Structural analysis based on unsupervised learning: Search for a characteristic low-dimensional space by local structures in atomistic simulations

Ryo Tamura [1,2,3,*], Momo Matsuda,[4] Jianbo Lin [1,5] Yasunori Futamura [4,5,6]
Tetsuya Sakurai [4,5,6,†] and Tsuyoshi Miyazaki [1,6,‡]

[1]*International Center for Materials Nanoarchitectonics, National Institute for Materials Science, Tsukuba 305-0044, Japan*
[2]*Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, Tsukuba 305-0044, Japan*
[3]*Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8568, Japan*
[4]*Department of Computer Science, University of Tsukuba, Tsukuba 305-8573, Japan*
[5]*Center for Artificial Intelligence, University of Tsukuba, Tsukuba 305-8573, Japan*
[6]*Master's/Doctoral Program in Life Science Innovation, University of Tsukuba, Tsukuba 305-8577, Japan*

Owing to the advances in computational techniques and the increase in computational power, atomistic simulations of materials can simulate large systems with higher accuracy. Complex phenomena can be observed in such state-of-the-art atomistic simulations. However, it has become increasingly difficult to understand what is actually happening and mechanisms, for example, in molecular dynamics (MD) simulations. We propose an unsupervised machine learning method to analyze the local structure around a target atom. The proposed method, which uses the two-step locality preserving projections (TS-LPP), can find a low-dimensional space wherein the distributions of data points for each atom or groups of atoms can be properly captured. We demonstrate that the method is effective for analyzing the MD simulations of crystalline, liquid, and amorphous states and the melt-quench process from the perspective of local structures. The proposed method is demonstrated on a silicon single-component system, a silicon-germanium binary system, and a copper single-component system.

I. INTRODUCTION

The recent progress in atomistic simulations is remarkable. This approach can be implemented to simulate large systems with higher accuracy with the help of advanced computational techniques and improved computational power. Compared with the classical molecular dynamics (MD) simulation, which uses empirical atomic force fields, the MD simulation based on the density functional theory (DFT), which is called first-principles MD (FPMD), is reliable even for new materials or unknown phases considering the lack of experimental information [1–9]. With the progress in large-scale DFT calculation techniques [10], FPMD simulations for larger system sizes can be realized. Machine-learning (ML) techniques can construct atomic force fields, which can reproduce the DFT results quite accurately [11–19]. With such ML force fields, the computation time is greatly reduced, and long-term MD simulations of large systems are possible [20,21].

As recent FPMD simulations can be utilized for explaining various phenomena and complex structural changes, it has become increasingly difficult to understand what is actually happening during such simulations. In some cases, a local or global unknown phase or structure is encountered. Even well-known materials may have a hidden structural order in

the middle range [22–24]. To analyze the structural properties of materials, radial distribution functions (RDFs) or pair correlation functions have been traditionally used [25–27]. Recently, analysis based on the persistent homology diagram has been employed [28–30]. However, these methods require statistics over many atoms or from a long-time profile. For FPMD simulations with a large system size, the system may not be in a single phase or can be irregular. Thus it is desirable to develop an analysis method with small statistics based on the local structure of the atoms.

To solve this problem, some methods for local structure detection have been proposed [31–37]. A common feature is that a local structure around each atom in a snapshot is expressed by a relatively high-dimensional descriptor (high-dimensional vector). The bond order parameter, smooth overlap of atomic positions (SOAP), and atom-centered symmetry functions (ACSF) are typically used as descriptors. The data sets, where the descriptors for many atoms are collected, are analyzed using supervised or unsupervised ML methods, such as neural networks to detect the differences in the local structures for each atom. These methods have been used to detect differences in local structures of crystal [31], amorphous [34], and supercooled liquid compounds [37].

In this study, we propose an unsupervised learning-based method to analyze local structures with the same motivation as that for previous studies. Specifically, the method aims to find a low-dimensional space, where the distributions of datapoints for each atom or groups of atoms can be accurately captured. The major difference between the proposed

*tamura.ryo@nims.go.jp
†sakurai@cs.tsukuba.ac.jp
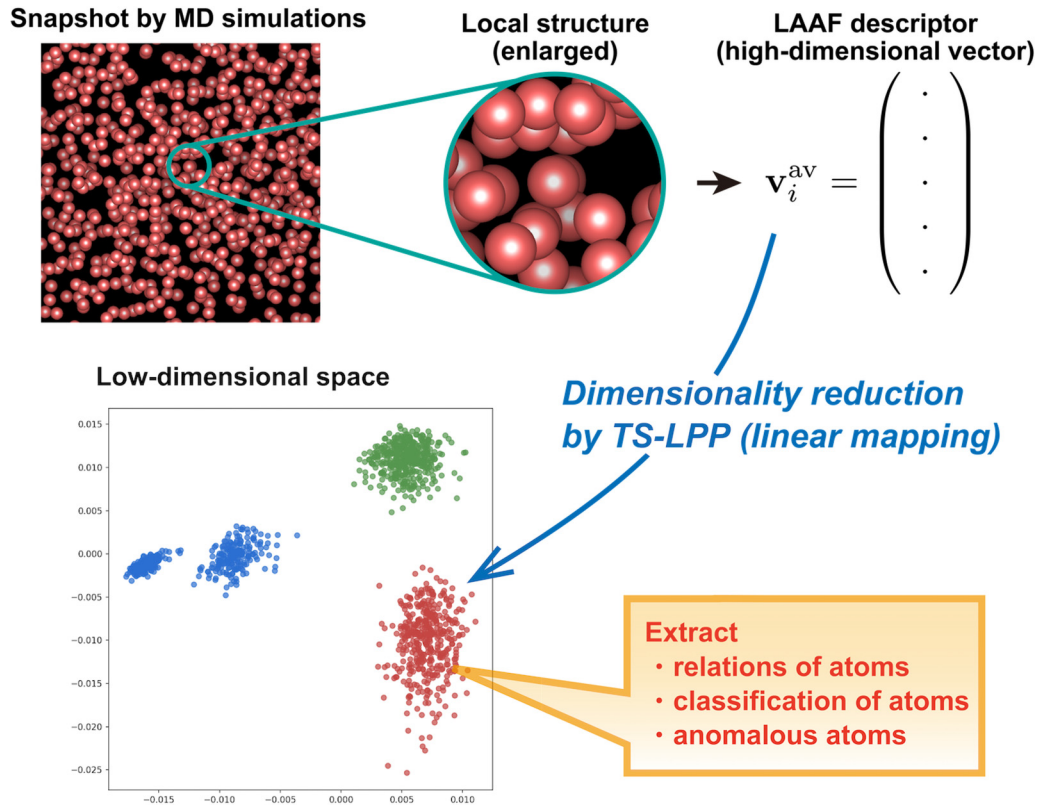‡miyazaki.tsuyoshi@nims.go.jp

FIG. 1. Schematic of unsupervised learning based structural analysis.

method and previous methods for local structure detection is that we use locality preserving projections (LPP) to perform dimensionality reduction. This allows us to obtain a linear transformation from a high-dimensional descriptor to a low-dimensional space, similar to that achieved through principal component analysis (PCA). The advantage to use linear transformation is that the exact linear transformation of new data points (out-of-samples) can be achieved. In addition, by a linear transformation, data mapping to low-dimensional space, which reflects the original data structure in the high-dimensional space, can be realized. We should note that there are many successful reports of the structural analyses, especially in the development of useful reaction coordinates, using nonlinear dimensionality reduction techniques such as a sketch-map method [38,39]. However, nonlinear transformation may sometimes significantly change the original data structure in the low-dimensional space.

The proposed method has three steps, as shown in Fig. 1 (detailed in Sec. II). Firstly, a high-dimensional descriptor, which is called as locally averaged atomic fingerprints (LAAF) descriptor, is introduced to represent the local structure. The LAAF descriptor shows a two-body correlation for the target atom and its neighboring atoms. Here, the LAAF descriptor using two-body correlation is the simplest descriptor belonging to the ACSF. Although many better descriptors have been developed, we adopt the simplest descriptor to verify the usefulness of our dimensionality reduction method. Secondly, a dimensionality reduction scheme is employed. As a dimensionality reduction method, the two-step locality preserving projections (TS-LPP) method, in which the calculations by conventional LPP [40] are repeated two times,

is proposed. It is important that the dimensionality reduction is unsupervised because it should detect the atoms in a new or unknown phase. Moreover, dimensionality reduction will ensure that a more reliable measure is introduced to evaluate the similarity between the two data. In the low-dimensional space, the classification of new data points (i.e., out-of-sample points) is expected to be more robust. Thirdly, in the low-dimensional space, we extract relations of each atom in terms of the local structure, perform the classification of new data points, and detect singular atoms which characterize the structure and drive phase transitions.

To validate the proposed unsupervised learning-based structural analysis method, we introduce the results of some examples. First, we perform unsupervised clustering of the local structure data sampled from snapshots of silicon in the crystalline, liquid, and amorphous phases in Sec. III A. To realize the unsupervised method, the information of the detected phases during dimensionality reduction is not used. The widely used PCA method cannot find a low-dimensional space where the three phases are clearly divided using their local structures. In contrast, in a low-dimensional space obtained by TS-LPP method, three well-separated groups, each of which corresponds to its correct phase, are obtained. Furthermore, we discuss why TS-LPP works well in comparison to PCA and LPP. Next, we focus on the structural analysis of the melt-quench process from liquid to amorphous formation in the silicon system in Sec. III B. Since linear mapping function for datapoints is available with the TS-LPP method, we can analyze new datapoints in the low-dimensional space. Thus each atom in the snapshots from the melt-quench process is plotted and analyzed in a low-dimensional space. We

address the behavior of the atoms in the process and the amorphous structures depending on the melt-quench processes. In addition, the silicon-germanium binary system and the copper system are targeted for the proposed unsupervised learning-based structural analysis in Secs. III C and III D. Discussion and Summary are given in Sec. IV.

II. METHODS

In this section, we explain an unsupervised learning-based method to analyze the local structure. The proposed structural analysis method uses the LAAF descriptor to express the local structure of each atom as high-dimensional vectors and the TS-LPP as the dimensionality reduction method. In addition, the setups of the FPMD and classical MD methods to generate snapshots of MD simulations are introduced.

A. LAAF descriptor

To express the local structure of the target atom using a high-dimensional vector, we use the atomic fingerprint proposed by Botu and Ramprasad [41]. For single-component systems, the atomic fingerprint vector of the ith atom is written as

$$V_i(\eta_m; R_c) = \sum_{j \neq i} \exp[-(r_{ij}/\eta_m)^2] f(r_{ij}; R_c), \quad (1)$$

$$f(r_{ij}; R_c) = \begin{cases} 0.5[\cos(\pi r_{ij}/R_c)^2 + 1] & (r_{ij} \leqslant R_c) \\ 0 & (r_{ij} > R_c) \end{cases}, \quad (2)$$

where $f(r_{ij}; R_c)$ is the cutoff function with cutoff R_c, which is related to the locality of the structural order in the two-body distribution function. Furthermore, $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between the position $\mathbf{r}_i$ of the ith atom and the position $\mathbf{r}_j$ of the jth neighbor atom, and $\eta_m (m = 1, \ldots, M)$ is the decay rate with distance. Thus an M-dimensional atomic fingerprint vector for ith atom is obtained as

$$\mathbf{V}_i(R_c) = (V_i(\eta_1; R_c), V_i(\eta_2; R_c), \ldots, V_i(\eta_M; R_c)). \quad (3)$$

The LAAF descriptor with cutoff R_a from ith atom corresponds to the locality with respect to the statistics (similarity) around the target atom and is calculated by

$$\mathbf{V}_i^{\mathrm{av}}(R_c, R_a) = \frac{1}{N_{\in R_a}} \sum_{j \in R_a} \mathbf{V}_j(R_c), \quad (4)$$

where $N_{\in R_a}$ is the number of atoms within the average radius R_a from the ith atom, and the sum is computed for the jth atom within R_a from the ith atom. If R_a and R_c are sufficiently small, the LAAF descriptor expresses the local structure around the ith atom. As R_a and R_c increase, locality decreases. In this study, we set $M = 100$, and η_m is defined by a logarithmic grid up to R_c.

