Energy Landscapes and Persistent Minima

Joanne M. Carr,[1] Dorian Mazauric,[2] Frédéric Cazals,[2, a)] and David J. Wales[1, b)]

[1)] *University Chemical Laboratories, University of Cambridge, Lensfield Road,*
Cambridge CB2 1EW, UK

[2)] *Inria Sophia Antipolis Méditerranée, 2004 route des Lucioles,*
F-06902 Sophia Antipolis, France

(Dated: 9 February 2016)

We consider a coarse-graining of high-dimensional potential energy landscapes based upon persistences, which correspond to lowest barrier heights to lower-energy minima. Persistences can be calculated efficiently for local minima in kinetic transition networks that are based on stationary points of the prevailing energy landscape. The networks studied here represent peptides, proteins, nucleic acids, an atomic cluster, and a glassy system. Minima with high persistence values are likely to represent some form of alternative structural morphology, which, if appreciably populated at the prevailing temperature, could compete with the global minimum (defined as infinitely persistent). Threshold values on persistences (and in some cases equilibrium occupation probabilities) have therefore been used in this work to select subsets of minima, which were then analysed to see how well they can represent features of the full network. Simplified disconnectivity graphs showing only the selected minima can convey the funnelling (including any multiple-funnel) characteristics of the corresponding full graphs. The effect of the choice of persistence threshold on the reduced disconnectivity graphs was considered for a system with a hierarchical, glassy landscape. Sets of persistent minima were also found to be useful in comparing networks for the same system sampled under different conditions, using minimum oriented spanning forests.

[a)]Electronic mail: frederic.cazals@inria.fr
[b)]Electronic mail: dw34@cam.ac.uk

I. INTRODUCTION

Visualisation often plays a key role in our ability to understand physical phenomena. Unfortunately, the high dimensionality of a potential energy landscape, which is a function of $3N$ coordinates for N atoms, makes visualisation problematic for most systems of interest. Projections onto low-dimensional surfaces can be highly misleading. For example, even relatively good order parameters are unlikely to distinguish the two sides of every barrier in configuration space between all pairs of local minima. This projection error generally leads to underestimation of free energy barriers in reduced dimensionality,[1–8] as highlighted recently for the 38-atom cluster considered again in the present contribution.[9] In fact, since the number of transition states (saddles with Hessian index one) connected to each minimum is expected to scale linearly with N,[10] any three-dimensional representation of the landscape for a many-body system is likely to be misleading. Low-dimensional embeddings are typically based on some dimensionality-reduction method, such as principal component analysis, multi-dimensional scaling, isomap, or diffusion maps. In all such methods, the low-dimensional embedding is tightly coupled to the spectral properties of a matrix.[11] If the eigenspectrum of that matrix does not exhibit a clear gap (e.g. two or three large eigenvalues, the remaining ones being small), no low-dimensional embedding can accurately represent the corresponding landscape.

Although the true complexity of a many-dimensional potential energy landscape cannot be represented by a three-dimensional surface, there are useful alternatives. In particular, disconnectivity graphs can be constructed that provide a faithful account of the potential or free energy barriers between local minima.[12–15] Such graphs can be constructed automatically from databases of local minima and the transition states that connect them, and provide a useful picture of the underlying kinetic transition network.[6,8,16] Quantitative analysis of thermodynamic properties and kinetics requires additional data, such as normal mode frequencies, associated with entries in the database. Nevertheless, certain emergent properties can be inferred from the appearance of the disconnectivity graph itself. For example, competition of morphologies corresponding to low energy structures separated by a high barrier is likely to result in multiple relaxation time scales, and associated features in the heat capacity.[17–24] Such landscapes can be considered "frustrated",[25,26] and quantitative measures are being actively developed to describe the corresponding topography. Recent work on an

intrinsically disordered protein suggests that landscapes with multiple competing potential energy funnels might be associated with multi-functional systems.[27]

Since the number of local minima is expected to grow exponentially with the number of atoms[10,28] it can be useful to consider further coarse-graining of the landscape and the corresponding disconnectivity graph. In the present contribution we consider the notion of "persistence", defined from the viewpoint of computational topology, and how it can be used to pick out subsets of minima from a kinetic transition network. The selected minima are analysed to see how well they can represent features of the full network, and also used as "indicators" when comparing two or more networks sampled under different conditions for the same molecule. Lumping local potential energy minima together to define free energy minima is another, intuitive, way to achieve a coarse-graining. Here, we also combine a self-consistent grouping scheme defined on the basis of a free energy barrier threshold[29] with the persistence analysis. We find that selection of local minima using persistence can provide an efficient, consistent, and complementary coarse-graining approach for high-dimensional landscapes.

II. METHODS

Following Cazals et al.,[30] we analyse a number of kinetic transition networks (KTNs) using concepts from topography and computational topology. The KTNs were obtained in previous work[19,31–35] using pathway based approaches, including discrete path sampling,[19,20,36] and consist of databases of stationary points on the prevailing potential energy landscape: local minima and the transition states[37] (TSs) that connect them.

A. Persistence diagrams

The persistence of each local minimum (apart from the global minimum) is defined here as the lowest barrier height to a lower-energy minimum,[38] i.e. the absolute value of the energy difference between the minimum itself and the highest transition state on the lowest-energy path that connects a minimum of lower energy. Using terminology from topography, this particular TS is referred to as the *key saddle* for the original minimum. The persistence of the global minimum is infinite, since this particular minimum does not possess a key saddle.

Persistences can alternatively be referred to as prominences.[38]

In the current work, persistences are obtained via a superbasin analysis,[12] which may be seen as a flooding process to find the mergers between basins of attraction[39] associated with local minima, since these events occur at key saddles. A superbasin analysis is also employed in the construction of disconnectivity graphs[12,13] (DGs) but in that case, the analysis is usually performed at an appropriately chosen series of equally spaced energies, rather than at the energies of the transition states themselves, in order not to obscure the overall structure of the landscape.

More formally, the sub-level set associated with an energy value is the portion of the landscape found below that threshold. Imagine sweeping the landscape with a horizontal hyper-plane, by increasing the energy. Whenever the hyper-plane hits a new local minimum, say m_1, a connected component (CC) of the corresponding sub-level set is born; while initially restricted to the local minimum m_1 itself, upon raising the threshold, this CC expands to the basin of attraction of m_1. Moreover, once the hyper-plane reaches the energy of the key saddle linking m_1 and another local minimum, say m_2, the CC of m_1 merges with that of m_2. Computing persistences in the current context therefore requires the energies at which CCs associated with local minima are born and merge with those of lower-energy minima. Equivalently, one needs to maintain the connected components of a graph whose nodes are the local minima, and whose edges correspond to the transition states. To do so, one processes stationary points (local minima and transition states) in order of increasing energy. Upon processing a local minimum, one adds a vertex to the graph, since a CC is born. Upon processing a transition state, if the two directly-linked local minima are not already in the same CC, then their CCs are merged and this transition state becomes the key saddle for whichever is the higher in energy of the lowest-energy minima from the two CCs before merging. The maintenance of the CCs of this evolving graph is optimally accomplished via a Union-Find algorithm,[40] which essentially scales linearly as a function of the number of nodes and edges in the graph, or by resorting to a more general (Morse) homology calculation.[41,42] The Union-Find approach is sufficient to compute the persistence of local minima, but not stationary points of other indices.

A persistence diagram[38] can then be plotted with the energy of the minimum on the x axis and the energy of the key saddle on the y axis, so that all points lie above the diagonal $y = x$. The most persistent metastable minima correspond to the points that lie

furthest from the diagonal. Normalised cumulative histograms can be used to illustrate the distribution of persistences.

B. Selection of minima using persistence

To identify particularly significant stationary points, we select minima whose persistence and equilibrium occupation probability (at a chosen temperature) exceed given thresholds, and add the global minimum to this selection. The aim of the condition on the occupation probability is to ensure that the chosen persistent minima are thermodynamically relevant at the prevailing temperature. For computational speed, the equilibrium occupation probabilities (P_{eq}) are estimated here using the harmonic approximation to the local density of vibrational states.[43,44]

Once the most significant minima have been selected, a reduced DG visualisation of the landscape can be employed: the relevant connected component of the original network is used in the superbasin analysis, but only the selected minima may be displayed in the graph. One of the principal objectives of the present contribution is to compare these reduced graphs with the originals, to see if they provide a helpful visualisation. This aim is similar in spirit to a coarse-graining based upon minima that lie at the bottom of basins defined by monotonic sequences,[45] which has previously been considered for atomic clusters.[46] Disconnectivity graphs based upon both persistent minima and monotonic sequence basins can be obtained using our `disconnectionDPS` program.[47]

Persistence can also be incorporated in our recursive scheme[29] to define free energy minima (and associated free energy transition states) as groups of minima in potential energy that can interconvert without encountering a barrier higher than a chosen threshold value. In an extension to the original method, minima selected using the criteria described above can be required to belong to separate groups, i.e. no two persistent minima can belong to the same group.

These analyses only include the stationary points in the single connected component of the network that addresses conformations of interest (usually defined by the global minimum). The KTNs may contain other separate "islands" as a consequence of the targeted sampling approaches used, or by exclusion of unphysical connecting TSs, but these are not considered here, precisely because they are disconnected and likely to be under-sampled. Stationary

points may also be excluded on the basis of thresholds for a maximum energy, a minimum number of direct connections, or (for TSs) a maximum barrier height.

C.   Comparing sets of minima via minimum oriented spanning forests

Two sets of structures, for example the minima from two KTNs constructed for the same molecule using different potential energy functions, may be compared using a minimum oriented spanning forest[30] (MSF), as described in Figure 1. Two values are then obtained: the average of the nearest-neighbour least root-mean-square distance (lRMSD) values from set 1 to set 2 ($\overline{w}_{1\to2}^{MSF}$) and the average from set 2 to set 1 ($\overline{w}_{2\to1}^{MSF}$). The sets can also be simplified prior to the MSF analysis using a threshold on the persistence. In the current work, we again only include minima in the single connected component of the prevailing network that contains the global minimum, as persistences for local minima in other components are likely to be unreliable due to under-sampling of their connections.

III.   RESULTS AND DISCUSSION

In the present contribution, the methods described in Section II are applied to KTNs for peptides, proteins, nucleic acids, an atomic cluster, and a glassy system, for illustration. In the results presented below, energies are in $kcal\,mol^{-1}$ and k_BT in the calculation of equilibrium occupation probabilities was set to 0.592 (corresponding to a temperature of $298\,K$) for the atomistic molecular mechanics force fields. Reduced units are employed for the Lennard-Jones potential[48] and the BLN protein bead model.[49–52] As a general rule (unless otherwise stated), the threshold on the persistence was chosen as the 90th percentile from the corresponding cumulative histogram, and the threshold on P_{eq} was set to 0.0001. In the case of dialanine, since a clear gap is visible in both of the persistence diagrams, only the large values were retained.

A.   Alanine Peptides

It was previously found[35] that for both dialanine and tetraalanine in vacuo, KTNs representing the CHARMM27[53] and AMBER99SB[54] landscapes exhibit significant differences: the global minima differ structurally, the heat capacity curves vary, and there are many

more local minima supported on the CHARMM27 landscapes than for AMBER99SB. In the present work, the four KTNs are analysed further using topological persistence. The aim is to see whether subsets of minima from one network provide a good representation of minima from the other force field, in terms of structural measures (lRMSD) and thermodynamic quantities (heat capacities).

For each molecule, the two sets of local minima are referred to as A (AMBER99SB) and C (CHARMM27). A set of minima obtained by simplifying A by selecting minima using a persistence threshold p_a is denoted $A \searrow$; the CHARMM27 equivalents are C, p_c and $C \searrow$.

The correlation between A and persistent CHARMM27 minima was studied via two comparisons:

- (1) $A \to C \searrow$. The average nearest-neighbour lRMSD from AMBER99SB minima to CHARMM27 minima was obtained, upon selecting a subset of CHARMM27 minima using persistence thresholds. On varying p_c, a subset of CHARMM27 minima that provides a good representation of AMBER99SB minima is evidenced by a plateau in the curve $\overline{w}_{A \to C}^{MSF}(p_c)$. That is, the minima corresponding to the persistence range defining such a plateau suffice to maintain the approximation quality of the AMBER99SB minima. This comparison is repeated for several thresholds p_a for AMBER99SB.

