Behavior of Linear and Nonlinear Dimensionality Reduction for Collective Variable Identification of Small Molecule Solution-Phase Reactions

Hung M. Le,[†] Sushant Kumar,[‡] Nathan May,[¶] Ernesto Martinez-Baez,[†] Ravishankar Sundararaman,[*,‡] Bala Krishnamoorthy,[*,¶] and Aurora E. Clark[*,†]

[†]*Department of Chemistry, Washington State University, Pullman, Washington 99164, United States*

[‡]*Materials Science and Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180, United States*

[¶]*Department of Mathematics, Washington State University, Vancouver, Washington 98686, United States*

E-mail: sundar@rpi.edu; kbala@wsu.edu; auclark@wsu.edu

Abstract

Identifying collective variables for chemical reactions is essential to reduce the $3N$ dimensional energy landscape into lower dimensional basins and barriers of interest. However in condensed phase processes, the non-meaningful motions of bulk solvent often overpower the ability of dimensionality reduction methods to identify correlated motions that underpin collective variables. Yet solvent can play important indirect or direct roles in reactivity and much can be lost through treatments that remove or dampen solvent motion. This has been amply demonstrated within principal component analysis, although less is known about the behavior of nonlinear dimensionality

reduction methods, e.g., UMAP, that have become more popular recently. The latter presents an interesting alternative to linear methods though often at the expense of interpretability. This work presents distance attenuated projection methods of atomic coordinates that facilitate the application of both PCA and UMAP to identify collective variables in solution, and further the specific identity of solvent molecules that participate in chemical reactions. The performance of both methods is examined in detail for two reactions where the explicit solvent plays very different roles within the collective variables. The first reaction consists of the dynamic exchange of a cation about a polyhydroxy anion that is facilitated by waters of solvation, while the second reaction consists of a nucleophilic attack of H_2O upon ethylene to initiate *cis/trans* isomerization. When applied to raw data, both PCA and UMAP representations are dominated by bulk solvent motions. On the other hand, when applied to data preprocessed by our attenuated projection methods, both PCA and UMAP identify the appropriate collective variables in solution.

Introduction

Identifying the collective variables (or reaction coordinates) that reduce the $3N$ dimensional energy landscape of a chemical system into the basins and barriers of interest is essential to understanding complex chemical transformations.[1,2] Dimensionality reduction techniques encompass a large area of theoretical and applied research, with significant emphases dedicated to understanding simultaneous processes in biochemical reactions/interactions[3] or the coupled nature of solvent dynamics to solute reactivity.[4,5] Importantly, reactions that occur in solvent often have contributions from the solvent degrees of freedom and/or associated collective motion. Thus, chemical intuition alone may be unsatisfactory for dimensionality reduction and identification of the "collective" variables under these conditions. Generally, "ideal" reaction coordinates should: (i) only rely upon the instantaneous set of all coordinates x, (ii) evolve monotonically as the reactant progresses toward the product, and (iii) be dynamically self-consistent.[6,7] Over the last decade several techniques have been adapted or developed to identify complex CVs, reduce energy landscape dimensionality, and enhance sampling. Amongst these, principal component analysis (PCA), as an eigenvalue decomposition of the correlated motions (covariance of configurations) of particles (atoms, molecules, or other coarse-grained units) has been widely employed to study the dynamics of biochemical systems.[8–11] The computational expedience of this method, combined with the orthogonality of coordinates in reduced hyperspace, is appealing for a number of applications. This includes identifying the most active chemical reaction sites in complex biochemical systems,[12–14] dimensionality reduction of protein folding,[8,10,11,15] data illustration and process tracking,[16,17] and most recently characterizing direct reaction pathways in small organic molecules in the gas phase.[18,19]

Part of the success of PCA relies upon high amounts of correlation within the participating chemical species. Although this may be an obvious trait of biochemical processes, or isolated molecules, solution phase chemical reactions can be much more nuanced. This is amply demonstrated when PCA is used for the purposes of enhanced sampling in solution-

phase reactions. For example, replica exchange molecular dynamics combined with PCA has been employed to study the dynamics of α-conotoxins and its two mutants in water and 1-ethyl-3-methyl-imidazolium acetate (50%), with the projected trajectories given by the first two PCs showing significant distinction in the dynamical pattern for the two solvents in use.[20] The dynamical entropy of solvating molecules usually distracts PCA in mapping the primary reaction of interest, and scaling the solvent mass by a small factor has been shown to be a good strategy to reduce solvent viscosity and enhance sampling.[21,22] These examples include transformations where large amounts of correlation are still present despite the solvent degrees of freedom. It is then useful to consider the more challenging case of small(er) molecule chemical reactions, in particular those where the solvent is known to participate either as an explicit reactant or where the direct solvent interactions influence reactant behavior. In these cases, if PCA is capable of identifying the CV, it may also be a useful tool to identify the explicit role of individual solvent molecules and collective solvent motions upon the reaction energetics and pathways.

Compared to linear approaches such as PCA, nonlinear dimensionality reduction techniques may be more effective at identifying the global structure in the input data set.[23] Although less studied, Das et al.[24] has examined low-dimensional free-energy landscapes of protein folding reactions produced by the nonlinear dimensionality reduction method called Isomap (and its variants),[23,25] demonstrating an ability to identify the transition-state ensemble. One issue, however, is that the reduced dimensions from nonlinear approaches may not be amenable to direct interpretation in terms of the input dimensions. More recently, Chen and Ferguson[26] have presented an approach based on the deep learning technique of autoencoders for collective variable discovery and accelerated free energy landscape exploration. The advantage of using autoencoders for nonlinear dimensionality reduction is that one typically obtains each CV as a differentiable function of the input atomic coordinates. Yet the framework of autoencoders requires the user to choose multiple parameters as well as appropriate error or loss function(s) to be optimized, and the computation may be quite

expensive.[27] The state-of-the-art nonlinear dimensionality reduction techniques are t-SNE[28] and UMAP.[29] While these tools do not compute the reduced dimensions as functions of the input dimensions, they have proven to be much more computationally efficient than autoencoders. Further, the user is not required to choose a variety of parameters (usually, just the number of nearest neighbors for each point needs to be specified). Highly efficient implementations of t-SNE[30] and UMAP[31] have been made available recently (late 2018), and the latter has seen widespread used in many application domains.[32–37] Within the context of these tools, it is interesting to consider their efficacy for: 1) identifying collective variables for small molecule reactions in the condensed phase (where solvent participates in the reaction coordinates), and subsequently 2) their ability to elucidate the importance of solvent degrees of freedom within the reduced energy landscape framework.

This work compares and contrasts linear and nonlinear dimensionality reduction methods to identify the collective reaction coordinates for two solution-phase reactions. In the first case, a highly symmetric, yet still chemically complex, reaction that consists of a "roaming reaction" of K^+ about hydrated aluminate $Al(OH)_4^-$, where the K^+ exchanges between different terminal -OH groups. The high symmetry of this roaming process that occurs in aqueous media provides an ideal scenario to use PCA to identify collective variables for the reaction, and to investigate rigorously the role of dynamically evolving and solvating water. In the second reaction, we study the reversible nucleophilic attack of a water molecule on ethylene in an aqueous solution, which prompts *cis* and *trans* isomerization. In this case, statistical sampling of the reversible reaction changes the solvent identities that participate in the reaction and presents a challenging case for dimensionality reduction methods. In combination with different preprocessing schemes to dampen extraneous solvent motion, we demonstrate the ability of PCA to both identify chemically reasonable solution-phase collective variables as well as the number of participating solvent molecules within the chemical reaction. Nonlinear dimensional reduction, specifically UMAP, is shown to be more sensitive to solvent motion than PCA. Under appropriate conditions UMAP can identify the key

structures occurring along the reaction coordinate within the reduced data.

Computational and Analysis Methods

Simulation of Cation Roaming Within a Contact Ion Pair. Density functional theory-based molecular dynamics (DFTMD) was employed within the CP2K software[38] to study the K^+ ... $Al(OH)_4^-$ ion-pair dynamics in bulk water, where the K^+ is observed to migrate systematically between different -OH groups. Ninety H_2O were included in a cubic periodic box with an edge length of 14.5 Å. The revised Perdew-Burke-Ernzerhof (revPBE)[39–41] density functionals were used with Goedecker-Teter-Hutter pseudo-potentials[42–44] for the auxiliary wave-functions, while the double-zeta valence-polarization (DZVP) basis set[45,46] was employed to construct the valence electrons for all atoms, with plane wave expansions around the Γ point. The simulation trajectory was performed at 300 K within the NVT canonical ensemble using the Langevin thermostat.[47] According to the potential of mean force for the Al-K distance in a previous study,[48] the contact ion pair has an equilibrium distance of 4.2 Å (Figure S1). Given competition between the K^+ roaming about $Al(OH)_4^-$ and exchange between the 1^{st} and 2^{nd} solvation shells, the Al-K distance was then constrained to a distance less than 4.2 Å to enable detailed study solely of the roaming reaction. The constraint was imposed using PLUMED,[49] with a predefined force constant of 150 kcal/Å^2 and the distance was chosen based upon umbrella sampling of the ion-pairing distance coordinate, as described in the Supplementary Information.

Over a 40 ps period, three K^+ transfer events among hydroxyls of $Al(OH)_4^-$ were observed (see Figure S2). To enhance sampling of the roaming pathway across all four -OH groups and provide reasonable statistics, a data cloning procedure was employed that is based upon rotation and reflection of the structural coordinates. First, translation and rotation of the $Al(OH)_4^-$ was removed. Then, subsampling was used to generate a set of 8,000 configurations wherein K^+ departed from O(1), made a visit to O(2), then back to O(1), and finally arrived

at O(3). The data was then replicated 24 times with multiple rotations and reflections to cover all transition possibilities, which included 24 departures and 24 arrivals at each of the four -OH present at the vertices of the tetrahedron of $Al(OH)_4^-$. More detailed information is presented in the Supplementary Information (Figures S3, S4 and Table S1). The enhanced sampling guarantees: (1) that the number of occasions K^+ visiting each O vertex is identical, (2) that the K^+ traveling along each O-O edge of the $Al(OH)_4^-$ tetrahedron is identical, and lastly, (3) that the motion of K^+ is continuous. As a result of the data replication procedure, after roaming all the way around the tetrahedron, the final K^+ position is the same as at the starting point of the trajectory. A total of 192,000 structural configurations were generated for subsequent PCA analysis.

