JCTC Journal of Chemical Theory and Computation

Article

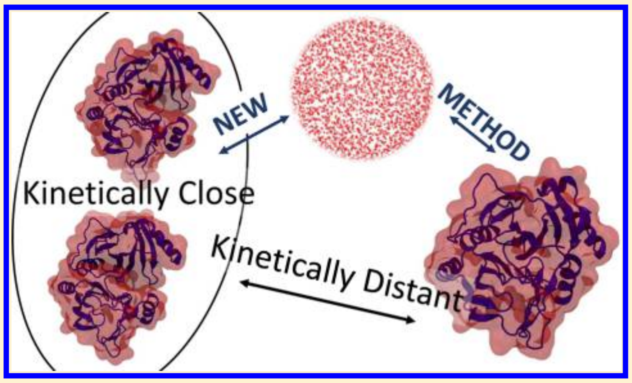
pubs.acs.org/JCTC

Conserve Water: A Method for the Analysis of Solvent in Molecular Dynamics

Matthew P. Harrigan,[†] Diwakar Shukla,[†,‡] and Vijay S. Pande*[,†,∥,§]

[†]Department of Chemistry, [∥]Department of Computer Science, and [§]Department of Structural Biology, Stanford University, Stanford, California 94305, United States

ABSTRACT: Molecular dynamics with explicit solvent is favored for its ability to more correctly simulate aqueous biological processes and has become routine thanks to increasingly powerful computational resources. However, analysis techniques including Markov state models (MSMs) ignore solvent atoms and focus solely on solute coordinates despite solvent being implicated in myriad biological phenomena. We present a unified framework called "solvent-shells featurization" for including solvent degrees of freedom in analysis and show that this method produces better models. We apply this method to simulations of dewetting in the two-domain protein BphC to generate a predictive MSM and identify functional water molecules. Furthermore, the proposed methodology could be easily extended for building MSMs of any systems with indistinguishable components.



1. INTRODUCTION

Changes in conformations of proteins and nucleic acids underlie the majority of emergent biological phenomena in daily life. Life, death, and disease are the result of molecules changing shape in dynamical processes such as protein folding, kinase activation, and signaling.[1−3] Understanding these dynamical processes is fundamental to our understanding of biology. Experimental probes such as X-ray crystallography and NMR can provide static pictures of macromolecules, and certain specialized methods can give limited information about dynamics.[3] For systems ill-suited to experimental characterization, molecular dynamics (MD) offers unparalleled atom-level detail of the dynamics of microscopic systems. Recent advances in computing including the use of GPUs,[4,5] specialized hardware,[6] and distributed computing[7−9] have enabled simulations to probe biologically relevant macromolecules at biologically relevant time scales. Additionally, increasing computational power has enabled simulations to probe molecules in biologically relevant solvent environments: explicit representation of water molecules[10] and lipid membranes[11] has become routine. With simulation times reaching milliseconds and the number of atoms approaching hundreds of thousands, some sort of dimensionality reduction is needed to make sense of this huge amount of data.[12,13]

One dimensionality reduction technique involves construction of a Markov state model (MSM) from the time series of atomic coordinates from MD.[14] MSMs parametrize a system by a set of states and rates. Snapshots from MD trajectories are grouped or clustered into *k* states. Some information is necessarily lost by lumping conformations, but with a sufficiently fine partitioning, we can resolve states with sufficient detail. [15] The ideal clustering for MSM construction groups conformations that interconvert rapidly as shown in Figure 1. Lacking an a priori measure of the interconversion rate of two given conformations (e.g., two frames of an MD trajectory), estimation of this kinetic closeness is approximated by a conformational distance metric. For example, root-mean-square deviation (RMSD) or euclidean distance between features such as dihedral angles or contact distances have

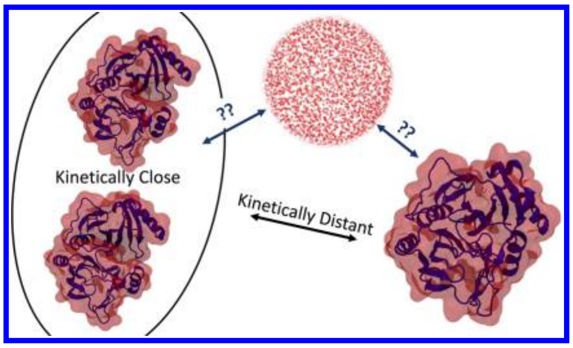


Figure 1. Building an MSM requires a distance metric to cluster kinetically close conformations. The enzyme BphC is depicted. Left: two extended conformations with a "wet" interfacial cavity. We estimate that these similar structures interconvert rapidly. Right: The collapsed, dewetted structure. This is kinetically distant from the extended conformations. Top: A sample of the water box solvating BphC. We lack a method for estimating kinetic distance for solvent degrees of freedom due to the large number and indistinguishability of solvent molecules.