- (2) $C \searrow \to A$. Here the average nearest-neighbour lRMSD from CHARMM27 minima to AMBER99SB minima was obtained, upon selecting CHARMM27 minima based on their persistence. The aim is to discover whether selected persistent CHARMM27 minima have a preferential location with respect to AMBER99SB minima. Upon varying p_c, such a set is characterised by a plateau or a decreasing section in the curve $\overline{w}_{C \to A}^{MSF}(p_c)$. This comparison was also repeated for several thresholds p_a for AMBER99SB.

We sought the presence of the features described above at common values of p_c for (1) and (2), to position AMBER99SB minima with respect to persistent CHARMM27 minima and vice versa. Similarly, the correlation between C and persistent AMBER99SB minima was studied via two comparisons: (3) $A \searrow \to C$, the complementary situation to (2); and (4) $C \to A \searrow$, the complementary situation to (1).

1. *Dialanine*

Figure 2 presents the persistence diagrams (top) and the normalised cumulative histograms of persistences (bottom) for AMBER99SB (left) and CHARMM27 (right). Thresholds of 1.8 for AMBER99SB and 4.1 for CHARMM27 clearly separate around 10 persistent minima in the main connected component for each of the force fields.

From the MSF analysis, average lRMSD values were obtained as a function of p_a and p_c (Figure 3):

- (1) $A \rightarrow C \searrow$ (top-left). By construction the average distance $\overline{w}_{A \rightarrow C}^{MSF}$ increases: the nearest neighbour of each local minimum from AMBER99SB moves away upon selecting persistent minima from C.

  Persistent AMBER99SB minima (curve for $p_a = 1.8$, 10 minima) are better represented by CHARMM27 minima than all AMBER99SB minima (curve for $p_a = 0$, 23 minima).

  In varying p_c, four regions of interest are identified. For low values of p_c, a steep increase is observed in moving from $p_c = 0$ (181 minima) to $p_c = 0.2$ (71 minima), showing that the 110 CHARMM27 minima removed are important to represent the AMBER99SB minima properly, despite their low persistence. Similarly, five other such CHARMM27 minima are singled out in moving from $p_c = 0.6$ (27 minima) to $p_c = 0.8$ (22 minima). From $p_c = 0.8$ (22 minima) to $p_c = 1.6$ (14 minima), the quality of the coverage of AMBER99SB minima by CHARMM27 minima hardly changes, showing that the 14 minima suffice to represent the AMBER99SB minima. From $p_c = 1.6$ (14 minima) to $p_c = 2.6$ (9 minima), the curves rise again, showing that the 5 CHARMM27 minima removed are significant in representing the AMBER99SB minima.

- (2) $C \searrow \rightarrow A$ (top-right). Three regions may be distinguished in the plot, i.e. $p_c \leq 0.3$, $0.3 < p_c \leq 1.6$, and $1.6 < p_c$. The thresholds 0.3 and 1.6 correspond to 55 minima and 14 minima, respectively.

  To assist this analysis, consider the Voronoi diagram of the AMBER99SB minima for the lRMSD, that is, the partition of the conformational space with one cell per AMBER99SB local minimum, such that all conformations nearest to this minimum are assigned to its cell. The first region, which corresponds to a rapid decrease of the number of CHARMM27 minima, shows that these non-persistent minima do not

have any clear location in the Voronoi cells. In the second region, the decreasing trend means that upon selecting persistent minima from CHARMM27, one preferably removes peripheral nodes from the Voronoi cells. In the third region, the increasing trend means that the top nine most persistent CHARMM27 minima are not represented well, relative to less persistent minima, by AMBER99SB minima. More precisely, all CHARMM27 minima and the most persistent CHARMM27 minima exhibit a comparable average distance to their respective AMBER99SB nearest neighbours (circa 1.09 Å).

- (3) $A \searrow \rightarrow C$ (bottom-left). On selecting persistent AMBER99SB minima and assigning them nearest neighbours in CHARMM27 sets (all minima i.e. $p_c = 0$, or persistent minima i.e. $p_c = 4.1$), no clear trend is observed. This result shows that the AMBER99SB minima do not have a clear preference for being either close to or far from the CHARMM27 minima.

- (4) $C \rightarrow A \searrow$ (bottom-right). Similarly to (1), persistent CHARMM27 minima ($p_c = 4.1$, black curve) are better represented by AMBER99SB minima, than all CHARMM27 minima ($p_c = 0$, purple curve). Furthermore, while the curve for all minima rises regularly, the curve for persistent minima exhibits two regions, namely up to $p_a = 1.4$ (11 minima), and beyond. In the first part, the moderate rise shows that the 11 AMBER99SB minima provide a good approximation of the persistent CHARMM27 minima and, to a lesser extent, all CHARMM27 minima. In the second part, upon moving from $p_a = 1.4$ (11 minima) to $p_a = 1.8$ (10 minima), the sharp rise shows that the AMBER99SB minimum removed plays a key role in representing two persistent CHARMM27 minima.

From comparisons (1) and (2), the plateau ending at $p_c \sim 1.6$ in (1) and the minimum observed in the second plot at $p_c \sim 1.6$ in (2) show that a subset of CHARMM27 minima providing a good representation of AMBER99SB minima can be obtained for persistences in the range $p_c \in (0.8, 1.6)$. Comparisons (3) and (4) show that a core set of 11 AMBER99SB minima provides a good approximation to the persistent CHARMM27 minima.

Heat capacities, within the harmonic superposition approximation,[43,44] for the CHARMM27 landscape are presented in the top left panel of Figure 4 and those for the AMBER99SB landscape in the top right panel. The full lines were calculated using all the minima in the

single component connected to the prevailing global minimum, and the dashed lines were obtained using only the most persistent minima in this component ($p_c \geq 4.1$ or $p_a \geq 1.8$, see Figure 2). To compare heat capacity curves for the same potential as the number of minima included is reduced (i.e. the persistence threshold is increased), the function is integrated between fixed temperatures that bracket the peak for the largest set of minima. Differences between the resulting internal energy change (ΔU) for a given persistence threshold and the reference for the largest set of minima, are presented for both potentials in the lower panel of Figure 4. These results suggest that subsets of persistent minima for CHARMM27 are insufficient to represent the full component, in terms of heat capacities, but sets of AM-BER99SB minima identifiable by their relatively high persistence can provide a reasonable representation.

2. *Tetraalanine*

The persistence diagrams for AMBER99SB and CHARMM27 (Figure 5, top) exhibit significant differences, in terms of the number of minima, the range of persistences and the clustering in energy of the minima themselves. The corresponding normalised cumulative histograms of persistences are given in Figure 5 (bottom). In the absence of a meaningful separation of scales in these persistence diagrams, in studying the relative positions of the minima of the two force fields using MSFs, we vary the persistence thresholds as follows: $p_a \in [0, 12]$, and $p_c \in [0, 20]$.

From the MSF analysis (Figure 6):

- (1) $A \rightarrow C \searrow$ (top-left). On selecting CHARMM27 minima, no plateaux are observed, suggesting that these subsets of CHARMM27 minima do not provide a good approximation of AMBER99SB minima (or any subset).

- (2) $C \searrow \rightarrow A$ (top-right). The flatness of the curves shows that no subsets of CHARMM27 minima have a preferential location with respect to the AMBER99SB minima.

- (3) $A \searrow \rightarrow C$ (bottom-left). A situation analogous to (2).

- (4) $C \rightarrow A \searrow$ (bottom-right). A situation analogous to (1).

Heat capacities, within the harmonic superposition approximation, for the CHARMM27 landscape are presented in the top left panel of Figure 7 and those for the AMBER99SB landscape in the top right panel. The full lines were calculated using all the minima in the single component that includes the global potential energy minimum, and the dashed lines were obtained using only a set of persistent minima in this component ($p_c, p_a \geq 1.0$). Differences between the resulting internal energy change for a given persistence threshold and the reference for the largest set of minima, for the two highest-temperature peaks, are presented for both potentials in the lower panels of Figure 7. The results suggest that subsets of AMBER99SB minima can provide a reasonable representation of the full set for the low and high temperature peaks, but capture the middle peak less well; for CHARMM27, there is more variation in the quality of the representation as the persistence threshold changes.

B. The LJ$_{38}$ Cluster

Here we consider a network[19] for a cluster of 38 Lennard-Jones particles[48] (LJ$_{38}$) containing 5310 minima and 7610 valid transition states in the connected component of interest. The persistence diagram and the distribution of persistences are plotted in Figure 8. Disconnectivity graphs (original and reduced) are presented in Figure 9 for the region of the landscape below an energy threshold of -169, in order to avoid clutter from insignificant higher-lying minima. Minima were selected using a threshold of 1.64 on the persistences, and $k_{\mathrm{B}}T = 0.12$ was used to calculate the P_{eq}, since this corresponds to the solid-solid transition temperature.[17,55] Application of the selection criteria picks out five out of the seven lowest-energy minima (excluding the global minimum). The two minima lower in energy but not selected lie higher up the individual "funnel" that has structure c (labelled in Figure 9) at the bottom, and therefore have persistences below the threshold. The global minimum is a face-centred-cubic truncated octahedron,[56] and the other selected minima are based on incomplete icosahedra, with the five-fold symmetry being somewhat distorted in structures a and b (labelled in Figure 9), according to the geometric Common Neighbour Analysis.[57–59] Thus, the persistence analysis has highlighted minima from the two competing structural

morphologies of this archetypal double-funnel landscape.[17]

The recursive grouping scheme including persistence (Section II) was tested on this KTN; for a persistence threshold of 1.64, and for barrier thresholds where both methods are successful, the results are identical to those for the original algorithm (data not shown). That is to say, the minima identified by other means as both persistent and thermodynamically relevant are also distinguished from one another via the recursive grouping algorithm.[29] The recursive grouping scheme is appropriate when local equilibration occurs more rapidly within groups than between them, i.e. there is a separation of timescales and free energy barrier heights. Persistent minima are therefore likely to belong to different groups when the effects of vibrational and configurational entropy on free energy at the prevailing temperature either enhance or do not detract significantly from the high potential energy barriers.

C. BLJ_{60}

Next we consider a system with a glassy landscape: a 60-atom binary Lennard-Jones mixture of 48 type A and 12 type B particles (BLJ_{60}) at number density 1.3 (in reduced units) and $k_BT = 0.71$, with periodic boundary conditions.[31] There are 11474 minima and 13051 transition states in the connected component of interest. The persistence diagram and the distribution of persistences are plotted in Figure 10. The distribution here has a long tail from a relatively small number of minima with very high persistences, different from the LJ_{38} cluster, which also has a smaller range of persistences. Original and reduced disconnectivity graphs are presented in Figure 11 (top and middle). The P_{eq} were calculated at $k_BT = 0.71$ and a threshold of 2.868 was applied on the persistences. Here the hierarchical nature of the landscape, arising from a range of barrier heights, is preserved in the reduced disconnectivity graph, but with the simplification that many of the local minima higher up in individual "funnels" are not selected.

In previous work,[31] features referred to as metabasins[60–63] were visualised via a disconnectivity graph in which only minima connected by particular mechanisms (non-cage-breaking transitions or cage-breaking transitions that include reversals of previous cage-breaking events) were plotted. Removing the high-barrier productive cage-breaking transitions causes the graph to fragment, as emphasised by the use of colour. Figure 11 from Ref. 31 is recreated here in Figure 11 (bottom), with a slightly different colour scheme, and the minima

also selected according to persistence and P_{eq} identified. As one might expect, most of the selected minima lie at or near the bottoms of metabasins. Five of the 74 selected minima are not plotted in Figure 11 because they are connected by productive cage-breaks. Some significant metabasins do not contain any of the selected minima. For example, the local minimum marked with an asterisk in Figure 11 (top and bottom) lies in a metabasin that is not represented in the reduced DG, Figure 11 (middle), due to a sufficiently high persistence but low P_{eq}, arising from a relatively high geometric mean normal mode (vibrational) frequency.

