SCIENCE AND TECHNOLOGY OF ADVANCED MATERIALS, 2017
VOL. 18, NO. 1, 756–765
https://doi.org/10.1080/14686996.2017.1378060

Taylor & Francis
Taylor & Francis Group

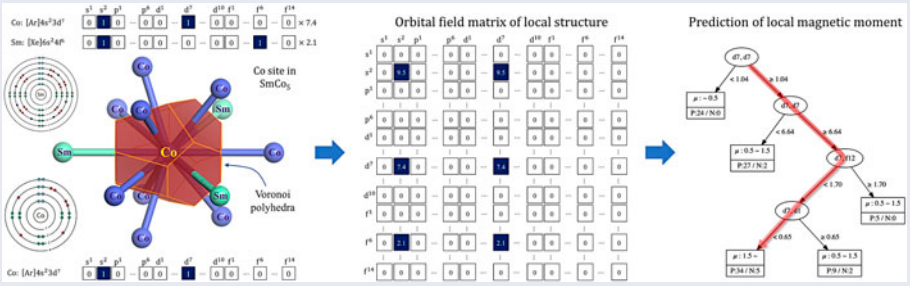
Machine learning reveals orbital interaction in materials

Tien Lam Pham[a,b], Hiori Kino[b,c], Kiyoyuki Terakura[a,c], Takashi Miyake[b,c,d], Koji Tsuda[c,g,h], Ichigaku Takigawa[e,f] and Hieu Chi Dam[a,c,e]

[a]Japan Advanced Institute of Science and Technology, Nomi, Japan;
[b]Elements Strategy Initiative Center for Magnetic Materials, National Institute for Materials Science, Tsukuba, Japan;
[c]Center for Materials Research by Information Integration, Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, Tsukuba, Japan;
[d]Research Center for Computational Design of Advanced Functional Materials, National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan;
[e]JST, PRESTO, Kawaguchi, Japan;
[f]Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan;
[g]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, University of Tokyo, Kashiwa, Japan;
[h]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

ABSTRACT

We propose a novel representation of materials named an 'orbital-field matrix (OFM)', which is based on the distribution of valence shell electrons. We demonstrate that this new representation can be highly useful in mining material data. Experimental investigation shows that the formation energies of crystalline materials, atomization energies of molecular materials, and local magnetic moments of the constituent atoms in bimetal alloys of lanthanide metal and transition-metal can be predicted with high accuracy using the OFM. Knowledge regarding the role of the coordination numbers of the transition-metal and lanthanide elements in determining the local magnetic moments of the transition-metal sites can be acquired directly from decision tree regression analyses using the OFM.



ARTICLE HISTORY
Received 30 June 2017
Revised 7 September 2017
Accepted 7 September 2017

KEYWORDS
Material descriptor; machine learning; data mining; magnetic materials; material informatics

CLASSIFICATION
60 New topics/Others; 404 Materials informatics / Genomics; 203 Magnetics / Spintronics / Superconductors

1. Introduction

Recently, the increasing volume of available experimental and quantum-computational material data, along with the development of machine learning techniques, has provided a new opportunity to develop methods for accelerating discoveries of new materials and physical and chemical phenomena. By using machine learning algorithms, hidden information on materials, including patterns, features, chemical rules, and physical laws, can be automatically discovered from both first-principles-calculated data and experimental data [1–8]. It is commonly known that, in a material dataset, the most important information for identifying a material is its structure. Information on the structure of a material is usually described using a set of atoms with their coordinates and periodic unit cell vectors, which are required for crystalline systems. From the viewpoint of data science, the material data using this primitive representation can be categorized as unstructured data, and the mathematical operations performed on such material data involve the algebra of sets only. Therefore, advanced quantitative machine learning algorithms cannot be applied directly and effectively to conventional material data, owing to the limitation of the algebraic operations of the primitive data representation. In order to apply well established machine learning methods, including predictive learning and descriptive learning, it is necessary to convert the

CONTACT Hieu Chi Dam dam@jaist.ac.jp
Supplemental data for this article can be accessed at https://doi.org/10.1080/14686996.2017.1378060.

primitive representation into fixed-dimensional vectors or matrices, such that the comparison and calculations using the new representation reflect the nature of the materials and the actuating mechanisms of the chemical and physical phenomena. Various methods for encoding materials have been developed in the field of materials informatics.

Previously, Behler and coworkers [9–15] utilized atom-distribution-based symmetry functions to represent the local chemical environments of atoms with a cutoff radius of approximately 1.0 nm, and employed a multilayer neural network to map these chemical environments to the associated local (atomic) energies. The global (total) energy of a given material was then calculated by taking the summation of its local energies. This descriptor is recognized as one of the most successful descriptors for fitting the atomic potential energy surfaces. Bartók and colleagues [16–18] employed the atomic density distribution to compare molecules and solids. Gaussian kernels were used to smoothly approximate atomic density in a local structure. And the similarity between two local structures was estimated by overlapping of their atomic densities which are expanded by spherical harmonic functions. Another successful descriptor was developed by Rupp and coworkers, and is known as the Coulomb matrix (CM) [19–21]. The CM descriptor includes all the pairwise structural information on the atoms in a system and is long range with a length dependence of $1/r$. The CM is used for predicting the atomization energies of small isolated organic molecules and obtained very successful results [20]. Complementary to the mentioned descriptors, there is an effort of combining many types of materials representation including atomic information, the partial radial distribution function, the generalized radial distribution function, etc., together with their covariances, to predict cohesive energies with high accuracy [22]. In spite of the advantage in some predictive analyses, the above descriptors cannot be effectively employed to other interesting mining tasks that require high interpretability of the learning results, for instance, the problems regarding pattern detection of materials behaviors, the extraction of hidden chemical/physical knowledge from a material dataset, the visualization of material datasets in a low dimensional space, etc.

