JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC

Article

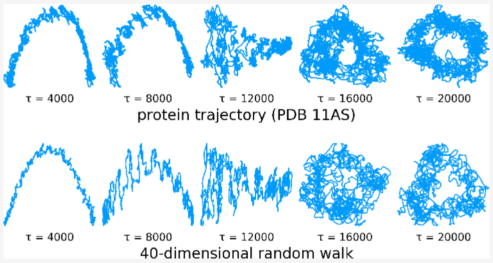
Time-Lagged Independent Component Analysis of Random Walks and Protein Dynamics

Steffen Schultze* and Helmut Grubmüller*

Cite This: J. Chem. Theory Comput. 2021, 17, 5766−5776

Read Online

ACCESS | Metrics & More | Article Recommendations



ABSTRACT: Time-lagged independent component analysis (tICA) is a widely used dimension reduction method for the analysis of molecular dynamics (MD) trajectories and has proven particularly useful for the construction of protein dynamics Markov models. It identifies those "slow" collective degrees of freedom onto which the projections of a given trajectory show maximal autocorrelation for a given lag time. Here we ask how much information on the actual protein dynamics and, in particular, the free energy landscape that governs these dynamics the tICA-projections of MD-trajectories contain, as opposed to noise due to the inherently stochastic nature of each trajectory. To answer this question, we have analyzed the tICA-projections of high dimensional random walks using a combination of analytical and numerical methods. We find that the projections resemble cosine functions and strongly depend on the lag time, exhibiting strikingly complex behavior. In particular, and contrary to previous studies of principal component projections, the projections change noncontinuously with increasing lag time. The tICA-projections of selected 1 μs protein trajectories and those of random walks are strikingly similar, particularly for larger proteins, suggesting that these trajectories contain only little information on the energy landscape that governs the actual protein dynamics. Further the tICA-projections of random walks show clusters very similar to those observed for the protein trajectories, suggesting that clusters in the tICA-projections of protein trajectories do not necessarily reflect local minima in the free energy landscape. We also conclude that, in addition to the previous finding that certain ensemble properties of nonconverged protein trajectories resemble those of random walks; this is also true for their time correlations.

1. INTRODUCTION

The atomistic dynamics of proteins, protein complexes, and other biomolecules is exceedingly complex, covering time scales from subpicoseconds to up to hours.[1,2] It is governed by a similarly complex high-dimensional free energy landscape or funnel,[3] characterized by a hierarchy of free energy barriers,[4] and has been widely studied computationally by molecular dynamics (MD) simulations.[5] With particle numbers ranging from several hundreds to hundreds of thousands or more,[6−9] the correspondingly high-dimensional configuration space of the system poses considerable challenges to a fundamental understanding of biomolecular function, for example, of the conformational motions of these biological "nanomachines",[10,11] protein folding,[12] or specific binding.

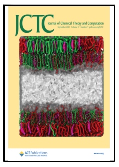
Several attempts to reduce the dimensionality of the dynamics have addressed this issue. Most notable approaches are principal component analysis (PCA) to extract the essential dynamics[13] of the protein that contributes most to the atomic fluctuations, and time-lagged independent component analysis (tICA), which identifies those collective degrees of freedom that exhibit the strongest time-correlations for a given lag-time.[14,15] Both dimension reduction techniques can yield information on the conformational dynamics of a protein, that is, how the protein moves through several conformational substates, which can be defined as metastable conformations characterized by local free energy minima.[16]

This property also renders these dimension reduction techniques highly useful as a preprocessing step to describing the conformational dynamics of macromolecules in terms of a discrete Markov process.[17−19] Currently tICA is most widely used, and it is preferred over PCA for this purpose[20] because it additionally uses time information on the input trajectory.

In this context, both PCA and tICA rely on MD trajectories as input, which raises the question how much of these analyses is determined by actual information on the protein dynamics, as opposed to noise due to the inherently stochastic nature of each trajectory, and, importantly, how these two can be quantified.

For PCA, this question has been answered by analysis of the principal components of a high-dimensional random walk in a flat energy landscape.[21,22] Unexpectedly, these turned out to

Received: March 18, 2021
Published: August 27, 2021



Downloaded via MPI MULTIDISCIPLINARY SCIENCES on April 26, 2022 at 21:17:08 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ACS Publications

© 2021 The Authors. Published by American Chemical Society

https://doi.org/10.1021/acs.jctc.1c00273

approximate cosine functions, thus providing a very powerful criterion for the convergence of MD trajectories: The more an MD trajectory resembles a cosine, quantified by the cosine content,[21] the more it resembles a random walk, and the less information it contains on the actual protein dynamics or the underlying free energy landscape.

These analyses[21,22] have also suggested that clusters observed in low-dimensional PCA projections do not necessarily imply the existence of conformational substates and, instead, may also be a stochastic and/or projection artifact. Particularly the latter finding is highly relevant for the use of PCA for the construction of Markov models,[19] which thus may also in part reflect the randomness of one or several trajectories. Note that this holds also true—albeit probably to a lesser extent—for the construction of Markov models from several or many trajectories, as these have to be spawned from a seeding trajectory or from starting structures generated from other advanced sampling methods.[16,23−25]

For tICA, no such analysis is available, but the inspection of several examples suggests that similar effects may also be at work.[26,27] To address this issue, here we will therefore analyze the tICA-projections of high dimensional random walks, and subsequently compare them to tICA-projections of selected protein trajectories. In particular, we will semianalytically derive an expression for random walk tICA-projections, which will prove analogous to the PCA cosine functions and thus can also serve as a criterion for convergence as well as for the quality of derived Markov models. Unexpectedly, and contrary to the regular behavior of random walk PCA projections, tICA-projections turn out to display much more complex behavior. In particular, we observed critical lag times at which the random walk projections change drastically and — for high dimensions — even discontinuously. The resulting much richer and more intricate structure of random walk projections renders the proper interpretation of tICA-projections of protein dynamics trajectories particularly challenging, and has profound implications for the proper constructions of Markov models.

2. THEORETICAL ANALYSIS AND METHODS

2.1. Definition of tICA. To establish notation, we briefly summarize the basic principle of tICA; for a more comprehensive treatment with particular focus on molecular dynamics applications, see ref 28.

Consider a d-dimensional trajectory $\mathbf{x}(t) = (x_1(t), ..., x_d(t))^{\mathrm{T}} \in \mathbb{R}^d$ with Cartesian coordinates $x_1, ..., x_d$, which for compact notation we assume to be mean-free, that is, the time average $\langle \mathbf{x}(t)\rangle_t$ is zero. TICA determines those "slowest" independent collective degrees of freedom $\mathbf{v}_k \in \mathbb{R}^d$, $k = 1, ..., d$, onto which the projections $y_k(t) = \mathbf{v}_k \cdot \mathbf{x}(t)$ have the largest time-autocorrelation

$$\frac{\langle y_k(t) y_k(t+\tau)\rangle_t}{\langle y_k(t)^2\rangle_t}$$

where τ is a chosen lag time. Equivalently, using the time-lagged covariance matrix

$$\mathbf{C}(\tau) = (\langle x_i(t) x_j(t+\tau)\rangle_t)_{ij} \in \mathbb{R}^{d\times d}$$

each degree of freedom $\mathbf{v}_k$ maximizes

$$\frac{\mathbf{v}_k^{\mathrm{T}}\mathbf{C}(\tau)\mathbf{v}_k}{\mathbf{v}_k^{\mathrm{T}}\mathbf{C}(0)\mathbf{v}_k}$$

under the constraint that it is orthogonal to all previous degrees of freedom. Hence, the $\mathbf{v}_k$ are the solutions of the generalized eigenvalue problem

$$\mathbf{C}(\tau)\mathbf{v}_k = \lambda_k \mathbf{C}(0)\mathbf{v}_k \tag{1}$$

We will use the term "tICA-eigenvector" for the $\mathbf{v}_k$ and "tICA-projection" for the projections y_k onto the tICA-eigenvectors. In the literature, the term "tICA-component" is often used, but it is somewhat ambiguous and we will therefore avoid it.

