in this work, the global dataset is divided into Cluster 1 and Cluster 2 based on the similarity between their structure descriptors.

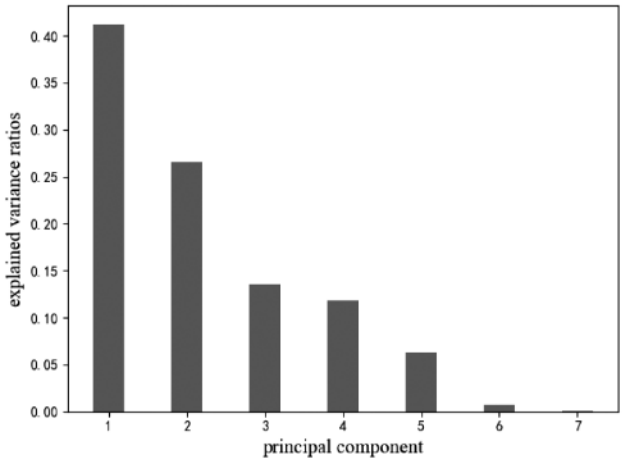

Figure 1: Bar chart of the explained variance ratio

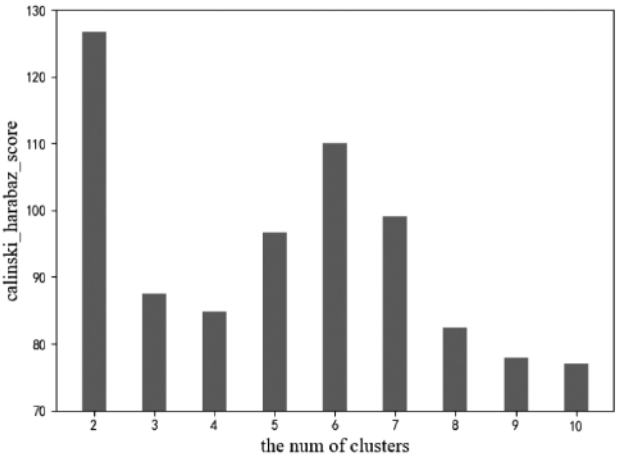

Figure 3: Bar chart of *CH* evaluation indicators

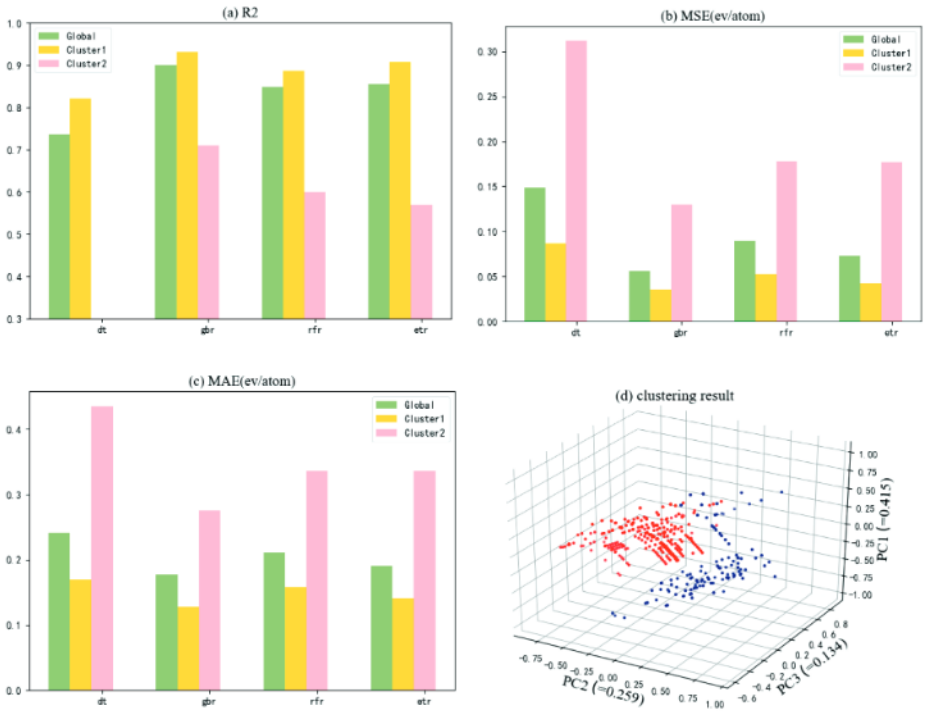

Figure 4: Comparison of the prediction performances of H_f by four regression models and three different datasets: a) R^2, b) *MSE,* c) *MAE,* d) PCA projection for 371 compounds – red for Cluster 1 and blue for Cluster 2