D. BLN$_{69}$

The largest network[33] we consider here is for a bead model of a protein. BLN$_{69}$ is a 69-bead protein where the beads are either hydrophobic (B), hydrophilic (L) or neutral (N).[49–52] The particular sequence was designed to fold into a six-strand β-barrel with a frustrated landscape.[64] There are 163871 minima and 221319 valid transition states in the connected component of interest. The persistence diagram and the distribution of persistences are plotted in Figure 12; the distribution shows a long tail with a few minima at high values. Disconnectivity graphs (original and reduced) are presented in Figure 13 for the region of the landscape below an energy threshold of -80, in order to avoid clutter from insignificant higher-lying minima. Persistent minima were selected using a threshold of 4.05 on the persistence, and $k_B T = 0.6$ was used to calculate the P_{eq}. The reduced disconnectivity graph clearly shows the distinct funnels leading to structures that differ from the global minimum in the arrangement of the β-strands, reflecting the frustration present at the bottom of the original landscape. Minima that are higher in energy are also selected.

E. Binding modes for influenza virus

This network[32] describes interconversion pathways between two binding modes of $\alpha 2, 6$ sialic acid to hemagglutinin glycoprotein of human H1N1 influenza virus, modelled using the AMBERff99SB molecular mechanics force field[54,65] and a generalised Born implicit solvent model.[66] It is the smallest KTN considered here, with 1200 minima and 1995 valid transition states in the connected component of interest, and it also exhibits a small range of persis-

tences (see Figure 14). Dead-end minima (i.e. minima with fewer than two connections via valid TSs to valid minima) were removed from the analyses. Disconnectivity graphs (original and reduced) are presented in Figure 15 for the region of the landscape below an energy threshold of -6598, in order to avoid clutter from insignificant higher-lying minima. Persistent minima were selected using a threshold of 2.17 on the persistence. The selected minima are spread throughout the lower half of this funnelling landscape, with many of the longer dangling branches picked out; the funnelling nature is simplified and preserved in the reduced DG.

F.   UUCG RNA tetraloop

This KTN is for a UUCG tetraloop,[34] a 10-nucleotide RNA hairpin modelled using an all-atom molecular mechanics force field (AMBER99/bsc0,[67] employing the latest torsional corrections[68]) and a generalised Born implicit solvent model.[66] There are 58291 minima and 72459 valid transition states in the connected component of interest. The persistence diagram and the distribution of persistences are plotted in Figure 16. Disconnectivity graphs (original and reduced) are presented in Figure 17 for the region of the landscape below an energy threshold of -2426. The selection criterion on the persistence was chosen as the 80th percentile from the cumulative histogram, rather than the 90th, because in this case the criterion on P_{eq} deselects many minima with high persistences. The chosen value is 3.12. The landscape here is steeply funnelled, but with a significant side-funnel, which constitutes a kinetic trap and includes structures with different loop conformations, but the same Watson–Crick base pairing as the global minimum. The persistence analysis has picked out the side funnel (though not via its lowest minimum, which has a larger geometric mean normal mode frequency and hence a smaller P_{eq}) and a set of structures at relatively low energy in the funnel leading to the global minimum.

G.   Effect of varying the persistence threshold

The BLJ_{60} landscape is used to illustrate the effects of changing the threshold on the persistence used, along with the usual threshold on the occupation probability, to select minima. A set of disconnectivity graphs is presented in Figure 18 as the threshold is increased

from zero. These graphs show, as expected, an increasingly coarse-grained picture due to the progressive loss of internal structure within the funnels of this hierarchical landscape.

IV. CONCLUSIONS

Persistences, or lowest barrier heights to lower-energy minima, were calculated for local minima in various kinetic transition networks that are based on stationary points of the potential energy landscape. The calculation is a computationally inexpensive post-processing analysis once the more time-consuming step, sampling the landscape, has been performed. Minima with high persistence values are likely to represent some form of alternative structural morphology which, if appreciably populated at the prevailing temperature, could compete with the global minimum (defined with infinite persistence). Threshold values on persistences and, in some cases, equilibrium occupation probabilities (within the harmonic approximation) have therefore been used in this work to select subsets of minima, which were then analysed to see how well they can represent the full network.

Disconnectivity graphs showing only the selected minima were simplified yet still convey the funnelling (including multiple-funnel) characteristics of the corresponding full graphs, though not necessarily via the structure of lowest potential energy in the funnel, due to the thermodynamic criterion. The effect of the choice of persistence threshold on the reduced disconnectivity graphs was analysed for a system with a hierarchical, glassy landscape (BLJ_{60}). Sets of selected minima were also used to compare different KTNs, representing energy landscapes for two molecular mechanics force fields, AMBER99SB and CHARMM27, for the same molecule in vacuum, which were found in previous work[35] to exhibit significant global differences. Using minimum oriented spanning forests, sets of persistent minima were found to be well correlated between the two landscapes for dialanine, but not for tetraalanine. Features of heat capacity curves were captured reasonably well using only a selected subset of minima for the dialanine and tetraalanine AMBER99SB landscapes, but less well for both of the CHARMM27 landscapes. In future work, it will be interesting to assess the degree of correlation between energy landscapes for more complex biochemical systems in solution.

The comparison of minima using spanning forests and selection of persistent minima might be useful in developing coarse-grained potentials based upon atomistic force fields.

Since the spanning-forests analysis only exploits local minima and distances between them, more elaborate comparisons can also be used, in particular those based on earth mover distances[30], which also employ occupation probabilities of local minima and possibly transition paths.

The reduction in complexity afforded by the selection of local minima on the basis of their persistence and equilibrium occupation probability could equally be achieved using free energies at the prevailing temperature, for groups of rapidly interconverting minima and ensembles of transition states that connect minima in different groups.[15,29] Hence, disconnectivity graphs plotted on the basis of free energy[14,15] could also be reduced as demonstrated here for potential energy.

REFERENCES

[1] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).

[2] S. V. Krivov and M. Karplus, Proc. Nat. Acad. Sci. USA **101**, 14766 (2004).

[3] S. V. Krivov and M. Karplus, J. Phys. Chem. B **110**, 12689 (2006).

[4] S. Muff and A. Caflisch, Proteins: Struct., Func. and Bioinf. **70**, 1185 (2008).

[5] S. V. Krivov and M. Karplus, Proc. Nat. Acad. Sci. USA **105**, 13841 (2008).

[6] F. Noé and S. Fischer, Curr. Op. Struct. Biol. **18**, 154 (2008).

[7] B. M. Dickson, D. E. Makarov, and G. Henkelman, J. Chem. Phys. **131**, 074108 (2009).

[8] D. J. Wales, Curr. Op. Struct. Biol. **20**, 3 (2010).

[9] D. J. Wales, J. Chem. Phys. **142**, 130901 (2015).

[10] D. J. Wales and J. P. K. Doye, J. Chem. Phys. **119**, 12409 (2003).

[11] J. Lee and M. Verleysen, *Nonlinear dimensionality reduction* (Springer Verlag, 2007).

[12] O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).

[13] D. J. Wales, M. A. Miller, and T. R. Walsh, Nature **394**, 758 (1998).

[14] S. V. Krivov and M. Karplus, J. Chem. Phys. **117**, 10894 (2002).

[15] D. A. Evans and D. J. Wales, J. Chem. Phys. **118**, 3891 (2003).

[16] D. Prada-Gracia, J. Gómez-Gardenes, P. Echenique, and F. Fernando, PLoS Comput. Biol. **5**, e1000415 (2009).

[17] J. P. K. Doye, M. A. Miller, and D. J. Wales, J. Chem. Phys. **110**, 6896 (1999).

[18] J. P. Neirotti, F. Calvo, D. L. Freeman, and J. D. Doll, J. Chem. Phys. **112**, 10340 (2000).

[19] D. J. Wales, Mol. Phys. **100**, 3285 (2002).

[20] D. J. Wales, Mol. Phys. **102**, 891 (2004).

[21] M. Picciani, M. Athenes, J. Kurchan, and J. Tailleur, J. Chem. Phys. **135**, 034108 (2011).

[22] V. A. Sharapov and V. A. Mandelshtam, J. Phys. Chem. A **111**, 10284 (2007).

[23] F. Calvo, Phys. Rev. E **82**, 046703 (2010).

[24] D. J. Wales, Chem. Phys. Lett. **584**, 1 (2013).

[25] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).

[26] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).

[27] Y. Chebaro, A. J. Ballard, D. Chakraborty, and D. J. Wales, Sci. Rep. **5**, 10386 (2015).

[28] F. H. Stillinger and T. A. Weber, Science **225**, 983 (1984).

[29] J. M. Carr and D. J. Wales, J. Phys. Chem. B **112**, 8760 (2008).

[30] F. Cazals, T. Dreyfus, D. Mazauric, C.-A. Roth, and C. H. Robert, J. Comp. Chem. **36**, 1213 (2015).

[31] V. K. de Souza and D. J. Wales, J. Chem. Phys. **129**, 164507 (2008).

[32] C. Shang, C. S. Whittleston, K. H. Sutherland-Cash, and D. J. Wales, J. Chem. Theor. Comput. **11**, 2307 (2015).

[33] M. T. Oakley, D. J. Wales, and R. L. Johnston, J. Phys. Chem. B **115**, 11525 (2011).

[34] D. Chakraborty, R. Collepardo-Guevara, and D. J. Wales, J. Am. Chem. Soc. **136**, 18052 (2014).

[35] S. Somani and D. J. Wales, J. Chem. Phys. **139**, 121909 (2013).

[36] D. J. Wales, Int. Rev. Phys. Chem. **25**, 237 (2006).

[37] J. N. Murrell and K. J. Laidler, Trans. Faraday. Soc. **64**, 371 (1968).

[38] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba, J. ACM **60**, 41:1 (2013).

[39] P. G. Mezey, Theo. Chim. Acta **58**, 309 (1981).

[40] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to algorithms* (MIT press, 2009 (3rd edition)).

[41] A. Banyaga and D. Hurtubise, *Lectures on Morse Homology* (Kluwer, 2004).

[42] F. Cazals and D. Cohen-Steiner, Computational Geometry Theory and Applications **45**, 1 (2011).

[43] D. J. Wales, Mol. Phys. **78**, 151 (1993).

[44] D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).

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

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

[47] D. J. Wales, "disconnectionDPS: A program for creating disconnectivity graphs," http://www-wales.ch.cam.ac.uk/software.html (2015).

[48] J. E. Jones and A. E. Ingham, Proc. R. Soc. A **107**, 636 (1925).

[49] J. D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. USA **87**, 3526 (1990).

[50] J. D. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).

[51] Z. Y. Guo and D. Thirumalai, Biopolymers **36**, 83 (1995).

[52] Y. Guo and D. L. Thompson, J. Chem. Phys. **104**, 576 (1996).

[53] X. Zhu, P. E. M. Lopes, and A. D. MacKerell, Wiley Interdisciplinary Reviews: Computational Molecular Science **2**, 167 (2012).

[54] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, Proteins: Structure, Function, and Bioinformatics **65**, 712 (2006).

[55] J. P. K. Doye and D. J. Wales, Phys. Rev. Lett. **80**, 1357 (1998).

[56] J. P. K. Doye, D. J. Wales, and R. S. Berry, J. Chem. Phys. **103**, 4234 (1995).

[57] J. D. Honeycutt and H. C. Andersen, J. Phys. Chem. **91**, 4950 (1987).

[58] D. Faken and H. Jónsson, Computational Materials Science **2**, 279 (1994).

[59] C. L. Cleveland, W. D. Luedtke, and U. Landman, Phys. Rev. B **60**, 5065 (1999).

[60] F. H. Stillinger, Phys. Rev. B **41**, 2409 (1990).

[61] F. H. Stillinger, Science **267**, 1935 (1995).

[62] B. Doliwa and A. Heuer, Phys. Rev. E **67**, 030501 (2003).

[63] B. Doliwa and A. Heuer, J. Phys. Cond. Matt. **15**, S849 (2003).

[64] S. A. Larrass, L. M. Pegram, H. L. Gordon, and S. M. Rothstein, J. Chem. Phys. **119**, 13149 (2003).

[65] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kellman, J. Am. Chem. Soc. **117**, 5179 (1995).

[66] A. Onufriev, D. Bashford, and D. A. Case, Proteins: Structure, Function, and Bioinformatics **55**, 383 (2004).

