

UC Berkeley

UC Berkeley Previously Published Works

Title

Using Diffusion Maps to Analyze Reaction Dynamics for a Hydrogen Combustion Benchmark Dataset

Permalink

<https://escholarship.org/uc/item/77h9r263>

Journal

Journal of Chemical Theory and Computation, 19(17)

ISSN

1549-9618

Authors

Ko, Taehee

Heindel, Joseph P

Guan, Xingyi

et al.

Publication Date

2023-09-12

DOI

10.1021/acs.jctc.3c00426

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Using Diffusion Maps to Analyze Reaction Dynamics for a Hydrogen Combustion Benchmark Dataset

Taehee Ko,* Joseph P. Heindel,* Xingyi Guan,* Teresa Head-Gordon,* David B. Williams-Young,* and Chao Yang*



Cite This: *J. Chem. Theory Comput.* 2023, 19, 5872–5885



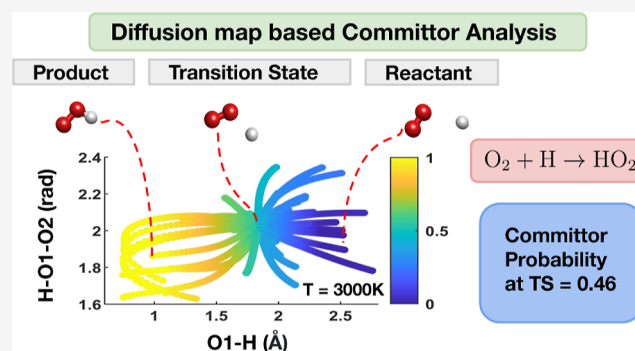
Read Online

ACCESS |

Metrics & More

Article Recommendations

ABSTRACT: We use local diffusion maps to assess the quality of two types of collective variables (CVs) for a recently published hydrogen combustion benchmark dataset¹ that contains ab initio molecular dynamics (MD) trajectories and normal modes along minimum energy paths. This approach was recently advocated in² for assessing CVs and analyzing reactions modeled by classical MD simulations. We report the effectiveness of this approach to molecular systems modeled by quantum ab initio MD. In addition to assessing the quality of CVs, we also use global diffusion maps to perform committor analysis as proposed in.² We show that the committor function obtained from the global diffusion map allows us to identify transition regions of interest in several hydrogen combustion reaction channels.



1. INTRODUCTION

The search for effective collective variables (CVs) for many-body molecular systems is paramount for characterizing the primary dynamic pathway(s) for conformational changes and chemical transformations in a tractable and/or more physically interpretable lower manifold.^{3–6} For some molecular systems such as the alanine dipeptide, physically intuitive CVs such as dihedral angles perform very well, and for chemical reactions simple CVs based on distances are often assumed to be adequate. In such cases, they can also be readily combined with enhanced sampling techniques such as metadynamics^{7,8} or umbrella sampling^{9,10} to identify reaction barriers and transition pathways.

Several methods have been developed to identify CVs, including those using machine learning,¹¹ but they are often validated using molecular systems for which a lot is known already, especially for the alanine dipeptide^{12–15} However, in general, it is not so clear how to identify relevant CVs for complex chemical or material systems in the condensed phase or when chemical reactions occur. For example, aqueous phase transformations can involve many solvent molecules that undergo concerted rearrangements that are mediated through long-range interactions.¹⁶ Furthermore, these studies usually utilize classical molecular dynamics (MD) which typically limits one to studying only conformational changes. For chemical reactions involving bond forming and breaking, a quantum mechanical treatment is needed but can be

computationally slow or prohibitive and the CVs can be unintuitive for complex reaction mechanisms.

Diffusion maps have been shown to be an effective tool for identifying good CVs and computing committor functions and have been widely used in many MD applications,^{2,17–19} and relevant to the topic explored here, for searching for rare transitions within dynamical information.^{2,19,20} While the diffusion coordinates (DCs) obtained from the eigenvectors of the diffusion map operator have been proposed as a means to obtain CVs, their interpretation in terms of the potential energy surface (PES) can be unclear. Recently, Trstanova and co-workers have suggested that DCs obtained from a “local” diffusion map can be constructed from samples along a trajectory to identify high-quality CVs as long as these samples satisfy a so-called quasi-stationary distribution (QSD).² However, this new diffusion map technique was applied and tested for toy models and molecular systems that can be well described by classical force fields. Even so, the CVs and reaction coordinates for simple systems can often be chosen by intuition, and such new formulations have not been fully tested on chemically complex reactive chemistry.

Received: April 17, 2023

Published: August 16, 2023



In this work, we apply the diffusion map technique proposed in² to a hydrogen combustion benchmark dataset that contains ab initio MD (AIMD) trajectories and normal modes along minimum energy paths (MEPs) traversing through transition states between reactant and products.¹ Unlike conformational coordinates for systems such as alanine dipeptide, the main reaction coordinates of the hydrogen combustion reaction channels are less obvious as bonds break and form. To the best of our knowledge, this is the first time this approach has been used to select CVs and identify the main reaction coordinate for realistic chemical systems in which quantum mechanical effects are important. We evaluate and compare CVs obtained from two commonly used approaches, i.e., internal coordinates (ICs) and principal component analysis (PCA) to identify potentially good CVs by computing the correlations between different CVs and the leading DCs obtained from a local diffusion map constructed from MD snapshots within a metastable region of the configuration space. We also examine the possibility of using diffusion maps to compute committor functions which describe the probability of reaching either of two local minima from a particular configuration using an ensemble of short trajectories.^{21,22} The committor function is of central importance because it generalizes the concept of a transition state by explicitly accounting for dynamics on a high-dimensional PES.

We first show that the constant scaling parameter used to define the kernel matrix from which the DCs are derived from the local diffusion maps, which works well for configurational problems like alanine dipeptide, can be made more robust for chemically reactive systems by making the scaling parameter a configuration-dependent quantity. With this, our results show that for most test problems, the leading DCs obtained from a local diffusion map and QSD criteria enable us to identify good PCA-based candidate CVs. However, we also observe that not all IC-based CVs that are highly correlated with the DCs are important ones for the hydrogen combustion reactions, although ICs might be an intuitive choice. Using a global diffusion map constructed from MD snapshots initiated from the transition state, we observe that the value of the committor function is close to 0.5 at the transition state for all systems, which is expected. And while the committor function separates the reactant and product regions well when projected into a low-dimensional space spanned by previously identified good CVs, for systems with larger number of degrees of freedom such a separation becomes less clear, indicating the potential limitation of such analysis for large systems.

The paper is organized as follows. In Section 2.1, we review several techniques to construct a diffusion map. In Section 2.2, we introduce the mathematical description of the overdamped Langevin dynamics and the QSD which is the required distribution of molecular configurations used to construct an effective diffusion map. In Section 2.3, we describe an algorithm to compute committor probabilities based on the diffusion map. We review two commonly used approaches for constructing CVs in Section 2.4 and provide details on the hydrogen combustion dataset in Section 2.5. In Section 3 of results, we report the good CVs for all hydrogen combustion reactions identified by diffusion maps and validate such an assessment by examining the PES of four representative reactions in different pairs of CVs. We also show the committor functions obtained from global diffusion maps for four representative reactions and discuss how well they

characterize the reaction mechanism of these reactions. Finally, we offer concluding remarks and future directions in Section 4.

2. THEORY AND METHODS

2.1. Diffusion Coordinates and Diffusion Map Construction. We consider an overdamped Langevin dynamics

$$dx_t = -\nabla U(x_t) dt + \sqrt{2\beta^{-1}} dW_t \quad (1)$$

where x is a molecular configuration, $U(x)$ is the potential energy at x that takes into account quantum mechanical forces among different atoms, β is the inverse temperature, and W_t is a standard Wiener process. The infinitesimal generator of the Markov process $(x_t)_{t \geq 0}$ associated with (1) is related to the Kolmogorov operator

$$L = -\nabla U \cdot \nabla + \beta^{-1} \Delta \quad (2)$$

For example, for any smooth observable $f: \Omega \rightarrow \mathbb{R}$ in $L_2(\Omega)$, the expected value of $f(x_t)$ along a trajectory governed by the overdamped Langevin dynamics (1), denoted by $g(x, t) := \mathbb{E}[f(x_t)]$, satisfies the backward Kolmogorov equation

$$\frac{\partial g}{\partial t} = Lg \quad (3)$$

where the region Ω is compact, the elliptic operator L has a discrete set of eigenvalues λ_j and the corresponding eigenfunctions $\psi_j(x)$. The solution to (3) can be expressed in terms of eigenpairs $(\lambda_j, \psi_j(x))$ of L , i.e.,

$$g(x, t) = \sum_{j=1}^{\infty} b_j e^{-\lambda_j t} \psi_j(x) \quad (4)$$

where the coefficient b_j is derived from the duality with the forward Kolmogorov operator¹⁹ and the eigenvalues of L satisfy that $0 = \lambda_1 < \lambda_2 \leq \lambda_3 \leq \dots$. The nonzero eigenvalues and the corresponding eigenfunctions of L evaluated at x can be used to define a set of DCs

$$(e^{-\lambda_2 t} \psi_2(x), e^{-\lambda_3 t} \psi_3(x), \dots) \quad (5)$$

As a result, the distance between x and y can be measured in terms of the L_2 norm of the DCs associated with x and y . This is often referred to as the diffusion distance $D_t^{(x,y)}$, where

$$D_t^2(x, y) = \sum_{j \geq 2} e^{-2\lambda_j t} |\psi_j(x) - \psi_j(y)|^2 \quad (6)$$

If there is a large spectral gap between the λ_{k+1} and λ_{k+2} , the diffusion coordinate (5) is dominated by the first k components

$$(e^{-\lambda_2 t} \psi_2(x), e^{-\lambda_3 t} \psi_3(x), \dots, e^{-\lambda_k t} \psi_k(x), e^{-\lambda_{k+1} t} \psi_{k+1}(x)) \quad (7)$$

As a result, the diffusion distance $D_t^{(x,y)}$ can be computed from (6) by keeping the first k terms. The use of a k -component diffusion coordinate for a relatively small k allows us to achieve significant dimension reduction.

However, finding the DCs in high dimensions by computing the eigenvalues and eigenvectors of a discretized (by, e.g., a finite-element method) L defined in (2) is computationally intractable in general. Alternatively, one can obtain approximate DCs (5) by constructing a transition probability matrix

from configurations sampled along a trajectory using a Gaussian kernel. This kernel is referred to as a diffusion map.^{2,23} In the following, we briefly describe the basic steps of a diffusion map construction and how DC can be obtained from a diffusion map.

