ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Fortuitous Correlations in Molecular Dynamics Simulations: Their Harmful Influence on the Probability Distributions of the Main Principal Components

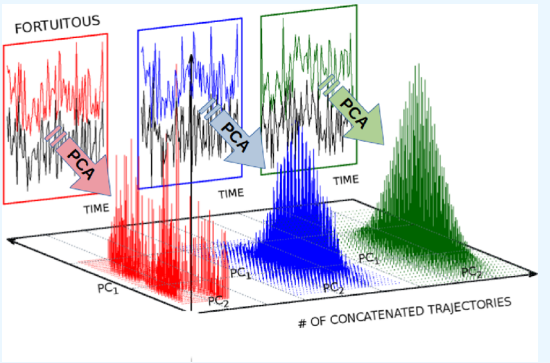
Juliana Palma* and Gustavo Pierdominici-Sottile

Cite This: *ACS Omega* 2024, 9, 20488−20501

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Nonsense correlations frequently develop between independent random variables that evolve with time. Therefore, it is not surprising that they appear between the components of vectors carrying out multidimensional random walks, such as those describing the trajectories of biomolecules in molecular dynamics simulations. The existence of these correlations does not imply in itself a problem. Still, it can present a problem when the trajectories are analyzed with an algorithm such as the Principal Component Analysis (PCA) because it seeks to maximize correlations without discriminating whether they have physical origin or not. In this Article, we employ random walks occurring on multidimensional harmonic potentials to evaluate the influence of fortuitous correlations in PCA. We demonstrate that, because of them, this algorithm affords misleading results when applied to a single trajectory. The errors do not only affect the directions of the first eigenvectors and their eigenvalues, but the very definition of the molecule's "essential space" may be wrong. Additionally, the main principal component's probability distributions present artificial structures which do not correspond with the shape of the potential energy surface. Finally, we show that the PCA of two realistic protein models, human serum albumin and lysozyme, behave similarly to the simple harmonic models. In all cases, the problems can be mitigated and eventually eliminated by doing PCA on concatenated trajectories formed from a large enough number of individual simulations.



INTRODUCTION

The Principal Component Analysis (PCA) is a procedure extensively employed in data science with diverse purposes. It has found widespread use in making sense of data collected from Molecular Dynamics (MD) simulations of biological molecules. In this context, its main goal is to define a low-dimensional vector space, typically with less than twenty degrees of freedom, where the most extensive and (hopefully) most relevant deformations of the molecule can be adequately described. García was the first to propose the implementation of PCA to analyze MD simulations of proteins after noting that the atomic fluctuations of Crambin were highly correlated.[1] Shortly afterward, Berendsen and co-workers popularized the procedure by stating it allows the definition of the "essential space" of proteins: a small subspace that contains the movements required for their functioning.[2] The main point in these two seminal articles was the realization that just a few eigenvectors of the covariance matrix account for the vast majority of the atomic fluctuations observed in a trajectory. Therefore, only this small set would be required to characterize the "essential" macromolecule's dynamics.

The projections of a trajectory onto the PCA eigenvectors are called the Principal Components (PCs). In an article published in 2000, Hess demonstrated that the PCs of a random walker moving on a multidimensional flat potential are cosine functions whose period equals twice the simulation time divided by the component index.[3] He cited several articles reporting protein simulations with this behavior and presented his own examples, suggesting in those cases the simulation time was so short that the protein could not feel the effect of the underlying Potential Energy Surface (PES). In addition, he demonstrated that, for proteins remaining within the same free energy well for a relatively long time, the cosine-like shapes disappear. More worrying but less discussed was Hess's finding that the PCA eigenvalues of the multidimensional random walker (on a flat potential) decay rapidly with the PC index, similar to those computed from protein MD simulations. The unavoidable conclusion from this result is that the fluctuations observed in such simulations are also highly correlated. However, because of the very nature of the model, these

Received: February 21, 2024
Revised: April 4, 2024
Accepted: April 9, 2024
Published: April 23, 2024

https://doi.org/10.1021/acsomega.4c01515
ACS Omega 2024, 9, 20488−20501

correlations must be arbitrary. They have no physical cause but occur just by chance.

The existence of arbitrary correlations in the random walks of pairs of independent variables is well-known in the field of applied statistics.[4] It was already described by Yule in 1926.[5] He called them "nonsense" correlations. Yule's claims remained unnoticed until 1974 when computer simulations provided further proof to the observations he had made by drawing playing cards from previously shuffled packs.[6] Since then, the analysis of these correlations has constituted an important topic in econometric and economic modeling, where it receives the names of "spurious regressions" or "spurious correlations".[7,8] Recent research was able to determine the analytical expression for the second moment of the correlation coefficient for a pair of independent variables carrying out long random walks; and even for short walks, a characterization of the expected correlations has been presented.[9] It seems intuitively clear that these spurious correlations, so well characterized when the random walk occurs in two dimensions, will also appear when it takes place in spaces of higher dimensionality. One would expect that the larger the dimensionality of the simulation space, the greater the chances of finding spurious correlations between certain components.

Since MD simulations do not occur on flat potentials, molecular dynamics practitioners could believe the arbitrary correlations mentioned above do not present any concerns to them. The current possibility to afford long-time simulations, where the system necessarily feels the shape of the PES, could even reinforce this view. However, an article by Antognini and Sohl-Dickstein published in 2018 casts doubts on this belief.[10] The authors characterized the PCA of multidimensional random walks on flat potentials and arrived at the same conclusion as Hess regarding the shape of the PCs. Also, they estimated that the first PC accounts for ∼60% of the total fluctuations, the first two components account for ∼80%, and so on. In other words, just a few PCs explained most of the global variance. The authors seem to be unaware of Hess's articles on the subject. This is not surprising as they do research in a completely different field (that of neural network training).

The derivation of Antognini and Sohl-Dickstein, very concise and elegant, starts from writing the whole random walk in matrix form. Then, they recognize that some of the matrices involved are banded Toeplitz matrices which, in the limit of a very large number of steps, n, approach circulant matrices.[11] The final result stems from the known properties of circulant matrices. Their derivation is also quite general. There are just two additional requirements on top of the already mentioned $n \gg 1$. One of them is that the random steps should be taken from a (any) probability distribution with zero mean and finite covariance. The other is that the dimension of the random walkers, d, should be much larger than the number of steps ($d \gg n$). This last assumption is not always fulfilled by MD simulations of biomolecules, but it is adequate for the trajectories of neural networks that move on the vector space of their parameters during the training process. In any case, we note that if the components of the random walker have zero covariance (i.e., they are uncorrelated), the $d \gg n$ condition can be replaced by $d \gg 1$ and the proof is still valid. This requirement seems more adequate, at least as a first approximation, for simulations of molecular systems.

The demonstration given by Antognini and Sohl-Dickstein can be easily extended to a multidimensional random walk occurring on a harmonic potential. They provided a detailed discussion of this extension for the case of an isotropic potential (i.e., when all of the force constants are the same). They found that, when the number of steps tends to infinity ($n \rightarrow \infty$), all of the eigenvalues achieve the same limit, which is determined by the force constant. This is the expected result for a bound system. Moreover, from statistical thermodynamics, we know this value equals k_BT/k, where k is the harmonic constant, k_B is theBoltzmann constant and T is the absolute temperature. But the surprising results appear when the number of steps, even if large, is not infinite. In this case, the PCA eigenvalues decay very rapidly with the PC index, similar to what they do in the case of the flat potential or the MD simulations of proteins. In other words, even though all directions present the same physical characteristics, when the trajectory passes through the filter of PCA, there seem to be some privileged directions along which the system is allowed to perform large fluctuations, and some other nonimportant directions where the movements are "near-constrained". We emphasize that this occurs in long enough trajectories that in no way can be assimilated to a random walk on a flat potential. It is just an artifact created by the PCA algorithm.

In the following sections, we provide examples of the impact of this artifact in analyzing the dynamics of model systems that resemble biomolecules in solution. We will argue that it may affect the very separation of the configurational space into an essential and a nearly constrained space. Also, we will demonstrate that the actual characteristics of the models can be recovered by doing PCA on fictitious trajectories, obtained by concatenating a large enough number of equivalent trajectories. We have already proposed this procedure to attain a reproducible definition of the main PCA eigenvectors of a protein. Here, we demonstrate that it affords a better characterization for all of the eigenvalues and eigenvectors and also provides the correct probability density functions of the principal components. These results could have noticeable consequences in the calculation of conformational entropies based on the quasi-harmonic approximation. Finally, we present the results of extensive PCAs of serum albumin[12] and lysozyme[13] that illustrate how the observations made on simple harmonic models assimilate to those of actual protein models.

SIMULATIONS

We have simulated random walks on simple potential energy models and analyzed them via PCA. Thus, the knowledge of the underlying PES allowed us to discriminate the influence of parameters such as the characteristic frequencies, the simulation time, or the system's dimensionality on the PCA results. In particular, we were interested to see how these parameters influence the definition of the essential space. We have also performed extensive MD simulations of serum albumin and lysozyme in a water solution and analyzed them with PCA, aiming to evaluate if the insights gained from the simpler models could help to understand the behavior of more realistic systems. In the following sections, we describe the protocols employed in the simulations and provide the numerical details required for reproducing the results.

Model Systems. Our model systems consist of vectors of dimension N_d. Each of their components, x_i, with $i = 1, ..., N_d$,

https://doi.org/10.1021/acsomega.4c01515
ACS Omega 2024, 9, 20488−20501

evolves according to the Langeving equation in the diffusion limit:

$$F_s(x_i) + F_r(t) - \zeta \dot{x}_i = 0 \tag{1}$$

Here, $\dot{x}_i$ is the time derivative of component x_i, ζ is a friction coefficient that is the same for all components, $F_s(x_i)$ is the systematic force acting on x_i, and $F_r(t)$ is a random force that fulfills

$$\langle F_r(t) \rangle = 0 \tag{2}$$

$$\langle F_r(t) F_r(t') \rangle = 2\zeta k_B T \delta(t - t') \tag{3}$$

where $\delta(t)$ is Dirac's delta function and $\langle ... \rangle$ denotes ensemble average. Note that eq 1 assumes that the systematic force acting on coordinate x_i does not depend on $x_j \neq x_i$. This is because the potential energy function ruling these coordinates was chosen as the sum of independent harmonic oscillators:

$$V_s(\{x_i\}) = \sum_{i=1}^{N_d} \frac{1}{2} k_i x_i^2 \tag{4}$$

so that $F_s(x_i) = -k_i x_i$. Therefore, by construction, our model systems are just sets of independent coordinates, assembled into a vector of dimension N_d.