Received: November 7, 2014
Published: February 12, 2015

DOI: 10.1021/ct5010017
J. Chem. Theory Comput. 2015, 11, 1094−1101

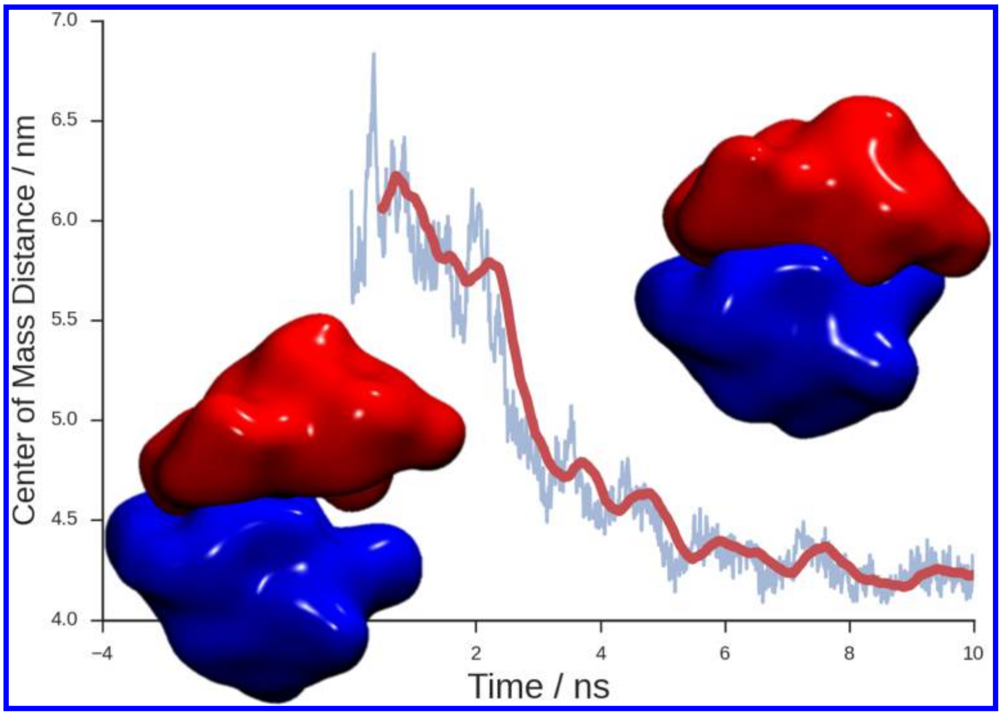


Figure 2. Two-domain enzyme BphC is started from an extended conformation and is allowed to dewet. The two domains are shown in a low-resolution surface representation at t = 0.7 ns (left) and t = 8.8 ns (right). The center of mass distance is plotted over time. We use this system as a model to test the new method presented. This molecule can be seen in "cartoon" representation in Figure 1.

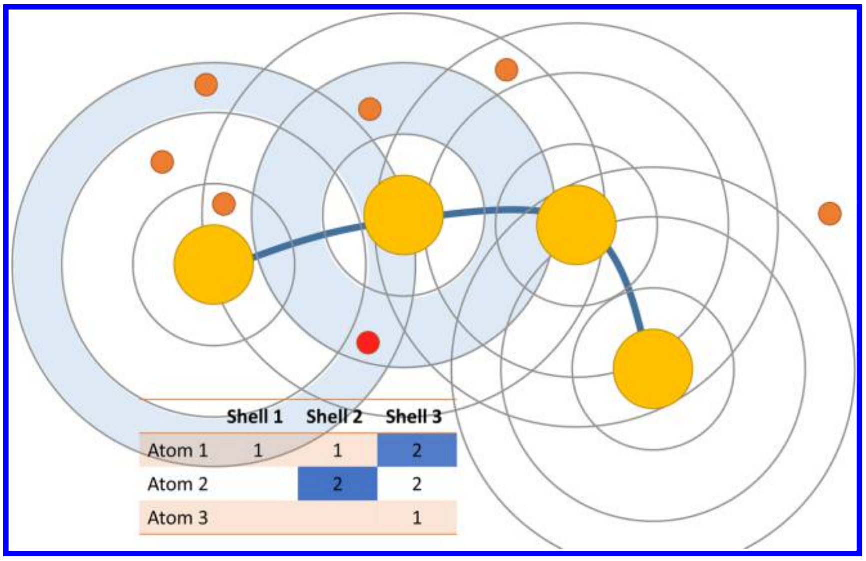
been used for state definition. [16] With the states defined and MD snapshots assigned to their proper states, we estimate k^2 transition probabilities among the k states. "Markovianity" means we idealize the dynamics of the biological system as a set of memoryless jumps between states. MSMs have been used successfully to reveal important biological structure and function in diverse biological applications such as folding,[17−20] kinase and protease activation,[21,22] GPCR signaling, [9] protein−ligand binding,[23] and self-assembly. [24]

A second, routine "dimensionality reduction" has been to discard all solvent atoms prior to analysis,[19,25,26] despite water and membranes being crucial to protein function and biology as a whole. In fact, solvent has been implicated in hydrophobic collapse,[27] protein stability,[28] protein−ligand binding,[29] modulation of ion channel function,[30] antifreeze proteins,[31] and aggregation.[32,33] There is understandable reason to discard solvent atoms. First, the number of atoms is large: there are often 50 times as many water molecules as protein residues (and 10,000 times as many water molecules as protein molecules) in an explicit-solvent MD simulation box. Second, solvent molecules—unlike proteins or nucleic acids—are indistinguishable. It does not matter if water #1525 or #19832 is solvating a particular side-chain; conformations in which these solvent atoms are exchanged should be treated as identical. Methods tuned for analysis of solutes are ill-suited to considering both the indistinguishability and the large number of solvent molecules in a typical simulation. MSMs are no exception: specifically, we lack an estimation of kinetic closeness in solvent degrees of freedom. Traditional treatments of solvent rely on aligning solute conformations and laying down a grid of voxels for which properties like density can be calculated and visualized.[23,34] These methods (1) fail for large conformational changes or folding when alignment of the solute is poor and (2) fail to interface with statistical tools (e.g., principle component analysis (PCA), MSMs) due to an overabundance of resultant features. If we consider a cubic simulation box of side-length 80 Å and voxels of side-length 3 Å, grid-based approaches would yield 19,000 features. Other work has focused on grouping solvent conformations based on truncated hydrogen-bonding networks.[35] This approach fails for general indistinguishable particles other than H_2O for which bonding criteria is not known or not relevant. This method does not afford a distance metric to relate conformations and, as such, requires one state per enumerated hydrogen-bonding network (k = 50,000 when only considering the first and second solvation shells). The lack of an Euclidean distance metric once again hinders interface with statistical tools like PCA or K-means clustering. A suitable transformation of solvent positions into tractable features ("featurization") like those available for solutes would yield a solvent distance metric that could be used during the clustering stage of MSM construction.

