Physical Chemistry Chemical Physics
RSC Publishing
Journals Physical Chemistry Chemical Physics Advance Articles DOI: 10.1039/b913802a

Phys. Chem. Chem. Phys., 2009 | DOI: 10.1039/b913802a | Paper

Principal component analysis of potential energy surfaces of large clusters: allowing the practical calculation of the master equation†

Nima Shariat Panahi and R. S. Berry *
The University of Chicago, 929 East 57th St., Chicago, Illinois 60637, USA. E-mail: berry@uchicago.edu

Received 10th July 2009 , Accepted 2nd October 2009

First published on the web 6th November 2009

The number of variables in many-particle systems is typically unmanageably large; some way to reduce that number and still retain access to the important information about the system of interest is one of the great challenges in the broad topic of complexity. Principal components and principal coordinates provide a powerful means to extract—from unwieldy, large data sets—a reduced collection of variables that provide the information one needs, in a relatively efficient way and useful form. We investigate the application of principal components to the analysis of kinetics of the atomic motions in atomic clusters, particularly of clusters that are large enough so that a full description in terms of the entire high-dimensional potential surface is entirely impractical. A specific application is the use of principal components linking minima with their adjacent saddles, permitting the evaluation of rate coefficients (in the context of transition state theory) as ratios of partition functions of only one or two key variables.

1 Introduction

1.1 PCO and PCA

The analysis of the potential energy surfaces (PES) of clusters has proven to be a useful tool in theoretical chemical physics,[1–4] especially for relating the interparticle forces to the dynamical and kinetic behavior of moderately complex systems. With current computational power and efficient algorithms, one can find all of the minima and the important saddles on a potential energy surface for systems composed of up to about 18 particles. However, the number of geometrically distinct minima grows at least exponentially with the number of particles in the system, and the number of saddles grows even faster than that.[5] And this does not yet take account of the number of permutational isomers. Consequently, cataloging all the minima and important saddles is, at very least, undesirable for 20 particles and more. Even if one were to do such an exhaustive search, most of the topographical information would be of negligible use for the analysis of dynamics.[3] It is thus more desirable—and, in effect, necessary—to reduce drastically the number of variables one uses to describe the system. It is clearly desirable to construct a *sample* PES from a statistical sampling of the full PES, but to do this in a way that yields a reasonably accurate representation of the dynamics or kinetics.[6] Such a statistical sample PES would, one hopes, be adequate, especially to reveal the most important, and, presumably, slow relaxation processes and their rates. In this paper, we explore one approach that may offer a means to achieve such a reduction, namely the use of principal component analysis (PCA) and its complement, principal coordinate analysis (PCO); specifically, we examine the use of these tools to evaluate the rate coefficients on a many-dimensional potential surface. We use these to avoid having to use the full set of coordinates to find the partition functions and rate coefficients for passages between local potential minima on complex surfaces. In the

end, we are able to estimate the minimum size of systems for which this approach would be a significant aid.

This investigation is an exploration of the possibility and potential difficulties in using PCA and PCO to reduce, in one particular way, the unmanageable complexity of the full master equation of a many-body system. Here, we do not address the sampling problem—the choice of criteria for constructing a suitable statistical sample of the full energy landscape. Nor do we address the related problem of identifying "pathological" landscapes, such as those of Ar_{38}, Ar_{75} or perhaps prions—landscapes with deep, narrow minima that have energies significantly lower than the readily accessible minima in large basins, and that typically have structures different from those in the large basins, the landscapes sometimes called "Stillinger nightmares", after that researcher's conjecture of the existence of such phenomena.

The methods of sampling pathways on complex potential surfaces have been studied extensively elsewhere.[6–8] Our focus here is primarily on making more efficient the evaluation of the many rate coefficients for well-to-well passages. In our study, we chose as our main vehicle an Ar_{13} cluster, approximated by pairwise Lennard-Jones potentials. We also examined Ar_{20} for some investigations, to get some sense of the difficulties of working with larger systems. For the 13-atom cluster, a size manageable for a relatively detailed study, we began with the database of 1505 geometrically distinct minima and 25 653 important saddles.[4,9,10] We selected 100 minima and the 1111 saddles connecting them as our sample PES. This was done by starting at the global minima and following the connectivity through saddles and branching out until we had reached our required 100 minima. We were careful to include the most connected minima and branch out from them to other well connected minima so that we obtained a good representation of the connectivity, in order to model the dynamics. In addition, we took care to include all transition states directly linking any two minima. The pathways were consistent with the "rough trajectory" criterion found previously to be the most successful guide for finding reasonably good estimates of the slowest rates of motion on the landscape.[6]

Once the sample PES has been obtained, then the rate constants for well-to-well passage of the argon cluster can be calculated from the topography of the sample PES. We then went on to calculate the master equation and other important features of the sample PES. Since principal component analysis (PCA) and principal coordinate analysis (PCO) offer ways to reduce the dimensionality of the problem in calculating rate constants and in visualizing the PES, respectively, PCA and PCO can, in principle, be used to calculate rate constants efficiently and portray the significant features of the PES for large clusters. Specifically, degrees of freedom that do not change significantly from the initial minimum to the saddle can be neglected in evaluating the rate constant for that transition, within the formalism of transition state theory, because the partition functions for such degrees of freedom are the same at both places on the potential surface.

Principal coordinate analysis was first developed by Gower in 1966.[11] Gower also showed the duality of principal coordinate analysis and principal component analysis. PCO is a variant or dual of PCA and both are means to select the most important out of the whole set of degrees of freedom.[11,12]

In our work, both PCO and PCA were used. Although there is a duality between the two methods, the two are suited for different tasks. For example, PCO is principally suited for visualizing purposes and does not provide analytic results.[12] On the other hand, PCA is not well suited for reducing the dimensionality of a PES representation but does provide analytic results for the reduced dimensions,[12] such as those needed for calculating rate constants. Here, we make quantitative use of PCA, and apply PCO primarily for building an intuitive picture of what the approximating approach is doing.

In this project, first, PCO is used to reduce the dimensionality of the PES of Ar_{20} in order to represent its essentials visually in three-dimensional space. Then, PCA is applied to the Ar_{13} cluster in order to reduce the vibrational analysis to a set of reduced dimensionality, in order to obtain rate constants for transitions between minima (using their connecting saddles) in the context of transition state theory. That way, when we obtain the rate constants, it becomes simple to construct and solve the master equation.

The crux of this approach is the way the rate coefficient for passage over a barrier is represented as the ratio of the partition functions of the transition state and the initial state. These partition functions are

products of the partition functions of the vibrational degrees of freedom. Any degree of freedom that is essentially the same in the initial and transition state contributes the same partition function to the numerator and denominator of the ratio, and hence, since these cancel, contributes only a factor of unity to the rate coefficient. Only those degrees of freedom that differ significantly in passage from initial to transition state contribute to the rate coefficient. These are precisely the degrees of freedom that PCO and PCA identify.

1.2 Master equation

The stochastic master equation[13] dynamics method has previously been applied to the study of the kinetics on a PES[13–16] Two general methods have been widely used to study near-equilibrium dynamics and thermodynamics, such as relaxations of argon clusters: stochastic master equations and molecular dynamics simulations (MD). The stochastic master equation does not reveal the atomic-level detail of MD, but that method has several advantages over MD. First, solving the master equation is much faster than solving the MD equations for the time intervals and time resolutions needed to reach equilibrium and achieve results comparable to master equation solutions. The time-consuming part of solving the master equation is constructing the transfer matrix (which contains the rate constants between minima) and diagonalizing the resulting $N \times N$ matrix (where N is the number of minima in the sample PES). This process yields the eigenvalues and eigenvectors that solve the equations.[14,16,17] The complexity of diagonalizing such a matrix is $O(N^3)$.

Second, by the method of master equation, we obtain average behavior of an ensemble, whereas MD simulations require many runs to obtain satisfactory averages.[4,18,19,20]

Third, aside from the limitations of diagonalizing possibly a large matrix (though in this study, the size is limited by choosing a sample PES that is as small as possible in order to describe the dynamics of the cluster and still retain the desired accuracy), the master equation method does not suffer the limitations of MD, such as large storage requirements, limited total simulation time, and establishing a time scale on which ergodicity is achieved in a simulation.

