Science and Technology of Advanced Materials: Methods

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tstm20

Atomic descriptors generated from coordination polyhedra in crystal structures

Yuki Inada, Yukari Katsura, Masaya Kumagai & Kaoru Kimura

To cite this article: Yuki Inada, Yukari Katsura, Masaya Kumagai & Kaoru Kimura (2021) Atomic descriptors generated from coordination polyhedra in crystal structures, Science and Technology of Advanced Materials: Methods, 1:1, 200-212, DOI: 10.1080/27660400.2021.1986359

To link to this article: https://doi.org/10.1080/27660400.2021.1986359

SCIENCE AND TECHNOLOGY OF ADVANCED MATERIALS: METHODS
2021, VOL. 1, NO. 1, 200–212
https://doi.org/10.1080/27660400.2021.1986359

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

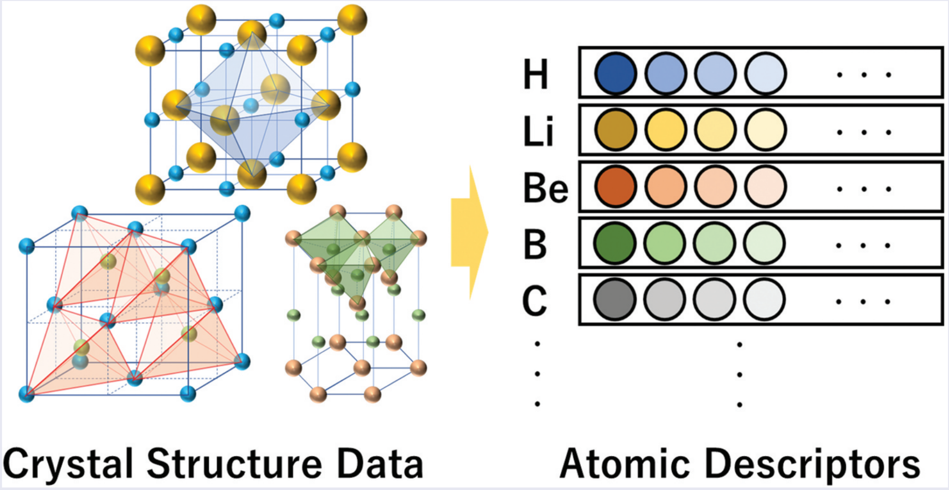
Atomic descriptors generated from coordination polyhedra in crystal structures

Yuki Inada[a], Yukari Katsura[a,b,c], Masaya Kumagai[c,d,e] and Kaoru Kimura[a]

[a]Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan; [b]Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science, Ibaraki, Japan; [c]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan; [d]SAKURA Internet Research Center, SAKURA Internet Inc, Osaka, Japan; [e]Institute for Integrated Radiation and Nuclear Science, Kyoto University, Osaka, Japan

ABSTRACT

We developed atomic descriptors from local crystal structures, which will facilitate researchers' use of machine learning to predict the properties of inorganic materials via materials informatics. We applied singular value decomposition to the occurrence matrix of local coordination polyhedra in crystal structures. We generated two atomic descriptors, each based on the coordination atoms and topology of the coordination polyhedra. As a result of atomic clustering using these descriptors, the composition descriptor proposed in previous research depends on the similarity between same-group atoms in the periodic table. In contrast, when using our original descriptors based on the coordination atoms and topology of the coordination polyhedra, the similarity between adjacent atoms in the periodic table as well as the similarity between same-group atoms was pertinent. When we used machine learning to predict the formation energy and band gap using these descriptors as inputs, the prediction accuracy and generalization ability increased compared with using a physical property descriptor.



ARTICLE HISTORY
Received 21 May 2021
Revised 9 September 2021
Accepted 22 September 2021

KEYWORDS
Materials informatics; machine learning; atomic descriptors; coordination polyhedra; formation energy prediction; band gap prediction

CLASSIFICATIONS
Data analysis (AI, machine learning, data-driven analysis, descriptor development, structure search/identification)

1. Introduction

Machine learning is pertinent to materials informatics (MI). In supervised machine learning, target values can be predicted from input feature vectors by using a program that infers the relationship between the two from a large quantity of data.

Because the structural chemistry of inorganic materials is an especially challenging line of work, MI may help researchers design inorganic materials. The Materials Project [1] and analogous open databases are repositories of crystal structure data based on experiments and theoretical predictions. By using these databases, researchers are developing advanced predictive methods for materials properties.

Chemical composition is readily accessible information for most inorganic materials. Crystal structure is needed for proper prediction of materials properties, but it requires a lot of effort to determine the crystal structure by experiment or calculation. In addition, for materials discovery, crystal structure is not always controllable and considering the crystal structure produces a result of large searching space.

CONTACT Yuki Inada inada-yuki021@g.ecc.u-tokyo.ac.jp Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

Supplemental data for this article can be accessed here.

Predicting materials properties from chemical composition is beneficial approach because chemical composition is most controllable parameter in experiment. Using the composition as the input for machine learning, researchers have predicted various pertinent materials properties, such as the formation energy [2–7], the band gap [8], the transition temperature of superconductivity [9,10], elastic properties [11], and thermoelectric properties [12]. Researchers have also related compositions to crystal structures by machine learning [13–15]. Materials properties predictions by machine learning are fast and facilitate screening of large datasets [2,12,16,17].

Appropriate selection of input descriptors is essential for reliable machine learning. When using chemical composition as input data for machine learning, atomic descriptors are often used to convert chemical composition into a set of numerical values (a vector).

To date, researchers have used knowledge of the physical and chemical properties of elements to develop atomic descriptors. Researchers commonly base the similarity between elements on the periodic table proposed by Mendeleev, which arranges elements based on, e.g. atomic electron configuration similarities. Researchers commonly use, e.g., group number or period number in machine learning. Pettifor proposed Mendeleev numbers [18] as another method of representing atomic similarities; this approach is also common among researchers. In Mendeleev numbers, elements are arranged in a one-dimensional sequence based on consolidating pertinent atomic properties.

Classical sets of atomic descriptors consist of physical properties of the elements. Major libraries – such as Pymatgen [19], XenonPy [20], Magpie [21], and Matminer [22] – include such empirical descriptors. Dozens of physical properties are available and researchers have reported hundreds of descriptors by corresponding cross-operations. However, most of these descriptors are physically similar. Many descriptors are related to electronegativity, calculated by various methods. Other descriptors are related to atomic size, such as atomic or ionic radius, calculated by various methods. Most of the remaining descriptors are non-fundamental physical properties of elemental crystals – such as elastic modulus, melting point, boiling point, electrical conductivity, thermal conductivity, and superconducting transition temperature – and strongly depend on the bonding type and crystal structure of the elements. Because these properties make no distinction between different crystals containing the same element, researchers should not use them as descriptors for the elements within crystals. Despite the large number of descriptors available, the chemical properties of the elements are not fully reflected by any of these descriptors. Accordingly, researchers need descriptors that represent atomic features and similarities in an alternative format.

There are approaches to extract atomic features from materials data by using data science and unsupervised machine learning [18,23–26]; for example, researchers have attempted to improve the representation of atomic features by data-driven reconstruction of the periodic table [23–25] and the Mendeleev number [18]. There was an approach to generate new atomic descriptors from materials statistical data [26]. Atom2Vec [26] by Zhou *et al.* is an elegant approach to attain chemical descriptors by only using an inorganic materials database. These researchers used the list of chemical formulas recorded in the Materials Project [1], and generated an occurrence matrix of the chemical formulas within the database. For example, from the information that there are compounds of compositions BiTe and Bi_2Te_3, the values of '(1)Te1' and '(2)Te3' for Bi become 1, and '(1)Bi1' and '(3)Bi2' for Te become 1. For the remaining composition patterns, the values were set to 0. This was performed for all compositions included in the Materials Project. Then they obtained an occurrence matrix and normalized it, to obtain a sparse matrix χ that has N (number of elements) rows and M (number of composition formula patterns) columns. By performing singular value decomposition, this matrix χ is converted to $\chi = UDV^T$, where U is an $N \times N$ square matrix, D is an $N \times M$ diagonal matrix, and V is an $M \times M$ square matrix. From this, a square matrix $F = UD = \{f_1, f_2, \ldots, f_N\}$ can be obtained. These horizontal vectors $\{f_1, f_2, \ldots, f_N\}$ of length N are a group of descriptors that represent the characteristics of each element. The importance of each descriptor is reflected in the magnitude of the corresponding singular value.