A model system for solvent dynamics is that of hydrophobic collapse in the BphC enzyme (1dhy).[36,37] This two-domain protein functions in degrading toxic polychlorinated biphenyls. Hydrophobic residues on the interface of the two domains promote both dewetting of the interfacial cavity and structural collapse from extended conformations. By starting simulations from artificially extended conformations, we can observe dewetting transitions. Figure 2 shows one such transition. We stress that the focus of this study is to demonstrate how to include solvent degrees of freedom in MSM analysis and not to provide novel insight into the function of BphC.

In this paper, we introduce a new method called the solvent-shells featurization for transforming solvent positions into suitable solvent features. We characterize and parametrize this method on 100 (10 ns each) MD simulations of the BphC enzyme, each initialized from an extended conformation. Through the use of an appropriate scoring function under cross-validation, we examine the hyperparameters in model

DOI: 10.1021/ct5010017
J. Chem. Theory Comput. 2015, 11, 1094−1101



	Shell 1	Shell 2	Shell 3
Atom 1	1	1	2
Atom 2		2	2
Atom 3			1

Figure 3. Schematic representation of the solvent-shells featurization. Solvent atoms (small, filled circles) are binned based on radial distance (concentric rings) from each solute atom (large, filled circles). By training an intermediate kinetic model such as tICA, "important" solvent-shell features can be identified (highlighted table entries). We can exploit overlapping shells (e.g., two shaded rings corresponding to highlighted table entries) to provide nonradially symmetric identification of regions of solvent (small, red circle).

construction and show that including solvent degrees of freedom in MSM construction gives better models. Finally, we interpret the resulting model by taking advantage of state-of-the-art MSM techniques and the new solvent features.

2. SOLVENT-SHELLS FEATURIZATION

In contrast to traditional conformational distance metrics used for clustering of solute (protein) states, we seek a distance metric suitable for solvents that (1) treats solvent molecules as indistinguishable and (2) is invariant under translation and rotation of the solvent box relative to the solute molecule. A solvent metric would be particularly desirable if it (3) can identify solvent molecules of interest and (4) is fast to compute.

Gu et al. defined a "solvent fingerprint"[38] which uses a sum of weighted solute–solvent distances for each solute atom to define a vector representation of the solvent configuration.

$$\mathrm{FP}(x \in \text{solute}) = \sum_{y\in\text{solvent}} \exp\left(-\frac{\|x-y\|^2}{2\alpha^2}\right) \qquad (1)$$

where $\|x - y\|$ is the Euclidean distance between solute atom position x and solvent atom position y, and α is a free parameter that defines a distance scale. The resulting feature vector is of length N_{solute} and can be used with an ordinary l^2 norm for clustering. Physically, we can interpret the feature values as the degree of solvation of each solute atom.

We propose an extension of this fingerprint where we seek to preserve spatial resolution of the solvent that is destroyed in the summation. We define the solvent-shell featurization:

$$\mathrm{SS}(x \in \text{solute};\, r) = (4\pi r_{\text{mid}}\, dr)^{-1} \sum_{y\in\text{solvent}} \mathbf{I}(r \le \|x-y\| < r + dr) \qquad (2)$$

parametrized by a set of spherical shells specified by distance r and width dr, and where $r_{\text{mid}} = r + dr/2$. The solvent-shell featurization gives the instantaneous solvent density in each of the shells. The resulting feature vector is of length $N_{\text{solute}}{\cdot}N_{\text{shells}}$ and can be used with an ordinary l^2 norm for clustering. For computational and cognitive convenience, we use an integer number of equal-width shells. The featurization is shown schematically in Figure 3.

These features bin solvent atoms without regard for their identity, satisfying (1), and only consider relative solute–solvent distances, satisfying (2). By recording the assignment of solvent atoms to shells, we can back-out individual solvent atoms corresponding to each feature with resolution at least as good as dr. This satisfies criterion (3) and is a powerful way to extract biophysical understanding by identifying functional waters, see Figure 4.

This featurization is implemented as a plug-in for the open-source software package MSMBuilder. Computation of the features is performed with SSE4.1 vectorized operations, thus satisfying (4). The resulting feature vector enables the use of fast clustering methods such as Mini-batch K-means,[39] further enhancing computational speed in contrast to the traditional RMSD metric.

2.1. Unified Framework for Solvent and Conformational Dynamics.

Having a protocol for computing solvent features, we wish to construct a model that captures both solvent and solute dynamics. To that end, we choose a set of solute conformational features (e.g., dihedral angles, raw Cartesian coordinates) to be used in conjunction with the solvent-shells features.