Nucleophilic Attack of H_2O on Ethylene. Reactive classical molecular dynamics was used to study the nucleophilic attack of H_2O on ethylene (C_2H_4) in pure water. To identify the resulting *cis* and *trans* isomerization reaction, two H-atoms were replaced with deuterium to yield $C_2H_2D_2$ (see Figure S5). Biased simulations were performed along a distance coordinate of a preselected H_2O within a periodic cubic simulation box containing 52 water molecules. Although H_2O is a weak nucleophile, at a biased distance of 1.5 Å it will add to $C_2H_2D_2$ to create $HOCH_2$-CD_2 with the release of H^+. The rehybridization of the hydroxylated C-atom from sp^2 to sp^3 enables dihedral rotation about the C-C bond. The reactive nature of the $HOCH_2$-CD_2 will lead to H^+ addition to re-form the H_2O reactant and re-formation of ethylene.

The potential between the atoms is modeled using the ReaxFF force field developed by Zhang and van Duin[50] which has been designed to describe hydrocarbon-water interactions. One of the C-atoms of $C_2H_2D_2$ is held fixed to its initial position while allowing the other atoms of the molecule to move freely. An H_2O is constrained throughout the course of the simulation and is moved closer and farther away from this fixed C-atom in accordance with a time-dependent sawtooth function. The distance between the fixed C-atom and constrained

O-atom varies between 1.5 Å and 2.5 Å with a time period of the oscillation of 5 ps that enables 20 reaction cycles (including the attack and dihedral rotation) for 200,000 time steps (using a time step of 0.5 fs). These MD simulations are performed in an NVT ensemble using the Nosé-Hoover thermostat as implemented in LAMMPS.[51] For each reaction cycle, the atomic positions are saved every 20 time steps, with 20 reaction cycles to create a total dataset of 10,000 snapshots.

Preprocessing of Simulation Trajectories. Within these solution-phase reactions much of the translational motion of H_2O will not be relevant to the reaction, while the motion of a select H_2O will participate. A method is thus needed to attenuate non-meaningful solvent translational motion. The Cartesian coordinates are first aligned with a reference; in the case of the nucleophilic attack this is composed of the two C-atoms and the H- and D-atoms lying on the same side of the fixed C-atom; in the case of the cation roaming reaction the Al-center was chosen to be the origin. We then introduce two techniques to map the effective water COM translation within a characteristic radius R_0. First, we develop an "inverse stereographic projection" that involves the transformation of the original Cartesian coordinates q$(N \times 3)$ into $q_{new}(N \times 4)$. The coordinates of each particle $i \in \{1, \ldots, N\}$ is transformed as

$$\vec{q}_{new,i} = \left(\frac{2\vec{q}_i R_0}{|\vec{q}_i|^2 + R_0^2}, \frac{|\vec{q}_i|^2 - R_0^2}{|\vec{q}_i|^2 + R_0^2} \right). \tag{1}$$

Here, the fourth coordinate is introduced to map data points within R_0 to the Southern hemisphere of a spherical distribution. An illustration of inverse stereographic projection is given in Figure 1, in which an arbitrary 2D configuration is mapped.

A second method, "cut-projection", was also considered, where the coordinates of each particle are transformed by

$$\vec{q}_{new,i} = \begin{cases} \left(\frac{\vec{q}_i}{|\vec{q}_i|} \sin \frac{\pi |\vec{q}_i|}{2R_0}, -\cos \frac{\pi |\vec{q}_i|}{2R_0} \right), & |\vec{q}_i| \leq 2R_0 \\ \left(\vec{0}, 1 \right), & |\vec{q}_i| > 2R_0 \end{cases}. \tag{2}$$

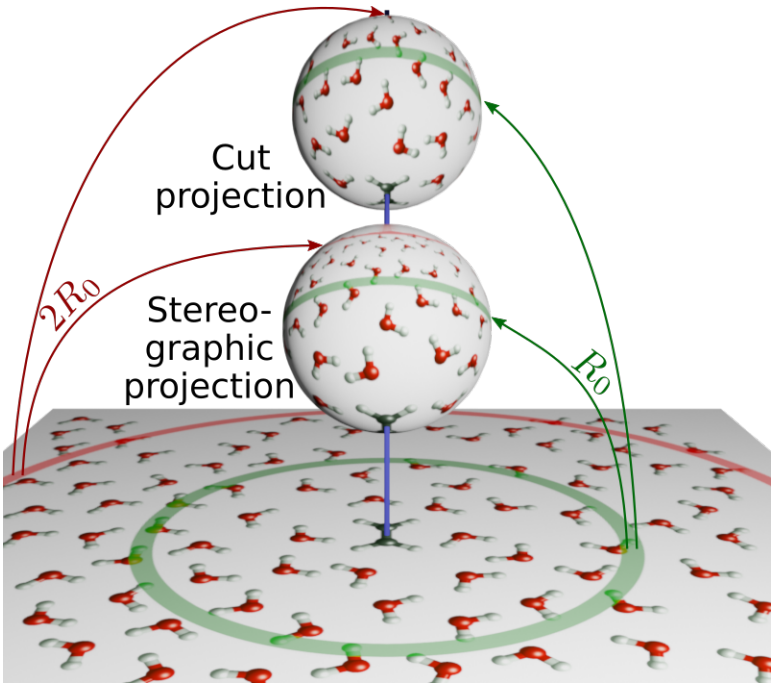

Figure 1: Inverse stereographic projection and cut projection of a 2D configuration of ethylene surrounded by water particles.

Intuitively, this map "wraps" the ball $|\vec{q}_i| \leq 2R_0$ around the unit sphere in one-dimension higher, where the boundary and exterior of the ball are all concentrated at the north pole of the sphere, as shown in Figure 1. This map completely damps out the influence of solvent atoms outside $2R_0$ on the subsequent dimensionality reduction, in contrast to the stereographic projection which diminishes the influence of solvent molecules outside R_0 asymptotically as $\sim 1/|\vec{q}_i|$, but never to exactly zero at any distance. As an example, in Figure 2 the root mean square displacement of all particles as a function of mean distance from the origin of the aligned atomic coordinates are shown and compared with those from the inverse stereographic and cut-projected coordinates for both the ion roaming and nucleophilic attack reactions.

Dimensionality Reduction Methods. Given input points in $3N$ dimensions, dimensionality reduction aims to find a representation of the points in d-dimensions for small values of d that "preserves most of the structure" of the input system. These methods are

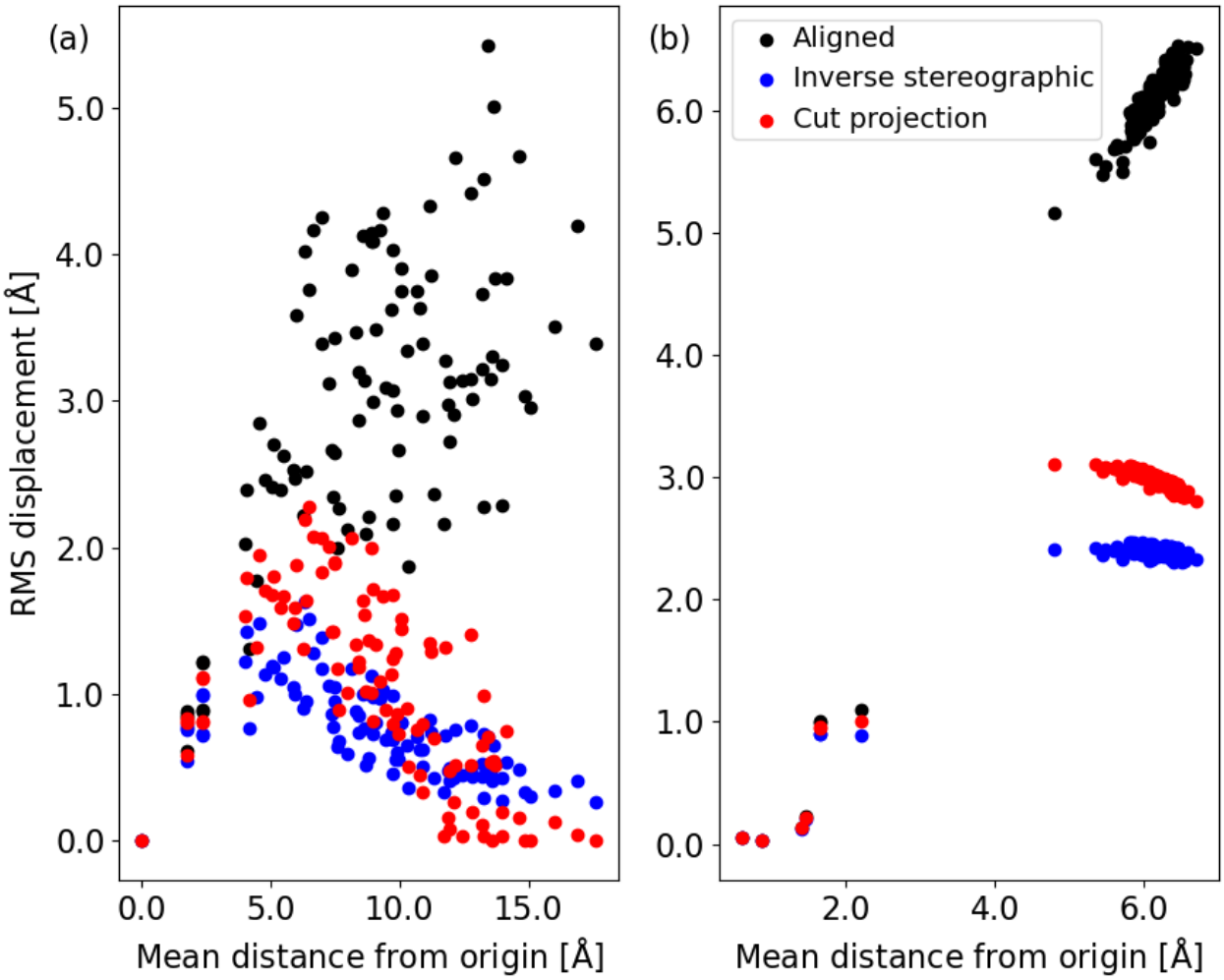

Figure 2: Root mean squared displacement as a function of mean distance from the origin of the aligned atomic coordinates along with inverse stereographic and cut projected coordinates for (a) ion roaming, and (b) nucleophilic attack, where $R_0 = 5$ Å.

classified into *linear* and *nonlinear* approaches.[23] The nonlinear methods are also referred to as *manifold learning* techniques, where the underlying assumption is that the input data in $3N$ dimensional space is sampled from a nonlinear manifold of much smaller dimension d. The goal of the method is to "learn" this manifold and produce a truthful representation of the same in d dimensions.