For an infinite trajectory of a time-reversible system the matrices in this eigenvalue problem are symmetric. However, for the finite trajectories considered here, with time steps $t = 1, ..., n$, the matrix $\mathbf{C}(\tau)$ is usually not symmetric. There are two slightly different symmetrization methods that circumvent this problem. The more popular one, which we denote the "main" method, uses an estimator that replaces the simple time-lagged averages mentioned by averages over all pairs $(\mathbf{x}_t, \mathbf{x}_{t+\tau})$ and $(\mathbf{x}_{t+\tau}, \mathbf{x}_t)$, following, for example, Noé[28] and the popular software package PyEMMA.[29] As a result, on the left-hand side of eq 1 $\mathbf{C}(\tau)$ is replaced with

$$\begin{aligned}\mathbf{C}_{\mathrm{sym}}(\tau) &= \frac{1}{2}(\mathbf{C}(\tau) + \mathbf{C}(\tau)^{\mathrm{T}})\\ &= \left(\frac{1}{2}\frac{1}{n-\tau}\left(\sum_{t=1}^{n-\tau} x_i(t)x_j(t+\tau) + \sum_{t=1}^{n-\tau} x_i(t+\tau)x_j(t)\right)\right)_{ij}\end{aligned}$$

and on the right-hand side $\mathbf{C}(0)$ is replaced with

$$\boldsymbol{\Sigma} = \left(\frac{1}{2}\frac{1}{n-\tau}\left(\sum_{t=1}^{n-\tau} x_i(t)x_j(t) + \sum_{t=1}^{n-\tau} x_i(t+\tau)x_j(t+\tau)\right)\right)_{ij}$$

yielding a symmetrized version of eq 1 with real eigenvalues,

$$\mathbf{C}_{\mathrm{sym}}(\tau)\mathbf{v}_k = \lambda_k \boldsymbol{\Sigma}\mathbf{v}_k \tag{2}$$

The second "alternative" symmetrized version of eq 1 only differs on the right-hand side, where $\mathbf{C}(0)$ is not replaced with $\boldsymbol{\Sigma}$,

$$\mathbf{C}_{\mathrm{sym}}(\tau)\mathbf{v}_k = \lambda_k \mathbf{C}(0)\mathbf{v}_k \tag{3}$$

Our analysis is very similar for both versions, though with unexpectedly different results.

2.2. Theory. To render this symmetrized generalized eigenvalue problem more amenable to analysis, and following ref 30, we define a matrix formed from the trajectory

$$\mathbf{X} = \begin{pmatrix} | & | & & | \\ \mathbf{x}(1) & \mathbf{x}(2) & ... & \mathbf{x}(n) \\ | & | & & | \end{pmatrix}$$

as well as a shorter time-lagged matrix

$$\mathbf{X}_{\mathrm{lag}} = \begin{pmatrix} | & | & & | \\ \mathbf{x}(\tau+1) & \mathbf{x}(\tau+2) & ... & \mathbf{x}(n) \\ | & | & & | \end{pmatrix}$$

and one that is cut off at the end

$$\mathbf{X}_{\mathrm{cut}} = \begin{pmatrix} | & | & & | \\ \mathbf{x}(1) & \mathbf{x}(2) & ... & \mathbf{x}(n-\tau) \\ | & | & & | \end{pmatrix}$$

The latter two matrices serve to rewrite the above left and right-hand sides,

$$\mathbf{C}_{\text{sym}}(\tau) = \frac{1}{2}\frac{1}{n-\tau}(\mathbf{X}_{\text{cut}}\mathbf{X}_{\text{lag}}^{\text{T}} + \mathbf{X}_{\text{lag}}\mathbf{X}_{\text{cut}}^{\text{T}})$$

and

$$\mathbf{\Sigma} = \frac{1}{2}\frac{1}{n-\tau}(\mathbf{X}_{\text{lag}}\mathbf{X}_{\text{lag}}^{\text{T}} + \mathbf{X}_{\text{cut}}\mathbf{X}_{\text{cut}}^{\text{T}})$$

and, hence, also the symmetrized tICA-equation,

$$(\mathbf{X}_{\text{cut}}\mathbf{X}_{\text{lag}}^{\text{T}} + \mathbf{X}_{\text{lag}}\mathbf{X}_{\text{cut}}^{\text{T}})\mathbf{v}_k = \lambda_k(\mathbf{X}_{\text{lag}}\mathbf{X}_{\text{lag}}^{\text{T}} + \mathbf{X}_{\text{cut}}\mathbf{X}_{\text{cut}}^{\text{T}})\mathbf{v}_k \tag{4}$$

This defining eq 4 for tICA can be converted into a more convenient form using the matrices

$$\mathbf{A} = \begin{pmatrix} 0 & \overset{\tau}{\cdots} & 0 & 1 & & \\ \vdots & \ddots & & \ddots^{n-\tau} & \ddots & \\ 0 & & \ddots & & \ddots & 1 \\ 1 & \ddots & & \ddots & & 0 \\ & \ddots & \ddots & & \ddots & \vdots \\ & & 1 & 0 & \cdots & 0 \end{pmatrix}$$

and

$$\mathbf{B} = \text{diag}(\underbrace{1, \ldots, 1}_{\tau}, \underbrace{2, \ldots, 2}_{n-2\tau}, \underbrace{1, \ldots, 1}_{\tau}).$$

Noting that

$$(\mathbf{X}_{\text{cut}}\mathbf{X}_{\text{lag}}^{\text{T}} + \mathbf{X}_{\text{lag}}\mathbf{X}_{\text{cut}}^{\text{T}}) = \mathbf{XAX}^{\text{T}},$$
$$(\mathbf{X}_{\text{lag}}\mathbf{X}_{\text{lag}}^{\text{T}} + X_{\text{cut}}\mathbf{X}_{\text{cut}}^{\text{T}}) = \mathbf{XBX}^{T}$$

and eq 4 reads

$$\mathbf{XAX}^{\text{T}}\mathbf{v}_k = \lambda_k\mathbf{XBX}^{\text{T}}\mathbf{v}_k \tag{5}$$

This can be transformed into a normal eigenvalue problem using the AMUSE-algorithm[31,32] as follows. First diagonalize the right-hand side by an orthogonal matrix $\mathbf{Q}$ and a diagonal matrix $\mathbf{\Lambda}$ such that

$$\mathbf{Q}^{\text{T}}\mathbf{XBX}^{\text{T}}\mathbf{Q} = \mathbf{\Lambda}$$

Substituting $\mathbf{v}_k = \mathbf{Wu}_k$, with $\mathbf{W} = \mathbf{Q\Lambda}^{-1/2}$, and assuming all diagonal elements of $\mathbf{\Lambda}$ are nonzero, yields

$$\mathbf{XAX}^{\text{T}}\mathbf{Wu}_k = \lambda_k\mathbf{XBX}^{\text{T}}\mathbf{Wu}_k$$

Note that this assumption is actually not necessarily true here, but since we are only interested in the nonzero eigenvalues and their eigenvectors the end results will still be correct. Since $\mathbf{W}$ is invertible, this equation is equivalent to

$$\mathbf{W}^{\text{T}}\mathbf{XAX}^{\text{T}}\mathbf{Wu}_k = \lambda_k\mathbf{W}^{\text{T}}\mathbf{XBX}^{\text{T}}\mathbf{Wu}_k$$

where the matrix on the right-hand side turns out to be the unit matrix,

$$\mathbf{W}^{\text{T}}\mathbf{XBX}^{\text{T}}\mathbf{W} = \mathbf{\Lambda}^{-1/2}\mathbf{Q}^{\text{T}}\mathbf{XBX}^{\text{T}}\mathbf{Q\Lambda}^{-1/2} = \mathbf{\Lambda}^{-1/2}\mathbf{\Lambda\Lambda}^{-1/2} = 1$$

Hence eq 5 simplifies to

$$\mathbf{W}^{\text{T}}\mathbf{XAX}^{\text{T}}\mathbf{Wu}_k = \lambda_k\mathbf{u}_k \tag{6}$$

Now consider the following "swapped" version:[30]

$$\mathbf{X}^{\text{T}}\mathbf{WW}^{\text{T}}\mathbf{XAy}_k = \lambda_k\mathbf{y}_k \tag{7}$$

