ADVANCED INTELLIGENT SYSTEMS
Open Access
www.advintellsyst.com

# Target Material Property-Dependent Cluster Analysis of Inorganic Compounds

*Nobuya Sato,* Akira Takahashi,* Shin Kiyohara, Kei Terayama, Ryo Tamura, and Fumiyasu Oba*

**The cluster analysis of materials categorizes them according to similarities based on the features of materials, providing insight into the relationship between the materials. Conventional cluster analyses typically use basic features derived from the chemical composition and crystal structure without considering target material properties such as the bandgap and dielectric constant. However, such approaches do not meet demands for grading materials according to properties of interest simultaneously with chemical and structural similarities. Herein, a clustering method grouping similar materials in terms of both the target properties and basic features is proposed. The clustering is compared considering the cohesive energy with that considering the bandgap of metal oxides, showing that their categorizations are clearly different. Further, several clusters classified by the bandgap are analyzed, and coordination environments related to each range of the bandgap are revealed. The clustering for the electronic static dielectric constant identifies a cluster involving several perovskite-type oxides and balancing with the bandgap near the Pareto front. The method enables analyses with different viewpoints from those of the conventional clustering and feature importance analyses by taking the relationship between the target property and the basic features into account.**

## 1. Introduction

In the field of materials science, it is quite common to classify substances into various categories based on their constituent elements and crystal structure characteristics. For instance, we often categorize substances into oxides, sulfides, nitrides, or others according to their composition and constituent elements, or classify them into other classes such as II–VI and III–V semiconductors employing other criteria. Furthermore, it is also a prevalent practice to classify materials into prototype crystal structures such as rock-salt or perovskite types. The fundamental characteristics of constituent elements and crystal structures that define materials are extremely diverse, and the appropriate classification varies depending on the intended application of the materials. It is essential to identify promising classes of materials by considering material properties and functions closely relevant to the target application.

Meanwhile, progress in machine learning has brought new approaches to materials science.[1–4] One of the most common applications of machine learning would be to predict material properties of interest from basic characteristics usually derived from constituent elements and/or crystal structures, which are often called descriptors. Interpretable or explainable machine learning techniques[5–7] can unveil chemical trends and identify controlling factors of material properties. For instance, the importance

N. Sato, A. Takahashi, S. Kiyohara, F. Oba
Laboratory for Materials and Structures
Institute of Innovative Research
Tokyo Institute of Technology
R3-7, 4259 Nagatsuta, Midori-ku, 226-8501, Japan
E-mail: nobuya.sato.000@gmail.com; takahashi.a.bb@m.titech.ac.jp

S. Kiyohara
Institute for Materials Research
Tohoku University
2-2-1 Katahira, Aoba-ku, Sendai 980-8577, Japan

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aisy.202400253.



**DOI: 10.1002/aisy.202400253**

K. Terayama
Graduate School of Medical Life Science
Yokohama City University
1-7-29 Suehiro-cho, Tsurumi-ku, 230-0045, Japan

K. Terayama
RIKEN Center for Advanced Intelligence Project
1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

K. Terayama, F. Oba
MDX Research Center for Element Strategy
International Research Frontiers Initiative
Tokyo Institute of Technology
SE-6, 4259 Nagatsuta, Midori-ku, Yokohama 226-8501, Japan

R. Tamura
Center for Basic Research on Materials
National Institute for Materials Science
1-1 Namiki, Tsukuba 305-0044, Japan

R. Tamura
Graduate School of Frontier Sciences
The University of Tokyo
5-1-5 Kashiwa-no-ha, Kashiwa 277-8568, Japan

of features in random forest or gradient boosting decision tree regression,[8–10] Shapley additive explanations,[9,11–14] variable importance in projection scores,[15] sure independence screening and sparsifying operator analysis regression,[16,17] and multiple linear regression with an expectation maximization algorithm[18,19] have been employed to extract such knowledge. Recently, natural language processing has enabled the extraction of materials knowledge even from the text corpora.[20,21] The cluster analysis is also such a technique aimed at categorizing data so that similar data are in the same groups. Clustering is typically utilized for categorizing materials based on their basic features, such as constituent chemical elements,[17,22,23] crystal structures,[24,25] and local structures.[26,27] The obtained groups summarize similarities between the input data points, giving us insight into their relationship and routes to further analysis within each group.

The conventional cluster analysis is an unsupervised learning technique, where the target property to predict does not exist in contrast to supervised learning, such as regression. However, grading materials according to a target property as a function of the features is often desired. For example, there is a case where we try to categorize semiconductors and insulators according to the width of the bandgap and investigate the chemical and structural characteristics of respective categories. This kind of clustering requires taking the relationship between the basic features and the target property into account and, therefore, differs from the clustering with only the target property, which is agnostic about the basic features. The analysis only with the target property is not straightforward because close target values in a cluster may originate from different chemical and structural natures. For example, the clustering of the bandgap may gather materials into a cluster where some bandgaps are determined primarily by the electronegativity difference of the constituent elements, which is a good measure of ionicity, and others are determined mainly by features relevant to covalency such as the average electronegativity. They cannot be distinguished only by the target values, and separate analyses with the basic chemical and structural features are required. In addition, the conventional clustering solely using basic features does not necessarily gather materials close to each other in terms of the property of our interest.

In this article, we propose a clustering method involving information about the target property as well as the basic features. We inject information about the target property into the clustering of the materials by the random forest (RF) regression.[28] Our method consists of three parts. First, an RF regression model is trained for predicting a given target property. Second, the feature vectors are transformed into "$z$-vectors" based on the paths in the decision trees when making predictions with the trained RF model. Finally, the cluster analysis is performed for the $z$-vectors.

The proposed method in this work is inspired by the past work by Breiman[29] on the analysis of the RF classification model, which divides the classes further by clustering with a data proximity obtained from the model. Our approach differs from such a conventional method in the following two aspects: 1) it is utilized for regression rather than classification problems and 2) not only cosine similarity but also any arbitrary similarity measure can be applied for clustering, the details of which are described later. Our method is also somewhat similar to several existing methods to extract features from the structures of machine learning models. For example, there exists the image classification method, where an ensemble model of classification trees is applied to transform an image into a vector for a subsequent classification.[30] Another example is the technique called feature learning or representation learning.[31,32] These works are similar to our method in that data are transformed into a different form beforehand. The main difference is that our purpose is to introduce information about the target variable into the clustering rather than finding a transformation specific to the subsequent task.

## 2. Results and Discussion

### 2.1. Comparison Between Different Target Properties

First, we compare the clustering of the same dataset between two different target properties, i.e., the cohesive energy per atom ($E_{\mathrm{coh}}$) and the bandgap ($E_{\mathrm{g}}$) from first-principles calculations, to confirm that the results of clustering are actually different and analyze how they differ. We applied the developed clustering method to the dataset which consists of 7981 oxides collected from the Materials Project database.[33,34] The distributions of the target properties are shown in Figure S1, Supporting Information. Note that $E_{\mathrm{g}}$ of this dataset tends to be underestimated compared to experimental values; see the Computational Section for details. We constructed RF models using feature values shown in Table S1, Supporting Information for $E_{\mathrm{coh}}$ and $E_{\mathrm{g}}$. The RF models are confirmed to be constructed accurately (Figure S2 and Table S2 and S3, Supporting Information). Applying the transformation into $z$-vectors and with the agglomerative hierarchical clustering, we obtained dendrograms for $E_{\mathrm{coh}}$ and $E_{\mathrm{g}}$, respectively (Figure S3 and S4, Supporting Information). In this study, we set the threshold of the distance in the $z$-space to divide the $E_{\mathrm{g}}$ and $E_{\mathrm{coh}}$ datasets into 30 groups, respectively. It should be noted that appropriate thresholds of the distance in the $z$-space can be applied to such dendrograms to divide the materials into any number of clusters. The higher (lower) the threshold, the finer (coarser) the classification becomes. This threshold, and therefore the number of clusters, can be adjusted to an appropriate cluster size.

