Towards the prediction of order parameters from molecular dynamics simulations in proteins

Juan R. Perilla and Thomas B. Woolf
Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

(Received 11 November 2011; accepted 22 March 2012; published online 23 April 2012)

A molecular understanding of how protein function is related to protein structure requires an ability to understand large conformational changes between multiple states. Unfortunately these states are often separated by high free energy barriers and within a complex energy landscape. This makes it very difficult to reliably connect, for example by all-atom molecular dynamics calculations, the states, their energies, and the pathways between them. A major issue needed to improve sampling on the intermediate states is an order parameter – a reduced descriptor for the major subset of degrees of freedom – that can be used to aid sampling for the large conformational change. We present a method to combine information from molecular dynamics using non-linear time series and dimensionality reduction, in order to quantitatively determine an order parameter connecting two large-scale conformationally distinct protein states. This new method suggests an implementation for molecular dynamics calculations that may be used to enhance sampling of intermediate states. [http://dx.doi.org/10.1063/1.3702447]

I. INTRODUCTION

Proteins represent complex dynamical systems with multiple stable states. Sampling protein motions has shown itself to be complicated by the multiple time scale problem[1] with many different wells and barrier heights being found. Despite considerable effort, there is currently no rapid way to determine the range of stable states from a single structure or from sequence alone. Because biological function is intrinsically linked to the large scale conformational change of proteins, an improved understanding of how conformational change in the complex energy landscape of the protein is determined will provide important insights on both biological and physical questions.

When the state change is connected by an obvious low-dimensional reaction coordinate, specialized sampling methods have been developed that can reliably enhance the collection of intermediate states and the understanding of the relative free energy change.[2–6] For example, the passage of an ion through a channel, simple alchemical changes, certain types of conformational change where the movement is mainly hinge-like or otherwise obvious on inspection fall into this category.[7,8] But, for many biological problems, the low-dimensional reaction coordinate that optimally predicts functional behavior is not at all obvious from the structure. As the set of solved x-ray structures has continued to grow, there has been an increasing number of situations where alternative structures for the same protein have been determined.[9] Ideally these alternate structures would also suggest the reaction coordinate that connects one conformation to the other. But, despite many outstanding efforts to design sampling methods, a strong limitation is that an order parameter to enable sampling has been difficult to determine from either single or pairs of static x-ray structures.

Some groups have suggested, in this situation, that harmonic analysis from diagonalization of the second derivative matrix (the Hessian) would be sufficient to find the most important collective modes.[10–16] The findings from the coarse-grained model community, in particular, have suggested that this approach can reveal important details about how a protein is connected to large conformational change.[17] Other groups have cautioned that the harmonic model is lacking and have suggested instead focusing on determining effective collective modes from the covariance fluctuation matrix (quasi-harmonic, essential dynamics, or principal component analysis).[18–22] These calculations has demonstrated that the low frequency collective motions inferred from the covariance matrix differ from the harmonic analysis. Recently there has been considerable efforts towards further improving on the linear assumptions in principal component analysis, using such tools as kurtosis and quasi anharmonic analysis to further elaborate on the deviations from Gaussian behavior that are found in (MD) molecular dynamics trajectories.[23–26]

To directly sample on large-scale conformational change there have been many methods proposed.[27–36] The most well known contemporary method is transition path sampling and uses a small number of conformations along a candidate transition pathway with a Monte Carlo move set to anneal an optimal prediction of intermediates in a conformationally changing system.[37,38] Alternative methods have used the RMS differences between states as an order parameter to control change,[39,40] directly adding a new force that biases motion along the root mean square (RMS) gradient. We developed a method, called dynamic importance sampling (DIMS)[41–50] that uses concepts from stochastic differential equations[51] to create a family of independent transitions that together define the likelihood of different pathways and the kinetics of the transition with sufficient sampling. However, similar to

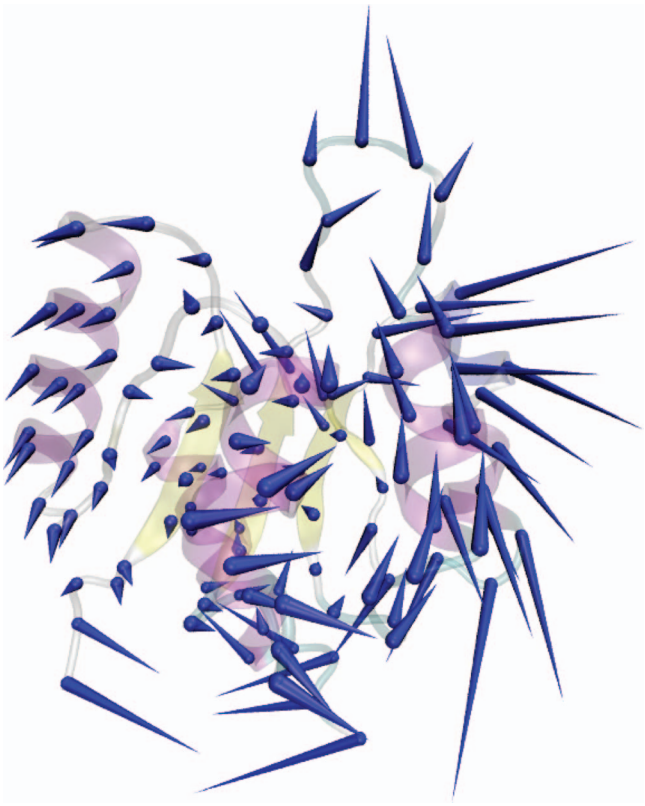

FIG. 1. Lowest frequency mode from PCA analysis for the inactive-state of NtrC.

the RMS based methods, the DIMS method requires a progress variable for use during the computed transitions and to create the biasing and its correction for an unbiased estimate of pathways, kinetics, and states.

