Data-driven classification of ligand unbinding pathways

Dhiman Ray[a,1] and Michele Parrinello[a,1]

Studying the pathways of ligand–receptor binding is essential to understand the mechanism of target recognition by small molecules. The binding free energy and kinetics of protein–ligand complexes can be computed using molecular dynamics (MD) simulations, often in quantitative agreement with experiments. However, only a qualitative picture of the ligand binding/unbinding paths can be obtained through a conventional analysis of the MD trajectories. Besides, the higher degree of manual effort involved in analyzing pathways limits its applicability in large-scale drug discovery. Here, we address this limitation by introducing an automated approach for analyzing molecular transition paths with a particular focus on protein–ligand dissociation. Our method is based on the dynamic time-warping algorithm, originally designed for speech recognition. We accurately classified molecular trajectories using a very generic descriptor set of contacts or distances. Our approach outperforms manual classification by distinguishing between parallel dissociation channels, within the pathways identified by visual inspection. Most notably, we could compute exit-path-specific ligand-dissociation kinetics. The unbinding timescale along the fastest path agrees with the experimental residence time, providing a physical interpretation to our entirely data-driven protocol. In combination with appropriate enhanced sampling algorithms, this technique can be used for the initial exploration of ligand-dissociation pathways as well as for calculating path-specific thermodynamic and kinetic properties.

molecular dynamics | dynamic time warping | machine learning | kinetics | rare events

Mechanistic understanding of protein–ligand interaction is of fundamental importance to the rational design of therapeutic drugs. Atomistic molecular dynamics (MD) simulations can provide insight into the mechanisms of protein–ligand association and dissociation, and are, therefore, used extensively in computer-aided drug design (1). With the development of enhanced sampling algorithms (2) and coarse-grained models (3), it is now possible to simulate physiologically relevant drug-receptor interactions at an affordable computational cost.

Most drug discovery efforts based on MD simulation have focused on calculating the binding free energy which indicates the thermodynamic stability of the ligand–receptor complex (1). However, the kinetics and pathways of target recognition are critical determinants of the efficacy and selectivity of small-molecule drugs. It has been demonstrated, in a variety of targets, that drug residence time, i.e., the ligand unbinding kinetics is better correlated to the in-vivo efficacy of pharmaceutical drug candidates (4–6). The significance of drug unbinding pathways has been realized in recent experimental and computational studies when Lyczek et al. and Shekhar et al. showed that a drug-resistant mutation in Abl kinase can reduce the activity of the anticancer drug Imatinib (Gleevec) by modifying only the drug unbinding pathway without changing its binding affinity (7, 8).

Despite the progress in developing physics-based algorithms for accurate prediction of free energy and kinetics of ligand–receptor binding, the study of molecular pathways is usually conducted at a qualitative level, through manual observation and interpretation of the MD trajectories. This can become impractical if screening of a large number of drug candidates for a given pharmaceutical target is required. Moreover, a large volume of MD simulation datasets are being made available recently for pharmaceutically relevant targets such as the proteins in the SARS-CoV-2 virus (9). Significant human effort is required to analyze these datasets and understand the molecular processes involved. For example, manual analysis of multiple microseconds of MD trajectories has led to the discovery of the dominant pathways for the binding of beta-blocker drug molecules to the G-protein-coupled receptors (GPCR) (10). Ansari et al. discovered, through visual inspection, two different pathways of benzamidine unbinding from trypsin protein (11). The two paths differ only in the role of the hydration water molecules in facilitating the ligand release. However, the kinetics of these pathways differ by almost one order of magnitude.

Significance

Present-day computational drug design primarily relies on ligand–receptor binding free energies, despite the growing realization that drug residence time and unbinding pathways play key roles in determining in-vivo efficacy. We introduce an automated approach to classify ligand dissociation pathways using a powerful speech recognition algorithm called dynamic time warping. Combining enhanced sampling atomistic simulations with our path-classification algorithm, we distinguish various ligand unbinding pathways with ~90% accuracy and find kinetically distinct dissociation channels that were indistinguishable through conventional analysis of the trajectories. We also predict exit-path-specific ligand unbinding kinetics in quantitative agreement with experiments. Incorporating this information can transform rational drug design and help combat the emergence of drug-resistant mutations.

[1]To whom correspondence may be addressed. Email: dhiman.ray@iit.it or michele.parrinello@iit.it.

Only recently, efforts are being devoted to designing automated approaches to analyze the transition pathways sampled from MD simulation. Notable examples include the use of a t-distributed stochastic neighbor embedding (t-SNE) algorithm to project the drug unbinding trajectories in a low dimensional space followed by agglomerative clustering to classify the sampled pathways (12). A variational autoencoder-based latent-space path clustering (LPM) algorithm could also distinguish molecular transition pathways into parallel kinetic channels (13), making it easier to decipher molecular mechanisms (14). In a recent study, it has been demonstrated that a standard autoencoder trained on molecular descriptor space can also be useful to distinguish the different pathways of ligand binding to T4-Lysozyme (15). Wolf and coworkers have attempted not only to classify the ligand release pathways but also to compute pathway-specific free energy profiles and kinetics. To this effect, they used principal component analysis (PCA) and constructed a dendrogram of molecular trajectories (16). Pathway-dependent kinetics and free energy profiles are then computed using dissipation-corrected targeted molecular dynamics (dcTMD) (17, 18) and a modified version of Jarzynski equality (19). Deep learning techniques such as self-organizing maps (SOM) have also been tested for classifying ligand unbinding pathways. The resulting approach, known as PathDetect-SOM, could successfully distinguish different pathways of inhibitor dissociation from heat shock protein 90 (HSP90) (20).

The currently available path classification algorithms have a few key limitations making them difficult to use in drug discovery. These methods often assume that different pathways are distinguishable when projected on a reduced dimensional space (12–16). However, during the process of binding or unbinding, a drug interacts with many residues of the target protein in a specific temporal order characteristic to the particular pathway. Therefore, it is not guaranteed that such complex pathways can be separated in a low-dimensional latent space. Moreover, considerable system-specific knowledge is necessary to successfully use these algorithms, making them unsuitable for large-scale applications. Often, it is also necessary to use Markov state modeling (MSM) and (or) transition path theory (TPT) to preprocess and coarse-grain the trajectory data before it can be used for classification, significantly increasing the complexity of the protocol.

In this work, we aim to address these challenges by constructing an automated data-driven algorithm for the classification of transition pathways for molecular systems. Our approach is based on the powerful speech recognition algorithm, dynamic time warping (DTW) (21), which also found applications in signal processing (22), gene sequence alignment (23), handwriting recognition (24), gesture recognition (25) and predicting trends in finance and econometrics (26–28). DTW is suitable for comparing time series or sequences of different lengths and clustering them based on their degree of similarity (29).