The obtained clusters are numbered in ascending order by the median of the target values (Figure S3, and S4, Supporting Information). As shown in **Figure 1**, the target values of each cluster are distributed within narrower ranges than the whole dataset, which suggests that information about the target properties is injected into the clustering as desired. Furthermore, there are clusters distributed in nearly the same range of the target value, suggesting that the dataset is divided by not only the target property but also the features as desired. A comparison between clusters of close target values is given later.

The clusters are separated better in the target property space than those by the conventional clustering method simply appending target properties to feature vectors. Figure S10, Supporting Information shows distributions of target values of clusters obtained by the conventional clustering, i.e., the agglomerative hierarchical clustering of the features and target variable (the number of features plus one variable), with the Euclidean distance and the Ward method. Note that the conventional clustering with the average method results in quite a few large clusters and many small clusters, which is the chaining effect, while

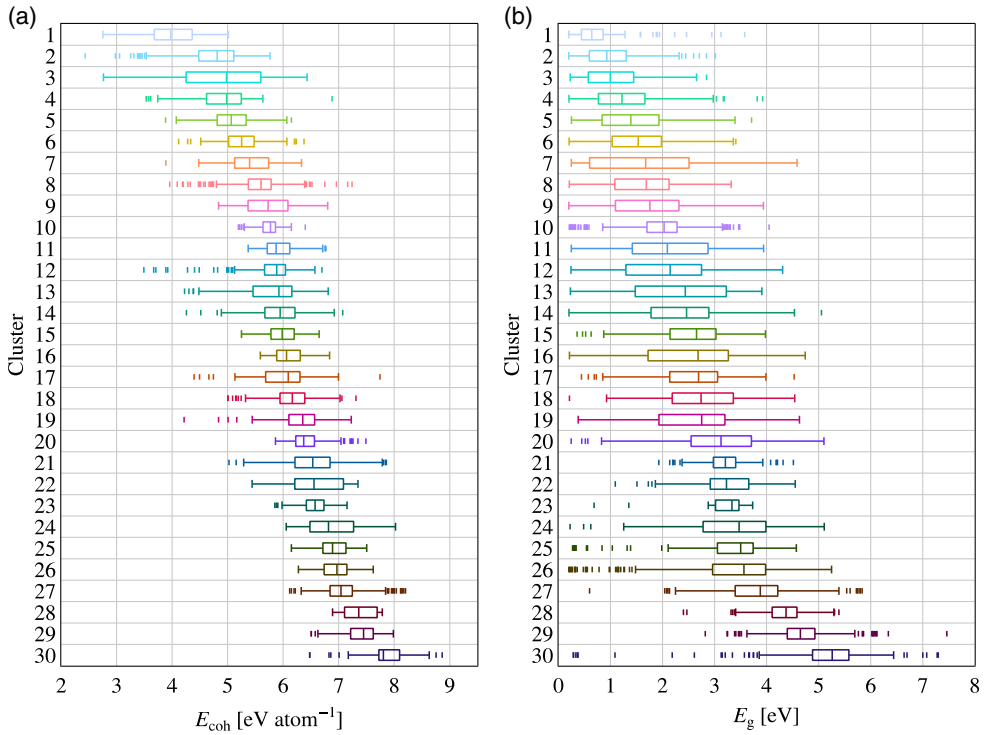


**Figure 1.** Distributions of clusters obtained by the developed clustering method. a) and b) show the clustering for $E_{coh}$ and $E_g$, respectively. The clusters are numbered in ascending order by the median of the target values.

our clustering does not exhibit such effects, as can be seen in Figure 1. The goodness of separation of clusters with respect to the target property is measured by the Calinski–Harabasz index,[35] which is defined as a ratio of the intercluster dispersion to the intracluster dispersion and the larger is the better. The Calinski–Harabasz indices of our method are 1180 for $E_{coh}$ and 508 for $E_g$, while those of the conventional method are 252 for $E_{coh}$ and 90 for $E_g$. The clusters of our method are separated better than those of the conventional method in terms of the target values because the target property is only one among hundreds of variables in the conventional method.

Although our dataset contains multication oxides, the distribution of the binary oxides in respective clusters would be helpful information to understand the chemical tendency. **Figure 2** shows which cluster each binary oxide belongs to; tabular form summaries are given in Table S4a and S5a, Supporting Information. As desired, the belonging of binary oxides depends on the target property. The difference is especially clear in the oxides of group 1 and 2 elements: the group 1 and 2 oxides other than $Li_2O$ and BeO belong to different clusters for $E_{coh}$ and the same cluster for $E_g$. The difference would be related to outliers. The formation energies of group 1 oxides (excluding $Li_2O$) are concentrated in a low range. Specifically, they show values of 2.79–3.31 eV atom$^{-1}$, which are bottom 0.4% or lower in the whole dataset (**Figure 3**a). The $E_g$ value of BeO is also an outlier, which is 7.46 eV, the largest in the whole dataset (Figure 3b). The lower $E_{coh}$ values of the group 1 oxides result partly from lower Madelung energies due to the smaller valence of cations. In contrast, the $E_g$ values of the group 1 and 2 oxides are both related to oxygen $p$ and cation $s$ orbital characteristics in the valence and conduction bands, respectively. The distributions of their $E_g$ values are almost overlapped, which would have led to clustering into the same cluster.

### 2.2. Detailed Analysis of the Bandgap

We now analyze the characteristics of several clusters in the results of the $E_g$ clustering to see how they are related to physical and chemical pictures. By inspecting the feature values within a cluster, we can reveal the chemical and structural tendencies of the cluster.

First, we take a closer look at cluster 3, which is the third lowest in the median of $E_g$ among the 30 clusters. This cluster includes the binary oxides of Cu(I) and Ag(III); namely, $Cu_2O$ and $Ag_2O_3$, and 82% of the oxides in this cluster involve at least one of Cu and Ag. Although this proportion is significantly high, the proportion of such oxides of cluster 3 in the entire dataset including the other clusters is only 34%. This suggests that cluster 3 is not classified solely based on elemental information and, therefore, further analysis is needed to characterize this cluster.