The initial step in the construction of a diffusion map associated with the operator (2) is to build a kernel matrix from configurations $\{x_i\}_{i=1}^n$ sampled along a trajectory of an overdamped Langevin dynamics (1), which is of the form

$$[A_\epsilon]_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{\epsilon}\right) \quad (8)$$

The local parameter ϵ scales as distances between the samples. The kernel matrix A_ϵ can be further normalized by, for example, configuration-dependent densities or the row sums of the kernel matrix, to yield¹⁹

$$[\hat{A}_{\epsilon,\alpha}]_{ij} := \frac{[A_\epsilon]_{ij}}{p_i^\alpha p_j^\alpha} \quad (9)$$

where $p_i := \sum_{k=1}^n [A_\epsilon]_{ik}$ is the i th row sum of A and $\alpha \in [0, 1]$ is an appropriate chosen parameter.

To obtain a transition probability matrix $P_{\epsilon,\alpha}$ from $\hat{A}_{\epsilon,\alpha}$, we scale it by a diagonal matrix D , i.e.,

$$P_{\epsilon,\alpha} = D^{-1} \hat{A}_{\epsilon,\alpha} \quad (10)$$

where the i th diagonal entry of D equals the i th row sum of $\hat{A}_{\epsilon,\alpha}$. Importantly, in the limit of $n \rightarrow \infty$ and $\epsilon \rightarrow 0$

$$\frac{P_{\epsilon,\alpha} - I}{\epsilon} \rightarrow L \quad (11)$$

This expression allows us to construct an approximation to L from the transition probability matrix of sampled configurations along a Langevin trajectory.²⁵ Therefore, we can obtain approximate DCs (7) by computing eigenvalues and eigenvectors of the matrix on the left-hand side of (11). The parameter $\alpha \in [0, 1]$ determines the type of the continuous operator on the right hand side of (11) in the limit $n \rightarrow \infty$ and $\epsilon \rightarrow 0$. In particular, when $\alpha = \frac{1}{2}$, $\hat{A}_{\epsilon,\alpha}$ converges to the backward Kolmogorov operator that appears in (2) and (3) (see²³).

Several variations of the diffusion map have been developed to improve the effectiveness of dimension reduction through the use of DCs. For MD simulations, one can use the energy-based kernel suggested in¹⁹

$$[\hat{A}]_{ij} := \frac{[A_\epsilon]_{ij}}{\sqrt{e^{-\beta U(x_i)} e^{-\beta U(x_j)}}} \quad (12)$$

where $U(x)$ represents the energy of the configuration x and β is the inverse temperature. A generalization of this kernel was established and proven theoretically to ensure a small set of DCs can capture the underlying manifold \mathcal{M} in.²⁶ Nevertheless, in situations where sampling is highly varied due to some structural instability, finding a proper scaling parameter ϵ in the kernel (8) can be difficult due to large variations in local scales associated with the sampling. In such situations, it may be better to define a kernel matrix with configuration-dependent scaling parameters $\{\epsilon_i\}_{i=1}^n$ instead of a single constant,^{17,27} for example

$$[A_{\{\epsilon_i\}_{i=1}^n}]_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{\epsilon_i \epsilon_j}\right) \quad (13)$$

A simple rule of selecting those configuration-dependent scaling parameters is to use the nearest neighborhood criterion, i.e.,

$$\epsilon_i = \text{the } r\text{-}n\text{-th smallest value of the distances} \\ \{\|x_i - x_j\|\}_{j=1}^n \quad (14)$$

where the parameter $0 < r < 1$ determines the size of neighborhood. A range of procedures for selecting the local scales have been proposed.^{27,28}

At an abstract level, the application of the normalization defined in (9) or (12) to the locally scaled kernel (13) can be viewed as a special case of the *weighted* kernel matrix²⁷

$$[\hat{A}_{\{\epsilon_i\}_{i=1}^n}]_{ij} = \sqrt{w_i w_j} \exp\left(-\frac{\|x_i - x_j\|^2}{\epsilon_i \epsilon_j}\right) \quad (15)$$

For example, the weighting factor used in (9) is $w_i = 1/p_i$. It is $w_i = e^{\beta U(x_i)}$ in (12). Due to the presence of configuration-dependent local scale parameter ϵ_i such as the one defined in (14), this formulation yields more robust diffusion maps than the ones constructed from a constant scale kernel (8). From our experiments with ab-initio MD and potential energy data, we found that the diffusion map constructed from the formulation (15) provided more reasonable information than diffusion maps defined in (9) or (12).

2.2. Local Diffusion Map and Quasi-Stationary Distribution. Trajectories from the Langevin dynamics (1) tend to stay in a metastable region for a very long time before exiting the region. It has been suggested that DCs obtained from a “local” diffusion map constructed from samples along such a trajectory can be used to identify high-quality CVs as long as these samples satisfy a so-called QSD.^{2,29}

By definition, a QSD for the stochastic process (1) is a probability measure ν on a metastable region Ω that satisfies

$$\lim_{t \rightarrow \infty} \mathbb{E}[f(x_t) | T > t] = \int_{\Omega} f(x) \nu(dx) \quad (16)$$

for any initial configuration $x_0 \in \Omega$ and any smooth function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ and the random variable T is the first exit time of the process x_t , i.e.,

$$T := \inf\{t \geq 0: x_t \notin \Omega\} \quad (17)$$

It follows from (16) that the expected value of $f(x)$ in Ω with respect to the measure ν can be approximately obtained from the expected value of $f(x_t)$ along a sufficiently long Langevin trajectory that stays within Ω .

A practical question we need to address in using samples along a Langevin trajectory to construct a local diffusion map is to ensure that these samples are within a metastable region Ω and satisfy a QSD. A practical procedure for achieving such a goal is based on the connection between the first eigenfunction u_1 of the Kolmogorov operator (eq 2) and $\nu(x)$ described by the equation²⁹

$$\nu(x) = \frac{u_1(x) e^{-\beta V(x)}}{\int_{\Omega} u_1(x) e^{-\beta V(x)} dx} \quad \text{for all } x \in \Omega \quad (18)$$

The expression given in (18) suggests that we can determine whether samples along a Langevin dynamics satisfy a QSD by monitoring the first eigenvalue of the operator L associated with the diffusion map constructed from these samples. When the eigenvalue does not change much, we can consider the samples to satisfy a QSD. In practice, we construct diffusion maps from a trajectory data iteratively and keep track of their first few eigenvalues instead of just the first one to determine whether samples along a trajectory satisfy the QSD.

2.3. Global Diffusion Map and Committor Function.

The diffusion map is an effective tool for computing the committor function.^{2,30} For two metastable regions A and B , the committor function assigns to snapshot $x \in \Omega$ the probability of a trajectory starting from x to reach B first rather than A , namely

$$q(x) := \mathbb{P}\{\tau_B < \tau_A | x_0 = x\} \quad (19)$$

where the random variable τ_A (τ_B) is the first time when a trajectory initialized at x_0 hits A (B). Note that $q(x) = 0$ if $x \in A$ and 1 if $x \in B$. Ideally, as a trajectory gets closer to B from A , the committor function continuously increases from 0 to 1 and the isocommittor surfaces ($q(x) = 0.5$) can be used to define transition regions.³¹

On a global region Ω that includes A and B , the committor function can be interpreted as a solution to the backward Kolmogorov equation with a boundary condition

$$\begin{cases} Lq = 0, & x \in \Omega - (A \cup B) \\ q(x) = 0, & x \in A \\ q(x) = 1, & x \in B \end{cases} \quad (20)$$

where L is the operator defined in (2). To find an approximate solution, one can exploit the asymptotic property of the diffusion map (11) and can compute an approximate committor function q by solving the following system of linear equations^{2,30}

$$(P_{e,\alpha} - I)[c, c]q[c] = -(P_{e,\alpha} - I)[c, b]q[b] \quad (21)$$

where c and b represent the set of indices for configurations in the complement of $A \cup B$ and in the metastable region B , respectively. Furthermore, if local distance scales are highly variable for given snapshots, one can use the kernel defined with different local scales (15) for solving the eq 21.

2.4. Dimension Reduction and the Choice of CVs.

There are many ways to choose CVs. In this work, we focus on two particular types of CVs, ICs and PCA, that are commonly used to define reaction coordinates of molecular systems. ICs refer to bond lengths, bond angles, and dihedral angles of a bonded molecule. Compared to Cartesian coordinates, ICs have the desirable advantage that they are invariant under an overall translation or rotation of the molecule which do not alter the potential energy of the molecule. Selected ICs are sometimes used as CVs. A well-known example is the use of two dihedral angles of an alanine dipeptide molecule as CVs to examine conformational changes of the molecule.^{2,27} These will be examined with respect to DCs in Section 3.

PCA is one of the most common and useful dimensionality reduction techniques to represent complex information as low-dimensional data. We apply the PCA to identify directions in which the data are varied most, [e.g., the projection of data onto the 1st principal component (PC) with the largest variance, the second largest variance in the second PC, and so

on]. Given $X \in \mathbb{R}^{3n \times N}$, where N is the number of samples, the PCs are defined as the eigenvectors of the covariance of X or the right-singular vectors via the singular value decomposition

$$X - \bar{X} = Y\Sigma Z^T \quad (22)$$

where \bar{X} is the mean overall samples, column vectors of Y are the PCs and the diagonal matrix Σ contains the singular values of $X - \bar{X}$. Each configuration can be expanded as a linear combination of the PCs contained in Y . The coefficients of the j th PC (y_j) for all configurations contained in X can be obtained from

$$C_j = X^T y_j \quad (23)$$

If the first few singular values in Σ are much larger than other singular values, each configuration in X can be well represented by the first few PCs. The coefficients associated with these PCs can be used as CVs.

2.5. Hydrogen Combustion Dataset. We consider the effectiveness of using a local diffusion map to assess the quality of CVs obtained for several reactions in the hydrogen combustion benchmark data published in.¹ Table 1 lists all

Table 1. 19 Reactions Contained in the Hydrogen Combustion Benchmark Dataset^a

no. reaction	atoms	DoF	DoF _{int}
Association/Dissociation			
5. H ₂ → 2H	2	6	1
6. O ₂ → 2O	2	6	1
7. OH → O + H	2	6	1
8. H + OH → H ₂ O	3	9	3
9. H + O ₂ → HO ₂	3	9	3
15. H ₂ O ₂ → 2OH	4	12	6
Substitution			
16. H ₂ O ₂ + H → H ₂ O + OH	5	15	9
O-transfer			
1. H + O ₂ → OH + O	3	9	3
11. HO ₂ + H → 2OH	4	12	6
12. HO ₂ + O → OH + O ₂	4	12	6
H-transfer			
2. O + H ₂ → OH + H	3	9	3
3. H ₂ + OH → H ₂ O + H	4	12	6
4. H ₂ O → 2OH	4	12	6
10. HO ₂ + H → H ₂ +O ₂	4	12	6
13. HO ₂ + OH → H ₂ O + O ₂	5	12	9
14. 2HO ₂ → H ₂ O ₂ + O ₂	6	18	12
17. H ₂ O ₂ + H → HO ₂ + H ₂	5	15	9
18. H ₂ O ₂ + O → HO ₂ + OH	5	15	9
19. H ₂ O ₂ + OH → H ₂ O + HO ₂	6	18	12

^aThe number of atoms involved in each reaction, the total number of degrees of freedom (DoF) in Cartesian coordinates, and total number of degrees of freedom in ICs (DoF_{int}).