Multiplying eq 1 by the time step, Δt, and reorganizing, one obtains

$$x_i(t + \Delta t) = \left(1 - \frac{k_i \Delta t}{\zeta}\right) x_i(t) + \frac{F_r(t)}{\zeta} \Delta t \tag{5}$$

when the systematic forces are calculated from the potential energy function of eq 4. In our models, we set $k_B T = \zeta = \Delta t = 1$. Therefore, the propagation equation was

$$x_i(t + \Delta t) = (1 - k_i) x_i(t) + R_r(t) \tag{6}$$

with $\langle R_r(t)\, R_r(t') \rangle = 2\delta(t - t')$. The trajectories were started with the coordinates at the minimum of their harmonic potentials. For these settings, the equipartition theorem states that, at thermodynamic equilibrium, the ensemble average of the squared displacement of coordinate x_i is

$$\langle \Delta x_i^2 \rangle_{eq} = k_i^{-1} \tag{7}$$

where $\Delta x_i(t) = x_i(t) - x_i(0) = x_i(t)$ measures the change of coordinate x_i from its initial value to the current one.

We ran simulations with two alternative models that employ the potential energy function of eq 4. Both of them have $N_d = 100$. In the D model (by degenerate) all k_i's equal 0.01. In the S model (by spectrum), $k_i = 0.001 \times i$, with $i = 1, ..., N_d$. Thus, the S model's smallest and largest constants are 10 times smaller and 10 times larger than those of the D model, respectively. Between these limits, the constants are equally spaced. To set an appropriate value for the simulation's length, we evaluated the time evolution of the ensemble-averaged squared displacement:

$$\langle \Delta x_i^2(t) \rangle = \frac{1}{N_{set}} \sum_{m=1}^{N_{set}} (x_i^{(m)}(t))^2 \tag{8}$$

where m is a label that identifies the alternative simulations used in the calculation. Figure S1, in the Supporting Information, shows the averages computed from $N_{set} = 5000$ simulations of a single coordinate that obeys eq 6 with either $k_i = 0.001$, 0.01, or 0.1. The resulting curves are compared with the behavior expected for the random walk on a flat potential. We note that all curves deviate soon from that corresponding to the flat potential. However, they require a somewhat longer time to reach their thermodynamic equilibrium values. For $k_i = 0.1$, the expected average is reached in about 15 steps, but it takes about 250 steps to do that when $k_i = 0.01$ and ~2500 steps when $k_i = 0.001$. Based on these results, we set the number of steps, $N_s = 1000$, for the simulations of the D model, and $N_s = 4000$ for the S model. This setting ensures that, at the end of each simulation, all coordinates have felt the curvature of the potential energy function acting on them, and that their ensembles approximate the thermodynamic equilibrium behavior. We note, however, that these simulation lengths are not enough to establish ergodicity since the time average of the square displacements,

$$\overline{(\Delta x_i(t))^2} = \frac{1}{N_s} \sum_{k=1}^{N_s} (\Delta x_i(t_k))^2 \tag{9}$$

where N_s is the number of steps of the simulation, $t_k = k \times \Delta t$ and $t = N_s \times \Delta t$, requires much longer times to approximate the thermodynamic equilibrium values. This can be seen in Figure S2, in the Supporting Information, where we plotted $\overline{(\Delta x_i(t))^2}$ for five equivalent trajectories of a single coordinate propagated using eq 6 with $k_i = 0.01$. In this Article, we will use brackets to denote ensemble averages and overbars for the time averages. However, for the sake of simplicity, we will hereafter not explicitly indicate the time dependence of these averages. According to the previous discussion, it is clear that the ensemble averages will be time-independent for the simulation lengths used in this work while the time averages will not.

We finally emphasize that the displacements Δx_i introduced above should not be confused with the fluctuations $x_i - \bar{x}_i$ employed in the calculation of the covariance matrix. When evaluating fluctuations, we subtract the average of x_i in the trajectory, $\bar{x}_i$, from its value at a given time. Instead, when calculating the displacements, we subtract the initial value, $x_i(0)$, which is zero for our simulations. For ergodic trajectories run on harmonic potentials, $\bar{x}_i$ equals the minimum of the potential. Therefore, if the simulations are started at the minimum, as we did, fluctuations and displacements should be the same. However, at intermediate times, these quantities can differ significantly, as we show in Figure S3, of the Supporting Information.

Protein MD Simulations. We carried out MD simulations of Human Serum Albumin (HSA) and lysozyme in their apo form. HSA is a globular protein with 609 amino acids that contains 17 disulfide bridges (UniProtKB: P02768). To build the computational model, we used the structure 1AO6 of the Protein Data Bank (PDB). It only contains 578 residues, from Ser29 to Ala606. The PDB file was fed into the LEAP module of AMBER18.[14] Water molecules were then introduced to fill an octahedral cell whose walls were 15 Å from the nearest protein atom. The model was neutralized by adding Na^+ ions. The ff14SB force field[15] was used for the protein, the TIP3P for the water molecules,[16] and the frcmod.ionsjc_tip3p parameters for the ions.[17] The PMEMD module of AMBER18 was used to run the simulations.

After model building, the system was minimized. Then, it was heated at a constant volume for 2 ns to reach a temperature of 310 K using the Langevin thermostat[18] with a collision frequency of 1.0 ps^{-1}. The SHAKE algorithm was

https://doi.org/10.1021/acsomega.4c01515
ACS Omega 2024, 9, 20488−20501

used to constrain the bonds involving hydrogen atoms,[19] and the Particle Mesh Ewald method[20] with a cutoff radius of 10.0 Å was applied to compute the nonbonded interactions. After the heating, we changed from NVT to NPT conditions and ran another 50 ns to allow density to relax and equilibrate. The pressure was controlled with a Berendsen barostat[21] with a coupling constant of 2.0 ps. The time step was set to 2.0 fs. We used the last snapshot of the equilibration stage as the starting point for 100 alternative production runs. All of them lasted for 10 ns, and samples were taken every 0.1 ns. The initial atomic velocities of the simulations were randomly taken from a Maxwellian distribution.

We followed the same procedure to conduct simulations of human lysozyme, an enzyme that has undergone various computational studies aimed at identifying its functional modes.[22−25] In particular, PCA studies have shown that the enzyme's most significant collective coordinate describes a hinge-bending oscillation that expands or contracts the volume of the catalytic cleft, a movement involved in substrate binding and product release.[24,25] Human lysozyme has 148 amino acids and four disulfide bridges (UniProtKB: P61626). We built the computational model using structure 1REX from the PDB, which is formed by 130 residues, from Lys19 to Val148.[13] The production stage of this system also consisted of 100 independent trajectories of 10 ns.

■ ANALYSIS

All simulations were analyzed via PCA using the standard procedure. For the D and S models, the analysis was carried out on the vectors of dimension $N_d = 100$, whose components evolved according to eq 6. For the MD simulations of HSA and lysozyme, the vectors were built with the Cartesian coordinates of the C_α atoms, after eliminating the effects of global translation and rotation.[26] Therefore, in these cases, $N_d = 3N_{at}$, where N_{at} is the number of C_α atoms in the computational model. Thus, N_d evaluates to 1734 for HSA and 390 for lysozyme.

The elements of the covariance matrix, of dimension $N_d \times N_d$, were calculated as

$$C_{ij} = \frac{1}{N_s}\sum_{k=1}^{N_s}(x_i - \overline{x_i})(x_j - \overline{x_j}) \tag{10}$$

where N_s is the number of snapshots in the trajectory under PCA scrutiny. Thus, when the analysis is carried out on data collected from an individual trajectory, C_{ij} is the time average of the product between the instantaneous fluctuation of coordinate x_i and that of coordinate x_j, $\overline{(x_i - \overline{x_i})(x_j - \overline{x_j})}$, and the diagonal elements contain the time average of the squared fluctuations, $\overline{(x_i - \overline{x_i})^2}$. We also analyzed fictitious trajectories obtained by concatenating more than one individual trajectory. In those cases, the calculation of C_{ij} implies an average on the set of simulations under consideration, in addition to the time average.[27] In any case, the columns of matrix $\mathbf{R}$ that diagonalizes $\mathbf{C}$,

$$\mathbf{R}^T\mathbf{C}\mathbf{R} = \Lambda \tag{11}$$

are the PCA eigenvectors, $\mathbf{v}_i$. The projection of the original vectors onto the $\mathbf{v}_i$'s are the principal components, PC_i, while the diagonal elements of matrix Λ, λ_i, are the square fluctuations of the PC_i's.

For both simple models and MD simulations of HSA and lysozyme, we found significant differences between the results of alternative, individual trajectories. An example of this is given in Figure S3, which depicts the fluctuations of the principal components as a function of the PC index, for five illustrative trajectories of the D model. In the same picture, we also plotted the mean squared fluctuations of the components of the original vectors, $\overline{(x_i - \overline{x_i})^2}$, and the squared displacements $(\Delta x_i)^2$. In the last two cases, we followed a decreasing order, instead of the index order, to facilitate the comparison with the PCA eigenvalues. The displacements reveal that some of the original components have traveled long distances while others have, apparently, hardly moved. We note, however, that the small displacements typically occur when the coordinate comes back over its steps after traveling some distance. The calculation of the time-averaged fluctuations levels off the differences observed in the displacements, while applying the PCA algorithm exacerbates the moderate disparities observed in the fluctuations. This is the expected result because PCA was designed to find the linear combinations of the original coordinates that maximize the squared fluctuations of the first components. As the procedure involves a similarity transformation that does not change the trace of the covariance matrix, the sum of all squared fluctuations remains constant. Therefore, maximizing the fluctuations along the first eigenvectors implies reducing those of the highest-order eigenvectors. Anticipating the arguments provided in the Discussion section, it seems appropriate here to highlight that all of the directions of the D model are physically equivalent. Therefore, the differences among fluctuations and displacements plotted in Figure S3 happened just by chance and were remarkably enlarged by applying PCA.