Fourth, the transfer matrix is a product of the theory, in this case Rice-Ramsperger-Kassel-Marcus (RRKM) transition state theory, used to describe the state-to-state kinetics,[21,22] and also it retains the characteristics of topography of the underlying PES, such as connectivity and energies of the stationary points. This allows us to study the effect of different underlying approaches to setting up and solving the master equation.[23]

Fifth, the master equation avoids the round-off error propagated in integrating the equations of motion by the MD algorithms. Finally, to obtain kinetic results for different temperatures, new ensembles of MD trajectories must be calculated, whereas for the master equation, the transition matrix can just be modified in a fast and simple manner to account for the temperature dependences of the rate coefficients.

Hence, this is an exploration into a possible way to make manageable the description of the motions of moderately complex systems by constructing and solving master equations based on suitably chosen samples of their energy landscapes, as an alternative to using more detailed methods such as molecular dynamics. It is not in any way a study of how to construct the statistical sample, only of examining a way that may be useful for evaluating the rate coefficients on that sample landscape.

For the Ar_{13} system, the transition matrix was calculated for a temperature equivalent in Lennard-Jones units to a conventional temperature of 30 K in this work. This temperature was an arbitrary choice that falls in the range of solid–liquid coexistence for this system. Other temperatures would certainly involve different rates and different equilibrium population distributions. A primary motivation for this effort is to try to find a way to speed up and simplify the calculation of the rate constants that comprise the transition matrix of the master equation. The rate coefficients that are the elements of that matrix are, within the context of transition state theory, ratios of partition functions, at saddles and at the initial states leading to those saddles. The goal of this effort was to see whether reducing the number of variables needed to compute the partition coefficients to just those that carry significant changes, *i.e.* the most important PCAs and PCOs, would simplify the construction of the master equation to a useful extent.

2 Theory

2.1 Calculating PCO and PCA

PCA is the first step in PCO, as both are duals of each other. PCA uses a correlation matrix formed from comparison between variables that set apart one data point from the others. Here, the coordinates are used as variables and different configurations as individual data points.

We start with an $n \times p$ matrix, Y, with n observables and p variables for each observable; here, the two methods diverge. For PCA, we start with the $p \times p$ matrix Y^TY, measuring the variance between variables. In contrast, PCO uses the $n \times n$ matrix YY^T, measuring the similarity between configurations. Both PCA and PCO together are the tools of a general method, far more than just our specialized application and exhibit great variety, not only in the fields in which they are applied, but also in how they are mathematically adapted to the need. For example, the distance matrix in PCO can be defined in many ways to obtain what one considers the correct measure of distance (or dissimilarity) for the application. Likewise, PCA's covariance matrix can be replaced with similar matrices, such as correlation or sums-of-squares-cross-products (SSCP),[24] depending on the application.

The first step in our use of PCO is setting up the distance matrix between configurations. The matrix is defined by $d_{ij} = \sum_{r=1}^{p} (X_{ir} - X_{jr})^2$, where i and j are the configuration indices, r is the coordinate index, X are Cartesian coordinates, and d_{ij} is the measure of dissimilarity between the ith and jth configurations. This describes the "distance" between configurations and is thus suitable for constructing the PES graph.

The second step is constructing a *centralized* distance matrix as follows:

1. Form the A matrix, $-d_{ij}/2$, of interparticle distances.

2. Centralize the A matrix and, from it, form the B matrix, $b_{ij} = a_{ij} - a_{i.} - a_{j.} + a_{..}$, where $a_{i.}$ is the average over the ith row, $a_{j.}$ is the average over the jth column, and $a_{..}$ is the average over the whole matrix A. We do this to remove averages from the calculations and results.

3. Diagonalize the B matrix and obtain the eigenvalues and eigenvectors.

4. Normalize the eigenvectors to eigenvalues. That is, normalize the eigenvectors so that their norms are the corresponding eigenvalues. This is done by dividing the eigenvectors by the square root of the corresponding eigenvalue.

5. The normalized eigenvalues give the percent of the total variance between the structures contained in the corresponding eigenvector.

We then proceed to pick the first few eigenvectors (or PCOs) with the largest eigenvalues which reveal that they are the ones with the largest variances contained in them, in order to derive the new data set. To construct a 3D graph, we pick the first two eigenvectors and plot the energy as a function of these variables, to get a picture of the main features of the PES.

The procedure for PCA is similar to that of PCO:[12,25]

1. One first subtracts the mean from the data, similar to centralizing our data for the PCO. That is, we subtract the average across each dimension from each datum in that dimension.

2. We calculate the covariance matrix (see ref. 25 and 26 as needed).

3. We then diagonalize the covariance matrix and obtain the eigenvectors and eigenvalues.

4. The eigenvectors give a linear combination of the data set. We have done this in two ways. In our first case, we chose the interatomic distances as the variables and configurations as the data points. In our second case, the $3N$ atomic coordinates are the variables and the configurations are the data points.

5. The normalized eigenvalues give the percentage variance contained in the corresponding eigenvectors.

6. Last, we pick out the most important variables, corresponding to the eigenvectors with the highest eigenvalues, using the linear combination as a weight.

In the central effort of this study, only one eigenvector was picked for the case based on interatomic distances, and three eigenvectors were picked for the atomic coordinates case. The calculations were also

done with more eigenvectors or PCAs, but this did not change the results, as the mathematics dictates exactly how many PCA coordinates we need (*i.e.* the 1-dimensional nature of the interatomic case requires one PCA while the 3-dimensional nature of the atomic coordinates requires three PCAs). This shows that the first few PCAs effectively contain all of the important variance in our data. In the linear combination, we proceeded to pick the most important interatomic distances by looking at their weights.

This reduction, in turn, told us what coordinates needed to be included in the computation of the reduced Hessian from which we obtained the frequencies required to compute the rate constants of the master equation. Finally, for the PCO and PCA methods, since we only needed the first few largest eigenvalues/eigenvectors, we used Lanczos methods[27] to cut down on computational cost. The computational cost saved by Lanczos method is significant, especially for large systems for which neither the computation time nor resources exist to deal with diagonalizing very large matrices. Our own tests showed that for very large systems, the Lanczos method can be orders of magnitude faster than a full diagonalization.

2.2 Constructing the master equation

The stochastic master equation formalism is an initial value problem whose solutions are time-dependent occupation probabilities for the minima in our sample PES.[13] The main component of the master equation is its transition matrix. The eigenvalues and eigenvectors of the transition matrix solve the master equation. The eigenvalues give the exponential rates of flow of the population distributions on the PES, and the eigenvectors describe these flows in terms of changes in the populations among the minima connected to the particular minimum. That is, the absolute value of the *j*th component of the *i*th eigenvector will determine the magnitude of the rate by which the mode *j* will change the population at state *i* and its sign will determine if the population will increase or decrease.

The first step in constructing the master equation for our system is obtaining the partition functions used in calculating the rate constants of the RRKM transition state theory. Realistic partition functions, including anharmonic corrections, have been studied in detail elsewhere.[23] For our purposes, it suffices for us to use a classical harmonic model to obtain the vibrational partition function:

$$Z^{\text{vib}} = \prod_{j=1}^{m} \frac{1}{\beta h \nu_j}$$

where m is the number of vibrational degrees of freedom in each type of configuration, with $m = 3N - 6 = 33$ for minima and $m = 3N - 7 = 32$ for transition states, $\beta = \frac{1}{k_B T}$ for k_B, the Boltzmann constant, and T is the temperature (30 K in these calculations), and ν_j is the jth normal mode vibrational frequency. The frequencies, ν_j, are obtained by diagonalizing the Hessian of a given stationary point. It is here that we use PCA to dramatically reduce the dimension of the Hessian matrix and thus speed up obtaining the frequencies. The rationale behind this is that the important quantities, in our case the rate constants and equilibrium solutions to the master equation, use the ratio of vibrational partition functions of the minimum and the saddle. This means that the partition function factors of similar normal modes effectively cancel out, leaving us with only those normal modes that change between the saddle and the minimum. We can then construct the partition function[23]

$$Z_i = n_s Z^{\text{vib}} \exp(-\beta V_i)$$

where V_i is the potential energy at the stationary point. The degeneracy factor, n_s, accounts for the number of distinct permutational isomers, and is given by

$$n_s = \frac{2n_p}{h_s}$$

where n_p is the total number of nuclear permutations, and h_s is the order of the point group for that configuration.[28]