Notably, for each $\mathbf{y}_k$ satisfying eq 7 there exists a corresponding eigenvector that solves eq 6. Indeed, choosing $\mathbf{u}_k = \mathbf{W}^{\text{T}}\mathbf{XAy}_k$ yields

$$\mathbf{W}^{\text{T}}\mathbf{XAX}^{\text{T}}\mathbf{Wu} = \mathbf{W}^{\text{T}}\mathbf{XAX}^{\text{T}}\mathbf{WW}^{\text{T}}\mathbf{XA}y = \mathbf{W}^{\text{T}}\mathbf{XA}\lambda_k\mathbf{y}_k = \lambda_k\mathbf{u}_k$$

Finally, up to normalization, $\mathbf{y}_k$ is the projection of the trajectory onto the corresponding $\mathbf{v}_k = \mathbf{Wu}_k$,

$$\mathbf{X}^{\text{T}}\mathbf{v}_k = \mathbf{X}^{\text{T}}\mathbf{Wu}_k = \mathbf{X}^{\text{T}}\mathbf{WW}^{\text{T}}\mathbf{XAy}_k = \lambda_k\mathbf{y}_k$$

In other words, the tICA-projections of the trajectory are the eigenvectors (with nonzero eigenvalues) of the matrix $\mathbf{M} = \mathbf{X}^{\text{T}}\mathbf{WW}^{\text{T}}\mathbf{XA}$.

We will use this reformulation of the tICA defining equation to calculate the tICA-projections of random walks of given finite dimension and length.

2.3. Random Walks. For the numerical and semianalytical evaluation of tICA components, random walk trajectories $\mathbf{x}(t) \in \mathbb{R}^d$ of dimension d were generated by carrying out n steps according to

$$\mathbf{x}(t+1) = \mathbf{x}(t) + \mathbf{r}(t), \mathbf{r}(t) \sim \mathcal{N}$$

where $\mathcal{N}$ is a d-dimensional univariate normal distribution centered at 0. Each trajectory was centered to zero before further processing. We verified empirically that other fixed probability distributions with mean 0 and finite variance yield similar results.

2.4. Molecular Dynamics Simulation. For two proteins a 1 μs molecular dynamics trajectory each was analyzed (Andreas Volkhardt, private communication). Both were generated using the GROMACS 4.5 software package[33] with the Amber ff99SB-ILDN force field[34] and the TIP4P-Ew water model.[35] The starting structures were taken from the PDB[36] entries 11AS[37] and 2F21,[38] respectively. From the latter, only a part of the structure (the WW-domain) was used. Energy minimization was performed using steepest descent for 5×10^4 steps. The hydrogen atoms were described by virtual sites. Each protein was placed within a triclinic water box using gmx-solvate, such that the smallest distance between protein surface and box boundary was larger than 1.5 nm. Natrium and chloride ions were added to neutralize the system, corresponding a physiological concentration of 150 mmol/L. Each system was first equilibrated for 0.5 ns in the NVT ensemble, and subsequently for 1.0 ns in the NPT ensemble at 1 atm pressure and temperature 300 K, both using an integration time step of 2 fs. The velocity rescaling thermostat[39] and Parrinello−Rahman pressure coupling[40] were used with coupling coefficients of τ = 0.1 ps and τ = 1 ps, respectively. All bond lengths of the solute were constrained using LINCS with an expansion order of 6, and water geometry was constrained using the SETTLE algorithm. Electrostatic interactions were calculated using PME,[41] with a real space cutoff of 10 Å and a Fourier spacing of 1.2 Å. The integration time step was 4 fs, and the coordinates of the alpha carbons were

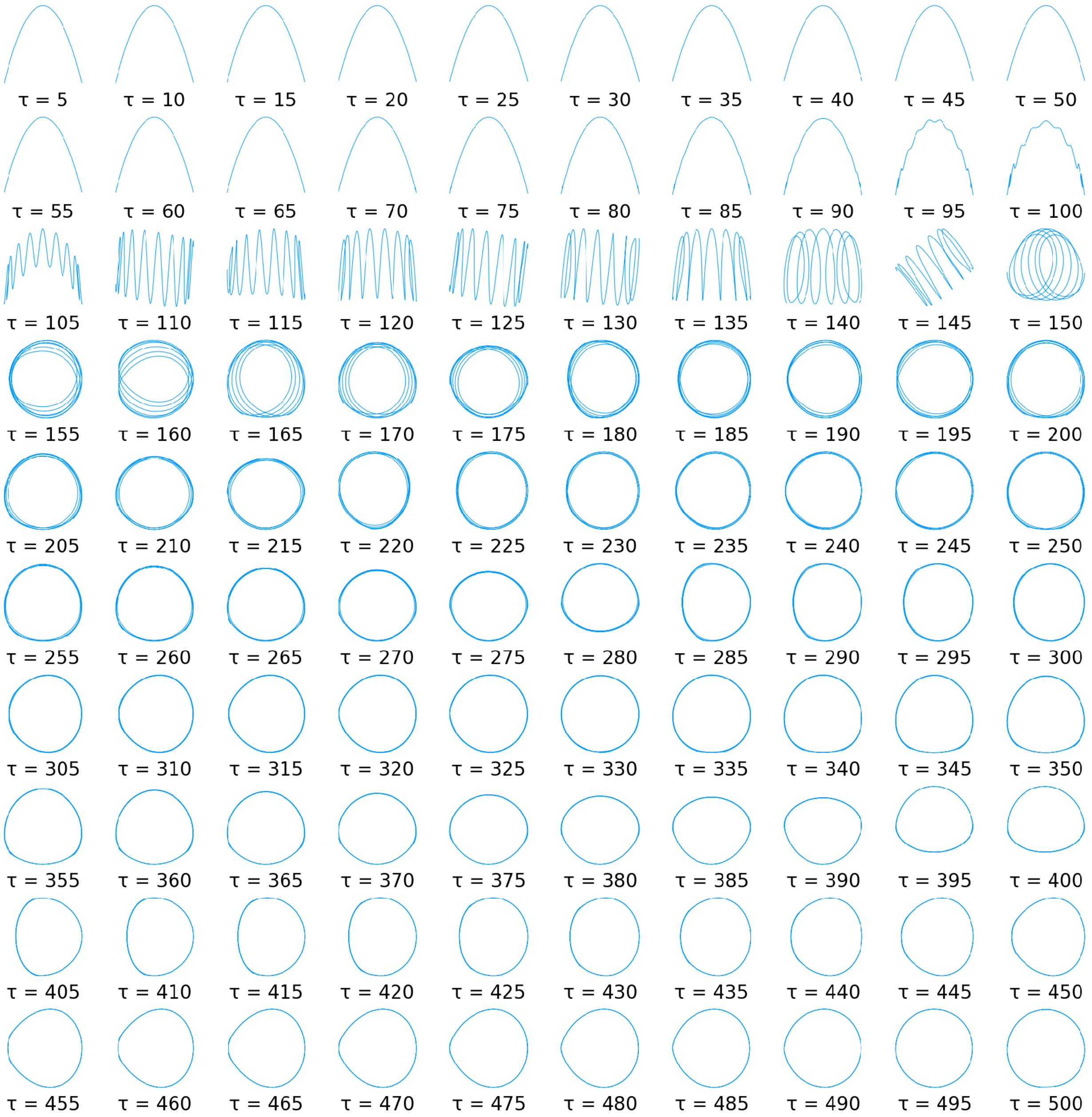


Figure 1. First two "expected" tICA-projections of random walks of dimension $d = 50$ with $n = 1000$ time steps for varying lag time τ, computed with the averaging method from section 3.2 using a sample of 20 000 random walks. For each τ, the first tICA-projection is shown on the x-axis and the second one on the y-axis.

saved every 10 ps, such that 10^5 snapshots were available for each trajectory. Of these we discarded the first 10^4 steps, leading to trajectories of length $n = 9 \times 10^4$.

3. RESULTS AND DISCUSSION

To characterize the tICA components and projections of random walks, we will proceed in two steps. We will first analyze a special case, for which some analytical results can be obtained. Second, we will use the obtained insights to generalize this result to random walks of arbitrary length n and dimension d using a combined analytical/numerical approach. Subsequently, we will compare the obtained random walk projections to tICA analyses of biomolecular trajectories.

