Explaining reaction coordinates of alanine dipeptide isomerization obtained from deep neural networks using Explainable Artificial Intelligence (XAI)

Kikutsuji, Takuma
Division of Chemical Engineering, Department of Materials Engineering Science, Graduate School of Engineering Science, Osaka University

Mori, Yusuke
Division of Chemical Engineering, Department of Materials Engineering Science, Graduate School of Engineering Science, Osaka University

Okazaki, Kei-ichi
Research Center for Computational Science, Institute for Molecular Science

Mori, Toshifumi
Institute for Materials Chemistry and Engineering, Kyushu University

他

https://hdl.handle.net/2324/4785205

KYUSHU UNIVERSITY

Explaining reaction coordinates of alanine dipeptide isomerization obtained from deep neural networks using Explainable Artificial Intelligence (XAI)

Takuma Kikutsuji, Yusuke Mori, (iD) Kei-ichi Okazaki, et al.

View Online   Export Citation   CrossMark

ARTICLES YOU MAY BE INTERESTED IN

AIP Publishing

Explaining reaction coordinates of alanine dipeptide isomerization obtained from deep neural networks using Explainable Artificial Intelligence (XAI)

Takuma Kikutsuji,[1] Yusuke Mori,[1] Kei-ichi Okazaki,[2,3,a]   Toshifumi Mori,[4,5,b]
Kang Kim,[1,c]   and Nobuyuki Matubayasi[1,d]

AFFILIATIONS

[1] Division of Chemical Engineering, Department of Materials Engineering Science, Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka 560-8531, Japan
[2] Research Center for Computational Science, Institute for Molecular Science, Okazaki, Aichi 444-8585, Japan
[3] The Graduate University for Advanced Studies, Sokendai, Okazaki, Aichi 444-8585, Japan
[4] Institute for Materials Chemistry and Engineering, Kyushu University, Kasuga, Fukuoka 816-8580, Japan
[5] Interdisciplinary Graduate School of Engineering Sciences, Kyushu University, Kasuga, Fukuoka 816-8580, Japan

[a] Electronic mail: keokazaki@ims.ac.jp
[b] Electronic mail: toshi_mori@cm.kyushu-u.ac.jp
[c] Author to whom correspondence should be addressed: kk@cheng.es.osaka-u.ac.jp
[d] Electronic mail: nobuyuki@cheng.es.osaka-u.ac.jp

ABSTRACT

A method for obtaining appropriate reaction coordinates is required to identify transition states distinguishing the product and reactant in complex molecular systems. Recently, abundant research has been devoted to obtaining reaction coordinates using artificial neural networks from deep learning literature, where many collective variables are typically utilized in the input layer. However, it is difficult to explain the details of which collective variables contribute to the predicted reaction coordinates owing to the complexity of the nonlinear functions in deep neural networks. To overcome this limitation, we used Explainable Artificial Intelligence (XAI) methods of the Local Interpretable Model-agnostic Explanation (LIME) and the game theory-based framework known as Shapley Additive exPlanations (SHAP). We demonstrated that XAI enables us to obtain the degree of contribution of each collective variable to reaction coordinates that is determined by nonlinear regressions with deep learning for the committor of the alanine dipeptide isomerization in vacuum. In particular, both LIME and SHAP provide important features to the predicted reaction coordinates, which are characterized by appropriate dihedral angles consistent with those previously reported from the committor test analysis. The present study offers an AI-aided framework to explain the appropriate reaction coordinates, which acquires considerable significance when the number of degrees of freedom increases.

I. INTRODUCTION

Identifying reaction coordinates (RCs) from a large number of collective variables (CVs) is important for appropriately describing the transition state (TS) distinguishing the reactant and product in various complex molecular systems.[1–9] The analysis of committor $p_B(r) \in [0, 1]$, that is, the probability of trajectories reaching the product B prior to the reactant A starting from any conformation r, is a promising statistical examination to explore proper RCs from transition path samplings using molecular dynamics (MD) simulations.[10,11] The distribution of p_B should be unimodal

with a sharp peak at $p_B \sim 0.5$ corresponding to TS because an appropriate RC is of such p_B near TS.[12–27]

Peters *et al.* proposed the likelihood maximization method to find RCs using transition path sampling in MD simulations,[28–30] specifically, one in which the committor is modeled as a sigmoidal function, $p_B(q) = (1 + \tanh(q))/2$, where q is given by linear combinations of CVs. A good RC is obtained such that the logarithmic form of the likelihood is maximized, corresponding to logistic regression. This likelihood maximization method has been applied to various complex systems.[31–53] The cross-entropy minimization method has recently been proposed with the help of the pre-evaluated committor values p_B^* ranging from 0 (toward state A) to 1 (toward state B).[54,55] This is an extension of the likelihood maximization in the sense that the cross-entropy is derived from the Kullback–Leibler divergence by considering the logarithmic form of the likelihood.[54] In general, these maximization or minimization methods can be categorized as the linear regression (LR) in the field of machine learning.

Machine learning models have been widely utilized to determine the dominant CVs from trajectories obtained using MD simulations.[56–73] Furthermore, a feasible application is the nonlinear regression based on a deep neural network (DNN), which is expected to have a performance beyond that of the LR in searching for an appropriate RC.[69,74–76] In particular, nonlinear functions of a DNN with hidden layers will provide richer expressions when the number of CVs is drastically increased in the system of interest. However, it remains difficult to obtain a human-interpretable explanation for DNN learning. Frassek *et al.* reported the application of an autoencoder consisting of an encoder, a reconstruction decoder, and a committor decoder to obtain a low-dimensional representation of RCs discovered from many input CVs.[69] Jung *et al.* proposed an advanced sampling scheme for rare events, in which the maximum likelihood method combined with deep learning was designed to identify the relevant RC. Symbolic regression was further utilized to provide human-interpretable forms for trained DNN models using mathematical expressions.[74,75] More recently, Neumann and Schwierz applied a related model of DNN to predict RCs of magnesium binding to RNA and used the permutation importance method to characterize the feature importances out of input CVs.[76]