As for characteristic structural features of cluster 3, two features denoted as $Q_4^{max}$ and $Q_8^{max}$ are concentrated in large values (**Figure 4**a,b). They are the maximums of the bond-orientational order parameters[36–38] among atoms defined as

$$Q_l^{max} = \max_i \left( \frac{4\pi}{2l+1} \sum_{m=-l}^{l} \left| \sum_{j \in A_i} \frac{\Omega_{ij}}{4\pi} Y_{lm}(\boldsymbol{r}_{ij}) \right|^2 \right)^{1/2} \quad (1)$$

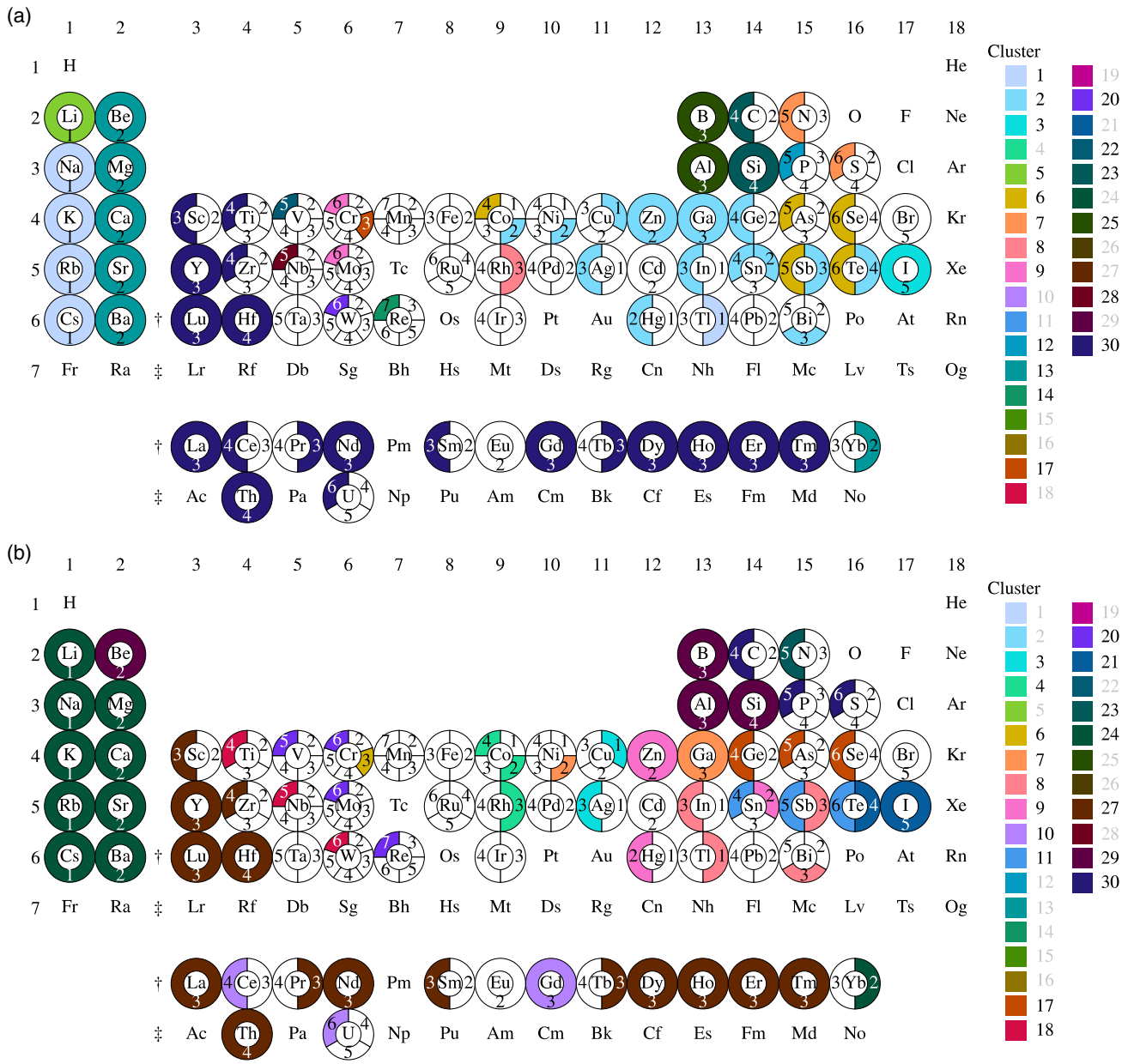


**Figure 2.** Distributions of binary oxides in clusters obtained by the developed clustering. a) and b) show the clustering for $E_{coh}$ and $E_g$, respectively. Cations in binary oxides are depicted as colored annulus sectors with their oxidation numbers, where the colors correspond to the respective clusters. Cations of colorless sectors are contained only in oxides consisting of more than two cation species in the present dataset. The clusters are numbered in ascending order by the median of the target values, where oxides consisting of multiple cations are also considered. The clusters with light-gray numbers do not contain binary oxides.

where $i$ is an index of an atom, $A_i$ is a set of indices of atoms adjacent to the $i$th atom in terms of the Voronoi tessellation, $\Omega_{ij}$ is a solid angle from the $i$th atom subtended by the face common to the Voronoi cells of the $i$th and $j$th atoms, $Y_{lm}(\cdot)$ are the spherical harmonics, and $\boldsymbol{r}_{ij}$ is a direction vector from the $i$th atom to the $j$th atom. Large $Q_4^{max}$ and $Q_8^{max}$ suggest that an oxide contains a roughly octahedrally coordinated atom.[36] For example, $ZnCu_2O_4$, which has the largest $Q_4^{max}$ in this cluster, consists of Cu atoms coordinated by six O atoms whose maximum distance difference is 0.90 Å (Figure S5a, Supporting Information). Also, $Li_4Co_3TeO_8$, which has the largest $Q_8^{max}$ in this cluster, consists of Li, Co, and Te atoms coordinated by six O atoms whose maximum distance differences are 0.11, 0.11, and 0.00 Å, respectively (Figure S5b, Supporting Information). Note that these features are relatively less important for the whole dataset: their permutation importances are only 7% and 3% of the highest. The developed clustering has thus identified a series of oxides characterized by these locally important features in a narrow range of $E_g$, unlike conventional feature importance analysis for the whole dataset.

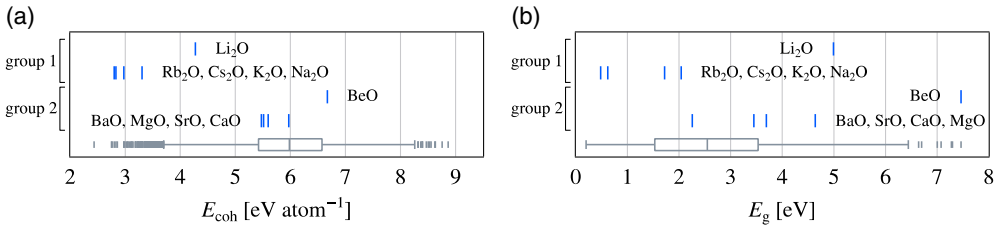


**Figure 3.** Positions of group 1 and 2 binary oxides in the distributions. a) and b) show the distributions of $E_{coh}$ and $E_g$, respectively. Box plots indicate the distributions of the whole dataset.