Transition matrix and master equation. The transition probabilities, W_{ij}, which make up the elements of the transition matrix, **W**, for passage from well j to well i ($i \neq j$) is the sum of the RRKM transition rates for each of the $N^{\dagger}_{ij}$ transition states, l, connecting the wells:

$$W_{ij} \equiv k^{j\to i}(\beta) = \sum_{l=1}^{N_{ij}^{\dagger}} k_l^{j\to i}(\beta) = \frac{1}{\beta h}\sum_{l=1}^{N_{ij}^{\dagger}} \frac{Z_l^{\dagger}}{Z_j}\exp(-\beta\Delta V_l)$$

where $\Delta V_l = V_l - V_j$ is the barrier height of the transition (V_l and V_j are the potential energies of the transition state l and minimum j, respectively). Z_j and Z_l are the partition functions for the local minimum j and the transition state l, respectively. Note that the above equation for W_{ij} is only used to obtain off-diagonal terms in the transition matrix. So, in order to write the master equation in matrix form, we need to define the diagonal terms. The diagonal terms are the combined rates for all transitions out of well i into wells connected to it:

$$W_{ii} = -\sum_{j\neq i} W_{ji}$$

Now we construct the master equation for the time-dependent probability vector, $\mathbf{P}(t)$ [with N (the number of minima in the sample PES) components with values $P_i(t)$, the probability of the system residing in well i at time t]; writing it in component form:

$$\frac{\mathrm{d}P_i(t)}{\mathrm{d}t} = \sum_{j\neq i}[W_{ij}P_j(t) - W_{ji}P_i(t)]$$

or writing it in matrix form:

$$\mathbf{P}(t) = \mathbf{W}\mathbf{P}(t).$$

The transition matrix, and thus the master equation, do not contain degenerate contributions; that is, transitions to the same well or permutational isomers do not affect the ensemble population of that geometry, and therefore are not counted.

The equilibrium, *i.e.* infinite time, solutions to the master equation are quite simple and are given by the Boltzmann distribution[23]

$$P_i^{\mathrm{eq}} = \frac{Z_i \exp(-\beta V_i)}{\sum_{j=1}^{N} Z_j \exp(-\beta V_j)}.$$

Again, Z_i is the partition function for the minimum i, and V_i is the potential energy at that minimum.

Solutions to the master equation. In our calculations, we used a Householder reduction to obtain a QR‡ decomposition and thus diagonalize the transition matrix **W**.[29,30] To use this, and to ensure a spanning set of eigenvectors with real eigenvalues, we need to symmetrize the transition matrix. We can obtain this by evoking the condition of detailed balance:

$$W_{ij}\sqrt{\frac{P_j^{\mathrm{eq}}}{P_i^{\mathrm{eq}}}} = W_{ji}\sqrt{\frac{P_i^{\mathrm{eq}}}{P_j^{\mathrm{eq}}}}$$

We therefore form a new symmetric matrix, $\tilde{\mathbf{W}}$:

$$\tilde{W}_{ij} = W_{ij}\sqrt{\frac{P_j^{eq}}{P_i^{eq}}};$$

$\tilde{\mathbf{W}}$ and $\mathbf{W}$ have the same eigenvalues, λ_i. Their eigenvectors, $\tilde{\mathbf{u}}^{(i)}$ and $\mathbf{u}^{(i)}$, respectively, are related by $\mathbf{u}^{(i)} = \sqrt{P_i^{eq}}\tilde{\mathbf{u}}^{(i)}$. The final solution to the master equation in terms of the eigenvectors of $\tilde{\mathbf{W}}$ is:

$$P_i(t) = \sqrt{P_i^{eq}}\sum_{j=1}^{N}\tilde{u}_i^{(j)}e^{\lambda_j t}\left[\sum_{m=1}^{N}\tilde{u}_m^{(j)}\frac{P_m(0)}{\sqrt{P_m^{eq}}}\right],$$

or in terms of the eigenvectors of $\mathbf{W}$:

$$P_i(t) = \sum_{j=1}^{N}u_i^{(j)}e^{\lambda_j t}\left[\sum_{m=1}^{N}u_m^{(j)}\frac{P_m(0)}{P_m^{eq}}\right].$$

In our work, we look at and compare the eigenvalue spectrum and ultimately the relaxation curves $P_i(t)$, which are the solutions to the master equation, to judge the validity of our method. However, also using eigenvector similarities have been studied elsewhere by Lu *et al*.[6]

3 Results

First, we examine the way PCO can give insights into the topography of a complex energy landscape. For this, we used the Ar_{20} cluster, one whose potential is far too complex to be envisioned in any relatively full way. We chose 75 points on the PES of Ar_{20} using the program OPTIM2 by David Wales, with the Lennard-Jones potential. Since we were interested in seeing how to apply PCO, and were not concerned here with finding good methods to choose the PES points to which we would apply PCO, we simply picked our 75 by starting from the global minimum and the second lowest minimum we could find and branched out from them by constructing monotonic sequences, but in random directions. We also added a few points just randomly from general non-monotonic sequences originating from the global minimum point to complete the picture. We then performed PCO with Cartesian coordinates and picked the most important two eigenvectors to obtain the graph of the PES (Fig. 1).

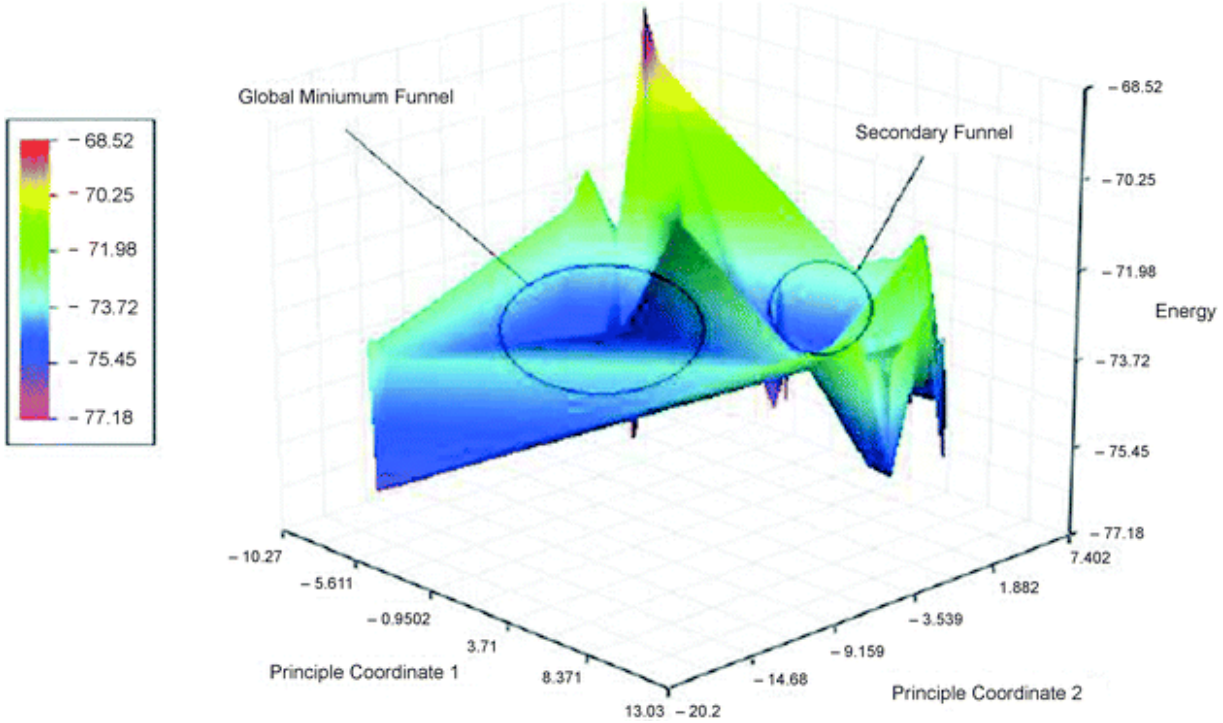

Fig. 1 Principal coordinate representation of the PES of Ar_{20} based on the two most important principal coordinates derived from the Cartesian coordinates of the particles.

We can see that the PES of Ar_{20} has two main funnels, at least in the portion we sampled. Thus, with only two PCO eigenvectors, we were able to get a good picture of the complex PES of Ar_{20}. Because almost all the variance between these two stable forms is encapsulated in the two principal coordinates, they are adequate for estimating the effective distance between them.