3.1. A Special Case.

To gain first insight into the tICA components of a random walk, first consider the special case $d = n$, which allows for an almost fully analytical approach. In this case, all matrices in eq 7 are square and, assuming that $\mathbf{X}$ is invertible,

$$\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{X} = \mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X} = \mathbf{X}^{\mathrm{T}}\mathbf{X}^{-\mathrm{T}}\mathbf{B}^{-1}\mathbf{X}^{-1}\mathbf{X} = \mathbf{B}^{-1}$$

such that eq 7 becomes independent of $\mathbf{X}$,

$$\mathbf{B}^{-1}\mathbf{A}\mathbf{y}_k = \lambda_k\mathbf{y}_k \tag{8}$$

Note that the assumption that $\mathbf{X}$ is invertible is not strictly correct, as it has one zero-eigenvalue associated with the eigenvector given by $\mathbf{y}_0 = (1, ..., 1)^{\mathrm{T}}$. This is also an eigenvector of $\mathbf{B}^{-1}\mathbf{A}$, but instead with eigenvalue 1. Therefore, all the eigenvectors and all but one eigenvalue of eq 7 are identical to those of eq 8, and the analysis can proceed using eq 8.

In the limit of large n, and using the above definitions for $\mathbf{A}$ and $\mathbf{B}$, the matrix $\mathbf{B}^{-1}\mathbf{A}$ approaches a circulant matrix with the

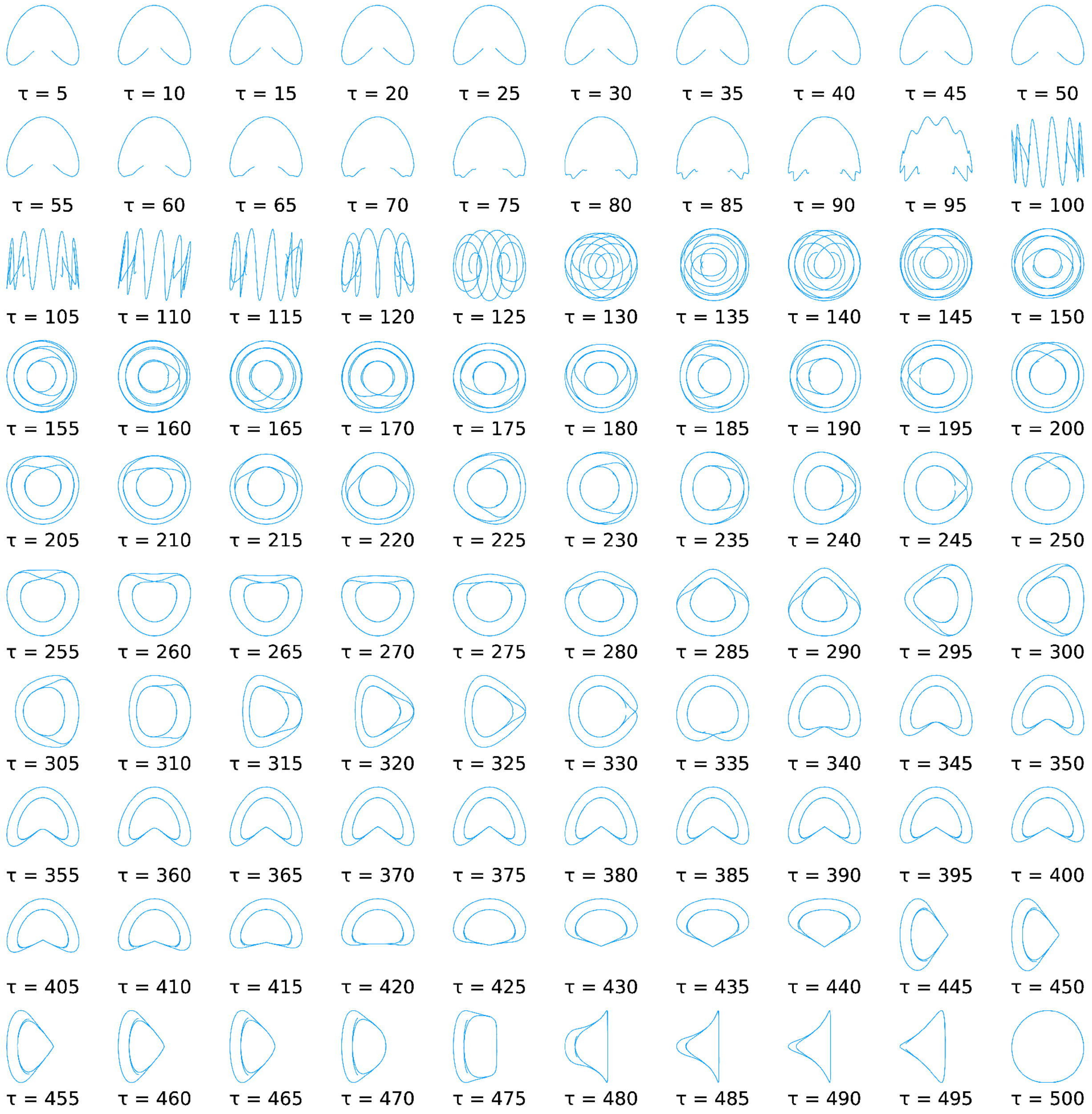


Figure 2. First two "expected" tICA-projections, for the alternative symmetrization method, of random walks of dimension $d = 50$ with $n = 1000$ time steps for varying lag time τ, computed with the averaging method from section 3.2 using a sample of 20 000 random walks. For each τ, the first tICA-projection is shown on the x-axis and the second one on the y-axis.

property that each of its columns is a cyclic permutation of the preceding one. It differs from a circulant matrix only at the four "corners" (of size τ) of the matrix, and for large $n = d$ these "corners" become small relative to the size of the matrix. More precisely, $\mathbf{B}^{-1}\mathbf{A}$ and the circulant matrix are asymptotically equivalent as defined in ref 42.

Circulant matrices are diagonalized by the Fourier transform,[43] yielding eigenvectors

$$\tilde{\mathbf{y}}_k = (1, \omega_k, \omega_k^2, ..., \omega_k^{n-1}), \qquad \omega_k = \exp\left(2\pi \mathrm{i}\frac{k}{n}\right)$$

and eigenvalues

$$\lambda_k = \frac{\omega_k^\tau + \omega_k^{n-\tau}}{2} = \cos\left(2\pi\frac{\tau k}{n}\right) \tag{9}$$

These eigenvectors are complex, but since $\lambda_k = \lambda_{n-k}$ and $\tilde{\mathbf{y}}_k = \tilde{\mathbf{y}}_{n-k}^*$, the real and imaginary part of $\tilde{\mathbf{y}}_k$ (cosine and sine) are real eigenvectors for the same eigenvalues. Depending on τ and n, many of these eigenvalues are equal, since they only depend on $\tau k \bmod n$.

This result implies that for large $n = d$ the eigenvalues of $\mathbf{B}^{-1}\mathbf{A}$ approach those of the circulant matrix. More precisely, their eigenvalues are asymptotically equally distributed.[42] In contrast, the eigenvectors are only preserved in limits or under small perturbations if the respective adjacent eigenvalues are well-separated from each other.[44] For the case at hand, however, this eigenvalue separation very quickly approaches zero for small k and large n (and for other k with $|\cos(2\pi\tau k/n)| \approx 1$). As a result, the eigenvectors of $\mathbf{B}^{-1}\mathbf{A}$ for small k (and other k as before) differ from those of the circulant matrix even in this limit. Rather, they need to be represented as approximate linear combinations of those eigenvectors of the circulant matrix with similar eigenvalues.

This subtlety contributes to the complexity of the problem as well as of the solution, and has so far has prohibited us from proceeding further purely analytically both for finite $d = n$ as well as for $d = n \to \infty$. Nevertheless, the eigenvalue problem eq 8 provides a good starting point for a numerical approach. Still, the degeneracy discussed above needs to be taken properly into account, as the numerical eigenvectors are essentially arbitrarily chosen from the eigenspaces.