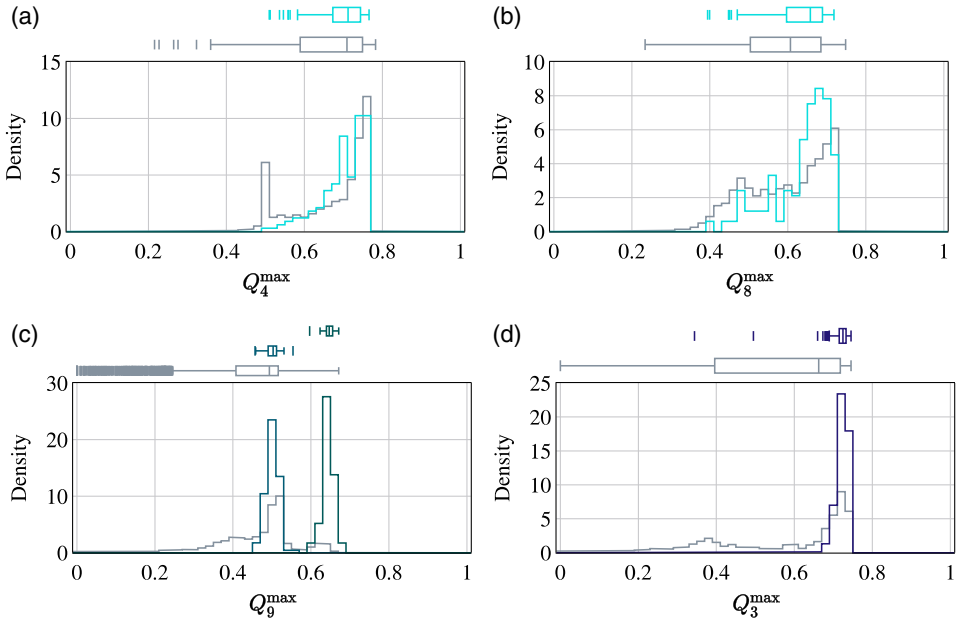


**Figure 4.** Distributions of the maximum bond-orientational order parameters among atoms ($Q_l^{max}$) in specific clusters for $E_g$. a) shows the $Q_4^{max}$ distribution in cluster 3, the $Q_8^{max}$ distribution in cluster 3, the $Q_9^{max}$ distributions in clusters 22 (the left peak) and 23 (the right peak), and the $Q_3^{max}$ distribution in cluster 30. The histograms in the back and the lower box plots indicate the distribution for the whole dataset. Each histogram is normalized so that the area is equal to one.

As for the target property, the $E_g$ of $Cu_2O$, $Ag_2O_3$, and $ZnCu_2O_4$ are related to the energy splitting of partially occupied $d$ states even in $Cu_2O$ with a formally Cu $d^{10}$ configuration: both the valence band maxima and conduction band minima are mainly characterized by the transition metal (Cu or Ag) $d$ states and O $p$ states (Figure S6a–c, Supporting Information). The valence band maximum of $Li_4Co_3TeO_8$ is also characterized by Co $d$ states and O $p$ states, while the conduction band minimum is characterized by Te $s$ states and O $p$ states (Figure S6d, Supporting Information). The main peaks of the unoccupied Co $d$ states in the density of states are a few eV higher in energy than the conduction band minimum.

Our clustering method also helps us to analyze and understand oxides with similar target values originating from different chemical and structural natures. For example, although clusters 22 and 23 in the $E_g$ case are distributed in $E_g$ ranges close to each other, whose medians are 3.2 and 3.3 eV, respectively, they are separated from each other in quite early stage of the agglomerative hierarchical clustering and their contents show clearly different tendencies in the chemical composition. All oxides of cluster 23 contain N, while none of the oxides in cluster 22 contain N and 76% of the oxides contain at least one of Si, Ge, and As. It would be useful to be able to classify substances with similar physical and chemical properties by different origins like this case, which is difficult by directly applying clustering methods to only the target properties.

Looking more closely at cluster 23, it can be seen that the values of $Q_9^{max}$ are concentrated at large values compared to the entire dataset (Figure 4c). It suggests that cluster 23 consists of nitrates because $Q_9^{max}$ is large if the trigonal planar coordination exists. Cluster 23 also involves compounds whose ratio of N atoms to O atoms is not 1:3 such as $YNO_4$, which has an $NO_3$ local structure though its composition ratio seems not to be a nitrate.

Cluster 30 in the $E_g$ case, which is the largest in the median of $E_g$ among the 30 clusters, includes binary oxides of C(IV), P(V), and S(VI), that is, $CO_2$, $P_2O_5$, and $SO_3$. $CO_2$ and $SO_3$ are molecular crystals. Almost all oxides, 97%, of this cluster contain at least one of B, P, and S. Oxides containing C are only 1% though its binary oxide is included in this cluster. In contrast, 34% of the

oxides in this cluster contain B even though its binary oxide is not included.

The inspection of the features shows that $Q_3^{max}$ is large for most entries of cluster 30 (Figure 4d). In this dataset, $Q_3^{max}$ is correlated with the maximum tetrahedral order parameter,[39] which measures the similarity of local structures to the tetrahedral coordination (with the Pearson correlation coefficient of 0.79), though the maximum tetrahedral order parameter has been eliminated from the regression by the feature selection. For example, $KAl(SO_4)_2$, which has the largest $Q_3^{max}$ in this cluster, contains S atoms coordinated by four O atoms whose maximum distance difference is 0.03 Å (Figure S5c, Supporting Information).

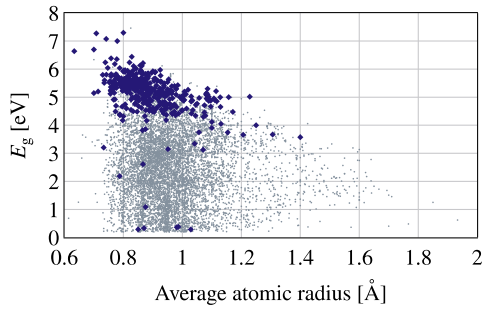


**Figure 5.** Distribution of oxides with respect to the average atomic radius and $E_g$. Blue diamonds indicate oxides included in cluster 30 for $E_g$.

The chemical formula and electronic structure of $KAl(SO_4)_2$ imply its strong ionic character. The valence band of $KAl(SO_4)_2$ is characterized predominantly by O *p* states, while a minor hybridization of S *s*, S *p*, O *s*, and O *p* states can be found in the conduction band (Figure S6e, Supporting Information). The ionic character of oxides in cluster 30 is also implied by the feature indicating the average atomic radius, which shows a weak negative correlation with $E_g$ (**Figure 5**). A small average atomic radius is likely to correspond to short interatomic distances even in oxides with a high ionicity, which causes strong Coulomb interactions and results in a large $E_g$ value.

### 2.3. Clustering of the Polycrystalline Average of the Electronic Static Dielectric Tensor

We also analyze the polycrystalline average of the electronic contribution to the static dielectric tensor ($\varepsilon_{el}$) and extract features of materials near the Pareto front in the $E_g$–$\varepsilon_{el}$ space. The dataset consists of 1301 oxides collected from the Materials Project database. Note that the $\varepsilon_{el}$ values in this dataset tend to be overestimated compared to experimental values, as detailed in the Computational Section. The dataset is divided into 20 clusters by our method with the agglomerative hierarchical clustering, where the clusters are numbered in ascending order by the median of $\varepsilon_{el}$ (Figure S7a, Supporting Information), as in the cases of $E_{coh}$ and $E_g$. The number of clusters is reduced from 30 to 20 because the dataset is smaller than that for $E_{coh}$ and