State-of-the-art MSM construction methods suggest training a kinetic model prior to the clustering step.[13,40] One such model is time-independent component analysis (tICA). tICA is similar to principle component analysis (PCA) in that it produces a set of linear combinations of input features which define "components" to serve as a new basis set for the data. Whereas PCA finds components which maximize variance among input degrees of freedom, tICA finds components which maximize autocorrelation of the time-series input.[41,42] We effectively find the slowest degrees of freedom for the system (subject to the constraint that the degrees of freedom be linear combinations of the input features). By projecting our input features on the top n slowest time-independent components (tICs), we introduce an intermediate dimensionality reduction which aligns our conformation-based estimate of kinetic interconversion rates even closer to the actual kinetics of the model. Because of the linearity constraint of tICA, we generally still need to build an MSM to capture the nonlinear dynamics of the system under study. This intermediate processing with tICA or PCA permits a unified framework for treating solvent and conformational degrees of freedom in which all features are fed as input, and the component analysis model selects those features deemed “relevant”.

Special care should be taken when building an MSM directly from a union of conformational and solvent features without an intermediate model such as tICA. Spherical clustering algorithms (e.g., k-centers) are sensitive to scaling of input features; the two sets of features should be normalized to have equal variance to ensure meaningful clustering. PCA includes normalization by variance. tICA includes normalization by either variance or autocorrelation time scale (i.e., slowness); in this study, the autocorrelation time scales were used for normalization.

Table 1. Hyperparameters Were Investigated and Tuned by Performing Two Grid Searches over the Values Given in This Table[a]

hyperparameter	grid search 1	grid search 2
features	DRID, solvent, both	
shell widths (dr), Å	1, 2, 3, 4	3
total extent ($dr \cdot N_{\text{shells}}$), Å	5, 10	
number of shells (N_{shells})		1, 2, 3, 4
tICA components	1, 2, 3	1
MSM microstates	50, 100, 200, 400, 800	
no. models	255	45

[a] For each parameter configuration, models were scored using three-fold cross validation using the GMRQ scoring functional. In grid search 1, the spacial extent of the featurization was explicitly specified. In cases where this would dictate a fractional number of shells, the nearest integer number was used. In grid search 2, the number of shells was explicitly specified which implicitly determined the spacial extent.

3. EVALUATION ON BPHC ENZYME

With this new method, we seek to create better MSMs by capturing the slow, biologically relevant dynamical processes with a generalizable model that uses a unified framework to include solute, solvent, membrane, and any other key degrees of freedom. Recently, McGibbon et al. introduced a scoring function based on the generalized matrix Rayleigh quotient (GMRQ) that quantifies the goodness-of-fit for an MSM.[43] The GMRQ is a scalar functional which measures the ability of a rank-m projection operator (in this case, the top m eigenvectors of the MSM) to capture the slow dynamics of a system.[44] In theory, the GMRQ is bounded by the sum of the first m eigenvalues of the true dynamical propagator: a nonperfect dimensionality reduction will always model dynamics which are too fast.[45] However, McGibbon et al. showed that this bound can be violated when the model was parametrized from statistically noisy inputs (e.g., a MD simulation with less than infinite sampling). This overconfidence in the model is a result of overfitting. It can be eliminated by evaluating the model on a different data set than the one on which it was trained, i.e., via cross-validation. Due to its generality, the GMRQ permits direct comparison among MSMs built with methods that could differ in hyperparameters, intermediate processing steps, and/or featurization.

We performed three-fold cross validation (trajectories kept whole across folds) over the grid of hyperparameters specified in Table 1 (Grid Search 1) on 100 (10 ns each) MD simulations of the BphC enzyme, each initialized from an extended conformation. Conformational degrees of freedom were included using distribution of reciprocal of interatomic distances (DRID) features.[46] For each solute atom, the reciprocal distance to every other solute atom is computed, forming a distribution. DRID characterizes this distribution by its first three moments. DRID is a translationally and rotationally invariant way to featurize solute molecules with no a priori knowledge of the system. Solvent degrees of freedom were included using the solvent-shells features introduced in this paper. Solute atoms were defined to be the α-carbons of the protein residues, and solvent atoms were defined to be the water oxygens. Pruning redundant parameter configurations (solvent-specific parameters do not matter when including only DRID features) yielded 255 models and associated scores over five dimensions. The models were evaluated using GMRQ based on fidelity to the two slowest dynamical processes and the equilibrium distribution (rank m = 3) at a lag time of 0.5 ns.

Due to the generality of the GMRQ, we can simply select the set of hyperparameters that yield the highest mean (over folds) test set score. This suggests using both solute (via DRID) and solvent (via solvent-shells) features, 4 shells each of width 3 Å, 1 tIC, and 100 MSM states. Further investigation of the marginal effects of specific hyperparameters offers insight into the new method.

Figure 4a shows scores as a function of the number of tICA components included in MSM construction for each of the three input-feature configurations. These scores were taken from the optimal solvent-shell parameters (given above) and marginalized over number of MSM microstates by taking the maximum score. Error bars represent standard deviation over folds. For configurations which include the solvent-shells metric, the score decreases with increasing number of components included in MSM construction. This is most likely due to overfitting to the extra degrees of freedom. These observations are consistent with these simulations, which are dominated by one coordinate (the dewetting conformational change). Similarly, including conformational degrees of freedom in addition to the solvent degrees of freedom does not appreciably increase the score of the MSM. The conformational degrees of freedom do not capture anything in the first tIC that the solvent metric does not. The conformational metric (DRID) behaves differently: its score peaks at 2 tICs (albeit with high variance across folds) before succumbing to overfitting. It makes sense that the solvent features reproduce this one coordinate better than a general conformational metric in this system where solvent change is the dominating characteristic. We expect more complicated systems to benefit from multiple tICs and a combination of solvent and conformational degrees of freedom.