In this work, the input data sets contains time-dependent simulations that sample a $3N$-dimensional energy landscape. In the case of PCA, the i^{th} PC is the unit vector along the *line* that best fits (in a sense of least squared error) the data while being orthogonal to the first $(i-1)$ PCs. Hence the first PC ($i = 1$) gives the direction of the best linear approximation of the data. For a data set with m points in $3N$ space (representing m snapshots from the MD simulation with N atoms), we start with the input matrix $X \in \mathbb{R}^{m\times 3N}$. The $3N$ PCs are computed as the eigenvectors of the covariance matrix

$$C = \frac{1}{m-1} X^T X \tag{3}$$

Let $W \in \mathbb{R}^{3N\times 3N}$ represent all the PCs arranged in the decreasing order of variances (from first to $3N$ columns). The reduced dimensional representation $Y \in \mathbb{R}^{m\times d}$ of the data is then obtained as

$$Y = XW_{:,1:d} \tag{4}$$

by choosing only the *first* d eigenvectors.

The transformation from Cartesian space to PC coordinates and *vice versa* is adopted using a linear transformation, in which all PC vectors are required to be orthogonal. In terms of n-dimensional space, a linear projection of Cartesian data into an imaginary hyperspace is executed, in which movements of atoms are coupled (or mapped) into the most effective number of condensed coordinates to describe a particular chemical behavior (i.e., reactions, structural rearrangements, or internal molecular vibrations). The PCA projection is similar to the transformation procedure used in normal mode analysis to study molecu-

lar vibrations.[52] While normal mode analysis takes into account the Hessian force-constant matrix at a specific geometry, PCA employs a Hessian matrix that accounts for structural transformation (variance) throughout the entire MD trajectory data.

Scikit-learn[53] was employed to project the MD geometries to the PC hyperspace and perform backward conversion of the selected trajectory of multiple PCs into an attenuated, non-mass weighted, Cartesian trajectory for visualization purposes. Several variants were tested, including using the H_2O COM versus all of the H- and O-atom coordinates. The results from the scikit-learn are consistent with those obtained from PathReducer.Hare et al. [19] Let m to be the number of chosen MD configurations for PCA, while N is the number of particles. The overall set of data has the size of $(m \times 3N)$, with coordinates q_i. Each element of the $3N \times 3N$ covariance matrix C is computed as[54]

$$c_{ij} = \langle (\vec{q_i} - \langle \vec{q_i} \rangle)(q_j - \langle q_j \rangle) \rangle. \tag{5}$$

The unitary transformation matrix U_C is obtained by executing eigenanalysis on C:

$$\Omega = U_C C U_C^T, \tag{6}$$

where $\Omega(\omega_i)$ is the diagonal matrix of eigenvalues ω_i and U_C are the eigenvectors for the forward and backward transformations between Cartesian and PC coordinates, respectively.

Uniform Manifold Approximation and Projection, or UMAP, is a state of the art nonlinear dimension reduction algorithm.[29,31] It is a non-parametric graph-based dimension reduction algorithm that uses techniques from applied Riemannian geometry and algebraic topology to find low-dimensional embeddings of high dimensional data. The UMAP algorithm comprises two main steps: (1) compute a graphical representation (as a "fuzzy simplicial complex") that captures the local relationships in the input point cloud of data; and (2) determine a low-dimensional embedding of the graphical representation using optimization such that "structure" of the graph is preserved.

One parameter that the user must choose when applying UMAP is an appropriate number of nearest neighbors (NN). The UMAP algorithm relies heavily on a nearest-neighbor graph, where similarity between data points (the Cartesian coordinates within each snapshot of the trajectory) is measured via proximity in local neighborhoods. For every data point, it's neighborhood consists of the closest NN distinct points. Generally speaking, the smaller the NN value, the more local the approximation, and likewise, the larger the NN, the more global the approximation. The trajectories within this study continuously sample the reaction coordinate, however the sampling is not evenly distributed (barrier regions of the energy landscape are sampled less than minima). Although the UMAP output should converge as NN is increased this also increases computational cost [29,31] and establishing the criteria for convergence is an important consideration. One strategy is to require the UMAP nearest-neighbor graph to sufficiently sample the transition state region as presumably this part of the trajectory is the most relevant toward identifying the reaction coordinates. Thus nearest neighbors of each representative transition snapshot within the trajectory must sufficiently sample the entire transition. Representative transition snapshots are chosen to be the midpoint of a transition, and we say that its neighbors sufficiently sample the transition if they're sampled uniformly. We test for uniformity using Pearson's Chi-Squared test for goodness of fit, requiring a p-value of .95 or higher. As illustrated in Figure 3, the simulation trajectory of the ion-roaming reaction (without data cloning) consists of three hops of the K^+ between -OH groups. Choosing the mid-point in the transition as the "transition state", the UMAP graph distance within the NN graph was compared against the Chi-squared P-value. The graph metric used is the mean-square error of edge-weight difference, taken over all shared edges. Therein, the nearest-neighbor graphs had sufficiently sampled transitions with NN $>$ 1600 both in the absence and presence of water (using both inverse-stereographic and cut-projections). For the nucleophilic attack trajectory, 1000 NN was sufficient, and for the cut projected data, 1200 NN was sufficient (Figure S6). This criteria is quite conservative, however, and smaller NN values may be sufficient in practice. For both of these data,

UMAP projections appear to be stable by 400 NN, by visual inspection.

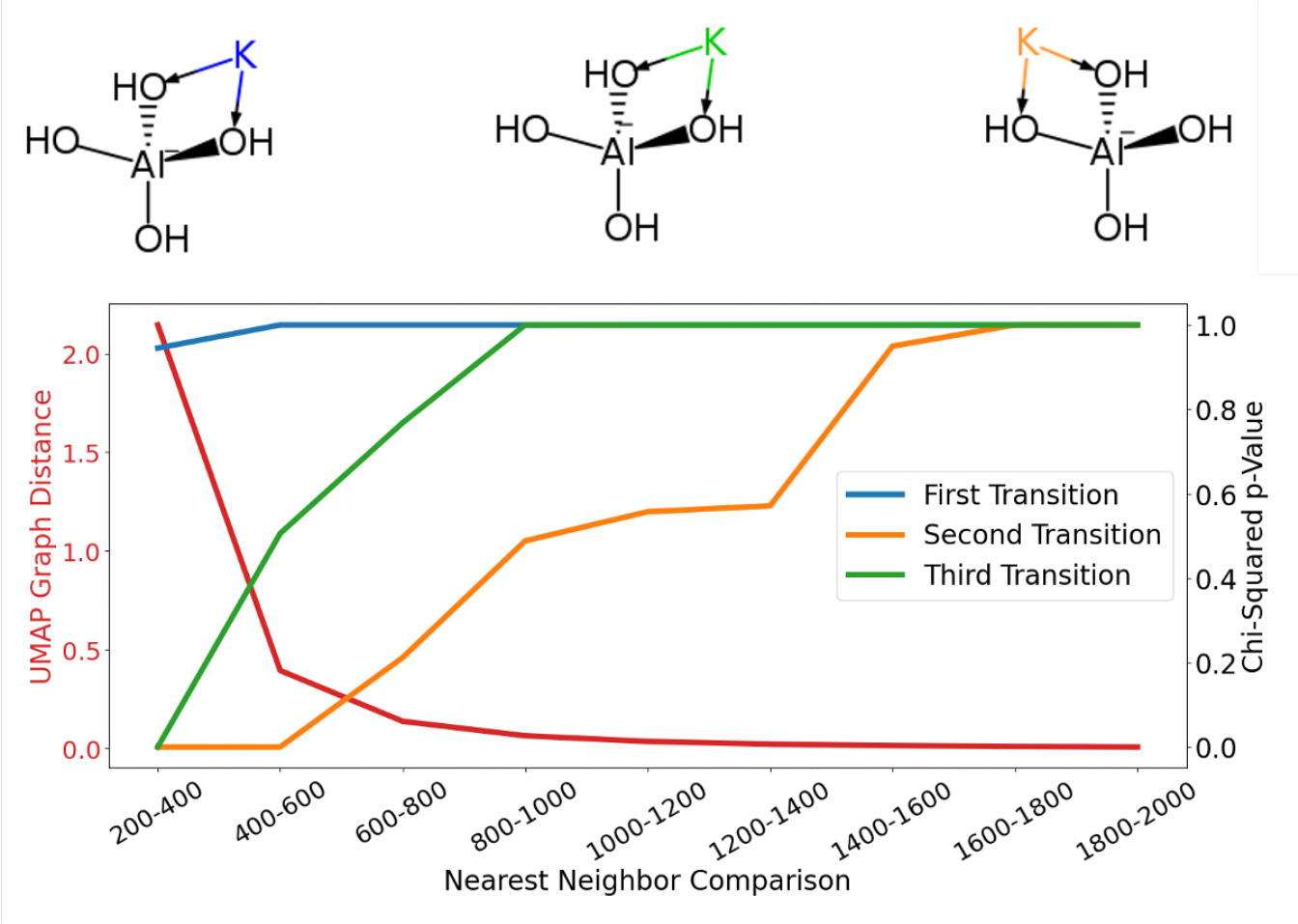

Figure 3: Three K^+ hopping transitions sampled within the AIMD trajectory (Top). Convergence of the UMAP graph distances between two different nearest neighbor lists (200 vs. 400, etc.) and the associated Chi-squared P-value (Bottom).

