

# Contact- and distance-based principal component analysis of protein dynamics

Cite as: J. Chem. Phys. **143**, 244114 (2015); <https://doi.org/10.1063/1.4938249>

Submitted: 23 October 2015 • Accepted: 07 December 2015 • Published Online: 28 December 2015

Matthias Ernst, Florian Sittel and Gerhard Stock



View Online



Export Citation



CrossMark

## ARTICLES YOU MAY BE INTERESTED IN

[Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates](#)

The Journal of Chemical Physics **141**, 014111 (2014); <https://doi.org/10.1063/1.4885338>

[Dihedral angle principal component analysis of molecular dynamics simulations](#)

The Journal of Chemical Physics **126**, 244111 (2007); <https://doi.org/10.1063/1.2746330>

[Principal component analysis on a torus: Theory and application to protein dynamics](#)

The Journal of Chemical Physics **147**, 244101 (2017); <https://doi.org/10.1063/1.4998259>



Webinar  
Quantum Material Characterization  
for Streamlined Qubit Development



Register now

# Contact- and distance-based principal component analysis of protein dynamics

Matthias Ernst, Florian Sittel, and Gerhard Stock<sup>a)</sup>

*Biomolecular Dynamics, Institute of Physics, Albert Ludwigs University, 79104 Freiburg, Germany*

(Received 23 October 2015; accepted 7 December 2015; published online 28 December 2015)

To interpret molecular dynamics simulations of complex systems, systematic dimensionality reduction methods such as principal component analysis (PCA) represent a well-established and popular approach. Apart from Cartesian coordinates, internal coordinates, e.g., backbone dihedral angles or various kinds of distances, may be used as input data in a PCA. Adopting two well-known model problems, folding of villin headpiece and the functional dynamics of BPTI, a systematic study of PCA using distance-based measures is presented which employs distances between  $C_\alpha$ -atoms as well as distances between inter-residue contacts including side chains. While this approach seems prohibitive for larger systems due to the quadratic scaling of the number of distances with the size of the molecule, it is shown that it is sufficient (and sometimes even better) to include only relatively few selected distances in the analysis. The quality of the PCA is assessed by considering the resolution of the resulting free energy landscape (to identify metastable conformational states and barriers) and the decay behavior of the corresponding autocorrelation functions (to test the time scale separation of the PCA). By comparing results obtained with distance-based, dihedral angle, and Cartesian coordinates, the study shows that the choice of input variables may drastically influence the outcome of a PCA. © 2015 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4938249>]

## I. INTRODUCTION

Classical molecular dynamics (MD) simulations facilitate a microscopic study of the structure, dynamics and function of biomolecular systems. To deal with the ever-growing amount of simulation data and obtain a concise but correct interpretation of simulation results, one often wants to systematically reduce the dimensionality by introducing a transformation from high-dimensional MD data  $\mathbf{r} = (r_1, \dots, r_N)$  to a low-dimensional reaction coordinate  $\mathbf{x} = (x_1, \dots, x_d)$ . While numerous methods have been suggested to this end,<sup>1–8</sup> probably the most simple and widely used technique is principal component analysis (PCA), which represents a linear transformation that diagonalizes the covariance matrix of  $\mathbf{r}$  and thus removes instantaneous linear correlations among the coordinates.<sup>9</sup> Ordering the eigenvalues decreasingly, it has been shown that a large part of the system's fluctuations in the high-dimensional vector space  $\{\mathbf{r}_n\}$  can be represented by the first few PCA eigenvectors or principal components (PCs)  $\{x_i\}$  of the system.<sup>10–16</sup>

Since the eigenvectors form a complete basis, the PC representation of the conformational space becomes exact when  $d = N$ . Providing a systematic means to approximate data by including only a few components, PCA is often used as a general preprocessing tool for high-dimensional data. When we find a time scale separation between the slow motion of the first few components (i.e., the “system”) and the fast motion of the remaining components (i.e., the “bath”), the first PCs may serve as a multidimensional reaction coordinate. In this way, the collective variables  $\{x_i\}$  may be

used to construct Langevin<sup>17–21</sup> or Markov state models<sup>22–27</sup> of the dynamics. Last but not least, PCs are often used to construct a free energy surface  $\Delta G(\mathbf{x}) = -k_B T \ln P(\mathbf{x})$ , where  $P$  is the probability distribution of the MD data along  $\mathbf{x}$ . Characterized by its minima (which represent the metastable conformational states of the system) and its barriers (which connect these states), the free energy landscape allows us to account for the pathways and their kinetics occurring in a biomolecular process.<sup>28–32</sup>

In a first step, we need to decide on suitable coordinates  $\{\mathbf{r}_n\}$  to represent the MD trajectory. Cartesian coordinates are convenient, because they are directly provided by the MD simulation, their kinetic energy is diagonal, and they allow us to calculate and easily illustrate any quantity of interest (e.g., the PCA eigenvectors). Commonly, either backbone atoms or  $C_\alpha$ -atoms are employed in a Cartesian coordinate PCA (cPCA). However, cPCA may yield spurious results in the case of large-amplitude motion (as occurring, e.g., in folding processes), since structural dynamics of flexible molecules necessarily results in a mixing of overall and internal motion.<sup>33</sup> To circumvent this problem, internal coordinates such as  $(\phi, \psi)$  backbone dihedral angles<sup>34–36</sup> may be used. Dihedral angles PCA (dPCA) has indeed proven useful, as it allows for a high resolution of metastable states in the dPCA free energy landscape.<sup>36–39</sup> However, dPCA may require many components, resulting in a relatively high dimensionality of the reaction coordinate.

Including only backbone atoms or backbone dihedral angles, standard cPCA and dPCA do not provide direct information on the side chains of a biomolecule. By considering distances or contacts between specific atoms, on the other hand, also structure and dynamics of side

<sup>a)</sup>Electronic address: stock@physik.uni-freiburg.de

chains may be taken into account. To this end, several authors have considered PCAs based on distances between closest lying atoms of each residue, hydrogen bonds or  $C_\alpha$ -atoms.<sup>40–45</sup> Since the dimensionality of distance-based PCA scales quadratically with the number of considered atoms, however, this approach is numerically expensive and thus prohibitive for larger systems. Moreover, the inclusion of a large number of distances may result in highly correlated coordinates, while it is advantageous for a PCA if relatively few and only weakly correlated input coordinates are used.<sup>39</sup>

In this work, we want to explore the virtues and shortcomings of distances as basis for a PCA description of protein dynamics. To this end, we employ distances between  $C_\alpha$ -atoms as well as distances between inter-residue contacts of the protein. To reduce the number of degrees of freedom, we focus on contacts that are present in the native state of a protein. While native contacts are obviously important to describe small-amplitude motions of a folded protein, they have been recently shown to also largely determine the folding pathways.<sup>46</sup> Moreover, several groups have successfully used the fraction of native contacts as one-dimensional reaction coordinate.<sup>46–48</sup> In a similar vein, we only include  $C_\alpha$ -distances that are shorter than a certain threshold in the native state. Adopting two well-known model proteins, villin headpiece (HP35) and bovine pancreatic trypsin inhibitor (BPTI), for which long (up to ms) all-atom MD trajectories are available from D. E. Shaw research,<sup>49,50</sup> we compare the performance of various versions of a contact-based PCA (henceforth, termed “conPCA”) and  $C_\alpha$ -distance-based PCA (termed “ $C_\alpha$ PCA”) to the more established methods cPCA and dPCA.

