JAIST Repository
https://dspace.jaist.ac.jp/

Title	Novel mixture model for the representation of potential energy surfaces
Author(s)	Pham, Tien Lam; Kino, Hiori; Terakura, Kiyoyuki; Miyake, Takashi; Dam, Hieu Chi
Citation	The Journal of Chemical Physics, 145(15): 154103-1-154103-6
Issue Date	2016-10-17
Type	Journal Article
Text version	publisher
URL	http://hdl.handle.net/10119/14785
Rights	Copyright 2016 American Institute of Physics. This article may be downloaded for personal use only. Any other use requires prior permission of the author and the American Institute of Physics. The following article appeared in Tien Lam Pham, Hiori Kino, Kiyoyuki Terakura, Takashi Miyake and Hieu Chi Dam, The Journal of Chemical Physics, 145(15), 154103 (2016) and may be found at http://dx.doi.org/10.1063/1.4964318
Description	

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

AIP | The Journal of Chemical Physics

Novel mixture model for the representation of potential energy surfaces

Tien Lam Pham, Hiori Kino, Kiyoyuki Terakura, Takashi Miyake, and Hieu Chi Dam

Citation: The Journal of Chemical Physics **145**, 154103 (2016); doi: 10.1063/1.4964318
View online: http://dx.doi.org/10.1063/1.4964318
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/145/15?ver=pdfcov
Published by the AIP Publishing

Articles you may be interested in

A compact and accurate semi-global potential energy surface for malonaldehyde from constrained least squares regression
J. Chem. Phys. **141**, 144310 (2014); 10.1063/1.4897486

An exchange-Coulomb model potential energy surface for the Ne–CO interaction. II. Molecular beam scattering and bulk gas phenomena in Ne–CO mixtures
J. Chem. Phys. **132**, 024308 (2010); 10.1063/1.3285721

Reproducing kernel Hilbert space interpolation methods as a paradigm of high dimensional model representations: Application to multidimensional potential energy surface construction
J. Chem. Phys. **119**, 6433 (2003); 10.1063/1.1603219

Global potential energy surfaces for the H 3 + system. Analytical representation of the adiabatic ground-state 1 1 A ′ potential
J. Chem. Phys. **112**, 1240 (2000); 10.1063/1.480539

An analytical representation of the ground potential energy surface (2 A ′) of the H+Cl 2 →HCl+Cl and Cl+HCl→HCl+Cl reactions, based on ab initio calculations
J. Chem. Phys. **108**, 3168 (1998); 10.1063/1.475713

Reuse of AIP Publishing content is subject to the terms: https://publishing.aip.org/authors/rights-and-permissions.

Novel mixture model for the representation of potential energy surfaces

Tien Lam Pham,[1,2,3] Hiori Kino,[3,4] Kiyoyuki Terakura,[2,4] Takashi Miyake,[3,4,5] and Hieu Chi Dam[1,2,4,6]

[1]*Institute for Solid State Physics, The University of Tokyo, Kashiwanoha 5-1-5, Kashiwa, Chiba 277-8581, Japan*
[2]*Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan*
[3]*ESICMM, National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan*
[4]*Center for Materials Research by Information Integration, National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan*
[5]*CD-FMat, AIST, 1-1-1 Umezono, Tsukuba 305-8568, Japan*
[6]*JST, PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan*

(Received 21 June 2016; accepted 22 September 2016; published online 17 October 2016)

We demonstrate that knowledge of chemical physics on a materials system can be automatically extracted from first-principles calculations using a data mining technique; this information can then be utilized to construct a simple empirical atomic potential model. By using unsupervised learning of the generative Gaussian mixture model, physically meaningful patterns of atomic local chemical environments can be detected automatically. Based on the obtained information regarding these atomic patterns, we propose a chemical-structure-dependent linear mixture model for estimating the atomic potential energy. Our experiments show that the proposed mixture model significantly improves the accuracy of the prediction of the potential energy surface for complex systems that possess a large diversity in their local structures. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4964318]

INTRODUCTION

Computational materials science encompasses a range of methods that are used to model materials and simulate their responses at different length and time scales. Among the many problems that are tackled by computational materials science, the development of methods for computing the potential energy surfaces (PESs), from which atomic forces are also obtained, with low computational cost is of primary importance. There are two main approaches in dealing with this issue: the first-principles approach and the empirical approach. The first-principles approach, which is based on the explicit modeling of electrons, is associated with high accuracy. However, the application of this approach is limited to small system sizes (i.e., typically a few hundred atoms), depending on the level of approximation and available computing power. On the other hand, in the empirical approach, computation methods (e.g., atomic-scale methods, including semi-empirical tight-binding and empirical force fields or reactive potential) are developed with less transferability, but with high efficiency. One of the major principles of this approach is to develop simple models to meet the unique characteristics of materials of interest. Therefore, these methods can yield not only impressive fidelity/accuracy by their inherent approximation but can also yield applicability to systems up to tens of billions of atoms.