In this paper, we describe the use of effective transfer entropy for the determination of a reduced set of degrees of freedom that can be used to define order parameters behind large scale conformational change. Our approach combines insights from the physics of non-linear time-series analysis, dimensionality reduction, and the chemical physics of protein motions on a complex energy surface to enable the dynamics of the complex system to define an order parameter candidate. This improves on other methods for the determination of order parameters where the candidate order parameter was inferred from empirical analysis of the static structure or simply assumed to correlate with the RMS between two different states. In the calculations to follow, we mainly use the receiver domain of nitrogen regulatory protein C (NtrC) (Fig. 1), in addition, we have performed steered molecular dynamics and checks of the implementation on the glucose-galactose binding protein (GGBP).

II. PRINCIPAL COMPONENT ANALYSIS

The method of principal component analysis (PCA) has been used in the analysis of protein motions for many years.[18,22,52–57] This approach depends on the determination of a set of effective collective modes that define the complex motions that have been seen in the dynamics.[52] While the initial excitement over the method as a way to sample on longer time-scales seems to have faded, there remains much effort to use this approach as a tool for the analysis of conformational change. A caution in that analysis has been the suggestion that PCA may lead to significant systematic error when there are multiple stable states separated by a large barrier.[53]

To compute the PCA modes a MD trajectory is used along with the determination of the average fluctuations in the simulation. Then, from the MD trajectory $\vec{q}(t) = (q_1(t), q_2(t), \ldots, q_{3N}(t)))$ of a protein with N atoms, the covariance matrix σ is built as follows:

$$\sigma = \langle(\vec{q}(t) - \langle\vec{q}(t)\rangle)(\vec{q}(t) - \langle\vec{q}(t)\rangle)\rangle, \tag{1}$$

where the brackets ($\langle\ldots\rangle$) denote time averages. The orthonormal basis vectors (principal components/PC) $\vec{\eta}_\alpha$ are determined by the eigenvalue problem $\lambda_\alpha\vec{\eta}_\alpha = \sigma$.

The lowest frequency modes from PCA are normally associated with slow, collective motions and have been used to try and predict intermediate states.[20] Figure 1 depicts the lowest frequency mode obtained by applying Eq. (1) and, solving the eigenvalue problem for our 600 ns trajectory of NtrC. On this plot the porcupine spines are located at the C_α atoms and their magnitude and direction shows the type of motion involved in the mode.

To connect the PCA modes with conformational transitions between two structures, we use the involvement coefficient. This is defined in the following way. For a given mode α, the involvement coefficients (IC) is

$$\nu_\alpha = \|\vec{\eta}_\alpha \cdot (\hat{q}^A - \hat{q}^B)\|, \tag{2}$$

where $\hat{q}^{A,B}$ indicates the set of normalized coordinates ($\hat{q}^{A,B} \cdot \hat{q}^{A,B} = 1$) that represent the active-state and inactive-state conformations, respectively. Therefore, the ICs measure the amount of overlap between a principal component and the direction defined by the displacement vector between structures. In the case of hinge-bending motions, PCA shows higher values for the ICs compared to those from more complex motions. For instance, in the case of Adenalyte Kinase (AdK), the ICs for the first two modes are 0.49 and 0.63, respectively,[58,59] thus it is possible to characterize most of the transition just by using these two modes. In a previous study, we explored the fact that the structural difference between the apo and the holo states of AdK are almost completely captured by linear correlations within our DIMS framework in order to elucidate ensembles of candidate pathways;[41] in a similar way, another study[60] was able to obtain intermediate states of the AdK transition by computing the normal modes from an elastic network model during short simulations ($\approx 10^1$ ps).

In the case of NtrC, the ICs are much lower (Fig. 4), in consequence the directions of the first PCs of both stable states are not pointing directly towards the other end state and therefore are not characterized by linear correlations. What is more, in another study, by using a set of order parameters based on observations of both stable structures, it was possible to obtain higher ICs values.[59] These order parameters involve only localized regions of the system and are proposed in an

orderly series of events, that is, by using a single order parameter it is not possible to characterize the whole transition between the two states.

One of the ideas behind our goal of looking for an order parameter is that a few degrees of freedom dominate part of or the entire transition, while the rest of the system would follow. Therefore finding an order parameter is equivalent to locating such leading modes. In this paper, we use an information theoretical approach to identify the leading modes by measuring the transfer entropy between pairs of residues. The more dominant residues are those that transfer the largest amount of entropy to the rest of the system.

III. INFORMATION FLOW IN PROTEINS

The networks of interactions between atoms and residues define the web of dependencies and patterns of dynamic coupling between domains in a protein, characterized by the directed flow of information spanning multiple spatial and temporal scales. An initial application of transfer entropy to DNA binding proteins was the first to apply the asymmetry of information transfer to protein molecular motions.[61] Let X be the time series for the center of mass of the ith residue and, $p(X)$ its probability distribution. Therefore it is possible to measure the average number of bits needed to optimally encode independent draws by using the Shannon entropy $H_X = -\sum_x p_x \log p(x)$,[62,63] where the sum extends over all the states that X can reach.

A. Transfer entropies

For a residue $j \neq i$ with a center of mass Y and, probability distribution $p(Y)$; one could say that its trajectory is independent of that of residue i if

$$p(y_{n+1}|y_n) = p(y_{n+1}|y_n, x_n), \tag{3}$$

where $p(y_{n+1}|y_n)$ is the conditional probability to find residue j at state y_{n+1} given its past $y_n, \ldots, y_1$ and $p(y_{n+1}|y_n, x_n)$ is the conditional probability to find residue j at state y_{n+1} given the past of both i and j. In the case where there is not a flux of information from X to Y then Eq. (3) is correct. On the other hand, in the event that there is flux of information in any direction, the divergence from correctness of Eq. (3) can be quantified by the Kullback-Leibler entropy[64] hence defining the transfer entropy,[65]

$$T_{X\to Y} = \sum p(y_{n+1}, y_n, x_n) \log \frac{p(y_{n+1}|y_n, x_n)}{p(y_{n+1}|y_n)}. \tag{4}$$

The transfer entropy between i and j is minimum and equal to zero when the two residues are independent and there is a maximum and equal to the entropy rate,

$$h_Y = -\sum p(y_{n+1}, y_n) \log p(y_{n+1}|y_n), \tag{5}$$

when the residues are completely coupled. In order to minimize artifacts within the time series, we use the normalized effective transfer entropy given by[66,67]

$$T^E_{X\to Y} = \frac{1}{h_Y}\left(T_{X\to Y} - \frac{1}{N_{\text{trials}}}\sum_{n=1}^{N_{\text{trials}}} T_{X_{\text{surrogate}}\to Y}\right), \tag{6}$$

where the second term is the average transfer entropy from N_{trials} surrogated samples of X, to Y.