Results and Discussion

Identification of CVs in the Absence of Water

Prior work has demonstrated significant success of PCA to identify small molecule organic phase reaction coordinates in the absence of water.[19,53] Much less effort has examined the similar fidelity of nonlinear DR techniques. We thus begin by first demonstrating that both PCA and UMAP are capable of identifying the collective variables for the reactions,

sampled from the MD data in solution but with all solvent removed as a post-processing of the trajectories. In the case of the nucleophilic attack, the nucleophilic water was retained in the analyzed data as it is a reactant.

PCA of the Cation Roaming Reaction. PCA was performed for the 40 ps AIMD trajectory that exhibited three K^+ hopping events across three vertices of $Al(OH)_4^-$, wherein the H_2O molecules within the simulation are removed. Two principal components are observed with significant eigenvalues of 2.04 and 1.44 (see Table S2). The explained-variance ratio for a particular PC is the ratio of its eigenvalue over the sum of all eigenvalues obtained from PCA.[19] Accordingly, PC1 and PC2 explain 46.8% and 32.9% of the variance, respectively and in total capture nearly 80% of the variance.

We then evaluate the terms λ_{cation} and λ_{anion} for the first and second eigenvectors that represent the contributions of the K^+ and $Al(OH)_4^-$ over the motion of the whole system. The λ for cation or anion is defined as:

$$\lambda(\langle PC_i \rangle) = \sum \omega_{i,j}^2, \tag{7}$$

where $\omega_{i,j}$ is the element of associated with the K^+ or the $Al(OH)_4^-$ within the i^{th} eigenvector. The contributions of K^+ motion in the PC1 and PC2 eigenvectors (see Figure 4) are 41% and 47%, respectively, and analysis of the vectors further indicates that PC1 and PC2 captures the two hopping events present in the DFTMD data set.

Executing PCA on the replicated data where K^+ migrates across all four -OH groups has the first three eigenvectors given by the U_C covariance with eigenvalues of 8.32, 7.85, and 3.06 and illustrated in Figure 4a (the full set of eigenvalues is presented in Table S2). In total, those major PCs jointly capture 90% of the variance and fully describe the motion of K^+ about the four -OH vertices of aluminate. The 3D plot of (PC1, PC2, PC3) for the fully-sample data can be found in Figure 4(b), in which a spherical shape formed by three reactive PCs can be observed. Note that the remaining PC's reflect the dynamic internal

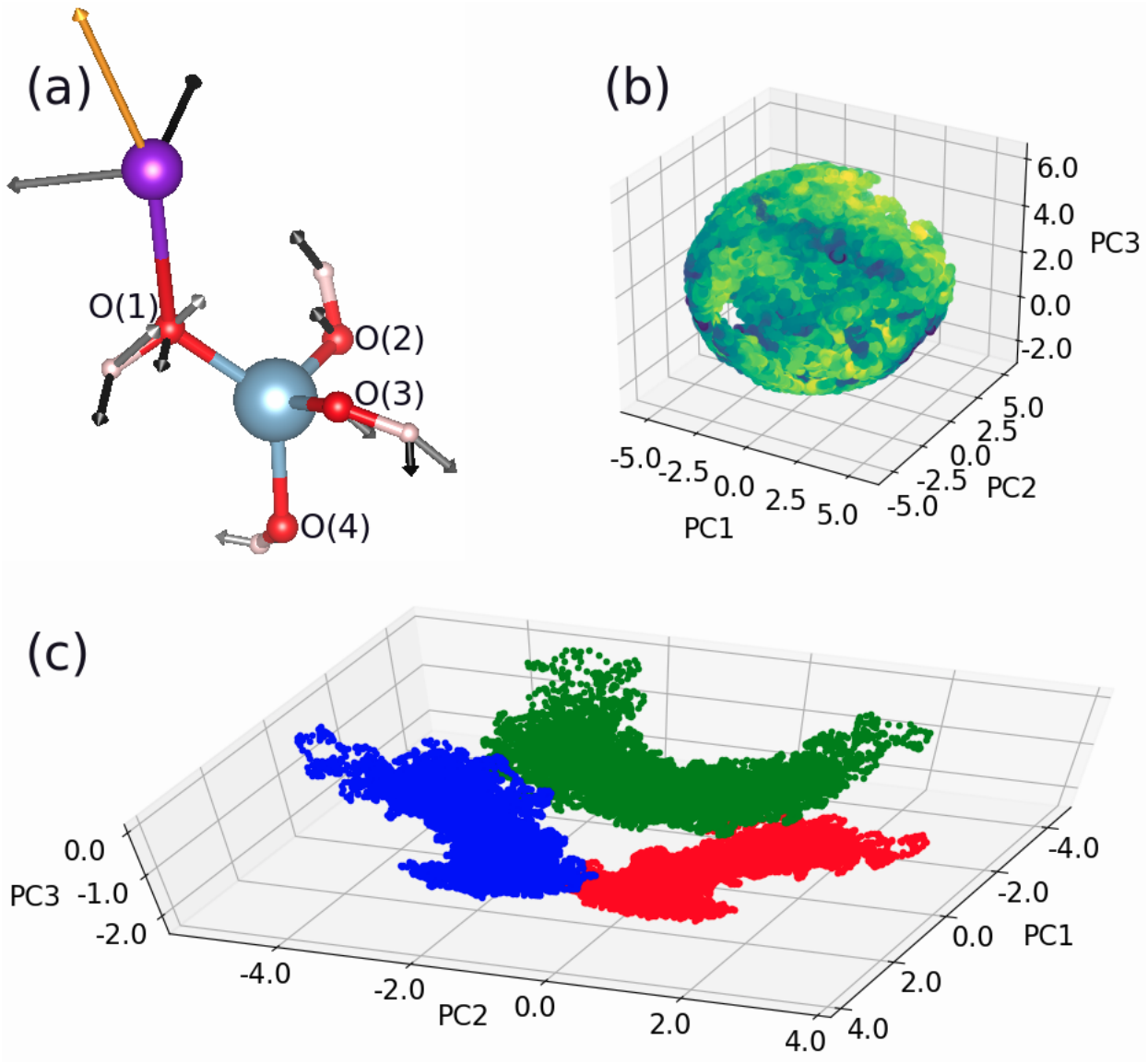

Figure 4: (a) PC vectors acting on each atom of the K^+-$Al(OH)_4^-$ cluster, (b) the shape of PC distribution in (PC1, PC2, PC3) space with dark points indicating transition state, (c) the shape of data distribution in (PC1, PC2, PC3) space when considering one fourth of the fully-sampled dataset, which describes the roaming of K^+ about the O(1) (red), O(2) (green), and O(3) (blue) vertices. The respective green, red, and blue colors indicate the identity of K-O(1), K-O(2), or K-O(3), respectively.

vibrations of OH^- groups and have small eigenvalues. To understand how the K^+ migrates on one of the four O-O-O facets, we then examine the distribution of (PC1, PC2, PC3) on a chosen set of data, which is constituted as one fourth of the fully-sampled dataset and depicts K^+ roaming on one O-O-O facet. In this subset, the configurations are sampled in such a way that K^+ travels back and forth on three O vertices equally. As a result, a triangular shape of PC distribution can be seen in Figure 4 The color-attributed region represents the roaming of K^+ around one specific O-vertex. Interestingly, those three regions share common borders, which allow the transition state to occur at the border within a continuous reaction process as a result of the linear projection method (PCA). The topological structure would be considered a mpermutahedron of order four. As might be anticipated, comparison of the PC's from the subsampled vs. full roaming trajectory indicates a higher contribution of the internal -OH vibrations and rotations in aluminate within the subsampled data.

PCA of the Nucleophilic Attack of H_2O on Ethylene. The expected topology of the nucleophilic attack in the principal component space as shown in Figure 5(b) and consists of the dihedral rotation of the ethylene molecule, which manifests itself as a circle. As the nucleophile moves away from the ethylene molecule, thereby locking it in either the *cis* or *trans* configuration, two distinct regions, corresponding to 0° and 180° dihedral angle, branch out along the reaction coordinate. Figure 3(c) shows the PCA of the trajectories consisting only of the ethylene molecule and the attacking O-atom. The configurations in the principal component space of the first three PCs are illustrated with the colors denoting the dihedral angle in the respective configuration. The topology obtained from the MD simulations in Figure 5b is very similar to the one based on intuition shown in Figure 5a. Within PC space, the configurations are symmetrically distributed into two halves, where each half corresponds to either the *cis* or *trans* orientation. Figure 5(b) also presents several configurations as visualized in real space for representative data points chosen from the PC space. As expected, the points in the circular region are the configurations when the