Inspecting the Fourier transforms of the numerical eigenvectors suggests that the eigenspaces of eq 8 for small k each contain an eigenvector that resembles a cosine function

$$y_k(t) \approx \cos\left(\pi \frac{tk}{n}\right)$$

with increasing accuracy for increasing n.

Another effect of the poor separation of the eigenvalues is that the above results are very sensitive to small changes to the matrix in eq 8. For example, when the alternative symmetrization method defined by eq 3 is used, the analysis in section 2.2 is unchanged, except that all diagonal entries of B become 2, and eq 8 reads

$$\frac{1}{2}\mathbf{A}\mathbf{y}_k = \lambda_k \mathbf{y}_k$$

For $n = d \to \infty$, the same circulant matrix is obtained, such that the eigenvalues, eq 9, are unchanged. The numerical solution however reveals that the first few eigenspaces instead contain eigenvectors given by

$$y_k(t) \approx \sin\left(2\pi \frac{tk}{n}\right)$$

This result is indeed strikingly different, in that the cosine functions are replaced by sine functions with twice the frequency.

3.2. General Solution.

Next, we will consider the general case, that is, a random walk of length n in $d < n$ dimensions. Unfortunately, we were unable to find analytical solutions similar to the above; however, the results of section 2.2 permit an elegant way for a numerical approach by computing the expectation value of the matrix $\mathbf{M}$. To this aim, $\mathbf{M}$ was computed for a sample of 20000 random walks of given fixed dimension d and number of time steps n, from which an average matrix $\langle \mathbf{M} \rangle$ was computed. The eigenvectors of $\langle \mathbf{M} \rangle$ served as the semianalytical solution for the general case. We note that this does not necessarily produce the same results as averaging the individual tICA-projections directly. We have, however, tested that the eigenvectors of $\langle \mathbf{M} \rangle$ are very similar to the averages of the tICA-projections. An exception to this is that averaging the tICA-projections can produce artifacts arising from to the fluctuating order of the eigenvectors, and these artifacts are not present in the eigenvectors of $\langle \mathbf{M} \rangle$.

As an illustration, Figure 1 shows the first two resulting tICA-projections for random walks with $n = 1000$ and $d = 50$, revealing a strong dependence on the lag time τ. For short lag times τ, $y_1(t) \approx \cos(\pi t/n)$ and $y_2(t) \approx \cos(2\pi t/n)$. With increasing τ, these low-frequency cosines are gradually replaced by higher-frequency components, first in $\mathbf{y}_2$ (starting at about $\tau = 90$) and for further increasing $\tau > 150$ also in $\mathbf{y}_1$. From then on, the frequencies of both $\mathbf{y}_1$ and $\mathbf{y}_2$ slowly decrease, maintaining a π phase shift.

In contrast to the special case considered above (section 3.1), our numerical studies suggest that for large lag times the averaged projections do not approach exact cosines for large n. Rather, "cosine like" functions appear, as can be seen for the high lag-times shown in Figure 1, where the circular shape that would be expected for exact cosines is noticeably distorted, even if n is further increased. In contrast, for short lag times, where the higher frequency components have not yet appeared (e.g., $\tau < 90$ in Figure 1), the projections do seem to approach exact cosines with increasing n.

For the alternative symmetrization method, eq 3, the same method can be applied, and the obtained projections are shown in Figure 2. Indeed, when the two figures are compared, even more dramatic differences are seen as a result of this very small change. In particular, for short τ values, the cosine-like functions seem to be replaced by sine-like functions of twice the frequency, just like we have already seen for the special case $d = n$. Also, for increasing τ a much richer and complex behavior is seen. Finally, the onset of higher frequencies occurs for somewhat smaller τ values (at $\tau \approx 100$) compared to that in Figure 1 (at $\tau \approx 110$). This abrupt emergence of higher frequencies deserves closer inspection.

3.3. Abrupt Changes.

To gain more insight into why these abrupt changes occur, Figure 3A shows the eigenvalues of $\langle \mathbf{M} \rangle$ as a function of τ for dimension $d = 30$, revealing a strikingly complex pattern. For small lag times τ all eigenvalues decrease with τ, with associated cosine-shaped eigenvectors of period lengths $2n$, $2n/2$, $2n/3$, ..., as annotated in the figure. The decrease of these curves reflects the sampling of the cosine-

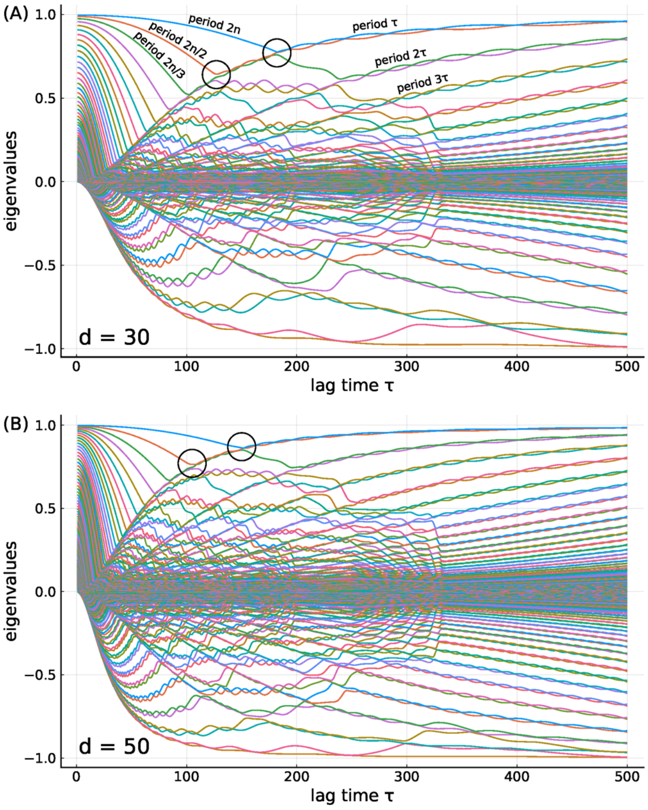


Figure 3. Eigenvalues of the averaged matrix $\langle \mathbf{M} \rangle$ as a function of the lag time τ at (A) dimension $d = 30$ and (B) dimension $d = 50$. The two abrupt changes are indicated using black circles. The colors indicate the order of the eigenvalues.