[67] A. Pérez, I. Marchán, D. Svozil, J. Šponer, T. E. Cheatham, C. A. Laughton, and M. Orozco, Biophys. J. **92**, 3817 (2007).

[68] M. Zgarbová, M. Otyepka, J. Šponer, A. Mládek, P. Banáš, T. E. Cheatham, and P. Jurečka, J. Chem. Theory Comput. **7**, 2886 (2011).
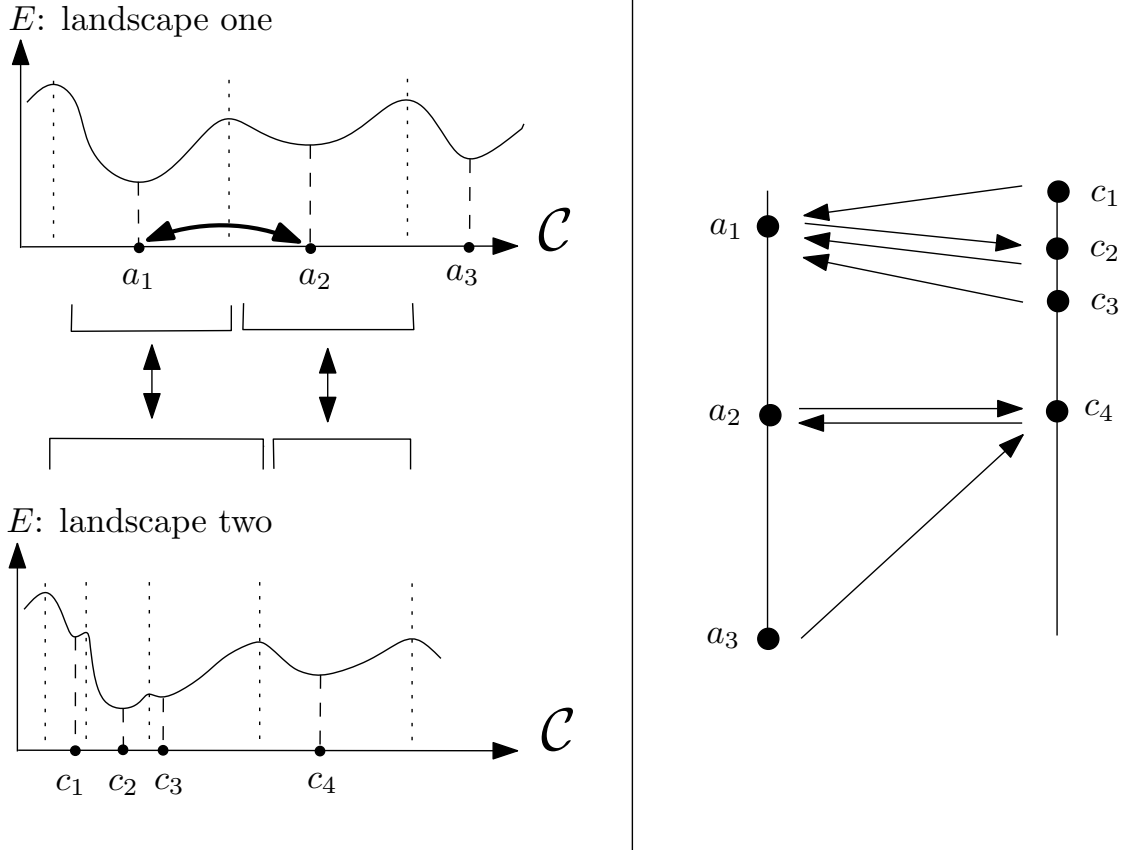
FIG. 1. Comparing sets of minima using a minimum oriented spanning forest (MSF). Given two sets of local minima, associated with two landscapes, each local minimum chooses its nearest neighbor in the other set. The resulting distance (e.g. the least root mean square deviation) defines the weight of the corresponding oriented edge connecting the two minima. In this fictitious example, every minimum from the second landscape has a close neighbor amongst minima from the first landscape, but the converse is false: minimum a_3 is (relatively) isolated. Due to this lack of symmetry, comparisons are performed both ways. The average weight of edges from the first landscape to the second one is denoted $\overline{w}_{1\to 2}^{MSF}$, and the average weight in the reverse direction is denoted $\overline{w}_{2\to 1}^{MSF}$.
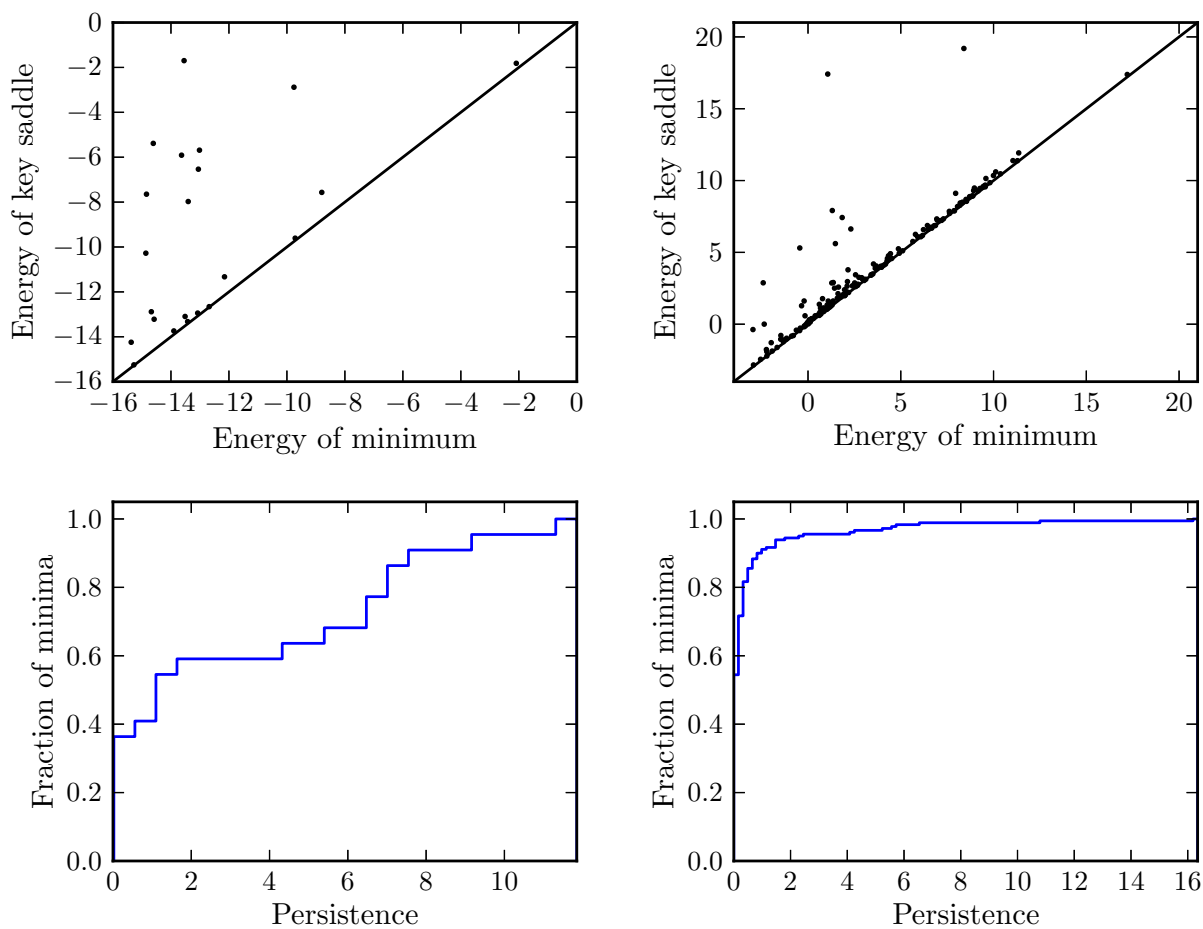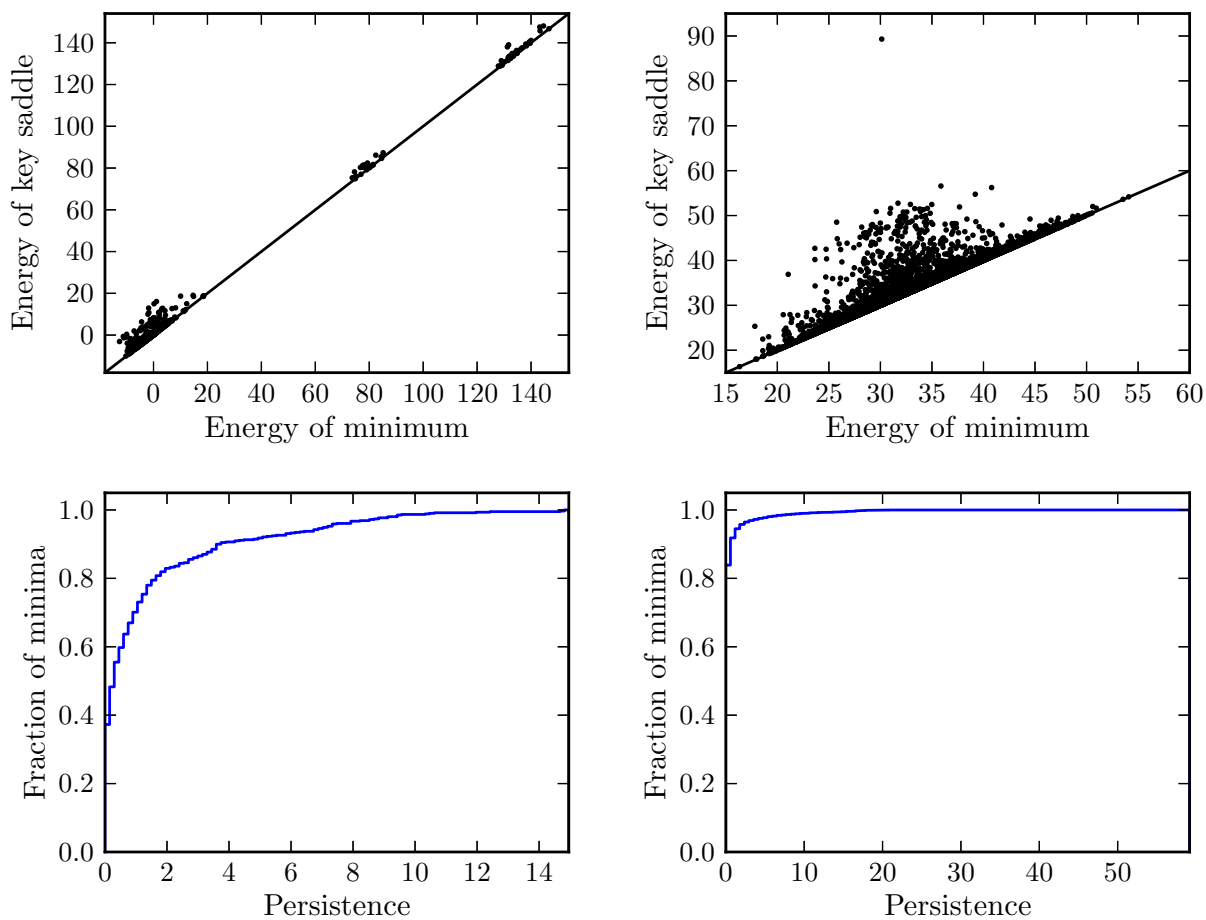