In addition, the LAAF descriptor for binary systems can be defined in a straightforward manner. We consider a system consisting of two elements, a and b. The atomic fingerprint vectors of the ith atom of element a are defined as

$$V_{i \in a}^a(\eta_m; R_c) = \sum_{j \in a. \mathrm{and}. j \neq i} \exp[-(r_{ij}/\eta_m)^2] f(r_{ij}; R_c), \quad (5)$$

$$V_{i \in a}^b(\eta_m; R_c) = \sum_{j \in b} \exp[-(r_{ij}/\eta_m)^2] f(r_{ij}; R_c), \quad (6)$$

where the former and latter equations capture the structure of elements a and b around the ith atom, respectively. When we use $\eta_m(m = 1, \ldots, M)$, a $2M$-dimensional atomic fingerprint vector for the ith a element atom is obtained as

$$\mathbf{V}_{i \in a}(R_c) = \big(V_{i \in a}^a(\eta_1; R_c), V_{i \in a}^a(\eta_2; R_c), \ldots, V_{i \in a}^a(\eta_M; R_c),$$
$$V_{i \in a}^b(\eta_1; R_c), V_{i \in a}^b(\eta_2; R_c), \ldots, V_{i \in a}^b(\eta_M; R_c)\big). \quad (7)$$

Using the cutoff R_a, the LAAF descriptor for the a-element atom can be calculated as

$$\mathbf{V}_{i \in a}^{\mathrm{av}}(R_c, R_a) = \frac{1}{N_{a, \in R_a}} \sum_{j \in R_a} \mathbf{V}_{j \in a}(R_c), \quad (8)$$

where $N_{a, \in R_a}$ is the number of a-element atoms within the average radius R_a from the ith atom and the sum is computed for the jth a-element atom within R_a from the ith atom. To calculate the LAAF descriptor for the b-element atom, a is swapped with b in the equations.

B. TS-LPP method

We propose the TS-LPP method as a dimensionality reduction method and an automated determination technique for the hyperparameters. The flow of this algorithm is shown in Fig. 2. Consider the case in which the dimension of the LAAF descriptors is M and the dimension of the target low-dimensional space is d_r. In TS-LPP, the initial dimension reduction from M-dimensional space to d_m-dimensional space ($d_r < d_m < M$) is performed by conventional LPP [40].

The LPP method is as follows (see Fig. 2). The input data matrix X is composed of M-dimensional feature vectors $\{\mathbf{x}_i\}_{i=1,\ldots,N}$, where N is the number of data points. In this study, $\mathbf{x}_i$ is the ith LAAF descriptor, which is standardized by the z score. The method begins with a calculation of the weighted adjacency matrix W where each component is defined as follows:

$$W_{ij} = \begin{cases} \exp[-\|\mathbf{x}_i - \mathbf{x}_j\|]/2\sigma^2 & (i \neq j) \\ 0 & (i = j) \end{cases}, \quad (9)$$

where σ is a hyperparameter, and this adjacency matrix is the $N \times N$ matrix. To create the k_{nn}-nearest-neighbor graph (similarity graph), for each column in W, the off-diagonal elements, except those related to the k_{nn}-nearest-neighbor data points, are forced to 0. Here, we fix $k_{\mathrm{nn}} = 7$. Next, to represent the adjacency matrix as a symmetric matrix, the following operation is performed: $W_{ij} = \max(W_{ij}, W_{ji})$ [42]. Furthermore, we introduce the degree matrix D, which is the $N \times N$ diagonal matrix given as

$$D = \begin{pmatrix} d_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_N \end{pmatrix}, \quad (10)$$

where each diagonal component is defined as

$$d_i = \sum_{j=1}^{N} W_{ij}. \quad (11)$$

Using W and D, the graph Laplacian is defined as
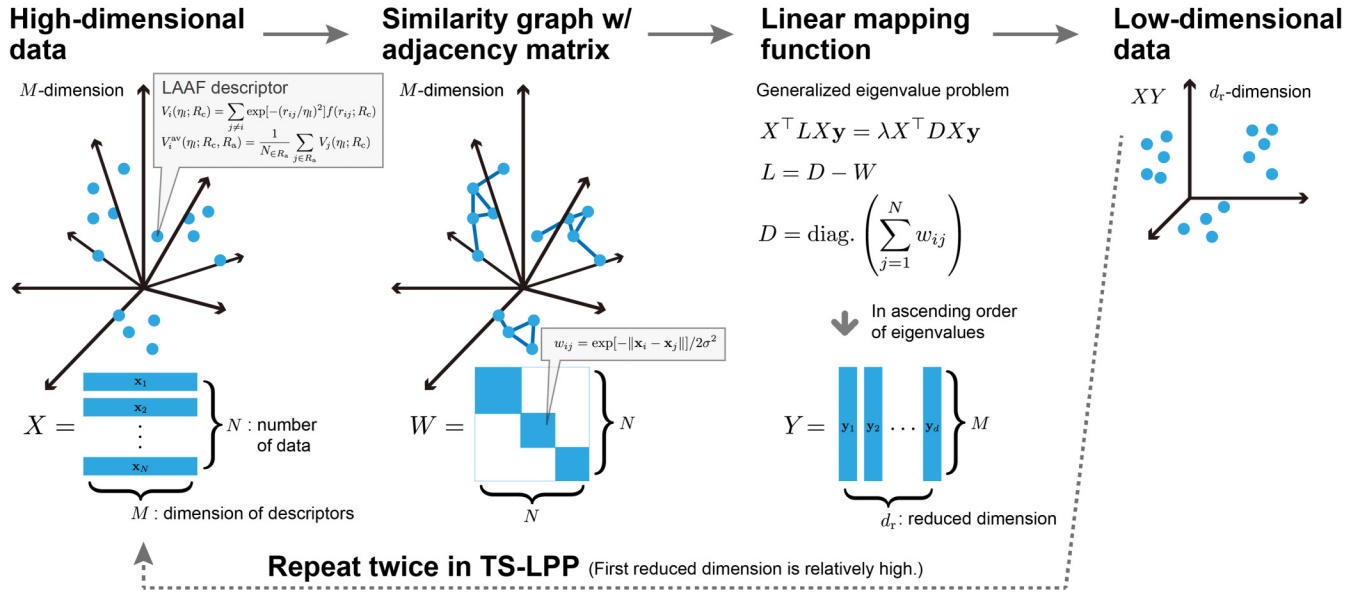
$$L = D - W. \quad (12)$$

FIG. 2. Procedure of the two-step locality preserving projection (TS-LPP) method. The input data matrix X is constructed by the M-dimensional feature vector $\mathbf{x}_i$ defined by the LAAF descriptor sampled from molecular dynamics simulations. Through a similarity graph with an adjacency matrix W, similar data are embedded to the closer positions in low-dimensional space, and a linear mapping matrix Y is obtained by solving the generalized eigenvalue problem. In TS-LPP, first, the embedded data in the middle-dimensional space, which is larger than the target dimension, is calculated. Subsequently, dimension reduction to the final target dimension is performed again. In the low-dimensional space, we extract the relations of each atom in terms of the local structure using the TS-LPP method.

In LPP, we solve the following generalized eigenvalue problem:

$$X^{\top}LX\mathbf{y} = \lambda X^{\top}DX\mathbf{y}. \tag{13}$$

Here, $\{\lambda_i, \mathbf{y}_i\}$ is the ith eigenvalue and eigenvector of this problem, which are arranged in ascending order.

Let d_{r} denote the target dimension for dimension reduction. Then, the mapping matrix by LPP from M-dimensional space to d_{r}-dimensional space is given by

$$Y = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{d_{\mathrm{r}}}). \tag{14}$$

Using this mapping matrix, the feature vector $\mathbf{x}$ is mapped to the low-dimensional vector $\mathbf{x}'$ with d_{r} dimensions as

$$\mathbf{x}' = \mathbf{x}Y. \tag{15}$$

In the numerical calculation for obtaining Y, the low-rank approximation is used to improve the numerical stability, which is explained in Appendix.

In TS-LPP, we conduct LPP from d_{m}-dimensional space to d_{r}-dimensional space. Thus d_{m} is a hyperparameter in this method. Furthermore, to perform LPP, another hyperparameter σ exists to prepare the weighted adjacency matrix. To determine the appropriate hyperparameters d_{m} and σ, the Calinski-Harabasz score (Pseudo-F) [43] is used. This score is defined as the ratio between the intra-cluster and inter-cluster dispersions. This score is large when the distances between clusters are large, and each cluster is dense. A grid search for hyperparameters d_{m} and σ is performed to maximize the Calinski-Harabasz score, using scikit-learn [44]. To evaluate the score, the labels obtained by the k-means method [45] with a fixed cluster number are used. Thus it is necessary to analyze the cluster number dependency of the score. However, as shown in the results, this dependency is not critical, and

the cluster number is not an important factor. Therefore we try several values of cluster numbers and adopt the cluster number with which the distributions of datapoints for each atom or groups of atoms can be properly captured in the low-dimensional space. The scheme mentioned above is unsupervised because information about the correct state of the training data is not used. Note that the common value of σ is used in the first and second LPP calculations.

C. Characteristic of TS-LPP method

To demonstrate the advantage of TS-LPP method, we introduce an example when a 2D synthetic data set is widely distributed along the y axis and the distance Δ between the two classes in the x axis is gradually decreased. The prepared synthetic data set is shown in Fig. 3. As we discuss later, similar data structure is observed in the distribution of the data generated by the MD simulations in this work. This 2D synthetic data set is randomly generated from the normal distribution with mean of 0 and standard deviation of 10^{-6} for x axis and 0.98 for y axis, respectively. Each class has 500 datapoints and the distance between these classes along x axis is changed as $\Delta = 1.2, 0.2$, and 0.0. In this demonstration, hyperparameter σ is fixed as 5. In Fig. 3(a), we show scatter plots in the case of $\Delta = 1.2$. Using dimensionality reduction methods, we intend to obtain 1D space that allow one to (linearly) separate class 1 (orange) and 2 (blue). In the panels representing PCA, LPP, and TS-LPP results, we show not only the first principal axis (shown as x axis) derived by a dimensionality reduction but also the second axis (y axis) in order to see the behaviors of the linear maps. As shown in the left most panel (original) of the figure, class 1 and 2 can be separated on x axis. The most widespread y
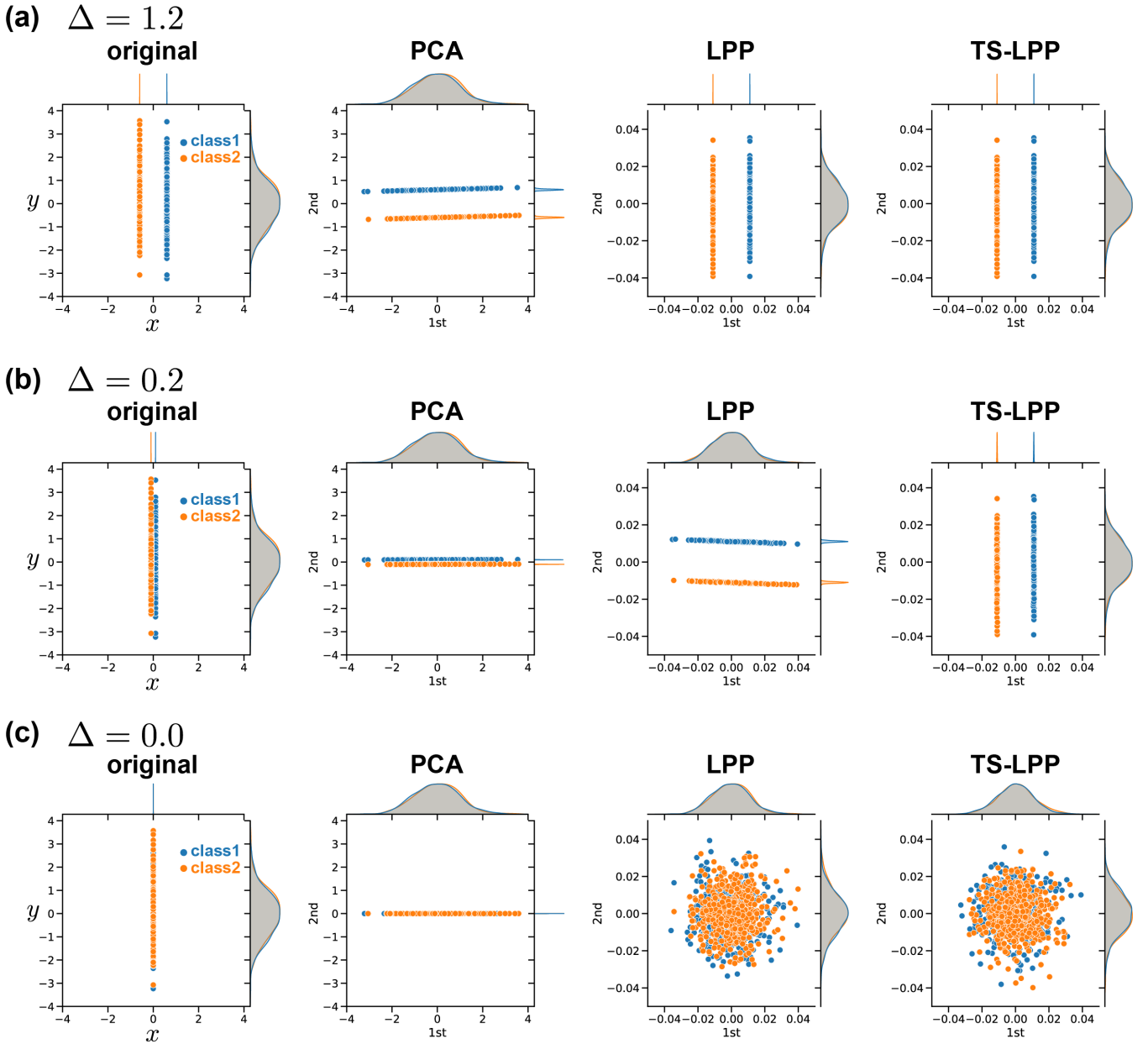
FIG. 3. Two dimensional synthetic data where the two classes are widely distributed along the y axis and the distance between the classes in the x axis is gradually decreased as (a) 1.2, (b) 0.2, and (c) 0.0. The original data, PCA, LPP, and TS-LPP results are shown.

axis cannot be used to distinguish class 1 and 2. This data distribution makes the first axis of PCA which fails to separate class 1 and 2. In contrast, both LPP and TS-LPP are successful to determine the first axis that can separate the classes.

