SCIENTIFIC REPORTS

OPEN

Pattern Learning Electronic Density of States

Byung Chul Yeo, Donghun Kim, Chansoo Kim & Sang Soo Han

Received: 13 July 2018
Accepted: 28 March 2019
Published online: 10 April 2019

Electronic density of states (DOS) is a key factor in condensed matter physics and material science that determines the properties of metals. First-principles density-functional theory (DFT) calculations have typically been used to obtain the DOS despite the considerable computation cost. Herein, we report a fast machine learning method for predicting the DOS patterns of not only bulk structures but also surface structures in multi-component alloy systems by a principal component analysis. Within this framework, we use only four features to define the composition, atomic structure, and surfaces of alloys, which are the d-orbital occupation ratio, coordination number, mixing factor, and the inverse of miller indices. While the DFT method scales as O(N^3) in which N is the number of electrons in the system size, our pattern learning method can be independent on the number of electrons. Furthermore, our method provides a pattern similarity of 91 ~ 98% compared to DFT calculations. This reveals that our learning method will be an alternative that can break the trade-off relationship between accuracy and speed that is well known in the field of electronic structure calculations.

Electronic density of states (DOS) plays a tremendously important role in determining the properties of metals[1]. Researchers in the fields of solid-state and condensed matter physics carefully diagnose density distributions of free electrons in metals to understand scientific concepts that are hidden in such density distributions (e.g., the *d*-band center theory)[2] and to develop new materials[3,4].

Quantum mechanical approaches (e.g., density functional theory) shed light on the nature of electrons in metals, and first-principles density functional theory (DFT) calculations are successful methods to develop the electronic DOS of metals. Although quantum mechanical methods provide a high accuracy, they have the disadvantage of a severe computational workload, which originates from the complexity of many-body systems[5]. Thus, many researchers are seeking a fast method to predict electronic structures of materials with a high accuracy[6–9].

Within quantum mechanical frameworks, their high computational cost limits the system size that can be studied. To circumvent such causality-based frameworks, an inductive method can be realized by utilizing data and statistical learning algorithms[10–17]. Recently, a machine-learning approach was pursued to address different quantum mechanical problems[18,19], and in particular, to predict the electronic structures of alloys, e.g., to predict the DOS values at the Fermi level[20] or the *d*-band centers[21]. However, to date, these attempts have been limited to the prediction of only single value, and no machine-learning technique is available for the prediction of DOS patterns that includes both the value and shape.

Herein, we propose a new perspective on the representation of DOS that has been regarded as multi-dimensional digital data from one-dimensional continuous curves. Using principal component analysis, we identified highly correlated DOS patterns for various metal systems and proposed features to determine the correlation between the DOS patterns and the atomic structures of materials in a linear subspace. We successfully reproduced the DOS patterns of alloys usually found by quantum mechanical approaches, which is independent of the number of electrons in the system. Furthermore, our method achieves a small loss of accuracy (Accuracy >90%) compared to DFT calculations. The DOS pattern learning method can provide a breakthrough in the trade-off relationship between accuracy and speed, which is well known in the field of electronic structure calculations. Moreover, the approach is applicable for predicting DOS patterns in not only bulk structures but also of surfaces in multi-component alloy systems.

Results

When mapping DOS patterns from the atomic structures of alloys, there is a mathematical puzzle, i.e., the number of input material labels (e.g., compositions, crystal structures, and lattice parameters) is much smaller than the number of output DOS values at the corresponding energy levels. Accordingly, we first compressed the output information by digitizing an analog signal of the DOS in a rectangular window to *one* multi-dimensional vector,

Computational Science Research Center, Korea Institute of Science and Technology (KIST), Seoul, 02792, Republic of Korea. Correspondence and requests for materials should be addressed to S.S.H. (email: sangsoo@kist.re.kr)

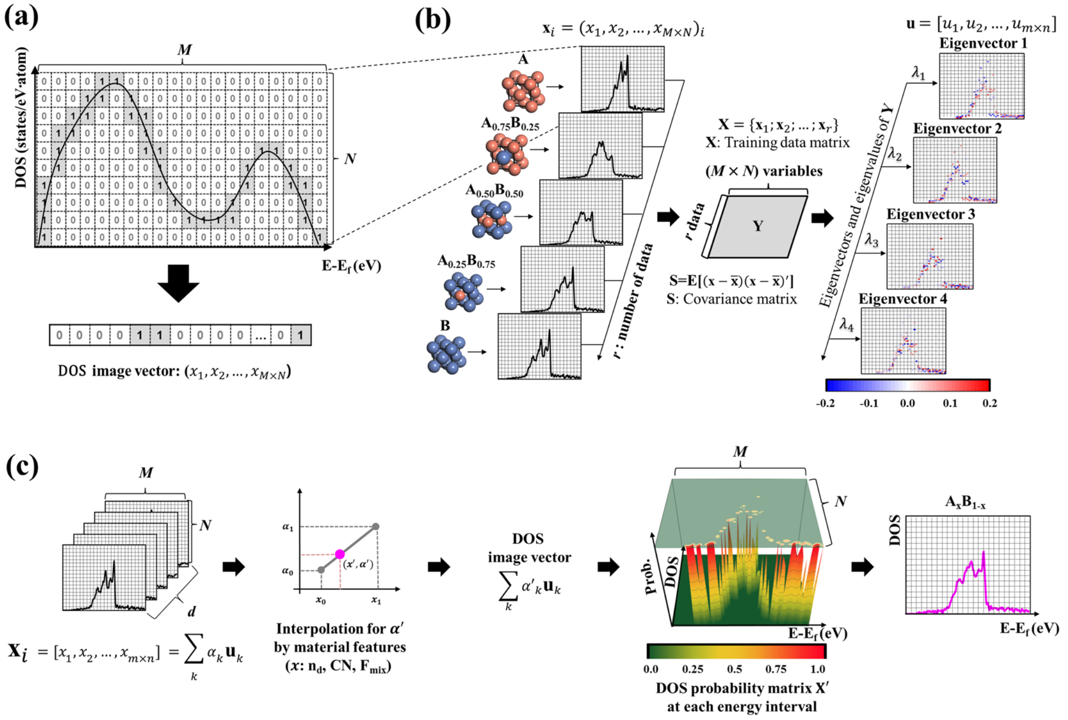
(a)
M
N
DOS (states/eV·atom)
E-E$_f$ (eV)
DOS image vector: $(x_1, x_2, \ldots, x_{M\times N})$
(b)
$\mathbf{x}_i = (x_1, x_2, \ldots, x_{M\times N})_i$
A
$A_{0.75}B_{0.25}$
$A_{0.50}B_{0.50}$
$A_{0.25}B_{0.75}$
B
r : number of data
$\mathbf{X} = \{\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_r\}$
X: Training data matrix
(M × N) variables
r data
Y
$\mathbf{S}=\mathbf{E}[(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})']$
S: Covariance matrix
Eigenvectors and eigenvalues of Y
$\mathbf{u} = [u_1, u_2, \ldots, u_{m\times n}]$
Eigenvector 1
λ_1
Eigenvector 2
λ_2
Eigenvector 3
λ_3
Eigenvector 4
λ_4
-0.2
-0.1
0.0
0.1
0.2
(c)
M
N
d
$\mathbf{X}_i = [x_1, x_2, \ldots, x_{m\times n}] = \sum_k \alpha_k \mathbf{u}_k$
α_1
α_0
(x', α')
x_0
x_1
Interpolation for α' by material features (x: n$_d$, CN, F$_{mix}$)
DOS image vector $\sum_k \alpha'_k \mathbf{u}_k$
M
N
Prob.
DOS
E-E$_f$ (eV)
0.0
0.25
0.5
0.75
1.0
DOS probability matrix X′ at each energy interval
A_xB_{1-x}
DOS
E-E$_f$ (eV)

