Computationally efficient characterization of potential energy surfaces based on fingerprint distances

Bastian Schaefer and Stefan Goedecker

Citation: The Journal of Chemical Physics **145**, 034101 (2016); doi: 10.1063/1.4956461
View online: http://dx.doi.org/10.1063/1.4956461
View Table of Contents: http://aip.scitation.org/toc/jcp/145/3
Published by the American Institute of Physics

Computationally efficient characterization of potential energy surfaces based on fingerprint distances

Bastian Schaefer and Stefan Goedecker[a)]
Department of Physics, University of Basel, Klingelbergstrasse 82, CH-4056 Basel, Switzerland

(Received 25 May 2016; accepted 27 June 2016; published online 15 July 2016)

An analysis of the network defined by the potential energy minima of multi-atomic systems and their connectivity via reaction pathways that go through transition states allows us to understand important characteristics like thermodynamic, dynamic, and structural properties. Unfortunately computing the transition states and reaction pathways in addition to the significant energetically low-lying local minima is a computationally demanding task. We here introduce a computationally efficient method that is based on a combination of the minima hopping global optimization method and the insight that uphill barriers tend to increase with increasing structural distances of the educt and product states. This method allows us to replace the exact connectivity information and transition state energies with alternative and approximate concepts. Without adding any significant additional cost to the minima hopping global optimization approach, this method allows us to generate an approximate network of the minima, their connectivity, and a rough measure for the energy needed for their interconversion. This can be used to obtain a first qualitative idea on important physical and chemical properties by means of a disconnectivity graph analysis. Besides the physical insight obtained by such an analysis, the gained knowledge can be used to make a decision if it is worthwhile or not to invest computational resources for an exact computation of the transition states and the reaction pathways. Furthermore it is demonstrated that the here presented method can be used for finding physically reasonable interconversion pathways that are promising input pathways for methods like transition path sampling or discrete path sampling. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4956461]

I. INTRODUCTION

Thermodynamic and kinetic properties of multi-atomic systems are encoded in the topology of their potential energy surfaces (PESs). For example, the folding of a protein into its native state seems to be impossible based on the sheer abundance of conformational possibilities (Levinthal's paradox).[1] However, a steep funnel-like shape of the PES results in driving forces that rapidly lead the system towards its stable configuration, independent of its initial denatured structure.[2] In contrast, multi-funnel PES can explain why a certain system might be observed in a metastable state. Glass formation can be identified with trapping in some disordered state.[3] Accurately assessing the shape of a PES usually requires not only the computation of local minima but also the network of possible transitions and the corresponding energy barriers.

There exist various methods such as transition path sampling (TPS),[4–10] discrete path sampling (DPS),[11,12] stochastic surface walking based reaction sampling (SSW-RS),[13] the activation relaxation technique nouveau (ARTn),[14–17] temperature accelerated dynamics (TAD),[18,19] or the minima hopping guided path sampling (MHGPS) approach,[20–22] which allow the rigorous sampling of reactive processes. Some of these methods can be even used at sophisticated levels of theory, for example, at the level of density functional theory (DFT) level. Nevertheless, these methods are computationally very demanding, typically even more costly than the already challenging global optimization[14–17,20,23–32] problem. Therefore, computationally lightweight methods that allow to obtain at least a qualitative impression of a PES are of high interest. To this end we recently introduced distance-energy (DE) plots that allow us to distinguish glassy from non-glassy systems.[33] In a DE plot the (atomization) energies per atom of metastable configurations are measured relatively to the global minimum and they are plotted versus their configurational distance to the global minimum. Structural fingerprints, which are based on the overlap matrix of Gaussian type orbitals, can be used for measuring the configurational distances.[33,34] The same structural fingerprints are also relevant for the present work. Therefore, for the sake of being self-contained, a brief introduction into these structural fingerprints can be found in the Appendix. Even on demanding levels of theory like DFT, it is computationally feasible to produce DE plots, because only the geometries and energies of a few hundred energetically low-lying local minima, including the global minimum, are needed.

In contrast to the disconnectivity graphs of Becker and Karplus,[3,35] DE plots contain different and complementary information. DE plots focus on the relation of metastable configurations to the global minimum and display the density of the structures both with respect to energies and with respect to configurational distances. This allows the deduction of a measure for the driving force towards the global minimum. However, DE plots give no relation between two arbitrary minima and, therefore, cannot display topological information beyond the driving force towards the global minimum. This is a consequence of the very modest requirements of DE plots:

a)stefan.goedecker@unibas.ch

only the energies and geometries of the global minimum and a few hundred energetically low-lying local minima are needed. There is no need for transition state energies or the information, which minima are connected with each other by only one intermediate transition state. However, in this contribution it is demonstrated that, based only on the data obtained during conventional MH runs, an approximation to this connectivity information is available. Furthermore, it is discussed that an empirical guess for the transition state energies can be obtained, which is based solely on fingerprint distances of local minima. The combination of the approximate connectivity information and the guess for the transition state energies allow us to generate a new type of disconnectivity graph that shows a remarkable resemblance to disconnectivity graphs which are based on exact transition state energies and exact connectivity information. The post-processing of the MH data for the generation of DE plots, for the extraction of the approximate connectivity information, and for the computation of the transition state energy guess can conveniently be performed on a single core of a standard personal computer within a negligible amount of wall-clock time. Therefore, DE plots and the method presented in this contribution give a useful and computationally very affordable overview of the characteristics of a PES. They can serve as a valuable aid for making a decision whether investing the computer time that is required for building a rigorous network of transitions and their corresponding barrier energies is worthwhile and expedient with respect to a certain research goal, or not. Furthermore, they provide a qualitative idea on the kinetics and thermodynamics of a system. Moreover, the method presented below is demonstrated to be a promising tool for isolating physically reasonable intermediate metastable structures of complicated reactions, which, for example, might be used for generating initial pathways that are needed for methods like TPS or its discrete variant, DPS.

II. CORRELATING TRANSITION STATE ENERGIES WITH STRUCTURAL DIFFERENCES

Often the energies of two structurally similar minima of a PES are very close to each other, whereas the energy differences between structurally very different minima can be large. Nevertheless, it is clear that structurally very different minima can have very similar energies, as well. In other words, it is expected that for small structural differences the probability to find large energy differences is small, whereas for large structural differences, both, small and large energy differences between two adjacent minima are likely. Indeed, this expectation is supported by the data shown in Fig. 1. For the neutral silicon cluster consisting of 20 atoms, this figure shows the density of the distribution of energy differences of minima pairs plotted versus the corresponding permutationally optimized root-mean-square distance (RMSD).[34] All minima pairs used for this plot are separated by only one intermediate transition state. It is seen from this plot, that for small RMSD values the density of the data points vanishes for large energy differences, whereas for large RMSD values,

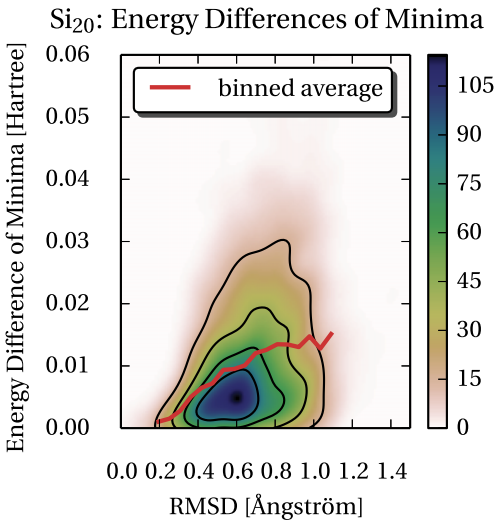

FIG. 1. Density plot of the energy differences of pairs of minima versus their RMSD distance for the Si_{20} system. The shown data sets consists of roughly 2900 minima pairs. Each pair of minima is connected by only one intermediate transition state. The structures, energies and the connectivity of the stationary points were determined at the DFT level of theory (PBE exchange correlation functional) by using the MHGPS method coupled to the BigDFT code.[21,22,36–38] The shown density was obtained from the corresponding scattered data by means of a Gaussian kernel density estimate as implemented in Python's scipy library. The red bold line shows the same data, but averaged within 25 bins along the RMSD axis. Only bins that contain at least 5% of the number of data points of the bin with the most data points are shown.

there is a significant density, both for small and large energy differences. Because the variance in the energy differences is larger for increasing RMSD values, also the average values of the energy differences rises, as is shown by the binned average of the energy differences (red line).

Except for degenerate rearrangements, the barrier energy of every transition state can be measured with respect to the lower or the higher energy minimum to which the transition state is connected to. In contrast to the distribution of the energy differences of neighboring minima in an energy difference versus RMSD plot, it can be expected that there is a stronger correlation in a plot of the uphill (larger) barriers versus the RMSD. Intuitively, this partially should result from a combination of the fact that the absolute values of the energy differences of two neighboring minima are a lower bound for the uphill barriers and the assumption that the average downhill barrier energy should rise if the distance between the minima increases. Therefore, the probability to find small uphill barriers between structurally very different minima should be expected to be small.