19 reactions contained in the hydrogen combustion benchmark data.¹ The benchmark data contain molecular configurations sampled along an MEP for each reaction, including the reactant, transition, and product states. The dataset contain 290,000 potential energies and 1,270,000 forces for hydrogen compounds in different molecular configurations. These configurations are generated through normal mode sampling and AIMD at different temperatures beginning from different points along the 0 K intrinsic reaction coordinate (IRC) or

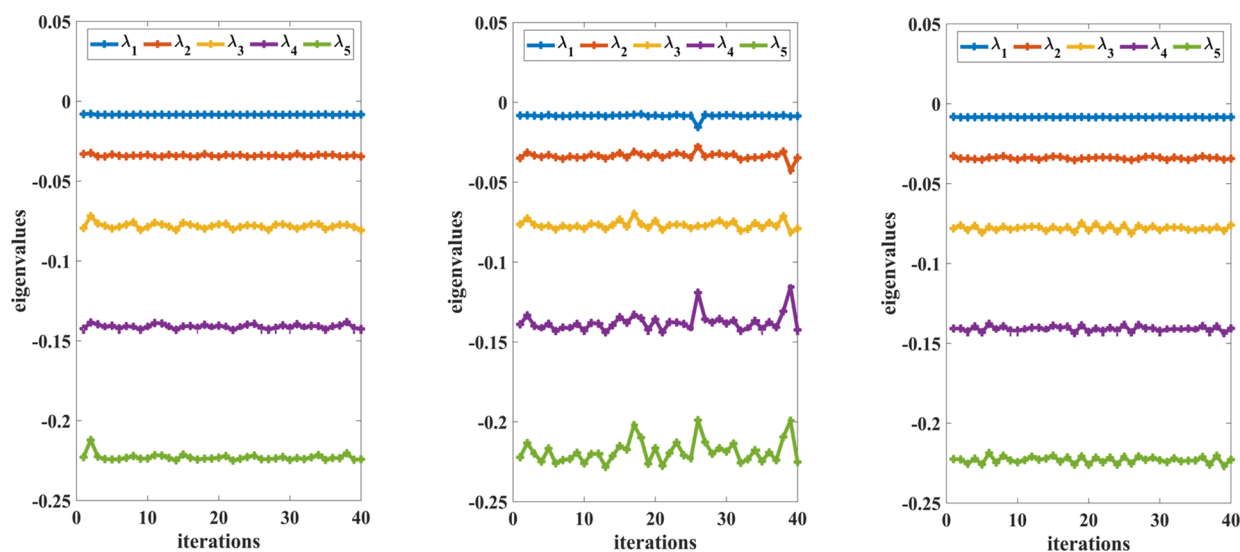


Figure 1. The x -axis and the y -axis represent the QSD checking iteration number and the amplitudes of eigenvalues of diffusion maps, defined in (7), respectively. The changes of the first few dominant eigenvalues of the diffusion maps with respect to the QSD checking iteration number are shown for reactions 11 (left), 14 (middle), and 16 (right). The diffusion maps are constructed every $m = 100$ AIMD steps of a 4000-step trajectory from snapshots sampled along the trajectory.

equivalently MEP, including the transition state. In regards the AIMD simulations for the 19 reaction channels, each reaction has 10,000 snapshots obtained at four different temperatures: 500, 1000, 2000, and 3000 K generated from running AIMDs using the Q-Chem software package³² with the transition state as the starting point. As reported in,¹ some of the reactions are relatively simple, e.g., reactions 5 and 6, involving the association and dissociation of two atoms along certain directions which constitutes the only degree of freedom; we will not examine this type of reaction in this paper. Among the 19 reaction channels, we focus on reactions 09, 11, 14, and 16 as representatives for association, O-transfer, H-transfer, and substitution, respectively. These reactions also include molecules with more atoms such that the number of degrees of freedom in ICs can be as large as 18.

We should also note that in,¹ the MEP is plotted in terms of two coordination numbers (CNs) that are considered as internal reaction coordinates (IRC). A CN is defined as the number of bonded or closest neighbors of a central atom or molecule of interest. In the hydrogen combustion study,¹ the i th atom in a molecule is defined as

$$CN_i = \sum_{j \neq i} \frac{2.0}{1 + \exp(\sigma(r_j - r_{0,j}))} \quad (24)$$

where r_j is the distance between atom i and atom j and $r_{0,j}$ is the equilibrium distance between atom i and j . Each Fermi-Dirac function in (24) is close to zero if $r_j - r_{0,j}$ is large and close to 1 if $r_j - r_{0,j}$ is small. Therefore, (24) effectively gives the number of neighboring atoms of atom i that are close to be in equilibrium positions. In,¹ one or two CNs were chosen as CVs that incorporated the IRC MEP that connects a local minimum of the PES (reactant) to another local minimum (product).

To generate additional samples near the reactant and product states (local minima), which are often available in a practical setting, we ran some additional MD simulations using QChem³² starting from either the reactant or product state. We used density functional theory (DFT), specifically the

ω B97X-V functional and the cc-pVTZ basis set to perform potential energy calculations and MD simulations. For all MD simulations, we performed Langevin dynamics at 300 K with a 0.12 femtosecond (fs) time step; we set the reactant provided in the IRC data¹ as the initial configuration of the dynamics. For reactions 11, 14, and 16, we generated 4000 snapshots. We could only generate 105 snapshots for reaction 09 before dissociation occurred.

3. RESULTS

3.1. Checking QSD for Local Diffusion Map Construction.

As we discussed in Section 2.1, a local diffusion

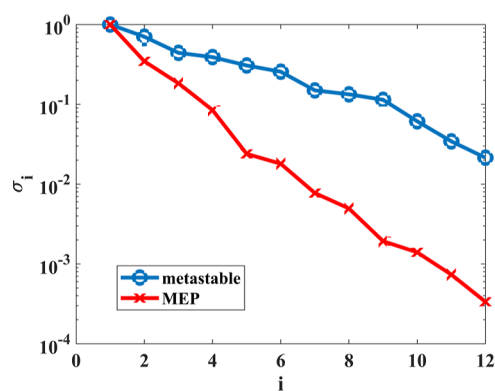


Figure 2. The x -axis and the y -axis represent the index and the magnitude of singular values (in a log scale) obtained from (22), respectively. Normalized singular values of two matrices containing configurations sampled within a metastable region (blue) and along the MEP (red) of reaction 19 are shown. The normalization is performed by dividing all singular values by the largest singular value.

map constructed from molecular configurations within a metastable region can be used to assess the quality of CVs as long as the configurations used to construct the diffusion map satisfies a QSD.³³

Due to the connection between the distribution of configurations along an MD trajectory and the eigenfunction

Table 2. Absolute Values of Correlation Coefficients between ICs and PCA-based CVs and DC2, DC3 for Representative Hydrogen Combustion Reaction Channels 9, 11, 14, and 16^a

	DC2	DC3
PC1	0.94	0.29
PC2	0.15	0.44
PC3	0.29	0.5
IC1 (O1-O2)	0.65	0.42
IC2 (O1-H)	0.94	0.29
IC3 (H-O1-O2)	0.98	0.15

(a) Association reaction 09

	DC2	DC3
PC1	0.99	0.03
PC2	0.08	0.82
PC3	0.12	0.44
IC1 (H1-O1)	0.00	0.01
IC2 (H1-O2)	0.03	0.02
IC3 (H1-H2)	0.99	0.04
IC4 (O2-H1-O1)	0.06	0.01
IC5 (H2-H1-O1)	0.62	0.65
IC6 (H2-H1-O1-O2)	0.82	0.47

(b) O-transfer reaction 11

	DC2	DC3
PC1	0.98	0.12
PC2	0.14	0.94
PC3	0.02	0.11
IC1 (H1-O1)	0.04	0.14
IC2 (H1-O2)	0.85	0.36
IC3 (H1-H2)	0.84	0.49
IC4 (O1-O3)	0.29	0.72
IC5 (O2-O4)	0.04	0.23
IC6 (O2-H1-O1)	0.88	0.28
IC7 (H2-H1-O1)	0.90	0.28
IC8 (O3-O1-O2)	0.1	0.69
IC9 (O4-O2-H2)	0.06	0.57
IC10 (H2-H1-O1-O2)	0.19	0.52
IC11 (O3-O1-O2-H2)	0.20	0.27
IC12 (O4-O2-H2-O3)	0.07	0.38

(c) Substitution reaction 16

	DC2	DC3
PC1	0.99	0.03
PC2	0.08	0.28
PC3	0.08	0.93
IC1 (H1-O1)	0.01	0.00
IC2 (H1-O2)	0.02	0.01
IC3 (H1-H2)	0.01	0.04
IC4 (O1-H3)	0.98	0.09
IC5 (O2-H1-O1)	0.03	0.02
IC6 (H2-H1-O1)	0.01	0.08
IC7 (H3-O1-O2)	0.61	0.77
IC8 (H2-H1-O1-O2)	0.01	0.08
IC9 (H3-O1-O2-H2)	0.79	0.08

(d) H-transfer reaction 14

^aWe highlight some of the relatively high correlation coefficients in red.

associated with the first eigenvalue of corresponding Kolmogorov operator (2) as discussed in Section 2.1, we can check whether the sampled MD snapshots satisfy QSD by monitoring the spectrum of the diffusion map constructed from these snapshots. To be specific, every $m = 100$ MD steps, which constitutes an iteration of the QSD checking procedure, we construct a diffusion map based on the weighted kernel matrix (15) with weights being free energies and local scale parameters from (14) with $r = 0.1$, using samples along the trajectories available at that point and compute the first few eigenvalues of the corresponding approximate Kolmogorov operator L_e . These eigenvalues are shown in Figure 1 for reactions 11, 14, and 16. Because each MD trajectory consists of $N = 4000$ snapshots, $N/m = 4000/100 = 40$ iterations were performed to monitor the change in the eigenvalues of L_e .

