JCTC Journal of Chemical Theory and Computation
Article
pubs.acs.org/JCTC

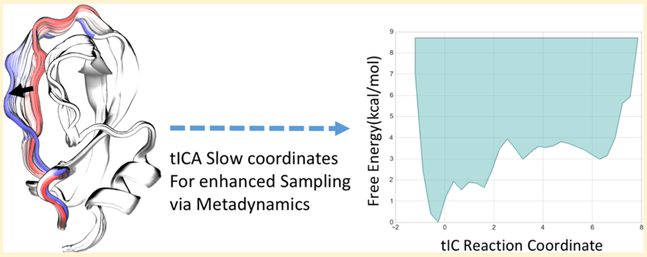
tICA-Metadynamics: Accelerating Metadynamics by Using Kinetically Selected Collective Variables

Mohammad M. Sultan and Vijay S. Pande*

Department of Chemistry, Stanford University, 318 Campus Drive, Stanford, California 94305, United States

Supporting Information

ABSTRACT: Metadynamics is a powerful enhanced molecular dynamics sampling method that accelerates simulations by adding history-dependent multidimensional Gaussians along selective collective variables (CVs). In practice, choosing a small number of slow CVs remains challenging due to the inherent high dimensionality of biophysical systems. Here we show that time-structure based independent component analysis (tICA), a recent advance in Markov state model literature, can be used to identify a set of variationally optimal slow coordinates for use as CVs for Metadynamics. We show that linear and nonlinear tICA-Metadynamics can complement existing MD studies by explicitly sampling the system's slowest modes and can even drive transitions along the slowest modes even when no such transitions are observed in unbiased simulations.



INTRODUCTION

Molecular dynamics (MD) is a powerful computational method used to explore protein free energy landscapes that provides structural, kinetic, and thermodynamic insights into complex biophysical processes.[1] However, due to the millisecond or greater time scales of biological processes, MD often, if not always, suffers from lack of convergence issues. While algorithmic improvements, use of smaller trajectories,[2] dedicated hardware,[3,4] and GPU[5,6] technologies have alleviated the problem, routine access to biologically interesting processes and time scales remains prohibitive to all but a few.[3,7]

To improve efficiency, researchers have developed a variety of enhanced sampling methods. Metadynamics[8−12] is one such method that "pushes" a system out of free energy basins by adding positive Gaussian potentials at local minima in the trajectory history. Metadynamics has been successfully used to understand various biophysical[13] processes and can be thought of as computational sand filling.[14] However, Metadynamics requires the identification of molecular descriptors or collective variables (CVs), and it is not immediately clear how the CVs should be chosen such that they are representative of a protein's complex state space.

Researchers have also leveraged information from hundreds of short (10−100s of ns) unbiased MD trajectories by turning to Markov state models[2] (MSMs), which parametrize a kinetic master equation whose spectral decomposition provides insight into the slow dynamics of a system. MSMs require only local equilibrium within the dynamics of each state, enabling trivial parallelization of the MD simulations. Currently, MSM construction involves partitioning of the proteins' accessible phase space into a set of nonoverlapping states and counting the transitions between them. The states can be kinetically defined along a reduced set of coordinates obtained via the tICA[15,16] method. tICA is a dimensionality reduction method that seeks to embed MD data into a lower dimensional subspace while minimizing the loss of kinetic information. It does this by finding the slowest variables within the data set.

In this paper, we propose combining ideas from Metadynamics with current MSM methods. In particular, we leverage the tICA[15−18] method to find slowly decorrelating variables for direct use as CVs in Metadynamics. The use of tICA eigenvectors (tICs) can be used either to complement unbiased MD sampling or to drive trajectories along their slowest reaction coordinates. To this end, we first restate what properties a "good" set of CVs poses. We then show that these properties are satisfied by the dominant eigenvectors of the transfer operator, which is approximated in various forms via both tICA[15−18] and MSMs. We conclude by demonstrating results for model systems.

MSMs have also been recently employed to improve convergence[19] of umbrella sampling[20] calculations, and using machine learning (ML) in combination with MD trajectories to build CVs is not new.[21−24] Previous work involved using Principal component analysis (PCA), Isomaps, and Sketch maps. These are dimensionality reduction techniques that seek to embed high dimensional MD data into lower dimensional subspaces by minimizing certain objective functions. The lower dimensional coordinates arguably provide superior CVs because they inherently include more information regarding, for example, in the case of PCA, large scale amplitude motion. However, in contrast to tICA, these ML methods neither use

Received: February 20, 2017
Published: April 6, 2017

DOI: 10.1021/acs.jctc.7b00182
J. Chem. Theory Comput. 2017, 13, 2440−2447

the time information contained in trajectories, nor are they capable of finding slow coordinates directly from the data. Furthermore, solutions employing differing internal parameters are not necessarily comparable to one another. For example, are the principal components/normal modes built using the dihedrals of a system *better* capable of describing the slow dynamics than those built from contact distances? Why or how are such dimensionality reductions capable of describing the system's slow dynamics? In contrast, a recently developed variational bound for tICA and MSM analysis[25] in combination with statistical cross-validation[26] allow modelers to quantitatively compare the goodness of tICA and MSMs across a MD trajectory data set, potentially integrating out a large set of modeling choices.

■ TICA SOLUTIONS PROVIDE SLOW CVS FOR METADYNAMICS

Although any well-defined single- or multidimensional CV has an associated free energy function, the CVs[14] for any enhanced sampling algorithm should ideally correlate with the slow dynamics of the systems. In other words, CVs are expected to be capable of identifying the transitions between metastable states in full phase space. In Abrahms et al.[9] and Laio et al.,[14] it was argued that a good set of CVs should represent a dimensionality reduction from the full phase space Ω to a small set of $\mathbb{R}$numbers. CVs should also be able to describe all the relevant slow events as well as distinguish start, end, and intermediate states.