Because Atom2Vec descriptors were generated from composition data, the descriptors mainly reflect the concept of valency. However, crystal structure data contain more information about atoms and their interactions. Accordingly, atomic descriptors based on the relationships between atoms and corresponding crystal structures may provide comparatively more information than composition descriptors, and represent atomic characteristics in a more useful manner.

In a crystal structure, researchers assume local structures to be affected largely by the characteristics of each atom. Villars [27] studied this relationship by plotting coordination polyhedra in binary crystals on a three-dimensional map, where the axes were the average number of valence electrons, the difference in electronegativity, and the difference in atomic radius. Using this map, the regions can be clearly divided in which each type of coordination polyhedra appears. Thus, determination of the coordination polyhedra largely depends on the properties of the atoms. Therefore, descriptor vectors that reflect the properties of the atoms can be generated more directly from information on constituent atoms or shape of the coordination polyhedra than from composition formula.

We applied the Atom2Vec approach to information on local structure obtained with ToposPro [28], software that analyzes coordination polyhedra in crystal structures. In ToposPro a coordination polyhedron is based on the Voronoi polyhedron of each atom. The shape of each coordination polyhedron is expressed by the topology of polyhedra, using point symbols. Point symbols are determined by the connections between vertices and edges of the polyhedra. A polyhedron with n vertices has $n(n-1)/2$ pairs of vertices, and the shortest cycle for each pair can be identified. Point symbols are determined by the result of counting the number of vertices of each cycle [29]. For example, in Figure 1, there are 12 pairs of vertices that are linked by a single edge on a cube; adding the two paths from the center atom, the path length of the cycle is 3. This information is represented as 3^{12}. The number of vertex pairs linked by two edges is 12 and that linked by three edges is four. Accordingly, the point symbol representation of the cube becomes $3^{12}.4^{12}.5^{4}$.

To confirm that the characters of the elements are reflected in the generated atomic descriptors, we performed unsupervised machine learning on the descriptor vectors to cluster the similar elements. To verify the effectiveness of these descriptors for

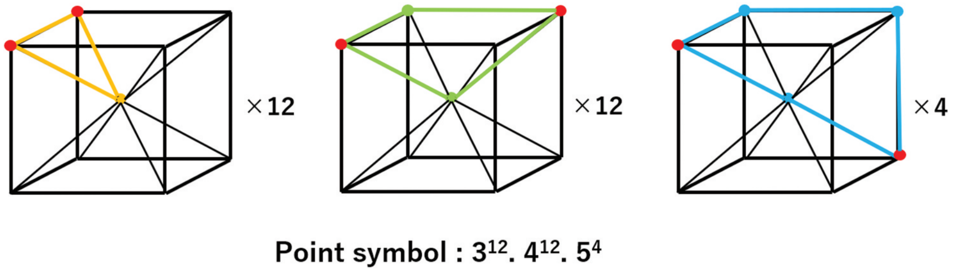


Figure 1. Representation of the topology of coordination polyhedra by point symbol.

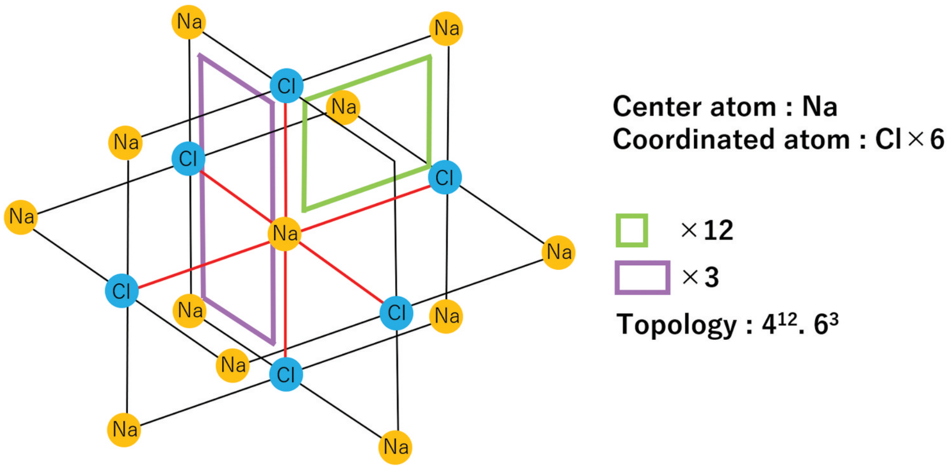
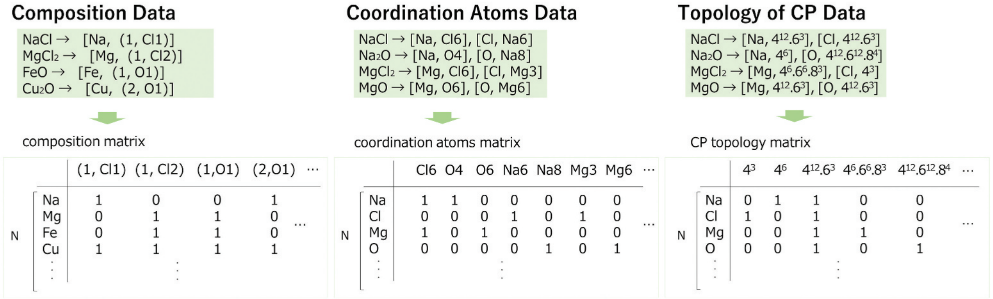


Figure2. Example of the topology of coordination polyhedra of Na in NaCl crystal.



Composition Data

NaCl → [Na, (1, Cl1)]
$MgCl_2$ → [Mg, (1, Cl2)]
FeO → [Fe, (1, O1)]
Cu_2O → [Cu, (2, O1)]

composition matrix

N		(1, Cl1)	(1, Cl2)	(1,O1)	(2,O1)	···
	Na	1	0	0	1	
	Mg	0	1	1	0	···
	Fe	0	1	1	0	
	Cu	1	1	1	1	
	⋮		⋮			

Coordination Atoms Data

NaCl → [Na, Cl6], [Cl, Na6]
Na_2O → [Na, O4], [O, Na8]
$MgCl_2$ → [Mg, Cl6], [Cl, Mg3]
MgO → [Mg, O6], [O, Mg6]

coordination atoms matrix

N		Cl6	O4	O6	Na6	Na8	Mg3	Mg6	···
	Na	1	1	0	0	0	0	0	
	Cl	0	0	0	1	0	1	0	···
	Mg	1	0	1	0	0	0	0	
	O	0	0	0	0	1	0	1	
	⋮				⋮				

Topology of CP Data

NaCl → [Na, $4^{12}.6^{3}$], [Cl, $4^{12}.6^{3}$]
Na_2O → [Na, 4^{6}], [O, $4^{12}.6^{12}.8^{4}$]
$MgCl_2$ → [Mg, $4^{6}.6^{6}.8^{3}$], [Cl, 4^{3}]
MgO → [Mg, $4^{12}.6^{3}$], [O, $4^{12}.6^{3}$]

CP topology matrix

N		4^{3}	4^{6}	$4^{12}.6^{3}$	$4^{6}.6^{6}.8^{3}$	$4^{12}.6^{12}.8^{4}$	···
	Na	0	1	1	0	0	
	Cl	1	0	1	0	0	···
	Mg	0	0	1	1	0	
	O	0	0	1	0	1	
	⋮			⋮			

Figure 3. Example of the matrix output by each environment (composition, coordination atoms, and topology of coordination polyhedra).

machine learning, we developed machine learning models that predict the formation energy from these descriptors by using the data of the Materials Project [1] as the training data. We compared the prediction accuracy among the descriptors generated by three datasets: the compositions, coordinating elements, and topologies of the coordination polyhedra.

