Hierarchical Clustering of Structural and Electronic Characteristics Obtained from Molecular Dynamics Simulation of Catalytic Reaction on Metal Nanoparticle

Monami Tsunawaki, Satoru Fukuhara and Yasushi Shibuta*

Department of Materials Engineering, The University of Tokyo, Tokyo 113-8656, Japan

Unsupervised machine learning (ML) is examined for the result of molecular dynamics (MD) simulation to extract characteristics of catalytic reaction. O–H bond dissociation of ethanol on Fe–Co nanoparticle in ab initio MD simulation [S. Fukuhara et al., Chem. Phys. Lett. 731 (2019) 136619] is employed as an example. Hierarchical clustering of radial distribution function successfully classifies coordinates on reaction in the dendrogram. Moreover, receiver operating characteristic curve reveals the distance to the farthest-neighbor atom from the target atom is a dominant descriptor for the clustering. An optimum structure of catalytic nanoparticle is predicted based on these automated analyses. This study shows a new way of post-process of results of MD simulations based on the unsupervised learning technique and it paves the way for a new possibility of ML-based materials design. [doi:10.2320/matertrans.MT-M2021032]

(Received February 25, 2021; Accepted March 29, 2021; Published May 14, 2021)

Keywords: molecular dynamics simulation, unsupervised machine learning, hierarchical clustering, descriptor extraction, catalytic reaction

1. Introduction

Thanks to the recent progress in high-performance computing, length scale of atomistic simulation is drastically expanding.¹ Now, more than billion atoms can be handled by large-scale molecular dynamics (MD) simulations performed on recent supercomputer.² Since only coordinates and velocity of atoms in the system (and electronic state from ab initio MD) are obtained from MD simulation, it is essential to translate primary data of MD simulation into collective properties of target materials in the post-analyses. For example, there are several established techniques for the identification of crystallinity of atoms such as common neighbor analysis³,⁴ and polyhedral template matching.⁵ On the other hand, it is not straightforward to analyze rare events in MD simulation including solid diffusion, phase transition and chemical reaction. In general, energy transition along kinetic path of the rare event is manually examined by an additional energy calculation in conjunction with nudged elastic band (NEB) method.⁶ Moreover, recent metadynamics-based techniques enable to estimate kinetic path and activation energy from dynamic viewpoints⁷–⁹ and also enable to accelerate MD simulation for specific reactions.¹⁰ However, manual analysis of specific rare event does not cover whole phase space of the simulation and therefore it does not show an overall picture of phenomena in the simulation in most case. Moreover, it takes a lot of costs and efforts to analyze primary data of MD simulation manually as the number of atoms increases, which is a pressing problem at present.

Therefore, automated analysis of the result of atomistic simulations has been widely considered on the basis of machine learning (ML)-based techniques.¹¹ In general, there are three categories in ML: supervised learning,¹² unsupervised learning¹³ and reinforced learning.¹⁴ In supervised learning, input and output data are labeled for classification to provide a learning basis for classifying future data. Development of the neural network potential¹⁵ is one of representative applications of supervised learning related to MD simulation. Moreover, we have performed automated analysis of local atomic configuration for the MD simulation of solid-liquid interface using convolutional neural network, in which bulk structures of each phase were employed as training data.¹⁶ On the other hand, main purposes of unsupervised learning are clustering and dimensionality reduction for unlabeled data. For example, unsupervised learning is used for clustering atomistic configurations of deoxyribonucleic acid (DNA),¹⁷ proteins¹⁸ and other bio molecules¹⁹ from MD simulations. For more information, various applications of ML on atomistic simulations are reviewed in recent literatures.¹¹,²⁰

In this study, we aim to clarify unique characteristics of catalytic-reactive site in metal nanoparticles from the results of ab initio MD simulation based on unsupervised learning. In our previous study of ethanol dissociation on a Fe–Co binary nanoparticle by ab initio MD simulation,²¹ there were many dissociation reactions including C–C, C–O and O–H bond dissociations on the Fe–Co nanoparticle. Based on our manual NEB analysis in conjunction with the d-band theory,²² we concluded that not only the electronic state of whole catalyst but also the local arrangement of atoms and the local electronic state at the reaction site were important for the reaction. However, it was not clear whether above discussion covered whole reaction processes in the simulation or not since several dominant reactions were manually picked up for the discussion. Therefore, we came up with the idea that automated clustering of results from MD simulation could extract unique features related to specific dissociation reaction. To this end, unsupervised learning based on clustering technique is performed for the results of MD simulation in this study. Specifically, O–H bond dissociation of ethanol on Fe–Co nanoparticle is employed as a study example of the unsupervised learning since O–H bond dissociation occurred most frequently compared to other dissociation reactions such as C–C and C–O bonds.²¹ Then, the detail of O–H bond dissociation reaction is discussed based on the receiver operating characteristic (ROC) curve.²³ Finally, optimum structure of catalytic nanoparticle is predicted based on the results of these automated analyses.