In order to analyze this idea more rigorously, a simple parabola model of the PES, as illustrated in Fig. 2, is used. In fact, similar parabola models can be used for the explanation of the Bell-Evans-Polanyi principle (a linear model is sufficient, though), the Marcus equation, Hammond's postulate, and the relationship of low-curvature directions with low barrier energies.[39–45] In such a parabola model, the transition state is given by the intercept $(\xi, E(\xi))$ of both parabolas. From Fig. 2 it is evident that the barrier energies should rise with increasing structural distances between the minima. Here both

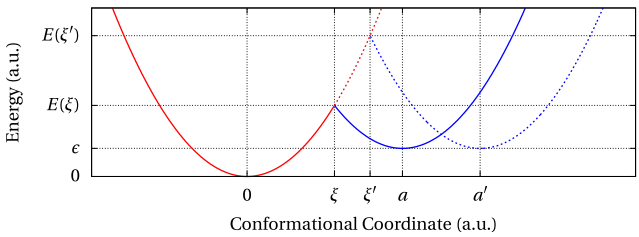

FIG. 2. Parabola model for the transition state energy. For increasing structural differences of both minima the transition state energy is rising. Here this is illustrated by shifting the minimum of the solid blue parabola from a to a'. The shifted parabola is visualized by a blue dashed line.

parabolas are assumed to have the same curvature k ("force constant"), and their minimum values are shifted by an amount of ϵ. The structural distance of both minima is denoted as a. Consequently, the transition state ξ and its corresponding uphill barrier $E_u = E(\xi)$ is given by

$$\xi = \frac{a}{2} + \frac{\epsilon}{2ak}, \tag{1}$$

$$E_u = k\left(\frac{a}{2} + \frac{\epsilon}{2ak}\right)^2. \tag{2}$$

For each pair of minima, this model is applied to the data of Fig. 1 and the result is visualized in Fig. 3 ($k = 0.08$ Ha/Å^2). In contrast to the energy differences of the minima in Fig. 1, this model predicts a clear correlation between the structural difference (RMSD) of two directly neighboring minima and the energy of the corresponding uphill barrier.

It remains to be verified if the energies of real (computed) uphill barriers between structurally very different minima also tend to be larger than the energies of the uphill barriers between structurally similar minima. If there is a breakdown in this hypothesis, it is expected that no correlation of the type shown in Fig. 3 is seen. For this verification, transition states and their directly connected minima were computed for Si_{20} and Au_{26}^- at the DFT level of theory as implemented in the

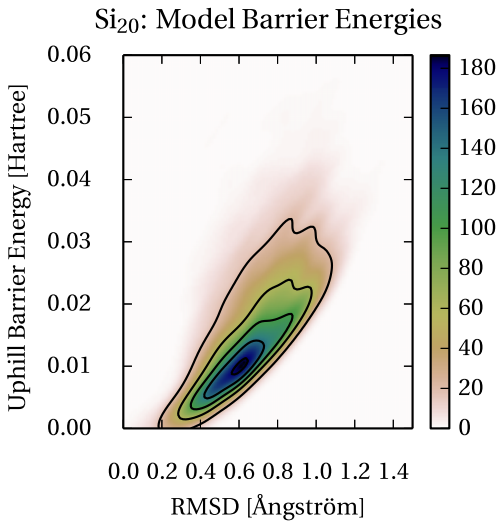

FIG. 3. Same as Fig. 1 but for model uphill barrier energies instead of energy differences of minima. Shown is the distribution of uphill barriers plotted versus the configurational distance of directly neighboring minima, as obtained by the model of Eq. (2). Here, the same pairs of minima are used that already were used for Fig. 1.

BigDFT[36–38] code and for $(NaCl)_{32}$ and $(NaCl)_{29}$ using the Born-Mayer-Huggins-Tosi-Fumi[46–50] (BMHTF) force field. For Si_{20} the PBE[51] functional was used, whereas for Au_{26}^- the LDA[52,53] functional was used and in case of the BMHTF force field the parameters of Ref. 54 were chosen. Furthermore, transition states and the directly connected neighbors were computed for the Lennard-Jones[55,56] clusters of sizes 19, 38, and 55. Except for Au_{26}^-, the geometries and energies of the minima, as well as their connectivity, were established using the MHGPS method as implemented in the BigDFT suite. In the case of Au_{26}^- the data were taken from a previous study and it is referred to this study for a description of its computation.[57] The density plots of the uphill barrier energies versus the RMSD are given in Fig. 4. As can be seen from this figure, there is indeed a good correlation between the structural difference (RMSD) and the uphill barrier.

Though the permutationally optimized RMSD is a very natural measure for structural differences, it is very time consuming to compute, which often makes it impracticable to use. For example, the computation of the roughly 58 000 RMSDs for the LJ_{55} plot in Fig. 4 took about 14 h (wall clock time), despite using 150 cores in parallel. Of course, actual wall clock times depend very strongly on the underlying computer hardware. Nevertheless, this example illustrates that computing large numbers of RMSDs can be problematic in practice. Therefore, the plots of Fig. 4 have been repeated using s- and p-orbital fingerprint distances instead of RMSDs and are shown in Fig. 5. Again, a correlation between the structural difference measured by the s- and p-orbital fingerprint distance and the uphill barrier energy can be observed. Using s- and p-orbital based fingerprint distances as a measure for structural differences, the LJ_{55} plot in Fig. 5 took on the order of minutes on a single core, which is a striking advantage over the RMSD and makes it much more useful in practice. Plots from fingerprint distances using only s-type orbitals have a very similar appearance and are given in the supplementary material.

Finally, a short comment seems to be appropriate on why it is almost exclusively focused on the uphill barriers. After all, as can be seen from Eq. (2), the same dependence of the downhill barriers on the structural distance as for the uphill barriers is predicted, except for a constant energy shift that is given by the energy difference of both minima. This, however, does not imply that necessarily a similar correlation as for the uphill barriers must be observed for the downhill barriers. The reason is that even though two minima might be far apart from each other, the downhill barrier can be vanishingly small if, in return, the energy difference between the two minima is comparatively large. Indeed, plotting the downhill barrier versus the structural difference results in a distribution that looks very similar to the distribution of the energy differences of the minima.

III. GENERATING ROUGH OVERVIEWS OF POTENTIAL ENERGY SURFACES

In this section, an empirical method suitable to generate trajectory-based connectivity databases is presented. This method is based on post-processing data obtained from one or

several MH runs. Once MH runs are done, the computational cost of this method is independent of the underlying level of theory that was used for the MH runs. On a single core of a standard office computer, this method allows the generation of trajectory-based connectivity databases within a negligible amount of wall clock time even if the trajectory-based connectivity database shall describe PESs that are defined by computationally demanding methods, for example, DFT. To introduce this novel method, first the term "trajectory-based connectivity database" is defined. A trajectory-based connectivity database is understood to contain three types of information. First, it contains all local minima visited during a certain number of MH runs. Second, it contains the information which minima were visited consecutively by the MH walkers and finally, also a qualitative measure for the energy needed to interconvert the consecutively visited minima is part of a trajectory-based connectivity database. Furthermore, a pair of minima visited consecutively by the MH walker will be denoted as "hopping pair."

In contrast to such a trajectory-based connectivity database, the stationary point database defined by Wales[3,11,12] contain minima, transition states, and the information to which minima the transition states are connected by minimum energy or energy minimized pathways. Thus, a trajectory-based connectivity database can be seen as an approximation to a stationary point database. The connectivity information is approximated by the information which minima were visited consecutively by the MH walker. This is a reasonable approximation, because the MH walkers explore the PES by means of short MD trajectories that, at most times, have relatively moderate initial kinetic energies. As a consequence, the geometries of hopping pair members typically are very similar to each other, a circumstance that is also used in the MHGPS scheme.[21] Quantitative evidence for the validity of this connectivity approximation is given in Fig. 6. In this figure, the probability distribution of the number of intermediate transition states needed by the MHGPS method to connect pairs of consecutively accepted minima is given. These numbers constitute an upper bound to the minimum number of intermediate transition states located in between two consecutively accepted minima. It can be seen from this figure that the majority of consecutively accepted minima