FIG. 2. Dialanine: persistence diagram and normalised cumulative histogram of persistences for AMBER99SB (left column) and CHARMM27 (right column).
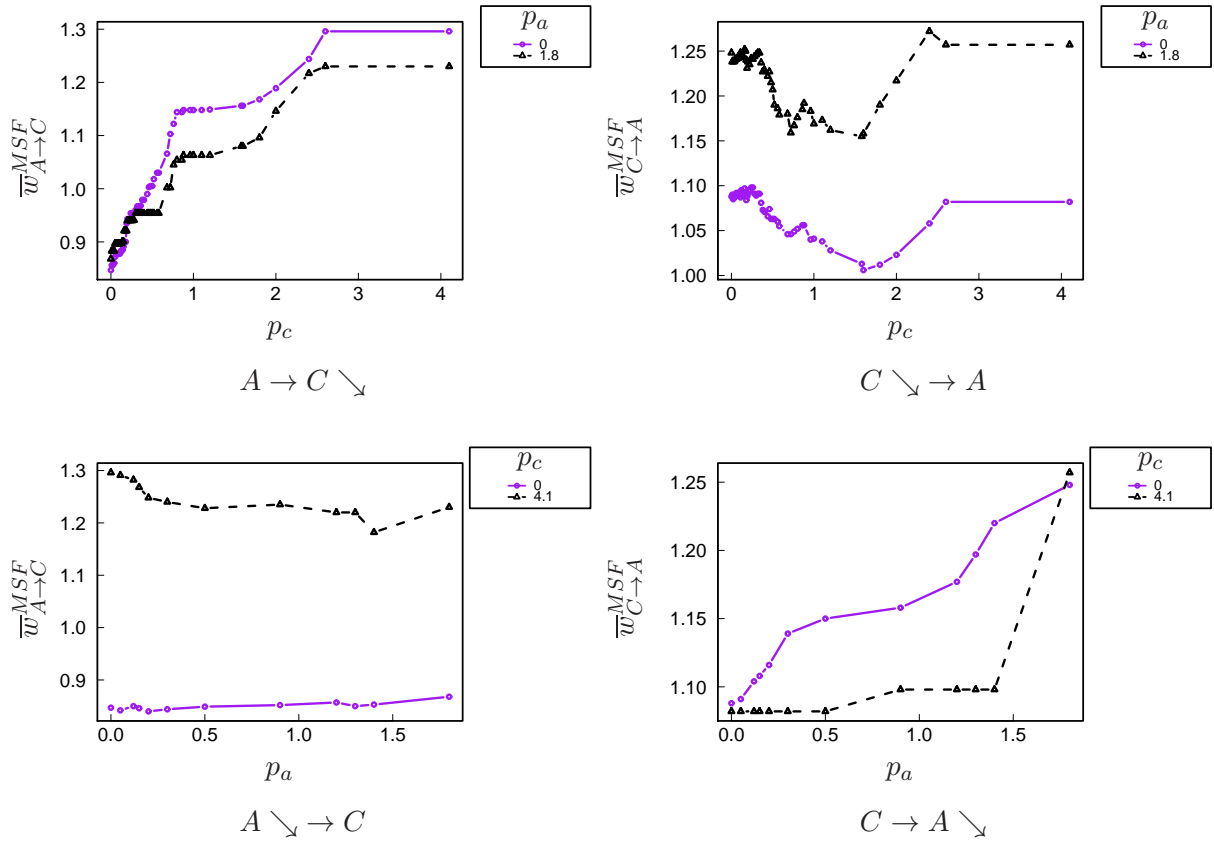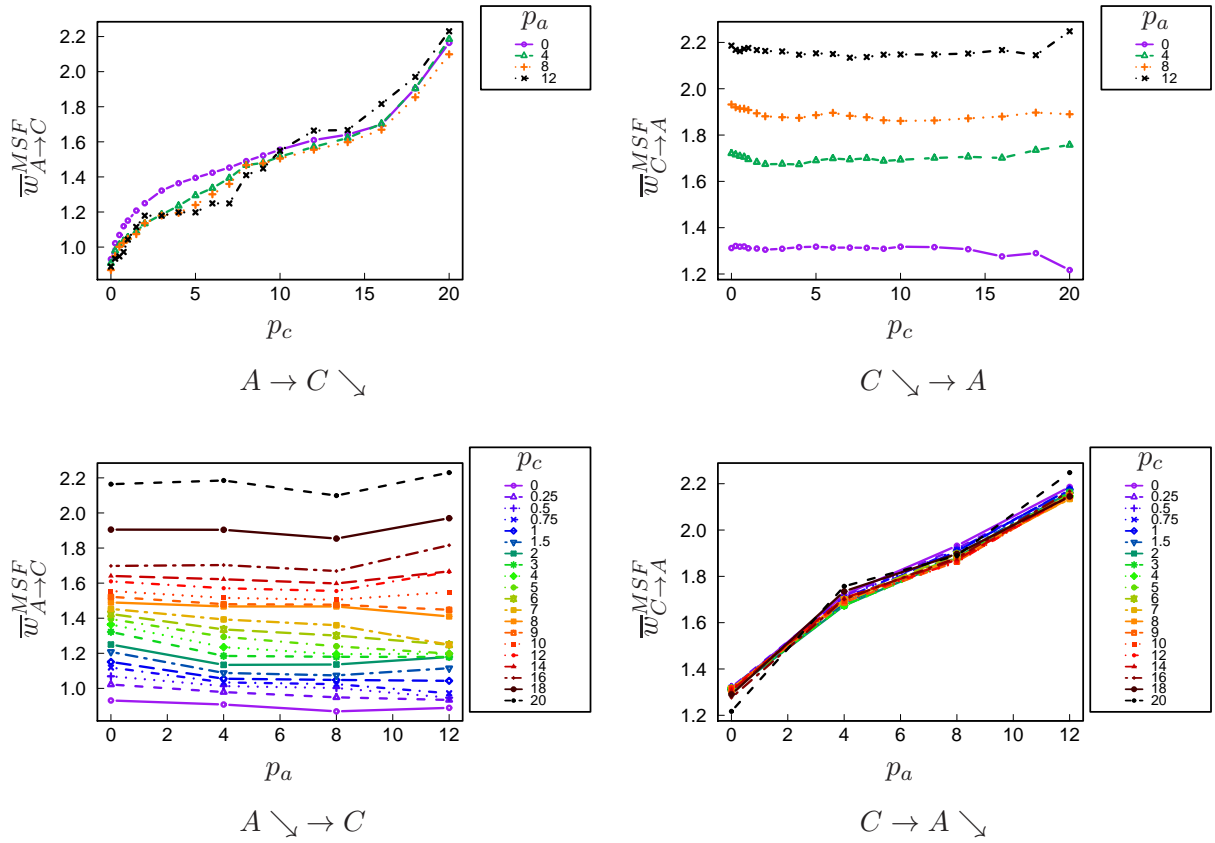
FIG. 3. Dialanine: comparing the set of (persistent) local minima using minimum oriented spanning forests. p_a and p_c are persistence thresholds for AMBER99SB and CHARMM27, respectively; $\overline{w}_{A \to C}^{MSF}$ and $\overline{w}_{C \to A}^{MSF}$ are average MSF weights.
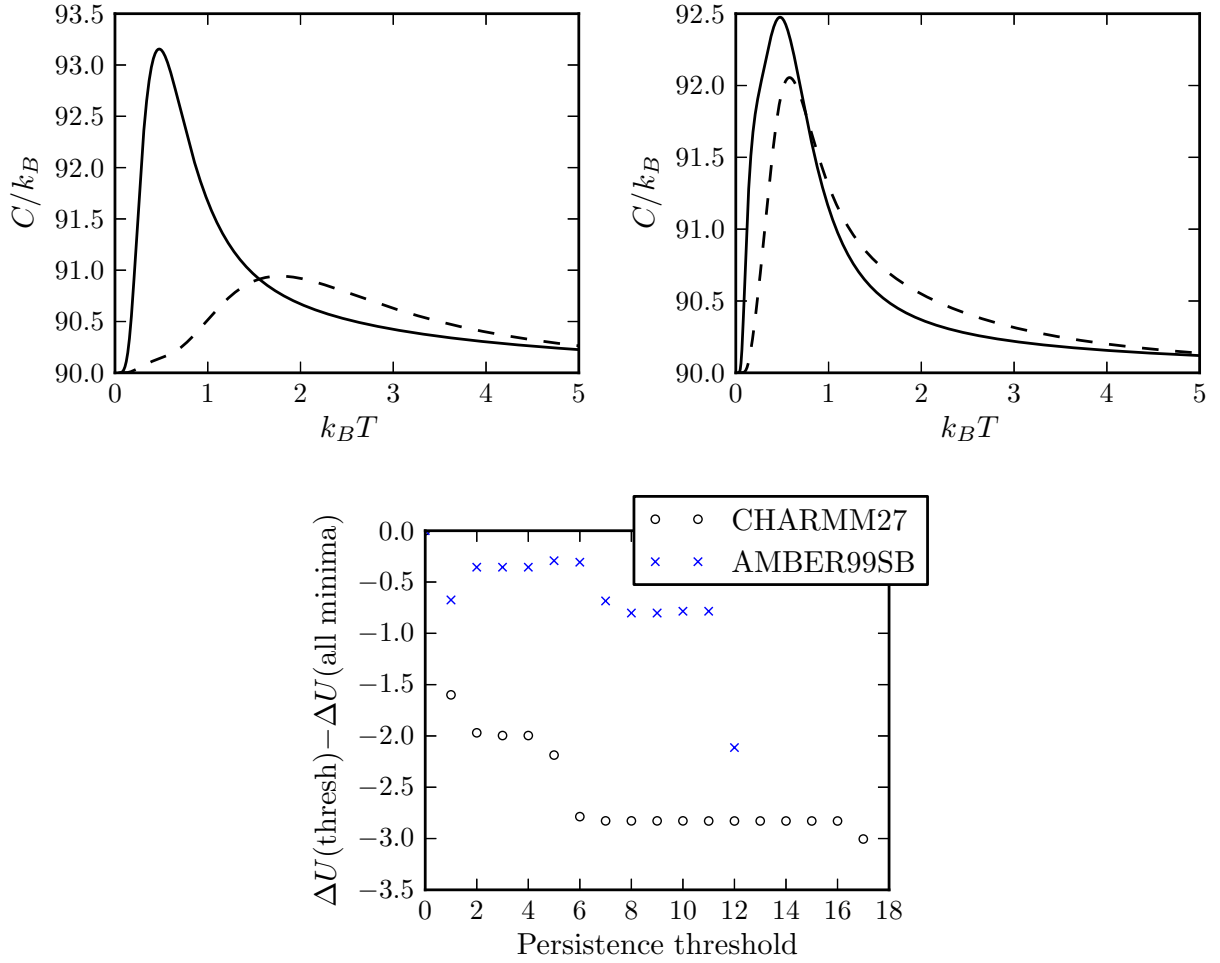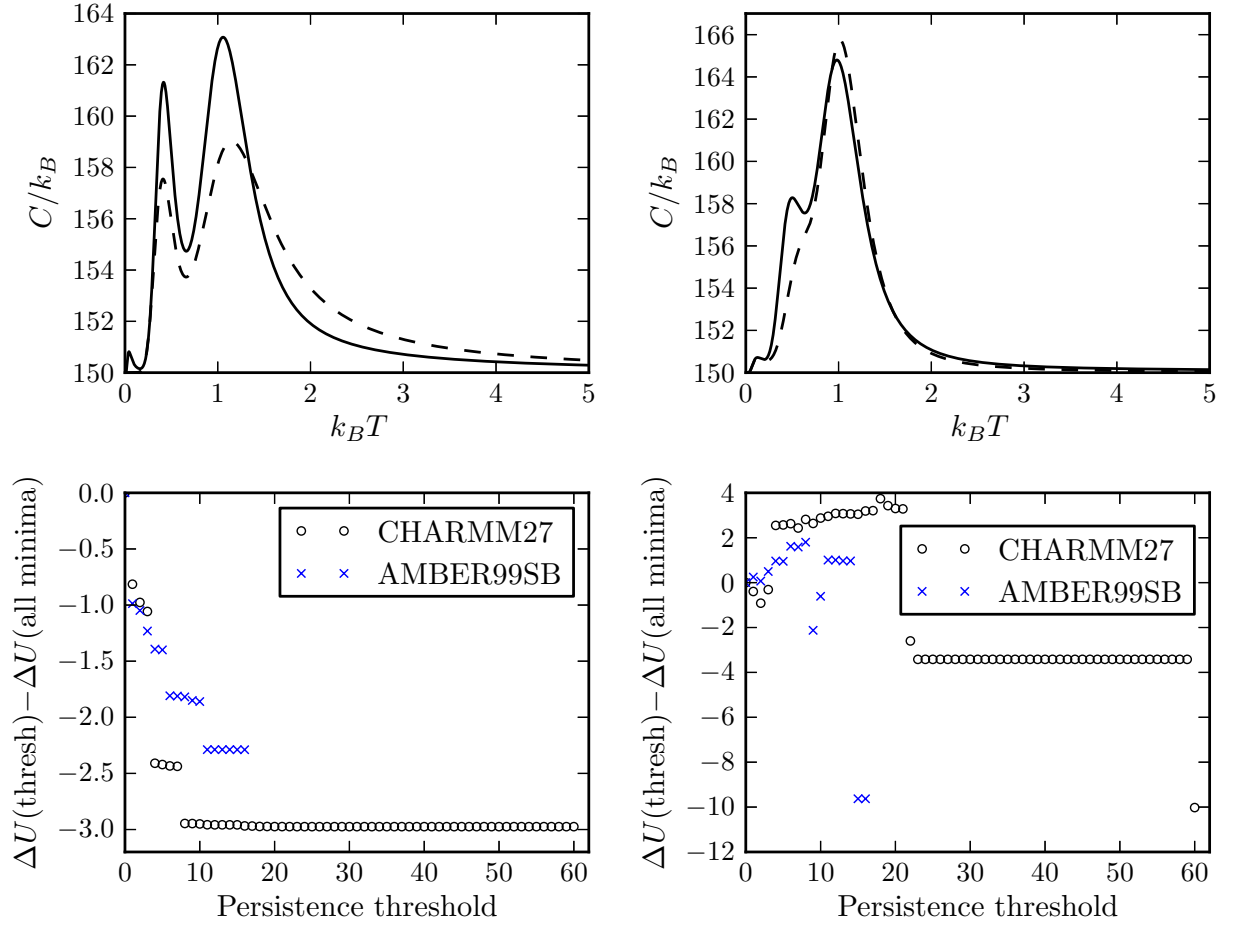
FIG. 4. Dialanine: heat capacity curves from the harmonic superposition approximation using CHARMM27 (top left) and AMBER99SB (top right). The full lines were calculated using all the minima in the single component connected to the prevailing global minimum, and the dashed lines, using only the minima in this component whose persistence exceeds 4.1 for CHARMM27 or 1.8 for AMBER99SB. Lower panel: internal energy changes (ΔU), plotted as the difference from the reference energy change using all the minima in the relevant connected component, as the persistence threshold is increased. The bracketing temperatures are [0.1,2.0] for CHARMM27 and [0.1,1.5] for AMBER99SB, in units of $k_\mathrm{B}T$.