Then we turned to the more quantitative problem of finding whether PCA can be used to reduce the dimensionality of the problem of finding rate coefficients. For Ar_{13}, we applied PCA. First, we formed the data with interatomic distances as the variables ($N(N-1)/2$ of them) and stationary points as configurations. We then performed PCA and picked the first eigenvector. Out of the first eigenvector, we picked the highest weights corresponding to the most important interatomic distances. For all rate constants, the first eigenvector effectively represented all the variance of the data. With those data, we then proceeded to do harmonic vibrational analysis using the Hessian matrix on the minima and saddles. Finally, we constructed and solved the master equation to get the eigenvalues. We repeated the calculation a second time with the modification of using atomic coordinates as the variables ($3N$ of them) and stationary points as configurations and taking the three largest eigenvectors instead of one. We compared the eigenvalue spectra obtained from a full Hessian and from our two-PCA reduced Hessian (Fig. 2 and 3).

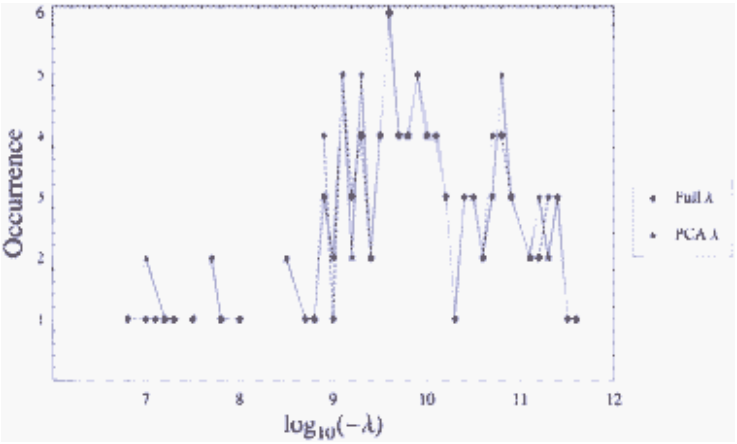

Fig. 2 The eigenvalue spectra of the full master equation compared with PCA using interatomic distances.

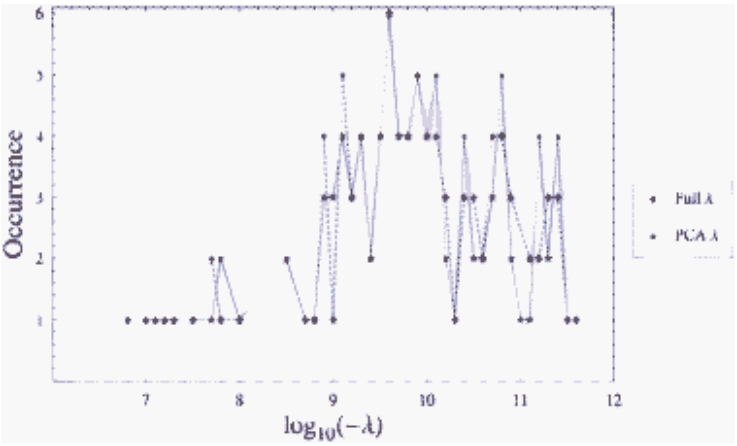

Fig. 3 The eigenvalue spectra of the full master equation compared with PCA using atomic coordinates.

As we can see from the spectral graphs, the PCAs give results for the eigenvalue spectrum nearly identical with those of the full treatment. However, the relaxation curves are an even better match than the spectra, as discussed below. Also, the PCA, using two principal components for interatomic distances, gave exactly the same spectrum as that from PCA with one component since, as stated above, the second PCA adds nothing to the variance (that is, the first PCA component represented 100% of the motions with the remaining components adding nothing, due to the fact that changes in the cluster are completely captured by the change in the bonds, which are 1-dimensional and thus the largest covariance eigenvector contains all the variance). A similar thing happened with more than three principal components based on the atomic coordinates for the reasons states above. For this case of three principal components, the variance divided between the three components varied from case to case, but the variance divided somewhat equally on average among the three; each contained about 1/3 of the variance. This behavior of the coordinate-based PCAs meant that in this representation, one must keep three components, since eliminating any of them would mean losing roughly 1/3 of the variance. Adding more than three components increased computations dramatically but added nothing. However, these three components gave us the same PCA spectrum as that based on interatomic distances, but with a much improved speed in computation compared to the interatomic method. Finally, it is very important to note that the similarity between the full master equation and one based on PCA reduced coordinates is especially strong for the highest (least negative) eigenvalues, which dominate the relaxation process. We see the numerical comparison for eigenvalues listed from highest to lowest (log nearest zero and hence slowest process

first), and thus with the important eigenvalues leading the list, in Table 1.

Table 1 Spectra of the $\log_{10}(-\lambda)$ of the eigenvalues, λ, and their comparison between full and PCA molecule

Full Hessian		PCA interatomic		PCA coordinates	
$\log_{10}(-\lambda)$	Occurrence	$\log_{10}(-\lambda)$	Occurrence	$\log_{10}(-\lambda)$	Occurrence
6.8	1	6.8	1	6.8	1
7.0	1	7.0	2	7.0	1
7.1	1	—	—	7.1	1
7.2	1	7.2	1	7.2	1
7.3	1	7.3	1	7.3	1
7.5	1	7.5	1	7.5	1
7.7	2	7.7	2	7.7	1
7.8	1	7.8	1	7.8	2
8.0	1	8.0	1	8.0	1
8.5	2	8.5	2	8.5	2
8.7	1	8.7	1	8.7	1
8.8	1	8.8	1	8.8	1
8.9	4	8.9	3	8.9	3
9.0	1	9.0	2	9.0	3
9.1	5	9.1	5	9.1	4
9.2	3	9.2	2	9.2	3
9.3	4	9.3	5	9.3	4
9.4	2	9.4	2	9.4	2
9.5	4	9.5	4	9.4	4
9.6	6	9.6	6	9.6	6
9.7	4	9.7	4	9.7	4
9.8	4	9.8	4	9.8	4
9.9	5	9.9	5	9.9	5
10.0	4	10.0	4	10.0	4
10.1	4	10.1	4	10.1	5
10.2	3	10.2	3	10.2	2
10.3	1	10.3	1	10.3	1
10.4	3	10.4	3	10.4	4
10.5	3	10.5	3	10.5	2
10.6	2	10.6	2	10.6	2
10.7	4	10.7	3	10.7	3
10.8	4	10.8	5	10.8	5
10.9	3	10.9	3	10.9	2
—	—	—	—	11.0	1

11.1	2	11.1	2	11.1	1
11.2	2	11.2	3	11.2	4
11.3	3	11.3	2	11.3	2
11.4	3	11.4	3	11.4	4
11.5	1	11.5	1	11.5	1
11.6	1	11.6	1	—	—

As one can see, there is a very close numerical relationship between the spectra, with almost identical results that differ in eigenvalue numbers by at most 0.1 and in occurrence by at most 2, with the vast majority not differing at all. Again, the relaxation curves are in an even better agreement with the full Hessian method, as seen below.

The master equation solutions given as relaxation curves, $P_i(t)$, give a better picture of how well the PCA methods did. We set up the relaxation by picking out the top tier of highest energy levels in our sample PES for Ar_{13} which were well separated from the rest. We then divided the starting population equally among them and let them relax. This seemed to be an initial condition that would be least likely to introduce a bias in the relaxation process, and, at this stage, a detailed exploration of the effects of the initial conditions seemed to be a "second-level" problem, something to address in later studies. In order to understand the results better, it is important to note that the global minimum for Ar_{13} is not only in a very deep well and its energy is significantly lower than even the next closest minimum, but it is also by far the most connected (almost three times more connected than the next closest one). Connectedness plays a very important role in the relaxation, as it dictates how much and how fast population is dumped in and out of a minimum. So, for example, a higher energy minimum that is well connected will populate fast initially but will quickly lose that population and go to zero fast. Also, a low energy minimum that is well connected, with most channels dumping into it from higher energy minima than depopulating it to lower energy minima, will build up population fast and then depopulate slower and have a higher population left in it at infinite time. Obviously, since the global minimum is the most connected, it has a much lower energy than anything else, and does not dump its population anywhere else, it will grow fast and approach a population fraction of 1 rapidly. Since the PCA methods gave results almost identical to each other and to the full Hessian method, we only graph either the $P_i(t)$ that play an important role and/or the PCA methods for them that deviate from the full Hessian or each other. All PCA curves are solid lines and all curves go to zero except for the global minimum.