3.2 Model selection

To better predict H_f, the prediction performance of four candidate regression models and three different datasets were compared. The four regression models included DTR, GBRT, RFR and ETR. And the three datasets were the global dataset, Cluster 1 and Cluster 2.

For each regression model, 80 % of the samples of the dataset were chosen for random training and the remaining 20 % of the samples were used as the test set. For each dataset, we built the above four regression models for comparison. Evaluation indicators were R^2, *MSE* and *MAE* obtained by calculating the average value of 100 runs on different test sets. The evaluation indicators of the test sets based on four regression models and three datasets are depicted in **Figures 4a** to **4c.** With respect to the models, the GBRT model of Cluster 1 has the highest R^2, and the lowest *MSE* and *MAE*, indicating the optimal prediction performance of H_f. For Cluster 1, R^2 is slightly improved compared to the global dataset, while both *MSE* and *MAE* are significantly reduced. For Cluster 2, R^2 is reduced greatly, while *MSE* and *MAE* become larger.

On the whole, we achieved a natural group by clustering structure descriptors of 371 perovskite-type oxides. All the models of Cluster 1 were better than those of the global dataset, while the models of Cluster 2 were poorer. The two reasons for this may be as follows: firstly, as shown in **Table 2**, for each structure descriptor, the standard deviation of Cluster 2 is the largest one and the standard deviation of Cluster 1 is the smallest one. Secondly, as shown in **Figure 4d,** the sample points in Cluster 2 are smaller than in Cluster 1, but they are more widely scattered. It is known to all that machine learning relies heavily on a proper dataset and proper regression model. The poor prediction result for Cluster 2 may be due to its large standard deviation, small training samples and a poor training model of Cluster 2. Structure-property relationships indicate that compounds with similar structure descriptors are more likely to exhibit similar properties. Clustering based on structural descriptors is thus likely to group compounds with similar properties.[24] However, there are several factors that affect the clustering results, such as the selected feature set, the heterogeneity of the dataset and the size of the dataset. Local models built of sub-clusters can outperform global models build of the global dataset, but there are exceptions when the local models can be equal to, or even poorer than, the global models,[25] such as Cluster 2 in our experiment. Therefore, it is important to select a proper dataset and proper regression model to improve the target property.

Table 2: Standard deviations of 7 descriptors of the global dataset, Cluster 1 and Cluster 2, respectively

Dataset	M_A	M_B	d_{AO}	d_{BO}	r_A/r_O	r_B/r_O	t
Global	24.691	20.882	0.211	0.165	0.142	0.130	0.112
Cluster1	11.004	19.678	0.135	0.130	0.128	0.093	0.083
Cluster2	25.933	22.661	0.239	0.207	0.153	0.150	0.154

3.3 Model application: new compounds

It is hypothesized that when a new sample was included into Cluster 1, we could build a GBRT model of

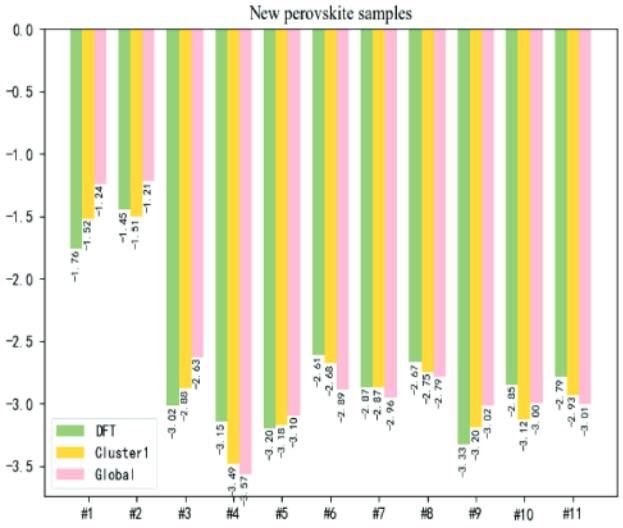

Figure 5: Comparison of DFT values, values predicted by Cluster 1 and values predicted by the global dataset for H_f of 11 previously unseen samples