However, the empirical approach depends strongly on prior knowledge concerning how to identify groups of materials with similar characteristics, which derives from the manual extraction of behavioral patterns of materials by researchers. The increasing volume of available experimental and quantum computational materials databases together with the development of machine learning techniques provides an opportunity to integrate the two approaches. According to this approach, prior knowledge of hidden patterns can be automatically extracted from both first-principles-calculated data and experimental data[1–5] by using machine learning techniques; such information can then be utilized to construct a simple empirical potential model.

Recently, methods wherein the PES of a system is derived from a large set of quantum calculation energies have been developed. According to these methods, information concerning local atomic structures is represented by using various types of descriptors and is then mapped to potential energy using a unique predictive modeling technique (i.e., supervised learning). Various methods have been applied to model this mapping using linear regression,[6] kernel ridge regression,[7–9] Gaussian process regression,[10,11] and neural networks.[12–21] Improvement of the original kernel regression approach using the predefined reference centers in atomic descriptor space, which is defined by K-mean algorithm, is also proposed.[22] Although significant research efforts have been directed towards the fast and accurate estimation of the PES, the performance is still limited for systems with a large diversity in local structure—e.g., multi-component systems.

In this study, we propose a novel method for computing the PES of multi-component systems by integrating together generative modeling and predictive modeling. Crystalline Si with (100) surface orientation and amorphous silicon hydride (a-Si:H) systems are used in demonstrating the validity of our approach. More specifically, atom-distribution-based descriptors are used to represent an atom in a certain local

Reuse of AIP Publishing content is subject to the terms: https://publishing.aip.org/authors/rights-and-permissions.

chemical environment.[23] A Gaussian mixture model (GMM) is then applied to identify groups and calculate the probability of an atom belonging to a particular identified group. With this prior knowledge, we show that the PES of crystalline Si with surfaces and the PES of a-Si:H systems, which contain diverse local structures, can be accurately predicted in the framework of a linear model (LM) by embedding the information about the groups of the local structures learnt from data.

DATA PREPARATION

Raw data preparation

Si crystal with (100) surface orientation was used in the first demonstration. This system was simulated by a repeated slab of 12 atomic layers and a 4×4 super-cell. Molecular dynamics (MD) calculations were performed using CPMD[24,25] in NVT ensembles at 1500 K. For a better evaluation on the rationality of the learned patterns of atomic local structures, we performed the simulation at the temperature below the melting point of Si for maintaining the surface states of the slab. We employed the PBE exchange-correlation[26] functional to approximate the exchange-correlation energy, plane wave basis set with a cutoff energy of 40 Ry, and ultra-soft pseudo-potential[27] for treating the interaction between valence electrons and core electrons. Eight thousand structures were extracted from these MD trajectories and were used as the database. The MD simulation with the same electronic calculation settings was applied to build the database for the a-Si:H system. A supercell with 120 Si atoms and 9 H atoms was used for representing a-Si:H. The amorphous structure was generated by quenching from the melt state, which is simulated at 2200 K. This supercell was equilibrated in an NPT ensemble at 300, 500, and 1500 K. Three thousand structures extracted from MD trajectories were used as the database. The total energies of these structures were recalculated using PWSCF code[28] with the PBE exchange-correlation function, an ultra-soft pseudo-potential, and a plane wave basis set with a cutoff radius of 40 Ry.

Data representation for learning process

The underlying hypothesis of this study is that the total energy E_T of a system can be calculated by the summation of the effective atomic energies of the constituent atoms:[6,10,12] $E_T = \sum_{i=1}^{N} E_i$, where E_i is the contribution of an atom with index i and N is the number of constituent atoms of the system. Further, the effective atomic energy E_i of atom i can be estimated based on information regarding its surrounding local chemical environment. The local chemical environment surrounding atom i is represented by using descriptor vector $\vec{x}_i$ of which components are calculated from the functional of distributions of the two-body central term (r_{ij}) and the three-body non-central terms $(\theta_{jik}, r_{ij}, r_{ik})$, where r_{ij} is the distance between atom i and atom j, and θ_{jik} the bond angle between the j,i and i,k bonds. We can extend the representation with descriptors to higher-order terms, such as four-body and five-body terms. In this study, we truncate the representation $\vec{x}_i$ up to three-body terms. The actual functional forms, basis functions, and parameters are provided in the supplementary material (Eqs. (S2)-(S7)). In this context, we should note that in the present work we aim at predicting the atomic energy of each atom by learning from the supervised-data with the explicitly calculated total energy of the system, which is equal to the sum of the atomic energies over all constituent atoms.