Furthermore, the results of LPP and TS-LPP are different when the distance between two classes is smaller ($\Delta = 0.2$) as shown in Fig. 3(b). While the first axis of LPP cannot separate the classes as with PCA, the distance between the classes on the second axis is larger. TS-LPP applies LPP again to the 2D space obtained by LPP. The resulting first axis successfully separates the classes. Note that even if LPP is performed after PCA, the classes are not separated in the principal axis. We also show the case of $\Delta = 0$ in Fig. 3(c). In this case, the distance between class 1 and 2 along the x axis is zero, and all

methods unsurprisingly fail to separate the classes. For PCA, the most widespread y axis is chosen as the first axis, which is also reasonable.

This validation on synthetic data shows that TS-LPP performs well when the intraclass variation is large and the interclass distance is fairly small. In addition, if the distance between the classes is large enough, the data can be separated even with a single LPP. These results can be explained by the locality preserving nature of LPP. In LPP, close data in the original space are mapped to close positions in the low-dimensional space. In other words, if the distance between two classes is too small, the data will be mapped to close positions in the first axis, and the two data sets will be mixed in the low-dimensional space. On the other hand, even if the distance between the two classes is not large in the original space, this
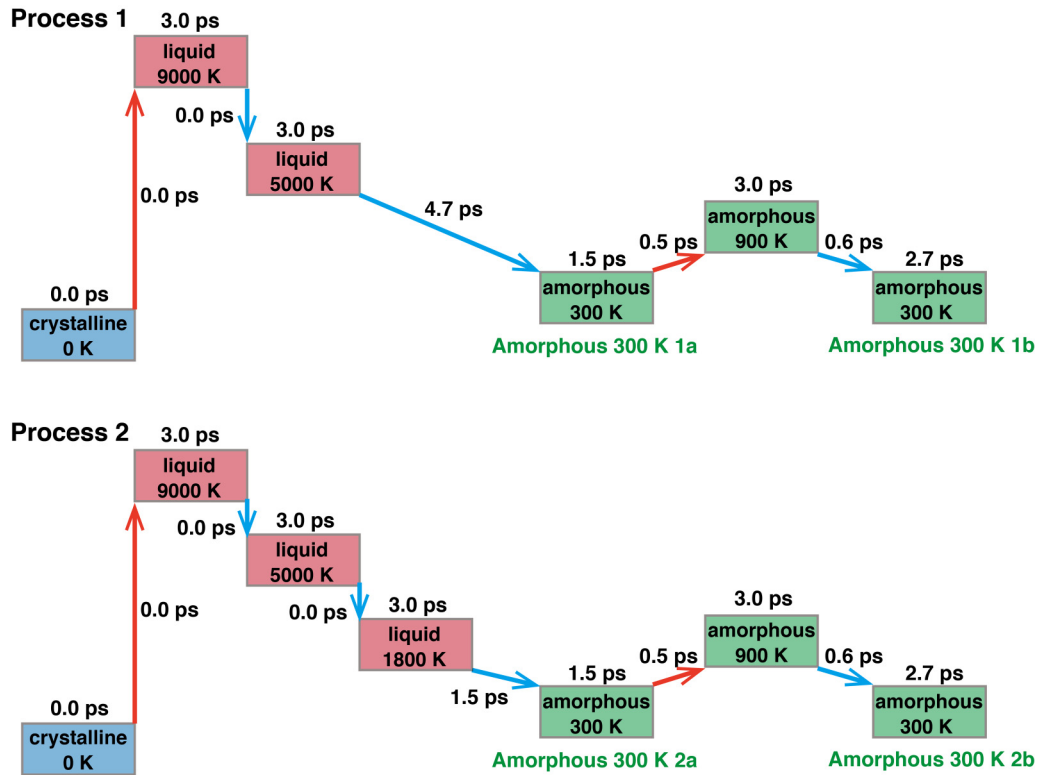
Process 1

Process 2

FIG. 4. Two melt-quench processes to prepare the amorphous states.

distance is extended in the second axis by LPP. This situation is not observed by PCA, and thus this is an advantage to use LPP. As will be demonstrated in Sec. III, this aspect of LPP or TS-LPP is important to find out the subtle differences between the similar data of local structures in different phases.

D. MD simulations

To generate MD simulation results for silicon single-component system and silicon-germanium binary system, we perform FPMD simulations using a linear-scaling method with the CONQUEST code [46,47]. The accuracy of this calculation method is explained in Refs. [48–50]. The local density approximation with the standard Ceperley-Alder exchange-correlation functional is used. Using the SIESTA code [51], Troullier-Martins-type norm-conserving pseudopotentials and the pseudo-atomic orbital basis sets are generated. We employ a minimal basis set with a cutoff energy for the charge density grid of 80 Hartree. The density matrix minimization (DMM) method realizes linear-scaling FPMD simulations [52]. The cutoff range of the auxiliary density matrix in the DMM method is 16.0 Bohr. Using the Nose-Hoover chain thermostats, canonical ensemble (NVT) simulations [53] are carried out.

For the copper single-component system, classical MD simulations with an effective medium theory (EMT) potential are performed using the atomic simulation environment (ASE) package [54]. We prepare initial structures for crystalline or liquid states with the Maxwell-Boltzmann distribution of the velocity at the targeted temperatures. NVT-MD simulations are conducted using the time step of 1 fs and

Langevin thermostat with a friction value of 0.01 atomic unit.

III. RESULTS

To show a potential of our method, the results of the proposed unsupervised structural analysis are presented in this section. The target systems are the silicon single-component system, silicon and germanium binary system prepared by FPMD simulations, and the copper single-component system prepared by classical MD simulations.

A. Structural analysis of crystalline, liquid, and amorphous states in Si system

To verify the efficiency of the proposed unsupervised structural analysis method, we work on a silicon system in the crystalline, liquid, and amorphous phases as the simplest case. We show that TS-LPP can find the low-dimensional space where the three states are completely divided even if R_a and R_c values are small.

The various structures in the three states are prepared by the FPMD simulations for a system containing 1000 silicon atoms at constant temperature and constant volume. FPMD simulations are performed for the crystalline states at 300 and 1200 K; liquid states at 3000, 5000, and 9000 K; and the four amorphous states labeled as 1a, 1b, 2a, and 2b at 300 K. The amorphous states are prepared using two different melt-quench processes, as shown in Fig. 4. The total time of the FPMD simulations is 19 ps in both cases. The RDFs for

these states are shown in Fig. S1 [55], and these results are consistent with those previously reported [56–60].

1. Distributions of datapoints in the 2D spaces obtained by TS-LPP, LPP, and PCA

To find a low-dimensional space, we prepare a target data set including atoms expressed by the LAAF descriptor in the crystalline states at 300 and 1200 K, the liquid states at 3000 and 9000 K, and the amorphous states 1a and 1b at 300 K. The number of datapoints in each state is 200, and the total number of data is 1200. Figure 5(a) shows the distributions in 2D space ($d_r = 2$) obtained by the TS-LPP method depending on two parameters R_a and R_c. R_c is related to the locality of the structural order in the two-body distribution function, and R_a corresponds to the locality with respect to the statistics (similarity) around the target atom. Thus local structure analysis can be performed when these parameters are sufficiently small. Note that, to determine the appropriate hyperparameters, the cluster number is fixed as three, because we assume that the data will be grouped into three clusters. However, as will be shown later, this value has little effects on the results.

By increasing R_a and R_c, the distributions of datapoints for the three states are distinguished. When $R_a = 4$ Å and $R_c = 4$ Å, the three distributions are well distinguished, although a small overlap between the crystalline and amorphous states is observed. The case of $R_a = 4$ Å and $R_c = 5$ Å is the best result when the three states are completely divided. The typical interatomic distance between Si atoms is 2.2 Å for the nearest and 3.8 Å for the next-nearest-neighbor atoms. This information is included in the LAAF descriptor when $R_c = 5$ Å at least. This means that when the LAAF descriptor of a target atom is analyzed, if the atomic fingerprints for the nearest-neighbor and next-nearest-neighbor atoms are also considered, the phase of the target atom can be identified perfectly.

We consider the results depending on the number of clusters for determining hyperparameters. Figure 5(b) shows the distributions in the reduced two dimensions for different cluster numbers. The values of the optimized hyperparameters are also provided. It is encouraging that the results are almost the same. This suggests the present analysis method can extract a useful low-dimensional space to classify the local structures without any preceding information.

To compare the results by other dimensionality reduction methods, where the linear transformation to a low-dimensional space can be obtained, the distributions by PCA and LPP are shown for the case of $R_a = 4$ Å and $R_c = 5$ Å in Fig. 6. The three distributions overlap with each other in both methods. The PCA method extracts a component (or dimension) showing the largest possible variance under the constraint that it is orthogonal to the preceding components. In both first and second principal components, the minimum and maximum data points originate from the liquid state. In the original LAAF descriptor space, the variation of the data points in the liquid state is probably much larger than that in the crystalline or amorphous state. Hence, efficient clustering to distinguish the data points of the amorphous or the crystalline states from the liquid states is difficult. Thus PCA

does not seem to work with such a data structure (distributions depending on the cutoff values are shown in Fig. S2 [55]). On the other hand, if the values of R_a and R_c increase, LPP can create a 2D space where the three states are completely divided (distributions depending on the cutoff values are shown in Fig. S3 [55]). We conclude that TS-LPP is an effective dimensionality reduction method for distinguishing atoms by local structures when R_a and R_c are sufficiently small.

In addition, the results obtained by other conventional dimensionality reduction methods are summarized in Fig. S4 [55], and it is found that these methods are not as efficient as TS-LPP. Recently, several reports have suggested that two-step analysis for dimension reduction is useful, such as PCA-LPP, PCA-t-SNE, and PCA-UMAP [61–64]. Our results show that there is a useful case even if the same dimensionality reduction method is repeated in the two-step analysis, although the different algorithms are usually combined. In Sec. III A 3, we will discuss why the TS-LPP can create a better 2D space than PCA and LPP.

2. Performance for out-of-sample points

For the proposed method to be capable of various applications, it is also important to locate new data points (out-of-sample points), which are not included in the data set used to create the low-dimensional space. Note that some dimensionality reduction methods, such as neural networks and t-SNE, cannot be applied to a new data set using a linear mapping function. For this purpose, we prepare a new data set from nine FPMD simulations and plot the new data points in the low-dimensional spaces shown in Figs. 5(a) and 6 for $d_r = 2$, using the linear mapping function obtained by the PCA, LPP, and TS-LPP methods when $R_a = 4$ Å and $R_c = 5$ Å. The new data set contains 200 data points for each MD trajectory. To prepare out-of-sample points, for the crystalline 300 and 1200 K, liquid 3000 and 9000 K, and amorphous 300 K 1a and 1b, the new data points in different snapshots from the training data are selected. On the other hand, none of the data from the trajectories of MD simulations of the liquid phase at 5000 K, 2a and 2b of the amorphous phase at 300 K are used as the training data, and thus all of the data from the MD trajectories are out-of-sample points. To evaluate the relationship between the new data and its correct state, we assign each new data point into the crystalline, liquid, or amorphous state, by the k-nearest-neighbor method with $k = 10$ using the data sets in the low-dimensional spaces. Then, for each FPMD simulation, we count the number of test data assigned to the correct state in the low-dimensional space. Figure 7 shows the ratio of the agreement in the case of $d_r = 2$. The classification using PCA does not agree well with the correct state, especially for the liquid states. The results of LPP are much better than those of PCA for liquid states, but the results for the amorphous states by these two methods are different, and they are both not reliable. In contrast, the TS-LPP results are perfect for all cases, both for the data from the MD trajectories used for creating the low-dimensional space and for those from other trajectories (i.e., liquid at 5000 K, amorphous 2a and 2b at 300 K). This result suggests that the reduced dimensions prepared by TS-LPP are transferable and useful, even for the analysis of new data.

