

# Visualizing Energy Landscapes with Metric Disconnectivity Graphs

Lewis C. Smeeton, Mark T. Oakley, and Roy L. Johnston

The visualization of multidimensional energy landscapes is important, providing insight into the kinetics and thermodynamics of a system, as well the range of structures a system can adopt. It is, however, highly nontrivial, with the number of dimensions required for a faithful reproduction of the landscape far higher than can be represented in two or three dimensions. Metric disconnectivity graphs provide a possible solution, incorporating the landscape connectivity information present in disconnectivity graphs with structural information in the form of a metric. In this study, we present a new software package, PyConnect, which is

capable of producing both disconnectivity graphs and metric disconnectivity graphs in two or three dimensions. We present as a test case the analysis of the 69-bead BLN coarse-grained model protein and show that, by choosing appropriate order parameters, metric disconnectivity graphs can resolve correlations between structural features on the energy landscape with the landscapes energetic and kinetic properties. © 2014 The Authors Journal of Computational Chemistry Published by Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23643

## Introduction

The potential energy surface,  $U(\mathbf{r})$ , of an  $N$  atom chemical system represents the potential energy as a function of  $3N$  atomic coordinates. The topography of  $U(\mathbf{r})$ , or energy landscape, determines its structure, kinetics, and thermodynamics<sup>[1,2]</sup> and its analysis has proved useful in studying a range of physical systems and phenomena, including glasses,<sup>[3]</sup> biomolecules,<sup>[4–6]</sup> and clusters.<sup>[7–9]</sup> For all but the simplest cases,  $U(\mathbf{r})$  has many more degrees of freedom than it is possible to visualize conventionally, making it impossible to assess the surface topography directly. One solution to the visualization problem is to partition the landscape into discrete regions, and then hierarchically cluster these regions according to some similarity measure. This clustering can then be represented as a tree-graph in either two or three dimensions (2D or 3D). There are a number of examples of hierarchical clustering methods in the literature, broadly based on either geometry, energetic barriers, or local ergodicity.

In geometrical clustering, regions are clustered according to their structural similarity, which is usually defined by the root-mean-square deviation (RMSd) between them. In this context, regions can either correspond to minima on  $U(\mathbf{r})$ <sup>[10]</sup> or points along a molecular dynamics trajectory.<sup>[11,12]</sup> The structures are clustered either by an iterative process, by which each structure is joined to its nearest neighbor until only a single cluster remains,<sup>[11]</sup> or clustering structures that are within a critical distance of one another.<sup>[10]</sup>

When clustering according to energetic barriers, the landscape is partitioned into basins of attraction whereby each point on the landscape  $U(\mathbf{r})$ , is mapped onto a local minimum  $\alpha$ , with coordinate,  $\mathbf{r}_\alpha$ , by a steepest-descent path.<sup>[3,13]</sup> Alternatively, the landscape can be partitioned using a lumping approach,<sup>[14]</sup> in which energy thresholds are used to group connected regions below the threshold. The similarity measure used for hierarchical clustering is the barrier energy that separates any two regions. Starting from the energy of the global

minimum,  $U_0$ , regions are clustered together if they are separated by a barrier with an energy lying in the interval  $U_{i+1} - U_i$ , where  $U_{i+1} = U_i + \Delta U$  and  $\Delta U$  is the width of the interval. This clustering is repeated until a particular energy threshold,  $U_t$ , is reached, or all the minima are clustered together. Such graphs have come to be referred to as disconnectivity graphs,<sup>[13,15]</sup> and have been used in a number of studies to visualize energy landscapes.<sup>[13,15–17]</sup> Disconnectivity graphs retain both the energies of minima on  $U(\mathbf{r})$ , and the barriers that separate them, making them a useful diagnostic in visually assessing the thermodynamic and kinetic behavior of a system.<sup>[18,19]</sup> They can also be used to represent free-energy surfaces by estimating the vibrational entropy of minima and transition states on the landscape from the harmonic superposition approximation.<sup>[20–22]</sup> Clustering landscapes by local ergodicity involves partitioning the landscape into basins about local minima. Equilibration between basins is determined by comparing forward and backward transition rates between states<sup>[23]</sup> or the time-dependent probability distributions of connected basins.<sup>[24]</sup>

A weakness of the disconnectivity graph method is that it does not retain any structural information on the minima and thus neglects a large portion of the information contained in the energy landscape. Metric disconnectivity graphs capture some of this structure by defining a metric, and then calculating an order parameter from the metric for each minimum of

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

L. C. Smeeton, M. T. Oakley, R. L. Johnston

School of Chemistry, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom

E-mail: r.l.johnston@bham.ac.uk

Contract grant sponsor: Engineering and Physical Sciences Research Council, UK (EPSRC); Contract grant number: EP/I001352/1

© 2014 The Authors Journal of Computational Chemistry Published by Wiley Periodicals, Inc.

interest on the landscape. The minima can then be plotted along a metric axis perpendicular to the energy axis. Metric information can be included in a number of other ways, such as by changing the color, or thickness of the nodes and edges.<sup>[25,26]</sup>

In this article, we will refer to metric disconnectivity graphs as those for which the nodes are organized along a metric axis. A judicious choice of metric captures overall structural trends in the system, while ignoring noisy or irrelevant information.

Here, we demonstrate the use of metric disconnectivity graphs, using several metrics, to visualize the energy landscapes of coarse-grained proteins. These disconnectivity graphs are plotted with our new energy landscape visualization package, PyConnect.<sup>[27]</sup>

## Methodology

### BLN model

Metric disconnectivity graph analysis was performed on a database of stationary points for a BLN model protein. This database was generated with discrete path sampling<sup>[8,28]</sup> as implemented in PATHSAMPLE.<sup>[29]</sup> The BLN model<sup>[30,31]</sup> is a coarse-grained, off-lattice protein model in which each protein residue is represented by one of three types of bead: hydrophobic, hydrophilic, or Neutral. Here, we use a version of the BLN potential in which the interresidue distances and angles are restrained with stiff springs.<sup>[32]</sup> The beads interact with each other according to

$$U_{\text{BLN}}(\mathbf{r}) = \frac{1}{2}K_r \sum_{i=1}^{N-1} (R_{i,i+1} - R_e)^2 + \frac{1}{2}K_\theta \sum_{i=1}^{N-2} (\theta_i - \theta_e)^2 + \varepsilon \sum_{i=1}^{N-3} \{A_i(1 + \cos\phi_i) + B_i(1 + 3\cos\phi_i)\} + 4\varepsilon \sum_{i=1}^{N-2} \sum_{j=i+2}^N C_{ij} \left(\frac{\sigma}{R_{ij}}\right)^{12} + D_{ij} \left(\frac{\sigma}{R_{ij}}\right)^6 \quad (1)$$

where  $R_{ij}$  is the distance between two beads  $i$  and  $j$ . The first term is a harmonic bond restraint with  $K_r = 231.2\varepsilon\sigma^{-2}$  and  $R_e = \sigma$ . The second term represents a harmonic angle constraint with  $K_\theta = 20\text{rad}^{-2}$  and  $\theta_e = 1.8326\text{rad}$ . The third term takes into account torsional angles along the chain and is defined by four consecutive beads. If two or more beads are N, then  $A = 0$  and  $B = 0.2$ , else  $A = B = 1.2$ . The fourth term represents long range, water-mediated hydrophobic interactions between nonbonded pairs. If both beads are B, then  $C = D = 1$ . If one residue is L and the other is L or B, then  $C = \frac{2}{3}$  and  $D = -1$ . If either residue is N, then  $C = 1$  and  $D = 0$ .<sup>[32]</sup>

Though other sequences exist and have been studied, we consider here BLN-69, which consists of 69 beads with the sequence<sup>[33]</sup> B<sub>9</sub>N<sub>3</sub>(LB)<sub>4</sub>N<sub>3</sub>B<sub>9</sub>N<sub>3</sub>(LB)<sub>4</sub>N<sub>3</sub>B<sub>9</sub>N<sub>3</sub>(LB)<sub>5</sub>L. BLN-69 has been designed to exhibit a frustrated energy landscape, with a 6-strand  $\beta$ -barrel structure as its global minimum. The model has been shown to have a number of low-energy  $\beta$ -barrel-like structures, which differ from the global minimum by a chain slip along the length of the barrel,<sup>[4]</sup> but are separated by large barriers. Such frustration is absent when considering the

“G $\ddot{o}$ ” version of the model (G $\ddot{o}$ -69), where attractive interactions between pairs of residues that are not in contact in the native state (i.e., the global minimum) are neglected.<sup>[34,35]</sup>