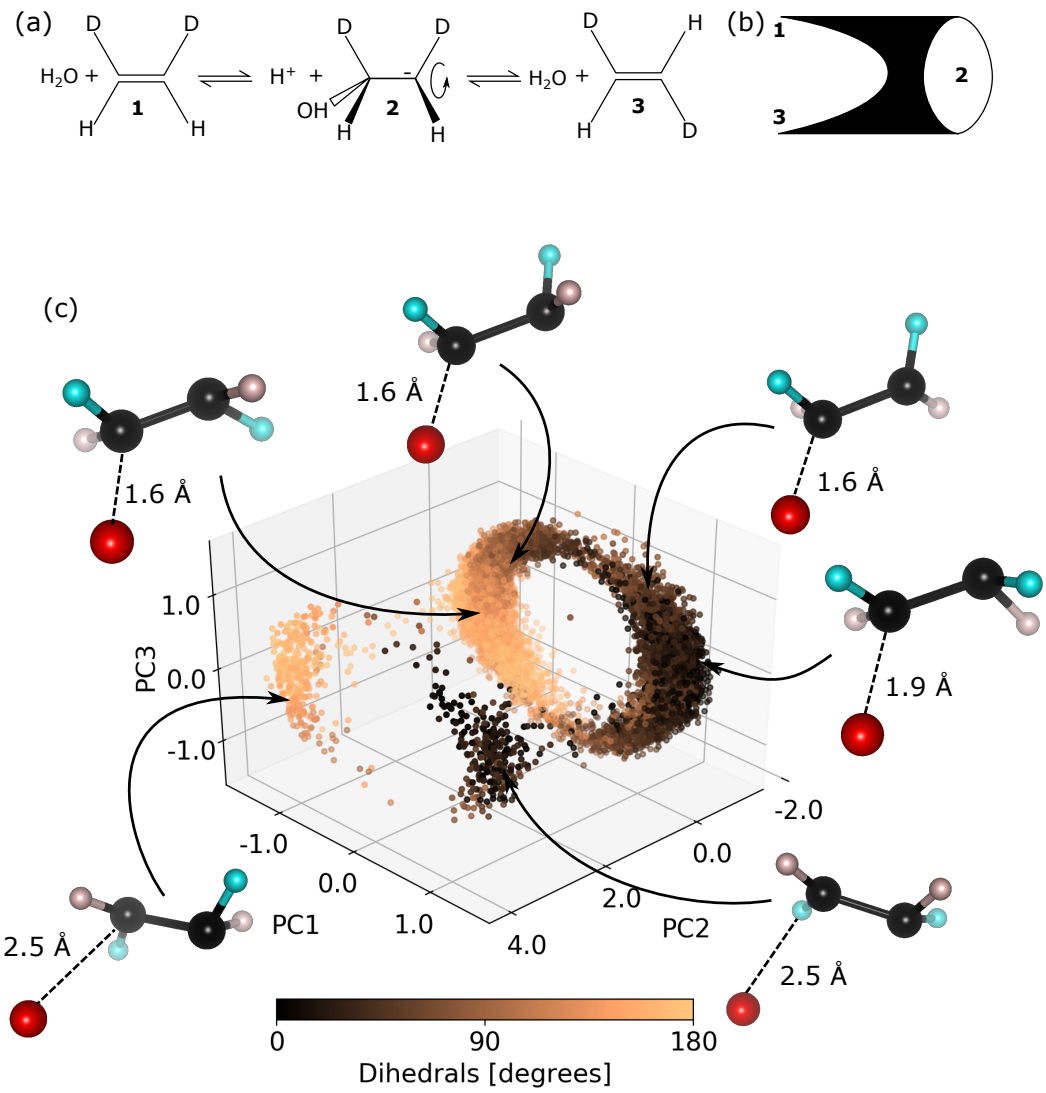

Figure 5: Principal component analysis (PCA) of the nucleophilic reaction in the absence of water. The colors denote the dihedral angle computed at each time step. The two parallel circular regions correspond to the *cis* and *trans* configurations of the ethylene molecule.

attacking O-atom is close enough such that the ethylene molecule becomes sp^3 hybridized and is free to rotate. It should be noted that the molecule, however, prefers to have the dihedral angle closer to 0° or 180°. With the O-atom moving away, the molecule locks itself into *cis* or *trans* orientation as observed at the two ends of the plot along PC2.

The first three eigenvectors have eigenvalues 1.20, 0.87 and 0.38 respectively (remaining eigenvalues in Table S3). Among these, PC1 captures 36.6%, PC2 captures 26.6% and PC3 explains 11.6% of the variance observed in the data. Together, these PCs capture 74.8% of the variance. Physically, PC1 separates the *cis* and *trans* regions in the PC space and hence, has the highest variance. PC1 and PC3 are essentially the projections of the dihedral angle and together form the circle seen in Figure 5(b). PC2, which has the second largest variance, is the reaction coordinate and represents the back and forth motion of the attacking O-atom.

Nonlinear DR of Cation Roaming and Nucleophilic Attack Reactions. In contrast to PCA, the reduced dimensions of the representation produced by nonlinear DR techniques such as UMAP cannot usually be expressed as closed form (even if nonlinear) functions of the input dimensions. A reverse map taking the reduced dimensions back to input space is also not readily available. While the nonlinear and global nature of UMAP makes it a more powerful DR technique, identification of CVs requires more detailed analysis of the UMAP output coupled with knowledge of the underlying chemical reaction. The goal is then to identify key structures along the appropriate collective variable of the UMAP space by examining the transitions and their associated snapshots from the MD trajectories within the reduced data.

For the roaming data with three K^+ hops, a NN value of 1600 is chosen when water is removed. After projecting into the plane the UMAP representation is shown to distinguish the data into 4 distinct components or "arms". Plotting the three transitions in Figure 6a, we see that each "jump" from one component to another, suggesting that these four components may correspond to the location of K^+ with respect to the four -OH groups. Within the

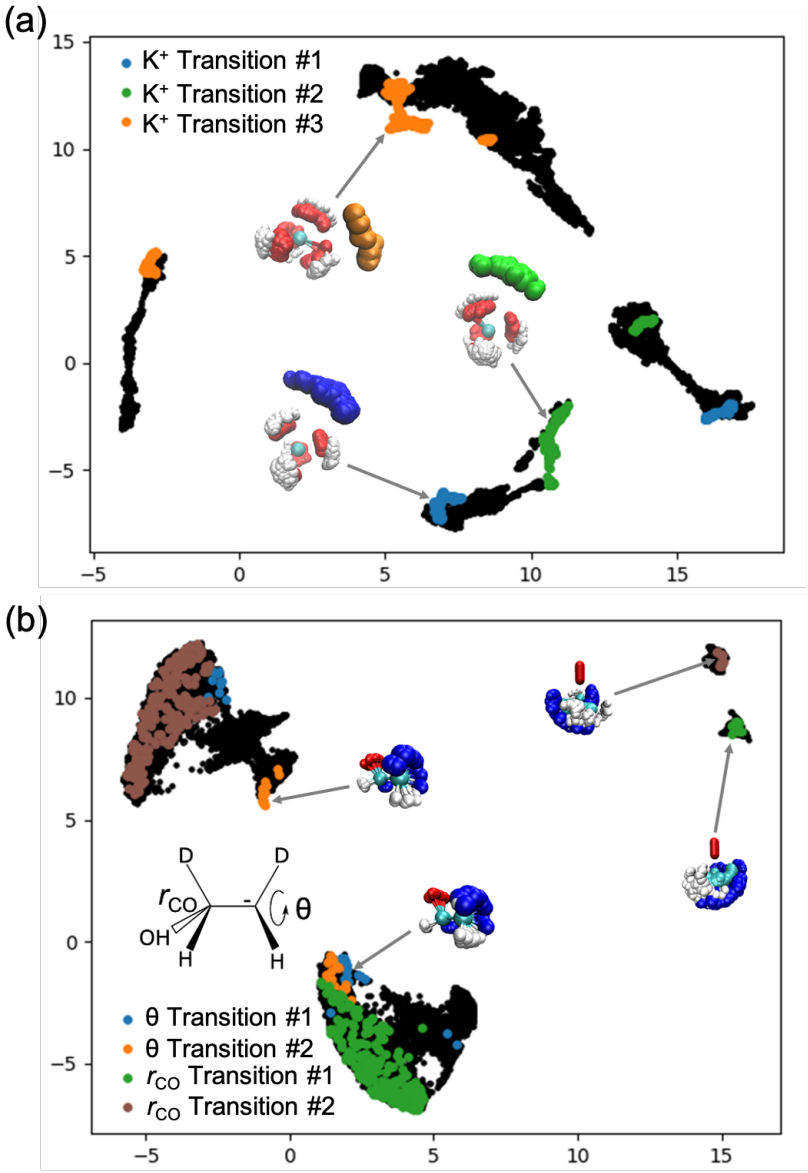

Figure 6: (a) UMAP representation of the cation roaming reaction with 1600 NN. (b) UMAP representation for the nucleophilic attack reaction with 400 NN.

nucleophilic attack data, 400 NN were chosen that sampled two different incidents of water attack where the C-OH_2 (r_{CO}) is decreased and where the dihedral angle θ indicates the isomerization reaction. The UMAP projection when no water is present also identifies 4 distinct components. Two components correspond to each dihedral angle configuration (*cis* and *trans*). For each dihedral configuration, we have two components that correspond the C...OH_2 distance (Figure 6b). Observe that for each isomerization process involving θ the component associated with r_{CO} is linked or connected within the reduced representation.

Collective Variables Identified in the Presence of Water

PCA of the Cation Roaming Reaction. When a reaction occurs in solution, the fluctuations of solvent can be significantly faster than the reaction itself. In the systems considered here, it is possible for different individual H_2O to contribute along different portions of the reaction coordinate. Further, the significant motion of bulk water in the simulation can dominate the PCA. Indeed, PCA performed on the raw simulation trajectory (with no preprocessing by inverse stereographic or cut projection) yields components that have little participation of cation or anion (λ_{cation} or λ_{anion} in Eqn. 7) as shown in Table 1 for the cation roaming and nucleophilic attack reactions. The principal components were then compared for the two preprocessing schemes using different R_0 (see Figure S7) that allow different numbers of H_2O to contribute to the covariance of molecular motion.

Starting with the cation roaming reaction, two predominant PCA eigenvalues and associated eigenvectors are observed until c.a. 24 H_2O are incorporated at a R_0 of 7 Å. A complete list of all eigenvalues are presented in Table S4. When employing the inverse stereographic projection a linearly increasing trend of variance is captured by PC1 as R_0 increases (Figure S8). As presented in Table 1, λ_{anion} is a dominant contributor to PC1 and PC2. This is unsurprising given that there are 9 atoms in $Al(OH)_4^-$ as compared to K^+. For PC1, the cation typically contributes 2 - 10 % of what the anion contributes, where the contribution of K^+ increases with increasing R_0. As the R_0 is increased, the cation contribution to PC2

Table 1: Eigenvalues ω_i and contribution factors λ for the cation roaming and nucleophilic attack reactions. The number of dynamical waters (n_{H_2O}) for datasets before and after inverse stereographic projection and cut-projection with various cutoff radii R_0. For the nucleophilic attack reaction, solute is defined to be the ethylene molecule + the attacking water.