Another interesting attempt at descriptor design involves the introduction of information on the electronic structures. Previously, Isayev et al. used band structures and density of states (DOS) fingerprint vectors as representations of materials to visualize the material space [5]. However, use of information on the electronic structure requires first-principles calculations, which have a high computational cost. We believe that it is a good direction if we can take into account information of the electronic states, and the atomic electron configuration may be regarded as the zeroth order approximation and could be considered as a viable substitution. Structural fragment arrangement has also been utilized to encode materials in order to predict their physical properties [5,23]. This kind of descriptor exhibits good performance for molecular systems, and important fragment patterns concerning a certain material property can be discovered from the learned results. Through consideration of these descriptors, the present authors obtained the concept of developing a descriptor for crystalline materials based on a local structure comprised of a center atom and its neighboring atoms (this local structure can also be regarded as a structural fragment), along with information on the atomic electronic structure (electronic configuration) of the constituent atoms.

To render data-driven approaches meaningful and useful for materials science studies, it is necessary to design material representations with which the results derived using machine learning methods can be interpreted in the language of physical chemistry. It has been well established in fundamental chemistry that certain important aspects of the electronic structure can be deduced from a simple description of the nearest atoms or valence electrons around an atom in a molecular or crystalline system; e.g. the Lewis theory provides powerful tools for studying molecular structure [24]. The ligand field and crystal field theories are examples of other theories developed based on this intuition to classify or categorize local atomic environments, and several fruitful results have been obtained using these theories [25]. Needless to say, within these theories, information regarding the long-range interactions can be included by embedding the information on the local chemical environment of the nearest atoms using a *convolutional manner*. We utilize this heuristic intuition to implement the above-mentioned concept of developing a novel representation by incorporating the information on the local structure and the number of valence orbitals (electrons) coordinating the valence orbital of the center atom. We name this type of descriptor the 'orbital field matrix (OFM)'.

In this work, with emphasis on the interpretability of the derived learning results, we design a material descriptor that (1) utilizes information on the local structure, (2) incorporates the valence atomic configuration, and (3) accepts algebraic operations to construct global descriptors from local descriptors. To verify the applicability of the proposed material representation, we focus on magnetic materials based on bimetal alloys of lanthanide metal and transition-metal (LAT) and LAT alloys including a light element X, which may be B, C, N, or O (LATX). We first examine the decision trees for predicting the magnetic moments of Mn, Fe, Co, and Ni in LAT alloys. The decision trees learned from the LAT alloy data show that the coordination numbers of the occupied *d* orbitals of the

transition-metals and the occupied f orbitals of the lanthanides play important roles in determining the local magnetic moments of the transition-metal sites. The obtained results confirm the interpretability of our OFM representation regarding structural and physical chemistry. In addition, kernel ridge regression (KRR) analyses using standard techniques and similarity measures are implemented in learning prediction models to quantitatively predict the local magnetic moments of transition-metal sites in LAT alloys, formation energies for LATX materials, and atomization energies for organic molecules. Our computational experiments show that the OFM representation can accurately reproduce the local magnetic moments of transition-metal sites in LAT alloys, formation energies of crystalline systems, and atomization energies of molecular systems. The high prediction accuracy confirms the practicability of our OFM representation.

2. Methodology

2.1. Representation of materials

To design the representation for a material, we start with the representation for an atom as a material building block. We utilize the standard notation for electron configuration to develop the representation for an atom; e.g. the electron configurations of Na and Cl are [Ne]$3s^1$ and [Ne]$3s^2 3p^5$, respectively. In order to convert this standard notation into a numerical vector, we borrow the concept of one-hot vector in the field of natural language processing, in which a word is represented by a bit vector having the dimension of the number of words in a dictionary. The vector consists of elements with values of 0, with the exception of a single element used uniquely to identify the word. The representation of an atom is then converted from the standard notation into a one-hot vector $\vec{O}_{atom}$ by using a dictionary comprised of the valence subshell orbitals: $D = \{s^1, s^2, p^1, p^2, \ldots, p^6, d^1, d^2, \ldots, d^{10}, f^1, f^2, \ldots, f^{14}\}$ (e.g. d^5 indicates the electron configuration in which the atomic valence d orbital holds five electrons), which consists of 32 elements (Figure 1).

Next, we design the representation of the coordination number. It is not easy to define the coordination number for realistic crystal structures and there exist a number of such definitions. In this study, we adopt the definition by O'Keeffe [26], which utilizes the solid angles determined by the faces of the Voronoi polyhedra. This method can give the same coordination numbers for the high-symmetry atomic environment and evaluate coordination numbers for the lower-symmetry atomic environment automatically and with no ambiguity. We implement this method using Python Materials Genomics (pymatgen) code [27].

We represent a local structure surrounding an atom by considering the sum of the weighted vector representations of all surrounding atoms in the local structure using $\vec{O}_{atom}$ and the coordination number. A central atom at site p in a local structure can be represented using the OFM with the elements X^p, which are defined as follows:

$$X^p = \sum_{k=1}^{n_p} \vec{O^p}^T \times \vec{O}_k \times w_k,$$
$$X^p_{ij} = \sum_{k=1}^{n_p} o^p_i o^k_j \frac{\theta^p_k}{\theta^p_{max}}, \quad (1)$$

where $i, j \in D$; k is the index of the nearest-neighbor atoms; n_p is the number of nearest-neighbor atoms surrounding site p; w_k is a weight that represents the contribution of atom k to the coordination number of the center atom, p; o^k_j and o^p_i are elements of the one-hot vectors of the kth neighboring atom and p (o^u_v is 1 if the valence orbitals of the atom at site u have electron configuration of type v; otherwise, it is 0) representing the electron configuration. Further, $w_k = \theta^p_k/\theta^p_{max}$, gives a weight of atom k in the coordination of the central atom at site p, where θ^p_k is the solid angle determined by the face of the Voronoi polyhedron separating k and p, and θ^p_{max} is the maximum among n_p of them. An element of OFM, X^p_{ij}, represents the number of orbitals j coordinating the center orbital i.