B. The set Γ of most dominant residues

The total flux between two residues X and Y, can be calculated by the equation,

$$D_{X\to Y} = T^E_{X\to Y} - T^E_{Y\to X}. \tag{7}$$

Residues are selected according to the following rules: i is selected if $D_{X\to Y} > 0$, residue j is selected if $D_{X\to Y} < 0$ and, if $D_{X\to Y} = 0$ then no residue is selected. The set of most dominant residues Γ is then defined as the set of residues that follow the rules above and also that are above a fixed cutoff $|D_{X\to Y}| \geq D_{\text{cutoff}}$.

IV. EXPERIMENTS WITH GGBP

To verify that our implementation was correct, we performed analysis of coupled chaotic Ulam maps, for Henon maps and for autoregressive processes. In addition, as a more challenging test case, we used the Glucose-galactose binding protein (GGBP).[68] The two domains of GGBP exhibit a 0.5 rad hinge opening motion from one state to the other. The structure of the open state for an unbound glucose-galactose binding protein (GGBP) was crystallized by Borrock *et al.* (PDBID:2FW0) (Ref. 68) at 1.55 Å. For the purpose of testing we used both DIMS transitions and we applied a constant pulling force along the line determined by residues Phe:142 and Leu:144 to create a system with a known directional change (highlighted in green in Figure 2). The size of this force was very small, sufficiently so that inspection of unsteered versus steered simulations in visual molecular dynamics (VMD)[90] would look identical. Thus, the applied force was meant only to enable us to simulate a situation with a clear set of degrees of freedom that lead and others that should lag, rather than a simulation that was dramatically and artificially shifted too strongly to a non-equilibrium situation.

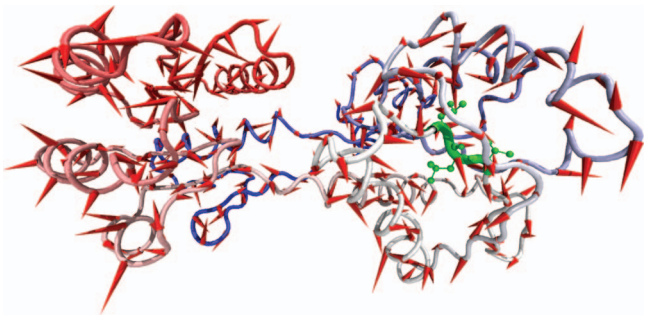

FIG. 2. A pulling force is applied along the line defined by residues Phe:142 and Leu:144 (highlighted in green). The arrows in red represent the modes determined from a PCA of the pulling trajectory.

As a comparison point for our methods, we performed a PCA analysis over the trajectory generated by this same pulling along the residues Phe:142 and Leu:144. With the transient nature of the pulling, it can be seen how PCA is unable to detect the pulling direction (Figure 2). We now describe the data treatment and some results from our initial testing for the transfer entropy analysis that we propose.

A. Time series treatment

The time series from MD describing the atomic motions of proteins are generally double precision real-valued entries. Previous work on the application of time-series analysis has shown that to determine the joint probability densities in Eq. (4), from real valued data is not only computationally expensive but unnecessary. For example, it has been shown that the amplitude of collective excitations, representing correlated global motions in the protein, samples multicentered distributions.[20] Therefore, although single or double precision arithmetic is necessary for the stability and accuracy of the simulations themselves, the accuracy of the analysis does not require this same level of precision. This can greatly aid the determination of the probability distributions while greatly reducing noise and increasing computational efficiency. We optimize our implementation by incorporating high performance computing techniques (massively parallel calculations extended over thousands of cores) and by applying dimensionality reduction and data mining techniques that we briefly describe in the following sections. In other applications of transfer entropies[61,66,67,69,70] discretization of the data is performed mainly by using symbolization techniques. In some cases the discretization maps the data to a single bit time series (spikes), for example in the situations where this analysis has been applied to data from neurophysiological in epilepsy patients.

1. Piecewise aggregate approximation (PAA)

A time series $\vec{q}(t) = (q_1(t), q_2(t), \ldots, q_{3N}(t)))$ of length n can be represented by a second time series $\vec{Q}(t') = (Q_1(t'), Q_2(t'), \ldots, Q_{3N}(t')))$ of length $w < n$, where each element $\vec{Q}(t')$ is computed according to[71]

$$\vec{Q}(t') = \frac{1}{\Delta t}\int_{t'}^{t'+\Delta t} \vec{q}(t)\mathrm{d}t, \tag{8}$$

where $\Delta t = n/w$. In other words, each vector of the time series $\vec{Q}(t')$ is simply the average, over a time range Δt, of the time series $\vec{q}(t)$. When Δt is constant, PAA can be seen as an attempt to approximate the original time series with a series of linear functions. Other approaches of PAA include using an adaptive mechanism to adjust Δt according to certain rules, i.e., defining a threshold such that $\sigma(t = T) < \langle q(t) - \langle q(t)\rangle_{t=1\ldots T}\rangle_{t=1\ldots T}$. For all calculations we set the time range $\Delta t = 0.1$ ns.