(a) TS-LPP results depending on R$_a$ and R$_c$ (Cluster number = 3)
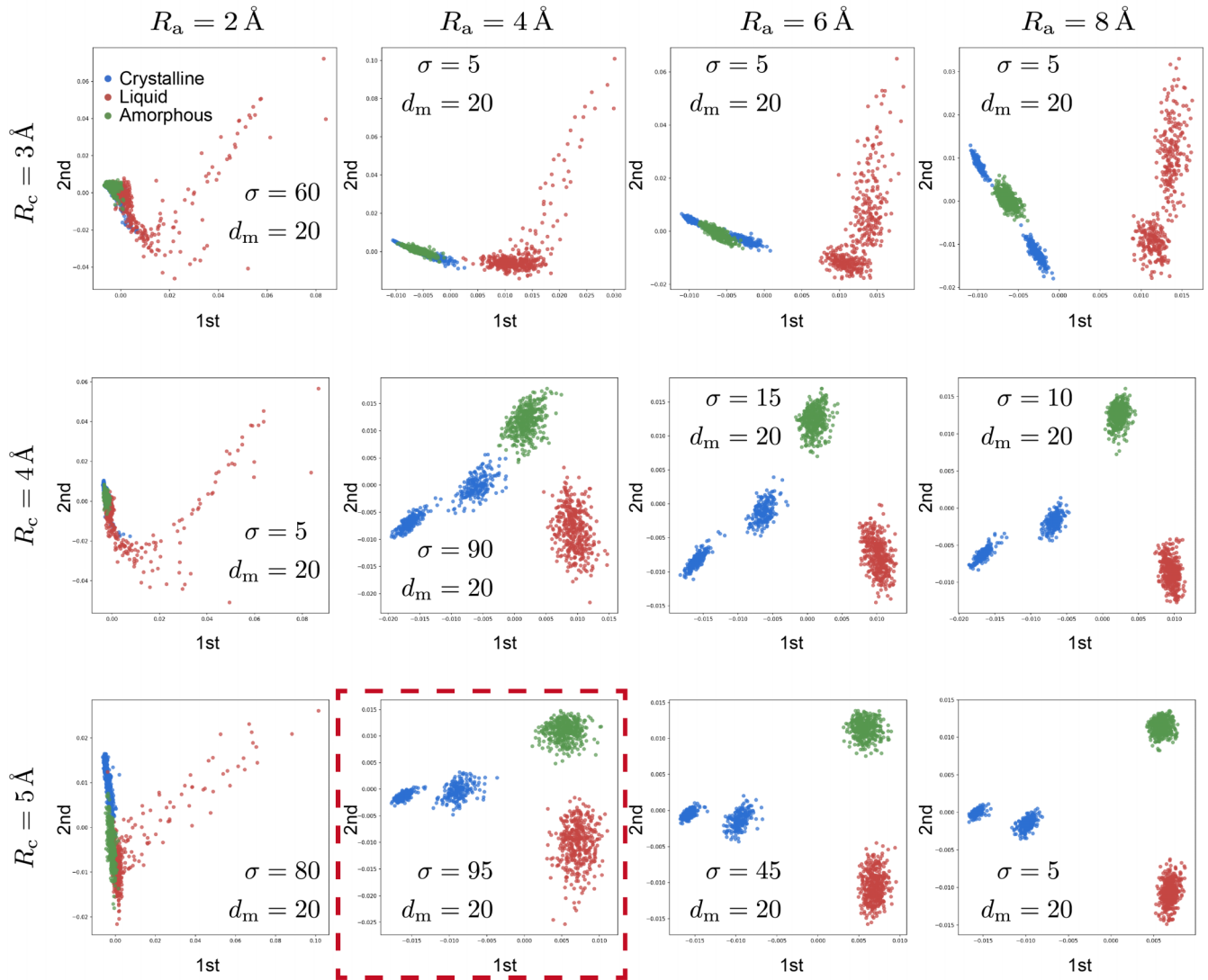
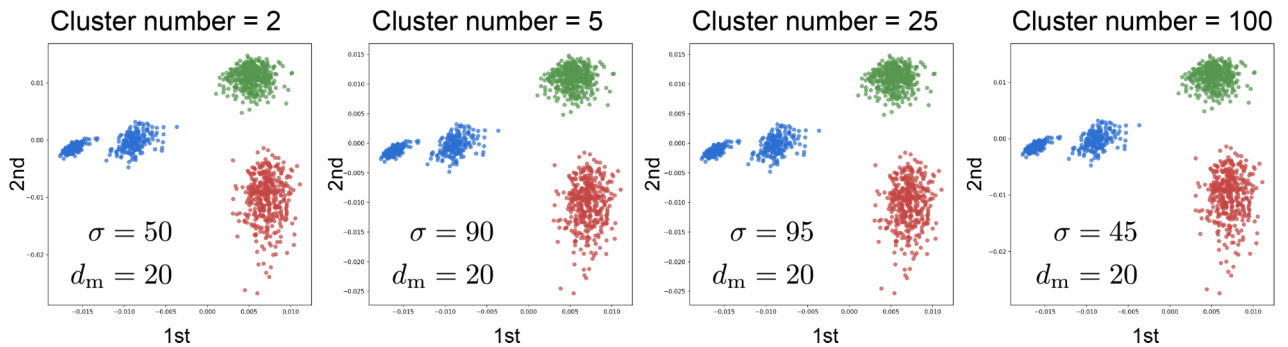## (b) TS-LPP results depending on cluster number ($R_a = 4$ Å and $R_c = 5$ Å)

FIG. 5. (a) Distributions after dimension reduction to $d_r = 2$ by the TS-LPP method depending on R_a and R_c. Blue, red, and green indicate crystalline, liquid, and amorphous states, respectively. The values of the hyperparameters σ and d_m are determined by the Calinski-Harabasz score when the cluster value is fixed at three. When $R_a = 4$ Å and $R_c = 5$ Å, perfect clustering is obtained, and the result is surrounded by a red dotted line. (b) Cluster number dependence on the distributions and hyperparameters when $R_a = 4$ Å and $R_c = 5$ Å by the TS-LPP method. The results are almost the same with as in the three-cluster case.
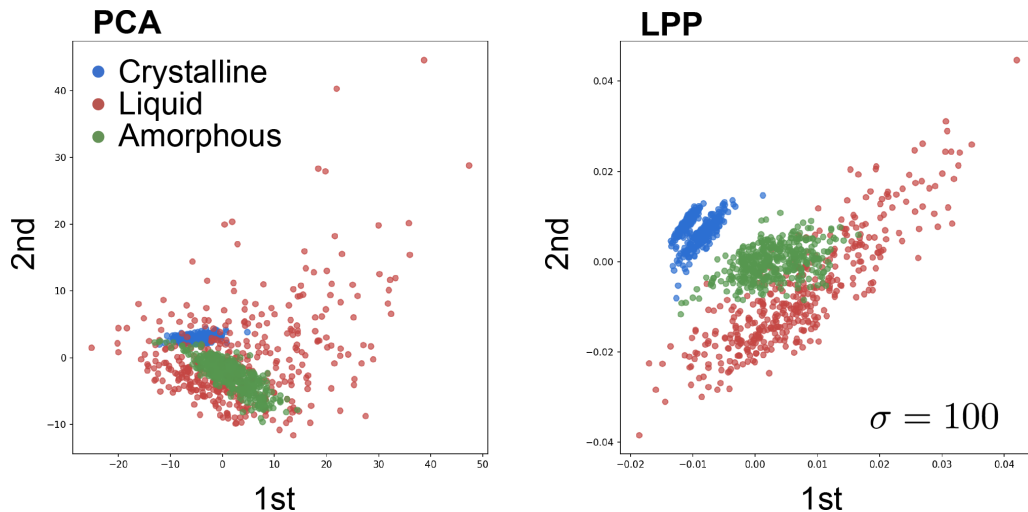
FIG. 6. Distributions after dimension reduction to $d_r = 2$ by the PCA and LPP methods when $R_a = 4$ Å and $R_c = 5$ Å. Blue, red, and green indicate the crystalline, liquid, and amorphous states, respectively. The optimal value of σ is denoted in the LPP case when the clustering number is fixed at three.

3. Advantage of TS-LPP method

In this section, we discuss why TS-LPP can find a better low-dimensional space for our structural analysis than PCA and LPP. To visualize the transformation of data distribution performed by dimensionality reduction methods, we perform a 3D to 2D dimensionality reduction. By a preliminary experiment, we identify and select three dominant descriptors from the 100-dimensional LAAF descriptor that provide a 3D space well-representing the characteristic of the original 100-dimensional space. A 3D plot for the three descriptors
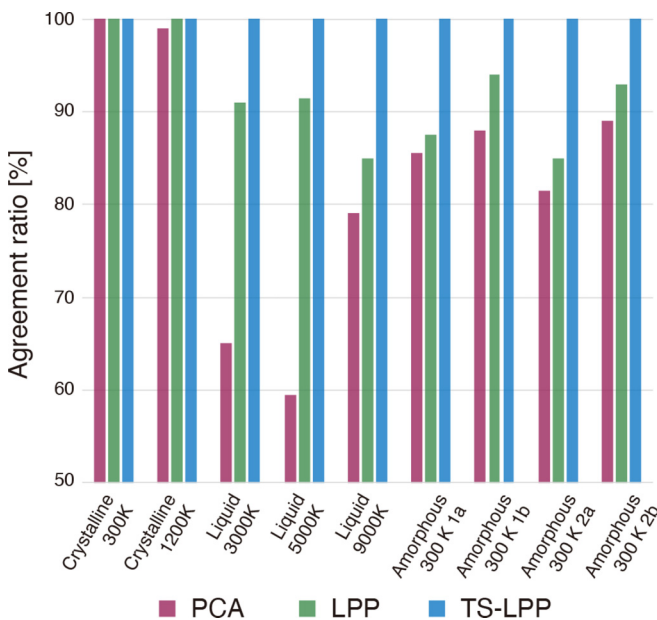

FIG. 7. Ratio of the agreement between the predicted and correct states for new data points by PCA, LPP, and TS-LPP for $R_a = 4$ Å and $R_c = 5$ Å by the k-nearest-neighbor method. The 2D spaces shown in Figs. 5(a) and 6 are used. Datapoints of the liquid state at 5000 K and the amorphous states 2a and 2b at 300 K are not included in the training data set.

is shown in Fig. 8(a). For each state (crystalline, liquid, and amorphous), the data points are distributed on a plane. The orientations of the planes look similar and the difference of the positions of the planes is subtle. For this 3D data set, PCA, LPP, and TS-LPP are performed to obtain the 2D spaces, and the results are shown in Figs. 8(b)–8(d). Only TS-LPP can separate the crystalline state from the liquid and amorphous states.

As we have seen in the analyses of synthetic data shown in Sec. II C, TS-LPP performs well when the data structure consists of multiple widely distributed data in a low-dimensional space. The data given by the present MD simulations have a similar data structure [Fig. 8(a)]. Such a data structure will appear, for example, when each feature in the high dimension is not independent but is correlated with each other. That is, there is a lower-dimensional space that represents the correlations in the features, and the data points will be widely distributed within this space. In this case, the differences between the classes appear as differences in this lower-dimensional space, and the data structure will be a pile of widely distributed data. In fact, since the LAAF descriptors used in this study are highly related to each other, TS-LPP would be effective for our target data sets. Because the descriptors used in materials science and condensed matter physics are often related to each other, TS-LPP is an effective method for dimensionality reduction in these fields.

B. Structural analysis of melt-quench process in Si system

We analyze the trajectories of the melt-quench process in the Si system are analyzed. From Sec. III A, we have two trajectories, that is, melt-quench processes 1 and 2, by FPMD simulations for melt-quench processes (see Fig. 4), and we show the structural analysis results using these trajectories.