The dependence of score on solvent shell width (dr) was investigated for 1 tIC and total extent $N_{\text{shells}} \cdot dr$ =10 Å, again marginalizing over number of MSM microstates (Figure 4b). Choosing an appropriate shell width balances statistical variance with spatial resolution: A large number of skinny shells provides a higher resolution description of the solvent environment, but wider shells occupied by more molecules provide a lower variance estimate of the local density. We observe that a shell width of 3 Å provides the best balance and

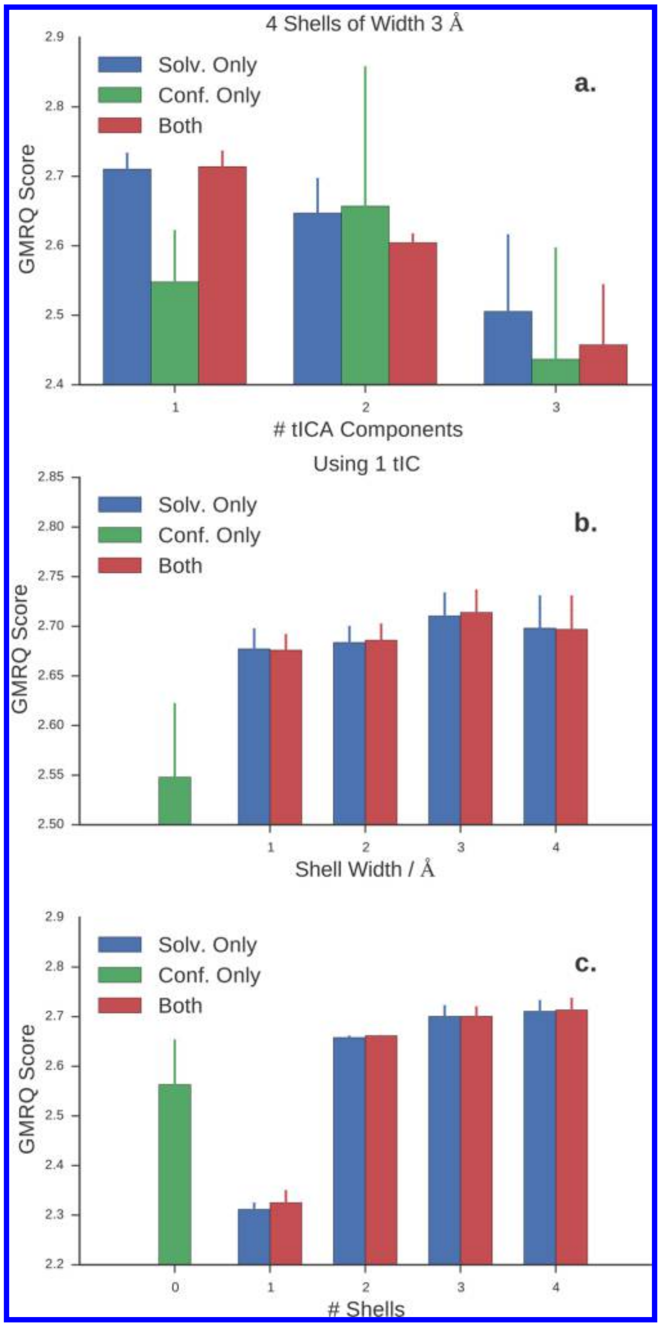


Figure 4. (a) Increasing the number of tICs biases the model toward over fitting. In this simple system, one coordinate is sufficient to provide a generalizable model. Using only the conformational features requires two tICs to maximize the score, whereas inclusion of solvent features which match the physics of the simulations maximizes the score with only one tIC. The score is maximized with the inclusion of the solvent-shells features introduced in this paper. (b) A shell width of 3 Å balances statistical variance with spatial resolution. (c) Increasing number of solvent shells results in a better score. By not extending the solvent featurization far enough (i.e., using only 1 shell), the model performs significantly worse than one fit only on conformational features.

maximizes the GMRQ score. This is physically reasonable as it corresponds to between one and two solvation shells in water.

A second grid search was run to investigate dependence of model score on the number of shells included in the featurization (Figure 4c). Shell widths and number of tICs were kept constant at optimal values from the first grid search (Table 1, grid search 2). While we might expect solvation to be a local effect, including more shells seems to improve the model without introducing overfitting. In fact, including only the closest shell results in a significantly worse model than one with just conformational features. We postulate that increasing the spatial extent of the featurization allows nonspherically symmetric localization of important regions of solvent when used in conjunction with tICA. For example, consider a region of solvent that partially occupies shell 1 of residue 1 and shell 2 of neighboring residue 2. The overlap of these two occupancies breaks the spherical symmetry of the featurization around each individual residue as shown schematically in Figure 3. These findings suggest that only considering hyper-local solvation, as in ref 38, may be misguided.

4. BIOPHYSICAL INTERPRETATION

A powerful feature of the solvent-shells featurization is its interpretability. Solvent molecules can be assigned to the shells which they occupy. One convenient way to exploit this property is to use the coefficients of the tICA model's slowest components. Each tIC is defined as a linear combination of input features. When the input features are the solvent-shell occupancies, the resulting coefficients can be used to assign "importance" (i.e., degree of contribution to slow dynamical processes) to the solvent shells and, by extension, individual solvent molecules which occupy those shells.