The most important relaxation curve is that of the global minima, where all the population quickly ends up, for reasons stated above. Only the atomic coordinate method of PCA differed from the full Hessian for the global minimum. As can be seen in Fig. 4, the atomic coordinate method differs very slightly from the full treatment. The low energy and connectivity of the global minima allow it to dominate even the next closest minimum, which is both at a much higher energy and less than 1/3 the connectivity, as shown in Fig. 5. The lowest tier of energy (excluding the global minimum), which are also the most connected, are shown in Fig. 6. They populate well initially and go to zero much slower than the other minima, for reasons stated above. Again, only the atomic coordinate PCA differs from the full Hessian method on two of the minima. The second tier of lowest energy levels are shown in Fig. 7. These have much lower maximum population levels than those in Fig. 6 and go to zero much faster. For comparison, we have the two minima that are the second most connected (*i.e.* lower than those in Fig. 6) shown in Fig. 8 and Fig. 10 of the ESI.† The difference in the atomic coordinate PCA is shown in Fig. 8, while the highest energy tier that start with equal populations are shown in Fig. 9. We have included more graphs in the ESI.†

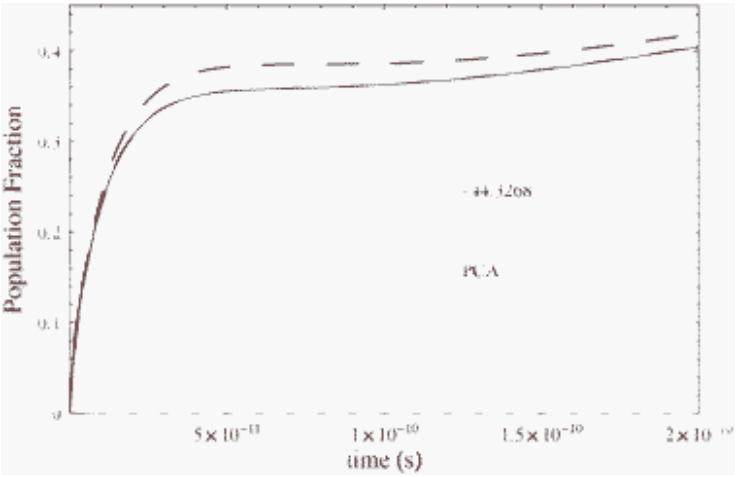

Fig. 4 Close-up view of the first part of the global minimum relaxation curve. Atomic coordinates PCA. Energies in ε.

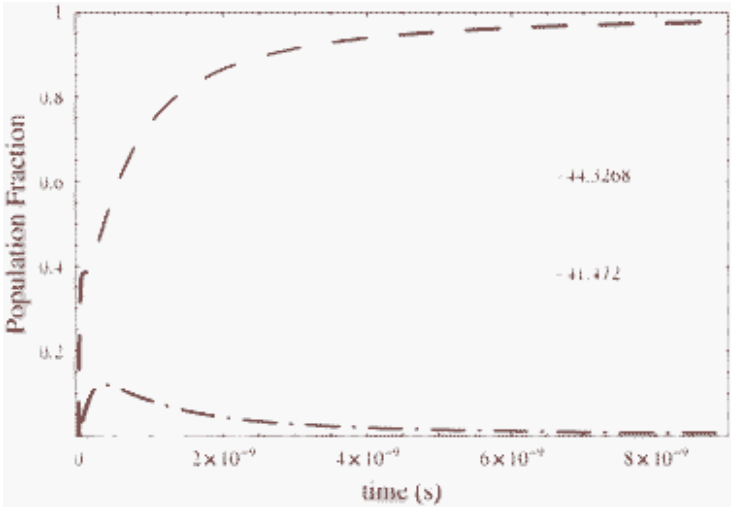

Fig. 5 Relaxation curves of the global minimum and the next lowest energy minimum. Energies in ε.

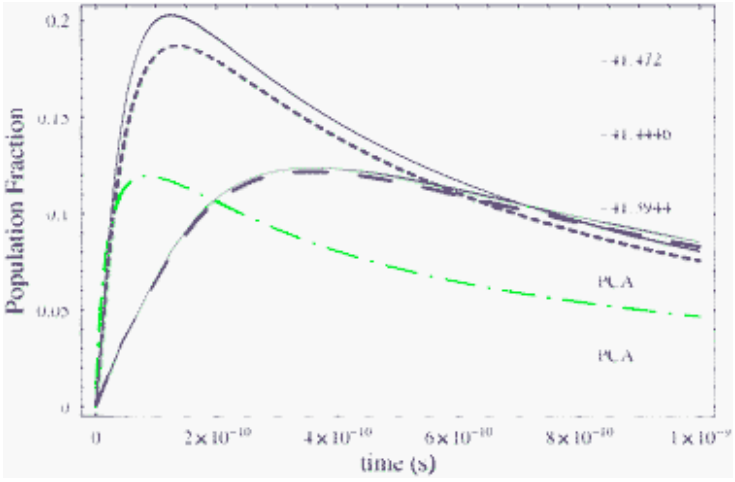

Fig. 6 Relaxation curves of the lowest energy levels (also the most connected), excluding the global minimum. Atomic coordinates PCA. Energies in ε.

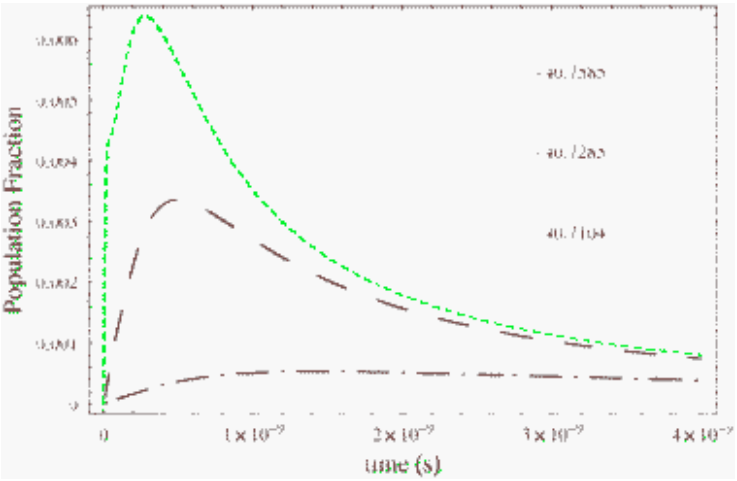

Fig. 7 Relaxation curves of the energy levels just above the lowest ones. Energies in ε.

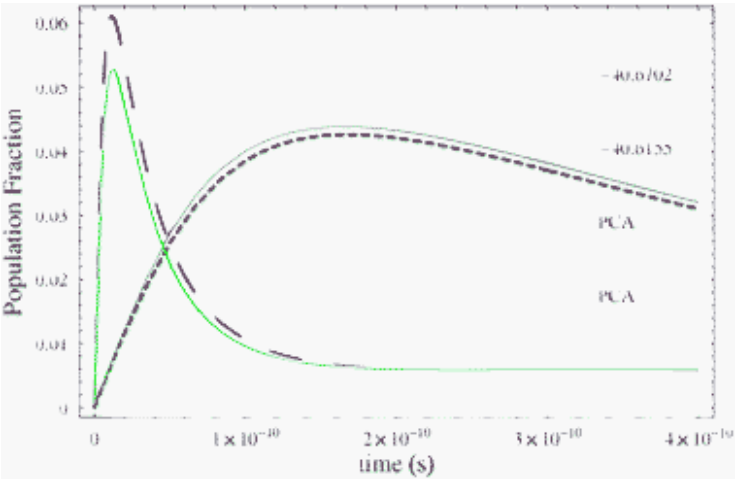

Fig. 8 Relaxation curves of second most connected minima (second to the lowest energy tier). Atomic coordinates PCA. Energies in ε.

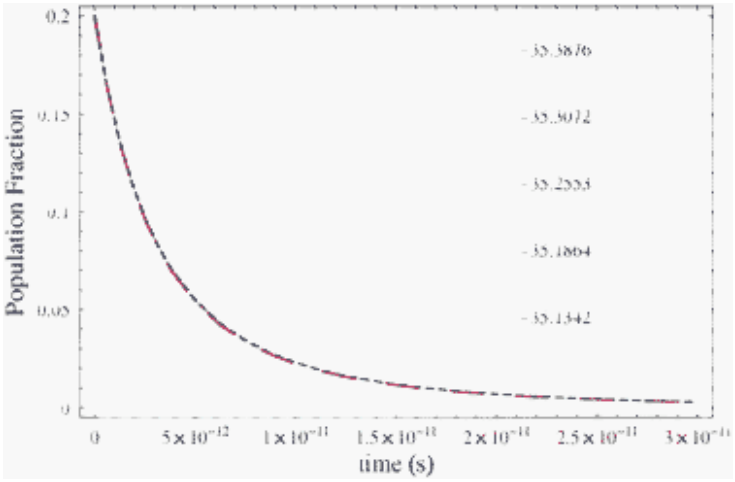

Fig. 9 Relaxation curves of the top energy tier, which were all equally populated at $t = 0$. Energies in ε.