Additionally, to incorporate the information on the sizes of the valence orbitals, the distance r_{pk} between p and k should be included in w_k. We propose the following form for the calculation of the OFM elements:

$$X'^p_{ij} = \sum_{k=1}^{n_p} o^p_i o^k_j \frac{\theta^p_k}{\theta^p_{max}} \zeta(r_{pk}), \quad (2)$$

where $\zeta(r_{pk})$ is a function representing the contribution of r_{pk} to w_k. In this work, we use the inverse of the distance as the distance-dependent weight function: $\zeta(r_{pk}) = 1/r_{pk}$. We use this $\zeta(r_{pk})$ to distinguish atoms of the same valence configuration with different core shells and to describe the length dependence between the atoms. (Note that we can add the information on the core shells to the hot vector without losing the algebraic operation.)

Composing the descriptor for a structure (a molecule or a crystal system) from its local structure representation requires careful consideration. From the data science viewpoint, the composed descriptors should include as much information as possible. On the other hand, from the materials science viewpoint, the descriptors should be composed so that the natures of the target physical properties are reflected appropriately. For simplicity, in this work, for the atomization energy of a molecule (which is proportional to the molecule size), we take the sum of the descriptors of the local structures as the descriptor for the entire structure:

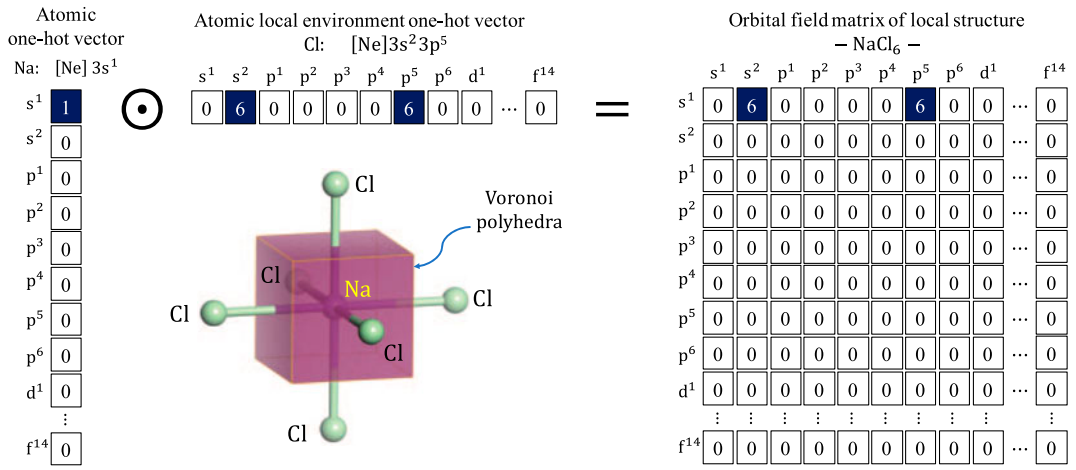


Figure 1. OFM representation for an Na atom in a regular octahedral site surrounded by six Cl atoms: atomic one-hot vector for Na (left), representation for the six Cl atoms surrounding the Na atom (middle), and representation for the Na atom surrounded by six Cl atoms (right).

$$F_{ij} = \sum_{p}^{N_p} X_{ij}^{'p}, \tag{3}$$

where F is the OFM representing the entire molecule. For the formation energy (per atom) of a crystal, which is not proportional to the system size, the descriptor for the entire structure is obtained by averaging the descriptors of the local structures:

$$F_{ij} = \frac{1}{N_p} \sum_{p}^{N_p} X_{ij}^{'p}, \tag{4}$$

where N_p is the number of atoms in the unit cell.

3. Results and discussion

3.1. Prediction of local atomic properties

We now examine how the OFM can be employed to predict the local atomic properties of materials. In this work, we focus 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 [28,29]. We select the structures by combining transition-metals and lanthanides from the 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}. We employ Vienna Ab Initio Simulation Package (VASP) 5.4.1 [30–33] with the generalized gradient approximation (GGA)/Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional[34,35] to calculate the local magnetic moments of these structures. We followed the Materials Project database regarding the selection of projector augmented wave (PAW) projectors [36,37], and employed pymatgen 4.3.0 [27] to prepare the VASP input files with 0.1eV Gaussian smearing of MITRelaxSet and a k-point mesh density of 150 $Å^{-3}$. The energy cutoff is 520 eV. The VASP-PAW includes scalar relativistic effects by default. We perform collinear spin calculations without spin-orbit coupling. The systematic simulations performed in this study were conducted with the assistance of the Organizing Assistant for Comprehensive and Interactive Simulations (OACIS) [38].

In LAT alloys, three types of exchange interactions exist, including the exchange interaction between transition-metal (T) atoms in the T sub-lattices (T–interaction), the exchange interaction between lanthanide metal (LA) atoms and the T sub-lattices (LA –T interaction), and the exchange interaction between lanthanide metal atoms in the LA sub-lattices (LA – interaction). The exchange interactions involving LA elements are mediated by their 5*d* states, because of the strong spatial localization of the 4*f* states. The LA–T interaction is weak and the LA–LA interaction is marginal, in comparison to the T–T interaction. Our description of the local structure in terms of the coordination of the valence electrons is expected to include a significant amount of information regarding these magnetic interactions, which are essential for predicting the local magnetic moment. We first examine which elements in the OFM determine the local magnetic moments of the Mn, Fe, Co, and Ni sites in the LAT dataset through decision tree regression analyses.

