Learning structure-property relationship in crystalline materials: A study of lanthanide–transition metal alloys

Tien-Lam Pham[1,2,3], Nguyen-Duong Nguyen[1,5], Van-Doan Nguyen[1,2], Hiori Kino[2,4], Takashi Miyake[2,3,4], and Hieu-Chi Dam[1,2,5]
[1] *Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan*
[2] *Center for Materials Research by Information Integration, National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan*
[3] *ESICMM, National Institute for Materials Science 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan*
[4] *CD-FMat, AIST, 1-1-1 Umezono, Tsukuba 305-8568, Japan*
[5] *JST, PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan*

(Dated: July 19, 2018)

We have developed a descriptor named Orbital Field Matrix (OFM) for material structures in datasets of multi-element materials. The descriptor is based on the information regarding atomic valence shell electrons and their coordination. In this work, we develop an extension of OFM called OFM1. We have shown that these descriptors are highly applicable in predicting the physical properties of materials and in providing insights on the materials space by mapping into a low embedded dimensional space. Our experiments with transition metal/lanthanide metal alloys show that the local magnetic moments and formation energies can be accurately reproduced using simple nearest-neighbor regression, thus confirming the relevance of our descriptors. Using kernel ridge regressions, we could accurately reproduce formation energies and local magnetic moments calculated based on first-principles, with mean absolute errors of 0.03 μ_B and 0.10 eV/atom, respectively. We show that meaningful low-dimensional representations can be extracted from the original descriptor using descriptive learning algorithms. Intuitive prehension on the materials space, qualitative evaluation on the similarities in local structures or crystalline materials, and inference in the designing of new materials by element substitution can be performed effectively based on these low-dimensional representations.

INTRODUCTION

Human beings have always paid significant attention to learning nature's "game" by observation and imagination of natural phenomena. In this respect, we have observed the vast diversity of nature and unified different natural phenomena in a small set of fundamental variables or laws. This consideration of science is strongly related to the field of data-mining, which is developed to discover hidden knowledge. Recently, the increasing volume of available experimental and quantum-computational material databases, together with the development of machine-learning techniques, has provided new opportunities for developing techniques that help researchers accelerate the discovery and comprehension of new materials and phenomena. Machine-learning algorithms can be used to automatically extract knowledge regarding materials, including their patterns and chemical and physical rules, using both first-principles-calculated data and experimental data [1–8].

It has been pointed out that using machine learning algorithms to extract knowledge from data requires appropriate data representation, appropriate knowledge representation, appropriate optimization algorithm, and appropriate evaluation criteria [9]. In a conventional materials dataset, a material is described by a set of atoms with their coordinates and periodic unit-cell vectors, which are required for crystalline systems. From the viewpoint of data science, materials data in primitive representation is categorized as unstructured data, in which mathematical reasoning follows the algebra of sets. Therefore, advanced quantitative machine-learning algorithms can hardly be applied directly to conventional materials data due to limitations of the algebra of the primitive data representation.

In order to apply well-established machine-learning methods including predictive learning and descriptive learning, an appropriate transformation from primitive representation to a structured representation, such as vectors or matrices, is required, such that comparisons and calculations using the new representation reflect the nature of the materials and the underlying mechanisms of chemical/physical phenomena. Various methods for encoding materials have been developed in the field of materials informatics. Behler et al. [10–16] utilized atom-distribution-based symmetry functions to represent the local chemical environment of atoms, and employed a multilayer perceptron to map this representation to

atomic energy. The arrangement of structural fragments has also been used to encode materials in order to predict the physical properties of molecular and crystalline systems [5, 17]. Isayev used the band structure and density of states (DOS) fingerprint vectors as a representation of materials to visualize material space [5]. Rupps et al. developed a representation known as the Coulomb matrix (CM) to predict atomization energies and formation energies [18–20].

Recently, we have proposed a novel descriptor named the orbital field matrix (OFM), which incorporates the valence atomic configuration to describe the local and the entire structure of materials. This descriptor was based on the consideration that certain essential aspects of the electronic structures can be deduced from a simple description of the valence electrons surrounding a central atom [21]. Since we focus on the representation of a local structure including the central atom and those surrounding it, the information on the central atoms should play an essential role in describing the characteristic of the local structure. However, our previous descriptor (OFM) does not explicitly contain the information of the center atom. In this study, we extended OFM to the new (OFM1), which explicitly includes the information on the central atom in each local structure as seen in Eq. 3. We demonstrate that these descriptors are highly applicable in predicting physical/chemical properties of materials, and in providing insights on materials space by mapping into a low-embedded dimensional space. Our experiments with transition metal/lanthanide metal alloys show that the local magnetic moments and formation energies can be accurately reproduced using simple nearest neighbors and kernel ridge regressions (KRR) based on our descriptors. Using KRR, the local magnetic moments and formation energies of the materials obtained by first-principles calculations could be predicted with mean absolute errors (MAE) of 0.03 μ_B and 0.10 eV/atom, respectively.

In materials science studies, along with the prediction of properties, the detection of the pattern of behaviors of materials is also an important task. Herein we demonstrate that OFM and OFM1 are also applicable for an unsupervised learning to extract the pattern of behaviors of an atom in a local environment (local look) and the materials (global look). We show that the new and meaningful low-dimensional representations can be extracted from the original descriptor using descriptive learning algorithms. Manifold learning techniques can be applied to the initial representation using OFM descriptors to discover hidden embedding features in the transition metal/lanthanide metal alloys data set. The dataset is then mapped to the embedding features into low-dimensional space. Intuitive prehension on the local structure space can be easily acquired based on low-dimensional representations. Qualitative evaluation of the similarity in local structures can be inferred directly from the Euclidean distance in the extracted low-dimensional space. Groups of local structures with similar symmetries and shapes can be easily identified in the form of trajectories in the low-dimensional representations. The extracted low-dimensional space of crystal structure of the transition metal/lanthanide metal alloys dataset shows an apparent separation between the two groups of materials having high and low formation energies. The obtained results demonstrate that one may obtain prehension on local structures and crystal structures, and be able to infer their properties from their proximities in the extracted low-dimensional spaces.