### Metric disconnectivity graphs

Disconnectivity graphs and metric disconnectivity graphs are plotted using PyConnect.<sup>[27]</sup> The PyConnect package comprises two components: PCA, which calculates the principal components of molecular systems from PATHSAMPLE<sup>[29]</sup> databases, and PyConnect, which constructs and displays metric disconnectivity graphs. Both of these programs were written in Python. The disconnectivity graphs are rendered with Matplotlib,<sup>[36]</sup> and users can choose to create disconnectivity graphs and metric disconnectivity graphs in 2D or 3D. PyConnect also provides some cosmetic features, including the ability to label minima, color minima according to an order parameter or according to their basin of residence. PyConnect can also be used to modify graphs interactively using the iPython<sup>[37]</sup> virtual environment. In the disconnectivity graphs produced by PyConnect, the position of nodes and minima along the x axis are determined by algorithms similar to those used in DISCONNECT,<sup>[38]</sup> another program for producing disconnectivity graphs from databases of minima and transition states. Full details of the algorithms used can be found on the PyConnect website.<sup>[27]</sup> Two-dimensional metric disconnectivity graphs are plotted with the position of the minima on the x axis defined according to a metric. In 3D disconnectivity graphs, two metrics are used. The positions of nodes on the disconnectivity graphs are defined as the mean of the metrics for all minima connected to that node.

**Native Contact Metric.** The native contact metric evaluates for each minimum,  $\alpha$ , the ratio  $N_\alpha/N_{NC}$ , where  $N_{NC}$  is the number of native contact pairs, and  $N_\alpha$  is the number of contact pairs in minimum  $\alpha$  that are also present in the native conformation. Here, contacts are defined as those beads which are within  $1.167\sigma$  of each other, excluding pairs that are within three beads of each other in the peptide sequence.<sup>[4]</sup>

Hydrogen bonding is important in protein folding, and native contact analysis can provide a useful analogy for coarse-grained protein models.  $N_\alpha/N_{NC}$  is commonly used as a progress variable in computational studies of protein folding to distinguish between the different degrees of partially folded protein.<sup>[39]</sup>

**RMSd Metric.** The RMSd,  $d_{\alpha\beta}$ , measures the distance between the conformation of minimum  $\alpha$  and  $\beta$ ,  $\mathbf{r}_\alpha$  and  $\mathbf{r}_\beta$  respectively, according to

$$d_{\alpha\beta} = \sqrt{|\mathbf{r}_\alpha - \mathbf{r}_\beta|^2} \quad (2)$$

Invariance under global translations and rotations is implicit if structures are represented in internal coordinates. When working in Cartesian coordinates, the Kabsch algorithm<sup>[40]</sup> was used to align structures to minimize  $d_{\alpha\beta}$ . In the RMSd metric,  $d_{\alpha\beta}$  is calculated between the conformation of each minimum and the conformation of the global minimum,  $\mathbf{r}_{GM}$ .

**Principal Component Metric.** The principal component metric is based on principal component analysis (PCA), a statistical procedure used to analyze large, high-dimensional data sets, which is commonly used in dimensional reduction and, or when the relevant degrees of freedom in a data set are not clear.<sup>[41]</sup> PCA attempts to reexpress a data set in terms of a new basis set, the principal components, which are a linear transform of the data sets original basis set. The principal components lie along the axes of greatest sample variance, with the first principal component, PC1, capturing the axis of greatest variance, the second principal component, PC2, capturing the axis of second greatest variance (orthogonal to the first) and so on.<sup>[42]</sup>

We performed PCA on the set of  $N_{sp}$  stable configurations  $\{\mathbf{r}_x\}_{U_t}$ , where  $\{\mathbf{r}_x\}_{U_t}$  are all local minima connected to the global minimum below a certain threshold energy  $U_t$ . The initial basis sets employed were the  $3N$  dimensional external Cartesian basis set,  $\{\mathbf{e}_i\}$ , and an internal basis set of dihedral angles,  $\{\psi_i\}$ . To remove the periodicity of  $\{\psi_i\}$ , we used the sines and cosines of the internal dihedrals,  $\{\cos\psi_i, \sin\psi_i\}$ .<sup>[43]</sup> Rotational and translational invariance of  $\{\mathbf{r}_x\}_{U_t}$  was enforced by implementing McLachlan's best fit procedure<sup>[44]</sup>:

1. A reference configuration defined as the ensemble average,  $\langle \mathbf{r} \rangle$  of  $\{\mathbf{r}_x\}$  was calculated, where  $\{\mathbf{r}_x\}$  is the set of  $N_{sp}$  minima of interest, and where each configuration in  $\{\mathbf{r}_x\}$  has its centroid centered on the origin.
2. Define a new set  $\{\mathbf{r}'_x\}$ , rotate each configuration about its origin to be as close to  $\langle \mathbf{r} \rangle$  as possible using the Kabsch algorithm, and thus minimize

$$s = \frac{1}{2} \sum_{x=1}^{N_{sp}} (\mathbf{r}_x - \langle \mathbf{r} \rangle)^2 \quad (3)$$

3. Replace  $\{\mathbf{r}_x\}$  with  $\{\mathbf{r}'_x\}$ .
4. Repeat steps 1–3 until the ensemble average converges to some threshold criterion.

In our study, we used the threshold criterion defined by Komatsuzaki et al.,<sup>[25]</sup>  $s \leq 10^{-8}$ .

Hereafter, whether discussing Cartesian or internal coordinates, we define  $\{\mathbf{r}_x\}$  as the translation-free, rotation-free set of configurations, and  $\{\mathbf{q}_i\}$  as the basis set, where  $i$  is the coordinate index.

To perform PCA, we begin with defining the  $3N \times N_{sp}$  mean-centered configuration matrix,  $\mathbf{R}$

$$\mathbf{R} = (\mathbf{r}_1 \quad \cdots \quad \mathbf{r}_x \quad \cdots \quad \mathbf{r}_N) \quad (4)$$

where each column of  $\mathbf{R}$  is a  $3N$  dimensional vector corresponding to a stable configuration in the set  $\{\mathbf{r}_x\}_{U_t}$ . The PCs are the eigenvectors of the  $3N \times 3N$  covariance matrix,  $\mathbf{C}$

$$\mathbf{C} = \mathbf{R}^\dagger \mathbf{R} \quad (5)$$

and are thus the basis set  $\{\mathbf{Q}_i\}$ , where  $i$  is the coordinate index, in which  $\mathbf{C}$  is diagonalized. The PCs are calculated using the sin-

gular value decomposition method, which states that a  $3N \times N_{sp}$  configuration matrix,  $\mathbf{R}$  can be written as the product

$$\mathbf{R} = \mathbf{W} \mathbf{S} \mathbf{V}^\dagger \quad (6)$$

where  $\mathbf{W}$  is the  $3N \times 3N$  matrix;

$$\mathbf{W} = (\mathbf{Q}_1 \quad \cdots \quad \mathbf{Q}_i \quad \cdots \quad \mathbf{Q}_{3N}) \quad (7)$$

whose columns are the PCs of  $\mathbf{C}$  and  $\mathbf{S}$  is a  $3N \times N_{sp}$  matrix with diagonal elements,  $S_{ji}$ , where (dropping the double index for clarity)  $S_i^2$  is the variance associated with  $\mathbf{Q}_i$ . The PCs are ordered so that  $\mathbf{Q}_1$  has the greatest variance,  $\mathbf{Q}_2$  has the second greatest variance, and so on. The  $i$ th principal component metric is calculated by transforming each member of  $\{\mathbf{r}_x\}_{U_t}$  into the basis set of  $\{\mathbf{Q}_i\}$ , and using the value of the  $i$ th PC for each minimum as the order parameter.

One can visualize the PCs of a given  $\{\mathbf{r}_x\}_{U_t}$  by choosing a reference structure,  $\mathbf{r}_{ref}$ , and adding the  $\mathbf{Q}_i$  of interest to it

$$\mathbf{r}_\lambda = \mathbf{r}_{ref} + \lambda \mathbf{Q}_i \quad (8)$$

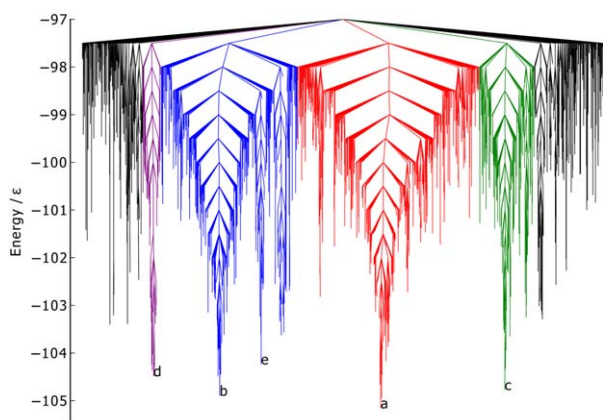
where  $\lambda$  is a progress variable.