To obtain the coordination information, we first employ Equation (1) to analyze the local magnetic moments, without considering the effects of different atomic orbitals having the same angular quantum numbers, but different principle quantum numbers. We

abbreviate $X_{i,j}^{p}$ to $(i,j)^{p}$, which represents the number of orbitals j surrounding orbital i. For instance, to encode the local structures of the metal sites in Sm-Fe alloys via OFM, we begin by representing the valence electron configuration of atomic Fe by s^2d^6 ([Ar] $3d^64s^2$) and that of Sm by s^2f^6 ([Xe] $4f^66s^2$). The $(d^6, s^2)^{\mathrm{Fe}}$ element in the derived OFM indicates the total coordination number of the Fe sites, as s^2 appears in both the Fe and Sm sites. The $(d^6, f^6)^{\mathrm{Fe}}$ element represents the number of Sm sites surrounding the Fe sites, and the number of Fe surrounding an Fe site can be found at the $(d^6, d^6)^{\mathrm{Fe}}$ element. For simplicity, we drop the superscript (p) hereafter. The decision tree regressions for the local magnetic moments of the Mn, Fe, Co, and Ni sites derived from the data are summarized in Figure 2. It is clearly apparent that the (d^n, d^n) elements dominate the decision trees while the (d^n, s^2), (d^n, f^n), or (d^n, d^m) elements decorate the trees. This is consistent with the fact that the local magnetic moment of a transition-metal site is determined mainly by the number of unpaired electrons of the d-orbitals of the central transition-metal atom as well as by the T–T interaction between the same element due to the energy level relation. The appearance of (d^n, s^2), (d^n, f^n), or (d^n, d^m) elements indicates that the LA–T interaction also plays a significant role in the determination of the local magnetic moment.

The tree for the Fe site cases shows that the magnetic moment is less than 2.2 μ_B when the (d^6, d^6) element is less than 6.6 or greater than 9.15. This result implies that the Fe atom appears to have a smaller magnetic moment when surrounded by less than seven Fe atoms or more than nine Fe atoms. The latter case reminds us of the anti-ferromagnetic ground state of face-centered cubic (fcc) Fe with 12 as (d^6, d^6). Further, the magnetic moments of the Fe sites may be greater than 2.5 μ_B when the (d^6, d^6) element is greater than 6.6, but the (d^6, s^2) element (namely, the total coordination number including the contribution of the lanthanide metal atoms) is less than 8.73. The decision tree for the Ni sites shows that those sites tend to have a small magnetic moment (less than 0.2 μ_B) when the (d^8, d^8) element is less than 7.22. However, a large magnetic moment (greater than 0.4 μ_B) can be obtained when the (d^8, d^8) element is greater than 8.25. This implies that the Ni atom has a large magnetic moment when surrounded by more than nine Ni atoms. The magnetic moments of the Ni sites may be greater than 0.4 μ_B when the (d^8, d^8) element is greater than 7.22, but the (d^8, s^2) element (namely, the total coordination number including the contribution of the lanthanide metal atoms), is less than 9.15. This result is also qualitatively consistent with the observation that Ni cannot sustain its magnetic moment alone in metals [39].

For the Co sites, we see that the decision tree uses (d^7, f^{12}) and (d^7, d^1), where f^{12} comes from Er and d^1 comes from La, Ce, Gd, and Lu at the lower branches. Careful analysis reveals that these branches are constructed to separate the cases of $\mu = 0.5$–$1.5\ \mu_B$ for LA-Er and (La, Ce, Gd, Lu)-Co from the case of $\mu > 1.5\ \mu_B$. There are five Er-Co with a local magnetic moment of $1.5\ \mu_B$ and five Er-Co with a local magnetic moment of less than $1.5\ \mu_B$, the criterion of which is $(d^7, f^{12}) = 1.7$. The (d^7, d^1) leaf separates Ce-Co and La-Co (the magnetic moments of which are 1.596 and 1.624 μ_B, respectively) from those with magnetic moments of less than $1.5\ \mu_B$. The maximum local magnetic moment of LA-Co is 1.74 μ_B. The leaf for which the local magnetic moment is largest, i.e. larger than $1.5\ \mu_B$, contains 36 positive and 19 negative cases, if we do not use (d^7, f^{12}) or (d^7, d^1). However, this leaf contains 34 positive and five negative cases if we use the information related to the T–LA interaction to cluster the cases appropriately.

For the case of the Mn sites, the trend is not as clear as for Fe, Co, or Ni. In fact, it is observed that the local magnetic moments for the Mn sites fall within a large range, i.e., from 0.0 to 3.2 μ_B. To obtain a local magnetic moment greater than 2.0 μ_B, a (d^5, d^5) element less than 7.34 and (d^5, s^2) greater than 8.69 are required. However, among the 12 cases satisfying these conditions, only six positive cases were found. Further, three of the six negative cases exhibit local magnetic moments of less than 1.0 μ_B, whereas the other three cases exhibit local magnetic moments of 1.0–2.0 μ_B. This observation can be attributed to the complex magnetic structures of the half-filled d orbitals.

These results show that clustering by the decision trees can determine important elements of the OFM that are consistent with the physical or chemical picture. Further, we can automatically derive quantitative relations between the elements of the OFM and the local magnetic moments using the developed method. Thus, we can expect that the OFM, which we employed as descriptors in the decision trees, can be good descriptors for the regression of the local magnetic moments of the LAT systems.

In the next step, we examine how the local magnetic moments can be represented by the OFM descriptors based on the fact that materials with higher similarity (as estimated by the descriptors) should possess similar local magnetic moments. For this purpose, we employ a simple nearest-neighbor regression method to predict the local magnetic moments, and the cross-validated root mean squared error (RMSE) is used to measure the performance of our descriptors. In the nearest-neighbor regression, a property of a data point is deduced from the properties of the nearest-neighbor points in the training data. For the quantitative prediction of physical properties, it is necessary to distinguish the valence orbital using a different principal quantum number, e.g., the $3d$ orbitals should differ from the $4d$ orbitals. Therefore, hereafter, we use Equation (2) with