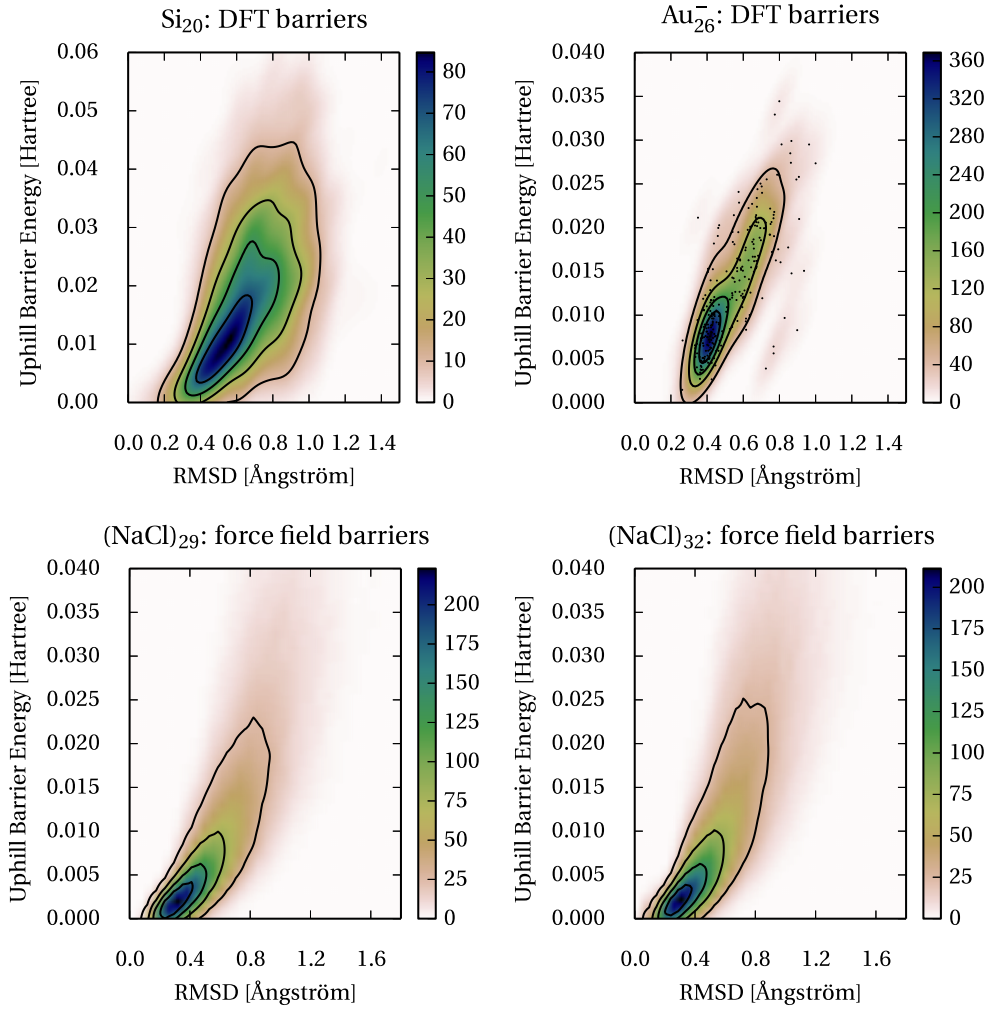

FIG. 4. Gaussian kernel density estimates of the uphill barrier energies versus the (permutationally and chirally optimized) RMSD distance of minima pairs that are separated by only one intermediate transition state. If two minima are connected by more than one intermediate transition state, only the transition state with the lowest energy was included in the data sets used for these plots. The plot for Au_{26}^- was obtained from only 259 transition states. It, therefore, is possible to show every single data point for Au_{26}^-, which allows us to demonstrate the soundness of the Gaussian kernel density estimate. The plot for Si_{20} was generated from roughly 3000 transition states and the plots for the systems described by force fields were obtained from roughly 50 000 to 70 000 transition states.

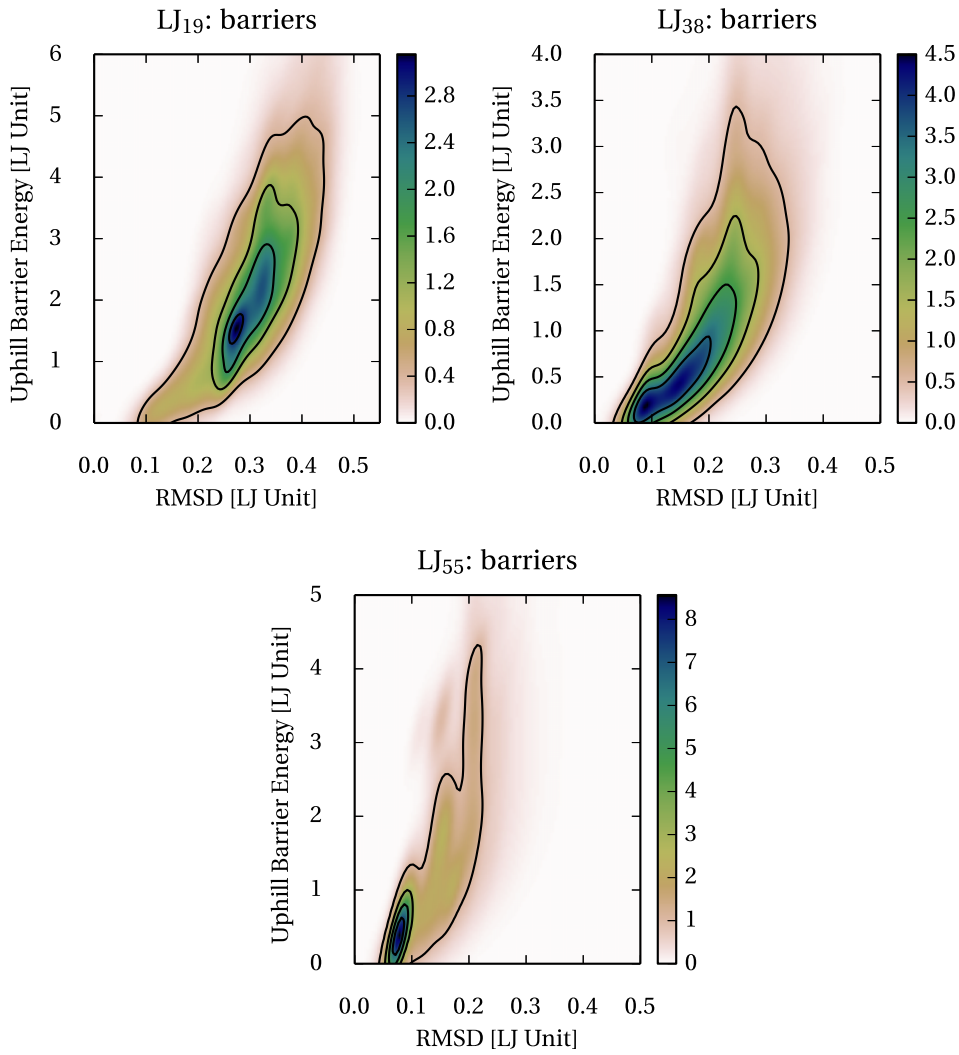

FIG. 4. (*Continued.*)

can be connected with each other by no more than two intermediate transition states.

What remains to be discussed is, how an educated guess for the energy, which is needed to interconvert the minima of a hopping pair, can be obtained. Before describing the actual method for obtaining such a guess, a different approach is discussed. From a theoretical point of view, it would be very satisfying if Eq. (2) could be used to obtain a guess for the transition state energy. Indeed, using a suitable value for the force constant k, it turned out to be possible to generate disconnectivity graphs of similar quality as those based on the method that is presented below. However, for us, it was only possible to choose good values for k, if the correct appearance of the disconnectivity graph was known. Unfortunately, a procedure that is able to reliably determine the force constant and that is able to give disconnectivity graphs of similar quality as those based on the method outlined below has yet to be found. In fact, using inappropriate values for k can produce completely misleading disconnectivity graphs. In contrast to this, in all tested cases, the approach discussed below produced qualitatively very reasonable disconnectivity graphs.

The remainder of this section focuses on describing the empirical method that was able to produce an educated qualitative guess for the transition state energies. In this approach the energy difference of the two minima of a hopping pair is compared to the average energy difference of minima of hopping pairs that are separated by a similar structural fingerprint distance. If the energy difference is larger than the average value at this fingerprint distance, the uphill barrier of a hopping pair is estimated as the absolute value of the energy difference of the two hopping pair members. Otherwise, the uphill barrier is estimated as the average absolute value of the energy differences at this fingerprint distance. In practice, this is done by plotting the absolute values of the energy differences of the minima of each hopping pair versus their fingerprint distance and computing binned averages of these data. A continuous function describing this binned average is obtained by means of a fitting procedure. Of course, this approach does not give a quantitative estimate of the energy of each single barrier, but it is intended to reproduce the energy scale and roughly the average trend in uphill barrier energies that was discussed in Sec. II. More explicitly, assuming the minima energies of a hopping pair to be E_1 and E_2 with

$E_1 \leq E_2$, the absolute energy E_t needed to interconvert the two minima is estimated as

$$E_t := \max\left(E_1 + E_u(a),\, E_2\right), \tag{3}$$

where the max-function returns the larger of its two arguments and the uphill barrier energy, E_u, is a function of the fingerprint distance a (see Fig. 2). E_u is defined as

$$E_u(a) := \alpha \exp(-\beta|a + \gamma|^{\delta}), \tag{4}$$

where the parameters α, β, γ, and δ are obtained by a fit to the binned averages of the energy differences of the minima of hopping pairs. The fitting function given in Eq. (4) is a heuristic and pragmatic choice that turned out to work well in all tested cases. The fitting itself is performed with the help of the nonlinear least-squares Marquardt-Lavenders algorithm as implemented in the gnuplot code.[58–60] Of course, other fitting methods can be used, because E_u is only required to provide a continuous function of the qualitative trends for the uphill barrier energies. A plot exemplifying such a fit is given in Fig. 7 for the case of $(NaCl)_{32}$.