Figure 1 shows that the first 5 eigenvalues of the diffusion map for reaction channels 11 and 16 do not change much during the entire simulation. For reaction channel 14, the dominant eigenvalues of the diffusion map begin to fluctuate around a few mean values, indicating that the subsequent trajectory samples no longer fulfill the QSD defined on the metastable region. Based on those measures, we pick the subset of the snapshots up to the point when the first 5 eigenvalues of the diffusion map begin to change more significantly. The configurations within such a subset are deemed to satisfy QSD. For reaction 9, because there are only 105 snapshots in the AIMD trajectory, we use all of them because they appear to be within a metastable region.

3.2. Implementation Details. In all experiments, we used ICs of sampled configurations to construct local diffusion maps. For global diffusion maps used in Section 3.4 to perform

committor analyses, we found that root mean-squared deviation aligned Cartesian coordinates, as defined in,³⁴ were sometimes more effective. When using ICs to construct a diffusion map kernel matrix, it is important to note that the absolute difference in two angles should never be larger than π when the difference of two ICs is used to evaluate a kernel matrix element defined in (8). For example, if $\theta_1 = \pi/6$ and $\theta_2 = 2\pi - \pi/6$, the absolute difference between θ_1 and θ_2 should be $\pi/3$ instead of $2\pi - \pi/3$, i.e., when the difference between two angles exceeds π , we need to subtract 2π from the difference. Otherwise, the Euclidean difference between the two sets of ICs can be artificially increased by the extra π in the angle difference. This increase can yield distorted DC.

We perform PCA on sampled ICs. Ideally, we would like to use configurations along or close to the MEP because singular values resulting from PCA are likely to decrease faster and the dominant PCs obtained from such an analysis are likely to represent the main reaction mechanism well. For example, Figure 2 shows that the singular values obtained from the PCA performed on the configurations sampled along the MEP decreases much faster than those obtained from the configurations sampled within a metastable state of reaction 19. However, because the MEP is unknown in general, this approach is not practical. In the following, we use MD snapshots sampled within a metastable region to perform the PCA even though such an analysis is not optimal in the sense that more PCs may be required to capture the reaction mechanism.

In addition to using diffusion maps to assess the quality of CVs, we validate the quality assessment by plotting the PES in these CVs to see if a saddle point can be observed on such a surface. The presence of a saddle point would indicate that the chosen CVs are good CVs for describing the reaction mechanism near the transition state. To plot the PES with respect to two CVs c_1 and c_2 , we evaluate the potential energy at

$$z = z_{\text{ts}} + c_1 u_1 + c_2 u_2 \quad (25)$$

where z_{ts} denotes coordinates of the transition state and u_1, u_2 are either two elementary basis vectors, i.e., columns of an identity matrix when ICs are chosen as the CVs, or two PCs when PCA-based CVs are to be examined. We plot the energies

$$E(c_1, c_2) \equiv E(z_{\text{ts}} + c_1 u_1 + c_2 u_2) \quad (26)$$

on a uniform sampled 2D domain $[c_1^{\text{lb}}, c_1^{\text{ub}}] \times [c_2^{\text{lb}}, c_2^{\text{ub}}]$ for appropriately chosen $c_1^{\text{lb}}, c_1^{\text{ub}}, c_2^{\text{lb}},$ and c_2^{ub} values.

3.3. Assessing the Quality of CVs. We construct the diffusion map and use it to obtain two DCs: DC2 and DC3. We use these DCs as a means to assess different definitions of CVs by examining the correlation of CVs with the DCs. To be specific, we compute the Pearson correlation coefficient between CVs and the DCs as follows

$$\rho(C, D) = \frac{\text{cov}(C, D)}{\sigma_C \sigma_D} \quad (27)$$

where $\text{cov}(C, D)$ denotes the covariance between random variables C and D , and $\sigma_C(\sigma_D)$ denotes the standard deviation of C (D). If C or D is chosen as an IC, it can be easily obtained directly from the Cartesian coordinates of the snapshots used to construct the diffusion map. For the PCA, we use the variable in eq 23.

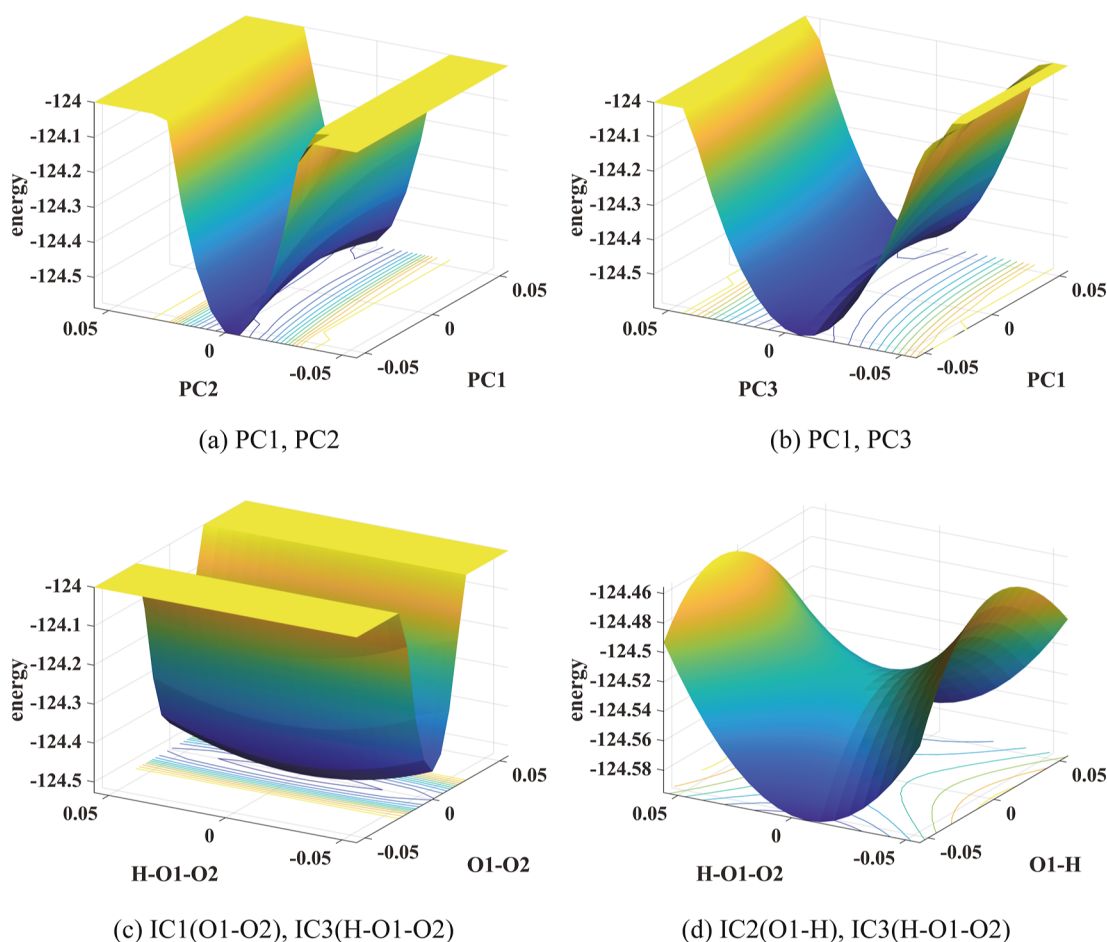


Figure 3. PESs of reaction 9 in pairs of PCs and pairs of ICs show relatively high (yellow) and low (blue) energy regions. The unit of energy is kcal mol⁻¹. In the lower panels, the unit of the angle H–O1–O2 is radian and that of the bond length is Å.

Table 2 reports the correlations of the IC and PCA-based CVs with the first two DCs (labelled as DC2 and DC3) for each representative reaction. In the construction of the diffusion map, we set $r = 0.1$ in (14), and use the kernel (9) for (a) and (b) (3 and 4 atoms) and the energy-based kernel (12) for (c) and (d) (5 and 6 atoms). We observe that for each reaction several ICs can be highly correlated with the same DC. For example, for reaction 9, the absolute values of the correlation coefficient between IC2 and DC2 and between IC3 and DC2 are over 0.9, respectively. For reaction 11, IC3 and IC6 are highly correlated with DC2. For reaction 16, IC4 and IC9 are highly correlated with DC2. For reaction 14, IC2, IC3, IC6, and IC7 are all highly correlated with DC2. By contrast, we find that each DC tends to be highly correlated with only one of the PCA-based CVs.

To validate the diffusion map-based assessment of CVs shown in Table 2, we plot PESs in different combinations of CVs. We evaluate the energy at each pair of CVs using the formula 26 where z_{st} is the transition state specified in the hydrogen combustion dataset.¹ For PCA-based CVs, u_1 and u_2 are chosen to be the first and second PC vectors or the first and the third ones. For IC-based CVs, u_1 and u_2 are standard basis vectors, i.e., columns of an identity matrix, with the position of 1 indicating which IC is chosen. When the projected transition state corresponds to a saddle point on this PES, we consider the CVs in which the PES is shown as good CVs. Figure 3a,b shows that a saddle point can be found in the

projected PES in both (PC1,PC2) and (PC1,PC3) for reaction 9. PC1 appears to define the main reaction coordinate for reaction 9 along which the barrier is located at the full transition state. This is consistent with the high correlation coefficient shown in Table 2. PC2 and PC3 define the direction orthogonal to the reaction coordinate along which the barrier is a local minimum. We also observe from Figure 3d that the main reaction coordinate appears to be along IC2 which corresponds to the O1–H bond length. This is consistent with the relatively high correlation coefficient between IC2 and DC2 in Table 2 although the correlation coefficients between IC3 (H–O1–O2 angle) and DC2 is slightly higher. A saddle point can clearly be seen in Figure 3d when the PES is plotted in terms of IC2 and IC3, whereas no such saddle point can be seen in Figure 3c where the PES is plotted in IC1 (O1–O2 bond length) and IC3 (H–O1–O2 angle).