The great disparity found between the results of the individual trajectories indicates that the problem must be treated statistically. Thus, for the simple models and the MD simulations of HSA, we ran a large number of trajectories and carried out a statistical analysis of their PCA results. In particular, we computed the probability distribution of the PCA eigenvalues, as well as the minimum, maximum, and average values observed in these large sets of simulations. We determined 100 000 trajectories for the simple models, while 100 MD simulations were performed for HSA and lysozyme. Also, we carried out a statistical analysis of the PCA results obtained from concatenated trajectories. In a previous article, we noted that this procedure improves the consistency of the subspace defined by the first eigenvectors (the so-called essential space).[28] Here, we analyze how it modifies the whole eigenvalue spectra, as well as the probability density functions of the first principal components. These modifications have important implications for the very definition of the essential space.

■ RESULTS

General Aspects of the Simple Models. In this section, we discuss the general behavior of the PCA of models D and S. Figure 1 presents the spectra of the normalized eigenvalues determined from the D model. The analysis was performed on the 100 000 individual trajectories and on 10 000, 1000, and 100 concatenated trajectories, formed by combining 10, 100, or 1000 individual simulations, respectively. However, the figure only shows the results obtained with the individual trajectories and the concatenated trajectories formed by 1000

https://doi.org/10.1021/acsomega.4c01515
ACS Omega 2024, 9, 20488−20501

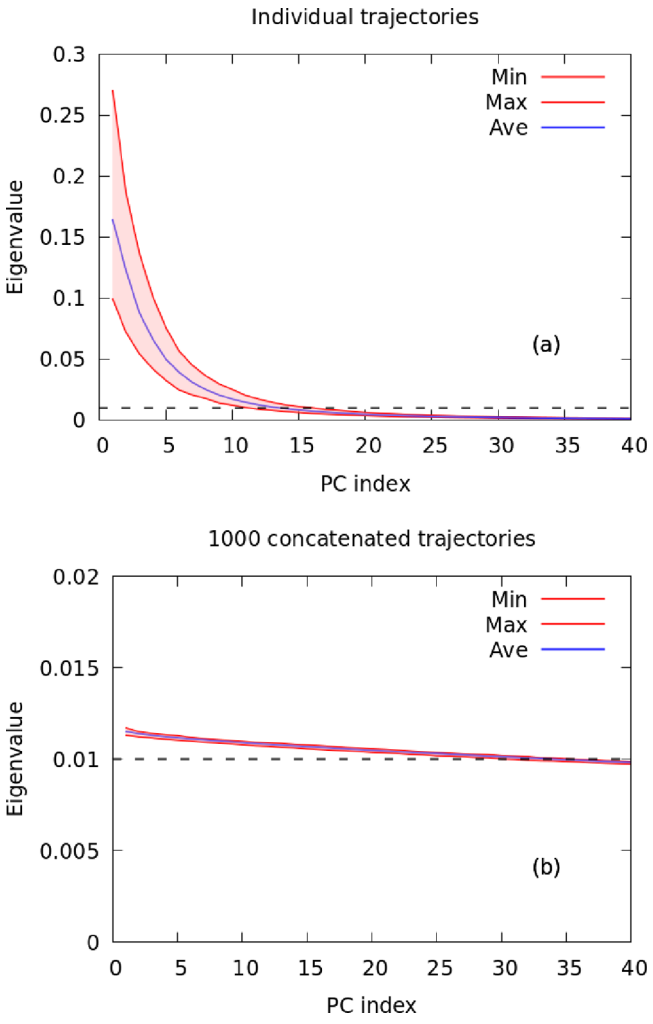


Figure 1. Normalized eigenvalues as a function of the PC index for the PCA performed on trajectories of the D model with k_i = 0.01. Panel a shows the results obtained with the 100 000 individual trajectories. Panel b presents the results computed from a set of 100 trajectories, each of them formed by combining 1000 individual elements. The red lines indicate the minimum and maximum values observed in the calculation while the blue line corresponds to the average. The dashed line indicates the value that would be obtained with an ergodic simulation.

elements. Plots for the whole data set are provided in Figure S4 of the Supporting Information. Panel a of Figure 1 depicts the results from the individual trajectories. We see that the eigenvalue spectra are typical scree plots, as those usually found when doing PCA on MD simulations of proteins. At first glance, this result is not surprising because we are used to those plots. However, it becomes confounding when we note that all directions of the N_d-dimensional configurational space should be equivalent. Besides, as we explained above, the components of the original vectors evolve independently of one another. In other words, the underlying PES does not impose correlations between them. Despite that, the shape of the spectra suggests the existence of significant underlying correlations that would allow the system to move almost freely in some selected directions while being mostly constrained in others. This suggestion is wrong. There are no selected directions, as can be confirmed by analyzing the first PCA eigenvectors, $\mathbf{v}_1$, calculated with alternative individual trajectories. As an example, we present in Figure S5 of the Supporting Information, the squared components for the $\mathbf{v}_1$ vectors obtained from five alternative trajectories. It can be readily seen that components that have important contributions to this vector in one trajectory have modest contributions in others. This evidences that the directions of the $\mathbf{v}_1$ vectors are random. This idea is further confirmed by the calculation of the scalar product between $\mathbf{v}_1$ vectors determined from alternative individual trajectories. Figure 2 depicts the probability distribution function for the absolute value of these products. Far from having a peak near 1, the function becomes negligible for values greater than 0.3, and the most likely scalar product is zero. We note that the larger the dimensionality of the system, the higher the peak at zero (data not shown). As the dimensionality of the degenerate space increases, obtaining a null scalar product becomes more likely.

Panel b of Figure 1 shows that, by doing PCA on a concatenated trajectory with a large enough number of individual simulations, the misleading appearance of the eigenvalue spectra is corrected. The aspect of a scree plot disappears, and all eigenvalues tend to the expected value. The first eigenvalues converge from above and the last ones from below. The more trajectories are combined, the better the results are, as can be noted in Figure S4 of the Supporting Information. This outcome is unsurprising if we consider that doing PCA on concatenated trajectories is equivalent to carrying out an ensemble average on top of the time average one performs in the PCA of individual trajectories. As we showed above, ensemble averages converge to their equilibrium values much more rapidly than the time averages. To further support this appreciation, we show in Figure 3 the probability distribution function of selected PCA eigenvalues, for both the individual trajectories and the concatenated trajectories formed with 1000 elements. We see that the first eigenvalues of the individual trajectories present broad distributions, whose maxima are between 5 and 15 times larger than those expected for an ergodic trajectory. At the same time, the distributions of the last eigenvalues collapse on the left side of the plot, having peaks below the expected value. For the sake of clarity, these data are not shown in Figure 3. On the other hand, the PCA eigenvalues computed by concatenating 1000 single trajectories exhibit thin distributions, just slightly shifted from their expected value of 100. The peak of the distribution corresponding to the first eigenvalue is 16% larger than expected while that of the last eigenvalue is 20% smaller. To conclude the presentation of the results of the D model, we note that the directions of the vectors obtained by concatenating 1000 single trajectories are, nonetheless, random. This is just a consequence that all the directions of the 100-dimensional space are equivalent.

The situation is different when we analyze the results of the S model. In this case, each of the original components has a distinct harmonic constant and therefore a different expectation for its average displacement. Nonetheless, the eigenvalue spectra from the individual trajectories have a shape similar to that of the D model, as can be seen in Figure 4a. In particular, we note that the first eigenvalues present broad distributions similar to those of Figure 1a. To emphasize this point, Figure 5 shows the probability distribution functions for the eigenvalues of the first five eigenvectors. The components of selected PCA eigenvectors, computed from a randomly chosen individual trajectory of the S model, are shown in Figure 6. It can be seen that the first eigenvector only has significant contributions

https://doi.org/10.1021/acsomega.4c01515
ACS Omega 2024, 9, 20488−20501

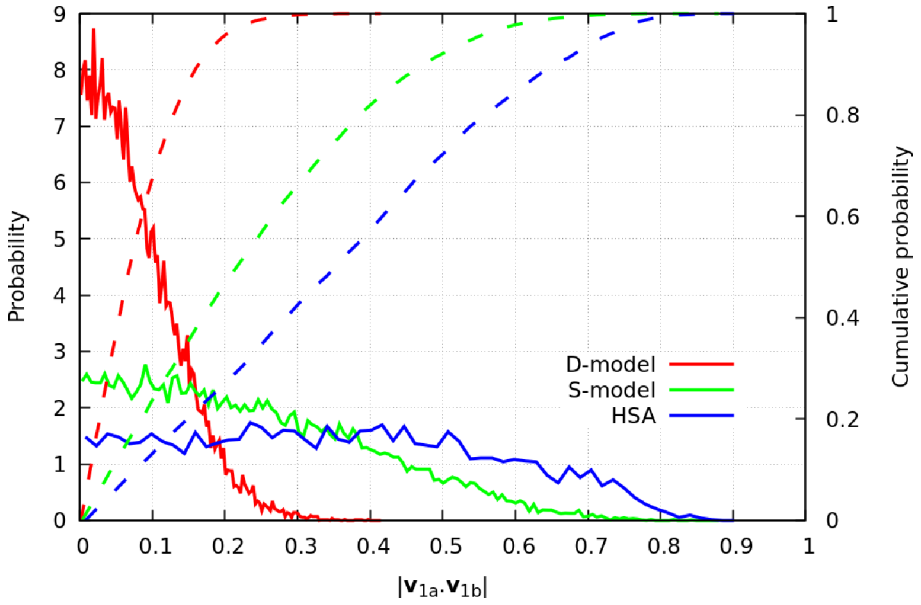


Figure 2. Probability distribution functions (solid lines) and cumulative probability (dashed lines) for the absolute value of the scalar product between the first PCA eigenvectors, $\mathbf{v}_1$, computed from different individual trajectories. Red lines are used for the results obtained with the D model, green lines for the S model, and blue lines for the MD simulations of HSA. For the simple models, the analysis was carried out with the 499 500 alternative pairs of vectors that can be obtained from 1000 individual trajectories. For HSA, we used the 4950 products formed from 100 individual trajectories.