https://doi.org/10.1021/acs.jctc.1c00273

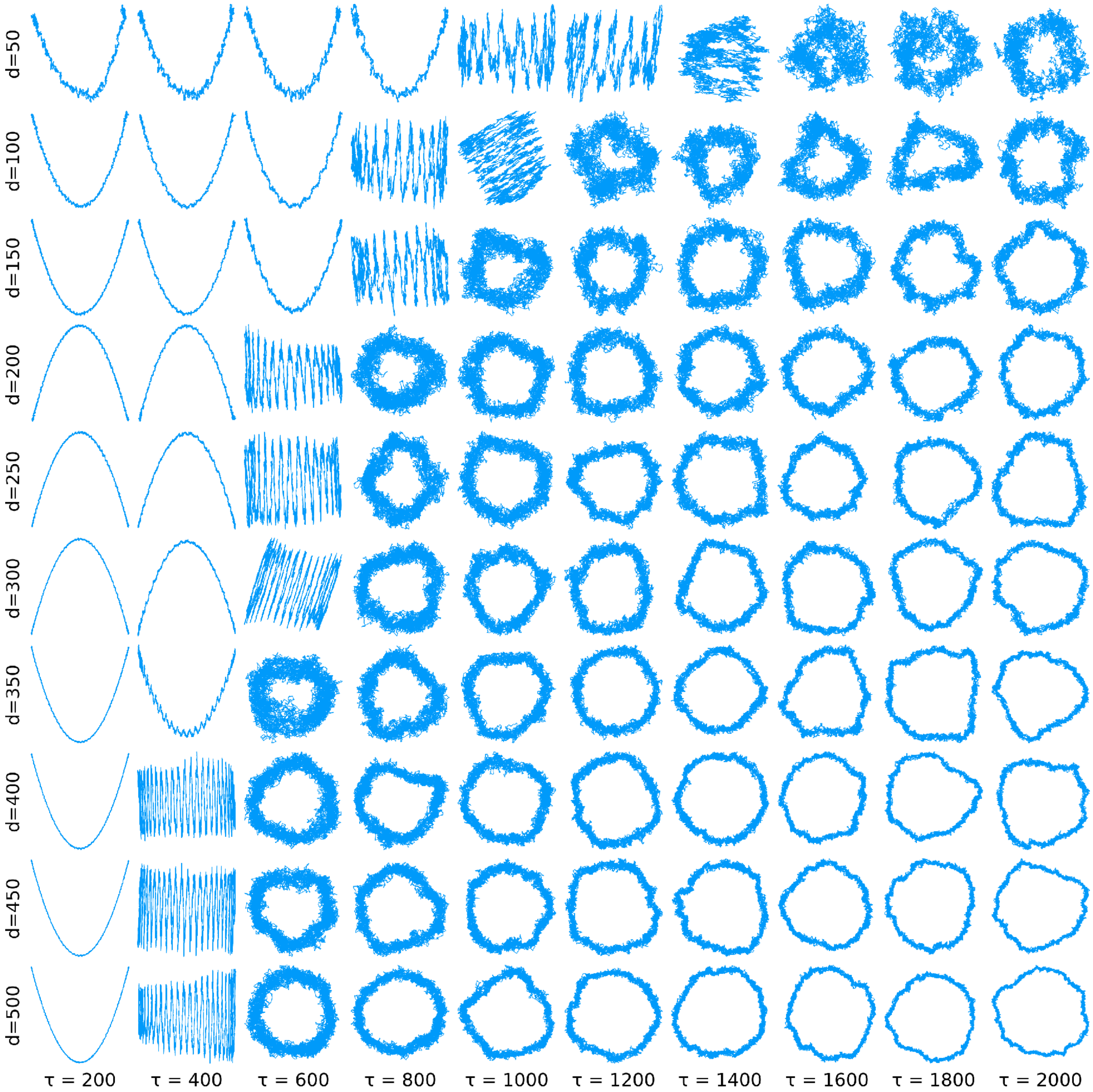


Figure 4. First two tICA-projections of random walks with varying dimensions d, each with n = 10000. The lag times of the abrupt changes decrease with increasing dimension.

shaped eigenvectors with increasing lag time τ and, hence, the respective autocorrelations also resemble cosine functions.

Also visible are several curves that monotonically increase with τ, each starting at zero for small τ. These curves represent two eigenvalues each, with cosine-shaped and sine-shaped eigenvectors of period lengths τ, 2τ, 3τ, ..., respectively, as also annotated in the Figure. Their increase is less obvious, as one might expect the autocorrelation of a τ-periodic function at lag time τ to be unity and, therefore, constant. Note, however, that the eigenvalue of $\langle \mathbf{M} \rangle$ does not strictly represent this autocorrelation; rather, it represents the average of the autocorrelations of many instances of this eigenvector for each single random walk—each of which is not strictly periodic. For increasing period lengths, the eigenvectors approach cosines or sines, such that their average autocorrelation increases and so do the corresponding eigenvalues of $\langle \mathbf{M} \rangle$.

At the intersections of these two sets of curves (black circles) the respective eigenvalues are degenerate and their order changes, which causes abrupt changes of the eigenvectors and, therefore, also of the projections onto these eigenvectors, the first two of which were discussed above.

For larger dimensions d, for example, for d = 50 as shown in Figure 3B, one would expect that the tICA-projections resemble cosine or sine functions increasingly closely, also at increasingly higher frequencies. As a result, the eigenvalues corresponding to the eigenvectors with period lengths τ, 2τ, 3τ, ... should increase with d at any given lag time τ, whereas the decreasing eigenvalue curves on the left side should remain unchanged. Therefore, the respective intersections should occur at smaller lag times τ. Comparison of the black circles in the two panels of Figure 3 shows that this is indeed the case. To illustrate this effect, Figure 4 shows the first two tICA-projections of random walks with dimensions ranging from 50 (top row) to 500 (bottom row) for increasing τ.

To quantify this behavior, we generated a large number of random walks and determined the lag times τ at which the abrupt changes occur. Figure 5 shows the first and second of these critical lag times as a function of dimension d and for n ranging from 1000 to 5000 (colors). To enable direct comparison, the lag times τ have been normalized by n. As can be seen, for d between ca. 150 and $n/2$ both the first (upper curves) and second (lower curves) approximate power laws $n/\tau \propto d^b$, as indicated by the respective fits (solid lines, the colors correspond to the values of n). For each fit, only dimensions d within the above range have been used.

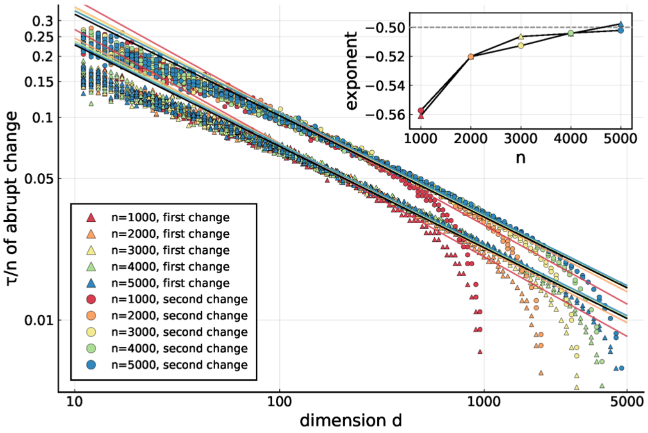


Figure 5. Lag time at which the abrupt changes occur in dependence of the dimension for various n. Each dot represents an independently generated random walk. Also shown are the power law fits $n/\tau = a{\cdot}d^b$ (colored lines), their exponents (inset), and the lines corresponding to $b = -0.5$ (black lines).

The inset of Figure 5 shows the power law exponents b for varying n and for the first and second abrupt change, both of which apparently approach $b = -1/2$ for large n (also represented by the black lines in the main figure). Although we were unable to find a rigorous proof, this finding suggests that in the limit of large n and d, with d markedly smaller than n, the first few lag times at which abrupt changes occur scale as $\tau \propto n/\sqrt{d}$.

3.4. Comparison of Random Walks and MD-Trajectories.

We next compared the tICA-projections of random walks with those of molecular dynamics trajectories of proteins in solution. To that end, we used two MD-trajectories of length 1 μs each (generated as described in section 2.4), one of a comparatively large protein (PDB 11AS, 330 amino acids)[37] and one of a smaller protein (WW-domain of PDB 2F21, 34 amino acids).[38]

As can be seen in Figure 6, the tICA-projections of the larger protein (top group) are indeed spectacularly similar to those of a random walk (bottom group). Even the strong dependence on the lag time is very similar, as are the abrupt changes discussed above.

Note that this striking similarity was obtained for a particular choice of $d = 40$ for the random walk; other dimensionalities yield less similar projections. Intriguingly, this finding thus suggests a new method of estimating an "effective" dimensionality of MD trajectories.

It is also worth noting that both the MD-trajectory and the random walk projections show apparent "clusters", for example, for $\tau = 500$ and $\tau = 8000$, which also look quite similar. The fact that such clusters are also seen for the random walk strongly suggests that these are mostly stochastic artifacts and do not point to minima of the underlying free energy landscape.

A closer inspection of the random walk projections offers an additional possible explanation for some of the clusters, which may also apply to the MD trajectory projections. Focusing, for example, at the averaged tICA-projections in Figure 1 immediately before the first abrupt change, one can see that the projection becomes overlaid with a cosine of higher frequency. Particularly at the ends of the curves, and in the presence of noise typical for single trajectories, this high frequency component can also produce apparent "clusters".

In contrast, for the smaller protein (Figure 7) no similarity to the tICA-projections of random walks is observed. In fact, the tICA-projections of the trajectory of the smaller protein show no resemblance to a cosine-like function at all. In light of the above analysis, this finding suggests that this trajectory is sufficiently long to explore one or several minima of the underlying free energy landscape, thereby deviating from a random walk. Further, one may infer that the three clusters seen in the figure actually point to conformational substates and, hence, can serve as proper Markov states.

It is an intriguing question whether or not, for given trajectory length, larger or more flexible proteins tend to more closely resemble random walks.