## II. THEORY AND METHODS

### A. MD details

#### 1. Villin headpiece

HP35 is a 35-residues protein fragment that represents a standard model of ultrafast protein folding.<sup>51–55</sup> It consists of a hydrophobic core with three helices (residues 3–10, 14–19 and 22–32) that are connected via two unstructured loops (Fig. 1(a)). To study the folding of HP35, extensive all-atom equilibrium MD simulations of wild-type HP35 and its mutants were carried out by Piana *et al.*<sup>49</sup> at various temperatures, employing the Amber ff99SB\*-ILDN force field<sup>56–58</sup> and the TIP3P explicit water model.<sup>59</sup> Here, a  $\approx 300 \mu\text{s}$  segment of the fast folding double mutant (HP35 NleNle) at 360 K was adopted. According to our definition below, we identified 53 native contacts from the crystal structure (pdb 2F4K),<sup>54</sup> which are depicted in Fig. 1(b).

#### 2. Bovine pancreatic trypsin inhibitor

BPTI is a well-studied 58-residue protein that exhibits small-amplitude functional motion. According to DSSP analysis<sup>60</sup> of the crystal structure (pdb 5PTI),<sup>61</sup> it contains a  $3_{10}$  helix (residues 3–6), two  $\beta$ -sheets (residues 18–24 and 29–35) connected by a turn (residue 25–28) and an

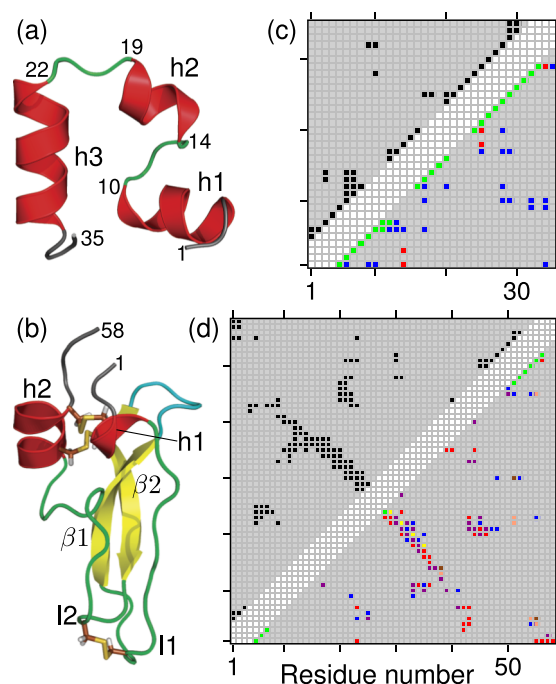


FIG. 1. Left: Structures of (a) HP35 and (b) BPTI, where secondary structure elements are color coded as helix (red),  $\beta$ -sheet (yellow), turn (cyan), loop (green) and termini (gray). Disulfide bridges are shown as sticks. Right: Contact maps of the reference structures for (c) HP35 and (d) BPTI. The upper triangle shows  $C_\alpha$ -contacts, the lower triangle heavy-atom contacts, using a distance threshold of 8.0 Å and 4.5 Å, respectively. Contacts indicated by white fields are found too close in sequence, gray fields too far away in distance. The type of contact is color coded as  $C_\alpha$  (black), helical (green), hydrogen bond (red), hydrophobic (blue),  $\beta$ -sheet (yellow), weak H-Bond (purple), and disulfide bridge (brown).

$\alpha$ -helix (residues 48–55), see Fig. 1(c). Three short bends and an isolated  $\beta$ -bridge can also be identified, but we consider them as part of the two long loop regions connecting the helices with the  $\beta$ -strands (spanning residues 7–17 and 36–47, respectively). The whole structure is stabilized by three disulfide bonds. To study the functional dynamics of BPTI, Shaw *et al.*<sup>50</sup> performed a  $\approx 1$  ms long all-atom equilibrium MD simulation at 300 K, using the AMBER ff99SB force field<sup>56</sup> and the TIP4P-Ew<sup>62</sup> water model. Using the reference structure of Shaw *et al.*<sup>50</sup> based on the crystal structure (pdb 5PTI),<sup>61</sup> we identified 108 native contacts for the present analysis (Fig. 1(d)).

### B. PCA

The correlated internal motion of a system with  $N$  degrees of freedom can be described by the covariance matrix,

$$\sigma_{mn} = \langle (r_m - \langle r_m \rangle) (r_n - \langle r_n \rangle) \rangle, \quad (1)$$

where  $r_1, \dots, r_N$  denote the input coordinates and  $\langle \dots \rangle$  represents the average over all sampled conformations. Diagonalization of this covariance matrix results in  $N$  eigenvectors ( $\mathbf{v}^{(i)}$ ) and eigenvalues ( $\lambda_i$ ) which describe the modes of the collective motion and their respective amplitudes. The PCs

$$x_i = \mathbf{v}^{(i)} \cdot \mathbf{r} \quad (2)$$

are the projections of the coordinates  $\mathbf{r}$  onto the eigenvectors and may be used to construct a reaction coordinate.

Instead of commonly employed Cartesian coordinates, one may also use  $(\phi, \psi)$  dihedral angles of the protein backbone as input coordinates  $\{r_n\}$ . Being circular variables, however, the angles first need to be transformed to a coordinate space with linear metric (e.g., a vector space with the usual Euclidean distance). This can be achieved by the transformation<sup>63</sup>

$$q_{2n-1} = \cos \varphi_n, \quad q_{2n} = \sin \varphi_n, \quad (3)$$

where  $n = 1, \dots, M$  with  $M$  being the total number of dihedral angles considered. The doubling of variables in dPCA can be explained by considering a complex-valued version (i.e.,  $q_n = z_n = e^{i\varphi_n}$ ), which also showed that dPCA amounts to a one-to-one representation of the original angle distribution.<sup>36</sup> Details of the cPCA and dPCA on HP35 and BPTI are given in Refs. 64 and 33, respectively.