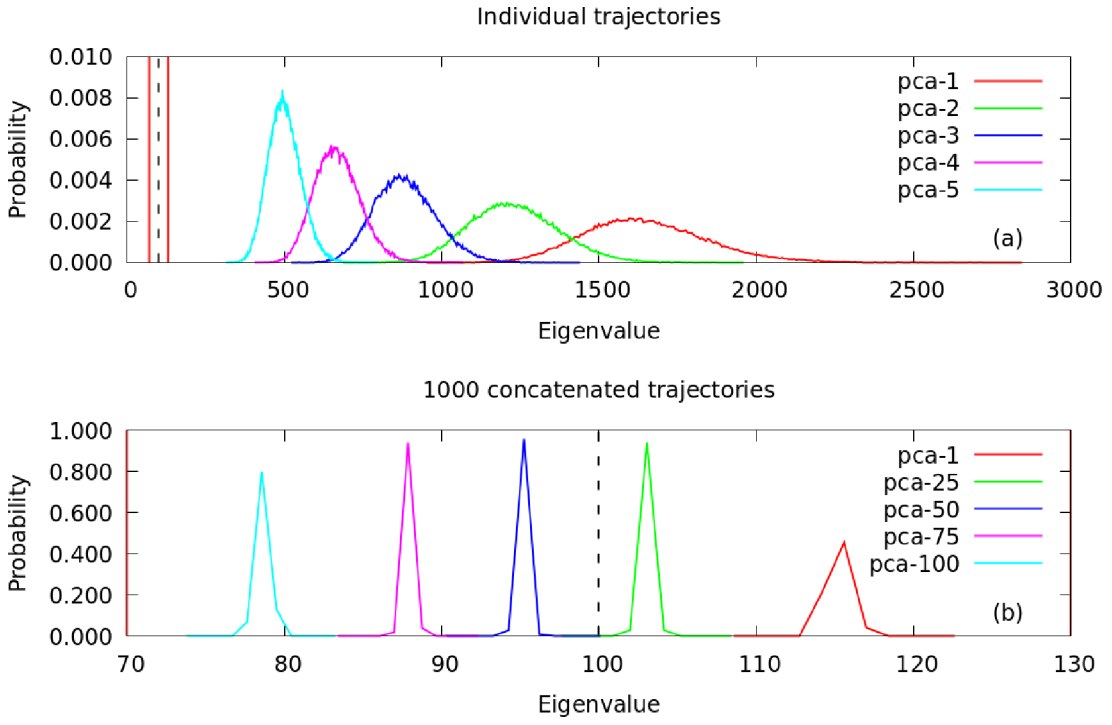


Figure 3. Probability distribution functions for selected PCA eigenvalues obtained with the D model. Panel a shows the results obtained with the 100 000 individual trajectories. Panel b presents the results computed from a set of concatenated trajectories, each of them formed by combining 1000 individual simulations. The dashed black lines indicate the expected value. The vertical red lines of panel a indicate the limits of the *x*-range of panel b so that the differences between their scales can be better appreciated.

from the original components with the smallest ∼20 or 30 k_i's. The following three to four eigenvectors have a similar appearance (data not shown). On the contrary, the last eigenvector has contributions from the original components with the largest harmonic constants. However, in this case, the number of contributing modes is larger and their contributions are more evenly distributed than those of the first mode. Eigenvectors in the middle of the eigenvalue range have noticeable contributions from the whole range of original components. A paradigmatic example of this is vector $\mathbf{v}_{50}$, also shown in Figure 6. In summary, even though, in general, the PCA eigenvectors are random mixtures of the original

https://doi.org/10.1021/acsomega.4c01515
ACS Omega 2024, 9, 20488−20501

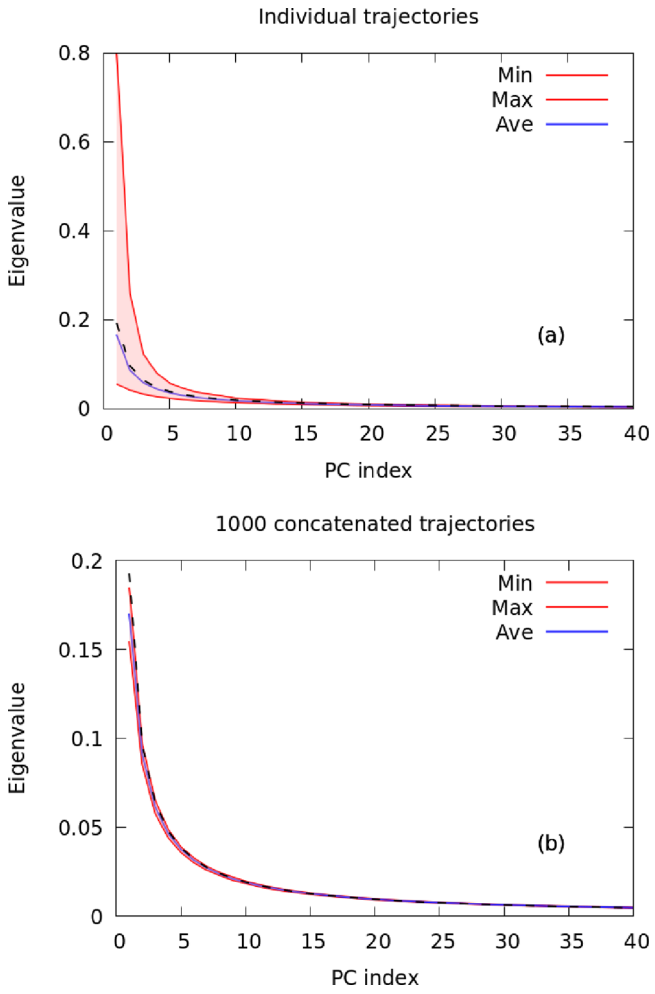


Figure 4. Normalized eigenvalues as a function of the PC index for the PCA performed on trajectories of the S model. Panel a shows the results obtained with the 100 000 individual trajectories. Panel b presents the results computed from a set of 100 concatenated trajectories, each of them formed with 1000 individual elements. The red lines indicate the minimum and maximum values observed in the calculation, while the blue line corresponds to the average. The dashed line depicts the value that would be obtained with an ergodic simulation.

components, the first eigenvectors only have contributions from the modes with the largest amplitudes. This scenario is more appropriate for the definition of an essential space than the one found with the D model.

Panel b of Figure 4 demonstrates that, by concatenating a large enough number of trajectories, the eigenvalue spectra of the S model get pretty close to the expected one. The distributions of the eigenvalues get thinner and their maxima closer to the correct values (see Figure 5). The more trajectories are combined, the better the results are, as can be seen in Figure S6 of the Supporting Information. Of course, this improvement in the eigenvalues is caused by a better definition of the eigenvectors. In Figure 7, we present the contribution of the original components to selected eigenvectors determined from a concatenated trajectory constructed with 1000 individual simulations. It can be seen that the first eigenvector almost agrees with the original component with the smallest k_i. Eigenvectors with higher indexes have noticeable contributions from several of the original modes. However, in all cases, there is a perfect correspondence between the eigenvector index and the order of the component that makes the largest contribution. For example, the largest contribution to vector $\mathbf{v}_{50}$ comes from the component with $k_i = k_{50} = 0.001 \times 50$.

General Aspects of Protein Models. We have observed that the simulations of HSA and lysozyme display comparable behavior in their general aspects. Accordingly, to be brief, we present here the results of HSA and provide those corresponding to lysozyme in the Supporting Information. Instead, these proteins differ in the probability distributions of their main principal components. We discussed these features in the next section.

Panel a of Figure 8 shows the eigenvalue spectra determined from the 100 individual trajectories of HSA. The similarity with the spectra derived from the simple models is considerable. In particular, we call attention to the broad range of possible values of the first eigenvectors, a feature that manifests its randomness. To further confirm this impression, we calculated the absolute value of the scalar product between the first PCA eigenvectors, determined from different individual simulations. We carried out this computation for the 4950 putative pairs that can be formed from 100 trajectories. Then, we evaluated its probability distribution function. The results are shown in Figure 2, where they can be compared with those of models D and S. We see that the results from HSA are somewhat better than those of the S model, which in turn are the best among the simple models. Nevertheless, the chances of obtaining low values of the scalar product are high. For example, the probability of getting a scalar product smaller than 0.5 is 65%. Panel b of Figure 8 contains the spectra obtained from two concatenated trajectories each formed with 50 individual simulations. We remark that there are no common elements between the two sets. The simulations included in one of them are excluded from the other. Therefore, the agreement between the two curves indicates that a fairly good convergence is already achieved with 50 individual trajectories. Figure S7, in the Supporting Information, depicts the curves of eigenvalues computed from individual and concatenated trajectories of lysozyme. The similarities with the curves of Figure 8 can be readily appreciated.

In contrast with the analysis of the simple models, for these realistic protein models, we *a priori* ignore the shape of the underlying PES. Accordingly, evaluating the convergence of the eigenvectors with plots such as those of Figures 6 and 7 requires some assumptions. We, therefore, considered the eigenvectors derived from the concatenated trajectory formed with 100 elements as the reference basis set. Then, we determined the components of the first five PCA eigenvectors of an individual simulation on this basis. We present the results in Figure 9, which reveals that these PCA eigenvectors are rather ill-defined. In addition, we projected the first five eigenvectors of one of the concatenated trajectories with 50 elements into the basis set of eigenvectors of the other. The results are presented in Figure 10. It can be readily noted that the convergence is much better in this case, although there seems to be some mixing between eigenvectors 1 and 2. This finding is not surprising because these vectors are almost degenerate. The first and second eigenvectors explain ∼20% and ∼18% of the total fluctuations, respectively, while the following one accounts for less than ∼10%. We recall degenerate vectors cannot be univocally defined. In any case,