REPRESENTATION OF MATERIALS

Encoding atom and local structure

From fundamental chemistry and physics, we learned that the number of atomic orbitals surrounding a central atomic orbital plays a significant role in determining many material properties such as magnetic properties. To embed this knowledge in material representation, we started with the representation of an atom with a one-hot row vector $\vec{O}_{atom}$ using a dictionary comprising the valence subshell orbitals: $D = \{s^1, s^2, p^1, p^2, ..., p^6, d^1, d^2, ..., d^{10}, f^1, f^2, ..., f^{14}\}$ (e.g., d^5 indicates the electron configuration in which the atomic valence d orbital holds 5 electrons). Based on this atom representation, we designed a matrix whose element, X_{ij}, represents the number of an atomic orbital, orbital j, coordinated with a central atomic orbital, orbital i, to encode a local structure including a central atom and the neighboring atoms. Here, i and j are in the dictionary comprising the valence subshell orbitals D. To build this matrix, we utilized the one-hot row vector representation for the central atom, $\vec{O}_{central}$, and the neighboring atoms, $\vec{O}_k$, where k is the index of the neighboring atoms. We then summed the vector of the coordinating atoms to form a vector that represents the environment surrounding a central atom, $\vec{O}_{env}$:

$$\vec{O}_{env} = \sum_{k}^{K} w_k \vec{O}_k \tag{1}$$

where the weight, w_k, measures the contribution of the k^{th} neighboring atom, and K is the number of the neighboring atoms. The representation matrix of a local structure now becomes:

$$X_{local} = \vec{O}^T_{central} \times \vec{O}_{env} \tag{2}$$

where $\vec{O}^T_{central}$, a column vector, is the transpose of $\vec{O}_{central}$.

In this study, we adopted the definition of neighboring atoms by O'Keeffe [22], which utilizes the solid angles, θ_k,

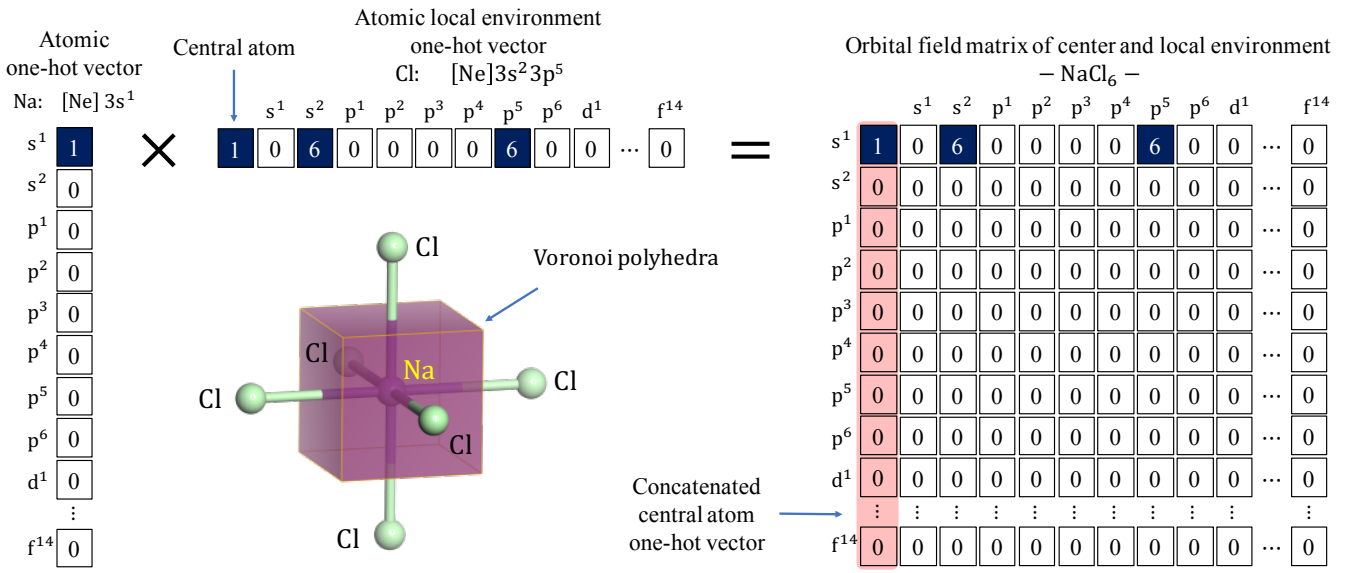

FIG. 1. Orbital field matrix representation for Na atom in an octahedral site surrounded by six Cl atoms: atomic one-hot vector for Na (left), representation of 6 Cl atoms surrounding the Na atom (middle), and representation of Na atom surrounded by 6 Cl atoms (right). The vector of the central atom is concatenated in the first column on the left-hand side [21].

determined by the faces of the Voronoi polyhedra. This method can give the same coordination numbers for the high-symmetry atomic environment, and evaluate the coordination numbers for the lower-symmetry atomic environment automatically and with no ambiguity. In this method, the weight, w_k, is determined by the solid angles: $w_k = \frac{\theta_k}{\theta_{max}}$, where θ_k is the solid angle determined by the face of the Voronoi polyhedra between the central atom and the neighboring atom, k, and θ_{max} is the maximum solid angle among those between the central atom and the neighbor atoms. Additionally, to incorporate information on valence orbital sizes, the distance r_k between the central atom and the k^{th} atom should be included in w_k.

Although the matrix X_{local} in Eq. 2 also includes information on the central atom, but it is not explicitly exploited in the similarity measure based on vector or matrix calculations (discussed in detail in the next section) [21]. To explicitly incorporate the information on the central atom, we simply concatenated $\vec{O}^T_{central}$ to matrix X_{local} as a new column. Finally, we propose the following form for representing a local structure:

$$X'_{local} = \vec{O}^T_{central} \times \left(1.0, \sum_k \vec{O}_k \frac{\theta_k}{\theta_{max}} \zeta(r_k)\right) \quad (3)$$

where $\zeta(r_k)$ is a function representing the contribution of r_k to w_k. In this study, we use the inverse of the distance as the distance-dependent weight function: $\zeta(r_k) = 1/r_k$. Using this formula, the central vector is concatenated to X_{local} in the first left column (Fig. 1); this descriptor is named OFM1.