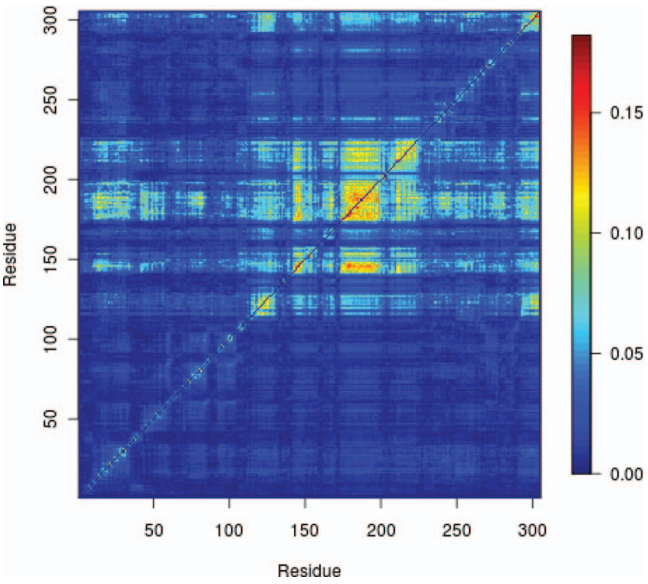

FIG. 3. Transfer entropies computed for DIMS trajectories for GGBP.

B. Transfer entropies from DIMS trajectories

In previous work, we generated a set of transitions for GGBP;[44] the simulations were carried out using CHARMM27FF with crossterm map (CMAP)[91] (Ref. 72) with our implementation of DIMS and using an implicit solvent model (ACE2).[73] The rotational and translational degrees of freedom were removed by rms fitting the target structure to the evolving system and, the alignment atoms were selected on the N-terminal domain (Residues 111 to 252 and, 293 to 305). By applying our transfer entropy analysis we were able to identify the key residues in the DIMS transition (Figure 3). The results show that the leading residues for the transition are located in the three-segment hinge that connects the N- and C-termini 3.

V. FINDING THE LEADING MODES ON NTRC

The structures of the inactive-state and active-state conformations of NtrC have been solved by NMR.[74–76] At room temperature NtrC samples both conformational states, however after phosphorylation the active states dominate the ensemble set of populations. Recent studies suggest that the transition pathway between the two conformations can be decomposed in a series of segmented progress variables (order parameters).[59] For this study both states were solvated in box of dimensions 20 Å × 20 Å × 20 Å with TIP3 waters, equilibrated for 15 ns; the total number of atoms, including solvent and ions, is 12 168 and 13 688 for the active and inactive states respectively. Production runs were performed for 600 ns using NAMD2.7 (Ref. 77) at NICS-Kraken. Analysis of the trajectories was executed using our code at NCSA-Abe/Lincoln.

A. Computing the modes

A key insight is that the atoms with the strongest leading effective transfer entropy can be used as a subset of degrees

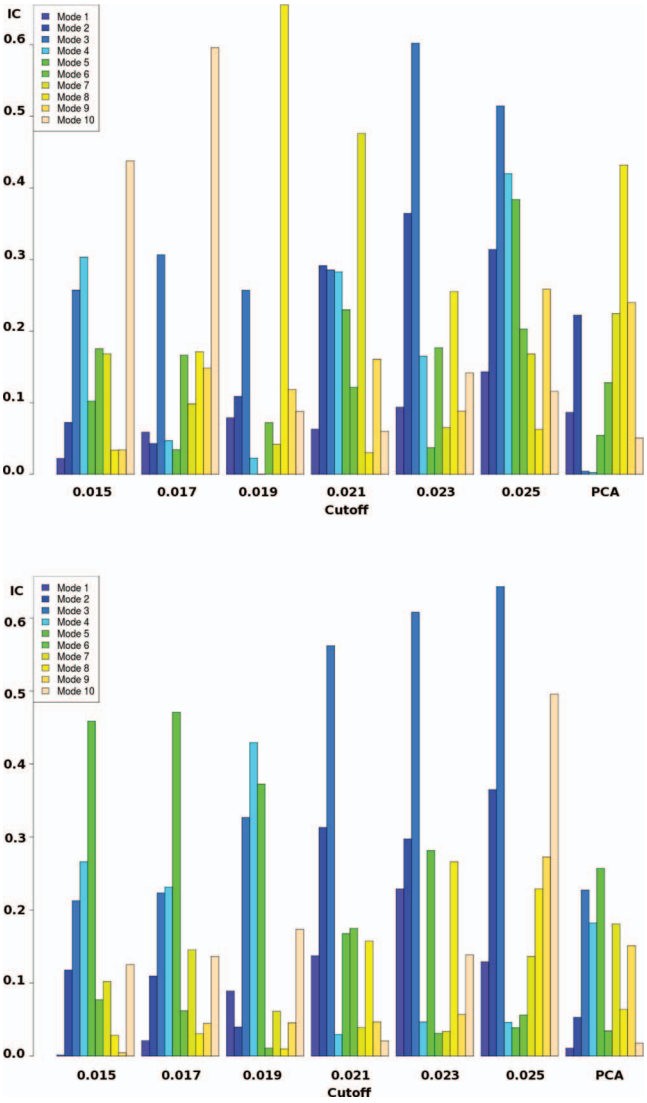

FIG. 4. Involvement coefficients for the two states of NtrC for different cutoffs D_{cutoff}.