https://doi.org/10.1021/acsomega.4c01515
ACS Omega 2024, 9, 20488−20501

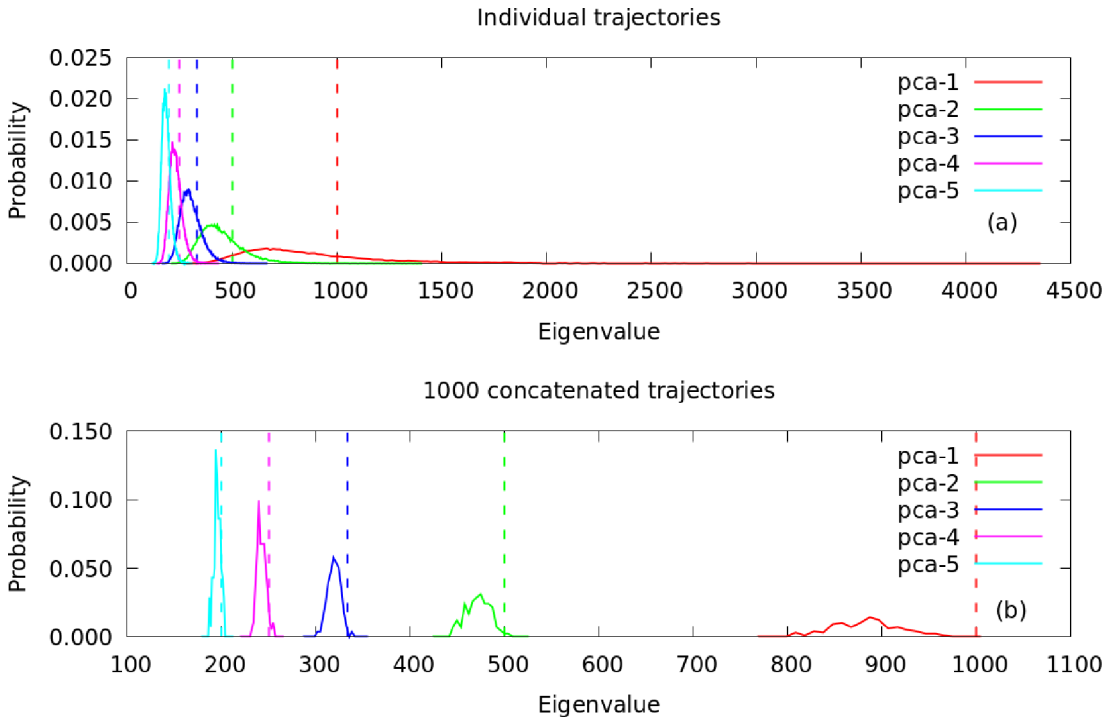


Figure 5. Probability distribution function for selected PCA eigenvalues obtained with the S model. Panel a shows the results obtained with the 100 000 individual trajectories. Panel b presents the results computed from a set of concatenated trajectories, each of them formed by combining 1000 individual simulations. In both panels, the dashed lines indicate the expected value.

the convergence of the PCA computed by concatenating 100 simulations would be even better than that of 50, giving support to the assumption made to create Figure 9.

Plot analogous to those of Figures 9 and 10, but computed from the simulations of lysozyme, are provided in Figures S8 and S9 of the Supporting Information. They are qualitatively similar to those of HSA, although there are differences in the details, as expected. In particular, for lysozyme, modes v_3 and v_4 of the concatenated trajectories are almost degenerate. For this reason, they get mixed up: v_3 of one of them is v_4 of the other one, and vice versa.

Finally, we believe an illustrative idea of how misleading the PCA of individual trajectories can be is obtained by comparing movies S1 and S2 of the Supporting Information. They show the animation of the first PCA eigenvector computed from an individual simulation (movie S1) and the concatenated trajectory with 100 elements (movie S2) of HSA. The difference between the two collective motions is readily noticeable. An analogous comparison can be done between movies S3 and S4, which correspond to lysozyme.

Effect on the Essential Space. As stated previously, one of the most extended applications of PCA in the analysis of MD simulations consists of determining the essential space of biomolecules: the small subspace whose directions describe the largest, and hopefully functional, molecule's deformations. One of the criteria followed to build such space is to include the first PCA eigenvectors until the sum of their eigenvalues reaches a certain fraction of the total fluctuations observed in the trajectory. As demonstrated by the results of Figures 1 and 4, the procedure can afford misleading results. The problem is particularly severe if the motions with the largest fluctuations are degenerate or near degenerate. In that case, the PCA algorithm can create the illusion that some directions of the degenerate space allow large fluctuations while others are almost restrained, producing an ill-defined essential space.

Another criterion consists of computing the first principal components for all the snapshots sampled from the simulation and then evaluating the probability density function of these components. Finally, the essential space is built with the modes whose components present non-Gaussian distributions. Panel a of Figure 11 shows the distributions of selected principal components computed from one of the trajectories of the S model. All the individual trajectories we analyzed afford qualitatively similar results but differ in the particular shapes of the curves. We can see that the first PCs have non-Gaussian distributions, typically showing more than one peak while, as the principal component's index increases, they assume an approximated Gaussian shape. In contrast, panel b of Figure 11 shows that the distributions calculated from a concatenated trajectory with a large enough number of elements present Gaussian shapes. Not only does the form of the distributions agree with the expectations, but also their standard deviations do. In other words, these distributions fully agree with the predicted ones.

We carried out the same analysis for HSA and lysozyme and found significant differences between them. Figure 12 shows the distributions for selected principal components obtained with one of the individual simulations of HSA (panel a) and the concatenated trajectory formed with 100 elements (panel b). We can see that the behavior of this realistic protein model almost replicates the one observed in the extremely simple S model. In particular, the first eigenvalue, which is highly anharmonic when computed from a single trajectory becomes almost a Gaussian function when evaluated with the concatenated trajectory. In turn, this result reveals that the PES of this protein can be adjusted to a sum of harmonic

https://doi.org/10.1021/acsomega.4c01515
ACS Omega 2024, 9, 20488−20501

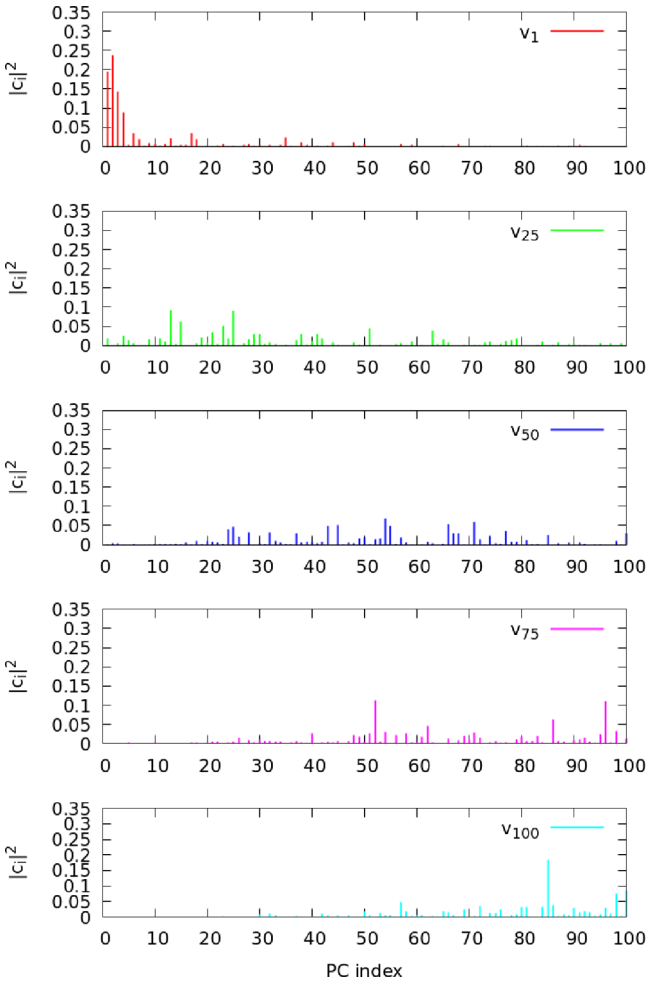


Figure 6. Squared components of selected PCA eigenvectors derived from an individual trajectory of the S model. To facilitate the comparison, all plots employ the same scale on the *y* axis.

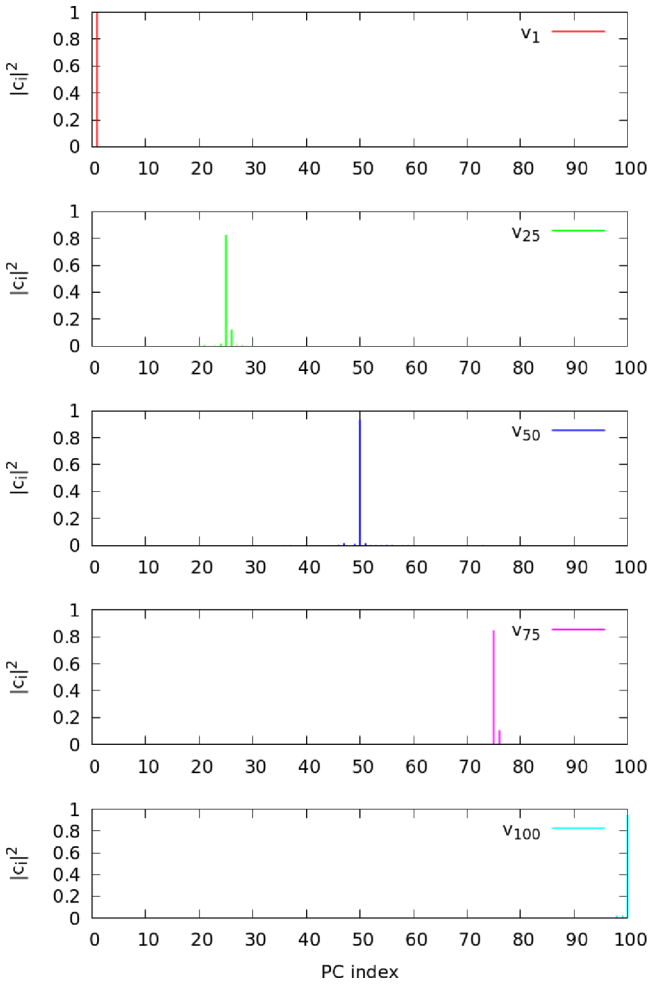


Figure 7. Squared components of selected PCA eigenvectors obtained from a concatenated trajectory formed with 1000 individual trajectories of the S model.

functions (at least the fraction of the PES sampled with our set of simulations).