Encoding molecular and crystal structures

Composing the descriptor for a structure (a molecule or crystal system) from its local structure representation requires careful consideration. From a data science viewpoint, the composed descriptors should include as much information as possible. However, from a materials science viewpoint, the descriptors should be composed such that they appropriately reflect the nature of the target physical properties. In this study, for the formation energy (per atom) of a crystal (which can be considered an accumulative quantity of the contribution of constituent local structures), we obtained the mean over the local structure descriptors as the descriptor for the entire structure:

$$F = \frac{1}{N_p} \sum_p^{N_p} X'_p, \quad (4)$$

where p and X'_p are the indices and representations of local structures surrounding atoms in a structure, respectively; N_p is the number of atoms the unit cell; and F is the OFM1 representing the entire crystalline material.

PREDICTIVE ANALYSES

Prediction of local magnetic moment

Here, we examine the use of OFM1 to predict the local atomic properties of materials. We implement this

method using the Python Materials Genomics (pymatgen) code [23]. We focused on the local magnetic moments of transition metals in LAT alloys (in ferromagnetic configuration), the dataset of which includes 658 structures collected from the Materials Project database [24, 25]. We selected the structures by combining transition metals and lanthanides from sets of {Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, Hf, Ta, W, Re, Os, Ir, Pt, Au} and {La, Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu}. Details of the data preparation can be found in [21].

Since the local magnetic moment of a transition-metal site is mainly determined by the number of unpaired electrons in the d-orbitals, our description of local structure encoding information on valence electron coordination should include a large amount of information on local magnetic moment. We examine whether the local magnetic moment information is significantly included in the descriptors by considering that materials with higher similarity, as estimated by the descriptors, should possess similar local magnetic moments. To this end, we test whether the local magnetic moments can be predicted by using a nearest-neighbor regression. The cross-validated root mean square error (RMSE), the mean absolute error (MAE), and the coefficient of determination R^2, were used to measure the performance of our descriptors.

The principle behind the nearest-neighbor method is to find a predefined number of training samples closest in distance to the new point, and predict the label from these samples. The number of samples can be a user-defined constant (k-nearest-neighbor learning), or can vary based on the local density of points (radius-based neighbor learning) [5]. The accuracy of nearest-neighbor regression therefore directly reflects the performance of data representation and similarity measurement. In nearest-neighbor regression, data properties are deduced from the properties of nearest neighbors in the training data. In this study, we employed a nearest-neighbor regressor implemented in the scikit-learn package [26]. The number of nearest neighbors was fixed at 5, and the nearest neighbors were determined by a brute-force search. The prediction was weighted by the distance to the nearest neighbors. We examined the following distance measurements for localizing the nearest-neighbor data:

Euclidean distance $d_{eucl}(X,Y) = \sqrt{\sum_{i,j}(X_{ij}-Y_{ij})^2}$,

Manhattan distance $d_{man}(X,Y) = \sum_{i,j}|X_{ij}-Y_{ij}|$,

Cosine distance $d_{cos}(X,Y) = 1 - \frac{\sum_{ij}X_{ij}Y_{ij}}{\sqrt{\sum_{i,j}X_{ij}^2}\sqrt{\sum_{i,j}Y_{ij}^2}}$,

Bray-Curtis distance $d_{bar}(X,Y) = \frac{\sum_{ij}|X_{ij}-Y_{ij}|}{\sum_{ij}|X_{ij}|+|Y_{ij}|}$,

Canberra distance $d_{can}(X,Y) = \sum_{ij}\frac{|X_{ij}-Y_{ij}|}{|X_{ij}+Y_{ij}|}$,

and Correlation distance $d_{cor}(X,Y) = \frac{Cov(X,Y)}{\sqrt{\sigma_X \cdot \sigma_Y}}$. Here, X and Y are two vectors representing two data points; $Cov(X,Y)$ is the covariance of X and Y; and σ_X and σ_Y are the variances of X and Y, respectively.

We also implemented a CM descriptor for a local structure, and used it to predict the local magnetic moments for comparison. The local structure determined by the Voronoi polyhedra was considered a molecule, and the CM descriptor of this local structure was calculated following the Rupp scheme [18]. We first examine the dependence of MAE for the test set on the number of training data. Fig. 2 shows the learning curves of the nearest regression by CM [18], OFM, and OFM1. It is clearly seen that the OFM and OFM1 yield the more accurate prediction than that given by CM, and OFM1 shows a slight improvement over OFM. Table I shows the results of the nearest-neighbor regression obtained by CM, OFM, and OFM1 by 10-times 10-fold cross-validation.

As mentioned above, matrix X in Eq. 2 also includes information on the central atom in the local structure. This information can be obtained manually by checking the indices of the non-zero column of the OFM representation. It can also be extracted automatically using similarity measures based on a comparison of direction differences between the vectors. However, information regarding valence-orbital coordination (encoded in the OFM) includes the coordinations of each type of valence subshell orbital. Therefore, similarity measures that consider comparisons of the magnitude of dimensions between vectors are preferred. However, both central atom information and valence-orbital coordination information are indispensable for learning the local structure. Consequently, as seen in Table I, the Manhattan, Bary-Curtis, and Canberra distances, which include the differences in both direction and magnitude between vectors, show an overwhelmingly superior prediction accuracy ($R^2 > 0.86$) than that of the Euclidean distance for the OFM ($R^2 = 0.53$) (which does not appropriately measure the difference in direction between vectors).

To explicitly incorporate central atom information, we simply concatenated $\vec{O}_c^T$ to matrix X as a new column. OFM1 with the Manhattan distance provided a remarkable improvement over OFM, while the OFM yielded a significantly better performance than CM. This result implies that our OFM1 can explicitly embed substantially more information about the local structure compared to both OFM and CM. Further, we achieved a reasonably good performance in the prediction of local magnetic moments using OFM1 with the Manhattan distance, with the best RMSE of approximately $0.151\mu_B$, MAE of $0.036\ \mu_B$, and R^2 of 0.948. This result indicates that closer materials in our description space of local structures yield similar local magnetic moments, which implies that our data representation includes significant information about local magnetic moments.