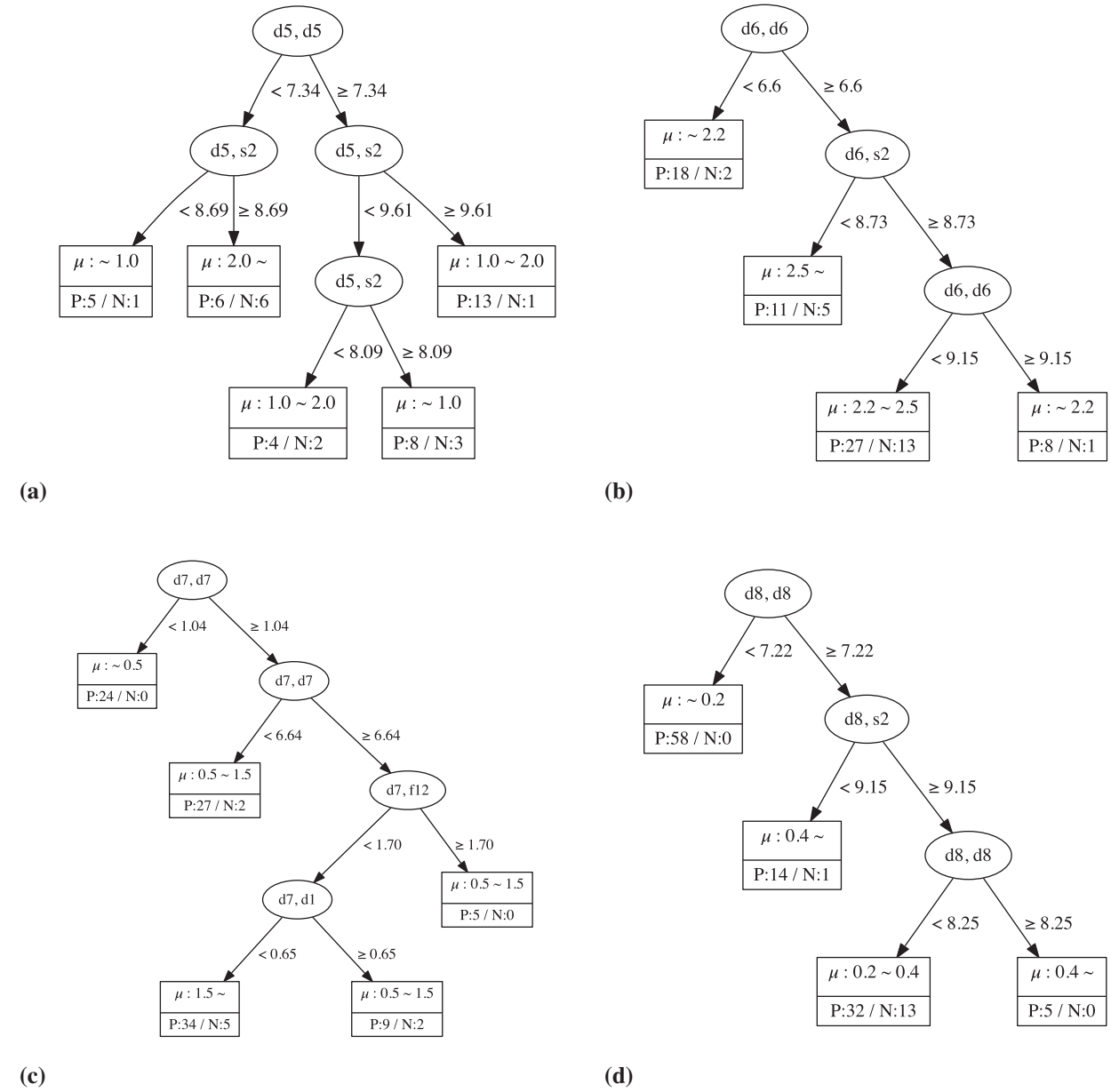


Figure 2. Decision tree regression for Mn (a), Fe (b), Co (c), and Ni (d). In each leaf, the upper part indicates the values of the local magnetic moments, whereas the lower part indicates the number of positive (P) and negative (N) examples.

Table 1. Cross-validated RMSE (μ_B) and R^2 for predicted local magnetic moments obtained via nearest-neighbor regression with selected distance measurements (enumerated in the supplemental information).

Distance	d_{eucl}	d_{man}	d_{cos}	d_{bar}	d_{can}	d_{cor}
RMSE	0.26	0.21	0.23	0.21	0.21	0.23
R^2	0.86	0.90	0.89	0.90	0.90	0.90

the distance weight to generate the descriptors for the local and global structures.

Table 1 summarizes the cross-validated RMSE and the coefficient of determination R^2 between the observed and predicted values obtained using our nearest-neighbor regression and different distance measurements. We achieve a reasonable performance as regards the prediction of the local magnetic moments, obtaining an RMSE of approximately 0.2 μ_B and an R^2 value of 0.9. This result indicates that close materials in our description space of a local structure yield similar local magnetic moments, which implies that our data representation includes significant information about the local magnetic moments. To further improve the prediction of the local magnetic moments, we apply KRR as the prediction model. We obtain a cross-validated RMSE of 0.18 μ_B, a cross-validated mean absolute error (MAE) of 0.05 μ_B, and an R^2 value of 0.93, as indicated in Table 2.

To assess the capability of the OFM descriptor (X' in Equation (2)), we compare its performance with that of the CM descriptor proposed by Rupp and coworkers [19–21]. We treat the local structures in the same

Table 2. Cross-validated RMSE (μ_B), cross-validated MAE (μ_B), and R^2 for predicted local magnetic moments obtained via KRR regression with OFM and CM descriptors.

Descriptor	OFM	CM
RMSE	0.18	0.21
MAE	0.05	0.11
R^2	0.93	0.90

manner as isolated molecules, and the calculated CM descriptors are used to predict the local magnetic moments using KRR regression. Using the CM descriptor, we obtain a cross-validated RMSE of approximately 0.21 μ_B, a cross-validated MAE of 0.11 μ_B, and an R^2 value of 0.90, as indicated in Table 2. The obtained results show that the OFM descriptor, which includes information on the coordination of valence electrons, is more informative and, consequently, yields a slight improvement in prediction accuracy compared to the CM descriptor for the local magnetic moments of the LAT alloys.