### C. Contact- and distance-based PCA

There are numerous definitions of protein residue-residue contacts, which differ in the choice of atoms (e.g.,  $C_\alpha$  atoms or closest lying atoms of each residue), the distance cutoff up to which a contact is considered to be formed (typically between 4 and 8 Å), and what type of contacts are included (e.g., all possible contacts, all hydrogen bonds, or tertiary contacts only).<sup>40–45</sup> As explained in the Introduction, we find it advantageous to restrict the PCA to the native contacts of the protein.<sup>46</sup> We consider a contact as formed if the distance between the closest lying heavy atoms of each residue is less than 4.5 Å (Ref. 65)

$$D_v = \min(|\vec{r}_{i,k} - \vec{r}_{j,l}|) \leq 4.5 \text{ Å}, \quad (4)$$

where the indices  $k$  and  $l$  run over all heavy atoms of the selected residue pair  $(i, j)$ . Moreover, we discard contacts between residues that are less than four residues apart, thereby omitting short-range contacts as, e.g., in helical structure elements. We note that distances according to (4) can be calculated for the reference structure (i.e., only once), or for every frame of the MD trajectory. As both methods give quite similar results for the considered systems, the former approach seems sufficient. All contact and distance calculations were done using the MDAnalysis framework.<sup>66</sup>

Adopting above definitions, Fig. 1 shows the contact maps of (c) HP35 and (d) BPTI, where we color-coded the type of the respective contact. HP35 clearly shows secondary structure contacts along the diagonal, which reflects the three  $\alpha$ -helices (residues 4–10, 15–19, and 23–32). Moreover, we find several tertiary contacts which are either contacts of the hydrophobic core or hydrogen bonds. The contact map of BPTI shows secondary structure contacts due to the  $\beta$ -sheets and the two short helices as well as tertiary contacts between both the region of clearly defined secondary structure and the less structured loop regions.

Alternatively, we also considered distances between  $C_\alpha$  atoms of the crystal structure, including all  $C_\alpha$ -distances that are shorter than 8 Å. Figure 1 reveals that the resulting contact maps based on  $C_\alpha$ -distances and heavy-atom contacts are very similar. This is especially the case for BPTI with its

rather stable and closely packed structure, where we find more  $C_\alpha$ -contacts than heavy-atom contacts. For HP35, on the other hand, some of the side-chain contacts forming the hydrophobic core are not found when we use our criteria for  $C_\alpha$ -contacts.

Using the distances defined by Eq. (4), we calculate the covariance matrix,

$$\sigma_{\mu,\nu} = \langle (D_\mu - \langle D_\mu \rangle) (D_\nu - \langle D_\nu \rangle) \rangle, \quad (5)$$

which defines the contact-based PCA (conPCA). Similarly, we employ  $C_\alpha$ -distances to calculate the corresponding covariance matrix that defines the  $C_\alpha$ -based PCA ( $C_\alpha$ PCA). We also tried various other variants. For example, we performed a  $C_\alpha$ -based PCA including *all* existing  $C_\alpha$ -distances (see supplementary material<sup>70</sup>). Moreover, we calculated the covariance matrix using *reciprocal* distances (termed iconPCA), which shifts the focus from the large-scale motions (preferably seen by conPCA) to small motions around the native contact distances.<sup>67</sup> Instead of directly using the distances  $D_\nu$  to calculate the covariance matrix, one may also discretize the distances used for conPCA by employing the same criterion as in (4) and setting  $D_\nu \equiv 1$  if  $|\vec{r}_{i,k} - \vec{r}_{j,l}| \leq 4.5 \text{ Å}$  and  $D_\nu \equiv 0$  otherwise.<sup>45</sup> As we found that the resulting jumps of the discretized trajectory typically introduce additional noise to the observables and result in a reduced resolution of the free energy surfaces (data not shown), we discarded this option.

## III. RESULTS

In the following, we adopt two well-established model problems, the folding of HP35 and the functional dynamics of BPTI, to study the performance of the various versions of PCA introduced above.

### A. HP35

We begin with considering the (normalized) cumulative fluctuations  $V_d = \sum_{i=1}^d \lambda_i / (\sum_j \lambda_j)$  covered by a PCA using  $d$  PCs, where  $\lambda_i$  denotes the  $i$ th eigenvalue of the PCA. Since a reaction coordinate  $\mathbf{x} = (x_1, \dots, x_d)$  should represent a sufficiently large part of the motion of the system, the variance  $V_d$  of the  $\{x_i\}$  should contain a large fraction of the collective variance of the  $\{r_n\}$ . Figure 2(a) shows that the cumulative fluctuations obtained by dPCA converge relatively slowly with the number of PCs included, e.g., it takes about 20 PCs to cover 60% of the overall variance. This might be caused by residual nonlinear correlations of the PCs,<sup>39</sup> as well as by the fact that dihedral angles contain very detailed structured information which may be difficult to cover by a small number of PCs. The cumulative fluctuations of the distance-based PCAs are found to increase much more rapidly with the number of included PCs. To cover 60% of the overall variance, it only takes four PCs for  $C_\alpha$ PCA and iconPCA, and only two for conPCA. We note that no results are shown for the Cartesian coordinate PCA, since cPCA breaks down in the case of large-amplitude folding processes due to the mixing of overall and internal motion.<sup>33</sup>



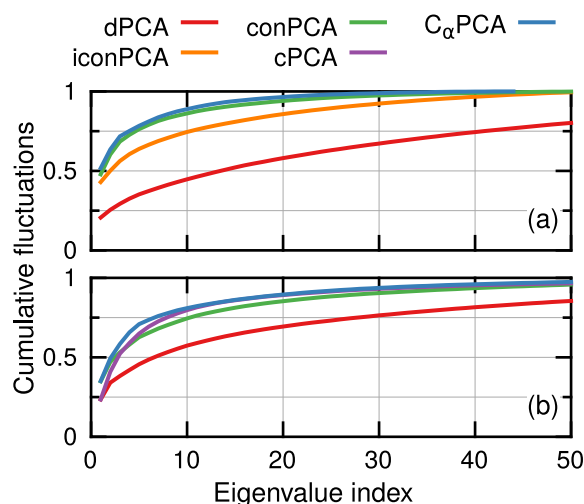


FIG. 2. Relative cumulative fluctuations of the first 50 PCs, obtained for (a) HP35 and (b) BPTI, using various versions of PCA.

The folding dynamics of a protein is typically associated with rare transitions between conformational states, which are separated by energy barriers that are significantly larger than the thermal energy  $k_B T$ . Hence, the free energy landscape of the folding of HP35 should discriminate several minima corresponding to these metastable states. To get an overview of the conformational distribution associated with the folding of HP35, Fig. 3 (top) shows two-dimensional free energy surfaces obtained by the different PCA variants, using the two PCs that yield the best structural resolution of the energy landscape. As discussed in Ref. 64, the energy landscape of HP35 consists of the entropic unfolded basin  $U$  where the restructuring of the protein takes place, the intermediate basin  $I$  which is connected to  $U$  via the rate-limiting  $U \rightarrow I$  transition state reflecting the formation of helix-1, and the native basin  $N$  containing a state close to the NMR structure. Employing recently developed clustering methods<sup>68,69</sup> Fig. 3 (bottom) demonstrates that all considered PCA methods

are able to discriminate the unfolded state ( $U$ ) from the folded state (comprising  $N$  and  $I$ ). Remarkably, the overall folding/unfolding is always mediated by the first PC, while higher PCs describe further substates in the folded state.