For a better representation of local magnetic moment, we applied a KRR [27] model to predict the local magnetic moment. KRR is a combination of the kernel method and ridge regression, and has recently proved

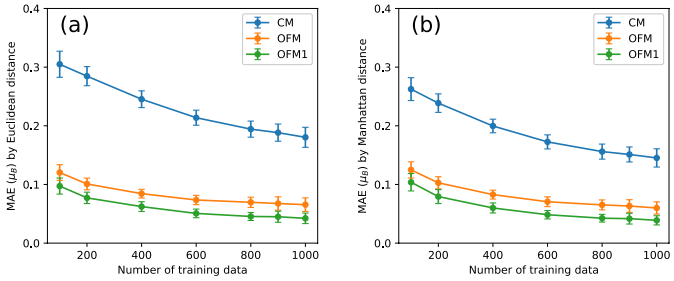

FIG. 2. The learning curve of the nearest-neighbor regression for the prediction of the local magnetic moment: the dependence of MAE on the number of fold in the cross-validation by OFM1 (green), OFM (yellow), and CM (blue) with Euclidean (a) and Bay-Curtis (b) distances. The error bar represent the margin error at the confidence level fo 95%.

TABLE I. Cross-validation RMSE (μ_B), cross-validation MAE (μ_B), and coefficient of determination R^2 values for prediction of local magnetic moments obtained by nearest-neighbor regression with different distance measurements.

Distance		d_{eucl}	d_{man}	d_{cos}	d_{bar}	d_{can}	d_{cor}
CM [18]	RMSE	0.405	0.354	0.483	0.352	0.245	0.483
	MAE	0.168	0.135	0.205	0.132	0.071	0.204
	R^2	0.639	0.724	0.486	0.727	0.868	0.487
OFM	RMSE	0.263	0.239	0.237	0.237	0.238	0.256
	MAE	0.062	0.057	0.070	0.058	0.071	0.069
	R^2	0.53	0.878	0.860	0.880	0.880	0.860
OFM1	RMSE	0.202	0.151	0.171	0.163	0.160	0.171
	MAE	0.042	0.036	0.039	0.037	0.045	0.039
	R^2	0.906	0.948	0.933	0.939	0.941	0.934

successful in materials and chemical science applications. In the KRR algorithm, the property of a system can be given by the weighted kernel function:

$$y = f(x, c) = \sum_{k \in D_{ref}} c_k K(x, x_k), \tag{5}$$

where k runs over all reference data (D_{ref}). We used a Laplacian function, $K(x, x_k) = e^{-\gamma d(x, x_k)}$, where $d(x, x_k)$ is the Euclidian distance between x and x_k. In order to minimize the prediction risk, the coefficients c_k were determined by minimizing the total square error regularized by L2 norm (ridge regression):

$$\arg\min_{c} \left(\sum_i [f(x_i) - y_i]^2 + \lambda \sum_k ||c_k||_2^2 \right). \tag{6}$$

We used stratified ten-fold cross validation for model selection and performance estimation. Parameters γ and λ were determined in an inner loop of the ten-fold cross validation by using a logarithmic scaling grid. This procedure is routinely applied in machine learning and statistics to avoid overfitting and overly optimistic error estimation.

TABLE II. Cross-validated RMSE (μ_B), MAE (μ_B), and coefficient of determination R^2 values for prediction of local magnetic moments obtained by KRR with CM, OFM, OFM1.

Descriptor	CM [18]	OFM	OFM1
RMSE	0.21	0.18	0.12
MAE	0.11	0.05	0.03
R^2	0.90	0.93	0.97

The prediction results of local magnetic moments are summarized in Table II. The OFM and OFM1 also show advantages compared to CM with KRR regression. We obtained RMSE, MAE, and R^2 values of approximately 0.12 μ_B, 0.03 μ_B, and 0.97, respectively. These results confirm that OFM and OFM1 can be useful for predicting the local magnetic moment in LAT alloys.

Prediction of formation energies

Next, we applied our descriptors, OFM and OFM1, to predict the formation energy of LATX alloy systems. First, we examined how our description of materials can represent the formation energies of these systems by using nearest-neighbor regression to predict the formation energies of materials. We focused on transition metal binary alloys (TT) and bimetal alloys of lanthanide metals and transition metals (LAT), as well as LATX and TTX, which are LAT and TT alloys including a light element X. We selected transition metals from {Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, Hf, Ta, W, Re, Os, Ir, Pt, Au}, lanthanides from {La, Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu}, and X elements from {B, C, N, O}. We collected the data of more than 4,000 alloys, including their structures and formation energies, from the Materials Project repository: 1510 LATX alloys, 1311 TTX alloys, 692 LAT alloys, and 707 TT alloys. For brevity, this dataset is referred to as LATX. It notes that although there is the error of PBE calculations on the formation energy for the f-metals, the magnitude of the MAE is comparable to the overall MAE [28]. Herein, we aim to obtain the materials descriptors (and their applications) that accurately reproduce a DFT description of systems. Thus, comprehensive studies to compare theoretical and experimental results are beyond the scope of this study.

The distance measurement was similarly selected for predicting local magnetic moment in order to determine the nearest-neighbor materials. The energy of a material is determined by its five nearest-neighbor materials in the training data weighted by their distances. The nearest-neighbor materials were determined by a brute-force search. We first investigate the dependence of MAE on the size of the train set. Fig. 3 depict the learning curves of the nearest neighbor regression by CM [19],

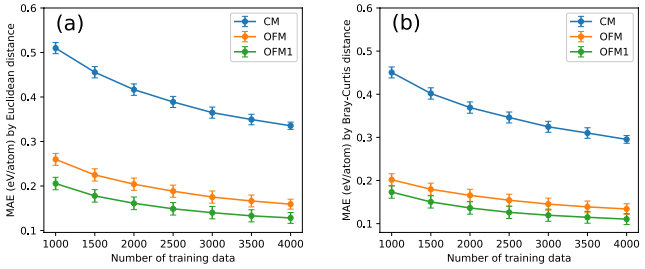

FIG. 3. The learning curve of the nearest-neighbor regression for the formation energy of LATX dataset: the dependence of MAE on the number of training data by OFM1 (green), OFM (yellow), and CM (blue) with Euclidean (a) and Bay-Curtis (b) distances. The error bar represent the margin error at the confidence level of 95%.