2. Methods

2.1. Dataset

The crystal structure data, including the optimized atomic coordinates and calculated formation energies, were obtained from the Materials Project [1]. Structures containing elements with almost no entries (noble gases and artificial elements other than Pu and Np) were excluded. From the remaining structures, 35,746 structures with energy_above_hull = 0 were selected as stable crystal structures. These data mainly composed of insulator, semiconductor, and intermetallic compounds. From these, 32,000 structures were randomly selected to generate atomic descriptors. The remaining 3,746 structures were used to evaluate the prediction accuracy of the machine learning models that used the generated descriptors.

2.2. Generation of atomic descriptors

From the selected crystal structure datasets, three occurrence matrices were generated from the information of the overall compositions, coordinating atoms, and topologies of the coordination polyhedra. The 'overall compositions' descriptors were designed to resemble the descriptors reported in Atom2Vec [26]. The coordinating atoms and topologies of coordination polyhedra were obtained with ToposPro 5. 3 [28]. In this process, some geometrical information such as bond length was ignored.

In Atom2vec, the 'environment' for each element was defined simply from the overall composition formulas [26]. In this study, more-detailed local environments were defined by listing the coordination environments expressed in two ways, constituent atoms of coordination polyhedra and topology of coordination polyhedra. The former one was based on the number and species of atoms coordinated to the central atom in the coordination polyhedra. Regarding the topology of coordination polyhedra, environments were defined as point symbol representations of polyhedra. When one or more of these environments were evident, the value of the position of [atom, environment] in the matrix was set to 1. For each descriptor, the environment that appeared only once was excluded. The values in each row were normalized by dividing by the square root of the sum. The matrix was subjected to singular value decomposition using the SciPy library [30]. As a result, $\chi = \mathrm{UDV}^{\mathrm{T}}$ (where U is an $N \times N$ square matrix, D is an $N \times M$ diagonal matrix, and V is an $M \times M$ square matrix) was obtained and $\mathrm{F} = \mathrm{UD} = \{\mathrm{f}_{\mathrm{H}}, \mathrm{f}_{\mathrm{Li}}, \ldots, \mathrm{f}_{\mathrm{Pu}}\}$ became the atomic descriptors. The descriptor vectors had 80 dimensions (equal to the number of atomic species used for generating the descriptors). The descriptors were as follows: from the composition, CMP; from the coordination atoms, CRD; and from the topology of the coordination polyhedra, TPL.

2.3. Similarity evaluation of elements by using atomic descriptors

Using each of the three descriptors, elements were hierarchically clustered by the Ward method [31]. The results of the clustering were represented by dendrograms, which indicate the discrepancy between the clusters by the length of the feet in the dendrograms.

We also evaluated the similarity of the elements by taking the cosine similarity between the descriptor vectors of the two elements. When defining the similarity between elements by using atomic physical properties, each property was defined in different scaling. At the descriptors generated by Atom2Vec method, the importance of each component is reflected in singular value and similarity of elements can be easily calculated without considering which component should be prioritized. These results were compared with the differences between the elements evaluated from the known physical properties of the elements (electronegativity, atomic radius, and maximum oxidation states).

2.4. Formation energy and band gap prediction

To compare the performance of each set of the atomic descriptors in machine learning, machine learning models were trained that predict the formation energy and band gap by using these descriptors. The following six cases were tested: (1) with atomic descriptors based on standardized physical properties (Pymatgen [19]); (2) without element descriptors; (3) with Atom2Vec-like atomic descriptors based on overall compositions (CMP); (4) with our original atomic descriptors generated from the coordinating atoms (CRD); (5) with our original atomic descriptors generated from the topology of the coordination polyhedra (TPL); and (6) with CMP, CRD, and TPL descriptors altogether (ALL). To convert the overall composition of the compound to the input vector, the weighted average of these descriptors was calculated. The reason that weighted average was employed was that this is the simplest way to convert composition to input descriptor. Our purpose is to evaluate the performance of elements descriptors, not to consider the method to convert composition to input descriptors

Table 1. Atomic properties obtained with Pymatgen.

Atomic properties	Classification Booleans	Physical properties of the elemental phases
atomic number	transition metal	molar volume
atomic mass	rare earth metal	thermal conductivity
atomic radius	Metal	boiling point
average ionic radius	alkali metal	melting point
electronegativity	alkaline earth metal	boiling point – melting point
max oxidation state	Halogen	
minimum oxidation state	Lanthanoid	
Mendeleev number	Actinoid	
group		
row		

by technical numerical operations between descriptors components. Table 1 shows the physical properties of each element obtained with Pymatgen [19]. Physical properties were selected that were available for all of the target elements.

There were 80 components of the descriptor vectors generated in this research. The number of descriptor components is important for prediction accuracy; either too many or too few are inappropriate for machine learning. To evaluate the influence of the number of descriptor components, each descriptor was changed with its number of components from 10 to 80 in 10 increments and prediction accuracies by using each descriptor were measured. In descriptors generated by using singular value decomposition, the importance of each descriptor component was represented in a corresponding singular value. When we use n components from each descriptor, descriptor components that have top n singular value were selected.

A four-layer neural network was used as a learning model with TensorFlow [32]. Weighted average of each descriptor component based on the composition were used for inputs. The structures of the neural network were the same for all of the descriptors. [Input layer (80)] – [hidden layer (128 – 64 – 16)] – [output layer (1)] (the numbers in brackets are the number of nodes in each layer), and ReLU was used for the hidden layer as the activation function. The mean square error was used for the loss function. Learning was performed at a rate of 0.001 with AdamOptimizer [33] for updating the weights. Five-fold cross validation was performed on 32,000 training data points. Training was performed 10× for each fold and the corresponding average was used as results. The prediction accuracy was evaluated based on the mean absolute error (MAE) of the prediction values of the formation energy in 3,746 data points (test data), which were not used for generating the descriptor.

In MI, a sufficient quantity of training data is not always obtainable. In addition, a search for new materials may require predicting the physical properties of materials, including elements, in which the training data are insufficient. If the atomic descriptors correctly represent the similarity between elements, machine learning models should predict the target values reasonably well even if the training dataset lacks some elements. For this purpose, there is a method to evaluate the prediction accuracy for data including unknown elements X [34]. The original dataset was the data from 35,746 compounds obtained from the Materials Project [1]. First, training datasets consisted of the compounds not including target elements X in their constituent elements. Test datasets consisted of the compounds including target elements X in their constituent elements. All 80 elements used for generating the descriptors were used as the target unknown elements X, and 80 datasets were made. The same neural network model as the aforementioned model was employed. The following six cases were tested: (1) with element descriptors based on standardized physical properties (Pymatgen [19]); (2) without element descriptors; (3) with CMP descriptor; (4) with CRD descriptor; (5) with TPL descriptor; and (6) with CMP, CRD, and TPL descriptors altogether (ALL). For each element, prediction models were trained by each training data point and the prediction accuracy was evaluated by quantitating the prediction error of each test data. Prediction accuracy was measured 10× for each target element and descriptor. The number of descriptor components was changed from 10 to 80 in 10 increments and components that have high singular values were used in priority.

3. Result and discussion

3.1. Results of atomic clustering using each created descriptor

Figure 4 shows the results of elements clustering by Ward method and the atomic descriptor vectors.