Cation Roaming Reaction							
R_0	ω_1	ω_2	$\lambda_{cation}\langle PC1\rangle$	$\lambda_{anion}\langle PC1\rangle$	$\lambda_{cation}\langle PC2\rangle$	$\lambda_{anion}\langle PC2\rangle$	n_{H_2O}
None	569	173	0.000	0.004	0.004	0.006	90
Stereographic projection							
3.0 Å	3.42	1.70	0.004	0.147	0.016	0.165	0.00
4.0 Å	4.41	2.09	0.004	0.083	0.017	0.086	3.17
5.0 Å	5.14	2.33	0.003	0.053	0.015	0.052	9.97
6.0 Å	5.60	2.44	0.002	0.036	0.013	0.036	16.28
7.0 Å	5.85	2.45	0.002	0.027	0.012	0.027	23.86
Cut projection							
3.0 Å	1.66	1.11	0.004	0.267	0.015	0.324	0.00
4.0 Å	3.73	2.23	0.004	0.076	0.017	0.058	3.17
5.0 Å	5.99	3.28	0.003	0.032	0.011	0.023	9.97
6.0 Å	8.15	3.71	0.001	0.017	0.009	0.014	16.28
7.0 Å	9.46	3.63	0.001	0.011	0.007	0.012	23.86
Nucleophilic Attack Reaction							
R_0	ω_1	ω_2	ω_3	$\lambda_{solute}\langle PC1\rangle$	$\lambda_{solute}\langle PC2\rangle$	$\lambda_{solute}\langle PC3\rangle$	n_{H_2O}
None	1110.69	1055.86	1011.89	0.003	0.002	0.001	52
Stereographic projection							
1.5 Å	106.87	94.55	92.34	0.264	0.026	0.030	0.00
2.0 Å	125.43	117.44	113.75	0.152	0.013	0.026	0.05
3.0 Å	151.78	147.36	142.95	0.048	0.005	0.023	2.75
4.0 Å	166.70	163.02	157.96	0.019	0.003	0.018	7.59
5.0 Å	172.91	169.18	163.82	0.010	0.003	0.013	15.15
6.0 Å	173.23	169.61	164.00	0.007	0.003	0.010	26.72
7.0 Å	170.04	166.71	160.63	0.005	0.003	0.008	39.42
Cut projection							
1.5 Å	71.36	43.60	38.93	0.803	0.622	0.902	0.00
2.0 Å	87.44	66.40	55.23	0.382	0.066	0.348	0.05
3.0 Å	122.75	112.41	100.48	0.076	0.006	0.000	2.75
4.0 Å	155.69	150.12	140.68	0.023	0.002	0.004	7.59
5.0 Å	171.81	168.75	159.51	0.011	0.000	0.005	15.15
6.0 Å	174.28	170.09	162.78	0.005	0.002	0.005	26.72
7.0 Å	168.77	164.97	158.37	0.004	0.002	0.005	39.42

grows dramatically λ_{cation} is c.a. 45% that of the anion at R_0 of 7 Å. Plotting PC1 vs. PC2 and overlaying the atomic configurations clearly reveals K^+ hopping between different -OH groups (Figure 7). The eigenvalues of the PCs using the cut-projection are larger than those from inverse stereographic projection. Two major PCs from the cut-projection with eigenvalues > 1 are observed only with $R_0 \geq 6$ Å. Both PC1 and PC2 capture less variance as R_0 increases (see Figure S8) and at $R_0 = 6$ Å they capture 32.3% and 21.5%, respectively. The sum of captured variance for PC1 and PC2 in this case is 53.8%, significantly less than the 69.3% observed with the inverse stereographic projection at the same R_0. Interestingly, K^+ never contributes more than 3% of what the anion contributes in PC1, while in PC2 at R_0 values of 6 - 7 Å it contributes a more reasonable 10 % of what the anion contributes.

We then use the inverse stereographic-projected data at $R_0 = 5$ Å to identify the number and identity of H_2O that participate in the roaming reaction. The contribution to the roaming process (ρ) was determined using the eigenvector matrix and eigenvalues based on Eqn. 7:

$$\rho(\langle PC_i \rangle) = \nu_i \sum \omega_{i,j}^2, \tag{8}$$

In the above equation, ν_i is the eigenvalue corresponding to the PC (here we only examine PC1 and PC2), while $\omega_{i,j}$ are the four elements of eigenvector i^{th} corresponding to each H_2O COM. The plots of water contribution in PC1 and PC2 are shown in Figure S16, where two groups of waters are clearly identified. Six H_2O contribute significantly to PC1 ($\rho > 0.1$). Among those, two H_2O solvate K^+ during the course of the roaming reaction, while five H_2O have hydrogen bonds with the hydroxyl groups. A single H_2O in particular both solvates the K^+ and has a HB with the -OH group.

PCA of the Nucleophilic Attack. We next look at the PCA of the nucleophilic attack trajectories as a function of varying water contribution to the variance (Table 1). As we increase the value of R_0, the number of dynamical water molecules increases, correspondingly decreasing the contribution factor of the solute atoms (ethylene + attacking water) to each

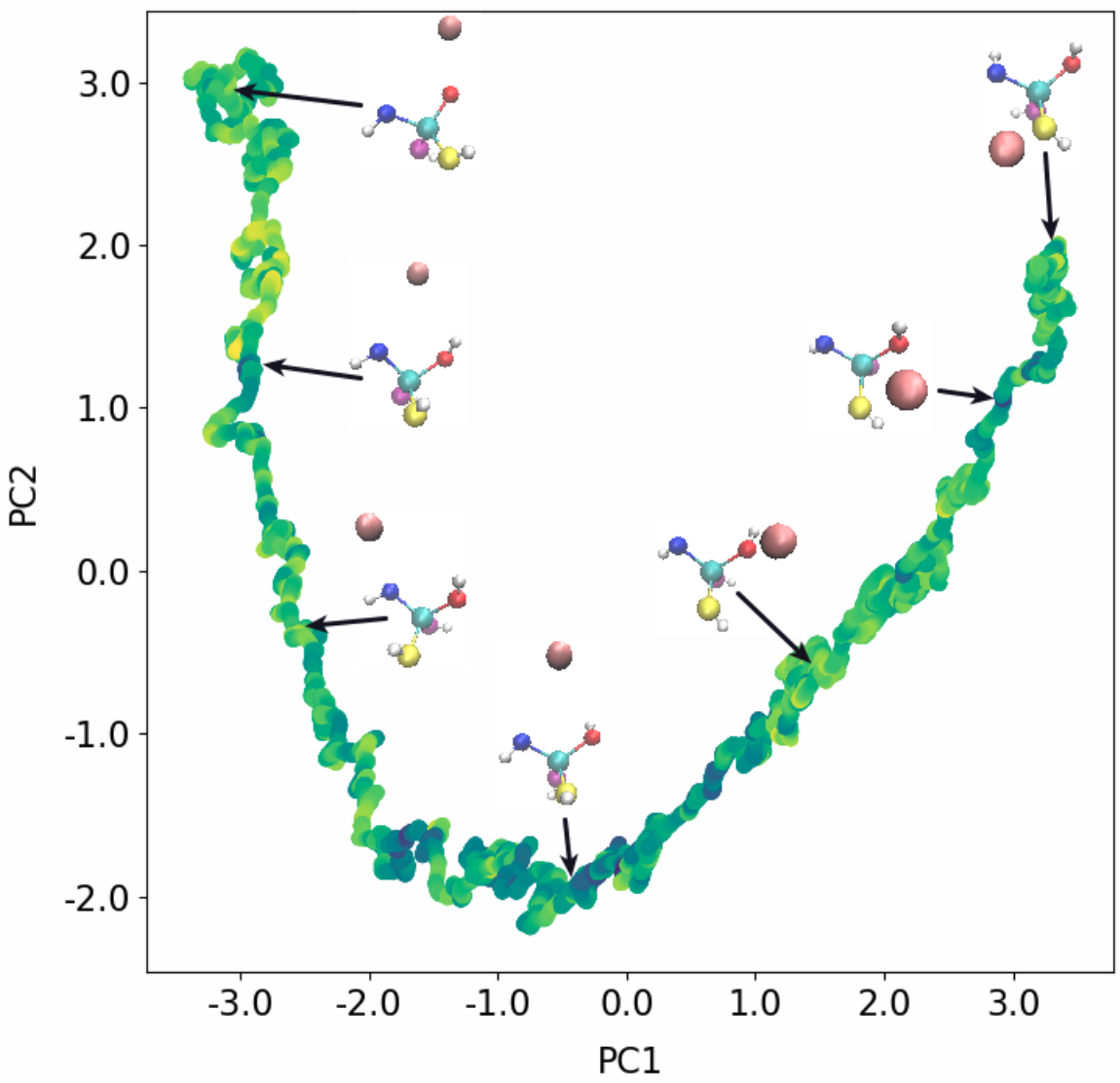

Figure 7: 2D plot of (PC1, PC2) from the PC analysis executed on the dataset given by inverse stereographic projection with $R_0 = 5$ Å. The color in PC plot is assigned based on K-O distance: bright indicates K-O interaction and dark indicates transition of K^+ between two O sites. This can be compared to the PC plot without water in Figure S9.

of the PCs. Figure S17 presents the contribution of water in the three PCs for each of the 52 water molecules in the simulation for different cut-off radii. We find that of the three PCs, the H_2O contribution is strongest in PC1. Correspondingly, $\lambda_{solute}\langle PC1\rangle$ in Table 1 changes most dramatically with the R_0 of the projection.

Figure 8 shows the PCA of the trajectories obtained after using the cut-projection technique for R_0 = 1.5 - 3 Å which has 0 - 3 H_2O. These predominantly include the reacting H_2O and the H_2O that picks up the proton from the first water once it reacts with ethylene. Similar figures showing the effect of increasing R_0 on the PCA results are provided in Figures S10 - S13. At R_0 = 1.5 Å, many of the features seen in the case without water molecules are discernible (Figure 5(b)), however the correlation with the water-absent data decreases as more H_2O are included within the cut-off, even as R_0 increases to 2 or 3 Å. For example, PC1, as seen previously for the nucleophilic reaction, separates the region of *cis* and *trans* configurations. PC3 and PC1 together represent the dihedral rotation of the molecule. A clear reaction coordinate is not identified in either of the PCs. As such, despite the fact that the solvent affect may be dampened using our projection methods, it is still a challenge to clearly identify all the CVs observed in the reaction in the presence of more than three H_2O. Further increasing R_0 adds more H_2O to the system which drowns out the important and meaningful features of the chemical reaction.