A closer examination of the resulting clusters reveals, however, that the underlying molecular structure of the conformational states may be different for the various PCAs. For example, the intermediate state  $I$  differs from the native state  $N$  mainly in residue 3, which hardly changes the distances of HP35 but results in a somewhat larger flexibility of this residue.<sup>64</sup> As a consequence, native and intermediate conformations are found to partially overlap in the energy landscapes of the distance-based PCAs, which therefore exhibit less details than the dPCA landscape. This finding is supported by the one-dimensional free energy profiles  $\Delta G(x_i)$  of the PCs shown in Fig. S1.<sup>70</sup> While the distance-based PCAs result in about three PCs with several minima, dPCA yields seven structured energy profiles. ConPCA and iconPCA are found to give quite similar results, although the reciprocal distances used by the latter method appears to somewhat enhance the overall resolution. Finally, we also considered a  $C_\alpha$ -based PCA, where *all* (not only the preselected)  $C_\alpha$ -distances are taken into account. Interestingly, Fig. S3<sup>70</sup> shows that the state resolution of the resulting energy landscapes is clearly minor compared to the results of  $C_\alpha$ PCA in Fig. 3(d). Hence, a reasonable preselection of the degrees of freedom may reduce the “noise” of the data, leading to an improved resolution of the PCA.

To get an intuitive picture of the motion described by a PC, it is instructive to draw molecular structures along this motion. This is straightforward when Cartesian coordinates are used (since the PCA eigenvector is expressed in terms of atomic coordinates) but more involved in the case of internal coordinates (which do not necessarily account for the position of all atoms). In the case of conPCA, the structural evolution along some PC can be easily illustrated by considering the contacts that mainly change during this motion. As an example, Fig. 4(a) shows the squared elements

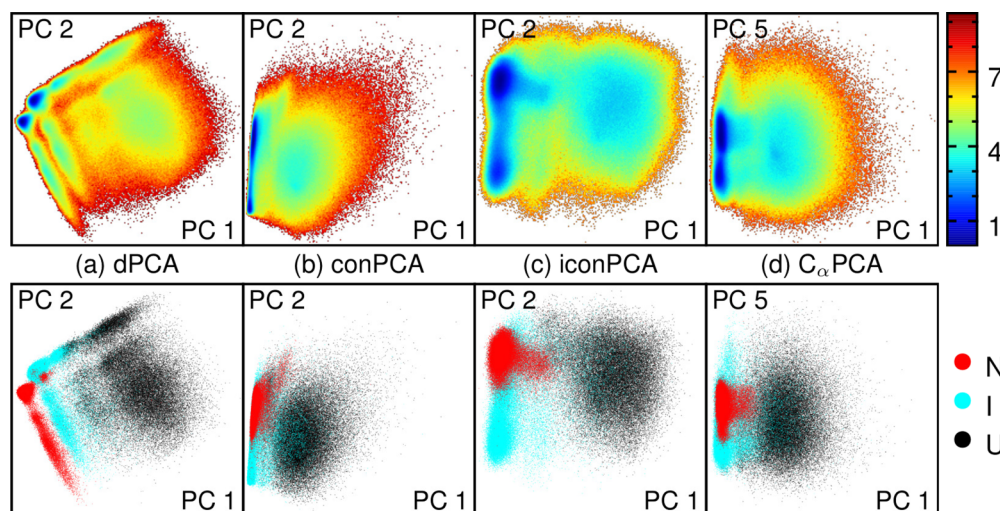


FIG. 3. Top: Two-dimensional representations of the free energy landscapes  $\Delta G(x_i, x_j)$  (in units of  $k_B T$ ) of HP35, as obtained from (a) dPCA, (b) conPCA, (c) iconPCA and (d)  $C_\alpha$ PCA, respectively. Bottom: Corresponding results of a dynamical clustering method based on dPCA of the folding trajectory of HP35, showing the entropic unfolded basin ( $U$ ) in black, the intermediate state ( $I$ ) in blue and the native state ( $N$ ) in red color.

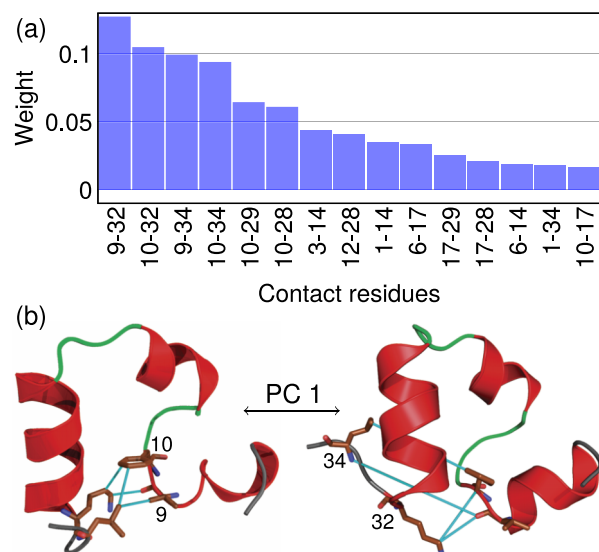


FIG. 4. (a) Squared elements of the normalized first eigenvector  $\{v_i\}$  of conPCA, with index  $i$  labeling the considered contacts of HP35. Ordered decreasingly, only components that constitute up to 80% of the norm are shown. (b) Representative structures of HP35 discriminated by PC1, showing the intermediate state  $I$  (left) and the unfolded state  $U$  (right). The four most important contacts that change along the first PC are highlighted in blue.

of the normalized eigenvector  $\{v_i\}$  of the first PC, where the index  $i$  labels the contacts considered in the conPCA. While numerous contacts vary slightly, the main changes occur for the contacts between helix-1 and helix-3 and between helix-1 and helix-2, that is, the tertiary contacts that keep the hydrophobic core of the protein together. This confirms that the first PC indeed describes the overall folding/unfolding transition of HP35. As the sign of all elements  $\{v_i\}$  is the same, all contacts are simultaneously formed or broken along this motion. (Opposite signs of the  $\{v_i\}$  indicate the formation of one contact while another one is broken.) As a further illustration, Fig. 4(b) shows representative molecular structures of folded and unfolded HP35 and indicates the most important contacts that change along the first PC.