LEARNING THE POTENTIAL ENERGY SURFACE (PES) FROM DATA

Behler *et al.* applied identical models to identical atomic species even if their local environments are quite different.[12–18] This required the use of a complex neural network to represent atomic energies. For complex systems, such as crystalline Si with surfaces and a-Si:H, which exhibit a large diversity in local structure and bonding nature (not only three types of *sp* hybridization but also the ionic bonding character in a-Si:H), the atomic interaction will be even more complicated. Therefore, a prediction model with higher complexity using a large neural network and hence, a large number of training data are needed, and the complexity in the training process will dramatically increase. As a result, it is difficult to carry out a statistical selection for the most appropriate learning model by performing regularization and cross-validation. In addition, the neural network functions as a black box that maps input to output values, and the atomic potential energy is learnt indirectly from a decomposition process of the total energy. Therefore, it is nearly impossible to interpret the physical meaning of the learnt results.

In the present work, we propose a new approach that is complimentary to the neural network approach. The basic idea of the present approach is that different atomic species, as well as the same atomic species embedded in different local chemical environments, are treated differently. For example, in crystalline Si with surfaces or in a-Si:H, not only different models used for Si and H but the models for predicting the energies of Si atoms (or H atoms) should also be different for different local chemical environments (such as Si in bulk versus Si at the surface). In our method, the prior knowledge concerning patterns of the local chemical environments is learnt automatically from data by using unsupervised machine learning. The advantage of using this approach is that simple prediction models (such as linear models) can be adopted for predicting the energies of atoms in rather complicated systems, wherein a separate simple prediction model is used for atoms embedded in each of the different patterns of the local chemical environments. Further, if the linear models can be adopted, owing to the linearity of the mapping from the descriptors (of the local chemical environment) space to the atomic energy space, the representation of the structure of the entire material system can be represented in the same manner (namely, by a vector), which is the sum of all descriptor vectors of the constituent atoms. This is an important by-product of this approach as explicitly shown in Eq. (9); it enables us to learn, directly and explicitly, the atomic potential energy from the total energy data.

To implement this approach, we first employ GMM[29] for learning the patterns of atomic local chemical environments

Reuse of AIP Publishing content is subject to the terms: https://publishing.aip.org/authors/rights-and-permissions.

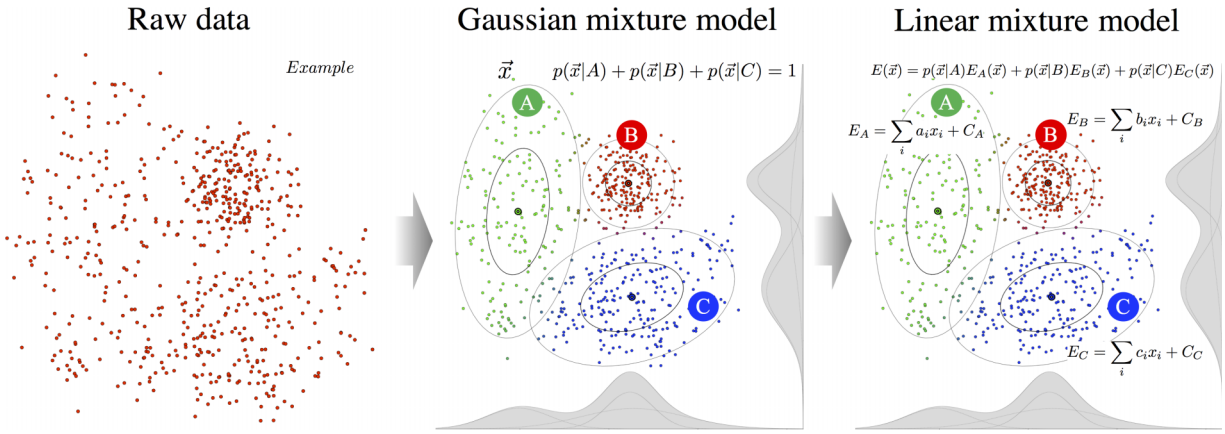

FIG. 1. Linear mixture model (LMM) for estimating atomic energy based on the clustering of atomic groups.