In McGibbon et al.,[18] these properties were codified under the definition of a *natural* reaction *coordinate*. They argued that a *natural* reaction *coordinate* a) is a dimensionality reduction b) is learned from the dynamics and c) is maximally predictive of future evolution. They then showed that these properties are satisfied by the leading eigenfunctions (ψ) of the system's Markovian dynamics in Ω. The first eigenfunction corresponds to the equilibrium distribution and is not relevant to our discussion. However, the second eigenfunction (ψ_2) is the system's most "slowly decorrelating collective variable" and is maximally predictive of its future dynamics. This formalism is thus well suited to automatically choosing slowly decorrelating CVs for Metadyanamics, and all that remains are finding ways to approximate the leading eigenfunctions/eigenvectors. MSM and tICA are two methods that seek to approximate these eigenvectors, albeit with a differing choice of basis.

tICA[15−17] is a kinetically motivated technique designed to find projections in the data that minimize loss of kinetic information by maximizing the autocorrelation function. Recall that high autocorrelation values imply slower structural degrees of freedom. In tICA,[15,16] we seek to approximate ψ_2 by finding a linear combination of input structural order parameters that decorrelate the slowest at a particular lag time. These order parameters could be all of the backbone or side chain dihedrals or contact distances, etc. The complete derivation for the tICA formalism is outside the scope of this paper, but we briefly describe it using the notation from ref 15. For a more complete derivation including tICA's connection to Transfer operator, we refer the readers to previous work.[18,25,26] Let $\{\chi_t^d\}_{t=0}^{N_t-1}$ be a multidimensional MD time series where each frame is represented via a column of d-dimensions for a system. We further assume that the data has a mean of 0. Then, tICA seeks to maximize the following function

$$f(|a_1\rangle) = \frac{\mathbb{E}[\langle a_1|\chi_t\rangle\langle a_1|\chi_{t+\Delta t}\rangle]}{\mathbb{E}[\langle a_1|\chi_t\rangle\langle a_1|\chi_t\rangle]}$$

where the maximization is over the autocorrelation function of the projection of the data onto a_1. By taking advantage of inner product symmetries, it can be then shown that

$$f(|a_1\rangle) = \frac{\langle a_1|C(\Delta t)|a_1\rangle}{\langle a_1|\Sigma|a_1\rangle}$$

Here, the symmetric matrix $C(\Delta t)$ is the time lagged correlation between the basis functions

$$C_{ij}(\Delta t) = \mathbb{E}[\chi_t^i \cdot \chi_{t+\Delta t}^j]$$

and $\sum$ is positive definite covariance matrix of the basis functions.

$$\Sigma_{ij} = \mathbb{E}[\chi_t^i \cdot \chi_t^j]$$

After applying unit variance constraints and setting up an optimization problem solvable via Langrage multipliers, it can be shown that a_1 (which is also the closest linear approximation to ψ_2)[18] is the solution to the generalized eigenvalue problem

$$C(\Delta t)|a_1\rangle = \lambda_1 \Sigma |a_1\rangle$$

Subsequent projections, $a_2 ... a_d$, can be similarly found by requiring that the a_n solution be uncorrelated with all solutions before it. An attractive property of the above formalism is that it generalizes to multiple orthogonal reaction coordinates (tICs), ordered by decreasing time scales.[15,26] This quality can be harnessed to prevent slow convergence due to "hidden" orthogonal degrees of freedom by simultaneously sampling multiple tICs individually and connecting them via replica exchange.[9,11] This means that if we accelerate the top "*m*" tICs for X-ns a piece such that X is larger than the time scale of the next (*m*+1th) tICs, then the system will naturally equilibrate and sample those additional degrees of freedom.

In practice, finding a set of slowest tICA components (tICs) requires the following steps.[15,16]

1. Run some MD simulations starting from all available crystal structures or homology models.

2. Pick a set of relevant structural order parameters (such as all the dihedrals or contacts or some combination thereof) to represent each frame of the trajectory. The choice of features remains but tICA can easily handle hundreds to thousands of order parameters, and we can pick the best set of features conditioned on available data via cross-validation.[26]

3. After picking a tICA lag time Δ*t*, compute the symmetric matrix $C(\Delta t)$ and positive definite matrix $\sum$. Heuristically, we have found that tICA is robust to choices in the 1−100 ns range, though this parameter can be automatically chosen.[26]

4. Solve $C(\Delta t) \cdot a = \lambda \sum \cdot a$.

5. Pick the top few eigenvectors (tICs) with the largest eigenvalues.

Previously, the tICA method has been used as an intermediate dimensionality reduction step in MSM construction.[15,16] Here, we add a final step to the recipe highlighted above.

6. Use Metadynamics to directly sample the slow degrees of freedom by using the tICs as CVs.

This is akin to using the principal components (PCs)[21,24] but is more advantageous because we do not implicitly assume that the slow degrees of freedom correlate with high variance.[15,24]