In this study, we propose an artificial intelligence (AI)-aided method to determine the nonlinear RC and then interpret the RC locally at the TS using an Explainable AI (XAI) framework. DNN is used to identify the appropriate RC from the CV dataset and pre-evaluated committor values p_B^* obtained from transition path samplings. The target reaction is the C_{7eq} and C_{7ax} isomerization of alanine dipeptide in vacuum [see Fig. 1(a)]. Empirically, the dihedral angle change in φ is thought to be coupled with the other major angle ψ. Bolhuis *et al.* revealed that an additional dihedral angle θ next to φ becomes relevant for describing the proper committor p_B distribution with a peak at $p_B \sim 0.5$.[14] An analogous result was reported by Ren *et al.* using the string method.[77] Furthermore, in a seminal study by Ma and Dinner, a neural network combined with a genetic algorithm was applied to the committor values, predicting that an appropriate RC involves the dihedral angle θ.[56] We also demonstrated that the cross-entropy minimization enabled elucidation of the importance of θ in the alanine dipeptide isomerization.[55] More recently, Manuchehrfar *et al.* reported persistent homology analysis results for characterizing configurations with committor values $p_B \sim 0.5$ on a two-dimensional plot of probability density using φ and θ.[78] Note that there are other target reactions that can be applied by the DNN, such as the rate-promoting vibrations model of enzyme catalysis,[79] of which RC was identified using the likelihood maximization by Peters.[33]

The purpose of XAI is to provide an explainable model for the black-box-type predictions of DNNs.[80,81] In other words, XAI can be regarded as a class of model-agnostic interpretation method, which is separated from predictions using DNN. XAI is further classified into local and global explanation methods. Global explanation methods, including the permutation importance method used in Ref. 76, characterize the average contribution of input variables to the prediction. In contrast, local explanation methods have an advantage of giving interpretable models to the individual predictions by estimating the contribution of each input variable to each prediction. As major implementations of local explanation methods, the Local Interpretable Model-agnostic Explanation (LIME)[82] and the game theory-based framework known as Shapley Additive exPlanations (SHAP)[83] are employed. We locally examined the feature contributions of the input CVs to an appropriate RC predicted from the
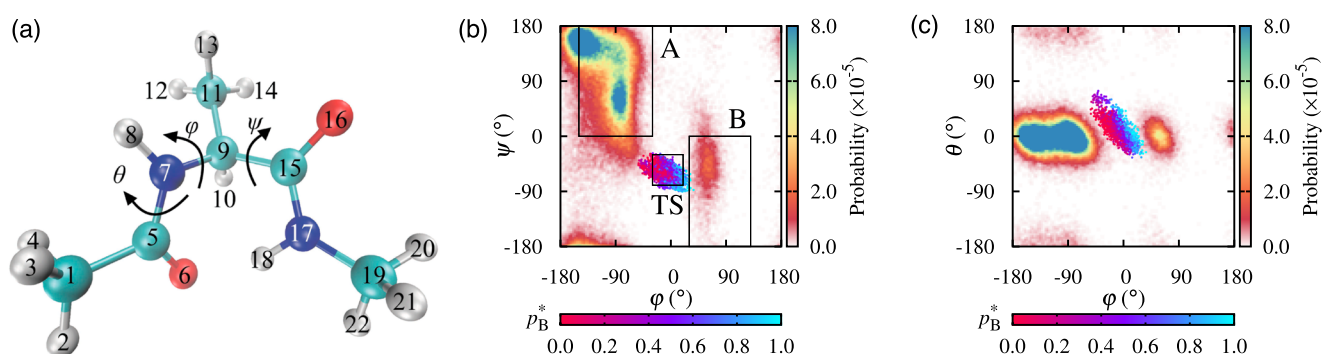
FIG. 1. (a) Index assigned to the alanine dipeptide atom. Major three dihedral angles φ, ψ, and θ are also described. (b) Ramachandran plot describing the probability distribution using φ and ψ. The black boxes describe states A $[(-150°, 0°) \leq (\varphi, \psi) \leq (-30°, 180°)]$, B $[(30°, -180°) \leq (\varphi, \psi) \leq (130°, 0°)]$, and TS $[(-30°, -80°) \leq (\varphi, \psi) \leq (20°, -30°)]$. (c) Contour plot of the probability distribution using φ and θ. In (b) and (c), the shooting points are colored by the p_B^* values given in the bottom color bar.

DNN using LIME and SHAP. In particular, it is important to obtain a local explanation model for conformations that exhibit $p_B^* \sim 0.5$ corresponding to TS. This information may not be readily accessed by the symbolic regression, which provides a global model for RC. Finally, the extracted dominant CVs are interpreted by a separatrix line distinguishing states A and B on the contour map of the probability distribution using two dihedral angles, φ and θ.

II. METHODS

A. Simulation details

We numerically examined the isomerization of alanine dipeptide in vacuum using MD simulations. The system is the same as that in our previous study, where one alanine dipeptide molecule was simulated (other numerical conditions are described in Ref. 55).

Figure 1(b) shows the Ramachandran plot of the alanine dipeptide using the major dihedral angles, $\varphi(C–N–C_\alpha–C)$ and $\psi(N–C_\alpha–C–N)$ [see also Fig. 1(a)]. We examined the transition paths between two energetically stable states, the β-sheet structure (C_{7eq} denoted as state A) and the left-handed α-helix structure (C_{7ax} denoted as state B), which are characterized in Fig. 1(b). We defined states A and B as A [$(-150°, 0°) \leq (\varphi, \psi) \leq (-30°, 180°)$] and B [$(30°, -180°) \leq (\varphi, \psi) \leq (130°, 0°)$], respectively. In addition, the intermediate region was regarded as TS [$(-30°, -80°) \leq (\varphi, \psi) \leq (20°, -30°)$].

In our previous study,[55] we sampled 2000 shooting points from the TS region using the aimless shooting method.[28] For each shooting point, the velocity was randomly assigned according to the Maxwell–Boltzmann distribution at 300 K, generating a trajectory of 1 ps. This was repeated 100 times for each point, and the committor value p_B^* was quantified from the number of transitions to states A or B. We also calculated a total of 90 CVs from all 45 dihedral angles in the molecule into cosine and sine forms. See Fig. 1(a) and Table I for details of the investigated dihedral angles. We used the same datasets of CVs and p_B^* as those used in Ref. 55.

TABLE I. Definition of the CV index. The atom index is represented in Fig. 1(a). Note that the dihedral angles are used in cosine and sine forms, i.e., x_1 to x_{45} and x_{46} to x_{90} are the cosine and sine forms, respectively.