TABLE III. Cross-validated RMSE (eV/atom), MAE (eV/atom), and coefficient of determination R^2 values for predicting the formation energies of LATX obtained by nearest-neighbor regression with different distance measurements.

Distance		d_{eucl}	d_{man}	d_{cos}	d_{bar}	d_{can}	d_{cor}
CM [19]	RMSE	0.593	0.534	0.577	0.519	0.632	0.577
	MAE	0.340	0.309	0.334	0.302	0.365	0.334
	R^2	0.805	0.842	0.815	0.85	0.778	0.815
OFM	RMSE	0.301	0.251	0.280	0.245	0.259	0.280
	MAE	0.161	0.137	0.144	0.134	0.148	0.144
	R^2	0.950	0.965	0.956	0.966	0.963	0.956
OFM1	RMSE	0.245	0.217	0.231	0.211	0.244	0.231
	MAE	0.128	0.114	0.118	0.109	0.136	0.118
	R^2	0.967	0.974	0.970	0.975	0.967	0.970

OFM, and OFM1. It is clearly seen that OFM and OFM1 both have an advantage of over CM, whereas OFM1 gives a slight improvement over OFM. Table III shows the cross-validated RMSE, MAE, and R^2 values for predicting the formation energies obtained by 10-times 10-fold cross-validation. It was also observed that the results of nearest-neighbor regression effectively predicted the formation energies. We obtained the best cross-validated RMSE value of approximately 0.211 eV/atom, and R^2 above 0.975 by using the Bary-Curtis distance. This result was substantially better than that given by KRR with the CM descriptor, which resulted in a cross-validated RMSE of 0.47 eV/ atom and R^2 of 0.87 obtained by KRR. The CM is implemented following the work of Faber and coworkers [19]; this implies that our materials description also includes a significant amount of information on the formation energies of these materials.

We also applied KRR to represent the formation energy of LATX alloys. The cross-validated RMSE, MAE, and R^2 values with OFM were approximately 0.190 eV/atom, 0.112 eV/atom, and 0.98, respectively, while those obtained by OFM1 were 0.18 eV/atom, 0.098 eV/atom, and 0.99, respectively (Table IV).

TABLE IV. Cross-validated RMSE (eV/atom), MAE (eV/atom), and coefficient of determination R^2 values for formation energies of LATX obtained by KRR using CM, OFM, and OFM1 descriptors.

Descriptor	CM [19]	OFM	OFM1
RMSE	0.470	0.190	0.180
MAE	0.390	0.112	0.098
R^2	0.87	0.98	0.99

DESCRIPTIVE ANALYSES

Dimensionality reduction with manifold learning

The results obtained from predictive analyses imply that our OFM and OFM1 embeds appropriate and significant information not only on local structure but also on local magnetic moments and formation energies of crystalline materials. The results also indicate that closer local structures and materials in our description space yield similar local magnetic moments and formation energies, respectively. This fact motivates us to introduce dimensionality reduction techniques to perform descriptive analyses on the LATX dataset.

Several dimensionality reduction algorithms, for e.g., principle component analysis (linear dimensionality reduction method) and manifold learning (non-linear dimensionality reduction method), have been developed and employed to discover low-dimensional structures from high-dimensional data. In this study, we focused on the ISOMAP [29] manifold learning technique. ISOMAP aims to extract a low-dimensional data representation that best preserves all pairwise distances between input points, as measured by their geodesic distances along the manifold. It approximates the geodesic distance as a series of hops between neighboring points. ISOMAP can be viewed as an adaptation of Classical Multidimensional Scaling (MDS) [30], in which geodesic distances replace Euclidean distances. The first step of ISOMAP is to construct the pairwise distance between all points in the original space, identify the neighbors of each point, and make a connection between the center points to their neighbors. In the second step, the geodesic distances between all the points are estimated. Finally, MDS is applied to the geodesic distance matrix to find the lower-dimensional representation embedded in the original high-dimensional representation. In this study, we applied ISOMAP to find a low-dimensional representation of valence-orbital coordination descriptors of material space for the visualization and detection of patterns of behaviors in the material space of LATX alloys. The visualization of materials space is expected to pro-