At a fundamental level molecular reaction pathways are high dimensional time sequences of uneven length since, due to the inherent stochasticity, transition paths may take different times in different independent trajectories. Dynamic time-warping can take into account this issue and can reduce the manual effort of trajectory pre-processing to a large extent. This feature motivated us to use DTW to classify various reactive pathways directly from MD trajectories. This comes with a few notable advantages. First, no dimensionality reduction or MSM construction is necessary, making it possible to compare directly different trajectories. Moreover, the DTW classification algorithm is independent of the enhanced sampling algorithm used to explore the pathways.

Therefore, by making a careful choice of the enhanced sampling algorithm, such as metadynamics or conformational flooding, one can recover pathway-specific properties (i.e., free energy or kinetics) without modifying the path-classification protocol.

Using a combination of enhanced sampling and dynamic time warping, we show that different reactive pathways can be accurately distinguished for a model Müller Brown potential, in the conformational transition of Alanine dipeptide, and for protein–ligand unbinding. We chose, as an example of ligand–receptor dissociation, the complex of benzene with the L99A mutant of T4 lysozyme, a system for which multiple ligand unbinding pathways have been reported in the literature (30–35). Therefore, it is possible to compare the performance of our automated path classification protocol against the analysis conducted by human experts. We could also calculate the unbinding timescale (τ_{off}) for individual pathways, with the fastest dissociation timescales in quantitative agreement with the experimental residence time. Although designed to study ligand–receptor binding or unbinding, our protocol can be generalized to any molecular conformational transition.

Dynamic Time Warping

To classify molecular trajectories into different pathways it is necessary to construct a distance metric to measure the similarity between each pair of sampled transitions. Although root mean square deviation (RMSD) or similar metrics are widely used to compare static conformations of molecules, measuring the difference between dynamic trajectories can be challenging. The trajectories of protein–ligand dissociation can be represented as high-dimensional time series in molecular descriptor space composed of interatomic distances and contacts. The trajectories can have very different lengths and the key events in the unbinding mechanism (e.g., interaction between the ligand and protein residues) can take place at different times in different trajectories. Therefore a direct comparison of sampled conformations between different trajectories is not possible without the loss of the temporal information and compromising the accuracy of the classification algorithm (Fig. 1A).

Dynamic Time Warping (DTW) is a data mining algorithm that is capable of comparing time series that are of unequal length (Fig. 1B). However, the utility of DTW in classifying molecular trajectories has not been exploited so far. A different metric named Fréchet distance (FD) has been used to monitor the convergence of MD trajectories (36). The computational complexity of FD is, however, higher than DTW and it only contains information about the maximum distance between trajectories (37), which limits its widespread adoption.

The DTW algorithm is primarily used to distinguish one-dimensional time series. Generalizations of DTW to higher dimensions can be accomplished through the decoupling of each dimension of the time series (independent DTW or DTW_I). However, such approaches are not suitable for analyzing MD trajectories, as molecular degrees of freedom are often correlated to one another. A recently proposed modification of DTW (or DTW_D, where the suffix D stands for dependent) can classify high-dimensional time series in correlated descriptor space (38). The DTW_D method has been demonstrated to have a higher success rate compared to DTW_I in gesture recognition from videos and for analyzing accelerometer data (38). Below we summarize the theoretical background of dynamic time warping, followed by a simple one-dimensional example.

Let's assume there are two one-dimensional time series $Q = q_1, q_2, q_3, ..., q_{m_1}$ and $C = c_1, c_2, c_3, ..., c_{m_2}$ where in general
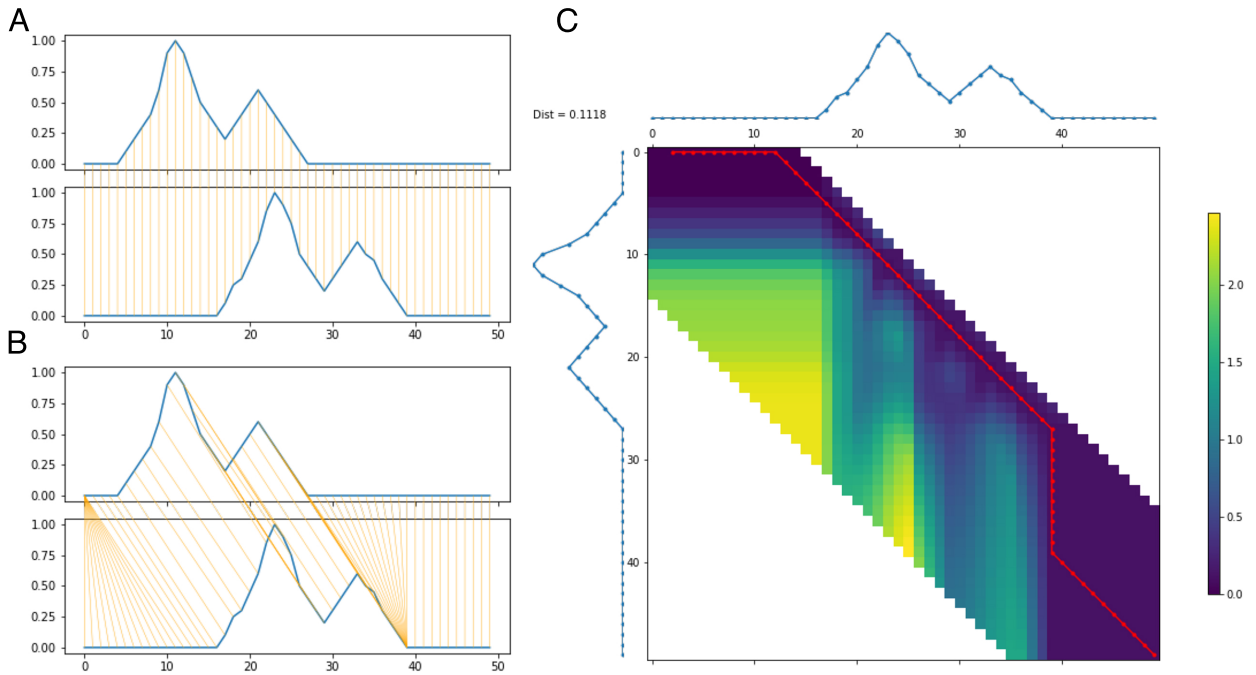
Fig. 1. A schematic representation of dynamic time warping of two 1D time series. Panel (*A*) denotes the situation of a Euclidean warping where the signal at time = *t* from time series 1 is compared directly with the signal at time *t* of signal 2. (*B*) In dynamic time warping an optimal comparison between the two time-series is performed by minimizing the warping cost. This helps in comparing specific features of time series 1 with the corresponding signals in time series 2. In both (*A*) and (*B*), the x-axis indicates time, and the y-axis indicates the signal of the time series. (*C*) The matrix **D** for comparing the two 1D time series. The colors indicate the value of the matrix elements. The minimum cost warping path is shown in red. In this example, the maximum deviation of the warping path from the diagonal is set to 15.