**Isomap Metric.** The Isomap metric is based on the Isomap algorithm,<sup>[45,46]</sup> a nonparametric, nonlinear dimensionality reduction technique. The aim of the Isomap algorithm is to define a low-dimensional embedding that as accurately as possible preserves geodesic distances between all pairs of points in the data cloud. The geodesic distance between a pair of points that lie on a manifold is the length of the shortest path between them that lies along that manifold. Isomap assumes that such a low-dimensional manifold exists, and that its shape can be estimated from the distribution of points in the data cloud. The Isomap algorithm approximates the geodesic distance between a given pair of points on the manifold by calculating the shortest possible path between them that can be found by stepping from one point to its neighbor.

We applied Isomap to the set of  $N_{sp}$  stable configurations  $\{\mathbf{r}_x\}_{U_t}$  using the Isomap implementation in the scikit-learn machine learning package.<sup>[47]</sup> As with the principal component metric, rotational and translational invariance of  $\{\mathbf{r}_x\}_{U_t}$  was enforced by implementing McLachlan's best fit procedure.

**Table 1.** The variance captured by the first three principal components in Cartesian,  $S_i^{\text{cart}}$ , and dihedral,  $S_i^{\text{di}}$ , bases of the  $N_{sp}$  structures in the sublevel sets of minima below threshold energy  $U_t$  for BLN-69.

$U_t/\epsilon$	$N_{sp}$	Cartesian PCA			Dihedral PCA		
		$S_1^{\text{cart}}$	$S_2^{\text{cart}}$	$S_3^{\text{cart}}$	$S_1^{\text{di}}$	$S_2^{\text{di}}$	$S_3^{\text{di}}$
-95.0	6891	25.0	9.2	8.0	12.3	8.5	8.2
-95.5	5973	25.7	8.9	8.1	12.5	8.6	8.4
-96.0	5135	26.0	8.7	8.2	12.3	8.9	8.6
-96.5	4353	27.4	8.5	8.2	12.7	9.2	9.8
-97.0	1611	37.0	10.0	7.9	12.8	10.6	9.7
-97.5	561	21.2	19.0	10.8	15.1	13.7	11.9
-98.0	409	25.8	16.6	10.8	15.6	14.1	12.0



**Figure 1.** Disconnectivity graph of BLN-69,  $U_t = -97.0\epsilon$ ,  $N_{sp} = 1611$ . The color scheme is chosen to distinguish between energetic funnels. Labeled minima correspond to the global minimum and low-energy minima separated from the global minimum and one another by large kinetic barriers and are shown in Figure 2.

The Isomap algorithm works in three steps;

1. A weighted graph,  $G$ , of  $\{\mathbf{r}_\alpha\}_{U_t}$  is built, where each conformation is a node and where the  $k$  nearest-neighbors of each conformation  $\alpha$  are joined by an edge with

weight  $d_{\alpha\beta}$ . Isomap has been shown to be fairly robust to the choice of  $k$ ,<sup>[46]</sup> and in this study we took  $k = 15$ .

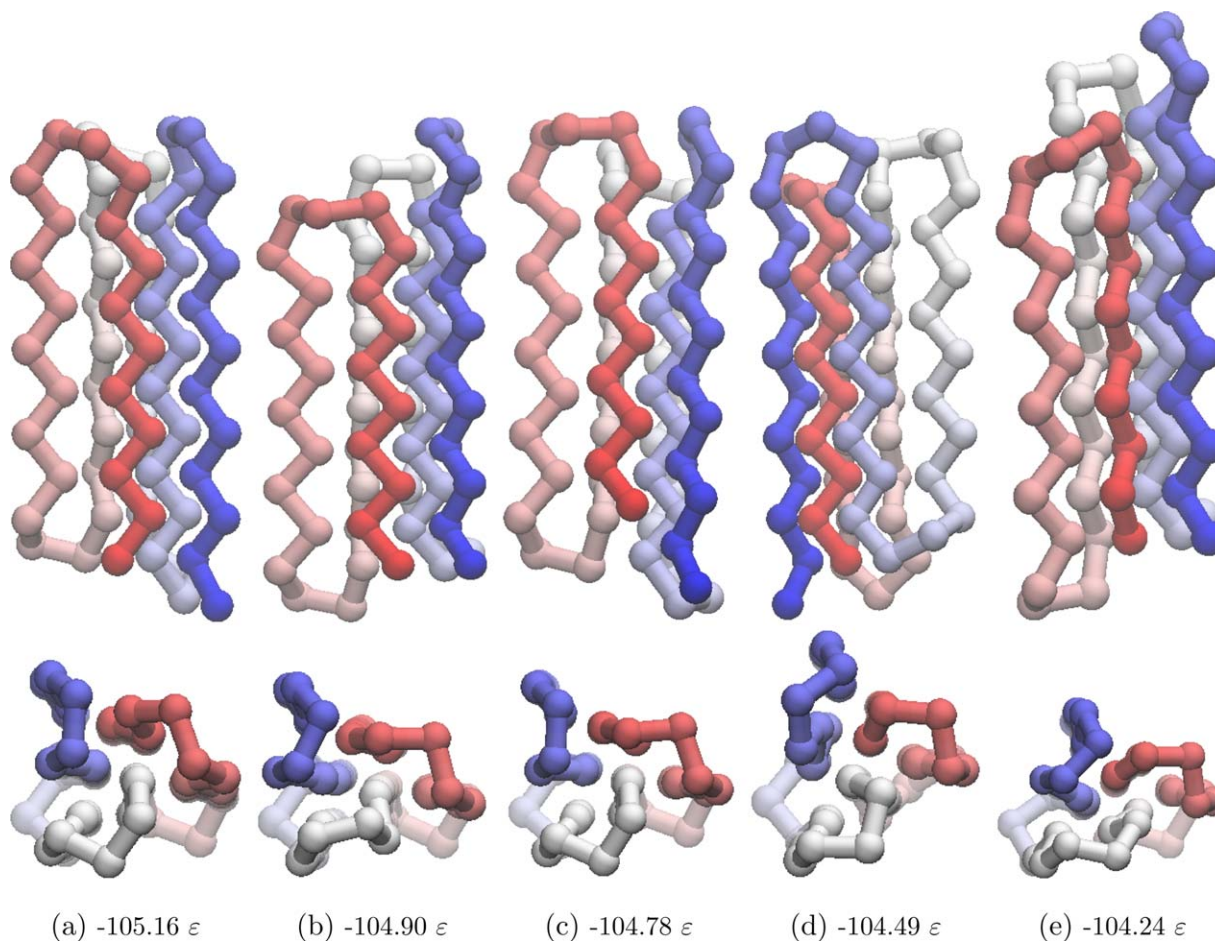
2. The shortest path between each conformation through the graph  $G$  is determined and a distance matrix,  $\mathbf{D}$ , is computed, where  $D_{\alpha\beta} = \min\{d_{\alpha\beta}, d_{\alpha\gamma} + d_{\gamma\beta}\}$  for  $\gamma = 1, \dots, N_{sp}$ . The elements  $D_{\alpha\beta}$  are the approximate geodesics between conformations  $\alpha$  and  $\beta$ .
3. Classical multidimensional scaling is applied to the matrix  $D$ , producing a low-dimensional embedding of the conformational coordinates that best preserves geodesic distances on the manifold.

The  $i$ th Isomap metric corresponds to the  $i$ th embedded dimension of the low-dimensional manifold of  $\{\mathbf{r}_\alpha\}_{U_t}$ .