The similarity of rare earth elements was high for all of the descriptors. The clustering program based on CMP and CRD descriptors divided rare earths into trivalent light rare earths cluster, trivalent heavy rare earths cluster, and Eu and Yb. Divalent rare earth elements Eu and Yb were classified into a cluster of alkaline earth metals. In the CMP descriptor, both Y and Sc were in the rare-earth cluster, whereas in the CRD and TPL descriptors Sc was classified into the Zr and Hf cluster. In the TPL descriptor, the clustering program divided rare earths elements into light rare earths and Gd cluster, and remaining heavy rare earths cluster.

Using the CMP descriptors, the similarity of valence electrons was high regarding typical metal elements. There was higher similarity of Be and Mg to group 12 elements compared with group 2 elements. The clustering program classified the following into close clusters: halogen and alkali metals, and

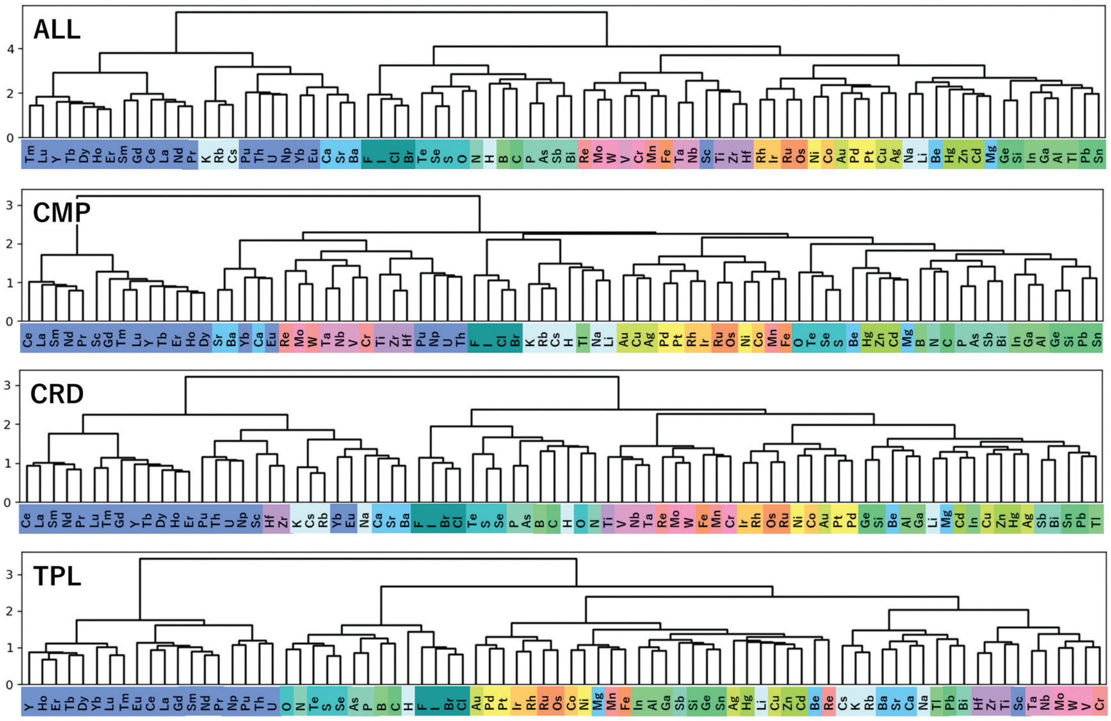


Figure 4. Results of elemental clustering by using each descriptor.

chalcogens and group 12 elements, probably because they appeared in the same ionic compound in many cases. However, in the CRD and TPL descriptors – based on the coordination structures – alkali metals and alkaline earth metals were classified into close clusters. In addition, Na was classified into alkaline earth clusters. Thus, the effect of atomic size was more appreciable than the valence state. In other typical elements, elements that have large atomic numbers and are close to each other on the periodic table.

For non-metallic elements, halogens formed one cluster, common to all descriptors. Chalcogens also formed one cluster. However, regarding the TPL descriptor, N was classified in this cluster because of the similarity to O, and regarding the CRD descriptor, O was classified into the N, H, B, and C cluster. H classification differed in accordance with the descriptor used. H was classified into alkali metals as per the CMP descriptor, O and N clusters as per the CRD descriptor, and halogen clusters as per the TPL descriptor.

For transition metals, in each descriptor, the elements were generally classified into the left and right halves of the periodic table into separate clusters. In each descriptor, the elements of groups 5 and 6 were classified into close clusters. The clustering results of iron-based elements from Cr to Ni and the noble metal elements differed in the three descriptors. In the CMP descriptor, Ru and Os were classified into the same cluster in addition to Mn to Ni. The noble metals were divided into Cu, Ag, Au and Rh, Pd, Ir, and Pt. In the CRD descriptor, Cr, Mn, and Fe formed one cluster. The noble metals formed a cluster of Ru, Rh, Os, and Ir; and a cluster of Co, Ni, Pd, Pt, and Au. Regarding the TPL descriptor, Mn to Ni elements formed one cluster, and the noble metals formed the following clusters: Ru, Rh, Os, and Ir; Pd, Pt, and Au; and Cu to Hg elements.

The result that CMP descriptor highly depended on valence electrons was same as original Atom2Vec and it is reasonable because features of valence electrons are one of the most easily obtainable information from composition formula. In CRD and TPL descriptors, information about valence electrons was not directly included. On the other hand, CRD and TPL descriptors were generated from information about the elements similarity based on behavior in crystal structures. CRD and TPL descriptors were able to capture the features which cannot be derived only from chemical composition data such as atomic size, and the effect of valence electrons were lesser than CMP descriptor.

3.2. Similarity evaluation of elements by cosine similarity between atomic descriptors

We measured the cosine similarity between the CMP, CRD, and TPL descriptors of each element. Tables 2–4 show the results of these calculations, where we show the similarity of the elements H through F (excluding He) to various elements.

There was high similarity between H and halogen elements as per all three descriptors. Regarding the CMP and TPL descriptors, the cosine

Table 2. Similarity of atoms as per the cosine similarity using CMP descriptor.

H	F	Cl	Br	I	Li	Au	Ag	Na
	0.433	0.271	0.251	0.188	0.179	0.177	0.156	0.134
Li	Na	Ag	K	Rb	Cu	Cs	Tl	Mg
	0.433	0.271	0.251	0.188	0.179	0.177	0.156	0.134
Be	Zn	Mg	Mn	Cd	Fe	Sc	Ti	Al
	0.250	0.174	0.152	0.123	0.101	0.099	0.099	0.098
B	Ga	Al	Si	P	Fe	Cr	Au	C
	0.170	0.164	0.147	0.117	0.113	0.108	0.105	0.103
C	N	Ge	Si	B	Se	Te	Ti	Os
	0.160	0.134	0.131	0.103	0.081	0.077	0.072	0.068
N	P	As	C	Sb	Bi	I	Rh	S
	0.217	0.177	0.160	0.138	0.127	0.110	0.104	0.094
O	S	Se	Te	N	Hg	Pt	As	Pd
	0.392	0.257	0.203	0.065	0.060	0.059	0.057	0.055
F	Cl	Br	I	H	Hg	Pd	Au	O
	0.509	0.351	0.249	0.233	0.074	0.058	0.057	0.051

Table 3. Similarity of atoms as per the cosine similarity using CRD descriptor.

H	Cl	Br	Li	F	I	Ba	Sr	Be
	0.099	0.099	0.079	0.078	0.078	0.076	0.062	0.062
Li	Mg	Na	Ag	Zn	Cd	Ca	Pd	In
	0.365	0.259	0.252	0.211	0.202	0.193	0.183	0.172
Be	Zn	Li	Mn	Al	Ga	Mg	Si	Sn
	0.171	0.166	0.164	0.157	0.146	0.138	0.126	0.111
B	P	C	As	Si	Be	Ge	Sb	Te
	0.137	0.124	0.107	0.107	0.098	0.074	0.060	0.050
C	N	B	P	As	Te	Rh	Be	Sb
	0.186	0.124	0.092	0.088	0.081	0.078	0.070	0.068
N	O	C	P	As	S	Se	Te	I
	0.196	0.186	0.163	0.143	0.124	0.106	0.104	0.088
O	S	F	Se	N	Cl	Br	I	Te
	0.318	0.235	0.228	0.196	0.177	0.161	0.139	0.116
F	Cl	Br	I	O	S	Se	H	N
	0.520	0.380	0.322	0.235	0.095	0.084	0.078	0.048

Table 4. Similarity of atoms as per the cosine similarity using TPL descriptor.