While the quantities studied so far are concerned with statistical properties of the MD data (such as the conformational distribution), we now wish to consider observables that describe the dynamics of the considered system. As explained in the Introduction, a desired property of a set of suitable reaction coordinates generated by PCA is a time scale separation between the slow dynamics of the system coordinates  $\{x_i\}$  and the fast dynamics of the remaining bath coordinates. This property can be tested via the decay times of the position autocorrelation function,

$$C_i(t) = \langle \delta x_i(t) \delta x_i(0) \rangle / \langle \delta x_i^2 \rangle, \quad (6)$$

where  $\delta x_i = x_i - \langle x_i \rangle$ . That is, the first few PCs representing the system coordinates should decay much slower than the remaining PCs representing the bath coordinates. To assess the ability of the various PCA variants to achieve such a time scale separation, Fig. 5 shows the autocorrelation function of the first seven PCs. In all cases, we find that the first PC reflecting the folding and unfolding of HP35 decays on a time scale of about  $2.5 \mu\text{s}$ . This decay is at least an order of magnitude

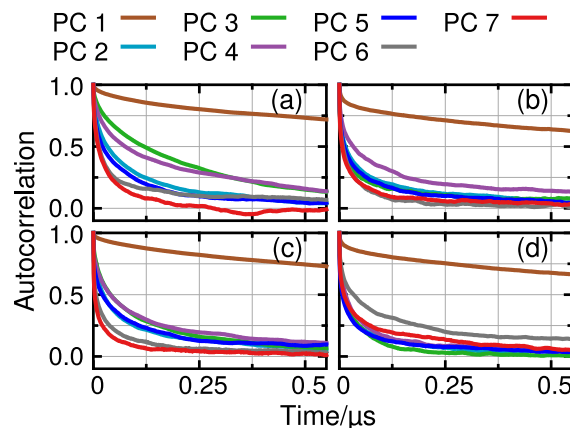


FIG. 5. Autocorrelation functions of the first seven PCs of HP35, as obtained from (a) dPCA, (b) conPCA, (c) iconPCA, and (d)  $C_\alpha$ PCA, respectively.

slower than the decay of the next few PCs, which account for transitions between native and intermediate conformational states in the folding of HP35.<sup>64</sup> While the decay times of these PCs are roughly the same for conPCA, iconPCA and  $C_\alpha$ PCA, the dPCA reveals a few somewhat slower PCs which are associated with structured free energy profiles shown in Fig. S1.<sup>70</sup> The autocorrelation functions  $C_i(t)$  of higher ( $i \geq 6$ ) PCs are found to decay on a much faster (ns) time scale.

Fig. 5 shows that various PCAs of the folding of HP35 yield quite similar autocorrelation decay times for the first PC. Moreover, the corresponding free energy curves (Figs. 3 and S1<sup>70</sup>) are roughly similar along this component, which also contains most of the variance of the system (Fig. 2(a)). This is remarkable in the light of the fact that dPCA is based on “local” coordinates (i.e., backbone dihedral angles) while the distance-based PCAs are based on “global” coordinates (i.e., residue-residue contacts). To highlight this similarity, Fig. 6 shows the time trace of the first PC as obtained for the various methods. Interestingly, we find that the time evolution is almost identical in all cases and, moreover, also matches the time evolution of the root mean square deviation (RMSD) of the system. Obviously, the overall folding/unfolding motion largely dominates the structural dynamics of HP35, such that it is recovered by any reasonable one-dimensional reaction coordinate. We note that this only holds for PC1, i.e., no comparable similarity is found for higher PCs.

## B. BPTI

While HP35 serves as a standard example of protein folding, BPTI is a well-established model to study functional dynamics. To investigate how this type of dynamics can be described by PCA, we again compare the above defined observables obtained from dPCA (using dihedral angles), conPCA (using native contacts), and  $C_\alpha$ PCA (using selected  $C_\alpha$ -distances). Since the small-amplitude motion of the functional dynamics of BPTI should allow for valid separation of overall and internal motion,<sup>33</sup> we also performed a cPCA (using Cartesian coordinates of backbone atoms). On the other hand, we did not include iconPCA (using reciprocal distances) in the discussion, since it again yields very similar results as conPCA (data not shown).

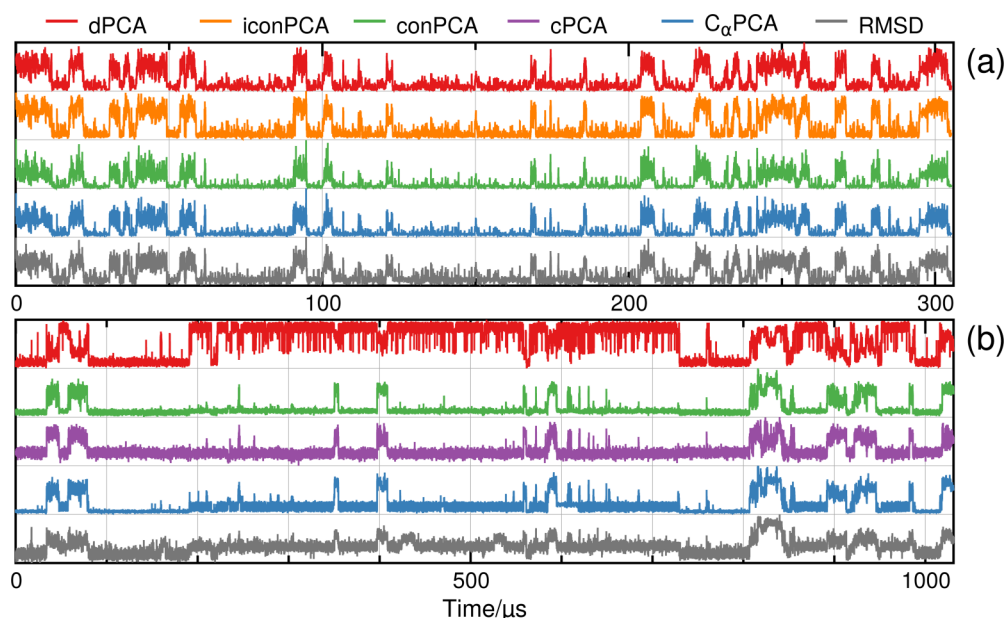


FIG. 6. Time evolution of the first PC obtained for (a) HP35 and (b) BPTI. Shown are results from dPCA (red), conPCA (green), iconPCA (orange), cPCA (purple), and  $C_\alpha$ PCA (blue). The gray line displays the RMSD with respect to the native structure of the system.

Beginning the discussion with the cumulative fluctuations shown in Fig. 2(b), we find that again dPCA converges relatively slowly with the number of PCs, i.e., 10 PCs are needed to cover 60% of the overall variance. To this goal,  $C_\alpha$ PCA needs 6 PCs, while conPCA and cPCA require only 4 PCs. While the overall trend is similar to the case of HP35, we note that for BPTI the difference between dPCA and the distance-based PCA variants is not as large.