Index of CVs	Index of atoms for dihedral angles		
1-3	2-1-5-6	2-1-5-7	3-1-5-6
4-6	3-1-5-7	4-1-5-6	4-1-5-7
7-9	1-5-7-8	1-5-7-9	6-5-7-8
10-12	6-5-7-9	5-7-9-10	5-7-9-11
13-15	5-7-9-15	8-7-9-10	8-7-9-11
16-18	8-7-9-15	7-9-11-12	7-9-11-13
19-21	7-9-11-14	10-9-11-12	10-9-11-13
22-24	10-9-11-14	15-9-11-12	15-9-11-13
25-27	15-9-11-14	7-9-15-16	7-9-15-17
28-30	10-9-15-16	10-9-15-17	11-9-15-16
31-33	11-9-15-17	9-15-17-18	9-15-17-19
34-36	16-15-17-18	16-15-17-19	15-17-19-20
37-39	15-17-19-21	15-17-19-22	18-17-19-20
40-42	18-17-19-21	18-17-19-22	1-7-5-6
43-45	5-9-7-8	9-17-15-16	15-19-17-18

The shooting points colored by the p_B^* values are shown in Figs. 1(b) and 1(c). Figure 1(c) shows the contour map of the probability distribution as a function of φ and θ. Although points with $p_B^* \sim 0.5$ are widely distributed on the (φ, ψ) plane, a clear separatrix line can be described on the (φ, θ) plane. This result indicates that the appropriate RC is related to θ instead of ψ.[55]

B. Linear regression and deep neural network

In this study, we used the LR and DNN to learn the relationship between the committor distribution p_B^* and candidate CVs, yielding the committor prediction. Figure 2 illustrates a schematic of the training of the LR and DNN. The procedure consists of two parts: one is the training part that transforms from the input layer given by the CVs into a one-dimensional variable q, and the other is the prediction part that transforms q into the sigmoidal function $p_B(q) = (1 + \tanh(q))/2$.

The training part for both LR and DNN is set up as follows: LR is implemented by a simple perceptron, and thus, the output can be described by $q = \sum_{m=1}^{M} w_m x_m + w_0$. Here, M is the number of CVs, x_m is the m-th CV, and w_m is the corresponding coefficient. In addition, w_0 denotes the bias term. By contrast, the DNN consisted of five hidden layers, of which the odd- and even-numbered layers had 400 and 200 nodes, respectively. We used the leaky rectified linear unit (Leaky ReLU) with a leaky parameter set to 0.01 as the default for the activation function.[84] The output is a one-dimensional variable q. For the prediction part for both the LR and DNN, the relationship between the output q and committor p_B is given by $p_B(q) = (1 + \tanh(q))/2$. In other words, the corresponding activation function can be described in a sigmoidal manner.

The dataset of CVs and committor values from 2000 shooting points were divided into training, validation, and test datasets at a ratio of 5:1:4. The variables of CVs were standardized. Optimization was performed using AdaMax.[85] The learning rate lr and two decay factors β_1 and β_2 were set to the default values of 0.001, 0.9, and 0.99, respectively. The L_2 norm regularization was set to both LR and DNN with a regularization parameter of 0.001 to prevent overfitting. In addition, the dropout was set to the hidden layers at a rate of 0.5 during DNN training. We used the TensorFlow library to implement the LR and DNN.[86]

The cross-entropy function[54,55]

$$\mathcal{H}(p_B^*, p_B) = -\sum_{k=1}^{N} p_B^*(r_k) \ln p_B(q) - \sum_{k=1}^{N} \left(1 - p_B^*(r_k)\right) \ln[1 - p_B(q)]$$

(1)

was used to derive the loss function for both the LR and DNN training. Here, r_k represents the kth conformation of the molecule with the number of shooting points N. The cross-entropy minimization principle was derived from the Kullback–Leibler divergence[54] and was applied to the LR to search for the appropriate RC for alanine dipeptide isomerization.[55]

C. LIME—Local Interpretable Model-agnostic Explanation

LIME was applied to explain the committor predictions obtained by the DNN training. In general, DNNs are very complex in terms of obtaining understandable features. LIME can explain the contribution of each input variable of any black-box-type classifier
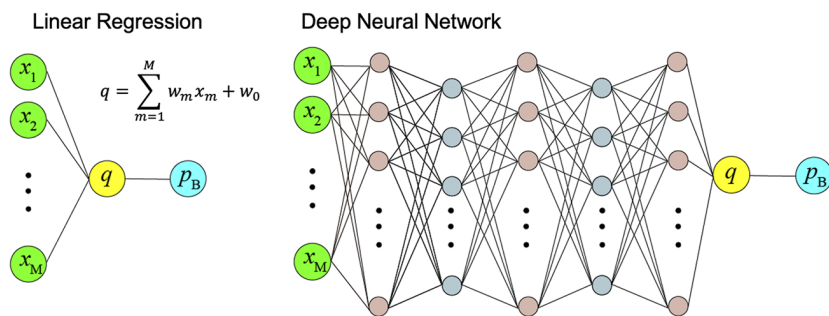
FIG. 2. Schematic diagram of LR and DNN learning the relationship between pre-evaluated committor value p_B^* and $M = 90$ CVs plus 1 bias term $\boldsymbol{x} = (1, x_1, x_2, \ldots, x_M)$ and predicting committor p_B as a sigmoidal function $p_B(q) = (1 + \tanh(q))/2$. Note that the node representing the bias term is omitted from the diagram.

using a more interpretable model.[82] It provides a linear regression function for the local behavior of a target instance explained by perturbation of input variables. In other words, features with large coefficients in the linear regression function provide a predictive interpretation. The interpretable feature $\xi(\boldsymbol{x})$ is obtained by a linear regression function g for the input data by the CV vector $\boldsymbol{x} = (1, x_1, x_2, \ldots, x_M)$ (90 dihedral angles plus 1 bias term in our case) and the black-box model f by the DNN. This is represented by the following equation:

$$\xi(\boldsymbol{x}) = \arg\min_{g \in G}\left(L^{\mathrm{LIME}}(f, g, \pi_{\boldsymbol{x}}^{\mathrm{LIME}}) + \Omega(g)\right), \qquad (2)$$

where $L^{\mathrm{LIME}}(f, g, \pi_{\boldsymbol{x}}^{\mathrm{LIME}})$ is the squared loss function that measures the distance between f and g, with G representing a class of explanation models ($g \in G$). $\pi_{\boldsymbol{x}}^{\mathrm{LIME}}$ represents the proximity measure around the input data $\boldsymbol{x}$ to be explained. In practice, the exponential kernel function is

$$\pi_{\boldsymbol{x}}^{\mathrm{LIME}}(\boldsymbol{z}) = \exp\left(-D(\boldsymbol{x}, \boldsymbol{z})^2/\sigma^2\right), \qquad (3)$$

where the distance function D and width σ are used for any perturbed instance $\boldsymbol{z}$, which is randomly generated around the data $\boldsymbol{x}$. In LIME, the perturbed data $\boldsymbol{z}$ around $\boldsymbol{x}$ weighted by the proximity measure is transformed into binary variables $\boldsymbol{z}' \in \{0, 1\}^M$ with the number of input variables M for the human-understandable presentation because the important components of the original input data are not always interpretable. Regularization $\Omega(g)$ is also used to reduce the complexity of the explanation for $g \in G$. We used the LIME package, which is available at https://github.com/marcotcr/lime.