Figure 1. Scheme of the pattern learning (PL) method for learning and predicting electronic DOSs. (**a**) Conversion of a DOS pattern from a continuous energy function in a rectangular window to a digital image vector with $M \times N$ entries. (**b**) Learning process of PCs of A_xB_{1-x} alloys with their DOS patterns. $\mathbf{x}_i$ is a row vector where M and N correspond to the grid size of the DOS window, and $\bar{\mathbf{x}}$ is the average value of the entries in the row vectors. As a training system for learning, five compositions (A, $A_{0.75}B_{0.25}$, $A_{0.50}B_{0.50}$, $A_{0.25}B_{0.75}$, and B) are considered on the left side. A covariance matrix, **Y**, is constructed in the middle. PCA determine the eigenvectors, which are PCs, and eigenvalues of the training data set, which are shown on the right side. (**c**) The prediction process of an unknown DOS pattern for an arbitrary alloy, A_xB_{1-x}. The process involves several steps: (1) estimation of PC coefficients using features, including n$_d$, CN, and F$_{mix}$; (2) estimation of a new DOS image vector; (3) production and utilization of the DOS probability matrix; and (4) prediction of the DOS pattern for the test alloy, A_xB_{1-x}, using a probability matrix.

as shown in Fig. 1a. Next, we applied principal component analysis (PCA), an unsupervised learning technique, to reduce the high-dimensional data to a low-dimensional data set[22,23]. Then, we could build a model to represent the DOS patterns.

Learning process of DOS patterns. In the learning process of the DOS pattern (ρ), PCA was employed in which, we implemented Python code with matrices operation package *NumPy*[24] for the analysis. Mathematically, this code finds the maximum variance of linearly independent eigenvectors. Prior to the analysis, DOS image vectors were digitized in a rectangular window. In our study, we considered an energy range from −10 eV to 5 eV and a DOS range from 0 to 3. We standardized the DOS image vectors of the training data by obtaining the normalized matrix **Y** in which the i^{th} row ($\mathbf{y}_i$) of **Y** is $\mathbf{x}_i - \bar{\mathbf{x}}$, where $\bar{\mathbf{x}}$ is the mean of each column vector of **X**. Then, we calculated the eigenvectors, $\mathbf{u}_p = (u_1, u_2, \ldots, u_{M\times N})_p$, and the corresponding eigenvalues, λ_p, were calculated by the covariance matrix, $\mathbf{S} = \mathbf{Y}^T\mathbf{Y}$, according to Eq. (1):

$$\mathbf{S}\mathbf{u}_p = \lambda_p \mathbf{u}_p \tag{1}$$

Here, the eigenvectors are called *principal components* (PCs), and the corresponding eigenvalues describe the data variance along the PCs.

The original vector **x** can be reconstructed by using the following Eq. (2):

$$\mathbf{x} \approx \sum_{p=1}^{P} (\mathbf{y}^T\mathbf{u}_p)\mathbf{u}_p + \sum_{p=1}^{P} (\bar{\mathbf{x}}^T\mathbf{u}_p)\mathbf{u}_p = \sum_{p=1}^{P} \alpha_p \mathbf{u}_p \tag{2}$$

where P is the number of PCs and p is their index. Thus, coefficient $\alpha_p$of the eigenvectors can be computed by $\mathbf{y}^T\mathbf{u}_p + \bar{\mathbf{x}}^T\mathbf{u}_p$, and it corresponds to the coordinate values on the linear subspace that is composed of PCs.

In the learning process using the PCA, we identified the linear subspace for which the orthogonal projections of the image vector, **x**, have a maximum variance, and we learned the eigenvectors, **u**, of the training systems in the linear subspace (Fig. 1b). The original image vectors can be reconstructed by $\sum_{p=1}^{P}\alpha_p\mathbf{u}_p$.

Predicting process of DOS patterns. During the predicting process for the DOS pattern (a new image vector, $\mathbf{x}'$) of a test alloy, as shown in Fig. 1c, we estimated the new coefficients, α'_p, via a linear interpolation between α_p of the two training systems that is most similar to the test composition, where features relevant to the electron occupation and atomic configuration were considered (Supplementary Figs S1 and S2). Using $\sum_{p=1}^{P}\alpha'_p\mathbf{u}_p$, we obtained a new image vector, $\mathbf{x}'$, and transformed from the $\mathbf{x}'$ to the DOS probability matrix, $\mathbf{X}'$, the elements of which are the probable values of each DOS levels at the given energy interval.

To predict DOS patterns, only a single DOS value must be determined by a given energy interval. Thus, we defined the DOS probability matrix originating from the DOS image vector. The DOS image vector, $\mathbf{x}' = (x'_1, x'_2, \ldots, x'_{M\times N})$, calculated by the PCs and the estimated coefficients, was transformed to the DOS image matrix, $\mathbf{I}'$, with M columns and N rows in a grid-based rectangular window, and its size is the same as the size used in the learning process (Supplementary Fig. S6). To define the DOS probability matrix, we considered only positive entries in the $\mathbf{I}'$, and the other entries were regarded as zero. Moreover, we normalized all of the entries of the DOS levels at each energy interval. Then, we defined the DOS probability matrix, $\mathbf{X}'$, with M columns and N rows, as given by

$$\mathbf{X}'_{m,n} = \frac{x'_{m,n}}{\sum_n x'_{m,n}} \tag{3}$$

where $x'_{m,n}$ is the positive entry value of the column vector in $\mathbf{X}'$, and m and n are the matrix indices.

To predict the DOS pattern with $\mathbf{X}'$, one should determine a single DOS value at each energy interval. Therefore, we obtained the estimated DOS, which is ρ' and is given by

$$\rho' = \sum_{m=1}^{M}\rho'(E_m) = \sum_{m=1}^{M}\sum_{n=1}^{N}\{\mathbf{X}'_{m,n}\cdot\rho_n(E_m)\} \tag{4}$$

where E_m is the m^{th} energy interval, and ρ_n is the n^{th} DOS level.