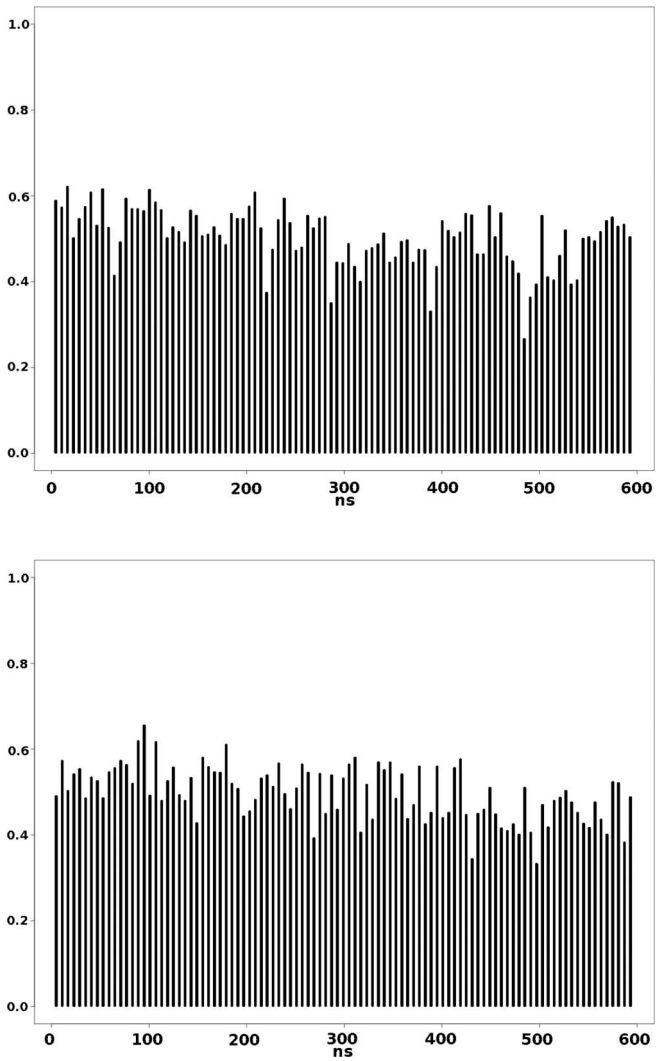

FIG. 5. Cumulative involvement coefficient as a function of time (ns) for the first $\alpha = 20$ modes.

of freedom to define collective modes that are new candidate order parameters. To accomplish this goal, once a cutoff and a time-length for the interrogation of the dynamics has been defined, is straightforward. The modes are determined by fluctuations of the leading effective transfer components and together describe a set of collective motions.

For the residues in the set Γ we compute the covariance matrix as in Eq. (1) over the full trajectory and obtain a set of modes $\vec{\eta}'_\alpha$. The involvement coefficients (Eq. (2)) for different values of the cutoff D_{cutoff} are presented in Figure 4. As the cutoff increases fewer residues are selected as dominant, however, the involvement coefficients are clearly increasing. This suggests that the most dominant modes $\vec{\eta}'_\alpha$ are pointing towards the end structure. Since the modes are transferring entropy to the entire system biasing along these modes would result in a collective bias for the entire system.

Since η_α is an orthonormal base we can define the cumulative involvement coefficient μ_α of the first α principal components as

$$\mu_\alpha = \sum_{i=1}^{\alpha} \nu_i^2, \tag{9}$$

and measure how much of the overall difference is accounted by the first α modes.

This last figure suggests that relatively short molecular dynamics simulations are converging onto the important degrees of freedom determined by the effective transfer entropy analysis (Fig. 5). It suggests that an algorithm for the use of the effective transfer entropy modes can be readily defined in CHARMM or other computer code. In that algorithm the lowest frequency modes would be the direction of biasing that is applied through DIMS or another approach (e.g., transition path sampling or targeted MD). The modes would be defined by a relatively short unbiased simulation and then followed by biasing for a similar amount of time to the mode determination. For example, this figure would suggest that 5 ns of sampling for the effective modes followed by 5 ns of sampling along the modes could

be used to improve the confidence that the most important intermediate states are being reached. This would then be repeated with unbiased sampling including light restraints on the backbone atoms to define a new set of effective transfer entropy modes. By continuing this process until the end state is reached, a transition pathway would be defined. If this process is then repeated for multiple starting points with various sampling windows and different random number seeds, along with a random selection of cutoffs and mode selections, then a good sampling of the intermediate space should be obtained.

VI. COMPARISON TO OTHER METHODS AND ASSUMPTIONS

It needs to be emphasized that the proposed method requires dynamic information for the calculation. The time-series comparisons that underlie the proposed method may be sensitive to the system choice and the total amount of information that is captured. Our results in this paper suggest that the total time needed to capture the leading and lagging degrees of freedom is much less than we might have initially assumed. But, the appropriate amount of time needed to collect dynamic information before a calculation of the effective transfer entropies is still an open question. It should be noted that there should be no pulling forces applied or other biasing if the suggestions for an order parameter are to correctly reflect motions from the unbiased state towards other directions. In a similar way, though we have assumed that two conformations are available for the calculation of the utility of the approach for conformational change, there is no restriction to two or more conformations for the calculation of the effective transfer entropy. Instead, the method outlined suggests that an effective order parameter that leads out from a particular conformational state may be defined by this approach and does not require that the order parameter, by itself, lead towards a specific endpoint. In that regard, then, the approach may also be helpful for sampling on multiple intermediate states that connect different larger conformational states. We have yet to fully test the utility of this thought, so the directions that the order parameter may lead could be coupled, as we outlined here, to computational efforts for understanding conformational change between states, or for the purposes of enhancing sampling away from one state and towards other, yet unexplored, states.

In addition, the method should be contrasted with other approaches that have attempted to determine subsets of states from long molecular dynamics data and then by extension to define intermediates and their connections to the states.[78] For example, the Head-Gordon group has suggested using instantaneous normal modes to define changes in the AdK system.[60] This relates to efforts using modes defined by essential dynamics analysis to sample on conformational change in the same system,[79] to work with Monte Carlo methods and collective modes[80,81] as well as to efforts using amplified collective modes.[82] Our work on AdK suggests that the conformational change is much simpler than in NtrC, and that the optimal order parameter may be easier.[41] Other groups have emphasized that cracking of secondary structural elements may be important for conformational change in AdK and should be considered in conformational change.[83] The current approach does not make any assumptions about the nature of the secondary structure of the domain motions needed for the conformational change.