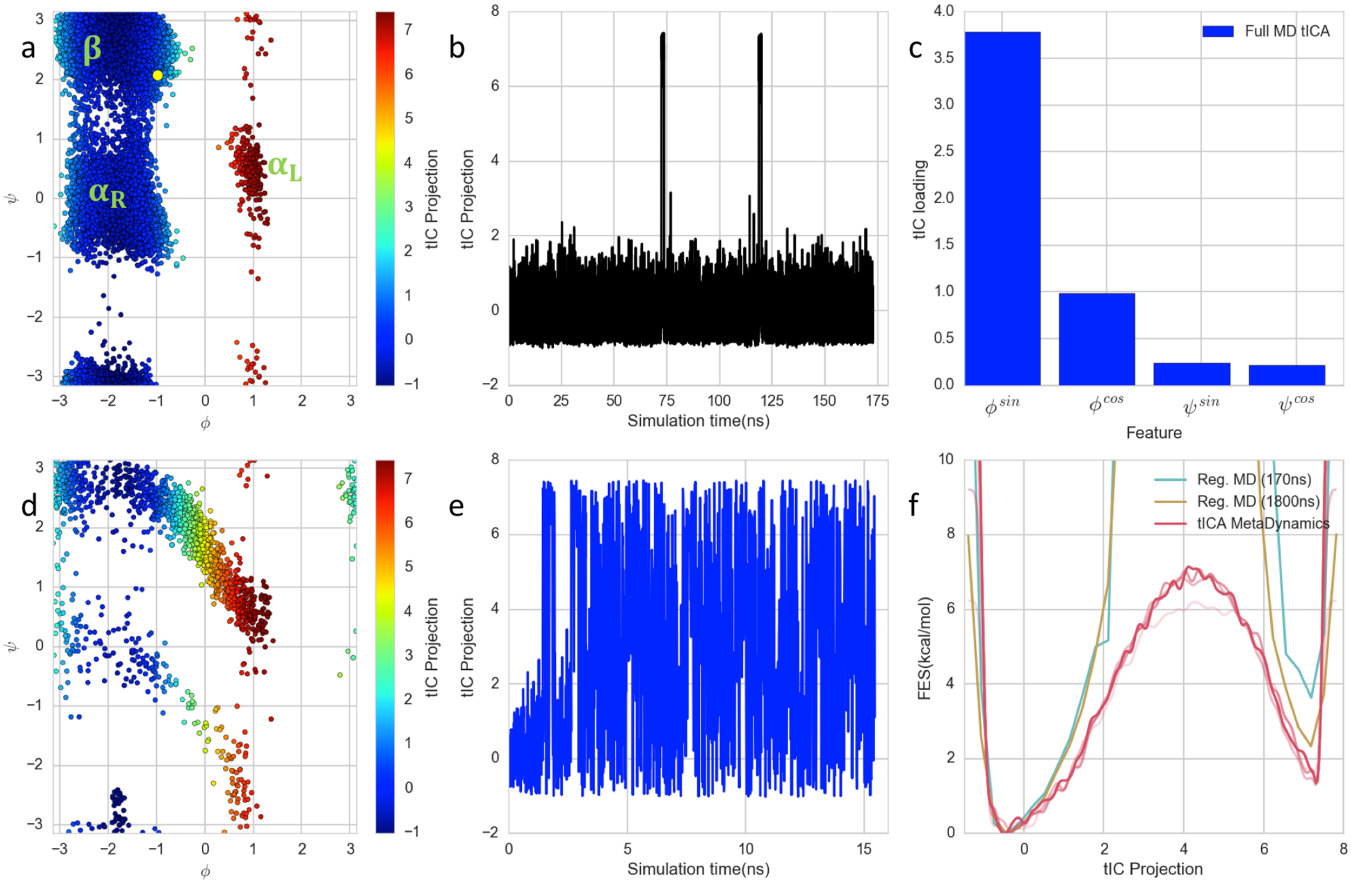


Figure 1. tICA-Metadynamics for alanine dipeptide helps to converge simulations where only a few slow transitions have been observed. Clockwise from top a). Ramachandran plot of the long trajectory colored according to the projection of each frame unto the dominant tIC (arbitrary units). The dominant tIC corresponds to movement along both (ψ) and (ϕ). The large yellow circle represents the starting frame for all simulations in this paper. b). Projection of the trajectory along the dominant tIC showing two slow transitions at ~75 and ~120 ns mark. c). The dominant tIC broken down according to its loadings on each input feature. Higher values indicate greater contribution to the slowest coordinate. See SI Figure 1 for a comparison to PCA. d). 15 ns Metadynamics run projected using (a). e). Projection of the Metadynamics simulation unto the tIC shows fast mixing. f). Comparison of the protein's thermodynamics from the MD (170 and 1800 ns) vs tICA-Metadynamics approach. The lighter red curves indicate estimate of the free-energy surface after every 2000 Gaussians.

This might happen in situations where the high variance modes might be movements within free-energy basins (see SI Figure 1 for a comparison of tICA vs PCA on alanine) rather than between them. We also note that our method is similar in spirit to the spectral gap maximization work of Tiwary et al.,[27] though we propose a simpler CV optimization strategy that can handle larger and more diverse feature inputs with realistic extensions to multiple nonlinear modes via the landmark approximation to the kernel trick[28] and Bias-Exchange.[9]

■ TICA-METADYNAMICS ON ALANINE DIPEPTIDE LEADS TO FAST MIXING ALONG THE SLOWEST MODE

We end this paper by providing simple examples with solvated alanine dipeptide and bovine pancreatic trypsin inhibitor (BPTI). Alanine dipeptide has been previously used as an interesting model system in Metadynamics because the choice of faster CV (ψ) leads to hysteresis and slow simulation convergence. Here, we show that not only the dominant tIC is capable of finding a slow coordinate but also it does so in an unbiased fashion.

All MD trajectories were generated in the NPT ensemble with a MonteCarlo Barostat (1 atm), a Langevin integrator (300 K), and a 2 fs time step. Long-range electrostatics were modeled using the PME[29] method. Protein dynamics were modeled using Amber99sb-ildn[30] with the Tip3P water model.[31] All simulations, including the Metadynamics runs, were started from the same initial coordinates and velocities and were carried out on GPUs using OpenMM[5] and Plumed.[32] For all Metadynamics runs, we limited sampling to only the dominant (slowest) tIC unless specified otherwise. The Gaussians were dropped every 2 ps with a height of 0.2 kJ/mol and a width of 0.1 (arb. tIC units). We also ran other trajectories where we increased the width and height of the Gaussians but found the results to be comparable. All tICA models were built using the sin-cosine transform of the backbone phi and psi dihedral angles at a tICA lag time of 1 ns. All MSMs were built at a MSM lag time of 1 ns after clustering the data to 50 states along the dominant tIC. We used MDTraj (version 1.8)[33] and MSMBuilder3 (version 3.6.1)[34] to build all tICA, cluster, and Markov state models. We used pyEMMA (version 2.2.7)[35] for the MBAR[36] and WHAM[37] calculations.