3.2. Prediction of material properties

With the aim of obtaining a prediction model with high prediction accuracy, the representation of materials is usually designed to include as much information as possible via a large number of descriptors, without considering their interpretability. In this work, as mentioned above, we focus on developing descriptors, taking both the applicability and interpretability into consideration. Therefore, instead of designing a complicated representation for materials, we choose a simple approach in which the descriptor of a material is derived by averaging or summing the descriptors for the local structures of its constituent atoms. Here, we implement the prediction models for the formation energies of crystalline systems and the atomization energies of molecular systems in order to examine the applicability of the OFM descriptors.

For crystalline systems, we focus on transition-metal binary alloys (TT), and bimetal alloys of lanthanide metal and transition-metal (LAT), as well as LATX and TTX, which are LAT and TT alloys that include a light element X. We select the transition-metals from the set 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}, the lanthanides from {La, Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu}, and the X elements from {B, C, N, O}. We collect the data of more than four thousand compounds, including their structures and formation energies, from the Materials Project repository: 1510 LATX compounds, 1311 TTX compounds, 692 LAT compounds, and 707 TT compounds. We use the average of the descriptors for their local structures to build the global descriptor for each of these materials.

Table 3. Cross-validated RMSE (eV/atom), cross-validated MAE (eV/atom), and R^2 for formation energy of LATX and atomization energy of QM7 dataset obtained using OFM and CM descriptors.

Dataset	LATX		QM7	
Descriptor	OFM	CM [20]	OFM	CM [19]
RMSE	0.190	0.470	0.043	0.040
MAE	0.112	0.390	0.027	0.020
R^2	0.98	0.87	0.98	0.99

For these crystalline systems, we compare the performance of our OFM descriptor (X' in Equation (4)) with that of the CM descriptor, which is based on the Ewald sum and which was developed by Faber and coworkers [20]. We use a KRR model with a Laplacian kernel for both the OFM and CM descriptors. A 10-fold cross-validated comparison between the DFT-calculated formation energies and the machine learning-predicted formation energies is shown for the OFM in Figure 3. The DFT-calculated and ML-predicted formation energies show good agreement, with an R^2 value of 0.98, a cross-validated RMSE of 0.19 eV/atom, and a cross-validated MAE of 0.11 eV/atom. This result is better than that obtained using the CM descriptor, which yields an R^2 value of 0.87, a cross-validated RMSE of 0.47 eV/atom, and a cross-validated MAE of 0.39 eV/atom, as summarized in Table 3. A similar relatively poor result of the CM descriptor has been already reported on the performance in the prediction of the formation energies of crystal systems [20].

For the molecular systems, we focus on the atomization energies of organic molecules. We use the QM7 dataset with 6915 organic molecules [19,40]. (Originally, the QM7 dataset contained 7195 molecules, but more than 100 molecules were removed because of a technical problem in determining Voronoi polyhedra for flat structures). As noted above, the descriptor of a molecule is built by summing over the descriptors of its local structures. Using our OFM representation, Equation (2), and KRR regression, we obtain a cross-validated RMSE of 0.043 eV/atom, a cross-validated MAE of 0.027 eV/atom, and an R^2 value of 0.98. In contrast, the CM yields a cross-validated RMSE of 0.040 eV/atom, a cross-validated MAE of 0.020 eV/atom, and an R^2 value of 0.99 [19–21], as indicated in Table 3. It is worth noting that although our dimension of our OFM seems to be high compared to CM for the small systems, the advantage of OFM is that its dimension is fixed regardless the size of the system. In fact, our OFM contains the only small number of non-zero elements depending on data set. For the QM7 data set, we only need 25 features.

This result confirms that the construction of the OFM of a material, which is achieved by averaging or summing the descriptors of all the local structures of the constituent atoms, yields superior prediction accuracy than the CM descriptor for the formation ener-

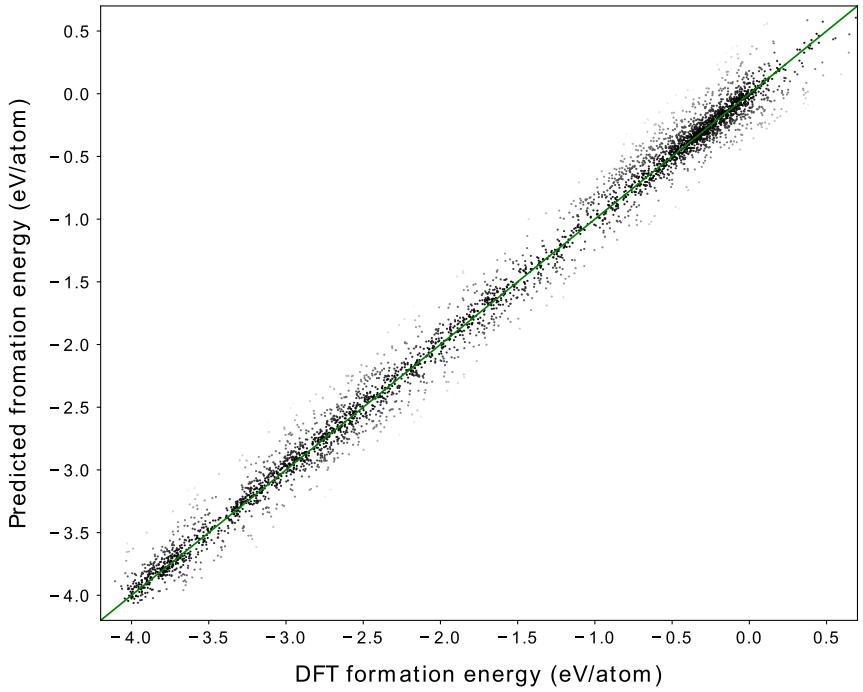


Figure 3. Comparison of formation energies calculated using DFT and those predicted through machine learning (ML-predicted), using OFM.