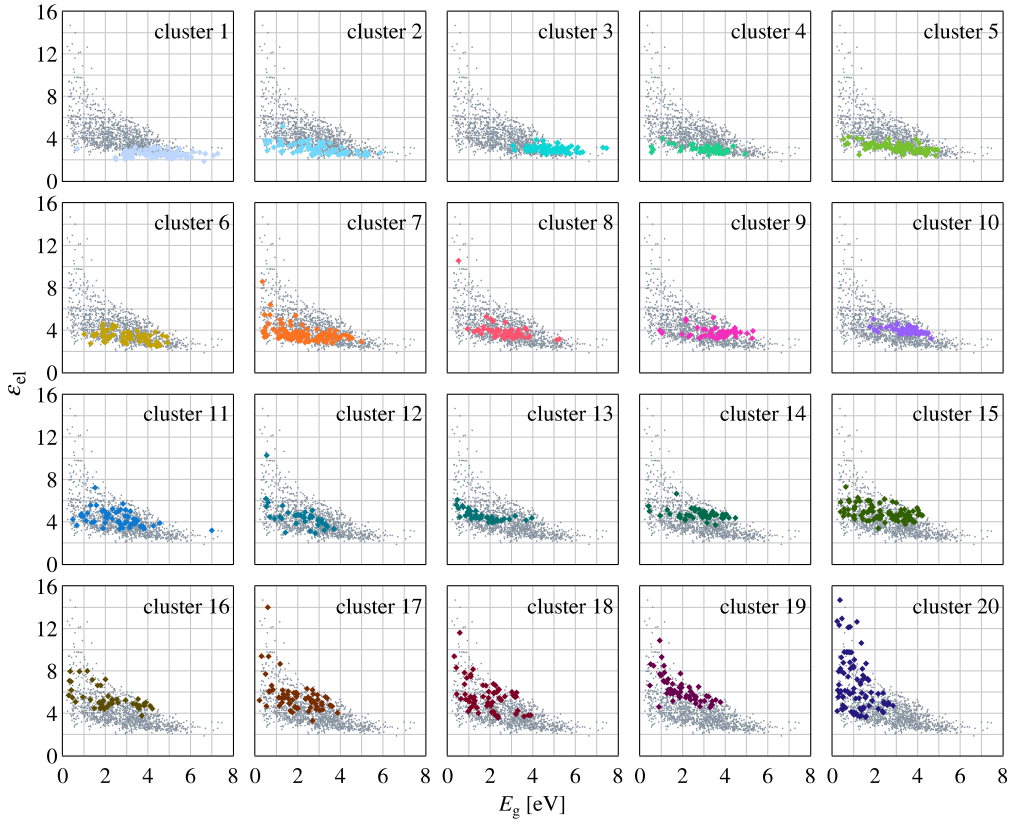


**Figure 6.** Distributions of oxides with respect to $E_g$ and $\varepsilon_{el}$. Each panel shows a cluster composed of oxides indicated by colored diamonds. The clustering is performed for $\varepsilon_{el}$. The clusters are numbered in ascending order by the median of $\varepsilon_{el}$.

$E_g$; we have confirmed that the clusters to which respective binary oxides belong do not change by reducing the number of clusters, except for $HfO_2$. The distributions of clusters with respect to the target property $\varepsilon_{el}$ and the clusters to which binary oxides belong are shown in Figure S7b and Table S6, Supporting Information, respectively.

**Figure 6** shows distributions of clusters with respect to $E_g$ and $\varepsilon_{el}$. We focus on a region where both $E_g$ and $\varepsilon_{el}$ values are relatively large because they are known to be in a trade-off relationship.[40] We find that cluster 19 is distributed near the Pareto front, which is depicted in **Figure 7** in more detail. A binary oxide in cluster 19 is only that of W(VI) or $WO_3$. All the other oxides in cluster 19 also contain a transition-metal atom. For the regression of $\varepsilon_{el}$, the features with the highest and second highest permutation importances are the mass density and the fraction of transition metal atoms. The importance of mass density and its chemical interpretations have been revealed in our previous study.[8] Although all oxides in cluster 19 contain a transition metal atom, the correlation between the fraction of transition metal atoms and $\varepsilon_{el}$ is not apparent, within the whole dataset or within this cluster (**Figure 8**a). Moreover, the mass density is not clearly correlated within cluster 19 (Figure 8b). The other features selected for the regression are also hardly correlated with $\varepsilon_{el}$ within this cluster: the feature measuring the average difference in the number of filled valence *s* electrons between an atom and its neighbors gives the maximum magnitude of the Pearson correlation coefficient with $\varepsilon_{el}$ (0.44); the feature measuring the average difference in the number of empty valence *s* electrons gives the same correlation coefficient because these two features are identical for all oxides in the cluster 19. Although they seem to be moderately correlated with $\varepsilon_{el}$, this correlation might be due to $NbRhO_4$ which is largest in both the features (0.60) and $\varepsilon_{el}$ (10.9). The correlation coefficient decreases to 0.37 if this oxide is omitted. The weak correlations to the features imply that the trade-off relationship within cluster 19 is related to multiple features complexly.

The feature measuring the maximum similarity of local structures to the octahedral coordination[41] is concentrated in large values for cluster 19 (**Figure 9**), though the feature is eliminated by the feature selection performed during the construction of the RF model. We have investigated all the structures in the cluster and found that they contain a metal atom octahedrally coordinated by O atoms. Notable entries in this cluster are perovskite-type oxides, some of which are known as ferroelectric compounds.[42] There are eight in this cluster: tetragonal $SrTiO_3$, rhombohedral $BaTiO_3$, tetragonal $PbTiO_3$, tetragonal $PbVO_3$, rhombohedral $BiFeO_3$, orthorhombic $NaTaO_3$, cubic $KTaO_3$, and rhombohedral $AgTaO_3$. The perovskite-type oxides have a

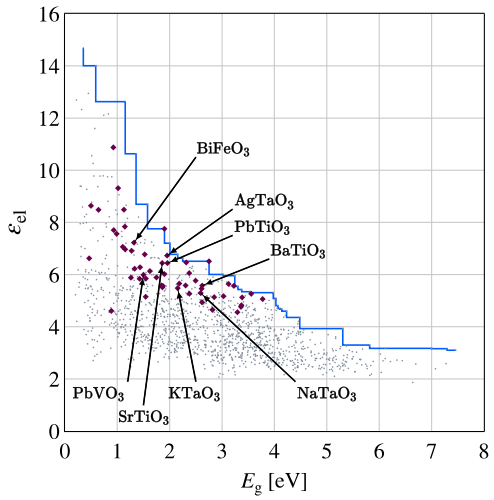


**Figure 7.** Detail of cluster 19 obtained by the clustering for $\varepsilon_{el}$. The step line shows the Pareto front. The diamonds indicate the oxides in cluster 19.

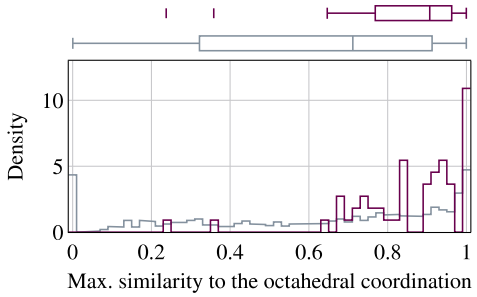


**Figure 9.** Distribution of the maximum similarity to the octahedral coordination in cluster 19 for $\varepsilon_{el}$. The gray histogram and the lower box plot indicates the distribution for the whole dataset. Each histogram is normalized so that the area is equal to one.