In a related way, there is research from the Pande and other groups that is attempting to define Markov models based on long-dynamics simulations. In principle, the Markov models should also define the reduced descriptors needed for transitions between the Markov defined states. In practice, the approach outlined in this contribution may help with improving sampling between the states defined by the Markov models, since the intermediates may well be undersampled relative to the states themselves.[84]

Work within the Thorpe group has suggested that conformational change can be considered in terms of the pebble game and degrees of freedom that are available from a static structure.[85] In that regard the current contribution may be thought of as finding those most important subsets of degrees of freedom that lead the change, as opposed to defining solely the available subset. It may be fruitful to define more fully what types of correlated motions are most likely to lead to order parameters and what less likely. This would be another interesting extension of this work and would complement work on static structural analysis. In a somewhat similar manner, simpler chemical systems have suggested that algorithms can be designed to follow peaks and valleys on adiabatic surfaces based on a single structure to define transition states.[86,87] Others have suggested that an improved understanding of the connections between temperature and transitions would aid an understanding of the intermediates.[88]

Finally, there is a growing body of work addressing the limitations of principal component analysis and this suggests that there may be connections between the effective transfer entropy and the improved resolution of non-Gaussian analysis of long molecular dynamics trajectories.[26] While the nature of these connections remains to be understood, it suggests that the non-Gaussian components of motion may be the most important determinants of change out from the system. This could also tie into work from the Clementi group that is trying to find dimensionally reduced representations of dynamic conformation space.[89]

VII. CONCLUSIONS

A molecular understanding of how protein function is related to protein structure will require an ability to understand large conformational changes between multiple states. Unfortunately these states are often separated by high free energy barriers and within a complex energy landscape. This makes it very difficult to reliably connect, for example, by all-atom molecular dynamics calculations, the states, their energies, and the pathways between them. A major issue needed to improve sampling on the intermediate states is an order parameter – a reduced descriptor for the major subset of degrees of freedom – that can be used to aid sampling for the large conformational change. In this paper, we present a way to combine information from molecular dynamics using non-linear

time series and dimensionality reduction, in order to quantitatively determine an order parameter connecting two large-scale conformationally distinct protein states. The results presented show that the leading modes can be computed from short simulations. This new method suggests an implementation for molecular dynamics calculations that may dramatically enhance sampling of intermediate states.

[1] B. J. Berne, *Molecular Dynamics and Monte Carlo Simulations of Rare Events* (Academic, 1985).
[2] P. A. Kollman, "Free-energy calculations: Applications to chemical and biochemical phenomena," Chem. Rev. **93**, 2395–2417 (1993).
[3] C. Bartels and M. Karplus, "Probability distributions for complex systems: Adaptive umbrella sampling of the potential energy," J. Phys. Chem. B **102**, 865–880 (1998).
[4] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, and K. A. Dill, "Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations," J. Chem. Theory Comput. **3**, 26–41 (2007).
[5] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method," J. Comput. Chem. **13**, 1011–1021 (1992).
[6] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," J. Chem. Phys. **129**, 124105 (2008).
[7] X. Kong and C. L. Brooks, "Lambda-dynamics: A new approach to free energy calculations," J. Chem. Phys. **105**, 2414 (1996).
[8] B. Roux, T. Allen, S. Berneche, and W. Im, "Theoretical and computational models of biological ion channels," Q. Rev. Biophys. **37**, 15–103 (2004).
[9] N. Echols, D. Milburn, and M. Gerstein, "Molmovdb: Analysis and visualization of conformational change and structural flexibility," Nucleic. Acids Res. **31**, 478–482 (2003).
[10] V. Alexandrov, U. Lehnert, N. Echols, D. Milburn, D. Engelman, and M. Gerstein, "Normal modes for predicting protein motions: A comprehensive database assessment and associated web tool," Protein Sci. **14**, 633–643 (2005).
[11] B. Brooks and M. Karplus, "Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor," Proc. Natl. Acad. Sci. U.S.A. **80**, 6571–6575 (1983).
[12] F. Tama and Y.-H Sanejouand, "Conformational change of proteins arising from normal mode calculations," Protein Eng. **14**, 1–6 (2001).
[13] N. Go, T. Noguti, and T. Nishikawa, "Dynamics of a small globular protein in terms of low-frequency vibrational modes," Proc. Natl. Acad. Sci. U.S.A. **80**, 3696–3700 (1983).
[14] L. Skjaerven, S. M. Hollup, and N. Reuter, "Normal mode analysis for proteins," J. Mol. Struct.: THEOCHEM **898**, 42–48 (2009).
[15] H. Wako, M. Kato, and S. Endo, "Promode: A database of normal mode analyses on protein molecules with a full-atom model," Bioinformatics **20**, 2035–2043 (2004).
[16] W. Zheng and S. Doniach, "A comparative study of motor-protein motions by using a simple elastic-network model," Proc. Natl. Acad. Sci. U.S.A. **100**(23), 13253–13258 (2003).