We next consider one-dimensional free energy profiles obtained for the various methods, in order to test which PCs show an energy landscape with several minima. Figure S2<sup>70</sup> reveals that dPCA yields ten,  $C_\alpha$ PCA four, conPCA five, and cPCA three PCs with structured free energy profiles. Choosing from Fig. S2<sup>70</sup> the two most important PCs for each method, Fig. 7 shows two-dimensional free energy surfaces which

reflect the conformational distribution of the 1 ms trajectory of BPTI. Judged by the number of well distinguishable states, dPCA provides the highest resolution. Employing again our clustering methodology<sup>68,69</sup> on the dPCA data set, we are able to discriminate twelve metastable conformational states. The two distance-based methods, conPCA and, in particular,  $C_\alpha$ PCA, discriminate most of the states monitored by dPCA but cannot resolve all details of the conformational distribution. We also considered again a PCA that includes *all*  $C_\alpha$ -distances (Fig. S4),<sup>70</sup> which for BPTI gave quite similar results as  $C_\alpha$ PCA including only selected  $C_\alpha$ -distances. Finally, cPCA discriminates only two out of twelve states, which is presumably due to residual mixing of overall and internal motion.<sup>33</sup> As discussed previously,<sup>33</sup> the metastable states of BPTI differ mostly in the conformations of the

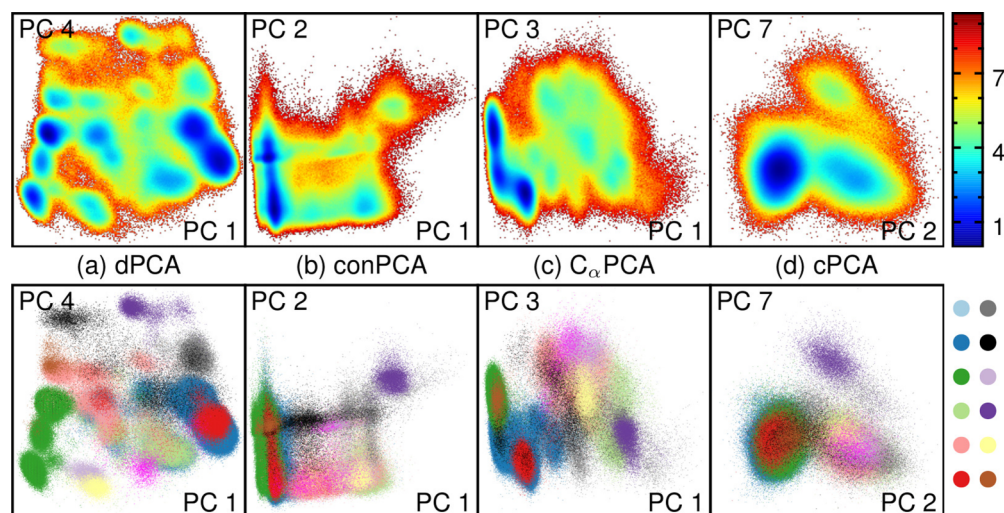


FIG. 7. Top: Two-dimensional representations of the free energy landscape (in units of  $k_B T$ ) of BPTI, as obtained from (a) dPCA, (b) conPCA, (c)  $C_\alpha$ PCA, and (d) cPCA, respectively. Bottom: dPCA-based clustering of the BPTI trajectory yields twelve metastable conformational states which are drawn in different colors in the respective PC spaces.



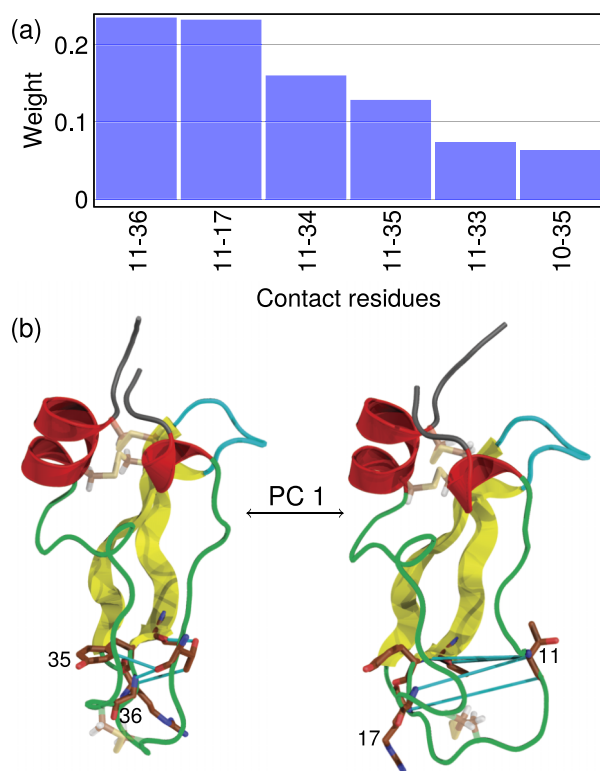


FIG. 8. (a) Squared elements of the normalized first eigenvector  $\{v_i\}$  of conPCA, including components up to 90% of total norm. The index  $i$  labels the considered contacts of BPTI. (b) Representative molecular structures before and after a jump in this PC, indicating the four most important contacts by blue lines.

first and second flexible loop, while the two  $\beta$ -sheets and the  $\alpha$ -helix remain relatively stable. Indeed, Fig. 8 nicely demonstrates that the first PC of conPCA describes the making and breaking of contacts between these two loops.

As seen in Fig. S2,<sup>70</sup> the metastable states of BPTI may be separated by large barriers, which render the transitions between these states a rare event. In fact, rare events exist which are not well sampled by the 1 ms trajectory of BPTI, e.g., the change of the RMSD at  $\sim 820 \mu\text{s}$  in Fig. 6(b) reflects a singular conformational transition. This nonstationarity requires some caution in the interpretation of the dynamics of slow PCs. Considering the autocorrelation functions, Fig. 9 indeed shows that the first few PCs of all considered PCAs decay only within several microseconds. The time scale separation achieved by the various methods, though, is found to differ significantly. While conPCA and cPCA show a single slowly decaying PC, dPCA, and  $C_\alpha$ PCA identify several slow PCs. In particular,  $C_\alpha$ PCA clearly shows four slow PCs which are well separated from the remaining degrees of freedom. In all cases, autocorrelation functions of higher ( $\geq 7$ ) PCs are found to decay on a much faster (ns) time scale.