1. Classification of atoms in melt-quench process

We apply the method, TS-LPP with the LAAF descriptor, to the analysis of the FPMD simulation of the melt-quench process for amorphous formation. Here, we classify all atoms
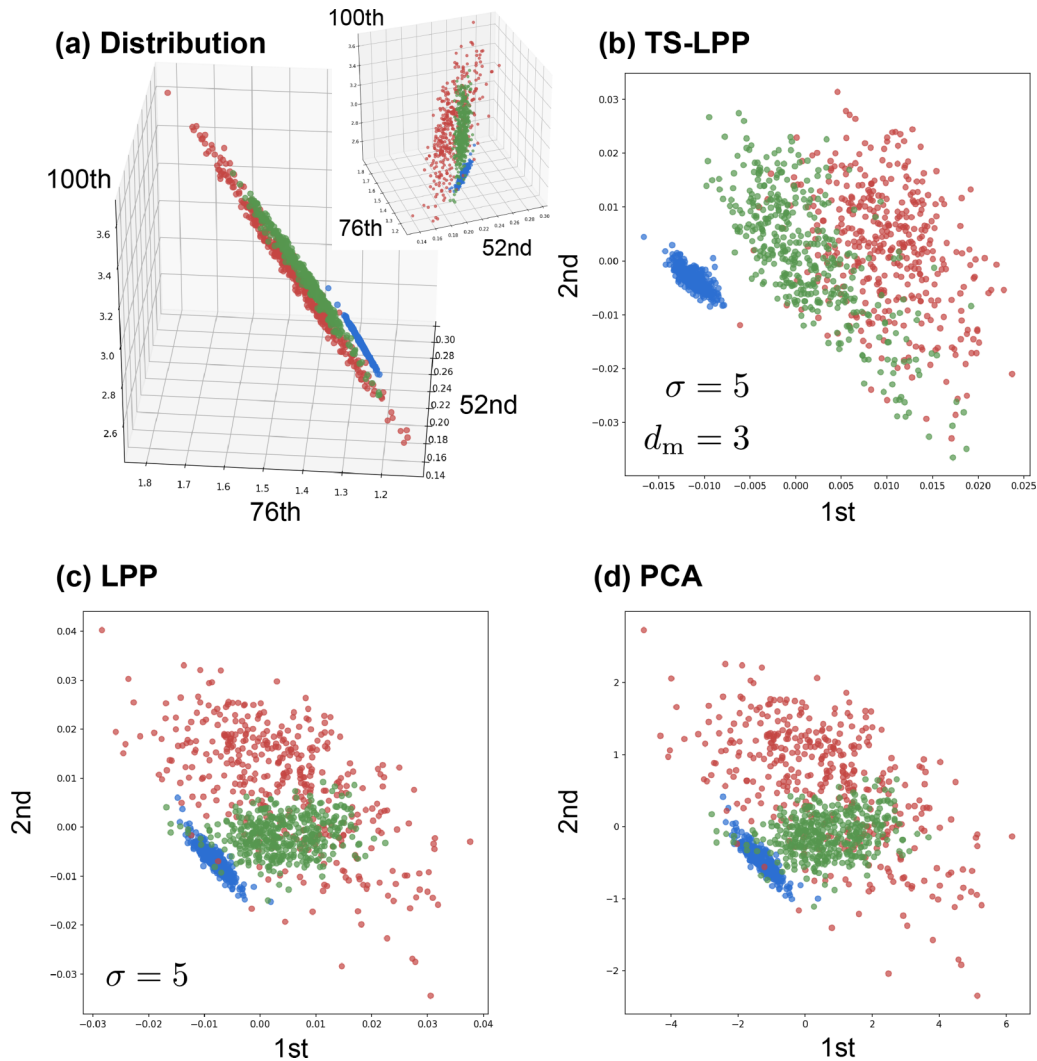
FIG. 8. (a) Distributions of the selected three features of 52nd, 76th, and 100th LAAF descriptors when $R_a = 4$ Å and $R_c = 5$ Å. Blue, red, and green indicate the crystalline, liquid, and amorphous states, respectively. The data points of each state are distributed on a plane. The orientations of the normal vectors of the planes look similar and the difference of the positions of the planes are small. Distributions in 2D space by (b) TS-LPP, (c) LPP, and (d) PCA are plotted. The optimal values of σ and d_m are denoted when the clustering number is fixed at three. For this data set, TS-LPP is better for distinguishing crystalline states from liquid and amorphous states.

at all snapshots in melt-quench process 1 (see Fig. 4) by using the k-nearest-neighbor method with $k = 10$ in the 2D space obtained by TS-LPP for $R_a = 4$ Å and $R_c = 5$ Å, as shown in Fig. 5(a).

Figure 9(a) summarizes the characteristic snapshot structures during the melt-quench process (lower figure) and the data points colored in orange for ten randomly selected atoms from each snapshot in the 2D space by TS-LPP, together with the points of the data set used to create low-dimensional space (upper figure). The blue, red, and green points denote the crystalline, liquid, and amorphous states, respectively. Figures 9(b) and 9(c) show the time profile of the temperature and the percentage of atoms belonging to each state, which is predicted by the k-nearest-neighbor method, respectively. The supplementary movie 1 [55] shows the time dependence of these results for the entire melt-quench process.

During the melt-quench process, the target temperature is sometimes set as constant for a few picoseconds. The system shows an almost equilibrated state during this period, for

example, cases III and VI in Fig. 9. In such cases, almost all atoms are classified as the corresponding state, indicating that the data points of all atoms are located near or in the region of the corresponding cluster in the 2D space obtained by TS-LPP. In contrast, when the system is cooled or heated, the classified states are mixed. For instance, when the temperature rapidly increases from the crystalline state (initial state, case I in Fig. 9), the three states are mixed (see case II in Fig. 9). In the snapshot of case IV, where the percentages of atoms classified as the liquid state and amorphous state are approximately the same (i.e., around 1280 K), atoms classified as the amorphous state are centered at the upper right part (dotted circle). The vacancy seems to act as a nucleus for the phase transition between the liquid and amorphous states.

More interestingly, in the cooling process from 5000 to 300 K, the selected ten atoms are observed to move along the line (mainly along the second axis, i.e., vertical axis) from the liquid to amorphous regions. This suggests that the line connecting the two states can be considered as a reaction
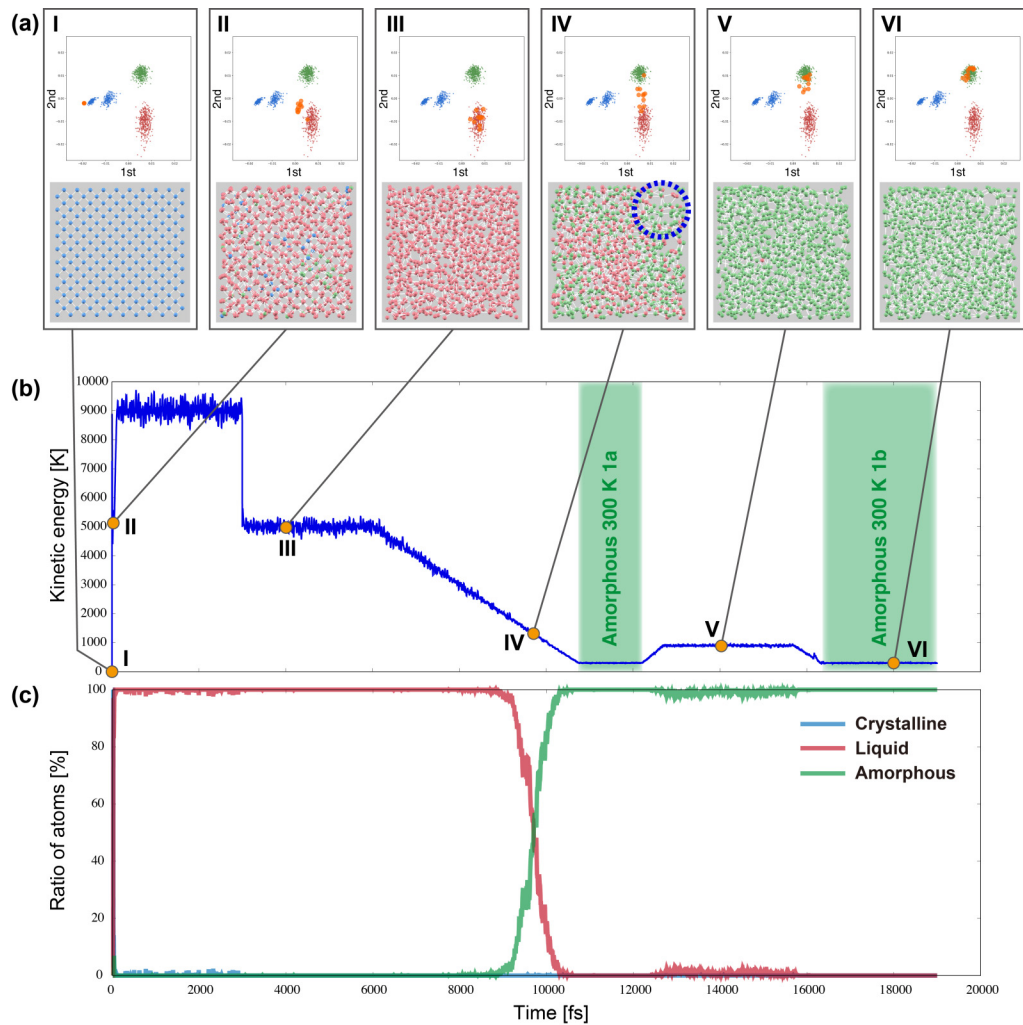
FIG. 9. (a) Data points for the ten randomly selected atoms in the 2D space by TS-LPP with $R_a = 4$ Å and $R_c = 5$ Å, which are colored in orange (upper figures), and characteristic snapshots (lower figures) in the melt-quench process. In the upper figures, the blue, red, and green points are regarded as crystalline, liquid, and amorphous states in the target data set, respectively. In the lower figures, to classify each atom, the k-nearest-neighbor method in the 2D space by TS-LPP is used. (b) Kinetic energy corresponding to the temperature of the system as a function of time. (c) Percentage of atoms belonging to each state.

coordinate for the "local" liquid-amorphous transition. In case V at 900 K, most atoms are classified as being in the amorphous state, but some atoms appear in the liquid state. Such detailed information can be found in the process of finding a more stable amorphous state (i.e., amorphous state 1b at 300 K) during the annealing process. These results demonstrate that by using our unsupervised method, we can visually understand the FPMD simulation results in detail. Note that such analysis would not be possible with PCA, because new data are often misclassified as shown in Fig. 7. For melt-quench process 2, the structural analysis of amorphous formation is also performed using the proposed method. The results are summarized in the supplementary movie 2 [55].

2. Detailed analysis of different amorphous structures

So far, we have investigated the clustering of the data points from three clearly different groups, crystalline, liquid, and amorphous states. Here, we focus on the amorphous phase and perform the detailed analysis of the local atomic structures in various amorphous states. For this purpose, we analyze the

five amorphous states: 1a, 1b, 2a, and 2b, which are made from the melt-quench process shown in Fig. 4, and the one made from the continuous random network model (CRN), which is an idealized model for perfectly coordinated amorphous semiconductors [65,66]. The quality of the amorphous structure generated by melt-quench methods strongly depends on process, especially on the quenching speed [21]. It is found that some detailed feature of the experimental structure factor cannot be reproduced by the amorphous structures made by the present melt-quench processes. On the other hand, the fifth structure is generated from a snapshot of the NPT simulation starting from the high-quality CRN structure previously reported, which agrees well with experimental results [67].