We began by generating a single 170 ns MD trajectory and building a tICA model to identify the slow dynamics within the system (Figure 1 a-c). Our results indicate that the slowest transition happens on the order of 50−75 ns and corresponds to the movement from the β/α_R basins to the α_L basin (Figure 1a). This movement involves transitions along both the faster ψ

DOI: 10.1021/acs.jctc.7b00182
J. Chem. Theory Comput. 2017, 13, 2440−2447

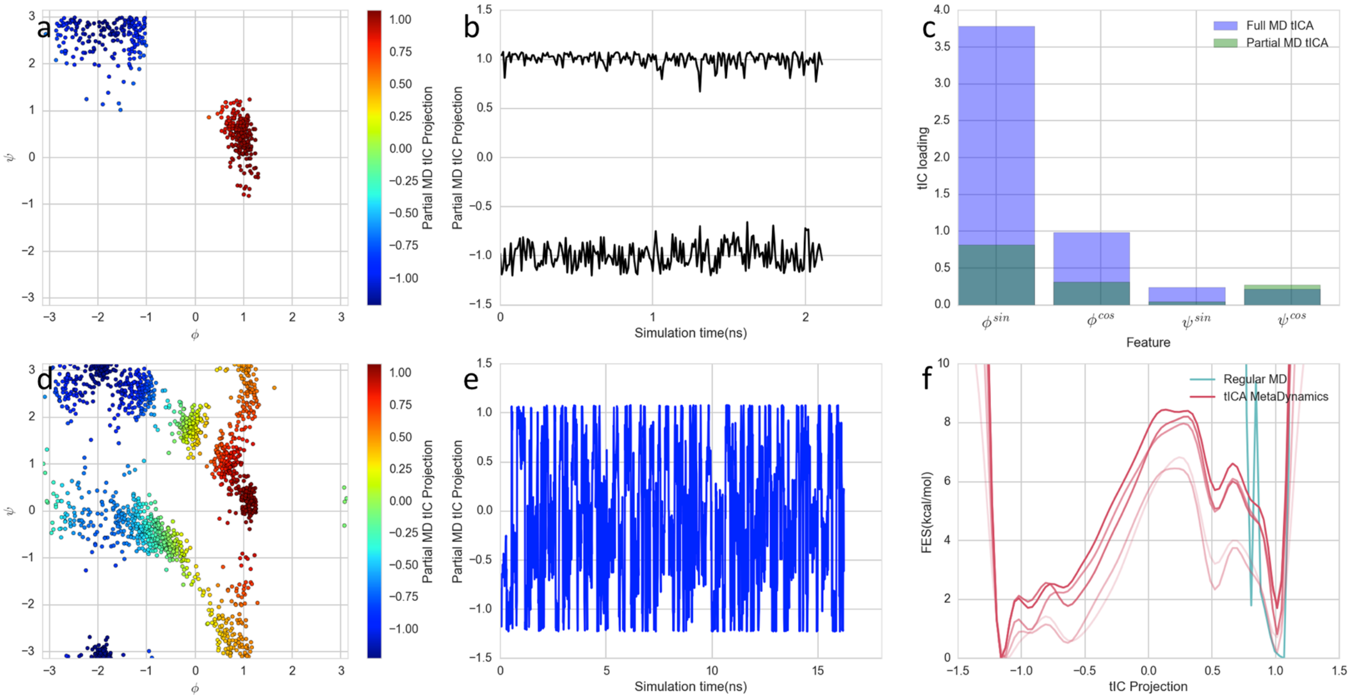


Figure 2. tICA-Metadynamics can drive alanine dipeptide simulations along slow coordinates even when no transitions are observed in the tICA training data. Clockwise from top a). Ramachandran plot of the subsections of the trajectory from Figure 1a colored according to the projection of each frame unto a new tICA model trained on the sliced trajectory data. The dominant tIC corresponds to movement along both (ψ) and (ϕ) b). Projection of the trajectory along the dominant tIC showing no forward or backward transitions in the 4 ns of simulation. c). The dominant tIC broken down according to its loadings on each input feature. d). 15 ns Metadynamics run projected using (a). e). Projection of the Metadynamics simulation onto the partial MD tIC coordinate shows fast mixing. f). Comparison of the protein's thermodynamics from the partial MD vs tICA-Metadynamics approach. The lighter red curves indicate estimate of the free-energy surface after every 2000 Gaussians. We explicitly note that this is a new tICA model and should not be numerically compared to the model from Figure 1.

and slower ϕ coordinate (Figure 1c). Within the unbiased MD simulation, we observed just two such transitions (Figure 1b), leading to large statistical uncertainties in the thermodynamics (Figure 1f, cyan curve). To complement this unbiased MD simulation, we sampled the tIC (starting from the same initial coordinates and velocities) using Metadynamics. Within the course of 15 ns, we were able to witness the same transition on the order of tens of times, leading to robust free energy estimates for the two basins (α_L vs α_R and β). For example, our simulation visited, defined as getting to a basin and continuously staying there for at least one frame (10 ps), the α_L basin (tIC value >5) 185 times over the course of the 15 ns in contrast to two times over 170 ns of traditional MD, presenting a potential speed up of close to a 1000×. It is worth noting that when we extended our regular MD simulations to an aggregate of 1800 ns (Figure 1f, gold curve) that the free energy difference between MSM and tICA-Metadynamics drops to less than 1 kcal/mol, indicating synergistic opportunities between the techniques.