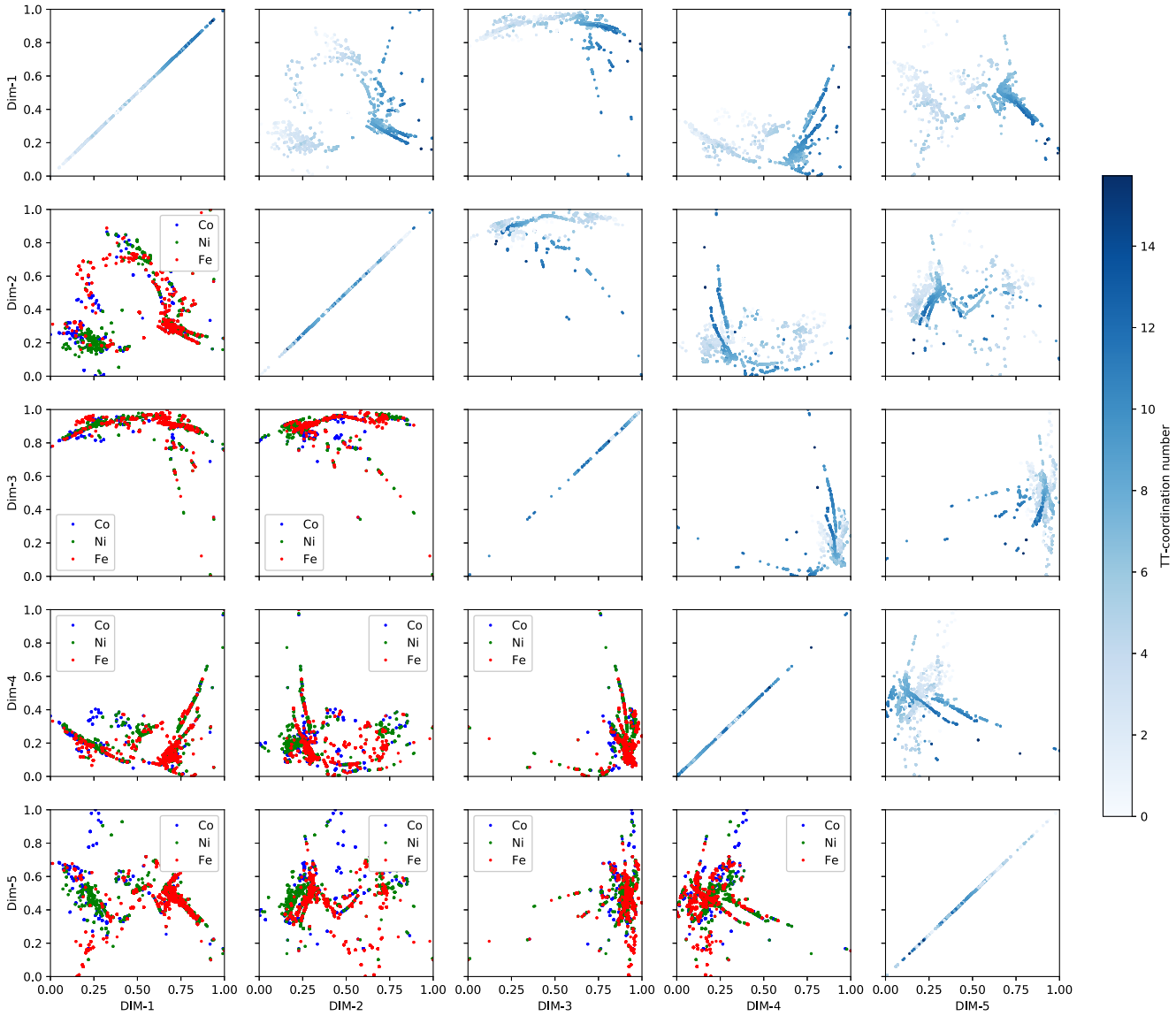

FIG. 4. Maps of the chemical environments of Fe, Co, and Ni based on the five embedded features extracted from the descriptors for local structures by ISOMAP. Diagonal panels show the histograms of the embedded features; the upper-right panels show scatter plots of all the environments, colored by the number of the transition metal atoms surrounding the central atom; the lower-left panels show scatter plots, colored according to chemical symbol: Fe (red), Ni (green), and Co (blue).

vide important insights supporting the inferences on the properties of new materials based on their locations and proximities.

Local structure space visualizations

Fundamental chemistry and physics have shown that elements prefer to reside in some particular chemical environments according to their valence states. As described above, our descriptor includes essential information on valence shell electrons and configurations, and the coordination of atoms around a central atom. Our descriptor is expected to be vital in providing insights and searching for the suitable environment for a specific element.

We collected all the chemical environment vectors (Eq. 1) of Fe, Co, and Ni in the LATX dataset, and applied nonlinear manifold learning to find the hidden features embedded in the dataset. The 36 nearest neighbors were used to build the graph for geodesic distance calculations. For visualization, we maintained the five major new dimensions extracted by the ISOMAP algorithm. Fig. 4 shows the maps (the pairplots) of the chemical environ-

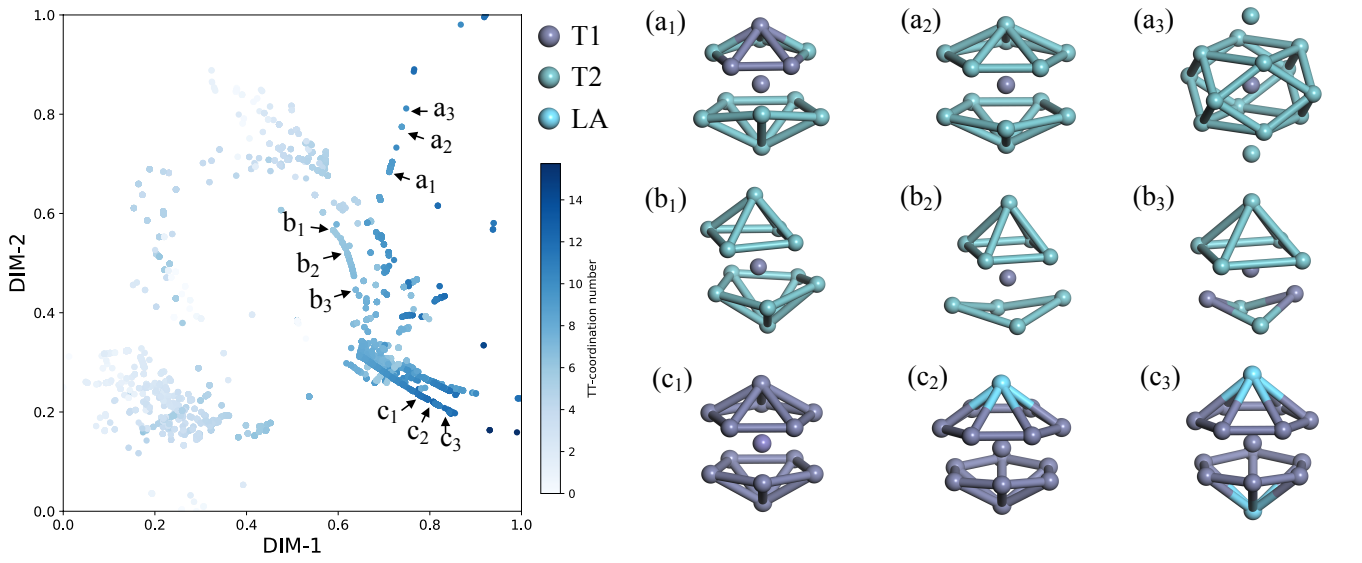

FIG. 5.

Groups of local structures with similar symmetries and shapes can be identified in the form of trajectories in the maps with embedded features obtained by ISOMAP algorithm. T1 indicates an atom of a fourth-period transition metal in the periodical table of elements. T2 indicates a fifth- or sixth-period transition metal atom. LA indicates a lanthanide metal atom.

ments of Fe, Co, and Ni in five hidden features obtained by ISOMAP. The upper right panels show scatter plots of all the environments, colored by the number of transition metal atoms surrounding the central atom; the lower left panels show scatter plots showing each chemical symbol in a different color : Fe (red), Ni (green), and Co (blue). The upper triangular panels of Fig. 4 show the maps of all the local environments of Fe, Ni, or Co, with continuously changing number of transition metal atoms surrounding the central atoms (TT-coordination number). The plots in the first row show the separation of high and low TT-coordination numbers. The upper left plot shows the continuing change in TT-coordination number, implying that the first hidden feature correlates with the TT-coordination number. We obtained the correlation coefficient for this feature and a TT-coordination of 0.7.