by clustering constituent atoms into groups so that atoms within a group have similar local chemical environments. The GMM is based on the assumption that the data consist of different groups and that the data in each group follow their own Gaussian distribution. In other words, in GMM, the distribution of data is fitted to a combination of a certain number M of Gaussian functions (M: number of data groups).[29] The probability distribution of an atom with index i having a representation of $\vec{\boldsymbol{x}}_i$, $f(\vec{\boldsymbol{x}}_i)$, can be approximated by

$$f(\vec{\boldsymbol{x}}_i) = \sum_{m=1}^{M} \alpha_m \Phi(\vec{\boldsymbol{x}}_i; \vec{\boldsymbol{\mu}}_m, \boldsymbol{\Sigma}_m), \tag{1}$$

where

$$\Phi(\vec{\boldsymbol{x}}; \vec{\boldsymbol{\mu}}_m, \boldsymbol{\Sigma}_m) = \frac{\exp\left[-(\vec{\boldsymbol{x}} - \vec{\boldsymbol{\mu}}_m)^T \boldsymbol{\Sigma}_m^{-1} (\vec{\boldsymbol{x}} - \vec{\boldsymbol{\mu}}_m)\right]}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_m|^{\frac{1}{2}}} \tag{2}$$

is a multivariate Gaussian distribution with mean $\vec{\boldsymbol{\mu}}_m$ and covariance matrix $\boldsymbol{\Sigma}_m$, and d is the dimension of the representation vector $\vec{\boldsymbol{x}}_i$. The coefficients α_m are the weights that satisfy the following constraint:

$$\sum_{m=1}^{M} \alpha_m = 1. \tag{3}$$

The probability that $\vec{\boldsymbol{x}}_i$ belongs to group m can be represented as follows:

$$p(\vec{\boldsymbol{x}}_i|m) = \frac{\alpha_m \Phi(\vec{\boldsymbol{x}}_i; \vec{\boldsymbol{\mu}}_m, \boldsymbol{\Sigma}_m)}{\sum_{m=1}^{M} \alpha_m \Phi(\vec{\boldsymbol{x}}_i; \vec{\boldsymbol{\mu}}_m, \boldsymbol{\Sigma}_m)}. \tag{4}$$

The model parameters, $\{\alpha_m, \vec{\boldsymbol{\mu}}_m, \boldsymbol{\Sigma}_m\}$, are determined by an expectation-maximization algorithm.[30] The number of data groups, M, is determined from the data by using information criteria such as the Akaike information criterion[31] or the Bayesian information criterion.[32] It is interesting to note that the GMM provides a "probabilistic image" of the pattern of the atomic local chemical environment, wherein instead of assigning an atom to a specific group, it provides the probability for an atom to stay in a group. The sum of probabilities for an atom to stay in either of the groups is 1.

The GMM is therefore expected to be discovered from the data distinctive patterns of atoms in local structures/chemical environments and to calculate the probability that an atom belongs to a group. These discovered groups are defined as the types of atoms; we can apply a different linear representation model to determine the atomic energy of each different type of atom. Figure 1 shows a schematic procedure of our approach as a demonstration for randomly generated data in the space spanned by two descriptors. In this case, the GMM discovers three types of atoms in the system: group A (green), group B (red), and group C (blue).

If an atom with a descriptor vector $\vec{\boldsymbol{x}}$ belongs exclusively to a specific group m, its atomic energy is expressed using a linear model as follows:

$$E(\vec{\boldsymbol{x}}) = \vec{\boldsymbol{c}}_m \cdot \vec{\boldsymbol{x}} + C_m, \tag{5}$$

where m is the index of the groups (A, B, or C), and $\vec{\boldsymbol{c}}_m$ and C_m are the slope vector and the intercept of the linear model, respectively. In the systems considered here, for an atom with index i, which is represented by a descriptor vector $\vec{\boldsymbol{x}}_i$, GMM can yield probabilities $w_i^m = p(\vec{\boldsymbol{x}}_i|m)$ with which the atom i belongs to group m. The atomic energy of this atom can be expressed as follows:

$$E_i = E(\vec{\boldsymbol{x}}_i) = \sum_{m=1}^{M} w_i^m (\vec{\boldsymbol{c}}_m \cdot \vec{\boldsymbol{x}}_i + C_m). \tag{6}$$

To include the prior knowledge learnt by GMM, the atom i can therefore be represented by an extended descriptor vector,

$$\vec{\boldsymbol{X}}_i = \left(w_i^A \vec{\boldsymbol{x}}_i, w_i^A, w_i^B \vec{\boldsymbol{x}}_i, w_i^B, w_i^C \vec{\boldsymbol{x}}_i, w_i^C\right). \tag{7}$$

The atomic energy with the above extended descriptor vector is formally expressed as

$$E(\vec{\boldsymbol{X}}_i) = \vec{\boldsymbol{\gamma}} \cdot \vec{\boldsymbol{X}}_i, \tag{8}$$

where $\vec{\boldsymbol{\gamma}}$ represents the expansion coefficient vector, $\{\vec{\boldsymbol{c}}_A, C_A, \vec{\boldsymbol{c}}_B, C_B, \vec{\boldsymbol{c}}_C, C_C\}$, which is to be learnt. At this stage, the total energy E_T of the system is given by

$$E_T = \sum_i E(\vec{\boldsymbol{X}}_i) = \vec{\boldsymbol{\gamma}} \cdot \sum_i \vec{\boldsymbol{X}}_i. \tag{9}$$