Cluster 1 rather than of the global dataset to predict H_f. As a further test of the hypothesis, we utilized 15 randomly reserved ABO3 samples that were previously unseen by the model. We computed the Euclidean distance between 15 new samples and the centers of sub-clusters, and then found the sub-cluster to which the new sample belonged. Taking a random trial as an example, 11 samples were clustered into Cluster 1, then GBRT models were built of Cluster 1 and of the global dataset to compare the prediction performances of those 11 samples. **Figure 5** shows the comparison of the DFT values, values predicted by Cluster 1 and values predicted by the global dataset for H_f. We can see that the errors between the values predicted by Cluster 1 and DFT values are more acceptable than the errors between the values predicted by the global dataset and DFT values. We implemented many random trials and also found that when a new sample is included into Cluster 1, it is proposed to build a GBRT model of Cluster 1 rather than of the global dataset to better predict its H_f.

5 CONCLUSIONS

In this work, we designed a two-stage machine-learning strategy based on a hierarchical clustering method and regression to predict H_f. A total of 371 compounds were clustered into Cluster 1 and Cluster 2 using cluster internal indicators based on structure descriptors. Then we built four different regression models of Cluster 1, Cluster 2 and the global dataset to compare the prediction performances for H_f. The values of R^2, *MSE* and *MAE* for the test sets showed that all the models of Cluster 1 were better than those of the global dataset, while all the models of Cluster 2 were poorer. Meanwhile, the GBRT model had a better prediction performance than the DT, RFR and ERT models. Finally, we utilized 11 previously unseen compounds to confirm that the GBRT model of Cluster 1 had a better prediction performance than that of the global dataset. The result suggested that if new compounds were clustered into Cluster 1, it was reasonable to build a GBRT model of Cluster 1 rather than of the global dataset to predict their H_f values. The QSPR model of Cluster 1 could capture the structure-property relationship more accurately than the global model, indicating a better prediction performance. However, larger standard deviations and fewer samples might have led to the poor training model of Cluster 2, so if compounds were clustered into Cluster 2, we could choose to build a GBRT model of the global dataset rather than Cluster 2 to predict H_f. Furthermore, it was of great significance for the acceleration of material discovery when all the models of sub-clusters had a better prediction performance than those of the global dataset. In fact, the clustering method can be used for data pre-processing required for predicting a target property, which is significant for predicting the QSPR in many cases, especially when the global dataset is heterogeneous or diverse.

6 REFERENCES

[1] W. Li, R. Jacobs, D. Morgan, Predicting the thermodynamic stability of perovskite oxides using machine learning models, Computational Materials Science, 150 (**2018**) 454–463, doi:10.1016/j.commatsci.2018.0-4.033

[2] W. K. Ye, C. Chen, Z. B. Wang, I. H. Chu, S. P. Ong, Deep neural networks for accurate predictions of crystal stability, Nature Communications, 9 (**2018**) 1, doi:10.1038/s41467-018-06322-x

[3] A. A. Emery, C. Wolverton, High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO3 perovskites, Scientific Data, 4 (**2017**) 170153, doi:10.1038/sdata.20-17.153

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

[5] J. Im, S. Lee, T. W. Ko, H. W. Kim, Y. Hyon, H. Chang, Identifying Pb-free perovskites for solar cells by machine learning, Computational Materials, 5 (**2019**) 1, doi:10.1038/s41524-019-0177-0

[6] S. H. Lu, Q. H. Zhou, Y. X. Ouyang, Y. Guo, Q. Li, J. L. Wang, Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning, Nature Communications, 9 (**2018**) 1, doi:10.1038/s41467-018-05761-w

[7] P. V. Balachandran, B. Kowalski, A. Sehirlioglu, T. Lookman, Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning, Nature Communications, 9 (**2018**) 1, doi:10.1038/s41467-018-03821-9