D. SHAP—Shapley Additive exPlanations

We also applied SHAP, which is a game-theory-based method for explaining the predictions of black-box-type models.[83] Note that LIME assumes that the local behavior can be described by the linear model, but this has no theoretical background. In contrast, SHAP guarantees that the prediction is fairly distributed among the input features. In fact, SHAP utilizes the Shapley value,[87] which is a value that fairly distributes the rewards given by the team collaboration to individual players. In SHAP, an additive feature attribution method provides a linear function consisting of binary variables,

$$g(\boldsymbol{z}') = \phi_0 + \sum_{i=1}^M \phi_i z_i', \qquad (4)$$

where the coefficient ϕ_i serves as the Shapley value, which explains the importance of the binary feature $\boldsymbol{z}' \in \{0, 1\}^M$. LIME can be understood as an additive feature attribution method that provides linear models in the binary vector space in the sense that the explanation model g of LIME can be expressed by Eq. (4).

We used one of several implementations for SHAP, known as the Kernel SHAP, which is designed as a model-agnostic estimation that provides a local explanation model using Shapley values and LIME. For the Kernel SHAP, the following specific equations are used with Eq. (2) of the LIME algorithm:

$$\Omega(g) = 0, \qquad (5)$$

$$\pi_{\boldsymbol{x}}^{\mathrm{SHAP}}(\boldsymbol{z}') = \frac{M-1}{{}_MC_{|\boldsymbol{z}'|}|\boldsymbol{z}'|(M - |\boldsymbol{z}'|)}, \qquad (6)$$

$$L^{\mathrm{SHAP}}(f, g, \pi_{\boldsymbol{x}}^{\mathrm{SHAP}}) = \sum_{\boldsymbol{z}' \in Z} \pi_{\boldsymbol{x}}(\boldsymbol{z}')\left(f(h_{\boldsymbol{x}}^{-1}(\boldsymbol{z}')) - g(\boldsymbol{z}')\right)^2, \qquad (7)$$

where $|\boldsymbol{z}'|$ is the number of non-zero features in $\boldsymbol{z}'$ and h_x is a mapping function of binary variables $\boldsymbol{z}'$ into the original input data $\boldsymbol{x}$. The local explanation model $g(\boldsymbol{z}')$ approximating $f(h_{\boldsymbol{x}}^{-1}(\boldsymbol{z}'))$ can be obtained using a weighted linear regression with Eq. (6), which differs from the weight function used in LIME [see Eq. (3)]. Note the contrast between Eqs. (3) and (6), which are used in LIME and SHAP, respectively. In practice, if we use SHAP's kernel of Eq. (6) as LIME's kernel, LIME will provide values similar to SHAP values. The SHAP package is available at https://github.com/slundberg/shap. We obtained SHAP values using this package, where the Akaike information criterion is used for the regularization as default.

III. RESULTS AND DISCUSSION

A. Training and prediction of committor

Figure 3 shows the learning process using a training dataset (1000 points) for both LR and DNN by plotting the loss function as a function of the iteration step. The loss function is the cross-entropy of Eq. (1) for both of LR and DNN. In parallel, to check for overfitting of the training model, we monitored the loss function values obtained from the validation dataset (200 points), which are also plotted in Fig. 3. To examine the robustness of the learning, we performed ten trials from randomly chosen coefficients and confirmed that there was no significant difference in the converged value of the loss function for both LR and DNN. The results are presented
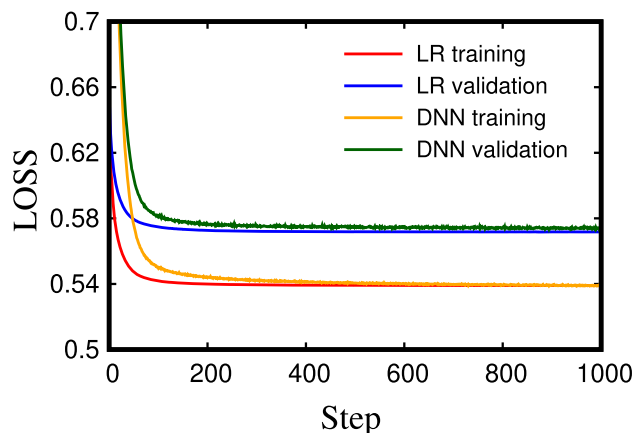
FIG. 3. Training process of cross-entropy minimization for both LR and DNN. Red and blue (orange and green) curves indicate the loss function [cross-entropy $\mathcal{H}$ in Eq. (1)] of LR (DNN) as a function of iteration steps with the training and validation dataset, respectively.

as an average of ten trials. It is demonstrated that the loss function decreases without any increase in those for the validation dataset and finally converges after several hundred iterations for both LR and DNN. The convergence of DNN is slower than that of LR, which can be regarded as a result of the number of parameters, that is, 91 for LR and ~360 000 for DNN.