## Results and Discussion

### BLN-69

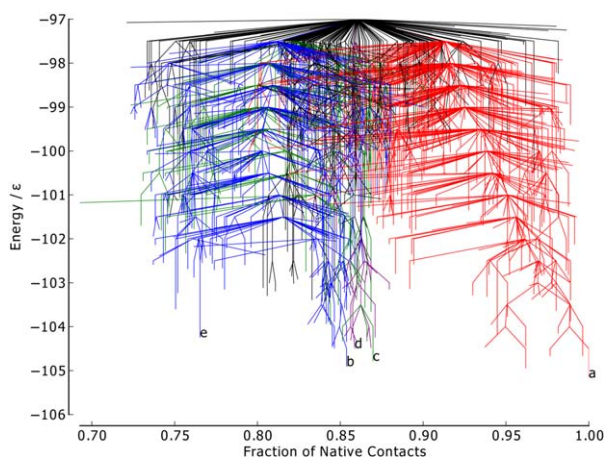
For BLN-69, a database containing 141,835 minima and 173,692 transition states was used in the study. The first three Cartesian and dihedral principal components for the sets,  $\{\mathbf{r}_\alpha\}_{U_t}$ , connected to the global minimum below energy  $U_t$ , where  $-95.0\epsilon \geq U_t \geq -98.0\epsilon$  are shown in Table 1.



**Figure 2.** Structures of the minima labeled in Figure 1, corresponding to the global minimum, Figure 2a and low-energy minima separated from the global minimum and one another by large kinetic barriers, Figures 2b–2e. Energetic and structural details are provided in Table 2. The beads are colored from red to blue (N-terminus to C-terminus).

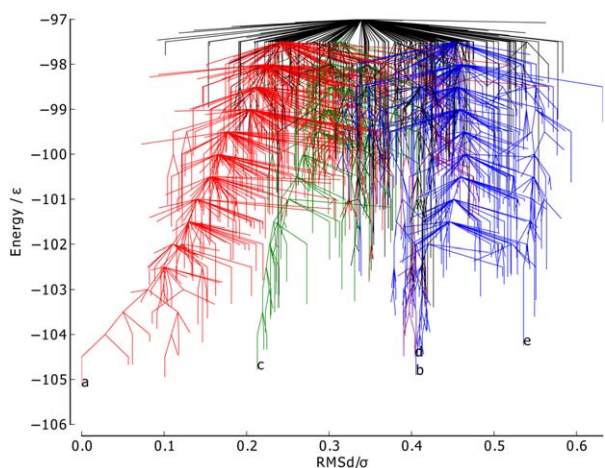
**Table 2.** Energy above the global minimum,  $\Delta U$ , fraction of native contacts,  $N_x/N_{NC}$ , RMSd from the global minimum and difference in PC1 and PC2 from the global minimum  $\Delta Q_1$  and  $\Delta Q_2$ , respectively, for minima *b–e*.

Minimum	$\Delta U/\epsilon$	$N_x/N_{NC}$	RMSd/ $\sigma$	$\Delta Q_1/\sigma$	$\Delta Q_2/\sigma$	Defect
<i>b</i>	0.26	0.85	0.41	-5.32	-1.25	Chain-slip
<i>c</i>	0.38	0.87	0.21	-0.02	-0.76	Reptation
<i>d</i>	0.67	0.86	0.40	-2.73	-3.51	Double chain-slip
<i>e</i>	0.92	0.77	0.54	-6.54	-2.61	Twist

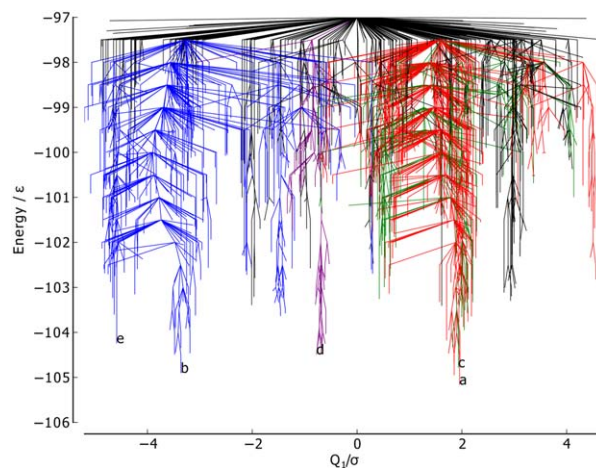


**Figure 3.** Metric disconnectivity graph of BLN-69,  $U_t = -97.0\epsilon$ ,  $N_{sp} = 1611$ , with fraction of native contacts used as an order parameter. The color scheme and labels are as used in Figure 1. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](#).]

For the Cartesian PCs, PC1 captures significantly more of the variance than PC2 for all data sets considered. PC1 for the sub-level set of minima connected to the global minimum below  $U_t = -97.0\epsilon$ ,  $\{\mathbf{r}_\alpha\}_{U_t = -97.0\epsilon}$  has the largest fractional variance and therefore this threshold was selected for all disconnectivity graphs. The dihedral PCs have a more uniform variance dis-



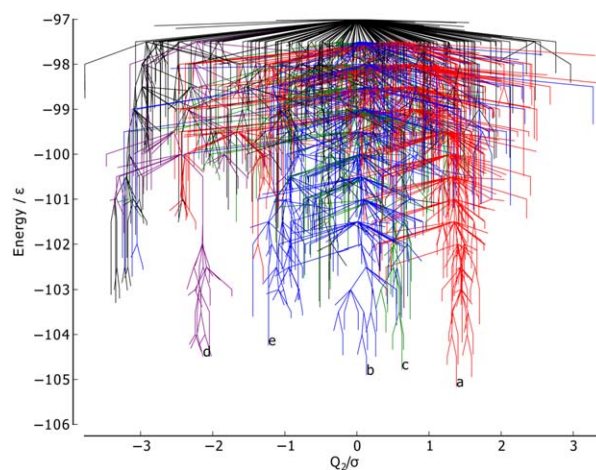
**Figure 4.** Metric disconnectivity graph of BLN-69,  $U_t = -97.0\epsilon$ ,  $N_{sp} = 1611$ , with RMSd of each structure from the global minimum used as an order parameter in units of  $\sigma$ . The color scheme and labels are as used in Figure 1. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](#).]



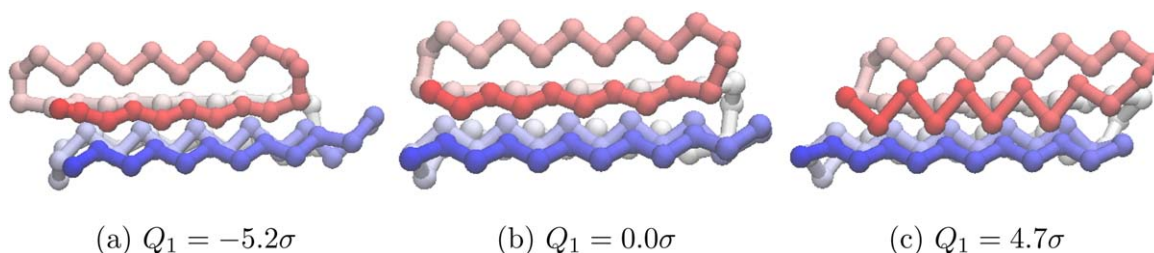
**Figure 5.** Metric disconnectivity graph of BLN-69,  $U_t = -97.0\epsilon$ ,  $N_{sp} = 1611$ , with PC1 for  $\{\mathbf{r}_\alpha\}_{U_t = -97.0\epsilon}$ ,  $Q_1$ , used as an order parameter in units of  $\sigma$ . The color scheme and labels are as used in Figure 1. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](#).]

tribution than the Cartesian PCs, with  $S_1^{di} \approx \frac{1}{2} S_1^{cart}$ , for all  $\{\mathbf{r}_\alpha\}_{U_t}$  considered in both BLN-69 and Gō-69. The dihedral PCs are thus not appropriate metrics for studying these systems, and have not been used to create metric disconnectivity graphs. The set of minima where  $U_t = -97.0\epsilon$  is represented as a disconnectivity graph in Figure 1. Figure 2 shows the low-energy minima labeled *a–e* in Figure 1. Minima *b–e* are all structurally similar to one another with each adopting compact  $\beta$ -barrel geometries and differing from global minimum *a* by either a chain-slip, chain-reptation, or twist in the turn regions, with further details given in Table 2.