Application into binary alloy systems. To test our pattern learning (PL) method, it was first applied to a Cu-Ni system. Thermodynamically, this alloy system shows a complete solid solution, indicating that the Cu and Ni atoms in the alloys are homogeneously mixed in a face-centered cubic (fcc) structure regardless of the composition. Thus, it is expected that the DOS of the alloy system follows intrinsic electronic structures of Cu and Ni crystals and their composition can be a key feature for the representation of their DOS patterns. Therefore, we define the *d*-orbital electron occupation rate ($\mathrm{n_d}$) as an alloy composition-dependent feature that represents local DOS patterns of the *d*-orbitals. Moreover, all of the pristine Cu and Ni and their alloys have a fcc crystal structure, indicating that the effect of the atomic structure on the DOS pattern is not significant. Accordingly, to predict the DOS patterns in this system, we considered only $\mathrm{n_d}$ as a feature. After training the DOS patterns for various Cu-Ni compositions {Cu, $Cu_{0.75}Ni_{0.25}$, $Cu_{0.5}Ni_{0.5}$, $Cu_{0.25}Ni_{0.75}$, Ni}, we predicted the DOS of $Cu_{0.375}Ni_{0.625}$ as the test alloy (Fig. 2a) by considering three PCs. A comparison with the DFT results revealed that our method obtained the pattern similarity of 95% ($\sigma = 0.95$). However, the calculation time is less than 1 minute even on 1 core of an Intel Xeon CPU, whereas the DFT method requires approximately 2 hours on 16 cores of the CPU.

In contrast to the Cu-Ni system, the Cu-Fe system was also considered because the crystal structures of Cu and Fe are different and their alloys do not exhibit a complete solid solution. This implies that features based on the atomic structures in addition to $\mathrm{n_d}$ are required for the DOS representation. We introduced the coordination number (CN) and a mixing factor ($\mathrm{F_{mix}}$) as features to distinguish the atomic structures. The CN was obtained by dividing the number of all bonds between two atoms by the total number of atoms in the system, where the bonds were calculated using the covalent atomic radii. $\mathrm{F_{mix}}$ indicates the ratio of the number of different pair bonds in the alloy system to the total number of bonds. Using $\mathrm{F_{mix}}$, one can distinguish the atomic distributions in alloy systems that have the same CN (Supplementary Fig. S2).

To represent the DOS patterns for the test data, the coefficient α'_p should be determined. Since the eigenvectors obtained after PCA correspond to the PC vectors, the distributed coefficients lying on identical eigenvectors were correlated with each other. Thus, we generated linear regression lines between the α_p of the training data in which we focused on the linear regression line between two training data near the test composition. Then, using the features of the training and test systems, we estimated the α'_k contributions of $\mathrm{n_d}$, CN, and $\mathrm{F_{mix}}$ ($\alpha_p^{'\mathrm{n_d}}$, $\alpha_p^{'\mathrm{CN}}$, $\alpha_p^{'\mathrm{F_{mix}}}$) for the test system using the linear regression line (Fig. 1c). We defined the set of features as $\Phi = \{\mathrm{n_d}, \mathrm{CN}, \mathrm{F_{mix}}\}$. Here, it was assumed that the three features have equal weights so that

$$\alpha'_p = \sum_{\varphi\in\Phi}\beta_\varphi\cdot\alpha'^{\varphi}_p \tag{5}$$

where β_φ is 1/3 for all the features. A detailed description of the estimation of the coefficients is also provided in Section 3 of the Supplementary Information (Table S1 and Fig. S7).

Using these three features ($\mathrm{n_d}$, CN, and $\mathrm{F_{mix}}$), the DOS of $Cu_{0.375}Fe_{0.625}$ was predicted (Fig. 2b). The use of only $\mathrm{n_d}$ leads to the pattern similarity of 78%, while the use of all three features improves the pattern similarity up to 95%. Even in the previously examined Cu-Ni system, consideration of CN and $\mathrm{F_{mix}}$ in addition to $\mathrm{n_d}$ can slightly improve the pattern similarity up to 96% for $Cu_{0.375}Ni_{0.625}$ (Supplementary Fig. S8). Furthermore, we calculated DOS patterns for new test data, $Cu_{0.625}Ni_{0.375}$ and $Cu_{0.625}Fe_{0.375}$, and then obtained the pattern similarities of 95% and 97%, respectively (Supplementary Fig. S9).

To highlight a novelty of our PL method, we additionally calculated the DOS patterns by a linear interpolation of the DOS patterns of the two nearest neighbors without PCA. Compared to the DFT calculation, the

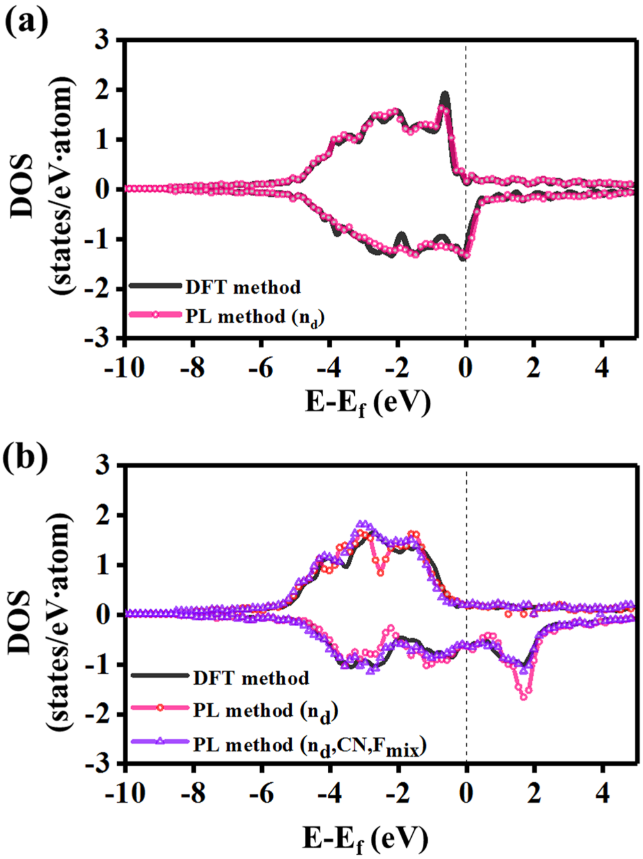

Figure 2. Prediction results of the PL method in binary alloy systems. (**a**) DOS pattern of $Cu_{0.375}Ni_{0.625}$ as a test of the Cu-Ni alloy system. Its atomic structure is shown in Fig. S3. The energy range ($E - E_f$) is from $E = -10$ eV to $E = 5$ eV, and the DOS range is from 0.0 to ± 3.0, where the positive region is for the up-spin, and the negative region is for the down-spin. Black corresponds to the DFT method, and pink corresponds to the learning method using only one feature of n_d. (**b**) DOS pattern of $Cu_{0.375}Fe_{0.625}$ as a test of the Cu-Fe alloy system. Its atomic structure is shown in Fig. S3. Black corresponds to the DFT method, pink corresponds to the learning method using the n_d feature, and violet corresponds to the learning method using all features including n_d, CN, and F_{mix}.

linear interpolation method shows the pattern similarities of 90% for $Cu_{0.375}Ni_{0.625}$ and 88% for $Cu_{0.375}Fe_{0.625}$ (Supplementary Fig. S10), which are lower accuracies than those predicted by our PL method.