The LR/DNN trained model provides a prediction with regard to the relationship between the committor value p_B and RC q from the test dataset (800 points). Figure 4(a) shows the committor distribution p_B^* as a function of q for the test dataset. It was confirmed that p_B^* values of the test dataset globally follow the sigmoidal function. To further investigate p_B^* values close to the TS at $q = 0$, the distribution of p_B^* in the range of $-0.2 < q < 0.2$ is shown in Fig. 4(b). The p_B^* distribution exhibits a sharp peak at ~0.5 for both LR and DNN trainings. Thus, the predicted q from both LR and DNN training can characterize the appropriate RC for the alanine dipeptide isomerization.
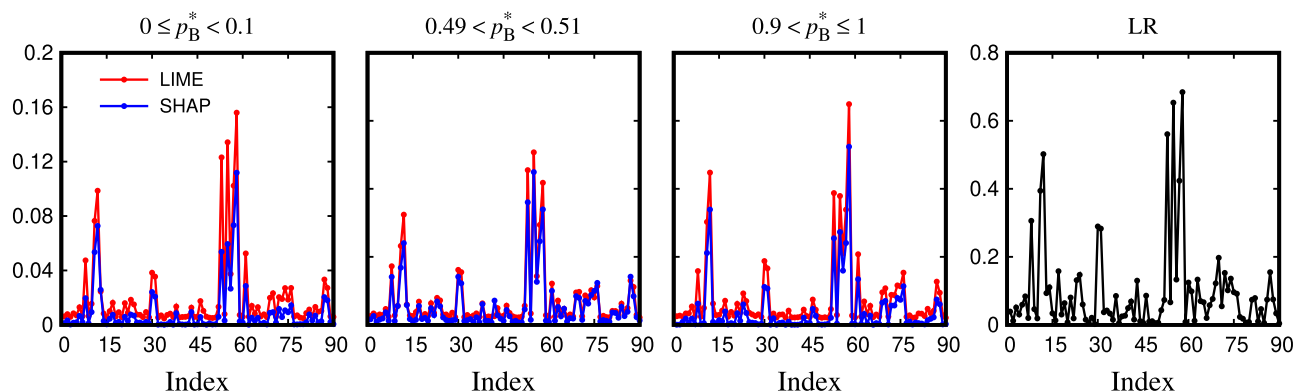
B. Contribution of collective variables to the prediction

We used XAI, that is, LIME and SHAP, to explain the contribution of the input CVs to the RC from our DNN training. In practice, the dataset (2000 points) is classified into three regions, $0 \leq p_B^* < 0.1$ (near state A), $0.49 < p_B^* < 0.51$ (near TS), and $0.9 < p_B^* \leq 1$ (near state B) to provide a local explanation. For each region, randomly chosen 30 points were analyzed by LIME and SHAP, from which the contributions of the input variables to q (the appropriate RC) were quantified. Figure 5 shows the absolute values of the feature contributions averaged over 30 points for the three p_B^* regions obtained using LIME ans SHAP. For comparison, the absolute values of the optimized coefficient w_m obtained from the LR training are also plotted in Fig. 5. The feature contribution values of LIME and SHAP have the same meaning as the coefficient w_m obtained by the LR model, although the values do not match among LIME, SHAP, and LR. The top five dominant contributions and their absolute values are listed in Table II.
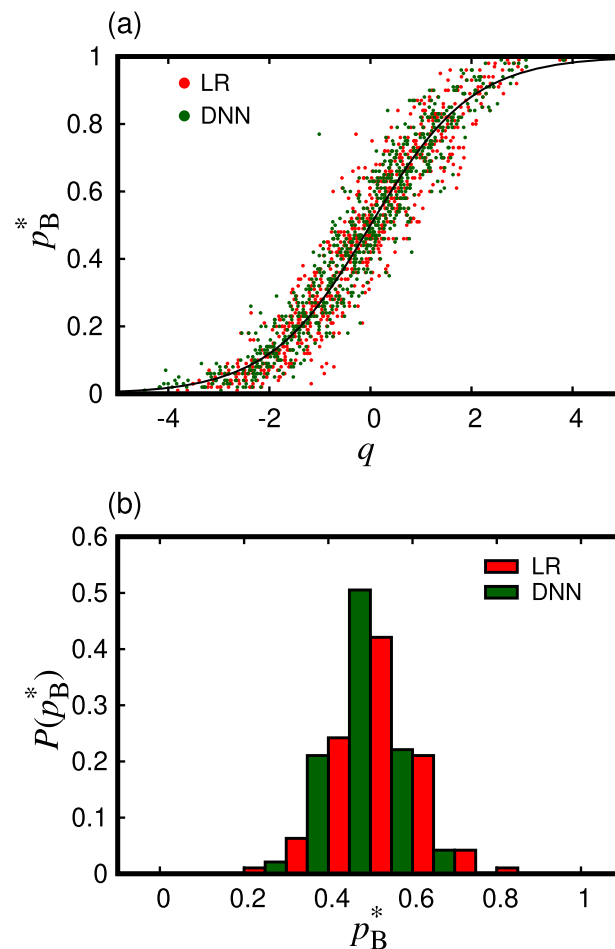
FIG. 4. (a) Relationship between committor distributions p_B^* and q obtained from the LR (red) and DNN (green) trained model using test dataset (800 points). The solid black line represents the sigmoidal function, $p_B(q) = (1 + \tanh(q))/2$. (b) Probability of p_B^* for $q \in (-0.2, 0.2)$ for LR (red) and DNN (green), where the points are extracted from the data in (a).

Indices 57 and 54, corresponding to $\sin\varphi$ (5-7-9-11) and $\sin\theta$ (6-7-8-9), respectively, have large contributions using the LR model, which is consistent with the results of our previous study.[55] This result indicates that the relevant angle to the rotation of φ is not ψ but instead θ. Notably, both LIME and SHAP reveal two dihedral angles, φ (indices 11, 56, and 57) and θ (indices 52 and 54), as major contributions to the p_B prediction obtained from the DNN. The results obtained by LIME and SHAP are similar over the three regions of p_B^*. This observation may justify LIME's linear model assumption by being consistent with the SHAP's result guaranteed by the game theory. However, the order is different between TS ($0.49 < p_B^* < 0.51$) and states near A or B ($0 \leq p_B^* < 0.1$ or $0.9 < p_B^* \leq 1$). This difference can be understood as follows: Near the two stable states, A and B, the contribution of variables relating to φ is larger than those corresponding to θ because a large increase in the angle φ is necessary for transitions A → B and B → A. By contrast, near TS, both LIME and SHAP explain that the contribution of θ (indices 52 and 54) becomes

FIG. 5. Feature contribution of each collective variable in absolute value obtained using LIME (red) and SHAP (blue) for the DNN training. The local explanation is provided for three regions corresponding to $0 \leq p_B^* < 0.1$, $0.49 < p_B^* < 0.51$, and $0.9 < p_B^* \leq 1$ from left to right. The rightmost panel shows optimized coefficients w_m of each variable x_m in absolute value obtained from the LR model.

TABLE II. Top five dominant contributions and those absolute values obtained using LIME and SHAP for the three regions, $0 \leq p_B^* < 0.1$, $0.49 < p_B^* < 0.51$, and $0.9 < p_B^* \leq 1$. The result of the LR model is also shown for comparison.