*Corresponding author, E-mail: shibuta@material.t.u-tokyo.ac.jp
2. Methodology and Data Preparation

2.1 Hierarchical clustering

Clustering is a popular technique for grouping a set of data in the several group (called a cluster) based on the similarity between data. There are many clustering algorithms such as centroid-based (k-means) clustering, density based clustering, hierarchical clustering and so on. In this study, hierarchical clustering is employed since it can provide an extensive hierarchy of clusters using a dendrogram. Figure 1 shows schematic image of hierarchical clustering with data points A to F. Standard algorithm of hierarchical clustering is as follows. At first, each data point is regarded as a cluster. The distance between clusters are calculated and the closest clusters are combined into one cluster. The schematic image, data D and E are combined (Fig. 1-1). The distance between newly defined clusters are again calculated and the closest clusters are combined into one cluster (Fig. 1-2). These processes are continued until all cluster are combined (Figs. 1-3, 1-4). Hierarchy of the relationship in each data point is visualized using a tree diagram called dendrogram. Vertical axis of the dendrogram represents the degree of dissimilarity. That is, similarity between clusters becomes lower as two clusters are connected at higher point in the dendrogram. Arbitrary clustering with a selected number of clusters is obtained by cutting the dendrogram at a certain height. For example, a clustering with \( k = 2 \) (cluster 1: A and B, cluster 2: C, D, E and F) is obtained by setting a threshold height of dissimilarity as \( L \) in Fig. 1-4.

A distance between data points should be defined in some way in order to apply the clustering technique to a practical data set. The distance between two data points can be directly defined by the Euclidean distance as far as data points are explicitly distributed in \( n \)-dimensional space. However, it is not intuitive to define the distance between distribution functions such as probability distribution and spectrum. In general, the similarity is used to discuss the distance between distribution functions. There are several definitions of the similarity including cosine distance, Pearson product-moment correlation coefficient and Wasserstein distance (also called the Earth Mover’s Distance, EMD). In this study, the Wasserstein distance is employed to estimate the similarity in distributions of radial distribution function (RDF) and electronic density of states (DOS) derived from coordinates and electronic structures of \( ab \) \( initio \) MD simulation. Wasserstein distance is one of established techniques for the estimation of similarity in distribution functions. Details on the Wasserstein distance is described in Appendix A. In addition, a distance between clusters (called the clustering linkage) should be defined in the hierarchical clustering. In this study, group average is employed, in which the average of distance between arbitrary two points from each cluster is defined as the distance between clusters.

2.2 Data preparation for hierarchical clustering

Although primary data (coordinates, velocity of atoms and electronic state) of the MD simulation provide a complete description of the simulation result, they have not been used in ML scheme as they are. They are usually translated into some representative properties. In this study, coordinates and electronic states for all steps of \( ab \) \( initio \) MD simulation of ethanol dissociation on Fe-Co binary nanoparticle were translated into RDF and DOS, respectively. Then, data set of RDF and DOS are classified into several clusters to extract unique characteristics of O–H bond dissociation reaction. Here, we briefly explain our previous study of \( ab \) \( initio \) MD simulation, which was aiming to understand mechanism of metal-catalyzed carbon nanotube growth.