23

FIG. 5. Tetraalanine: persistence diagram and normalised cumulative histogram of persistences for AMBER99SB (left column) and CHARMM27 (right column).

FIG. 6. Tetraalanine: comparing the set of (persistent) local minima using minimum oriented spanning forests. p_a and p_c are persistence thresholds for AMBER99SB and CHARMM27, respectively; $\overline{w}_{A \to C}^{MSF}$ and $\overline{w}_{C \to A}^{MSF}$ are average MSF weights.

FIG. 7. Tetraalanine: heat capacity curves from the harmonic superposition approximation using CHARMM27 (top left) and AMBER99SB (top right). The full lines were calculated using all the minima in the single component connected to the prevailing global minimum, and the dashed lines, using only the minima in this component whose persistence exceeds 1.0. Lower panels: internal energy changes (ΔU) for the two highest-temperature peaks, plotted as the difference from the reference energy change using all the minima in the relevant connected component, as the persistence threshold is increased. In units of k_BT, the bracketing temperatures are peak 2 (left): [0.16,0.66] for CHARMM27 and [0.2,0.63] for AMBER99SB; peak 3 (right): [0.66,2.5] for CHARMM27 and [0.63,2.0] for AMBER99SB.
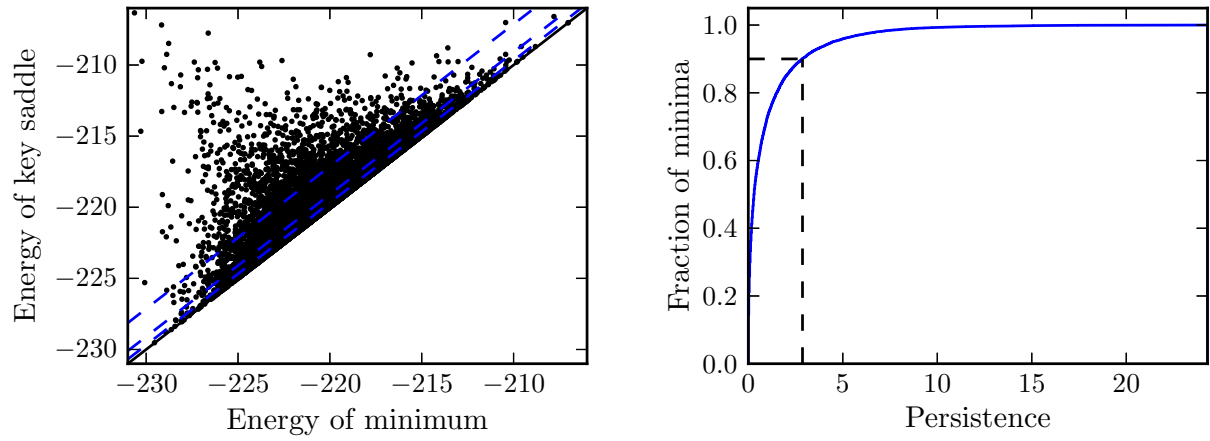
26
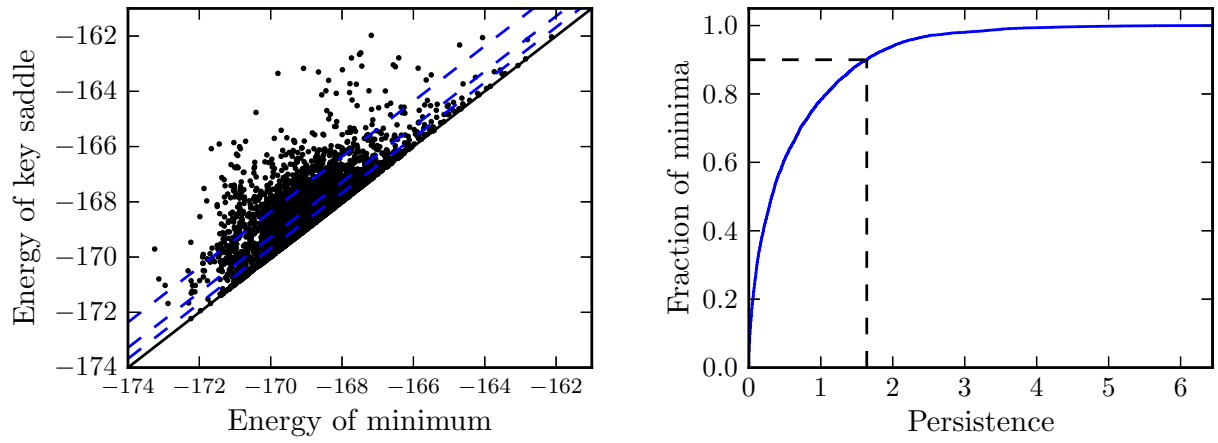
FIG. 8. LJ$_{38}$: persistence diagram (left). The dashed lines mark the persistences corresponding to the 50th, 70th and 90th percentiles. Normalised cumulative histogram of persistences, with the 90th percentile marked (right).
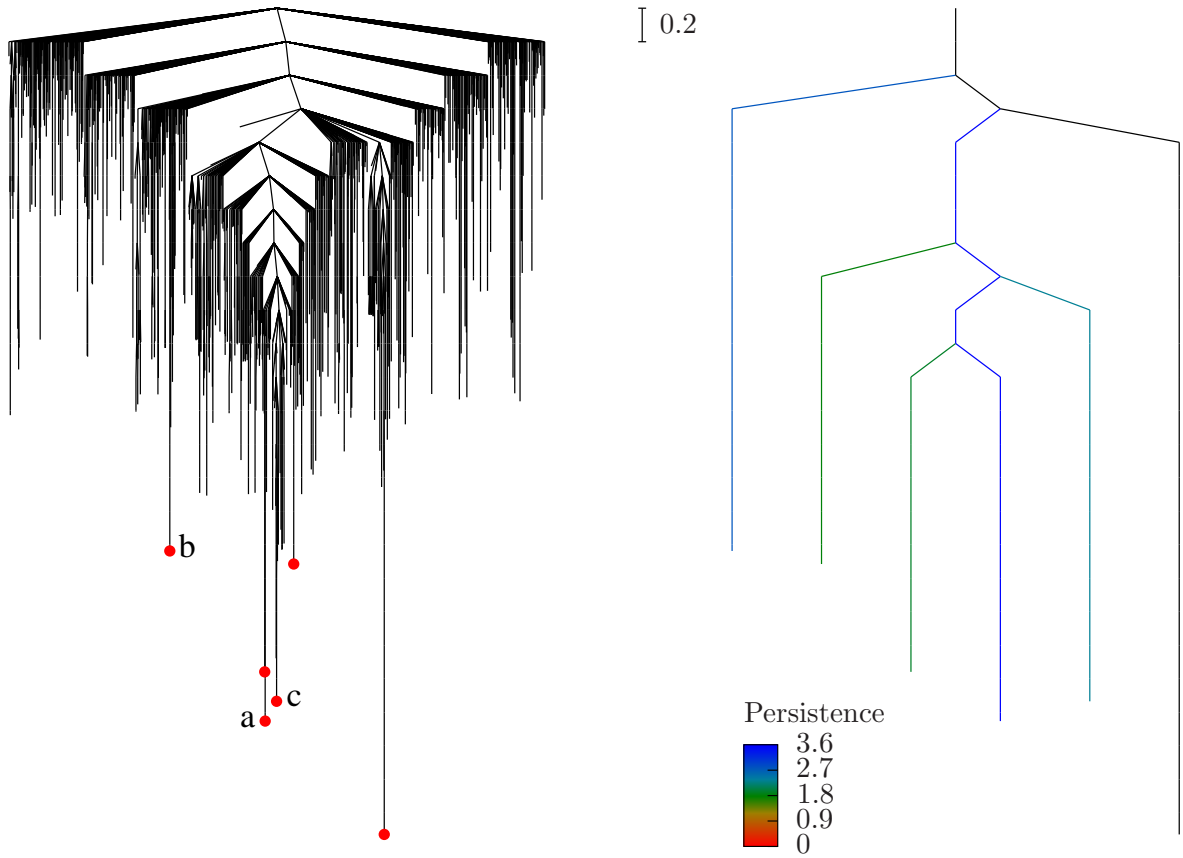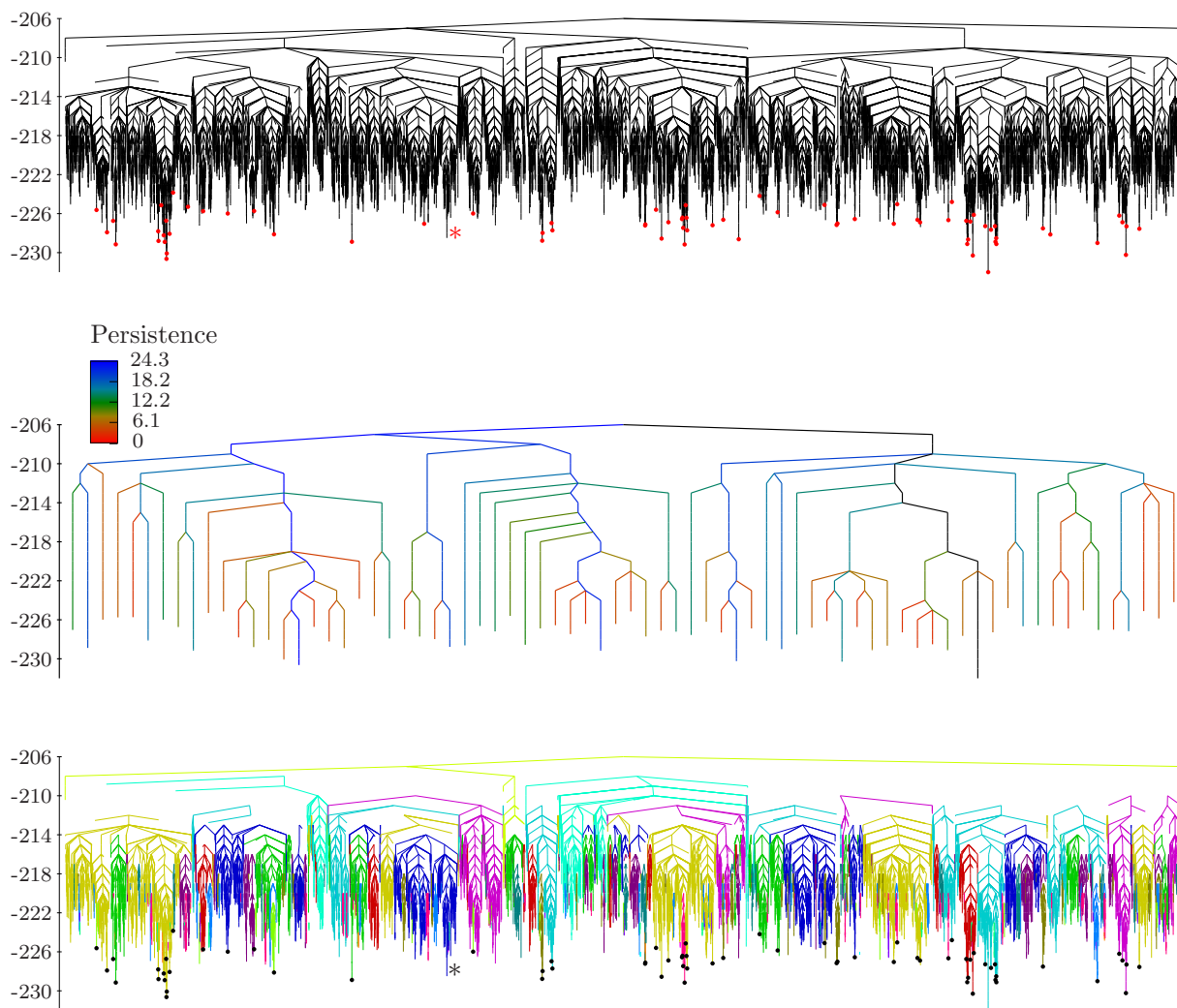
FIG. 9. LJ$_{38}$: disconnectivity graphs. The vertical scale, denoted by the scale bar, is potential energy. Left: original graph. The red circles mark the minima that meet the selection criteria described in the text. Right: reduced graph, showing only those minima that meet the selection criteria. The lines are coloured according to the persistence of the corresponding local minimum; the line to the global minimum, defined to have infinite persistence, is black.