$m_1 \neq m_2$. DTW creates a one-to-many alignment between these sequences by constructing a matrix $\mathbf{D} \in \mathbb{R}^{m1 \times m2}$, whose elements are given by

$$D_{ij} = d(q_i, c_j)$$
$$+ \min\{D(i-1, j-1), D(i-1, j), D(i, j-1)\}, \quad [1]$$

where $d(q_i, c_j)$ is the squared Euclidean distance between q_i and c_j: $d(q_i, c_j) = (q_i - c_j)^2$. A warping path P of length T is defined as a contiguous set of matrix elements representing a mapping between time series Q and C:

$$P = p_1, p_2, p_3, ..., p_T$$
$$\text{such that} \quad \max(m_1, m_2) \leq T \leq m_1 + m_2 - 1. \quad [2]$$

The warping path is subject to the following constraints: 1) P must start and finish in diagonally opposite corners of the matrix *A*. 2) The steps in the warping path are restricted to adjacent cells. 3) The points in the warping path are monotonically ordered in time. In addition, a restriction may also be imposed on how far the path is allowed to deviate from the diagonal. The cost $\mathcal{C}(P)$ associated with the warping path P is defined as the sum of the matrix elements that are part of the warping path.

$$\mathcal{C}(P) = \sum_{t=1}^{T} p_t = \sum_{D_{ij} \in P} D_{ij}. \quad [3]$$

The DTW distance between two-time series is measured from the path that minimizes the warping cost:

$$DTW(Q, C) = \min_{P_k \in \mathbb{P}} \left\{ \mathcal{C}(P_k) \right\}, \quad [4]$$

where $\mathbb{P}$ is the set of all possible warping paths for the matrix **D**. The best possible alignment between the two time series can be obtained as the warping path with the least warping cost. It is represented by a valley of low D_{ij} in the space of the matrix elements of **D** (Fig. 1*C*). After evaluating Eq. **1** for all elements of matrix **D**, the DTW distance between the series C and Q is computed by tracing backward the pathway of least D_{ij} values starting from the last element of the matrix $D_{m_1 m_2}$.

The dynamic time-warping algorithm is demonstrated with a simple example in Fig. 1. We compare two 1D time series where the characteristic signal is virtually identical but out of phase. By directly comparing the two signals at every time point, one obtains the Euclidean warping (Fig. 1*A*). This, however, fails to capture correctly the similarity between the two signals as the phase difference is not taken into account. DTW aims to circumvent this problem by finding an optimal alignment between the sequences to make a more meaningful comparison (Fig. 1*B*). To accomplish this, one constructs a matrix whose elements measure the pairwise Euclidean distance between every point in the two time-series. The dynamic programming algorithm (Eq. **1**) then identifies a continuous path (colored red in Fig. 1*C*) of the smallest elements through the matrix. This process can be thought to be analogous to identifying a 1D minimum energy manifold in a 2D potential energy surface. The DTW distance is then measured by summing all the elements of this optimal warping path.

The DTW$_D$ generalizes this algorithm to multidimensional situation by modifying the distance measure $d(q_i, c_j)$ from Eq. **1** as:

$$d(q_i, c_j) = \sum_{m=1}^{M} (q_{i,m} - c_{j,m})^2. \quad [5]$$

Here, $q_{i,m}$ is the i-th data point in the m-th dimension of an M dimensional time series Q, and $c_{j,m}$ is the j-th data point in the m-th dimension of an M dimensional time series C.

Based on the distances between trajectories computed from DTW_D algorithm, a pairwise distance matrix is generated. K-medoid clustering is performed using this distance matrix to cluster the trajectories corresponding to different pathways. The choice of the clustering algorithm is inspired by the data mining and time series clustering literature (39–42). K-medoid clustering identifies a physical trajectory (medoid) that is nearest to each cluster center, thereby increasing the interpretability of the overall protocol (43). To identify the optimum number of clusters, we used the silhouette score metric (44). A schematic of the overall protocol is shown in Fig. 2. Analysis of the computational complexity of the overall protocol is provided in *SI Appendix*.

Pathway Classification

Müller–Brown Potential. To test the performance of DTW_D in classifying MD trajectory data, we first applied it to the model system of Müller–Brown potential. We could sample two different transition pathways using either the x or the y coordinate as CV in the exploratory variant of the On-the-fly Probability Enhanced Sampling (45)(OPES Explore or $OPES_E$). The trajectory data from two pathways are presented together to the DTW_D algorithm, which could accurately distinguish the paths from the entire ensemble of trajectories represented as 2D time series $(x(t), y(t))$ (Fig. 3). Despite the simplicity of the problem, this serves as a proof of concept of our protocol and sets the stage for its application to more complex systems.

Alanine-Dipeptide. Next, we tested our approach to distinguish conformational transition paths in the gas phase alanine dipeptide. We sampled 100 transitions from the $C7_{eq}$ state to the $C7_{ax}$ state using $OPES_E$ algorithm with the ϕ torsion angle as CV. These transitions took place along 3 distinct pathways (Fig. 4). Two of them proceed through the middle of the free energy surface (Fig. 4 *C* and *D*), while the third is a high energy path involving a reverse rotation around the ϕ torsion angle (Fig. 4*B*). To avoid the inclusion of pre-existing system knowledge, we used a descriptor set consisting of 45 pair-wise distances between all nonbonded heavy atoms (46, 47). Even without any direct information about the ϕ and ψ torsion angles that best describes the transitions, the DTW_D algorithm successfully classified the pathways into three different clusters. In addition, the silhouette score metric indicated that the optimum number of clusters in

this trajectory ensemble is 3. This result demonstrates that our protocol can correctly classify molecular paths using a generic descriptor set, and can automatically identify the number of pathways involved in the process.

Protein–Ligand Unbinding. Now we investigated the accuracy of our approach in classifying unbinding pathways of benzene ligand from the L99A mutant of T4 Lysozyme. Using $OPES_E$ simulation, we could sample 100 dissociation events within a cumulative computational time of tens of nanoseconds. The ligand dissociation pathways can be manually classified into eight pathways based on the interaction of the ligand with eight different helices of the protein (labeled c to j in Fig. 5). The transition paths sampled in the present study (Fig. 6) are consistent with those of Capelli et al. (31), Rydzensky and Valsson (32), and Nunes-Alves et al. (30, 33). To provide a general description of the unbinding process, the coordination number of the ligand heavy atoms with the C_α atoms of each residue of the protein is included in the descriptor set resulting in a feature space with a dimensionality of 161. It should be noted here that the DTW algorithm has rarely been applied to such a high-dimensional time series clustering problem. However, our DTW_D protocol could identify seven clusters among the 100 ligand release trajectories, out of which five clusters directly correspond to five pathways (A, C, D, E, and H) identified through manual classification. The pathways F and G are combined within one cluster presumably because in these pathways benzene interacts with the g, j, and h helices in a virtually identical manner except for changing course following the exit from the binding site. Instead, path B could not be identified as a separate cluster, likely due to its low population and its similarity with Path A and C in terms of the interaction of benzene with the protein residues in helix c (Fig. 6). To visualize the classification accuracy of the DTW_D algorithm, we constructed a confusion matrix with the columns representing the pathways identified through manual classification and the rows representing those clustered by our automated protocol. The precision, recall, and F1 score values for every cluster are above 0.75 and an overall weighted F1 score of 0.86. This indicates a classification accuracy of 86% which, considering the complexity of the protein–ligand dissociation mechanism, is noteworthy.