The data points sampled from these five amorphous states are mapped to the low-dimensional spaces obtained by the TS-LPP method [Fig. 5(a)] and the PCA method (Fig. 6) when $R_a = 4$ Å and $R_c = 5$ Å. The distributions of 200 data points for each amorphous state are shown in Figs. 10(a) and 10(b), and an enlarged view of TS-LPP results is shown in Fig. 10(c). Note that, different from Figs. 5 and 6, the same phase with
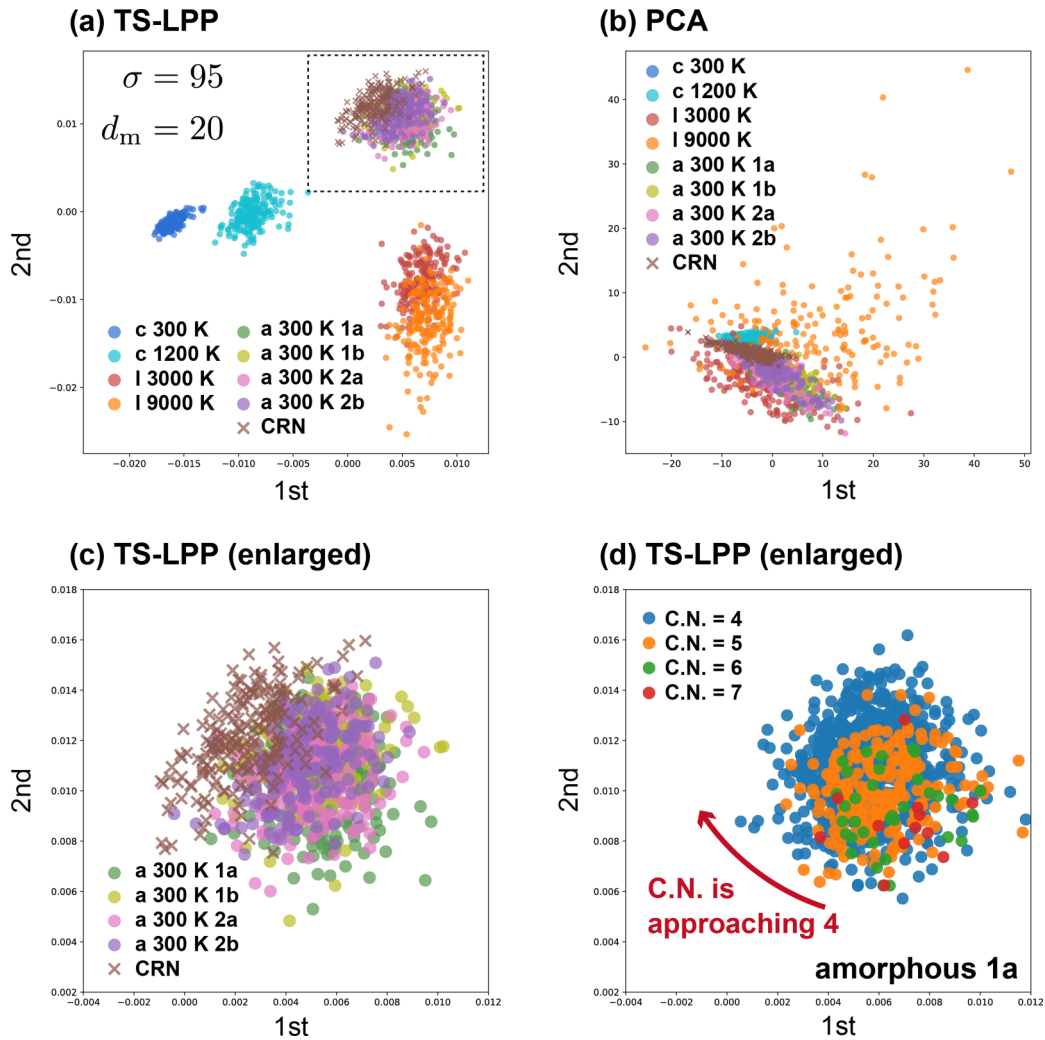
FIG. 10. Distributions of data points for amorphous states 2a and 2b at 300 K, and the CRN results mapped to the 2D space constructed by (a) TS-LPP method and (b) PCA, when $R_a = 4$ Å and $R_c = 5$ Å. (c) Enlarged view of the amorphous states of TS-LPP. The region is denoted by the dotted line in (a). (d) Distributions of data points for 1000 atoms in amorphous 1a depending on the coordination number (C.N.). The region is denoted by the dotted line in (a).

different temperatures are distinguished by different colors. In the low-dimensional space created by PCA [Fig. 10(b)], the distribution of the data points sampled from the crystalline, liquid and the five amorphous states are all overlapped. Although the distribution of the CRN model is localized in the space by PCA, it is difficult to extract its unique structural properties from those of other states.

On the other hand, in the space created by TS-LPP, the distribution of the three phases are well separated even with the newly added amorphous data points. From Fig. 10(a), it can be seen that the new data points sampled from the amorphous 2a and 2b states are distributed close to the results of 1a and 1b in both cases, which is consistent with the prediction of states using the k-nearest-neighbor method (see Fig. 7). The data points from the CRN model are also distributed close to the region of 1a and 1b, but it is found that the center of its distribution is slightly shifted to upper left from those of 1a, 1b, 2a, and 2b. Furthermore, as can be seen in Fig. 10(c), the distribution of 1b is closer to the region of CRN model than 1a. This tendency is observed also in the data points from the

melt-quench process 2; the distribution of 2b is closer to the one for the CRN model than 2a. It supports that the annealing process to 900 K (see Fig. 4) is effective for the modeling of the amorphous structure. Furthermore, we also analyze the coordination number of the data points of amorphous state 1a, whose distribution is most deviated from that of CRN. The mapping of all 1000 atoms in one snapshot of amorphous state 1a is shown in Fig. 10(d), where the color differences represent the coordination number of each atom. Note that all of the atoms are four coordinated in the CRN model. While most of the atoms in the 1a state are also four coordinated, there are some atoms whose coordination number is larger than 4. It is found that such atoms tend to distribute in the lower right region, deviated from the center of the distribution of the CRN model. Considering that the number of atoms in this region decrease by the annealing process, one of the effects by annealing process is to reduce the highly coordinated atoms. However, it should be noted that there are also many atoms whose coordination number is 4 in the same lower right region. Thus the low-dimensional space constructed by
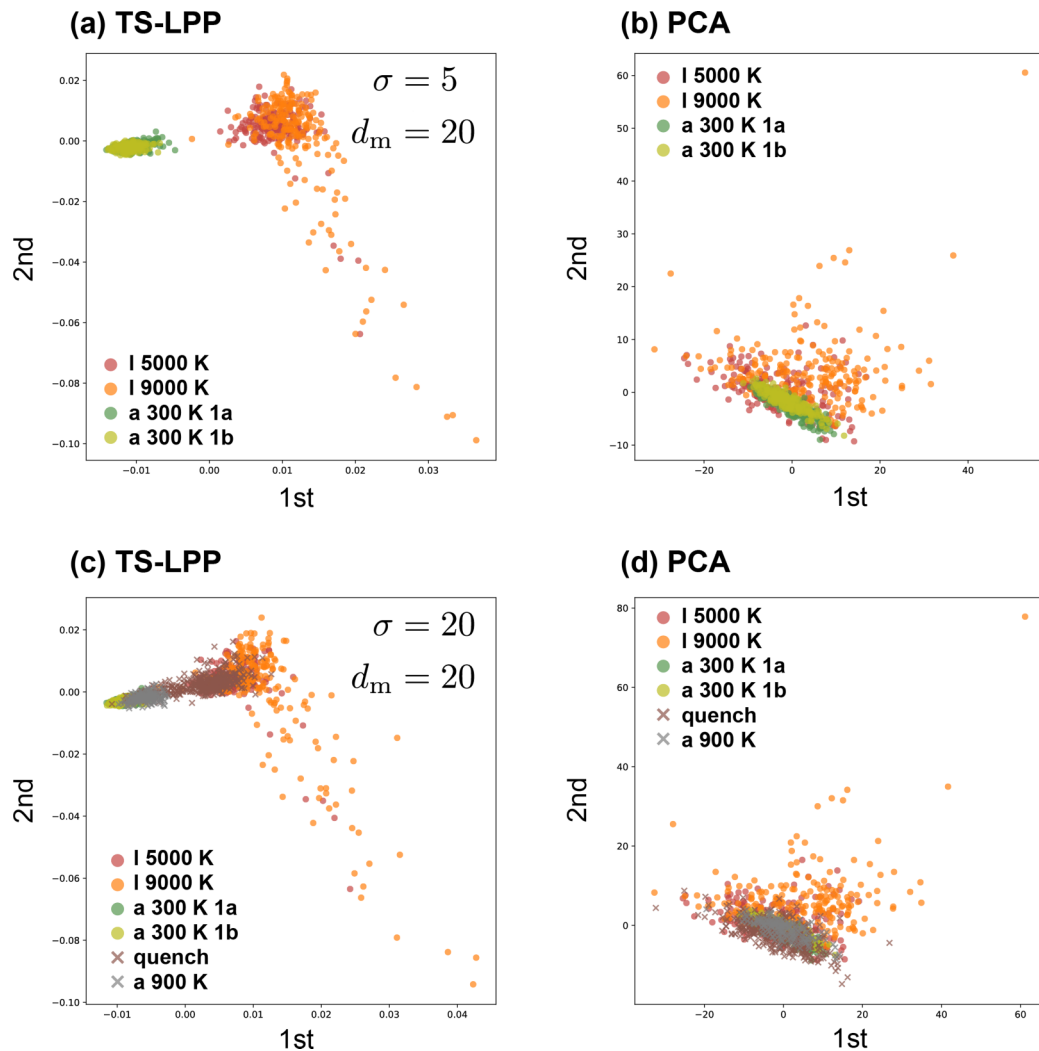
FIG. 11. Distributions of datapoints for liquid states at 5000 and 9000 K, amorphous states 1a and 1b at 300 K in the low-dimensional space constructed by (a) TS-LPP and (b) PCA, when $R_a = 4$ Å and $R_c = 5$ Å. Distributions of datapoints for quench states and amorphous states at 900 K in addition to the cases in (a) and (b) by (c) TS-LPP and (d) PCA.

the TS-LPP method is not simply related to the coordination number and we expect the present analysis would be useful to extract more detailed information in the structural change during the annealing process. Such attempts are now under way, and further reports will be presented in the future.

3. Analysis of the process from liquid to amorphous formation

So far, we have searched for low-dimensional spaces where crystalline, liquid, and amorphous states are well distinguished. Here, we address the structural analysis when there are only liquid and amorphous states. Figures 11(a) and 11(b) show the 2D space obtained by TS-LPP method and PCA, respectively, when liquid states at 5000 and 9000 K and amorphous states 1a and 1b at 300 K are adopted. In TS-LPP, it can be seen that the liquid and amorphous states can be separated on the 1st axis. Furthermore, the amorphous state 1a at 300 K is closer to the liquid states than to 1b. Although the number of clusters for determining hyperparameters is set to three, it is confirmed that even if the cluster number is increased, the obtained results are almost the same. In contrast, when using

PCA, all the results overlap, and it is difficult to extract the relations of atoms with each other.

Next, we generate a data set from the entire melt-quench process and consider the low-dimensional space created by the entire data set of the melt-quench process (see Fig. 4). The data set includes 200 points each from liquid states at 5000 and 9000 K, amorphous states 1a and 1b at 300 K, and amorphous state at 900 K. In addition, 400 points during the quench process from liquid state at 5000 K to amorphous state 1a at 300 K are also included. The results of the low-dimensional space obtained by the TS-LPP method and PCA are shown in Figs. 11(c) and 11(d). In this way, by the TS-LPP method, the liquid and amorphous states are separated even when the quench state is added, and the quench state is distributed to connect the liquid and amorphous states. Thus we can consider that the 1st axis is apparently related to the temperature. In contrast, in the PCA results, all the data are overlapped. Therefore it is shown that by using TS-LPP, we can find an appropriate low-dimensional space even if the entire melt-quench process is included in the data set.
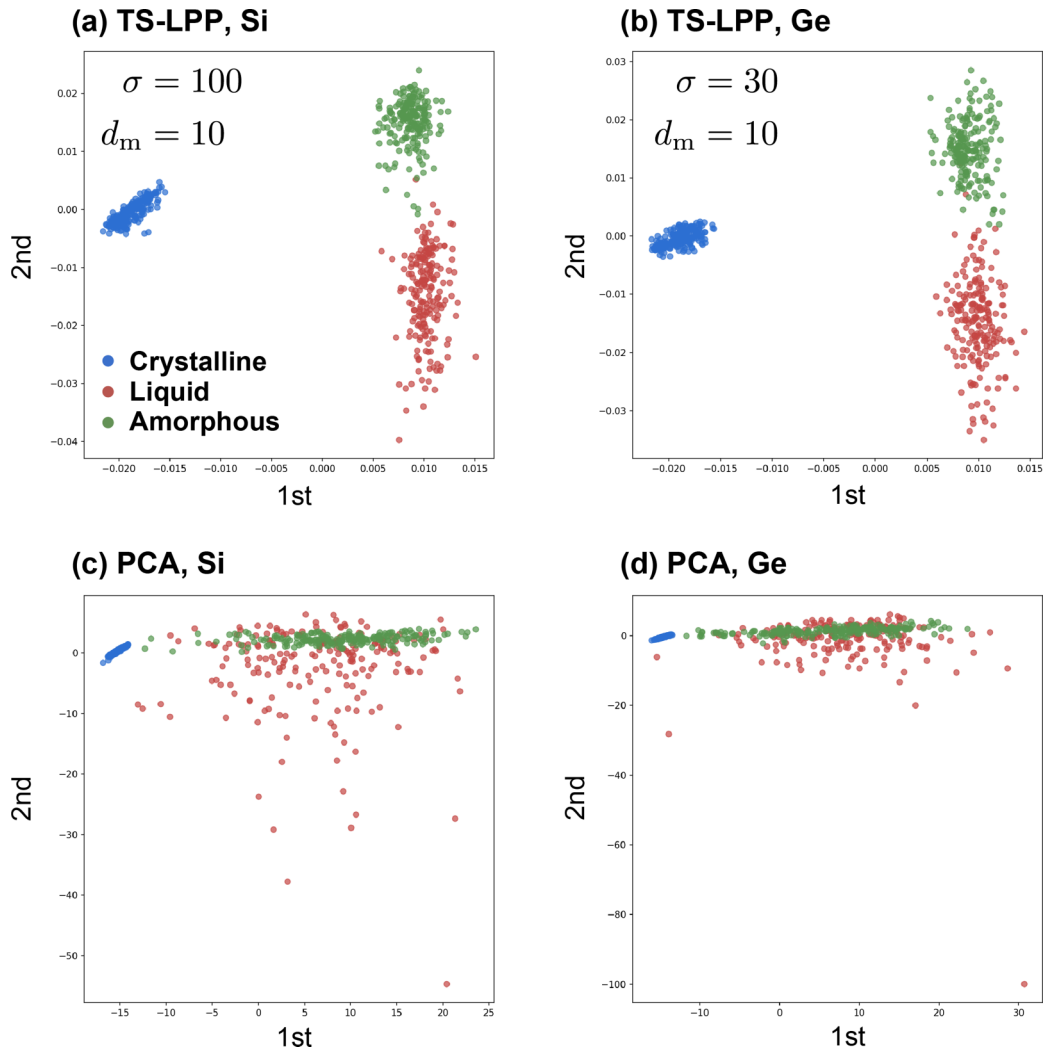
FIG. 12. Distributions after dimension reduction to $d_r = 2$ by the TS-LPP method for (a) silicon and (b) germanium. Distributions obtained by PCA for (c) silicon and (d) germanium. LAAF descriptors for binary systems with $R_a = 4$ Å and $R_c = 5$ Å are used.