It is obvious that $\sum_i \vec{\boldsymbol{X}}_i$ can be considered to be the representation of the entire material system; the atomic energy

Reuse of AIP Publishing content is subject to the terms: https://publishing.aip.org/authors/rights-and-permissions.

and the total energy of the system can be predicted by using the same linear model. Consequently, the learning process is now a standard linear regression, and we can employ regularization techniques for carrying out the model selection process to validate and improve the predictability of the model properly. In this work, we employ L2 regularization (known as ridge regression) in which the sum of squared errors is minimized with an additional penalty term: $\lambda\|\vec{\gamma}\|_2^2$. The regularization parameter, λ, is determined by a cross-validation search in 1D logarithmic grid of 100 points. It is important to note that although this approach (hereafter referred to as the linear mixture model (LMM)) is reduced to a standard linear regression, the dimension of the vector $\vec{\boldsymbol{X}}_i$ is M times larger than that of $\vec{\boldsymbol{x}}_i$ for the simple linear regression (hereafter referred to as LM) and the information concerning the patterns of local chemical structure is already taken into account in the weights, w_i^m, which may absorb non-linear effects in the regression. Furthermore, the proposed linear model benefits a great advantage in computational cost for the prediction process; the sparse modeling techniques (such as the LASSO with L1 regularization[29]) can also be easily employed to reduce the dimensionality of the descriptor space and to discover which features of the local chemical environment essentially contribute to the atomic energy.

RESULTS AND DISCUSSION

We first apply the GMM to identify the pattern of local structures of silicon atoms in crystalline Si with (100) surface (i.e., slab Si).

For the GMM analysis, we employ 100-dimensional vectors (as defined by Eq. (S11) in the supplementary material) to represent the atomic information of an atom in its chemical environment. However, as strong correlation exists between components in the vectors, principal component analysis (PCA) is applied to reduce the dimension of the descriptor vectors, which lead to GMM analysis with better performance.

Figure 2(a) depicts the density of Si atoms of the Si slab in the principal component space. Based on the evaluation using statistical information criteria,[31,32] Si atoms can be classified into four groups. The distribution of the data is fitted to a combination of four Gaussian distributions (GMM) to obtain the probabilities that an Si atom belongs to group A, group B, group C, and group D, which are given by $w_a = p(\mathrm{Si}\,|A)$, $w_b = p(\mathrm{Si}\,|B)$, $w_c = p(\mathrm{Si}\,|C)$, and $w_d = p(\mathrm{Si}\,|D)$, respectively. In order to understand the physical meaning of the GMM results, Si atoms are assigned to the group with the highest probability, and the geometrical structures of the four Si types are shown in Fig. 2(b). We found that atoms of the first type (type A) are Si atoms in the bulk; atoms of the