H	F	Br	Cl	I	Be	C	Au	O
	0.256	0.178	0.154	0.148	0.119	0.114	0.106	0.103
Li	Mg	Al	Zn	Mn	Cu	Ga	Co	V
	0.434	0.431	0.423	0.386	0.385	0.378	0.375	0.367
Be	Zn	F	B	Au	Ge	Li	Co	Al
	0.369	0.365	0.327	0.320	0.315	0.314	0.312	0.307
B	P	Ge	S	Si	As	Te	C	Se
	0.420	0.366	0.362	0.361	0.361	0.360	0.360	0.340
C	S	Se	B	N	Te	P	As	Cl
	0.384	0.377	0.360	0.357	0.332	0.300	0.299	0.298
N	O	S	Se	Cl	Br	I	F	Te
	0.539	0.501	0.489	0.443	0.428	0.416	0.389	0.389
O	N	S	Cl	Se	I	Br	Te	F
	0.539	0.539	0.534	0.507	0.492	0.440	0.404	0.390
F	Br	Cl	I	O	N	Be	S	Se
	0.534	0.514	0.473	0.390	0.389	0.365	0.316	0.315

similarities between H and F were >0.2. However, regarding the CRD descriptor, the cosine similarity between H and Cl – which showed the highest value – were <0.1 and the similarities were low compared with that obtained by using the other descriptors.

Li was similar to the elements that have one valence electron when CMP descriptor was used. Mg and Zn were in high rank when we used CRD descriptor, and elements that have one valence electron were in comparatively much lower rank when we used TPL descriptor. This may be because atomic properties other than the number of valence electrons, such as atomic radius, were highly pertinent.

Be exhibited high similarity to Zn in all three of the descriptors. When we used CMP descriptor, the similarity to the same-group element Mg was the second highest; and when we used CRD descriptor, the similarity to Li – which is the adjacent to Be in the periodic table – was the second highest. When we used TPL descriptor, another adjacent element, B, exhibited high similarity. There was high similarity between Be and F when using TPL descriptor; we require further research to explain this finding.

B exhibited high similarity to Al and Ga, which have the same number of valence electrons when using the CMP descriptor. However, these elements were not ranked in the top eight when using either the

CRD or TPL descriptors. Non-metallic or semi-metallic elements – such as P, C, Si, and As – exhibited high similarity to B.

C and N were most similar to one another when using CMP and CRD descriptors. The same-group elements Ge and Si exhibited the second- and third-highest similarity to C when using CMP descriptor. In contrast, Ge and Si were not similar to C when we used CRD and TPL descriptors, and B – which is adjacent to C in the periodic table – exhibited high similarity to C. When using TPL descriptor, S and Se were also similar to C.

When using CMP descriptor, N and O exhibited high similarity with same-group elements. When using CRD and TPL descriptors, elements that are adjacent to N and O in the periodic table exhibited correspondingly high similarity.

F exhibited higher similarity to halogen elements when using any of the three descriptors and those values were sometimes greater than 0.5.

Figure 5 shows our results for all element combinations in comparison with differences in electronegativity, atomic radius, and maximum oxidation state between each element.

The CMP descriptor exhibited the same trends as the CRD descriptor. When the cosine similarity between two atomic descriptors was high, the physical properties of each element were also similar to one another. When the cosine similarity was low, there were no clear relationships between the similarity of the descriptors and physical properties. The TPL descriptor was the exception and had a comparatively higher correlation with the atomic radius. The relationships between the TPL descriptor, electronegativity, and maximum oxidation state were less clear than the CMP and CRD descriptors and the aforementioned properties. These results may be because electronegativity and oxidation state are determined by relationships between atoms in compounds, but the TPL descriptor encompasses only information on the shape of coordination polyhedra, and does not include information about relationships between atoms.

3.3. Formation energy and band gap prediction using each descriptor

When using 10 components of descriptors, formation energy prediction accuracies were worse than those obtained with the Pymatgen physical property descriptor (23 elements). When using 20 components of descriptors, most of the prediction accuracies were almost the same as those obtained by using the physical property descriptor, and the CRD descriptor exhibited the highest accuracy (MAE = 0.077 eV/atom) for predicting the formation energy. In accordance with increasing number of components of each CMP, CRD, and TPL descriptor, the prediction accuracies increased and then converged at about

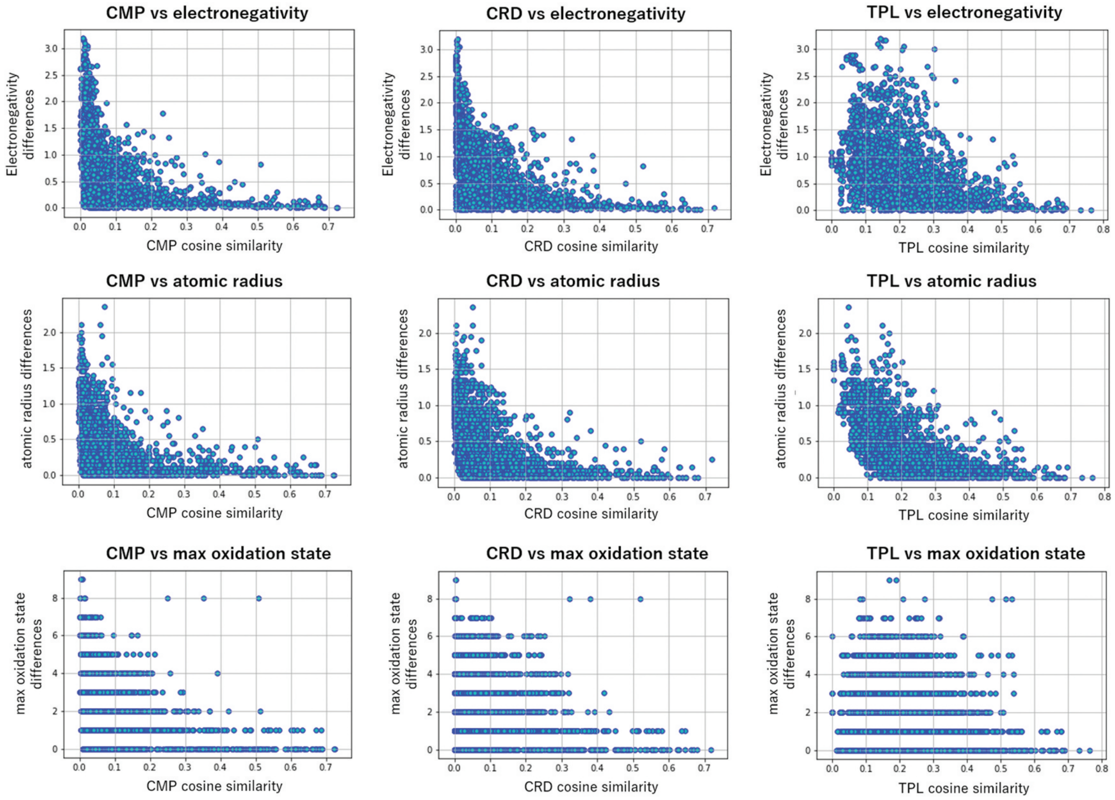


Figure 5. Relationships between the cosine similarity of each descriptor and the differences of electronegativity, atomic radius, and maximum oxidation state between two atoms.