C. Structural analysis of crystalline, liquid, and amorphous states in Si-Ge system

To confirm that our method is widely usable, we perform structural analysis for a silicon-germanium binary system in the crystalline, liquid, and amorphous phases, which are prepared by FPMD simulations using CONQUEST. To prepare crystalline and liquid phases, a system containing 500 silicon and 500 germanium atoms at a constant temperature of 300 K for crystalline and 5000 K for liquid and constant volume is simulated. The alloy where silicon and germanium atoms are arranged alternately in the lattice structure is considered as a crystalline state. On the other hand, the amorphous structures are simulated by a system with 550 silicon and 450 germanium atoms at 300 K prepared by the melt-quench process 2, as shown in Fig. 4.

Figure 12 shows the distributions of data points for silicon and germanium atoms in the 2D space ($d_r = 2$) by TS-LPP and PCA when $R_a = 4$ Å and $R_c = 5$ Å. For TS-LPP, the number of clusters for determining hyperparameters is set to three, but the cluster number is not an important factor. In the case of TS-LPP, the three states are well separated in both

silicon and germanium. The 1st axis divides the crystalline states from the liquid and amorphous states, while the liquid and amorphous states are separated by the second axis. This is the same result obtained in silicon single-component system (see Fig. 5). On the other hand, in the case of PCA, the liquid and amorphous states completely overlap, and the crystalline states are located near other states. Thus we conclude that even if a binary system is considered in a straightforward manner, the TS-LPP method can find a well-defined low-dimensional space, where the distributions of data points for each atom or group of atoms can be properly captured. This result indicates that the method has the potential to be applied to more complex systems.

D. Structure analysis for Cu system

So far, only covalent bond systems are considered. We next analyze the structures of metallic systems, fcc, hcp, and liquid states of the single component copper system. Note that fcc and hcp are both close-packed structures and it is difficult to distinguish them from the present LAAF descriptor, which includes only the radial term without the angle information
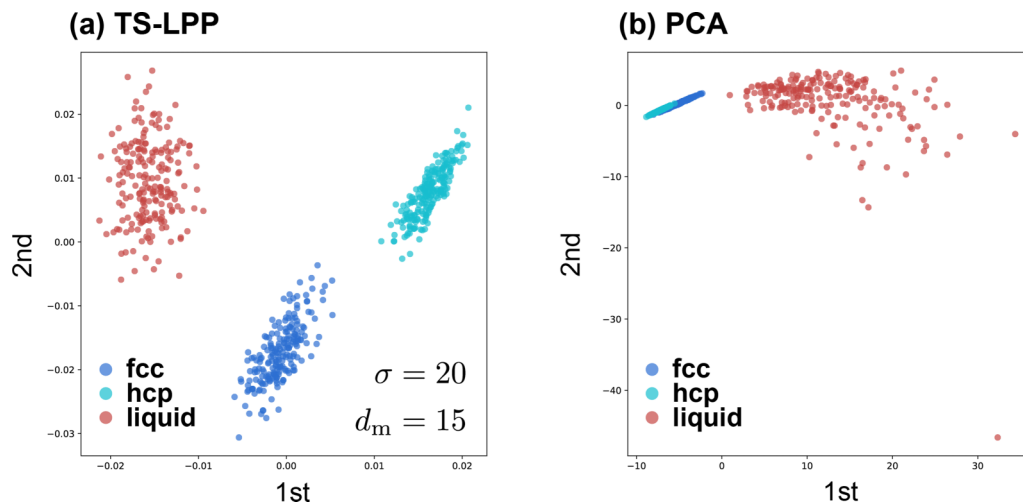
FIG. 13. Distributions in 2D space obtained by (a) TS-LPP and (b) PCA for $R_a = 3$ Å and $R_c = 3$ Å for Cu system. The fcc and hcp structures are prepared as a crystalline states.

of three atoms. We investigate the ability of TS-LPP method for this analysis. To generate the local structures of the copper system, classical MD simulations with the EMT potential are performed using the atomic simulation environment (ASE) package [54]. The number of atoms is 500. MD calculations are performed at a constant temperature of 300 K (fcc and hcp crystalline phases) and 5000 K (liquid) and constant volume. Data is generated by extracting 200 points from each state, and structural analysis is performed using 600 data points in total.

The obtained 2D spaces for $R_a = 3$ Å and $R_c = 3$ Å are shown in Figs. 13(a) and 13(b) when TS-LPP and PCA are used. For TS-LPP, the number of clusters for determining hyperparameters is set to three, although we confirm that the distributions are almost the same even when the cluster number is changed. In both TS-LPP and PCA, it can be seen that the crystalline 300 K and liquid 5000 K states are separated. The fcc and hcp data can be distinguished by TS-LPP, while these data are overlapped in PCA result.

IV. DISCUSSION AND SUMMARY

In summary, we have proposed a novel structural analysis method based on an unsupervised ML technique and applied it to structural analysis problems. Our method can find a low-dimensional space where the distributions of data points for each atom or group of atoms can be properly captured using the proposed LAAF descriptor showing the local structure without labels. The key to our method is to perform dimensionality reduction from a high-dimensional descriptor space to a lower dimension using TS-LPP, where conventional LPP is performed twice. Our method is much more reliable than PCA or other conventional dimensionality reduction methods. Furthermore, our method is a parameter-free unsupervised technique, since the hyperparameters in the method can be determined automatically. Our implementation of the TS-LPP method is available on GitHub [68]. We have used the method to analyze the FPMD simulations of a silicon single-component system, silicon and germanium binary system, and classical MD simulations for copper single-component system. Through these applications, we have demonstrated

that the method is effective for understanding the phenomena in detail and for elucidating the local phenomena during the simulations.

We believe that the new method is useful in many kinds of applications. As the method is unsupervised, it can be used to analyze the structures of unknown phases or materials. In addition, we can control the resolution of the structural analysis depending on the degree of dimension. Here, we discuss two possible applications. The first application is the development of highly accurate and transferable ML force fields. In many cases, the problem of the ML forces is the transferability [18,69]. If we can construct the force fields from multiple ML forces based on different sets of training data, depending on the local environment, the accuracy of ML forces can be improved easily. The classification method proposed in this study should be effective for this purpose as well as the Gaussian mixture model [13]. For example, in many systems that include irregular regions, such as surface or interface structures, there are structural variations, and it is difficult to determine how the atoms in different regions should be classified. Even in such cases, since our method is unsupervised, we can simply collect the data from a short FPMD simulation and classify the atoms into several groups depending on the data structure. Note that to obtain the low-dimensional space by our method, an eigenvalue problem depending on the dimension of the descriptors should be solved. Since the dimension of the descriptor is usually not large, the computational cost of this eigenvalue problem is not expensive. Furthermore, the cost can be further reduced as the TS-LPP does not require all eigenvalues or eigenvectors. If we construct the ML forces based on this classification, the accuracy of the ML forces should be high. The proposed method can also detect a new structures during a long MD simulation using the constructed ML forces. Therefore on-the-fly development of ML forces [70–73] of large, irregular, and complex systems should be possible with a large-scale DFT code, such as the CONQUEST code that we used in this work. It should be noted that large systems can be treated using the recent active-learning algorithms [70,73,74], even without large-scale DFT codes [74].

The second application is to find a useful coordinate to characterize the local phenomena in complex dynamical processes, chemical reactions, or phase transformations. As we have pointed out in the analysis of the melt-quench process, the component in the low-dimensional space can be considered as a good reaction coordinate to express the "local" transformation from a liquid to an amorphous state. Note that although the present descriptor is "local" and specific to each atom, the descriptor also includes the information of its neighbors. With the LAAF descriptor, we can follow the phenomenon of a group of atoms. If TS-LPP with the LAAF descriptor can find useful coordinates for many kinds of phenomena, we expect that it can be helpful in calculating the energy barrier for the local reactions or for the efficient structural search of new phases. We intend to explore these two applications in future work.

APPENDIX: LOW-RANK APPROXIMATION FOR LPP

For our target data, the input data matrix X has a number of small nonzero singular values. Thus, to improve the numerical stability for calculations in LPP, we use the low-rank approximation of X using the singular value decomposition $U \Sigma V^\top = X$. Here, U is $N \times M$, the columns of which are the left singular vectors of X. $\Sigma = \text{diag}(s_1, s_2, \ldots, s_M)$ is an $M \times M$ diagonal matrix with the descending ordered singular values $\{s_i\}_{i=1,\ldots,M}$ of X. V is $N \times M$ matrix, the columns of which are the right singular vectors of X. By introducing a parameter $0 < \delta < 1$, the numerical rank of ρ is determined by:

$$\frac{s_{\rho+1}}{s_1} < \delta < \frac{s_\rho}{s_1}. \tag{A1}$$

Let U_0 and V_0 be matrices with the ρ leftmost columns of U and V, respectively, and Σ_0 be the leading principal submatrix of order ρ of Σ. Using these matrices, instead of solving Eq. (13), we solve the generalized eigenvalue problem as:

$$U_0^\top L U_0 \mathbf{z} = \lambda U_0^\top D U_0 \mathbf{z}, \tag{A2}$$

where $\{\lambda_i, \mathbf{z}_i\}$ is the ith eigenvalue and eigenvector. The alternative mapping matrix $Y = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{d_r})$ is obtained by

$$\mathbf{y}_i = V_0 \Sigma_0^{-1} \mathbf{z}_i, \quad (i = 1, \ldots, d_r). \tag{A3}$$

[1] R. Car and M. Parrinello, Unified Approach for Molecular Dynamics and Density-Functional Theory, Phys. Rev. Lett. **55**, 2471 (1985).

[2] R. Car and M. Parrinello, Structural, Dymanical, and Electronic Properties of Amorphous Silicon: An *Ab Initio* Molecular-Dynamics Study, Phys. Rev. Lett. **60**, 204 (1988).

[3] P. E. Blöchl and M. Parrinello, Adiabaticity in first-principles molecular dynamics, Phys. Rev. B **45**, 9413 (1992).

[4] A. Pasquarello, M. S. Hybertsen, and R. Car, Interface structure between silicon and its oxide by first-principles molecular dynamics, Nature (London) **396**, 58 (1998).

[5] A. R. Oganov, J. P. Brodholt, and G. D. Price, The elastic constants of MgSiO₃ perovskite at pressures and temperatures of the Earth's mantle, Nature (London) **411**, 934 (2001).

[6] A. Hirata, P. Guan, T. Fujita, Y. Hirotsu, A. Inoue, A. R. Yavari, T. Sakurai, and M. Chen, Direct observation of local atomic order in a metallic glass, Nat. Mater. **10**, 28 (2011).

[7] X. Sang, X. Li, W. Zhao, J. Dong, C. M. Rouleau, D. B. Geohegan, F. Ding, K. Xiao, and R. R. Unocic, *In situ* edge engineering in two-dimensional transition metal dichalcogenides, Nat. Commun. **9**, 2051 (2018).