We finally compare again the time evolution of the first PC as obtained by the various methods. Figure 6(b) shows that the time traces of the distance-based PCAs are very similar and also match the RMSD of the system (except for minor details, e.g., at  $420 \mu\text{s}$ ). While most features are also observed by dPCA, this method additionally shows a prominent transition at 190 and  $720 \mu\text{s}$ . A closer analysis reveals that this transition

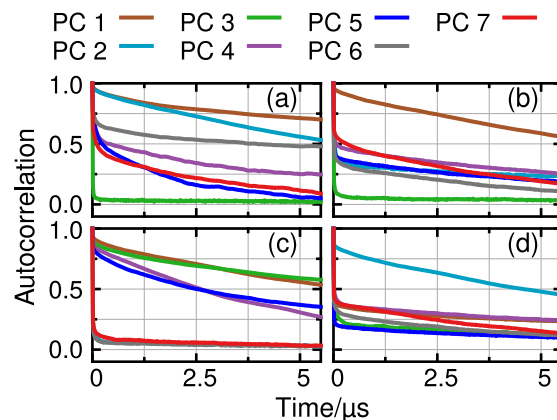


FIG. 9. Normalized autocorrelation functions of first 7 PCs of BPTI, as obtained from (a) dPCA, (b) conPCA, (c)  $C_\alpha$ PCA, and (d) cPCA, respectively.

is caused by a flipping of the  $C_{14}^\beta-S_{14}-S_{38}-C_{38}^\beta$  dihedral angle of the disulfide bridge between residues 14 and 38, which hardly affects the considered distances of BPTI.

#### IV. CONCLUDING REMARKS

Adopting two well-established biomolecular model problems, it has been demonstrated that the choice of input coordinates may drastically influence the outcome of a PCA. As an alternative to Cartesian coordinate PCA and backbone dihedral angle PCA considered previously, we have performed a systematic study of PCAs using distance-based measures. While this approach seems prohibitive for larger systems due to the quadratic scaling of the number of distances with the size of the molecule, we have shown that it is sufficient to include only relatively few selected distances as input data. In particular, we have chosen to consider distances associated with native contacts (conPCA) or, alternatively,  $C_\alpha$ -distances that are less than 8 Å apart in the native structure ( $C_\alpha$ PCA). Besides considerably reducing the numerical effort, this preselection of the degrees of freedom reduces the “noise” of the MD data, which typically results in better resolved conformational distributions obtained from the PCA. Moreover, this reduction avoids the apparent overrepresentation of the system (by using  $\sim N^2$  rather than  $\sim N$  variables), which may result in a “double counting” of the underlying degrees of freedom and affects the physical interpretation of the resulting free energy landscape.

To compare the various PCA methods, we have considered the number of PCs needed to cover a substantial amount of the overall variance (Fig. 2), two-dimensional free energy landscapes (Figs. 3 and 7) to identify metastable conformational states and barriers, and the PC autocorrelation function (Figs. 5 and 9), which reflects the time scale separation achieved by the PCA. For the considered systems HP35 and BPTI, we have generally found that the free energy landscapes of the distance-based PCAs give a somewhat minor state resolution than obtained for dPCA. This can be explained by the fact that certain important conformation rearrangements (e.g., residue 3 in HP35 or the disulfide bridge between residues 14 and 38 in BPTI) result in substantial



changes of some dihedral angles but hardly affect the distances of the system. On the other hand, the distance-based PCAs show a significantly better convergence of the cumulative fluctuations that necessitate less PCs. A further appealing feature is that the structural evolution along some PC can be easily illustrated by considering the main contacts that change during this motion (Figs. 4 and 8). Interestingly, the generally better performance of the  $C_\alpha$ PCA compared to the conPCA suggests that the structure of the backbone (in particular, the considerable restriction of possible conformations in a Ramachandran plot) is more important for a PCA description of the overall motion than the structure of the side-chains. Finally, it is important to note that in all considered cases the PCA (which sorts the PCs according to variance) also results in a suitable separation of time scale, that is, the first few PCs representing the system coordinates decay much slower than the remaining PCs representing the bath coordinates. While the optimal choice of internal coordinates certainly depends on the specific molecule and the process considered, our study has shown that distance-based PCAs, particularly  $C_\alpha$ PCA, represent a versatile approach towards this end.

## ACKNOWLEDGMENTS

We thank D. E. Shaw Research for sharing their trajectories of HP35 and BPTI and Sebastian Buchenberg and Abhinav Jain for numerous instructive and helpful discussions.