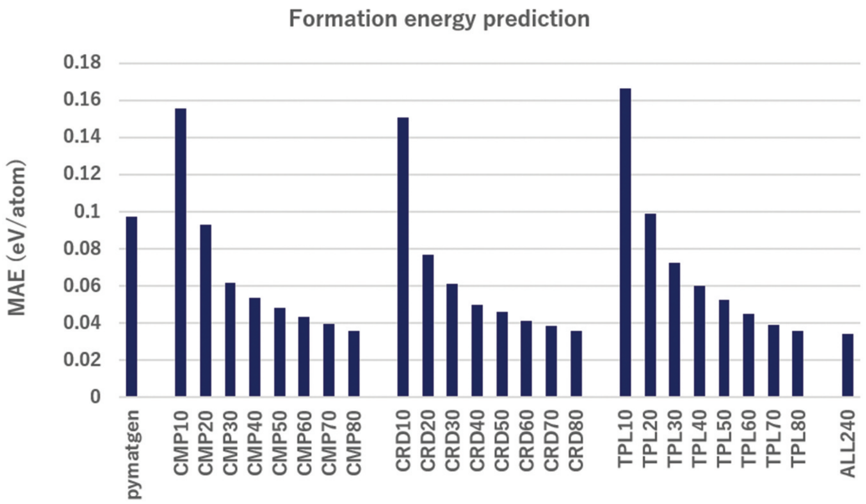


Figure 6. Mean absolute errors (MAE) of the formation energy prediction by using each descriptor. The numbers after each descriptor name are the number of descriptor components used for the input descriptors of machine learning.

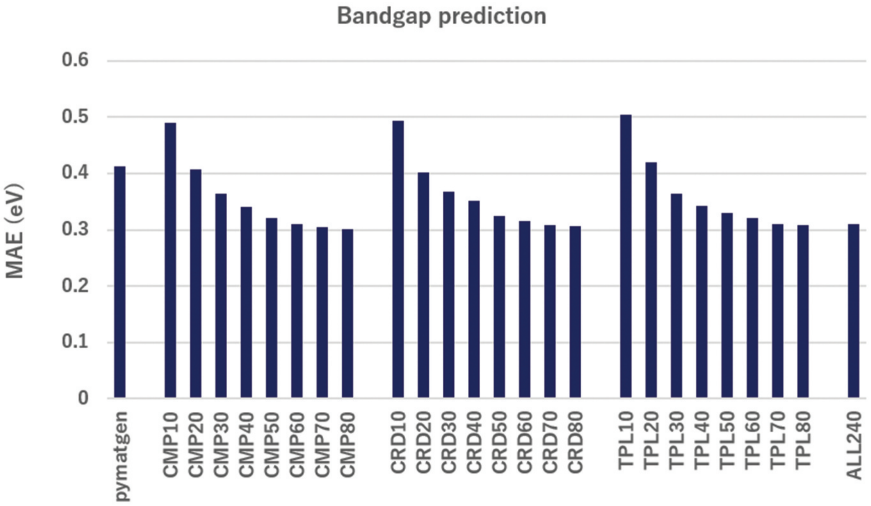


Figure 7. Mean absolute errors (MAE) of the band gap prediction by using each descriptor.

MAE = 0.035 eV/atom. When we used all of the CMP, CRD, and TPL descriptors simultaneously, the prediction accuracy did not substantially improve. When we did not use any atomic descriptors, MAE was about 0.048 eV/atom. Descriptors which have sufficient number of components provided superior prediction accuracies than prediction by using atomic physical properties and prediction without any descriptors.

There were no significant differences between CMP, CRD, and TPL descriptors in prediction accuracies of band gap. The band gap prediction accuracies converged at about MAE = 0.30 eV and superior than Pymatgen atomic physical property descriptor. Also, prediction without any descriptors could achieve MAE of 0.32 eV.

When using a lot of components of descriptors, MAE was converged both in formation energy and band gap prediction. Crystal structures are required for calculation of formation energy and band gap, but in this research, input descriptor is generated only from weighted average of atomic descriptors components based on chemical composition. This may limit the improvement in prediction accuracy. When using all descriptors, prediction accuracy was not significantly improved. This may be because components of three descriptors were not completely independent and representation ability of input descriptors was not improved.

Figures 8 and 9 show the average prediction accuracy of the formation energy and band gap when we choose all elements as unknown element X. Using too many or two few input components decreased the prediction accuracy. When the number of input components was insufficient, the expression ability of the prediction model decreased. When there were too many input components, the representation ability of

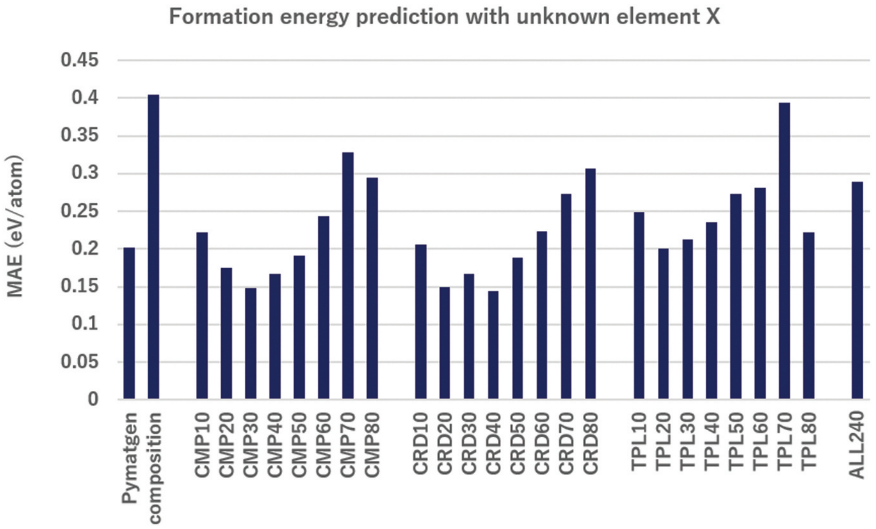


Figure 8. Average mean absolute errors (MAE) of the formation energy prediction on the data including unknown elements X by each descriptor.

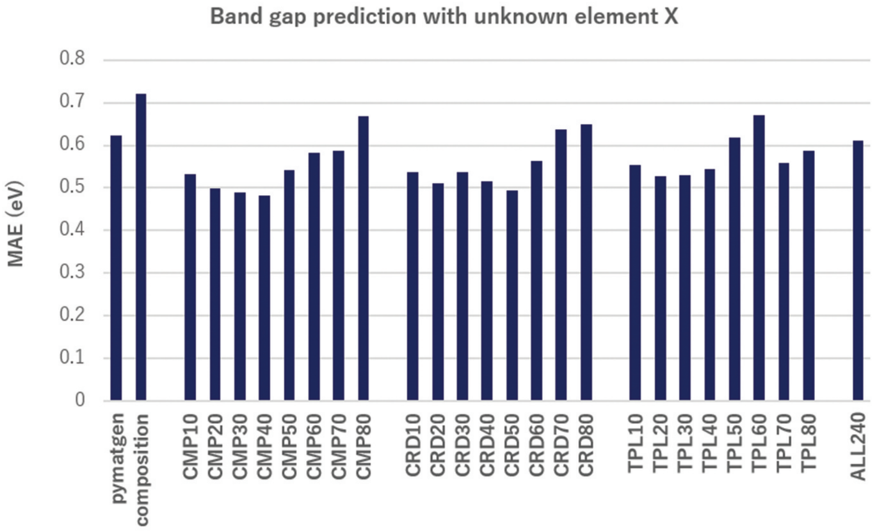


Figure 9. Average mean absolute errors (MAE) of the band gap prediction on the data including unknown elements X by each descriptor.

the prediction model became too high. Excessively high representation ability made the prediction model over-fit to training data and its generalization ability was lost. There were differences between the distribution of test data and that of training data and the effect of deterioration of generalization ability became obvious.