[17] C. Chennubhotla and I. Bahar, "Signal propagation in proteins and relation to equilibrium fluctuations," PLOS Comput. Biol. **3**, 1716–1726 (2007).
[18] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, "Essential dynamics of proteins," Proteins: Struct., Funct., and Bioinf. **17**, 412–425 (1993).
[19] B. L. de Groot, D. M. F. van Aalten, A. Amadei, and H. J. C. Berendsen, "The consistency of large concerted motions in proteins in molecular dynamics simulations," Biophys. J. **71**, 1707–1713 (1996).
[20] A. E. Garcia, "Large-amplitude nonlinear motions in proteins," Phys. Rev. Lett. **68**(17), 2696–2699 (1992).
[21] S. Hayward, A. Kitao, and N. Go, "Harmonicity and anharmonicity in protein dynamics: A normal mode analysis and principal component analysis," Proteins: Struct., Funct., and Genet. **23**, 177–186 (1995).
[22] D. M. F. van Aalten, A. Amadei, A. B. M. Linssen, V. G. H. Eijsink, G. Vriend, and H. J. C. Berendsen, "The essential dynamics of thermolysin: Confirmation of the hinge-bending motion and comparison of simulations in vacuum and water," Proteins: Struct., Funct., and Bioinf. **22**, 45 (1995).
[23] J. S. Hub and B. L. de Groot, "Detection of functional modes in protein dynamics," PLOS Comput. Biol. **5**(8), e1000480 (2009).
[24] O. F. Lange and H. Grubmller, "Full correlation analysis of conformational protein dynamics," Proteins: Struct., Funct., and Bioinf. **70**(4), 1294–1312 (2008).
[25] A. Ramanathan, A. J. Savol, C. J. Langmead, P. K. Agarwal, and C. S. Chennubhotla, "Discovering conformational sub-states relevant to protein function," PLoS ONE **6**(1), e15827 (2011).
[26] A. J. Savol, V. M. Burger, P. K. Agarwal, A. Ramanathan, and C. S. Chennubhotla, "QAARM: quasi-anharmonic autoregressive model reveals molecular recognition pathways in ubiquitin," Bioinformatics **27**(13), i52–i60 (2011).
[27] R. Crehuet and M. J. Field, "A temperature-dependent nudged-elastic-band algorithm," J. Chem. Phys. **118**(21), 9563–9571 (2003).
[28] P. Eastman, N. G. Jensen, and S. Doniach, "Simulation of protein folding by reaction path annealing," J. Chem. Phys. **114**(8), 3823–3841 (2001).
[29] H. Huang, E. Ozkirimli, and C. B. Post, "Comparison of three perturbation molecular dynamics methods for modeling conformational transitions," J. Chem. Theory Comput. **5**(5), 1304–1314 (2009).
[30] S. Huo and J. E. Straub, "The MaxFlux algorithm for calculating variationally optimized reaction paths for conformational transitions in many body systems at finite temperature," J. Chem. Phys. **107**(13), 5000–5006 (1997).
[31] M. K. Kim, G. S. Chirikjian, and R. L. Jernigan, "Elastic models of conformational transitions in macromolecules," J. Mol. Graphics Modell. **21**(2), 151–160 (2002).
[32] L. R. Pratt, "A statistical method for identifying transition states in high dimensional problems," J. Chem. Phys. **85**(9), 5045–5048 (1986).
[33] W. Ren, E. Vanden-Eijnden, P. Maragakis, and W. E, "Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide," J. Chem. Phys. **123**(13), 134109 (2005).
[34] B. W. Zhang, D. Jasnow, and D. M. Zuckerman, "Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin," Proc. Natl. Acad. Sci. U.S.A. **104**(46), 18043–18048 (2007).
[35] W. Zheng, B. R. Brooks, and G. Hummer, "Protein conformational transitions explored by mixed elastic network models," Proteins **69**(1), 43–57 (2007).
[36] W. Zheng and B. R. Brooks, "Normal-modes-based prediction of protein conformational changes guided by distance constraints," Biophys. J. **88**(5), 3109–3117 (2005).
[37] P. G. Bolhuis, C. Dellago, and D. Chandler, "Reaction coordinates of biomolecular isomerization," Proc. Natl. Acad. Sci. U.S.A. **97**(11), 5877–82, 5 (2000).
[38] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, "Transition path sampling and the calculation of rate constants," J. Chem. Phys. **108**(5), 1964–1977 (1998).
[39] P. Maragakis and M. Karplus, "Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase," J. Mol. Biol. **352**(4), 807–822, 9 (2005).
[40] A. van der Vaart and M. Karplus, "Simulation of conformational transitions by the restricted perturbation–targeted molecular dynamics method," J. Chem. Phys. **122**(11), 114903 (2005).
[41] O. Beckstein, E. J. Denning, J. R. Perilla, and T. B. Woolf, "Zipping and unzipping of adenylate kinase: atomistic insights into the ensemble of open↔closed transitions," J Mol. Biol. **394**(1), 160–176 (2009).
[42] E. J. Denning and T. B. Woolf, "Cooperative nature of gating transitions in K^+ channels as seen from dynamic importance sampling calculations," Proteins: Struct., Funct., and Bioinf. **78**(5), 1105–1119 (2010).
[43] H. Jang and T. B. B. Woolf, "Multiple pathways in conformational transitions of the alanine dipeptide: An application of dynamic importance sampling," J. Comput. Chem. **27**(11), 1136–1141 (2006).
[44] J. R. Perilla, O. Beckstein, E. J. Denning, and T. B. Woolf, "Computing ensembles of transitions from stable states: Dynamic importance sampling," J. Comput. Chem. **32**(2), 196–209 (2011).
[45] T. Shimamura, S. Weyand, O. Beckstein, N. G. Rutherford, J. M. Hadden, D. Sharples, M. S. P. Sansom, S. Iwata, P. J. F. Henderson, and A. D. Cameron, "Molecular basis of alternating access membrane transport by the sodium-hydantoin transporter Mhp1," Science **328**(5977), 470–473 (2010).
[46] P. J. Stansfeld and M. S. P. Sansom, "Molecular simulation approaches to membrane proteins," Structure (London) **19**(11), 1562–1572 (2011).
[47] T. Woolf, "Path corrected functionals of stochastic trajectories: towards relative free energy and reaction coordinate calculations," Chem. Phys. Lett. **289**(5-6), 433–441 (1998).