It turned out that by using Eq. (3) for obtaining transition state energy guesses, it is possible to produce disconnectivity graphs that reflect reasonably well the characteristics of a PES. Before presenting these disconnectivity graphs, it is appropriate to discuss the reasonable performance of Eq. (3). To see this, first it is realized that Eq. (3) splits up the hopping pairs into two sets.

In the first set, the uphill barrier of a hopping pair is guessed by means of Eq. (4). In Fig. 5, the fitting function Eq. (4) is plotted on top of the uphill barrier distributions of Si_{20}, $(NaCl)_{29}$, $(NaCl)_{32}$, LJ_{19}, LJ_{38}, and LJ_{55}. From these plots it is evident that the binned average of the absolute values of the energy differences of hopping pair minima is a reasonable guess for the uphill barrier energy. Eq. (4) prevents the assignment of low transition state energies to hopping pairs with structurally very different minima and, therefore, is in agreement with the results of Sec. II. This agreement is essential for an acceptable reproduction of the characteristics of a PES. Otherwise, as will be seen from the disconnectivity graphs that are presented below, superbasins are likely to be merged, which can result into a completely misleading appearance of a PES. Furthermore it can be seen from Fig. 5 that the uphill barrier energy which is assigned to a hopping pair corresponds in most cases to a not completely unlikely

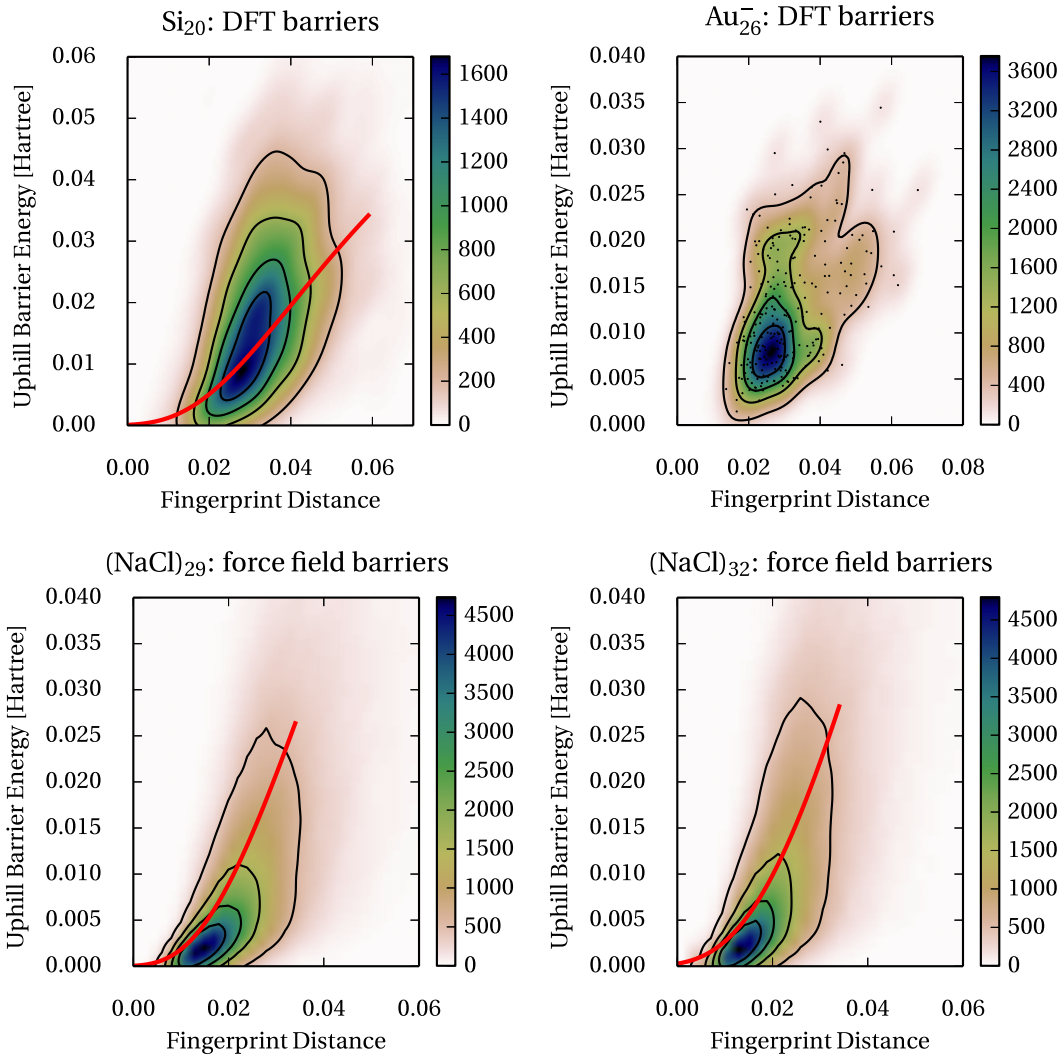

FIG. 5. Same as Fig. 4, but using s- and p-orbital fingerprint distances instead of the permutationally optimized RMSD. Plots from fingerprint distances using only s-type orbitals have a very similar appearance and are given in the supplementary material. The red lines are graphs of Eq. (4) and are discussed in Sec. III.

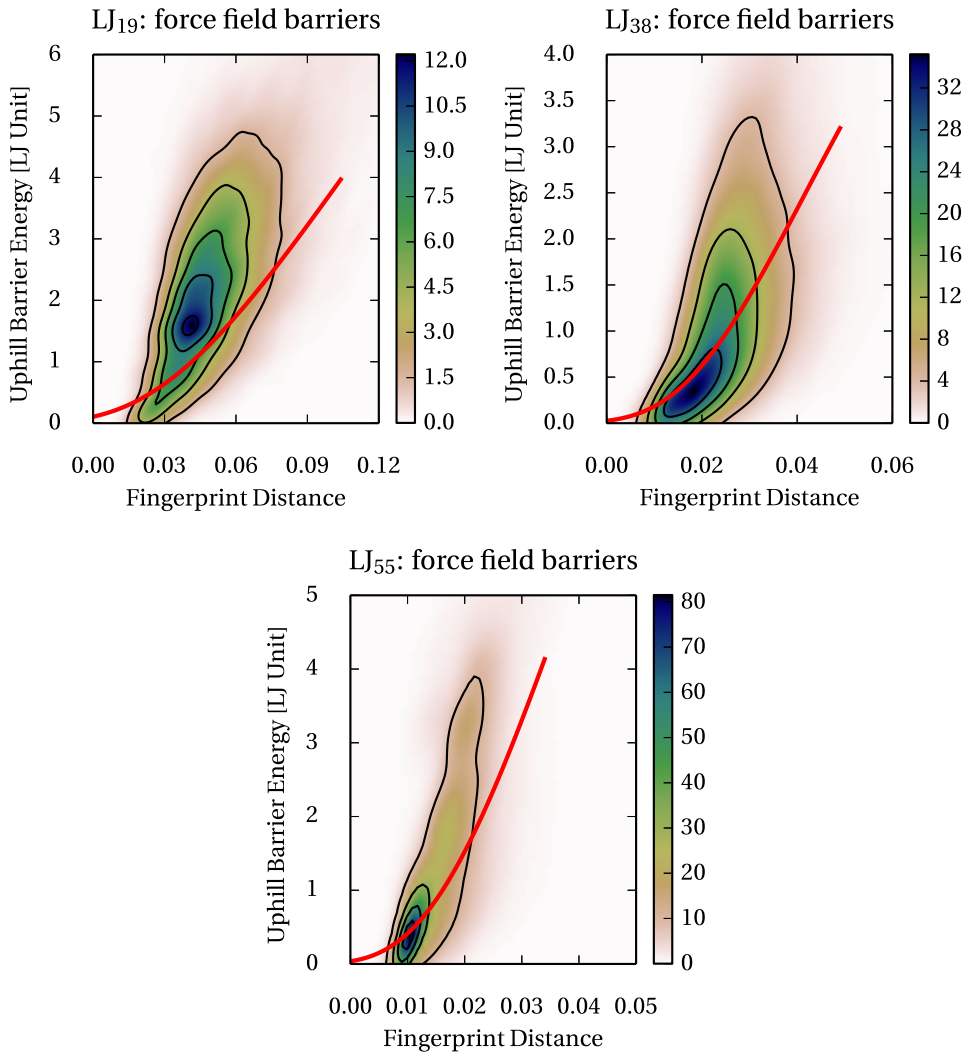

FIG. 5. (*Continued.*)

uphill barrier energy at a given structural distance. As was demonstrated by Fig. 6, the minima of many hopping pairs are separated by only one intermediate transition state and it is clear that the trend of increasing uphill barrier energies with increasing structural distances that was described in Sec. II can be applied to these hopping pairs. However, there is no strict guarantee for the minima of a hopping pair to be in a close neighborhood to each other. Despite this