4. CONCLUSIONS

Here we have analyzed projections of random walks on tICA subspaces and subsequently compared those to tICA-projections of molecular dynamics trajectories of proteins. Our combined analytical and numerical study revealed a staggering complexity of the random walk tICA-projections, which showed a much richer mathematical structure than projections of random walks on principal components (PCA).[21,22]

We attribute this complexity primarily to the fact that, in contrast to PCA, tICA components encode time information on the trajectory and, therefore, extract and process significantly more information. Mathematically, the complex behavior originates from the noncontinuous switch of the order of eigenvalues for increasing lag time τ, when passing through points of eigenvalue degeneracy. At these points, the associated eigenvectors change abruptly, and so do the corresponding projections of both random walks and molecular dynamics simulations. We also find that tICA can be very sensitive to very small changes in the definitions of the involved matrices. In particular, the projections of random walks are very different for the two discussed symmetrization methods.

A particularly striking example is the first abrupt change of the projections onto the two largest eigenvalues. Here, a closer inspection revealed an approximate square root relationship between the lag times at which this occurs and the dimensionality of the random walk. A similar square root law is already known for PCA: Approximately the first $\sqrt{d}$ principal components of random walks resemble cosines.[21]

Comparison of tICA-projections of random walks with those of a large protein (PDB 11AS) revealed striking similarities. This remarkable finding suggests that not only the ensemble properties of the finite protein trajectory resemble those of a random walk, as has been shown earlier via PCA,[21] but also the time correlations of the underlying protein dynamics. Here, the appearance of cosine-like functions in the projections onto the tICA-vectors associated with the longest correlation times

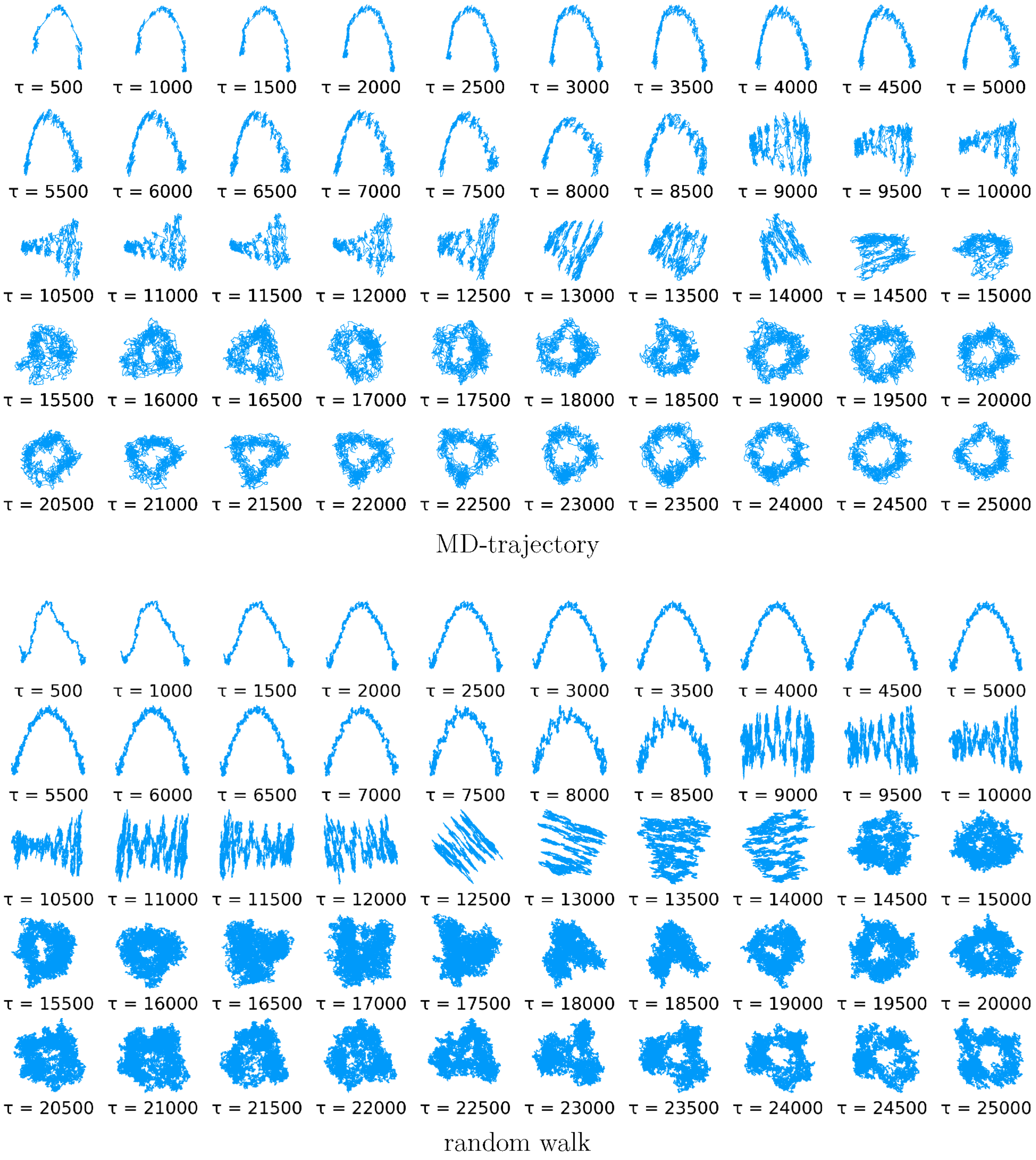


Figure 6. First two tICA-projections of an MD-trajectory of PDB-entry 1IAS (upper group) and those of a 40-dimensional random walk (lower group) for varying lag time τ. In this plot those of the MD-trajectory are smoothed using a moving average to improve readability.

clearly points to a nonconverged trajectory. For the comparatively small lag times typically used, the tICA-projections of random walks almost exactly resemble cosine functions, such that the cosine-content[22] of the tICA-projections should serve as a good quantifier of this.

In contrast, no resemblance to a random walk was seen for the second, smaller protein studied here, indicating that the projection reflects actual features of the underlying conformational dynamics of the protein.

The example in Figure 6 also illustrates the risk of overinterpreting apparent "clusters" seen in the tICA-projections as actual conformational substates,[4,16] which are defined as local minima of the protein free energy landscape that are sufficiently deep for the system to stay there for a certain amount of time.[16] Clearly, it is tempting to also see "clusters" in the random walk projections, which, however, by the definition of the random walk as a diffusion on a flat energy landscape, cannot represent conformational substates. This finding raises concerns for using automated clustering algorithms to identify, for example, folding intermediates or to characterize conformational motions from tICA-projections.[45]

Because the additional parameter of a varying lag time provides a much richer structure and many instead of only one projection (as is the case for PCA), we speculate that the tICA

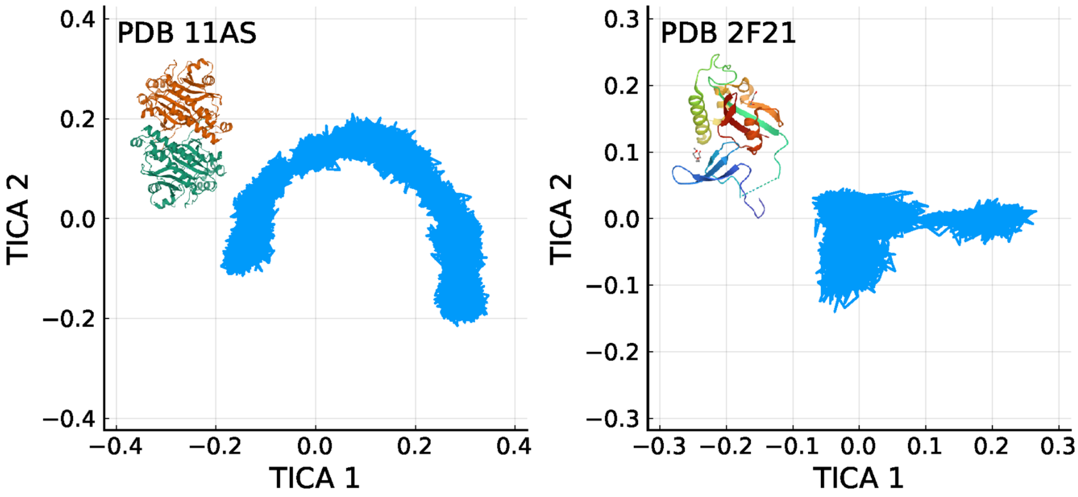


Figure 7. First two tICA-projections of trajectories of the PDB-entries 11AS (on the left) and 2F21 (on the right). The larger protein (11AS) produces a cosine-like shape, while the smaller one does not.

resemblance to a random walk offers a much more sensitive tool to detect lack of convergence in MD trajectories of large biomolecules. Further, by adjusting the dimension of the random walk such as to maximize the similarity to a given MD trajectory, one can estimate the effective dimensionality of the underlying dynamics. The latter idea, as well as precisely how this "effective dimensionality" can be defined, clearly deserves further exploration.