Figure 13, on the other hand, shows the distributions for selected principal components obtained with one of the individual simulations of lysozyme (panel a) and the concatenated trajectory formed with 100 elements (panel b). The results corresponding to a single trajectory are similar to those of HSA. However, something different occurs with the concatenated trajectory. In this case, PC_1 does not acquire a Gaussian shape. Instead, it presents a broad multimodal distribution with the largest maximum at ∼7.0 Å, another one with a small shoulder at the origin, and a long tail that extends to −16.0 Å. The distributions of PC_1 computed from two alternative concatenated trajectories, each consisting of 50 elements, are highly comparable to one another and also to the concatenated trajectory formed with 100 elements. This implies that running 50 individual simulations is adequate to attain convergence of this function. We note that the probability distribution of PC_2 also somewhat deviates from the Gaussian shape. Figure S10, in the Supporting Information, shows the free energy landscape (FEL) of lysozyme as a function of PC_1 and PC_2. These coordinates describe the well-known hinge bending motion of this enzyme.[24,25] The observation of the structures at the alternative regions of the FEL reveals that the closed structure is the most stable. Nevertheless, the system only needs to surmount barriers of ∼1−2 k_BT to reach the widest open conformation.

DISCUSSION

PCA is a procedure extensively employed in the analysis of MD simulations of proteins, to separate collective movements that carry out the largest fluctuations from those that are nearly restrained. It is believed the amplitude of these motions is dictated by the underlying potential energy surface and that their directions span a subspace that contains the functional movements of the molecule. This is the so-called essential space of the protein. The essential space is built with the directions of the PCA eigenvectors with the largest eigenvalues so that they account for a given (large) fraction of the total fluctuations observed in the simulation. Typical values are around ∼80%. Alternatively, the snapshots collected from the simulation are projected onto the first PCA eigenvectors to determine the principal components. Then, the probability density functions of these components are determined, and all directions with non-Gaussian distributions are considered to belong to the essential space.

In this work, we scrutinized the assumptions that guide the use of PCA in the analysis of MD simulations and the practices employed to determine the essential space. With that aim, we applied the PCA algorithm to random walks that occur on very

https://doi.org/10.1021/acsomega.4c01515
ACS Omega 2024, 9, 20488−20501

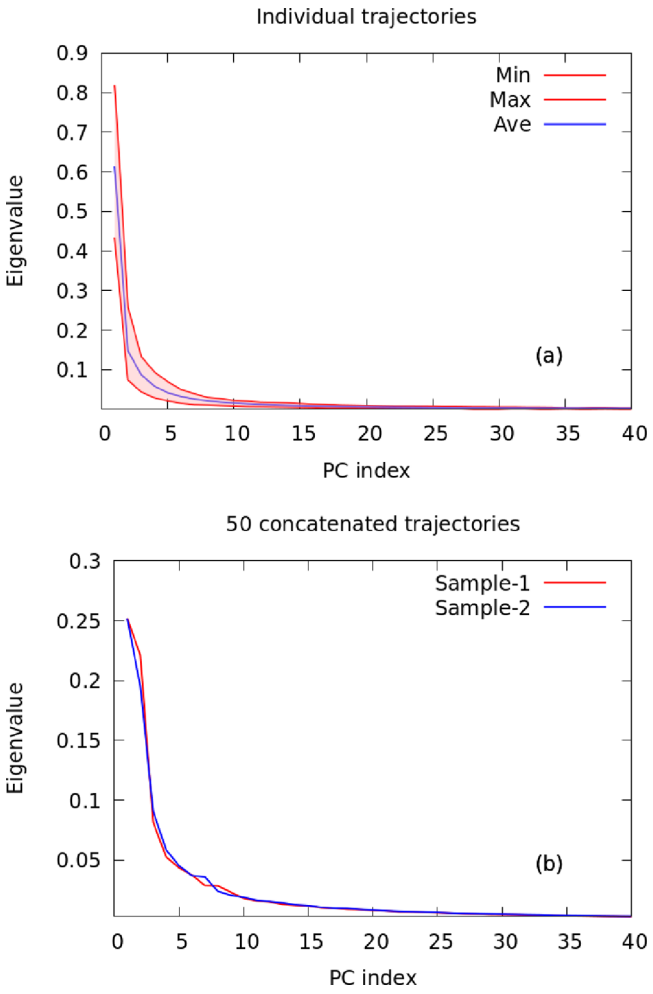


Figure 8. Normalized eigenvalues as a function of the PC index calculated from MD simulations of HSA. Panel a shows the results obtained from the individual trajectories. The red lines indicate the minimum and maximum observed in the 100 trajectories, while the blue line depicts their average. Panel b presents the results obtained from concatenated trajectories with 50 elements.

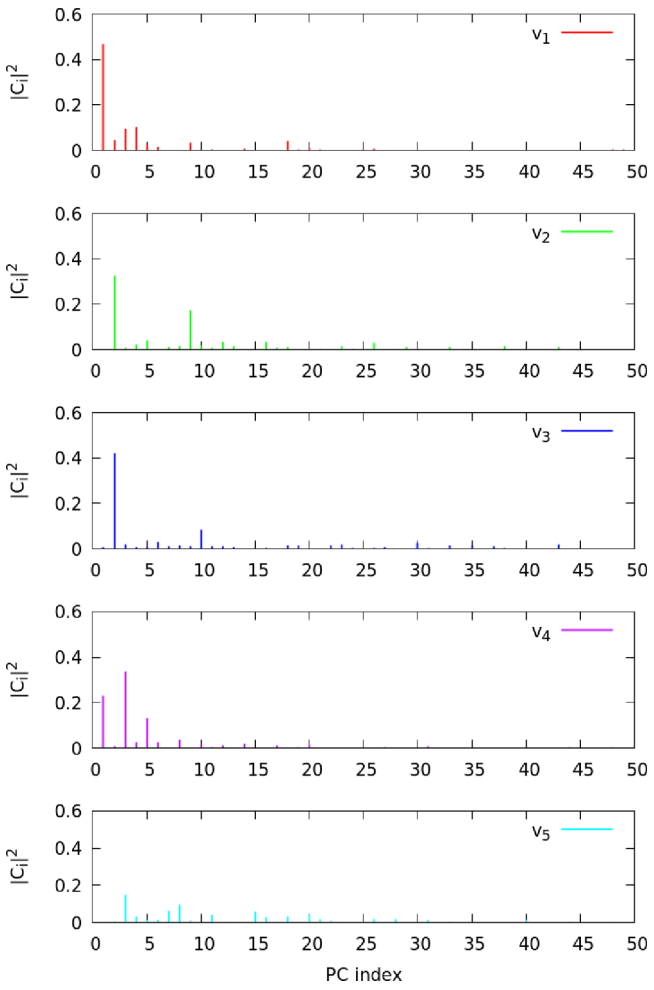


Figure 9. Squared components of selected PCA eigenvectors computed from a single trajectory of HSA. The components were determined by projecting the eigenvectors of the individual trajectory onto the basis set of the eigenvectors of the concatenated trajectory with 100 elements. To facilitate the comparison, all plots employ the same scale on the *y* axis. Only the first 50 components are shown. The total amount is 1728.

simple multidimensional PESs. We built the D model as a 100-dimensional degenerate harmonic well and the S model as a 100-dimensional harmonic well with all different, equally spaced harmonic constants. The movement of the walkers was determined with the Langevin equation in the diffusion limit. The values of the harmonic constants and the simulation times were selected so that the walkers experience the curvature of the PES during a single trajectory, but the trajectories are far from ergodic. Nonetheless, the average displacement observed in a large enough ensemble of walkers approximates the expected value at thermodynamic equilibrium.

The results of the D model demonstrate that there are serious difficulties in defining the essential space from the results of a single MD trajectory if the underlying PES has many directions with the same or similar constants. In this case, the PCA algorithm creates the illusion of a few directions where large fluctuations are allowed, while other equally important motions appear as completely irrelevant. Thus, if one builds the essential space from the first eigenvectors until they add up a given amount of the total fluctuations, important directions of movement can be left aside. This is a serious drawback that, to the best of our knowledge, has not been discussed before. The S model performs better than the D model, as the first eigenvectors are linear combinations of the directions of movement with the smaller harmonic constants. However, the coefficients of these combinations are also random, to a good extent. This finding is consistent with the results of Hess, who ran 100 simulations using a harmonic model with 30 degrees of freedom, which was slightly more complex than the S model examined here. Hess discovered that, for simulations that were not too short and not extremely long, the first PCA eigenvectors were a combination of the original coordinates with the lowest harmonic constants.[29]

Additionally, we found that individual trajectories corresponding to the simple models typically afford non-Gaussian probability distributions for their first PCs, despite the fact that the underlying PES is harmonic. These non-Gaussian shapes are, therefore, an artifact of the algorithm when applied to a single nonergodic trajectory. For the two models, all artifacts can be gradually attenuated by including more and more individual trajectories in a concatenated one, which is then subjected to a PCA. Eventually, the procedure affords eigenvectors with the correct directions and eigenvalues.

https://doi.org/10.1021/acsomega.4c01515
ACS Omega 2024, 9, 20488−20501

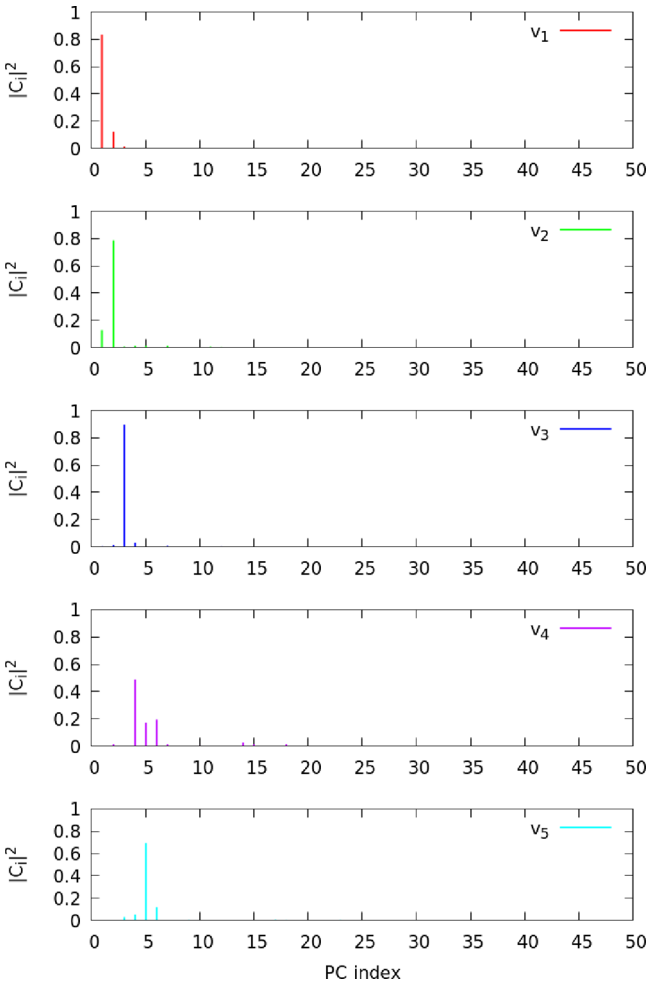


Figure 10. Squared components of selected PCA eigenvectors computed from a concatenated trajectory of HSA with 50 elements. The components were determined by projecting those eigenvectors onto the basis set of eigenvectors obtained from the alternative concatenated trajectory with 50 elements. Only the first 50 components are shown. The total amount is 1728.