TICA-METADYNAMICS ON ALANINE DIPEPTIDE CAN DRIVE TRANSITIONS ALONG THE SLOWEST MODE

We next investigated whether we could use the tICA-Metadynamics method when no slow transitions occur in the data, but a tICA model is able to partially approximate the true time lagged correlation matrix and covariance matrix from simulations. This is reasonable because only a few data points are lost when we do not include the rare crossover region in the correlation calculation. We also note that this is a common problem for MD simulations in which the modeler has access to multiple starting crystal structures, but the free energy differences, the lowest free energy path connecting those states, or the knowledge of intermediate or off-pathway states remains elusive. For instance, in the case of alanine dipeptide, this would refer to the transition between β/α_R to the α_L basin (Figure 1a-c) and is dominated by the slowly evolving ϕ dihedral.

To test our hypothesis, we sliced two ~2 ns (2.5% of original 170 ns) worth of trajectories from the regular long simulation, taking care to prevent any crossovers between the β and α_L basins (Figure 2 a−c). We next trained a *new* tICA model (Figure 2c, partial MD tICA). Again, the dominant tIC corresponded to transitions in and out of the α_L basin, though the associated time scale is now meaningless. More importantly, when we performed Metadynamics along this coordinate, our simulation was able to cross from β to the α_L basin. Due to the length of the simulations, our trajectories naturally discovered the α_R basin which is separated from the β region by a much smaller barrier. We obtained 10s of crossovers along the dominant mode in ~15 ns of simulation. Again, for example, our simulation visited the α_L basin (defined as tIC value >0.6) 122 times over the course of the 15 ns in contrast to zero times in the 4 ns of training data.

TICA-METADYNAMICS CAN BE EXTENDED TO MULTIPLE COORDINATES BY EXPLOITING NONLINEARITY AND BIAS-EXCHANGE

It is possible that for larger more complex biophysical systems, no single linear combination of input features can provide a good approximation to the system's slowest dynamics. This is a

DOI: 10.1021/acs.jctc.7b00182
J. Chem. Theory Comput. 2017, 13, 2440−2447

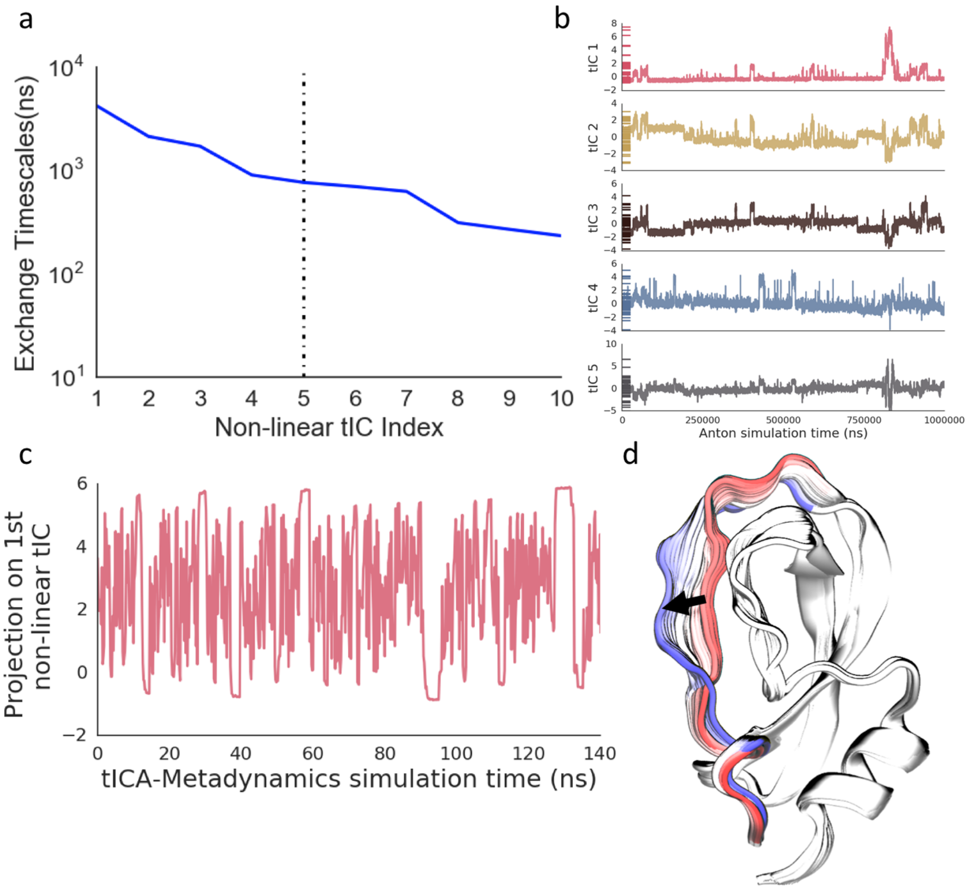


Figure 3. Accelerating BPTI along its slow modes enables efficient sampling of the opening and closing of the protein's core. a). Inferred exchange time scales for the nonlinear tICA model. We limited Metadynamics sampling to the top 5 modes (dashed line). b). Projection of the 1 ms Anton trajectory unto the top 5 modes. The horizontal marks at the start correspond to the projection of the 50 pdbs that served as the landmark coordinates. The opening of the protein core corresponds to the large scale motion at about 800 μs mark. c). Projection of 140 ns of one of the Metadynamics trajectories showing that this motion can be repeatedly sampled by employing nonlinear tICA and bias exchange. d). Pictorial representation of the first tIC primarily corresponding to outward movement of the N-terminal loop. The protein image was generated using VMD,[44] the secondary structure was assigned using STRIDE,[45] and graphs were made using Matplotlib,[46] IPython,[47] MSMBuilder, and MSMExplorer.[48]

generic problem within most CV based enhanced sampling algorithms and is usually solved by using either alternative and/or multiple CVs. In contrast, we propose using the kernel trick[28,38] to significantly improve tICA's ability to estimate the eigen functions of the transfer operator.[28]