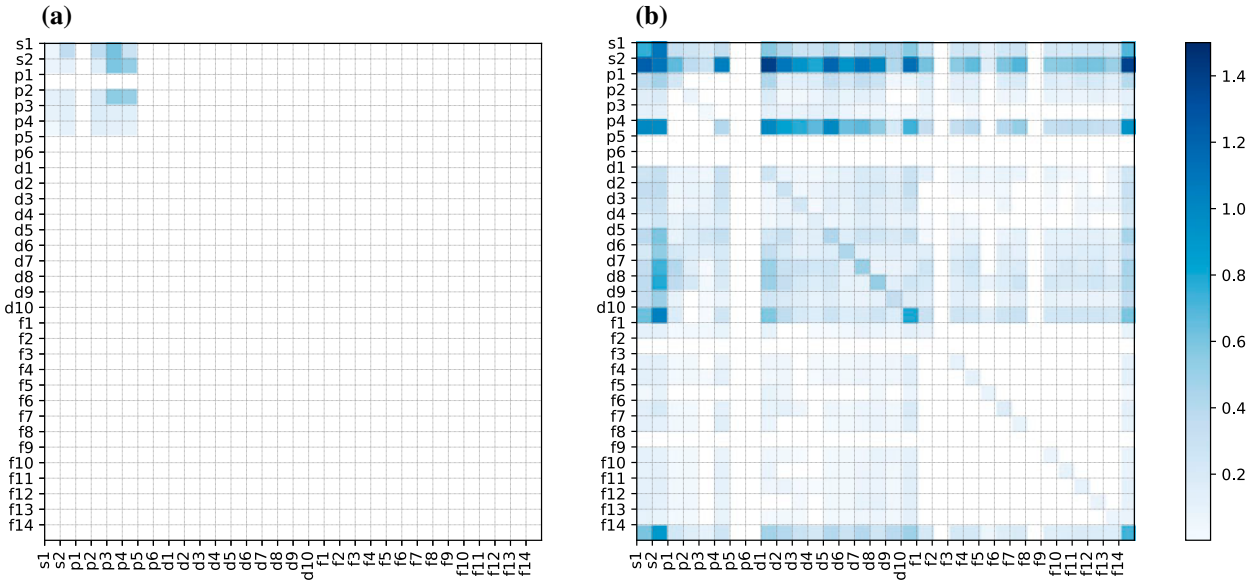


Figure 4. Standard deviations of local OFMs of QM7 (a) and LATX (b) datasets.

gies of LATX systems, and comparable accuracy to the CM descriptor for the atomization energies of organic molecular systems in the QM7 dataset. It may be noted that, for molecular systems (the QM7 dataset contains light elements such as C, H, O, N, and S only), the CM descriptor yields a slightly better result than our OFM. However, for LATX systems with a variety of elements (the LATX dataset contains transition-metals, lanthanides, and light elements), our OFM exhibits superior prediction ability.

Figure 4(a) and (b) depicts the standard deviations of the OFMs of all local structures for the QM7 and LATX datasets, respectively. It is apparent that the QM7 dataset contains only a small number of non-zero OFM elements, whereas the LATX dataset exhibits a large variety of OFMs. Moreover, the QM7 dataset exhibits a small deviation of the OFM, whereas the LATX dataset has a greater deviation. These differences arise because the QM7 dataset is comprised of organic molecules, where the covalent bonding formed by the *sp*

hybridization is a major factor determining the geometry of the nearest neighbor atoms or the coordination number, in principle. Therefore, the QM7 dataset appears to be less divergent as regards the OFM descriptors. On the other hand, the crystalline materials in the LATX dataset, which includes so-called ionic bonding as well as covalent bonding, have considerably higher diversity in terms of both composition and structure. Interestingly, our OFM yields a better result for these complex and divergent systems. The important point to note is that the OFM can describe large diversity in atomic composition and structure more clearly, facilitating the learning and prediction of the properties of both crystalline and molecular systems. The OFM includes the effect of the nearest neighbor sites chosen by the Voronoi polyhedra only. However, for a molecule, the OFM can yield performance equivalent to that of the CM, the descriptors of which are based on long-range power-law decay. Further, the OFM results are of considerably better quality than those of the CM for the periodic LATX systems. Thus, our results indicate that our developed OFM technique offers an essential basis for the theoretical design of materials properties, via an approach similar to building blocks.

4. Conclusions

We have proposed a novel representation of crystalline materials named as 'orbital-field matrix (OFM)', which is based on the distribution of valence shell electrons. We have demonstrated that this new representation can be highly useful in describing and measuring the similarities of materials or local structures in bimetal alloys of lanthanide metal and transition-metal (LAT) as well as LATX (X: light element) ternary alloys. Our experiments show that our OFM can accurately reproduce the DFT-calculated local magnetic moments of transition-metal sites in LAT alloys with a cross-validated RMSE of 0.18 μ_B and an R^2 value of 0.93. Moreover, the results can be interpreted in the language of physical chemistry; that is, the ligand field theory for the local magnetic moment. Decision tree regression shows the importance of the coordination numbers of the occupied d orbitals of the transition-metals and the occupied f orbitals of the lanthanides in determining the local magnetic moments of the transition-metal sites. Further, the formation energies of crystalline systems and the atomization energies of molecular systems can be well predicted using our OFM. That is, with KRR representation, the formation energies of the crystalline systems and atomization energies of the molecular systems can be accurately reproduced with an R^2 value of approximately 0.98. Incorporating information on the atomic orbital coordination, OFM exhibits superior applicability to systems with high diversity in atomic composition and structure in LATX compared to the CM approach. The acquired results suggest that OFM could be useful for mining chemical/physical information on materials from available datasets using modern machine learning algorithms.

Details of the methods and the model parameter optimization are summarized in the supplemental materials.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported in part by Precursory Research for Embryonic Science and Technology from Japan Science and Technology Agency (JST), by the Elements Strategy Initiative Project under the auspice of MEXT, by 'Materials research by Information Integration' Initiative (MI2 I) project of the Support Program for Starting Up Innovation Hub from Japan Science and Technology Agency (JST), by MEXT as a social and scientific priority issue (Creation of new functional devices and high-performance materials to support next-generation industries; CDMSI) to be tackled by using post-K computer, and also by JSPS KAKENHI [Grant Numbers 17K19953 and 17H01783].