We applied the solvent-shell featurization to the same ensemble of 100 (10 ns each) simulations of the model two-domain protein BphC. Figure 5 shows that projection along the two slowest tICs provides separation into at least two regions of high population. The trajectory paths suggest that the first tIC is highly correlated with dewetting of the interdomain cavity. We use the trained tICA model to enhance our biophysical understanding by visualizing each solvent molecule colored

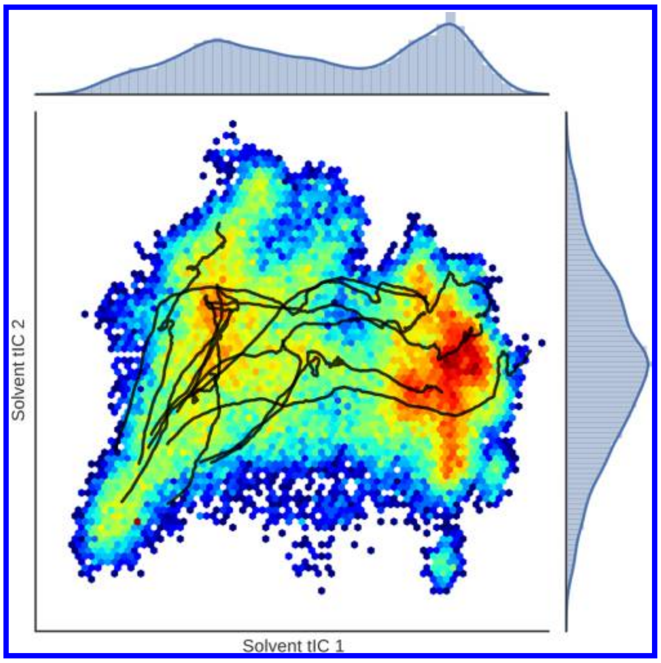


Figure 5. tICA analysis on solvent-shell features successfully identifies the slow degree of freedom corresponding to dewetting. Several trajectories (black lines) projected onto the first two tICs are overlaid on a 2D histogram of solvent conformations to show the general progression from wet (low tIC1, left) to dry (high tIC1, right).

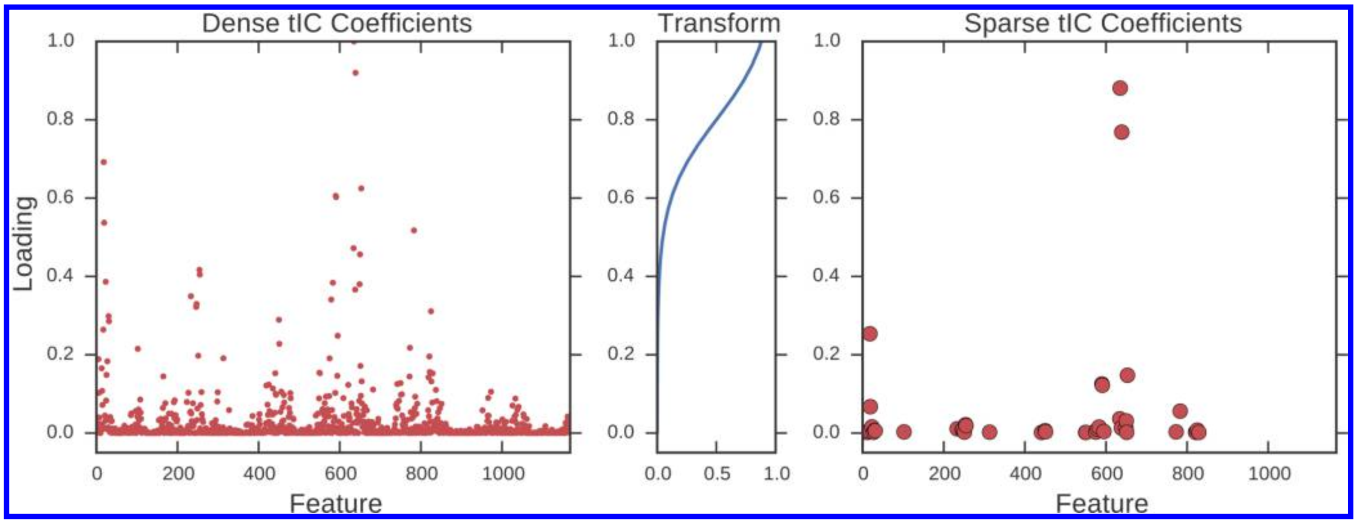


Figure 6. tICA independent component coefficients are dense with regards to input features. A logistic function is applied to up-weight the significant features while reducing noise. When performed on the solvent-shells features, the sparse coefficients can be used to visualize solvent molecules of importance.

according to its tIC coefficient. We applied a logistic function as depicted in Figure 6 and summed the coefficients of water-oxygen atoms occupying overlapping solvent shells. Altering the logistic function parameters did not qualitatively affect the resulting visualization. With VMD,[47] these values could be used to color, show, or hide solvent molecules of importance. As seen in Figure 7, the solvent-shells features allow the automated discovery of the interesting, slow-dynamical solvent features. The water molecules in BphC's hydrophobic cavity are discovered a priori.

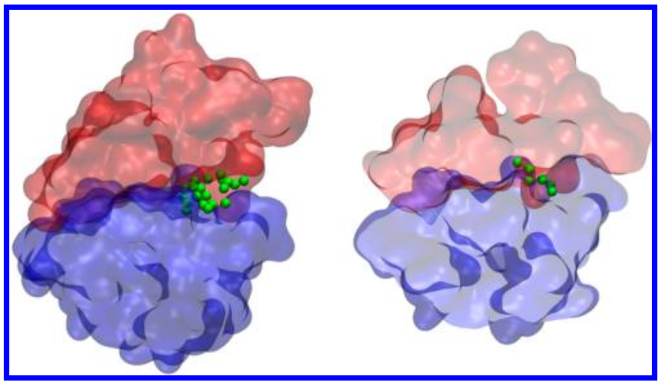


Figure 7. By including only the solvent molecules with tIC coefficients (learned by this method) above a cutoff value, individual water molecules that comprise the slowest degree of freedom are revealed. Here, the two domains of BphC (colored red and blue) are shown in a surface representation, and water molecules are represented by spheres centered on the water-oxygens. (Left) A wet structure at $t = 1$ ns contains many waters in the interdomain cavity. (Right) The same trajectory at $t = 9$ ns. The water has mostly been expelled from the cavity. The front half of the BphC molecule is not drawn to show the few trapped waters within the enzyme. The solvent molecules participating in the dewetting are automatically identified by the method due to their kinetic relevance.