Kernel methods[24,28,39] are algorithms from ML literature. In these methods, the input features are implicitly expanded onto a higher dimensional space. This is done by computing the inner product between all pairs of data. In our case, this nonlinear transform is followed by linear tICA, making our final solution linear combinations of nonlinear functions. The most commonly used kernel is the Gaussian kernel

$$k(x, y) = e^{-\|x-y\|^2/2\sigma^2}$$

where x and y are two protein conformations, and σ is a user parameter that is dependent on the chosen distance metric. Thus, the slowest coordinate in our system is the linear combination of exponential distances that decorrelates the slowest. Here, the exponential function implicitly encodes nonlinear behavior. The full kernel operation would require us to compute the distance of any newly generated conformation to *all* of the previously generated training data points. Empirically, we found that the use of the complete kernel-tICA solution slowed down our Metadynamics simulations too much, and we thus were forced to use the landmark approximation. In it, the kernel $k(x,y)$ is only computed for a set of preselected landmark conformations in the unbiased MD data set.[40]

To test out the landmark kernel-tICA-Metadynamics simulation protocol, we chose to look at the millisecond long simulation of 58-residue protein BPTI by Shaw et al.[41] performed on the Anton machine.[3] In their simulation, BPTI remains stable over the course of the 1 ms simulation save for a large scale conformational change occurring about 800 μs into the simulation. This high free energy state corresponds to the opening and hydrating of the protein's core. We tested if landmark-kernel-tICA could find a coordinate along which Metadynamics could accelerate this opening and closing. As a side note, we also tested whether tICA could find an approximation to the dominant tIC coordinate without the need for full sampling by only building the models using data from the open and closed states (SI Figure 2), finding that tICA could distinguish between the basins using a variety of input features.

For our experiments on BPTI, the original Anton trajectories were subsampled to 25 ns for faster analysis. We used 50 landmark points (chosen by picking the centroids of a 50 state

DOI: 10.1021/acs.jctc.7b00182

clustering model) combined with the all atom RMSD distance metric and a σ value of 0.3 nm. It is worth noting here that the all atom RMSD landmark kernel tICA solution operates in the full-phase space and is free from many of the problems that plague traditionally selected CVs. Similar to previous work,[42] our tICA experiments with other tICA parameters and distance metrics led to the discovery of kinetically disconnected data sets, indicating the need for validating the tICA model before performing tICA-Metadynamics. We believe this process could be a force field or sampling artifact and chose to focus on the results of the all atom RMSD metric that correctly described the opening of the protein core. The tICA model was built using all of the available data at a lag time of 100 ns (4 frames) and using the kinetic mapping[17] scaling. We projected the data onto the top five slow modes (Figure 3a-b).

The results of projecting the entire 1 ms Anton simulation onto the five tICs are shown in Figure 3b. The nonlinear landmark kernel tICA model captures the large scale conformational change (Figure 3d) in the protein (low to high values in Figure 3b, top panel). We decided to accelerate the sampling of this process by performing Bias-Exchange[9] well-tempered Metadynamics along the five tICA coordinates. In well-tempered Metadynamics,[10] the Gaussian heights are scaled as the simulation progresses leading to a smoothly converged free energy profile. In Bias-Exchange[9,11,43] Metadynamics, several coordinates are accelerated at once, and at set simulation times the coordinates are swapped according to a Monte Carlo criterion.

All of our BPTI simulations were performed in the NPT ensemble using the Amber99sb-ildn[30] protein model in conjunction with the TIP3P[31] water model. We ran each replica for ∼140 ns attempting swaps every 6 ps. We set the Gaussian drop rate to once per 2 ps, the Gaussian height was set to 1 kJ/mol, the width was set to 0.2, and a bias factor consisted of 1000. As shown in the results for the first tIC-Metadynamics replica (Figure 3c), we were able to accelerate BPTI's opening and closing rather significantly, capturing many core opening events.

It is possible that for more complex systems kernel-tICA proves insufficient. For example, for BPTI, there are still several tIC coordinates whose exchange time scales are on the order of hundreds of nanoseconds (Figure 3a, indices >5). It is likely given the length of our trajectories that our simulations did not adequately sample those dimensions (SI Figure 3), and we will need to further extend the Metadynamics simulations. In such scenarios, we recommend accelerating the system along *all* tICA coordinates whose inferred time scales are greater than the simulation length of any single replica. These trajectories can then be reweighted unto full state space using the weighted histogram technique (WHAM)[43,49] or multistate Bennett Acceptance Ratio (MBAR).[36] This is highly desirable in situations where one would like to estimate the free energies along coordinates that were not biased in the simulation. The SI contains an example of alanine dipeptide where we used nonlinear tICA in conjunction with Bias-Exchange well-tempered Metadynamics to accelerate the top two slow modes, reweighted the trajectories unto full phase space using MBAR, and compared the results against several microseconds of regular MD.

To conclude, we propose using the tICA method from the MSM literature as an unbiased way for discovering Metadynamics CVs (tICA-Metadynamics). This is superior to ML-based or heuristic engineering because tICA, by construction, attempts to find the slowest decorrelating combination of input structural order parameters. We further show that even in the low data regime, tICA's approximate solution can provide CVs for Metadynamics.

It is worth noting some limitations of tICA-Metadynamics. In the examples where we would like to complement the MSM's estimates of thermodynamics and kinetics,[2] we require unbiased sampling of the slow transitions via MD. This might require hundreds of microseconds of aggregate MD sampling, making it unfeasible for all but the smallest systems. In the incomplete MD case, where we would like to drive the transition along the slowest tICs, we assume that enough MD is run starting from high and low free energy states that the tICA model is able to approximate the time lagged correlation function. Depending on the system, this might be very difficult; for example, the proper rank ordering of the set of slowest coordinates might not be possible. In both cases, the choice of the number of tICA coordinates to accelerate remains. While the kernel extension and simultaneously accelerating along multiple coordinates can greatly attenuate the situation, convergence will likely require sampling of all tICA coordinates whose exchange time scales are slower than the individual simulation lengths and coupled to one another via a replica scheme. Lastly, we inherently assume that the tICs are biologically meaningful and not an artifact of improper choice of basis, sampling, and/or force field inaccuracies.[18]

Before we end, we note that while we have limited our results to improving CV selection for Metadynamics simulations, it is possible to use this method to find better CVs for other enhanced sampling methods as well. For example, it is entirely possible to directly sample these tICs using Umbrella sampling[19,20] or its variants. It could also be possible to couple the tICs to differing thermostats similar to temperature-accelerated MD.[50,51] The latter method could potentially allow for the use of a larger number of tICs though reconstructing the exact free energy surface in that case is an unsolved problem. We also note that while we have used two of the main variants of tICA, other approximations and improvements[17,18] exist that might be better suited for tICA-Metadynamics.