As the calculation condition, 30 ethanol molecules and an Fe\(_{16}\)Co\(_{16}\) nanoparticle were placed into a cubic cell with \( 15 \times 15 \times 15 \) \( \text{Å}^3 \) and the system was kept at 1500 K during 20000 steps with time step of 0.242 fs (i.e., 10 a.u.) to examine ethanol dissociation on the nanoparticle. The projector augmented wave (PAW) method within the framework of the density functional theory (DFT) was employed for the calculation of the electronic states. The Perdew–Burke–Ernzerhof (PBE) generalized gradient approximation (GGA) with spin polarization was used for the exchange-correlation energy. The electronic wave functions were expanded in an atomic-orbital basis set on the basis of a formulation generalized to the PAW method. There were many dissociation reactions of ethanol including C–C, C–O, C–H and O–H bonds dissociations. Among them, O–H bond dissociation is the most frequently found. The O–H dissociation found in the simulation generally consisted of three stages. First, ethanol molecule drifted in gas phase...
(stage 1) and it attached on the nanoparticle by forming O–M (M: metal atom, Fe or Co) bond (stage 2). After a short duration, O–H bond in ethanol dissociated on the metal nanoparticle (stage 3) while O–M bond is kept binding during the process. It is considered that the metal atom bonding to O atom in ethanol plays a key role for the O–H bond dissociation both from geometric and electronic viewpoints. Therefore, RDF and DOS for the metal atom connecting with O atom (denoted as target atoms) are employed for hierarchical clustering. RDF of all metal atoms in the nanoparticle from the target atom are examined. The target atom is not limited to either Fe or Co atom since there were no significant difference in the feature in Fe and Co by our preliminary check as far as examining RDF and DOS in above way. Note that there are many other choices of representative properties for the hierarchical clustering. We choose them as representative properties as an example of a sprouting study.

Figure 2 shows flowchart of hierarchical clustering in this study. Snapshots of every 20 steps in stage 2 are extracted and time average of RDF and electronic DOS are used for hierarchical clustering. O–H bond dissociation is defined to occur when O–H bond length becomes larger than 1.5 Å since O–H bond in ethanol generally fluctuates between 0.9 and 1.1 Å in our simulation. O–H bond dissociation for all (i.e., 30) ethanol molecules are examined. The final snapshot in stage 2 for each dissociation reaction is labeled as the “coordinate on reaction”. According above criterion, 2451 coordinates are regarded as ones in stage 2 and 12 out of 2451 coordinates are labeled as “coordinate on reaction”. RDF and DOS from 2451 coordinates are separately classified by hierarchical clustering according to the methodology in Section 2.

3. Results and Discussions

Figure 3 shows the dendrogram of the hierarchical
clustering of DOS and RDF from 2451 snapshots of the \textit{ab initio} MD simulation. Vertical axis represents the Wasserstein distance showing the degree of dissimilarity. 2451 DOS are classified into 8 clusters with the threshold distance of 0.015. The same number of RDF are classified into 6 clusters with the threshold distance of 0.039. Interestingly, 9 out of 12 coordinates on reaction are classified into the specific cluster (i.e., Cluster 6) in the dendrogram of RDF. On the other hand, coordinates on reaction do not gather at a specific cluster but distribute in 5 clusters in the dendrogram of DOS. Figure 4 shows DOS and RDF of two representative coordinates indicated by arrows (i) and (ii) in Fig. 3. Coordinate (ii) is a representative coordinate on reaction and coordination (i) is chosen from other than the coordinates on reaction. It should be stressed that both coordinates belong to the same cluster (Cluster 8) in the dendrogram of DOS whereas these are classified into different clusters (Clusters 5 and 6) in the dendrogram of RDF. Indeed, there is a similar trend in peaks of DOS in these two cases. That is, DOS of up spin has a large peak below Fermi level with a small peak at right shoulder and that of down spin has a large peak above Fermi level with a small peak a left shoulder. On the other hand, there is a different trend in peaks of RDF. That is, the second peak is the smallest among three typical peaks in RDF of Fe atoms of coordinate (ii) whereas peaks become small monotonically with increasing radius in the case of coordinate (i).