From the relaxation curves, we see that the two PCA methods give almost exact curves to the full Hessian methods and to each other. The only difference of the PCA methods are shown above, which are few in number and differ only slightly. The only real difference between the two PCA methods is fact that the coordinate method is much faster (requiring $3N \times 3N$ covariance matrices *versus* $\frac{N(N-1)}{2} \times \frac{N(N-1)}{2}$ matrices), while the interatomic has a slight advantage in accuracy. The enormous speed advantage of the atomic coordinate method, especially for medium sized and above systems, means that it is the preferred method. The accurate eigenvalues and master equation solutions mean that PCAs give the same set of important dynamics as the case where we did not use them. However, their usefulness is currently limited since their speed advantage comes into play for systems much larger than those currently being studied computationally, as discussed in the conclusion.

4 Conclusion

Both PCA and PCO offer a way to reduce the dimensionality of a data set. PCO was successfully used to provide a 3D graph of the PES of Ar_{20}. The graph showed that the PES of Ar_{20} is made up of two main funnels. PCA was used to reduce the dimensionality of the vibrational partition function, for purposes of computing rate coefficients for the master equation. The rate constants were then used to get the eigenvalues of the master equation. The eigenvalue spectrum and master equation solutions obtained by using PCA were almost identical to using the full Hessian. These proved that PCA is a valid way to reduce the dimensionality of the Hessian and obtain effective partition functions for use in the transition matrix of the master equation. The striking similarity of the eigenvalues from PCA and the full set of coordinates gave clear evidence that passage between a minimum and a specific saddle above that minimum can be expressed very effectively in terms of a single variable—but a different variable for every minimum–saddle pair. Although PCO and PCA were both successful for their intended purposes, PCA is nonetheless not recommended for use in calculating rate constants for many situations, as explained below.

First, in constructing a PES for a cluster, all the Hessian eigenvalues, at least at the stationary points, are typically calculated as part of the search methods. For example, the database provided by Mark Miller *et al.*[10] included the products of all the eigenvalues of the individual minima and saddles, this being suitable for calculating the vibrational partition functions. Therefore, the only potential use for any method that calculates Hessian eigenvalues from the PES database (including PCA) is a case in which one does not have the eigenvalues used in constructing the PES, which is rare (if not an impossibility).

Second, calculations using the full Hessian methods require that we only diagonalize the Hessian once for each stationary point. However, the PCA method means we have to compare each saddle–minimum

pair to find the most significant changes and calculate the Hessian for the minimum and the saddle in that pair based on the result of the comparison. Therefore, while for a minimum–saddle–minimum, the full Hessian method diagonalizes three $3N \times 3N$ matrices, the atomic coordinates PCA method first needs to diagonalize two $3N \times 3N$ covariance matrices and then diagonalize four reduced Hessian matrices of about the size of $N \times N$.

Even with the computational cost advantages when one uses the Lanczos method to cut down on the computation of obtaining the three largest eigensystems of the two $3N \times 3N$ covariance matrices and diagonalizing the smaller $N \times N$ Hessians, the sheer number of extra calculations involved for the PCA means it will only see advantages for very large PES databases. At those large PES database sizes, the Lanczos method becomes many orders of magnitude cheaper than full diagonalization and the smaller size of the PCA reduced Hessians outpace the fact that more of them need to be calculated. The point at which those two computational savings of the PCA overtake the full Hessian method is very dependent on the algorithm and computational platform. Processor speed and architecture, RAM availability, operating system, and parallel computing make the most significant impact on the turn-over point. Therefore, our results must be taken as very rough guides and the only way to get a more accurate estimate of what the turning point would be for a computing system is to run speed tests on that system.

In calculating the crossover point for the PCA methods, we biased the calculation toward the PCA method by using very conservative estimates and equations for things such as the number of saddles in the database *versus* the number of minima and the advantage of using Lanczos method. Based on the timings obtained on the computational resources available to us, our estimates give us a rough point of a PES the size of that of the full PES of Ar_{200} to Ar_{250}. Note that the important factors are the size of the PES database, the relative ratio of the number of saddles to the number of minima, and the size of the Hessian matrices, so if one uses a sampled PES, it needs to be at least the size of the full PES of the above argon systems with the number of saddles not significantly larger than those in the argon systems. Since we estimate the Ar_{200} to have at least 7×10^{47} minima and 4×10^{50} saddles connecting those minima, and the largest current databases are not even in the millions, the PCA method will not be of any use for a very long time.

References

1 R. S. Berry and R. Breitengraser-Kunz, *Phys. Rev. Lett.*, 1995, **74**, 3951 [Links].

2 R. E. Kunz and R. S. Berry, *J. Chem. Phys.*, 1995, **103**, 1904 [Links].

3 K. D. Ball, R. S. Berry, R. E. Kunz, F. Y. Li, A. Proykova and D. J. Wales, *Science*, 1996, **271**, 963 [Links].

4 D. Wales, *Energy Landscapes*, Cambridge University Press, Cambridge, 2003.

5 C. J. Tsai and K. D. Jordan, *J. Phys. Chem.*, 1993, **97**, 11227 [Links].

6 J. Lu, C. Zhang and R. S. Berry, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3443 [Links].

7 K. D. Ball and R. S. Berry, *J. Chem. Phys.*, 1999, **111**, 2060 [Links].

8 D. J. Wales, *Mol. Phys.*, 2002, **100**, 3285 [Links].

9 D. J. Wales, J. P. K. Doye, M. A. Miller, P. N. Mortenson and T. R. Walsh, *Adv. Chem. Phys.*, 2000, **115**, 1.

10 J. P. K. Doye, M. A. Miller and D. J. Wales, *J. Chem. Phys.*, 1999, **111**, 8417 [Links].

11 J. C. Gower, *Biometrika*, 1966, **53**, 325.

12 I. T. Jolliffe, *Principal Component Analysis*, Springer, Berlin, 2002.

13 N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, Elsevier, Amsterdam, 1981.

14 K. D. Ball and R. S. Berry, *J. Chem. Phys.*, 1998, **109**, 8557 [Links].

15 M. A. Miller, J. P. K. Doye and D. J. Wales, *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.*, 1999, **60**, 3701 [Links].

16 Y. Levy, J. Jortner and R. S. Berry, *Phys. Chem. Chem. Phys.*, 2002, **4**, 5052 [Links].
17 O. M. Becker and M. Karplus, *J. Chem. Phys.*, 1997, **106**, 1495 [Links].
18 D. C. Rapaport, *The Art of Molecular Dynamics Simulation*, Cambridge University Press, 2004.
19 P. B. Balbuena and J. M. Seminario, *Molecular Dynamics (Theoretical and Computational Chemistry)*, Elsevier Science, Amsterdam, 1999.
20 Andrew Leach, *Molecular Modeling: Principles and Applications*, Prentice Hall, Englewood Cliffs, New Jersey, 2nd edn, 2001.
21 P. J. Robinson and K. A. Holbrook, *Unimolecular Reactions*, Wiley-Intersceince, London, 1972.
22 R. G. Gilbert and S. C. Smith, *Theory of Unimolecular and Recombination Reactions*, Blackwell Scientific, Oxford, 1990.
23 Keith D. Ball and R. Stephen Berry, *J. Chem. Phys.*, 1998, **109**, 8541 [Links].
24 F. Murtagh and A. Heck, *Multivariate Data Analysis*, Kluwer Academic, Dordrecht, 1987.
25 Lindsay I. Smith, *A Tutorial on Principal Components Analysis*, URL http://csnet.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf, 2002.
26 J. Edward Jackson, *A User's Guide to Principal Components*, John Wiley & Sons Inc., New Jersey, 2003.
27 Louis Komzsik, *The Lanczos Method: Evolution and Application*, Society for Industrial & Applied Mathematics, New York, 2003.
28 J. P. K. Doye and D. J. Wales, *J. Chem. Phys.*, 1995, **102**, 9659 [Links].
29 Alston S. Householder, Unitary Triangularization of a Nonsymmetric Matrix, *J. Assoc. Comput. Mach.*, 1958, **5**(4), 339–342 [Links].
30 Roger A. Horn and Charles R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

Footnotes

† Electronic supplementary information (ESI) available: Relaxation curves of second most connected minima; relaxation curves of some middle energy tier minima which have better maximum population achieved than the rest of the middle tier; relaxation curves of some middle energy tier minima which have better maximum population achieved than the rest of the middle tier; relaxation curves of some middle energy tier minima; some of the relaxation curves of the energy tier below the highest energy tier. See DOI: 10.1039/b913802a

‡ QR decomposition is the transformation of a matrix A into the product of an upper triangular matrix Q and an orthogonal matrix R.