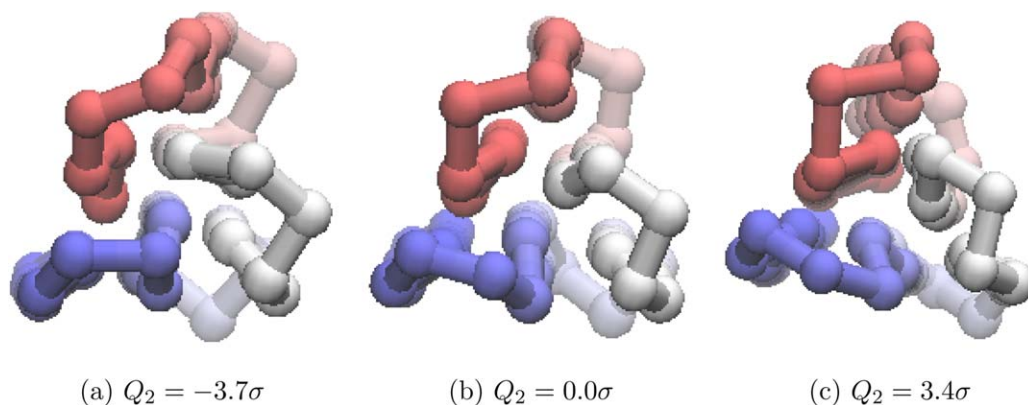
The native contact metric (Fig. 3) splits the two largest funnels, with each having distinct fractions of native contacts (the mean fraction of native contacts for the funnels containing minima *a* and *b* is 0.91 and 0.81, respectively). Though the native contact metric differentiates between kinetically separated minima, it does not differentiate according to their



**Figure 6.** Metric disconnectivity graph of BLN-69,  $U_t = -97.0\epsilon$ ,  $N_{sp} = 1611$ , with PC2 for  $\{\mathbf{r}_\alpha\}_{U_t = -97.0\epsilon}$ ,  $Q_2$ , used as an order parameter in units of  $\sigma$ . The color scheme and labels are as used in Figure 1. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](#).]



**Figure 7.** Different values of  $Q_1$  for  $U_t = -97.0\epsilon$  projected onto the structure of the global minimum of BLN-69. For the global minimum,  $Q_1 = 1.96\sigma$ . The beads are colored from red to blue (N-terminus to C-terminus). An animated version of this projection is available as Supporting Information. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

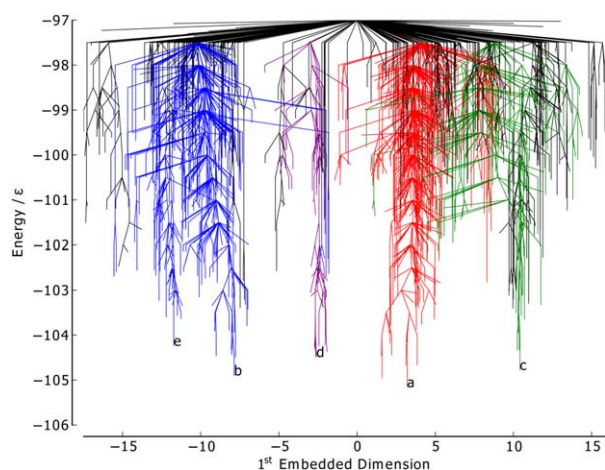


**Figure 8.** Different values of  $Q_2$  for  $U_t = -97.0\epsilon$  projected onto the structure of the global minimum of BLN-69. For the global minimum,  $Q_2 = -5.5\sigma$ . The beads are colored from red to blue (N-terminus to C-terminus). An animated version of this projection is available as Supporting Information. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

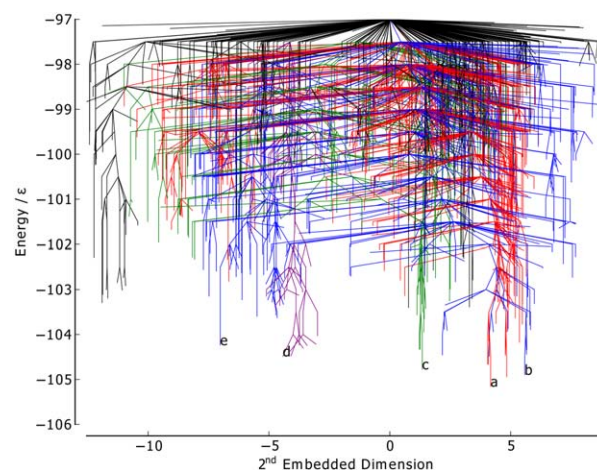
energies. There are a number of unstable, high-energy minima with energetically unfavorable turns in the flexible  $N$  bead regions, but otherwise with almost all native contacts satisfied. Minimum  $a$  by definition satisfies all native contacts. Minima  $b$ – $d$  are all very similar according to this metric, with each satisfying  $\approx 85\%$  of possible native contacts, in spite of their geometries being relatively dissimilar.

The RMSd metric (Fig. 4) is capable of distinguishing between the different major funnels on the surface, with each

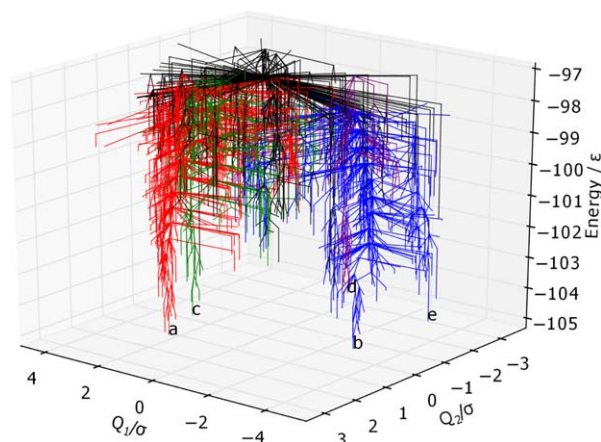
having its own mean value of the metric (mean RMSd for the funnels containing minima  $a$ ,  $c$ , and  $b$   $0.24\sigma$ ,  $0.31\sigma$ , and  $0.45\sigma$ , respectively). There is also some relation to the minima energy in the green and red funnels, where lower energy corresponds to RMSd metric values closer to 0. The RMSd metric differentiates the basin minima into four groups, with minimum  $c$  being most similar to the global minimum, which is as expected from a visual inspection of the structures.



**Figure 9.** Metric disconnectivity graph of BLN-69,  $U_t = -97.0\epsilon$ ,  $N_{sp} = 1611$ , with the first embedded dimension for  $\{\mathbf{r}_x\}_{U_t = -97.0\epsilon}$  from Isomap analysis used as an order parameter in units of  $\sigma$ . The color scheme and labels are as used in Figure 1. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]



**Figure 10.** Metric disconnectivity graph of BLN-69,  $U_t = -97.0\epsilon$ ,  $N_{sp} = 1611$ , with the second embedded dimension for  $\{\mathbf{r}_x\}_{U_t = -97.0\epsilon}$  from Isomap analysis used as an order parameter in units of  $\sigma$ . The color scheme and labels are as used in Figure 1. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]



**Figure 11.** 3D metric disconnectivity graph of BLN-69,  $U_t = -97.0\epsilon$ ,  $N_{sp} = 1611$ , plotted with the first two principal components of  $\{\mathbf{r}_x\}_{U_t = -97.0\epsilon}$ ,  $Q_1$  and  $Q_2$ , used as order parameters in units of  $\sigma$ . The color scheme and labels are as used in Figure 1. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

The PC1 metric (Fig. 5) splits the blue funnel (mean  $3.29\sigma$ ) from the red and green funnels, which sit on top of one another (mean  $1.59\sigma$  and  $1.60\sigma$ , respectively). The purple funnel lies at the boundary of the two, with a mean of  $-0.86\sigma$ . Minima *a* and *c* have almost identical values of  $Q_1$ , while minima *d*, *b*, and *e* have increasingly dissimilar values. Given that PC1 corresponds to a chain-slip between the C and N termini, and that minima *d*, *b*, and *e* have chains that have shifted relative to the global minimum in the same direction, it gives confidence that PCA is capable of identifying structural features of the energy landscape.

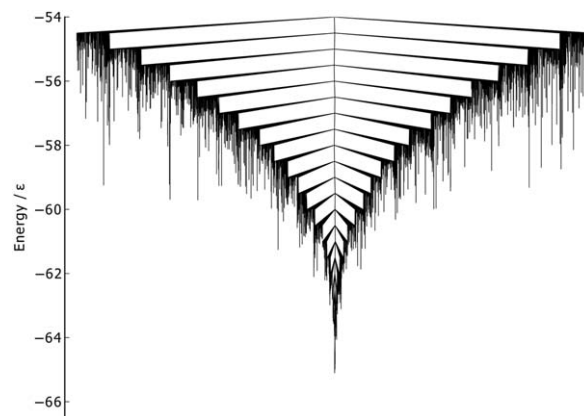
The PC2 metric (Fig. 6) does not reveal any obvious correlation between structure and energetics or kinetics, with no distinction made between the funnels and with the points reasonably evenly distributed along the order parameter. Thus, this PC corresponds to variations within all of the funnels rather than structural differences between the funnels.