CRD descriptor consistently exhibited the highest prediction accuracy for the formation energy. When we used 40 components in the CRD descriptor, the prediction accuracy reached a maximum (MAE = 0.144 eV/atom). The CMP descriptor exhibited the second-highest accuracy. When we used 30 components in the CMP descriptor, the prediction accuracy (MAE = 0.148 eV/atom) was comparable with that obtained by using the CRD descriptor. The prediction accuracy using TPL descriptor were consistently poor, but when we used all 80 elements in the descriptors, the prediction accuracy was higher than that in corresponding 80-component experiments that used CMP and CRD descriptors.

The CMP descriptor exhibited the highest accuracy for predicting the band gap. We achieved the highest prediction accuracy by using 40 components for the CMP descriptor (MAE = 0.482 eV), and the CRD descriptor exhibited nearly the same MAE when we used 50 components (MAE = 0.494 eV). The appropriate descriptor changes depended on the prediction targets, even when we used data-driven descriptors.

Figure 10 shows the MAEs for predicting the formation energy when we selected each element as unknown element X. The number of input CMP, CRD, and TPL descriptors was 40. When metal elements were unknown element X, high prediction accuracies (MAE was about 0.10 eV/atom) are achieved. We correctly predicted the formation energy

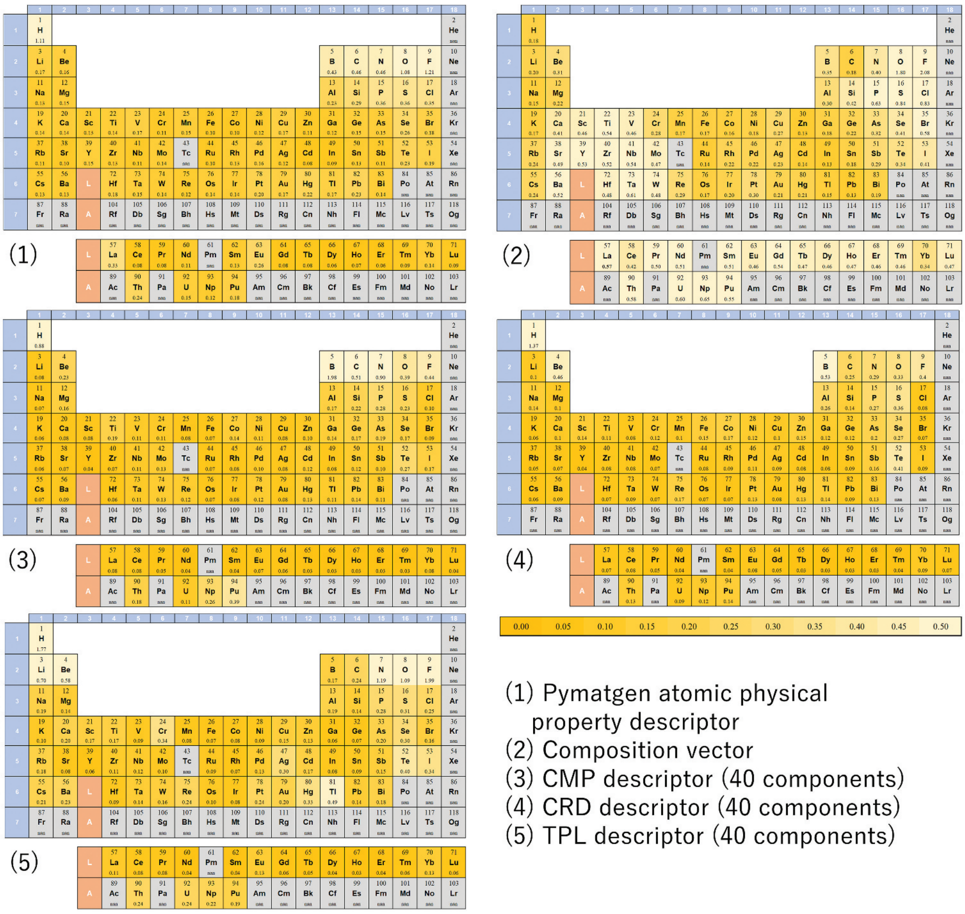


Figure 10. Prediction accuracies for the formation energy when we selected each atom as unknown atom X. Input was as follows: (1) Pymatgen atomic physical property descriptor, (2) composition vector, (3) CMP descriptor, (4) CRD descriptor, and (5) TPL descriptor, and for (3)-(5), we used 40 components of descriptors.

when we selected metal elements as the unknown element X. Most of metal elements, especially transition metal elements, were similar to other metal elements. In contrast, the accuracy of the predicted formation energy for compounds that included light and non-metal elements was insufficient.

When using vectors that simply represent the composition, most of the results were poor because there was no information on the similarity between elements. Thus atomic descriptors must be used when using machine learning on a dataset that has insufficient information on some elements. Pymatgen atomic physical property descriptor afforded a sufficient MAE on metal elements and achieved higher accuracy than the CRD and TPL descriptors on some elements, but predictions for non-metal elements were much worse than those for metal elements. The CMP descriptor consistently afforded high prediction accuracy, but the accuracies of some elements – such as H, B, C, N, O, and F – were poor. This means that the important features of these elements for formation energy prediction were not completely captured only from chemical composition data. We improved some of these prediction accuracies by using CRD and TPL descriptors. The CRD descriptor provided the best prediction accuracy on N and O. The TPL descriptor afforded insufficient scores for metals, but the prediction accuracy for some elements – such as B – was higher than that obtained for the other two descriptors.

Figure 11 shows the prediction accuracy for the band gap when we selected each element as unknown element X. The distribution of the prediction accuracy was larger than that obtained by using the formation energy prediction. In contrast to the formation energy prediction, the MAE for some metal elements was poor, such as Ca by using the CMP descriptor and Cr by using the CRD descriptor. Prediction accuracies for noble metals were higher than that obtained for other elements. This is because most compounds that include these elements have a small or no band gap, and the MAEs for these compounds were comparatively small.

The best descriptor which provided the highest prediction accuracies depended on target properties and focused elements. Machine learning is used when the relationship between input descriptor and target values is unclear. In such situation, diverse options of descriptors which represent the features of elements from different viewpoints are desirable, and our descriptors can be the powerful options for this purpose.