Nonlinear DR of Cation Roaming and Nucleophilic Attack Reactions. The UMAP projections were examined as a function of the number of H_2O that are included with various R_0. Starting with the cut-projection, the more solvent that's introduced, the more the variation of the H_2O coordinates dwarf the changes to geometry associated with the chemical reaction. UMAP is highly sensitive to solvent variation, and including too many H_2O results in a UMAP representation that captures the time trajectory of the data rather than the reaction coordinates. For the roaming reaction, $R_0 < 4$ Å removed all H_2O and as anticipated produced similar reductions to that of the original data with no water. Variations in the

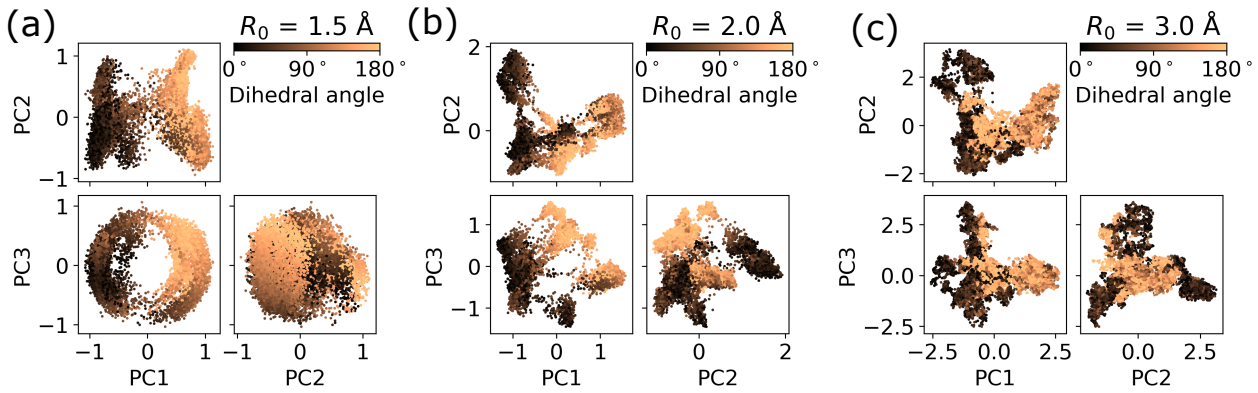

Figure 8: 2D plot of the first three principal components–PC1, PC2 and PC3 as obtained from the PCA executed on the nucleophilic attack dataset given by cut projection with (a) $R_0 = 1.5$ Å, (b) $R_0 = 2$ Å, and (c) $R_0 = 3$ Å. Note that projection with a small R_0 is critical in suppressing solvent contributions and recovering a clear separation of configurations by dihedral angle into *cis* and *trans* isomers.

reductions are thus due to the affect of the cut- and inverse stereographic projection upon the reaction coordinate. An $R_0 > 5$ Å led to ca. 10 H_2O being included within the data and the H_2O contributions dominate the UMAP space (Figure S14). Only with a cut-radius of 4 Å, are three distinct components observed that correspond to the location of K^+ with respect to the hydroxyls (Figure 9) and incorporate the contributions of 3 H_2O within the trajectory. Similar behavior is observed for the nucleophilic attack reaction trajectory. Here we see a somewhat similar structure as with the no solvent data, where we see the dihedral transitions (θ transitions 1 - 2) "hopping" between two central components, and the two attacks (r_{CO} transitions 1 - 2) "hopping" between the two inner components and the two outer components (Figure 9) The net result is that with suitable cut-radii, UMAP representations of the cut-projected data successfully capture the desired CVs in both data at 1200 and 400 nearest neighbors, respectively(Figure 9). UMAP projections on the inverse stereographic projection data for both reactions ultimately did not alleviate time-dependency regardless of the radius or NN value (Figure S15).

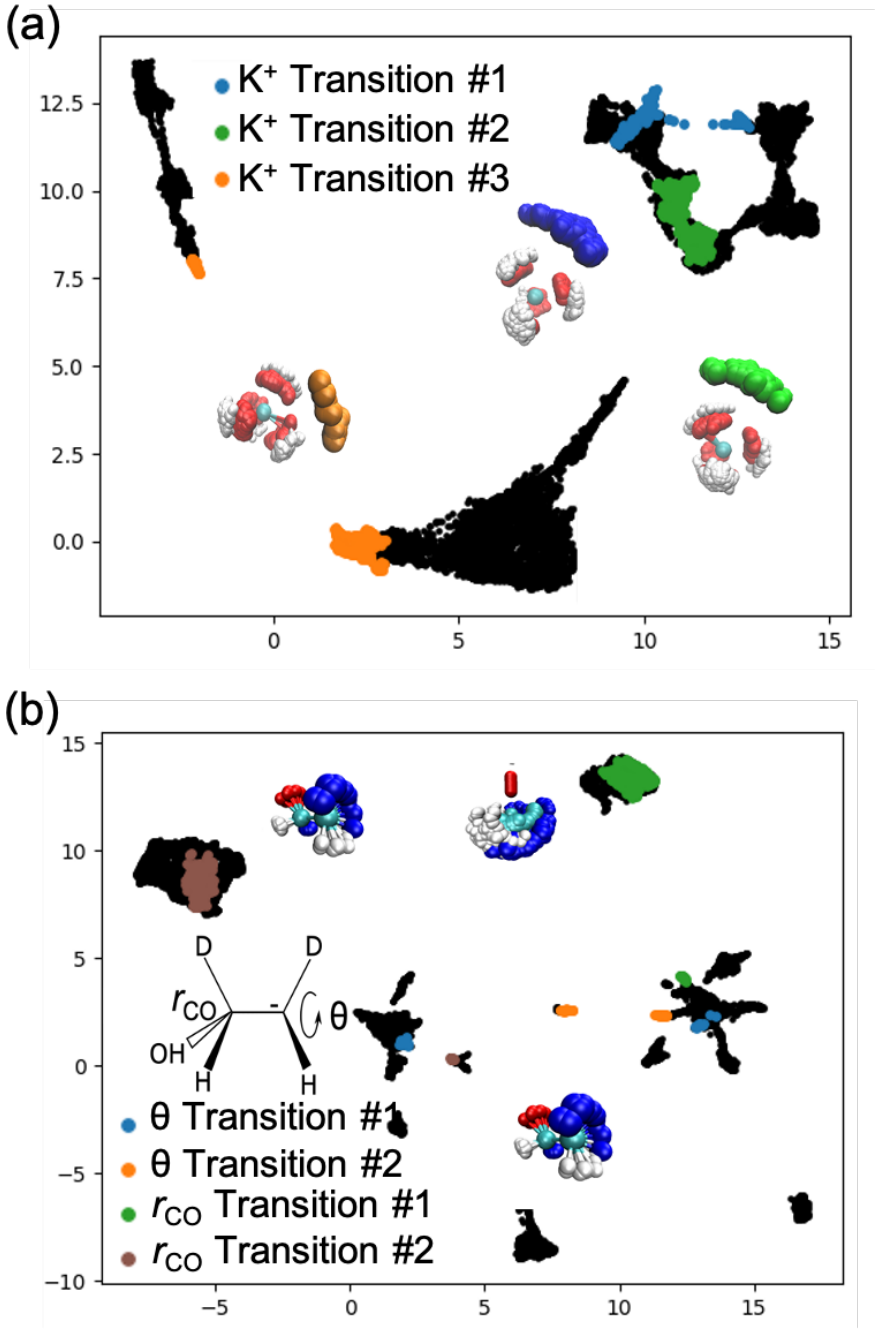

Figure 9: (a) UMAP representation of the cation roaming reaction with 1200 NN. (b) UMAP representation for the nucleophilic attack reaction with 400 NN.

Conclusions

A detailed comparison has been made of the ability of linear versus nonlinear dimensionality reduction methods (PCA vs. UMAP) to identify the collective variables of simple reactions in solution when the solvent is an active participant. Although PCA has been amply studied for reactive systems where there is significant correlated motion of the participating chemical species, the nuanced role of solvent within solution phase reactions is often ignored. To delineate between solvent that plays an active role within the solution phase reactions and non-participating solvent molecules, two different projection methods have been developed that attenuate translational motion of solvent as a function of distance from the reacting species. In the absence of such projection, both PCA and UMAP are dominated by the bulk solvent motion within the reduced representation. However with this attenuation in place, we demonstrate that both PCA and UMAP are capable of identifying the appropriate collective variables in solution. Using the distance dependence of the attenuation of solvent motion, we can further identify the specific solvent molecules that participate in the reaction. These methods are not without significant computational considerations, and even attenuation of solvent motion does not guarantee the ability to identify the appropriate collective variables. In some cases, even the addition of a single extra solvent molecule to the projected coordinates inhibits PCA or UMAP from yielding meaningful reduced representations. Further, it is important to recognize that closed form expressions are usually not available to express either the reduced dimensions produced by nonlinear DR techniques such as UMAP as functions of the input dimensions, or the input dimensions as functions of the reduced ones. Despite these issues, the ability of PCA and UMAP to identify the reaction collective variables within these systems represents a promising step forward for their application and the development of new dimensionality reduction methods that can interpret complex reaction spaces.

Acknowledgment

This material is based upon workpartially supported by the National Science Foundation under Grants 1934725, 1661348, and 1819229. This research used resources from the Center for Institutional Research Computing at Washington State University. We are thankful to S. R. Hare (School of Chemistry, University of Bristol) for providing the *PathReducer* code.

Supporting Information Available

Relevant plots that demonstrate distance dependencies upon the cut-ff radius within the projection methods, comparisons of behavior for the different projection methods, and supporting information for the computational methods are provided.

References

(1) Wales, D. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*; Cambridge Molecular Science; Cambridge University Press, 2004.

(2) Heuer, A. Energy Landscapes. Applications to Clusters, Biomolecules and Glasses. By David J. Wales. *Angew. Chem. Int. Ed.* **2005**, *44*, 1756–1757.