Furthermore, we observe that the three components of the first PC, which correspond to the contribution of IC1, IC2, and IC3 to the PC vector, respectively, are -0.088 , 0.996 , and 0.0142 . This indicates that IC2 contributes the most to the first PC and consistent with the observation that the CVs defined in terms of PC1 and IC2 correspond to the same reaction coordinate. No clear saddle point can be seen in Figure 3c where the PES is plotted in terms of IC1 (the O–O bond distance) and IC3 (H–O1–O2 angle). This observation indicates that the O1–O2 bond distance is not a good CV for

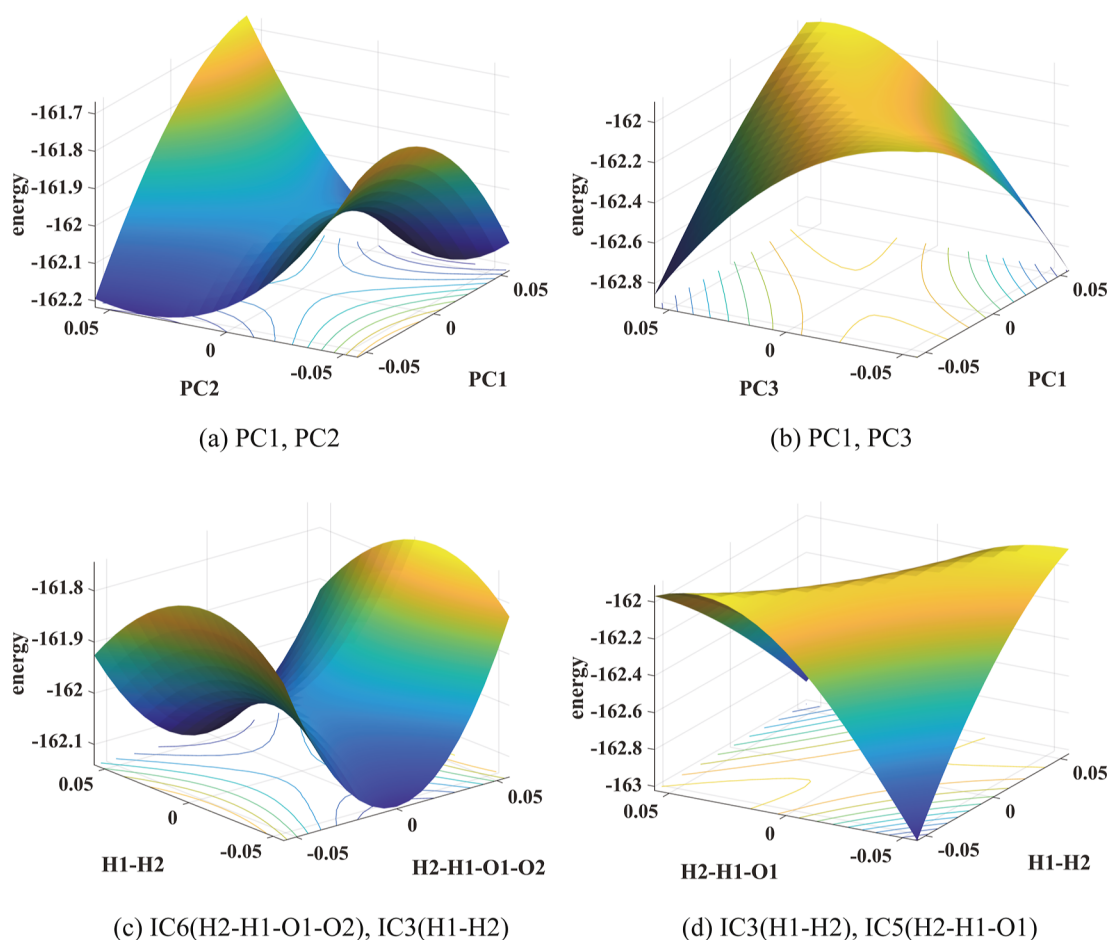


Figure 4. PESs of reaction 11 in pairs of PCs and pairs of ICs show relatively high (yellow) and low (blue) energy regions. The unit of energy is kcal mol^{-1} . In the lower panels, the unit of angles is radian and that of the bond length is \AA .

characterizing the reaction. This is also consistent with observation that IC1 has a relatively low correlation coefficient with DC2 and DC3.

For reaction 11, Table 2b shows that PC1 and PC2 are highly correlated with DC2 and DC3, respectively, indicating that they are good CVs. This observation is consistent with the PES shown in Figure 4a where a saddle point can be clearly seen at the transition state. Furthermore, it appears that the main reaction coordinate is mostly aligned with PC1, although it is not strictly parallel to PC1, and PC2 also contributes to the reaction coordinate. We see from Table 2b that the correlation coefficients between PC3 and DC2, and between PC3 and DC3 are relatively smaller, indicating that PC3 may not be a good CV. This is also evident from Figure 4b which shows that no clear saddle point can be seen from the PES plotted in PC1 and PC3. Table 2 also shows that both IC3 (H1–H2 bond length) and IC6 (H2–H1–O1–O2 dihedral angle) are highly correlated with DC2, and IC5 is moderately correlated with both DC2 and DC3. We see in Figure 4c that a saddle point is present in the PES plotted in IC3 and IC6, and the main reaction coordinate appears to be well described by IC3 (H1–H2 bond length), which is also the largest component in PC1 in magnitude. No clear saddle point is observed in Figure 4d where the PES is plotted in IC3 and IC5. These plots are consistent with the observed correlation coefficients between different ICs and DCs. They indicate that ICs that are highly correlated with DCs are indeed good CVs.

Table 2c shows that PC1 and PC3 are highly correlated with DC2 and DC3 for reaction 16, which is somewhat surprising. However, this may be explained by the observation that second and third singular values of the mean subtracted snapshot matrix are $\sigma_2 = 2.45$ and $\sigma_3 = 1.55$, which are very close. Both are an order of magnitude smaller than $\sigma_1 = 24.28$ and two to three times larger than $\sigma_4 = 0.83$. Figure 5b shows that PC1 and PC3 are indeed better CVs to describe the reaction mechanism. In fact, the main reaction coordinate seems to be along the PC1 direction. Such information cannot be obtained from the PES plotted in PC1 and PC2 shown in Figure 5a where no saddle point can be found. Table 2c also shows that IC4 (O1–H3 bond length), IC7 (H3–O1–O2 torsion angle), and IC9 (H3–O1–O2–H2 dihedral angle) have relatively high correlations with either DC2 or DC3 indicating that they may be good CVs. Figure 5d confirms that the main reaction coordinate appears to be in the direction of IC4 (O1–H3 bond length) with some contribution from IC7 (H3–O1–O2 angle). A saddle point can clearly be identified at the transition state. On the contrary, no reaction mechanism can be inferred from the PES plotted in IC1 and IC3 in Figure 5c. Both IC1 and IC3 have low correlations with DC2 and DC3.

Finally, Table 2d shows that PC1 and PC2 are highly correlated with DC2 and DC3, respectively, for reaction 14, indicating that they may be good CVs. This prediction is confirmed in Figure 6a in which the PES is plotted in PC1 and PC2. A saddle point can be observed at the full-dimensional

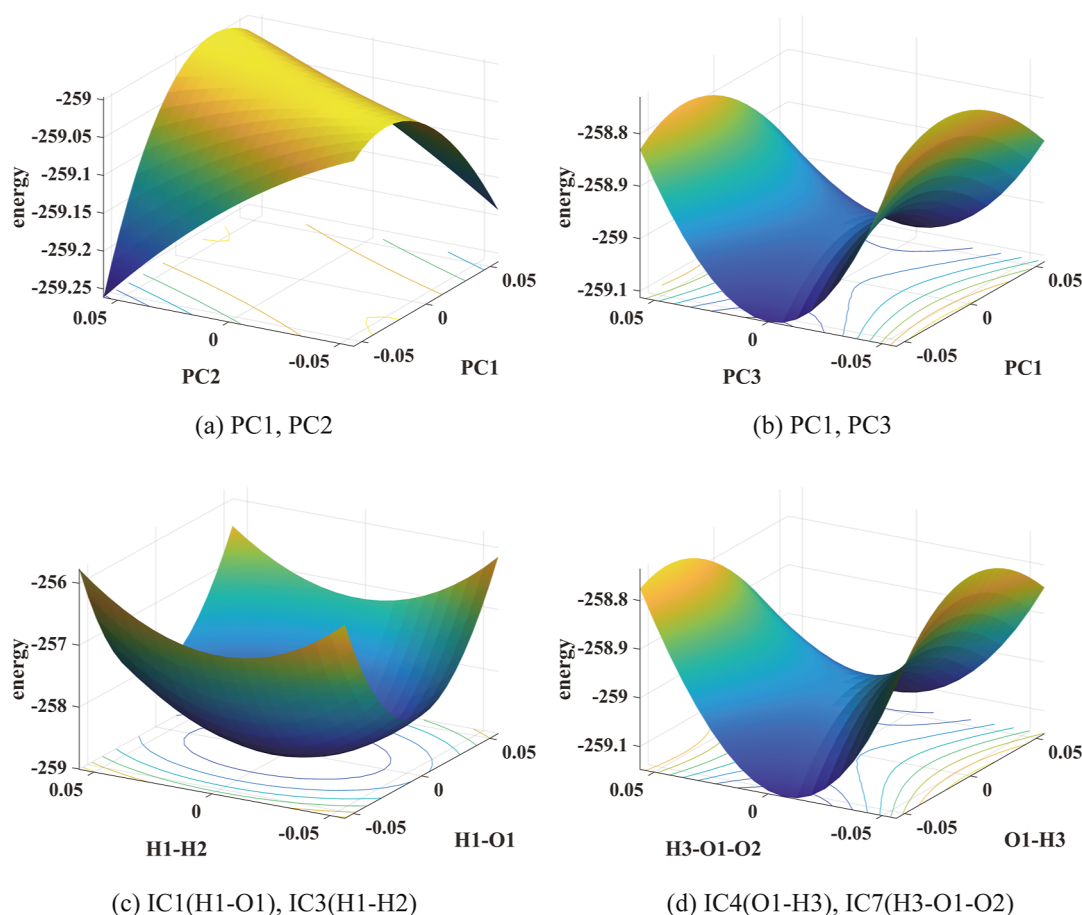


Figure 5. PESs of reaction 16 in pairs of PCs and pairs of ICs show relatively high (yellow) and low (blue) energy regions. The unit of energy is kcal mol^{-1} . In the lower panels, the unit of angles is radian and that of the bond length is Å.

transition state. However, the main reaction coordinate does not appear to be aligned with either PC1 or PC2 but a combination of them. For this reaction, the first 4 singular values of the mean subtracted snapshot matrix are $\sigma_1 = 9.41$, $\sigma_2 = 5.59$, $\sigma_3 = 3.99$, and $\sigma_4 = 2.26$. Because σ_1 is not significantly larger than σ_2 , both PC1 and PC2 (and possibly PC3 also) are important in describing the main reaction coordinate. We see from Figure 6b that no saddle point can be clearly observed in the PES plotted in PC1 and PC3, which indicates that PC2 is a better CV than PC3, an observation that is consistent with the conclusion drawn from the correlation coefficients reported in Table 2.