Application into multi-component alloy systems. To extend our method to multi-component alloy systems, we also developed a method to represent the DOS patterns of ternary systems, using the example of the Cu-Ni-Pt system (Fig. 3). Figure 3a shows a triangular composition diagram of the Cu-Ni-Pt system, where a total of 15 compositions were considered as the training set: pure 3, binary 9, and ternary 3. Similar to the previous binary cases, by determining the coefficients (α_p) of the PCs for a ternary test composition, one can represent its DOS pattern. First, we selected three training compositions that were located closest to the test composition and calculated the distances (d) between the test composition and the three training compositions (Fig. 3b). Then, the α'_p for the DOS representation were estimated by using the features and α_p at the training compositions, where it was assumed on physical grounds that the DOS pattern of the test composition is represented by the highest weight for the training composition that is nearest to the test composition. According to coherent-potential approximation (CPA)[25,26] that has been extensively used in calculating electronic structures of various alloy systems, the effective or coherent potential lattice can be represented by the average behavior of the A-B binary alloy. However, the single-site nature of the CPA limits its applicability to systems with negligible short-range order and local lattice relaxation effects[27]. When estimating the α'_p for the DOS representation of the test ternary alloy on the basis of the theory, we consider the three training compositions that is nearest to the test composition, rather than the pure compositions. The atomic structures at the training compositions that is nearest to the test composition include more similar atomic distribution information (e.g., atomic ordering or lattice relaxation) to the test alloy than those of the pure metals or others.

The basic idea for an A-B-C ternary case is similar to that of the binary case. In this case, we define the set of features as $\Phi = \{n_{d,A}, n_{d,B}, n_{d,C}, CN_{norm}, F_{mix}\}$. The number of feature values for n_d depends on the number

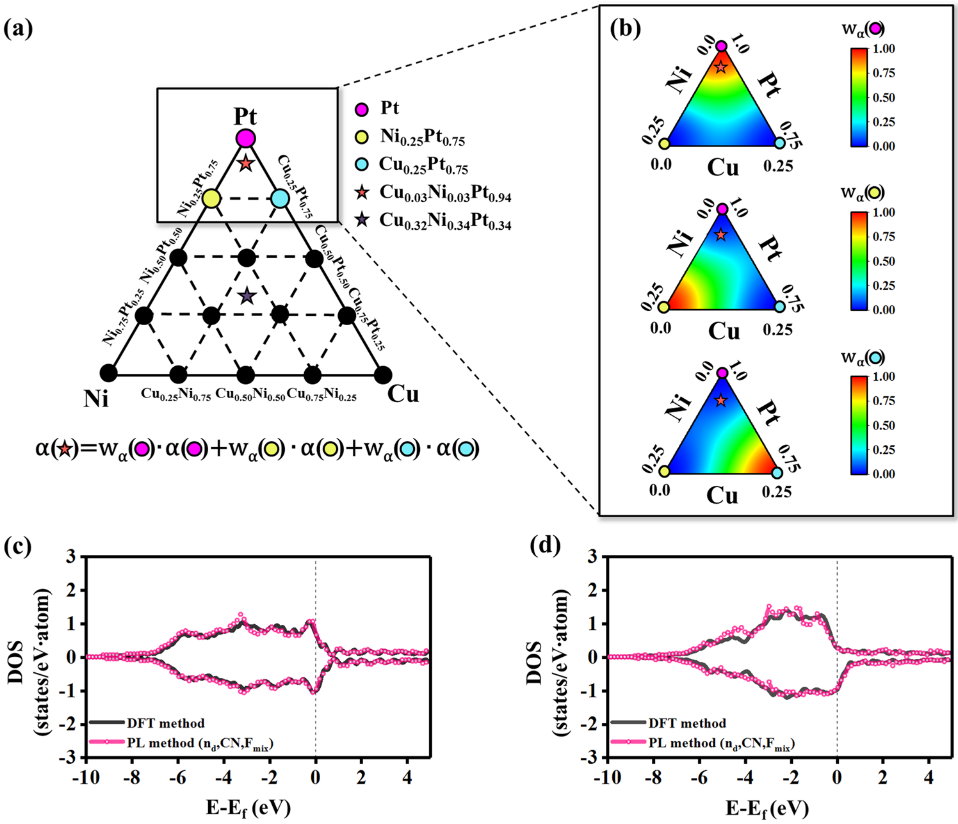

Figure 3. Estimation of coefficients and prediction results of the PL method in ternary alloy systems. (**a**) Triangular diagram of the Cu-Ni-Pt system representing the training data (circle) and test data (star). The equation for the calculation of the PCs coefficients for the test data is shown at the bottom of the figure: the equation is based on the coefficients and their weights for training alloys that most closely match the test alloy composition. (**b**) Maps of the weights of the coefficients of the PC vectors for the test composition ($Cu_{0.03}Ni_{0.03}Pt_{0.94}$). The weights depends on the distance between the test composition and each training composition, and they also depend on the difference of three features (n_d, CN, and F_{mix}) between the training and test data. (**c**) DOS pattern of the $Cu_{0.03}Ni_{0.03}Pt_{0.94}$ test alloy. (**d**) DOS pattern of the $Cu_{0.32}Ni_{0.34}Pt_{0.34}$ test alloy. Their atomic structures are shown in Supplementary Fig. S3. In (**c**,**d**), black corresponds to the DFT method, and pink corresponds to the learning method using all features including n_d, CN, and F_{mix}.

of elements in the multi-components case. Here, we also considered the differences (d_{ij}) in the feature values between the test and the adjacent three training compositions (Supplementary Fig. S11) as given by:

$$d_{ij} = \sum_{\varphi \in \Phi} (\varphi^i - \varphi^j)^2 \tag{6}$$

where *i* and *j* are the selected data of the A-B-C alloy system, and $n_{d,A}$, $n_{d,B}$, $n_{d,C}$, CN_{norm}, and F_{mix} are feature values corresponding to the data. Of the three material features (n_d, CN, and F_{mix}), the n_d and F_{mix} values range from 0 to 1, whereas CN is greater than 1. To obtain units in the same range, we considered the normalized value of CN (CN_{norm}) by dividing the CN value by 12, which is based on the fact that the maximum CN value in the alloy system is 12 for a fcc structure. When the composition and crystal structure of a test alloy are more similar to the training data, the differences in the feature values decreases. We defined Ω as the set of the nearest three training data, ν as the test data, and ν' as the training data. To estimate the PC coefficient ($\alpha'_{k,\nu}$) for the test data, three weights (w) of the coefficients of the three training systems were calculated based on d_{ij} using Eq. (7):

$$w_{\nu\nu'} = \frac{d_{\nu\nu'}^{-1}}{\sum_{\nu' \in \Omega} d_{\nu\nu'}^{-1}} \tag{7}$$

The range of $w_{\nu\nu'}$ is from 0 to 1. Then, the estimated PC coefficients for the test alloy were calculated by

$$\alpha'_{p,\nu} = \sum_{\nu' \in \Omega} w_{\nu\nu'} \cdot \alpha_{k,\nu'} \tag{8}$$

For the example of the Cu-Ni-Pt system shown in Fig. 3, the n_d, CN, and F_{mix} of the training and test data are summarized in Supplementary Table S2. This approach was tested for two compositions: $Cu_{0.03}Ni_{0.03}Pt_{0.94}$ (Fig. 3c) and $Cu_{0.32}Ni_{0.34}Pt_{0.34}$ (Fig. 3d), and we determined that our method obtains the pattern similarity of 96%.