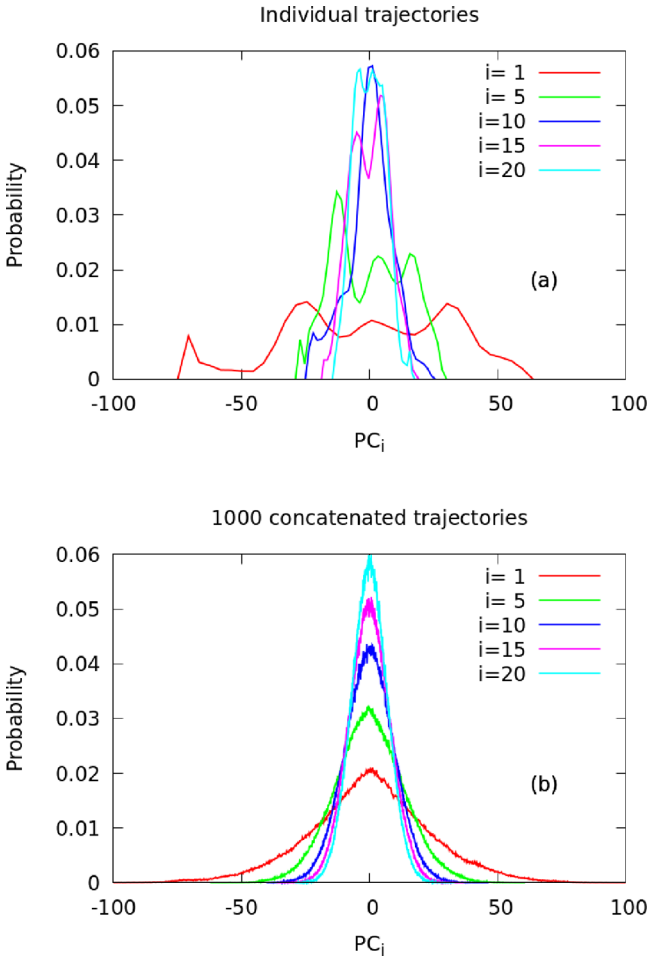


Figure 11. Probability distribution functions for selected principal components computed from the trajectories of the S model. Panel a depicts the results obtained with one individual trajectory. Panel b shows the distributions obtained with a concatenated trajectory formed with 1000 elements.

Furthermore, the probability distributions of their PCs become almost perfect Gaussians with the correct standard deviations.

We tested if the observations made on the simple models also apply to realistic protein models. To that aim, we performed MD simulations of HSA and lysozyme and analyzed them with analogous procedures. We found the behavior of both proteins very much resembles that of the simple models. In particular, if a large enough number of individual simulations is employed to build a concatenated trajectory, the PCA algorithm affords consistent eigenvectors and eigenvalues, a result we have already presented elsewhere.[28] However, in this case, we also analyzed what occurs with the probability distributions of the first PCs, an aspect that has not been discussed before. For HSA, we found that, while the distributions computed from individual trajectories have non-Gaussian shapes, those obtained from large enough concatenated trajectories are almost perfect Gaussians. The situation is different for lysozyme, an enzyme characterized for having a large domain motion that opens/closes its catalytic cleft. In this case, while most of the PC distributions acquire a Gaussian shape by increasing the number of individual trajectories in the concatenated one, the distribution of the first mode remains multimodal and that of the second mode slightly deviates from Gaussian.

A critical examination of the results of the simple models, and in particular of the D model, allowed determining that the problem with the PCA applied to single trajectories stems from the fact they are plagued with spurious correlations. This conclusion is implicit in the article Hess published more than 20 years ago[3] where he showed that the PCA eigenvalue spectra of multidimensional random walks occurring on flat potentials are the typical "scree plots". Hess's result was recently rederived by Antognini and Sohl-Dickstein using a completely different formalism.[10] The interesting point about this newer approach is that it could be extended to random walks on degenerate harmonic potentials. The analysis showed that, for those cases, too, the eigenvalue spectra are scree plots, so that a few of the first PCA eigenvectors account for the vast majority of the fluctuations observed in the trajectory. However, since there is nothing in the underlying PES that causes the observed correlations, the inexorable conclusion is that they must be random. After recognizing that the trajectories of multidimensional random walkers present random/fortuitous correlations, even for bound systems, the problem of analyzing them via PCA becomes clear. PCA is a linear transformation designed to maximize the fluctuations of

https://doi.org/10.1021/acsomega.4c01515
ACS Omega 2024, 9, 20488−20501

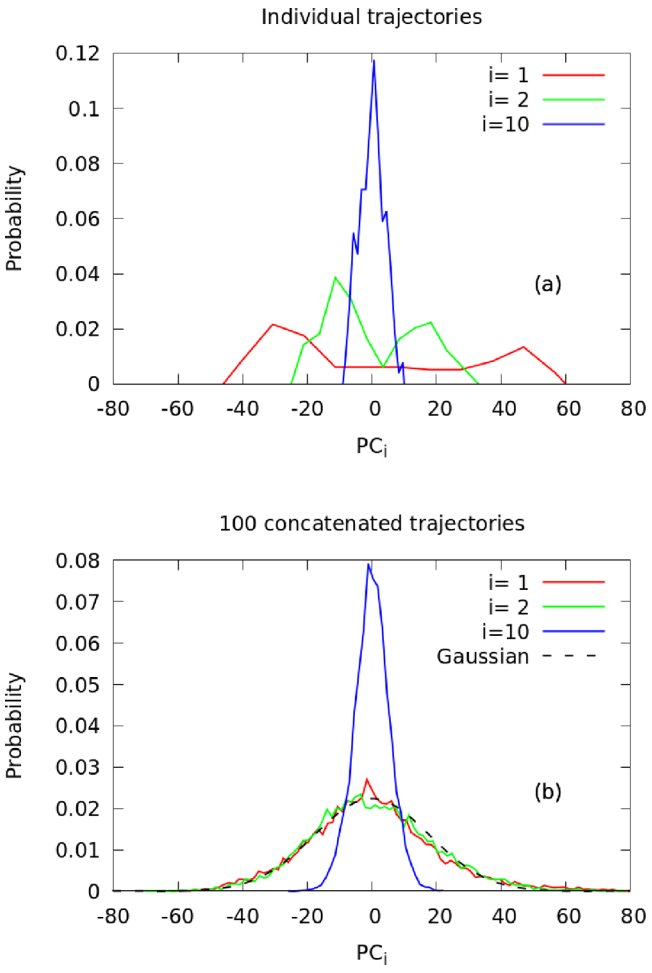


Figure 12. Probability distribution functions for selected principal components computed from the MD simulations of HSA. Panel a depicts the results obtained with one individual trajectory. Panel b shows the distributions obtained with a concatenated trajectory formed with 100 elements.

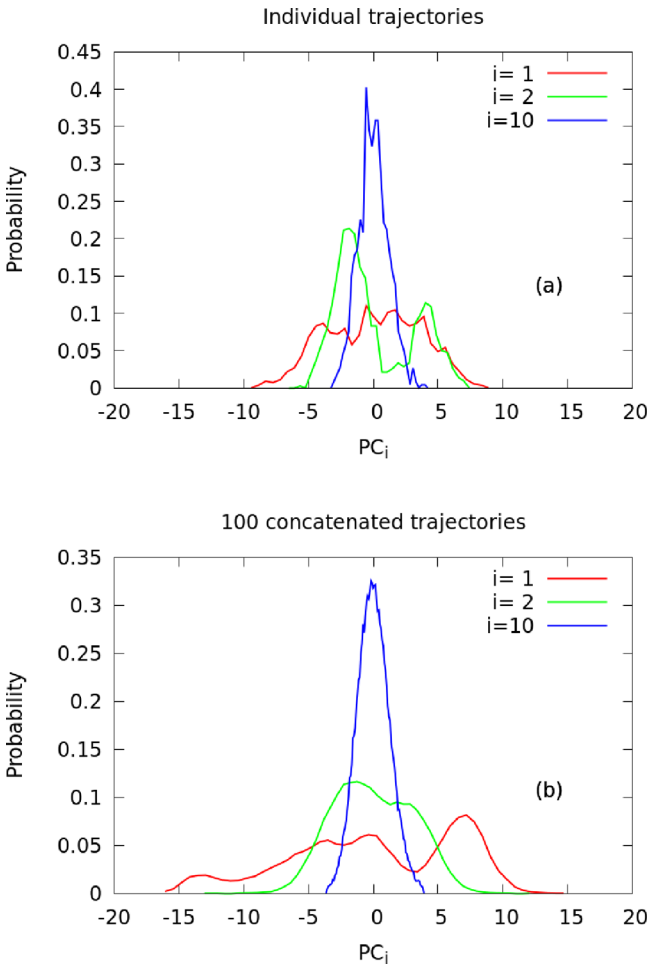


Figure 13. Probability distribution functions for selected principal components computed from the MD simulations of lysozyme. Panel a depicts the results of an individual trajectory. Panel b shows the distributions from a concatenated trajectory formed with 100 elements.

the new variables. Thus, it detects all correlations in the original set, without discriminating if they have a physical meaning or are fortuitous. Then, it combines the original coordinates to produce the directions of larger fluctuations.