FIG. 10. BLJ$_{60}$: persistence diagram (left). The dashed lines mark the persistences corresponding to the 50th, 70th and 90th percentiles. Normalised cumulative histogram of persistences, with the 90th percentile marked (right).

FIG. 11. BLJ$_{60}$: disconnectivity graphs. The vertical axes are potential energy. Top: original graph. The red circles mark the minima that meet the selection criteria described in the text. The red asterisk marks a metabasin that is not represented in the reduced graph. Middle: reduced graph, showing only those minima that meet the selection criteria. The lines are coloured according to the persistence of the corresponding local minimum; the line to the global minimum, defined to have infinite persistence, is black. Bottom: fragmented graph, showing only the minima connected by particular types of mechanism. This graph employs a different colouring scheme: here, changes in colour show isolated subsets of minima. The black circles and asterisk have the same meaning as the red versions in the top panel.
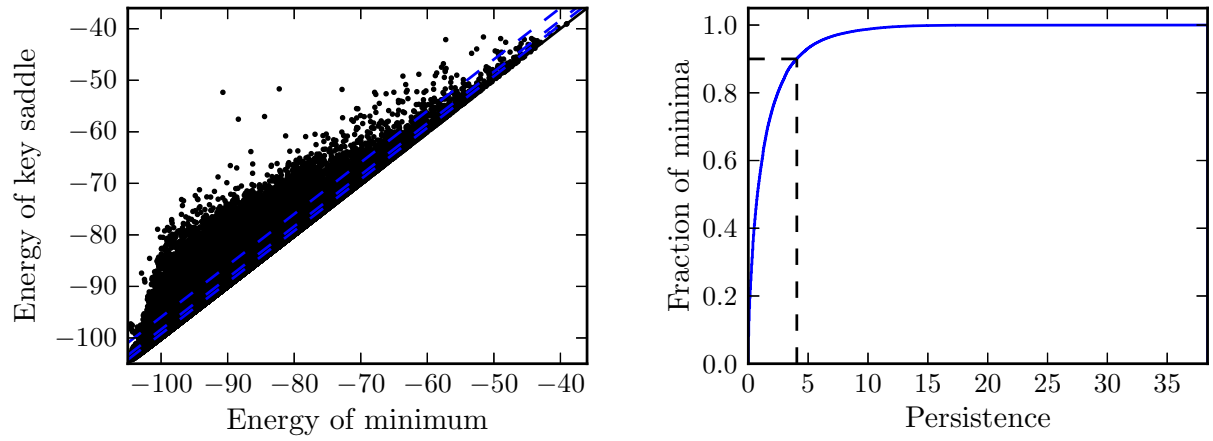
FIG. 12. BLN$_{69}$: persistence diagram (left). The dashed lines mark the persistences corresponding to the 50th, 70th and 90th percentiles. Normalised cumulative histogram of persistences, with the 90th percentile marked (right).
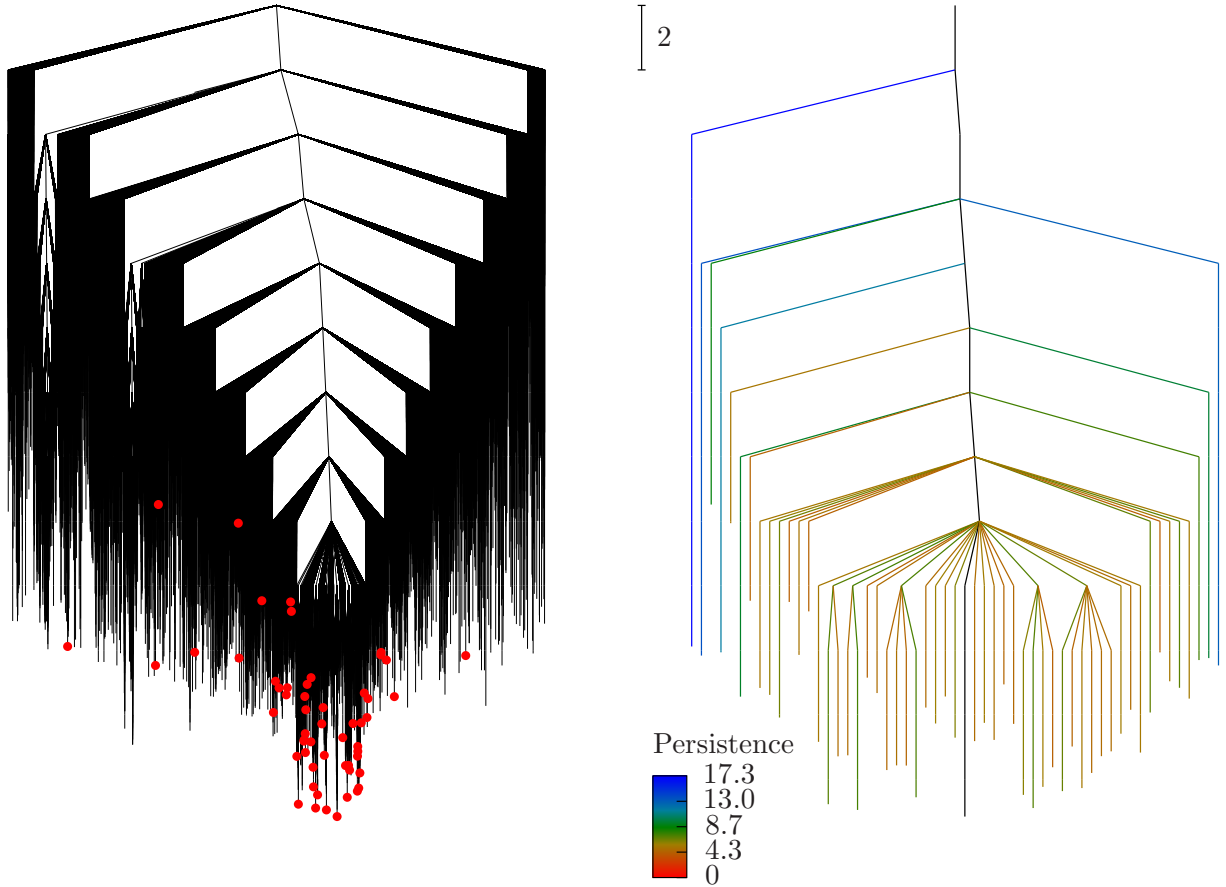
FIG. 13. BLN$_{69}$: disconnectivity graphs. The vertical scale, denoted by the scale bar, is potential energy. Left: original graph. The red circles mark the minima that meet the selection criteria described in the text. Right: reduced graph, showing only those minima that meet the selection criteria. The lines are coloured according to the persistence of the corresponding local minimum; the line to the global minimum, defined to have infinite persistence, is black.
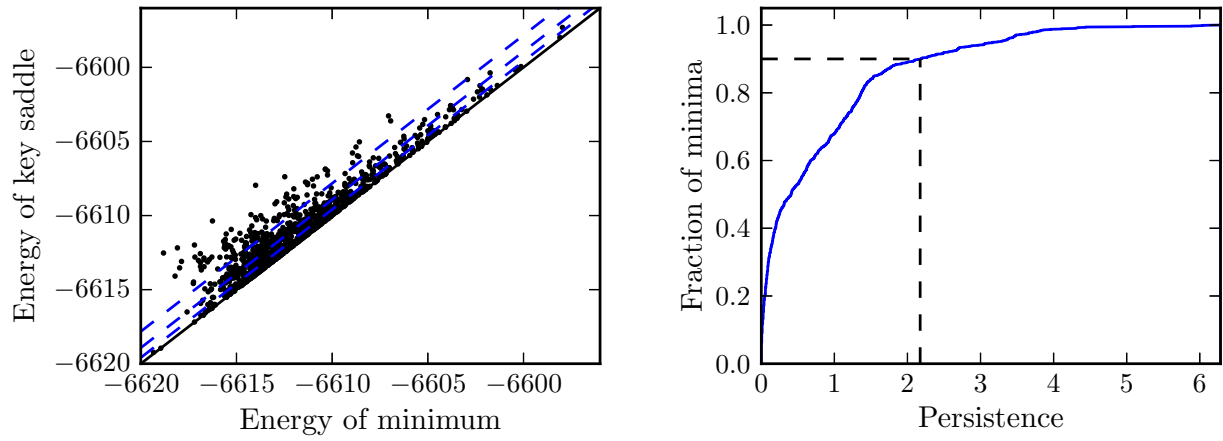
32