[48]D. M. Zuckerman and T. B. Woolf, "Dynamic reaction paths and rates through importance-sampled stochastic dynamics," J. Chem. Phys. **111**(21), 9475–9484 (1999).
[49]D. M. Zuckerman and T. B. Woolf, "Efficient dynamic importance sampling of rare events in one dimension," Phys. Rev. E **63**(1), 016702 (2000).
[50]D. M. Zuckerman and T. B. Woolf, "Transition events in butane simulations: Similarities across models," J. Chem. Phys. **116**(6), 2586–2591 (2002).
[51]W. Wagner, "Unbiased Monte Carlo evaluation of certain functional integrals," J. Comput. Phys. **71**(1), 21–33 (1987).
[52]I. Andricioaei and M. Karplus, "On the calculation of entropy from covariance matrices of the atomic fluctuations," J. Chem. Phys. **115**, 6289–6292 (2001).
[53]M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten, "Principal component analysis and long time protein dynamics," J. Phys. Chem. **100**(7), 2567–2572 (1996).
[54]T. Horiuchi and N. Go, "Projection of Monte Carlo and molecular dynamics trajectories onto the normal mode axes: Human lysozyme," Proteins: Struct., Funct., and Bioinf. **10**(2), 106–116 (1991).
[55]T. Ichiye and M. Karplus, "Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations," Proteins: Struct., Funct., and Bioinf. **11**, 205–217 (1991).
[56]M. Karplus and J. Kushick, "Method for estimating the configurational entropy of macromolecules," Macromolecules **14**, 325–332 (1981).
[57]J. Schlitter, "Estimation of absolute and relative entropies of macromolecules using the covariance matrix," Chem. Phys. Lett. **215**(6), 617–621 (1993).
[58]K. Henzler-Wildman and D. Kern, "Dynamic personalities of proteins," Nature (London) **450**(7172), 964–972 (2007).
[59]M. Lei, J. Velos, A. Gardino, A. Kivenson, M. Karplus, and D. Kern, "Segmented transition pathway of the signaling protein nitrogen regulatory protein C," J. Mol. Biol. **392**(3), 823–836 (2009).
[60]C. Peng, L. Zhang, and T. H-Gordon, "Instantaneous normal modes as an unforced reaction coordinate for protein conformational transitions," Biophys. J. **98**(10), 2356–2364 (2010).
[61]H. Kamberaj and A. van der Vaart, "Extracting the causality of correlated motions from molecular dynamics simulations," Biophys. J. **97**(6), 1747–1755 (2009).
[62]F. M. Reza, *An Introduction to Information Theory* (Dover, 1994).
[63]C. E. Shannon, "A mathematical theory of communication," MD Comput. Comp. Med. Pract. **27**(4), 306–317 (1948).
[64]S. Kullback and R. A. Leibler, "On information and sufficiency," Ann. Math. Stat. **22**(1), 79–86 (1951).
[65]T. Schreiber, "Measuring information transfer," Phys. Rev. Lett. **85**(2), 461–464 (2000).
[66]B. Gourévitch and J. J. Eggermont, "Evaluating information transfer between auditory cortical neurons," J. Neurophysiol. **97**(3), 2533–2543 (2007).
[67]R. Marschinski and H. Kantz, "Analysing the information flow between financial time series. An improved estimator for transfer entropy," Eur. Phys. J. B **30**(2), 275–281 (2002).
[68]J. J. Borrok, L. L. Kiessling, and K. T. Forest, "Conformational changes of glucose/galactose-binding protein illuminated by open, unliganded, and ultra-high-resolution ligand-bound structures," Protein Sci. **16**(6), 1032–1041 (2007).
[69]M. Lungarella, A. Pitti, and Y. Kuniyoshi, "Information transfer at multiple scales," Phys. Rev. E **76**(5), 056117 (2007).
[70]M. Staniek and K. Lehnertz, "Symbolic transfer entropy," Phys. Rev. Lett. **100**(15), 158101 (2008).
[71]J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD 03* (ACM, 2003), p. 2.
[72]B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "CHARMM: The biomolecular simulation program," J. Comput. Chem. **30**(10), 1545–1614 (2009).
[73]M. Schaefer, C. Bartels, and M. Karplus, "Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model," J. Mol. Biol. **284**(3), 835–848 (1998).
[74]C. A Hastings, S.-Y. Lee, Ho. S. Cho, D. Yan, S. Kustu, and D. E. Wemmer, "High-resolution solution structure of the beryllofluoride-activated NtrC receiver domain," Biochemistry **42**(30), 9081–9090 (2003).
[75]D. Kern, B. F. Volkman, P. Luginbühl, M. J. Nohaile, S. Kustu, and D. E. Wemmer, "Structure of a transiently phosphorylated switch in bacterial signal transduction," Nature (London) **402**(6764), 894–898 (1999).
[76]B. F. Volkman, D. Lipson, D. E. Wemmer, and D. Kern, "Two-state allosteric behavior in a single-domain signaling protein," Science **291**(5512), 2429–2433 (2001).
[77]J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, "Scalable molecular dynamics with NAMD," J. Comp. Chem. **26**(16), 1781–1802 (2005).
[78]A. Jain, R. Hegger, and G. Stock, "Hidden complexity of protein free-energy landscapes revealed by principal component analysis by parts," J. Phys. Chem. Lett. **1**, 2769–2773 (2010).
[79]C. Snow, G. Qi, and H. Steven, "Essential dynamics sampling study of adenylate kinase: Comparison to citrate synthase and implications for the hinge and shear mechanisms of domain motions," Proteins: Struct., Funct., and Bioinf. **67**, 325–337 (2009).
[80]N. Kantarci-Carsibasi, T. Haliloglu, and P. Doruker, "Conformational transition pathways explored by Monte Carlo simulation integrated with collective modes," Biophys. J. **95**, 5862–5873 (2008).
[81]G. Miloshevsky and P. Jordan, "Open-state conformation of the KcsA K^+ channel: Monte Carlo normal mode following simulations," Structure **15**(12), 1654–1662 (2007).
[82]Z. Zhang, Y. Shi, and H. Liu, "Molecular dynamics simulations of peptides and proteins with amplified collective motions," Biophys. J. **84**, 3583–3593 (2003).
[83]O. Miyashita, J. N. Onuchic, and P. G. Wolynes, "Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins," Proc. Natl. Acad. Sci. U.S.A. **100**, 12570–12575 (2003).
[84]G. R. Bowman and V. S. Pande, "Protein folded states are kinetic hubs," Proc. Natl. Acad. Sci. U.S.A. **107**, 10890–10895 (2010).
[85]M. Lei, M. I. Zavodsky, L. A. Kuhn, and M. F. Thorpe, "Sampling protein conformations and pathways," J. Comput. Chem. **25**, 1133–1148 (2004).
[86]J. M. Bofill and J. M. Anglada, "Finding transition states using reduced potential-energy surfaces," Theor. Chem. Acc. **105**(6), 463–472 (2001).
[87]C. J. Cerjan and W. H. Miller, "On finding transition states," J. Chem. Phys. **75**(6), 2800–2806 (1981).
[88]R. Elber and D. Shalloway, "Temperature dependent reaction coordinates," J. Chem. Phys. **112**(13), 5539–5545 (2000).
[89]M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, "Determination of reaction coordinates via locally scaled diffusion map," J. Chem. Phys. **134**, 124116 (2011).
[90]W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," J. Mol. Graphics **14**, 33–38 (1996).
[91]A. D. MacKerell, Jr., M. Feig, C. L. Brooks III, "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations," J. Comput. Chem. **25**, 1400–1415 (2004).