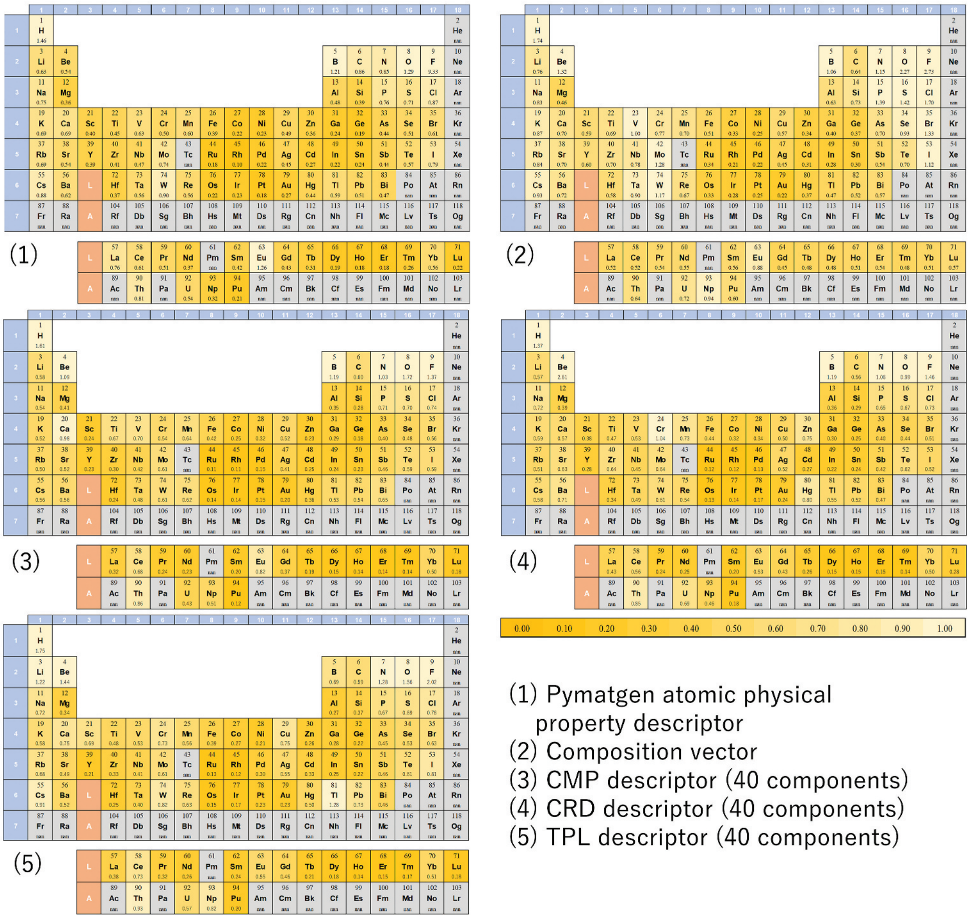


Figure 11. Prediction accuracies for the band gap when we selected each atom as unknown atom X. Input was as follows: (1) Pymatgen atomic physical property descriptor, (2) composition vector, (3) CMP descriptor, (4) CRD descriptor, and (5) TPL descriptor, and for (3)-(5), we used 40 components of descriptors.

4. Conclusion

Using data obtained by analyses of the local coordination structure in crystal structure data, we developed two atomic descriptors and compared with descriptors generated from composition data proposed in previous research. Each descriptor represented the similarity of elements in different contexts. Nearly all of our results were in accordance with conventional knowledge of inorganic chemistry. The similarity between elements can be easily calculated by using our descriptors. Inferring representations of elements from materials data and making similarity comparisons can accelerate discovery of new materials by assisting our expectation of elements replacement probability. In addition, formation energy and band gap of stable compounds were successfully predicted with higher accuracy using the descriptors reported in our research, compared with those using atomic physical property descriptors. The atomic descriptors reported in our research will be useful in machine learning toward designing high-performance materials. Suitable descriptor differs depending on target property and focused elements, and it is important for researchers that there are various options when materials data is converted to input descriptors of machine learning. Our new descriptors are based on the behavior of elements in crystal structures and will be powerful choices.

Acknowledgements

We thank Michael Scott Long, PhD, from Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

Disclosure statement

The descriptors data generated in this study are available within the supplementary materials of the article.

Funding

This work was supported by Japan Science and Technology Agency (JST) CREST grant number [JPMJCR19J1], Japan.

ORCID

Yuki Inada http://orcid.org/0000-0002-9104-1320

References

[1] Jain A, Ong SP, Hautier G, et al. The Materials Project: a materials genome approach to accelerating materials innovation. APL Mater. 2013;1(1):011002.

[2] Meredig B, Agrawal A, Kirklin S, et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. Phys Rev B. 2014;89:094104.

[3] Ghiringhelli LM, Vybiral J, Levchenko SV, et al. Big data of materials science: critical role of the descriptor. Phys Rev Lett. 2015;114:105503.

[4] Faber FA, Lindmaa A, Lilienfeld A, et al. Machine learning energies of 2 million elpasolite (ABC2D6) crystals. Phys Rev Lett. 2016;117:135502.

[5] Ye W, Chen C, Wang Z, et al. Deep neural networks for accurate predictions of crystal stability. Nat Commun. 2018;9:3800.

[6] Jha D, Ward L, Paul A, et al. ElemNet: deep learning the chemistry of materials from only elemental composition. Sci Rep. 2018;8:17593.

[7] Zhang Z, Li M, Flores K, et al. Machine learning formation enthalpies of intermetallics. J Appl Phys. 2020;128:105103.

[8] Dey P, Bible J, Datta S, et al. Informatics-aided bandgap engineering for solar materials. Comput Mater Sci. 2014;83:185–195.

[9] Stanev V, Oses C, Kusne AG, et al. Machine learning modeling of superconducting critical temperature. Npj Comput Mater. 2018;4:29.

[10] Li S, Dan Y, Li X, et al. Critical temperature prediction of superconductors based on atomic vectors and deep learning. Symmetry. 2020;12:262.

[11] Zhao XP, Huang HY, Wen C, et al. Accelerating the development of multi-component Cu-Al-based shape memory alloys with high elastocaloric property by machine learning. Comput Mater Sci. 2020;176:109521.

[12] Gaultois MW, Oliynyk AO, Mar A, et al. Perspective: web-based machine learning models for real-time screening of thermoelectric materials properties. APL Mater. 2016;4:053213.

[13] Oliynyk AO, Adutwum LA, Rudyk BW, et al. Disentangling structural confusion through machine learning : structure prediction and polymorphism of equiatomic ternary phases ABC. J Am Chem Soc. 2017;139(49):17870–17881.

[14] Ryan K, Lengyel J, Shatruk M. Crystal structure prediction via deep learning. J Am Chem Soc. 2018;140 (32):10158–10168.

[15] Zhao Y, Cui Y, Xiong Z, et al. Machine learning-based prediction of crystal systems and space groups from inorganic materials compositions. ACS Omega. 2020;5:3596–3606.

[16] Hautier G, Fischer CC, Jain A, et al. Finding nature's missing ternary oxide compound. Chem Mater. 2010;22:3762–3767.

[17] Li Z, Achenie LEK, Xin H. An adaptive machine learning strategy for accelerating discovery of perovskite electrocatalysts. ACS Catal. 2020;10: 4377–4384.

[18] Pettifor DG. A chemical scale for crystal-structure maps. Solid State Commun. 1984;51:31–34.

[19] Ong SP, Richards WD, Jain A, et al. Python materials genomics (pymatgen): a Robust, open-source Python library for materials analysis. Comput Mater Sci. 2013;68:314–319.

[20] Yamada H, Liu C, Wu S, et al. Predicting materials properties with little data using shotgun transfer learning. ACS Cent Sci. 2019;5(10):1717–1730.

[21] Ward L, Agrawal A, Choudhary A, et al. A general-purpose machine learning framework for predicting properties of inorganic materials. NPJ Comput Mater. 2016;2:16028.

[22] Ward L, Dunn A, Faghaninia A, et al. Matminer: an open source toolkit for materials data mining. Comput Mater Sci. 2018;152:60–69.

[23] Broderick SR, Rajan K. Designing a periodic table for alloy design: harnessing machine learning to navigate a multiscale information space. JOM. 2020;72(12):1–10.

[24] Willatt MJ, Musil F, Ceriotti M. A data-driven construction of the periodic table of the elements. arXiv:1807.00236v1 [physics.chem-ph].

[25] Kusaba M, Liu C, Koyama Y, et al. Recreation of the periodic table with an unsupervised machine learning algorithm. Sci Rep. 2021;11:4780.

[26] Zhou Q, Tang P, Liu S, et al. Learning atoms for materials discovery. PNAS. 2018;115(28):E6411–E6417.

[27] Villars P, Hulliger F. Structural-stability domains for single-coordination intermetallic phases. J Less Common Met. 1987;132:289–315.

[28] Blatov VA, Shevchenko AP, Proserpio DM. Applied topological analysis of crystal structures with the program package ToposPro. Cryst Growth Des. 2014;14:3576–3586.

[29] Blatov VA, O'Keeffe M, Proserpio DM. Vertex-, face-, point-, Schlafli-, and Delaney-symbols in nets, polyhedra and tilings: recommended terminology. CrystEngComm. 2010;12:44–48.

[30] Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–272.

[31] Ward JR. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963;58:236–244.

[32] Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous systems. 2015.

[33] Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv:1412.6980 [cs.LG].

[34] Zhao Y, Yuan K, Liu Y, et al. Predicting elastic properties of materials from electronic charge density using 3D deep convolutional neural networks. J Phts Chem C. 2020;124(31):17262–17273.