Importantly, the use of leading eigenvectors of the transfer operator as CVs opens up interesting avenues for further research. Could we potentially use the full MSM state-space as the driving force for Metadynamics, allowing us to integrate out the arduous problem of selecting CVs completely? The kernel tICA example partially solves this problem but suffers from the need for selection of representative landmark coordinates. On the Metadynamics side, could having partial or complete knowledge of the MSM free energy landscape help to accelerate convergence or reduce the number of researcher degrees of freedom such as the Gaussian width, shape, and drop rate? Could we potentially use the tICA coordinates from coarse grained simulations for atomistic simulations or vice versa? We anticipate that these questions could provide interesting avenues for future research into directly coupling reaction coordinate based methods with enhanced sampling algorithms.

■ DATA AVAILABILITY

All the data, models, and scripts needed to reproduce the main results of this paper are freely available at https://github.com/msultan/tica_metadynamics_paper_1

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b00182.

> Comparisons of tICA to the PCA model for alanine dipeptide (SI Figure 1), the calculated slow tICA coordinates of BPTI using a variety of different feature sets (SI Figures 2 and 3), and MSM thermodynamics comparison to tICA-metadynamics thermodynamics after reweighting via MBAR (SI Figure 4) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Pande@stanford.edu.

ORCID

Mohammad M. Sultan: 0000-0001-5578-6328

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank Ariana Peck and Brooke E. Husic for a critical reading of the manuscript and providing valuable feedback. M.M.S. acknowledges support from NSF-MCB-0954714. The Pande Group is broadly supported by grants from the NIH (R01 GM062868 and U19 AI109662) as well as gift funds and contributions from Folding@home donors. We thank D.E. Shaw research for generously providing us with the 1 ms BPTI trajectory.

■ REFERENCES

(1) Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* **2002**, *9* (9), 646−652.

(2) Bowman, G. R.; Pande, V. S.; Noé, F. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*; 2014; Vol. 797, DOI: 10.1007/978-94-007-7606-7.

(3) Shaw, D. E.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossváry, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Deneroff, M. M.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. In *Proceedings of the 34th annual international symposium on Computer architecture - ISCA '07*; ACM Press: New York, New York, USA, 2007; Vol. 35, p 1.

(4) Shaw, D. E.; Grossman, J. P.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H.; Forte, A.; Gagliardo, J.; Gill, G.; Greskamp, B.; Ho, C. R.; Ierardi, D. J.; Iserovich, L.; Kuskin, J. S.; Larson, R. H.; Layman, T.; Lee, L.-S.; Lerer, A. K.; Li, C.; Killebrew, D.; Mackenzie, K. M.; Mok, S. Y.-H.; Moraes, M. A.; Mueller, R.; Nociolo, L. J.; Peticolas, J. L.; Quan, T.; Ramot, D.; Salmon, J. K.; Scarpazza, D. P.; Schafer, U. B.; Siddique, N.; Snyder, C. W.; Spengler, J.; Tang, P. T. P.; Theobald, M.; Toma, H.; Towles, B.; Vitale, B.; Wang, S. C.; Young, C. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. In *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*; IEEE, 2014; pp 41−53.

(5) Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L.-P.; Shukla, D.; Tye, T.; Houston, M.; Stich, T.; Klein, C.; Shirts, M. R.; Pande, V. S. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **2013**, *9* (1), 461−469.

(6) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8* (5), 1542−1555.

(7) Shirts, M.; Pande, V. S. Screen Savers of the World Unite! *Science* **2000**, *290*, 1903−1904.

(8) Salvalaglio, M.; Tiwary, P.; Parrinello, M. Assessing the Reliability of the Dynamics Reconstructed from Metadynamics. *J. Chem. Theory Comput.* **2014**, *10* (4), 1420−1425.

(9) Abrams, C.; Bussi, G. Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy* **2014**, *16* (1), 163−199.

(10) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **2008**, *100* (2), 020603.

(11) Pfaendtner, J.; Bonomi, M. Efficient Sampling of High-Dimensional Free-Energy Landscapes with Parallel Bias Metadynamics. *J. Chem. Theory Comput.* **2015**, *11* (11), 5062−5067.

(12) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562.

(13) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. Free-Energy Landscape for β Hairpin Folding from Combined Parallel Tempering and Metadynamics. *J. Am. Chem. Soc.* **2006**, *128* (41), 13435−13441.

(14) Laio, A.; Gervasio, F. L. Metadynamics: A Method to Simulate Rare Events and Reconstruct the Free Energy in Biophysics, Chemistry and Material Science. *Rep. Prog. Phys.* **2008**, *71* (12), 126601.

(15) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9* (4), 2000−2009.

(16) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F.; Perez-hernandez, G.; Paul, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139* (1), 015102.

(17) Noé, F.; Clementi, C. Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation. *J. Chem. Theory Comput.* **2015**, *11* (10), 5002−5011.

(18) McGibbon, R. T.; Husic, B. E.; Pande, V. S. Identification of Simple Reaction Coordinates from Complex Dynamics. *J. Chem. Phys.* **2017**, *146* (4), 044109.