Reaction 14 has many more degrees of freedom (12) compared to other reactions considered earlier. In Figure 6c, we plot the PES in IC4 (O1–O3 bond length), which is highly correlated with DC3, and IC7 (H2–H1–O1 angle) which is highly correlated with DC2, as we can see from Table 2. A saddle point can clearly be seen at the transition state indicating that both IC4 and IC7 are good CVs as predicted by their high correlations with DC3 and DC2, respectively. The main reaction coordinate does not seem to be aligned with either one of them but a combination of the two. On the contrary, no saddle point can be observed in Figure 6d where the PES is plotted in IC12 (O4–O2–H2–O3 dihedral angle) and IC5 (O2–O4 bond length). Neither one of these ICs is highly correlated with DC2 or DC3 as we can see from Table 2d.

3.4. Committor Analysis. We use the AIMD dataset in¹ to build global diffusion maps and compute committor functions as described in (21). In our experiments, we set $r = 0.002$ in (14) to determine local scale parameters and construct a global diffusion map from snapshots at a given temperature. We verified that values of the computed committor functions at transition states are close to 0.5. Figure 7a shows the computed committor function for reaction 9 evaluated at configurations along AIMD trajectories generated at $T = 3000$ K. The committor function is plotted in the plane formed by two ICs: IC2 (O1–H distance) and IC3 (H–O1–O2 angle) that are highly correlated with the first diffusion coordinate as reported in Table 2. The variation of the IC1 (O1–O2 distance) among all the snapshots is less than 0.16 Å; therefore, the 2D view of the committor function in Figure 7a fully captures the main features of committor function. We can clearly see that the values of the committor function are closer to 1 to the left of the transition state, while they are closer to 0 to the right of the transition state.

We also plot the projected locations of the reactant and product in this two-dimensional space and observe that they are in reactant and product regions that are clearly separated by the transition state. From this figure, we can clearly see that O1–H distance is globally the main reaction coordinate, which is consistent with the observation we made in Figure 3d in a local region around the transition state. We also see that the information provided in the committor function is consistent with the free energy surface plot shown in Figure 7b in which

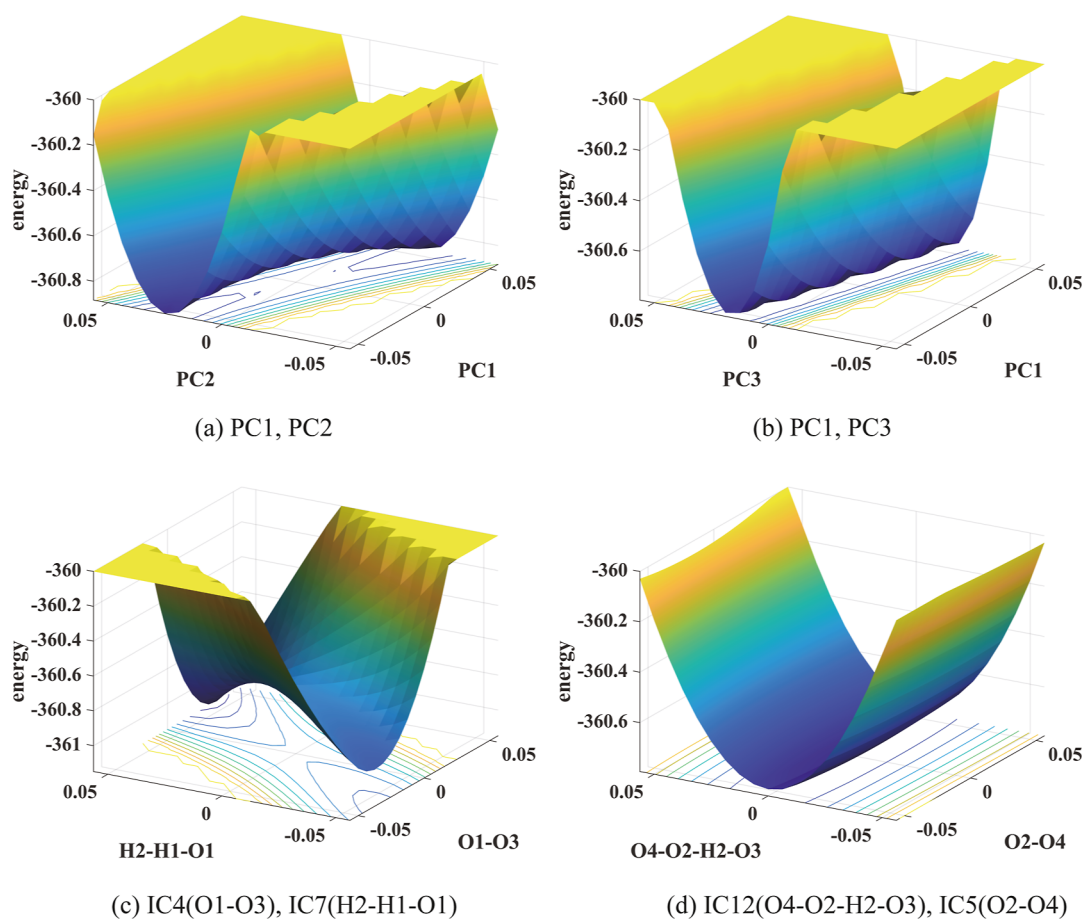


Figure 6. PESs of reaction 14 in pairs PCs and pairs of ICs show relatively high (yellow) and low (blue) energy regions. The unit of energy is kcal mol⁻¹. In the lower panels, the unit of angles is radian and that of the bond length is Å.

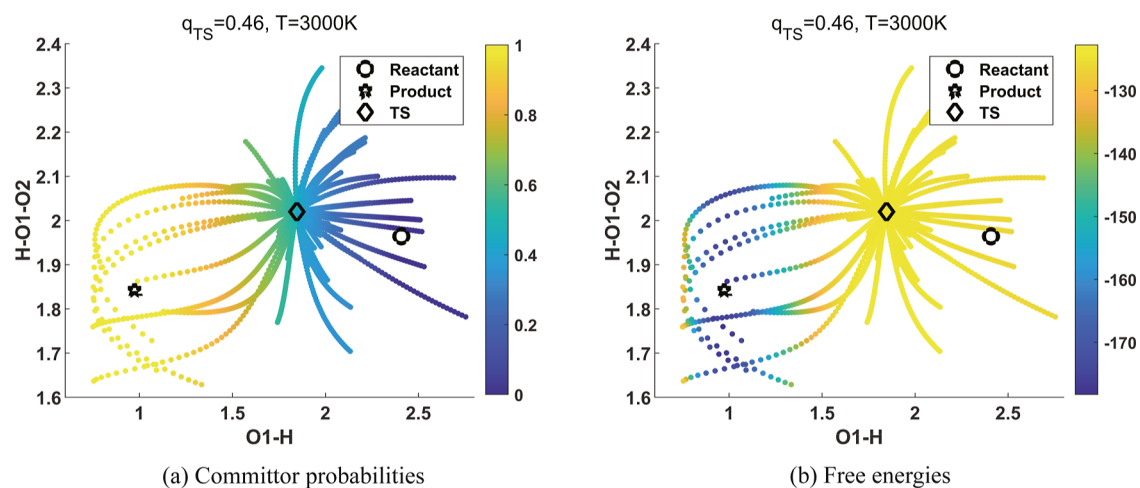


Figure 7. Committor analysis for reaction 9. The unit of angle H–O1–O2 is radian and that of bond length O1–H is Å. (a) Committor probabilities and (b) free energies at 3000 K are plotted in the space of (IC2, IC3) listed in Table 2. We used eq 9 to construct the diffusion map with $\alpha = \frac{1}{2}$ and computed the committor probabilities.

the configurations to the left of the transition state have much lower energies, whereas the configurations to the right of the transition state have slightly lower energies compared to the free energy at the transition state.

Figure 8a shows the committor functions in the PC1–PC2 plane at selected configurations along the AIMD trajectories of reaction 11 generated at $T = 500$ K. The configurations are

selected by restricting the coefficients of the other four PCs (out of a total of six) to be close to those obtained at the transition state. We again observe that the committor function has similar behavior near the transition state as before, and that in this case, PC1 is the main reaction coordinate as is consistent with the PES plot as shown in Figure 4b. However, we also observe that away from the transition state, the main

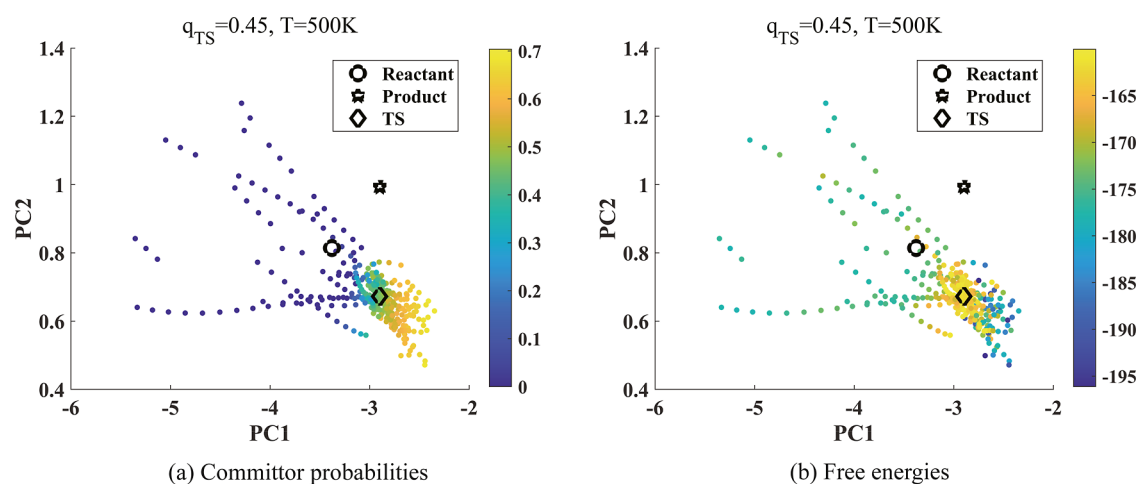


Figure 8. Committor analysis for reaction 11. (a) Committor probabilities and (b) free energies at 500 K are plotted in the space of (PC1, PC2). We used eq 9 to construct the diffusion map with $\alpha = \frac{1}{2}$ and computed the committor probabilities.