$0 \leq p_B^* < 0.1$				$0.49 < p_B^* < 0.51$				$0.9 < p_B^* \leq 1$				$0 \leq p_B^* \leq 1$	
LIME		SHAP		LIME		SHAP		LIME		SHAP		LR	
Index	Value	Index	Value	Index	Value	Index	Value	Index	Value	Index	Value	Index	Value
57	0.156	57	0.112	54	0.127	54	0.112	57	0.162	57	0.131	57	0.684
54	0.134	56	0.073	52	0.114	52	0.090	11	0.112	11	0.085	54	0.653
52	0.123	11	0.073	57	0.104	57	0.085	52	0.097	54	0.068	52	0.561
56	0.102	54	0.059	11	0.081	56	0.061	54	0.095	52	0.064	11	0.502
11	0.099	52	0.054	56	0.074	11	0.060	56	0.085	56	0.060	56	0.424

larger, as the change in θ has more influence than the change in φ. In Fig. 1(c), we show the two-dimensional probability distribution map of φ and θ, where the distribution of shooting points having $p_B^* \sim 0.5$ are located not perpendicular to φ but tilted in the direction of θ. Thus, the θ-angle change becomes important when crossing the separatrix distinguishing states A and B, whereas the φ-angle change is globally important for the target isomerization. Such a detailed process cannot be described by coefficients globally optimized using the LR model but is revealed by the local explanation model using DNN with XAI.

IV. CONCLUSIONS

In this paper, we proposed an AI-aided method, in which DNN is used to identify the appropriate RC for alanine dipeptide isomerization in vacuum. We trained the DNN to predict the committor function in a sigmoidal manner from the dataset of committor p_B^* and dihedral angles in the cosine and sine forms. The DNN revealed the importance of θ rather than ψ along the rotation about φ, which is consistent with various studies using the committor test analysis.[14,56,77]

Furthermore, LIME and SHAP were used as XAI tools to provide a local explanation model for black-box-type DNN prediction,

in contrast to the symbolic regression method for obtaining the human-interpretable model.[74,75] It was demonstrated that LIME and SHAP enabled the explanation for three p_B regions, $0 \leq p_B^* < 0.1$ (near state A), $0.49 < p_B^* < 0.51$ (near TS), and $0.9 < p_B^* \leq 1$ (near state B). In particular, the feature contribution of θ becomes more evident near TS than near state A or B, indicating the necessity of θ-angle change for crossing the barrier between the two stable states A and B. In fact, the influence of θ on the φ-angle change is in accordance with the separatrix line described by shooting points with $p_B^* \sim 0.5$ on the probability distribution of φ and θ, which is tilted in the θ-axis direction.

Finally, it should be noted that understanding the detailed mechanism for complex molecular systems becomes complicated because of the large numbers of CVs, regardless of the use of DNNs. Important targets are solute–solvent systems, such as alanine dipeptide in explicit water, which was examined using neural networks to predict the committor by Ma and Dinner.[56] The DNN predicting the committor is further applied to various solute–solvent systems,[69,74–76] and the current AI-aided method combined with XAI will be used as a practical tool to provide a local explanation for an appropriate RC obtained from DNN training, particularly near the TS of the transition connecting the stable states.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.6392326.

REFERENCES

[1] B. Peters, *Reaction Rate Theory and Rare Events* (Elsevier, Amsterdam, 2017).

[2] S. V. Krivov, "On reaction coordinate optimality," J. Chem. Theory Comput. **9**, 135–146 (2013).

[3] W. Li and A. Ma, "Recent developments in methods for identifying reaction coordinates," Mol. Simul. **40**, 784–793 (2014).

[4] B. Peters, P. G. Bolhuis, R. G. Mullen, and J.-E. Shea, "Reaction coordinates, one-dimensional Smoluchowski equations, and a test for dynamical self-consistency," J. Chem. Phys. **138**, 054106 (2013).

[5] D. J. Wales, "Perspective: Insight into reaction coordinates and dynamics from the potential energy landscape," J. Chem. Phys. **142**, 130901 (2015).

[6] B. Peters, "Reaction coordinates and mechanistic hypothesis tests," Annu. Rev. Phys. Chem. **67**, 669–690 (2016).

[7] P. V. Banushkina and S. V. Krivov, "Optimal reaction coordinates," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **6**, 748–763 (2016).

[8] F. Pietrucci, "Strategies for the exploration of free energy landscapes: Unity in diversity and challenges ahead," Rev. Phys. **2**, 32–45 (2017).

[9] P. G. Bolhuis and D. W. H. Swenson, "Transition path sampling as Markov chain Monte Carlo of trajectories: Recent algorithms, software, applications, and future outlook," Adv. Theory Simul. **4**, 2000237 (2021).

[10] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, "Transition path sampling: Throwing ropes over rough mountain passes, in the dark," Annu. Rev. Phys. Chem. **53**, 291–318 (2002).

[11] J. Rogal, "Reaction coordinates in complex systems-a perspective," Eur. Phys. J. B **94**, 223 (2021).

[12] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, "On the transition coordinate for protein folding," J. Chem. Phys. **108**, 334–350 (1998).

[13] P. L. Geissler, C. Dellago, and D. Chandler, "Kinetic pathways of ion pair dissociation in water," J. Phys. Chem. B **103**, 3706–3710 (1999).

[14] P. G. Bolhuis, C. Dellago, and D. Chandler, "Reaction coordinates of biomolecular isomerization," Proc. Natl. Acad. Sci. U. S. A. **97**, 5877–5882 (2000).

[15] M. F. Hagan, A. R. Dinner, D. Chandler, and A. K. Chakraborty, "Atomistic understanding of kinetic pathways for single base-pair binding and unbinding in DNA," Proc. Natl. Acad. Sci. U. S. A. **100**, 13922–13927 (2003).

[16] G. Hummer, "From transition paths to transition states and rate coefficients," J. Chem. Phys. **120**, 516–523 (2004).

[17] A. C. Pan and D. Chandler, "Dynamics of nucleation in the Ising model," J. Phys. Chem. B **108**, 19681–19686 (2004).

[18] Y. M. Rhee and V. S. Pande, "One-dimensional reaction coordinate and the corresponding potential of mean force from commitment probability distribution," J. Phys. Chem. B **109**, 6780–6786 (2005).

[19] A. Berezhkovskii and A. Szabo, "One-dimensional reaction coordinates for diffusive activated rate processes in many dimensions," J. Chem. Phys. **122**, 014503 (2005).