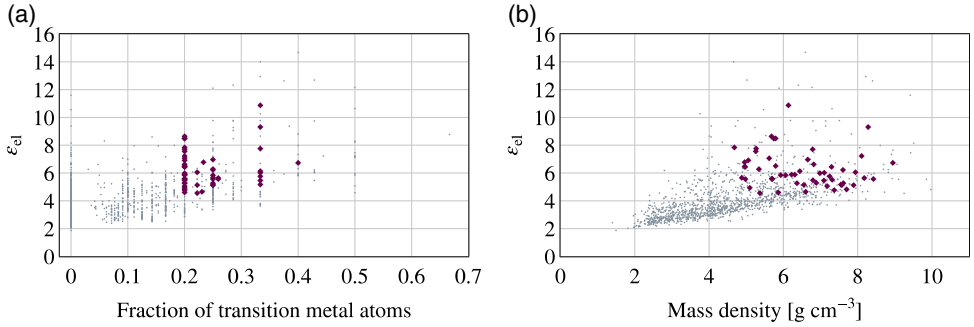


**Figure 8.** Distribution of oxides with respect to the $\varepsilon_{el}$ and a) fraction of transition metal atoms (the most important feature) and b) mass density (the second most important feature). The importance of feature is ranked by random forest permutation importance. The diamonds indicate the oxides included in cluster 19 for $\varepsilon_{el}$.

chemical composition of $ABO_3$ and consist of corner-sharing $BO_6$ octahedra and $A$ atoms surrounded by eight $BO_6$ octahedra (Figure S8a–h, Supporting Information). The $ReO_3$-type structure taken by $WO_3$ is similar to the perovskite-type structure, which consists of corner-sharing $BO_6$ octahedra without $A$ atoms (Figure S8i, Supporting Information). Note that the octahedra of tetragonal $PbTiO_3$ and tetragonal $PbVO_3$ are significantly distorted and exhibit small values of the feature for the maximum similarity of local structures to the octahedral coordination.

$BaTiO_3$ is a transition metal oxide such that the $d$ states of Ti are formally empty: the valence and conduction bands are dominated by O $p$ states and Ti $d$ states, respectively (Figure S9a, Supporting Information). The density of states is high and steep around both the valence band maximum and conduction band minimum, which complements an increase in the $\epsilon_{el}$ with many electronic transition routes from the valence to the conduction band states.[40] Consequently, both $E_g$ and $\epsilon_{el}$ are relatively large.

Although $PbVO_3$ also shows large $E_g$ and $\epsilon_{el}$, it is located rather far from the Pareto front. As well as $BaTiO_3$, the density of states of $PbVO_3$ is high and steep around the band edges (Figure S9b, Supporting Information). The difference is that the $d$ states of V are partially filled: the conduction band is dominated by V $d$ states, while the valence band is characterized by O $p$ states and a comparable contribution of V $d$ states. Since the $d$-$d$ electronic transitions are dipole-forbidden, this band structure makes $\epsilon_{el}$ slightly smaller than that of $BaTiO_3$ despite a narrower $E_g$.[40]

## 3. Conclusion

We have developed a clustering method involving the information about the target property, where the information is injected through a transformation of the features by the RF regression model. A comparison between the clustering for $E_{coh}$ and $E_g$ demonstrates injecting the target property information. An analysis of a narrow-$E_g$ cluster has revealed that features that are characteristic of each cluster are reasonable from the perspective of conventional physical and chemical pictures, but they are not necessarily important for the whole dataset. We have also analyzed a cluster near the Pareto front in the $E_g$–$\epsilon_{el}$ space. The cluster consists mainly of transition metal oxides, and they show a common structural characteristic that a metal atom is octahedrally coordinated by O atoms. The cluster includes several perovskite-type oxides, the electronic structures of which can explain a balance of relatively wide $E_g$ and large $\epsilon_{el}$. Our method enables analyses from viewpoints that are different from the conventional clustering and feature importance analyses by taking the relationship between the target property and the features into account. While we focus on single target property cases in this article for conciseness, our method is extendable to clustering with respect to multiple target properties: we can construct the RF model for each target variable, concatenate the vectors transformed by these models, and perform the cluster analysis for the concatenated vector.

## 4. Computational Section

*Clustering Assisted by the Random Forest*: Our method consists of constructing a RF model for predicting the target property, transforming the features using the model, and clustering the transformed data. A focal point is how the feature vectors are transformed. A schematic view is shown in **Figure 10**.

The RF regression model consists of a set of decision trees. The decision tree divides the feature space into two regions recursively, decomposing the feature space into an irregular grid (Figure 10a). The separation is conducted so that training data within the same region in the feature space have close values of the target variable. The training set of the decision tree is a bootstrap sample, i.e., a random sample from the training set of the RF model with replacement. Each decision tree in the RF model differently separates the feature space due to randomness in the training set and divided features. The RF regression model predicts a target value, $y$, for a feature vector, $\boldsymbol{x}$, by averaging predictions by the decision trees

$$y(x) = \frac{1}{T}\sum_{t=1}^{T} y_t(x) \tag{2}$$

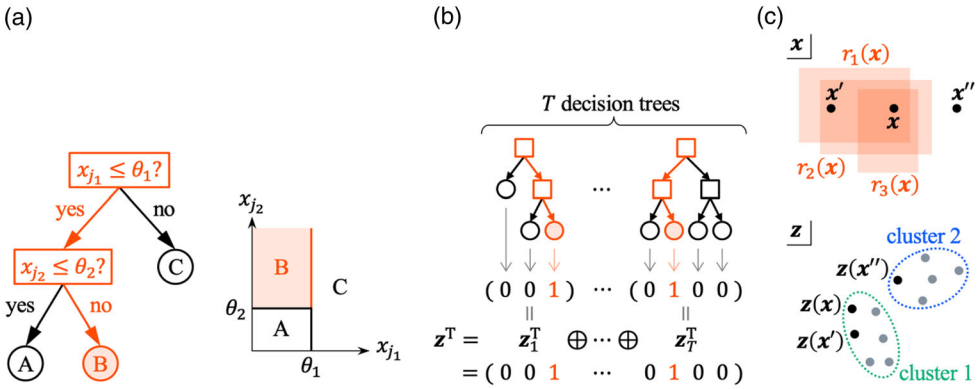


**Figure 10.** Schematic views of the clustering assisted by the RF regression. a) Region of the feature space to which a feature vector, $\boldsymbol{x}$, belongs. For the $t$th decision tree of the RF model, if the $j_1$th component of $\boldsymbol{x}$, $x_{j_1}$, is less than or equal to $\theta_1$ and $x_{j_2}$ is greater than $\theta_2$, $\boldsymbol{x}$ belongs to the region labeled B, hence $r_t(\boldsymbol{x}) = \text{B}$. b) Transformation of $\boldsymbol{x}$. For the $t$th decision tree, the region to which $\boldsymbol{x}$ belongs is represented by $\boldsymbol{z}_t$ whose component is one if corresponding to $r_t(\boldsymbol{x})$ and zero otherwise. The transformed vector, $\boldsymbol{z}$, is a concatenation of all $\boldsymbol{z}_t$. c) Clustering in the $\boldsymbol{z}$-space. In the feature space, $\boldsymbol{x}$ belongs to the regions labeled $r_1(\boldsymbol{x})$, $r_2(\boldsymbol{x})$, $r_3(\boldsymbol{x})$, and so on. The more (less) regions a data point is contained in, the more similar (dissimilar) to $\boldsymbol{x}$ the data point is. For example, $\boldsymbol{x}'$ is more similar to $\boldsymbol{x}$ than $\boldsymbol{x}''$. The relationship between the number of the containing regions and the similarity (dissimilarity) measure is defined as a closeness (distance) in the $\boldsymbol{z}$-space. A set of $\boldsymbol{z}$ is divided into clusters so that close ones are in the same clusters and distant ones are in different clusters.

where $T$ is the number of decision trees and $y_t(\boldsymbol{x})$ is a target value predicted by the $t$th decision tree. The prediction by the decision tree is constant within each separated region, which is an average target value of the training data belonging to the region

$$y_t(x) = \frac{1}{|\Lambda_{tr_t(x)}|} \sum_{i \in \Lambda_{tr_t(x)}} y_i \tag{3}$$

where $r_t(\boldsymbol{x})$ is a label of the region of the $t$th decision tree to which $\boldsymbol{x}$ belongs, $\Lambda_{tr}$ is a bag of indices of training data used by the $t$th decision tree and belonging to the region labeled $r$, and $y_i$ is a target value of the $i$th data point in the training set of the RF model.