(19) Jo, S.; Suh, D.; He, Z.; Chipot, C.; Roux, B. Leveraging the Information from Markov State Models to Improve the Convergence of Umbrella Sampling Simulations. *J. Phys. Chem. B* **2016**, *120* (33), 8733−8742.

(20) Kästner, J. Umbrella Sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (6), 932−942.

(21) Spiwok, V.; Lipovová, P.; Králová, B. Metadynamics in Essential Coordinates: Free Energy Simulation of Conformational Changes. *J. Phys. Chem. B* **2007**, *111* (12), 3073−3076.

(22) Hashemian, B.; Millán, D.; Arroyo, M. Modeling and Enhanced Sampling of Molecular Systems with Smooth and Nonlinear Data-Driven Collective Variables. *J. Chem. Phys.* **2013**, *139* (21), 214101.

(23) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Demonstrating the Transferability and the Descriptive Power of Sketch-Map. *J. Chem. Theory Comput.* **2013**, *9* (3), 1521−1532.

(24) Rohrdanz, M. A.; Zheng, W.; Clementi, C. Discovering Mountain Passes via Torchlight: Methods for the Definition of Reaction Coordinates and Pathways in Complex Macromolecular Reactions. *Annu. Rev. Phys. Chem.* **2013**, *64* (1), 295−316.

(25) Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noé, F. Variational Approach to Molecular Kinetics. *J. Chem. Theory Comput.* **2014**, *10* (4), 1739−1752.

(26) McGibbon, R. T.; Pande, V. S. Variational Cross-Validation of Slow Dynamical Modes in Molecular Kinetics. *J. Chem. Phys.* **2015**, *142* (12), 124105.

(27) Tiwary, P.; Berne, B. J. Spectral Gap Optimization of Order Parameters for Sampling Complex Molecular Systems. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (11), 2839−2844.

(28) Schwantes, C. R.; Pande, V. S. Modeling Molecular Kinetics with tICA and the Kernel Trick. *J. Chem. Theory Comput.* **2015**, *11* (2), 600−608.

(29) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089.

(30) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins* **2010**, *78* (8), 1950−1958.

(31) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926.

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

(33) Mcgibbon, R. T.; Beauchamp, K. A.; Schwantes, C. R.; Wang, L.-P.; Hernández, C. X.; Harrigan, M. P.; Lane, T. J.; Swails, J. M.; Pande, V. S.; Hern, C. X.; Herrigan, M. P.; Lane, T. J.; Swails, J. M.; Pande, V. S. MDTraj: A Modern, Open Library for the Analysis of Molecular Dynamics Trajectories. *bioRxiv* **2014**, 9−10.

(34) Harrigan, M. P.; Sultan, M. M.; Hernández, C. X.; Husic, B. E.; Eastman, P.; Schwantes, C. R.; Beauchamp, K. A.; Mcgibbon, R. T.; Pande, V. S. MSMBuilder: Statistical Models for Biomolecular Dynamics. *Biophys. J.* **2017**, *112* (1), 10−15.

(35) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J. H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11* (11), 5525−5542.

(36) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129* (12), 124105.

(37) Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. Efficient Reconstruction of Complex Free Energy Landscapes by Multiple Walkers Metadynamics. *J. Phys. Chem. B* **2006**, *110* (8), 3533−3539.

(38) Müller, K. R.; Mika, S.; Rätsch, G.; Tsuda, K.; Schölkopf, B. An Introduction to Kernel-Based Learning Algorithms. *IEEE Trans. Neural Netw.* **2001**, *12* (2), 181−201.

(39) Hofmann, T.; Schölkopf, B.; Smola, A. J. Kernel Methods in Machine Learning. *Annals of Statistics* **2008**, *36*, 1171−1220.

(40) Harrigan, M. P.; Pande, V. S. Landmark Kernel tICA For Conformational Dynamics. 2017, http://biorxiv.org/content/early/2017/04/04/123752.

(41) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science (80-.)* **2010**, *330* (6002), 341−346.

(42) Pérez-Hernández, G.; Noé, F. Hierarchical Time-Lagged Independent Component Analysis: Computing Slow Modes and Reaction Coordinates for Large Molecular Systems. *J. Chem. Theory Comput.* **2016**, *12* (12), 6118.

(43) Biarnés, X.; Pietrucci, F.; Marinelli, F.; Laio, A. METAGUI. A VMD Interface for Analyzing Metadynamics and Molecular Dynamics Simulations. *Comput. Phys. Commun.* **2012**, *183* (1), 203−211.

(44) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33−38.

(45) Frishman, D.; Argos, P. Knowledge-Based Protein Secondary Structure Assignment. *Proteins: Struct., Funct., Genet.* **1995**, *23* (4), 566−579.

(46) Hunter, J. D. Matplotlib: A 2D Graphic Environment. *Comput. Sci. Eng.* **2007**, *9* (3), 90−95.

(47) Pérez, F.; Granger, B. E. IPython: A System for Interactive Scientific Computing. *Comput. Sci. Eng.* **2007**, *9* (3), 21−29.

(48) Hernández, C. X.; Harrigan, M.; Pande, V. S. MSMExplorer: Data Visualizations for Biomolecular Dynamics. *J. Open Source Softw.* **2017**, *2* (12).

(49) Han, W.; Schulten, K. Characterization of Folding Mechanisms of Trp-Cage and WW-Domain by Network Analysis of Simulations with a Hybrid-Resolution Model. *J. Phys. Chem. B* **2013**, *117* (42), 13367−13377.

(50) Maragliano, L.; Vanden-Eijnden, E. A Temperature Accelerated Method for Sampling Free Energy and Determining Reaction Pathways in Rare Events Simulations. *Chem. Phys. Lett.* **2006**, *426* (1−3), 168−175.

(51) Abrams, C. F.; Vanden-Eijnden, E. Large-Scale Conformational Sampling of Proteins Using Temperature-Accelerated Molecular Dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (11), 4961−4966.