5. CONCLUSIONS

The inclusion of solvent degrees of freedom in MD trajectory analysis has the potential to be a boon for biophysical understanding both by enhancing interpretability of the models as well as improving the models themselves. As seen from the GMRQ scores for MSMs of the model protein BphC, the solvent-shells featurization in conjunction with a structural metric yields more generalizable dynamical models. With the aid of tICA, these models are also more interpretable. Projection onto solvent tICs allows visual inspection of candidate metastable states. Visualizing solvent molecules with high tIC coefficients shows important solvent features learned entirely from the MD data in their appropriate biological context. This enables the discovery of functional solvent molecules from simulations without prior knowledge. We anticipate that the ease of incorporating the solvent-shells metric into the standard MSM analysis paradigm combined with the benefits of treating *all* degrees of freedom in our analysis will have broad applications beyond water. Whereas the water-oxygen atoms were considered in this study, we emphasize that this method is general to any indistinguishable particles including ions and lipids. Including indistinguishable particles in dynamical analysis allows MSMs to more naturally model a much broader range of phenomena including protein−lipid interactions, membrane dynamics, colloidal systems, docking, and many-body protein simulations.[24]

6. SIMULATION DETAILS

The simulations were started from the crystal structure of BphC (PDB id: 1dhy.pdb).[36,37] The crystal structure contains two domains (residues 1−135 and 135−292) at a center of mass distance of 18.72 Å. These structures were solvated in a TIP3P[48] water box containing ∼16,500 molecules such that the minimum distance between the boundary and the protein is 12 Å. The system was neutralized by adding 8 Na^+ ions. The Amber99sb-ildn[49] force field was used for protein and ions. The structures obtained after an initial equilibration for 1 ns at constant temperature and pressure and with constraints on the heavy atom positions were used as the starting conformation for the subsequent simulations. The interdomain distance of the crystal structure was increased by ∼6 Å along the direction of two domain centers of geometry to create a gap between the two domains using the steered molecular dynamics[50] method. The constant velocity pulling method was used by restraining residues 1−135 and applying a force on residues 136−292. The resulting structure was then used for running 100 dewetting simulations of 10 ns each for a total simulation time of 1000 ns. The GROMACS[51] simulation package was used for running these simulations. Covalent bonds involving hydrogen atoms were constrained with LINCS,[52] and particle mesh Ewald[53] was used to treat long-range electrostatic interactions. Production MD simulations were carried out at constant temperature and pressure of 300 K and 1 atm, respectively, with a time step of 2 fs.

The code used for computing the solvent-shells featurization is available at http://github.com/mpharrigan/wetmsm and depends on MDTraj (mdtraj.org) and MSMBuilder (msmbuilder.org).

AUTHOR INFORMATION

Corresponding Author

*E-mail: pande@stanford.edu.

Present Address

‡Department of Chemical & Biomolecular Engineering, University of Illinois at Urbana−Champaign, Urbana, IL 61801

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We would like to thank Robert T. McGibbon and Christian R. Schwantes for useful discussion throughout the preparation of this manuscript. D.S. is supported by the Biomedical Data Science Initiative Fellowship from Stanford School of Medicine. D.S. and V.S.P. acknowledge support from the SIMBIOS NIH National Center for Biomedical Computation through the NIH Roadmap for Medical Research Grant U54 GM07297. In addition, we acknowledge funding from NSF-MCB-0954714 and NIH-R01-GM062868.

REFERENCES

(1) Grant, B. J.; Gorfe, A. A.; McCammon, J. A. *Curr. Opin. Struct. Biol.* **2010**, *20*, 142−147.

(2) Taylor, S. S.; Kornev, A. P. *Trends Biochem. Sci.* **2011**, *36*, 65−77.

(3) Henzler-Wildman, K.; Kern, D. *Nature* **2007**, *450*, 964−972.

(4) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. *J. Comput. Chem.* **2009**, *30*, 864−872.

(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. *J. Chem. Theory Comput.* **2013**, *9*, 461−469.

(6) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; 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.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. *Commun. ACM* **2008**, *51*, 91−97.

(7) Shirts, M.; Pande, V. S. *Science* **2000**, *290*, 1903−1904.

(8) Buch, I.; Harvey, M. J.; Giorgino, T.; Anderson, D. P.; De Fabritiis, G. *J. Chem. Inf. Model.* **2010**, *50*, 397−403.

(9) Kohlhoff, K. J.; Shukla, D.; Lawrenz, M.; Bowman, G. R.; Konerding, D. E.; Belov, D.; Altman, R. B.; Pande, V. S. *Nat. Chem.* **2014**, *6*, 15−21.

(10) Wang, L.-P.; Head-Gordon, T.; Ponder, J. W.; Ren, P.; Chodera, J. D.; Eastman, P. K.; Martinez, T. J.; Pande, V. S. *J. Phys. Chem. B* **2013**, *117*, 9956−9972.