[20] R. B. Best and G. Hummer, "Reaction coordinates and rates from transition paths," Proc. Natl. Acad. Sci. U. S. A. **102**, 6732–6737 (2005).

[21] D. Moroni, P. R. ten Wolde, and P. G. Bolhuis, "Interplay between structure and size in a critical crystal nucleus," Phys. Rev. Lett. **94**, 235703 (2005).

[22] B. Peters, "Using the histogram test to quantify reaction coordinate error," J. Chem. Phys. **125**, 241101 (2006).

[23] D. Branduardi, F. L. Gervasio, and M. Parrinello, "From A to B in free energy space," J. Chem. Phys. **126**, 054103 (2007).

[24] S. L. Quaytman and S. D. Schwartz, "Reaction coordinate of an enzymatic reaction revealed by transition path sampling," Proc. Natl. Acad. Sci. U. S. A. **104**, 12253–12258 (2007).

[25] D. Antoniou and S. D. Schwartz, "The stochastic separatrix and the reaction coordinate for complex systems," J. Chem. Phys. **130**, 151103 (2009).

[26] B. Peters, "$p(\mathrm{TP}|q)$ peak maximization: Necessary but not sufficient for reaction coordinate accuracy," Chem. Phys. Lett. **494**, 100–103 (2010).

[27] M. Ernst, S. Wolf, and G. Stock, "Identification and validation of reaction coordinates describing protein functional motion: Hierarchical dynamics of T4 lysozyme," J. Chem. Theory Comput. **13**, 5076–5088 (2017).

[28] B. Peters and B. L. Trout, "Obtaining reaction coordinates by likelihood maximization," J. Chem. Phys. **125**, 054108 (2006).

[29] B. Peters, G. T. Beckham, and B. L. Trout, "Extensions to the likelihood maximization approach for finding reaction coordinates," J. Chem. Phys. **127**, 034109 (2007).

[30] B. Peters, "Recent advances in transition path sampling: Accurate reaction coordinates, likelihood maximisation and diffusive barrier-crossing dynamics," Mol. Simul. **36**, 1265–1281 (2010).

[31] G. T. Beckham, B. Peters, C. Starbuck, N. Variankaval, and B. L. Trout, "Surface-mediated nucleation in the solid-state polymorph transformation of terephthalic acid," J. Am. Chem. Soc. **129**, 4714–4723 (2007).

[32] G. T. Beckham, B. Peters, and B. L. Trout, "Evidence for a size dependent nucleation mechanism in solid state polymorph transformations," J. Phys. Chem. B **112**, 7460–7466 (2008).

[33] B. Peters, "Transition-state theory, dynamics, and narrow time scale separation in the rate-promoting vibrations model of enzyme catalysis," J. Chem. Theory Comput. **6**, 1447–1454 (2010).

[34] J. Vreede, J. Juraszek, and P. G. Bolhuis, "Predicting the reaction coordinates of millisecond light-induced conformational changes in photoactive yellow protein," Proc. Natl. Acad. Sci. U. S. A. **107**, 2397–2402 (2010).

[35] W. Lechner, J. Rogal, J. Juraszek, B. Ensing, and P. G. Bolhuis, "Nonlinear reaction coordinate analysis in the reweighted path ensemble," J. Chem. Phys. **133**, 174110 (2010).

[36] B. Pan and M. S. Ricci, "Molecular mechanism of acid-catalyzed hydrolysis of peptide bonds using a model compound," J. Phys. Chem. B **114**, 4389–4399 (2010).

[37] G. T. Beckham and B. Peters, "Optimizing nucleus size metrics for liquid–solid nucleation from transition paths of near-nanosecond duration," J. Phys. Chem. Lett. **2**, 1133–1138 (2011).

[38] B. Peters, "Inertial likelihood maximization for reaction coordinates with high transmission coefficients," Chem. Phys. Lett. **554**, 248–253 (2012).

[39] L. Xi, M. Shah, and B. L. Trout, "Hopping of water in a glassy polymer studied via transition path sampling and likelihood maximization," J. Phys. Chem. B **117**, 3634–3647 (2013).

[40] S. Jungblut, A. Singraber, and C. Dellago, "Optimising reaction coordinates for crystallisation by tuning the crystallinity definition," Mol. Phys. **111**, 3527–3533 (2013).

[41] R. G. Mullen, J.-E. Shea, and B. Peters, "Transmission coefficients, committors, and solvent coordinates in ion-pair dissociation," J. Chem. Theory Comput. **10**, 659–667 (2014).

[42]R. G. Mullen, J.-E. Shea, and B. Peters, "Easy transition path sampling methods: Flexible-length aimless shooting and permutation shooting," J. Chem. Theory Comput. **11**, 2421–2428 (2015).

[43]L. Lupi, B. Peters, and V. Molinero, "Pre-ordering of interfacial water in the pathway of heterogeneous ice nucleation does not lead to a two-step crystallization mechanism," J. Chem. Phys. **145**, 211910 (2016).

[44]H. Jung, K.-i. Okazaki, and G. Hummer, "Transition path sampling of rare events by shooting from the top," J. Chem. Phys. **147**, 152716 (2017).

[45]M. N. Joswiak, M. F. Doherty, and B. Peters, "Ion dissolution mechanism and kinetics at kink sites on NaCl surfaces," Proc. Natl. Acad. Sci. U. S. A. **115**, 656–661 (2018).

[46]G. Díaz Leines and J. Rogal, "Maximum likelihood analysis of reaction coordinates during solidification in Ni," J. Phys. Chem. B **122**, 10934–10942 (2018).

[47]K.-i. Okazaki, D. Wöhlert, J. Warnau, H. Jung, Ö. Yildiz, W. Kühlbrandt, and G. Hummer, "Mechanism of the electroneutral sodium/proton antiporter PaNhaP from transition-path shooting," Nat. Commun. **10**, 1742 (2019).

[48]T. A. B. Arjun and P. G. Bolhuis, "Unbiased atomistic insight in the competing nucleation mechanisms of methane hydrates," Proc. Natl. Acad. Sci. U. S. A. **116**, 19305–19310 (2019).

[49]Y. Liang, G. Díaz Leines, R. Drautz, and J. Rogal, "Identification of a multi-dimensional reaction coordinate for crystal nucleation in Ni_3Al," J. Chem. Phys. **152**, 224504 (2020).

[50]J. R. Rogers and P. L. Geissler, "Breakage of hydrophobic contacts limits the rate of passive lipid exchange between membranes," J. Phys. Chem. B **124**, 5884–5898 (2020).