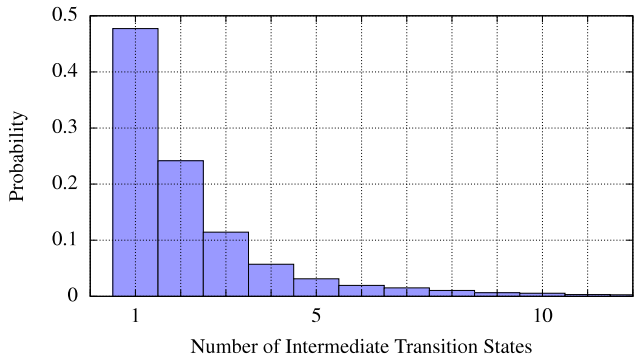

FIG. 6. Shown for the LJ_{55} system is the probability distribution of the number of intermediate transition states needed by the MHGPS approach as implemented in the BigDFT-suite to connect pairs of consecutively accepted minima. The data set consists of more than 20 000 connection attempts that were stopped if the connection could not be established within 30 transition state computations.

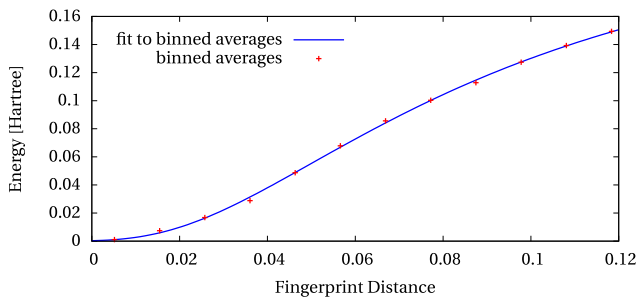

FIG. 7. Fit of E_u as defined in Eq. (4) to the binned averages of the energy differences of $(NaCl)_{32}$ hopping pairs, as modeled by the BMHTF force field, versus their structural difference measured by the overlap matrix fingerprint distance using s- and p-type orbitals. 25 bins were used for grouping the roughly 28.000 data points. Of those 25 bins, only those that contain at least 5% of the data points of the bin with the most data points are shown and were used for the fit. The values of the fitting parameters are $\alpha = 0.2449$ Ha, $\beta = 0.0128$, $\gamma = 0.0445$, and $\delta = -2.0159$.

fact, it is still the trend that was described in Sec. II that is used to obtain a guess for the barrier energies of those hopping pairs. At a first glance, this might be surprising since two structurally very different minima, which only can be interconverted into each other by crossing many intermediate transition states, might very well be separated by a low overall barrier. For example, this can be the case if the pairwise structural distances of all intermediate minima are small. Using a measure for the transition state energies that is based on the correlation discussed in Sec. II, a high barrier energy will be assigned to the direct transition between such minima. However, this is not a disadvantage, but rather a desirable effect. Typically, the analysis of a trajectory-based connectivity database will focus on low energy pathways. In such an analysis, the direct interconversion of those far apart minima is disfavored due to the high energy that is assigned to their direct interconversion. In contrast, low barrier energies are properly assigned to the pathway that leads over the large number of pairwise structurally similar minima, which allows for its identification.

In the second set, the transition state energies of hopping pairs are approximated by the energy of the energetically higher minimum. For transitions with downhill barriers that are small compared to the uphill barrier, this is a sufficient approximation. However, if the energy difference between two minima is small and their structural difference large, this approximation is not only quantitatively but also qualitatively very inaccurate. Fortunately, Eq. (4) rigorously prevents the latter hopping pairs from being included into this second set. This second set only contains hopping pairs with above-average energy differences with respect to a given structural distance. Therefore, for those hopping pairs for which a significant underestimation of the transition state energy endangers a reasonable reproduction of the overall PES characteristics in a disconnectivity graph, the uphill barriers are not estimated by the energy difference of the involved minima.

Fig. 8 displays disconnectivity graphs for Si_{20}, $(NaCl)_{29}$, $(NaCl)_{32}$, LJ_{19}, LJ_{38}, and LJ_{55}. As above, the PES of Si_{20} was computed at the DFT level of theory as implemented in the BigDFT code (PBE exchange correlation functional). For the sodium chloride clusters, again the BMHTF force field was used. No disconnectivity graphs are presented for Au_{26}^- because only the local minima, but not the complete minima hopping history, were archived from the previous work.[57] The panel labels of Fig. 8 follow the scheme (x.*n*), where "x" is one of a, b, c, d, e or f and represents the system (a = Si_{20}, b = $(NaCl)_{29}$, c = $(NaCl)_{32}$, d = LJ_{19}, e = LJ_{38}, and f = LJ_{55}) and *n* runs from one to three. Disconnectivity graphs in the panels (x.1) and (x.2) (the left and middle column of Fig. 8) are based on trajectory-based connectivity databases, where for the (x.1) panels the barrier energies were set to the energy of the higher minimum and for the (x.2) panels the barrier energies were approximated by Eq. (3) and the above described fitting procedure. The (x.2) disconnectivity graphs will also be denoted as "fingerprint disconnectivity graphs." For the center column of Fig. 8, fingerprint distances based on *s*- and *p*-orbitals were used. Disconnectivity graphs in the rightmost column of Fig. 8 (panels (x.3)) are based on stationary point databases that were generated by means of the MHGPS approach.[21] These standard disconnectivity graphs are considered as the reference for the present purpose. For each system, all three disconnectivity graphs show the same number of minima, however, not necessarily the identical minima. This is, because the stationary point databases are usually much more detailed, because they were thoroughly sampled by the MHGPS approach in order to generate exact reference disconnectivity graphs. In Table I rough sizes of the underlying databases are given.

It should be pointed out that according to both the conventional disconnectivity graph (Fig. 8(a.3)) and the novel fingerprint disconnectivity graph (Fig. 8(a.2)) the energy landscape of Si_{20} has a distinct double-funnel character with a high barrier separating the global minimum from the second lowest local minimum. Based on both the conventional and the new fingerprint disconnectivity graph it therefore seems conceivable that at moderate temperatures Si_{20} can be trapped kinetically in the funnel that belongs to the second lowest minimum. Furthermore, one can see that there are far more minima in the funnel that belongs to the second lowest minimum. Therefore, at finite temperatures, entropic effects might render the second lowest minimum to be thermodynamically stable. Even if only using the connectivity as provided by the trajectory-based connectivity database, but eliminating all barriers, the double-funnel landscape of Si_{20} is clearly visible (Fig. 8(a.1)), nevertheless, the appearance of the disconnectivity graph is improved by using the fitting procedure for approximating transition state energies (Fig. 8(a.2)).

Though, for Si_{20}, the most important feature of the system is already visible in the (a.1) panel, the same is not true for the remaining systems. Except for Si_{20}, completely eliminating the barriers results in disconnectivity graphs that correspond to extreme structure seekers and the true topology of the PESs is not visible in the (x.1) panels. In contrast to this, the fingerprint disconnectivity graphs in the (x.2) panels exhibit a remarkable resemblance to the standard disconnectivity graphs shown in the (x.3) panels of Fig. 8. For example, NaCl clusters are exceptional clusters because their most stable structures usually resemble fragments of the NaCl crystal lattice.[61] In the case of $(NaCl)_{32}$ the number of NaCl units allows to form a NaCl crystal lattice without any vacancies. The disconnectivity graphs of $(NaCl)_{32}$ (Figs. 8(c.2) and 8(c.3)) clearly identify this cluster to be a structure seeker. The downhill barriers from all metastable structures towards the global minimum are much smaller than the corresponding uphill barriers. Therefore, there is a strong driving force towards the global minimum for this cluster. In contrast to this, the energy landscape of $NaCl_{29}$ (Figs. 8(b.2) and 8(b.3)) does not show such a strong structure seeker character, which is a consequence of the number of NaCl units that is prime.

The fingerprint disconnectivity graphs based on *s*- and *p*-orbital fingerprints are slightly more similar to the standard disconnectivity graphs than those based only on *s*-orbitals and shown in the supplementary material. Nevertheless, the fingerprint disconnectivity graphs based on the *s*-only fingerprints also provide a striking resemblance to the standard

disconnectivity graphs, in particular if taken into account that generating fingerprint based disconnectivity graphs is a quasi-free lunch post-processing of MH data.