Using the similar procedure, our method can be readily extended to quaternary or quinary alloy systems. As an example, we considered the quinary system of Cu-Ni-Pt-Fe-Cr that is on an extension line of the ternary Cu-Ni-Pt system discussed in Fig. 3. We trained the DOS patterns for the 15 ternary Cu-Ni-Pt compositions in Fig. 3 and the 3 quaternary/quinary Cu-Ni-Pt-Fe-Cr compositions {$Cu_{0.315}Ni_{0.315}Pt_{0.25}Fe_{0.12}$, $Cu_{0.315}Ni_{0.315}Pt_{0.25}Cr_{0.12}$, and $Cu_{0.315}Ni_{0.315}Pt_{0.25}Fe_{0.06}Cr_{0.06}$}, and then predicted the DOS patterns of $Cu_{0.315}Ni_{0.315}Pt_{0.25}Fe_{0.03}Cr_{0.09}$ and $Cu_{0.315}Ni_{0.315}Pt_{0.25}Fe_{0.09}Cr_{0.03}$ as test quinary systems, in which the atomic structures of the training and test quaternary/quinary systems are shown in Supplementary Fig. S4. Compared to the DFT results, our method reveals the superior pattern similarity: 97% for $Cu_{0.315}Ni_{0.315}Pt_{0.25}Fe_{0.03}Cr_{0.09}$ and 96% for $Cu_{0.315}Ni_{0.315}Pt_{0.25}Fe_{0.09}Cr_{0.03}$ (Supplementary Fig. S12). This reveals a high utility of our method. In particular, it can be applied to the field of not only a metallic catalysis but also a high entropy alloy (HEA), in which the HEA is usually consisted of five or more metallic elements. Here, we need to compare our PL method with the CPA[25,26] that has been extensively used in calculating the electronic structure calculations in multicomponent random solid solutions. Recently, the electronic structures of various HEAs were calculated by the exact muffin-tin orbitals (EMTO) method in combination with the CPA[27,28]. The EMTO-CPA is undoubtedly an accurate and efficient method. However, the computational complexity of the EMTO-CPA scales linearly with the number of atoms in the supercell[29], while our PL method is independent on the number of atoms in the supercell (the details will be discussed in the Discussion section). This implies that as a supercell size increases our PL method can be more efficient than the EMTO-CPA method.

Application into surface structures. By expanding the scheme that was applied to bulk structures, we studied the representation of the DOS patterns for surface structures of alloys. In particular, we used method to represent the DOS patterns of high-index surfaces based on those of low-index surfaces. Here, it is important to find a feature to define the surface structures, with which we can estimate the PC coefficients for a test surface. In Fig. 4a, a high-index surface, (211), can be regarded as a surface to connect two low-index surfaces, (011) and (111). Moreover, the step alignment of atoms on the (211) surface plane is generated by a combination of the atom alignments on the (011) and (111) surface plane. In this regard, we employed a lattice plane vector by using the miller indices, which consist of three integers, *h*, *k*, and *l*, and where the notation of the lattice plane vector is written by (*hkl*). Then, we defined the lattice plane that intercepts three points, $\overrightarrow{\boldsymbol{L}}_1/h$, $\overrightarrow{\boldsymbol{L}}_2/k$, and $\overrightarrow{\boldsymbol{L}}_3/l$, where $\overrightarrow{\boldsymbol{L}}_1$, $\overrightarrow{\boldsymbol{L}}_2$, and $\overrightarrow{\boldsymbol{L}}_3$ are the lattice vectors in a conventional unit cell. Therefore, we considered the inverses of the miller indices, $1/h$, $1/k$, and $1/l$, as the features regarding the surface plane orientations, and they are denoted by h', k', and l', respectively. If one of the miller indices is zero, the feature value is set to be zero to avoid an infinity value.

Figure 4b shows the three-dimensional (h', k', l') vector space representing the inverses of miller indices for the training and test surfaces. The training samples include seven lattice plane vectors where all of the miller indices are lower than 2; {(001), (010), (100), (011), (101), (110), (111)}. The vectors correspond to low-index surface plane vectors. Then, after adding the origin vector, (000), and connecting all vectors of the training samples and the origin, a cubic geometrical figure can be obtained. The test sample is the vector of which the miller indices is larger than one, which corresponds to a high-index surface plane vector. Thus, the high index surface plane vectors can lie on an edge or face in the cube figure. In Fig. 4a, the alignments of atoms on the (211) surface plane is a combination of the atom alignments on the (011) and (111) surface planes. Therefore, we can estimate the DOS pattern for the (211) surface with those for the (011) and (111) surfaces. Here, for the three vectors, the miller indices *k* and *l* are the same, indicating that one can be distinguished only using the miller index *h*. During the predicting process of our method (Fig. 1c), we only used the h' value to determine the PC coefficients, $\alpha'_{p(211)}$, by the linearly interpolating between the two coefficients for the (011) and (111) surfaces after performing the PCA using all training DOS data.

To validate our method for surface structures, we tested (211) surfaces of the pure Cu metal (Fig. 4c) and the $Cu_{0.375}Ni_{0.625}$ alloy (Fig. 4d). For the Cu case, our method provided the pattern similarity of 93% compared to DFT calculation (Fig. 4c), where we considered only three DOS data for Cu(001), (011), and (111) surfaces as the training data. When predicting the DOS pattern for the (211) surface of the $Cu_{0.375}Ni_{0.625}$ alloy (see Supplementary Fig. S13), we considered (001), (011), and (111) surfaces of five Cu-Ni alloys; {Cu, $Cu_{0.75}Ni_{0.25}$, $Cu_{0.5}Ni_{0.5}$, $Cu_{0.25}Ni_{0.75}$, Ni}. We first predicted the DOS patterns for (001), (011) and (111) surfaces of the $Cu_{0.375}Ni_{0.625}$ alloy with three features (n_d, CN, and F_{mix}), which is similar to the method used in the bulk case. Then, we performed PCA one more time for the DOS patterns for the low-index surfaces of the $Cu_{0.375}Ni_{0.625}$ alloy that were obtained after the first pattern learning method. Then, using the inverse value of the miller index as a feature, we predicted the DOS pattern for the (211) surface of the $Cu_{0.375}Ni_{0.625}$ alloy, and obtained the pattern similarity of 97% (Fig. 4d). In DFT calculations, the slab calculations are much more time-consuming than bulk calculations. However, our method provides a similar calculation speed for both bulk and surface systems. For example, for the Cu(211) and $Cu_{0.375}Ni_{0.625}$(211) surface, the DFT calculation takes 2 hours on 36 cores of the CPU, while our method is still less than 5 minute even on 1 core of the CPU.