The progression of PC1 of  $\{\mathbf{r}_x\}_{U_t = -97.0\epsilon}$  from  $\lambda = -5.2\sigma$  to  $\lambda = 4.7\sigma$  (Fig. 7) corresponds to a chain-slip between the C and N termini. The progression of PC2 of  $\{\mathbf{r}_x\}_{U_t = -97.0\epsilon}$  from  $\lambda = -3.7\sigma$  to  $\lambda = 3.4\sigma$  (Fig. 8) corresponds to a twisting of the internal chain sequences.

The first embedded dimension of the Isomap metric (Fig. 9) clearly differentiates between all the colored funnels on the land-

**Table 3.** The variance captured by the first three principal components in Cartesian,  $S_i^{\text{cart}}$ , and dihedral,  $S_i^{\text{di}}$ , bases of the  $N_{sp}$  structures in the sublevel sets of minima below threshold energy  $U_t$  for Gō-69.

$U_t/\epsilon$	$N_{sp}$	Cartesian PCA			Dihedral PCA		
		$S_1^{\text{cart}}$	$S_2^{\text{cart}}$	$S_3^{\text{cart}}$	$S_1^{\text{di}}$	$S_2^{\text{di}}$	$S_3^{\text{di}}$
-52.0	5529	38.6	14.3	12.6	14.4	10.6	7.3
-53.0	4364	38.0	13.7	12.0	13.7	11.1	7.7
-54.0	3188	24.5	16.0	8.0	13.3	11.6	8.2
-55.0	2386	24.4	16.0	8.1	13.3	11.3	8.7
-56.0	1691	23.7	15.7	7.6	14.0	12.1	9.4
-57.0	1185	21.6	16.3	7.7	13.9	12.7	10.4
-58.0	739	21.5	15.3	7.7	14.9	12.9	11.4



**Figure 12.** Disconnectivity graph of Gō-69,  $U_t = -54.0\epsilon$ ,  $N_{sp} = 3189$ .

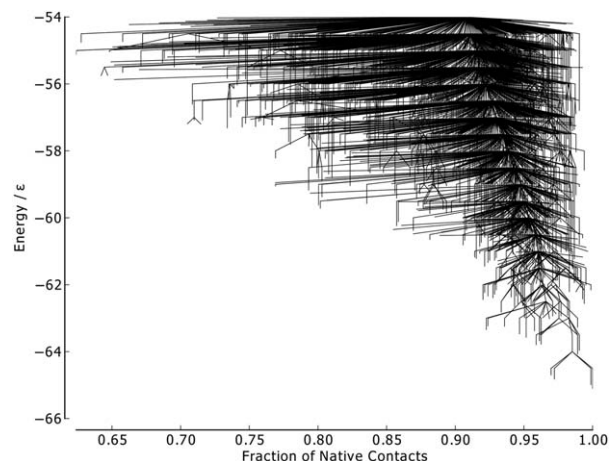
scape. The mean value of the blue, purple, red, and green funnels are  $-10.23\sigma$ ,  $-2.96\sigma$ ,  $4.12\sigma$ , and  $8.80\sigma$ , respectively. The structure of the graph is similar to the PC1 graph (Fig. 5), with the order of the colored funnels and labeled low-energy minima along the metric axis matching. The agreement between these two metrics suggests that the first embedded dimension of the Isomap metric is fairly linear, and that, as with the PC1 metric, it corresponds to a chain-slip between the C and N termini.

As with the PC2 metric, the disconnectivity graph for the second embedded dimension of the Isomap metric (Fig. 10) is difficult to interpret. The overlapping of the colored funnels suggests that the second embedded dimension corresponds to some structural variation common to each funnel.

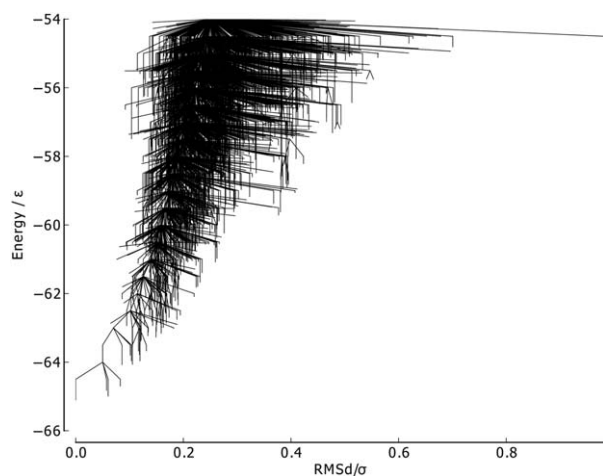
The information in Figures 5 and 6 is visualized on a single 3D metric disconnectivity graph of  $\{\mathbf{r}_x\}_{U_t = -97.0\epsilon}$  projected onto the plane of maximal variance in Figure 11. The plot shows  $\{\mathbf{r}_x\}_{U_t = -97.0\epsilon}$  for BLN-69 plotted against its first two principal components. Clear separation of minima *a*–*e* is discernible in this 3D metric disconnectivity graph.

## Gō-69

For Gō-69, a database containing 75,666 minima and 113,101 transition states was used. The first three Cartesian principal components for the sets,  $\{\mathbf{r}_x\}_{U_t}$ , connected to the global



**Figure 13.** Metric disconnectivity graph of Gō-69,  $U_t = -54.0\epsilon$ ,  $N_{sp} = 3189$ , with fraction of native contacts used as an order parameter.

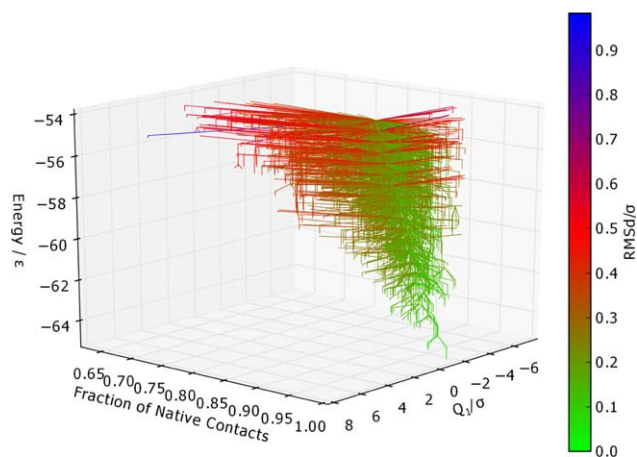


**Figure 14.** Metric disconnectivity graph of Gō-69,  $U_t = -54.0\epsilon$ ,  $N_{sp} = 3189$ , with RMSd of each structure from the global minimum used as an order parameter in units of  $\sigma$ .

minimum below energy  $U_t$ , where  $-52.0\epsilon \geq U_t \geq -58.0\epsilon$  are shown in Table 3.

As with the Cartesian PCs of BLN-69, PC1 captures significantly more of the variance than PC2 for all data sets considered. PC1 for the sublevel set of minima connected to the global minimum below  $U_t = -52.0\epsilon$  and  $U_t = -53.0\epsilon$ , have the largest fractional variance, though these large variances are due to a comparatively small number of unstable, high-energy minima in which one end of the chain has peeled away from the barrel and become unbound. For these systems, PC1 is no longer representative of the distribution of minima on  $U_r$ . For this reason, we consider the sublevel set of minima connected to the global minimum below  $U_t = -54.0\epsilon$ , for which all the minima have densely packed geometries. This set is represented as a disconnectivity graph in Figure 12. The results for the Isomap metric were fairly ambiguous for Gō, with no obvious pattern correlation between the embedded dimensions and the kinetic or energetic structure of the graph, so they have not been included in this work.