0.5 Supplementary

The the difference in the interatomic distance PCA is shown in Figure 10. For the middle energy tier, we list the ones that have the best maximum population for the tier in Figure 11/Figure 12 and we list some sample other middle tier minima in Figure 13/Figure 14. The difference in the atomic coordinate PCA are shown in Figure 11 and Figure 13, while the interatomic distance PCA are shown in Figure 12 and Figure 14. Finally, the highest energy tier that start with equal population are shown in Figure 9 and the energy tier just below them are shown in Figure 15. Note that the top high energy tier minima all relax exactly the same, probably since they have very similar energies and connectivities.

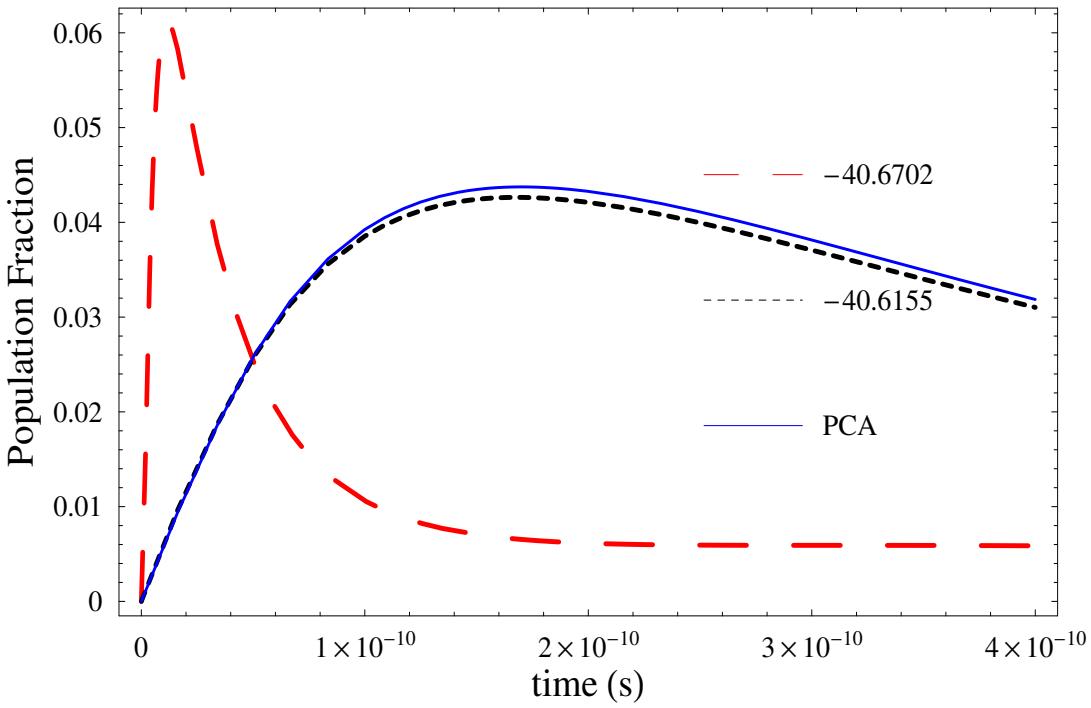

Figure 10: Relaxation curves of second most connected minima (second to the lowest energy tier). Interatomic PCA. Energies in ϵ.

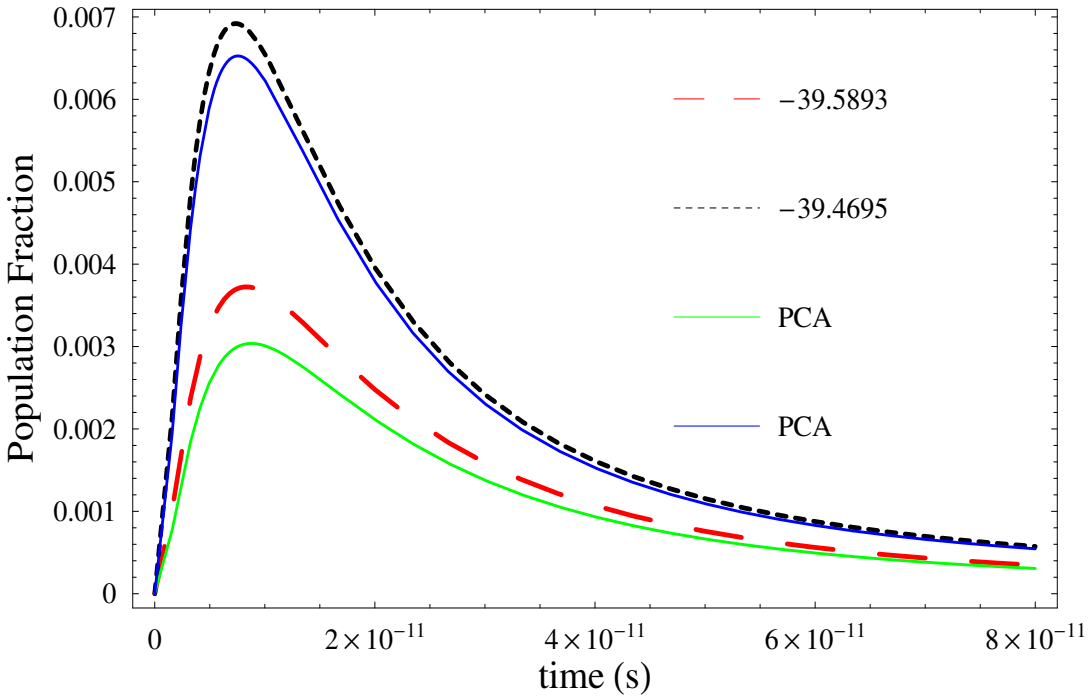

Figure 11: Relaxation curves of some middle energy tier minima which have better maximum population achieved than the rest of the middle tier. Atomic Coordinates PCA. Energies in ϵ.

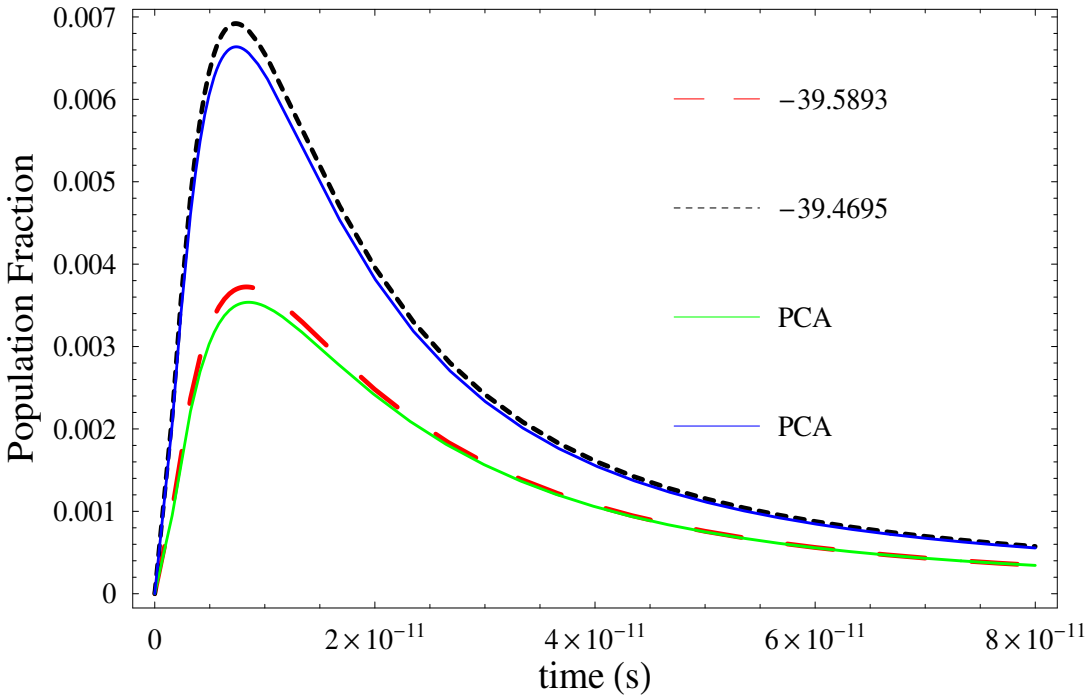

Figure 12: Relaxation curves of some middle energy tier minima which have better maximum population achieved than the rest of the middle tier. Interatomic PCA. Energies in ϵ.