Besides for generating disconnectivity graphs and qualitatively judging the kinetics and thermodynamics of PESs, trajectory-based connectivity databases can also be used to extract well aligned sequences of minima. These well aligned sequences of minima can be hoped to lie on a low-energy pathway between two given states. Such minima sequences are of great importance, because they provide promising starting points for generating initial pathways that are needed for methods like TPS or its discrete variant, DPS.[4–12] For non-trivial reactions involving large structural changes such a generation of initial pathways is in itself a very difficult task and no generally applicable solution seems to exist, so far.[62] Isolating a suitable sequence of minima from a trajectory-based connectivity database can be done by applying a modified Dijkstra's algorithm which in a first round searches for a path that minimizes the maximum barrier at any of its transitions and in a second round minimizes with

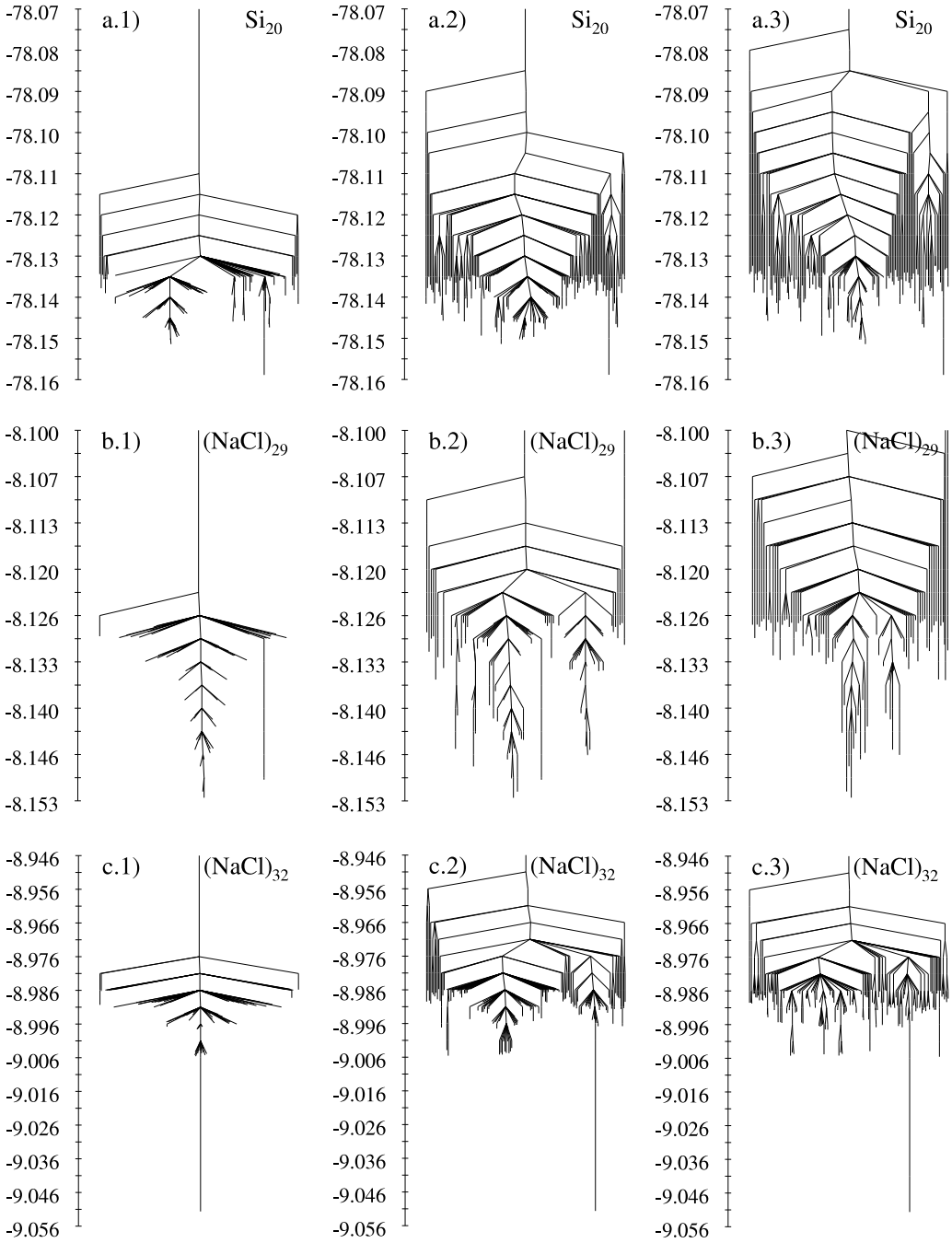

FIG. 8. Disconnectivity graphs for Si_{20} (panels (a.n)), $(NaCl)_{29}$ (panels (b.n)), $(NaCl)_{32}$ (panels (c.n)), LJ_{19} (panels (d.n)), LJ_{38} (panels (e.n)) and LJ_{55} (panels (f.n)). The graphs in panels (x.1) and (x.2) are based on trajectory-based connectivity databases. For the (x.1) panels, the barriers were eliminated, whereas the approximations to the barrier energies described in Sec. III were used for the (x.2) panels. Reference graphs based on stationary point databases that were sampled by the MHGPS approach are shown in the rightmost column (panels (x.3)). The energy scale is in Hartree (Si_{20}, $(NaCl)_{29}$, $(NaCl)_{32}$) and in Lennard-Jones units (LJ_{19}, LJ_{38}, LJ_{55}).

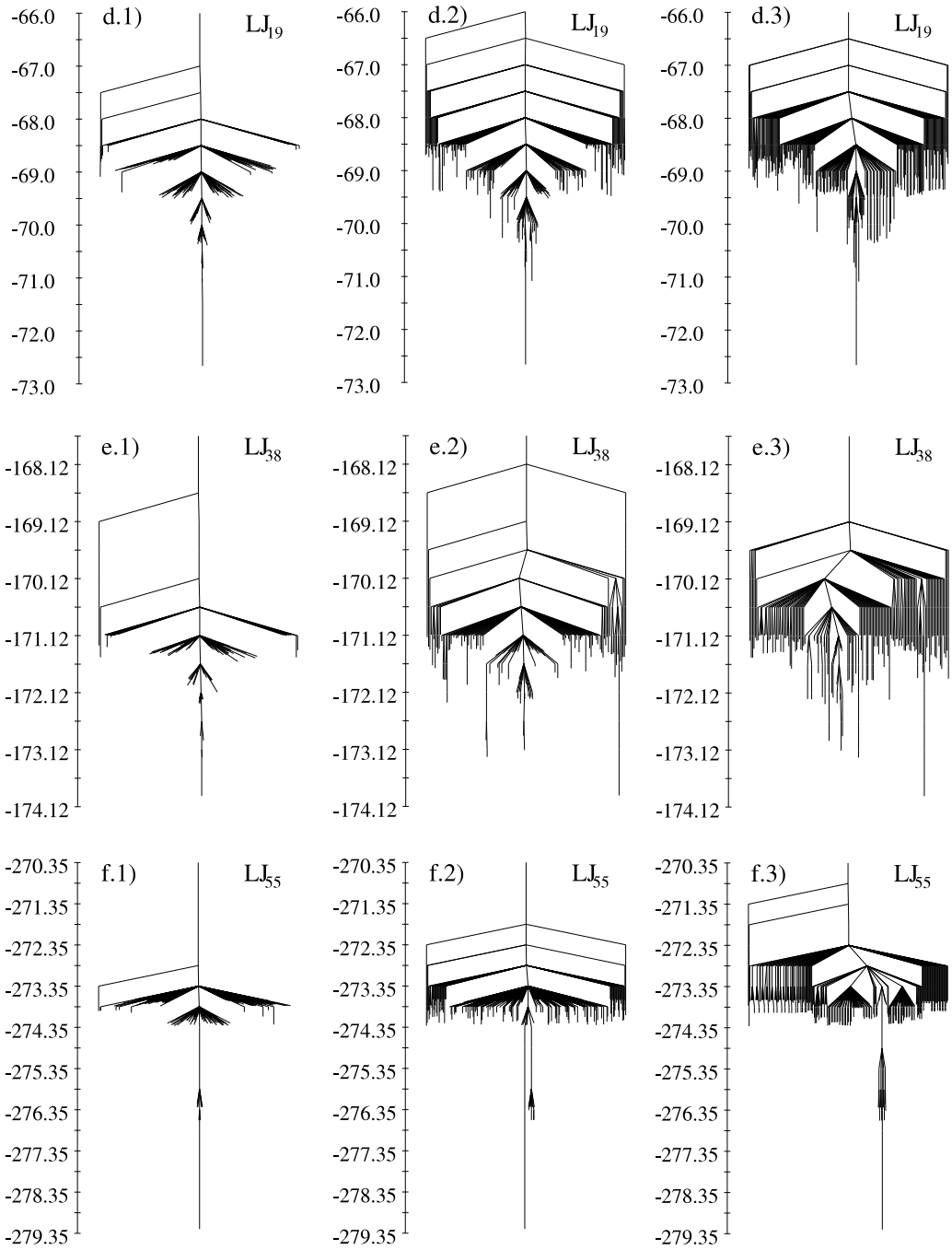

FIG. 8. (*Continued.*)

respect to the number of intermediate transitions.[21] Of course, the thus isolated pathways are not necessarily complete in the sense that it might not be possible to connect the two minima of a hopping pair by only one single intermediate transition state. However, the isolated sequence of minima represents minima that were visited in consecutive order by an MH walker. Therefore, they are suitable for getting connected by the connectivity finder module of the MHGPS code (instead of the usual sequence of accepted MH configurations).