Pathway Dependent Kinetics of Ligand Unbinding

To calculate the pathway-specific kinetics of ligand dissociation, we sampled 137 benzene unbinding events using OPES flooding ($OPES_f$) (48) algorithm (See *Materials and Methods* and *SI Appendix* for details). The dissociation timescale (τ_{off}) predicted from the entire set of trajectories (2.5 ms) is in good agreement with the experimental residence time of 1.1 ms. However, the P-value of the Kolmogorov–Smirnov (KS) test comparing the cumulative distribution of the ligand unbinding times with a perfect Poisson distribution is as low as 0.02. This indicates a deviation from a single exponential kinetic model, due to the different timescales involved in the exit of the ligand along different pathways.

An identical path classification protocol based on the 161-dimensional molecular descriptor set could cluster the $OPES_f$ trajectories into 16 clusters, with one or more clusters assigned to each ligand release pathway. The classification accuracy, measured in terms of weighted F1 score is 96%, an improvement over the results of the $OPES_E$ simulations. The increase in
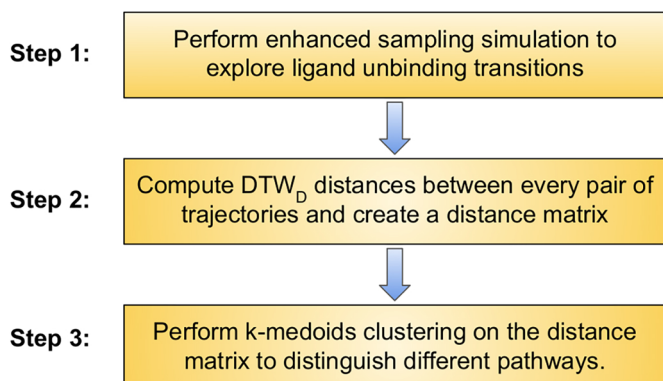

Step 1: Perform enhanced sampling simulation to explore ligand unbinding transitions

Step 2: Compute DTW_D distances between every pair of trajectories and create a distance matrix

Step 3: Perform k-medoids clustering on the distance matrix to distinguish different pathways.

Fig. 2. A flowchart of the dynamic-time-warping-based protocol for classifying ligand unbinding pathways.
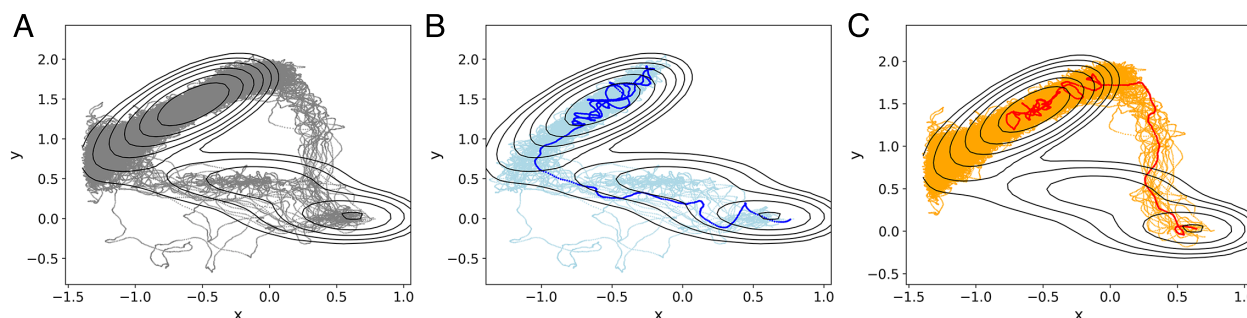
Fig. 3. (A) All 40 transitions, sampled using OPES$_E$ simulation, for the 2D Müller–Brown potential. Panels (B) and (C) represent the two clusters of pathways identified by the DTW$_D$ algorithm. The cluster medoid of each pathway is highlighted.

classification accuracy can be attributed to the fact that in OPES$_f$ trajectories spend more time in the transition region leading to better sampling of the exit paths. The pathways D, F, and H, for which there are only 1 to 3 samples, could also be classified correctly by our algorithm.

However, The most remarkable outcome of this work, is that we could calculate the exit-path-specific dissociation kinetics for this protein–ligand complex. The τ_{off} values differed significantly for different path clusters even within the same pathway. For example, the two clusters corresponding to Path A led to τ_{off} values that are different from each other by a factor of ∼3. Similarly, the unbinding timescales of four clusters for Path C range between ∼1.5 ms and ∼5 ms. This observation is noteworthy as these kinetically distinct path clusters are difficult to distinguish through manual visualization alone (Fig. 7).

Most notably, the dissociation channels, identified through DTW$_D$ analysis, showed single exponential kinetics, demon-

strated by KS test P-values significantly higher than 0.05 (between 0.3 and 0.99) (SI Appendix, Fig. S6). Our classification protocol, therefore, distinguishes between kinetically distinct pathways without using any kinetic information in the training data. The fastest ligand dissociation path has a τ_{off} of 1.27 ms, which is in better agreement with the experimental unbinding rate constant (k_{off} = 950 ± 20 s^{-1}), compared to the estimate obtained by taking together the dissociation timescales along different channels. This finding corroborates the pre-existing knowledge that the fastest unbinding pathway dominates the ligand residence time. We, however, avoided performing a weighted average of the residence times across different dissociation channels due to lack of sufficient statistics across all pathways. We could also identify the role of individual residues that distinguish between these dissociation channels within individual unbinding pathways, as shown in SI Appendix.
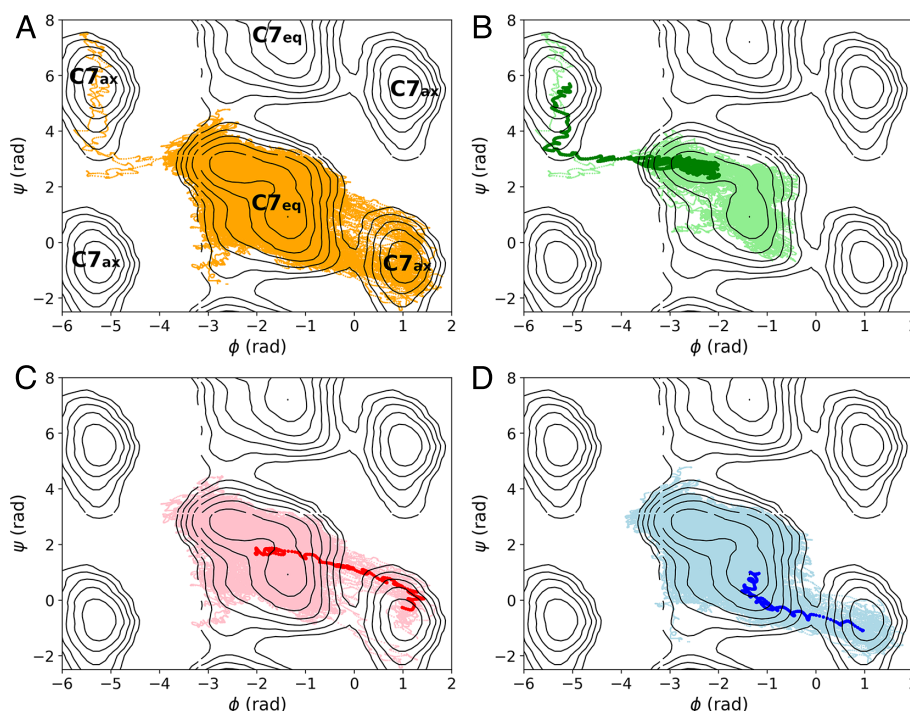