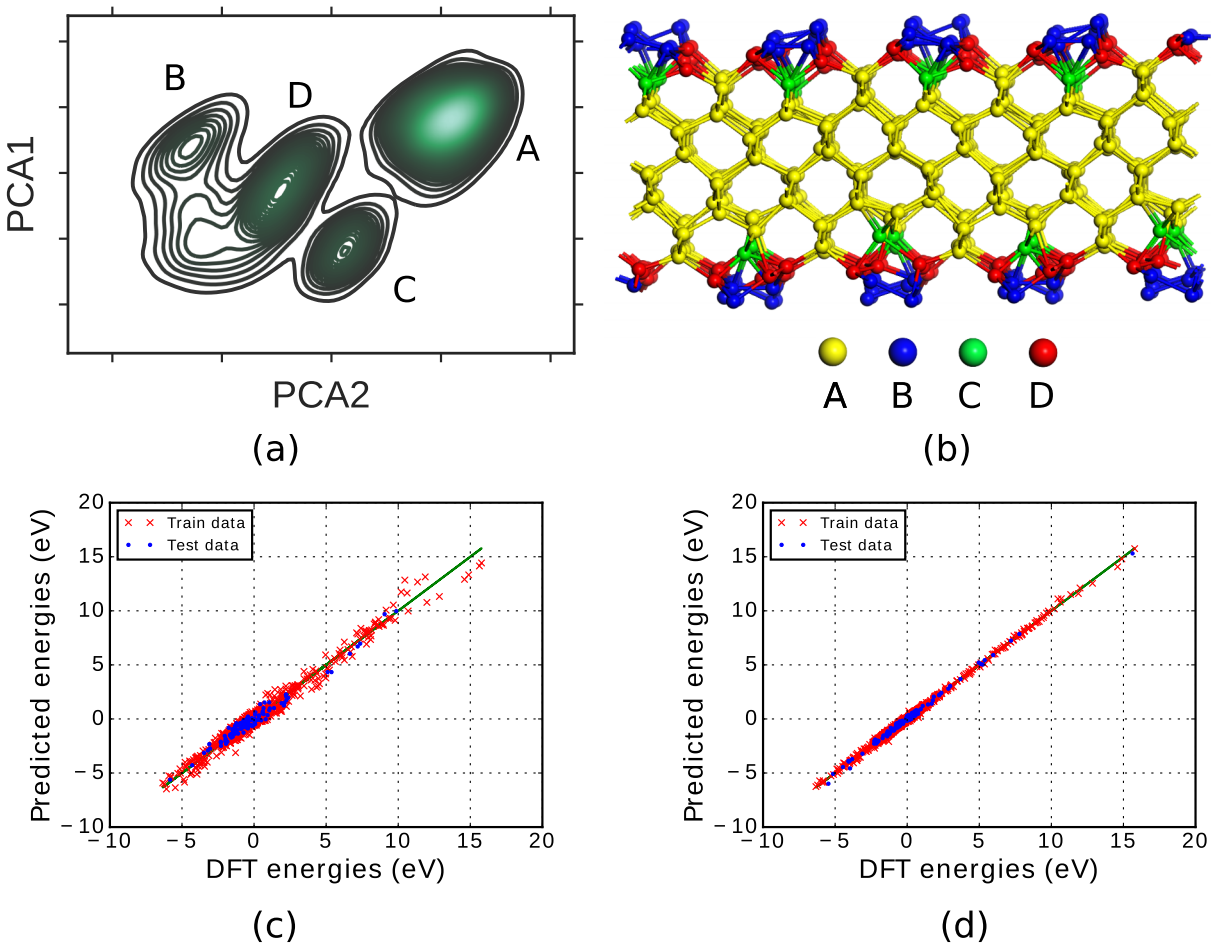

FIG. 2. (a) Density distribution of Si atoms in principal component space. (b) Geometrical structure of four types of Si atoms. (c) Comparison of DFT energies and predicted energies by simple linear model. (d) Linear mixture model (LMM).

Reuse of AIP Publishing content is subject to the terms: https://publishing.aip.org/authors/rights-and-permissions.

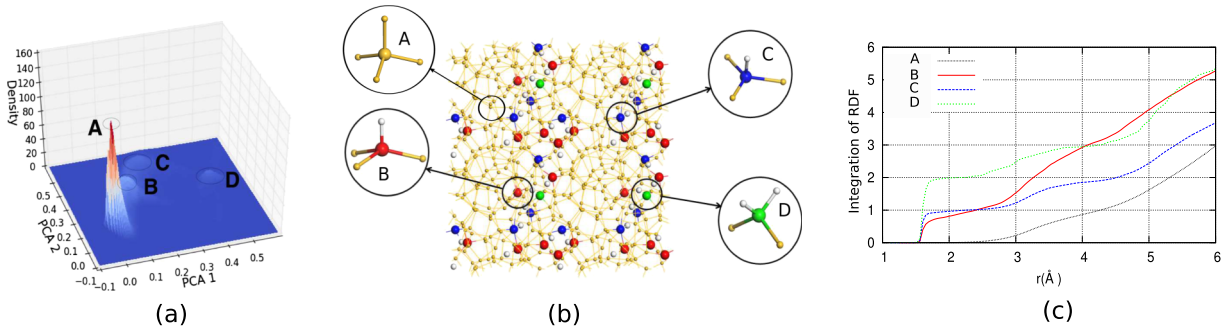

FIG. 3. (a) Density distribution in PCA space. (b) Local structure of four types of Si atoms: Si of the first, second, third, and fourth type are yellow, red, blue, and green, respectively. (c) Integral of radial distribution function (RDF) of the Si–H pair.

second type (type B) are buckled dimer atoms; atoms of the third type (type C) are the third layer atoms, which lie under buckled dimer atoms; atoms of the fourth type (type D) are Si atoms of the second layer from the surface. The obtained grouping is obviously in good agreement with current knowledge regarding the structure of bare Si surfaces. This result confirms the reliability of the prior knowledge of the atomic local chemical environment automatically learnt by GMM.