Discussion

Although high-performance computing machines have been used practically thus far, we could still tackle large-scale first-principles calculations of more than hundreds atoms using the limited computing power. Regarding the computation cost, it is well-known that the DFT method scales as $O(N^3)$, where N is the number of electrons in the system[30]. Indeed, a similar trend was observed in this work (Fig. 5a). However, our method remarkably shows a higher speed than DFT and requires only 1 minute regardless of N, although it depends on

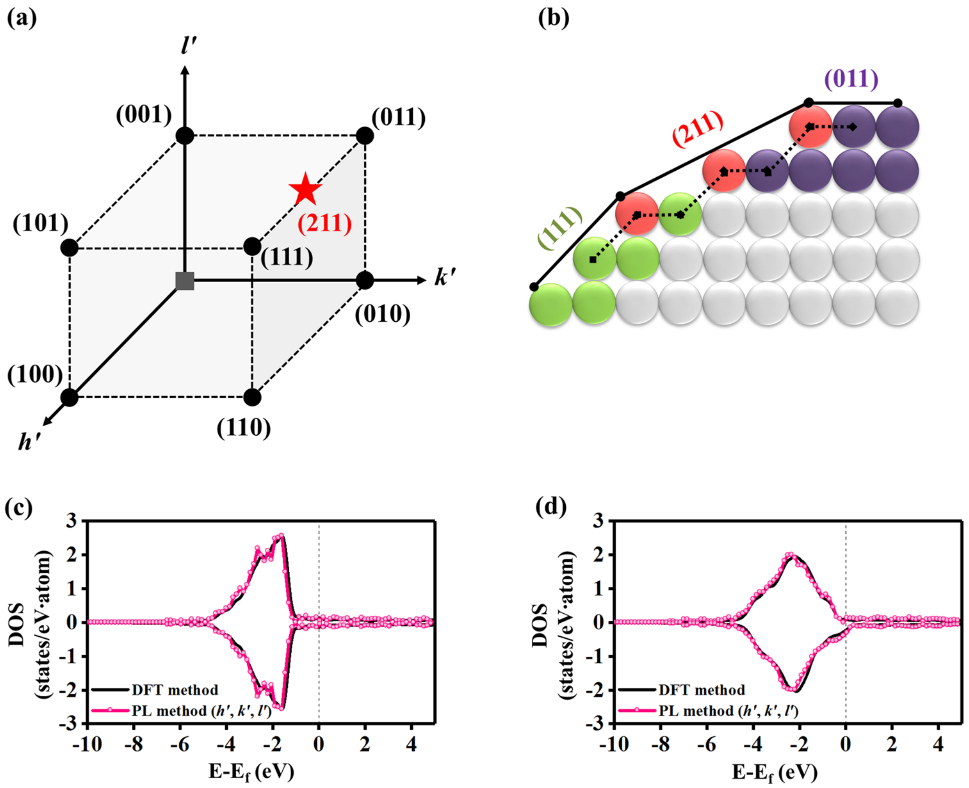

Figure 4. Scheme of the PL method for the DOS representation of surface structures and the predicted results. (**a**) Two-dimensional cleaved lattice structure to represent the plane vectors of the (011) and (111) low-index surface and the (211) high-index surface. Here, red, green, and violet nodes represent the atoms on the surface layer for the (211), (111), and (011) plane vector, respectively. The dotted lines represent the periodicities of the (111) and (011) lattice vectors. (**b**) Three-dimensional cubic diagram of the lattice plane vectors in the coordinate system for the inverse of miller indices h', k', and l' representing the training (black circle) and test (red star) data. (**c**) DOS pattern of the Cu (211) surface. (**d**) DOS pattern of the $Cu_{0.375}Ni_{0.625}$ (211) surface. In (**c**,**d**), black corresponds to the DFT method, and pink corresponds to the learning method (use of three PCs) using the inverse value of the miller index as a feature. The atomic structures of each surfaces can be found in Supplementary Fig. S4.

the training data size since our method needs to scan the entire training data. In addition, tight binding (TB) or density functional TB (DFTB) methods that are an approximate quantum mechanical approach are undoubtedly an efficient method to calculate electronic structures of materials. However, they require massive calculations for eigenvalue distributions of very large matrices[31–34]. Thus, their computational complexity scales linearly with a dimension of the matrix, which depends on the number of atoms or orbitals in the alloy system[35]. However, our PL method is independent of the system size, indicating that our method can show a higher speed than the TB or DFTB method.

Moreover, Fig. 5b shows the accuracy of our method for various binary alloy systems composed of 5 transition metals (Cu, Ni, Ru, Pd, Pt), and found that the pattern similarity is as high as 91~98%. Our method outperforms DFT calculations in terms of the calculation speed, and it loses little information compared to the DFT electronic structures. This clearly reveals that our method will be an alternative to break the trade-off relationship between accuracy and speed, which is well known in the field of electronic structure calculations. And, compared to TB or DFTB method, our PL method has another competence. Such TB approaches basically need a number of training DFT data to determine many TB parameters. For example, to accurately calculate the electronic structure of bulk Rh, 29 TB parameters fitted by ~7,400 training data were required[36]. However, our PL method used the limited training data. As an example, the DOS (96% similarity) of an arbitral Cu-Ni binary alloy can be obtained with only 5 training data (Supplementary Fig. S8).

One of the novelties of this work is that the electronic DOS that was originally a function of an energy level can be expressed with a simple model in a linear combination form of few (three or four) PC bases. Here, we highlight the use of only three or four PCs. Although the available number of PCs in a learning process of DOS patterns is as many as the dimension of the DOS image vectors, the number of PCs highly contributing to the representation of the DOS is few, where the contribution can be evaluated with the eigenvalue for each PC. Thus, four PCs are enough to recognize the diversity of the DOS patterns in the training data. For example, in the binary alloy systems of Fig. 2 where five training data are considered, the contribution of each PC to the representation of the DOS pattern is 32.9% for the 1st PC, 25.7% for the 2nd PC, 22.2% for the 3rd PC, and 19.1% for the 4th PC, which