[8] R. Yuan, Z. Liu, P. V. Balachandran, D. Q. Xue, Y. M. Zhou, X. D. Ding, J. Sun, D. Z. Xue, T. Lookman, Accelerated Discovery of Large Electrostrains in BaTiO3-Based Piezoelectrics Using Active Learning, Advanced Materials, 30 (**2018**) 7, 1702884, doi:10.1002/adma.-201702884

[9] L. Shi, D. P. Chang, X. B. Ji, W. C. Lu, Using Data Mining to Search for Perovskite Materials with Higher Specific Surface Area, Journal of Chemical Information and Modeling, 58 (**2018**) 12, doi:10.1021/acs.jcim.8b00436

[10] F. Faber, A. Lindmaa, R. Armiento, Crystal structure representations for machine learning models of formation energies, International

Journal of Quantum Chemistry, 115 (**2015**) 16, 1094–1101, doi:10.1002/qua.24917

[11] H. Yuan, Y. Y. Wang, Y. Y. Cheng, Local and Global Quantitative Structure – Activity Relationship Modeling and Prediction for the Baseline Toxicity, Journal of Chemical Information and Modeling, 47 (**2017**) 1, 159–169, doi:10.1021/ci600299j

[12] E. Stevens, D. R. Dixon, M. N. Novack, D. Granpeesheh, T. Smith, E. Linstead, Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning, International Journal of Medical Informatics, 29 (**2019**) 29–36, doi:10.1016/j.ijmedinf.2019.05.006

[13] Y. Liu, J. M. Wu, Z. C. Wang, X. G. Lu, M. Avdeev, S. Q. Shi, C. Y. Wang, T. Yu, Predicting creep rupture life of Ni-based single crystal superalloys using divide-and-conquer approach based machine learning, Acta Materialia, doi:10.1016/j.actamat.2020.05.001

[14] P. V. Balachandran, A. A. Emery, J. E. Gubernatis, T. Lookman, C. Wolverton, A. Zunger, Predictions of new ABO3 perovskite compounds by combining machine learning and density functional theory, 2 (**2018**) 4, doi:10.1103/PhysRevMaterials.-2.043802

[15] K. W. Johnson, P. M. Langdon, M. F. Ashby, Grouping materials and processes for the designer: an application of cluster analysis, Materials and Design, 23 (**2002**) 1, 1–10, doi:10.1016/s0261-3069(01)000358

[16] H. Abdi, L. J. Williams, Principal component analysis, Wiley Interdisciplinary Reviews: Computational Statistics, 2 (**2010**) 4, 433–459, doi:10.1002/wics.101

[17] J. H. Ward, Hierarchical Grouping to Optimize an Objective Function Journal of the American Statistical Association, 58 (**1963**) 301, 236, doi:10.2307/2282967

[18] T. Calinski, J. Harabasz, A dendrite method for cluster analysis, Communications in Statistics, 3 (**1974**) 1, 1–27, doi:10.1080/03610927408827101

[19] J. R. Quinlan, Induction on decision tree, Machine Learning, 1 (**1986**) 1, 81–106, doi:10.1007/BF00116251

[20] J. H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, Annals of Statistics, 29 (**2001**) 5, 1189–1232, doi:10.2307/2699986

[21] L. Breiman, Random Forests, Machine Learning, 45 (**2001**) 1, 5–32, doi:10.1023/A:101093-3404324

[22] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Machine Learning, 63 (**2006**) 1, 3–42, doi:10.1007/s10994-006-6226-1

[23] J. Zhao, R. Plagge, N. M. Ramos, M. L. Simoes, J. Grunewald, Application of clustering technique for definition of generic objects in a material, Journal of Building Physics, 39 (**2015**) 2, 124–146, doi:10.1177/1744259115588013

[24] G. M. Downs, J. M. Barnard, Clustering Methods and Their Uses in Computational Chemistry, John Wiley & Sons, Inc., New Jersey 2002, 1–40

[25] G. Piir, S. Sild, U. Maran, Comparative analysis of local and consensus quantitative structure-activity relationship approaches for the prediction of bioconcentration factor, SAR and QSAR in Environmental Research, 24 (**2013**) 3, 175–199, doi:10.1080/1062936x.2012.762426