It is not possible to eliminate spurious correlations from single trajectories of multidimensional random walks occurring on flat potentials. However, for bound systems such as the computational models of biological macromolecules, the situation differs. In those cases, the available configurational space has a finite volume. Accordingly, the structures sampled from a long enough simulation eventually fill that volume and the simulation becomes ergodic. Under such conditions, the density of samples within any small volume element of configurational space occurs in proportion to the equilibrium probabilities. Therefore, the underlying PES determines the values of the covariance matrix elements, and all correlations detected by PCA are actual correlations between the molecule's constituents. The problem is that the time demanded to reach this condition is extremely long. In particular, it is much longer than the total time multiple trajectories require to fill the same volume. This conclusion is well-known by MD practitioners.[30−32] Applying the PCA algorithm to a concatenated trajectory is equivalent to carrying out an ensemble average on top of the time average used to obtain the covariance matrix from a single simulation. For this reason, it affords much better converged results for the same total simulation time.

There has been a lot of discussion around the convergence and consistency of the eigenvalues and eigenvectors obtained from the PCA of MD simulations of biomolecules.[28,29,33,34] However, not much attention has been given to how it impacts the probability distributions of the principal components. This is surprising because these distributions provide valuable insights into the thermodynamic properties of the molecular system. An important conclusion of our study is that the shapes of these curves are likely to be wrong if the PCA is carried out with a unique MD trajectory. Typically, they appear structured and multimodal despite the fact that they should be simple Gaussians according to the underlying PES. Regardless of their actual shapes, we have shown that concatenation washes out the false correlations present in individual trajectories. As a result, the final distributions faithfully depict the form of the underlying PES, allowing an accurate evaluation of the molecule's free energy landscape. Determining the correct distributions is also crucial for the calculation of the configurational entropy. Cases such as human serum albumin, with fully harmonic PC distributions, support the use of the quasi-harmonic approximation (QHA).[35] But even for

https://doi.org/10.1021/acsomega.4c01515
ACS Omega 2024, 9, 20488−20501

more complicated systems, with one or a few anharmonic modes such as lysozyme, the entropy could be accurately determined by applying QHA to the harmonic modes and computing the required integrals for the few anharmonic ones.

CONCLUSIONS

By employing simple models, we demonstrated that PCA can afford fallacious results when applied to a single trajectory. The directions of the eigenvectors can be ill-defined, and their eigenvalues bear no relation with the actual amplitudes of the underlying modes. We noted that these drawbacks can lead to a poor definition of the essential space. Additionally, we showed that they lead to incorrect probability distribution functions for the main principal components, which then translates into misleading free energy landscapes and inaccurate configurational entropies.

We argued that these limitations stem from fortuitous correlations that typically show up in trajectories of independent random variables, even for bound systems. PCA maximizes the squared fluctuations of the new collective variables by combining the original coordinates with the highest correlations. However, PCA cannot discriminate whether those correlations are spurious or have a physical meaning. In addition, we illustrated how all of these problems are mitigated and, eventually, eliminated by doing PCA on a concatenated trajectory built by combining several individual simulations.

We also analyzed via PCA the simulations of two realistic protein models. We found that they behave very much like the simple harmonic models. From this observation, we infer that the lessons learned from the harmonic models also apply to MD simulations of other globular proteins. It remains to be tested to what extent they are also valid for more flexible biomolecules such as RNA or intrinsically disordered proteins.

To conclude, we note that many procedures used in the analysis of MD simulations of biomolecules employ PCA in some of their intermediate stages. The examples and discussions presented in this Article suggest that they would also be affected by the presence of fortuitous correlations between the many degrees of freedom considered in the analysis. We did not prove this hypothesis, but it seems the most plausible. Thus, we believe that the PCA algorithm should always be applied to concatenated trajectories, as this procedure eliminates fortuitous correlations revealing those with a physical origin.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c01515.

Figures S1−S10 (PDF)

All the input files, structures, scripts, and Fortran codes required to replicate this article's results (ZIP)

Movie S1: Animation of the first PCA eigenvector of an individual trajectory of 10 ns of HSA (MPG)

Movie S2: Animation of the first PCA eigenvector of a concatenated trajectory formed by 100 independent trajectories of 10 ns of HSA (MPG)

Movie S3: Animation of the first PCA eigenvector of an individual trajectory of 10 ns of lysozyme (MPG)

Movie S4: Animation of the first PCA eigenvector of a concatenated trajectory of 100 independent 10 ns-trajectories of lysozyme (MPG)

AUTHOR INFORMATION

Corresponding Author

Juliana Palma − *Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Buenos Aires B1876BXD, Argentina; Consejo Nacional de Investigaciones Científicas y Técnicas, Ciudad Autónoma de Buenos Aires C1425FQB, Argentina;* orcid.org/0000-0002-0761-0253; Email: juliana@unq.edu.ar

Author

Gustavo Pierdominici-Sottile − *Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Buenos Aires B1876BXD, Argentina; Consejo Nacional de Investigaciones Científicas y Técnicas, Ciudad Autónoma de Buenos Aires C1425FQB, Argentina;* orcid.org/0000-0003-3792-4509

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c01515

Funding

This research was supported by Universidad Nacional de Quilmes (UNQ; ID:1292/19), ANPCyT (PICT 2020-SerieA-00192), and CONICET (ID:11220130100260CO).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank CONICET, UNQ, and ANPCyT for their financial support. Both authors also want to acknowledge their parents (Gustavo, Angela, Marcos, and Elsa) for their unconditional support.

REFERENCES

(1) García, A. E. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* **1992**, *68*, 2696−2699.

(2) Amadei, A.; Linssen, A. B.; Berendsen, H. J. Essential dynamics of proteins. *Proteins: Struct. Funct. Genet.* **1993**, *17*, 412−425.

(3) Hess, B. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E* **2000**, *62*, 8438−8448.

(4) Witte, R. S.; Witte, J. S. *Statistics*; Wiley, 1992.

(5) Yule, G. U. Why do we sometimes get nonsense-correlations between Time-Series?−a study in sampling and the nature of time-series. *Journal of the royal statistical society* **1926**, *89*, 1−63.

(6) Granger, C.; Newbold, P. Spurious regressions in econometrics. *Journal of Econometrics* **1974**, *2*, 111−120.

(7) Greene, W. H. *Econometric Analysis*; Pearson, 2017.

(8) Newbold, P.; Carlson, W. L.; Thorne, B. *Statistics for Business and Economics*; Pearson, 2018.

(9) Hassler, U.; Hosseinkouchack, M. Understanding nonsense correlation between (independent) random walks in finite samples. *Statistical Papers* **2022**, *63*, 181−195.

(10) Antognini, J.; Sohl-Dickstein, J. PCA of high dimensional random walks with comparison to neural network training. *Advances in Neural Information Processing Systems*; MIT Press, 2018; vol *31*.

(11) Gray, R. *Toeplitz and Circulant Matrices: A Review; Foundations and Trends in Technology*; Now Publishers, 2006.

(12) Sugio, S.; Kashima, A.; Mochizuki, S.; Noda, M.; Kobayashi, K. Crystal structure of human serum albumin at 2.5 Å resolution. *Protein Engineering, Design and Selection* **1999**, *12*, 439−446.

(13) Muraki, M.; Harata, K.; Sugita, N.; Sato, K.-i. Origin of Carbohydrate Recognition Specificity of Human Lysozyme Revealed by Affinity Labeling. *Biochemistry* **1996**, *35*, 13562−13567.

(14) Case, D. et al. *Amber 18*; University of California: San Francisco, 2018.
(15) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(16) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(17) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112*, 9020−9041.
(18) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*; Academic Press, 2002.
(19) Miyamoto, S.; Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952−962.
(20) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(21) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(22) Brooks, B.; Karplus, M. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc. Natl. Acad. Sci. U. S. A.* **1985**, *82*, 4995−4999.
(23) Hayward, S.; Berendsen, H. J. Systematic analysis of domain motions in proteins from conformational change: New results on citrate synthase and T4 lysozyme. *Proteins: Struct., Funct., Bioinf.* **1998**, *30*, 144−154.
(24) de Groot, B.; Hayward, S.; van Aalten, D.; Amadei, A.; Berendsen, H. Domain motions in bacteriophage T4 lysozyme: A comparison between molecular dynamics and crystallographic data. *Proteins: Struct., Funct., Bioinf.* **1998**, *31*, 116−127.
(25) Hub, J. S.; de Groot, B. L. Detection of Functional Modes in Protein Dynamics. *PLoS Comput. Biol.* **2009**, *5*, e1000480.
(26) Palma, J.; Pierdominici-Sottile, G. On the Uses of PCA to Characterise Molecular Dynamics Simulations of Biological Macromolecules: Basics and Tips for an Effective Use. *ChemPhysChem* **2023**, *24*, No. e202200491.
(27) Pierdominici-Sottile, G.; Palma, J. New insights into the meaning and usefulness of principal component analysis of concatenated trajectories. *J. Comput. Chem.* **2015**, *36*, 424−432.
(28) Cossio-Pérez, R.; Palma, J.; Pierdominici-Sottile, G. Consistent Principal Component Modes from Molecular Dynamics Simulations of Proteins. *J. Chem. Inf. Model.* **2017**, *57*, 826−834.
(29) Hess, B. Convergence of Sampling in Protein Simulations. *Phys. Rev. E* **2002**, *65*, No. 031910.
(30) Pranami, G.; Lamm, M. H. Estimating Error in Diffusion Coefficients Derived from Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2015**, *11*, 4586−4592.
(31) Perez, J. J.; Tomas, M. S.; Rubio-Martinez, J. Assessment of the Sampling Performance of Multiple-Copy Dynamics versus a Unique Trajectory. *J. Chem. Inf. Model.* **2016**, *56*, 1950−1962.
(32) Yan, S.; Peck, J. M.; Ilgu, M.; Nilsen-Hamilton, M.; Lamm, M. H. Sampling Performance of Multiple Independent Molecular Dynamics Simulations of an RNA Aptamer. *ACS Omega* **2020**, *5*, 20187−20201.
(33) Amadei, A.; Ceruso, M. A.; Di Nola, A. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins: Struct. Funct. Genet.* **1999**, *36*, 419−424.
(34) Grossfield, A.; Zuckerman, D. In *Annual Reports in Computational Chemistry*; Wheeler, R., Ed.; Elsevier Ltd., 2009; pp 23−48.
(35) Karplus, M.; Kushick, J. N. Method for estimating the configurational entropy of macromolecules. *Macromolecules* **1981**, *14*, 325−332.

https://doi.org/10.1021/acsomega.4c01515
ACS Omega 2024, 9, 20488−20501