(3) Prada-Gracia, D.; Gómez-Gardeñes, J.; Echenique, P.; Falo, F. Exploring the Free Energy Landscape: From Dynamics to Networks and Back. *PLOS Comput. Biol.* **2009**, *5*, 1–9.

(4) Kachmar, A.; Goddard, W. A. Free Energy Landscape of Sodium Solvation into Graphite. *J. Phys. Chem. C* **2018**, *122*, 20064–20072.

(5) Patel, L. A.; Kindt, J. T. Simulations of NaCl Aggregation from Solution: Solvent Determines Topography of Free Energy Landscape. *J. Comput. Chem.* **2019**, *40*, 135–147.

(6) Peters, B.; Bolhuis, P. G.; Mullen, R. G.; Shea, J.-E. Reaction Coordinates, One-Dimensional Smoluchowski Equations, and a Test for Dynamical Self-Consistency. *J. Chem. Phys.* **2013**, *138*, 054106.

(7) Peters, B. Reaction Coordinates and Mechanistic Hypothesis Tests. *Annu. Rev. of Phys. Chem.* **2016**, *67*, 669–690.

(8) Komatsuzaki, T.; Hoshino, K.; Matsunaga, Y.; Rylance, G. J.; Johnston, R. L.; Wales, D. J. How Many Dimensions Are Required to Approximate the Potential Energy Landscape of a Model Protein? *J. Chem. Phys.* **2005**, *122*, 084714.

(9) Papaleo, E.; Mereghetti, P.; Fantucci, P.; Grandori, R.; Gioia, L. D. Free-Energy Landscape, Principal Component Analysis, and Structural Clustering to Identify Representative Conformations from Molecular Dynamics Simulations: The Myoglobin Case. *J. Mol. Graph. Model.* **2009**, *27*, 889 – 899.

(10) Jain, A.; Hegger, R.; Stock, G. Hidden Complexity of Protein Free-Energy Landscapes Revealed by Principal Component Analysis by Parts. *J. Phys. Chem. Lett.* **2010**, *1*, 2769–2773.

(11) Sicard, F.; Senet, P. Reconstructing the Free-Energy Landscape of Met-Enkephalin Using Dihedral Principal Component Analysis and Well-Tempered Metadynamics. *J. Chem. Phys.* **2013**, *138*, 235101.

(12) Ota, N.; Agard, D. A. Enzyme Specificity under Dynamic Control II: Principal Component Analysis of α-Lytic Protease Using Global and Local Solvent Boundary Conditions. *Protein Sci.* **2001**, *10*, 1403–1414.

(13) Karamzadeh, R.; Karimi-Jafari, M. H.; Sharifi-Zarchi, A.; Chitsaz, H.; Salekdeh, G. H.; Moosavi-Movahedi, A. A. Machine Learning and Network Analysis of Molecular Dynamics Trajectories Reveal Two Chains of Red/Ox-specific Residue Interactions in Human Protein Disulfide Isomerase. *Sci. Rep.* **2017**, *7*, 3666.

(14) Yang, L.-W.; Eyal, E.; Bahar, I.; Kitao, A. Principal Component Analysis of Native Ensembles of Biomolecular Structures (PCA_NEST): Insights into Functional Dynamics. *Bioinformatics* **2009**, *25*, 606–614.

(15) David, C. C.; Jacobs, D. J. In *Protein Dynamics: Methods and Protocols*; Livesay, D. R., Ed.; Humana Press: Totowa, NJ, 2014; pp 193–226.

(16) Kosanovich, K. A.; Dahl, K. S.; Piovoso, M. J. Improved Process Understanding Using Multiway Principal Component Analysis. *Ind. Eng. Chem. Res.* **1996**, *35*, 138–146.

(17) Quinn, K. N.; Clement, C. B.; De Bernardis, F.; Niemack, M. D.; Sethna, J. P. Visualizing Probabilistic Models and Data with Intensive Principal Component Analysis. *Proc. Natl. Acad. Sci.* **2019**, *116*, 13762–13767.

(18) Lange, O. F.; Grubmüller, H. Can Principal Components Yield a Dimension Reduced Description of Protein Dynamics on Long Time Scales? *J. Phys. Chem. B* **2006**, *110*, 22842–22852, PMID: 17092036.

(19) Hare, S. R.; Bratholm, L. A.; Glowacki, D. R.; Carpenter, B. K. Low Dimensional Representations along Intrinsic Reaction Coordinates and Molecular Dynamics Trajectories Using Interatomic Distance Matrices. *Chem. Sci.* **2019**, *10*, 9954–9968.

(20) Sajeevan, K. A.; Roy, D. Principal Component Analysis of a Conotoxin Delineates the Link among Peptide Sequence, Dynamics, and Disulfide Bond Isoforms. *J. Phys. Chem. B* **2019**, *123*, 5483–5493, PMID: 31095380.

(21) Lin, I.-C.; Tuckerman, M. E. Enhanced Conformational Sampling of Peptides via Reduced Side-Chain and Solvent Masses. *J. Phys. Chem. B* **2010**, *114*, 15935–15940, PMID: 21077595.

(22) Michielssens, S.; van Erp, T. S.; Kutzner, C.; Ceulemans, A.; de Groot, B. L. Molecular

Dynamics in Principal Component Space. *The Journal of Physical Chemistry B* **2012**, *116*, 8350–8354, PMID: 22263868.

(23) de Silva, V.; Tenenbaum, J. B. Global versus Local Methods in Nonlinear Dimensionality Reduction. Proceedings of the 15th International Conference on Neural Information Processing Systems. Cambridge, MA, USA, 2002; pp 721–728.

(24) Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proceedings of the National Academy of Sciences* **2006**, *103*, 9885–9890.

(25) Tenenbaum, J. B.; de Silva, V.; Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319–2323.

(26) Chen, W.; Ferguson, A. L. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *Journal of Computational Chemistry* **2018**, *39*, 2079–2102.

(27) Chen, W.; Tan, A. R.; Ferguson, A. L. Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design. *The Journal of Chemical Physics* **2018**, *149*, 072312.

(28) van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*, 2579–2605.

(29) McInnes, L.; Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints* **2018**, https://arxiv.org/abs/1802.03426.

(30) van der Maaten, L. Accelerating T-SNE Using Tree-Based Algorithms. *Journal of Machine Learning Research* **2014**, *15*, 3221–3245.

(31) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software* **2018**, *3*, 861.

(32) Allaoui, M.; Kherfi, M. L.; Cheriet, A. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. Image and Signal Processing. Cham, 2020; pp 317–325.

(33) Yang, Y.; Sun, H.; Zhang, Y.; Zhang, T.; Gong, J.; Wei, Y.; Duan, Y.-G.; Shu, M.; Yang, Y.; Wu, D.; Yu, D. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. *bioRxiv* **2021**,

(34) Diaz-Papkovich, A.; Anderson-Trocmé, L.; Ben-Eghan, C.; Gravel, S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLOS Genetics* **2019**, *15*, 1–24.

(35) Dorrity, M. W.; Saunders, L. M.; Queitsch, C.; Fields, S.; Trapnell, C. Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nature Communications* **2020**, *11*, 1537.

(36) Becht, E.; McInnes, L.; Healy, J.; Dutertre, C.-A.; Kwok, I. W. H.; Ng, L. G.; Ginhoux, F.; Newell, E. W. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology* **2019**, *37*, 38–44.

(37) Diaz-Papkovich, A.; Anderson-Trocmé, L.; Gravel, S. A review of UMAP in population genetics. *Journal of Human Genetics* **2021**, *66*, 85–91.

(38) Hutter, J.; Iannuzzi, M.; Schiffmann, F.; VandeVondele, J. CP2K: Atomistic Simulations of Condensed Matter Systems. *Comput. Mol. Sci.* **2014**, *4*, 15–25.

(39) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(40) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Phys. Rev. Lett.* **1997**, *78*, 1396–1396.

(41) Zhang, Y.; Yang, W. Comment on "Generalized Gradient Approximation Made Simple". *Phys. Rev. Lett.* **1998**, *80*, 890–890.

(42) Goedecker, S.; Teter, M.; Hutter, J. Separable Dual-Space Gaussian Pseudopotentials. *Phys. Rev. B* **1996**, *54*, 1703–1710.

(43) Hartwigsen, C.; Goedecker, S.; Hutter, J. Relativistic Separable Dual-Space Gaussian Pseudopotentials from H to Rn. *Phys. Rev. B* **1998**, *58*, 3641–3662.

(44) Krack, M. Pseudopotentials for H to Kr Optimized for Gradient-Corrected Exchange-Correlation Functionals. *Theor. Chem. Acc.* **2005**, *114*, 145–152.

(45) Chiodo, S.; Russo, N.; Sicilia, E. Newly Developed Basis Sets for Density Functional Calculations. *J. Comput. Chem.* **2005**, *26*, 175–184.

(46) VandeVondele, J.; Hutter, J. Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases. *J. Chem. Phys.* **2007**, *127*, 114105.

(47) Davidchack, R. L.; Handel, R.; Tretyakov, M. V. Langevin thermostat for rigid body dynamics. *The Journal of Chemical Physics* **2009**, *130*, 234101.

(48) Pouvreau, M.; Martinez-Baez, E.; Dembowski, M.; Pearce, C. I.; Schenter, G. K.; Rosso, K. M.; Clark, A. E. Mechanisms of Al3+ Dimerization in Alkaline Solutions. *Inorganic Chemistry* **2020**, *59*, 18181–18189.

(49) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.* **2014**, *185*, 604 – 613.

(50) Zhang, W.; van Duin, A. C. Improvement of the ReaxFF description for functionalized hydrocarbon/water weak interactions in the condensed phase. *The Journal of Physical Chemistry B* **2018**, *122*, 4083–4092.

(51) Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics* **1995**, *117*, 1–19.

(52) Wilson, E.; Decius, J.; Cross, P. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*; Dover Books on Chemistry Series; Dover Publications, 1980.

(53) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(54) Altis, A.; Otten, M.; Nguyen, P. H.; Hegger, R.; Stock, G. Construction of the Free Energy Landscape of Biomolecules via Dihedral Angle Principal Component Analysis. *J. Chem. Phys.* **2008**, *128*, 245102.

Supporting Information Available

~~Supporting Information can go here~~