- <sup>1</sup>M. A. Rohrdanz, W. Zheng, and C. Clementi, *Annu. Rev. Phys. Chem.* **64**, 295 (2013).
- <sup>2</sup>P. Das, M. Moll, H. Stamati, L. E. Kaviraki, and C. Clementi, *Proc. Natl. Acad. Sci. U. S. A.* **103**, 9885 (2006).
- <sup>3</sup>O. F. Lange and H. Grubmüller, *Proteins* **62**, 1053 (2006).
- <sup>4</sup>R. Hegger, A. Altis, P. H. Nguyen, and G. Stock, *Phys. Rev. Lett.* **98**, 028102 (2007).
- <sup>5</sup>S. V. Krivov and M. Karplus, *Proc. Natl. Acad. Sci. U. S. A.* **105**, 13841 (2008).
- <sup>6</sup>S. V. Krivov, *J. Chem. Theory Comput.* **9**, 135 (2013).
- <sup>7</sup>J. S. Hub and B. L. de Groot, *PLoS Comput. Biol.* **5**, e1000480 (2009).
- <sup>8</sup>G. Perez-Hernandez, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noe, *J. Chem. Phys.* **139**, 015102 (2013).
- <sup>9</sup>I. T. Jolliffe, *Principal Component Analysis* (Springer, New York, 2002).
- <sup>10</sup>T. Ichiye and M. Karplus, *Proteins* **11**, 205 (1991).
- <sup>11</sup>A. E. Garcia, *Phys. Rev. Lett.* **68**, 2696 (1992).
- <sup>12</sup>A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, *Proteins* **17**, 412 (1993).
- <sup>13</sup>A. Kitao and N. Gō, *Curr. Opin. Struct. Biol.* **9**, 164 (1999).
- <sup>14</sup>B. L. de Groot, X. Daura, A. E. Mark, and H. Grubmüller, *J. Mol. Biol.* **309**, 299 (2001).
- <sup>15</sup>A. Altis, M. Otten, P. H. Nguyen, R. Hegger, and G. Stock, *J. Chem. Phys.* **128**, 245102 (2008).
- <sup>16</sup>G. G. Maisuradze, A. Liwo, and H. A. Scheraga, *Phys. Rev. Lett.* **102**, 238102 (2009).
- <sup>17</sup>G. Hummer, *New J. Phys.* **7**, 34 (2005).
- <sup>18</sup>O. F. Lange and H. Grubmüller, *J. Chem. Phys.* **124**, 214903 (2006).
- <sup>19</sup>C. Micheletti, G. Bussi, and A. Laio, *J. Chem. Phys.* **129**, 074105 (2008).
- <sup>20</sup>R. Hegger and G. Stock, *J. Chem. Phys.* **130**, 034106 (2009).
- <sup>21</sup>N. Schaudinnus, B. Bastian, R. Hegger, and G. Stock, *Phys. Rev. Lett.* **115**, 050602 (2015).
- <sup>22</sup>F. Rao and A. Caflisch, *J. Mol. Biol.* **342**, 299 (2004).
- <sup>23</sup>N.-V. Buchete and G. Hummer, *J. Phys. Chem. B* **112**, 6057 (2008).
- <sup>24</sup>F. Noe, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. Weikl, *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19011 (2009).
- <sup>25</sup>G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, *J. Chem. Phys.* **131**, 124101 (2009).
- <sup>26</sup>J.-H. Prinz *et al.*, *J. Chem. Phys.* **134**, 174105 (2011).
- <sup>27</sup>D. Shukla, C. X. Hernandez, J. K. Weber, and V. S. Pande, *Acc. Chem. Res.* **48**, 414 (2015).
- <sup>28</sup>K. D. Ball *et al.*, *Science* **271**, 963 (1996).
- <sup>29</sup>J. N. Onuchic, Z. L. Schulten, and P. G. Wolynes, *Annu. Rev. Phys. Chem.* **48**, 545 (1997).
- <sup>30</sup>K. A. Dill and H. S. Chan, *Nat. Struct. Biol.* **4**, 10 (1997).
- <sup>31</sup>M. Gruebele, *Curr. Opin. Struct. Biol.* **12**, 161 (2002).
- <sup>32</sup>D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).
- <sup>33</sup>F. Sittel, A. Jain, and G. Stock, *J. Chem. Phys.* **141**, 014111 (2014).
- <sup>34</sup>D. M. D. van Aalten, B. L. de Groot, J. B. C. Finday, H. J. C. Berendsen, and A. Amadei, *J. Comput. Chem.* **18**, 169 (1997).
- <sup>35</sup>N. Elmáci and R. S. Berry, *J. Chem. Phys.* **110**, 10606 (1999).
- <sup>36</sup>A. Altis, P. H. Nguyen, R. Hegger, and G. Stock, *J. Chem. Phys.* **126**, 244111 (2007).
- <sup>37</sup>L. Riccardi, P. H. Nguyen, and G. Stock, *J. Phys. Chem. B* **113**, 16660 (2009).
- <sup>38</sup>A. Jain, R. Hegger, and G. Stock, *J. Phys. Chem. Lett.* **1**, 2769 (2010).
- <sup>39</sup>S. Omori, S. Fuchigami, M. Ikeguchi, and A. Kidera, *J. Chem. Phys.* **132**, 115103 (2010).
- <sup>40</sup>R. Abseher and M. Nilges, *J. Mol. Biol.* **279**, 911 (1998).
- <sup>41</sup>A. Kloczkowski *et al.*, *J. Struct. Funct. Genomics* **10**, 67 (2009).
- <sup>42</sup>N. Hori, G. Chikenji, R. S. Berry, and S. Takada, *Proc. Natl. Acad. Sci. U. S. A.* **106**, 73 (2009).
- <sup>43</sup>L. R. Allen, S. V. Krivov, and E. Paci, *PLoS Comput. Biol.* **5**, e1000428 (2009).
- <sup>44</sup>I. V. Kalgin, A. Caflisch, S. F. Chekmarev, and M. Karplus, *J. Phys. Chem. B* **117**, 6092 (2013).
- <sup>45</sup>M. K. Scherer *et al.*, *J. Chem. Theory Comput.* **11**, 5525 (2015).
- <sup>46</sup>R. B. Best, G. Hummer, and W. A. Eaton, *Proc. Natl. Acad. Sci. U. S. A.* **110**, 17874 (2013).
- <sup>47</sup>R. B. Best and G. Hummer, *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1088 (2010).
- <sup>48</sup>E. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus, *Phys. Rev. Lett.* **67**, 1665 (1991).
- <sup>49</sup>S. Piana, K. Lindorff-Larsen, and D. E. Shaw, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17845 (2012).
- <sup>50</sup>D. E. Shaw *et al.*, *Science* **330**, 341 (2010).
- <sup>51</sup>Y. Duan and P. A. Kollman, *Science* **282**, 740 (1998).
- <sup>52</sup>C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele, *Nature* **420**, 102 (2002).
- <sup>53</sup>D. L. Ensign, P. M. Kasson, and V. S. Pande, *J. Mol. Biol.* **374**, 806 (2007).
- <sup>54</sup>J. Kubelka, T. K. Chiu, D. R. Davies, W. A. Eaton, and J. Hofrichter, *J. Mol. Biol.* **359**, 546 (2006).
- <sup>55</sup>A. Rajan, P. L. Freddolino, and K. Schulten, *PLoS One* **5**, e9890 (2010).
- <sup>56</sup>V. Hornak *et al.*, *Proteins* **65**, 712 (2006).
- <sup>57</sup>R. B. Best and G. Hummer, *J. Phys. Chem. B* **113**, 9004 (2009).
- <sup>58</sup>K. Lindorff-Larsen *et al.*, *Proteins* **78**, 1950 (2010).
- <sup>59</sup>W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. Klein, *J. Chem. Phys.* **79**, 926 (1983).
- <sup>60</sup>W. Kabsch and C. Sander, *Biopolymers* **22**, 2577 (1983).
- <sup>61</sup>A. Wlodawer, J. Walter, R. Huber, and L. Sjölin, *J. Mol. Biol.* **180**, 301 (1984).
- <sup>62</sup>H. W. Horn *et al.*, *J. Chem. Phys.* **120**, 9665 (2004).
- <sup>63</sup>Y. Mu, P. H. Nguyen, and G. Stock, *Proteins* **58**, 45 (2005).
- <sup>64</sup>A. Jain and G. Stock, *J. Phys. Chem. B* **118**, 7750–7760 (2014).
- <sup>65</sup>J. Heringa and P. Argos, *J. Mol. Biol.* **220**, 151 (1991).
- <sup>66</sup>N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, *J. Comput. Chem.* **32**, 2319 (2011).
- <sup>67</sup>T. Zhou and A. Caflisch, *J. Chem. Theory Comput.* **8**, 2930–2937 (2012).
- <sup>68</sup>A. Jain and G. Stock, *J. Chem. Theory Comput.* **8**, 3810 (2012).
- <sup>69</sup>F. Sittel and G. Stock, “Robust density-based clustering to identify metastable conformational states of proteins” (to be published).
- <sup>70</sup>See supplementary material at <http://dx.doi.org/10.1063/1.4938249> for details on the one-dimensional free energy landscapes and the  $C_\alpha$ PCA using all residues.