For the Si_{20} system a sequence of minima between the putative global minimum and the putative second lowest minimum was extracted from the trajectory-based connectivity database. For this sequence of minima, all intermediate transition states and further emerging intermediate minima were determined by means of the connectivity finder module of the MHGPS code as implemented in the BigDFT suite. A pathway given by the trajectories of the stabilized quasi-Newton minimizer (SQNM)[22] is shown in Fig. 9(a). This pathway consists of 27 intermediate transition states. Fig. 9(b) shows a lowest barrier pathway that was extracted from the stationary point database which was sampled by means of unbiased MHGPS runs and already used for the standard disconnectivity graphs in Fig. 8(a.3). The pathway in Fig. 9(b) consists of 20 intermediate transition states. Remarkably, both paths exhibit the same highest energy transition state which is highlighted by the red arrows in Fig. 9. Still, the path extracted

TABLE I. Rough sizes of the databases used for Fig. 8.

Database type	System	n[a]	m[b]
TBCD[c]	Si_{20}	7 000	5 000
	$(NaCl)_{29}$	82 000	71 000
	$(NaCl)_{32}$	28 000	25 000
	LJ_{19}	1 800	1 100
	LJ_{38}	87 000	64 000
	LJ_{55}	35 000	33 000
SPD[d]	Si_{20}	3 400	2 000
	$(NaCl)_{29}$	200 000	171 000
	$(NaCl)_{32}$	68 000	61 000
	LJ_{19}	65 000	14 000
	LJ_{38}	68 000	45 000
	LJ_{55}	59 000	49 000

[a]Number of minima.
[b]Number of hopping pairs in case of trajectory-based connectivity databases or number of transition states in case of stationary point databases.
[c]Trajectory-based connectivity database.
[d]Stationary point database.

from the stationary point database (Fig. 9(b)) is shorter than the path in Fig. 9(a), both in terms of the integrated path length and in terms of the number of intermediate transition states.

There is no guarantee that extracting a sequence of minima from a trajectory-based connectivity database and connecting these minima by searching intermediate transition states will result in a pathway that has the same highest barrier as the pathway with the lowest highest barrier that is contained in a thoroughly sampled stationary point database. However, computer experiments performed for the LJ_{38} cluster indicate that physically reasonable pathways can be extracted from trajectory-based connectivity databases. Using the modified Dijkstra's algorithm, a sequence of minima was extracted from the complete trajectory-based connectivity database for LJ_{38}. By successively removing the highest energy transition along the lowest barrier pathway from the trajectory-based connectivity database, this process was repeated four more

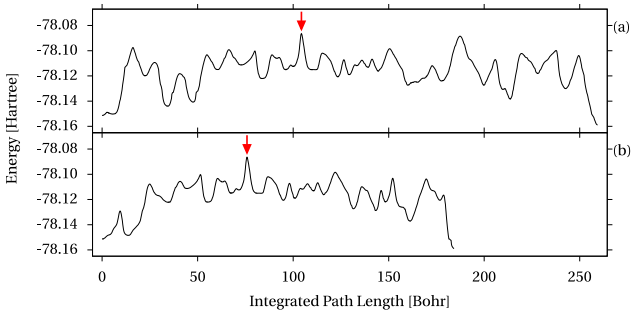

FIG. 9. Two energy minimized pathways connecting the two lowest minima of Si_{20} (DFT, PBE). The pathway in panel (a) was obtained by extracting a sequence of minima from the trajectory-based connectivity database and using this sequence of minima as input for the connectivity finder module of the MHGPS[21] code. Panel (b) shows a pathway that was extracted from a stationary point database sampled by entirely unbiased MHGPS runs. The shown pathways are SQNM[22] trajectories obtained by relaxations from the transition states after stepping away a small distance in both directions of the negative eigenmode. The transition states in the MHGPS runs were tightly converged by means of the stabilized quasi-Newton saddle (SQNS)[22] finding method. The red arrows indicate the highest energy transition states along the pathways. In both pathways, the highest energy transition states are identical.

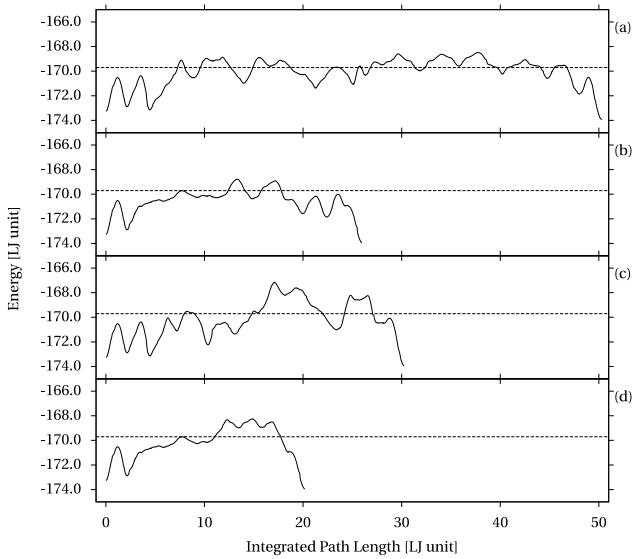

FIG. 10. Energy minimized pathways connecting the two lowest minima of LJ_{38}. The pathway in panel (a) was obtained by extracting a sequence of minima from the complete trajectory-based connectivity database. Panels (b)–(d) show pathways that were obtained by successively removing the highest energy transition along the lowest-barrier pathway from the trajectory-based connectivity database. Using the sequences of the extracted minima as input for the MHGPS[21] method, complete pathways were reconstructed. The SQNS[22] and SQNM[22] methods were used for converging to transition states and relaxing to the connected minima.

times. In this way, five different sequences of minima were obtained. Again, for each sequence, missing intermediate minima and transition states were added by means of the connectivity finder module of the MHGPS code.This procedure resulted in four pathways with non-identical highest barriers, which are shown in Fig. 10. The dashed line at an energy of −169.708 LJ units indicates the highest barrier along the lowest-known barrier pathway.[63,64] Considering the fact that, for instance, in the case of argon 1 LJ energy unit corresponds to roughly 10 meV,[65–67] one sees that the highest barriers along the pathways in Fig. 10 are not much higher than this lowest-known barrier.

IV. CONCLUSION

Based on Lennard-Jones, silicon, sodium-chloride, and gold clusters, it was found that uphill barrier energies of transition states between directly connected minima tend to increase with increasing structural differences of the two minima. Based on this insight it also turned out that post-processing MH data at a negligible computational cost allows us to obtain qualitative topological information on PESs that is stored in a so called trajectory-based connectivity database. The trajectory-based connectivity database can be used for generating fingerprint disconnectivity graphs that allow us to obtain a qualitative idea on thermodynamic and kinetic properties of a system of interest. Besides allowing us to assess system properties without the need of a computational expensive explicit sampling of transition states and the assessment of the PES's connectivity based on minimum energy or energy minimized pathways, this method also serves as a valuable tool that can help to decide whether a certain

multi-atomic system may exhibit desired properties before investing significant resources for assessing these properties more rigorously. Furthermore it was demonstrated that it is possible to extract from a trajectory-based connectivity database well aligned sequences of minima that can be used to generate initial pathways that are needed for methods like TPS or DPS.

SUPPLEMENTARY MATERIAL

See the supplementary material for the Gaussian kernel density estimates of the uphill barrier energies versus the *s*-orbital fingerprint distances and the fingerprint disconnectivity graphs based on *s*-orbitals, only.

ACKNOWLEDGMENTS

This research was supported by the NCCR MARVEL, funded by the Swiss National Science Foundation. Computer time was provided by the Swiss National Supercomputing Centre (CSCS) under Project ID No. s499.