Fig. 4. Panel (A) shows all 100 transitions from C7$_{eq}$ state to the C7$_{ax}$ state of gas phase alanine dipeptide, sampled using OPES$_E$ algorithm. The trajectories are projected along the ϕ and ψ torsion angles. The free energy surface for the system is shown in black contour lines. The adjacent periodic images of the free energy contour are also shown for ease of visualizing continuous pathways. The C7$_{eq}$ and C7$_{ax}$ states are labeled in Panel (A). The Panel (B–D) show the trajectories belonging to the three path clusters identified by DTW$_D$ algorithm. The medoid of each cluster is highlighted in dark green, red, and dark blue colors respectively.
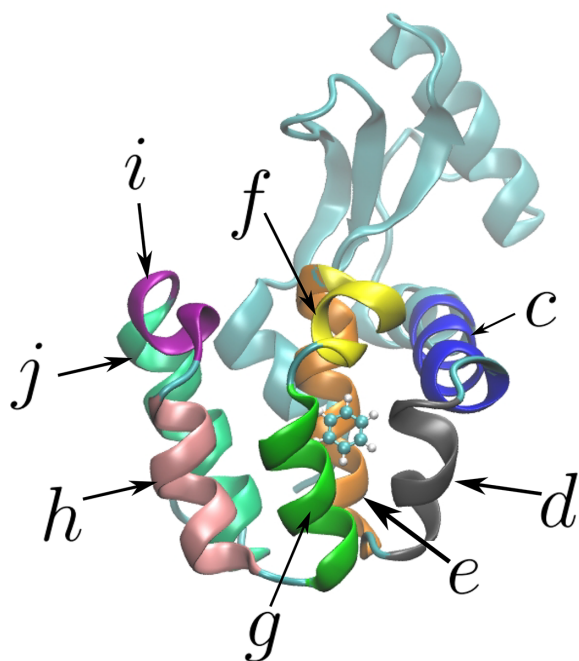
Fig. 5. The prototypical protein–ligand complex of benzene bound to L99A mutant of T4 Lysozyme. The 8 α-helices necessary to distinguish the various ligand release pathways are labeled *c* through *j*.

Discussions and Conclusions

We demonstrate that a multidimensional dynamic time warping algorithm can accurately classify molecular transitions into different pathways in an automated data-driven manner. This protocol is suitable for the study of protein–ligand interaction where the kinetic properties are known to be dependent on the transition path. Our approach treats MD trajectories as time series in a high dimensional generic descriptor space, thereby reducing the manual effort and intuition involved in

understanding complex ligand dissociation mechanisms. The DTW$_D$ protocol does not require any ad-hoc dimensionality reduction or the use of approximations like transition path theory, making it applicable to a wide range of molecular processes without system-specific knowledge.

To explore all possible pathways of ligand binding/unbinding, we use OPES$_E$ algorithm to sample transitions that are clustered afterward based on the DTW$_D$ distance metric. This exercise provides only a qualitative description, but it can have important ramifications in computer-aided drug discovery. For example, we utilized DTW to analyze the trajectories from different clusters to identify the key residues that differentiate the mechanism of ligand unbinding along different pathways. As shown by Shekhar et al. (8), such information can provide an intuition about potential mutations that can develop resistance against potential therapeutic candidates.

For a more quantitative insight, one can perform OPES$_f$ or OPES simulations, which, although computationally expensive, can compute the kinetics and free energy landscape, respectively, for individual pathways discovered by DTW$_D$. We demonstrated this by computing exit-path-specific τ_{off} for the benzene T4 lysozyme complex. The observed variance among the unbinding kinetics calls for a careful path classification analysis whenever attempting a ligand residence time calculation. Out of many possible transitions sampled from biased simulations, the kinetics along the fastest unbinding pathway is in excellent agreement with the experimental residence time.

Ansari et al. have demonstrated that despite the apparent simplicity of the trypsin–benzamidine complex there are two different pathways of ligand unbinding with different dissociation timescales (11). We see similar but more complex behavior in the case of benzene dissociation from T4 Lysozyme. The timescales for different pathways only vary by a factor of ∼3 indicating a difference in barrier heights by less than 1 kcal/mol, which is within the reputed uncertainties of the classical force fields. However, in situations where the difference between τ_{off} along different paths is higher, such an analysis can also provide a thermodynamic insight into the unbinding process.

Manual Classification

	A	B	C	D	E	F	G	H
A	8		1					
B								
C	1	2	25					
D	1			18				2
D2					1			
E					3			
F+G						7	13	2
H				4				12

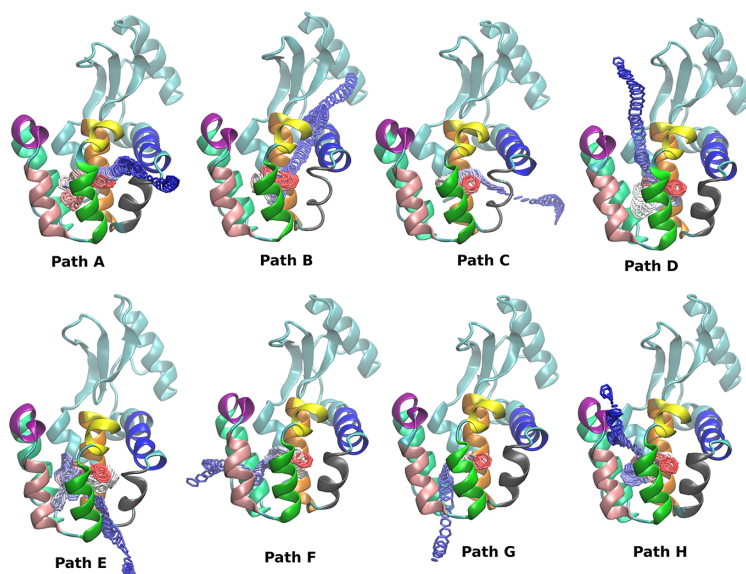| Class	Precision	Recall	F1 score
A	0.89	0.8	0.84
B	-	0	-
C	0.89	0.96	0.92
D	0.85	0.78	0.81
E	1.00	1.00	1
F+G	0.91	1.00	0.95
H	0.75	0.75	0.75