In the second experiment, the GMM was used to identify the atomic pattern in the a-Si:H system. The 100-dimensional descriptor vectors of Eq. (S11) in the supplementary material were also used to represent the local atomic information for both Si and H atoms. This representation was also fed into the PCA algorithm to extract the principal components. We only focus on Si atoms in the present work. Figure 3(a) shows the density distribution of the data in the reduced-dimensional space. The density distribution shows four separate peaks: one high peak (A), two medium peaks (B and C), and one small peak (D). Peaks A and D are well separated from others, while peaks B and C overlap to some extent. This observation also agrees with the evaluation using statistical information criteria.[31,32] By fitting the data for Si atoms to the GMM with four Gaussian distributions and calculating the probabilities of a Si atom belonging to each group, we are able to label the Si atoms.

A visualization of the local geometrical structures shows that the first type (type A, colored yellow in Fig. 3(b)) do not directly bond to any H atoms; Si atom of second type (type B, colored red) and the third type (type C, colored blue) bond to an H atom; Si atoms of the fourth type (type D, colored green) bond to two H atoms. These results are consistent with the fact that the density peaks of B and C are not completely separated. The radial distribution function (RDF) was calculated to confirm this observation. Figure 3(c) shows the RDF of Si–H pair. Within a radius of 2.5 Å, the RDF results are consistent with the above assignment of four-types of Si atoms. The RDF analysis provides more detailed information about the difference between types B and C. In particular, type B has another H atom in its chemical environment at a distance larger than 3.0 Å, whereas type C does not have such an atom in its vicinity. To understand the local environment more clearly, we analyze the histogram of Si and H coordinated around each type of Si atom (see Fig. S2 in the supplementary material). The results imply that type-B Si atoms appear to be more chemically active than type-C Si atoms. Because of this chemical and geometrical difference, different functional forms should be used for different types of Si atoms.

We now demonstrate that with the prior knowledge of the atomic pattern obtained by unsupervised machine learning, the total energies of the above two systems can be predicted with improved accuracy. As discussed already, different types of Si atoms exhibit different geometrical structures. Within each type, a type-dependent linear model (LMM) is applied to predict the atomic energies directly from the calculated total energy data. All the evaluations are carried out by using 10-times 10-fold-cross-validation.

As for the Si slab, the unsupervised GMM learning shows four types of Si atoms (groups A, B, C, and D). We used LMM Eq. (6) to represent the atomic energies of this system. Figures 2(c) and 2(d) illustrate the comparison between the density functional theory (DFT)-calculated total energies and the energies predicted by the simple LM and the LMM of Eq. (9), respectively. It is evident that with the prior knowledge of the atomic patterns in combination with a linear model, more accurate energies can be predicted as compared to only using a simple linear model. Table I shows the root mean square error (RMSE) of the training set and that of the test set for LM and LMM. The best score (R-factor) given by the LM is 0.960, while that of the LMM is approximately 0.999.

For the a-Si:H system, unsupervised GMM shows that there are four types (A, B, C, and D) of Si atoms with different chemical environments. A procedure similar to that for the slab Si was also performed for the a-Si:H system. The LMM was applied for silicon atoms, while LM is used for

TABLE I. RMSE (eV/structure) for training set and test set obtained by linear model (LM) and linear mixture model (LMM) for slab Si with (100) surfaces.

# basis function		100	200	300
LM	Train RMSE	0.587	0.382	0.221
	Test RMSE	0.591	0.339	0.230
LMM	Train RMSE	0.235	0.138	0.073
	Test RMSE	0.245	0.142	0.078

Reuse of AIP Publishing content is subject to the terms: https://publishing.aip.org/authors/rights-and-permissions.

TABLE II. RMSE (eV/structure) for training set and test set obtained by linear model (LM) and linear mixture model (LMM) for the a-Si:H system.

# basis function		100	200	300
LM	Train RMSE	18.3	4.91	1.94
	Test RMSE	19.8	5.90	2.86
LMM	Train RMSE	3.77	0.831	0.414
	Test RMSE	4.45	1.22	0.636

H atoms. It was observed that applying LM to all Si atoms yields the RMSE of 22 meV/atom for the test set, while LMM significantly reduces the RMSE to approximately 5 meV/atom for the test set (Table II).

These examples clearly show that the proposed LMM that utilizes the prior knowledge concerning the patterns of the local chemical environment performs much better than the simple LM without such prior knowledge. Further, it is important to note that the obtained accuracy confirms the linearity of the mapping from the representation space to the atomic energy and the total energy spaces. Consequently, the representation of the entire material system by using the sum of all descriptor vectors of the constituent atoms (Eq. (9)) is adequate. The obtained results verify the validity of our proposed scheme to learn, directly and explicitly, the atomic potential energy from the total energy data. Our experiments show that when an atom is embedded into different local chemical environments, its energy should be calculated by environment-dependent functions. The difficulty lies in the discovery of the patterns of the chemical environments or the identification of the atomic pattern of the system. We have demonstrated that the unsupervised machine learning GMM can be employed to find the atomic patterns of Si atoms in the crystalline Si with surfaces and a-Si:H, as seen in Figs. 2 and 3. The demonstrated integration of unsupervised learning and supervised learning for constructing an empirical potential can therefore be considered as a promising approach.