APPENDIX: CONFIGURATIONAL FINGERPRINTS BASED ON GAUSSIAN TYPE ORBITALS

The recently introduced configurational fingerprints[34] that are used throughout this work are given by the eigenvalues of an overlap matrix

$$O_{ij} := \int \Phi_i^{\mathbf{l}}(\mathbf{r})\Phi_j^{\mathbf{l}'}(\mathbf{r})d\mathbf{r}. \tag{A1}$$

Here, the Φ_i are Gaussian type orbitals centered on the atom at position $\mathbf{r}_i$

$$\Phi_i^{\mathbf{l}}(\mathbf{r}) \propto (x-x_i)^{l_x}(y-y_i)^{l_y}(z-z_i)^{l_z}\exp(-\alpha_i\|\mathbf{r}-\mathbf{r}_i\|^2), \tag{A2}$$

where $\mathbf{l} = (l_x, l_y, l_z)$ is a multi-index indicating the angular momentum $L = l_x + l_y + l_z$. Depending on the value of L, the orbitals are classified as *s*-type orbitals ($L = 0$), *p*-type orbitals ($L = 1$), *d*-type orbitals ($L = 2$), and so on. The orbital width α_i is usually chosen inversely proportional to the square of the covalent radius of the atom species on which the orbital is centered on. The set of sorted overlap matrix eigenvalues for a given cluster can be considered to form a vector which defines the fingerprint of the configuration. The structural difference between two clusters is given by the root mean square of the difference vector between the two fingerprint vectors and throughout this work, this distance measure is denoted as "fingerprint distance."

[1]C. Levinthal, in *Mössbauer Spectroscopy in Biological Systems: Proceedings of a Meeting held at Allerton House, Monticello, Illinois*, edited by P. Debrunner, J. C. M. Tsibris, and E. Münck (University of Illinois Press, Urbana, 1969), pp. 22–24.
[2]K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997).
[3]D. Wales, *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses* (Cambridge University Press, Cambridge, 2003).
[4]C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, J. Chem. Phys. **108**, 1964 (1998).
[5]C. Dellago, P. G. Bolhuis, and D. Chandler, J. Chem. Phys. **108**, 9236 (1998).
[6]P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).
[7]C. Dellago, P. G. Bolhuis, and P. L. Geissler, "Transition path sampling," in *Advances in Chemical Physics* (John Wiley & Sons, Inc., 2003), pp. 1–78.
[8]M. Grünwald, C. Dellago, and P. L. Geissler, J. Chem. Phys. **129**, 194101 (2008).
[9]M. Grünwald and C. Dellago, Nano Lett. **9**, 2099 (2009).
[10]W. Lechner, C. Dellago, and P. G. Bolhuis, Phys. Rev. Lett. **106**, 085701 (2011).
[11]D. J. Wales, Mol. Phys. **100**, 3285 (2002).
[12]D. J. Wales, Mol. Phys. **102**, 891 (2004).
[13]X.-J. Zhang and Z.-P. Liu, Phys. Chem. Chem. Phys. **17**, 2757 (2015).
[14]N. Mousseau and G. T. Barkema, Phys. Rev. E **57**, 2419 (1998).
[15]G. Wei, N. Mousseau, and P. Derreumaux, J. Chem. Phys. **117**, 11379 (2002).
[16]E. Machado Charry, L. K. Béland, D. Caliste, L. Genovese, T. Deutsch, N. Mousseau, and P. Pochet, J. Chem. Phys. **135**, 34102 (2011).
[17]N. Mousseau, L. K. Béland, P. Brommer, J.-F. Joly, F. El Mellouhi, E. Machado Charry, M.-C. Marinica, and P. Pochet, J. At., Mol., Opt. Phys. **2012**, 925278.
[18]M. R. Sørensen and A. F. Voter, J. Chem. Phys. **112**, 9599 (2000).
[19]D. Perez, B. P. Uberuaga, Y. Shim, J. G. Amar, and A. F. Voter, "Accelerated Molecular Dynamics Methods: Introduction and Recent Developments," in *Annual Reports in Computational Chemistry* (Elsevier, 2009), Vol. 5, pp. 79–98.
[20]S. Goedecker, J. Chem. Phys. **120**, 9911 (2004).
[21]B. Schaefer, S. Mohr, M. Amsler, and S. Goedecker, J. Chem. Phys. **140**, 214102 (2014).
[22]B. Schaefer, S. A. Ghasemi, S. Roy, and S. Goedecker, J. Chem. Phys. **142**, 034112 (2015).
[23]M. Amsler and S. Goedecker, J. Chem. Phys. **133**, 224104 (2010).
[24]A. R. Oganov and C. W. Glass, J. Chem. Phys. **124**, 244704 (2006).
[25]C. W. Glass, A. R. Oganov, and N. Hansen, Comput. Phys. Commun. **175**, 713 (2006).
[26]D. J. Wales and J. P. K. Doye, J. Phys. Chem. A **101**, 5111 (1997).
[27]J. P. K. Doye and D. J. Wales, Phys. Rev. Lett. **80**, 1357 (1998).
[28]J. P. K. Doye, D. J. Wales, and M. A. Miller, J. Chem. Phys. **109**, 8143 (1998).
[29]C. J. Pickard and R. J. Needs, Phys. Rev. Lett. **97**, 045504 (2006).
[30]C. J. Pickard and R. J. Needs, Nat. Phys. **3**, 473 (2007).
[31]C. J. Pickard and R. J. Needs, Phys. Rev. B **76**, 144114 (2007).
[32]C. J. Pickard and R. J. Needs, Nat. Mater. **7**, 775 (2008).
[33]S. De, B. Schaefer, A. Sadeghi, M. Sicher, D. G. Kanhere, and S. Goedecker, Phys. Rev. Lett. **112**, 083401 (2014).
[34]A. Sadeghi, S. A. Ghasemi, B. Schaefer, S. Mohr, M. A. Lill, and S. Goedecker, J. Chem. Phys. **139**, 184118 (2013).
[35]O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).
[36]L. Genovese, A. Neelov, S. Goedecker, T. Deutsch, S. A. Ghasemi, A. Willand, D. Caliste, O. Zilberberg, M. Rayson, A. Bergman *et al.*, J. Chem. Phys. **129**, 014109 (2008).
[37]S. Mohr, L. E. Ratcliff, P. Boulanger, L. Genovese, D. Caliste, T. Deutsch, and S. Goedecker, J. Chem. Phys. **140**, 204110 (2014).
[38]A. Willand, Y. O. Kvashnin, L. Genovese, Á. Vázquez Mayagoitia, A. K. Deb, A. Sadeghi, T. Deutsch, and S. Goedecker, J. Chem. Phys. **138**, 104109 (2013).
[39]R. P. Bell, Proc. R. Soc. A **154**, 414 (1936).
[40]M. G. Evans and M. Polanyi, Trans. Faraday Soc. **32**, 1333 (1936).
[41]G. S. Hammond, J. Am. Chem. Soc. **77**, 334 (1955).
[42]R. A. Marcus, J. Phys. Chem. **72**, 891 (1968).
[43]F. Jensen, *Introduction to Computational Chemistry* (John Wiley & Sons, New York, 2007).
[44]S. Roy, S. Goedecker, and V. Hellmann, Phys. Rev. E **77**, 056707 (2008).
[45]M. Sicher, S. Mohr, and S. Goedecker, J. Chem. Phys. **134**, 044106 (2011).
[46]M. Born and J. Mayer, Z. Phys. **75**, 1 (1932).
[47]J. E. Mayer, J. Chem. Phys. **1**, 270 (1933).
[48]M. L. Huggins and J. E. Mayer, J. Chem. Phys. **1**, 643 (1933).
[49]F. Fumi and M. Tosi, J. Phys. Chem. Solids **25**, 31 (1964).
[50]M. Tosi and F. Fumi, J. Phys. Chem. Solids **25**, 45 (1964).
[51]J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[52]W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[53]R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, Oxford, 1994).
[54]D. J. Adams and I. R. McDonald, J. Phys. C: Solid State Phys. **8**, 2198 (1975).
[55]J. E. Jones, Proc. R. Soc. A **106**, 463 (1924).
[56]J. E. Jones and A. E. Ingham, Proc. R. Soc. A **107**, 636 (1925).
[57]B. Schaefer, R. Pal, N. S. Khetrapal, M. Amsler, A. Sadeghi, V. Blum, X. C. Zeng, S. Goedecker, and L.-S. Wang, ACS Nano **8**, 7413 (2014).

[58]K. Levenberg, Q. Appl. Math. **II**, 164 (1944).
[59]D. W. Marquardt, J. Soc. Ind. Appl. Math. **11**, 431 (1963).
[60]T. Williams and C. Kelley, and many others, Gnuplot 4.6: An Interactive Plotting Program, 2014, http://gnuplot.sourceforge.net.
[61]S. Zhang and N. Chen, Phys. B **325**, 172 (2003).
[62]C. Dellago, in *Free Energy Calculations*, Springer Series in Chemical Physics Vol. 86, edited by C. Chipot and A. Pohorille (Springer, Berlin, Heidelberg, 2007), pp. 249–276.
[63]J. P. K. Doye, M. A. Miller, and D. J. Wales, J. Chem. Phys. **110**, 6896 (1999).
[64]J. P. K. Doye, M. A. Miller, and D. J. Wales, J. Chem. Phys. **111**, 8417 (1999).
[65]A. Rahman, Phys. Rev. **136**, A405 (1964).
[66]L. Rowley, D. Nicholson, and N. Parsonage, J. Comput. Phys. **17**, 401 (1975).
[67]J. A. White, J. Chem. Phys. **111**, 9352 (1999).