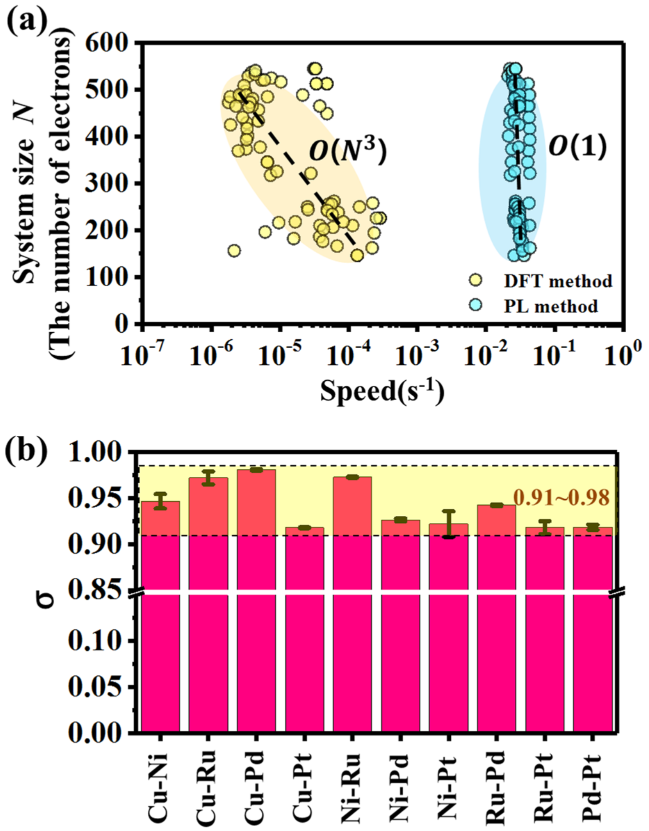

Figure 5. Performances of the PL method compared with the DFT method. (**a**) Comparison of the calculation speeds of the learning method (cyan) and DFT (yellow) as a function of the number of electrons in the alloy systems. The learning method scales as O(1), indicating no dependence on the system size, whereas the DFT scales as O(N^3). The calculation times for 1 core of CPU time and 80 alloy systems were considered. (**b**) The pattern similarity (σ) of the learning method for 10 test alloys in various binary alloy systems: Cu-Ni, Cu-Ru, Cu-Pd, Cu-Pt, Ni-Ru, Ni-Pd, Ni-Pt, Ru-Pd, Ru-Pt, and Pd-Pt. The yellow region is highlighted to show the ρ range of our learning model (91~98%).

indicates that the contribution of the remaining PCs is very miniscule (less than 0.1%). This clearly shows that only four PCs in the PCA are sufficient to represent DOS patterns.

The performance (calculation speed and accuracy) of our PL method is affected by the number of PCs and the grid size. Interestingly, due to the overfitting problem[37], the use of three PCs provides the most accurate DOS patterns in Cu-Ni alloy systems (Supplementary Fig. S14), although at least four PCs are required to fully represent the DOS patterns in training data as discussed in the above paragraph. The grid size also affects the pattern similarity (accuracy) and calculation time of our method (Supplementary Fig. S14). The use of a higher (or finer) grid size provides a higher accuracy, even though an improvement in the pattern similarity for a grid with a higher density than a 100 × 100 grid ($M = 100$) is not significant. However, for $M > 100$, the calculation time is significantly high. Thus, we should employ appropriate values with respect to the number of PCs and grid size to guarantee high pattern similarity (>90%) and low calculation time (<1 min).

To our knowledge, in this work, we presented the first machine-learning approach for calculating electronic DOS patterns (both of value and shape) with a strong accuracy and a fast speed. Moreover, our approach can handle a variety of spectrum image data of materials (X-ray photoelectron spectroscopy, X-ray diffraction, Raman spectrum, etc.). Toward an era of data-driven material design, the importance of material databases will continue to increase; however, the accumulation of data will be a serious bottleneck. In this regard, the fast generation of material databases will be a key in the future. Application of our PCA-based method into various image-type data will provide rapid and accurate prediction of various material properties in place of DFT calculations or other experimental measurements. Therefore, it is anticipated that our model will accelerate the construction of large-scale material databases as well as the design of materials in various fields such as catalysts and electronic devices[38,39].

Methods

Data selection.

To represent the DOS pattern for a test alloy system with the learning model developed in this work, the relevant training data are required. The data include alloy compositions, crystal structures, and DOS patterns. In general, the more training data would provide a more accurate representation. However, since the main purpose of this work is to introduce a new scheme for obtaining DOS patterns by a machine-learning method, we used a limited data set. In a binary *A-B* system, we considered five data sets (two pure structures and three alloy ones). For the pure cases, we used the experimental crystal structure. For the alloy systems, the $A_{0.25}B_{0.75}$, $A_{0.5}B_{0.5}$, and $A_{0.75}B_{0.25}$ compositions were considered. Here, based on the thermodynamic phase diagram of the alloy system, the crystal structures of the three compositions were determined. If there exists an intermetallic phase at the alloy composition, we preferentially considered an intermetallic crystal structure (e.g., $L1_0$ for $Pt_{0.5}Ni_{0.5}$, and $L1_2$ for $Pt_{0.25}Ni_{0.75}$ and $Pt_{0.75}Ni_{0.25}$). On the other hand, for the cases where the intermetallic phase does not exist, atomically randomly mixed structures (i.e., solid-solution phases) were considered with two crystal structures of pure A and B. Among these two structures, we selected the more stable structure as determined by the DFT calculations. The compositions and atomic structures considered in bulk and slab structures study as training and test data are described in Supplementary Figs S3 and S4, respectively. Then, in slab structures study, the compositions of surface layers are considered as same as A_xB_{1-x}. The DOS patterns of the training structures were also obtained from the DFT calculations.

DFT calculation of electronic structures. All electronic structure calculations were performed using the Vienna *ab initio* simulation package (VASP)[40,41]. The exchange-correlation energy was described by the revised Perdew-Burke-Ernzerhof (RPBE) exchange functional[42,43]. The electronic wave functions were expanded in the plane-wave basis set with a kinetic energy cutoff of 520 eV. The effect of the core electrons was modeled by projector augmented-wave (PAW) potentials[44]. The Brillouin zone was sampled using a Monkhorst-Pack *k*-point mesh, and the *k*-point sampling was set to 8 × 8 × 8 for bulk structures and 4 × 4 × 1 for slab structures. The bulk crystal structures were modeled using a 2 × 2 × 2 supercell (e.g., fcc: 32 atoms and bcc: 16 atoms), and the slab crystal structures were simulated periodically with four layer cells. In slab structures calculations, a large vacuum spacing >15 Å was used to prevent inter-slab interactions, and the top most surface layer and sub-surface layer of the computational cell were geometrically relaxed such that the maximum force on each atom was less than 0.05 eV $Å^{-1}$. Their DOS patterns were obtained after a geometry optimization process. We focused on the local DOS of the *d* orbitals in metals for simplicity, and every DOS normalized by the number of atoms in a periodic system was described as $DOS = f(E - E_f)$, where $E - E_f$ is the relative energy shift from the Fermi level (E_f). In addition, during the DFT calculations, we turned on the spin polarization effect to consider the magnetic properties of the metals. In representing the DOS patterns via our learning model, we applied our model separately for the up spin and the down spin.

Pattern similarity calculation. The pattern similarity of our learning model was calculated through a comparison with the DFT results, in which the l^2-norm was used. The pattern similarity σ is defined as follows:

$$\sigma = 1 - \frac{\sqrt{\sum_{m=1}^{M}|\rho'(E_m) - \rho(E_m)|^2}}{\sqrt{\sum_{m=1}^{M}|\rho(E_m)|^2}}$$

where ρ′ and ρ are the DOS patterns obtained by our learning method and calculated by the DFT method, respectively. When σ is closer to 1, our method becomes more accurate.

Data Availability

The data that support the plots within this paper and other findings of this study are available from the corresponding author on request.