Fig. 6. The eight different pathways (*A–H*) of benzene unbinding from T4 lysozyme sampled using OPES$_E$ simulations. The benzene ring is colored according to the simulation time (red to blue in increasing order). The hydrogen atoms of benzene and the solvent molecules are omitted for clarity. The pathways shown in this figure are distinguished by visual inspection. In the *Right* panel, we show the confusion matrix between the manual classification and DTW$_D$ classification. The true positive cells are colored green. Path (*F*) and Path (*G*) are considered together as they were not distinguished by DTW$_D$.
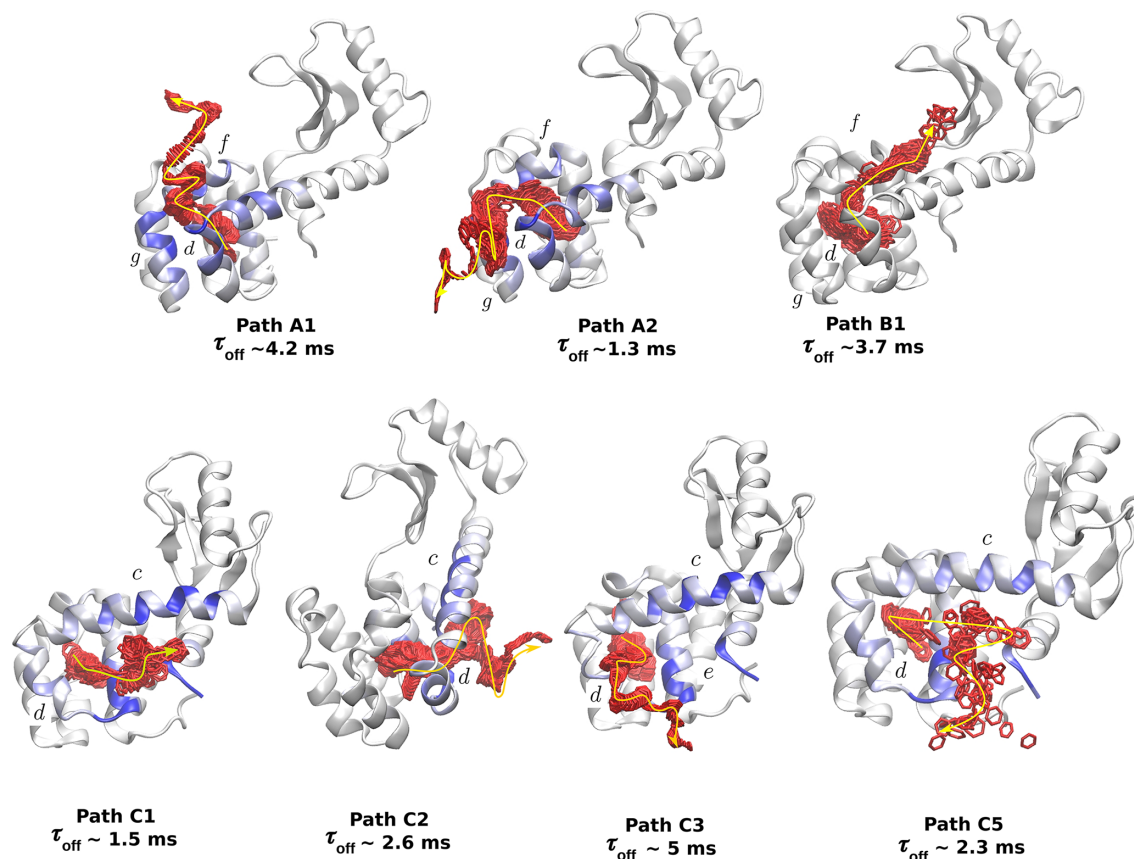
Fig. 7. The medoids of the different dissociation channels within the Path *A*, Path *B*, and Path *C* (*Text*) as identified by the DTW$_D$ analysis of the OPES$_f$ trajectories of the benzene-lysozyme complex. The unbinding timescales associated with each path cluster are denoted below each panel. The protein residues are colored according to their importance in distinguishing the path clusters within the same pathway (*Text*). Some of the helix labels are also included to aid visualization.

After an initial exploration it should also be possible to construct a path-specific collective variable (49) to focus the sampling along specific pathways to obtain a more detailed understanding of the ligand unbinding mechanism. This is a potential avenue for future research. We also envisage the possibility of reducing the computational complexity of our path classification protocol using deep neural networks following the suggestions of Seshan (50).

The present study on automated discovery and classification of ligand unbinding pathways will facilitate the incorporation of residence time and mechanistic descriptions of drug–target interactions within the realm of computer-aided drug discovery, a field that is heavily reliant on binding free energy calculations. This is an urgent necessity when limitations of free-energy-based screening are becoming increasingly apparent (4, 5). Our approach can facilitate a transition from the static free-energy-based screening of therapeutic candidates to the modeling of drug efficacy in the inherent out-of-equilibrium environment of the intracellular matrix. Moreover, the identification of protein residues that determine the ligand unbinding pathways, will help the pharmaceutical industry to adapt to the emergence of drug-resistant mutations (7, 8). Therefore, this work and follow-up studies in this area will be key steps toward potentially transforming the field of rational drug design.

Materials and Methods

Müller Brown Potential. Langevin dynamics simulations for the 2D Müller Brown potential are performed using the `ves_md_linearexpansion`

module of PLUMED v2.9, with a setup identical to ref. 51. Twenty OPES$_E$ simulations, with $\Delta E = 20\ k_B T$, were performed with either x or y as CV to sample transitions along the two pathways. Trajectories were stopped as soon as the system transited from the left to right minima. Trajectory data from the transition region ($x > -0.5$ or $y < 1.0$) were analyzed using DTW$_D$.

Alanine Dipeptide. Gas phase Alanine Dipeptide is modeled using AMBER99SB-ILDN force field and simulated at 300 K using the GROMACS v2021.5 package patched with PLUMED v2.9. All simulation parameters are identical to ref. 48. OPES$_E$ simulations are performed along the ϕ torsion angle with a barrier parameter of $\Delta E = 40$ kJ/mol to sample 100 transition events. Trajectories were terminated upon reaching C_{7ax} state from C_{7eq} state. The trajectory data of the 45-dimensional descriptor space (46) from the transition region ($\phi > -1.0$ rad) is analyzed using DTW$_D$.