Interestingly, these maps show some continuous trajectories, which, upon careful analyses, can be assigned to the gradual deformation of specific structure prototypes. We focus on the two-dimensional map plotted using the first and the second major dimensions extracted by the ISOMAP algorithm (left panel of Fig. 5). A series of dots, or look lines if the density of dots is higher, can be observed in the map. We pick up three characteristic and easy-to-recognize series named (a), (b), and (c), and investigate the corresponding local structures to clarify the manner in which the present descriptor recognize the similar structures. The representative local structures in the three series are shown in the right-hand-side panel of Fig. 5. T1 indicates an atom of a fourth-period transition metal atom in the periodical table of elements. T2 indicates a fifth- and sixth-period transition metal atom. LA indicates a lanthanide metal atom.

For the case of the local structures in the series (a), most of the neighboring atoms are T2 and two of the neighboring atoms are T1 in the local structure (a1). The local structure (a2) is the same as (a1), except that all the neighboring atoms are T2. The larger atomic radius of T2 than that of T1, leads to a variation in the local symmetry and an elongation in the bond lengths of (a2). The coordination number of the transition metal atoms surrounding the central atoms increases continuously from the local structures (a1) to (a2). Furthermore, in order to obtain a higher coordination number, the local structure (a2) changes drastically to the local structure (a3) with higher symmetry. The trend in the series (b) is almost the same as that of the series (a). Two T1 atoms are neighboring atoms in the local structure (b3), and become T2 atoms in (b2). The coordination number of transition metal atoms surrounding the central atom increases continuously from the local structures (b1) through (b2) to (b3).

The trend in the series (c) is slightly different from that in the series (a) and (b). All the neighboring atoms are T1 in (c1), but one atom in the top edge is LA in (c2), and another atom on the bottom edge is also LA in (c3). The replacement of T1 by LA increases the bond length and disrupts local symmetry significantly. Because the corresponding solid angles of the LA atom are small, only tiny deformations are counted in the OFM representation. Consequently, the three structures are close to each other in the ISOMAP, though the one-hot vector of LA is significantly different from that of T1. We can see the similar series of dots or lines in other pan-

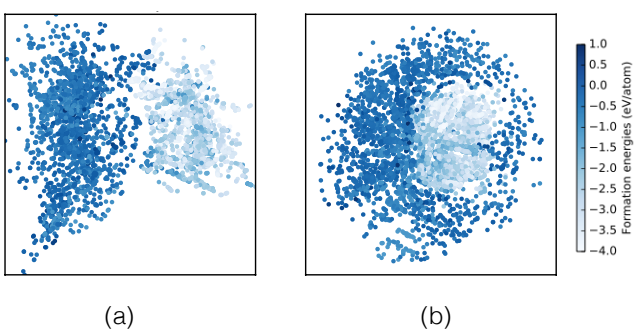

FIG. 6. Maps of the LATX dataset with two new embedded features extracted from the descriptors for entire crystalline materials by ISOMAP (a) and MDS (b).

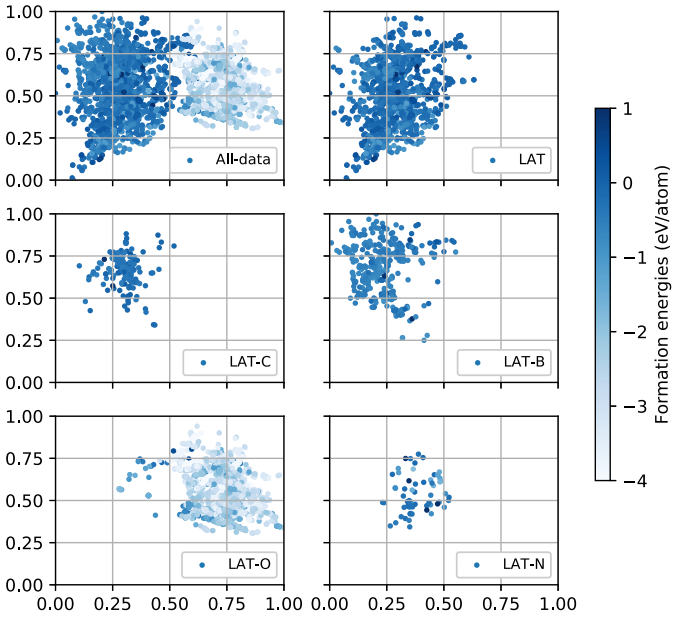

FIG. 7. Different parts of the LATX dataset: LAT, LAT-C, LAT-B, LAT-O, and LAT-N in maps by two embedded features extracted from the descriptors for entire crystalline materials by ISOMAP.

els in Fig. 4, which illustrate the the manner in which the OFM represents the structure if they are examined in detail.

The lower left panels of Fig. 4 show that distributions of the chemical environment of Fe, Ni, and Co can be distinguished. It should be noted that in the LATX dataset, both the alloys with single and multiple species of the transition metal were included, and more than half of the data belonged to alloys with multiple species of the transition metal. The obtained result explicitly indicates the differences in the preferred chemical environments of Fe, Co, and Ni. However, an overlap is also observed in the preferred chemical environment of these transition metals, which indicates that these transition metals have similar chemical environments in some materials. This is consistent with the results suggesting that our descriptor can be very useful for measuring the similarity between local structures in the materials. Therefore, we suggest that Fe, Co, or Ni can replace one another in the chemical environments belonging to the common preferred environment regions. Consequently, hypothetical structures for new stable alloys can be obtained automatically from a known alloy by partially substituting its transition metal sites with another transition metal that shares similar preferred chemical environments. Further application of this method for materials design is promising.