As far as comparing these two factors, RDF seems to be the dominant characteristic for O–H bond dissociation since most of RDF of the coordinates on reaction are classified into the specific cluster as shown in Fig. 3(b). On the other hand, DOS of the coordinates on reaction do not gather on a specific cluster in the dendrogram as shown above. Therefore, it is considered there is no strong correlation between clusters in dendrograms of DOS and RDF. In general, cross correlation in clusters defined by different features is discussed using a heat map. Figure 5 shows heat maps of cross correlation between clusters in the dendrograms of DOS and RDF in Fig. 3 for all 2451 coordinates and for only coordinates on reaction (i.e., 12 coordinates), respectively. Labels of cluster correspond to those in Fig. 3. There is some moderate correlation between DOS and RDF clusters in the heat map of all 2451 coordinates. For example, majority of coordinates in Cluster 8 of the dendrogram of DOS belong to Clusters 5 and 6 of the dendrogram of RDF. However, many other coordinates in other DOS clusters also belong to Clusters 5 or 6 of the dendrogram of RDF. Therefore, there is no clear one-to-one correlation between DOS and RDF clusters as far as considering all coordinates. Especially, difference in coordinates of Clusters 5 and 6 of the dendrogram of RDF is not correlated with the classification of DOS in the heat map of all coordinates. On the other hand, most of coordinates on reaction are classified into Cluster 6 of the dendrogram of RDF and there is no coordinate on reaction classified into Cluster 5 of the dendrogram of RDF. Therefore, it is considered that RDF is more suitable characteristic to extract coordinate on reaction than DOS as far as classifying our MD result. Although there are many other choices of representative properties for the hierarchical clustering, it is significant that coordinates with the specific feature (i.e., O–H bond dissociation in this case) in the MD simulation are successfully extracted by the unsupervised learning technique.

Now, most of the coordinates on reaction are successfully classified into Cluster 6 in the dendrogram of RDF by the automated analysis of hierarchical clustering. However, it is not clear what factor (called the descriptor) plays a major role for the classification of the coordinates. Therefore, separate evaluation of descriptors is needed to extract the descriptor for the target classification. Here, a descriptor dividing
Cluster 5 and 6 in the dendrogram of RDF is discussed based on ROC curve. ROC curve is a major tool to visualize their performance of binary classification. The basic concept of ROC curve is summarized in Appendix B. In this study, following three descriptors are listed as candidate classifiers for dividing coordinates in Cluster 5 and 6 of the dendrogram of RDF: (i) distance to the farthest-neighbor atom from the target atom, (ii) distance to the nearest-neighbor atom from the target atom and (iii) coordination number of the second-neighbor atom. All coordinates belong to Cluster 5 (defined as negative) and Cluster 6 (defined as positive) in the dendrogram of RDF are classified into two groups using these descriptors. True positive rate (TPR) and false positive rate (FPR) are calculated by changing threshold values for each descriptor and ROC curves for three descriptors are drawn as shown in Fig. 6. ROC curves for cases (i) and (iii) become convex upward whereas that for the case (ii) almost fits to the diagonal line of the graph area. Area under ROC curve (AUC) is calculated by accumulating integrated values of the ROC curve. Values of AUC for three descriptors are (i) 0.83, (ii) 0.51 and (iii) 0.72, respectively. Therefore, the distance to farthest-neighbor atom well explains the difference in Clusters 5 and 6 and the coordination number of the second nearest-neighbor atom follows. However, distance to the nearest-neighbor atom is not a good descriptor to separate Clusters 5 and 6.

Finally, an optimum structure of metal nanoparticle as a suitable catalyst for O–H bond dissociation reaction of ethanol is briefly discussed. Figure 7 shows schematic image of the specific feature of the nanoparticle as a catalyst of dissociation reaction predicted by hierarchical clustering. According to above evaluation of descriptors, distance to the farthest-neighbor atom from target atom of coordinates in Cluster 5 is basically larger than that in Cluster 6. That means the target atom tends to exist on top of major axis of elliptical nanoparticle in Cluster 5. Conversely, target atom tends to exist on top of minor axis of the nanoparticle for coordinates in Cluster 6. It is expected in the case of Cluster 6 that hydrogen atom of O–H bond is easily adsorbed on another metal atom to form additional H–M bond after the ethanol adsorption. This H–M bond enhance the possibility of O–H dissociation reaction due to the charge transfer from O–H to H–M bond. Actually, same reaction path is found in the snapshot of primary data of the MD simulation in Fig. 2. On the other hand, hydrogen atoms are hardly adsorbed on other metal atoms even if oxygen atoms are adsorbed on metal atoms in the case of Cluster 5. This difference causes the ability of O–H bond dissociation from geometric viewpoint.

4. Conclusion

In this study, unsupervised machine learning is examined for the results of ab initio MD simulation to extract unique
characteristics for O–H bond dissociation reaction of ethanol on Fe–Co nanoparticle. The hierarchical clustering is performed for representative properties (RDF and DOS), which are translated from primary data of \textit{ab initio} MD simulation. Hierarchical clustering of RDF successfully classifies the coordinates on reaction in a specific cluster of the dendrogram. Moreover, descriptor dominant for this classification is estimated using ROC curve. In summary, it is significant in this study to show a new way of post-process of results of MD simulations based on the unsupervised learning technique. It paves the way for a new possibility of ML-based design of optimum catalysts for various chemical reactions.