[8] R. Réocreux, É. Girel, P. Clabaut, A. Tuel, M. Besson, A. Chaumonnot, A. Cabiac, P. Sautet, and C. Michel, Reactivity of shape-controlled crystals and metadynamics simulations locate the weak spots of alumina in water, Nat. Commun. **10**, 3139 (2019).

[9] L. Shi, A. Xu, D. Pan, and T. Zhao, Aqueous proton-selective conduction across two-dimensional graphyne, Nat. Commun. **10**, 1165 (2019).

[10] D. R. Bowler and T. Miyazaki, O(N) methods in electronic structure calculations, Rep. Prog. Phys. **75**, 036503 (2012).

[11] J. Behler and M. Parrinello, Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, Phys. Rev. Lett. **98**, 146401 (2007).

[12] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, J. Chem. Phys. **134**, 074106 (2011).

[13] T. L. Pham, H. Kino, K. Terakura, T. Miyake, and H. C. Dam, Novel mixture model for the representation of potential energy surfaces, J. Chem. Phys. **145**, 154103 (2016).

[14] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, Machine learning unifies the modeling of materials and molecules, Sci. Adv. **3**, e1701816 (2017).

[15] T. Suzuki, R. Tamura, and T. Miyazaki, Machine learning for atomic forces in a crystalline solid: Transferability to various temperatures, Int. J. Quantum Chem. **117**, 33 (2017).

[16] C. Chen, Z. Deng, R. Tran, H. Tang, I.-H. Chu, and S. P. Ong, Accurate force field for molybdenum by machine learning large materials data, Phys. Rev. Materials **1**, 043603 (2017).

[17] R. Kobayashi, D. Giofré, T. Junge, M. Ceriotti, and W. A. Curtin, Neural network potential for Al-Mg-Si alloys, Phys. Rev. Materials **1**, 053604 (2017).

[18] R. Tamura, J. Lin, and T. Miyazaki, Machine learning forces trained by Gaussian process in liquid states: Transferability to temperature and pressure, J. Phys. Soc. Jpn. **88**, 044601 (2019).

[19] W. Li and Y. Ando, Dependence of a cooling rate on structural and vibrational properties of amorphous silicon: A neural network potential-based molecular dynamics study, J. Chem. Phys. **151**, 114101 (2019).

[20] Z. Li, J. R. Kermode, and A. De Vita, Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces, Phys. Rev. Lett. **114**, 096405 (2015).

[21] V. L. Deringer, N. Bernstein, A. P. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, and G. Csányi, Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics, J. Phys. Chem. Lett. **9**, 2879 (2018).

[22] F. H. Stillinger and T. A. Weber, Hidden structure in liquids, Phys. Rev. A **25**, 978 (1982).

[23] S.-H. Lee, J. Kim, S.-J. Kim, S. Kim, and G.-S. Park, Hidden Structural Order in Orthorhombic Ta_2O_5, Phys. Rev. Lett. **110**, 235502 (2013).

[24] H. Tong and H. Tanaka, Revealing Hidden Structural Order Controlling Both Fast and Slow Glassy Dynamics in Supercooled Liquids, Phys. Rev. X **8**, 011041 (2018).

[25] R. L. McGreevy and L. Pusztai, Reverse monte carlo simulation: A new technique for the determination of disordered structures, Mol. Simul. **1**, 359 (1988).

[26] F. Li and J. S. Lannin, Radial Distribution Function of Amorphous Carbon, Phys. Rev. Lett. **65**, 1905 (1990).

[27] N. Lačević, F. W. Starr, T. B. Schrøder, V. N. Novikov, and S. C. Glotzer, Growing correlation length on cooling below the onset of caging in a simulated glass-forming liquid, Phys. Rev. E **66**, 030101(R) (2002).

[28] Y. Hiraoka, T. Nakamura, A. Hirata, E. G. Escolar, K. Matsue, and Y. Nishiura, Hierarchical structures of amorphous solids characterized by persistent homology, Proc. Natl. Acad. Sci. USA **113**, 7035 (2016).

[29] M. Saadatfar, H. Takeuchi, V. Robins, N. Francois, and Y. Hiraoka, Pore configuration landscape of granular crystallization, Nat. Commun. **8**, 15082 (2017).

[30] Y. Onodera, S. Kohara, S. Tahara, A. Masuno, H. Inoue, M. Shiga, A. Hirata, K. Tsuchiya, Y. Hiraoka, I. Obayashi, K. Ohara, A. Mizuno, and O. Sakata, Understanding diffraction patterns of glassy, liquid and amorphous materials via persistent homology analyses, J. Ceram. Soc. Japan **127**, 853 (2019).

[31] W. Lechner and C. Dellago, Accurate determination of crystal structures based on averaged local bond order parameters, J. Chem. Phys. **129**, 114707 (2008).

[32] P. Geiger and C. Dellago, Neural networks for local structure detection in polymorphic systems, J. Chem. Phys. **139**, 164105 (2013).

[33] P. M. Piaggi and M. Parrinello, Entropy based fingerprint for local crystalline order, J. Chem. Phys. **147**, 114112 (2017).

[34] M. A. Caro, A. Aarva, V. L. Deringer, G. Csányi, and T. Laurila, Reactivity of amorphous carbon surfaces: Rationalizing the role of structural motifs in functionalization using machine learning, Chem. Mater. **30**, 7446 (2018).

[35] E. Boattini, M. Ram, F. Smallenburg, and L. Filion, Neural-network-based order parameters for classification of binary hard-sphere crystal structures, Mol. Phys. **116**, 3066 (2018).

[36] E. Boattini, M. Dijkstra, and L. Filion, Unsupervised learning for local structure detection in colloidal systems, J. Chem. Phys. **151**, 154901 (2019).

[37] E. Boattini, S. Marín-Aguilar, S. Mitra, G. Foffi, F. Smallenburg, and L. Filion, Autonomously revealing hidden local structures in supercooled liquids, Nat. Commun. **11**, 5479 (2020).

[38] M. Ceriotti, G. A. Tribello, and M. Parrinello, Simplifying the representation of complex free-energy landscapes using sketch-map, Proc. Natl. Acad. Sci. USA **108**, 13023 (2011).

[39] G. A. Tribello, M. Ceriotti, and M. Parrinello, Using sketch-map coordinates to analyze and bias molecular dynamics simulations, Proc. Natl. Acad. Sci. USA **109**, 5196 (2012).

[40] X. He and P. Niyogi, Locality preserving projections, in *Advances in Neural Information Processing Systems 16*, edited by S. Thrun, L. K. Saul, and B. Schölkopf (MIT Press, 2004), pp. 153–160.

[41] V. Botu and R. Ramprasad, Adaptive machine learning framework to accelerate ab initio molecular dynamics, Int. J. Quantum Chem. **115**, 1074 (2015).

[42] U. von Luxburg, A tutorial on spectral clustering, Stat. Comput. **17**, 395 (2007).

[43] T. Calinski and J. Harabasz, A dendrite method for cluster analysis, Commun. Stati. Theory Methods **3**, 1 (1974).

[44] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html.

[45] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[46] http://www.order-n.org/.

[47] A. Nakata, J. S. Baker, S. Y. Mujahed, J. T. L. Poulton, S. Arapan, J. Lin, Z. Raza, S. Yadav, L. Truflandier, T. Miyazaki, and D. R. Bowler, Large scale and linear scaling dft with the conquest code, J. Chem. Phys. **152**, 164112 (2020).

[48] D. R. Bowler, T. Miyazaki, and M. J. Gillan, Recent progress in linear scaling ab initio electronic structure techniques, J. Phys.: Condens. Matter **14**, 2781 (2002).

[49] T. Miyazaki, D. R. Bowler, R. Choudhury, and M. J. Gillan, Atomic force algorithms in density functional theory electronic-structure techniques based on local orbitals, J. Chem. Phys. **121**, 6186 (2004).

[50] D. R. Bowler, R. Choudhury, M. J. Gillan, and T. Miyazaki, Recent progress with large-scale *ab initio* calculations: the CONQUEST code, Phys. Stat. Sol. (b) **243**, 989 (2006).

[51] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal, The SIESTA method for ab initio order-N materials simulation, J. Phys.: Condens. Matter **14**, 2745 (2002).

[52] M. Arita, D. R. Bowler, and T. Miyazaki, Stable and efficient linear scaling first-principles molecular dynamics for 10000+ atoms, J. Chem. Theory Comput. **10**, 5419 (2014).

[53] T. Hirakawa, T. Suzuki, D. R. Bowler, and T. Miyazaki, Canonical-ensemble extended Lagrangian Born–Oppenheimer molecular dynamics for the linear scaling density functional theory, J. Phys.: Condens. Matter **29**, 405901 (2017).

[54] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson *et al.*, The atomic simulation environment—a Python library for working with atoms, J. Phys.: Condens. Matter **29**, 273002 (2017).

[55] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevB.105.075107 for Fig. S1. Radial distribution functions; Fig. S2. Distributions after dimension reduction by the PCA; Fig. S3. Distributions after dimension reduction by the LPP; Figure S4. Distributions after dimension reduction by different dimensionality reduction methods; supplementary movie 1. Time dependence for the entire melt-quench process 1; supplementary Mmovie 2. Time dependence for the entire melt-quench process 2.

[56] M. Ishimaru, Atomistic simulations of structural relaxation processes in amorphous silicon, J. Appl. Phys. **91**, 686 (2001).

[57] F. Alvarez, C. C. Díaz, A. A. Valladares, and R. M. Valladares, Radial distribution functions of *ab initio* generated amorphous covalent networks, Phys. Rev. B **65**, 113108 (2002).

[58] P. Ganesh and M. Widom, Liquid-Liquid Transition in Supercooled Silicon Determined by First-Principles Simulation, Phys. Rev. Lett. **102**, 075701 (2009).

[59] J. Wang, F. Fang, and X. Zhang, Nanometric cutting of silicon with an amorphous-crystalline layered structure: A molecular dynamics study, Nanoscale Res. Lett. **12**, 41 (2017).

[60] A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, Machine Learning a General-Purpose Interatomic Potential for Silicon, Phys. Rev. X **8**, 041048 (2018).

[61] L. Lin, J. Frelinger, W. Jiang, G. Finak, C. Seshadri, P.-A. Bart, G. Pantaleo, J. McElrath, S. DeRosa, and R. Gottardo, Identification and visualization of multidimensional antigen-specific T-cell populations in polychromatic cytometry data, Cytometry A **87**, 675 (2015).

[62] D. Lu, C. Ding, J. Xu, and S. Wang, Hierarchical discriminant analysis, Sensors **18**, 279 (2018).

[63] D. Kobak and P. Berens, The art of using t-SNE for single-cell transcriptomics, Nat. Commun. **10**, 5416 (2019).

[64] S. Sakaue, J. Hirata, M. Kanai, K. Suzuki, M. Akiyama, C. Lai Too, T. Arayssi, M. Hammoudeh, S. Al Emadi, B. K. Masri, H. Halabi, I. W. Uthman, R. Saxena, L. Padyukov, M. Hirata, K. Matsuda, Y. Murakami, Y. Kamatani, and Y. Okada, Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction, Nat. Commun. **11**, 1569 (2020).

[65] W. H. Zachariasen, The atomic arrangement in glass, J. Am. Chem. Soc. **54**, 3841 (1932).

[66] F. Wooten, K. Winer, and D. Weaire, Computer Generation of Structural Models of Smorphous Si and Ge, Phys. Rev. Lett. **54**, 1392 (1985).

[67] G. T. Barkema and N. Mousseau, High-quality continuous random networks, Phys. Rev. B **62**, 4985 (2000).

[68] https://github.com/rtmr/TS-LPP.

[69] N. Kuritz, G. Gordon, and A. Natan, Size and temperature transferability of direct and local deep neural networks for atomic forces, Phys. Rev. B **98**, 094109 (2018).

[70] E. V. Podryabinkin and A. V. Shapeev, Active learning of linearly parametrized interatomic potentials, Comput. Mater. Sci. **140**, 171 (2017).

[71] R. Jinnouchi, J. Lahnsteiner, F. Karsai, G. Kresse, and M. Bokdam, Phase Transitions of Hybrid Perovskites Simulated by Machine-Learning Force Fields Trained on the Fly with Bayesian Inference, Phys. Rev. Lett. **122**, 225701 (2019).

[72] R. Jinnouchi, F. Karsai, and G. Kresse, On-the-fly machine learning force field generation: Application to melting points, Phys. Rev. B **100**, 014105 (2019).

[73] J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, and B. Kozinsky, On-the-fly active learning of interpretable bayesian force fields for atomistic rare events, npj Comput. Mater. **6**, 20 (2020).

[74] M. Hodapp and A. Shapeev, In operando active learning of interatomic interaction during large-scale simulations, Mach. Learn.: Sci. Technol. **1**, 045005 (2020).