■ AUTHOR INFORMATION

Corresponding Authors

Steffen Schultze − *Max Planck Institute for Biophysical Chemistry, Göttingen 37077, Germany*; orcid.org/0000-0002-0136-8013; Email: sschult@mpibpc.mpg.de

Helmut Grubmüller − *Max Planck Institute for Biophysical Chemistry, Göttingen 37077, Germany*; orcid.org/0000-0002-3270-3144; Email: hgrubmu@gwdg.de

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.1c00273

Funding

Open access funded by Max Planck Society.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Nicolai Kozlowski, Malte Schäffner, and Andreas Volkhardt for very helpful discussions; and Andreas Volkhardt for providing the MD-trajectories for our analysis. This work was supported by the German Ministry of Education and Research, BMBF project 05K20EGA and the German Science Foundation, Grant SFB 1456. This analysis has been implemented using the Julia programming language.[46]

■ REFERENCES

(1) Henzler-Wildman, K.; Kern, D. Dynamic Personalities of Proteins. *Nature* **2007**, *450*, 964−972.

(2) Lewandowski, J. R.; Halse, M. E.; Blackledge, M.; Emsley, L. Direct Observation of Hierarchical Protein Dynamics. *Science* **2015**, *348*, 578−581.

(3) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Proteins: Struct., Funct., Genet.* **1995**, *21*, 167−195.

(4) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The Energy Landscapes and Motions of Proteins. *Science* **1991**, *254*, 1598−1603.

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

(6) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of Folded Proteins. *Nature* **1977**, *267*, 585−590.

(7) de Groot, B. L.; Grubmüller, H. Water Permeation Across Biological Membranes: Mechanism and Dynamics of Aquaporin-1 and GlpF. *Science* **2001**, *294*, 2353−2357.

(8) Zink, M.; Grubmüller, H. Mechanical Properties of the Icosahedral Shell of Southern Bean Mosaic Virus: A Molecular Dynamics Study. *Biophys. J.* **2009**, *96*, 1350−1363.

(9) Perilla, J. R.; Schulten, K. Physical Properties of the HIV-1 Capsid from All-Atom Molecular Dynamics Simulations. *Nat. Commun.* **2017**, *8*, 15959.

(10) Perilla, J. R.; Goh, B. C.; Cassidy, C. K.; Liu, B.; Bernardi, R. C.; Rudack, T.; Yu, H.; Wu, Z.; Schulten, K. Molecular Dynamics Simulations of Large Macromolecular Complexes. *Curr. Opin. Struct. Biol.* **2015**, *31*, 64−74.

(11) Bock, L. V.; Blau, C.; Schröder, G. F.; Davydov, I. I.; Fischer, N.; Stark, H.; Rodnina, M. V.; Vaiana, A. C.; Grubmüller, H. Energy Barriers and Driving Forces in tRNA Translocation through the Ribosome. *Nat. Struct. Mol. Biol.* **2013**, *20*, 1390−1396.

(12) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Protein Folding Kinetics and Thermodynamics from Atomistic Simulation. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17845−17850.

(13) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412−425.

(14) Molgedey, L.; Schuster, H. G. Separation of a Mixture of Independent Signals Using Time Delayed Correlations. *Phys. Rev. Lett.* **1994**, *72*, 3634−3637.

(15) Naritomi, Y.; Fuchigami, S. Slow Dynamics of a Protein Backbone in Molecular Dynamics Simulation Revealed by Time-Structure Based Independent Component Analysis. *J. Chem. Phys.* **2013**, *139*, 215102.

(16) Grubmüller, H. Predicting Slow Structural Transitions in Macromolecular Systems: Conformational Flooding. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1995**, *52*, 2893−2906.

(17) Perez-Hernandez, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noe, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139*, 015102.

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

(19) de Groot, B. L.; Daura, X.; Mark, A. E.; Grubmüller, H. Essential Dynamics of Reversible Peptide Folding: Memory-Free Conformational Dynamics Governed by Internal Hydrogen bonds11Edited by R. Huber. *J. Mol. Biol.* **2001**, *309*, 299−313.

(20) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140*, 2386−2396.

(21) Hess, B. Similarities between Principal Components of Protein Dynamics and Random Diffusion. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2000**, *62*, 8438−8448.

(22) Hess, B. Convergence of Sampling in Protein Simulations. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2002**, *65*, 031910.

(23) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(24) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 826−843.

(25) Faradjian, A. K.; Elber, R. Computing Time Scales from Reaction Coordinates by Milestoning. *J. Chem. Phys.* **2004**, *120*, 10880−10889.

(26) Olsson, S.; Noé, F. Mechanistic Models of Chemical Exchange Induced Relaxation in Protein NMR. *J. Am. Chem. Soc.* **2017**, *139*, 200−210.

(27) Xiao, J.; Salsbury, F. R. Na + -Binding Modes Involved in Thrombin's Allosteric Response as Revealed by Molecular Dynamics Simulations, Correlation Networks and Markov Modeling. *Phys. Chem. Chem. Phys.* **2019**, *21*, 4320−4330.

(28) Wu, H.; Nüske, F.; Paul, F.; Klus, S.; Koltai, P.; Noé, F. Variational Koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations. *J. Chem. Phys.* **2017**, *146*, 154104.

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

(30) Antognini, J. M.; Sohl-Dickstein, J. PCA of High Dimensional Random Walks with Comparison to Neural Network Training. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY, USA, 2018; p 10328−10337.

(31) Hyvärinen, A.; Karhunen, J.; Oja, E. *Independent Component Analysis*; John Wiley & Sons, Ltd: New York, 2001; Chapter 18, pp 341−354.

(32) Tong, L.; Liu, R.-w.; Soon, V.; Huang, Y.-F. Indeterminacy and identifiability of blind identification. *IEEE Trans. Circuits Syst.* **1991**, *38*, 499−509.

(33) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29*, 845−854.

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

(35) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120*, 9665−9678.

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

(37) Nakatsu, T.; Kato, H.; Oda, J. Crystal Structure of Asparagine Synthetase Reveals a Close Evolutionary Relationship to Class II Aminoacyl-tRNA Synthetase. *Nat. Struct. Biol.* **1998**, *5*, 15−19.

(38) Jager, M.; Zhang, Y.; Bieschke, J.; Nguyen, H.; Dendle, M.; Bowman, M. E.; Noel, J. P.; Gruebele, M.; Kelly, J. W. Structure-Function-Folding Relationship in a WW Domain. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 10648−10653.

(39) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(40) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.

(41) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(42) Gray, R. M. Toeplitz and Circulant Matrices: A Review. *Found. Trends Commun. Inf. Theory* **2005**, *2*, 155−239.

(43) Davis, P. J. *Circulant Matrices*; Wiley: New York, 1979.

(44) Davis, C.; Kahan, W. M. The Rotation of Eigenvectors by a Perturbation. III. *SIAM J. Numer. Anal.* **1970**, *7*, 1−46.

(45) Sengupta, U.; Carballo-Pacheco, M.; Strodel, B. Automated Markov State Models for Molecular Dynamics Simulations of Aggregation and Self-Assembly. *J. Chem. Phys.* **2019**, *150*, 115101.

(46) Bezanson, J.; Edelman, A.; Karpinski, S.; Shah, V. B. Julia: A Fresh Approach to Numerical Computing. *SIAM Rev.* **2017**, *59*, 65−98.

https://doi.org/10.1021/acs.jctc.1c00273