(11) Dickson, C. J.; Madej, B. D.; Skjevik, Å. A.; Betz, R. M.; Teigen, K.; Gould, I. R.; Walker, R. C. *J. Chem. Theory Comput.* **2014**, *10*, 865−879.

(12) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. *Curr. Opin. Struct. Biol.* **2013**, *23*, 58−65.

(13) Schwantes, C. R.; McGibbon, R. T.; Pande, V. S. *J. Chem. Phys.* **2014**, *141*, 090901.

(14) Bowman, G. R., Pande, V. S., Noé, F. In *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*; Bowman, G. R., Pande, V. S., Noé, F., Eds.; Springer: Dordrecht, the Netherlands, 2014; Vol. 797.

(15) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105.

(16) Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods* **2010**, *52*, 99 − 105.

(17) Jayachandran, G.; Vishal, V.; Pande, V. S. *J. Chem. Phys.* **2006**, *124*, 164902−164902−12.

(18) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2011**, *133*, 18413−18419.

(19) Voelz, V. A.; Jäger, M.; Yao, S.; Chen, Y.; Zhu, L.; Waldauer, S. A.; Bowman, G. R.; Friedrichs, M.; Bakajin, O.; Lapidus, L. J.; Weiss, S.; Pande, V. S. *J. Am. Chem. Soc.* **2012**, *134*, 12565−12577.

(20) Baiz, C. R.; Lin, Y.-S.; Peng, C. S.; Beauchamp, K. A.; Voelz, V. A.; Pande, V. S.; Tokmakoff, A. *Biophys. J.* **2014**, *106*, 1359−1370.

(21) Shukla, D., Meng, Y., Roux, B., Pande, V. S. *Nat. Commun.* **2014**, *5*, DOI: 10.1038/ncomms4397.

(22) Sadiq, S. K.; Noé, F.; Fabritiis, G. D. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 20449−20454.

(23) Buch, I.; Giorgino, T.; Fabritiis, G. D. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 10184−10189.

(24) Perkett, M. R.; Hagan, M. F. *J. Chem. Phys.* **2014**, *140*, 214101.

(25) Lapidus, L. J.; Acharya, S.; Schwantes, C. R.; Wu, L.; Shukla, D.; King, M.; DeCamp, S. J.; Pande, V. S. *Biophys. J.* **2014**, *107*, 947−955.

(26) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517−520.

(27) Chandler, D. *Nature* **2002**, *417*, 491−491.

(28) Sorin, E. J.; Pande, V. S. *J. Am. Chem. Soc.* **2006**, *128*, 6316−6317.

(29) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 808−813.

(30) Faure, É; Thompson, C.; Blunck, R. *J. Biol. Chem.* **2014**, *289*, 16452−16461.

(31) Sun, T.; Lin, F.-H.; Campbell, R. L.; Allingham, J. S.; Davies, P. L. *Science* **2014**, *343*, 795−798.

(32) Krone, M. G.; Hua, L.; Soto, P.; Zhou, R.; Berne, B. J.; Shea, J.-E. *J. Am. Chem. Soc.* **2008**, *130*, 11066−11072.

(33) Wei, G.; Shea, J.-E. *Biophys. J.* **2006**, *91*, 1638−1647.

(34) Fabritiis, G. D.; Geroult, S.; Coveney, P. V.; Waksman, G. *Proteins* **2008**, *72*, 1290−1297.

(35) Rao, F.; Garrett-Roe, S.; Hamm, P. *J. Phys. Chem. B* **2010**, *114*, 15598−15604.

(36) Zhou, R.; Huang, X.; Margulis, C. J.; Berne, B. J. *Science* **2004**, *305*, 1605−1609.

(37) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(38) Gu, C.; Chang, H.-W.; Maibaum, L.; Pande, V. S.; Carlsson, G. E.; Guibas, L. J. *BMC Bioinf.* **2013**, *14*, S8.

(39) Sculley, D. *Web-scale K-means Clustering*. Proceedings from the 19th International Conference on World Wide Web, Raleigh, NC, April 26−30, 2010; ACM: New York, 2010; pp 1177−1178.

(40) Chodera, J. D.; Noé, F. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135−144.

(41) Schwantes, C. R.; Pande, V. S. *J. Chem. Theory Comput.* **2013**, *9*, 2000−2009.

(42) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; Fabritiis, G. D.; Noé, F. *J. Chem. Phys.* **2013**, *139*, 015102.

(43) McGibbon, R. T., Pande, V. S. arXiv:1407.8083 [q.bio.BM].

(44) Noé, F.; Nüske, F. *Multiscale Model. Simul.* **2013**, *11*, 635−655.

(45) Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noé, F. *J. Chem. Theory Comput.* **2014**, *10*, 1739−1752.

(46) Zhou, T.; Caflisch, A. *J. Chem. Theory Comput.* **2012**, *8*, 2930–2937.
(47) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.
(48) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. *J. Chem. Phys.* **1983**, *79*, 926.
(49) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins: Struct., Funct., and Bioinf.* **2010**, *78*, 1950–1958.
(50) Izrailev, S., Stepaniants, S., Isralewitz, B., Kosztin, D., Lu, H., Molnar, F., Wriggers, W., Schulten, K. *Computational molecular dynamics: challenges, methods, ideas*; Springer: Berlin, 1999; pp 39–65.
(51) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. J. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.
(52) Hess, B.; Bekker, H.; Berendsen, H.; Fraaije, J. *J. Comput. Chem.* **1997**, *18*, 1463–1472.
(53) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.

DOI: 10.1021/ct5010017
J. Chem. Theory Comput. 2015, 11, 1094–1101