Material space visualizations

As reported above, a reasonably good performance of our nearest-neighbor regression in the prediction of formation energies indicates that a material and its neighboring materials in our representation space are "close" in terms of energy. On the other hand, we know that the formation energy of a material can be calculated using information of atomic positions in the optimal structure model. Therefore, the features of an optimal structure model of materials, as well as its derived formation energy, results in a series of optimizing processes and has strong correlations to one another. In other words, considering descriptors that can express all the degrees of freedom of the material structures, the structures of stable materials lie on a hyper surface of the space spanned by these descriptors. Based on this aspect, we expect that LATX dataset is on a manifold in the material structure space described in the orbital field matrix space. Hence, we applied nonlinear manifold learning to find the hidden features embedded in the dataset. ISOMAP with geodesic distance was employed in this study. The ten nearest neighbors were used to build the graph for geodesic distance calculations. For visualization, we only present the two major new dimensions extracted by the ISOMAP algorithm.

Fig. 6 (a) and (b) depict the map of LATX in the space with two new embedded features obtained by ISOMAP and MDS, respectively. The ISOMAP image shows two separate groups of alloys with high formation energy (left) and low formation energy (right). Interestingly, the MDS image does not show this separation. However, low-formation energy alloys tend to cluster in the center of the map. As seen in Fig. 7, LAT, LATB, LATC, and LATN alloys mainly lie in the high-energy region, while the LATO alloys mainly lie in the low-energy region. This observation can be attributed to the high affinity of LAT metals to oxygen. It is noted that using our orbital field matrix combined with manifold learning, datasets with transition metal and lanthanide metal alloys can be mapped to low dimensional maps with a

meaningful pattern.

CONCLUSION

Herein, we have developed descriptors for material structures for datasets of multi-element materials, based on the information on atomic valence shell electrons and their coordination. Our experiments with transition metal/lanthanide metal alloys show that the local magnetic moments and formation energies can be accurately reproduced using simple nearest neighbors and kernel ridge regressions based on our descriptors. Using kernel ridge regressions, we could accurately reproduce DFT formation energies and local magnetic moments with MAE values of 0.03 μ_B and 0.098 eV/atom, respectively. ISOMAP and MDS can be applied using the OFM and OFM1 descriptors to discover the hidden embedding features in the local structures and crystal structures of materials in the LATX dataset. The dataset is then mapped to transform the embedding features into low-dimensional spaces. Intuitive prehension on the local structure space can be easily acquired using these hidden embedding features. Qualitative evaluation of the similarities in local structures can be inferred directly using the Euclidean distance, and groups of local structures with similar symmetries and shapes can be easily identified in the form of trajectories in the extracted low-dimensional space. The extracted low-dimensional space of the crystal structure of the LATX dataset shows an apparent separation between the two groups of materials having high and low formation energies. The obtained results suggest a guideline for designing new materials by element substitution.

ACKNOWLEDGMENTS

This work was supported by Precursory Research for Embryonic Science and Technology from Japan Science and Technology Agency (JST), the Elements Strategy Initiative Project under the auspice of MEXT, the 'Materials Research by Information Integration' Initiative (MI^2I) project of the Support Program for Starting Up Innovation Hub from JST, and MEXT as a social and scientific priority issue employing the post-K computer (creation of new functional devices and high-performance materials to support next generation industries; CDMSI).

[1] S. Yousef, G. Da, N. Thanh, B. Scotty, C. J. R., and A. Wanda, Phys. Rev. B **85**, 104104 (2012).
[2] S. Yang, M. Lach-hab, I. I. Vaisman, and E. Blaisten-Barojas, Phys. Chem. C **113**, 21721 (2009).
[3] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, Chem. Mater **22**, 3762 (2010).
[4] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, Phys. Rev. Lett **108**, 253002 (2012).
[5] O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, and S. Curtarolo, Chem. Mater **27**, 735 (2015).
[6] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. **114**, 105503 (2015).
[7] H. C. Dam, T. L. Pham, T. B. Ho, A. T. Nguyen, and V. C. Nguyen, J. Chem. Phys. **140**, 044101 (2014).
[8] T. L. Pham, H. Kino, K. Terakura, T. Miyake, and H. C. Dam, J. Chem. Phys. **145**, 154103 (2016).
[9] P. Domingos, Commun. ACM **55**, 78 (2012).
[10] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[11] J. Behler, J. Phys. Chem **134**, 074106 (2011).
[12] N. Artrith and A. M. Kolpak, Nano Letters **14**, 2670 (2014).
[13] H. Eshet, R. Z. Khaliullin, T. D. Kuhne, J. Behler, and M. Parrinello, Phys. Rev. B **81**, 184107 (2010).
[14] H. Eshet, R. Z. Khaliullin, T. D. Kuhne, J. Behler, and M. Parrinello, Phys. Rev. Lett **108**, 115701 (2012).
[15] N. Artrith, T. Morawietz, and J. Behler, Phys. Rev. B **83**, 153101 (2011).
[16] N. Artrith and J. Behler, Phys. Rev. B **85**, 045439 (2012).
[17] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, Scientific Reports **3**, 2810 (2013).
[18] M. Rupp, A. Tkatchenko, K.-R. Muller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[19] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, Int. J. Quantum Chem. **115**, 1094 (2015).
[20] M. Rupp, Int. J. Quantum Chem. **115**, 1058 (2015).
[21] T. Lam Pham, H. Kino, K. Terakura, T. Miyake, K. Tsuda, I. Takigawa, and H. Chi Dam, Sci Technol Adv Mater **18**, 756 (2017).
[22] M. O'Keeffe, Acta. Cryst **A35**, 772 (1979).
[23] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Comp. Mater. Sci **68**, 314 (2013).
[24] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson, APL Materials **1**, 011002 (2013).
[25] S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, and K. A. Persson, Comput. Mater. Sci **97**, 209 (2015).
[26] 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, and E. Duchesnay, J Mach Learn Res **12**, 2825 (2011).
[27] K. P. Murphy, "Machine learning: A probabilistic perspective," (MIT Press, 2012).
[28] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, Npj Computational Materials **1**, 15010 EP (2015).
[29] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, Science **290**, 2319 (2000).
[30] A. J. Izenman, ed., "Modern multivariate statistical techniques: Regression, classification, and manifold learning," (CRC Press, 2009).