Benzene T4 Lysozyme Complex. The topology and parameters for the benzene-bound L99A mutant of T4 Lysozyme were identical to ref. 31 and are obtained from PLUMED NEST repository with plumID:19.017. The protein is modeled with the AMBER ff14SB force field and the ligand has been parametrized using the Generalized Amber Force field (GAFF) with RESP charges computed at the HF/6-31G(d) level of theory. All simulation conditions are identical to ref. 31.

For the OPES$_E$ simulations, we used the same collective variables as Capelli et al. namely the spherical coordinates (ρ, θ, ϕ) of the center of mass of the ligand with respect to the center of mass of the binding pocket. The bias was deposited along all three components of the CV space. The barrier parameter was set to $\Delta E = 50$ kJ/mol. The simulations are terminated when the ligand reached $\rho > 2.5$ nm, which we designate as the unbound state.

For calculating kinetics, OPES$_f$ simulations (See details in *SI Appendix*) have been performed with a 2D CV space comprising of the spherical distance ρ as

well as the total number of contacts between the protein and the ligand defined as

$$c = \sum_{i \in A} \sum_{i \in B} \frac{1 - (r_{ij}/d)^6}{1 - (r_{ij}/d)^{12}}. \qquad [6]$$

Here A and B are the sets of nonhydrogen atoms of the protein and the ligand respectively and $d = 0.45$ nm. The barrier parameter was set to $\Delta E = 30$ kJ/mol and an excluded region is applied at $\rho > 0.9$ nm to avoid bias deposition in the transition state. To ensure the sampling of independent pathways, the simulation starting points were picked from a 20 ns unbiased simulation in the bound configuration. All enhanced sampling MD simulations were performed using GROMACS v2021.5 patched with PLUMED v2.9.

For each trajectory, the coordination number between the ligand heavy atoms and the C-α atoms of each of the 161 protein residues is stored at 1 ps interval. The contacts are evaluated using Eq. 6 using $d = 0.8$ nm. Due to the asymptotic behavior of the coordination number only the residues near the ligand contribute to the calculation, leading to the removal of noisy signals from distant residues and introducing sparsity into the DTW matrix **D**. This makes the coordination number an optimal descriptor for our pathway analysis. Only the portion of the trajectory with $\rho > 0.3$ nm and $c > 3$ was used for path classification to focus the emphasis on the transition region. The resulting 161-dimensional time series were subsequently analyzed using DTW$_D$.

It has been shown that at least 10 transitions are to be sampled to estimate the rates to gain statistical reliability (8). Seven out of 16 clusters, identified by our approach, contained 10 or more unbinding events. This allowed

us to calculate the ligand dissociation kinetics along these seven channels (Fig. 7).

Dynamic Time Warping. The DTW$_D$ distance (38) between all pair of trajectories were computed using the dtw_ndim module of the python package DTAIDistance (52). The pairwise distance matrix is used to perform a k-medoids clustering of the trajectories using the FasterPAM (53) algorithm employed in the kmedoids package (54) in Python. Silhouette scores (44) were also computed using the kmedoids code.

Data, Materials, and Software Availability. The simulation input files and analysis scripts are available from the GitHub repository: https://github.com/dhimanray/DTW_Path_Classification.git (55) and on the PLUMED NEST repository: plumID:23.030 (56). All dynamic time warping analysis has been performed using the open source package DTAIDistance which can be accessed at https://github.com/wannesm/dtaidistance.git (57). The raw data for dynamic time-warping clustering and manual classification are provided in *SI Appendix*.

Author affiliations: ªSimulations Research Line, Italian Institute of Technology, Via Enrico Melen 83, Genova GE 16152, Italy