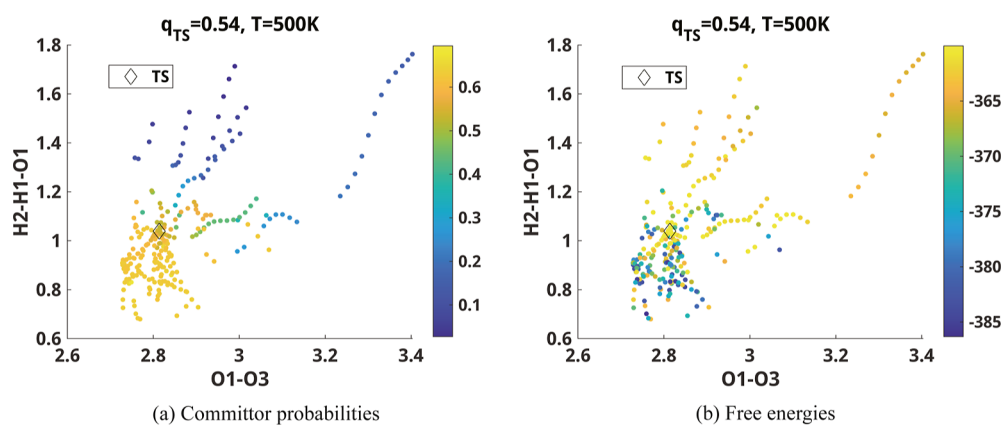


Figure 9. Committor analysis for reaction 14. The unit of angle H2–H1–O1 is radian and that of bond length H1–O3 is Å. (a) Committor probabilities and (b) free energies are shown in the space of IC4 and IC7 listed in Table 2. Snapshots near the TS are thresholded with respect to the ICs so that the two-dimensional projection of snapshots is performed approximately. We used eq 12 to construct the diffusion map and computed the committor probabilities. For reaction 14, the committor value at the TS is close to 0.5 only at 500 K.

reaction coordinate is not completely determined by PC1 as observed when we plot the projected locations of the reactant and product in the PC1–PC2 plane, in which the reactant and product PC coefficients are not close to those at the transition state. Therefore, the plotted locations of the reactant and product in the PC1–PC2 plane do not completely describe their proximity to the transition state in this figure. Because the committor function is only evaluated at configurations along the AIMD trajectory, we do not have the committor function value at configurations in which the PC1 and PC2 coefficients are fixed at those associated with the reactant or product while the coefficients of all other PCs' are fixed at those associated with the transition state.

In the case of reaction channel 14, the selected configurations with ICs close to those at the transition state except IC4 and IC7 have committor function values close to zero to the upper right of the transition state and close to one to the lower left of the transition state in the IC4–IC7 plane as shown in Figure 9a. This is consistent with the PES around the transition state plotted in Figure 6b. Near the transition state, the main reaction coordinate appears to be mostly determined by IC7 with a small contribution from IC4. However, away from the transition state, the main reaction path is determined

by a different linear combination of IC4 and IC7 depending on the location of the path.

Figure 10 shows the committor function and free energies of selected molecular configurations along the AIMD trajectories associated with reaction 16 at $T = 1000$ K. The configurations are selected to either have their ICs close to those associated with the transition state except IC4 and IC7 or have their PC coefficients close to those associated with the transition state except the coefficients of PC1 and PC3. The committor function and free energy are plotted in the plane of IC4 and IC7 and in the plane of PC1 and PC3 which are determined to be good CVs based on their correlation with the first two DCs as shown in Table 2. This is consistent with the corresponding potential energy plot near the transition state shown in Figure 5d although the number of configurations shown in these plots is limited.

4. CONCLUSIONS

Inspired by the recent work presented in,² we considered the use of diffusion maps to analyze several reaction channels involved in a hydrogen combustion system. For this purpose, the hydrogen combustion dataset is ideal for analyzing diffusion maps and committor functions as the molecular

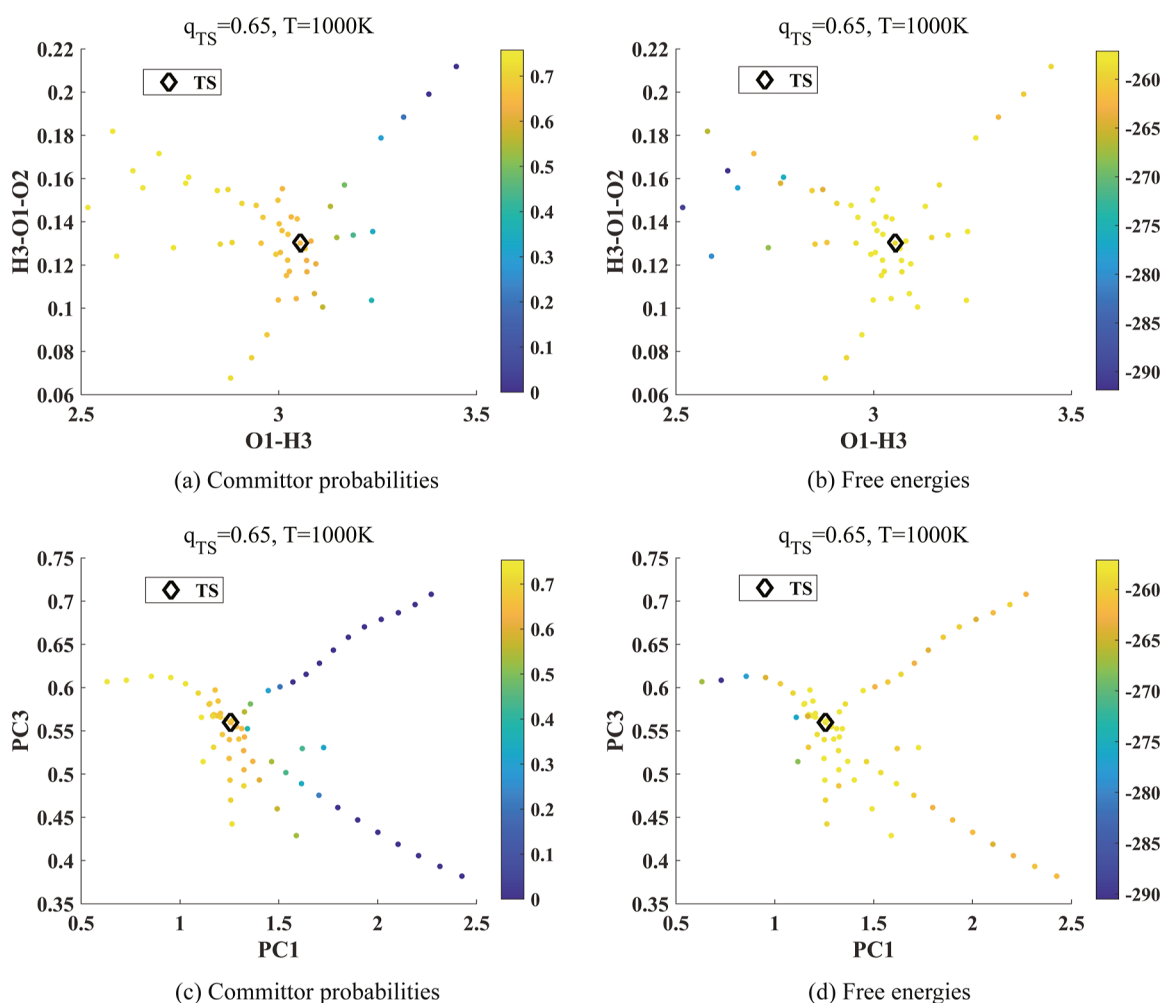


Figure 10. Committor analysis for reaction 16. The unit of angle H3-O1-O2 is radian and that of bond length O1-H3 is Å. (a,c) Committor probabilities and (b,d) free energies at 1000 K are projected onto the spaces of (IC4, IC7) and (PC1, PC3) listed in Table 2. We used eq 12 to construct the diffusion map and computed the committor probabilities.

species are relatively small, the energies (and forces) are generated from reliable DFT using the ω B97X-V DFT functional³⁵ with the cc-pVTZ basis set, and configurations near reaction barriers as well as configurations in metastable regions are well sampled for a variety of reaction channels. In particular, we use local diffusion maps constructed from configurations sampled along AIMD trajectories within the metastable region to identify and assess CVs obtained from ICs and PCA for a chemically reactive system. Unlike a system such as alanine dipeptide for which we can use a constant scaling factor in the construction of a diffusion map kernel, we found that configuration-dependent scaling was important in the construction of local diffusion maps for reactive systems such as hydrogen combustion.

We found that the correlations between the first few PCs and the first two DCs tend to be high. The two PCs that have the highest correlations with respect to the first two DCs were found to be good CVs for characterizing the reaction path near the transition state and validated by the presence of a saddle point on the PES. Although diffusion maps are in general considered as nonlinear dimension reduction techniques, whereas PCA is a linear dimension reduction technique, in a local diffusion map the leading DCs can correlate well with the leading PCs as can be seen in Figures 3–6, and also indicated

by the fast decay of the singular values. However, globally, the leading PCs from a single PCA may not be able to fully characterize the entire reaction path, and even a local diffusion map may not be sufficient to characterize the main reaction coordinate. We also observed that, while several ICs can be highly correlated with the first two DCs, they are not uniformly all good CVs. Nonetheless, the use of diffusion maps allows us to narrow down the IC choices and provides some alternatives that we otherwise would not have considered.

Using AIMD trajectories initiated from the transition states, we also constructed global diffusion maps that can be used to compute approximations to the committor functions associated with different reaction channels. When we examined the committor function in the plane of two CVs identified by a diffusion map, we found that for all reaction channels, the value of the committor function at the transition state is close to 0.5. For reaction channels that contain only a few degrees of freedom, the reactant and product are clearly separated by the selected CVs, and the main reaction path characterized by the committor function agrees well with that identified from the free energy surface in the selected CVs. However, for reaction channels that involve more degrees of freedom, the global reaction coordinate appears to be more complicated, and the reactant and product may not be easily separated on a two-

dimensional slice of committor function defined in a high-dimensional space. Even so, insights into the CVs provided by the diffusion maps have proved useful enough such that we will report the transition state free energies found for all hydrogen combustion reaction channels in a related publication. Finally, the results presented in this work provide an important baseline for comparison for future work.

AUTHOR INFORMATION

Corresponding Authors

Taehee Ko – Department of Mathematics, Penn State University, University Park, Pennsylvania 16802, United States; orcid.org/0000-0003-2522-0125; Email: tuk351@psu.edu

Joseph P. Heindel – Kenneth S. Pitzer Theory Center and Department of Chemistry, University of California, Berkeley, California 94720, United States; Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States; orcid.org/0000-0002-9748-1730; Email: heindelj@lbl.gov

Xingyi Guan – Kenneth S. Pitzer Theory Center and Department of Chemistry, University of California, Berkeley, California 94720, United States; Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States; Email: nancy_guan@berkeley.edu

Teresa Head-Gordon – Kenneth S. Pitzer Theory Center and Department of Chemistry and Departments of Bioengineering and Chemical and Biomolecular Engineering, University of California, Berkeley, California 94720, United States; Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States; orcid.org/0000-0003-0025-8987; Email: thg@berkeley.edu

David B. Williams-Young – Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States; orcid.org/0000-0003-2735-3706; Email: dbwy@lbl.gov

Chao Yang – Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States; orcid.org/0000-0001-7172-7539; Email: CYang@lbl.gov

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.jctc.3c00426>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing, and Office of Basic Energy Sciences, via the Scientific Discovery through Advanced Computing (SciDAC) program. This work used computational resources provided by the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract DE-AC02-05CH11231.