References

1. Martin, R. M. Electronic Structure: Basic Theory and Practical Methods (Cambridge Univ. Press, 2004).
2. Nørskov, J. K., Bligaard, T., Rossmeisl, J. & Christensen, C. H. Towards the computational design of solid catalysts. *Nat. Chem.* **1**, 37–46 (2009).
3. Seo, D., Shin, H., Kang, K., Kim, H. & Han, S. S. First-principles design of hydrogen dissociation catalysts based on isoelectronic metal solid solutions. *J. Phys. Chem. Lett.* **5**, 1819–1824 (2014).
4. Ma, X., Li, Z., Achenie, L. E. K. & Xin, H. Machine-learning-augmented chemisorption model for CO_2 electroreduction catalyst screening. *J. Phys. Chem. Lett.* **6**, 3528–3533 (2015).
5. Ratcliff, L. E. *et al.* Challenges in large scale quantum mechanical calculations. *WIREs Comput Mol Sci* **7**, 1–24 (2017).
6. Galli, G. Quantum molecular dynamics simulations. *Curr. Opin. Solid State Mater. Sci.* **1**, 864–874 (1996).
7. Saad, Y., Chelikowsky, J. R. & Shontz, S. M. Numerical methods for electronic structure calculations of materials. *SIAM Rev.* **52**, 3–54 (2010).
8. Goedecker, S. Linear scaling electronic structure methods. *Rev. Mod. Phys.* **71**, 1085–1123 (1999).
9. Ordej, P. Order-*N* tight-binding methods for electronic-structure and molecular dynamics. *Comput. Mater. Sci.* **12**, 157–191 (1998).
10. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
11. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
12. Carleo, G. & Troyer, M. Solving the quantum many-body problem with artificial neural networks. *Science* **606**, 602–606 (2017).
13. Biamonte, J. *et al.* Quantum machine learning. *Nature* **549**, 195–202 (2017).
14. Carrasquilla, J. & Melko, R. G. Machine learning phases of matter. *Nat. Phys.* **13**, 431–434 (2017).
15. Nieuwenburg, E. P. L. V., Liu, Y. & Huber, S. D. Learning phase transitions by confusion. *Nat. Phys.* **13**, 435–440 (2017).
16. Snyder, J. C., Rupp, M., Hansen, K., Mu, K. & Burke, K. Finding density functionals with machine learning. *Phys. Rev. Lett.* **108**, 253002 (2012).
17. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: Critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
18. Brockherde, F. *et al.* Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **8**, 872 (2017).
19. Arsenault, L., Lopez-bezanilla, A. & Millis, A. J. Machine learning for many-body physics: The case of the Anderson impurity model. *Phys. Rev. B* **90**, 155136 (2014).
20. Schütt, K. T., Glawe, H., Brockherde, F., Sanna, A. & Gross, E. K. U. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 205118 (2014).

21. Takigawa, I., Shimizu, K., Tsuda, K. & Takakusagi, S. Machine-learning prediction of the d-band center for metals and bimetals. *RSC Adv.* **6**, 52587–52595 (2016).
22. Mueller, T., Kusne, A. G. & Ramprasad, R. Machine learning in materials science: Recent progress and emerging applications. *Rev. in Comput. Chem.* **29**, 186–273 (2016).
23. Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning 534–552 (Springer, 2009).
24. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nat. Meth.* **14**, 263–266 (2017).
25. Soven, P. Coherent-potential model of substitutional disordered alloys. *Phys. Rev.* **156**, 809–813 (1967).
26. Gyorffy, B. L. Coherent-potential approximation for a nonoverlapping-muffin-tin potential model of random substitutional alloys. *Phys. Rev. B* **5**, 2382–2384 (1972).
27. Tian, F., Varga, L. K., Chen, N., Delczerg, L. & Vitos, L. *Ab initio* investigation of high-entropy alloys of 3d elements. *Phys. Rev. B* **87**, 075144 (2013).
28. Tian, F., Varga, L. K., Chen, N., Shen, J. & Vitos, L. *Ab initio* design of elastically isotropic TiZrNbMoV$_x$ high-entropy alloys. *J. Alloys Compd.* **599**, 19–25 (2014).
29. Peil, O. E., Ruban, A. V. & Johansson, B. Self-consistent supercell approach to alloys with local environment effects. *Phys. Rev. B* **85**, 165140 (2012).
30. Whitfield, J. D., Love, P. J. & Aspure-Guzik, A. Computational complexity in electronic structure. *Phys. Chem. Chem. Phys.* **15**, 397–411 (2013).
31. Cleri, F. & Rosato, V. Tight-binding potentials for transition metals and alloys. *Phys. Rev. B* **48**, 22–33 (1993).
32. Usman, M., Broderick, C. A., Lindsay, A. & O'Reilly, E. P. Tight-binding analysis of the electronic structure of dilute bismide alloys of GaP and GaAs. *Phys. Rev. B* **84**, 245202 (2011).
33. Mukherjee, S., Morán-López, J. L., Kumar, V. & Bennemann, K. H. Electronic theory for surface segregation in Cu$_x$Ni$_{1-x}$ alloy. *Phys. Rev. B* **25**, 730–737 (1982).
34. Wahiduzzaman, M. *et al.* DFTB parameters for the periodic table: Part 1, electronic structure. *J. Chem. Theory Comput.* **9**, 4006–4017 (2013).
35. Hams, A. & Raedt, H. D. Fast algorithm for finding the eigenvalue distribution of very large matrices. *Phys. Rev. E* **62**, 4365–4377 (2000).
36. Barreteau, C. & Spanjaard, D. Electronic structure and total energy of transition metals from an spd tight-binding method: Application to surfaces and clusters of Rh. *Phys. Rev. B* **58**, 9721–9731 (1998).
37. Dominggos, P. A few useful things to know about machine learning. *Communications of the ACM* **55**, 78–87 (2012).
38. Kolb, B., Lentz, L. C. & Kolpak, A. M. Discovering charge density functionals and structure-property relationships with PROPhet: A general framework for coupling machine learning and first- principles methods. *Sci. Rep.* **7**, 1–9 (2017).
39. Hill, J. *et al.* Materials science with large-scale data and informatics: Unlocking new opportunities. *MRS Bulletin* **41**, 399–409 (2017).
40. Kresse, G. & Joubert, D. From Ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
41. Kresse, G. & Furthmiiller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
42. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **3**, 3865–3868 (1996).
43. Hammer, B., Hansen, L. B. & Nørskov, J. K. Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals. *Phys. Rev. B* **59**, 7413–7421 (1999).
44. Blochl, P. E. Projector augmented-wave. *Phys. Rev. B* **50**, 17953–17979 (1994).

Acknowledgements

This work was supported by Creative Materials Discovery Program through the National Research Foundation of Korea (NRF-2016M3D1A1021140). We acknowledge the financial supports of the Korea Institute of Science and Technology (Grant No. 2E28000). This work was supported by Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-MA1801-03.

Author Contributions

B.C.Y. and S.S.H. conceived and designed the research. D.K. and C.K. provided theoretical support. B.C.Y. performed the research. B.C.Y., D.K., C.K., and S.S.H. analyzed the data. B.C.Y. and S.S.H. wrote the manuscript with feedback from all authors. S.S.H. managed the project.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-42277-9.

Competing Interests: The authors declare no competing interests.