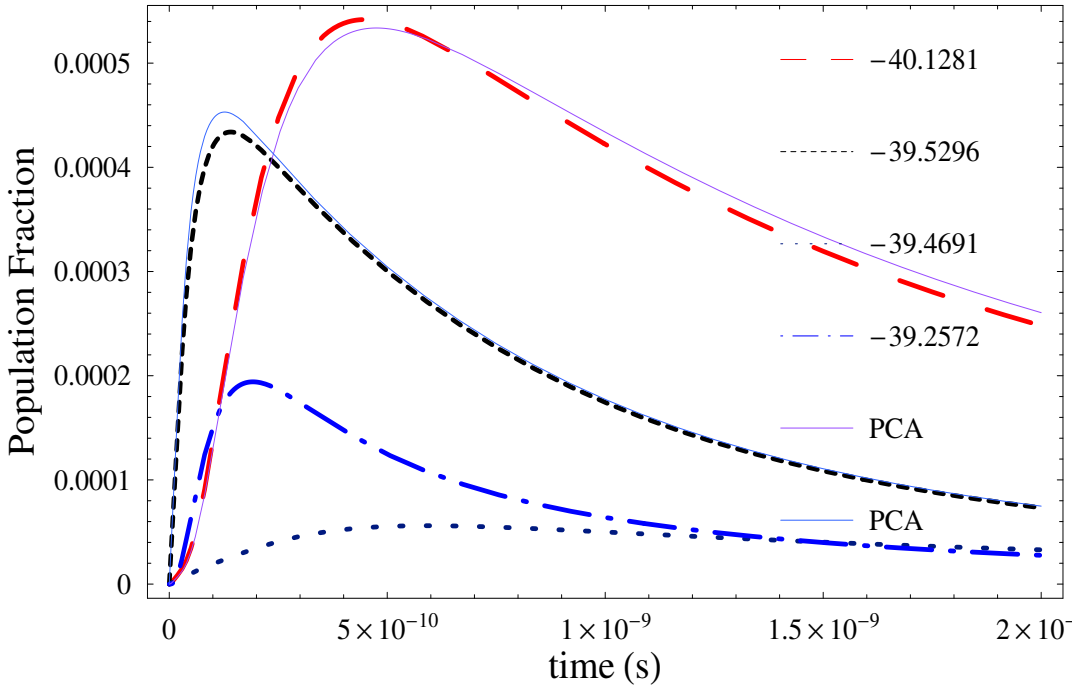

Figure 13: Relaxation curves of some middle energy tier minima. Atomic Coordinates PCA. Energies in ϵ.

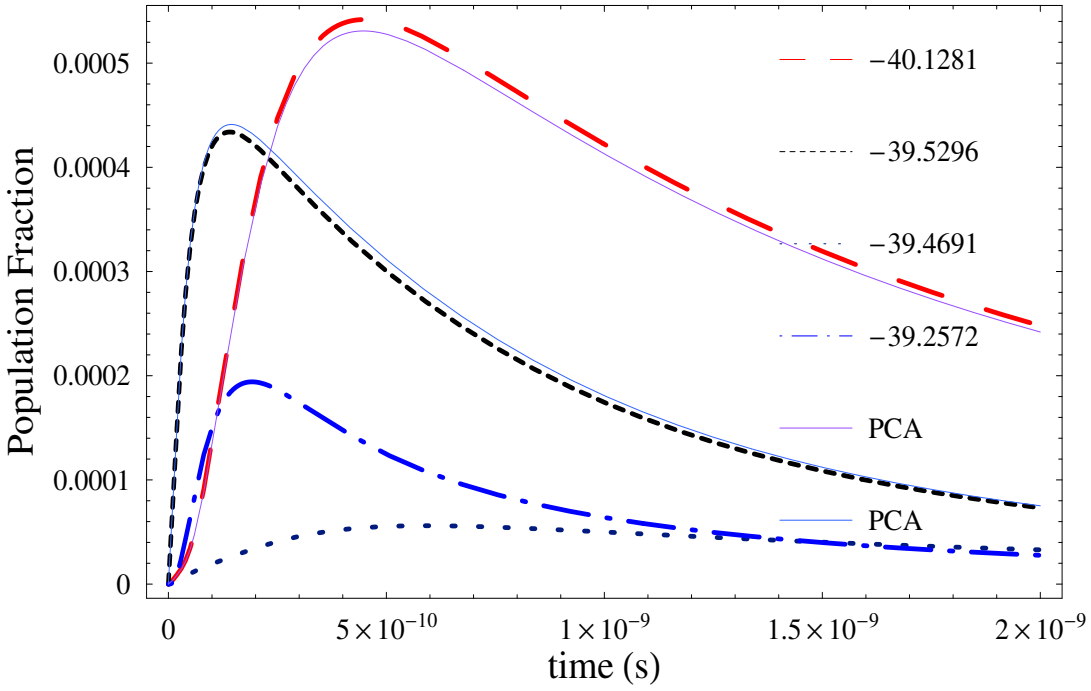

Figure 14: Relaxation curves of some middle energy tier minima. Interatomic PCA. Energies in ϵ.

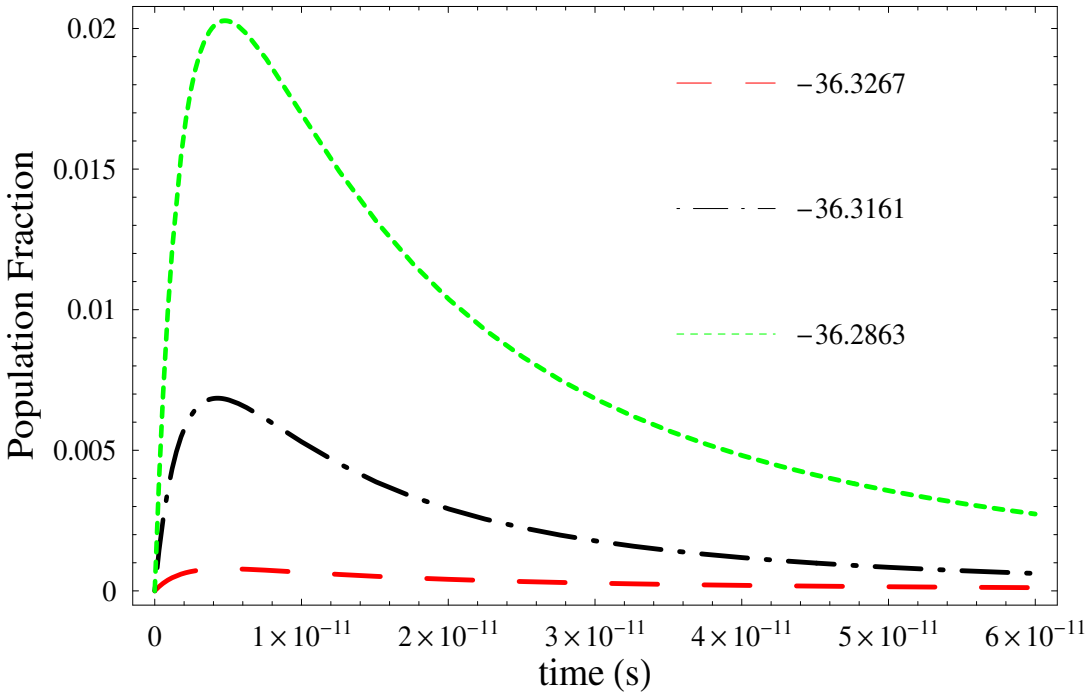

Figure 15: Some of the relaxation curves of the energy tier below the highest energy tier. Energies in ϵ.