Note that there are many other representative properties to be examined for the classification of dissociation reaction (and many other reactions). Therefore, suitable properties for classification will be examined in the next step as well as improvement of the ML-technique.

Appendix A. Wasserstein Distance

The Wasserstein distance,\textsuperscript{30} also called Earth Mover’s Distance (EMD),\textsuperscript{31} is an established method to define the distance between probability distributions on a given metric space. Intuitively, when the area of a unit element of the distribution is regarded as a certain amount of package, the Rubner distance is defined as the minimum cost for moving packages from one distribution ($P$ in Fig. A1) to the other one ($Q$ in Fig. A1). Namely, it is regarded as the optimized transport problem. Now, let us consider the case when a package in distribution $P$ ($P_i$) is moved to form distribution $Q$ ($Q_j$) (see Fig. A1). When a package of amount $f_{ij}$, is moved from $P_i$ to $Q_j$ in a distance $d_{ij}$, the work for this process is defined as $d_{ij} \times f_{ij}$. Therefore, the total work $W$ for moving all package from one to the other distribution is defined as

$$W = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}. \tag{A.1}$$

Now, the problem comes down to finding a set of $f_{ij}$ to minimize $W$. An optimum solution should be found under the following constraint conditions.

(i) Transportation is always performed in one direction (only from $P$ to $Q$):

$$f_{ij} \geq 0 \quad (1 \leq i \leq m, \ 1 \leq j \leq n). \tag{A.2}$$

(ii) A package larger than the amount in $P$ cannot be transported to $Q$:

$$\sum_{j=1}^{n} f_{ij} \leq w_{pi} \quad (1 \leq i \leq m). \tag{A.3}$$

(iii) A package exceeding the capacity of $Q$ will not be accepted:

$$\sum_{i=1}^{m} f_{ij} \leq w_{qj} \quad (1 \leq j \leq n). \tag{A.4}$$

(iv) The lesser value of the total volume of the package or the total capacity is adopted as the upper limit of the total value of the transport.

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min \left( \sum_{i=1}^{m} w_{pi}, \sum_{j=1}^{n} w_{qj} \right). \tag{A.5}$$

Solving this transport problem yields the optimal $f_{ij}^*$, and the Wasserstein distance (i.e., EMD) is defined using an optimum set of $f_{ij}^*$ as follows,

$$\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}^* \bigg/ \left( \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}^* \right). \tag{A.6}$$

The term in denominator is employed for the normalization. The optimum set of $f_{ij}^*$ is practically derived using a general linear programming technique. It is known that the Wasserstein distance estimates the distance between distribution functions more accurately than the distance in a vector space such as the Cosine distance.

Appendix B. Receiver Operating Characteristic Curve

Receiver operating characteristic (ROC) curve is a major.
tool to visualize their performance of binary classification.\textsuperscript{23} The basic concept of ROC curve is as follows. When some data are classified into two clusters (positive and negative) using a classifier, there are four possible patterns in the classification: true positive (TP, positive data classified into positive cluster correctly), false negative (FN, positive data classified into negative cluster incorrectly), true negative (TN, negative data classified into negative cluster correctly) and false positive (FP, negative data classified into positive cluster incorrectly). The true positive rate (TPR) is defined as $TP/(TP + FN)$, which represents the ratio of positive data classified into positive cluster to all positive data. Here, the number of data in each pattern is denoted by with Italics. The false positive rate (FPR) is defined as $FP/(FP + TN)$, which represents the ratio of negative data classified into positive cluster incorrectly to all negative data. ROC curve is a two-dimensional graph, in which TPR is plotted on the vertical axis and FPR is plotted on the horizontal axis. TPR and FPR change in conjunction with each other when the threshold dimension of graph, in which TPR is plotted on the vertical axis and FPR is plotted on the horizontal axis.

Acknowledgments

This work was partially supported by Grant-in-Aid for JSPS Research Fellow (No. 18J22727) from Japan Society for Promotion of Science (JSPS), Japan. S.F. was supported by JSPS through the Program for Leading Graduate Schools (MERIT).

REFERENCES

29) K. Pearson: Philo. Mag. 5th Ser. 50 (1900) 157–175.