Using the RF model, $\boldsymbol{x}$ can be nonlinearly transformed into a vector of the region labels to which it belongs: $[r_1(\boldsymbol{x}), ..., r_T(\boldsymbol{x})]^{\mathrm{T}}$. This categorical vector can be converted into a binary vector by the one-hot encoding. We denote the binary vector by $\boldsymbol{z}(\boldsymbol{x})$

$$\mathbf{z}(\mathbf{x}) = \mathbf{z}_1(\mathbf{x}) \oplus \ldots \oplus \mathbf{z}_T(\mathbf{x}) \tag{4}$$

where $\boldsymbol{z}_t(\boldsymbol{x})$ is a one-hot representation of $r_t(\boldsymbol{x})$, that is, the number of dimensions is equal to the number of leaf nodes of the $t$th decision tree, and a component is one if corresponding to $r_t(\boldsymbol{x})$ and zero otherwise (Figure 10b). The numbers of leaf nodes of the decision trees are determined by hyperparameters of the RF model, which would be tuned to give accurate predictions by the RF model. Note that information about the target variable is involved in $\boldsymbol{z}$: the transformation from $\boldsymbol{x}$ requires $r_t(\boldsymbol{x})$, hence a regression of the target variable.

The clustering is performed in the $\boldsymbol{z}$-space. The clustering method can be anything applicable to binary vectors. The clustering divides a set of data points so that similar ones are in the same clusters and dissimilar ones are in different clusters (Figure 10c). The (dis)similarity measure depends on the clustering method: the Euclidean distance is one of them. The clustering in the $\boldsymbol{z}$-space gathers data points similar in the target variable as well as the features because the RF model is trained so that data in the same region has close target values. The (dis)similarities among $\boldsymbol{z}$ can be different even if those among $\boldsymbol{x}$ are the same. It is a consequence of the nonlinear transformation or involving the target variable.

The clustering in the $\boldsymbol{z}$-space is also encouraged because the cosine similarity between $\boldsymbol{z}$ is identical to the proximity measure defined in ref. [29]. The proximity measure is inherent in a trained RF model, and that between two feature vectors is defined as the proportion of decision trees at which the feature vectors belong to the same region:[29] it is written for the feature vectors of $\boldsymbol{x}$ and $\boldsymbol{x}'$ as $\sum_{t=1}^{T} I[r_t(x) = r_t(x')]/T$, where $I[\cdot]$ is the indicator function. Using $\boldsymbol{z}$, the proximity can be rewritten as $\mathbf{z}(\mathbf{x})^{\mathrm{T}}\mathbf{z}(\mathbf{x}')/T$, which is the cosine similarity between $\boldsymbol{z}(\boldsymbol{x})$ and $\boldsymbol{z}(\boldsymbol{x}')$. In other words, the clustering in the $\boldsymbol{z}$-space is a generalization of the clustering described in ref. [29]. If the RF model is for the classification problem and the clustering in the $\boldsymbol{z}$-space is performed with the cosine similarity and it is equivalent to the method of ref. [29].

In our demonstrations for different target properties, the RF regression model is constructed for each target property by the scikit-learn library.[43] Hyperparameters are tuned by the fivefold cross-validation. The number of features is reduced by the recursive feature elimination:[44] the number is determined by the fivefold cross-validation using the tuned hyperparameters. The features are ranked by the permutation feature importance,[28] and those in the lowest 20% are removed recursively for each fold. The model used for the clustering is trained using the whole dataset with the tuned hyperparameters and the selected number features. Reducing the number of features facilitates the investigation into clusters concerning the features. The $\varepsilon_{\mathrm{el}}$ values are transformed by the common logarithm beforehand because large numerical errors are expected for large $\varepsilon_{\mathrm{el}}$ values.

The clustering in the $\boldsymbol{z}$-space is performed by the agglomerative hierarchical clustering implemented in the SciPy library.[45] The agglomerative hierarchical clustering recursively merges a closest pair of clusters into a new cluster, where a data point is treated as a cluster consisting only of the data point. The closest pair is determined by the distances between data points. The way of determining the closest pair is called linkage criterion. In short, the method of the agglomerative hierarchical clustering is specified by the distance metric for data points and the linkage criterion. We perform the clustering with four methods: combinations of two distance metrics and two linkage criteria. We use the cosine distance and Jaccard distance as the distance metric, and the average method (the unweighted pair group method with arithmetic mean), and the Ward method as the linkage criteria. We show only results with the combination of the cosine distance and the average method in the main text because we find that the four combinations end in similar groupings of binary oxides. Results with the other combinations are presented in the Supporting Information.

*Datasets*: We prepare two datasets for different target properties: one for $E_{\mathrm{coh}}$ and $E_{\mathrm{g}}$, and the other for $\varepsilon_{\mathrm{el}}$. Both datasets are collected from the Materials Project database,[33,34] a collection of properties computed based on density functional theory with the Perdew–Burke–Ernzerhof parametrization of the generalized gradient approximation[46] and Hubbard $U$ corrections.[47] The dataset for $E_{\mathrm{coh}}$ and $E_{\mathrm{g}}$ consists of 7981 compounds satisfying the following conditions: 1) O atoms are contained, 2) H and noble gas atoms are not contained, 3) anions are only $O^{2-}$, 4) the total energy is the lowest among polymorphs, 5) the formation energy is less than 0.1 eV atom$^{-1}$ against that of a mixture of competing phases, 6) $E_{\mathrm{g}}$ is determined, and 7) $E_{\mathrm{g}}$ is larger than or equal to 0.2 eV. Oxidation states of atoms (ions) are determined by the pymatgen library[48] based on given chemical compositions. The dataset for $\varepsilon_{\mathrm{el}}$ consists of 1301 compounds satisfying the following conditions in addition to above (1–7): (8) the static dielectric tensor is computed, and (9) $\varepsilon_{\mathrm{el}}$ is less than 50. Note that the collected $E_{\mathrm{g}}$ and $\varepsilon_{\mathrm{el}}$ values tend to be underestimated[49,50] and overestimated[49] compared to experimental values, respectively, owing to the generalized gradient approximation, especially for compounds treated without the Hubbard $U$ corrections. The distributions of the target properties are shown in Figure S1, Supporting Information.

We generate features of oxides from crystal structures by the matminer library[38] (see Table S1, Supporting Information for the generated features). After omitting nonnumerical features, constant features, and features not available for all oxides in the datasets, we obtain 696 features for $E_{\mathrm{coh}}$ and $E_{\mathrm{g}}$, and 662 features for $\varepsilon_{\mathrm{el}}$.

## Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

## Acknowledgements

This work was supported by JST CREST grant no. JPMJCR17J2, JSPS KAKENHI grant no. JP20H00302, and 21K14401, MEXT Data Creation and Utilization Type Material Research and Development Project grant no. JPMXP1122683430, MEXT Design and Engineering by Joint Inverse Innovation for Materials Architecture Project, and KISTEC Project. The computing resource of the Academic Center for Computing and Media Studies at Kyoto University was used for part of this work.

## Conflict of Interest

The authors declare no conflict of interest.

## Data Availability Statement

The data that support the findings of this study are openly available in Materials project database at https://legacy.materialsproject.org, reference number [31]. The underlying code for this work is available at https://github.com/nbsato/forestcluster.

## Keywords

clustering, inorganic compounds, interpretable artificial intelligence, random forest

Received: March 28, 2024
Revised: July 1, 2024
Published online:

[1] K. Rajan, *Mater. Today* **2005**, *8*, 38.
[2] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.
[3] W. Sha, Y. Guo, Q. Yuan, S. Tang, X. Zhang, S. Lu, X. Guo, Y.-C. Cao, S. Cheng, *Adv. Intell. Syst.* **2020**, *2*, 1900143.
[4] C. Yan, G. Li, *Adv. Intell. Syst.* **2023**, *5*, 2200243.
[5] F. Oviedo, J. L. Ferres, T. Buonassisi, K. T. Butler, *Acc. Mater. Res.* **2022**, *3*, 597.
[6] G. Pilania, *Comput. Mater. Sci.* **2021**, *193*, 110360.
[7] X. Zhong, B. Gallagher, S. Liu, B. Kailkhura, A. Hiszpanski, T. Y.-J. Han, *NPJ Comput. Mater.* **2022**, *8*, 204.
[8] A. Takahashi, Y. Kumagai, J. Miyamoto, Y. Mochizuki, F. Oba, *Phys. Rev. Mater.* **2020**, *4*, 103801.
[9] N. T. P. Hartono, J. Thapa, A. Tiihonen, F. Oviedo, C. Batali, J. J. Yoo, Z. Liu, R. Li, D. F. Marrón, M. G. Bawendi, T. Buonassisi, S. Sun, *Nat. Commun.* **2020**, *11*, 4172.
[10] K. Choudhary, K. F. Garrity, V. Sharma, A. J. Biacchi, A. R. Hight Walker, F. Tavazza, *NPJ Comput. Mater.* **2020**, *6*, 64.
[11] S. M. Lundberg, S.-I. Lee, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., Glasgow, Scotland **2017**.
[12] K. Morita, D. W. Davies, K. T. Butler, A. Walsh, *J. Chem. Phys.* **2020**, *153*, 024503.
[13] S. Fujii, Y. Shimizu, J. Hyodo, A. Kuwabara, Y. Yamazaki, *Adv. Energy Mater.* **2023**, *13*, 2301892.
[14] T. A. R. Purcell, M. Scheffler, L. M. Ghiringhelli, C. Carbogno, *NPJ Comput. Mater.* **2023**, *9*, 112.
[15] Y. Noda, M. Otake, M. Nakayama, *Sci. Technol. Adv. Mater.* **2020**, *21*, 92.
[16] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L. M. Ghiringhelli, *Phys. Rev. Mater.* **2018**, *2*, 083802.
[17] L. Sbailò, Á. Fekete, L. M. Ghiringhelli, M. Scheffler, *NPJ Comput. Mater.* **2022**, *8*, 250.
[18] F. R. Burden, D. A. Winkler, *QSAR Comb. Sci.* **2009**, *28*, 645.
[19] P. Mikulskis, M. R. Alexander, D. A. Winkler, *Adv. Intell. Syst.* **2019**, *1*, 1900045.
[20] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, *Nature* **2019**, *571*, 95.
[21] T. Gupta, M. Zaki, N. M. A. Krishnan, Mausam, *NPJ Comput. Mater.* **2022**, *8*, 102.
[22] I. E. Castelli, K. W. Jacobsen, *Model. Simul. Mater. Sci. Eng.* **2014**, *22*, 055007.
[23] W. Sun, C. J. Bartel, E. Arca, S. R. Bauers, B. Matthews, B. Orvañanos, B.-R. Chen, M. F. Toney, L. T. Schelhas, W. Tumas, J. Tate, A. Zakutayev, S. Lany, A. M. Holder, G. Ceder, *Nat. Mater.* **2019**, *18*, 732.
[24] E. L. Willighagen, R. Wehrens, P. Verwer, R. de Gelder, L. M. C. Buydens, *Acta Crystallogr. B* **2005**, *61*, 29.
[25] B. Meredig, C. Wolverton, *Chem. Mater.* **2014**, *26*, 1985.
[26] T. L. Pham, H. Kino, K. Terakura, T. Miyake, H. C. Dam, *J. Chem. Phys.* **2016**, *145*, 154103.
[27] R. Tamura, M. Matsuda, J. Lin, Y. Futamura, T. Sakurai, T. Miyazaki, *Phys. Rev. B* **2022**, *105*, 075107.
[28] L. Breiman, *Mach. Learn.* **2001**, *45*, 5.
[29] L. Breiman, https://www.stat.berkeley.edu/users/breiman/wald2002-2.pdf (accessed: January 2023).
[30] F. Moosmann, B. Triggs, F. Jurie, in *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA **2007**.
[31] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, Cambridge, MA **2016**.
[32] G. Zhong, L.-N. Wang, X. Ling, J. Dong, *J. Finance Data Sci.* **2016**, *2*, 265.
[33] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, *1*, 011002.
[34] S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, K. A. Persson, *Comput. Mater. Sci.* **2015**, *97*, 209.
[35] T. Calinski, J. Harabasz, *Commun. Stat. Theory Methods* **1974**, *3*, 1.
[36] P. J. Steinhardt, D. R. Nelson, M. Ronchetti, *Phys. Rev. B* **1983**, *28*, 784.
[37] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, I. Tanaka, *Phys. Rev. B* **2017**, *95*, 144110.
[38] L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, A. Jain, *Comput. Mater. Sci.* **2018**, *152*, 60.
[39] N. E. R. Zimmermann, A. Jain **2017**, in progress.
[40] F. Naccarato, F. Ricci, J. Suntivich, G. Hautier, L. Wirtz, G.-M. Rignanese, *Phys. Rev. Mater.* **2019**, *3*, 044602.
[41] N. E. R. Zimmermann, M. K. Horton, A. Jain, M. Haranczyk, *Front. Mater.* **2017**, *4*, 34.
[42] R. E. Cohen, *Nature* **1992**, *358*, 136.
[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825.
[44] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, *Mach. Learn.* **2002**, *46*, 389.
[45] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, et al., *Nat. Methods* **2020**, *17*, 261.
[46] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865.
[47] V. I. Anisimov, J. Zaanen, O. K. Andersen, *Phys. Rev. B* **1991**, *44*, 943.
[48] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2013**, *68*, 314.
[49] F. Oba, Y. Kumagai, *Appl. Phys. Express* **2018**, *11*, 060101.
[50] Y. Hinuma, A. Grüneis, G. Kresse, F. Oba, *Phys. Rev. B* **2014**, *90*, 155405.