As there is only a single funnel on the Gō-69 landscape, there are no large kinetic barriers for any of the metrics to differentiate between. The native contact metric (Fig. 13) is able

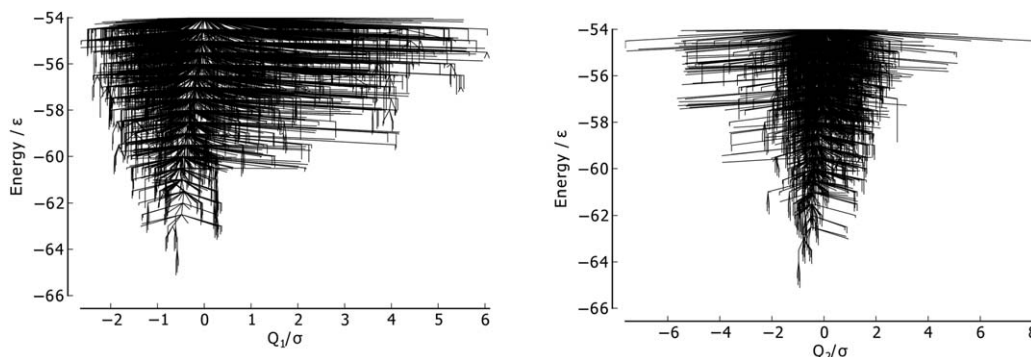


**Figure 16.** Metric disconnectivity graph of Gō-69,  $U_t = -54.0\epsilon$ ,  $N_{sp} = 3189$ , with PC1 for  $\{\mathbf{r}_x\}_{U_t = -54.0\epsilon}$ ,  $Q_1$ , and fraction of native contacts used as order parameters, and colored according to RMSd of each structure from the global minimum.  $Q_1$  and RMSd are in units of  $\sigma$ . [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

to partially distinguish between the structures of high- and low-energy minima. As with BLN-69, minima across the whole energy range examined were able to satisfy nearly full native contacts, including unstable, high-energy minima with energetically unfavorable turns in the flexible  $N$  bead regions. The converse is not true; however, as all low-energy minima have a high number of native contacts and low numbers of native contacts are only found for high-energy minima.

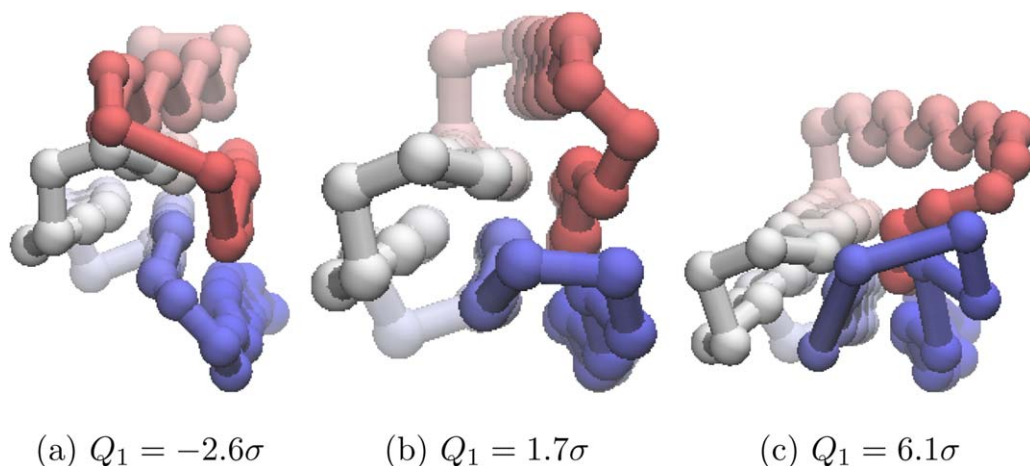
For the RMSd metric (Fig. 14), similar behavior to BLN-69 is exhibited, albeit with a single funnel, with RMSd from the global minimum increasing with increasing energy.

The metric disconnectivity graphs in Figure 15 use PC1 and PC2 as order parameters. In the PC1 graph, the majority of minima are centered about the global minimum, with a smaller number of high-energy minima extending to higher values of PC1. PC1 and the fraction of native contacts are well-correlated, as can be seen in the 3D metric disconnectivity graph (Fig. 16). The PC2 metric orders all but a few minima tightly in a rough column about  $Q_2 \approx 0.1$ . Those unstable, higher energy minima that are not in that column are structures with a partly unbound C-terminus chain-portion.



**Figure 15.** Metric disconnectivity graphs of Gō-69,  $U_t = -54.0\epsilon$ ,  $N_{sp} = 3189$ , with PC1 (left) and PC2 (right) for  $\{\mathbf{r}_x\}_{U_t = -54.0\epsilon}$ ,  $Q_1$  and  $Q_2$ , used as order parameters in units of  $\sigma$ .





**Figure 17.** Different values of  $Q_1$  for  $U_t = -54.0e$  projected onto the structure of the global minimum of G $\ddot{o}$ -69. For the global minimum,  $Q_1 = -0.6\sigma$ . The beads are colored from red to blue (N-terminus to C-terminus). An animated version of this projection is available as Supporting Information. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

The use of color allows an additional metric to be included on a metric disconnectivity graph. For example, Figure 16 shows the PC1, native contact, and RMSd metrics for G $\ddot{o}$ -69.

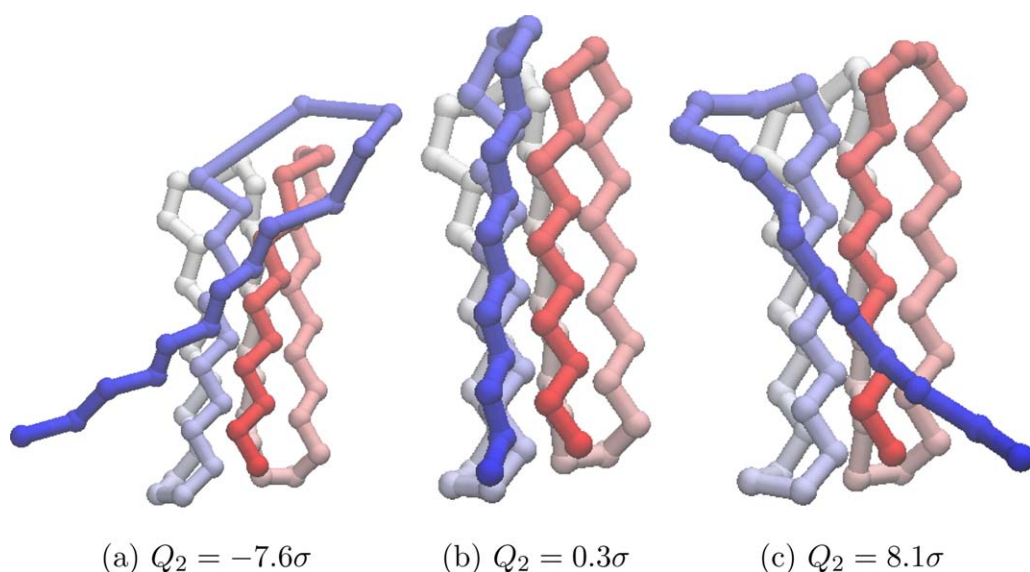
Figure 17 shows the progression of PC1 of  $\{\mathbf{r}_\alpha\}_{U_t = -54.0e}$  from  $\lambda = -2.6\sigma$  to  $\lambda = 6.1\sigma$ , with a view along the axis of the barrel and corresponds to a sweeping action of the red chain-portion across the face of the white chain-portion.

Figure 18 shows the progression of PC2 of  $\{\mathbf{r}_\alpha\}_{U_t = -54.0e}$  from  $\lambda = -7.6\sigma$  to  $\lambda = 8.1\sigma$ , and corresponds to a “can-can” like sweeping motion of the free C-terminus end of the chain.

## Conclusions

In this study, we have demonstrated how an appropriate order parameter can elucidate the connection between structures in the energy landscape of BLN-69 and G $\ddot{o}$ -69, such as funnels,

with certain structural motifs of the protein, including chain slips and twists in the turn regions. However, there are still shortcomings to the metrics proposed. Fraction of native contacts and RMSd metrics relied on having prior knowledge of the system. PCA provides a means to study systems without resorting to chemical intuition, but still assumes that the point cloud is approximately linear, and cannot be directly implemented for angular coordinates. Also, it considers all structures to be of equal importance, regardless of energy, leading to situations such as with G $\ddot{o}$ -69, where all the variance in structure was provided by a small number of high energy, unstable minima. Isomap allows one to discern low-dimensional, nonlinear manifolds in the data, and does not make the same assumptions of linearity as PCA. This is clearly a successful strategy, with Isomap distinguishing between the different kinetic structures on the landscape. A useful feature of PCA is the ease



**Figure 18.** Different values of  $Q_2$  for  $U_t = -54.0e$  projected onto the structure of the global minimum of G $\ddot{o}$ -69. For the global minimum,  $Q_2 = -0.9\sigma$ . The beads are colored from red to blue (N-terminus to C-terminus). An animated version of this projection is available as Supporting Information. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

with which one can project the principal components back into the original space, making it possible to visualize what these directions correspond to. In principal, it should be possible to do the same with Isomap, projecting the approximate geodesics of the manifold back into the original space, though we have not implemented this in the work presented here. Other nonlinear dimensionality reduction methods exist in the literature, such as sketch-map,<sup>[48]</sup> locally scaled diffusion map,<sup>[46,49]</sup> and spectral methods,<sup>[50]</sup> which are good candidate metrics for further study.