1. M. De Vivo, M. Masetti, G. Bottegoni, A. Cavalli, Role of molecular dynamics and related methods in drug discovery. *J. Med. Chem.* **59**, 4035–4061 (2016).
2. W. Rocchia, M. Masetti, A. Cavalli, "Enhanced sampling methods in drug design" in *Physico-Chemical and Computational Approaches to Drug Discovery*, J. Luque, X. Barril, Eds. (The Royal Society of Chemistry, 2012), pp. 273–301.
3. P. C. Souza *et al.*, Protein-ligand binding with the coarse-grained martini model. *Nat. Commun.* **11**, 3714 (2020).
4. R. A. Copeland, D. L. Pompliano, T. D. Meek, Drug-target residence time and its implications for lead optimization. *Nat. Rev. Drug Discov.* **5**, 730–739 (2006).
5. R. A. Copeland, The drug-target residence time model: A 10-year retrospective. *Nat. Rev. Drug Discov.* **15**, 87–95 (2015).
6. D. Guo, T. Mulder-Krieger, A. P. IJzerman, L. H. Heitman, Functional efficacy of adenosine A 2A receptor agonists is positively correlated to their receptor residence time. *Br. J. Pharmacol.* **166**, 1846–1859 (2012).
7. A. Lyczek *et al.*, Mutation in Abl kinase with altered drug-binding kinetics indicates a novel mechanism of imatinib resistance. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2111451118 (2021).
8. M. Shekhar, Z. Smith, M. A. Seeliger, P. Tiwary, Protein flexibility and dissociation pathway differentiation can explain onset of resistance mutations in kinases. *Angew. Chem. Int. Ed.* **61**, e202200983 (2022).
9. R. E. Amaro, A. J. Mulholland, A community letter regarding sharing biomolecular simulation data for COVID-19. *J. Chem. Inf. Model.* **60**, 2653–2656 (2020).
10. R. O. Dror *et al.*, Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 13118–13123 (2011).
11. N. Ansari, V. Rizzi, M. Parrinello, Water regulates the residence time of Benzamidine in Trypsin. *Nat. Commun.* **13**, 1–9 (2022).
12. J. Rydzewski, W. Nowak, Machine learning based dimensionality reduction facilitates ligand diffusion paths assessment: A case of cytochrome P450cam. *J. Chem. Theory Comput.* **12**, 2110–2120 (2016).
13. L. Meng, F. K. Sheong, X. Zeng, L. Zhu, X. Huang, Path lumping: An efficient algorithm to identify metastable path channels for conformational dynamics of multi-body systems. *J. Chem. Phys.* **147** (2017).
14. Y. Qiu, M. S. O'Connor, M. Xue, B. Liu, X. Huang, An efficient path classification algorithm based on variational autoencoder to identify metastable path channels for complex conformational changes. *J. Chem. Theory Comput.* **19**, 4728–4742 (2023).
15. S. Bandyopadhyay, J. Mondal, A deep encoder-decoder framework for identifying distinct ligand binding pathways. *J. Chem. Phys.* **158**, 194103 (2023).
16. S. Bray, V. Tanzel, S. Wolf, Ligand unbinding pathway and mechanism analysis assisted by machine learning and graph methods. *J. Chem. Inf. Model.* **62**, 4591–4604 (2022).
17. S. Wolf, G. Stock, Targeted molecular dynamics calculations of free energy profiles using a nonequilibrium friction correction. *J. Chem. Theory Comput.* **14**, 6175–6182 (2018).
18. S. Wolf, B. Lickert, S. Bray, G. Stock, Multisecond ligand dissociation dynamics from atomistic simulations. *Nat. Commun.* **11**, 2918 (2020).
19. S. Wolf, M. Post, G. Stock, Path separation of dissipation-corrected targeted molecular dynamics simulations of protein-ligand unbinding. *J. Chem. Phys.* **158**, 124106 (2023).
20. S. Motta, L. Callea, L. Bonati, A. Pandini, PathDetect-SOM: A neural network approach for the identification of pathways in ligand binding simulations. *J. Chem. Theory Comput.* **18**, 1957–1968 (2022).
21. H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **26**, 43–49 (1978).
22. M. Müller, *Dynamic Time Warping* (Springer, Heidelberg, 2007), pp. 69–84.
23. H. Skutkova, M. Vitek, P. Babula, R. Kizek, I. Provaznik, Classification of genomic signals using dynamic time warping. *BMC Bioinform.* **14**, 1–7 (2013).
24. C. Tappert, C. Suen, T. Wakahara, The state of the art in online handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**, 787–808 (1990).
25. A. Kuzmanic, V. Zanchi, "Hand shape classification using DTW and LCSS as similarity measures for vision-based gesture recognition system" in *EUROCON 2007 -The International Conference on Computer as a Tool* (2007), pp. 264–269.
26. L. Mastroeni, A. Mazzoccoli, G. Quaresima, P. Vellucci, Decoupling and recoupling in the crude oil price benchmarks: An investigation of similarity patterns. *Energy Econ.* **94**, 105036 (2021).
27. G. Orlando, M. Bufalo, R. Stoop, Financial markets' deterministic aspects modeled by a low-dimensional equation. *Sci. Rep.* **12**, 1693 (2022).
28. G. Orlando, M. Bufalo, Modelling bursts and chaos regularization in credit risk with a deterministic nonlinear model. *Financ. Res. Lett.* **47**, 102599 (2022).
29. S. Aghabozorgi, A. S. Shirkhorshidi, T. Y. Wah, Time-series clustering-a decade review. *Inf. Syst.* **53**, 16–38 (2015).
30. A. Nunes-Alves, D. M. Zuckerman, G. M. Arantes, Escape of a small molecule from inside T4 lysozyme by multiple pathways. *Biophys. J.* **114**, 1058–1066 (2018).
31. R. Capelli, P. Carloni, M. Parrinello, Exhaustive search of ligand binding pathways via volume-based metadynamics. *J. Phys. Chem. Lett.* **10**, 3495–3499 (2019).
32. J. Rydzewski, O. Valsson, Finding multiple reaction pathways of ligand unbinding. *J. Chem. Phys.* **150**, 221101 (2019).
33. A. Nunes-Alves, D. B. Kokh, R. C. Wade, Ligand unbinding mechanisms and kinetics for T4 lysozyme mutants from τRAMD simulations. *Curr. Res. Struct. Biol.* **3**, 106–111 (2021).
34. J. Mondal, N. Ahalawat, S. Pandit, L. E. Kay, P. Vallurupalli, Atomic resolution mechanism of ligand binding to a solvent inaccessible cavity in T4 lysozyme. *PLoS Comput. Biol.* **14**, e1006180 (2018).
35. Z. Smith, D. Pramanik, S. T. Tsai, P. Tiwary, Multi-dimensional spectral gap optimization of order parameters (SGOOP) through conditional probability factorization. *J. Chem. Phys.* **149**, 234105 (2018).
36. M. Fajer, Y. Meng, B. Roux, The activation of c-Src tyrosine kinase: Conformational transition pathway and free energy landscape. *J. Phys. Chem. B* **121**, 3352–3363 (2017).
37. Y. Tao *et al.*, A comparative analysis of trajectory similarity measures. *GIsci. Remote Sens.* **58**, 643–669 (2021).
38. M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, E. Keogh, Generalizing DTW to the multi-dimensional case requires an adaptive approach. *Data Min. Knowl. Disc.* **31**, 1–31 (2017).
39. V. Hautamaki, P. Nykanen, P. Franti, "Time-series clustering by approximate prototypes" in *2008 19th International Conference on Pattern Recognition* (IEEE, Tampa, FL, USA, 2008), pp. 1–4.
40. L. Kaufman, P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis* (John Wiley & Sons, 2009).
41. T. Liao *et al.*, "Understanding and projecting the battle state" in *23rd Army Science Conference, Orlando, FL* (2002), vol. 25.
42. T. W. Liao, C. F. Ting, P. C. Chang, An adaptive genetic clustering method for exploratory mining of feature vector and time series data. *Int. J. Prod. Res.* **44**, 2731–2748 (2006).
43. V. Vuori, J. Laaksonen, "A comparison of techniques for automatic clustering of handwritten characters" in *2002 International Conference on Pattern Recognition* (IEEE, Quebec City, QC, Canada, 2002), vol. 3, pp. 168–171.
44. P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
45. M. Invernizzi, M. Parrinello, Exploration vs. convergence speed in adaptive-bias enhanced sampling. *J. Chem. Theory Comput.* **18**, 3988–3996 (2022).
46. L. Bonati, V. Rizzi, M. Parrinello, Data-driven collective variables for enhanced sampling. *J. Phys. Chem. Lett.* **11**, 2998–3004 (2020).

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

48. D. Ray, N. Ansari, V. Rizzi, M. Invernizzi, M. Parrinello, Rare event kinetics from adaptive bias enhanced sampling. *J. Chem. Theory Comput.* **18**, 6500–6509 (2022).

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

50. A. Seshan, Using machine learning to augment dynamic time warping based signal classification. arXiv [Preprint] (2022). http://arxiv.org/abs/2206.07200 (Accessed 14 June 2022).

51. D. Ray, E. Trizio, M. Parrinello, Deep learning collective variables from transition path ensemble. *J. Chem. Phys.* **158**, 204102 (2023).

52. W. Meert *et al.*, Dtaidistance (2020).

53. E. Schubert, P. J. Rousseeuw, Fast and eager K-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. *Inf. Syst.* **101**, 101804 (2021).

54. E. Schubert, L. Lenssen, Fast K-medoids clustering in rust and python. *J. Open Source Softw.* **7**, 4183 (2022).

55. D. Ray, DTW_Path_Classification. GitHub. https://github.com/dhimanray/DTW_Path_Classification. Deposited 4 August 2023.

56. D. Ray, Data Driven Classification of Ligand Unbinding Pathways. PLUMED NEST. https://www.plumed-nest.org/eggs/23/030/. Deposited 4 August 2023.

57. W. Meert *et al.*, DTAIDistance (Version v2). Zenodo. http://doi.org/10.5281/zenodo.5901139. Deposited 25 January 2022.