[51]N. Schwierz, "Kinetic pathways of water exchange in the first hydration shell of magnesium," J. Chem. Phys. **152**, 224106 (2020).

[52]L. Levintov, S. Paul, and H. Vashisth, "Reaction coordinate and thermodynamics of base flipping in RNA," J. Chem. Theory Comput. **17**, 1914–1921 (2021).

[53]R. L. Silveira, B. C. Knott, C. S. Pereira, M. F. Crowley, M. S. Skaf, and G. T. Beckham, "Transition path sampling study of the feruloyl esterase mechanism," J. Phys. Chem. B **125**, 2018–2030 (2021).

[54]T. Mori and S. Saito, "Dissecting the dynamics during enzyme catalysis: A case study of Pin1 peptidyl-prolyl isomerase," J. Chem. Theory Comput. **16**, 3396–3407 (2020).

[55]Y. Mori, K.-i. Okazaki, T. Mori, K. Kim, and N. Matubayasi, "Learning reaction coordinates via cross-entropy minimization: Application to alanine dipeptide," J. Chem. Phys. **153**, 054115 (2020).

[56]A. Ma and A. R. Dinner, "Automatic method for identifying reaction coordinates in complex systems," J. Phys. Chem. B **109**, 6769–6779 (2005).

[57]M. M. Sultan and V. S. Pande, "Automated design of collective variables using supervised machine learning," J. Chem. Phys. **149**, 094106 (2018).

[58]C. Wehmeyer and F. Noé, "Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics," J. Chem. Phys. **148**, 241703 (2018).

[59]A. Mardt, L. Pasquali, H. Wu, and F. Noé, "VAMPnets for deep learning of molecular kinetics," Nat. Commun. **9**, 5 (2018).

[60]A. Bittracher, R. Banisch, and C. Schütte, "Data-driven computation of molecular reaction coordinates," J. Chem. Phys. **149**, 154103 (2018).

[61]W. Chen and A. L. Ferguson, "Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration," J. Comput. Chem. **39**, 2079–2102 (2018).

[62]J. M. L. Ribeiro, P. Bravo, Y. Wang, and P. Tiwary, "Reweighted autoencoded variational Bayes for enhanced sampling (RAVE)," J. Chem. Phys. **149**, 072301 (2018).

[63]J. Rogal, E. Schneider, and M. E. Tuckerman, "Neural-network-based path collective variables for enhanced sampling of phase transformations," Phys. Rev. Lett. **123**, 245701 (2019).

[64]L. Bonati, V. Rizzi, and M. Parrinello, "Data-driven collective variables for enhanced sampling," J. Phys. Chem. Lett. **11**, 2998–3004 (2020).

[65]Y. Wang, J. M. Lamim Ribeiro, and P. Tiwary, "Machine learning approaches for analyzing and enhancing molecular dynamics simulations," Curr. Opin. Struct. Biol. **61**, 139–145 (2020).

[66]D. Wang and P. Tiwary, "State predictive information bottleneck," J. Chem. Phys. **154**, 134111 (2021).

[67]H. Sidky, W. Chen, and A. L. Ferguson, "Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation," Mol. Phys. **118**, e1737742 (2020).

[68]J. Zhang, Y.-K. Lei, Z. Zhang, X. Han, M. Li, L. Yang, Y. I. Yang, and Y. Q. Gao, "Deep reinforcement learning of transition states," Phys. Chem. Chem. Phys. **23**, 6888–6895 (2021).

[69]M. Frassek, A. Arjun, and P. G. Bolhuis, "An extended autoencoder model for reaction coordinate discovery in rare event molecular dynamics datasets," J. Chem. Phys. **155**, 064103 (2021).

[70]F. Hooft, A. Pérez de Alba Ortíz, and B. Ensing, "Discovering collective variables of molecular transitions via genetic algorithms and neural networks," J. Chem. Theory Comput. **17**, 2294–2306 (2021).

[71]L. Bonati, G. Piccini, and M. Parrinello, "Deep learning the slow modes for rare events sampling," Proc. Natl. Acad. Sci. U. S. A. **118**, e2113533118 (2021).

[72]M. Chen, "Collective variable-based enhanced sampling and machine learning," Eur. Phys. J. B **94**, 211 (2021).

[73]Z. Belkacemi, P. Gkeka, T. Lelièvre, and G. Stoltz, "Chasing collective variables using autoencoders and biased trajectories," J. Chem. Theory Comput. **18**, 59–78 (2022).

[74]H. Jung, R. Covino, and G. Hummer, "Artificial intelligence assists discovery of reaction coordinates and mechanisms from molecular dynamics simulations," arXiv:1901.04595 (2019).

[75]H. Jung, R. Covino, A. Arjun, P. G. Bolhuis, and G. Hummer, "Autonomous artificial intelligence discovers mechanisms of molecular self-organization in virtual experiments," arXiv:2105.06673 (2021).

[76]J. Neumann and N. Schwierz, "Artificial intelligence resolves kinetic pathways of magnesium binding to RNA," J. Chem. Theory Comput. **18**, 1202–1212 (2022).

[77]W. Ren, E. Vanden-Eijnden, P. Maragakis, and W. E, "Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide," J. Chem. Phys. **123**, 134109 (2005).

[78]F. Manuchehrfar, H. Li, W. Tian, A. Ma, and J. Liang, "Exact topology of the dynamic probability surface of an activated process by persistent homology," J. Phys. Chem. B **125**, 4667–4680 (2021).

[79]D. Antoniou, M. R. Abolfath, and S. D. Schwartz, "Transition path sampling study of classical rate-promoting vibrations," J. Chem. Phys. **121**, 6442–6447 (2004).

[80]A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access **6**, 52138–52160 (2018).

[81]C. Molnar, *Interpretable Machine Learning* (Lulu Press, Morisville, NC, 2020).

[82]M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computational Linguistics, San Diego, CA, 2016), pp. 97–101.

[83]S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates, Red Hook, NY, 2017), pp. 4768–4777.

[84]A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in ICML Workshop on Deep Learning for Audio, Speech and Language Processing (2013).

[85]D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980 (2014).

[86]M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," arXiv:1603.04467 (2016).

[87]L. S. Shapley, "17. A value for n-person games," in *Contributions to the Theory of Games (AM-28), Volume II*, edited by H. W. Kuhn and A. W. Tucker (Princeton University Press, 1953), pp. 307–318.