Equally, though the data produced by PyConnect is of a high-quality, the data analysis is still fairly qualitative, and further efforts are being taken to quantify the observations, such as using graph-theoretic techniques to analyze and compare tree graphs.<sup>[51]</sup>


Future work should also focus on investigating more realistic, small protein systems, such G-protein<sup>[52]</sup> or cyclic peptides.<sup>[5,22]</sup>

## Acknowledgments

The computations described in this paper were performed using the University of Birmingham's BlueBEAR HPC service, which provides a High-Performance Computing service to the University's research community. See <http://www.birmingham.ac.uk/bear> for more details. The authors thank Prof. David Wales for helpful discussions, and Dr. Victor Ruhle and Dr. Jacob Stevenson for advice about implementation of our Python code.

**Keywords:** collective variables · protein · coarse-grained models · software · Python

How to cite this article: L. C. Smeeton, M. T. Oakley, R. L. Johnston. *J. Comput. Chem.* **2014**, 35, 1481–1490. DOI: 10.1002/jcc.23643

 Additional Supporting Information may be found in the online version of this article.

- [1] D. J. Wales, *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*, Cambridge Molecular Science; Cambridge University Press: Cambridge, UK, **2003**.
- [2] J. C. Schön, M. Jansen, *Zeitschrift für Krist.* **2001**, 216, 307.
- [3] F. H. Stillinger, T. A. Weber, *Phys. Rev. A* **1982**, 25, 978.
- [4] M. T. Oakley, D. J. Wales, R. L. Johnston, *J. Phys. Chem. B* **2011**, 115, 11525.
- [5] M. T. Oakley, R. L. Johnston, *J. Chem. Theory Comput.* **2013**, 9, 650.
- [6] B. Strodel, D. J. Wales, *Chem. Phys. Lett.* **2008**, 466, 105.
- [7] R. L. Johnston, *Dalton Trans.* **2003**, 4193.
- [8] D. J. Wales, *Mol. Phys.* **2004**, 102, 891.
- [9] G. Cox, R. S. Berry, R. L. Johnston, *J. Phys. Chem. A* **2006**, 110, 11543.
- [10] O. M. Becker, *Proteins* **1997**, 27, 213.
- [11] A. García, R. Blumenfeld, *Phys. D* **1997**, 107, 225.
- [12] J. M. Troyer, F. E. Cohen, *Proteins* **1995**, 23, 97.
- [13] O. M. Becker, M. Karplus, *J. Chem. Phys.* **1997**, 106, 1495.
- [14] K. H. Hoffmann, P. Sibani, *Phys. Rev. A* **1988**, 38, 4261.
- [15] D. J. Wales, M. A. Miller, T. R. Walsh, *Nature* **1998**, 394, 758.
- [16] R. Czerminski, R. Elber, *J. Chem. Phys.* **1990**, 92, 5580.
- [17] T. Middleton, J. Hernández-Rojas, P. Mortenson, D. Wales, *Phys. Rev. B* **2001**, 64, 184201.
- [18] D. J. Wales, *Curr. Opin. Struct. Biol.* **2010**, 20, 3.
- [19] D. J. Wales, T. V. Bogdan, *J. Phys. Chem. B* **2006**, 110, 20765.
- [20] D. A. Evans, D. J. Wales, *J. Chem. Phys.* **2003a**, 118, 3891.
- [21] D. A. Evans, D. J. Wales, *J. Chem. Phys.* **2003b**, 119, 9947.
- [22] M. T. Oakley, E. Oheix, A. F. A. Peacock, R. L. Johnston, *J. Phys. Chem. B* **2013**, 117, 8122.
- [23] N. Lempeis, G. C. Boulougouris, D. N. Theodorou, *J. Chem. Phys.* **2013**, 138, 12A545.
- [24] P. Sibani, J. C. Schön, *Euro. Phys. Lett.* **1993**, 22, 479.
- [25] T. Komatsuzaki, K. Hoshino, Y. Matsunaga, G. J. Rylance, R. L. Johnston, D. J. Wales, *J. Chem. Phys.* **2005**, 122, 84714.
- [26] G. J. Rylance, R. L. Johnston, Y. Matsunaga, C.-B. Li, A. Baba, T. Komatsuzaki, *Proc. Natl. Acad. Sci. USA* **2006**, 103, 18551.
- [27] L. C. Smeeton, M. T. Oakley, R. L. Johnston, PyConnect (**2014**), available at: <https://github.com/lsmeeaton/pyconnect>. Accessed on 17 March 2014.
- [28] D. J. Wales, *Mol. Phys.* **2002**, 100, 3285.
- [29] D. J. Wales, PATHSAMPLE: A Program for Refining and Analyzing Kinetic Transition Networks; **2013**, available at: <http://www-wales.ch.cam.ac.uk/PATHSAMPLE/>. Accessed on 29 July 2011.
- [30] J. D. Honeycutt, D. Thirumalai, *Proc. Natl. Acad. Sci. USA* **1990**, 87, 3526.
- [31] D. Thirumalai, Z. Guo, *Biopolymers* **1995**, 35, 137.
- [32] R. S. Berry, N. Elmaci, J. P. Rose, B. Vekhter, *Proc. Natl. Acad. Sci. USA* **1997**, 94, 9520.
- [33] S.-Y. Kim, *J. Chem. Phys.* **2010**, 133, 135102.
- [34] Y. Ueda, H. Taketomi, N. Gō, *Biopolymers* **1978**, 17, 1531.
- [35] J. Kim, T. Keyes, *J. Phys. Chem. B* **2008**, 112, 954.
- [36] J. D. Hunter, *Comput. Sci. Eng.* **2007**, 9, 90.
- [37] F. Pérez, B. E. Granger, *Comput. Sci. Eng.* **2007**, 9, 21.
- [38] M. A. Miller, DISCONNECT: A Program for Producing Disconnectivity Graphs; **2013**, available at: <http://www-wales.ch.cam.ac.uk/DISCONNECT/>. Accessed on 29 July 2011.
- [39] J. Wang, R. J. Oliveira, X. Chu, P. C. Whitford, J. Chahine, W. Han, E. Wang, J. N. Onuchic, V. B. P. Leite, *Proc. Natl. Acad. Sci. USA* **2012**, 109, 15763.
- [40] W. Kabsch, *Acta Crystallogr. A* **1978**, 34, 827.
- [41] J. Shlens, *Systems Neurobiology Laboratory, Salk Institute for Biological Studies*; **2005**, available at <http://riek-server.physiol.washington.edu/People/Fred/Classes/545/shlens-pca.pdf>. Accessed 23rd March 2012.
- [42] L. Riccardi, P. H. Nguyen, G. Stock, *J. Chem. Theory Comput.* **2012**, 8, 1471.
- [43] A. Altis, P. H. Nguyen, R. Hegger, G. Stock, *J. Chem. Phys.* **2007**, 126, 244111.
- [44] A. McLachlan, *Biopolymers* **1984**, 23, 1325.
- [45] J. B. Tenenbaum, V. de Silva, J. C. Langford, *Science* **2000**, 290, 2319.
- [46] P. Das, M. Moll, H. Stamati, L. E. Kaviraki, C. Clementi, *Proc. Natl. Acad. Sci. USA* **2006**, 103, 9885.
- [47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, 12, 2825.
- [48] M. Ceriotti, G. A. Tribello, M. Parrinello, *J. Chem. Theory Comput.* **2013**, 9, 1521.
- [49] M. A. Rohrdanz, W. Zheng, M. Maggioni, C. Clementi, *J. Chem. Phys.* **2011**, 134, 124116.
- [50] F. Cazals, F. Chazal, J. Giesen, in *Nonlinear Computational Geometry*; I. Z. Emiris, F. Sottile, T. Theobald, Eds.; Springer: New York, **2010**.
- [51] S. N. Dorogovstev, *Lectures on Complex Networks*; Oxford University Press; New York **2010**.
- [52] D. J. Wales, T. Head-Gordon, *J. Phys. Chem. B* **2012**, 116, 8394.

Received: 20 December 2013

Revised: 12 March 2014

Accepted: 14 April 2014

Published online on 28 May 2014