CONCLUSION

We have demonstrated that hidden knowledge concerning crystalline Si with surfaces and a-Si:H systems, i.e., the type of atoms in the system, can be learnt from the simulation data by using unsupervised machine learning techniques. Based on this prior knowledge, we proposed a novel LMM to learn the atomic energy for the estimation of PESs. In our implementation, the GMM was employed to cluster the atoms in the systems, and a probabilistic image of the atomic pattern was discovered. Adopting this prior knowledge to linear models, we built a local-chemical-structure-dependent mixture model for predicting the atomic energy and estimating the PESs. More specifically, each Si atom can be assigned to a group with a certain probability obtained by GMM, and the atomic energy of atoms in each class can then be represented by a distinct linear representation. Our experiments show that the model with mined prior knowledge can predict the PESs much more accurately than that without prior knowledge. Our approach is therefore expected to offer new opportunities in the automation of learning empirical potential from data with high accuracy, efficiency, and transferability.

SUPPLEMENTARY MATERIAL

See the supplementary material for details of the linear representation of atomic energies and statistical analysis of the GMM results.

ACKNOWLEDGMENTS

This work was partly supported by PRESTO and by Materials research by Information Integration Initiative (MI^2I) project of the Support Program for Starting Up Innovation Hub, both from Japan Science and Technology Agency (JST), Japan.

[1]Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Phys. Rev. B **85**, 104104 (2012).
[2]S. Yang, M. Lach-hab, I. I. Vaisman, and E. Blaisten-Barojas, J. Phys. Chem. C **113**, 21721 (2009).
[3]G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, Chem. Mater. **22**, 3762 (2010).
[4]J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, Phys. Rev. Lett. **108**, 253002 (2012).
[5]O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, and S. Curtarolo, Chem. Mater. **27**, 735 (2015).
[6]A. Seko, A. Takahashi, and I. Tanaka, Phys. Rev. B **90**, 024101 (2014).
[7]V. Botu and R. Ramprasad, Int. J. Quantum Chem. **115**, 1074 (2014).
[8]R. Matthias, Int. J. Quantum Chem. **115**, 1058 (2015).
[9]O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, and A. Knoll, Int. J. Quantum Chem. **115**, 1084 (2015).
[10]A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[11]A. P. Bartk and G. Csnyi, Int. J. Quantum Chem. **115**, 1051 (2015).
[12]J. Behler and M. Parrinello, Phys. Rev. Lett. **6**, 146401 (2007).
[13]H. Eshet, R. Z. Khaliullin, T. D. Kuhne, J. Behler, and M. Parrinello, Phys. Rev. B **81**, 184107 (2010).
[14]H. Eshet, R. Z. Khaliullin, T. D. Kuhne, J. Behler, and M. Parrinello, Phys. Rev. Lett. **108**, 115701 (2012).
[15]N. Artrith and J. Behler, Phys. Rev. B **85**, 045439 (2012).
[16]T. Morawietz and J. Behler, Z. Phys. Chem. **227**, 1559 (2013).
[17]N. Artrith, T. Morawietz, and J. Behler, Phys. Rev. B **83**, 153101 (2011).
[18]R. Z. Khaliullin, H. Eshet, T. D. Kuuhne, J. Behler, and M. Parrinello, Phys. Rev. B **81**, 100103 (2010).
[19]N. Artrith and A. M. Kolpak, Nano Lett. **14**, 2670 (2014).
[20]M. Sergei, D. Richard, and C. Tucker, Int. J. Quantum Chem. **115**, 1012 (2015).
[21]N. Artrith and A. Urban, Comput. Mater. Sci. **114**, 135 (2016).
[22]R. Fournier and S. Orel, J. Chem. Phys. **139**, 234110 (2013).
[23]J. Behler, J. Phys. Chem. **134**, 074106 (2011).
[24]R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).
[25]CPMD, 2011, http://www.cpmd.org/.
[26]J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[27]D. Vanderbilt, Phys. Rev. B **41**, 7892 (1990).
[28]P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, J. Phys.: Condens. Matter **21**, 395502 (2009).
[29]*Machine Learning: A Probabilistic Perpective*, edited by K. P. Murphy (MIT Press, 2012).
[30]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011); available at http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.
[31]H. Akaike, IEEE Trans. Autom. Control **19**, 716 (1974).
[32]B. G. Schwarz, Ann. Stat. **6**, 461 (1978).

Reuse of AIP Publishing content is subject to the terms: https://publishing.aip.org/authors/rights-and-permissions.