FIG. 14. Influenza virus binding: persistence diagram (left). The dashed lines mark the persistences corresponding to the 50th, 70th and 90th percentiles. Normalised cumulative histogram of persistences, with the 90th percentile marked (right).
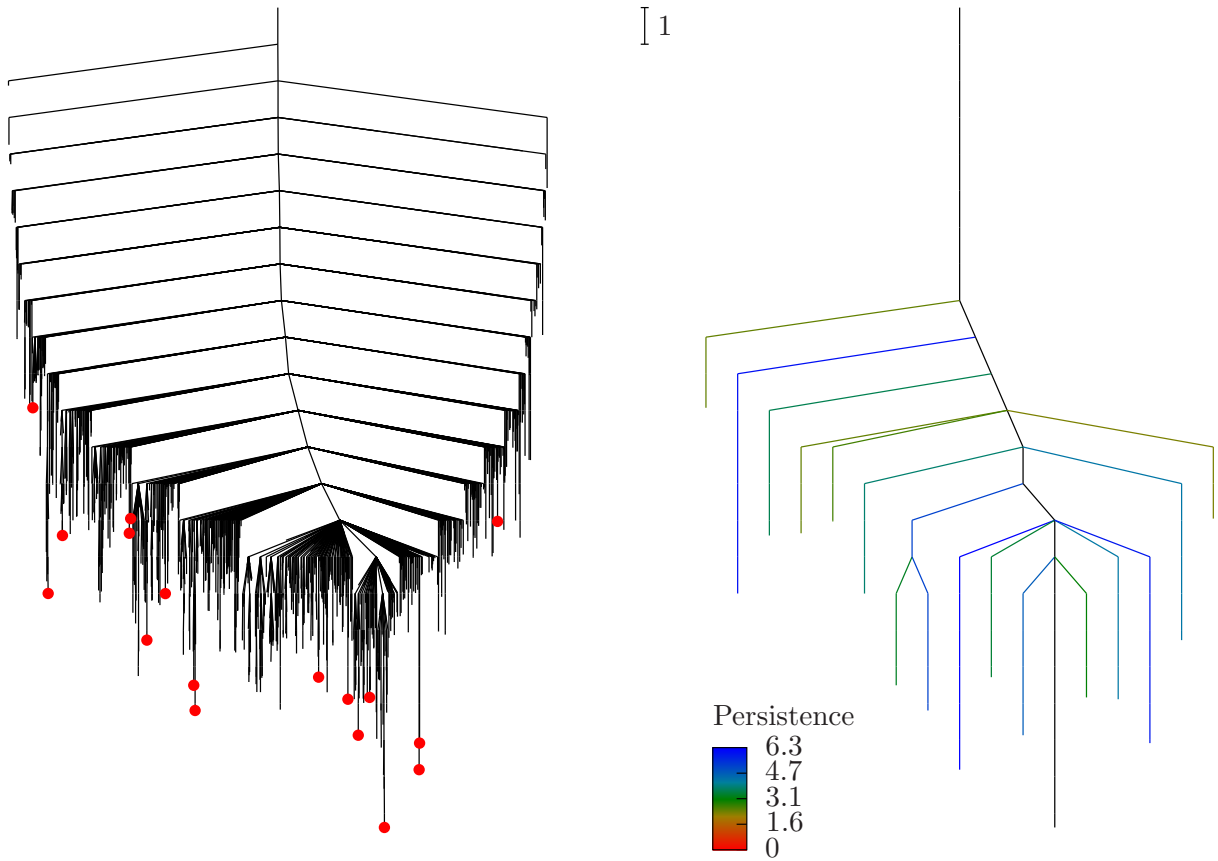
FIG. 15. Influenza virus binding: disconnectivity graphs. The vertical scale, denoted by the scale bar, is potential energy. Left: original graph. The red circles mark the minima that meet the selection criteria described in the text. Right: reduced graph, showing only those minima that meet the selection criteria. The lines are coloured according to the persistence of the corresponding local minimum; the line to the global minimum, defined to have infinite persistence, is black.
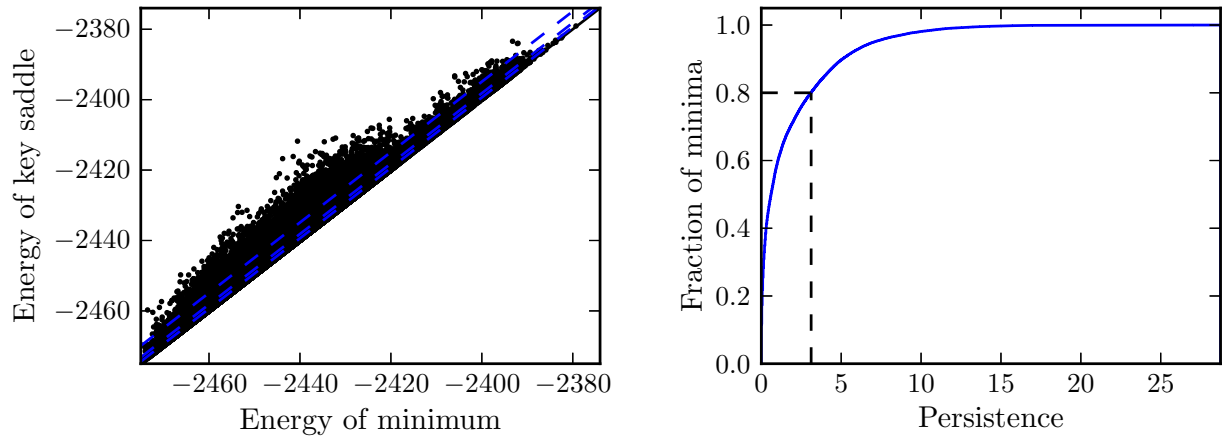
FIG. 16. UUCG RNA tetraloop: persistence diagram (left). The dashed lines mark the persistences corresponding to the 70th, 80th and 90th percentiles. Normalised cumulative histogram of persistences, with the 80th percentile marked (right).
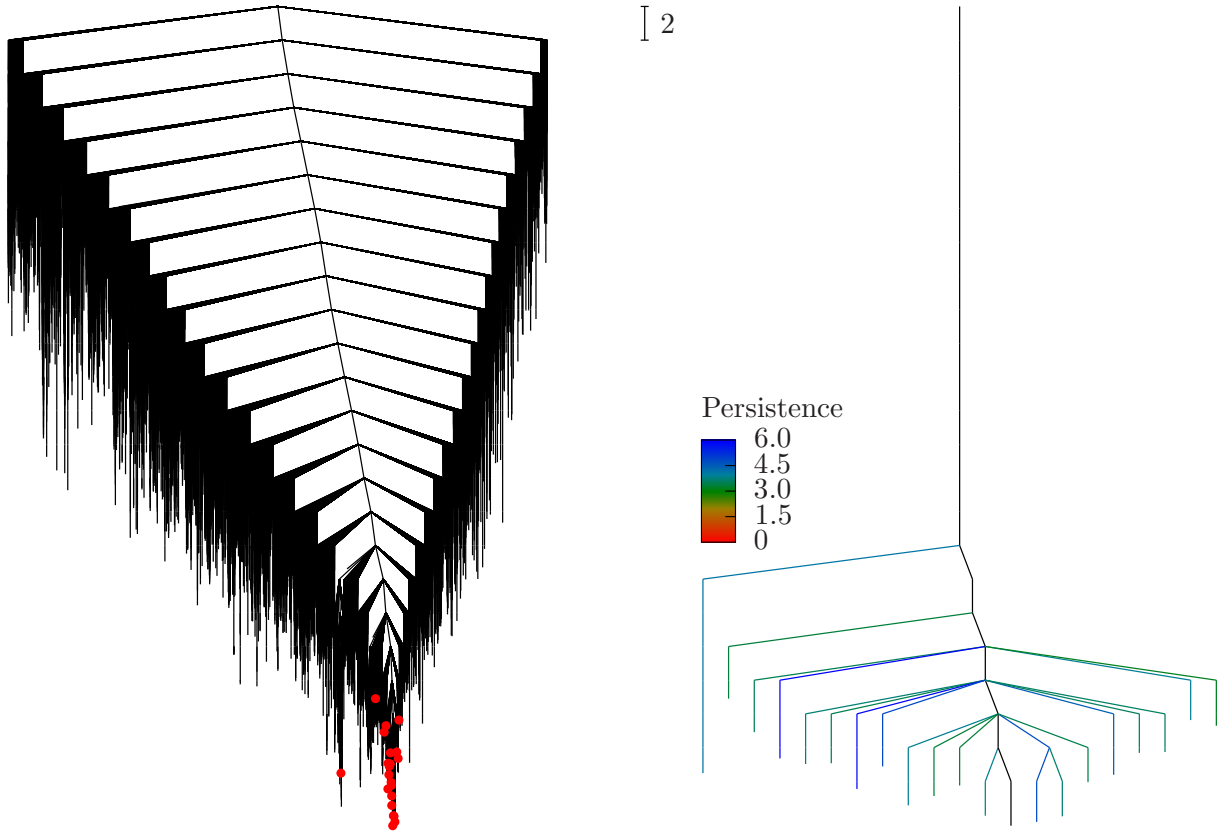
FIG. 17. UUCG RNA tetraloop: disconnectivity graphs. The vertical scale, denoted by the scale bar, is potential energy. Left: original graph. The red circles mark the minima that meet the selection criteria described in the text. Right: reduced graph, showing only those minima that meet the selection criteria. The lines are coloured according to the persistence of the corresponding local minimum; the line to the global minimum, defined to have infinite persistence, is black.
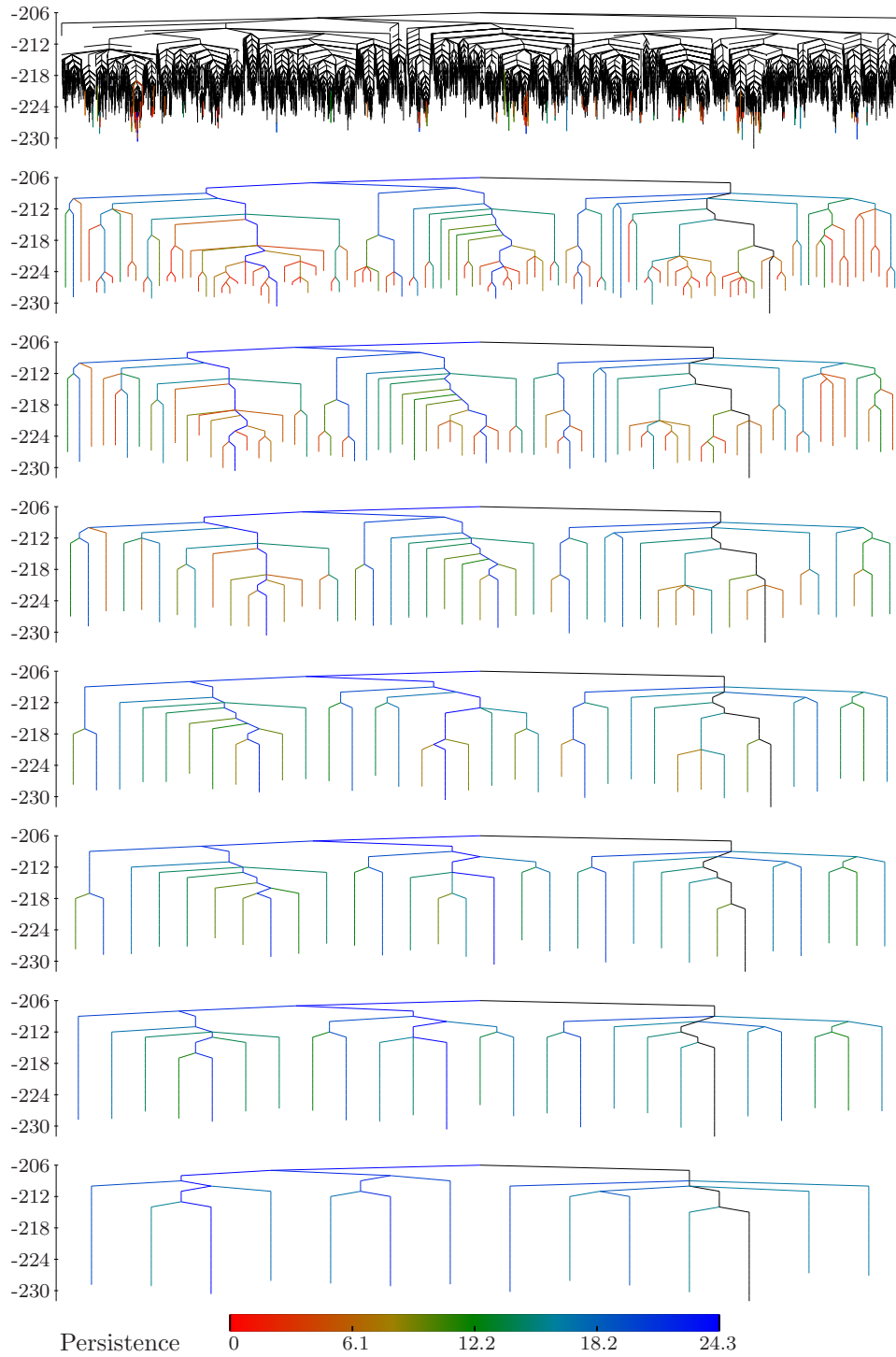
FIG. 18. Reduced disconnectivity graphs for BLJ_{60} as the persistence threshold is varied as follows (from top to bottom): 0, 1, 3, 5, 7, 9, 11, 15. Lines are coloured according to the persistence of the corresponding local minimum. In the full graph (no persistence cutoff; top panel), lines to minima that do not feature in the lower graphs are black. In all cases, the line to the global minimum is black. The vertical axes are potential energy.

37