REFERENCES

(1) Guan, X.; Das, A.; Stein, C. J.; Heidar-Zadeh, F.; Bertels, L.; Liu, M.; Haghighatlari, M.; Li, J.; Zhang, O.; Hao, H.; Leven, I.; Head-

Gordon, M.; Head-Gordon, T. A benchmark dataset for Hydrogen Combustion. *Sci. Data* **2022**, *9*, 215.

(2) Trstanova, Z.; Leimkuhler, B.; Lelièvre, T. Local and global perspectives on diffusion maps in the analysis of molecular systems. *Proc. R. Soc. A* **2020**, *476*, 20190036.

(3) Fiorin, G.; Klein, M. L.; Héning, J. Using collective variables to drive molecular dynamics simulations. *Mol. Phys.* **2013**, *111*, 3345–3362.

(4) Noé, F.; Clementi, C. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Curr. Opin. Struct. Biol.* **2017**, *43*, 141–147.

(5) Karmakar, T.; Invernizzi, M.; Rizzi, V.; Parrinello, M. Collective variables for the study of crystallisation. *Mol. Phys.* **2021**, *119*, No. e1893848.

(6) Mendels, D.; Piccini, G.; Parrinello, M. Collective variables from local fluctuations. *J. Phys. Chem. Lett.* **2018**, *9*, 2776–2781.

(7) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 826–843.

(8) Sutto, L.; Marsili, S.; Gervasio, F. L. New advances in metadynamics. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 771–779.

(9) Kästner, J. Umbrella sampling. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 932–942.

(10) You, W.; Tang, Z.; Chang, C.-E. A. Potential mean force from umbrella sampling simulations: What can we learn and what is missed? *J. Chem. Theor. Comput.* **2019**, *15*, 2433–2443.

(11) Belkacemi, Z.; Gkeka, P.; Lelièvre, T.; Stoltz, G. Chasing Collective Variables using Autoencoders and biased trajectories. *J. Chem. Theory Comput.* **2021**, *18*, 59–78.

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

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

(14) Sultan, M. M.; Pande, V. S. Automated design of collective variables using supervised machine learning. *J. Chem. Phys.* **2018**, *149*, 094106.

(15) Kikutsuji, T.; Mori, Y.; Okazaki, K.-i.; Mori, T.; Kim, K.; Matubayasi, N. Explaining reaction coordinates of alanine dipeptide isomerization obtained from deep neural networks using Explainable Artificial Intelligence (XAI). *J. Chem. Phys.* **2022**, *156*, 154108.

(16) Ma, A.; Dinner, A. R. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B* **2005**, *109*, 6769–6779.

(17) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.* **2011**, *134*, 03B624.

(18) Boninsegna, L.; Gobbo, G.; Noé, F.; Clementi, C. Investigating molecular kinetics by variationally optimized diffusion maps. *J. Chem. Theor. Comput.* **2015**, *11*, 5947–5960.

(19) Coifman, R. R.; Kevrekidis, I. G.; Lafon, S.; Maggioni, M.; Nadler, B. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Model. Simul.* **2008**, *7*, 842–864.

(20) Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. Integrating diffusion maps with umbrella sampling: Application to alanine dipeptide. *J. Chem. Phys.* **2011**, *134*, 04B606.

(21) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. TRANSITIONPATHSAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.

(22) Khoo, Y.; Lu, J.; Ying, L. Solving for high-dimensional committor functions using artificial neural networks. *Research in the Mathematical Sciences* **2019**, *6*, 1–13.

(23) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7426–7431.

- (24) Coifman, R. R.; Lafon, S. Diffusion maps. *Appl. Comput. Harmon. Anal.* **2006**, *21*, 5–30.
- (25) ref 24, Theorem 2.
- (26) Berry, T.; Harlim, J. Variable bandwidth diffusion kernels. *Appl. Comput. Harmon. Anal.* **2016**, *40*, 68–96.
- (27) Preto, J.; Clementi, C. Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. *Phys. Chem. Chem. Phys.* **2014**, *16*, 19181–19191.
- (28) Little, A. V.; Jung, Y.-M.; Maggioni, M. Multiscale estimation of intrinsic dimensionality of data sets. *2009 AAAI Fall Symposium Series*, 2009;.
- (29) Le Bris, C.; Lelievre, T.; Luskin, M.; Perez, D. A mathematical formalization of the parallel replica dynamics. *Monte Carlo Methods Appl.* **2012**, *18*, 119–146.
- (30) Prinz, J.-H.; Held, M.; Smith, J. C.; Noé, F. Efficient computation, sensitivity, and error analysis of committor probabilities for complex dynamical processes. *Multiscale Model. Simul.* **2011**, *9*, 545–567.
- (31) Weinan, E.; Ren, W.; Vanden-Eijnden, E. Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes. *Chem. Phys. Lett.* **2005**, *413*, 242–247.
- (32) Epifanovsky, E.; Gilbert, A. T. B.; Feng, X.; Lee, J.; Mao, Y.; Mardirossian, N.; Pokhilko, P.; White, A. F.; Coons, M. P.; Dempwolff, A. L.; Gan, Z.; Hait, D.; Horn, P. R.; Jacobson, L. D.; Kaliman, I.; Kussmann, J.; Lange, A. W.; Lao, K. U.; Levine, D. S.; Liu, J.; McKenzie, S. C.; Morrison, A. F.; Nanda, K. D.; Plasser, F.; Rehn, D. R.; Vidal, M. L.; You, Z.-Q.; Zhu, Y.; Alam, B.; Albrecht, B. J.; Aldossary, A.; Alguire, E.; Andersen, J. H.; Athavale, V.; Barton, D.; Begam, K.; Behn, A.; Bellonzi, N.; Bernard, Y. A.; Berquist, E. J.; Burton, H. G. A.; Carreras, A.; Carter-Fenk, K.; Chakraborty, R.; Chien, A. D.; Closser, K. D.; Cofer-Shabica, V.; Dasgupta, S.; de Wergifosse, M.; Deng, J.; Diedenhofen, M.; Do, H.; Ehlert, S.; Fang, P.-T.; Fatehi, S.; Feng, Q.; Friedhoff, T.; Gayvert, J.; Ge, Q.; Gidofalvi, G.; Goldey, M.; Gomes, J.; González-Espinoza, C. E.; Gulania, S.; Gunina, A. O.; Hanson-Heine, M. W. D.; Harbach, P. H. P.; Hauser, A.; Herbst, M. F.; Hernández Vera, M.; Hodecker, M.; Holden, Z. C.; Houck, S.; Huang, X.; Hui, K.; Huynh, B. C.; Ivanov, M.; Jász, Á.; Ji, H.; Jiang, H.; Kaduk, B.; Kähler, S.; Khistyayev, K.; Kim, J.; Kis, G.; Klunzinger, P.; Koczor-Benda, Z.; Koh, J. H.; Kosenkov, D.; Koulidas, L.; Kowalczyk, T.; Krauter, C. M.; Kue, K.; Kunitsa, A.; Kus, T.; Ladjántski, I.; Landau, A.; Lawler, K. V.; Lefrançois, D.; Lehtola, A.; Li, R. R.; Li, Y.-P.; Liang, J.; Liebenthal, M.; Lin, H.-H.; Lin, Y.-S.; Liu, F.; Liu, K.-Y.; Loipersberger, M.; Luenser, A.; Manjanath, A.; Manohar, P.; Mansoor, E.; Manzer, S. F.; Mao, S.-P.; Marenich, A. V.; Markovich, T.; Mason, S.; Maurer, S. A.; McLaughlin, P. F.; Menger, M. F. S. J.; Mewes, J.-M.; Mewes, S. A.; Morgante, P.; Mullinax, J. W.; Oosterbaan, K. J.; Paran, G.; Paul, A. C.; Paul, S. K.; Pavošević, F.; Pei, Z.; Prager, S.; Proynov, E. I.; Rák, Á.; Ramos-Cordoba, E.; Rana, B.; Rask, A. E.; Rettig, A.; Richard, R. M.; Rob, F.; Rossomme, E.; Scheele, T.; Scheurer, M.; Schneider, M.; Sergueev, N.; Sharada, S. M.; Skomorowski, W.; Small, D. W.; Stein, C. J.; Su, Y.-C.; Sundstrom, E. J.; Tao, Z.; Thirman, J.; Tornai, G. J.; Tsuchimochi, T.; Tubman, N. M.; Veccham, S. P.; Vydrov, O.; Wenzel, J.; Witte, J.; Yamada, A.; Yao, K.; Yeganeh, S.; Yost, S. R.; Zech, A.; Zhang, I. Y.; Zhang, X.; Zhang, Y.; Zuev, D.; Aspuru-Guzik, A.; Bell, A. T.; Besley, N. A.; Bravaya, K. B.; Brooks, B. R.; Casanova, D.; Chai, J.-D.; Coriani, S.; Cramer, C. J.; Cserey, G.; DePrince, A. E.; DiStasio, R. A.; Dreuw, A.; Dunietz, B. D.; Furlani, T. R.; Goddard, W. A.; Hammes-Schiffer, S.; Head-Gordon, T.; Hehre, W. J.; Hsu, C.-P.; Jagau, T.-C.; Jung, Y.; Klamt, A.; Kong, J.; Lambrecht, D. S.; Liang, W.; Mayhall, N. J.; McCurdy, C. W.; Neaton, J. B.; Ochsenfeld, C.; Parkhill, J. A.; Peverati, R.; Rassolov, V. A.; Shao, Y.; Slipchenko, L. V.; Stauch, T.; Steele, R. P.; Subotnik, J. E.; Thom, A. J. W.; Tkatchenko, A.; Truhlar, D. G.; Van Voorhis, T.; Wesolowski, T. A.; Whaley, K. B.; Woodcock, H. L.; Zimmerman, P. M.; Faraji, S.; Gill, P. M. W.; Head-Gordon, M.; Herbert, J. M.; Krylov, A. I. Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package. *J. Chem. Phys.* **2021**, *155*, 084801.
- (33) Collet, P.; Martínez, S.; San Martín, J. *Quasi-stationary Distributions: Markov Chains, Diffusions and Dynamical Systems*; Springer Berlin: Heidelberg, 2013; Vol. 1.
- (34) Kearsley, S. K. On the orthogonal transformation used for structural comparisons. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1989**, *45*, 208–210.
- (35) Mardirossian, N.; Head-Gordon, M. ω B97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9904–9924.