References

[1] Yousef S, Da G, Thanh N, et al. Data mining for materials: computational experiments with *ab* compounds. Phys Rev B. 2012;85:104104.
[2] Yang S, Lach-hab M, Vaisman II, et al. Identifying zeolite frameworks with a machine learning approach. Phys Chem C. 2009;113:21721–21725.
[3] Hautier G, Fischer CC, Jain A, et al. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. Chem Mater. 2010;22:3762–3767.
[4] Snyder JC, Rupp M, Hansen K, et al. Finding density functionals with machine learning. Phys Rev Lett. 2012;108:253002.
[5] Isayev O, Fourches D, Muratov EN, et al. Materials cartography: representing and mining materials space using structural and electronic fingerprints. Chem Mater. 2015;27:735–743.
[6] Ghiringhelli LM, Vybiral J, Levchenko SV, et al. Big data of materials science: critical role of the descriptor. Phys Rev Lett. 2015;114:105503.
[7] Dam HC, Pham TL, Ho TB, et al. Data mining for materials design: a computational study of single molecule magnet. J Chem Phys. 2014;140(4):044101.
[8] Pham TL, Kino H, Terakura K, et al. Novel mixture model for the representation of potential energy surfaces. J Chem Phys. 2016;145(15):154103.
[9] Behler J, Parrinello M. Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys Rev Lett. 2007;98:146401.
[10] Behler J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. J Phys Chem. 2011;134:074106.
[11] Artrith N, Kolpak AM. Understanding the composition and activity of electrocatalytic nanoalloys in aqueous

solvents: a combination of dft and accurate neural network potentials. Nano Lett. 2014;14(5):2670–2676.
[12] Eshet H, Khaliullin RZ, Kuhne TD, et al. *Ab initio* quality neural-network potential for sodium. Phys Rev B. 2010;81:184107.
[13] Eshet H, Khaliullin RZ, Kuhne TD, et al. Microscopic origins of the anomalous melting behavior of sodium under high pressure. Phys Rev Lett. 2012;108:115701.
[14] Artrith N, Morawietz T, Behler J. High-dimensional neural-network potentials for multicomponent systems: applications to zinc oxide. Phys Rev B. 2011;83:153101.
[15] Artrith N, Behler J. High-dimensional neural network potentials for metal surfaces: a prototype study for copper. Phys Rev B. 2012;85:045439.
[16] Bartók AP, Payne MC, Kondor R, et al. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. Phys Rev Lett. 2010;104:136403.
[17] Bartók AP, Csányi G. Gaussian approximation potentials: a brief tutorial introduction. Int J Quantum Chem. 2015;115(16):1051–1057.
[18] De S, Bartók AP, Csanyi G, et al. Comparing molecules and solids across structural and alchemical space. Phys Chem Chem Phys. 2016;18:13754–13769.
[19] Rupp M, Tkatchenko A, Muller KR, et al. Fast and accurate modeling of molecular atomization energies with machine learning. Phys Rev Lett. 2012;108:058301.
[20] Faber F, Lindmaa A, von Lilienfeld OA, et al. Crystal structure representations for machine learning models of formation energies. Int J Quantum Chem. 2015;115(16):1094–1101.
[21] Matthias R. Machine learning for quantum mechanics in a nutshell. Int J Quantum Chem. 2015;115(16):1058–1073.
[22] Seko A, Hayashi H, Nakayama K, et al. Representation of compounds for machine-learning prediction of physical properties. Phys Rev B. 2017;95:144110.
[23] Pilania G, Wang C, Jiang X, et al. Accelerating materials property predictions using machine learning. Sci Rep. 2013;3:2810 (1–6).
[24] Kotz JC, Treichel P, Townsend J, editors. Chemistry and chemical reactivity. Belmont, CA: Brooks/Cole; 2008.
[25] Jean Y, Marsden C, editors. Molecular orbitals of transition metal complexes. New York: Oxford University Press; 2005.
[26] O'Keeffe M. A proposed rigorous definition of coordination number. Acta Cryst. 1979;A35:772–775.
[27] 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.
[28] Jain A, Ong SP, Hautier G, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater. 2013;1(1):011002.
[29] Ong SP, Cholia S, Jain A, et al. The materials application programming interface (api): a simple, flexible and efficient (api) for materials data based on representational state transfer (rest) principles. Comput Mater Sci. 2015;97:209–215.
[30] Kresse G, Hafner J. *Ab initio* molecular dynamics for liquid metals. Phys Rev B. 1993;47:558–561.
[31] Kresse G, Hafner J. *Ab initio* molecular-dynamics simulation of the liquid-metal-amorphous-semiconductor transition in germanium. Phys Rev B. 1994;49:14251–14269.
[32] Kresse G, Furthm Ĭ. Efficiency of *ab-initio* total energy calculations for metals and semiconductors using a plane-wave basis set. Comput Mat Sci. 1996;6(1):15–50.
[33] Kresse G, Furthmüller J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. Phys Rev B. 1996;54:11169–11186.
[34] Perdew JP, Burke K, Ernzerhof M. Generalized gradient approximation made simple. Phys Rev Lett. 1996;77:3865–3868.
[35] Perdew JP, Burke K, Ernzerhof M. Generalized gradient approximation made simple [phys. rev. lett. 77, 3865 (1996)]. Phys Rev Lett. 1997;78:1396–1396.
[36] Blöchl PE. Projector augmented-wave method. Phys Rev B. 1994;50:17953–17979.
[37] Kresse G, Joubert D. From ultrasoft pseudopotentials to the projector augmented-wave method. Phys Rev B. 1999;59:1758–1775.
[38] Murase Y, Uchitane T, Ito N. A tool for parameter-space explorations. Phys Procedia. 2014;57:73–76.
[39] Terakura K, Hamada N, Oguchi T, et al. Local and nonlocal spin susceptibilities of transition metals. J Phys F Metal Phys. 1982;12(8):1661–1678.
[40] Blum LC, Reymond JL. 970 million druglike small molecules for virtual screening in the chemical universe database gdb-13. J Am Chem Soc. 2009;131:8732–8733.